Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

Diego Da Silva (dluiz.silva@utoronto.ca)
University of Toronto

Feras Elsaid
University of Toronto

Amer Shalaby
University of Toronto

Research Article

Keywords: Wi-Fi, Origin Destination Matrix, Urban Rail Transit, Subway

Posted Date: August 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3287115/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

Diego Da Silva\textsuperscript{1*}, Feras Elsaid\textsuperscript{1} and Amer Shalaby\textsuperscript{1}

\textsuperscript{1}Department of Civil and Mineral Engineering, University of Toronto, 35 St. George St., Toronto, M5S 1A4, ON, Canada.

*Corresponding author(s). E-mail(s): dluiz.silva@utoronto.ca; feras.elsaid@mail.utoronto.ca; amer.shalaby@utoronto.ca;

Abstract

Transportation systems planning and management rely heavily on Origin-Destination (OD) demand matrices. However, the traditional approach of creating these matrices using household travel survey data is not only time-consuming but also expensive, making it challenging to apply them to detailed and time-sensitive analyses. Automated fare collection (AFC) systems can provide a solution to these challenges. However, many transit fare systems are entry-only with no requirement to tap at the exit station, which makes it challenging to determine the location of alighting stations and analyze spatial demand patterns. This study proposes a framework that uses Wi-Fi traces and passenger counts at AFC entry/exit gates to construct OD matrices for entry-only Urban Rail Transit (URT) systems. The City of Toronto’s subway system was used as a case study, and the framework was compared to 2016 Transportation Tomorrow Survey (TTS), which is the primary source of OD-matrix estimation in the city. The generated OD-matrices were found to be very close to the OD-matrix from the household travel survey, with a cosine similarity close to one for most subway regions. Our estimated OD-matrices offer several advantages over traditional methods. They have low matrix sparsity and fast computational time and convergence, and they exhibit strong capability of recognizing demand patterns at the station-level.

Keywords: Wi-Fi, Origin Destination Matrix, Urban Rail Transit, Subway
1 Introduction

Reliable Origin–Destination (OD) matrices are crucial for understanding the demand patterns across transit systems, and the variability in ridership between different stations. The OD-matrix is also a major input for network planning, scheduling and operational management of transit systems, including Urban Rail Transit (URT) systems. Traditional approaches for constructing OD matrices rely predominantly on data from cross-sectional household travel surveys. In the City of Toronto, the main source of OD demand data is the Transportation Tomorrow Survey (TTS), which is conducted every five years on a 5% sample of all households in the Greater Golden Horseshoe (R.A. Malatest and Associates et al., 2018).

The increasing availability of individual travel data obtained through automated fare collection (AFC) systems presents significant opportunities to help us understand individual travel behavior and spatial patterns across transit systems. This is particularly true for “closed” AFC systems, where passengers tap in and tap out their smart cards. However, in “open” AFC systems (i.e., entry-only systems), such as the system implemented by the Toronto Transit Commission (TTC), passengers are not required to tap out while leaving the system. Due to the non-mandatory “tap-out” in such systems, trip trace-ability becomes a challenge to estimate alighting locations of transit trips and analyzing individual travel behavior of passengers over time.

Furthermore, while the TTS provides extensive information at the station-level, its OD-matrix is sparse and unable to capture the same level of detail as Wi-Fi data. The use of Wi-Fi data in close proximity to passenger boarding and alighting, via network connections, yields a more precise count in comparison to the traditional TTS statistical estimation method.

Motivated by previous research efforts which utilized various data sources for OD-matrix estimation and travel behavioral analysis, this study explores the potential of using the Wi-Fi passenger traces and passenger counts at AFC entry/exit gates for constructing OD-matrices. Specifically, this study has three main objectives:

1. provide a pipeline for generating OD matrices for “entry-only” rail transit systems, using Wi-Fi and AFC gate count data;
2. demonstrate the application of the pipeline in a case study of the TTC system in Toronto and conduct a comparative analysis of the OD-matrix generated by our model with the OD-matrix obtained from the TTS.
3. assess the use of Wi-Fi data at the station-level and how it can facilitate thorough demand analysis.

The paper is structured in the following manner: first, we provide a literature review on the topic. This is followed by a detailed description of the data used and our proposed method. After this, we present the application of our method through a case study on the Toronto subway. Finally, the concluding section offers a summary of our study and highlights key insights.
2 Literature Review

2.1 Travel Data Sources in Transit Systems

OD matrices can be at the route or network levels. The route-level OD matrices represent passenger flows between origin and destination stops along the same transit route, while network-level OD matrices represent passenger flows across the entire network, including passengers making one or more transfers. The network-level matrices provide a comprehensive understanding of passenger travel patterns, which are often used to support planning transit services and policies. Traditionally, planners undertake large-scale household travel surveys every few years to collect information on how household members travel (e.g., mode choice), where they travel (e.g., origins and destinations), when they travel (e.g., time of day), who travels (e.g., socio-demographics), and why they travel (e.g., trip purpose). This information typically collected through phone interviews (e.g., Computer-Aided Telephone Interview), self-administered questionnaires (paper or online), or trip diaries.

Due to a variety of challenging factors over the past two decades, household travel surveys have experienced increasingly declining participation rates and reduced funding. To overcome these challenges, transportation researchers have been exploring alternative methods for acquiring travel behavioral data using technologies such as Global Positioning Systems (GPS) data collected via smartphones (Lawson et al., 2023). Many smartphone travel survey apps were deployed as pilot studies. In the Greater Golden Horseshoe Area (GTHA), a study used The City Logger smartphone app after recruiting 1550 participants (389 via crowd-sourcing and 1191 via email invitations). However, the results found that participants recruited via crowd-sourced methods are considerably different in terms of demographic and travel behavior characteristics than the email invitation group and the observed population in the 2016 TTS (Imani et al., 2020). Another pilot study was deployed by the City of Montreal using a smartphone app, MTL Trajet. The interface of travel behavior required participants to confirm whether they reached their desired destination considering the mode taken and trip purpose. The City of Montreal reached their goal of 10,000 respondents (Patterson et al., 2019). Other pilot studies were also reported in other cities around the world. Although these active smartphone-based travel surveys showed promising results, they still require effort and have high costs associated with recruiting participants, organizing, deploying, and collecting the results.

Advancements in information and communication technologies have enabled the use of passive data collection methods such as automatic vehicle location (AVL), automatic fare collection (AFC), automatic passenger counter (APC), monitoring cameras, and mobile phone data. Such technologies can continuously collect, store, and transmit high resolution big data, which can be leveraged to analyze passenger travel patterns at much lower cost and in greater detail than was previously possible.
AFC: These systems rely on the smart card technology as the foundation technology. It was introduced to transit systems in the late 1990s. Since then, the smart card system has been applied in many cities around the world. When tapping the card at a station or vehicle, the location and time are recorded. In closed systems, passengers need to tap their cards when boarding and alighting, which provides accurate OD information. However, in open systems, passengers tap their cards only when entering. In such systems, the destination information is not readily available.

AVL: These systems collect the location of vehicles frequently, typically using GPS devices for buses and track sensors for railways. AVL systems can now produce real-time data streams with the presence of continuous communication technology.

APC: These systems record passenger activities, such as boarding and alighting, based on door counters or weight sensors installed in vehicles.

Mobile phone data: The foundation of this data source is the Call Data Records (CDR), which can provide the location of the nearest cell towers that serve a given user. These data are always being collected as long as the device is operational. Since the information they provide is limited to time and location of the nearest tower, further estimation techniques are required to process the CDR data to infer important trip details (e.g., mode). Several recent studies have explored their potential in modeling travel patterns in different transportation systems (Calabrese et al., 2011; Iqbal et al., 2014; Alexander et al., 2015; Toole et al., 2015).

The three major data sources of AVL, AFC, and APC can provide accurate and reliable information when used together as they cover different aspects of the transit trips and service. However, there are some challenges involved as these sources require further processing since they are not intended to be used in an integrated way for frequent OD estimation.

In parallel to the ongoing efforts on integrating and harnessing AVL, AFC, and APC, there have been a considerable amount of research dedicated to human mobility modeling based on individual digital traces such as GPS (Liu et al., 2020; Munizaga and Palma, 2012; Nassir et al., 2011; Sánchez-Martínez, 2017; Barry et al., 2009; Ji et al., 2015), Bluetooth (Dunlap et al., 2016), and Wi-Fi (Danalet et al., 2014; Mishalani et al., 2016, 2011).

2.2 Wi-Fi Probe Data in Transit Systems

Many transit agencies have recently provided their customers with free Wi-Fi connectivity across their transit networks, and as a result many passengers have started connecting their smart devices (e.g. phones, tablets) to the Wi-Fi network when riding transit. As the coverage of the Wi-Fi signals in the URT network grows, the quantity and positional accuracy of the data can be sufficient to support the construction of passengers’ individual trips. The Wi-Fi probe data typically includes details on the interaction between fixed
Access Points (APs) and individual mobile devices. The path of a person carrying a Wi-Fi enabled mobile device can be tracked. Several studies explored the potential of Wi-Fi technology in occupancy estimation of indoor environments (e.g., halls, classrooms) Wang et al (2017); Chen et al (2018). (Yoon et al, 2006) presented a framework capable of building mobility models for the simulation studies of mobile systems utilizing coarse-grained Wi-Fi traces. In that study, the OD demand matrix for trips between buildings in the Dartmouth College is estimated. (Danalet et al, 2014) proposed a methodology for detecting the different activity locations visited at the EPFL campus using the Wi-Fi network traces supported by the knowledge of pedestrian network map and attractive locations.

In transit networks, (El-Tawab et al, 2017) proposed passive scanning for Wi-Fi enabled devices to estimate the waiting time of passengers at a specific bus station at James Madison University Campus. (Shlayan et al, 2016) launched a pilot project that explored the potential benefits of using passenger data collected via Bluetooth and Wi-Fi detectors for estimating time-dependent OD demand and station waiting times of passengers. Moreover, (Myrvoll et al, 2017) proposed a framework that predicts the number of passengers travelling on buses via statistical approaches based on Wi-Fi users logging data and manual passenger counts. At the scale of large networks, the Wi-Fi probe data provides an opportunity to reconstruct the spatio-temporal trajectory of a passenger through the network. (Shang et al, 2019) utilized Wi-Fi probe data, along with complete smart card data for each passenger at their origin and destination and time-dependent passenger counts in several key transfer corridors, for estimating the passenger flow states in the Beijing subway network.

Previous studies have examined using various data sources to estimate the OD matrix, but “entry-only” payment systems pose challenges in tracking passengers at stations, either in data fusion or in data pre-processing, as highlighted previously. Toronto is an example of such a system, and the city relies on collecting data from households every five years to estimate the OD matrix - a process that tend to be slow and expensive for the city. However, Wi-Fi data is readily available and can serve as a dependable proxy for demand. This is particularly true since it is closely related to the events generated at each device connection in the network. Wi-Fi data can provide a more detailed analysis of passenger flow, with a temporal resolution of minutes. This can lead to more accurate estimates of the OD-matrix, as shown in the following sections.

3 The OD Demand Matrix Estimation

Usually, the data structure representing the distribution of trips in an observed area is a two-dimensional matrix indexed by \( i = 0, 1, 2, 3 \ldots m \) origin locations in rows, and \( j = 0, 1, 2, 3 \ldots m \) destination locations in columns, as seen in Table 1. For the purpose of this discussion, we consider only aggregate trips.
irrespective of mode, purpose, cost or passengers’ characteristics. For each
element \( t_{ij} \) in the \( m \times m \) matrix, we have the number of trips originating in \( i \) and
attracted to \( j \). The \( o_i \) corresponds to the total trips at the origin calculated
by Eq. \((1a)\), whereas \( d_j \), obtained by Eq. \((1b)\), represents the total trips per
destination. The sum of all matrix elements \( t_{ij} \) is given by \( T \) and obtained
according to Eq. \((1c)\). The equations \((1a)\) and \((1b)\) also form the conditions
that the trip estimation models must satisfy to obtain a balanced OD-matrix
that can be singly or doubly constrained depending on data availability.

\[
\sum_{j=1}^{m} t_{ij} = o_i \quad \text{(1a)}
\]
\[
\sum_{i=1}^{m} t_{ij} = d_j \quad \text{(1b)}
\]
\[
\sum_{i=1}^{m} \sum_{j=1}^{m} t_{ij} = T \quad \text{(1c)}
\]

<table>
<thead>
<tr>
<th>( t_{ij} )</th>
<th>( o_i )</th>
<th>( d_j )</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_{11} )</td>
<td>( t_{12} )</td>
<td>( t_{13} )</td>
<td>( \ldots t_{1j} )</td>
</tr>
<tr>
<td>( t_{21} )</td>
<td>( t_{22} )</td>
<td>( t_{23} )</td>
<td>( \ldots t_{2j} )</td>
</tr>
<tr>
<td>( t_{31} )</td>
<td>( t_{32} )</td>
<td>( t_{33} )</td>
<td>( \ldots t_{3j} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( t_{ij} )</td>
<td>( t_{ij} )</td>
<td>( t_{ij} )</td>
<td>( \ldots t_{ij} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( t_{m1} )</td>
<td>( t_{m2} )</td>
<td>( t_{m3} )</td>
<td>( \ldots t_{mj} )</td>
</tr>
<tr>
<td>( \sum_{i=1}^{m} t_{ij} )</td>
<td>( d_1 )</td>
<td>( d_2 )</td>
<td>( d_3 )</td>
</tr>
</tbody>
</table>

In our approach, the OD-matrix estimation pipeline consists of four
sequential steps, as illustrated in Figure 1. These steps include:

- **Data collection**: the first step consists of collecting trip data recorded in
Wi-Fi database and the accumulated ground-truth values per station in the
subway entry/exit gates;
- **Data engineering**: pre-processing is a crucial phase where we perform
the necessary transformations on the raw data so that it can be consumed
effectively in later stages. In this step, we run an imputation heuristic to
reduce the number of discarded trips with missing start or end timestamp values;

- **Seed Matrix Generator**: In this step, we build a reference matrix, called the seed matrix, so that the trip distributions for each origin-destination pair are equivalent to those captured by the Wi-Fi traces;

- **Matrix Balancing**: the last step involves matrix balancing through an iterative process, applying the doubly constrained growth-factor algorithm.

In the following subsections, we describe the details of the pipeline steps.

In the following subsections, we described the details of the pipeline steps.

![Sequential data pipeline](image)

**Fig. 1** Sequential data pipeline

### 3.1 Toronto’s Subway Network Wi-Fi Data

Recently, Wi-Fi coverage has become a ubiquitous technology in URT stations in many cities worldwide, and its primary purpose is to provide public internet service access. In parallel, Wi-Fi connectivity can be harnessed to collect travelers’ location-tracking data based on the interaction between their devices and the Wi-Fi Access Points (APs) network within stations. In the City of Toronto, the communications infrastructure of the TCONNECT Wi-Fi network is provided and managed by BAI Communications, which permitted our research team to access their Wi-Fi database for this research under a strict data sharing agreement that ensures data privacy.

Passengers wishing to use the Wi-Fi network for the first time must connect and authenticate their devices with the service provider, which will store the Service Set Identifier (SSID). After this initial connection, the registered device will always connect to the provider’s APs regardless of future authentication or active internet connection. As a result, the device will associate with any AP once it comes within the coverage range.

The provider creates a record for each device’s AP association through the metadata collected, including the timestamp, the AP identification, and the device Media Access Control (MAC) address, which allows us to identify the device on the network. Due to users’ privacy protection, the persisted MAC addresses are hashed and salted at the time of collection. Therefore, along the journey a series of association and disassociation from various APs throughout the URT system is stored.
The metadata gathered for each user’s device are grouped into origin-destination segments from the first arrival of the device within the service range, besides all interactions with APs along the path, until it exceeds more than 15 minutes away from the limits of the network. After, the segments are processed into trips by ordering the timestamps and stations visited, whereby it is possible to populate the fields `origin_start_time`, `platform_start_time`, `platform_end_time`, and `destination_end_time`. In practice, not all APs can establish a connection, leaving gaps in the device traceability, mainly at the platform level between train arrivals and departures.

In this research, we are more interested in the origin and destination than in the intermediate stations. In the Wi-Fi database, the `user_id` field represents an anonymized device associated with the Wi-Fi network and the `journey_id` uniquely identifies a single passenger’s journey. In most users’ sessions collected, device association and disassociation occur at three levels: street, concourse, and platform. We considered the recorded timestamps of non-platform and platform APs to identify the true origin and destination stations.

An ideal journey obeys an ordered sequence of non-repeating timestamps following a path that starts at the `origin_start_time`, passing through the origin’s `platform_start_time`, destination’s `platform_end_time`, and finally disconnecting at the concourse or street level of the destination station `destination_end_time`, as shown in Figure 2. Due to connection loss, turned-off or airplane mode devices, APs cannot track the device, and some inconsistencies in timestamp values may happen. In this way, the repetition of timestamps values for each of the time intervals `time_1`, `time_2`, `time_3`, and
Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

(time), or even the incorrect ordering of these values, were considered inconsistencies. Thus, we categorized the Wi-Fi trajectories into four classes for further data treatment:

- **Complete**: Both origin and destination are time-consistent.
- **Complete origin and incomplete destination**: Origin is time-consistent and destination is inconsistent.
- **Complete destination and incomplete origin**: Destination is time-consistent and origin is inconsistent.
- **Incomplete origin and destination**: Both origin and destination are time-inconsistent.

Except for the class where origin and destination are complete, all other classes are candidates for an imputation step. The imputation method is useful for reducing the number of discarded records due to uncertainty at the start or end of the trip. Imputed trips benefit the OD matrix modeling results by reducing sparsity and variability.

### 3.2 Imputation of Incomplete Wi-Fi Traces

The imputation step of incomplete trips consists of a user’s inference of frequent origin and destination stations. The method followed in this paper is based on easy-to-implement logical rules. It is inspired by other methods used in detecting home and work locations from mobile phone data Calabrese et al (2011); Liu et al (2016); Çolak et al (2015). For our study, we assumed that each passenger would only have one device. While there may be exceptions to this, we deemed it reasonable to accept this assumption for the purposes of our system-wide analysis.

Classifying trips as complete offers the advantage of maintaining consistency, but it also results in the loss of numerous trips that could be identified from the passenger history. To mitigate this issue, we utilize an imputation method that corrects inconsistencies by utilizing some of the discarded pairs where either the origin or destination was missing in the Wi-Fi trip. This preprocessing step is expected to increase the signal in certain origin-destination pairs and potentially even add counts to pairs that were previously reported with zero quantity.

We implemented three models to determine if there was an improvement in the quality of the OD-matrix estimation. These models only vary based on the assumptions made about the trips analyzed. Figure 3 illustrates the data flow diagram we used to generate the three models ($M_1$, $M_2$, and $M_3$). $M_1$ was considered the most restrictive, as the time sequence must strictly follow the temporal flow of events between the entrance and exit stations. The result of the query from Wi-Fi database for the category of trips called complete corresponded to the input of $M_1$. On the other hand, $M_3$ input was the least restrictive and included all trips during the given period regardless of whether the events were completed or not.
Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

Also, we looked for trips that included inconsistencies either in origin or in the destination, which were candidates for the imputation method. The inconsistency identified in the database occurred when the timestamp value on the platform was identical to the value recorded at the concourse or street level, which is physically impossible.

The imputation method consists of two logical verification steps. The first step examines whether, in the last 30 days, the user had at least one origin-destination pair equal to the trip classified as inconsistent. If the trip was identified in the historical list of origin-destination pairs, the respective station is imputed. If the pair was not identified, we generate a disaggregated historical list; that is, for each user, we have a list of origins and another of destinations. The second step was to check whether the origin or destination encompasses the known stations within the disaggregated list. After going through the imputation method, the concatenation operator aggregated the *imputed* and *complete* trips to be the input of $M_2$. Thus, trips that failed to verify the user's origins and destination history were added to the discarded pool.
3.3 Seed Matrix Generator

After pre-processing the Wi-Fi data and imputing missing trip details, the next step is the generation of a reference matrix, also called a seed matrix. The central role of the seed matrix was to provide the trip distribution proportions in the same layout as the OD-matrix presented in Table 1, but based only on Wi-Fi data that we denote with the prime symbol (').

We denote by $S$ the seed matrix with dimension $m \times m$ that contains the quotient $q_{ij}$ given by Eq. 2:

$$q_{ij} = \frac{t'_{ij}}{o'_i}$$

where the element $t'_{ij}$ are trips and $o'_i$ the sum of trips in the row from the Wi-Fi database.

The $S$ matrix form presented in Table 2 is one of the inputs for the balancing function we will describe in the following sub-Section. The advantage of using the $S$ matrix is to preserve trip proportions during the balancing function iterations. As more users connect to the Wi-Fi network, the $S$ matrix naturally produces total entry and exit counts that approach the ground-truth count data at the AFC gates.

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destinations</th>
<th></th>
<th></th>
<th>...</th>
<th>...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$q_{11}$</td>
<td>$q_{12}$</td>
<td>$q_{13}$</td>
<td>...</td>
<td>$q_{1j}$</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>$q_{21}$</td>
<td>$q_{22}$</td>
<td>$q_{23}$</td>
<td>...</td>
<td>$q_{2j}$</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>$q_{31}$</td>
<td>$q_{32}$</td>
<td>$q_{33}$</td>
<td>...</td>
<td>$q_{3j}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$i$</td>
<td>$q_{i1}$</td>
<td>$q_{i2}$</td>
<td>$q_{i3}$</td>
<td>...</td>
<td>$q_{ij}$</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m$</td>
<td>$q_{m1}$</td>
<td>$q_{m2}$</td>
<td>$q_{m3}$</td>
<td>...</td>
<td>$q_{mj}$</td>
<td>...</td>
</tr>
</tbody>
</table>

3.4 Growth-Factor Doubly Constrained Model

The balancing function is the last step in generating the OD-matrix. Several methods exist in the literature for balancing the trip matrix, such as the Singly and Doubly Constrained Growth-Factor methods Furness (1965), Gravity models Bouchard and Pyers (1965); Hansen (1962), Intervening Opportunities Model Stouffer (1940), Entropy-Maximization Murchland (1966), and Partial
Matrix Techniques Kirby (1979), each with its pros and cons. We selected the Doubly Constrained Growth-Factor as the balancing method due to the volume of data we had available for analysis, allowing us to use a method with a more straightforward implementation, few assumptions, interpretable, and rapid convergence. The limitations of the doubly constrained method noted by (de Dios Ortúzar and Willumsen, 2011) are overcome by the volume and speed at which data are captured on the Wi-Fi network, which allows us to map the amount and pattern of trips of the entire system in different temporal resolutions and without additional high costs. In addition, over time, we can map demand correction factors indicating fluctuations between peaks, seasonality, and service disruptions. This would allow us to perform analysis and planning in both the short and long terms. As the data comes from the users’ devices connected to the Wi-Fi network, the granularity and quality of the data reduce the dependencies of behavioral inferences used in gravitational models. Another benefit of the method is aggregating data at different system levels, e.g., regional, station, routes, and lines, without changing model assumptions. This makes the model extremely flexible and without the need for recurring calibrations. We assume that Wi-Fi and gate count databases record volumes for the entire system, i.e., large samples. In the case of small samples and several missing data, it is challenging to identify emerging origin-destination patterns from the data. Models with assumptions about trip patterns can handle better than the Growth-Factor method, which relies on a more accurate seed matrix.

The balancing function has two arguments: the seed matrix \( S \) and the gate count data. The gate count data encompasses the counting values of valid entries and exits denoted by the column-vector \( \mathbf{o} = (o_1, o_2, o_3 \ldots x_m) \) and row-vector \( \mathbf{d} = (d_1, d_2, d_3 \ldots d_m) \), respectively. The elements \( o_i \) and \( d_j \) correspond to the ground-truth data aggregated for each origin and destination and are the values we want to target in the iterative process of matrix balancing.

The Growth-Factor method has a unique constraint: the sum of the trips at the origins must be equal to the destination, as described in Eq.1a, Eq.1b, and Eq.1c. However, in practical applications, those sums are rarely equal. In this case, we applied a verification step, and if necessary, a correction on the vectors as shown in Eq.3 below:

\[
\langle r_{\mathbf{o}}, r_{\mathbf{d}} \rangle = \begin{cases} 
\langle \mathbf{o}, \sum_{i=1}^{m} d_j \sum_{i=1}^{m} o_i \rangle & \text{if } \sum_{i=1}^{m} o_i > \sum_{j=1}^{m} d_j \\
\langle \sum_{i=1}^{m} o_i \sum_{j=1}^{m} d_j, \mathbf{d} \rangle & \text{if } \sum_{j=1}^{m} d_j > \sum_{i=1}^{m} o_i \\
\langle \mathbf{o}, \mathbf{d} \rangle & \text{otherwise}
\end{cases}
\]

where \( \langle r_{\mathbf{o}}, r_{\mathbf{d}} \rangle \) is a 2-tuple that contains the vectors after verification and denoted by \( r_{\mathbf{o}} \) and \( r_{\mathbf{d}} \). Eq.3 shows that if the summation of \( o_i \) is greater than \( d_j \), then correct \( \mathbf{d} \) by redistributing the summation of \( o_i \), but maintaining the
proportion of trips on original \( d \), and vice versa. Otherwise, we maintain the vectors the same.

Before starting the iterative step of the matrix balancing function, we need to convert the matrix \( S \) into an unbalanced trip matrix that we denoted by \( U^{(0)} \) of the exact \( m \times m \) dimensions. We called the \( U^{(0)} \) the matrix as the initial state and it is the result of the element-wise product operation shown in Eq. 4. The following notation assumes the vector \( r_o \) choice, although it is not mandatory. The \( t_{ij} \) elements of the \( U^{(0)} \) matrix are the values updated by iterating the balancing function. Therefore, the correction performed on each origin (rows) and destination (columns) connects matrix \( U^{(0)} \) from its initial to the balanced final state.

\[
U^{(0)} = S \otimes r_o
\]  

(4)

For each iteration \( k = 0, 1, 2, 3, \ldots n \) we compute a sequence of steps that alternate between rows and columns of the matrix as illustrated in Figure 4. All even iterations are corrections on the columns, while the odd iterations are for the rows. As we started our matrix with \( r_o \), corrections must be made to the columns in the first iteration \( (k = 0) \). First, we calculate the totals that correspond to the sums of the rows or the columns, observing the \( k \) index of the iteration, according to Eq. 5a and Eq. 5b.

\[
o_i^{(k)} = \sum_{j=1}^{m} t_{ij}^{(k)} \quad \text{for} \quad i = 1, 2, \ldots, m, \quad k = 1, 3, 5, \ldots, n - 1 \tag{5a}
\]

\[
d_j^{(k)} = \sum_{i=1}^{m} t_{ij}^{(k)} \quad \text{for} \quad i = 1, 2, \ldots, m, \quad k = 0, 2, 4, \ldots, n \tag{5b}
\]

The central operation on Growth-Factor model consists of calculating correction ratios, \( \phi \). Since our model is doubly constrained, We need two correction factors, one exclusive for the rows and another for columns of matrix. We obtain the correction factor by applying Eq. 6a and Eq. 6b as follows:

\[
\phi_o^{(k)} = (r_o^\top \otimes o^{(k)})^\top \quad \text{for} \quad k = 1, 3, 5, \ldots, n - 1 \tag{6a}
\]

\[
\phi_d^{(k)} = r_d \otimes d^{(k)} \quad \text{for} \quad k = 0, 2, 4, \ldots, n \tag{6b}
\]
where the column-vector $\phi^{(k)}_o = (\phi^{(k)}_{o_1}, \phi^{(k)}_{o_2}, \ldots, \phi^{(k)}_{o_m})$ are the correction factors for the origin, and the row-vector $\phi^{(k)}_d = (\phi^{(k)}_{d_1}, \phi^{(k)}_{d_2}, \ldots, \phi^{(k)}_{d_m})$ for the destination in the iteration $k$.

The last step of the iteration is to update the matrix $U^{(k+1)}$ through the element-wise product operation, according to Eq. 7.

$$U^{(k+1)} = \begin{cases} U^{(k)} \otimes \phi^{(k)}_o & \text{if } k \in \{1, 3, 5, \ldots, n-1\} \\ (U^{(k)^T} \otimes \phi^{(k)^T}_d)^T & \text{if } k \in \{0, 2, 4, 6, \ldots, n\} \end{cases}$$

The iteration is interrupted when the exit criteria are satisfied. We have configured two exit criteria listed by priority:

- Satisfy the model conditions described on equations (1a), (1b), and (1c);
- Reach the maximum number of iterations.

Ideally, the model should converge to the theoretical values of the conditions; however, in practice, we are subject to computational rounding errors. Thus, we can satisfy the first exit criterion of the model by setting an error threshold that, when reached, the iteration is interrupted.

$$MAPE^{(k)}_o = \frac{1}{m} \sum_{i=1}^{m} \left| \frac{r_{o_i} - o^{(k)}_i}{r_o} \right| \times 100 \quad \text{for } k = 1, 3, 5, \ldots, n-1 \quad (8a)$$

$$MAPE^{(k)}_d = \frac{1}{m} \sum_{j=1}^{m} \left| \frac{r_{d_j} - d^{(k)}_j}{r_d} \right| \times 100 \quad \text{for } k = 0, 2, 4, \ldots, n \quad (8b)$$
At each iteration \( k \) we calculated the MAPE, Eq. (8a) and (8b), and compared if both errors, row and column, were below 0.5\%, otherwise the exit criterion is to execute until the last iteration term \( n \).

4 Case Study: The Subway Network of the City of Toronto

In this section, we present a case study of the subway system in the City of Toronto subway. The analysis is divided into three sub-Sections: Data Collection, Modeling, and Results.

4.1 Data Collection

Our data sample comes from two primary sources: data collected from devices connected to the Wi-Fi network and the cumulative number of trips per origin and destination at the TTC subway stations. The sample spans the period between 09/24/2021 to 09/30/2021, except the weekend, during the morning peak (6 a.m. to 9 a.m.).

The TTC subway system in the year 2021 had 75 stations distributed in four lines:

- **Line 1**: Yonge-University (Yellow Line)
- **Line 2**: Bloor-Danforth (Green Line)
- **Line 3**: Scarborough (Blue Line)
- **Line 4**: Sheppard (Purple Line)

The AFC data set, obtained from the Toronto Transit Commission, provided the passenger counts at the gates of each station on the same dates of the Wi-Fi data and grouped by hours. However, the raw data are unlinked origin and destination counts per station, which means we cannot derive OD pairs directly. If we inspect Table 1 where we described the OD-matrix form, we can infer that the sum of the absolute numbers of entries and the sum of absolute numbers of exits per station at the TTC data set would be equivalent to the vectors \( o_i \) and \( d_j \), respectively.

The cumulative number of trips extracted from the Wi-Fi database for the morning peak throughout the five days is 134,973. In contrast, the gate count data recorded 389,461 valid exits, corresponding to an approximate ratio of one trip recorded by Wi-Fi for every three gate counts, (see Figure 5). The result of the imputation method added 21,877 trips which brings the total number of trip observations to approximately 60\% of the data captured by the Wi-Fi network.

4.1.1 Transportation Tomorrow Survey (TTS)

Since 1986, the TTS has surveyed five percent of households in the Greater Golden Horseshoe (GGH) area within southern Ontario, Canada every five years R.A. Malatest and Associates et al (2018). The TTS is a household travel
survey that collects information such as the mode of travel and trip purpose, along with the demographics of household members. The most recent travel survey was conducted in 2016. The survey data were assigned weights such that the expanded survey numbers were representative of the Census counts of private dwellings occupied by usual residents (with further adjustments to better represent all households by dwelling type, household size, and householder gender and age groups). The 2016 TTS surveyed 162,708 households, 395,885 persons, and 798,093 trips. One of the key outputs of 2016 TTS is the OD-matrix for the TTC subway system.

To validate our models, we obtained the AM peak subway OD-matrix from the 2016 TTS. The AM peak was defined from 5:30 a.m. to 8:30 a.m., which represents a 30-minute shift from the AM peak definition used for the Wi-Fi and gate count data. This is because the TTS reports trip start times from respondents’ locations of the trip origins not the transit stations. We have included an arbitrary 30-minute buffer for traveling from the trip origin to the subway station. This value is not expected to introduce major errors or bias, as the trips start times reported in the TTS are not exact because respondents tend to round those times. Also, the shift should not affect the peak-within-the-peak period.

4.2 Modeling

We developed three models, one with complete trips only which we denoted $M_1$ for complete, $M_2$ for complete and imputed trips combined, and $M_3$ without any rigid criteria for validating the temporal ordering of event timestamps. Thus, we can assess the quality of trip imputation step through a comparative analysis, the similarity to the TTS OD-matrix, and the ability to estimate both at the regional-level and at the station-level the volume of trips in the selected period.
For adequate comparison with the TTS-based OD-matrix, we present the results on an aggregate form. This is because the 5% sample of the TTS produces a relatively sparse matrix which is not suitable for comparisons on a station-by-station basis. Therefore, we divided the subway network into 10 regions as follows:

- **Region 1 (R1)**: Spadina (transfer station), Museum, Queen’s Park, St. Patrick, Osgoode, St. Andrew, Union, King, Queen, Dundas, College, Wellesley, Bloor-Yonge (transfer station), Bay, St. George (transfer station), and Bathurst;
- **Region 2 (R2)**: Christie to Kipling;
- **Region 3 (R3)**: Sherbourne to Kennedy;
- **Region 4 (R4)**: Dupont to Eglinton West;
- **Region 5 (R5)**: Rosedale to Eglinton;
- **Region 6 (R6)**: Glencairn to Wilson;
- **Region 7 (R7)**: Lawrence to Finch;
- **Region 8 (R8)**: Bayview to Don Mills;
- **Region 9 (R9)**: Lawrence East to McCowan; and
- **Region 10 (R10)**: Sheppard West to Vaughan Metropolitan Centre;

![Regional classification for the Toronto subway system](image)

**Fig. 6** Regional classification for the Toronto subway system

In Figure 6, we illustrate the map of the Toronto subway system and the division into 10 regions. R1, highlighted on the map, includes all subway stations in the downtown area, which is the leading trip attraction zone in the morning peak and also the zone with the highest number of transfers and integrations with other modes in the network. An important caveat is that in 2016
When the TTS was carried out, R10 only had the Sheppard West station, formerly Downsview. Therefore, in our analyses, the models generated from the Wi-Fi data aggregated the six northernmost new stations into R10, and the TTS OD-matrix contained only the Sheppard West station in R10.

### 4.3 Results and Discussion

#### Regional-Level Analysis

For each of the models, we obtained the respective balanced OD-matrix. We can contrast the trip distribution represented as the proportion of trips between regional origin and destination pairs across models $M_1$, $M_2$, $M_3$ and TTS, as illustrated in Figure 7. Visually we can see that the heatmap of the models obtained proportionally similar trip distributions. As expected, the concentration of travel destinations in the morning peak is mostly on R1. However, the concentration of trips to R1 in the TTS is higher than in the models’ result, as in the pair R7-R1, which is approximately 2.5 times that of the balanced matrices. In the modelled matrices (e.g. M2), Regions R1, R2, and R3 produced and attracted approximately 50% of all system trips, while the corresponding value in the TTS matrix is 46%. One way to measure the OD-matrix’s quality is by its sparsity, which means a matrix with mostly zero elements is considered sparse. We can observe a smaller matrix sparsity in the models $M_1$ (22%), $M_2$ (24%), and $M_3$ (25%) to the TTS (31%) at the region-level. When analyzing Wi-Fi data, the sparsity of certain regions can indicate difficulties in capturing signals. This may be caused by low demand or limitations in the network infrastructure. However, for TTS, sparsity is caused by an estimation error. This will be more evident when we discuss it at the station-level.

While visual inspection can facilitate a qualitative comparison between the models and TTS, more is needed to assess the quality of the OD-matrix estimation models. For this, we performed a linear regression analysis, and our goal was based on the assumption that our method should be at least equivalent to the TTS-based matrix. Furthermore, we could quantify the information gained across models with the regression performance statistics.

Since the OD-matrix data are of different scales, we normalized the trip distributions as shown in Table 2. Next, we paired each estimated $t_{ij}$ element in the model’s OD-matrix with its corresponding element in the TTS matrix. The result of the linear regression is summarized in Tables 3 and 4. We can observe that the three models compared reasonably well to the TTS, with a high coefficient of determination, above 70%, particularly $M_1$ which has an Adjusted $- R^2$ of 78.7% and a lower RMSE. Numerically, this shows that the more restrictive model produces an OD matrix with similar structure and variability to the TTS-based matrix.

In Table 4, the summary of linear regression statistics for both models shows that both estimators are statistically significant. The models’ coefficients are close to 1, demonstrating that the regression line’s slope has an angle close to 45°, the blue dashed line at the top of the Figure 8. The $\arctangent$ of the
Fig. 7 Proportion of trips (%) per Region

Table 3 Linear Regression to compare OD-matrix estimation models

<table>
<thead>
<tr>
<th>Model</th>
<th>( R )</th>
<th>( R^2 )</th>
<th>( \text{Adj. } R^2 )</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>0.888</td>
<td>0.789</td>
<td>0.787</td>
<td>0.0892</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>0.882</td>
<td>0.779</td>
<td>0.776</td>
<td>0.0914</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>0.861</td>
<td>0.741</td>
<td>0.739</td>
<td>0.0989</td>
</tr>
</tbody>
</table>

trip (%) coefficients of \( M_1 \), \( M_2 \), and \( M_3 \) are approximately 49°, 49°, and 47°, respectively.

At the bottom of the Figure 8, the charts show the MAPE reduction rate calculated at each balancing iteration of the OD-matrix. As we use the Doubly Constrained Growth-Factor model, the first iteration starts in the column, that is, in the rows of the matrix. We identified it as index 0, the multiplication step of the gate count data, and the seed matrix. Afterward, the algorithm adjusts the values in the columns and alternately performs the balancing by calculating the growth factor for each row and column of the matrix. The models converged quickly, and the error stabilized in less than ten iterations. The
Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

Table 4  Linear Regression to compare OD-matrix estimation models

<table>
<thead>
<tr>
<th>Model</th>
<th>Coeff.</th>
<th>SE</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>t-test</th>
<th>p-value</th>
<th>Std. Estimate</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>Intercept</td>
<td>-0.0160</td>
<td>0.0109</td>
<td>-0.0376</td>
<td>0.0050</td>
<td>-1.48</td>
<td>0.143</td>
<td>0.888</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>Trips (%)</td>
<td>1.1605</td>
<td>0.0606</td>
<td>1.0403</td>
<td>1.2806</td>
<td>19.16</td>
<td>&lt; .001</td>
<td>0.888</td>
<td>0.796</td>
</tr>
<tr>
<td>$M_2$</td>
<td>Intercept</td>
<td>-0.0151</td>
<td>0.0111</td>
<td>-0.0371</td>
<td>0.0069</td>
<td>-1.36</td>
<td>0.177</td>
<td>0.882</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>Trips (%)</td>
<td>1.1511</td>
<td>0.0620</td>
<td>1.0281</td>
<td>1.2741</td>
<td>18.57</td>
<td>&lt; .001</td>
<td>0.882</td>
<td>0.788</td>
</tr>
<tr>
<td>$M_3$</td>
<td>Intercept</td>
<td>-0.0102</td>
<td>0.0120</td>
<td>-0.0339</td>
<td>0.0136</td>
<td>-0.851</td>
<td>0.397</td>
<td>0.861</td>
<td>0.759</td>
</tr>
<tr>
<td></td>
<td>Trips (%)</td>
<td>1.1018</td>
<td>0.0658</td>
<td>0.9713</td>
<td>1.2323</td>
<td>16.75</td>
<td>&lt; .001</td>
<td>0.861</td>
<td>0.759</td>
</tr>
</tbody>
</table>

average execution time on an Operational System Ubuntu 22.04.2 LTS 64-bit, processor Intel® Core™ i7-8550U CPU @ 1.80GHz × 8, and memory 16Gb was 0.44±0.12 seconds, which indicates its feasibility for practical implementation. In the last iteration, the $M_1$ MAPE for origin and destination was $240 \times 10^{-18}$ and $1.5 \times 10^{-3}$, respectively. For the $M_2$, the computed error was $230 \times 10^{-18}$ and $1.5 \times 10^{-3}$, while $M_3$ was $240 \times 10^{-18}$ and $1.5 \times 10^{-3}$.

The second way to analyze the quality of the models is to validate the similarity of column and row-vector space. Unlike $R^2$, similarity validation tells us how geometrically similar the matrices are without assuming any assumptions about the distribution of the values. We employed the cosine similarity to compare the vectors of each model OD-matrix output to the TTS. For contextualisation, let $a$ and $b$ be two vectors we are interested in comparing, the cosine similarity is obtained through the Eq. 9:

\[ \text{cosine similarity} = \frac{a \cdot b}{\|a\| \|b\|} \]
Constructing Origin-Destination Matrix using Wi-Fi and AFC Data

\[
similarity(a, b) = \frac{a \cdot b}{ab}
\]  

(9)

where \( a \) and \( b \) are Euclidean norms which normalize the dot product between the two vectors. The closer this measure is to 1, the more similar the vectors are. If the vectors are orthogonal to each other, they are considered unrelated.

In Table 5 we present the results of the cosine similarity for the row (i.e., origin) vectors and column (i.e., destination) vectors for each model. Except for R10, the other regions obtained values above 0.71 (45°), which shows a geometric similarity between our models’ results and the TTS. R10 was expected to have a lower similarity since, as mentioned previously, the northernmost stations on the city’s west side had not opened in 2016. However, comparatively, \( \mathcal{M}_1 \) had the best similarity scores among the models, mainly in the regions at the system’s edge.

### Table 5 OD Matrices cosine similarity

<table>
<thead>
<tr>
<th>Region</th>
<th>Model ( \mathcal{M}_1 )</th>
<th>Model ( \mathcal{M}_2 )</th>
<th>Model ( \mathcal{M}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Origin</td>
<td>Destination</td>
<td>Origin</td>
</tr>
<tr>
<td>R1</td>
<td>0.996</td>
<td>0.898</td>
<td>0.996</td>
</tr>
<tr>
<td>R2</td>
<td>0.937</td>
<td>0.966</td>
<td>0.940</td>
</tr>
<tr>
<td>R3</td>
<td>0.971</td>
<td>0.997</td>
<td>0.973</td>
</tr>
<tr>
<td>R4</td>
<td>0.987</td>
<td>0.946</td>
<td>0.987</td>
</tr>
<tr>
<td>R5</td>
<td>0.936</td>
<td>0.962</td>
<td>0.941</td>
</tr>
<tr>
<td>R6</td>
<td>0.993</td>
<td>0.975</td>
<td>0.995</td>
</tr>
<tr>
<td>R7</td>
<td>0.939</td>
<td>0.981</td>
<td>0.943</td>
</tr>
<tr>
<td>R8</td>
<td>0.943</td>
<td>0.988</td>
<td>0.935</td>
</tr>
<tr>
<td>R9</td>
<td>0.876</td>
<td>0.727</td>
<td>0.853</td>
</tr>
<tr>
<td>R10</td>
<td>0.635</td>
<td>0.423</td>
<td>0.609</td>
</tr>
</tbody>
</table>

At the system-wide level, the models closely match the OD-matrix generated by the TTS. Generally, the models displayed similar travel patterns in attractive regions but exhibited greater sensitivity in pairs of regions during the morning peak counterflow. This is because the Wi-Fi network collects data at the exact event locations, creating a Seed Matrix that better represents the proportion of trips without requiring estimation, as in the TTS. Comparing the models, it appears that \( \mathcal{M}_1 \) has a slight advantage, even though the performance is quite similar between all models. The imputation method did not substantially improve the comparison between the models due to data grouping at the regional level that smooths out minor differences in OD pairs.

The following analysis is to verify the temporal stability of our results. For this evaluation, we wanted to assess if the daily ratio value, has a high variability. If the values are stable, then it is possible to extrapolate our results over time using only one correction factor, which we call **Expansion Factor**.
The Expansion Factor can be global when using the cumulative totals at the system-level or at the regional-level, as shown in Eq. 10:

\[ \text{Expansion Factor} = \frac{T}{T'} \]  

(10)

where \( T \) and \( T' \) is obtained according to Eq. 1c and represent the total sum of trips for the balanced and Wi-Fi matrices, respectively.

In Table 6, we present the values of the global daily expansion factors and the descriptive statistics of the time series. We can see that the values remained stable during the five-day analysis for all models. In addition, we can highlight that \( M_3 \) obtained approximately half the values of \( M_1 \), in addition to a Coefficient of Variation (\( CV(\%) \)) of 0.66\%, which shows that the dispersion around the mean is relatively low. The result indicated that the models are stable for at least a short time horizon.

**Table 6** Comparison between global expansion factors per day for \( M_1, M_2 \) and \( M_3 \)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sep-24</th>
<th>Sep-27</th>
<th>Sep-28</th>
<th>Sep-29</th>
<th>Sep-30</th>
<th>( \bar{\tau} )</th>
<th>( \pm s )</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>6.56</td>
<td>6.88</td>
<td>6.87</td>
<td>6.75</td>
<td>6.93</td>
<td>6.80</td>
<td>0.13</td>
<td>1.93</td>
</tr>
<tr>
<td>( M_2 )</td>
<td>4.77</td>
<td>4.89</td>
<td>4.96</td>
<td>4.87</td>
<td>5.01</td>
<td>4.90</td>
<td>0.08</td>
<td>1.65</td>
</tr>
<tr>
<td>( M_3 )</td>
<td>3.29</td>
<td>3.31</td>
<td>3.35</td>
<td>3.31</td>
<td>3.29</td>
<td>3.31</td>
<td>0.02</td>
<td>0.66</td>
</tr>
</tbody>
</table>

For the regional-level OD Expansion Factors calculation, we aggregated the values from the five days of analysis and obtained the \( CV(\%) \) to visualize them, see Figure 9. We can see that both models, in most cells, presented a relatively low CV\%, highlighting some origin-destination pairs with higher dispersion, mainly in the morning peak counterflow. In addition, we can see that dispersions in \( M_1 \) are numerically higher than in \( M_2 \) and \( M_3 \). This can be either due to the daily variation in demand in some regions that were not captured by the Wi-Fi network, the most restrictive trip assumption on \( M_1 \), adjustments in the trip balancing, or changing in demand pattern behavior due to pandemic.

**Station-Level Analysis**

We also assessed the usability of Wi-Fi data as a proxy for the OD-matrix also at the station level, given that the sparsity of the TTS matrix is 51.5\%, which means that more than half of the origin-destination pairs have a null value. In our models, the balanced matrix revealed a sparsity at the station level of 27.6\%, 23.9\%, and 17.7\% for \( M_1, M_2 \), and \( M_3 \), respectively.
Examining the sparsity of the models at the station level allows us to understand the variations in the models more thoroughly. We selected 16 destination stations, the minimum index that approaches 50% of the matrix sparsity for each model, as described in Table 7. The sparsity is mostly concentrated in R8, R9, and R10, regions in the system’s counterflow in the morning peak.

In models $M_1$ and $M_2$, Dundas (R1) appears as the sixth most destination that contributed to the sparsity of the matrices; however, from observation and experience, we know that it is one of the busiest stations in the morning peak, with a concentration of commercial buildings, leisure options and a critical intersection with other modes of transport in the city. As the station does not appear in the $M_3$ list, the strictest criterion to eliminate trip inconsistencies reduced to zero some pairs heading to Dundas Station. Thus, the station-level analysis also serves as a reference to diagnose possible issues in capturing events on the Wi-Fi network.

### Table 7 OD-Matrix Sparsity per Destination

<table>
<thead>
<tr>
<th>Destination (Region)</th>
<th>% Null</th>
<th>Destination (Region)</th>
<th>% Null</th>
<th>Destination (Region)</th>
<th>% Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellesmere (R9)</td>
<td>4.3</td>
<td>Ellesmere (R9)</td>
<td>4.7</td>
<td>Ellesmere (R9)</td>
<td>6.1</td>
</tr>
<tr>
<td>Midland (R9)</td>
<td>3.9</td>
<td>Midland (R9)</td>
<td>4.4</td>
<td>McCowan (R9)</td>
<td>5.5</td>
</tr>
<tr>
<td>McCowan (R9)</td>
<td>3.8</td>
<td>McCowan (R9)</td>
<td>4.2</td>
<td>Midland (R9)</td>
<td>5.4</td>
</tr>
<tr>
<td>Bessarion (R8)</td>
<td>3.7</td>
<td>Highway 407 (R10)</td>
<td>4.0</td>
<td>Highway 407 (R10)</td>
<td>5.0</td>
</tr>
<tr>
<td>Highway 407 (R10)</td>
<td>3.6</td>
<td>Dundas (R1)</td>
<td>3.4</td>
<td>Dundas (R1)</td>
<td>3.5</td>
</tr>
<tr>
<td>Dundas (R1)</td>
<td>3.3</td>
<td>Bayview (R8)</td>
<td>3.1</td>
<td>Bayview (R8)</td>
<td>3.3</td>
</tr>
<tr>
<td>Bayview (R8)</td>
<td>3.0</td>
<td>Leslie (R8)</td>
<td>2.9</td>
<td>Leslie (R8)</td>
<td>3.5</td>
</tr>
<tr>
<td>Leslie (R8)</td>
<td>2.8</td>
<td>Glenccairn (R6)</td>
<td>2.8</td>
<td>Glenccairn (R6)</td>
<td>2.7</td>
</tr>
<tr>
<td>Glenccairn (R6)</td>
<td>2.8</td>
<td>Downsview Park (R10)</td>
<td>2.6</td>
<td>Downsview Park (R10)</td>
<td>2.6</td>
</tr>
<tr>
<td>Downsview Park (R10)</td>
<td>2.6</td>
<td>Old Mill (R2)</td>
<td>2.6</td>
<td>Old Mill (R2)</td>
<td>2.2</td>
</tr>
<tr>
<td>Old Mill (R2)</td>
<td>2.6</td>
<td>Rosedale (R5)</td>
<td>2.4</td>
<td>Rosedale (R5)</td>
<td>2.2</td>
</tr>
<tr>
<td>Rosedale (R5)</td>
<td>2.4</td>
<td>Chester (R3)</td>
<td>2.3</td>
<td>Chester (R3)</td>
<td>2.0</td>
</tr>
<tr>
<td>Chester (R3)</td>
<td>2.3</td>
<td>Scarborough (R9)</td>
<td>2.1</td>
<td>Scarborough (R9)</td>
<td>2.0</td>
</tr>
<tr>
<td>Scarborough (R9)</td>
<td>2.1</td>
<td>Eglington West (R4)</td>
<td>1.9</td>
<td>Eglington West (R4)</td>
<td>2.0</td>
</tr>
<tr>
<td>Eglington West (R4)</td>
<td>1.9</td>
<td>Finch West (R10)</td>
<td>1.8</td>
<td>Finch West (R10)</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Cumulative Sum (%) 46.7 48.1 53.4

Fig. 9 CV(%) at the regional-level
Also, we can apply the cosine of similarity to identify stations with similar trip patterns. Figure 10 illustrates the origin stations in descending order of similarity to Christie Station. Keele, Landsdowne, and Ossington stations are the top three in this ordering above the similarity of 0.8, which means that the morning peak demands of these stations are relatively similar. Since we are measuring the cosine of the angle of the origin vectors, the result of the origin-destination pair is undirected. Interestingly, the cosine of similarity identified the spatial patterns of the subway system. The bar colors represent the lines, Bloor-Dufferin green (Line 2 in the subway system), Yonge-University yellow (Line 1), Scarborough blue (Line 3), and Sheppard purple (Line 4). We have a separation between the similarities of Christie station and the rest of the system, with the stations on the westernmost side tending to be more similar to it than the similarity of the line to which it belongs to the others. Identifying symmetries through demand similarity is an alternative to apply more flexible models for system state prediction.

![Fig. 10 Similarity to Christie station. The colors represent the system lines, and the gray bars the transfer stations](image)

**5 Conclusion**

We used Wi-Fi and gate count data to develop a pipeline for generating OD matrices for entry-only (i.e., “tap-on” only fare systems) URT systems. OD-matrix is a fundamental element in public transport planning and management. The developed matrices used highly available, reliable, and more granular data, and produced OD patterns similar to the those displayed in the TTS-based OD-matrix.
We developed three models and compared them to assess their performance and validate their quality using the results reported by TTS. The first model used trips identified as complete, mainly trips with both trip ends reliably recorded in the database. The second model used a seed matrix formed by trips imputed by a heuristic method in addition to complete trips to minimize the number of trips discarded by the first model. The third model used data from a query extraction without a rigid criterion of logical temporal ordering of events. Our models showed they could be a reasonable proxy for the TTS in producing accurate OD matrices. Furthermore, the results show the possibility of generating OD demand matrices for the TTC subway system in at fine time scales (e.g. daily), if the gate count data are promptly available. This execution time fundamentally alters how public transport analysis could benefit from the shorter horizon to reduce planning uncertainty.

In the data pre-processing step, we applied an imputation method to address concerns regarding the quality of Wi-Fi data. We established three criteria, ranging from strict to relaxed assumptions, and observed that the model’s performance at the regional level was consistent across all criteria. Upon analyzing station-level data, we found that including approximately 40% of trips through the imputation method reduced matrix sparsity by roughly four percentage points in $M_2$. Therefore, for the Toronto subway Wi-Fi network, we concluded that trip imputation had only marginal gains and was thus unnecessary as the data quality allowed for analysis of the entire system without significant difference across models. It is important to mention that we strongly recommend executing the imputation step and comparing the resultant models in other subway systems with public Wi-Fi access to ensure data quality and enable consistent analysis.

Moreover, we showed that the matrices generated by our models were geometrically similar to the TTS by calculating the cosine similarity on row and column vector spaces and showing mostly values were close to one. The coefficients of determination for the linear regressions were above 70% and the matrix balancing convergence was less than 10 iterations with an average running time of only 0.44 seconds. The analysis of the models indicated a slight edge for $M_1$. However, it is safe to say that both models achieved comparable results, and we do not have compelling evidence to favor one over the other. This outcome is primarily due to the volume, approximately 35% of the gate count, and quality of the data gathered.

In addition, the Expansion Factors used to assess model stability over time had low CV values ($< 2\%$), which means the possibility of extrapolating our results, at least for the five-day window considered in our analysis. A possible application would be to use the series of Expansion Factors to generate an OD-matrix adjusted to demand fluctuation, large-scale events, transportation disruptions, or more accurate simulations. The Expansion Factors at the regional-level had the lowest CV ($\leq 15\%$) in downtown area (Region 1). The higher CV values can be explained by changes in the travel pattern of passengers during the pandemic, disruptions in services during the analysis period,
lack of data captured by the Wi-Fi network, and trip misclassification by the imputation heuristic. The reasons are non-exhaustive and require more granular analysis.

We demonstrated how the matrix we generated can be utilized to analyze individual stations. Our study indicated that our approach is more sensitive in estimating passenger demand during the morning peak counterflow, as the matrix had a lower sparsity than the one generated by TTS. Additionally, we identified patterns in demand among stations. A more thorough analysis of all stations can uncover symmetries in demand at various times of the day. These insights can be leveraged in advanced methods of learning algorithms or time series forecasting for tactical and operational decisions.

Future work will also include expand the analysis to other modes of public transport with Wi-Fi connection service available in Toronto to generate a multi-modal OD-matrix.

Acknowledgments. The authors confirm contribution to the paper as follows: study conception and design: D. Da Silva and A. Shalaby; data collection: D. Da Silva and F. Elsaid; validation and interpretation of results: D. Da Silva, F. Elsaid, and A. Shalaby; draft manuscript preparation: D. Da Silva, F. Elsaid. Project oversight: A. Shalaby. All authors reviewed the results and approved the final version of the manuscript.

References


Constructing Origin-Destination Matrix using Wi-Fi and AFC Data


Constructing Origin-Destination Matrix using Wi-Fi and AFC Data


Murchland JD (1966) Some remarks on the gravity model of traffic distribution, and an equivalent maximization formulation. London School of Economics and Political Science, Transport Network Theory Unit


