Understanding the temporal and spatial interactions between transit ridership and urban land-use patterns: an exploratory study

Article *in* Public Transport · June 2022

DOI: 10.1007/s12469-022-00296-2

Merkebe Demissie

The University of Calgary

31 PUBLICATIONS **437** CITATIONS

Lina Kattan

The University of Calgary

127 PUBLICATIONS **2,096** CITATIONS

Public Transport
https://doi.org/10.1007/s12469-022-00296-2

CASE STUDY AND APPLICATION

Understanding the temporal and spatial interactions between transit ridership and urban land-use patterns: an exploratory study

Merkebe Getachew Demissie[1] · Lina Kattan[1]

Accepted: 24 February 2022

Abstract

The land-use characteristics of urban areas continually change, and thus the activity patterns of the significant trip generators evolve. Efficient public transit planning needs to perform frequent estimates of the spatio-temporal distribution and dynamics of different activities in urban areas and measure the likely consequences of changes. Automated data collection systems usually collect transit ridership data (e.g., automated passenger count (APC)). Many transit agencies also generate General Transit Feed Specification (GTFS) data and make them publicly available. This study explores the use of APC, GTFS, and land-use data to examine various land-use and transit ridership interactions at the stop, route, and zonal levels using visualization, data mining, and statistical analysis techniques. Results show that transit ridership at the bus stop level gives a better understanding of each bus stop's unique land use. Zonal-level transit ridership patterns reveal the different trip generations and attraction roles of the neighboring land usage. This study could provide additional insights on the interaction between the temporal changes in population from the perspective of transit use and the associated land uses.

Keywords Transit ridership · Land use · Automated Passenger Count Data (APC) · General Transit Feed Specification data (GTFS) · Data mining · Transit demand

✉ Merkebe Getachew Demissie
merkebe.demissie@ucalgary.ca

Lina Kattan
lkattan@ucalgary.ca

[1] Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB, Canada

Published online: 08 April 2022

1 Introduction

An urban area can be described in terms of its form, such as buildings and other physical infrastructures, and its functions, such as the spatial interaction, movement and distribution of population to perform shopping, working, residing, etc. The set of relationships emerging out of its urban form and its underlying movement and interaction of population is best explained by land use and transportation (Rodrigue et al. 2016). There is a reciprocal interaction between land use and transportation that has been well investigated and modelled in the past (Padeiro 2014; TRB 1996). For example, land-use models consider the impact of transit services, in particular via accessibility, to measure development potential and the attractiveness of land (Waddell 2002; Deal and Schunk 2004); these will affect the pattern of trips and, therefore, impact the performance of the transport system (Ortúzar and Willumsen 2011).

Despite the interdependencies between land use and transportation, planning for land use and transportation (especially transit systems) often happens independently of one another. Badoe and Miller (2000) argue that one of the reasons for the lack of planning coordination is the scale of land use and transit planning, which usually happens at different levels of authority. While land-use planning and regulation is done by the jurisdiction of the local government, the scale of transit planning depends on spatial coverage, the mix of different modes, and the availability to be funded by the transit agency (Chakraborty and Mishra 2013). A number of previous studies have focused on the issue of transportation and land-use interaction with the objective of identifying the impact of land-use policies on travel behavior (Badoe and Miller 2000; Boarnet and Crane 2001); and investigated the connections between transit ridership and land-use and socio-economic variables within urban, suburban, and rural settings (Chakraborty and Mishra 2013; Cervero et al. 2010; Chakour and Eluru 2016; Kim et al. 2014; Gutiérrez et al. 2011).

The general consensus that has been widely reflected in transportation planning assumes spatial structures of significant trip generators are long lasting, which masks the short-term variability of activities people perform at these locations (Batty 2002; Bertaud 2004; Mungthanya et al. 2019). Urban areas continually change and the same urban structure could generate different levels of trips over time, for instance, a transit station when some event occurs, a recently renovated building that attracts new businesses (Demissie et al. 2015). Thus, public transit agencies need to perform frequent estimates of the spatio-temporal distribution of different activities in urban areas and measure the likely consequences of those changes upon transit uses. Many researchers have studied the interaction between transit ridership and land characteristics (Du and Mulley 2012; TRB 1996; Padeiro 2014; Tsai et al. 2012). These studies underlined the influence of three main land-use features on transit ridership such as urban structure, land-use density and urban design. However, the aforementioned studies do not provide adequate mechanisms on the specific topic of how transit ridership varies when the existing land use changes over time and what that means for transit planning.

The gap in our understanding of the transport and land-use interaction was primarily the result of data limitations (Badoe and Miller 2000; Hu et al. 2016). In recent years, public transit agencies collect large volumes of data. These data are obtained from on-board sensors introduced by Automated Vehicle Location (AVL), Automated Passenger Counting (APC), and Automated Fare Collection (AFC) systems. These datasets enable us to capture the diverse activity patterns of urban areas, detect emerging trends in a timely fashion, and unveil the social functions of different land uses. While many previous studies examined these data and provide an excellent starting point for this study (Sun et al. 2012; Pei et al. 2014; Yu and He 2017), their main focus was on capturing the spatio-temporal transit ridership pattern, efforts to infer the relationship between transit ridership patterns with different land-use characteristics is still overlooked.

In this study, we define the temporal and spatial interactions between transit demand and urban land-use patterns as the interplay between transit demand and land use in a way that they mutually affect each other across time and space. The functional characteristics of the land in terms of economic activities and people's daily activities occurring at different locations in a city affect transit ridership. We propose to capture the diverse profiles of urban areas at the transit stop, route, and zonal levels from the perspective of transit ridership data from APC and GTFS. The primary assumption is that transit ridership data can be used as a proxy to characterize people's activity patterns, thereby detecting emerging trends in a timely fashion, unveiling the evolution of land-use patterns, and the social functions of different land-use types. Transit agencies could use this information to identify the surge of emerging ridership and change in the land-use pattern across the city that may need attention in the upcoming service changes (e.g., schedule and route changes).

Two major contributions are made to the literature. First, we bring together GTFS data, APC data, and parcel-level land-use data to enrich the set of variables available for understanding the spatio-temporal distribution of transit ridership at the transit stop, route and zone levels. We capitalize on the untapped potentials of GTFS data. GTFS provides a static look at schedules, routes, trips, stops, and other moving parts of transit operations which was used to examine the temporal and spatial transit characteristics. Second, to study the transit ridership and land-use interactions, we have developed a set of methods to represent land-use types at the transit stop, route and zone levels. We use a new metric such as the Difference of Alighting's and Boarding's (DAB) to represent the time of day transit demand pattern at the transit stop and zone levels. Thus, this is a timely study showing the opportunities of various datasets and how effectively such datasets can be utilized to support the development of more quantitative and direct measures to understand the interaction between the demand for transit service at a specific time of the day and the trip generation and attraction roles of the associated land uses.

This paper is structured as follows. Section 2 surveys previous studies on the same topic. Section 3 describes the methodology including dataset description, data mining and statistical analysis techniques. Section 4 discusses the results and the final section outlines the conclusions and future works.

2 Literature review

The movement of people in an urban area is influenced by the distribution of their home, work and other activity destinations as well as in the transportation network connecting them (Decraene et al. 2013; Demissie et al. 2020). For instance, an urban transportation network is partially dependent on the commuting trips of citizens as determined by their respective origin locations (e.g. residential areas) to destinations (e.g. employment areas). On the other hand, the traditional perception of urban land use is rapidly changing due to the widespread use of information and communication technologies, mainly cellphone and the Internet (Demissie 2014; Soliman et al. 2017). For instance, a residential or touristic place could function as location for education or employment because of the ubiquitous use of networked communication. The individual land-use elements collectively influence city level dynamics. Thus, land-use characteristics of urban areas continually change and activity patterns of the significant trip generators evolve over short and long-time periods (Demissie et al. 2015).

Urban land use can be distinguished based on its physical characteristics (such as reflexivity and texture) or based on social function (such as residential, commercial, recreational, etc.) (Pei et al. 2014). Surveys have been used as traditional sources of information to classify land-use types. However, travel surveys are costly, time consuming, suffer from small samples, and are not frequently available to planners. A high-resolution remote sensing data can also be used for urban mapping. However, methods that rely on remote sensing information are limited to monitoring land cover since land usage is difficult to infer from physical infrastructure, specifically in mixed urban areas (Soliman et al. 2017).

Planners use service (catchment) areas of public transport stations (stops) to estimate the potential number of travelers. Two of the most common approaches are the circular buffer approach and the service area approach (Andersen and Landex 2008). However, the circular buffer approach can be unrealistic in the presence of indirect paths and impediments. Lam and Morrall (1982) suggested the importance of taking the detour factor during the delineation of catchment area boundaries. Theoretically, catchment (service) areas need to be mutually exclusive because transit users will use only one (closest) station, even if there are multiple stations within their origin or destination stations to access transit service. However, in the real world, generating non-overlapping service areas is challenging because of several reasons, including closely spaced transit stations at some parts of a city, possible overlapping regions, and different service areas for different directions of travel (Furth et al. 2007; Upchurch et al. 2004).

A number of previous works have employed various opportunistic data to characterize urban land-use types such as cellular network data (Calabrese et al. 2010; Demissie et al. 2015; Pei et al. 2014; Reades et al. 2007; Soto and Frias-Martinez 2011; Toole et al. 2012); micro-blogging services such as Twitter and Foursquare (Frias-Martinez and Frias-Martinez 2014; Wakamiya et al. 2011; Zhan et al. 2014); and GPS data (Liu et al. 2012; Demissie and Kattan 2022; Chang et al. 2019; Phithakkitnukooon et al. 2021; Kinjarapu et al. 2021). The aforementioned datasets have been combined with clustering (e.g., DBSCAN, k-means, fuzzy c-mean), and classification (e.g., tree-based models, support vector machine, artificial neural network) algorithms to identify land-use types of different urban areas. However, no studies exist that employ transit ridership data generated by the bus APC system to study land use and transit ridership interactions. The public transit agencies hold large volumes of data such as AVL, APC, AFC, and GTFS (Ge et al. 2021). The AFC and APC data have been used mainly for origin–destination trips estimation (Alsger et al. 2015; Ji et al. 2015; Li et al. 2011; Wang et al. 2011). The AVL data have been primarily used for transit reliability studies (Barabino et al. 2013; Mazloumi et al. 2010; Guido et al. 2016). The GTFS data have been used for a variety of purposes, including visualization (Kunama et al. 2017; Postsavee et al. 2020; Stewart et al. 2016); transit performance measurement (Wong 2013); real-time information and trip planning (Bast et al. 2015); and developing an analysis tool for planning (Lee et al. 2013).

Previous studies by (Du and Mulley 2012; TRB 1996; Lee et al. 2013; Tsai et al. 2012) investigated the relationship between transit ridership and urban forms. These studies underlined the interdependent nature of land use and transit system through which each supports and shapes the other (Padeiro 2014). Transit, especially high order public transport, can influence compact, mixed-use development, which in return can induce greater transit ridership. The Transit Cooperative Research Program report provided an insight regarding the importance of a high land-use mix to make transit an attractive choice; however, it is not easy to determine specific thresholds at which one has the right balance of land-use types that significantly influence transit ridership (TRB 1996). Tsai et al. (2012) and Hu et al. (2016) applied a set of analytical models to identify the impact of land-use variables on public transport demand. A number of other studies have analyzed transit ridership data (Sun et al. 2012; Pei et al. 2014; Yu and He 2017). However, the main focus of these studies was capturing the spatio-temporal transit ridership pattern, which lacks the ability to link the inferred transit patterns with different land-use characteristics. This work utilizes transit ridership data generated by bus users to measure the spatio-temporal changes in population from the perspective of transit use. In the process, we identify the relationship between the transit ridership and land use over the course of a typical week. Our analysis would allow to understand (i) Which spatio-temporal patterns of transit ridership exist? and (ii) what types of urban land use are generally associated with these patterns?

3 Methodology

The aim of the developed methodology is to investigate the spatial and temporal patterns of bus transit ridership and its association with land-use types at the transit stop, route and zonal levels. We follow the following three steps: (i) develop a set of metrics to characterize the land-use mix at the stop, route and zone levels; (ii) generate a week-long transit ridership pattern for each level of analysis; and (iii) apply a visualization, data mining, and statistical analysis techniques to examine the interactions between the transit ridership and the land-use patterns for each level of analysis. These methods are applied to a case study using bus transit GTFS, APC, and land-use data from the City of Calgary, Canada.

We use different types of Traffic Analysis Zones (TAZ) for different stages of the study. The stop level analysis will be based on information obtained from bus stops located within 500-by-500 m grid cells. While the 500-by-500 m grid cells function well to generate mutually exclusive TAZs and cover the entire city, they are limited. For instance, a bus stop falls into a given grid cell but may have a significant portion of its catchment area outside that cell leading to a misrepresentation of coverage. The route level analysis will be carried out based on the service area of transit stops along the path of each transit route. We define the service area of a bus stop using a circular buffer. The zonal level analysis will be carried out based on the zoning term given to parcels of land by the City of Calgary.

3.1 Data and case study area description

The city of Calgary has around 1.23 million inhabitants with a total area of 825.3 km^2 (Calgary 2016). Calgary Transit provides the public transportation services in Calgary, which is owned and operated by the City of Calgary. Calgary Transit currently operates two light rail transit (LRT) routes, 433 bus routes, over 6000 stops, and the total transit routes cover 4369 km. Figure 1 shows Calgary Transit's system

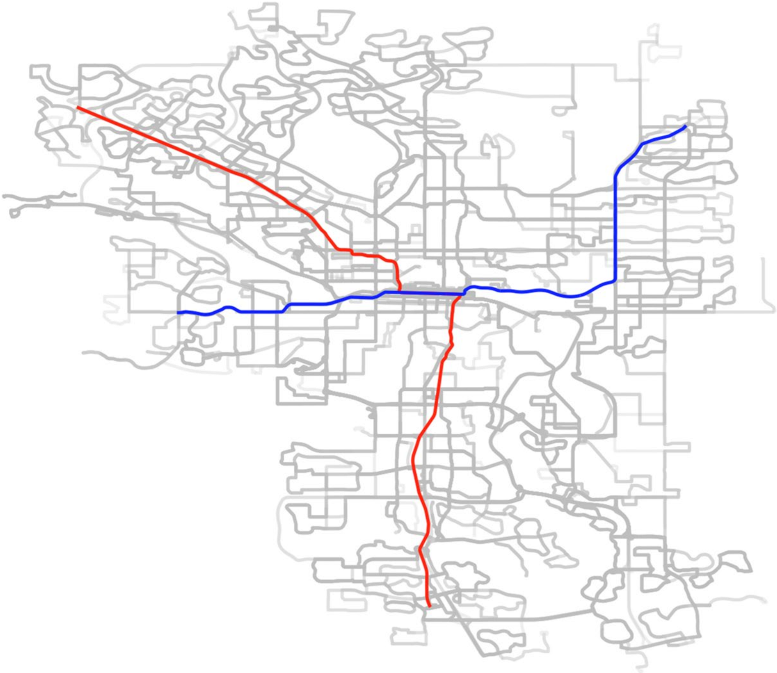

Fig. 1 Calgary Transit system map

map. There are two LRT routes (red and blue routes) spanning four lines. Figure 1 also shows the bus routes (grey color). This study is carried out using APC and GTFS data, which is obtained from Calgary Transit. The land-use data is obtained from the city of Calgary open data catalogue (Calgary 2016).

The availability of transit service within a reasonable walking distance to one's origin and destination is one of the primary factors for the decision to use public transit. Another important factor is the temporal transit service coverage. To show some measures of transit service coverage, the city of Calgary is divided into 500-by-500 m grid cells. Then, the number of bus stops and scheduled bus services (i.e. departures or arrivals) in each cell is calculated, and a Kriging function is used to create a smooth interpolated surface (Fig. 2). Darker green in Fig. 2a indicates areas with a high level of transit spatial coverage with more than ten bus stops. Darker red indicates a lower level of transit supply with no bus stops. Figure 2b is generated based on the aggregate number of scheduled bus services (i.e. temporal coverage) in each grid cell. The number of scheduled services are calculated for one week, and it refers to the number of times a bus departed/arrived at the bus stops in each cell. This result provides additional information regarding service frequency coverage.

3.1.1 GTFS data

GTFS is a common format arrangement established for the sharing of public transportation schedules and associated geographic information. A GTFS feed contains at least six, and up to 13 CSV files representing routes, stops, stop times, calendar, and shapefiles, etc. GTFS feeds allow public transit agencies to publish their transit data and share them with the general public and application developers that can develop

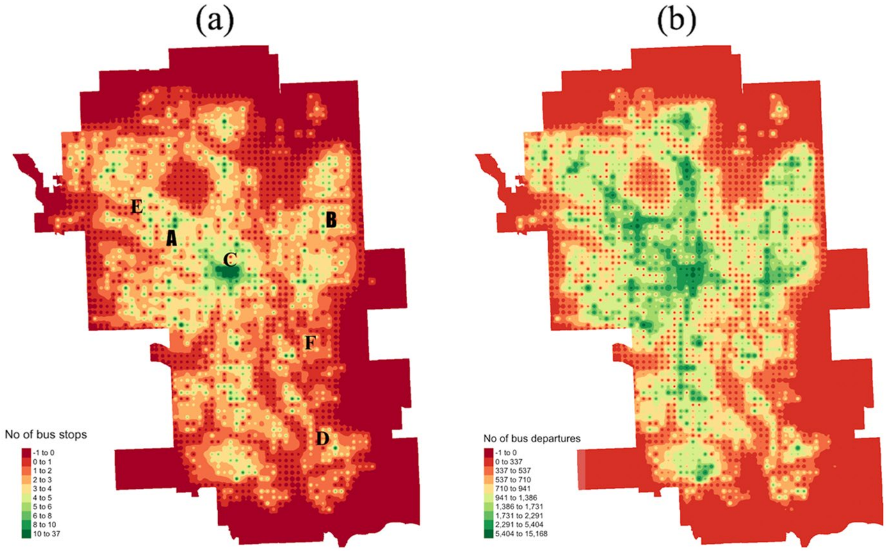

Fig. 2 Transit service coverage. **a** Spatial service coverage, **b** Temporal service coverage

different types of smart phone-based services. GTFS represents fixed route and scheduled public transit operations. Along with the GTFS-static feed, google recently started GTFS-realtime specification. A detailed explanation of all GTFS files and data fields is available on the GTFS reference website (Google 2016). The analysis in this study is based on the openly available GTFS data between August 2016 and December 2016. A relational table of the GTFS data is shown in Appendix A.1.

3.1.2 APC data

Calgary Transit has installed APC devices in more than 30 percent of its bus fleet, which automatically count the number of passengers as they board and alight at each bus stop during a trip. Since only 30 percent of the buses are equipped with the APC system, not all the transit routes are surveyed throughout the year. APC-equipped buses have to operate on selected routes over different time periods along with the inbound and outbound directions for a given period until the required APC data are collected. This study uses APC data collected between September 6/2016 to October 3/2016. Section 3.2 shows how the 27 days of APC data are reduced to a week-long ridership pattern. The data aggregation procedure developed in Sect. 3.2 allows us to expand the sample data and obtain a fuller version of ridership data for each level of analysis (transit stop, route, and zone). A relational table of the APC data is shown in Appendix A.2.

3.1.3 Land-use data

The dataset contains the land-use polygons and their designation based on the city of Calgary's land-use bylaw (1P2007) that has been in effect since June 1, 2008 (Calgary 2016). The land uses that are either permitted or discretionary include the following major categories: Residential (low-, medium-, and high-density residential areas), commercial (commercial- neighborhood, corridor and community; commercial-office), institutional (community institution, large-scale health, religious, and educational centers), recreational (community parks, school parks, public education, special purpose-school, community reserve and urban nature), industrial (industrial-business, industrial-outdoor, industrial-edge, industrial-general, industrial-heavy, industrial-commercial), major infrastructure (special purpose city and regional infrastructures), future urban development (lands that are awaiting urban development and is largely limited to uses that can easily be removed to allow for future urban development), direct control (direct control is a customized land-use designation for developments that require specific regulations unavailable in other land-use districts), mixed, and the rest is labeled as Other.

3.2 Transit stop aggregation, land-use mix, and a week-long transit ridership pattern generation for each level of analysis

In the first part of our analysis, we seek to merge the GTFS data which provides a pre-planned look at schedules, routes, and trips with the actual historic trips data derived from the APC system. We aggregate transit stops based on some common features and common service areas for the transit stops. The next step is determining the land-use mix that is likely to influence transit use within the vicinity of a bus stop and along the path of each transit route. Then, we generate a week-long transit ridership pattern for each level of analysis. These components are explained below.

In Fig. 3, the primary APC, GTFS and land-use data are arranged in the top row. Queries and output tables are used to process and store values for subsequent steps.

The raw APC data do not have a timestamp associated to the boarding and alighting passengers recorded at the bus stop. The time variables associated to these records are the date and the trip start and end time. Query 1 assigns a bus stop level time value (time_of_day) by taking the average of the trip start and trip end times.

Output 1 contains the number of boarding and alighting passengers associated with each trip. However, it is not possible to distinguish between different patterns of services (e.g. direction, short/long trips and trip destinations) running in the same route. Query 2 merges output 1 with GTFS trip.txt data where each trip can be augmented with additional variables such as trip direction and headsign (direction_id, and trip_headsign). In addition, the date format (September 6/2016 to October 3/2016) is transformed to a day_of_week variable (Monday, Tuesday, etc.).

Query 3 calculates the total number of scheduled trips for a week across the Calgary bus network. GTFS's calender.txt file holds information about service start and end dates and determines the set of dates on which a given service can be active. Each service is identified by the service_id field that can only appear once and can be defined for one or more routes. Trips are uniquely associated with the service_id. Thus, one can know which trips (identified by the trip_id) are available on particular days of the week between valid start and end dates of the service_id. This information is checked against GTFS's calender_dates.txt file, which holds information about specific days when a trip is not available (e.g. because of holidays).

Query 4 does two calculations: (i) it calculates what percentage of the scheduled trips is surveyed by the APC system between September 6/2016 and October 3/2016. This will allow us to measure the representativeness of the sample; and (ii) it selects the transit routes that have a complete one-week APC data (these routes will be later used for the route level ridership and land-use interaction analysis). In Query 2, each trip is associated with a day_of_week variable. Then, the average number of boarding and alighting passengers are calculated for scheduled trips that have similar trip characteristics (day_of_week, trip_id, route_id, and trip_headsign fields). This procedure improves the chance of obtaining representative week-long ridership data from each of the scheduled trips across all the transit routes. This way, the 27 days of APC data are reduced to a set of week-long ridership data. This analysis allows us to expand the sample into a fuller version of ridership data (number of boarding and alighting passengers) for each level of analysis (transit stop, route, and zone).

In addition to the number of boarding and alighting passengers, we use a new metric to represent the transit demand at the transit stop and zonal level analysis. The new metric is obtained by subtracting the number of boarding passengers from the number of alighting passengers. This metric will forthwith be referred to as DAB (the Difference of Alighting's and Boarding's). The time-of-day DAB variable is used to understand the interaction between the demand for transit service at specific times of the day and the trip generation and attraction roles of the associated

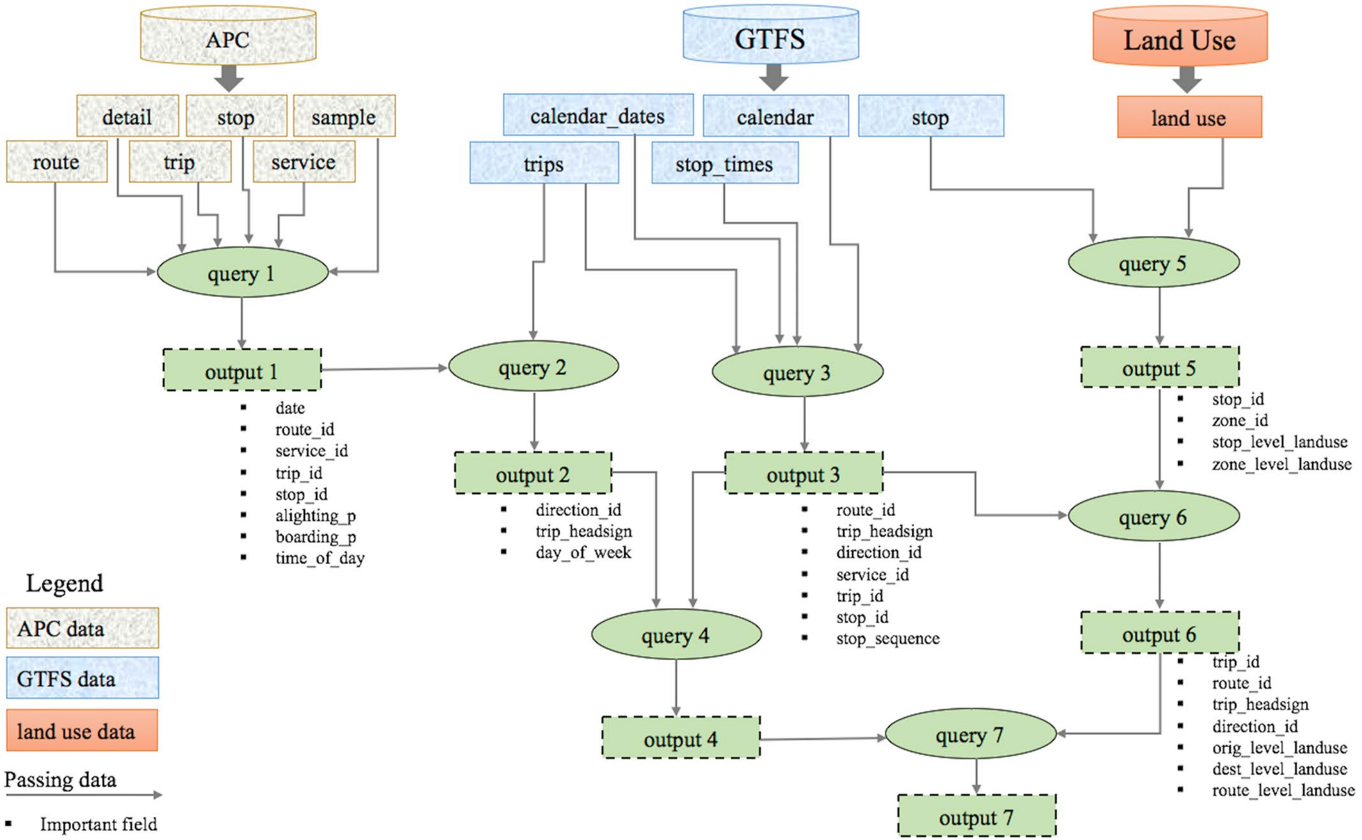

Fig. 3 Data preparation flow chart

land uses. A positive DAB value shows the presence of a higher number of alighting passengers than boarding passengers indicating the trip attraction nature of the associated land use at that time of the day. On the other hand, a negative DAB value suggests the trip generation nature of the associated land use at that time of the day.

Query 5 calculates the land-use mix that is likely to influence transit use within the vicinity of a bus stop. The first step is to determine the service area of a stop. A previous study by Lam and Morrall (1982) found that the average walking distance to a bus stop in Calgary is 327 m and the 75th-percentile walking distance is 450 m. Calgary Transit use a design guideline based on the 75th-percentile walking distance. For this study, the radiuses of the catchment areas that are used for major transfer, suburban, and commercial business district (CBD) stations are 700 m, 450 m, and 350 m, respectively. Once the service area of each bus stop is approximated, the next step is to calculate the land-use types covered by the service area. As shown in Fig. 8b, the city of Calgary is represented by ten major land-use types. With respect to the land-use polygon data, the land-use mix within each catchment area is calculated based on the portion of land encompassed by the area. For instance, for a given bus stop *b*, the portion of land covered by its catchment area is calculated as $Area_b \cap Area_z$, where $Area_b$ is the catchment area of bus stop *b*, and $Area_z$ is the area of land-use polygon *z* within the vicinity of bus stop *b*. Fractions of areas with the same land-use type are summed up to understand the land-use mix patterns within the vicinity of a bus stop.

The next step is to determine the land-use mix along the path of each route. For each route, the total number of stops are counted. Based on the direction of each route and bus stop sequence information, the stops are categorized into two groups. Transit stops in the first half are labeled as origin stops, which shows the land-use characteristics of a route from the origin side. Transit stops in the second half are labeled as destination stops, which reflects the land-use characteristics of a route at the destination side. Query 6 calculates three different types of transit route land-use variables: (i) orig_level_landuse: is obtained based on the land covered by the catchment areas of the origin side transit stops; (ii) dest_level_landuse: is obtained based on the land covered by the catchment areas of destination side transit stops; and (iii) route_level_landuse: is obtained based on the land covered by the catchment areas of all the transit stops along the path of a transit route.

Finally, Query 7 assigns the three different types of transit route land-use variables to the transit routes with a complete one week APC data (output 4). The output from this analysis is used for the route level land use and transit ridership interaction analysis.

3.3 Clustering analysis

A clustering technique is used to understand the interaction between the transit demand at specific times of the day and the associated land uses that may influence the timing of that demand at a zonal level. A k-means clustering technique is applied to create groups of TAZs that are similar in terms of a week-long transit ridership pattern. The k-means clustering algorithm identifies clusters of behavior and returns a typical member of that cluster represented by the mean behavior in that group.

TAZ is based on the zoning term given to a parcel of land by the city of Calgary, which describes land use that is either permitted or discretionary. The process of associating transit ridership and land-use characteristics at the zonal level follows the following steps: (i) generate a week-long transit ridership pattern for each zone; (ii) cluster the week-long transit ridership patterns using the k-means clustering technique; and (iii) associate each of the resulting cluster to the corresponding predominant land-use type.

The number of boarding and alighting passengers recorded at the bus stop level are assigned to their respective TAZ. The TAZs are used to aggregate the number of boarding and alighting passengers. This aggregate number estimates the zonal level week-long hourly transit ridership data. To initiate the clustering process, the hourly transit ridership data associated to each zone is transformed to a common scale to make a proper comparison between zones. For each TAZ i, we normalize the hourly transit data (V_i) over time for each day of the week. Then, each TAZ is represented by the normalized hourly transit data, $V_i = v_{i,d_1,h_1}, v_{i,d_1,h_2}, v_{i,d_1,h_3}, \ldots, v_{i,d_5,h_{24}}$, where d and h are the day of the week (5 weekdays), and hour of the day (24 h), respectively. Using the k-means clustering algorithm, the normalized hourly transit data are classified into $k\,(\leq n)$ clusters, where n is the total number of TAZs. The k-means clustering technique is used to segment the TAZs into k clusters. Each cluster is characterized by its centroid, and the algorithm aims to minimize an objective function, in this case a squared error function in Eq. (1):

$$\sum_{j=1}^{k} \sum_{V_i \in C_j} |distance(V_i, Centroid_j)|^2 \tag{1}$$

where C_j is the set of TAZs related to cluster j, and $Centroid_j$ is the mean of all the points in cluster j. The distance between the normalized week-long hourly transit data of two zones (V_1, V_2) is calculated using squared Euclidean distance in Eq. (2):

$$distance\left(V_1, V_2\right) = \left(\sum_{H=1}^{120} |V_{1H} - V_{2H}|^2 \right)^{1/2} \tag{2}$$

H is the total number of features that are used to characterize each TAZs (24 hourly transit data × 5 weekdays = 120 in this case).

There are a number of methods suggested to determine the number of clusters in a data set (Tan et al. 2005). In this study, we use the “elbow” method to determine the appropriate number of clusters in our data set. The idea of the elbow method is to run k-means clustering on the datasets for a range of values (e.g., 2–20). For each value of k, calculate an error measure (the within-cluster sums of squares) and plot them against the number of clusters. Then, take the number of clusters associated to the “elbow” on the plot.

In our analysis, different clustering techniques are possible candidates to segment zones based on their time of day transit ridership profile. K-means is a simple unsupervised machine-learning algorithm and is chosen because of its simplicity in implementation. Previous studies have also shown that the k-means clustering technique can be used to identify clusters of locations with similar-zoned uses based on activity patterns generated from opportunistic datasets (Reades et al. 2007; Becker et al. 2011).

3.4 Statistical analysis

We apply a statistical test to investigate the presence of enough evidence to support the assumption that transit routes with different levels of mixed land uses are likely to influence different levels of transit ridership. The underlying assumption is that the demand for public transit service at a specific time of the day may reveal how a particular land is used at that time of the day. For example, residential areas are known to have a high number of boarding passengers in the morning period and college or university areas exhibit a reverse pattern.

Statistical tests can be classified as parametric and nonparametric tests. Parametric tests assume underlying statistical distributions in the data, whereas, nonparametric tests do not rely on any distributions. In the case when the underlying assumptions for a parametric test are met (e.g., the assumption about normally distributed data), the t-test is a commonly used parametric test for the comparison of two population means. However, the t-test can perform well with continuous data that are not normally distributed if a sufficiently large sample is obtained (Lumley et al. 2002). The decision to choose between parametric and non-parametric tests would rather depend on a particular summary measure, for example, whether the mean or the median appropriately represents the center of the data's distribution (Lumley et al. 2002). The idea behind testing the difference between two means is based on selecting samples from two groups and comparing the means of the group. For samples that are independent of each other, two types of t-test can be used depending on whether the sample variances are equal or not.

4 Results and discussions

4.1 Aggregated stop level land use and transit ridership interactions

4.1.1 Results from selected locations

To conduct the aggregated bus stop level analysis, six locations presumably with different land-use characteristics are selected. Each of these locations has a homogeneous land use and is chosen based on prior knowledge of the city of Calgary. The locations are shown on Fig. 2a (A, B, C, D, E, F). Proximity and bus stop name (text) measures are considered to aggregate bus stops relevant to the selected locations:

Commercial Core (C): represents the city center accommodating various office towers, banks, legal firms, shopping centers, investment advisors, governments.
Commercial (D): area around the South Trail Crossing shopping in the southeast part of the city.

Institutional (A): area around the University of Calgary in the northwest part of the city.

Residential (B): medium-density residential area in the northeast part of the city.

Recreational (E): area around Varsity estates park in the northwest part of the city.

Industrial (F): around the 70th Ave SE part of the city with major industrial employment centers.

Figure 4 shows the number of boarding and alighting passengers for each sample location shown in Fig. 2a. The commercial core (C) and institutional (A) areas are associated with a high number of boarding and alighting passengers and the commercial (D), residential (B), recreational (E) and industrial (F) areas have a relatively low number of boarding and alighting passengers. The preliminary results indicate that the stop level transit ridership pattern reflects a detailed and unique sight-specific profile for a particular land-use type. For example, the commercial core, institutional and residential areas have clear morning and afternoon peaks. The distinct peaking in the number of boarding and alighting passengers in the morning and afternoon peak periods indicates a significant commuting activity in both directions. On the other hand, the commercial, recreational, and industrial areas show a diverse and distributed usage along the hours of the day.

The demand for transit in each of the six locations is further examined using the DAB metric. Figure 5 shows the patterns of the DAB values for the six locations displayed in Fig. 2a (A, B, C, D, E, and F). The commercial core and institutional areas are associated with a positive DAB (high number of alighting passengers) in the morning and a negative DAB (high number of boarding passengers) in the afternoon peak periods. The residential area exhibits a reverse pattern to the commercial core and institutional areas. The industrial area exhibits three daily peaks with a high number of passenger arrivals in the early morning, a high number of passenger departures in the afternoon peak, and a high number of passenger arrivals in the evening again. A high number of boarding passengers is observed before the opening hours of the South Trail Crossing shopping center. This pattern can be associated to the mall-walking program, where elderly people use malls in the early morning for indoor exercise because of the cold weather in Canada. The differences in the transit ridership patterns in each of the six locations reflect how the associated land uses influence the timing of the transit demand along the hours of the day.

4.1.2 Citywide transit ridership pattern

The catchment area of a bus stop can be determined on a stand-alone basis or as part of a group of bus stops located close to each other. Lee et al. (2013) emphasized the importance of combining catchment areas of bus stops that provide service in all directions to capture the trip generation and attraction characteristics of the area. For example, transit users who make home-based-work trips tend to return to their origin (home), indicating the necessity of aggregating transit stops in both directions to capture departure and arrival ridership records and obtain the complete daily ridership pattern of the area.

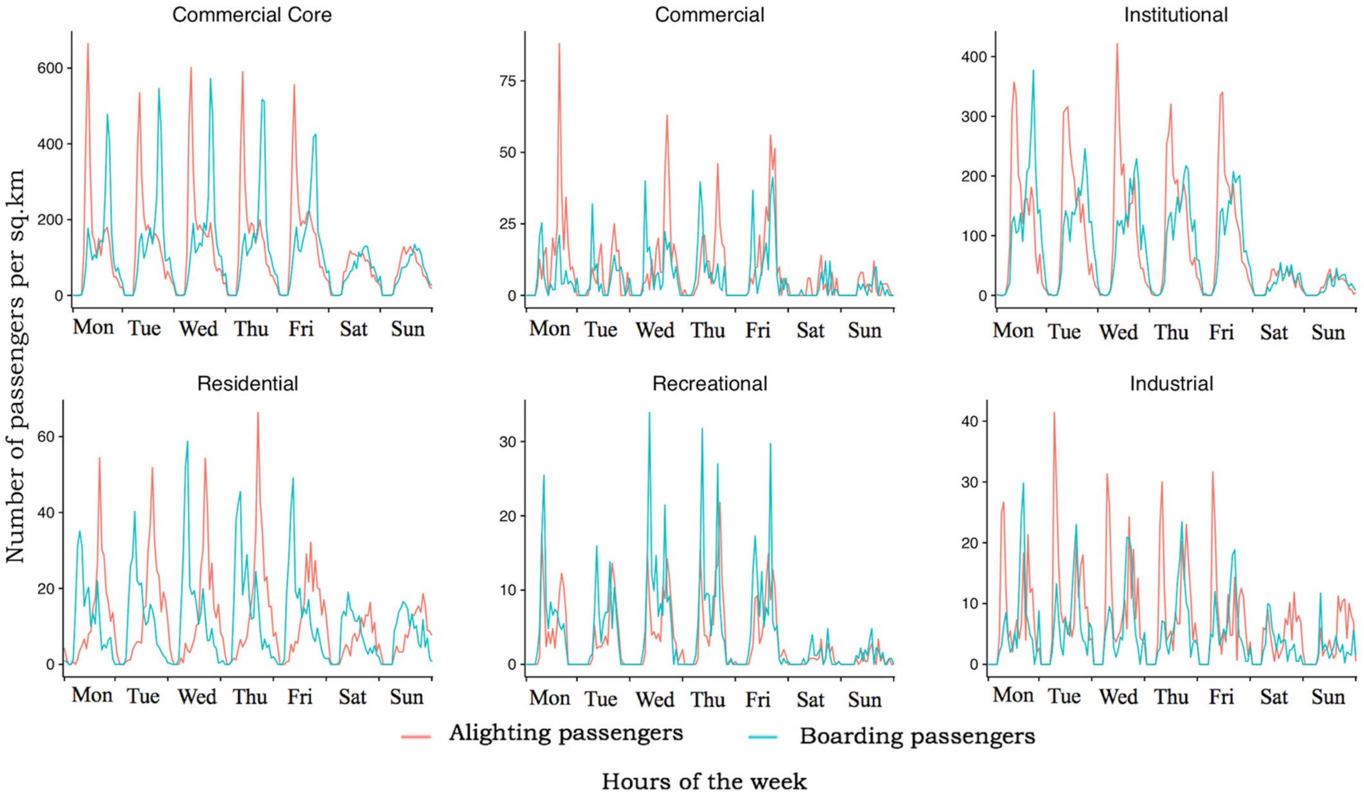

Fig. 4 Temporal variation of the number of boarding and alighting passengers for the selected six sample locations shown in Fig. 2a (A, B, C, D, E, and F)

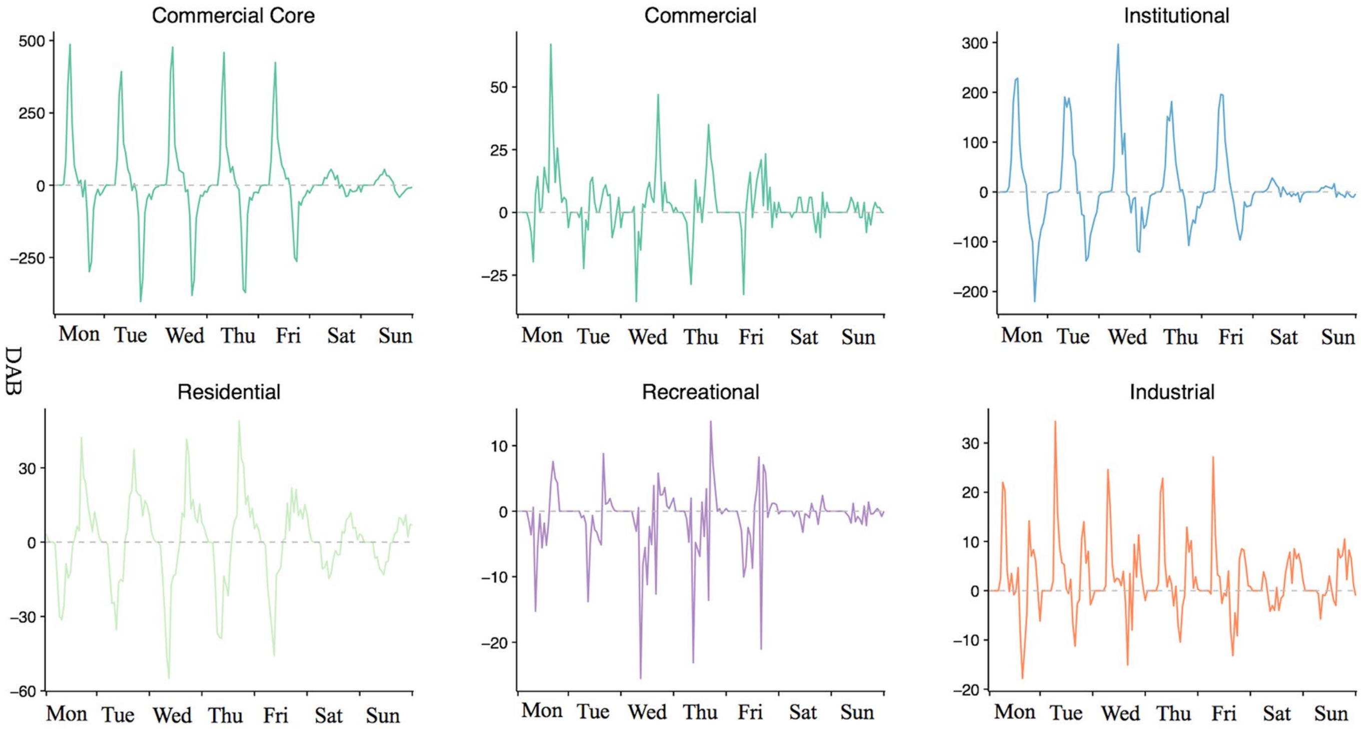

Fig. 5 The DAB values for the selected six sample locations shown in Fig. 2a (A, B, C, D, E and F)

We further analyzed the spatio-temporal evolution of transit ridership data based on information obtained from a group of bus stops that are located within 500-by-500 m grid cells. The hourly week-long stop level ridership data (calculated in Query 4) are aggregated for each grid cell. Then, we calculated hourly transit ridership values for the following four periods: morning peak (8am to 9am, representing time between 8:00 a.m. and 9:59 a.m.); day off-peak (10 a.m. to 3 p.m.); afternoon peak (4 p.m. to 6 p.m.); and evening and night off-peak (7 p.m. to 7a.m.). The peak and off-peak periods are obtained based on the citywide transit ridership trend. The citywide ridership pattern is fairly the same across the week days. Figure 6 illustrates Monday's spatio-temporal patterns of boarding and alighting passengers. During the morning and afternoon peak, a high number of boarding and alighting passengers are observed across the city, especially in the downtown, Northwest (NW), Northeast (NE) and Southwest (SW) parts of the city. There is also a significant number of boarding and alighting passengers during the day time off-peak period. The transit ridership decreases in the evening and night off-peak period. A similar

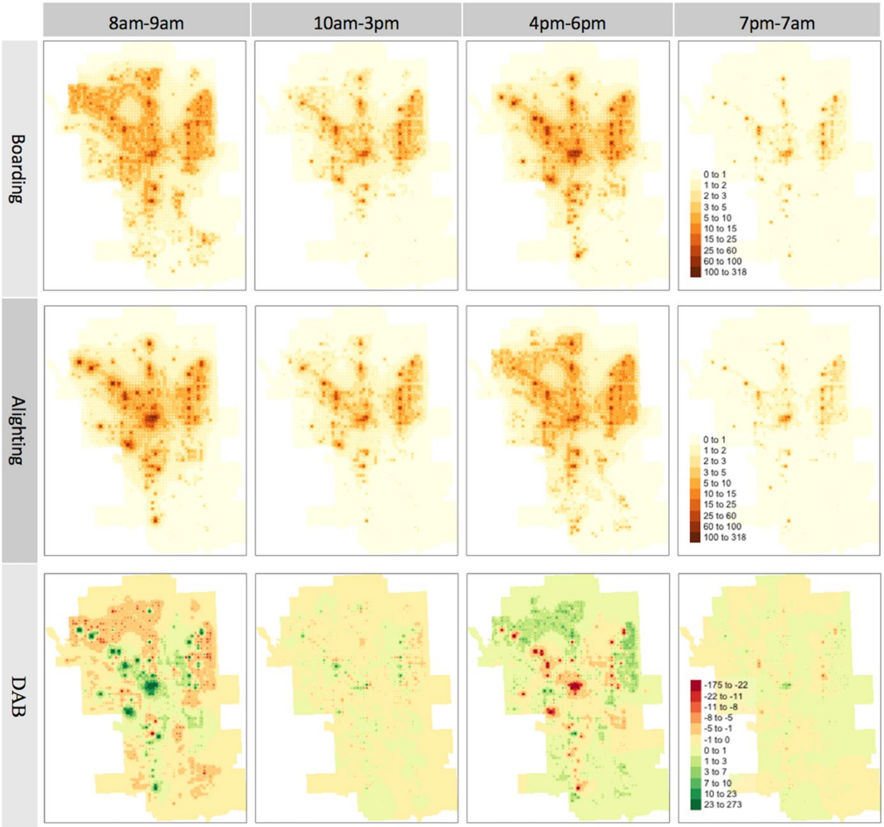

Fig. 6 Transit ridership pattern on a weekday, Monday

pattern across the four periods is the low transit ridership in the Southeast (SE) part of the city.

Figure 6 also reveals the major trip departure and arrival locations using the DAB metric. The darker green color indicates more arrivals of transit users and darker red indicates more departures of transit users. The major trip departure and arrival locations are more noticeable during the morning and afternoon peak hours, but in a reverse pattern. However, there are fewer major departure and arrival locations in the day, and evening and night time off-peak periods.

4.2 Zonal level land use and transit ridership interactions

The clustering technique discussed in Sect. 3.3 is used to understand the interaction between the transit demand at a specific time of the day and the associated land uses that may influence the timing of that demand at a zonal level. Zones (TAZs) are based on the zoning term given to a parcel of land by the city of Calgary, which describes land use that is either permitted or discretionary. This study uses ten major land-use groups such as residential, recreational, commercial, industrial, institutional, major infrastructure, future development, mixed use, direct control, and “other”. The number of boarding and alighting passengers recorded at the bus stop level are assigned to their respective TAZ. This results in zonal level week-long hourly transit ridership data. For each zone, a DAB vector is obtained.

The DAB metric is used for zonal level analysis to understand the relationship between the zonal level transit ridership patterns and the trip generation and attraction roles of the associated land uses. The magnitude of the differences between zones DAB values makes it hard to compare them in a more detailed way, so the DAB values should be transformed into a common scale. Feature scaling or unity-based normalization is used to bring the DAB values into the range (0,1). The DAB values are normalized over time for each day of the week. For each TAZ, the hourly DAB values are scaled relative to TAZ’s peak DAB value. Using this approach, we can identify zones with similar transit ridership patterns. However, the normalized DAB values are not measures of total transit demand. Since values are relative, they should not be used to compare transit demand between TAZs. Thus, zones associated with different levels of transit ridership intensities may belong to the same land-use type (e.g., low-density residential, medium-density residential, and high-density residential zones).

Because of the low sample size of the APC data on Saturdays and Sundays, weekend data are excluded for the zone level analysis. There are 2170 zones considered for this analysis. Thus, a matrix of 2170 TAZ × 120 DAB (5 days × 24 h) size is generated as input for the clustering analysis. Using the normalized DAB values, zones are classified into different groups using the k-means clustering technique. Figure 7a shows the centers of the six clusters. Figure 7b shows the geographical distribution of zones grouped into the six clusters.

A thorough examination of the transit ridership patterns of the six clusters shows a strong constancy of peak points over the period of each day. Each of the six patterns has two major peak points along the course of each day, but of different

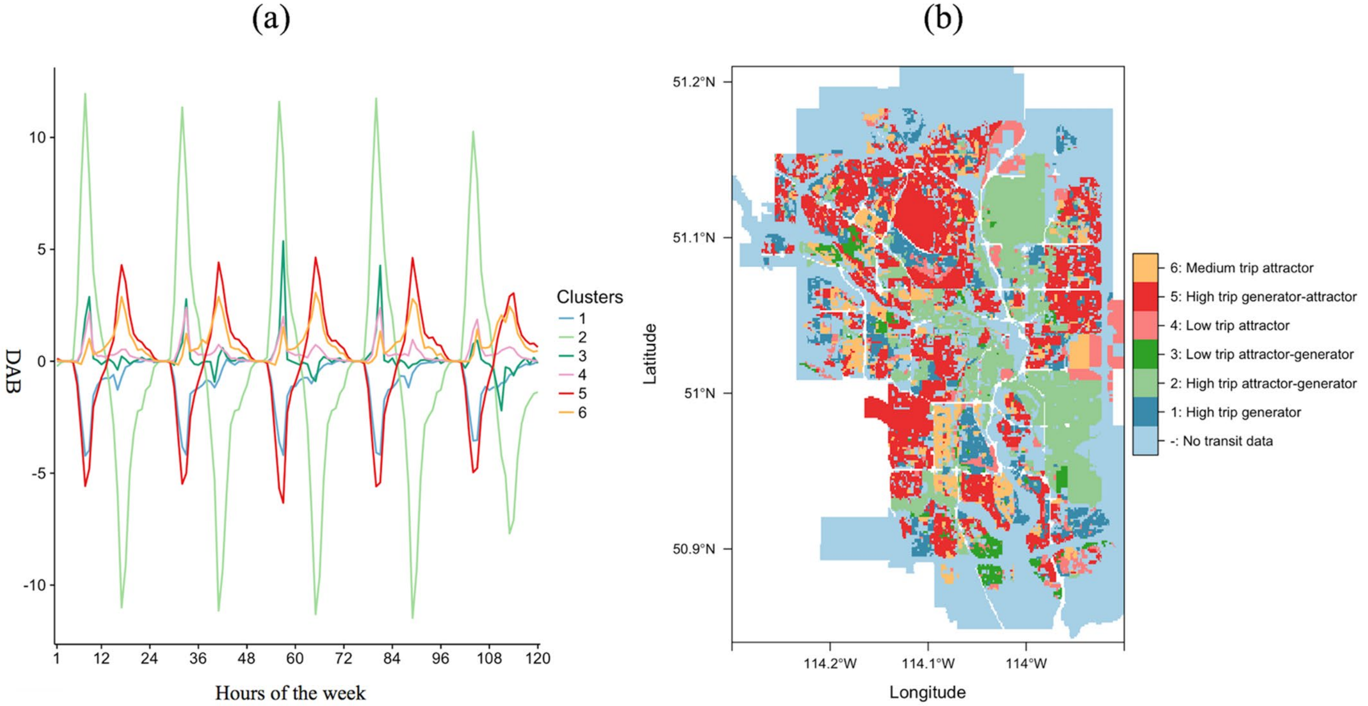

Fig. 7 **a** Patterns based on the mean values of DAB in each cluster, **b** geographical distribution of zones grouped into the six clusters

patterns, suggesting that the corresponding land uses may have different trip generations and attraction roles. The peak points mainly occur in the morning and afternoon peak periods. Thus, we assign an easy-to-understand description to each zone based on their morning/afternoon/day-long trip attraction and generation roles:

Cluster 1-High trip generator (a negative DAB throughout the day suggesting a day-long trip generation).
Cluster 2-High trip attractor-generator (a positive DAB in the morning and a negative DAB in the afternoon suggesting a morning trip attraction and afternoon trip generation roles).
Cluster 3-Low trip attractor-generator (exhibits a similar pattern with Cluster 2 but with lower intensity).
Cluster 4-Low trip attractor (exhibits a reverse pattern with Cluster 1).
Cluster 5-High trip generator-attractor (exhibit a reverse pattern with Cluster 2).
Cluster 6-Medium trip attractor (a positive DAB throughout the day suggesting day-long trip attraction).

Generally, a high percentage of riders tends to use transit for the two legs of a round trip (the outward leg and return leg). In addition, passengers access a transit service through the nearest station, and they tend to end their last trip at the station where they begin their first trip of the day, which translates to zero DAB for that station or TAZ. This is especially true for commuting trips, school trips, or other home-based trips. However, in Fig. 7, the DAB of Cluster 1, Cluster 4, and Cluster 6 are larger or smaller than zero for all hours of a day. This may be due to several reasons. For instance, in addition to commuting and school trips, passengers use transit for various reasons and activities. A school child, for example, might be dropped by the parent at the school in the morning while returning home in the afternoon using transit or on foot. A recent Calgary Transit survey showed that only 67% of their customers use transit for commuting trips and going to/from school (Calgary Transit 2020). Passengers may not complete both parts of their round trip using transit as some of these trips may constitute other modes (e.g., Uber, passenger car, bike, walk). A recent Calgary Transit survey showed that monthly and day pass tickets sale generated around 61.8% of the fare revenue. The remaining 38.2% was due to single ticket sales, which shows that some passengers may be using transit only for one leg of their trip (Calgary Transit 2020). In addition, due to service unavailability, the second leg of passengers' journey may not start at the destination station of their previous trip, or the second leg of their journey may not end strictly at the same station where they begin their first trip of the day.

To further understand the association of the six clusters with zonal land uses, we observe the proportion of zones that fall into one of the six clusters. The result is shown in Fig. 8a. For example, for the residential areas, 32.4% fall into Cluster 5, which is a high trip generator in the morning peak and a high trip attractor in the afternoon peak; 47% fall into Cluster 1 and Cluster 6, which have a high trip generation and medium trip attraction roles, respectively. In the case of commercial areas, 25% are characterized as medium trip attractors, 21% high trip attraction in the morning and high trip generation in the afternoon, 21% high trip generator attractor.

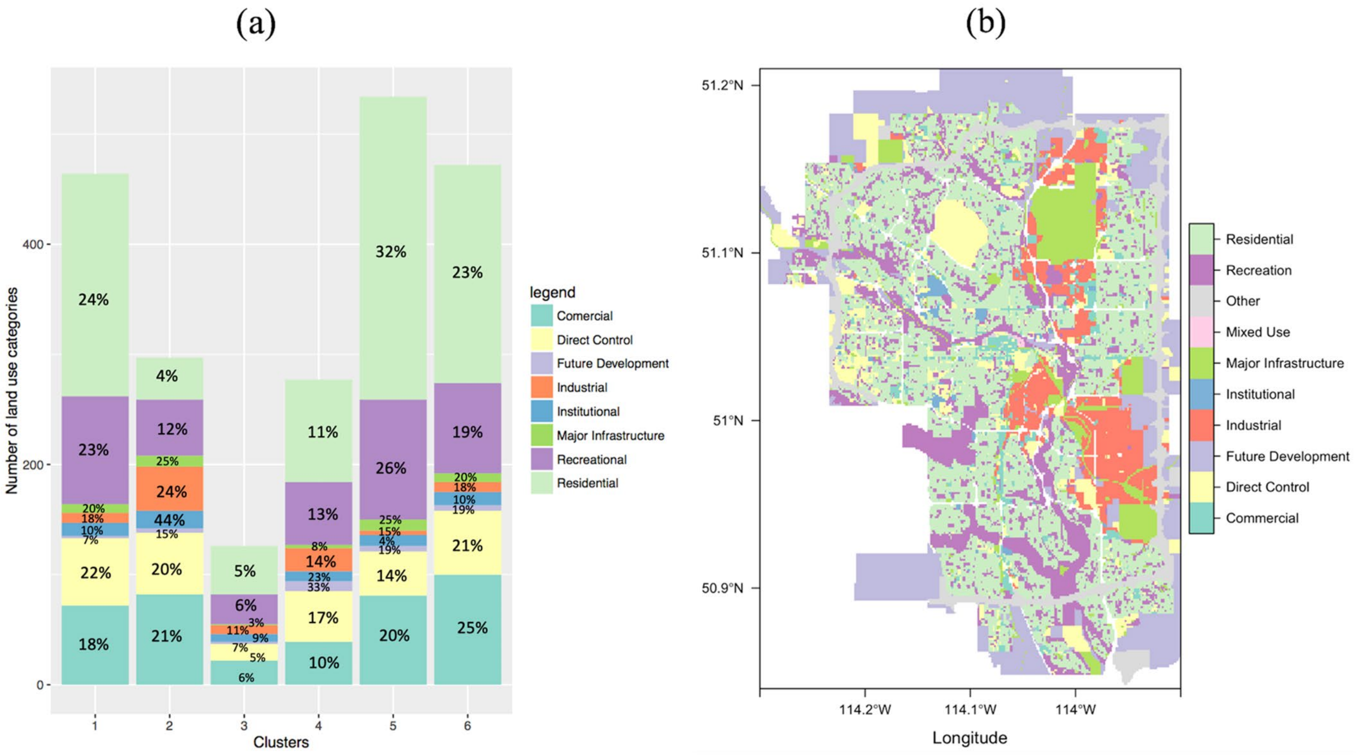

Fig. 8 **a** Number of land-use categories in each cluster, the proportion of land-use categories that fall into each cluster (for example, for residential land-use category, 24%, 4%, 5%, 11%, 32%, and 23% fall into Cluster 1 to 6, respectively), and **b** city of Calgary land-use polygons

On the other hand, 43% of industrial areas and 24% institutional areas are characterized as high trip attractors in the morning and high trip generators in the afternoon (cluster 2). Major infrastructure areas fall into the categories of Cluster 2 (25%) and Cluster 5 (25%), which exhibit a reverse pattern with Cluster 2.

Figure 8a shows the land-use type attributed to each cluster showing that more than one land use category fall into each cluster. The observation is that no single land-use category shows a unique cluster pattern. Figure 8b shows the distribution of land-use polygons. According to the city of Calgary's land-use bylaw (1P2007), how a particular area of a city is used is determined by the zoning term given to a parcel of land by the city of Calgary. In practice, however, many zones may feature different appears which might also differ somewhat from intended use. The deviation from the intended usage can contribute to the existence of more than one land-use category in a single cluster. The other reason may be the fuzzy nature of urban land-use types, where one can observe mixed land uses in a single zone. According to the city of Calgary's land-use bylaw (1P2007), a land-use polygon designated as commercial can have residential units on upper floors of buildings, and predominantly residential areas also have commercial sectors. The mixed usage within a TAZ causes difficulty to generate a unique pattern for each land-use category.

4.3 Route level land use and transit ridership interactions

One of the primary focuses of transit planning is the land-use mix around the major origin and destination stops (terminus locations). In addition to the terminus land-use mix, what is equally important is the land-use mix along the path of a transit route. A high land-use mix at the transit station supports higher off-peak ridership and provides access to a wide range of activities and uses (Krizek 2003). It also allows transit users to accomplish most of their activities along a single route and reduce the number of transfers transit users have to make (TRB 1996). For the route level land use and transit ridership interaction analysis, an observation of 56 routes is considered. Transit routes are considered if they have at least ten scheduled trips per week.

For the purpose of this study, fractions of land-use polygons (in terms of their areas) within the vicinity of bus stops catchment areas along each route are aggregated to measure the land-use mix of a route (Query 6). A transit route can serve up to ten different land-use types based on the ten land-use categories adopted in this paper. Within the Calgary Transit network a transit route can provide short and long trips depending on the time of day and day of week. Thus, the information on land-use composition of a transit route can be different depending on the direction of operation, time of day, and day of week. Therefore, separate land-use composition information is obtained for each trip associated to the selected 56 routes. For example, route 105 operates between Lions Park and Dalhousie terminuses. The land-use type served based on the scheduled trip on Tuesday resulted in the following land-use categories: residential (54%), recreational (19%), direct control (14%), commercial (6%), institutional (5%) and the remaining five land-use types have a sum of 2%. Figure 9 shows the frequency of the proportion of the first five predominant land-use types along the paths of the 56 examined transit routes. The averages of the proportions of the first, second, third, fourth, and fifth land-use type across the paths of the transit routes are 64.11%, 15.34%, 9.10%, 4.64%, and 2.46%, respectively (each group can have up to ten different land-use types).

By average, about 80% of the land-use fabric along a transit route is represented by its primary and secondary land use. For simplicity, we only use the primary and secondary land use to make up the land-use mix that is likely to influence transit use along the path of a transit route. We counted the number of times a land-use type appears in the first five predominant land-use types. The result shows that the residential land-use type appears as primary land-use for 93% of the time. Thus, we rely on the secondary use to separate transit routes of different land-use mixes. For example, there are two cases to regard a transit route which is predominantly commercial: (i) if the primary use of a transit route is residential and commercial is its secondary use; and (ii) if the primary use of a transit route is commercial and any other land use is its secondary use. A similar measure is taken to delineate associated predominant land uses to transit routes as predominantly industrial or predominantly institutional. On the other hand, because of the high similarity of the residential and recreational land uses, a transit route that has residential and recreational as primary/secondary use is regarded as predominantly residential.

The aggregated transit stop and zonal level analysis reveal the importance of the patterns of the morning and afternoon peak points in terms of revealing the trip attraction and trip generation roles of the corresponding land. We apply statistical methods based on the comparison of two means to quantitatively examine the difference in the significance of transit ridership behavior of transit routes for different land-use mixes.

Table 1 shows the results of twelve hypothesis tests such as the difference between two population means. First, in order to determine which t-test to use, an F-test is conducted to test if the variances of the two samples are equal. In every case, we use a two-tailed test with an assumption that variances are equal and a

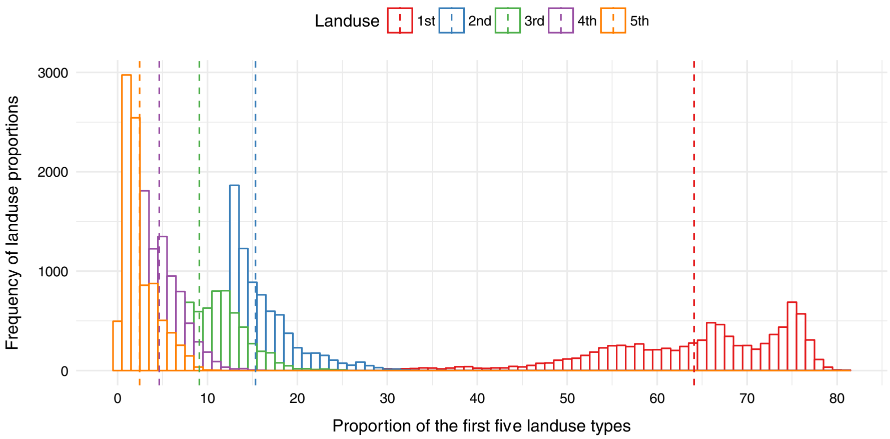

Fig. 9 Percentage of land-use types along the paths of transit routes

Table 1 Results of the hypothesis tests

No	Groups	Origin/Destination based Primary and secondary land uses		Data	Time period	Mean	Var	n	t-Test P-value	F-Test P-value
1	I	Residential and Recreation Recreation and Residential	Origin	Boarding Passengers	7 a.m. to 9 a.m.	0.798	0.186	1340	< 0.001	< 0.001
	II	Institutional and other Land Uses Residential and Institutional				0.497	0.078	80		
2	I	Residential and Recreation Recreation and Residential				0.798	0.186	1340	0.283	< 0.001
	II	Commercial and other Land Uses Residential and Commercial				0.780	0.114	127		
3	I	Residential and Recreation Recreation and Residential				0.798	0.186	1340	< 0.001	< 0.001
	II	Industrial and other Land Uses Residential and Industrial				0.393	0.061	98		
4	I	Institutional and other Land Uses Residential and Institutional	Destination	Alighting passengers	8 a.m. to 10 a.m.	0.772	0.191	92	< 0.001	< 0.001
	II	Residential and Recreation Recreation and Residential				0.441	0.123	1115		
5	I	Commercial and other Land Uses Residential and Commercial				0.734	0.107	323	< 0.001	0.071
	II	Residential and Recreation Recreation and Residential				0.441	0.123	1115		
6	I	Industrial and other Land Uses Residential and Industrial				0.686	0.125	109	< 0.001	0.023
	II	Residential and Recreation Recreation and Residential				0.441	0.123	1115		

test if they are unequal. Four of the twelve tests resulted in unequal variances and in those occasions, a t-test for unequal variances is used. In every case, a significance level of 0.05 is chosen and the one-tailed t-test is used to determine whether the difference between means found in the sample is significantly different from the hypothesized difference between means.

The following discussion explains the first two hypothesis tests. Based on the aggregated stop level land use and transit ridership interaction analysis, during a morning peak period, we observe that the residential area generates more trips when compared to the institutional area. The first hypothesis test aims to examine the presence of enough evidence to support this finding. The F-test shows unequal variances. Thus, a t-test for unequal variances is used. The variable we use is the average number of boarding passengers at a bus stop per trip (x). We calculated the mean value of x from both groups: (I) transit routes that are serving predominantly residential areas at the origin section of the route; and (II) transit routes that are serving predominantly institutional areas at the origin section of the route (the transit route land-use mix at the origin stops and at the destination stops is calculated using Query 6). The null hypothesis is that the mean value of x obtained from group I is less than or equal to the mean value of x obtained from group II. The alternative hypothesis is that the first mean is bigger than the second mean. Since the comparison is trip generation in the morning peak, the number of boarding passengers between 7 to 9am is used. According to the p-value, there are only $<0.1\%$ chances the null hypothesis is true given our sample. This shows that the alternative hypothesis is true at the observed and more specific confidence level of 99.9%. This supports the assumption that, in the morning peak, residential areas generate more trips when compared to institutional areas.

The second hypothesis test aims to examine the presence of a significant difference between the number of boarding passengers obtained from the following two groups during the morning peak period: (I) transit routes that are serving predominantly residential areas at the origin section of the route; and (II) transit routes that are serving predominantly commercial areas at the origin section of the route. The F-test shows unequal variances. Thus, a t-test for unequal variances is used. According to the p-value, there are 28.3% chances the null hypothesis is true given our sample. This shows that we cannot reject the null hypothesis, since the p-value (0.283) is greater than the significance level (0.05). From a total of twelve hypothesis tests, in two cases (2nd and 11th tests) we cannot reject the null hypothesis, since the p-values are greater than the significance level (0.05).

5 Conclusions

In recent years, advancements in information and communication technologies enable transit agencies to generate a variety of data on a high frequency and volume basis. These datasets offer transit agencies an unprecedented chance to examine the spatio-temporal patterns of passengers and compute multiple transit performance measures to continuously evaluate and plan their transit services. However, these datasets have yet to be fully exploited for the purpose of land use and transit

Table 1 (continued)

No	Groups	Origin/Destination based Primary and secondary land uses		Data	Time period	Mean	Var	n	t-Test P-value	F-Test P-value
7	I	Institutional and other Land Uses Residential and Institutional	Origin	Boarding passengers	4 p.m. to 6 p.m.	0.703	0.228	76	0.0270	0.025
	II	Residential and Recreational Recreational and Residential				0.594	0.136	1353		
8	I	Commercial and other Land Uses Residential and Commercial				0.750	0.141	238	< 0.001	0.352
	II	Residential and Recreational Recreational and Residential				0.594	0.136	1353		
9	I	Industrial and other Land Uses Residential and Industrial				0.663	0.135	112	0.028	0.494
	II	Residential and Recreational Recreational and Residential				0.594	0.136	1353		
10	I	Residential and Recreational Recreational and Residential	Destination	Alighting passengers	5 p.m. to 7 p.m.	0.749	0.135	1254	< 0.001	< 0.001
	II	Institutional and other Land Uses Residential and Institutional				0.586	0.064	72		
11	I	Residential and Recreational Recreational and Residential				0.749	0.135	1254	0.035	0.049
	II	Commercial and other Land Uses Residential and Commercial				0.708	0.160	217		
12	I	Residential and Recreational Recreational and Residential				0.749	0.135	1254	< 0.001	0.003
	II	Industrial and other Land Uses Residential and Industrial				0.583	0.089	108		

ridership interaction studies. This study used datasets such as GTFS and APC that most transit agencies collect and provide to examine land use and transit ridership interactions at the stop, route and zone levels.

The stop level analysis provides insight into the transit ridership and land-use patterns of a particular transit service area. A DAB metric is developed by subtracting the number of boarding passengers from the number of alighting passengers. The variation of the DAB metric along the hours of day provides an indication regarding the trip attraction and generation roles of a particular location. This study also attempted to examine the linkage between zonal level land-use characteristics and transit ridership patterns. Using the zonal level DAB values, six distinct transit ridership patterns are derived representing the existing land-use characteristics in the city. A thorough examination of these patterns shows the existence of strong regularity of morning and afternoon peak points over the course of each day. Based on the different patterns of these peak points, we were able to differentiate the trip attraction and generation roles of the corresponding zonal land uses. To carry out the route level analysis, fractions of land-use polygons (in terms of their areas) within the vicinity of bus stops catchment areas along each route are aggregated to measure the land-use mix of a route. Then, the statistical method based on the comparison of two sample means is used to quantitatively examine the route level land use and transit ridership interactions.

The proposed methods and results of the study offer several potential applications for urban and transit planning. First, the results of this study could provide additional insights on the interaction between the spatio-temporal changes in a population from the perspective of transit ridership and the associated land uses that influence the temporal distribution of transit demand at the stop, route, and zonal levels. The proposed route-level stop aggregation method provides a foundation that can simplify the complexity of the transit network, where a single metric can represent multiple transit routes that serve a similar land-use mix. The aggregated transit stop and zonal level analysis reveal the corresponding land's spatio-temporal trip attraction and trip generation roles. Second, transit agencies can use this approach to make valuable recommendations to better plan transit services and learn about emerging land-use patterns. For example, Calgary Transit makes four primary service changes within a year. These changes are usually made based on feedback from passengers and drivers, political decisions, changing ridership levels, and new development areas. In this regard, our study provides a framework to track the frequent fluctuation of transit ridership and land-use patterns and analyze their interaction. Transit agencies could use this information to diagnose the presence of emerging spatio-temporal ridership and land-use patterns across the city that may need attention in the upcoming service changes (e.g., schedule and route changes).

Despite the significance of our analysis, we should emphasize some limitations of the study. In our analysis, we define the service area of a bus stop using a circular buffer. Future studies should explore the design of the catchment areas of bus stops by considering the actual feeder routes or adding additional resistance to specific points in the road network (e.g., stairways) to represent the real world. Another limitation is related to the issue of overlapping catchment areas. In our approach, we define the catchment area of a bus stop using a circular buffer with a predefined radius, and an overlapping of catchment areas can happen in places where bus stops are less spaced (e.g., downtown areas). Future studies should explore the implementation of mutually exclusive service areas for each transit station (Upchurch et al. 2004).

Another limitation is related to the route level analysis. Routes are broken arbitrarily by the number of stations (e.g., the first half of stations are labeled as origin stops, while the second half are labeled as destination stops). Normally, bus stops are less spaced in downtown areas compared to outskirt areas, and, thereby, for radial routes, the center of the line (in terms of the first half of the stops for outbound trips) can be shifted towards the downtown which can decrease the accuracy of results. While the method to calculate the land-use mix along with the route path works well for many route structures (e.g., bi-directional or two-way linear), it may have a limitation when dealing with circular or loop routes. Future studies should incorporate route features (e.g., a two-way circular route, a route connecting low-density suburban places, or a circular route connecting major key destinations) to improve an origin-based and destination-based stops categorization. Also, some routes may have a major transfer station at a specific point along the route, which will mean that these routes may not follow the pattern that has been assumed in our methodology. Therefore, future studies should incorporate bus stop spacing information by the transit agencies and transfer volumes at the interchange points in their methodology.

There are a number of future avenues that might be pursued: (i) The route level analysis considers ridership data of individual trips of a transit route. Depending on the availability of full network ridership data, this analysis can be extended to a transit corridor level. For example, if there is more than one route serving the same corridor (or arterial) or serving two nearby streets, each of the routes will have less frequency and, thereby, less ridership. However, the lesser ridership is not related to the land uses around the stops, but rather to the bus transit system design; (ii) The stop and zonal level analysis attempted to address the interactions of transit ridership and land-use patterns. The route level analysis investigates the effect of land-use mix that is likely to influence the intensity of transit use along the path of a transit route. This approach can be improved by incorporating demographic and socioeconomic data that influence transit demand; and (iii) Several variables describe land-use factors, including accessibility, land-use density, and land-use mix. Future studies should explore the influence of these variables on transit demand and its spatial and temporal distribution, and its relationship with several land-use types and features of the built environment.

6 Appendix

See Fig. A1 and A2.

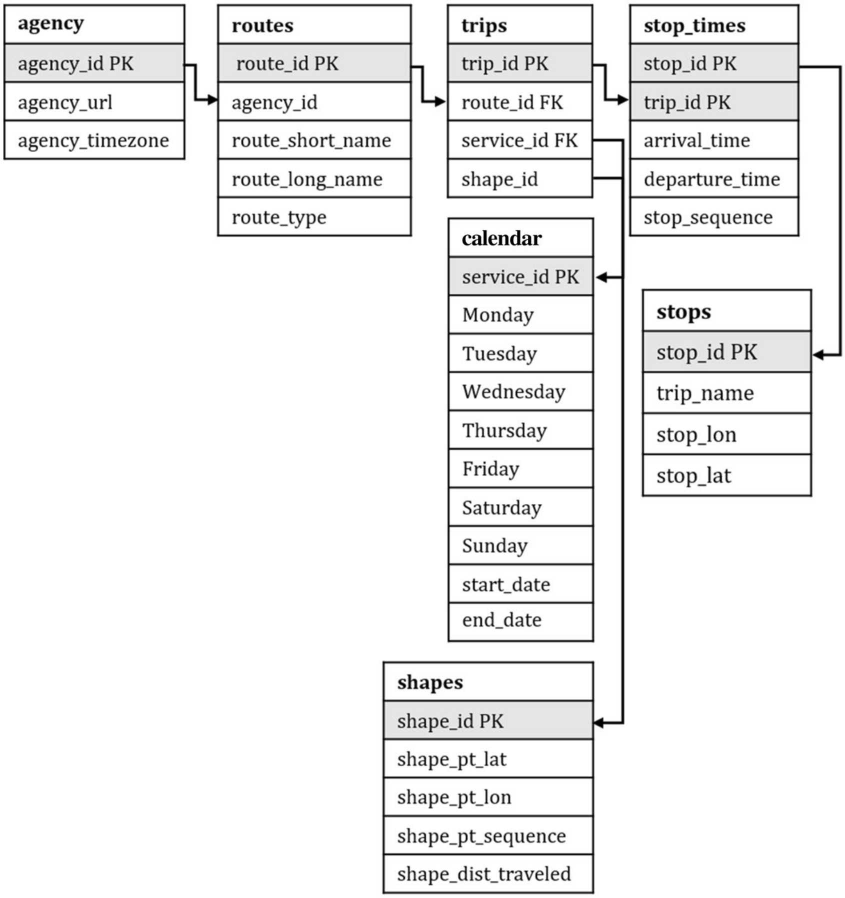

Fig. A1 Relational table of GTFS data (Kaeoruean et al. 2020)

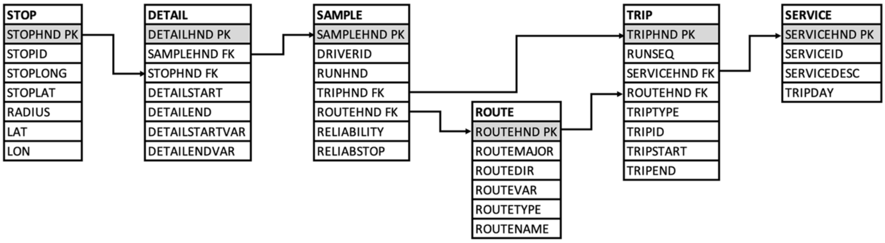

Fig. A2 Relational table of APC data

Acknowledgements We thank Calgary Transit and the City of Calgary for letting us use their data. This work is funded by the NSERC Discovery grant RT735236 (RGPIN/03942-2020); the Alberta Innovate Strategic Grant on Integrated Urban Mobility 10024748 (G2018000894); the NSERC CREATE on Integrated Infrastructure for Sustainable Cities (IISC) RT1002081 (CREATE/511060-2018); the Urban Alliance Chair in Transportation Systems Optimization RT756247; and the Schulich School of Engineering Startup Grant.

References

Alsger AA, Mesbah M, Ferreira L, Safi H (2015) Use of smart card fare data to estimate public transport origin-destination matrix. Transp Res Record 2535(1):88–96. https://doi.org/10.3141/2535-10

Andersen JLE, Landex A (2008) Catchment areas for public transport. WIT Trans Built Environ 101:175–184. https://doi.org/10.2495/UT080171

Badoe DA, Miller EJ (2000) Transportation-land-use interaction: empirical findings in North America, and their implications for modeling. Transp Res Part D 5(4):235–263. https://doi.org/10.1016/S1361-9209(99)00036-X

Barabino B, Di Francesco M, Mozzoni S (2013) Regularity analysis on bus networks and route directions by automatic vehicle location raw data. IET Intel Transport Syst 7(4):473–480. https://doi.org/10.1049/iet-its.2012.0182

Bast H, Brosi P and Storandt S (2015) Real-time movement visualization of public transit data. In: 22nd ACM SIGSPATIAL International Conference, pp 331–340. https://doi.org/10.1145/2666310.2666404

Batty M (2002) Thinking about cities as spatial events. Environ Plann B Plann Des 29:1–2. https://doi.org/10.1068/b2901ed

Becker RA, Cáceres R, Hanson K, Loh JM, Urbanek S, Varshavsky A, Volinsky C (2011) A tale of one city: using cellular network data for urban planning. IEEE Pervasive Comput 10(4):18–26. https://doi.org/10.1109/MPRV.2011.44

Bertaud A (2004) The spatial organization of cities: deliberate outcome or unforeseen consequence? IURD Working Paper Series. https://doi.org/10.1017/9781316271377.004.

Boarnet M, Crane R (2001) The influence of land use on travel behavior: specification and estimation strategies. Transp Res Part A 35(9):823–845. https://doi.org/10.1016/S0965-8564(00)00019-7

Calabrese F, Reades J, Ratti C (2010) Eigenplaces: segmenting space through digital signatures. IEEE Pervasive Comput 9(1):78–84. https://doi.org/10.1109/MPRV.2009.62

Calgary (2016) Open Calgary: The City of Calgary's Open Data Portal

Calgary Transit (2020) Calgary Transit Ridership, Revenue and RouteAhead update. https://pub-calgary.escribemeetings.com/filestream.ashx?DocumentId=139367

Cervero R, Murakami J, Miller M (2010) Direct ridership model of bus rapid transit in Los Angeles County, California. Transp Res Record 2145:1–7. https://doi.org/10.3141/2145-01

Chakour V, Eluru N (2016) Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. J Transp Geogr 51:205–217. https://doi.org/10.1016/j.jtrangeo.2016.01.007

Chakraborty A, Mishra S (2013) Land use and transit ridership connections: implications for state-level planning agencies. Land Use Policy 30(1):458–469. https://doi.org/10.1016/j.landusepol.2012.04.017

Chang AYJ, Miranda-Moreno L, Clewlow R, and Sun L (2019) Trend or Fad? Deciphering the enablers of micromobility in the U.S. SAE International

Deal B, Schunk D (2004) Spatial dynamic modeling and urban land use transformation: a simulation approach to assessing the costs of urban sprawl. Ecol Econ 51(1–2):79–95. https://doi.org/10.1016/j.ecolecon.2004.04.008

Decraene J, Monterola C, Lee GKK, Hung TGG (2013) A quantitative procedure for the spatial characterization of urban land use. Int J Modern Phys C 24(01):1250092. https://doi.org/10.1142/S0129183112500921

Demissie MG, Correia G, Bento C (2015) Analysis of the pattern and intensity of urban activities through aggregate cellphone usage. Transportmetrica A 11:502–524. https://doi.org/10.1080/23249935.2015.1019591

Demissie MG (2014) Combining datasets from multiple sources for urban and transportation planning: emphasis on cellular network data. PhD thesis. Coimbra University

Demissie MG, Kattan L (2022) Estimation of truck origin-destination flows using GPS data. Transp Res Part E 145:102621. https://doi.org/10.1016/j.tre.2022.102621

Demissie MG, Kattan L, Phithakkitnukoon S, Correia GHA, Veloso M, Bento C (2020) Modeling location choice of taxi drivers for passenger pick-up using GPS data. IEEE Intell Transp Syst Mag 13(1):70–90. https://doi.org/10.1109/MITS.2020.3014099

Du H, Mulley C (2012) Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. J Transp Land Use 5(2):46–59. https://doi.org/10.5198/jtlu.v5i2.225

Frias-Martinez V, Frias-Martinez E (2014) Spectral clustering for sensing urban land use using Twitter activity. Eng Appl Artif Intell 35:237–245. https://doi.org/10.1016/j.engappai.2014.06.019

Furth P, Mekuria M, SanClemente J (2007) Parcel-level modeling to analyze transit stop location changes. J Public Transp 10(2):73–91. https://doi.org/10.5038/2375-0901.10.2.5

Ge L, Sarhani M, Voß S, Xie L (2021) Review of transit data sources: potentials, challenges and complementarity. Sustainability 13(20):11450. https://doi.org/10.3390/su132011450

Google (2016) "GTFS." 2016. https://developers.google.com/transit/gtfs

Guido G, Vitale A, and Rogano D (2016) Assessing public transport reliability of services connecting the major airport of a low density region by using AVL and GIS technologies. In: EEEIC 2016—International Conference on Environment and Electrical Engineering. https://doi.org/10.1109/EEEIC.2016.7555483

Gutiérrez J, Cardozo OD, García-Palomares JC (2011) Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. J Transp Geogr 19:1081–1092. https://doi.org/10.1016/j.jtrangeo.2011.05.004

Hu N, Legara EF, Lee KK, Hung GG, Monterola C (2016) Impacts of land use and amenities on public transport use, urban planning and design. Land Use Policy 57:356–367. https://doi.org/10.1016/j.landusepol.2016.06.004

Ji Y, Mishalani RG, McCord MR (2015) Transit passenger origin-destination flow estimation: efficiently combining onboard survey and large automatic passenger count datasets. Transp Res Part C 58(1):178–192. https://doi.org/10.1016/j.trc.2015.04.021

Kaeoruean K, Phithakkitnukoon S, Demissie MG, Kattan L, Ratti C (2020) Analysis of demand-supply gaps in public transit systems based on census and GTFS data: a case study of Calgary, Canada. Public Transport 12:483–516. https://doi.org/10.1007/s12469-020-00252-y

Kim K, Kyuhyup Oh, Lee YK, Kim SH, Jung JY (2014) An analysis on movement patterns between zones using smart card data in subway networks. Int J Geogr Inf Sci 28:1781–1801. https://doi.org/10.1080/13658816.2014.898768

Kinjarapu A, Demissie MG, Kattan L, Duckworth R (2021) Applications of passive GPS data to characterize the movement of freight trucks—a case study in the Calgary region of Canada. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/tits.2021.3093061

Krizek KJ (2003) Operationalizing neighborhood accessibility for land use—travel behavior research and regional modeling. J Plan Educ Res 22(3):270–287. https://doi.org/10.1177/0739456X02250315

Kunama N, Worapan M, Phithakkitnukoon S, Demissie M (2017) GTFS-VIZ: tool for preprocessing and visualizing GTFS data. UbiComp/ISWC, pp 388–396. https://doi.org/10.1145/3123024.3124415

Lam W and Morrall J (1982) Bus passenger walking distances and waiting times: a summer–winter comparison. Transport Quart 36(3):407–421

Lee SG, Hickman M, Tong D (2013) Development of a temporal and spatial linkage between transit demand and landuse patterns. J Transport Land Use 6(2):33–46. https://doi.org/10.5198/jtlu.v6i2.268

Li D, Lin Y, Zhao X, Song H and Zou N (2011) Estimating a transit passenger trip origin-destination matrix using automatic fare collection system. Lect Notes Comput Sci 6637:502–513. https://doi.org/10.1007/978-3-642-20244-5_48

Liu Y, Wang F, Xiao Y, Gao S (2012) Urban land uses and traffic ‘source-sink areas’: evidence from GPS-enabled taxi data in Shanghai. Landsc Urban Plan 106(1):73–87. https://doi.org/10.1016/j.landurbplan.2012.02.012

Lumley T, Diehr P, Emerson S, Chen Lu (2002) The importance of the normality assumption in large public health data sets. Annu Rev Public Health 23:151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Mazloumi E, Currie G, Rose G (2010) Using GPS data to gain insight into public transport travel time variability. J Transp Eng 136(7):623–631. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000126

Mungthanya W, Phithakkitnukoon S, Demissie MG, Kattan L, Veloso M, Bento C, Ratti C (2019) Constructing time-dependent origin-destination matrices with adaptive zoning scheme and measuring their similarities with taxi trajectory data. IEEE Access 7:77723–77737. https://doi.org/10.1109/ACCESS.2019.2922210

Ortúzar JD, Willumsen LG (2011) Modelling transport, 4th edn. Wiley. https://doi.org/10.1002/9781119993308

Padeiro M (2014) The influence of transport infrastructures on land-use conversion decisions within municipal plans. J Transport Land Use 7(1):79–93. https://doi.org/10.5198/jtlu.v7i1.373

Pei T, Sobolevsky S, Ratti C, Shaw SL, Li T, Zhou C (2014) A new insight into land use classification based on aggregated mobile phone data. Int J Geogr Inf Sci 28:1988–2007. https://doi.org/10.1080/13658816.2014.913794

Phithakkitnukooon S, Patanukhom K, Demissie MG (2021) Predicting spatiotemporal demand of dockless E-scooter sharing services with a masked fully convolutional network. ISPRS Int J Geo Inf 10(11):773, 17 pages. https://doi.org/10.3390/ijgi10110773

Postsavee P, Phithakkitnukoon S, Demissie MG, Kattan L, Ratti C (2020) Visualizing public transit system operation with GTFS data: a case study of Calgary, Canada. Heliyon 6(4):e03729

Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. IEEE Pervasive Comput 6(3):30–38. https://doi.org/10.1109/MPRV.2007.53

Rodrigue JP, Comtois C, Slack B (2016) The geography of transport systems, 4th edn. Routledge, London. https://doi.org/10.4324/9781315618159

Soliman A, Soltani K, Yin J, Padmanabhan A, Wang S (2017) Social sensing of urban land use based on analysis of Twitter users’ mobility patterns. PLoS ONE 12(7):e0181657, 16 pages. https://doi.org/10.1371/journal.pone.0181657

Soto V and Frias-Martinez E (2011) Robust land use characterization of urban landscapes using cell phone data. In: Proceedings of the 1st Workshop on Pervasive Urban Applications, in Conjunction with 9th Int. Conf. Pervasive Computing.

Stewart C, Diab E, Bertini R, El-Geneidy A (2016) Perspectives on transit: potential benefits of visualizing transit data. Transp Res Record 2544(1):90–101. https://doi.org/10.3141/2544-11

Sun L, Lee DH, Erath A, Huang X (2012) Using smart card data to extract passenger’s spatio-temporal density and train’s trajectory of MRT system. Proc ACM SIGKDD Int Conf Knowl Discover Data Min, pp 142–148. https://doi.org/10.1145/2346496.2346519

Tan PN, Steinbach M, Kumar V (2005) Cluster analysis: basic concepts and algorithms, chapter 8 in introduction to data mining. Addison-Wesley

Toole JL, Ulm M, González MC, Bauer D (2012) Inferring land use from mobile phone activity. Proc ACM SIGKDD Int Workshop Urban Comput. https://doi.org/10.1145/2346496.2346498

TRB (1996) Transit and urban form. http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_rpt_16-1.pdf

Tsai CH, Mulley C, and Clifton G (2012) The spatial interactions between public transport demand and land use characteristics in the Sydney greater metropolitan area. 35th Australasian Transport Research Forum, Perth, Australia

Upchurch C, Kuby M, Zoldak M, Barranda A (2004) Using GIS to generate mutually exclusive service areas linking travel on and off a network. J Transport Geograph 12(1):23–33. https://doi.org/10.1016/j.jtrangeo.2003.10.001

Waddell P (2002) Urbansim: modeling urban development for land use, transportation, and environmental planning. J Am Plann Assoc 68:297–314. https://doi.org/10.1080/01944360208976274

Wakamiya S, Lee R, Sumiya K (2011) Urban area characterization based on semantics of crowd activities in Twitter. Lecture Notes Comput Sci 6631:108–123. https://doi.org/10.1007/978-3-642-20630-6_7

Wang W, Attanucci JP, Wilson NHM (2011) Bus passenger origin-destination estimation and related analyses using automated data collection systems. J Public Transp 14 (4):131–150. https://doi.org/10.5038/2375-0901.14.4.7

Wong J (2013) Use of the general transit feed specification (GTFS) in transit performance measurement. MSc thesis, Georgia Institute of Technology

Yu C, He ZC (2017) Analysing the spatial-temporal characteristics of bus travel demand using the heat map. J Transp Geogr 58:247–255. https://doi.org/10.1016/j.jtrangeo.2016.11.009

Zhan X, Ukkusuri SV, Zhu F (2014) Inferring urban land use using large-scale social media check-in data. Netw Spat Econ 14:647–667. https://doi.org/10.1007/s11067-014-9264-4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.