Data Science for Transportation (2023) 5:10
https://doi.org/10.1007/s42421-023-00072-z

RESEARCH

Explaining and Predicting Station Demand Patterns Using Google Popular Times Data

Teethat Vongvanich[1] · Wenzhe Sun[1] · Jan-Dirk Schmöcker[1]

Received: 9 February 2023 / Revised: 18 May 2023 / Accepted: 13 June 2023 / Published online: 26 June 2023

Abstract
Google Popular Times (GPT) data are a novel data source that is open to the public, accessible in real time and available in many cities around the world. We aim to explain and predict travel demand patterns for train stations in Kyoto city with these data. Stepwise multiple linear regression models are developed using popularity data to analyze the correlation of the station demand patterns and point of interest (POI) visitation rates in the station vicinity. Our linear regression models aim to identify POIs and POI types that have the highest impact on the demand at each station. To predict station demand, we compared different machine learning models with the multiple linear regression model and concluded that the best prediction performance is obtained by Gradient Boosting. We were able to identify influential POIs and quantify their impacts given that there are a sufficient number of POIs in the vicinity of the station. Our findings suggest that GPT data can enable transit planners and transit users to predict station demand in real time. City planners would also gain valuable insights into the activity types highly related to transit station demand. Moreover, the method can be scaled and applied to other types of transit stations in other cities.

Keywords Transport demand modeling · Point of interest · Google Popular Times · Transit station · Trip purpose · Activity

Introduction

Public transport demand estimation and forecasting is an essential aspect of infrastructure planning and transport policy. Demand models estimate line loads as well as which stations are frequented by passengers. Two important questions travel demand models need to answer are why and when people visit each station. Both "why" and "when" will help the planner understand the effects of proposed service changes, as well as the impact of land-use changes on travel demand. In this research, we propose a new method of explaining and predicting station demand using crowd-sourced real-time demand information. We aim to explain what activities are related to visiting a train station. Related to this, we also aim to forecast a change in station demand if activity levels at surrounding "Points of Interest" (POIs) change.

We have two reasons to explore the usefulness of this crowd-sourced data. First, it is costly and time-consuming to obtain station demand distribution using the traditional survey data. Additionally, these surveys usually ask for trip purposes without details about the final destinations, so that inferring the importance of a particular infrastructure on transit demand remains difficult. Second, during events, such as natural disasters, human-made hazards, or viral pandemics, it is difficult to obtain firsthand information from transit users. However, especially in disruptive scenarios, it is crucial for transit demand models to adapt to sudden changes.

Our models are going to use Google Popular Times (GPT) data as the input. As will be explained, the data have a number of limitations but also advantages. For one, the data are accessible via webpages for a wide range of POIs. Furthermore, the data are available in real time. We therefore propose that GPT data can help address some travel demand modeling challenges by incorporating POI data into station prediction models. With this crowd-sourced data, travel

✉ Wenzhe Sun
wz.sun@trans.kuciv.kyoto-u.ac.jp

Teethat Vongvanich
vongvanich.teethat.68c@st.kyoto-u.ac.jp

Jan-Dirk Schmöcker
schmoecker@trans.kuciv.kyoto-u.ac.jp

[1] Department of Urban Management, Kyoto University, Kyoto, Japan

demand models can better capture the complex relationships between urban activity and transportation demand.

Following our motivations, the two main objectives of this study are to explain station demand and to predict station demand. To explain station demand, we aim to identify which POIs affect the busyness at each station, focusing on POIs that are within walking distance from the station. The other objective of this research is to predict the busyness at each station 1 or 2 h ahead in the future. Multiple linear regression models can provide a simple yet effective way of modeling the relationship between station demand and POI popularity. In addition, machine learning algorithms can offer more sophisticated models capable of handling complex non-linear relationships between the input and output variables.

The remainder of the paper is organized as follows. The next section reviews the various data types that have been used to model station demand, including previous research utilizing GPT data. The section "Google Popular Times (GPT) Data" describes the data collection process, specifically the use of Google API and GPT data. This section also includes statistics on the available data for our case study city, Kyoto. The section "Methodology" introduces our multiple linear regression models and machine learning models for explaining and predicting demand. In the "Results" section, we present and discuss our findings. Finally, we conclude the study in the section "Conclusion" with discussions on how the results can be helpful to transportation systems and offer suggestions for future work.

Literature Review

Data Types for Station Demand Modeling

Station demand has traditionally been estimated through household surveys, which have been carried out using various methods and generally at high costs (Miller 2014; Zhang and Mohammadian 2010). The data from these surveys can be extrapolated to estimate the number of travelers at each station. Additionally, as surveys often include a question regarding trip purposes, one can gain a good understanding of whether stations are predominantly used by commuters or leisure travelers.

To obtain data in a more cost-effective manner and increase the sample size, several alternative data collection methods have been recently developed, specifically utilizing various "passive data" methodologies (Sun and Schmöcker 2021).

There is now a substantial body of literature that explores various types of transit data, primarily electronic ticketing data, especially smart card data. These resulting databases store the daily trip information of bus or train passengers, providing valuable continuous data collection that offers a complete and real-time travel diary (Pelletier et al. 2011). Wu et al. (2019) discuss that smart card data can be used to monitor the overall transportation system state of a city. Kurauchi and Schmöcker (2017) provide an overview of the uses of these data, including route flow estimation (which in turn provides station demand information) and activity type estimation. However, since the data are limited to boarding and alighting, it is clear that detailed inference regarding the specific buildings or facilities visited by users between transit stations is not available (Longley et al. 2018).

For obtaining station crowding information only, a range of other data sources are available. Simple turnstile numbers often provide a good estimate if the station is used solely for travel purposes. Other researchers have utilized video records or Wi-Fi sensors to obtain station crowding information. Ryu et al. (2020) demonstrated that a Wi-Fi sensing system can estimate passengers' origin–destination demand and their waiting times at bus stops. Aggregate information from mobile phone providers can also be used to understand the level of crowding in "mesh areas" around stations. These data allow for differentiation based on the country where the mobile phone is registered (Ahas et al. 2008).

Obtaining more detailed records of where people come from and where they go after visiting a transit station requires disaggregate tracking data. This could come from call detail records (CDR) collected by mobile phone service providers for customer billing purposes. For example, Breyer et al. (2022) use CDR to learn the modes of interregional trips. However, these data have limitations, such as accuracy and uniformity issues arising from the market share of a particular telecommunication company (Willumsen 2021) or the degradation of usefulness due to privacy-related re-anonymization (Aguilera and Boutueil 2018).

More commonly, GPS tracking data are used. GPS data provide detailed digital traces with both location and time, allowing for the inference of station visits as well as activities before and after. It is worth noting that GPS datasets capture short or secondary trips to "small points of interest" (POIs), such as shops along the way, which are often omitted in trip diary-type surveys (Aschauer et al. 2018). However, accessing such data requires either convincing a major service provider to share the data, which is challenging, or asking a large enough group of people to download and provide data through a dedicated mobile application (Gao and Schmöcker 2021). Additionally, the sample of people who opt in to providing GPS data can be biased (Lue and Miller 2019; Nishigaki et al. 2023). Jee et al. (2022) collected "activity transition" points from individuals within an open-source transit planning application to create meaningful trip and tour records along with related activities.

For various reasons, the usage of other crowd-sourced data has been increasing. For example, Rajput and Chaturvedi (2019) used crowd-sourced accelerometer data to measure bus crowding. Luo et al. (2023) and Osorio-Arjona et al. (2021) utilized social media data to identify heterogeneities in public transport travel satisfaction and the spatial distribution of complaints. Gao and Schmöcker (2023) demonstrated that Wi-Fi sensors placed at POIs, in combination with a limited sample of GPS traces from a public transport planning app, can be used to infer travelers' routes.

In particular, social media data are increasingly employed to infer the activities of travelers, including public transport users (Bi et al. 2023). Social media platforms allow people to share their activity locations or "check-ins," announcing their location when they are at restaurants, shopping malls, movie theaters, and so on (Golder and Macy 2014; Tasse and Hong 2014). In addition to the location, the data also provide geo-tagged associated text data in the posts. It has been found that data from platforms such as Twitter can assist in studying highly dynamic and disruptive events, such as natural disasters (Efthymiou and Antoniou 2012). However, extracting daily and typical behavior from such data sources remains challenging, as few people report routine activities. Furthermore, processing social media data for demand estimation is cumbersome (Cramer et al. 2011; Maghrebi et al. 2015).

Methods for Station Demand Modeling

Activity-based models (ABMs) consider travel as a demand derived from the need to perform activities distributed in space, which provides advantages over more aggregated travel demand models when evaluating traffic management policies (Axhausen and Gärling 1992). Instead of analyzing individual trips, ABMs focus on the individuals who perform them and view trips as a result of individuals' desires to perform specific activities (Ortúzar and Willumsen 2011). The connection between trips and individuals makes ABMs more adaptable to station demand forecasting applications. However, ABMs require a full diary of activities for each user to represent the population of the study area, which is not always available (Bassolas et al. 2019).

Agent-based models were developed in the 2000s for large-scale microscopic traffic simulation, with TRANSIMS being one of the earliest approaches (Cetin et al. 2002). Nowadays, the open-source software MATSim (Horni et al. 2016) is commonly used in academic applications. MATSim combines supply and demand in a network equilibrium, where individual travelers (agents) start with pre-defined day plans, search for routes through networks, and adjust their travel mode and time-of-day choices. A national-scale model was developed since Balmer et al. (2008). While agent-based models can be used to estimate station demand, the focus of this research is to understand the relationship between stations and their surrounding environment, which is better suited to a regression analysis.

Regression analysis methods are fit for predicting future values of the dependent variable based on the values of the independent variables, but also for identifying important factors that influence the dependent variable. Different studies also use different regression models to construct the relationship between independent variables and the station-level ridership. Linear regression modeling is the most commonly used method (Liu et al. 2016; Sung et al. 2014), while the negative binomial regression is another popular model (Zhu et al. 2019). In recent years, non-linear models, such as machine learning models and polynomial statistical models, have been widely employed (Wang and Ross 2018; Ding et al. 2021).

GPT Data as a Potential Data Source

In contrast to the above crowd-sourced data, GPT uses aggregated and anonymized data from Google location histories to provide visitation data for various locations and commercial venues. Thus, the location-based service provides information about a wide range of geographic places.

Tafidis et al. (2017) used GPT data to predict traffic conditions, demonstrating its potential for estimating environmental impacts and traffic performance in specific areas. Capponi et al. (2019) used GPT data to predict categories of local businesses (e.g., bars and restaurants), their attractiveness, and temporal demand patterns. Möhring et al. (2020) conceptually and empirically demonstrate the practicality and value of GPT to better understand, analyze, and predict tourist consumer behavior. MacKenzie and Cho (2020) estimated the number of walking trips, vehicle miles traveled, and greenhouse gas emissions associated with traveling to dog parks using GPT data. Poom et al. (2020) utilized “mobile big data”, including GPT data, to examine the spatial effects of the COVID-19 pandemic, emphasizing the need for privacy considerations and transparent methodologies in assessing societal impact. Arnal et al. (2020) examined the behavior and interrelations of GPT data to understand the changes in human mobility during the COVID-19 pandemic, providing insights into policy efficiency and the "new normal" in Spain. Timokhin et al. (2020) investigated the possibility of using auxiliary information from Google Maps, Yelp, and OpenStreetMap to model venue popularity and occupancy, finding promising similarities between Wi-Fi-based ground truth data and GPT data in measuring venue popularity. Bandeira et al. (2020) explore the potential of using GPT as a crowdsourcing tool to predict traffic-related impacts, revealing clear relationships between GPT and traffic volumes, travel times, pollutant emissions, and noise levels in different areas and periods through linear

regression models. Mahajan et al. (2021) found a correlation between POI popularities and COVID-19 lockdowns in Munich. They also found that POI type and distance-to-station had a significant impact on POI popularity. Vitello et al. (2023) used GPT data and electronic ticketing data to predict ridership. The following aims to continue this line of literature by showing that the temporal and spatial profiles of GPT data can be used to show spatial correlation in activity patterns with a focus on transit stations.

Google Popular Times (GPT) Data

Description

The acquisition of the GPT data occurs by first accessing the Google API to retrieve information about the POIs in Kyoto. This general information includes the name, location, node type, address, rating score, and number of ratings of each POI. Among all the POIs obtained from the Google API, only a subset of them have GPT data available. To collect the GPT data for these POIs, we developed a script that accesses the public webpages of each POI every hour and downloads its GPT data.

GPT data are a collection of visitation data for various POIs on Google Maps. It is sourced from individuals who have chosen to share their location history through Google Location History, and the data are aggregated and anonymized (Google 2022). These data are primarily intended to assist users in planning their visits to businesses. However, it should be noted that not all POIs have GPT data available, as it requires a sufficient amount of visit data to be collected by Google.

GPT data has four main information for each POI:

(1) Popular times graph: The blue section of the graph in Fig. 1 is the popular times graph. It depicts how busy the POI usually is at different hours of a certain day. The data are based on average popularity over the last few months, which we will refer to as the "historical average". The popularity is given relative to the typical peak popularity of the business within a period of 1 week. In Fig. 1, we see that 8–9 pm on Saturdays are one of the more active hours of this business.
(2) Live visit data: The red bar of the graph in Fig. 1 is the live visit data, that is, how busy the location is now compared to its regular level of busyness.
(3) Stay duration: These data show how much time people usually spend at the POI. The estimates are based on patterns of visits over the past several weeks.
(4) Wait time estimates: These data show how much time a customer must wait before receiving service. It is only available for few locations and we do not collect this information.

It is important to note that GPT data provide a relative measure of the busyness at a POI. The popularity of a POI for a specific hour is presented relative to its typical peak popularity within a 1-week period. The data are scaled from 0 to 100, with 100 representing the typical peak popularity. It should be emphasized that live visit data, which is updated in real time, may exceed 100 if there is an exceptionally high level of activity at a given time.

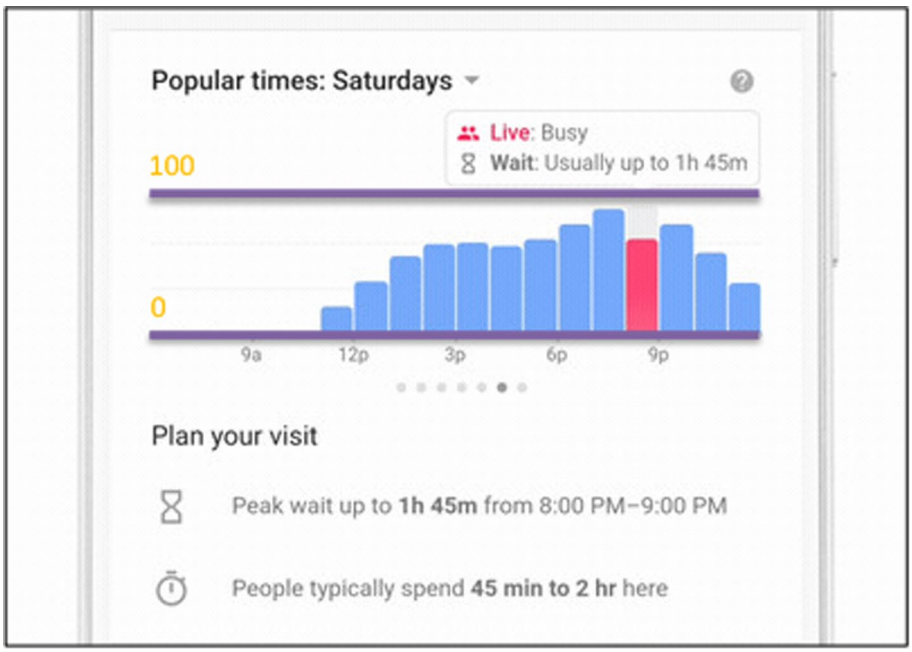


Fig. 1 GPT graph. Adapted from (Google 2022)

POIs in Consideration

As of April 2021, Google recognized 60,492 locations in Kyoto as Google POIs. Out of these Google POIs, 10,121 have GPT data available. Among the POIs with GPT data, we identified that 1524 POIs have live visitation data during the period of our data collection.

Our data collection for GPT data of POIs in Kyoto began on November 4, 2020, and extends until April 20, 2021. For the purpose of this study, we consider only the hours between 6 am and midnight, as stations in Kyoto do not provide services during the night. This time range provides us with a sample size of 3012 h.

Throughout our data collection period, we generate a CSV file containing GPT data for all POIs in Kyoto with available GPT data on an hourly basis. An example entry of GPT data is illustrated in Table 1. The "Historical average" represents the typical busyness of the POI at that specific hour of the week, while the "Live popularity" indicates the real-time visitation data for that hour.

Out of the 1524 POIs with live data, 64 of them are categorized as "train stations." The locations of these train stations are indicated in Fig. 2. In Kyoto, there are two local train lines that connect the central area to residential and tourist destinations in the outskirts: the Randen Tram Line and the Eizan Railway. These lines serve small local stations, and the trains and trams operating on these lines typically have a maximum of two carriages. On the other hand, there are additional train lines that connect Kyoto to other regions, namely the JR Japan Railway, Hankyu Railway, Keihan Railway, and Kintetsu Railway. The stations served by these lines are considerably larger in size and experience higher demand. It is worth noting that not all stations on a train line have GPT data available. While almost all stations on the Hankyu and Keihan Main lines have GPT data, only two stations on the Randen Tram Line have GPT data.

Table 1 An example of an entry GPT data of a POI in Kyoto

Date and time	November 05, 2020, 12:30:11
Name	National Museum of Modern Art, Kyoto
Latitude	35.0124
Longitude	135.782
POI type	Museum
Address	26-1 Okazaki Enshojicho, Sakyo Ward, Kyoto
Historical average	80
Live popularity	57
Rating score	4.1
Number of ratings	2686
Average time spent (min)	120

To provide an example of a larger station, Fig. 3 showcases Karasuma Station, which is one of the busiest stations in Kyoto. Additionally, it displays the POIs within an 800-m radius of the station.

POIs that are used in our models are POIs that have live GPT data and are within 800 m of the train stations. Table 2 provides a breakdown of the POIs under consideration, categorized into seven groups. The left columns indicate the groups and its count. The right columns show three POI types with the highest counts within each POI group.

Validation

Wi-Fi data are used as a source of ground truth to compare and validate the accuracy of the GPT-generated data. The Wi-Fi data are collected from sensors installed in specific locations throughout Kyoto city, including popular tourist attractions, central business areas, and the main railway station. The data collected from the sensors reflect the busyness of a location based on the number of probe requests sent by portable electronic devices for Wi-Fi access (Jee et al. 2021; Namulindwa 2023).

In Fig. 4a, we present the Live GPT data for Fushimi Inari Shrine, a famous tourist attraction, by hour. Figure 4b shows the Wi-Fi count data, which is an absolute measure of the number of mobile phone signals sent by the visitors at the site. In Fig. 4c, we illustrate the correlation between GPT and Wi-Fi data in different hours during the timeline. The overall correlation between the two datasets is fairly high and an R^2 of 0.82 is obtained. This validation process helps to increase confidence in the accuracy of GPT data and provides a means to address any biases or discrepancies that may arise. We note that we do not have any information to validate the GPT "stay duration" information, we can only observe that the estimates appear to be in line with the authors' experiences.

Methodology

Overview and Notation

Two linear regression models are proposed to explain travel demand: "Station live popularity" model and "Difference to historical-average" model. Further, for the case of prediction, five kinds of machine learning models are tested. Here, deriving explainable coefficients is not our main concern, but we show that these models can lead to higher model fit and prediction accuracy.

The notations used in this section are shown in Table 3, grouped by models.

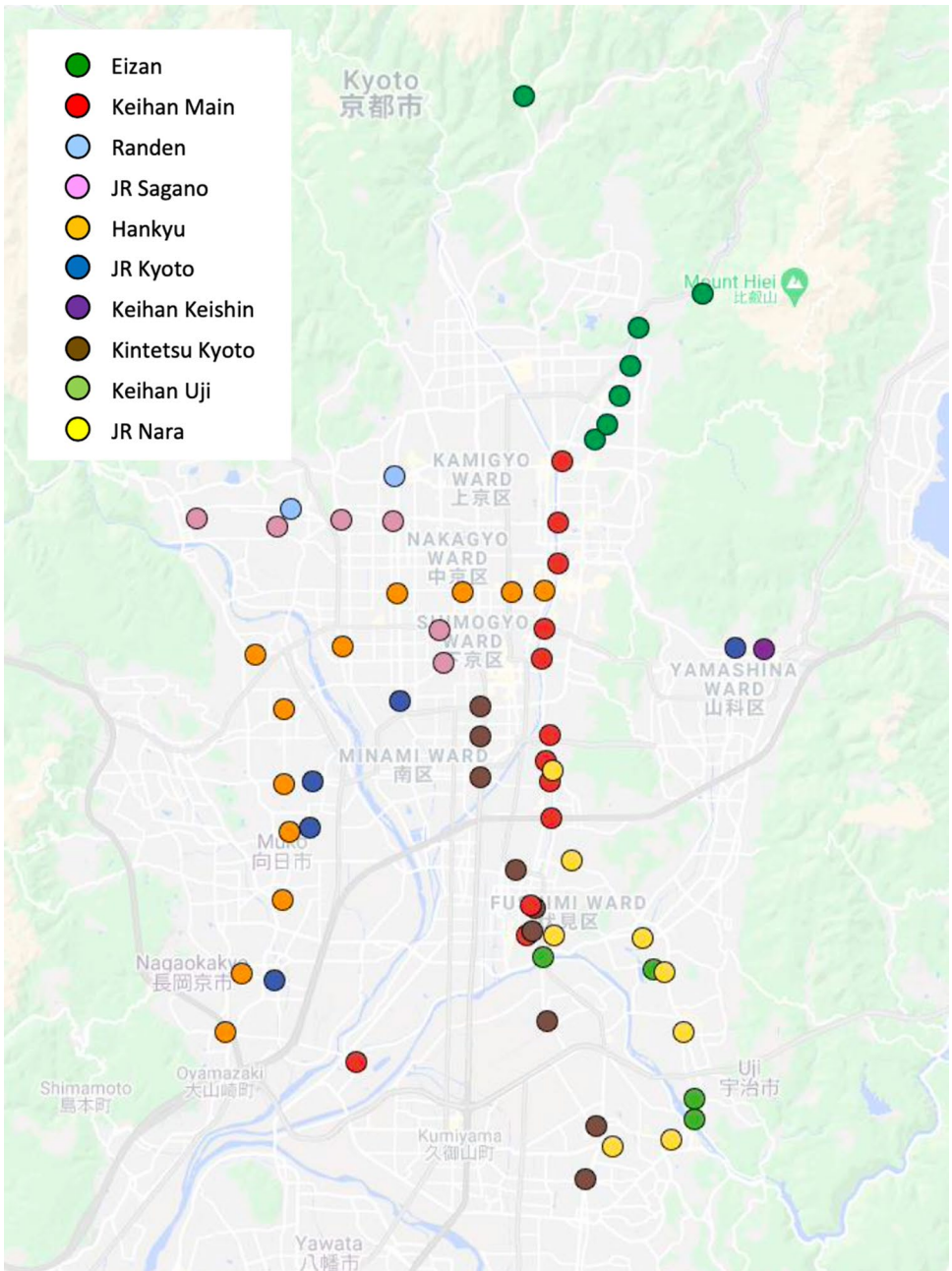


Fig. 2 Train stations in Kyoto with GPT data; colors denote different railway lines

Models to Explain Travel Demand

Clearly, not all station demand can be attributed to the set of POIs for which we have obtained GPT data. Home locations, for instance, are not considered as POIs, which means that we are likely to miss the start or end of commuting trips in our dataset. Furthermore, some individuals may not visit any POI or visit a POI significant time before or after their station visit, making it challenging to establish a direct link between station usage and POI visits. It is crucial to keep these limitations in mind when interpreting our subsequent models.

Even if the GPT data were presented as absolute numbers, a strong correlation between station busyness and POI busyness cannot be directly translated into the number of visits to that POI by transit users. It merely indicates the likelihood of people visiting that particular POI. Given this constraint, we will now describe the process of model creation.

First, we define a bounding circle around each station. Assuming that the individuals using the station are likely to have activities in the vicinity, the threshold distance for including POIs should not be too large. In this study, we consider a walking distance of 800 m around each station. POIs located within 800 m of the station are considered as

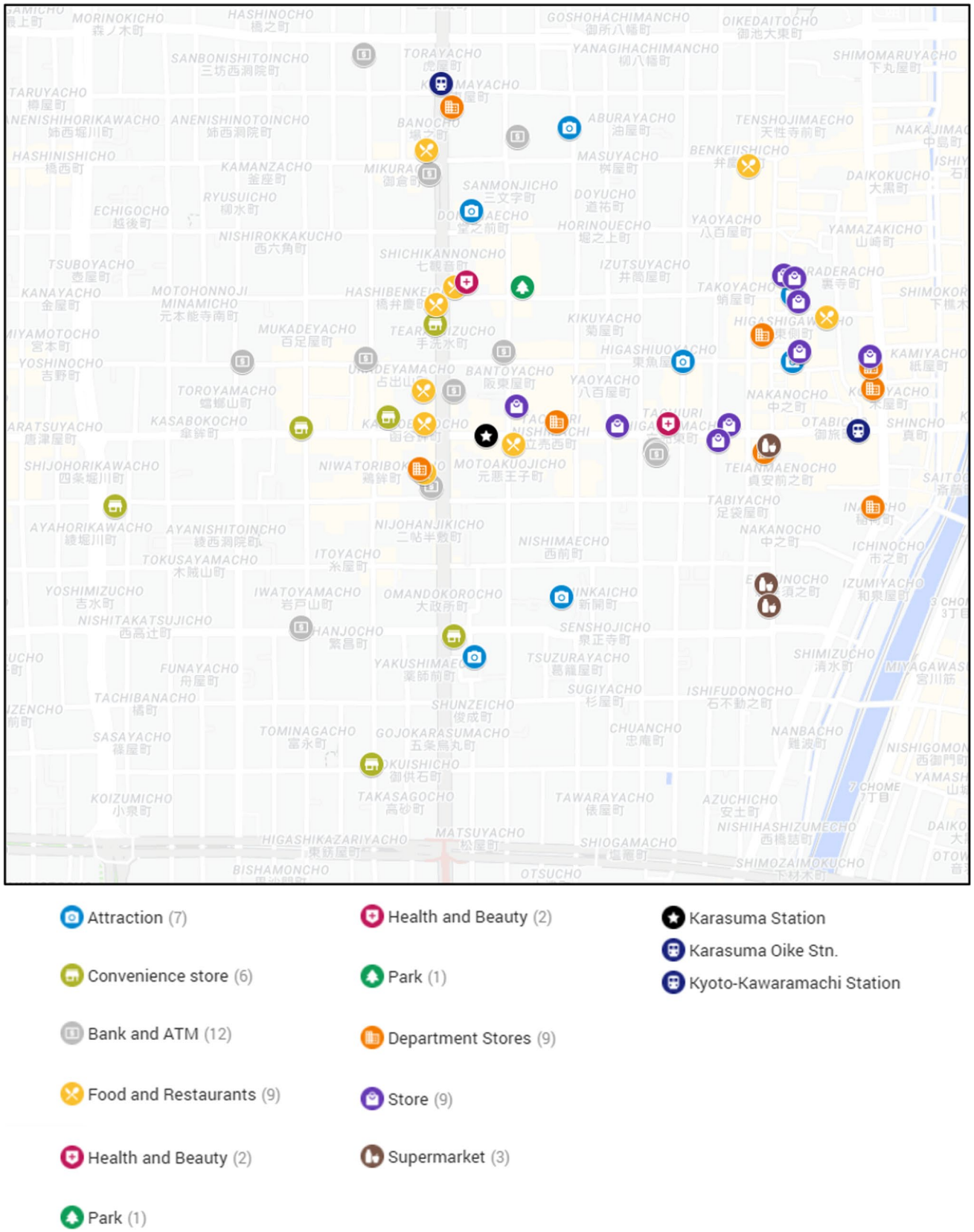


Fig. 3 POIs with GPT data that are within 800 m of Karasuma Station

Table 2 Distribution of POI types

POI group	Count	Top 3 POI types in group	Count
Shopping	163	Supermarket	51
		Book store	21
		Shopping mall	21
Food	99	Restaurant	38
		Café	33
		Meal takeaway	19
Tourist	62	Tourist Attraction	37
		Park	18
		Museum	7
Transit	47	Train station	40
		Subway station	6
		Taxi stand	1
Public facility	27	City hall	8
		Library	6
		Bank	5
Local service	22	Gym	18
		Hair care	2
		Beauty salon	1
Entertainment	5	Amusement park	2
		Bowling alley	2
		Aquarium	1

POIs within the station's range. We have tested other thresholds, but found that 800 m provide the most reasonable and generally best-fitting results. The haversine formula, which calculates distances on a sphere, is used to determine which POIs fall within this range.

Second, after identifying the POIs within the station's range, we collect data on live popularity, historical popularity, and the duration of time people spend at each POI. The stay duration l_i for each POI i is defined as the average amount of time people spend at that specific POI. We introduce a minimum stay duration to reduce the number of "secondary" POIs in our explanatory models. "Primary" POIs are considered the main destinations that people plan to visit on their trips, representing the purpose of their trip. Conversely, "secondary" POIs are visited by people as additional stops or "add-ons" during their trip and are not considered the main purpose of their journey. For example, after visiting the zoo, a traveler might purchase drinks from a convenience store located near the zoo. In this scenario, the zoo is considered a primary POI, representing the main purpose of the trip, while the convenience store is classified as a secondary POI. In transport planning scenarios, our aim is to identify the primary POIs of individuals.

To account for multicollinearity, a phenomenon where predictor variables in a regression model are highly correlated, we employed stepwise multiple linear regression for our explaining models. Stepwise regression is a technique that iteratively selects the most statistically significant predictor variables while considering both forward and backward steps. It helps mitigate the impact of multicollinearity by including only the most relevant variables in the final model. In our study, we utilized stepwise regression

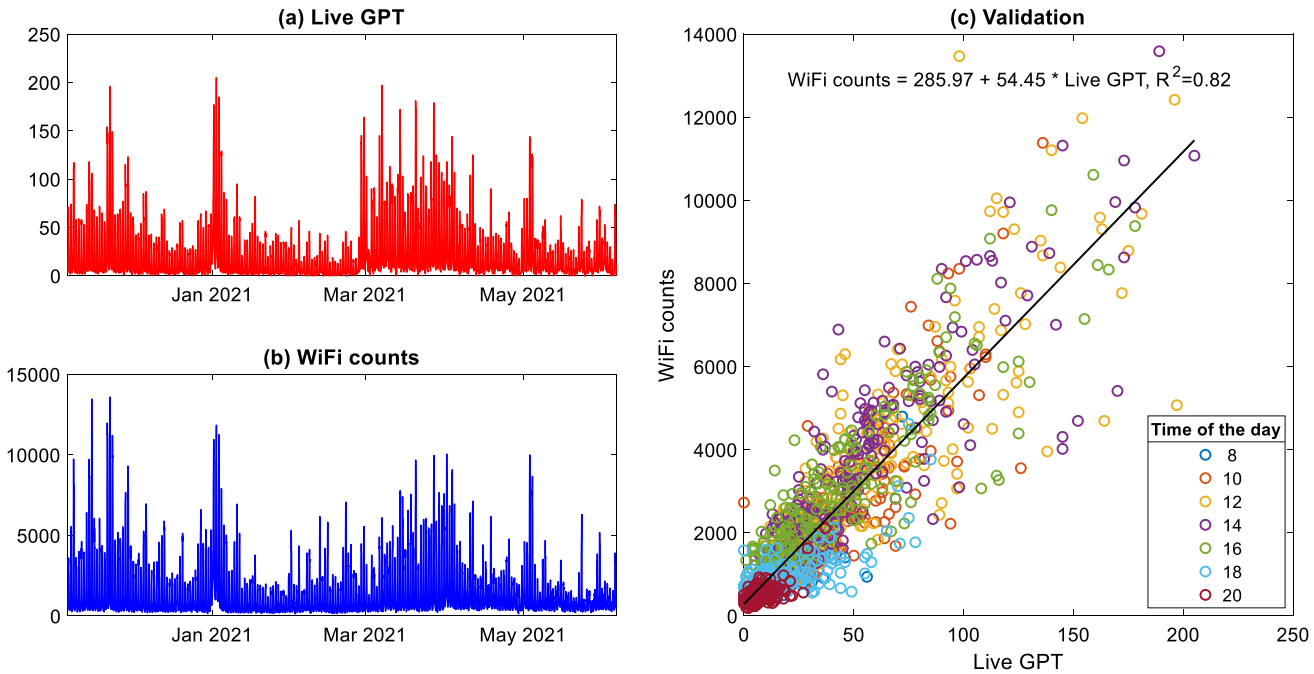


Fig. 4 Validation of live GPT data with Wi-Fi counts

Table 3 Notations used for regression models

"Station live popularity" model in the section "Models to Explain Travel Demand" and "Real-time prediction" models in the section "Models to Predict Travel Demand"	
$y_s(t)$	Live popularity of station s at time t
$y_s'(t)$	Historical popularity of station s at time t
$x_i(t)$	Live popularity of POI i at time t
β_i	Regression coefficient of POI i
α_s	Regression coefficient of live popularity of station s
α_s'	Regression coefficient of historical popularity of station s
l_i	Stay duration for POI i
l_M'	Prediction horizon of M minutes
"Difference to historical-average" model in the section "Models to Explain Travel Demand"	
$\widetilde{y}_s(t)$	Difference between live and historical popularity of station s at time t
$\widetilde{x}_i(t)$	Difference between live and historical popularity of POI i at time t
$\widetilde{\beta}_i$	Regression coefficient of POI i
l_i	Stay duration for POI i

to identify the subset of explanatory variables that have the strongest relationship with station demand, while excluding variables that introduce multicollinearity.

"Station Live Popularity" Model

This first model examines the relationship between the popularity of the station and the popularity of POIs near the station. In this model, the explained variable $y_s(t)$ is the live popularity data of the station at time t, while the explanatory variable $x_i(t)$ is the live popularity data of POI i that is within range of the station s at time t

$$y_s(t) = C + \sum\nolimits_{i \in S} \beta_i \big(x_i(t)\big) + \varepsilon_s(t). \tag{1}$$

The influence of the POIs will hence be represented by the set of regression coefficients β. For each of the 64 train stations in Kyoto with GPT data, a separate model is constructed. We note that we further tested models combining all stations and obtaining coefficients for POI categories such as "entertainment place", but omit these results here for brevity. Generally, the results of those models were reasonable but do not explain the variation in station busyness well.

In our regression analysis, we considered minimum stay durations l_i of 0, 15, 30, and 45 min. As the minimum stay duration increase, POIs with shorter durations, such as convenience stores, are removed from the regression models. Consequently, the model fit decreases as there are fewer explanatory variables. However, the remaining POIs are more likely to be the primary destinations of the trip, as they are the ones where people typically spend more time.

"Difference to Historical-Average" Model

The difference to historical-average model observes and predicts unusual changes in demand. Like the station live popularity model, it focuses on each station and the POIs around the station. However, the explained variable $\widetilde{y}_s(t)$ is the difference between the live popularity and the historical average of the station at time t. The explanatory variable $\widetilde{x}_i(t)$ is the difference between the live popularity and the historical average of POI i that is within range of the station s at time t

$$\widetilde{y}_s(t) = C + \sum\nolimits_{i \in S} \widetilde{\beta}_i \big(\widetilde{x}_i(t)\big) + \varepsilon_s(t). \tag{2}$$

The difference to historical-average model focuses on capturing demand abnormalities at the POIs and stations, where abnormalities refer to any deviation from the usual popularity. For instance, festivals and events can attract large crowds to POIs that are typically less busy during that time of day. Conversely, due to concerns about COVID-19, people may avoid going to supermarkets during peak hours. These examples illustrate how differences between live data and historical GPT information can be explained.

Similar to the previous model, we have 64 cases for the difference to historical-average model, corresponding to the 64 train stations in Kyoto with GPT data. Additionally, we will conduct regression analysis with minimum stay durations l_i of 0, 15, 30, and 45 min to minimize the inclusion of secondary POIs in the model.

Models to Predict Travel Demand

Overview and Linear Regression Model

In this study, six models were used to predict travel demand. These models included one linear regression model and five machine learning models. The data were split into training and testing sets using a ratio of 70:30.

In all models, the explained variable $y_s(t)$ is the popularity of the station at time t. The explanatory variables are the POI popularities; they are taken at time $t - l_M$, where l_M is the prediction horizon which is constant throughout the POIs. This model aims to predict the busyness at the station with the most up-to-date data available. The prediction horizon l_M represents the time into the future the model aims to predict.

The explanatory variables in all models are shown in Eq. 3 and include: (1) $x_i(t - l_M)$ the live popularity of POI i at time$t - l_M$, (2) $y_s(t - l_M)$ the live popularity of the station

at time $t - l_M$,and (3) $y'_s(t)$ the historical popularity of the station at time$t$.

We investigate the prediction capabilities when the prediction horizon l_M is 30, 60, 90, 120, 150, and 180 min. If the GPT reflect activities, then clearly we expect to see lower R^2 the longer the prediction horizon.

The resulting linear regression model is shown in Eq. (3)

$$y_s(t) = C + \sum_{i \in S} \beta_i \left(x_i\left(t - l_M\right)\right) + \alpha_s\left(y_s\left(t - l_M\right)\right) + \alpha'_s\left(y'_s(t)\right) + \varepsilon_s(t). \quad (3)$$

Machine Learning Models

Machine learning models are more flexible, can handle nonlinear relationships, manage categorical variables without encoding them, and handle high-dimensional data and missing values. We test the following common models:

Decision tree is a supervised machine learning model that is used for both classification and regression tasks. It uses a tree-like structure, where each internal node represents a feature or attribute of the data, each branch represents a decision based on the value of that attribute, and each leaf node represents a prediction or outcome.

Random Forest is an ensemble method that builds multiple decision trees and combines their predictions to improve the overall performance. It is known for its ability to handle high-dimensional data and to prevent overfitting.

Extra Trees, also known as Extremely Randomized Trees, are similar to Random Forest but with more randomness in the splitting process. It is less computationally expensive than Random Forest and often works well with high-dimensional data.

AdaBoost is an algorithm that combines multiple weak learners to create a strong ensemble model. It is known for its ability to handle a large number of features and to work well with data that contains noise or outliers.

Gradient Boosting is an ensemble method that builds multiple decision trees and combines their predictions to improve the overall performance. It is known for its ability to handle high-dimensional data, to prevent overfitting and also to handle missing values.

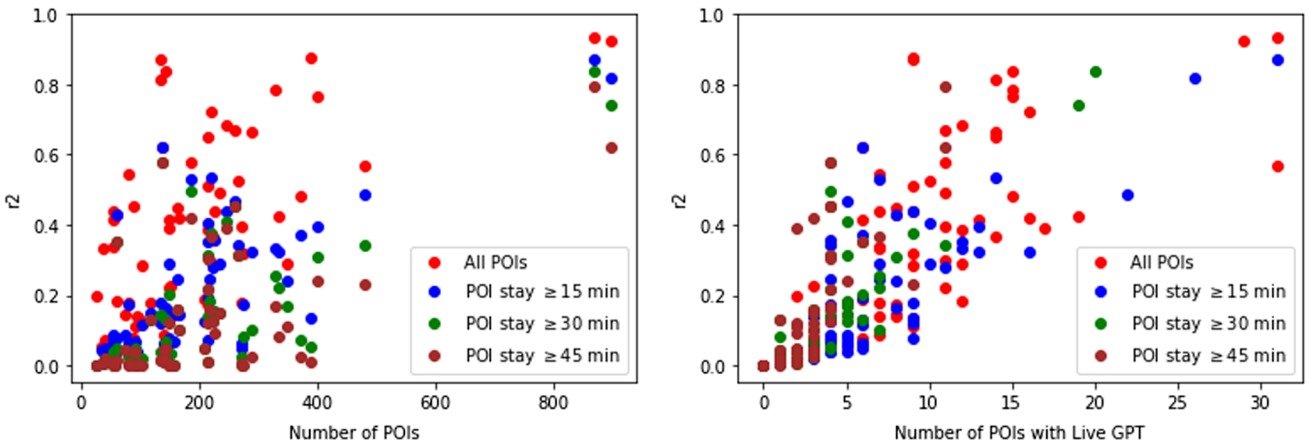
Results

Explaining Travel Demand

Station Live Popularity Model

The adjusted R^2 of the regression model of all 64 stations with different minimum stay durations are shown in Fig. 5a and b. For simplicity, we will refer to the adjusted R^2 as R^2. The left figure displays all POIs within the vicinity on the x-axis, while the right figure shows the number of available GPT data points for a subset of these POIs. Although some differences can be observed, there is a general correlation between points with a larger total number of POIs and a higher number of POIs with live GPT information. This correlation suggests that the GPT data reasonably reflect the overall distribution of POIs in the area. The absence of data points below the 45-degree diagonal line in the right-hand figure further demonstrates how the introduction of the minimum stay duration constraint impacts the model fit by reducing the number of data points.

In general, both Fig. 5a and b indicate that stations with fewer POIs have lower R^2 values. This implies that our linear



(a) R^2 against number of POIs existing within walking distance

(b) R^2 against number of POIs used in the regression

Fig. 5 The relationship between model fit and number of nearby POIs: station live popularity model

regression model is not able to effectively explain the station demand at stations with a limited number of POIs. Conversely, stations with a higher number of POIs tend to yield higher R^2 values. Increasing the minimum stay duration filters out more POIs, resulting in a smaller but potentially more meaningful subset of POIs in the regression model, albeit at the cost of reducing the R^2 values.

Tables 4 and 5 provide a detailed examination of two stations that have a low number of POIs but a high R^2 value. The first example is Kyoto-Kawaramachi station, which maintains a high R^2 value across different minimum stay durations, despite a decrease in the number of POIs included in the model. On the other hand, Nagaokatenjin station exhibits a high R^2 value in only one of the cases. To shed light on these differences, we delve into the analysis of the top three most influential POIs in each scenario.

In the case of Kyoto-Kawaramachi station, the three POIs with the highest regression coefficients represent places where people are likely to spend time before using the station. Pontocho Alley is a popular tourist area known for its numerous restaurants. "Round One" is a multi-story entertainment center offering various recreational activities. "Good Nature Station" is a shopping mall located near the station. These POIs effectively capture the range of activities that people engage in before arriving at Kyoto-Kawaramachi station.

On the other hand, for Nagaokatenjin station (Table 6), we observe varying magnitudes of coefficients. Unlike Kyoto-Kawaramachi station, Nagaokatenjin station only exhibits a high R^2 value when the minimum stay duration is set to 0 min. The most influential POI for Nagaokatenjin station is Nagaokakyo station, which is a nearby train station on a different train line. Many individuals transfer between these two stations, resulting in a strong correlation between their levels of busyness. However, since there are no time-spent data available for train stations, the influence of "Nagaokakyo Station" is disregarded when a minimum stay duration is introduced. Therefore, we cannot draw meaningful conclusions about the specific activities that individuals engage in upon disembarking at Nagaokatenjin station.

Table 4 Station live popularity model: number of POIs and R^2 of Kyoto-Kawaramachi station and Nagaokatenjin station

Stations	Minimum stay duration (minutes)			
	0	15	30	45
Kyoto-Kawaramachi Station				
Number of POIs in the model	31	31	20	11
R^2	0.933	0.870	0.836	0.797
Nagaokatenjin Station				
Number of POIs in the model	9	9	4	2
R^2	0.877	0.135	0.051	0.010

Table 5 POIs with highest regression coefficient for the Kyoto-Kawaramachi station live popularity model with a minimum stay duration of 45 min

POI	POI type	β_i	t value
Pontocho Alley	Attraction	0.263	24.41
Round One	Arcade center	0.237	19.93
Good Nature Station	Department store	0.211	17.20

Results of "Difference to Historical-Average" Model

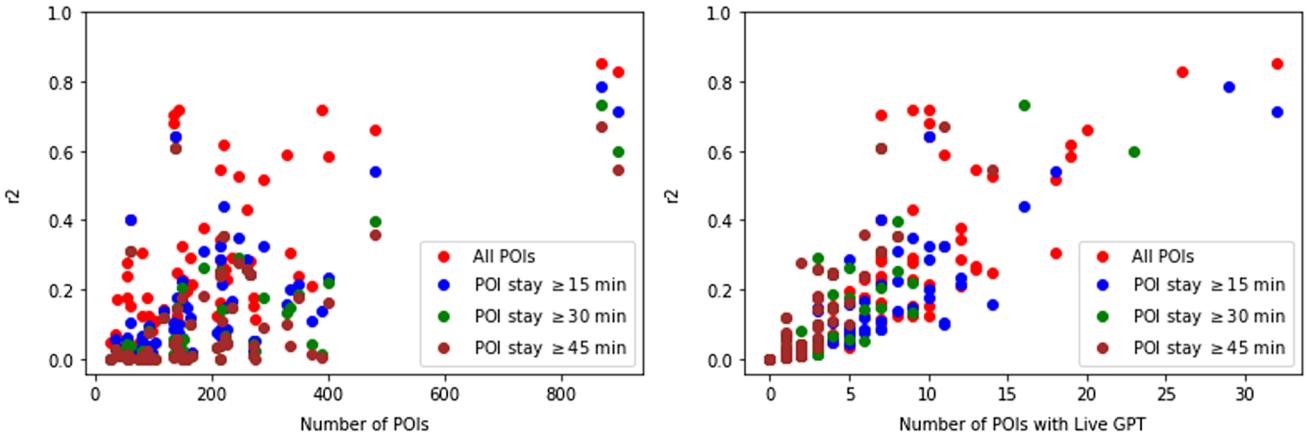
Figure 6 displays the R^2 values of the regression model for all 64 stations with various minimum stay durations. Consistent with the previous analysis, stations with a limited number of POIs tend to exhibit lower R^2 values, while stations with a higher number of POIs tend to have higher R^2 values. This pattern suggests that the model's ability to explain station demand improves as the number of relevant POIs increases.

Table 7 highlights Saga-Arashiyama station as an example of a station that possesses a relatively low number of POIs with live GPT data but demonstrates a comparatively high R^2 value. Despite the limited availability of GPT data, the model still manages to capture a significant portion of the station's demand variation.

In this model, the regression coefficient $\widetilde{\beta}_i$ represents how much of an impact changes in popularity of POI i has on changes in the popularity of the station. Table 8 provides insights into the regression coefficients for Saga-Arashiyama station, highlighting the top three POIs with the highest coefficients. The most influential POI is the Arashiyama Rilakkuma Tea House, a popular tourist cafe located in the bustling Arashiyama shopping strip. Following closely is Tenryuji temple, a prominent tourist destination in Arashiyama known for its cultural and historical significance. Finally, Arashiyama Park Nakanoshima area, another picturesque and tourist-friendly location, ranks third on the list.

Table 6 POIs with highest regression coefficient for the Nagaokatenjin Station live popularity model with a minimum stay duration of 0 min

POI	POI type	β_i	t value
Nagaokakyo Station	Station	0.699	137.85
Seiyu Nagaoka	Supermarket	0.080	8.90
Bank of Kyoto Nagaoka Branch Office	Bank	0.063	5.50



(a) R^2 against number of POIs existing within the walking distance

(b) R^2 against number of POIs used in the regression

Fig. 6 The relationship between model fit and number of nearby POIs: difference to historical-average model

Table 7 Difference to historical-average model: number of POIs and R^2 of Saga-Arashiyama station

Saga-Arashiyama Station	Minimum stay duration (minutes)			
	0	15	30	45
Number of POIs in the model	10	10	7	7
R^2	0.644	0.644	0.610	0.610

Table 8 POIs with highest regression coefficient for the Saga-Arashiyama Station difference to historical-average model with a minimum stay duration of 45 min

POI	POI type	$\widetilde{\beta}_i$	*t* value
Arashiyama Rilakkuma Tea House	Café	0.362	15.63
Tenryuji Temple	Attraction	0.248	20.50
Arashiyama Park	Park	0.171	17.18

These findings align with expectations, as these POIs attract significant visitation during festivals or seasonal tourist activities, which are often facilitated by the train service. The coefficients indicate the degree to which changes in the popularity of these specific POIs impact the overall popularity of Saga-Arashiyama station, providing valuable insights into the relationship between POI popularity and station demand.

Predicting Travel Demand

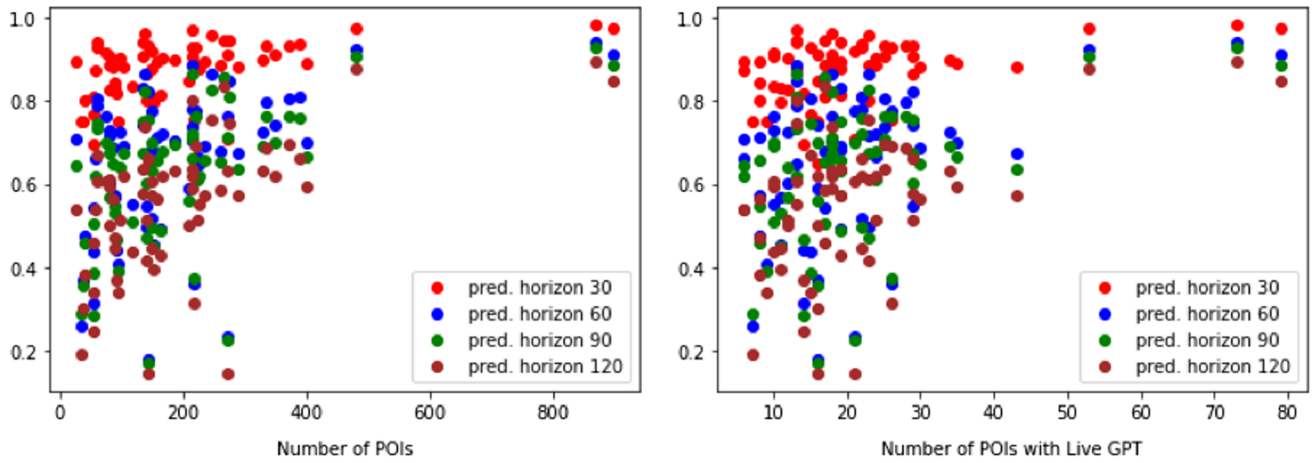
Results of Multiple Linear Regression

Figure 7 presents the results obtained with different prediction horizons for linear regression. The plots align with the previous section, with the minimum stay time now being replaced by the prediction horizon. Each station maintains a constant number of POIs for all four models, resulting in four vertical data points for each station in Figure (b). As anticipated, larger prediction horizons lead to lower R^2 values. The trade-off is that prediction accuracy decreases as predictions are made further into the future.

In general, most stations demonstrate high R^2 values, even when they have only a few POIs with live GPT data. However, it is important to highlight a case of "prediction failure" illustrated in Tables 9, 10. Keihan-Yamashina station possesses a substantial amount of POI data but exhibits a low R^2 value, indicating that the linear regression model fails to accurately predict station demand in this particular scenario.

In most stations, the most influential variables would be that of the station itself at time $t - l_M$ or its historical data at time t. However, in the case of Keihan-Yamashina station, the largest coefficient corresponds to Yamashina Station, which is a nearby train station on a different train line. This indicates that for stations that serve as transfer points from nearby stations, predicting demand further into the future becomes more challenging.

Unlike other stations where predictions are primarily derived from the station's own historical data and live popularity, stations with significant transfer activity require additional consideration of the popularity and demand at neighboring stations. The reliance on transfer passengers and their associated activities introduces complexities and



(a) R^2 against number of POIs existing within the walking distance

(b) R^2 against number of POIs used in the regression

Fig. 7 The relationship between R^2 and number of nearby POIs: multiple linear regression

Table 9 Multiple linear regression model: number of POIs and R^2 of Keihan-Yamashina station

Keihan-Yamashina Station	Prediction horizon (minutes)			
	30	60	90	120
Number of POIs in the model	21			
R^2	0.710	0.235	0.225	0.147

Table 10 POIs with highest regression coefficient for the Keihan-Yamashina Station prediction model with a minimum stay duration of 60 min

Explanatory variable	POI type	β_i	t value
Yamashina station	Train station	0.383	8.83
Historical data	Train station	0.322	8.78
Keihan-Yamashina station	Train station	0.308	17.45

uncertainties into the prediction process, making it more difficult to accurately forecast demand for an extended period.

Improvements by Machine Learning Models

In Fig. 8, the 50th percentile of the results from all 64 stations is displayed, comparing the multiple linear regression model with the five introduced machine learning models across different prediction horizons from 0 to 180 min. The figures illustrate a consistent downward trend in model fit as the prediction horizon increases, but the application of machine learning approaches can significantly mitigate this trend.

In Fig. 8a and c, it is evident that some machine learning models exhibit significant overfitting when applied to test data, despite performing well on the training data. For instance, AdaBoost demonstrates high training R^2 values close to 1 for all prediction horizons, but the testing R^2 gradually decreases from 1 to 0.6. Similarly, Random Forest, Extra Tree, and Decision Tree models also display overfitting tendencies, and varying hyperparameters did not effectively address this issue. In contrast, Gradient Boosting exhibits relatively less overfitting.

The overfitting phenomenon can be attributed to the nature of ensemble methods, such as Decision Tree, Random Forest, and Extra Trees, which utilize decision trees as their base learners. While these methods aim to mitigate overfitting by averaging the predictions of multiple trees instead of relying on a single decision tree, they still show signs of overfitting in this scenario. On the other hand, Gradient Boosting and AdaBoost are boosting algorithms that enhance the performance of weak learners by combining them in different ways. Compared to other machine learning models, Gradient Boosting demonstrates less susceptibility to overfitting due to its ability to adjust the weights of each feature. Consequently, we select Gradient Boosting as the best-performing machine learning model and further investigate its parameter settings.

Upon examining the hyperparameter "the number of estimators", we find that values larger than 200 do not result in a decrease in the test mean squared error. Therefore, we choose this value as the optimal parameter setting for the Real-time prediction model for travel demand.

Figure 9 provides insights into the relationship between R^2 values and the number of POIs. When predicting 60 min into the future (Fig. 9a), both the Multiple Linear Regression

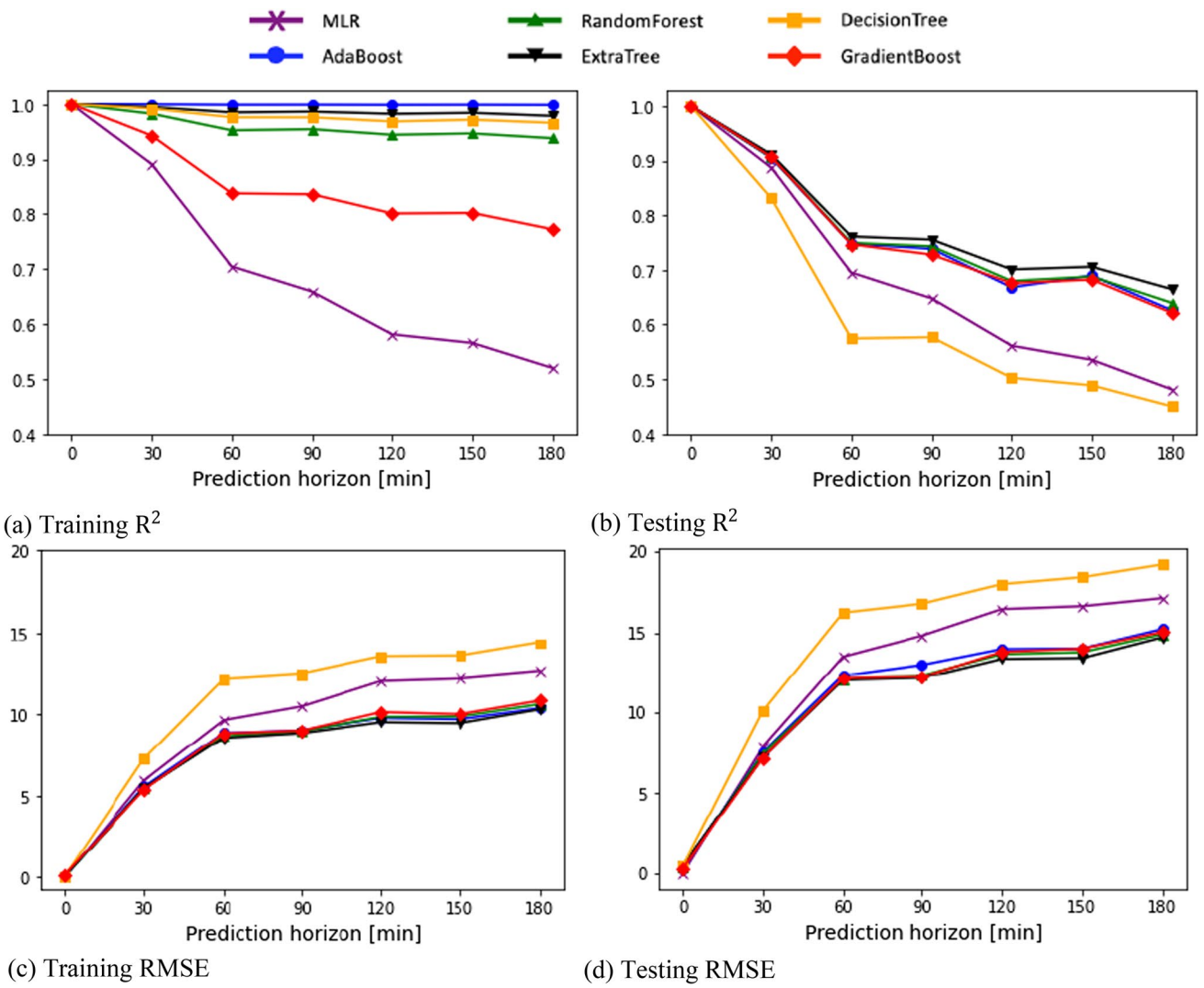


Fig. 8 Real-time prediction results from the six ML models

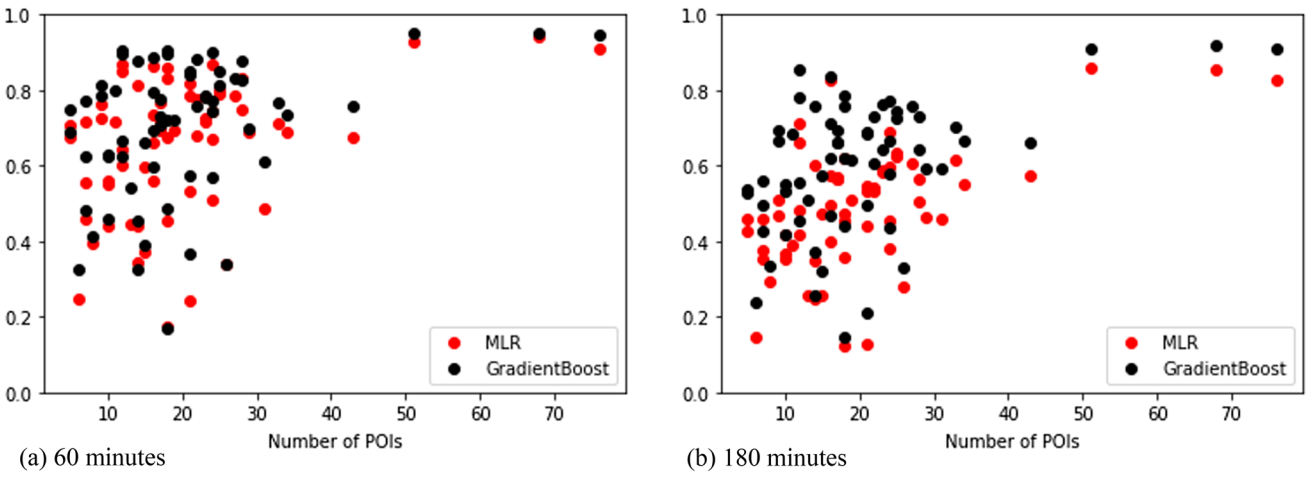


Fig. 9 The relationship between R^2 and number of nearby POIs under different prediction horizons: comparing multiple linear regression and gradient boosting models

(MLR) and Gradient Boosting models exhibit similar model fit, with Gradient Boosting slightly outperforming MLR. However, when extending the prediction horizon to 180 min (Fig. 9b), the MLR models do not perform as well, yielding R^2 values of approximately 0.5. On the other hand, the machine learning model, particularly Gradient Boosting, maintains a higher level of prediction accuracy, with an R^2 of approximately 0.7.

Notably, the machine learning model significantly improves predictions for stations with low MLR model fit. Upon further investigation, it was discovered that these stations exhibited irregular patterns in live and historical station demand. Unlike most stations, these underperforming stations often lacked typical infrastructure or had unique characteristics. This discrepancy likely affects the reliability of GPT data, as it becomes more challenging to determine whether individuals are present within the station or simply passing by in close proximity. For instance, one station had terminals situated on opposite sides of a large road, while others were smaller stations or lacked a roof, resulting in demand variations influenced by weather conditions.

Conclusion

Our motivation for this research was to develop models that can explain and predict station demand using GPT data. The station live popularity model and the difference to historical-average model aim to explain station demand. The station live popularity model identifies the POIs that have the highest influence on the demand of the stations. The difference to historical-average model observes and captures unusual changes in demand in relation to the surrounding POIs. Finally, the real-time prediction models focus on predicting busyness at stations.

GPT data are publicly available and accessible in real time, not only for our studied city but also for many other cities worldwide. The models proposed in this study can be applied and scaled to other cities. While our focus was on "train stations" to study station demand patterns, the models can also be applied to other types of POIs for demand explanation and prediction. For instance, one could explore how a specific activity type in an area relates to other activities and transit usage.

As expected, we found that models are most accurate at stations with a high number of nearby POIs. We believe that this correlation is not merely spurious, as even in cases with multiple POIs as explanatory variables, the types of POIs with large coefficients generally reflect the range of activities carried out near those stations. The models can be useful for short-term real-time prediction of unusual demand, as they often yield good model fit. However, the real-time prediction models struggle with stations that have irregular GPT data, which we suspect may be due to the stations' unusual infrastructures. Among the machine learning models, Gradient Boosting showed the best results and outperformed multiple regression significantly.

There are several challenges that need to be overcome for practical implementation of these models. Our explanatory models can only capture general activities that users engage in before or after visiting the station. It is reasonable to assume a high proportion of people visiting very large POIs, such as major department stores near a station, directly before or after their journey. However, the popularity presented in GPT data is not absolute, and we cannot obtain the actual number of POI visitations from this dataset alone. All regression results have been relative to the stations' peak busyness.

It is worth noting that since August 2021, Google has imposed a limit of 100 on live visit data. This constraint will also impact future work that aims to use live popularity as a measure of busyness at POIs. The data used in this study spanned from November 2020 to April 2021, allowing us to capture more variance in relative busyness, including some readings with live popularity exceeding 100.

All regression models introduced in this study represent mappings of the complete response of station time series with the complete time series of the POIs. However, not all POIs are active during the entire operating hours of the train stations. For example, bakeries are active in the morning, while bars and some restaurants are only active at night. Therefore, even if a certain bakery has a high influence in the morning, its regression coefficient is averaged throughout the entire day. To address this issue, one possible solution is to classify POI types and create regression models for different time periods or incorporate time-of-day dummy variables into the existing models.

Future research should expand the sampling to include other cities, countries, and various types of stations. This expansion would provide further insights into geographic constraints and their influence on station demand. Additionally, there are many aspects of the Google API and GPT data that were not incorporated into our methodology. For example, we attempted to reduce the number of secondary POIs by introducing the minimum stay duration. Another approach to address secondary POIs is to use a rating score and number of ratings to measure the attractiveness of each POI. With an additional source of data such as Wi-Fi or mobile phone data, we could potentially convert the relative popularity of GPT data into absolute visitation data.

Author Contributions TV: Content design, data analysis, and manuscript writing and editing. WS: Content design, data analysis, and manuscript writing and editing. J-DS: Content design, data analysis, and manuscript writing and editing.

Funding This work was supported by the EIG CONCERT-Japan DARUMA project, Grant No. JPMJSC20C4 funded by JST SICORP (Japan Science and Technology Agency), Japan.

Availability of Data and Materials The authors do not have the permission to share the data.

Declarations

Conflict of Interest The authors declared that there is no conflict of interest.

References

Aguilera A, Boutueil V (2018) Using cell phone data to understand travel behavior and transportation. Urban mobility and the smartphone: transportation. Elsevier

Ahas R, Aasa AR, Mark U, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: an Estonian case study. Tour Manage 29(6):469–486

Arnal RP, Conesa D, Alvarez-Napagao S, Suzumura T, Català M, Alvarez E, Garcia-Gasulla D (2020) Private sources of mobility data under COVID-19. arXiv preprint arXiv:2007.07095

Aschauer F, Hossinger R, Axhausen K, Schmid B, Gerike R (2018) Implication of survey mothods on travel and non-travel activities: a comparison of the Austrian national travel survey and an innovative mobility-activity-expenditure diary (MAED). Transp Infrastruct Res 18(1):4–35

Axhausen K, Gärling T (1992) Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transp Rev 12(4):323–341

Balmer M, Meister K, Rieser M, Nagel K, Axhausen KW (2008) Agent-based simulation of travel demand: structure and computational performance of MATSim-T. Arbeitsberichte Verkehrs-und Raumplanung, 504

Bandeira JM, Tafidis P, Macedo E, Teixeira J, Bahmankhah B, Guarnaccia C, Coelho MC (2020) Exploring the potential of web based information of business popularity for supporting sustainable traffic management. Transp Telecommun J 21(1):47–60

Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X (2019) Hierarchical organization of urban mobility and its connection with city livability. Nat Commun 10:4817

Bi M, Sun W, Schmöcker J-D, Ma Y, Moya-Gomez B, Nakao S, Yamada T (2023) Using geo-tagged tweets for understanding temporal and spatial activity distribution in Kyoto. In: 15th international conference of Eastern Asia Society for Transportation Studies (EASTS), Kuala Lumpur, Malaysia, 4–7 September

Breyer N, Rydergren C, Gundlegård D (2022) Semi-supervised mode classification of inter-city trips from cellular network data. J Big Data Anal Transp 4:23–39

Capponi A, Vitello P, Fiandrino C, Cantelmo G, Kliazovich D, Sorger U, Bouvry P (2019) Crowdsensed data learning-driven prediction of local businesses attractiveness in smart cities. IEEE symposium on computers and communications

Cetin N, Nagel K, Raney B, Voellmy A (2002) Large-scale multi-agent transportation simulations. Comput Phys Commun 147(1–2):559–564

Cramer H, Rost M, Holmquist L (2011) Performing a check-in: emerging practices, norms and conflicts' in location-sharing using foursquare. In: 13th International Conference on Human Computer Interaction with Mobile Devices and Services

Ding C, Cao X, Yu B, Ju Y (2021) Non-linear associations between zonal built environment attributes and transit commuting mode choice accounting for spatial heterogeneity. Transp Res Part A: Policy Pract 148:22–35

Efthymiou D, Antoniou C (2012) Use of social media for transport data. Procedia Soc Behav Sci 48:775–785

Gao Y, Schmöcker J-D (2021) Estimation of walking patterns in a touristic area with Wi-Fi packet sensors. Transp Res Part C 128:103219

Gao Y, Schmöcker J-D (2023) Inferring travel patterns and the attractiveness of touristic areas based on fusing Wi-Fi sensing data and GPS traces with a Kyoto Case study. Transportation Research Board. Washington D.C.

Golder SA, Macy MW (2014) Digital footprints: opportunities and challenges for online social research. Sociology 40(1):129

Google (2022) Popular times, wait times, and visit duration. Retrieved 1 11, 2022, from https://support.google.com/business/answer/6263531?hl=en

Horni A, Nagel K, Axhausen W (2016) The multi-agent transport simulation MATSim. Ubiquity Press, London

Jee H, Sun W, Schmöcker J-D (2021) Estimation of bus line specific waiting times using Wi-Fi signal data. In: 7th International symposium on the use of public transit automated data for planning and operations (TransitData2021). Held online

Jee H, Schmöcker J-D, Barbeau S, Lozano W, Watkins KE (2022) Constructing passenger trips and tours using data from an experiment with the "OneBusAway" application. CASPT, Tel Aviv,Israel

Kurauchi F, Schmöcker JD (eds) (2017) Public transport planning with smart card data. CRC Press

Liu C, Erdogan S, Ma T, Ducca FW (2016) How to increase rail ridership in Maryland: direct ridership models for policy guidance. J Urban Plan Dev. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000340

Longley PA, Singleton A, Cheshire J (2018) Smart card data and human mobility. Consumer data research. UCL Press, p 111

Lue G, Miller E (2019) Estimating a Toronto pedestrian route choice model using smartphone GPS data. Travel Behav Soc 14:34–42

Luo S, He SY, Grant-Muller S, Song L (2023) Influential factors in customer satisfaction of transit services: using crowdsourced data to capture the heterogeneity across individuals, space and time. Transp Policy 131:173–183

MacKenzie D, Cho H (2020) Travel demand and emissions from driving dogs to dog parks. Transp Res Rec 2674(6):291–296

Maghrebi M, Abbasi A, Rashidi TH, Waller ST (2015) Complementing travel diary surveys with Twitter data: application of text mining techniques on activity location, type and time. In: IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)

Mahajan V, Cantelmo G, Antoniou C (2021) Explaining demand patterns during COVID-19 using opportunistic data: a case study of the city of Munich. Eur Transp Res Rev 13:26

Miller EJ (2014) A framework for urban passenger data collection. In: 10th International conference on transport survey methods. Leura, Australia

Möhring M, Keller B, Schmidt R, Dacko S (2020) Google Popular Times: towards a better understanding of tourist customer patronage behavior. Tour Rev 76:533–569

Namulindwa S (2023) Crowd estimation and prediction using wi-fi sensor data: an experiment on katsura campus. Master's Thesis, Department of Urban Management, Kyoto University

Nishigaki T, Schmöcker JD, Yamada T, Nakao S (2023) Estimating the number of tourists in Kyoto based on GPS traces and aggregate mobile statistics. In: Proceedings of the 12th international scientific conference on mobility and transport: mobility innovations for growing megacities. Springer Nature, Singapore, pp 221–243

Ortúzar J, Willumsen L (2011) Modelling transport, 4th edn. Wiley, Hoboken

Osorio-Arjona J, Horak J, Svoboda R, García-Ruíz Y (2021) Social media semantic perceptions on Madrid metro system: using Twitter data to link complaints to space. Sustain Cities Soc 64:102530

Pelletier M-P, Trepanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C 2011(19):557–568

Poom A, Järv O, Zook M, Toivonen T (2020) COVID-19 is spatial: ensuring that mobile big data is used for social good. Big Data Soc 7(2):2053951720952088

Rajput P, Chaturvedi M (2019) Automatic detection of bus-stops and bus-crowdedness using crowdsourced data. In: IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand. pp 740–745

Ryu S, Park B, El-Tawab S (2020) WiFi sensing system for monitoring public transportation ridership: a case study. KSCE J Civ Eng 24(10):3092–3104

Sun W, Schmöcker JD (2021) Demand estimation for public transport network planning. In: The Routledge handbook of public transport (Chapter 21). Routledge, pp 289–305

Sung H, Choi K, Lee S, Cheon S (2014) Exploring the impacts of land use by service coverage and station-level accessibility on rail transit ridership. J Transp Geogr 36:134–140

Tafidis P, Teixeira J, Bahmankhah B, Macedo E, Coelho MC, Bandeira J (2017) Exploring crowdsourcing information to predict traffic-related impacts. In: IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe, pp 1–6

Tasse D, Hong J (2014) Using social media data to understand cities. In: Proceedings of NSF Workshop on Big Data and Urban Informatics

Timokhin S, Sadrani M, Antoniou C (2020) Predicting venue popularity using crowd-sourced and passive sensor data. Smart Cities 3(3):42

Vitello P, Fiandrino C, Connors R, Viti F (2023) Exploring the potential of Google Popular Times for transit demand estimation. Transportation Research Board (TRB) 102nd Annual Meeting

Wang F, Ross CL (2018) Machine learning travel mode choices: comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transp Res Rec 2672(47):35–45

Willumsen L (2021) Use of big data in transport modelling. In: International Transport Forum Discussion, 2021(5)

Wu L, Kand J, Chung Y, Nikolaev A (2019) Monitoring multimodal travel environment using automated fare collection data: data processing and reliability analysis. J Big Data Anal Transp 1:123–146

Zhang Y, Mohammadian A (2010) Bayesian updating of transferred household travel data. Transp Res Rec: J Transp Res Board 2049:111–118

Zhu Y, Chen F, Wang Z, Deng J (2019) Spatio-temporal analysis of rail station ridership determinants in the built environment. Transportation 46:2269–2289