Public Transport (2023) 15:287–319
https://doi.org/10.1007/s12469-022-00309-0

ORIGINAL RESEARCH

A supervised machine learning model for imputing missing boarding stops in smart card data

Nadav Shalit[1] · Michael Fire[1] · Eran Ben-Elia[2]

Accepted: 4 November 2022 / Published online: 7 December 2022

Abstract

Public transport has become an essential part of urban existence with increased population densities and environmental awareness. Large quantities of data are currently generated, allowing for more robust methods to understand travel behavior by harvesting smart card usage. However, public transport datasets suffer from data integrity problems; boarding stop information may be missing due to imperfect acquirement processes or inadequate reporting. This study introduces a supervised machine learning method to impute missing boarding stops based on ordinal classification using GTFS timetable, smart card, and geospatial datasets. A new metric, Pareto Accuracy, is suggested to evaluate algorithms where classes have an ordinal nature. The results are based on a case study in the city of Beer Sheva, Israel, consisting of one month of smart card data. We show that our proposed method is robust to irregular travelers and significantly outperforms well-known imputation methods without the need to mine any additional datasets. The data validation from another Israeli city using transfer learning shows the presented model is general and context-free. The implications for transportation planning and travel behavior research are further discussed.

Keywords Machine learning · Smart card · Boarding stop imputation · Public transport · Missing data · Pareto accuracy

✉ Michael Fire
mickyfi@bgu.ac.il

Nadav Shalit
nadshalit@gmail.com

Eran Ben-Elia
benelia@bgu.ac.il

[1] Data4Good Lab, Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[2] GAMESLab, Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer-Sheva, Israel

1 Introduction

Public transport (PT) is an integral part of everyday life in many cities. The gradual shift of the global population over the past century to urban areas is markedly increasing people's dependence on PT for their daily mobility needs (Petrović et al. 2016). PT is a complex system that is based on physical elements of stops, vehicles, routes, and other temporal and spatial elements (Ceder 2016). The PT system consists of regularly scheduled vehicle trips open to all paying passengers, with the capacity to carry multiple passengers whose trips may have different origins, destinations, and purposes (Walker 2012). PT is ideal when passengers regard its service as punctual and regular (Walker 2012). With the growth in the number of cars on urban roads, PT improvements have become an essential part of traffic congestion mitigation strategies and are vital in promoting sustainable transportation (Al Mamun and Lownes 2011). Although understanding the patterns of PT use is crucial to its planning, this task remains a significant challenge in practice and research.

Numerous studies in recent years have examined the behavior of PT travelers (Li et al. 2018) in efforts to address this challenge. Habitual travel behavior is of great interest to transportation planners, and its analysis can help improve demand predictions and justify necessary upgrades to PT supply (Briand et al. 2017). This analysis can also contribute to improvements in PT (service/planning/upgrades) with respect to the management of COVID transmission, in terms of providing better information on crowded areas, such as bus stops, which is critically important to the global issue of public health. To this end, transportation planners typically use travel behavior surveys (Stopher and Greaves 2007). While these surveys statistically reflect travel behavior correctly, they are also expensive, time-consuming, and often unable to generate sufficient amounts of data relative to the size of the population, and would need significant changes in scope to cover recent COVID concerns.

Conversely, data harvested from smart cards can generate millions of records compared to a typical sample ranging from 2,500 to 10,000 households using surveys (Maeda et al. 2019). Smart cards, also known as automatic fare collection (AFC), provide an efficient and cost-saving alternative to the manual fare collection method (Jang 2010; Chen and Fan 2018). In addition to fulfilling fare collection needs, as a bi-product, smart card transactions also generate geocoded timestamps that record every passenger's boardings, line transfers, and sometimes alightings for a wide range of PT vehicles (bus, tram, train, or metro) (Pelletier et al. 2011; Faroqi et al. 2018). These records are generated for almost the entire passenger population (Pelletier et al. 2011; Faroqi et al. 2018). Such information is a treasure trove for travel behavior analyses, especially for extracting passengers' spatio-temporal travel patterns (e.g., Origin–Destination matrices or path choice (Wang et al. 2011)). Nevertheless, common statistical inference methods applied in surveys are of little practical use for understanding the travel patterns of an entire population. Therefore, different methods are required.

Kandt and Batty (2021) proclaimed a new area of urban research defined by advances in big data analytics, with smoother decision making and a deeper understanding of urban systems. A massive increase in the *volumes*, *velocities* and *varieties* of big data have also been paralleled by recent developments in the data science field. New data mining tools and robust cloud computing capabilities (Li et al. 2015, 2018) create new opportunities to analyze travel behavior patterns at the individual level, over extended periods, and in large urban areas (Ma et al. 2017). The availability of big data has a vast potential to improve the quality of transportation planning and research, and by applying big data analytics and data mining methods, this task has become much more feasible (Ma et al. 2013).

However, similar to the case in other domains, the *veracity* of such datasets remains questionable (Ben-Elia et al. 2018). Smart card datasets, in particular, may suffer from integrity problems, such as incorrect or missing values, e.g., when operators only record partial data. For example, in Yan et al. (2019), boarding stop information was completely missing, and only time stamps remained intact in the dataset. A common solution for such problems is to replace the missing or erroneous data by utilizing alternative publicly accessible data. One possible solution is to use official PT timetables to impute the missing data in missing boarding stop information. One popular source for such data comes from the General Transit Feed Specification (GTFS), first created in 2006 by Google (Google, 2016), defined as a standard file format for storing PT schedules and associated geographic information (Ma et al. 2012). GTFS contains the complete schedules and routes of every PT line planned for each day of the month in tabular formats together with corresponding geographic shapefiles and is widely used in over 750 urban regions across the world (Hadas 2013; Antrim et al. 2013).

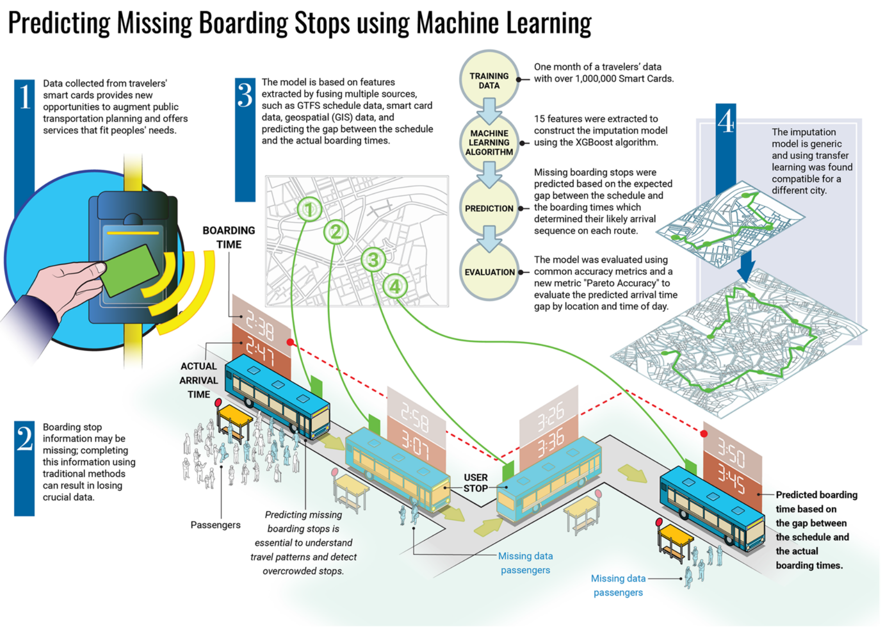
Nonetheless, PT running times and arrival times at stops are never perfectly aligned with their official timetables, where PT is not always punctual, even in developed countries. For example, Cats and Loutos (2016) found that only 10% of all arrivals were within an interval of 15 s. This issue becomes more acute, especially when PT vehicles–mainly buses–also share the same road space with private and commercial vehicles (i.e., mixed traffic). While this issue is less severe in major urban areas in developed countries where rail-based and PT bus preemption infrastructure is widespread and right-of-way strongly enforced, this is not the reality everywhere. For example, in Israel (an official OECD member), buses accounted nationally for 85% of PT trips in 2019, with more than 2M passengers served daily. The country suffers from a shortage of adequate PT infrastructure (namely, too few priority lanes—14 m per capita, compared to 300 m in the EU), thus resulting in poor PT service punctuality (Ceder 2004). As shown later, this fact makes schedule-based imputation a poor substitute for boarding stop prediction. A second solution is to discard such data by simply removing missing records or those that do not align with a prescribed hypothesis (Tao et al. 2014). Nonetheless, discarding data can be regarded as a reasonable solution only when that share of the missing data is small. However, when the missing portion is substantial, the whole dataset could be compromised and discarded. This scenario can render certain urban areas effectively blind vis-a-vis smart card data. A third option is to complement the missing data by combining different datasets. In this respect, either automatic vehicle location (or AVL), which uses installed GPS transponders to locate PT vehicles and estimate real-time arrival times at designated stops; or automatic passenger counters (or

APC), which use infrared or laser technologies to estimate boarding and alighting passenger numbers, have been used in combination with smart card data (Shalaby and Farhan 2004; Mazloumi et al. 2010; Khiari et al. 2016). Yet, such data is neither always available (Chen and Fan 2018; Yan et al. 2019), nor efficient, as considerably more errors may well be introduced in the process (Luo et al. 2018). These two facts likely reduce their suitability for data imputation. Moreover, even when such data sources exist, matching between them is somewhat challenging. For example, Lou et al. (2018) had no vehicle trip identification (ID), making it impossible to match with AVL records. A further difficulty is that missing data can vary by city or between different operators (Laña et al. 2018). In some cities, data integrity is regarded as very strong, and consequently, boarding and alighting imputation tasks are very good (Munizaga et al. 2014). In contrast, in other cities where data sources are lacking, data integrity can also be flawed.

The lack of common standards and methods for data handling and processing across the PT modes and sectors has been identified as a main problem hindering efficient utilization of smart cards and other PT-related data. Such a data interoperability requires developing a standardized application approach that will allow data mining tools and models to be tested and implemented as asserted by Covic and Voß (2019). In this respect heuristic methods, such as ML, can be regarded as a viable solution to perform data imputation tasks (Yan et al. 2019). To this end, our aim is a general and context-free boarding stop imputation method. Specifically, we address use cases where data quality is considered too insufficient to impute by cross-inference and without the need to harvest any other data than what is necessary. While still providing valuable insights for transportation planners, we consider this of particular relevance for developing countries where the traveler population is mostly PT-dependent. We established a general boarding stop imputation method to improve the quality and integrity of PT datasets by predicting missing or corrupted travelers' records in smart card data.

Namely, to the best of our knowledge, we developed the first machine learning (ML) algorithm for predicting passengers' boarding stops (see Fig. 1). Our algorithm is based on features extracted by harvesting three big data sources, the planned GTFS schedule data, smart card (AFC) data, and geospatial (GIS) data. We applied a machine learning model to these features to predict boarding stops based on the notion of embedding (see Sect. 3). To train and evaluate our algorithm's performance, we utilized a real-world smart card dataset from the city of Beer Sheva in Israel that consists of over a million trips taken by more than 85,000 passengers. Since the boarding stops are embedded, they also become ordered, and therefore, the problem we addressed is ordinal classification. Accordingly, we also propose a new method of evaluation that shows the percentage in each error dimension that we define as Pareto Accuracy, which is more interpretable and allows for better comparison between imputation models. We show that our model performed significantly better than a naïve prediction model based on harvesting GTFS data alone (aka schedule-based) and other imputation methods.

In this study, we succeeded to generate a model which is both wholly generic and has considerably higher accuracy and recall values than other tested imputation methods (see Sect. 4). Additionally, we demonstrated that we obtain similar



Predicting Missing Boarding Stops using Machine Learning

1 Data collected from travelers' smart cards provides new opportunities to augment public transportation planning and offers services that fit peoples' needs.

2 Boarding stop information may be missing; completing this information using traditional methods can result in losing crucial data.

3 The model is based on features extracted by fusing multiple sources, such as GTFS schedule data, smart card data, geospatial (GIS) data, and predicting the gap between the schedule and the actual boarding times.

4 The imputation model is generic and using transfer learning was found compatible for a different city.

TRAINING DATA: One month of a travelers' data with over 1,000,000 Smart Cards.

MACHINE LEARNING ALGORITHM: 15 features were extracted to construct the imputation model using the XGBoost algorithm.

PREDICTION: Missing boarding stops were predicted based on the expected gap between the schedule and the boarding times which determined their likely arrival sequence on each route.

EVALUATION: The model was evaluated using common accuracy metrics and a new metric "Pareto Accuracy" to evaluate the predicted arrival time gap by location and time of day.

Predicting missing boarding stops is essential to understand travel patterns and detect overcrowded stops.

Predicted boarding time based on the gap between the schedule and the actual boarding times.

Fig. 1 Predicting missing boarding stops algorithm overview

prediction results in an entirely different city using our method. Moreover, we show how other imputation methods are not always applicable, while our methodology can be applied with a broader scope.

Our study's overall focus is to improve the integrity of public transport data. Specifically, our study provides the following two main contributions:

1. We present a novel prediction model for imputing missing boarding stops using supervised learning. Moreover, our proposed model is generic and transferable, i.e., it can be trained on one city's data and then impute missing data in another municipality.
2. We propose a new metric—Pareto Accuracy—for evaluating public transport metrics that are more interpretable and allow broader comparisons between imputation models.

The rest of the paper is organized as follows: In Sect. 2, we review related work on smart card usage, missing data imputation, and ML applications in transportation research. Section 3 describes the use case, experimental framework, and methods used to develop the ML model and the extraction of its features. In Sect. 4, we present the results of the ML model and compare its performance to other known solutions. In Sect. 5, we discuss the implications of the findings and the study's limitations and present our conclusions and future research directions.

2 Related work

We provide an overview of relevant studies by first presenting smart card research in general, followed by studies that have utilized smart card data with machine learning to perform predictive analytics. We then give an overview of the field of missing data imputation. Lastly, we present studies in the field of ordinal classification.

2.1 Smart card analytics

The smart card system was introduced as a smart and efficient AFC system in the early 2000s (Chien et al. 2002) and has since become an increasingly popular payment method (Bagchi and White 2005; Trépanier et al. 2007). In particular, smart cards have also become an increasingly popular source of big data for research and policy making (Agard et al. 2006; Jang 2010). For example, smart card data is used for exploring travel behavior, determining travel patterns, measuring the performance of PT services, locating critical transfer points, and analyzing crowdedness effects on route choice (Bryan and Blythe 2007; Jang 2010; Alguero 2013; Zhao et al. 2017; Li et al. 2018; Yap et al. 2020). Recently, smart card datasets were used to study travel behavior changes to travel behavior as a result of the COVID-19 pandemic (Almlöf et al. 2020; Orro et al. 2020; Zhang et al. 2021). Comprehensive literature reviews of smart card usage were provided by Pelletier et al. (2011), Schmöcker et al. (2017), and Faroqi et al. (2018).

Initially, smart card research applied rather classic statistical methods and descriptive analytics. Devillaine et al. (2012) inferred the location, time, duration, and designation of PT users' activities using rules derived from smart card data and work and study schedules. The main research challenge evident in the literature was to estimate origin–destination (OD) matrices which describe the spatial distribution of travel demand between locations during different periods of the day (Chu and Chapleau 2008; Wang et al. 2011; Munizaga and Palma 2012; Gordon et al. 2013). OD matrices are also crucial inputs to perform the three stages in PT network design, namely: route design, frequency (headway) setting, and timetabling (Guihaire and Hao 2008). Before the advent of smart cards, these matrices were only derived and validated based on some representative sample of travelers (Chen et al. 2016). However, as noted, surveys often lack sufficient spatial and temporal coverage. Various studies have demonstrated the advances in OD estimation with smart card data (Chu and Chapleau 2008; Wang et al. 2011; Munizaga and Palma 2012; Gordon et al. 2013).

Nevertheless, with the introduction of smart cards, new problems in OD estimation appeared. Namely, many PT agencies adopted a TAP (Transit Access Protocol) IN system where only boarding stop information is recorded. In contrast, the availability of alighting stop information "TAP IN+TAP OUT" systems allows for the OD matrix to be derived using more straightforward approaches. Alighting stop information is necessary for many tasks such as route loading profiles, market research, and improvements in service planning. However, under TAP IN, the destination must be somehow predicted (Trépanier et al. 2007; Faroqi et al. 2018).

In addition, combining smart card data with a smaller scale travel behavior survey for validation purposes is a useful approach to better understand passengers' daily travel patterns (Wang et al. 2011). Nonetheless, OD analyses inherently assume that PT passengers travel routinely back and forth from/to the same locations. Recent findings suggest this assumption does not necessarily hold, and some share of PT passengers are quite flexible (Huang et al. 2018) or use PT infrequently (Benenson et al. 2019). Therefore, a simple OD estimation will possibly result in PT planning that is mismatched with actual demand patterns.

Traditional analysis methods do not take advantage of the full potential of the added value of big data. At the same time, rapid growth in power and cost reduction in computational technologies provide new opportunities, both in terms of the availability of the massive amount of data collected and the development of more novel algorithms (Welch and Widita 2019). Agard et al. (2006) obtained travel behavior indicators that identify daily travel patterns and clustering of major user groups. Kieu et al. (2015) applied a density-based spatial clustering application with noise (DBSCAN) algorithm to cluster passengers and identify classes of passengers for strategic planning improvements. Ma et al. (2013) used smart card data to cluster the travel patterns of PT riders to characterize commuter profiles.

In this respect, the literature shows a shift toward harvesting the prognostic nature of ML to yield better predictive analytics highlighting the growing emphasis on using smart card data for analytical purposes. This shift underscores the change from the more straightforward analyses conducted in the past to the more comprehensive analysis done today. Hagenauer and Helbich (2017) compared several ML classifiers and showed both their predictive power and ability to uncover travelers' mode choices via feature importance analysis. For example, they showed that the trip distance was the most important predicting factor, while the temperature was only a key feature for predicting bicycle use. In 2018, Palacio (2018) showed that ML predictions are much more accurate than traditional linear models that were sub-optimal both in terms of R-square and MSE. In the following year, Traut and Steinfeld (2019) combined smart card data with crime records to assist agencies in identifying insecure and dangerous PT stops. Chen et al. (2016), who inferred mode and route choices, stress the need for cross-disciplinary collaborations between data scientists and transportation planners to exploit the information withheld in the data. Further evidence of the prominence of big data analytics in PT research can be found in several review papers such as Fonzone et al. (2016), Namiot and Sneps-Sneppe (2017), Anda et al. (2017), Li et al. (2018), and Milne and Watling (2019).

Deep learning algorithms have also been utilized to address PT issues using smart card data. Deep learning is a sub-field of ML that automatically creates feature engineering, and its methods are state-of-the-art in many domains. Examples of such implementations include forecasting passenger destinations (Toqué et al. 2016; Jung and Sohn 2017), predicting multimodal passenger flows (Toqué et al. 2017), improving passenger segmentation (Dacheng et al. 2018), inference of passenger employment status (Zhang and Cheng 2020), and using standard deep network and long- and short-term memory networks, inference of demographics using convolutional neural networks (Zhang et al. 2019).

2.2 Missing data imputation

Incomplete data is a universal problem, and the application of different imputation methods will often yield different results. Therefore, to preserve reproducibility, they must be adequately addressed (Saunders et al. 2006). This problem is, notably, relevant for transportation planning, e.g., in the case of road traffic analysis (Qu et al. 2009). Incomplete data is a well-known problem in the data mining literature where a significant amount of data can be missing or incorrect. Lakshminarayan et al. (1996) elucidated both the severity of this issue as well as recommended applying ML techniques toward its solution rather than classical statistical methods. Batista and Monard (2003) assert that missing data imputation must be carefully handled to prevent bias from being introduced. Moreover, they show that the most common methods, such as mean or mode imputation, are not always optimal. One example we found in the PT literature is from Kusakabe and Asakura (2014). They used a Naïve Bayesian model for data imputation and analysis of PT to understand continuous long-term changes in trip attributes. They showed both the power of smart card data and the usefulness of missing data imputation in this field. Their method of imputation, however, is not reported in sufficient detail to be understood or replicated.

Several techniques to optimize missing data imputation showed the importance attributed to this area of research (Bertsimas et al. 2017). Moreover, even state-of-the-art deep learning methods have been applied to this problem (Garg et al. 2018; Costa et al. 2018; Camino et al. 2019). These implementations were performed on a variety of datasets and problems, such as classification of continuous attributes (breast cancer and default credit card classification); images (Camino et al. 2019; Garg et al. 2018); and regression (Camino et al. 2019). Insofar as this field of study has not been operationalized for PT data, further examination is warranted, particularly when considering the issue of completing missing data to provide better information on crowded PT areas, as it pertains to the spread of COVID.

In many imputation tasks, including PT, ML methods outperform standard methods significantly when the missing portion increases (Saunders et al. 2006; Laña et al. 2018; Echaniz et al. 2020; Yan et al. 2019). Additionally, standard imputation methods are too sensitive to the ratio of missing data and infrequent or ‘irregular’ users of the PT network (Van Lint et al. 2005). Conversely, ML-based imputation showed stable results regardless of the missing ratio (Laña et al. 2018). As noted previously, one solution is to impute the missing boarding stops using complementary datasets such as AVL or APC. However, AVL data are not always available (Chen and Fan 2018), whereas combining several datasets (i.e., AVL, AFC, APC, GTFS, etc.) can introduce more errors, and make it much harder to match them perfectly (Luo et al. 2018).

2.3 Ordinal classification

Classification is a form of supervised ML that aims to generalize a hypothesis from a given set of records. It learns to create $h(x_i) \rightarrow y_i$ where y has a finite number of classes (Kotsiantis et al. 2007). The basic metrics for classification are sensitivity, specificity, and accuracy (Jiao and Du 2016). Accuracy is the percentage of observations classified correctly, specificity is the percentage of true negatives classified correctly, and sensitivity is the percentage of true positives classified correctly. A classification task becomes ordered when the classes have some inherent order between them. There are a variety of metrics to evaluate supervised learning algorithms (Liu et al. 2014). Each metric has its advantages and limitations. This study introduces the Pareto Accuracy (see Sect. 3.3), suitable for assessing the constructed classifiers' performances in imputing missing boarding stops based on ordinal classification.

Ordinal classification is a form of multi-class classification where the classes exhibit some natural ordering (such as cold, warm, and hot), but not necessarily numerical traits for each class. Rather than being chosen based on the traditional metrics discussed above, a classifier may be chosen based on the severity of its errors (Gaudette and Japkowicz 2009). Additionally, classic modeling techniques will sometimes perform suboptimally since ML models assume there is no order between classes. In such tasks, e.g., the well-known Boston housing and breast cancer datasets, different models that take advantage of ordinal information are preferred (Frank and Hall 2001). In this case, additional metrics are proposed to calculate such tasks differently, such as regression metrics like Mean Absolute Error (MAE) and the Mean Square Error (MSE) and even their own metric, the Ordinal Classification Index (Cardoso and Sousa 2011). Notwithstanding, as noted below, these approaches neither fit our data nor our needs. Therefore, we developed a different and novel performance metric (see Sect. 3.3).

3 Methods

The main goal of our study is to use ML algorithms to improve the integrity of PT data. Specifically, we develop a supervised learning-based model to impute missing boarding stops in any given smart card dataset. Moreover, our goal is to construct a generic model that will be fully transferable to other datasets to impute missing data in different contexts without further adjustments.

To maintain these generic objectives, we had to contend with two significant challenges: First, we could only incorporate generic properties in our model. For instance, our model cannot include the actual line number of a bus route specific to a particular city. Moreover, since supervised ML algorithms can only predict classes they were initially trained upon, classification classes must remain the same across datasets, e.g., bus stop #14 in a specific city is an irrelevant feature for other cities. Therefore, once more, a different numerical representation is applied by embedding (see Sect. 3.2). Second, we develop a genuinely generic model that can also be applied to other geographical contexts in which it was not initially trained. The

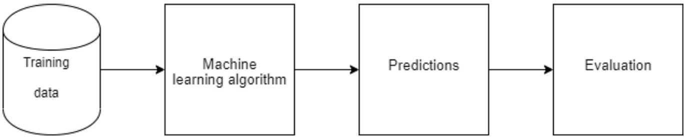


Fig. 2 Modeling methodology overview

model must also undergo a process of transfer learning (Torrey and Shavlik 2010) that entails the transfer of relevant knowledge by fine-tuning a model on a "novel" dataset. In our case, our model underwent the process of transfer learning using a dataset on which it had not been trained before.

The missing boarding stop values were imputed using the following methodology (see Figs. 1 and 2): First, we preprocessed and cleaned the smart card dataset that we utilized in this study (see Sect. 3.1). Next, we extracted various features from two other datasets: (a) the GTFS timetable data; and (b) open municipal geospatial data. In addition, we converted boarding stops from their original identifiers to embedded numerical representations based on GTFS data (see Sect. 3.2). Afterward, we applied ML algorithms to estimate a model that can predict the missing boarding stops. We used SHAP (SHapley Additive exPlanations) values for determining feature importance (Lundberg and Lee 2017),[1] i.e., which features make the most substantial contribution to the predictive power of the model (see Fig. 8). We also evaluated the performance of our model using a novel performance metric called Pareto Accuracy. Then, based also on common metrics, we evaluated our model relative to a schedule-based model estimated only on GTFS timetable data. Finally, we compared our model to several other comparative models (e.g., passenger history, temporal proximity, or semi-random guessing) that were previously used in the literature. Below, we describe each step of our approach in more detail.

3.1 Datasets and data preprocessing

As noted above, we used three datasets:

1. *The Smart card dataset*—"Rav Kav" is the Israeli AFC system applying the TAP protocol, allowing PT passengers to pay for their trip using their smartcards anywhere in the country. Rav-Kav operates a nationwide TAP IN for buses and rail that codes information on unique passenger identifiers, traveler types (such as student or senior travelers), boarding stops, boarding timestamps, fares, discount attributes, and unique trip identifiers of the line at that time. For rail trips only, TAP OUT also records alighting stops and times. During the period 2018/9, circa 2M boardings were recorded per day in the entire country.

[1] "SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and related extensions" https://github.com/slundberg/shap.

2. *GTFS*—a GTFS feed, as described above, consists of rail/bus schedules and timetables, stops, and routes of every PT trip planned for every day of the month. In Israel since 2012, the GTFS feed has been published daily online by the Ministry of Transport, providing schedules of 36 bus and rail operators, encompassing 7,800 route-direction-pattern alternatives served by 28,000 bus and rail stations. The GTFS feed aligns with the smart card dataset as described below. This study utilized the GTFS dataset to enrich the feature space and convert boarding stop records into an embedded numerical value.
3. *Geospatial information*—we derived a variety of geospatial attributes from municipal GIS databases.

To obtain a dataset suitable for constructing the prediction model, we were required to remove any record that lacked a boarding stop or a trip ID (a unique identifier of a trip provided by a specific and unique PT operator) from the smart card dataset. Next, we joined the smart card dataset with the GTFS dataset by matching the trip ID attributes. Lastly, we joined the geospatial dataset with the smart card dataset using the GTFS dataset, which contains all the geographic coordinates of each PT route.

3.2 Feature extraction and machine learning model construction

ML performance is highly correlated to the quality of the feature space, and therefore, including more features results in better model performance (Gudivada et al. 2017). While the smart card data contains the PT line and boarding time of each passenger, it lacked several essential data, such as the duration that had elapsed since that line left the origin depot, the time remained until arrival to the final destination, the total number of stops, and other relevant trip attributes. Moreover, the smart card data is missing physical geospatial characteristics, such as the number of traffic lights on the PT route that more likely increases traffic congestion and consequent delays, and could well strengthen model performance.

Overall, three features were extracted using the smart card dataset, five features using the GTFS dataset, three features using the geospatial dataset and four from combined GTFS and smart card datasets. From the 41 features we initially tested in total, we selected 15 to include in our model based on stepwise selection and feature importance analysis (see Table 1) by exploiting the SHAP values Lundberg and Lee (2017).

To construct the prediction model, we used the GTFS dataset to create a schedule-based prediction. This naive prediction reflects the transit vehicle's position along a line according to the GTFS schedule. Namely, let S_i be the sequence number of the boarding stop based on the GTFS schedule and let A_i be the actual boarding stop sequence number. Then, we define D_i as $D_i = A_i - S_i$. Our prediction model goal was to predict D_i by utilizing the variety of features presented in the previous section.

For instance, consider a passenger who boarded a line at the third stop, i.e., $A_i = 3$, but the transit vehicle was scheduled to arrive at the second stop at the designated

Table 1 Extracted features

Dataset	Feature	Explanation
Municipal geospatial records	Addresses_average	The number of addresses listed on the route
	Street_light_average	The number of streetlights on the route
	Traffic_Lights_average	The number of traffic lights on the route
GTFS	Number_of_points	The number of points in a shapefile in GTFS per route
	Average_distance_per_stop	The total length of the route divided by the number of points
	Average_time_per_stop	The total expected travel time of the route divided by the number of points
	Average_points_to_stops	The number of points in a shapefile in GTFS per route divided by the number of points
	Time_diff_of_trip	Total travel time
GTFS and smart card	Time_from_boarding_to_last_stop	Time from boarding time to expected last stop of the route
	Time_from_departure_to_boarding	Time from route departure time to boarding time
	Predicted_sequence	GTFS prediction sequence of the most likely stop
	Hourly_expected_lateness	The average lateness per hour (based on training data)
Smart card	Boardingtime_Seconds_from_midnight	Timestamp of boarding to a numerical value in seconds from midnight
	Boardingtime_weekday	The day of the week in which the boarding occurred
	Is_weekend	Is it a weekend?

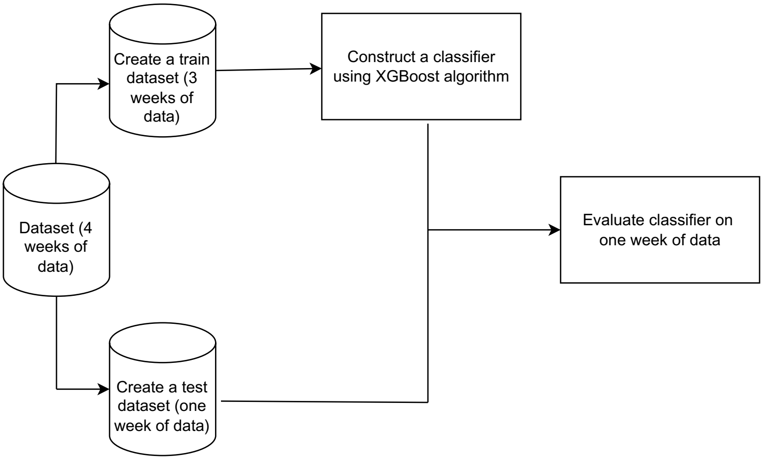

Create a train dataset (3 weeks of data)
Construct a classifier using XGBoost algorithm
Dataset (4 weeks of data)
Evaluate classifier on one week of data
Create a test dataset (one week of data)

Fig. 3 Evaluation process overview

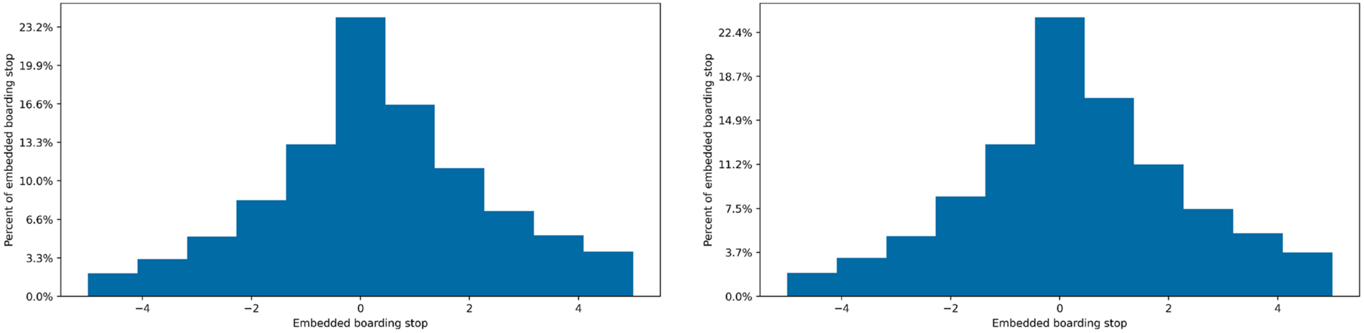
time. The schedule-based prediction would be 2, i.e., $S_i = 2$, the stop where it was supposed to be at that time. Then, the difference is $D_i = A_i - S_i = 3 - 2 = 1$, and this is the class the algorithm will predict.

Subsequently, we performed the following steps to construct the prediction model: First, we selected several well-known classification algorithms. Namely, we used Random Forest (Singh et al. 2016), Logistic Regression (Singh et al. 2016), and XGBoost (Chen and Guestrin 2016). Second, we split our dataset into training and testing datasets (see Fig. 3). Due to the temporal nature of the data, we used a logical splitting, the training dataset consisted of the first three weeks of data (75%), and a testing dataset consisted of the last week of data (25%).[2] Figure 4 shows the distributions of the embedded boarding stops by the computed difference between the actual and schedule-based sequences (D_i) for the training and testing subsets. No apparent differences between the two distributions are evident. Third, for both the training and testing datasets, we extracted all the 15 features mentioned above. Fourth, we constructed the prediction models using each one of the selected algorithms. Lastly, we compared the generated models and selected the one with the best performance based on the Pareto Accuracy metric (see Sect. 3.3).

3.3 Model evaluation

We evaluated each model and compared it to the schedule-based method on the test dataset using common metrics: accuracy, recall, precision, F1 (see Appendix A for

[2] It is frequent practice to split large-scale datasets into test and train datasets (Guyon 1997), where a split of 80%/20% is regarded a common practice. In our case, due to the temporal nature of the dataset, we find it practical and logical to train the classifier on three weeks (75%) of data and evaluate the classifier's performance on a week (25%) of data.



(a) Train boarding stop histogram

(b) Test boarding stop histogram

Fig. 4 Train and test boarding stop histogram

definitions), and the new metric we developed: Pareto Accuracy. We used the following variables for our novel Pareto Accuracy metric: Let p_i be the predicted sequence of stop$_i$, a_i be the actual sequence, and d_i be the absolute difference between them. Let l be the limit of acceptable difference for imputation, i.e., if an error of one stop is tolerated, such as for neighborhood segmentation, then $l = 0$. Let X_i be an indicator defined as:

$$X_i = \begin{cases} 1 & \text{if } di <= l \\ 0 & \text{otherwise.} \end{cases}$$

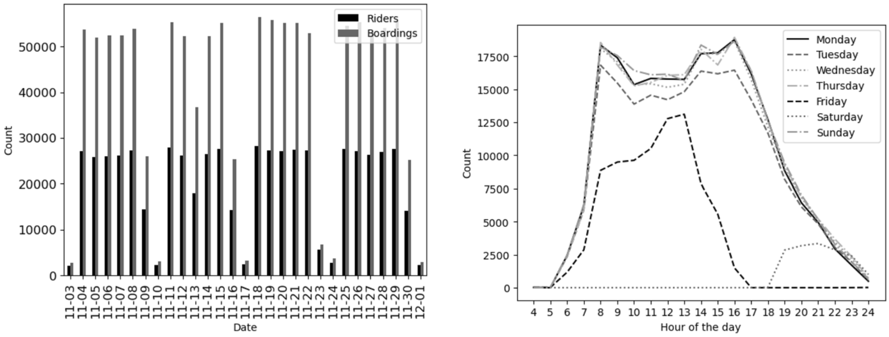
We define Pareto Accuracy as follows:

$$\mathrm{PA}_l = \frac{\sum_{i=1}^{n} X_i}{n}.$$

The PA metric is a generalization of the accuracy metric. Namely, PA_0 is the well-known accuracy metric. Unlike other ordinal classification methods, the primary advantage of using the PA metric is to evaluate the accurate dimension of error while being extremely robust to outliers (by setting parameter l). Moreover, this metric is highly informative since its outcome value can be interpreted easily; for example, 0.6 means that 60% of the predictions had at most l difference from true labels.

For example, let us consider a set of eight observations of embedded boarding stops $\{-2, 0, 3, 20, -3, 4, 3, 2\}$, where each observation is a simulated boarding by a passenger where each number (D_i) in the set represents the difference between expected (S_i) and actual boarding stops (A_i). With a value of 20, the fourth observation is an outlier, which might occur due to some fault in the decoder device of the public transport operator. We do not want to predict it, as it is naturally unpredictable. We seek a metric that will be both resilient to outliers, as they are unpredictable, and still account for the true dimension of the errors (see Sect. 2.3). Let us compare two classifiers, A and B. Classifier A predicted the following boarding stops $\{-2, 0, 4, 3, -2, 3, 2, 2\}$, while Classifier B predicted {3,0,3,7,1,1,3,2}. Classifier A is a more useful classifier since, in general, its predicted values are closer to the actual values, i.e., its variance is very small, which makes it more reliable. However, when using the classical accuracy and RMSE metrics, Classifier B has a higher accuracy and RMSE values than Classifier A, with accuracy values of 50% vs. 37.5%, and RMSE values of 5.2 vs. 6. By using the Pareto Accuracy (PA_1), we obtain a more accurate picture in which Classifier A clearly outperforms Classifier B (87.5% vs. 50%). Here, we see a case where metrics used for both classical classification (accuracy) and ordinal classification (RMSE) do not reflect the actual performance of each classifier.

In addition to the metrics, to evaluate the performance of our model and to compare it to the schedule-based model, we also performed a spatial analysis by plotting heatmaps and a temporal analysis using hours and day of the week (see Sect. 4.3). The analysis entailed comparing boarding stops that were predicted well, i.e., at accuracies of 50% or above. Lastly, to enrich our understanding of the nature and patterns of PT, we produced and analyzed feature importance by exploiting the SHAP values method, considered as constituting a unified framework for



(a) The number of boarding and users per day.

(b) Boarding profile by days of the week.

Fig. 5 Average usage profile for Beer Sheva users

interpreting predictions based on game theory (Lundberg and Lee 2017). The values are the average of the marginal contributions across all permutations.

3.4 Procedure

We evaluated the above methodology by applying it to the smart card data of the city of Beer Sheva, Israel. With about 200,000 inhabitants, Beer Sheva is the largest city in the southern part of Israel. It presents an interesting use case given its relatively remote location, making it more isolated from a traffic perspective. Additionally, it has a sparse PT network that is easier to model. Furthermore, it has complete passenger boarding stop information, and road traffic in the city is not prone to heavy congestion. We utilized a smart card dataset consisting of over 1M records (after preprocessing, about 92% of the smart card records remained) from over 85,000 distinct travelers for one month during November and December 2018. Based on pre-analysis of the smart card data, the boarding profile per day of the week, the number of boardings, and the number of users recorded show regular patterns of use throughout this period, both for weekdays (Sun–Thu) and weekends (Fri–Sat).

As evident from Fig. 5a, b the average usage profile for Beer Sheva users is quite stable across weeks and working days (Sun–Thu). The top figure shows the boardings and users per day for the one month of data. The bottom figure shows the boarding profile by day of the week.

Next, we used a GTFS feed containing over 27,000 stops and over 200,000 PT trips in Israel for the equivalent period as the smart card data included all the operators (or agencies in GTFS tables) in the country. The dataset also included a detailed timetable for every PT trip. Lines and stops for the city of Beer Sheva were sorted by operator and geographic coordinates. All selected routes were bus lines. In total, there were about 650 stops selected in the study area.

Based on the literature, there are different reports on the sizes of data sets used from several months of data for one city (e.g., Agard et al. 2006; Hasan et al. 2013) to more common studies between one week to four weeks of data (e.g., Chu and Chapleau 2010; Munizaga et al. 2014). Usually, longer studies tend to focus on more limited scales—lines or stations or small cities. Faroqi et al. (2018) noted this problem, especially if only one week of data is used. In this case, an inherent assumption is that travels are regular between days. However, as we have shown there is also the problem of irregular travelers that are commonly discarded (e.g., in OD analysis, Munizaga et al. 2014). Given the above, we consider that one month of data is most likely sufficient for our purposes.

We also used a geospatial dataset from the municipal open GIS portal that contained a variety of geographical attributes of the city of Beer Sheva, such as traffic light locations, built-area densities, and more. We then extracted the 15 features from the above datasets. We converted the boarding stops from their Beer Sheva identifiers to numerical values (i.e., embedding). Lastly, we estimated an ML algorithm to classify the boarding stops and evaluated the classifier's performance as described earlier.

3.5 Model validation, comparative imputation methods and robustness

As mentioned, one of our primary goals was to develop a generic model that can be applied in any city. To that end, we validated our model based on the data of the peripheral city of Kiryat Gat situated 43 km north of Beer Sheva and outside of the metropolitan region. We applied the method of transfer learning (Torrey and Shavlik 2010), entailing the transfer of relevant knowledge by fine-tuning a model on a "novel" dataset, i.e., a set of data on which it did not train. Other than allowing our model to train more to prove our hypothesis, we split the data initially into intervals of 10 days for the transfer learning task and then into intervals of 20 days for the evaluation.

The main advantage of our modeling approach is that no ground truth is necessary to apply the model. This advantage is related to the fact that training is enabled without using domain-specific labels, i.e., when data integrity is poor, and no complementary data is available. We test this assertion by comparing the ML model to other possible imputation models. Such methods, specifically passenger history and temporal closeness, can, in some cases, provide very accurate predictions, mainly when data integrity is high. However, it is important to note that they have some essential limitations. The passenger history method requires passengers have multiple observations in the dataset, which is not always available when dealing with irregular travelers or to split the research data and utilize fewer data records. Additionally, the temporal closeness method is susceptible to data integrity and sparse rides. Passenger history and temporal closeness were applied using the following two algorithms:

Algorithm 1 Predicting boarding stop based on passenger history

```
for each observation $i \in S$ do
    IF $\exists H_{i,r,t}$ return $H_{i,b,r,t}$
      Else return $P_i$
end for
```

Algorithm 2 Predicting boarding stop based on temporal closeness

```
for each observation $i \in S$ do
    IF $\exists\, j\, |T_j - T_i| < 30$ return $B_j$
      Else return $P_i$
end for
```

In addition, we also evaluated a semi-random classifier as a lower end imputation method using the following algorithm:

Algorithm 3 Predicting boarding stops based on semi-random predictions

```
for each boarding stop $j$ do
    $P_j \leftarrow \frac{\sum_{i=1}^{n} B_i = j}{n}$
end for
for each observation $i \in S$ do
    Sample $B_j$
      Return $B_j$
end for
```

Where:

1. *S*—Smart Card dataset
2. $H_{i,r,t}$—is the history of passenger *i* in route *r* and time period *t*
3. $H_{i,b,r,t}$—is the most frequent boarding stop *b* of passenger *i* in route *r* and time period *t*
4. P_i—is the ML prediction for observation *i*
5. T_j—is the timestamp of observation *j*
6. B_j—is the boarding stop of observation *j*

Model robustness was validated by examining model performance on irregular passengers in comparison to the comparative imputation methods, given that simple imputation methods are ineffective when considering irregular travelers (Van Lint et al. 2005). Therefore, we examined model performance for predicting the boarding stop of one-time travelers in Beer Sheva, i.e., passengers who boarded once and did not return with PT on the same day. These observations are usually discarded because they do not contribute to OD estimation (Munizaga et al. 2014).

4 Results

The results are presented in the following order: First, we describe some properties of the data we used, showing its suitability for the developed methodology. Second, we describe the estimated ML model and its performance in comparison to the schedule-based model. Third, we analyze the performance between the two models both temporally and spatially. Fourth, we show the validation of the ML model on the use case of the city of Kiryat Gat, using transfer learning. Fifth, we compare our

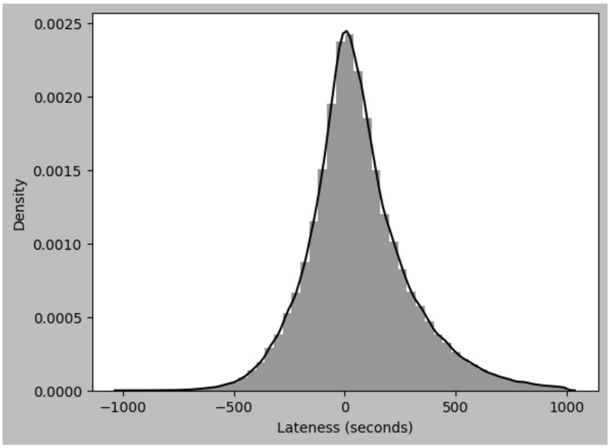


Fig. 6 The density of lateness in seconds in Beer Sheva

model to other alternative imputation model specifications mentioned earlier. Lastly, we examine prediction robustness.

4.1 Data properties

We began the analyses by exploring the processed data. First, we examined the degree of lateness in the smart card data compared to the timetable data in the GTFS feed for the city of Beer Sheva. For every PT trip, the time difference between planned and actual arrival times was computed for every stop on each line (see Fig. 6). As can be observed in Fig. 6, the density function shows both incidents of early arrival and lateness between about 500 s (8 min) early to 1000 s (16 min) late. This result suggests that the data is very suitable for applying our method. Moreover, it can be estimated that the schedule-based model using only GTFS timetable data will be less accurate.

Second, we investigated the distribution of the missing boarding stop information in the smart card data. Figure 7 presents the mean proportion of missing boarding stops per trip of the top three PT operators in Israel. This distribution is not random. If boarding stops were missing at random, the mean would be expected to be around 0 with a long tail. However, as the density function is far from that shape, we can deduce that boarding stops are indeed not missing at random.

4.2 Model training and performance

We trained several classifiers and evaluated their performances. Among the trained classifiers, the XGBoost classifier presented the best performance (see Table 2). We compared the classifiers using the common metrics as described before. Additionally, we evaluated our Pareto Accuracy metric based on error

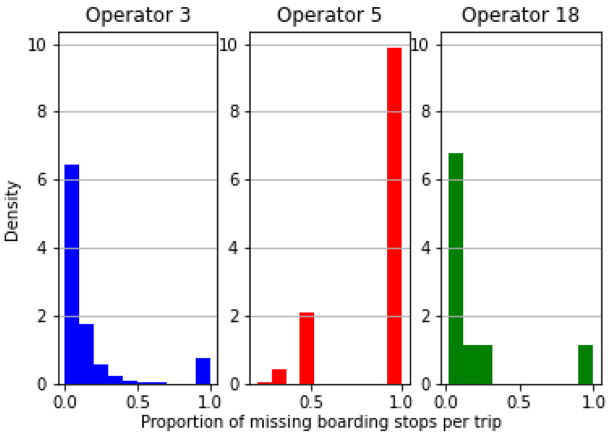


Fig. 7 The ratio of missing boarding stops per operator. Operator 3 is the largest, and 5, 18 are the second and third largest PT operators

Table 2 Classifier performances (test)

Algorithm	Accuracy	Recall	Precision	F1	AUC	PA_1	PA_2
Schedule based	0.209	0.209	0.212	0.209	0.590	0.470	0.643
Logistic regression	0.205	0.205	0.097	0.102	0.573	0.474	0.654
Random forest	0.368	0.368	0.348	0.353	0.666	0.672	0.818
XGBoost	**0.410**	**0.410**	**0.393**	**0.394**	**0.765**	**0.712**	**0.843**

The highest obtained result for each metric is marked in bold

sizes of 1, 2, i.e., PA_1, PA_2. Any larger gap would typically be deemed unacceptable in terms of level-of-service and because these error sizes are highly correlated with PA_i for $i > 2$. One significant advantage of embedding is the calculation speed, which was an average of 15.9 ± 0.023 s on about 300 K observations.

The SHAP values to evaluate the effect of each feature are presented in Fig. 8 (see also definitions in Table 1) . Here we can note: (a) by far the most important feature for the prediction is created by the predicted sequence, which shows it is highly correlated to actual patterns and is very useful for classification (i.e., schedule-based); (b) other than the first two SHAP features, the following four are temporal, which is commonsensical given that the different periods have varied impacts on traffic (such as the morning peak) and as a bus progresses along its route, stochastic events accumulate and the variance increases; (c) although geospatial features are not of the highest importance, they are not trivial, and thus, we conclude that certain physical attributes can influence the nature of our problem, e.g., denser areas can engender more congestion; and (d) the two least significant features pertain to the day of the week, from which we can assert that daily PT routines remained quite stable in our case study.

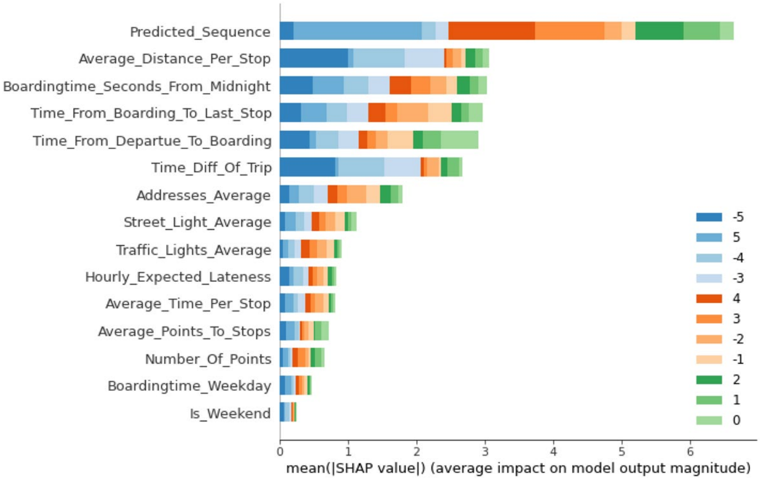


Fig. 8 Feature importance using SHAP values

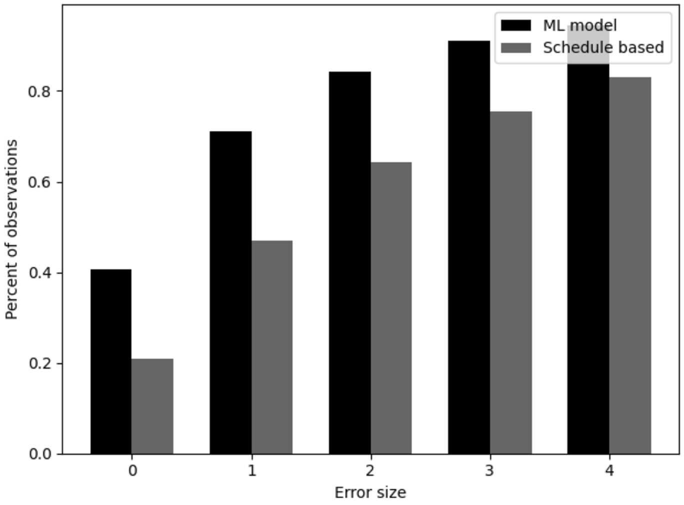
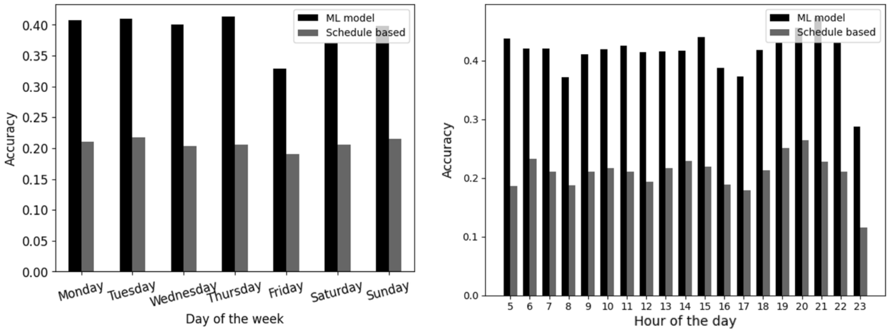


Fig. 9 Pareto accuracy comparison between ML and schedule-based models (test)

In Fig. 9, we present Pareto Accuracy between the ML model and the schedule-based one. It shows that the results are stable even for higher values than 1. Therefore, we can conclude that the proposed model outperforms the schedule-based model.

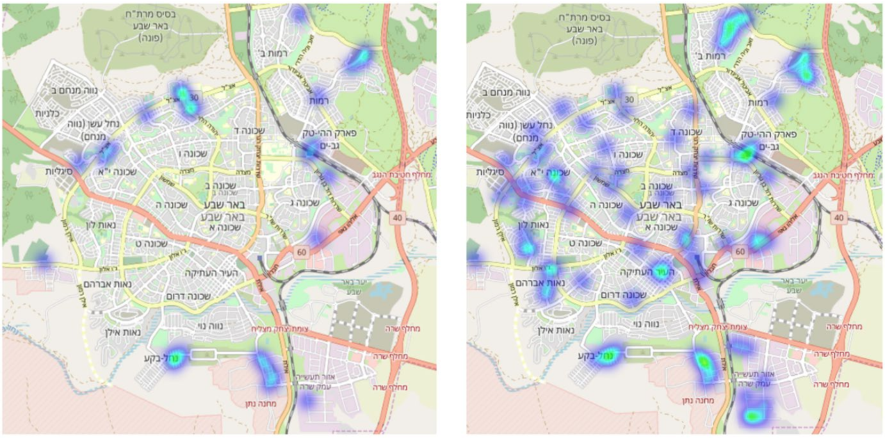
4.3 Spatial and temporal analyses

In addition to the aggregated results, we analyzed the model performance both temporally (see Fig. 10) and spatially (see Fig. 11). The temporal analysis shows that, in terms of accuracy, our proposed model outperformed the schedule-based method



(a) Daily performance (b) Hourly performance

Fig. 10 Temporal performance of models (test)—**a** daily, **b** hourly



(a) Schedule-based predictions (b) Proposed model predictions

Fig. 11 Heatmaps of boarding stops with prediction accuracy of over 50% (test)

on both a daily and an hourly basis.[3] Moreover, the spatial analysis showed similar results, and the stops where the predictions were ranked 'good', i.e., over 50% accuracy, were plotted.

Two major insights can be derived from these analyses: First, the ML model predicts considerably more stops than the schedule-based model. Second, the schedule-based model renders good predictions mainly for the central stops (train stations, main roads, or industrial zones). However, when the model is applied to non-central

[3] The hourly analysis was done on weekdays when traffic congestion makes the prediction of PT service punctuality likely more complex.

Table 3 Classifier performances for model validation

Algorithm	Accuracy	Recall	Precision	F1	AUC	PA_1	PA_2
Schedule based	0.253	0.253	0.224	0.234	0.599	0.404	0.550
Logistic regression	0.202	0.202	**0.789**	0.317	0.392	0.388	0.521
Random forest	0.221	0.221	0.594	0.246	0.578	0.423	0.588
XGBoost	**0.438**	**0.438**	0.441	**0.419**	**0.685**	**0.668**	**0.802**

The highest obtained result for each metric is marked in bold

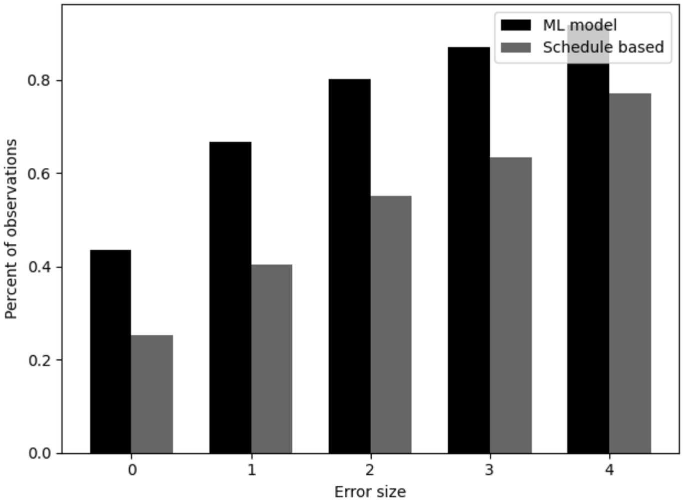
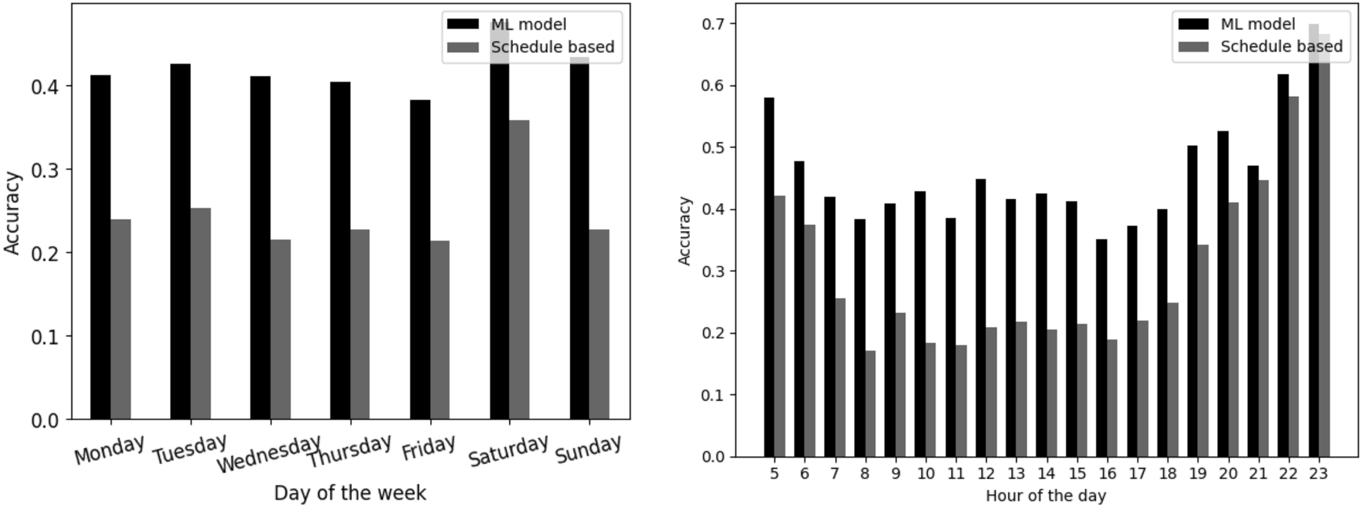


Fig. 12 Pareto accuracy comparison of models between ML and schedule-based models (validation)

locations, it is suboptimal, in stark contrast to the ML model, making good predictions across all locations.

4.4 Model validation

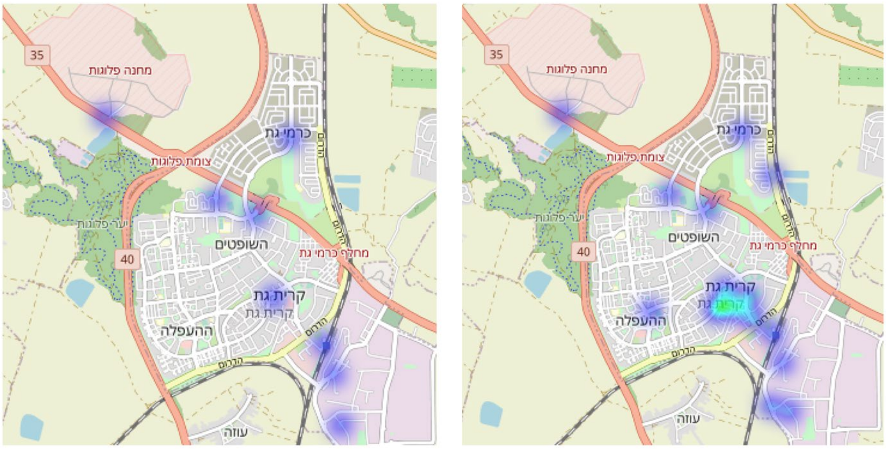
As noted, we performed the model validation for the nearby city of Kiryat Gat. Evidently, the ML model performed remarkably better than the schedule-based model (see Table 3). Figure 12 showing the Pareto Accuracy for different values of error size showing the ML model is consistently better. Figure 13 presents a performance comparison in Kiryat Gat showing the temporal analysis—accuracy by day of the week and on an hourly basis for weekdays which shows similar properties to the trained model, the ML model demonstrated higher accuracy compared to the



(a) Daily performance comparison.

(b) Hourly performance comparison.

Fig. 13 Temporal performance of models (validation)



(a) Schedule-based predictions.

(b) Proposed model predictions.

Fig. 14 Heatmaps of boarding stops with predicted accuracy of over 50% (validation)

schedule-based model. Figure 14 presents the spatial analysis revealing once more that the ML model predicts more stops with higher accuracy.

4.5 Comparative imputation methods

Table 4 shows the results of the comparisons to alternative imputation methods. While the predicted accuracy of the two alternative methods is similar, the disadvantages of the aforementioned methods are more evident in the lower share of the population than can be predicted compared to the ML model. The semi-random classifier naturally demonstrates that it is far from trustworthy in the case of hierarchical PT networks.

It is important to note that while the accuracy of our proposed method is lower, it is far more robust, both in terms of percentage of population predicted and on irregular travelers, which other suggested methods are incapable of predicting (see Sect. 4.6). For example, in predicting using historical records, we cannot predict a new passenger or a new route. For using temporal closeness, the prediction will be extremely sensitive to sparse routes.

4.6 Robustness to irregular travelers

While the personal history method can indeed be relevant as evident in Table 4, as noted above (see Sect. 3.5) model robustness was evaluated by examining performance for predicting the boarding stop of one-time travelers. As shown in Table 5,

Table 4 Results of comparative imputation methods

Method	Percent predicted (%)	Accuracy for predicted observations (%)
Proposed XGBoost	100	41
Passenger history	82	59
Temporally close passengers	52	59
Semi-random guessing	100	11

Table 5 Results of the ML model (XGBoost) on one-time travelers and passengers not predicted by comparative methods

Passenger type	Accuracy	Recall	Precision	F1	AUC	PA_1	PA_2
One-time	0.408	0.408	0.394	0.390	0.767	0.703	0.838
Not predicted by method 1	0.419	0.419	0.407	0.402	0.772	0.706	0.835
Not predicted by method 2	0.348	0.348	0.336	0.330	0.744	0.671	0.822

the results clearly show (see the first row in Table 5) that the ML model is robust and capable of predicting missing stops even for irregular or new passengers that have no historical pattern. Additionally, as noted earlier, the suggested methods are very limited. The evaluation of passengers they do not predict is clearly shown below in Table 5 (see the second and the third row).

5 Discussion and conclusions

In this study, we showed that by mining smart card data and extracting timetable data, we could construct a passenger boarding stop prediction model, which surpasses the traditional schedule-based method. Our research revealed that applying machine learning techniques improves the integrity of PT data, which can significantly benefit the field of transportation planning and operations. From the results, we can deduce the following conclusions: First, our methodology for feature extraction and machine learning model construction demonstrates several noteworthy advantages: (a) the ML algorithm generates a *generic model* that can be used with other smart card datasets since the labels (i.e., numeric representations) are always aligned in all datasets; (b) by embedding the boarding stops, our method ensures that the number of distinct labels is relatively small and a significant *computation time reduction* can be accomplished; (c) boarding stop use is inherently imbalanced, as some stops are frequently used while others are used rarely. Our proposed methodology is able to accurately classify many classes despite the inherent imbalances, thus contributing to *unpredictability reduction*; (d) the method is *data lean* and requires only mining a smart card dataset and a GTFS feed (or any compatible timetable dataset) without the need to process any other datasets; (e) the ML model is entirely *complementary* to other imputation methods including the schedule-based method as well as passenger history or temporal closeness; and (f) the method *provides a robust model* capable of dealing even with irregular or unpredictable passengers.

Second, our model (applying the XGBoost algorithm) produced the highest performance, with 41% accuracy and 71% PA_1, whereas the schedule-based method achieved only 21% accuracy and 47% PA_1. Even for larger error sizes, the ML model outperformed the schedule-based one. Moreover, the schedule-based method was able to render good predictions only for a few main stops compared to the ML model, which predicted well across all stops. This dependency on centrality was clearly visible in the spatial analysis of the stops that were well-predicted. This result confirms our conjecture that the schedule-based imputation approaches can be significantly improved by using ML methods. Furthermore, we also found that complex methods, such as ensemble, resulted in much better model performance than simple algorithms, such as logistic regression. In future research, we intend to test the performance of additional prediction algorithms, such as Deep Neural Networks (Jung and Sohn 2017; Liu and Chen 2017).

Third, from the SHAP values (Fig. 8), the following can be noted: The temporal features (created by the timetable from the GTFS feeds) are indeed crucial for the operation of the ML model. Geospatial features, however, were less important. Accordingly, we estimated a model trained without the geospatial features (see Table B.1 in Appendix B). In comparison to the richer model, the performance is somewhat worse. Therefore, we assert that such information is considered useful: Firstly, to understand patterns in a given city, for instance, which spatial attribute is more closely correlated with lateness or earliness. Secondly, it can help the transfer learning process in a new city, i.e., if the model was trained on city A, and will be used to predict city B, using the spatial features will produce a more robust model to the difference between those cities.

Fourth, we showed that the ML model is transferable (see Sect. 3.2) and able to provide strong and consistent results when validated on another city while outperforming the schedule-based imputation method. Nonetheless, our method, given its generic nature, is not entirely comparable with methods of dissimilar nature, such as those presented in Table 4 which cannot be straightforwardly transferred to another context. Since, to the best of our knowledge, no other imputation method shows such transferability, robustness, and generic nature other than the schedule-based imputation, the latter should be regarded as the comparative benchmark until another imputation method is developed.

Fifth, we recommend using our model when the lack of data does not allow for other more accurate methods to be used, such as passenger history or temporal closeness. Nonetheless, our model can complement these methods, especially for those records that are overlooked, as shown in Table 4, and thus can utilize more of the scarce data at hand. As noted, our method does not require mining or accessing any additional datasets (like AVL or APC), which are not always available and can increase the extent of errors in the prediction. This observation makes our method extremely suitable for planning purposes in non-auto-dependent and less technologically-orientated societies in developing countries and the Global South (Sohail et al. 2006).

Lastly, we introduced a new generalized accuracy metric which we named Pareto Accuracy that allows to better compare between classifiers for ordinal classification problems. This metric is more robust to outliers, easier to interpret, and accounts for the true dimension of errors. In addition, the metric is easy to implement. In the future, we hope to understand how Pareto Accuracy can improve additional ordinal classification use cases.

There are a few limitations to the study worth noting. One is that our method requires several constraints to succeed, such as timestamps, trip IDs, and existing trip timetables. These constraints potentially reduce the number of relevant datasets and the number of observations that could be imputed. However, these constraints also preclude the use of the schedule-based method; hence, in practice, our method has little effect on the ability to impute missing data. In addition, the generality of our method can increase bias, as it ignores features that cannot be transferred between datasets. These features, such as having each PT line as a categorical feature, can reduce bias when imputing a specific dataset.

Possible extensions include: predicting alighting stops (when the operator does not record TAP out), imputing other attributes of interest such as trip ID or time of day, etc. In the future, we would like to test our model in other cities to verify its generalizations. In addition, we also suggest testing the influence of transfer learning on new datasets.

Following a suggestion by one of our Reviewers, we consider it important that researchers also carry out transnational studies where models trained on data from one country are validated on similarly structured data from a least one other country to ensure geographical and cultural robustness. In addition, we suggest that researchers test the method with data from urbanities of different spatial scales to verify robustness to the public network dimensions.

To summarize, missing data imputation is a difficult and complex task. On the one hand, one wants as much data as possible for analyses, while on the other hand, data integrity is of critical importance and demands the availability of imputation methods that work well. We assert that the commonly used schedule-based method suffers from a subpar performance in terms of accuracy and other key metrics. It is highly dependent on the centrality of boarding stops. In contrast, we showed that our model outperformed the schedule-based method in all metrics over different temporal periods. It was more robust to the centrality of the imputed stops and irregularity of recorded trips. This makes it a much more suitable method for imputation as it improves data integrity. In addition, our method is based on generic classification and thus can be used in a wide variety of use cases.

Appendix A

Metrics presented in this paper:

- *Accuracy*—Percent of observations that were correctly classified
- *Recall*—The number of observations for each class that were correctly classified divided by the total number of distinct observations from this class. Final *Recall* is the weighted average of the above on all classes.
- *Precision*—The number of observations for each class that were correctly classified divided by the total number of observations that were predicted within this class. Final *Precision* is the weighted average of the above on all classes.
- *F1*—2 × (Precision × Recall)/(Precision + Recall)
- *AUC*—Area under curve (AUC) is the area under the receiver operating characteristic (ROC) curve. This curve, for each class, is the true positive rate as a function of the false positive rate. A weighted average of the areas under the curves of all classes is calculated as the AUC metric.
- *RMSE*—Root mean square error (RMSE) is a method for ordinal classification and regression. It sums the square difference from prediction to actual label, then returns the root of the above average.

Appendix B

See Table 6.

Table 6 Model performance with and without geospatial features (XGBoost)

Algorithm	Accuracy	Recall	Precision	F1	AUC	PA_1	PA_2
Without	0.409	0.409	0.392	0.393	0.770	0.712	0.842
With	0.410	0.410	0.393	0.394	0.765	0.712	0.843

Acknowledgements This research was supported by the Ministry of Science & Technology, Israel, and The Ministry of Science & Technology of the People's Republic of China (Grant No. 3-15741). We want to thank Prof. Itzhak Benenson (Tel Aviv University) for the preliminary discussions of the research question. We also want to thank Valfredo Macedo Veiga Junior (Valf) for designing the infographic illustration and Sandra Falkenstein for editing and proofreading this paper. Special thanks to Data Scientist Raz Vais, advisor to the Israeli National Public Transport Authority, for help in obtaining and processing the data. Lastly, we thank the Editor in Chief and two anonymous Reviewers for their constructive comments to our paper.

References

Agard B, Morency C, Trépanier M (2006) Mining public transport user behaviour from smart card data. IFAC Proc Vol 39(3):399–404

Al Mamun MS, Lownes NE (2011) A composite index of public transit accessibility. J Public Transp 14(2):69–87

Alguero P (2013) Using smart card technologies to measure public transport performance: data capture and analysis. Technical Report Industrial Engineering, Universitat Politecnica de Catalunya, Barcelona

Almlöf E, Rubensson I, Cebecauer M, Jenelius E (2020) Who is still travelling by public transport during Covid-19? Socioeconomic factors explaining travel behaviour in Stockholm based on smart card data. Working Paper, Integrated Transport Research Lab (ITRL), KTH - Royal Institute of Technology, Stockholm (September 8, 2020)

Anda C, Erath A, Fourie PJ (2017) Transport modelling in the age of big data. Int J Urban Sci 21(sup1):19–42

Antrim A, Barbeau SJ (2013) The many uses of GTFS data—opening the door to transit and multimodal applications. Location-Aware Information Systems Laboratory at the University of South Florida

Bagchi M, White PR (2005) The potential of public transport smart card data. Transp Policy 12(5):464–474

Batista GE, Monard MC (2003) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5–6):519–533

Ben-Elia E, Lyons G, Mokhtarian PL (2018) Epilogue: the new frontiers of behavioral research on the interrelationships between ICT, activities, time use and mobility. Transportation 45(2):479–497

Benenson I, Marinov M, Ben Elia E (2019) Is servicing commuters the goal of the public transport system? In: Geocomputation 2019. The University of Auckland

Bertsimas D, Pawlowski C, Zhuo YD (2017) From predictive methods to missing data imputation: an optimization approach. J Mach Learn Res 18(1):7133–7171

Briand A-S, Côme E, Trépanier M, Oukhellou L (2017) Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transp Res Part C Emerg Technol 79:274–289

Bryan H, Blythe P (2007) Understanding behaviour through smartcard data analysis. In: Proceedings of the Institution of Civil Engineers-Transport, vol 160. Thomas Telford Ltd, pp 173–177

Camino RD, Hammerschmidt CA, State R (2019) Improving missing data imputation with deep generative models. arXiv preprint arXiv:1902.10666

Cardoso JS, Sousa R (2011) Measuring the performance of ordinal classification. Int J Pattern Recognit Artif Intell 25(8):1173–1195

Cats O, Loutos G (2016) Real-time bus arrival information system: an empirical evaluation. J Intell Transp Syst 20(2):138–151

Ceder A (2004) New urban public transportation systems: initiatives, effectiveness, and challenges. J Urban Plan Dev 130(1):56–65

Ceder A (2016) Public transit planning and operation: modeling, practice and behavior. CRC Press, Boca Raton

Chen Z, Fan W (2018) Extracting bus transit boarding stop information using smart card transaction data. J Mod Transp 26(3):209–219

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794

Chen C, Ma J, Susilo Y, Liu Y, Wang M (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp Res Part C Emerg Technol 68:285–299

Chien H-Y, Jan J-K, Tseng Y-M (2002) An efficient and practical solution to remote authentication: smart card. Comput Secur 21(4):372–375

Chu KKA, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transp Res Rec 2063(1):63–72

Chu KKA, Chapleau R (2010) Augmenting transit trip characterization and travel behavior comprehension: multiday location-stamped smart card transactions. Transp Res Rec 2183(1):29–40

Costa AF, Santos MS, Soares JP, Abreu PH (2018) Missing data imputation via denoising autoencoders: the untold story. In: International symposium on intelligent data analysis. Springer, pp 87–98

Covic F, Voß S (2019) Interoperable smart card data management in public mass transit. Public Transp 11(3):523–548

Dacheng C, Ruizhi Y, Lei S, Kiat T, Hui D, Whye JKH, Kiong N (2018) Traveler segmentation using smart card data with deep learning on noisy labels. In: Proceedings of ACM KDD conference, vol 10, New York

Devillaine F, Munizaga M, Trépanier M (2012) Detection of activities of public transport users by analyzing smart card data. Transp Res Rec 2276(1):48–55

Echaniz E, Ho C, Rodriguez A, dell'Olio L (2020) Modelling user satisfaction in public transport systems considering missing information. Transportation 47(6):2903–2921

Faroqi H, Mesbah M, Kim J (2018) Applications of transit smart cards beyond a fare collection tool: a literature review. Adv Transp Stud 45:107–122

Fonzone A, Schmöcker J-D, Viti F (2016) New services, new travelers, old models? Directions to pioneer public transport models in the era of big data. J Intell Transp Syst 20:311–315

Frank E, Hall M (2001) A simple approach to ordinal classification. In: European conference on machine learning. Springer, New York, pp 145–156

Garg A, Naryani D, Aggarwal G, Aggarwal S (2018) DL-GSA: a deep learning metaheuristic approach to missing data imputation. In: International conference on sensing and imaging. Springer, New York, pp 513–521

Gaudette L, Japkowicz N (2009) Evaluation methods for ordinal classification. In: Canadian conference on artificial intelligence. Springer, New York, pp 207–210

Gordon JB, Koutsopoulos HN, Wilson NH, Attanucci JP (2013) Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp Res Rec 2343(1):17–24

Gudivada V, Apon A, Ding J (2017) Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int J Adv Softw 10(1):1–20

Guihaire V, Hao J-K (2008) Transit network design and scheduling: a global review. Transp Res Part A Policy Pract 42(10):1251–1273

Guyon I (1997) A scaling law for the validation-set training-set size ratio. AT &T Bell Laboratories 1(11)

Hadas Y (2013) Assessing public transport systems connectivity based on Google transit data. J Transp Geogr 33:105–116

Hagenauer J, Helbich M (2017) A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst Appl 78:273–282

Hasan S, Schneider CM, Ukkusuri SV, González MC (2013) Spatiotemporal patterns of urban human mobility. J Stat Phys 151(1):304–318

Huang J, Levinson D, Wang J, Zhou J, Wang Z-J (2018) Tracking job and housing dynamics with smartcard data. Proc Natl Acad Sci 115(50):12710–12715

Jang W (2010) Travel time and transfer analysis using transit smart card data. Transp Res Rec 2144(1):142–149

Jiao Y, Du P (2016) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant Biol 4(4):320–330

Jung J, Sohn K (2017) Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. IET Intell Transp Syst 11(6):334–339

Kandt J, Batty M (2021) Smart cities, big data and urban policy: towards urban analytics for the long run. Cities 109:102992

Khiari J, Moreira-Matias L, Cerqueira V, Cats O (2016) Automated setting of bus schedule coverage using unsupervised machine learning. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 552–564

Kieu LM, Bhaskar A, Chung E (2015) Passenger segmentation using smart card data. IEEE Trans Intell Transp Syst 16(3):1537–1548

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160:3–24

Kusakabe T, Asakura Y (2014) Behavioural data mining of transit smart card data: a data fusion approach. Transp Res Part C Emerg Technol 46:179–191

Lakshminarayan K, Harp SA, Goldman RP, Samad T (1996) Imputation of missing data using machine learning techniques. In: Proceedings of ACM KDD conference, pp 140–145

Laña I, Olabarrieta II, Vélez M, Del Ser J (2018) On the imputation of missing data for road traffic forecasting: new insights and novel techniques. Transp Res Part C Emerg Technol 90:18–33

Li H, Li F, Song C, Yan Y (2015) Towards smart card based mutual authentication schemes in cloud computing. KSII Trans Internet Inf Syst (TIIS) 9(7):2719–2735

Li T, Sun D, Jing P, Yang K (2018) Smart card data mining of public transport destination: a literature review. Information 9(1):18

Liu L, Chen R-C (2017) A novel passenger flow prediction model using deep learning methods. Transp Res Part C Emerg Technol 84:74–91

Liu Y, Zhou Y, Wen S, Tang C (2014) A strategy on selecting performance metrics for classifier evaluation. Int J Mobile Comput Multimed Commun (IJMCMC) 6(4):20–35

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, pp 4765–4774

Luo D, Bonnetain L, Cats O, van Lint H (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. Transp Res Rec 2672(8):175–186

Ma X-L, Wang Y-H, Chen F, Liu J-F (2012) Transit smart card data mining for passenger origin information extraction. J Zhejiang Univ Sci C 13(10):750–760

Ma X, Wu Y-J, Wang Y, Chen F, Liu J (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C Emerg Technol 36:1–12

Ma X, Liu C, Wen H, Wang Y, Wu Y-J (2017) Understanding commuting patterns using transit smart card data. J Transp Geogr 58:135–145

Maeda TN, Shiode N, Zhong C, Mori J, Sakimoto T (2019) Detecting and understanding urban changes through decomposing the numbers of visitors' arrivals using human mobility data. J Big Data 6(1):4

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

Milne D, Watling D (2019) Big data and understanding change in the context of planning transport systems. J Transp Geogr 76:235–244

Munizaga MA, Palma C (2012) Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp Res Part C Emerg Technol 24:9–18

Munizaga M, Devillaine F, Navarrete C, Silva D (2014) Validating travel behavior estimated from smartcard data. Transp Res Part C Emerg Technol 44:70–79

Namiot D, Sneps-Sneppe M (2017) A survey of smart cards data mining. In: AIST (Supplement), pp 314–325

Orro A, Novales M, Monteagudo Á, Pérez-López J-B, Bugarín MR (2020) Impact on city bus transit services of the Covid-19 lockdown and return to the new normal: the case of A Coruña (Spain). Sustainability 12(17):7206

Palacio SM (2018) Machine learning forecasts of public transport demand: a comparative analysis of supervised algorithms using smart card data. XREAP WP, available at SSRN

Pelletier M-P, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C Emerg Technol 19(4):557–568

Petrović N, Bojović N, Petrović J (2016) Appraisal of urbanization and traffic on environmental quality. J CO2 Util 16:428–430

Qu L, Li L, Zhang Y, Hu J (2009) PPCA-based missing data imputation for traffic flow volume: a systematical approach. IEEE Trans Intell Transp Syst 10(3):512–522

Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R (2006) Imputing missing data: a comparison of methods for social work researchers. Soc Work Res 30(1):19–31

Schmöcker J, Kurauchi F, Shimamoto H (2017) An overview on opportunities and challenges of smart card data analysis. Public transport planning with smart card data. CRC, Boca Raton, pp 2–12

Shalaby A, Farhan A (2004) Prediction model of bus arrival and departure times using AVL and APC data. J Public Transp 7(1):41–61

Singh A, Thakur N, Sharma A (2016) A review of supervised machine learning algorithms. In: 2016 3rd International conference on computing for sustainable global development (INDIACom). IEEE, pp 1310–1315

Sohail M, Maunder D, Cavill S (2006) Effective regulation for sustainable public transport in developing countries. Transp Policy 13(3):177–190

Stopher PR, Greaves SP (2007) Household travel surveys: where are we going? Transp Res Part A Policy Pract 41(5):367–381

Tao S, Rohde D, Corcoran J (2014) Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J Transp Geogr 41:21–36

Toqué F, Côme E, El Mahrsi MK, Oukhellou L (2016) Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In: 2016 IEEE 19th international conference on intelligent transportation systems (ITSC). IEEE, pp 1071–1076

Toqué F, Khouadjia M, Come E, Trepanier M, Oukhellou L (2017) Short & long term forecasting of multimodal transport passenger flows with machine learning methods. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC). IEEE, pp 560–566

Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, pp 242–264

Traut EJ, Steinfeld A (2019) Identifying commonly used and potentially unsafe transit transfers with crowdsourcing. Transp Res Part A Policy Pract 122:99–111

Trépanier M, Tranchant N, Chapleau R (2007) Individual trip destination estimation in a transit smart card automated fare collection system. J Intell Transp Syst 11(1):1–14

Van Lint J, Hoogendoorn S, van Zuylen HJ (2005) Accurate freeway travel time prediction with state-space neural networks under missing data. Transp Res Part C Emerg Technol 13(5–6):347–369

Walker J (2012) Human transit: how clearer thinking about public transit can enrich our communities and our lives. Island Press, Washington

Wang W, Attanucci JP, Wilson NH (2011) Bus passenger origin-destination estimation and related analyses using automated data collection systems. J Public Transp 14(4):131–150

Welch TF, Widita A (2019) Big data in public transportation: a review of sources and methods. Transp Rev 39(6):795–818

Yan F, Yang C, Ukkusuri SV (2019) Alighting stop determination using two-step algorithms in bus transit systems. Transportmetrica A Transp Sci 15(2):1522–1542

Yap M, Cats O, van Arem B (2020) Crowding valuation in urban tram and bus transportation based on smart card data. Transportmetrica A Transp Sci 16(1):23–42

Zhang Y, Cheng T (2020) A deep learning approach to infer employment status of passengers by using smart card data. IEEE Trans Intell Transp Syst 21(2):617–629

Zhang Y, Cheng T, Sari Aslam N (2019) Deep learning for demographic prediction based on smart card data and household survey. In: Proceedings of the 27th conference on GIS research UK (GISRUK), vol 2019. Geographic Information Science Research UK (GISRUK)

Zhang N, Jia W, Wang P, Dung C-H, Zhao P, Leung K, Su B, Cheng R, Li Y (2021) Changes in local travel behaviour before and during the COVID-19 pandemic in Hong Kong. Cities 112:103139

Zhao J, Qu Q, Zhang F, Xu C, Liu S (2017) Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans Intell Transp Syst 18(11):3135–3146

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