Data & Policy (2023), 5: e40
doi:10.1017/dap.2023.39

DATA PAPER

Analysis of spatial–temporal validation patterns in Fortaleza's public transport systems: a data mining approach

Kaio G. de Almeida Mesquita, Luan P. de Holanda Barros and Francisco Moraes de Oliveira Neto

Transport Engineering, Universidade Federal do Ceará, Fortaleza, Brazil
Corresponding author: Kaio G. de Almeida Mesquita; Email: kaio@det.ufc.br

Received: 27 April 2023; **Revised:** 25 October 2023; **Accepted:** 06 November 2023

Keywords: data mining; public transportation; smart card; spatial and temporal patterns; big data

Abstract

Understanding the spatio-temporal patterns of users' travel behavior on public transport (PT) systems is essential for more assertive transit planning. With this in mind, the aim of this article is to diagnose the spatial and temporal travel patterns of users of Fortaleza's PT network, which is a trunk-feeder network whose fares are charged by a tap-on system. To this end, 20 databases were used, including global positioning system, user registration, and PT smart card data from November 2018, prior to the pandemic. The data set was processed and organized into a database with a relational model and an Extraction, Transformation, and Loading process. A data mining approach based on Machine Learning models was applied to evaluate travel patterns. As a result, it was observed that users' first daily use has a higher percentage of spatial and temporal patterns when compared to their last daily use. In addition, users rarely show spatial and temporal patterns at the same time.

Policy Significance Statement

Understanding how users' demand for public transport varies and how their behavior in the system adjusts, favors, and supports a more assertive understanding of demand, that is, knowing the demand avoids unnecessary expenses with supply and even a reduction in the fleet or an increase in same at certain points, depending on the situation, in order to optimize demand. This is beneficial for decision makers, whether government or operating company, who will be able to develop an optimally operating network, in addition to the benefits to the user, who will have a better-quality service and possibly lower tariffs.

1. Introduction

Ortúzar and Willumsen (2011) discuss the role of transportation planning, where the supply of the system should be consistent with the demand, and therefore, understanding the users' commuting pattern and how this demand varies over time can help in the proposition of a supply planning (Mesquita et al., 2017). These patterns can vary with the influence of factors such as the type of activity performed by the user, public transport (PT) network, user's knowledge of the network, and their habits. By knowing these patterns and the distribution of demand, it is possible to perform analyses and make more assertive decisions regarding the supply and service level of the system, such as defining the most productive and attractive travel zones, predicting variations throughout the day and between days, defining future costs, proposing fare values appropriate to users' needs, and providing quality information in real-time to users, supported by available information systems (Hora et al., 2017).

The analysis of demand for public transportation systems advanced significantly as of the late 1990s when smart card payment systems were incorporated into PT systems in cities such as Washington D.C. and Tokyo, also known as Automated Fare Collection (AFC; Electronic Ticketing System—SBE), allowing fare payment (i.e., trip validation) through smart cards and reading equipment installed in the vehicles (Zhao et al., 2007; Pelletier et al., 2011; Munizaga and Palma, 2012). Fare charging has evolved from closed systems, in physical terminals, to open systems (with the possibility of validating at network stops outside the integration terminals), ensuring greater user accessibility. Besides SBE, many cities in the world have also been adopting automatic vehicle location (AVL) systems, composed of global positioning system (GPS), to locate vehicles in real-time, with a logistic aspect, but that is currently being used to understand the variability of supply and demand of PT. Other information systems cited in the literature and that have gained notoriety in the last decade are the general transit feeds specification (GTFS) and automated counters (APC). In many urban systems, including Fortaleza, variability in commuting patterns (routes, schedules, destinations, and transfers) is expected to affect how the system is used and how fares are paid. In particular, factors that can contribute are knowledge about the system (regular users), level of service (crowding and delays), purpose, and use of other modes, among other aspects (Mesquita and Neto, 2021).

In addition to fare collection, smart cards also continuously collect passenger behavior. In this way, the size of the data can become so large that it can exceed the processing capacity of conventional means, adding to the volume, the speed of extraction, the variety of the data, and the value of the information, making up the so-called Big Data of Public Transportation (BD-TP) (Kurauchi and Schmocker, 2016). It is therefore essential to treat Big Data for inconsistencies due to human error (e.g., swiping the card more than once) and equipment (e.g., not identifying the coordinates due to interference from tunnels and buildings). The explosion of data generated adds to the challenge of how to store and manage it.

Cheng et al. (2021) proposed a method in which they characterized and described two categories of users, the regular and the irregular, pointing out that the behavior in the choice of the route could and should be modeled in different ways. Thus, what sets up a commuting pattern are usually trips that go from an origin zone to a destination zone with high frequency, often through the same set of lines, with types of users, times, and reasons that are repeated over a long time scale. This characterization, although essential to define the modeling method and understand the demand, is often neglected. It is worth noting that regular users are different from users who have well-defined commuting patterns (Cats and Ferranti, 2022). Although this difference is not clear in the literature, this work will test the hypothesis that even the user, not being regular regarding the use of the system, can present a pattern or set of patterns in at least one specific day of the week, because of his type of activity.

Although there are works in the literature focused on prediction, they do not present a concern in defining and identifying the patterns and the factors that affect them, arising the need to deepen in how the systems work, besides evaluating the users' behavior in space and time during a series of validations. Thus, this work starts from the premise that understanding the PT System under analysis, from its contextualization to the "invisible" patterns present in the data, can help in modeling the reconstruction of trips and management of the PT System. In Fortaleza, some mishaps in the system that hinder the analysis of these patterns are: (i) The boarding and alighting points are not the same as the validation points (the card passes through the validator), so they are not known; (ii) there is no information whether the validation configures a transfer or short activity; (iii) the reason for the trip is not known; and (iv) the process of exploratory analysis should be preceded by prior treatment, although the studies evaluated ignore this important step (Mesquita and Neto, 2021). In this case, the pre-treatment consists of checking whether the data is suitable for assessing the problem, checking for inconsistencies such as outliers, transforming data types, and reducing dimensionality when necessary. Given the above, the motivating question of the work is presented: How to define and identify spatial–temporal patterns in Open Systems (possibility of validation and transfer outside an integration terminal), tap-on (validation right after boarding, but not at

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

disembarkation), and trunk-feeders (interchange lines that connect the terminals to the commercial center, fed by lines that connect the neighborhoods to the terminals) of PT?

The objective of this work is to identify the temporal and spatial displacement pattern of users of the Public Transportation network of Fortaleza, configured as tap-on, open, and trunk-feeder. Therefore, it was contextualized and defined what configures a user as regular and irregular. The types of patterns found to validate the hypotheses raised were also defined, varying mainly by the type of user (influence of the activity to be carried out) and travel time. The hypothesis was verified that there is a validation centroid (Spatial center point considering all validations of the same user in a time frame) representing the frequent points of access to the system by the user, which can be both identified and used to reconstruct the travel chain. To make the diagnosis feasible, data mining techniques using *k*-means, machine learning algorithms, scikit-learn libraries, numpy, pandas, and tensorFlow were reformulated for the type of problem mentioned. The article is divided into four more sections: (i) data mining for PT demand analysis; (ii) methodological proposal for pattern diagnosis; (iii) applications and results; and (iv) final considerations.

2. Travel patterns based on PT big data and data mining techniques

It is known in the literature that when starting from a dataset generated by automated systems and one wishes to understand the system characteristics for reconstructing trips, steps such as data treatment, destination inference, differentiation between transfers and activities, aggregation of point information into zonal information, and method validation are explored separately (Zhao et al., 2007; Chu and Chapleau, 2008; Chen et al., 2016; Kurauchi and Schmocker, 2016; Hussain et al., 2021).

The analysis of validation patterns is not cited as an integrated part of the TP demand analysis methods in these studies. Li et al. (2018) reviewed the literature on methods for reconstructing trips in open TP systems and classified these models as probability, trip chaining, and machine learning. In the literature, there is a wide range of work using trip chaining, based on the commuting of trips (the first validation of the day is considered the origin, while the last validation is considered close to the destination of the first trip) (Mesquita et al., 2017; Arbex and da Cunha, 2020).

The study by Trépanier et al. (2007) suggested improvements in the assumptions of the pioneering trip chaining method proposed by Barry et al. (2002), obtaining a correct destination inference rate of 66% (improving just over 10% over the original), but as a disadvantage, it had a high rate of discarded data (13%), because these held only one daily validation, and since the patterns are unknown, a route cannot be inferred from this type of user. Cats and Ferranti (2022) proposed a study to evaluate temporal mobility patterns using smart card data in the TP tap-on/tap-off system in Stockholm, Sweden. They relied on classification techniques (*k*-means) and hierarchical classification, using a Gaussian mixture model. In this way, they were able to find 10 commuting patterns ranging from regular peak-time commuters to regular early morning commuters. However, in this study, the characteristics that define regular and irregular users are not clear, and the objective is not focused on urban mobility.

2.1. Regression, classification, and clustering models

This new travel reconstruction proposal puts the understanding of travel patterns ahead of predictive modeling. The use of machine learning (ML) and statistical learning techniques proves superior to traditional trip reconstruction. This is because it does not require multiple prediction rules and assumptions, the latter being a functional analysis ML framework for dealing with statistical inference problems and finding a prediction function based on the data. In short, it is the process of extracting information from a data set, with the aim of discovering unknown properties of the data set. Some data extraction models that have been developed are association rules, classification, clustering, sequential patterns, and similarity patterns (Géron, 2019). Machine learning methods, on the other hand, focus on prediction based on known characteristics. These models can be supervised (with training data), unsupervised (grouping of data, without training data), or by reinforcement (alteration of the environment by external stimuli).

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

Finally, the models contain training, testing, and validation data sets, so that it is possible to infer the best coefficients and compatibility between variables in an optimized way in relation to other models. Its disadvantages are the dependence on large amounts of data, high levels of machine processing, and the possibility of overfitting the models, that is, the loss of generalization for new data sets (Cats and Ferranti, 2022).

2.2. Data mining algorithms

Mining algorithms can be classified according to their particular task or purpose, varying their implementations and adapting to new purposes, as is the case in these studies. Among the main techniques are: (i) Decision tree—Based on stages of decisions (nodes) and the separation of classes and subsets in a hierarchical way (e.g., CART, CHAID, and ID-3). (ii) Neural networks—Models inspired by the physiology of the brain, in which "knowledge" is the result of the map of neural connections, represented by probabilistic functions, to a large extent, and the weights of these connections (e.g., Perceptrons, MLP, and CNN). (iii) Genetic Algorithms—Optimization methods, inspired by the theory of evolution, in which with each new generation, better solutions are more likely to have "descendants" because they are better adapted (Ex). (iv) Fuzzy sets—A form of multivalued logic in which the truth values can be any real number between 0 (false) and 1 (true), moving away from Boolean logic (e.g., *k*-means and FCMdd) (Fayyad et al., 1996). In addition to some of the algorithms mentioned above, this work also used techniques from Natural Language Processing (NLP), Structured Data Association Rules, Database Normalization, Ticketing Spatialization with GPS data, and Extract, Transform, and Load (ETL). It is worth noting that the *k*-means algorithm is fundamental to the clustering of validations, comprising a process of partitioning the elements of a database into sets or clusters, in such a way that records that are similar are grouped together, differentiating them from other subsets. In this task, there are no predefined classes, but you can define the number of clusters that will be checked, as well as the verification parameters, in this case, coordinates.

3. Method for diagnosing spatio-temporal patterns of validation

The proposal of this work is directly linked to the models mentioned above (with an emphasis on Machine Learning and Statistical Learning), but for a new approach, based on Data Mining. This topic will present the method for diagnosing spatio-temporal validation patterns used in this work, but not excluding patterns relating to the number of validations per specific day of the week or activity-related issues. Figure 1 summarizes the method in its technical and phenomenological aspects with regard to the micro interactions of the stages, which can be reorganized independently depending on the future purposes of this research. It is worth noting that although the case study for applying the proposed method is for the municipality of Fortaleza, it can be expanded to other municipalities, a priori with tap-on characteristics, ticketing, and trunk-feeding, these being some of the restrictions.

The global method, proposed for pattern mining, is composed of the macro steps of Contextualization, Characterization, and Diagnosis. During the contextualization process the characteristics of the public transportation system of Fortaleza were presented, as well as information about integration and pricing (Step 01), besides the arrangement of lines and stops in the network. Subsequently, the data made available by the Urban Transportation Company of Fortaleza (ETUFOR) were presented, composing data for the year 2018, but tapering to the month of November, since it is a typical month, with low influence of seasonality, and without influence of the pandemic period (Step 02). A total of 20 databases were extracted, sorted, and grouped (according to the normalization criteria of structured databases, ensuring that there are no redundancies) for further treatment (Step 03). Among the databases are the ticketing base, GPS, GTFS, user registry, shape files of lines and stops, shape files of terminals, and vehicle identification codes between the ticketing and GPS bases, among others. Among the main transformations are the use of Natural Language Processing (NLP) models to make the names of the lines in several bases compatible, spatial extraction of coordinates from the GPS base outside the established zoning, and mining of the validations in union with the GPS using the time, date, line, direction, and vehicle identifier between the

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

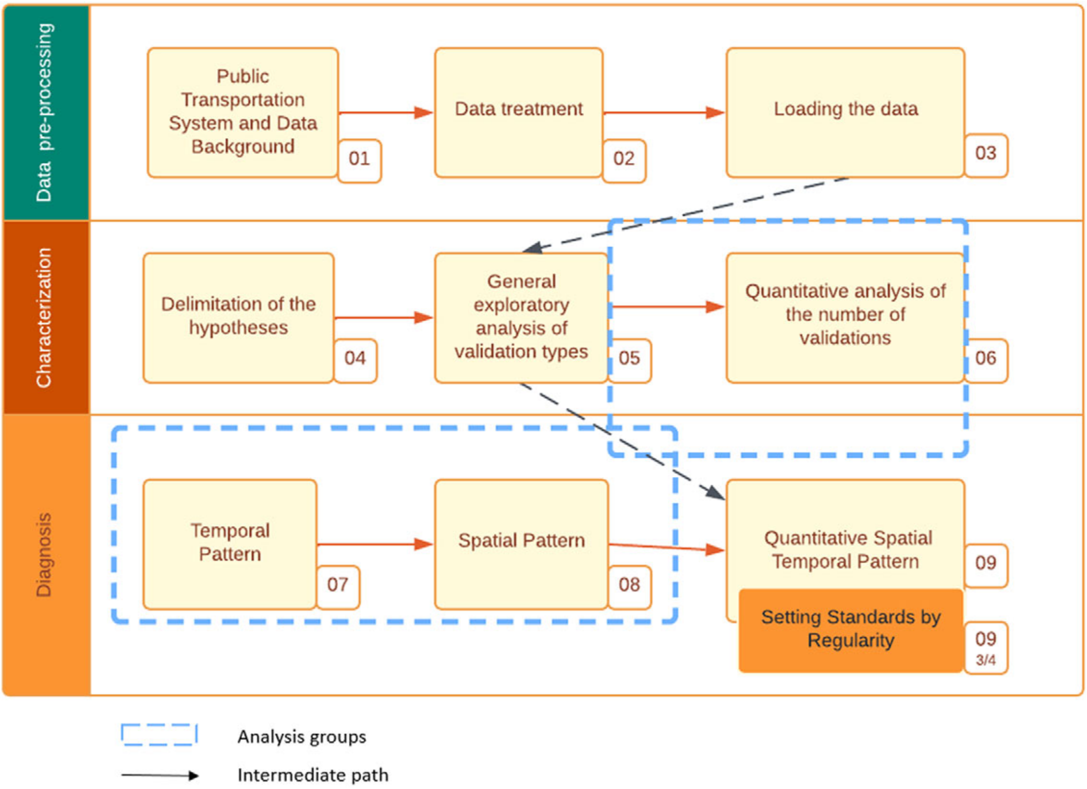
Data pre-processing
Characterization
Diagnosis
Public Transportation System and Data Background
01
Data treatment
02
Loading the data
03
Delimitation of the hypotheses
04
General exploratory analysis of validation types
05
Quantitative analysis of the number of validations
06
Temporal Pattern
07
Spatial Pattern
08
Quantitative Spatial Temporal Pattern
09
Setting Standards by Regularity
09 3/4
Analysis groups
Intermediate path

Figure 1. *Method to diagnose spatiotemporal validation patterns of public transportation.*

bases (Braga, 2019), a process that held a high computational cost because it is a Public Transportation Big Data (BD-TP) (Han et al., 2011; Kurauchi and Schmocker, 2016). Finally, finalizing the first macro stage, the algorithms were defined (Stage 04) for extracting the regularity patterns from the system.

In Macro Stage 2 (Characterization), the exploratory analysis of the validations (Stage 05) is available, as well as information about the location of the residence of registered users and more frequent lines, because it is assumed that there is an influence of the type of line, whether feeder, trunk, conventional, or complementary, since they perform specific operational activities segregating the possible patterns (Mesquita and Neto, 2021). Starting the analysis of the patterns, a quantitative characterization of the number of validations per working day and per week was established (Step 06).

Finally, the last macro stage (Diagnosis) that composes the diagnosis of spatiotemporal patterns, is divided between mining through *k*-means of the spatial differences from the centroids of validation to the respective validations that compose it, and mining of the time deviation of validations (Stages 7 and 8). The *k*-means algorithm is the ideal method for the project, as it enables a joint analysis with PCA (Principal Component Analysis). Each variable in a user group has a growth vector in relation to the components, making it easier to understand which variables most influence the characteristics of the groups. Another important point is the use of *k*-means++ to initialize the centroids in a nonrandom way. This work required modeling in Python language and SQL using the Pandas libraries for data analysis, Scikit-Learn, TensorFlow, and Numpy for modeling the grouping of validations into classes, and PyMySQL for storage and querying. *The* analyses were divided into first, last, and intermediate validations. In the validation data for each day in the month of November (average of 1,080,000 daily validations) were grouped and sorted by smartcard ID, day, and time of validation, respectively. Using daily and conditional arrays, the first and last validations for each user were separated. Inside each array was a dictionary containing as key the card identifier and as value the coordinates of the validation. The same process was repeated for the last validations and concurrently for the moments of the first and last validations. In total, it took 21 arrays

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

composing the working days, repeated between the four initial categories (First Validation—Spatial and Temporal, Last Validation—Spatial and Temporal), totaling 84 arrays. The data were saved in the relational database, composing 118 thousand valid users with validations that could be analyzed (first, last, and intermediate validations), among the 330 thousand available in the database. It is worth pointing out that the transformation process of a two-dimensional vector of coordinates into a one-dimensional vector of distances is actually a dimensionality reduction process, carried out by algebraic linear transformations. It is worth noting that only 118,000 users were used, because it is necessary to have consistency and reliability in the data, and these were the ones who had a true place of residence and user information.

The importance of analyzing the intermediate validations is due to the fact that the behavioral patterns vary according to how the user deals with the characteristics of the offer, the place where he lives, and the place to which he is moving. Thus, the remaining validations that were not configured as first or last (daily) were grouped by ID, day, and time. Validations with null latitude and longitude were disregarded, since for some cases it was not possible to identify the vehicle by GPS base.

Subsequently, after the segregated analyses for each user, the values for each working day in the month of November were aggregated by querying the database with respect to quantitative, spatial, and temporal patterns (Step 09). For each user, a categorization analysis was performed of which patterns and set of patterns it fit. Finally, a global analysis is presented and is the final product of this work, being the basis of subsequent analyses for modeling the travel chain of open, tap-on, and trunk-fed systems from the most recurrent patterns of travel (Step 93/4). Thus, some steps make up specific parts that can be evaluated differently, such as Steps 1, 2, and 3 that represent the data understanding and transformation aspect. Stages 6, 7, and 8 represent the diagnosis of the various types of patterns, while there is a strong influence of the treatment step on the patterns arranged from Stage 9, as the way the data is treated and stored governs 80% of the effort of this type of analysis and consequently influences the final work product (Géron, 2019), a relationship represented by the black arrows in Figure 1.

4. Application and results

In Fortaleza, the system is open, allowing validation during the journey and tap-on. The network follows a trunk-feeder distribution, so that the feeder lines take demand from the neighborhoods to the terminals and the trunk lines collect this demand and take it to the central regions. These central regions have a high concentration of businesses. The Integrated System of Fortaleza (SIT-FOR) has almost all the routes with fare payment by smart card, and the payment when in cash is handled directly with the driver and without the possibility of receiving change by the user, a gradual process of innovation of the system. It had more than a million daily validations (pre-pandemic), 279 regular lines, and 22 complementary lines that cover the entire city, according to the GTFS data of 2021, with a network of approximately 5650 km in length and an average of 11 km per line, with 14 companies managing the regular lines and 320 cooperative members managing the complementary lines (Mesquita and Neto, 2021). The city of Fortaleza currently has seven closed integrated terminals with GPS travel time control and two nonintegrated open terminals in the city center, with just over 5000 stops distributed in the network and an approximate fleet of 2700 vehicles (Braga, 2019). The network operates with a full fare of R$3.90, and half fare of R$1.80. Since its creation, the network has been operationally and physically integrated. Since 2013, the temporal integration was incorporated throughout the system, making it possible to make an unlimited number of transfers at any point of the network, in a period of up to two hours.

4.1. Data extraction, treatment, and compatibilization

The bases used in this work comprise the GTFS, ticketing, GPS, and user registry, as well as other complementary bases. The GTFS files used were the routes, trips, shape, stop times, and stops. The ticketing base is composed of the card identifier, line, direction, car number, date, time, and card type. The GPS base contains coordinates, date, time, and vehicle identifier. For the registration, the data refers to the user's identification (name and age), address of residence, name of the requesting company, and address of the

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

requesting company. The data used constitute 20 databases in csv, txt, JSON, and shp format. To formulate the relational model for the bank, the data was separated into three groups: I—Schedule data (GTFS); II—data passively collected by equipment in the vehicles; and III—Complementary data. Group I is composed of the five GTFS files, while Group II corresponds to the November 2018 Ticketing and GPS data. This year was chosen because it is pre-pandemic, and because it contains enough data from all the bases. Group III corresponds to user registration data, terminal and zoning shapes, and vehicle code dictionary (link between identifiers in the GPS and Ticketing bases) built from data from previous years. All the data (except the dictionaries and the terminal shapes that were formulated by the authors) were provided by ETUFOR, responsible for control, regulation, and inspection.

After creating the relational model, the relational priority was defined and, consequently, the same order should be maintained for creating the tables and loading them into the database. This relational order favors the normality of the database (a set of rules that aims to reduce redundancy). For all the bases a data cleansing was done, excluding null, duplicated, and incoherent values. The user registration, dictionary, ticketing, and GPS data, after transformation, were grouped in a single file each. Since the ticketing data did not contain identification of the validation location, it was necessary to use the dictionary that correlates the vehicle identifier in the GPS file with the car number in the ticketing file. Thus, for each validation the coordinate was identified, using as parameters the vehicle code, line, direction, time, and date of travel. In the base register, it was also necessary to identify the coordinates related to the user's home address, using the R language and the Google maps API.

4.2. *Exploratory validations analysis*

In the characterization stage, an exploratory data analysis was performed, obtaining more than one million average daily validations (2018), with around 20% corresponding to cash payments, that is, without the use of a smartcard. Given these records around 328,000 users use the system daily, equivalent to 12% of the population of Fortaleza in the year in question, and only 47% of these users are with valid registration in the ETUFOR system, being possible to obtain 110,000 reliable records about the exact place of residence of these users, according to the method presented in Section 3. Among these users 76% hold the most frequent line of use of the day as the same as the first validation of the day, and 27% have the first validation of the day near the terminals. We also evaluated the average daily, weekly, and monthly validations of each user during the month of November, identifying that 27% of them validate only once daily, while 72% validate up to twice, and 86% up to three times. While for the weekly validations, considering that the user needs at least two validations to close a daily trip chain and considering at least one integration per direction of travel, there is an acceptable limit of up to 20 validations per week, representing 97% of the users. It was possible to notice that the number of weekly validations has higher peaks at the beginning and end of the month, showing a tendency of users to use the PT more in these 2 weeks. Evaluating on a monthly scale, 52% presented up to 30 monthly validations, and 90% up to 52 monthly validations (working days only).

Finally, we verified the existence of a pattern given the number of validations and the spatialization by line type and validation order. In this work, what is considered as a pattern is related to the frequency of an act in at least one of the working days of the week in more than 50% of the analyzed cases of the same user. Thus, Figure 2 shows that users are more likely to follow a pattern on Mondays and Fridays (considering the classes of 75% and 100%), and it was evidenced that 87% of users present a frequency of the number of identical validations above 75% on at least one day of the week, that is, it is possible to show at least one pattern related to the frequency of use for almost all users if we consider that the displacement pattern can be modified by day, and not only by location and time as expected.

4.3. *Temporal and spatial pattern diagnosis*

Continuing the diagnosis, temporal and spatial analyses of each user were performed, separated between the first, last, and intermediate validations, in order to verify possible patterns that help understand the variability of demand, given the different times of day. However, following the vector of displacement at

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

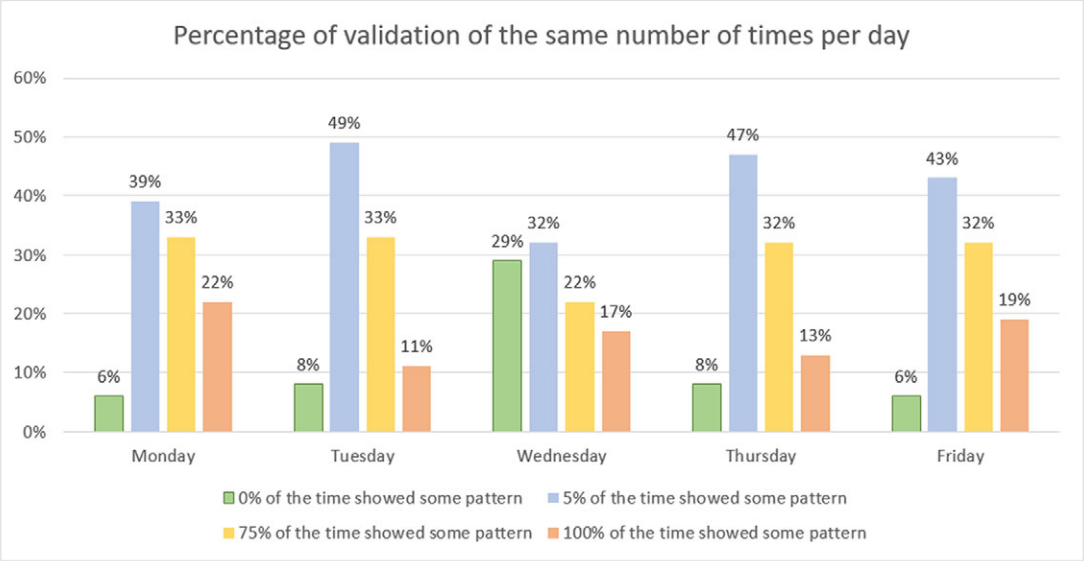

Figure 2. *Relationship of the probability of validating the same amount of times per day.*

peak hours, where users move from the peripheral regions to the commercial center in the morning and have a strong concentration of validations in the afternoon peak hour period, when they move from the commercial centers to the peripheral regions. This factor is explained by the urban segregation of large cities. We evaluated the feeder and trunk lines for student smartcards and transportation vouchers, both of which validated the previous statement. Figure 3 represents the concentration of validations by feeder lines and transport voucher, showing a concentration of spots in the peripheral region during the beginning of the day and concentration in the commercial center near Papicu terminal (Northeast) and Messejana commercial center (Southeast) in the afternoon peak. It is worth noting that the maps in Figures 3 and 4 do not show gradations (intervals), but only represent the places where these categories were validated. The more concentrated spots indicate a greater occurrence in that region (usually near terminals). The lighter spots indicate low use of the system in that region. The idea behind these analyses is simply to identify the locations with the highest concentration of users.

As found in the previous analyses, there is a direct influence of the weekday in relation to the pattern, so the spatio-temporal analyses were also performed following this logic. To calculate the spatial distances (Figure 3), as presented in the method, the validation distance between the centroid and the respective validations that compose it was verified, distances above 1000 m (Mesquita et al., 2017) were disregarded and the centroid was recalculated by an interactive process to avoid interference from outliers. Evaluating the classes of users of Transport Ticket, Students and Gratuity, they presented average distances less than 600 m for first and last validations, with gratuity having the highest variation (585 m) and Transport Ticket the lowest (528 m), contributing to the validation of the hypothesis that the reason for displacement (activity) influences the patterns, and especially, the more rigorous the schedule of the type of activity, the lower the variation of validation locations, possibly being more practical to infer the real locations of boarding and alighting of the same. It is worth noting that 60% of the users have ephemeral distances of up to 490 m, lower even than the average distance between stops in the PT network of Fortaleza (500 m). We also performed the analysis of the distribution of validation distances per cluster of each user's validations (Figure 4). This distance configures the boarding location and the validation location, considered as the stop of the user's most frequent line and closest to his residence (boarding location). The greatest distances of validation are concentrated near the terminals (red dots on the map), but students had shorter distances in these cases (blue dots), showing that these users tend to validate soon after boarding. The data used in Figures 3 and 4 refer to student cards and transport vouchers.

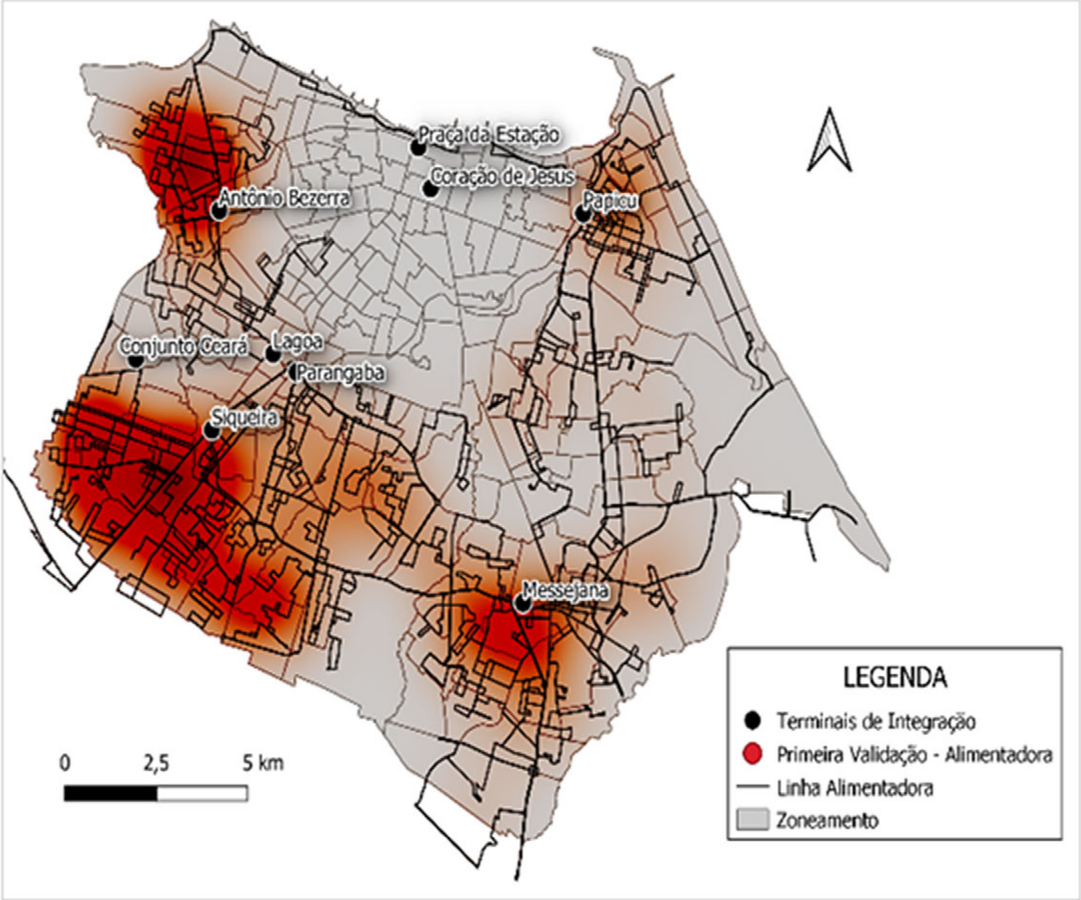

***Figure 3.** Heatmap of the first validations on feeder lines.*

Finally, the temporal distances between validations and the average times per weekday that the user usually validates were analyzed, considering a temporal deviation of up to 60 minutes for the first and last validations. The limit was established by the maximum headway of the vehicles found in the GTFS data. As for temporal distances, the afternoon peak hours presented higher values than the morning peak hour values, with a strong concentration of up to 40 s between one validation and another in the afternoon and 25 s in the morning. The inter-peak hours showed average distances of up to 100 s between validations. It is worth noting that these averages double on weekends and holidays. Another issue is that there are factors during peak hours that strongly influence the temporal validation distance, such as the crowding of vehicles during these hours. Finally, it was verified that the spatial pattern curve during the week resembles a convex arc with peak frequencies at the extremities, while for the temporal patterns, it resembles a concave curve where the peak occurs in the center of the week, more specifically on Wednesdays. Thus there is evidence that all the user categories evaluated appear to have strong spatial patterns at the beginning and end of the week, with greater temporal flexibility at these two extremes. It is worth noting that all the points in Figure 5 are on bus lines, following a route. Some validations just do not follow the validation characteristics of the majority of users, and these may be irregular us.

Figure 6 shows the spatial relationship between a regular user and an irregular user. The different colors indicate the lines most used by this user. In blue is the line with the highest usage on the first validation and in red the line with the highest usage on the last validation of the day. In the first case, the user always uses the same lines on different days of the month. The irregular user, on the other hand, uses several lines

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

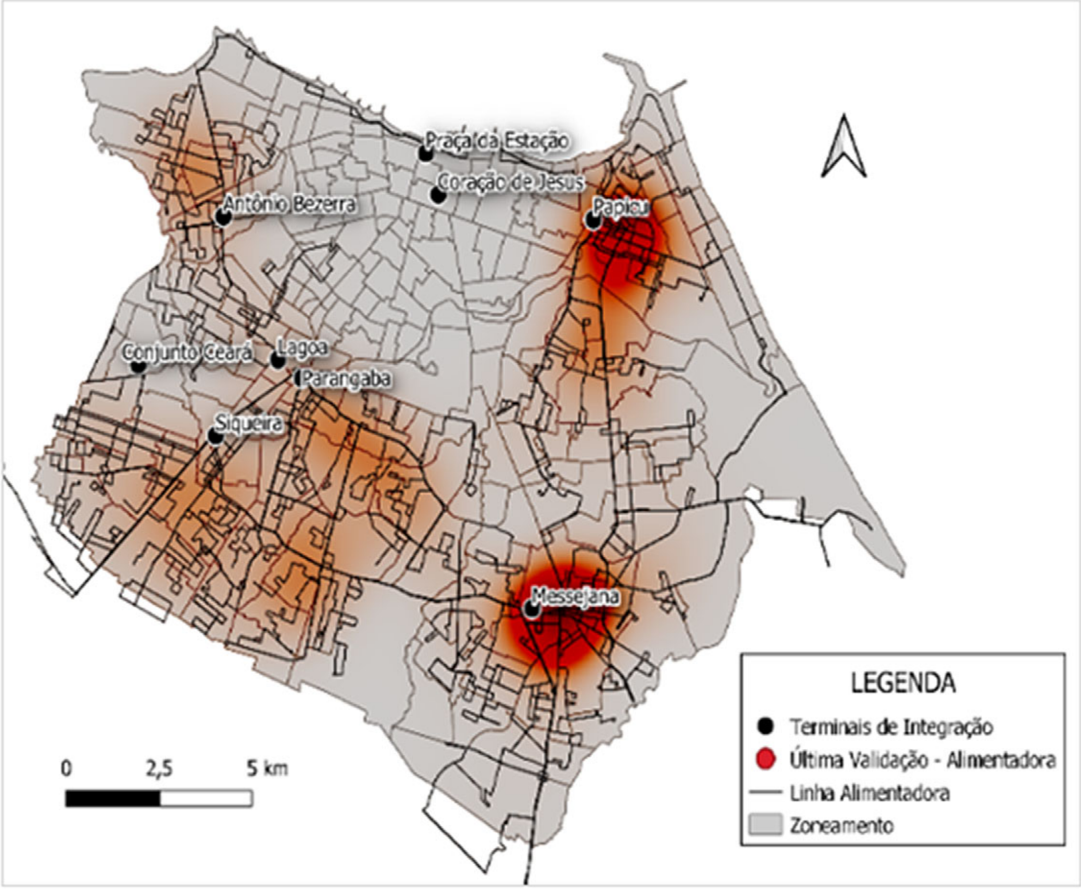

Figure 4. *Heatmap of the latest validations on feeder lines.*

throughout the month without apparently showing a pattern of behavior, at least spatially. The nonblue and red lines indicate lines with a medium frequency of use by irregular users, indicating other journeys in addition to the most frequent ones. The classification of irregular users in this context is linked to the lack of a pattern in the use of the same line for the start and end of an activity, and in many cases there is also no temporal pattern of use, making it difficult to identify a pattern in these cases. It is worth noting that this is just one representation of each class of user, using 2 different users with data from one month of validation.

4.4. Global diagnosis of patterns and the shifting phenomenon

Thus, it was evidenced that 81.2% of users present some kind of spatial pattern only in the first validation of the day and 65.2% in the last validation. While 66.9% of the users present a spatial and temporal pattern in at least one weekday in the first validation, a spatial–temporal pattern in the last validations occurs in only 20.56% of the users, that is, to close the travel chain it would be necessary to rely on a spatial or temporal pattern and not on them simultaneously, for the vast majority of users. Trip chaining techniques can be used. In addition, up to grade three of the intermediate validations were analyzed and the analyses in question were added. Thus, 48%, 33%, and 15% of the users showed some kind of pattern in the first, second, and third intermediate validations, respectively. Being the vast majority, in relation to spatial patterns, that is, users tend to perform transfers near the same location.

Patterns of use of the PT system can change based on sociodemographic characteristics. Another important point is how the operation and regulation of the system can impact on decisions about how to

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

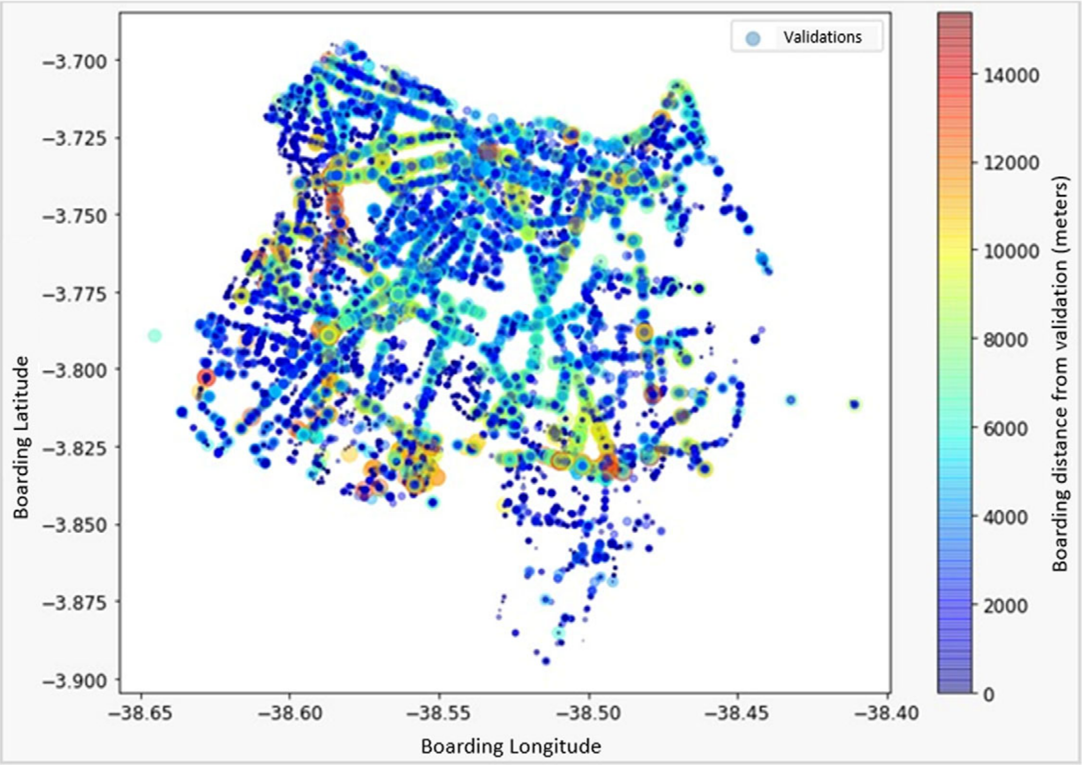

Figure 5. *Distance from first validation.*

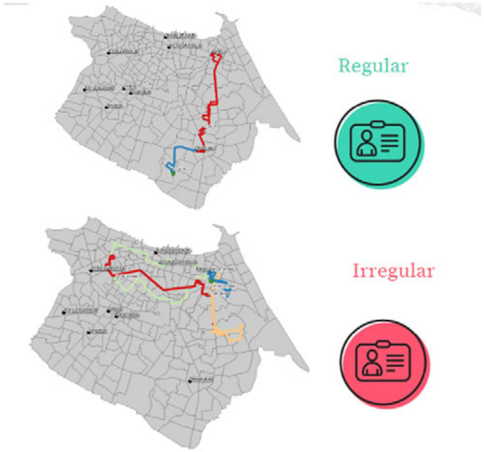

Figure 6. *Regular versus irregular user.*

get around. Temporal aspects, as presented in the text, have a strong impact on these patterns during the week, but hardly at close times.

5. Final considerations

Therefore, for this work, we conclude that demand studies can benefit from the analysis of patterns once it is known when and where the user usually validates on the network and thus would enable a more assertive supply to the system. It is also verified that the concept of frequency is closely linked to that of

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

pattern, however, the inverse cannot be said, since users may present a pattern in a specific day or time due to some not very recurrent activity, and even so cannot be considered a frequent or captive user of the system and need a different modeling when compared to frequent users. In addition, after the analysis, it was clear that to characterize the patterns of users it should be taken into account separately the spatial and temporal patterns, since it is difficult for users to have both patterns simultaneously. Among the main limitations of this work are the difficulty in obtaining and processing massive data and the partial loss of data due to problems in the charging system in identifying the card ID.

At the end of the study, we could reach the initial goal proposed for the work to diagnose the temporal and spatial displacement pattern of users of the public transportation network in Fortaleza, among the main patterns found are the users whose validations are concentrated very close to the centroid of validations at the beginning and end of the day, but their intermediate route has a high spatial–temporal deviation, as well as it was identified users with assertive patterns throughout the month, this assertiveness corresponding to the users of transport vouchers. Students presented validation patterns closer to the terminals and with shorter validation distances, since they largely use the temporal integration opportunity present in the system. Irregular patterns (in smaller numbers) were also verified in the study. For these, further study is needed, which can be supported by Machine Learning techniques for modeling the probability of validating when boarding, classification of the boarding location, and validation distance.

In future work, we intend to use the patterns found to reconstruct the public transportation travel chain, identifying through a time series the places that users usually validate, as well as the modeling of the boarding location and transfer at terminals.

Data availability statement. The authors thank Empresa de Transporte Urbano de Fortaleza (ETUFOR) for the availability of data regarding the validation of network users and GPS of fleet vehicles, which enabled the development of this research.

Acknowledgments. The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the authors are grateful to the foundation for the research incentive.

Author contribution. Conceptualization: K.G.A.M., L.P.H.B.; Data curation: K.G.A.M., L.P.H.B.; Formal analysis: K.G.A.M., L.P.H.B.; Funding acquisition: K.G.A.M., L.P.H.B.; Investigation: K.G.A.M., L.P.H.B.; Methodology: K.G.A.M., L.P.H.B.; Project administration: K.G.A.M., L.P.H.B.; Resources: K.G.A.M., L.P.H.B.; Software: K.G.A.M., L.P.H.B.; Supervision: K.G.A.M., L.P.H.B.; Validation: K.G.A.M., L.P.H.B.; Visualization: K.G.A.M., L.P.H.B.; Writing—original draft: K.G.A.M., L.P.H.B.; Writing—review and editing: K.G.A.M., L.P.H.B.

Funding statement. No specific funding exists.

Competing interest. The authors declare none.

References

Arbex RO and da Cunha CB (2020) Estimating the influence of crowding and travel time variability on accessibility to jobs in a large public transport network using smart card big data. *Journal of Transport Geography 85*, 0966–6923. https://doi.org/10.1016/j.jtrangeo.2020.102671.

Barry J, **Newhouser R**, **Rahbee A and Sayeda S** (2002) Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board 1817*, 183–187.

Braga CKV (2019) Big data de transporte público na análise da variabilidade de indicadores da acessibilidade às oportunidades de trabalho e educação. 108 f. Dissertação (Mestrado) – Curso de Engenharia Civil, Universidade Federal do Ceará, Fortaleza.

Cats O and Ferranti F (2022) Unravelling individual mobility temporal patterns using longitudinal smart card data. *Research in Transportation Business & Management 43*, 100816.

Chen C, **Ma J**, **Susilo Y**, **Liu Y and Wang M** (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies 68*, 285–299.

Cheng Z, **Trépanier M and Sun L** (2021) Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation 48*, 2035–2053. https://doi.org/10.1007/s11116-020-10120-0.

Chu KA and Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board 2063*, 63–72.

Fayyad UM, **Patesky-Shapiro G and Smyth P** (1996) From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*. Washington, DC: AAAI Press.

https://doi.org/10.1017/dap.2023.39 Published online by Cambridge University Press

Géron A (2019) *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*, Vol. *1*. Sebastopol, CA: O'Reilly Media, p. 856.

Han J, **Jian PEI and Kamber M** (2011) *Data Mining: Concepts and Techniques*. Amsterdam: Elsevier.

Hora J, **Galvão Dias T**, **Camanho A and Sobral T** (2017) Estimation of origin-destination matrices under automatic fare collection: The case study of Porto transportation system. *Transportation Research Procedia 27*, 664–671.

Hussain E, **Bhaskar A and Chung E** (2021) Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research 125*, 103044. https://doi.org/10.1016/j.trc.2021.103044.

Kurauchi F and Schmocker JD (2016) Public transport planning with smartcard data.

Li T, **Sun D**, **Jing P and Yang K** (2018) Smart card data mining of public transport destination: A literature review.

Mesquita HC, **Amaral MJ and Carvalho WL** (2017) *Matriz O/D com Base nos Dados do Sistema de Bilhetagem Eletrônica*. Recife: Congresso Nacional de Pesquisa em Transportes – ANPET.

Mesquita KGA and Neto FMO (2021) Método de Identificação dos Embarques em Viagens Big Data de Transporte Público. 35 Congresso Nacional de Pesquisa em Transportes da ANPET, assíncrono.

Munizaga MA and Palma C (2012) Estimation of disaggregate multimodal public transport origin–destination matrix from passive smart card data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies 24*, 9–18.

Ortúzar JD and Willumsen LG (2011) *Modelling Transport*, 4th Edn. West Sussex, UK: Wiley.

Pelletier M-P, **Trépanier M and Morency C** (2011) Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies 19*(4), 557–568.

Trépanier M, **Tranchant N and Chapleau R** (2007) Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems 11*(1), 1–14.

Zhao J, **Rahbee A and Wilson NH** (2007) Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering 22*(5), 376–387.

Cite this article: de Almeida Mesquita KG, de Holanda Barros LP and de Oliveira Neto FM (2023). Analysis of spatial–temporal validation patterns in Fortaleza's public transport systems: a data mining approach. *Data & Policy*, 5: e40. doi:10.1017/dap.2023.39