Using smart card data to model public transport user profiles in light of the COVID-19 pandemic

Maximiliano Lizana [a,b], Charisma Choudhury [a,*], David Watling [a]

[a] *Institute for Transport Studies, University of Leeds, LS2 9JT, United Kingdom*
[b] *Department of Civil Engineering, Universidad de La Frontera, Francisco Salazar, 01145 Temuco, Chile*

ABSTRACT

The COVID-19 pandemic caused an unprecedented impact on public transport demand. Even though several studies have investigated the change in the use of public transport during the pandemic, most existing studies where large passive datasets have been considered focus on the drop in ridership at the aggregate level. To address this gap, this research aims to identify and model profiles of passengers considering their public transport recovery after the long-term lockdown in Santiago, Chile, during the early stage of the pandemic. The methodology proposed a three-stage approach associated with the analysis of smart card records. First, cardholder residential areas were identified to enrich the available data by integrating demographic information from the census. Then, a clustering analysis was applied to recognise distinctive classes of users based on their public transport usage change between the pre-pandemic and the post-lockdown phase. Finally, two different models were implemented to uncover the relationships between class membership and travellers' characteristics (i.e. travel history and demographic characteristics of their residential area). Results revealed a heterogeneous recovery of public transport usage among passengers, summarising them into two recognisable classes: those who mainly returned to their pre-pandemic patterns and those who adapted their mobility profiles. A statistically significant association of travel history with the mobility adaptation profile was found, as well as with aggregate socio-demographic attributes. These insights about the extent of heterogeneity and its drivers can help in the formulation of specific policies associated with public transport supply in the post-pandemic era.

1. Introduction

The outbreak of COVID-19 in the world caused a significant change in people's mobility patterns as a result of people's fear of the virus's consequences, government measures, changes in transport provision and the emergence of new trends, such as teleworking and online shopping (Abdullah et al., 2020; Bin et al., 2021; Zannat et al., 2021). Although the existing literature suggests that the demand for all modes was affected by the pandemic, the evidence shows that public transport usage was the most negatively impacted (Przybylowski et al., 2021; Vickerman, 2021; Wielechowski et al., 2020).

Many studies to date have investigated the impact of COVID-19 on travel behaviour, focusing on the consequences on mode choices and risk perception (Abduljabbar et al., 2022). The evidence suggests that public transport has lost attractiveness while people prefer individual modes such as private car and non-motorized modes (Eisenmann et al., 2021). The negative perception toward public transport has also been

associated with high contagion risk and an increase in crowding aversion (Kolarova et al., 2021). Most of the current analyses have, however, been conducted using online surveys, either cross-sectional (Bucsky, 2020) or considering a limited number of waves (Beck and Hensher, 2021; Molloy et al., 2021). Additionally, studies where passive data have been used have focused mainly on drops in ridership (mostly in aggregate levels) without exploring the linkage with the characteristics of the individual, their travel history and/or spatial attributes (Abduljabbar et al., 2022). Due to this limitation, the characterization of the recovery in mobility patterns of public transport users that continued travelling after lockdowns remains limited.

This prompts this research, where we aim to identify and characterise profiles of public transport passengers who continued travelling after a critical disruption in mobility caused by a long-term lockdown, considering the recovery in their public transport usage. Therefore, we hypothesise that in response to the COVID-19 pandemic and associated restrictions, groups of passengers have experienced heterogeneous

---

* Corresponding author.
E-mail addresses: tsmel@leeds.ac.uk (M. Lizana), C.F.Choudhury@leeds.ac.uk (C. Choudhury), D.P.Watling@its.leeds.ac.uk (D. Watling).

changes in their travel behaviour. Moreover, we postulate that adopting a particular mobility profile in the post-lockdown period can be explained by the characteristics of the travellers – their pre-pandemic and lockdown travel history and attributes of the home location. A three-stage approach was proposed to describe and model public transport users' profiles based on an analysis of smart card data for Santiago de Chile.

The research thus aimed to expand the findings of previous works related to the impact of COVID-19 on individual public transport use by:

1. Using individual-level smart card data records with extensive coverage over the population to study the changes in public transport usage of those who continued travelling in a post-lockdown phase.
2. Proposing a comprehensive set of indicators to describe passengers' public transport usage change between the pre-pandemic and the post-lockdown.
3. Revealing hidden mobility adaptation profiles of frequent pre-pandemic passengers that illustrate the variability in the public transport usage recovery.
4. Associating explanatory factors to each profile to obtain insights as to which policies are most suitable for implementation in public transport systems in a post-pandemic era.

The remainder of this paper is structured as follows. First, in Section 2, an overview is given of the impact of COVID-19 on public transport and the role of smart card data in travel behaviour analysis. Then, Section 3 describes the data used, including a description of the context of the pandemic in the period analysed associated with the study case. The methodology followed in this study is described in Section 4, divided into three subsections: residence estimation, clustering analysis and modelling. Section 5 presents the main results, followed by the conclusions in Section 6.

2. Literature review

2.1. Impact of COVID-19 on public transport usage

The COVID-19 pandemic has had substantial impacts on human mobility. The effect of COVID-19 on public transport ridership, in particular, was dramatic, with the greatest reduction during the lockdown periods. In fact, during the most challenging periods of the pandemic, the drop in ridership was as much as 70%-90% in the major cities of Sweden (Almlöf et al., 2021), Germany (Kolarova et al., 2021), Belgium (Tori et al., 2023), Greece (Politis et al., 2021), Chile (Gramsch et al., 2022), US (Wang and Noland, 2021) and Hungary (Bucsky, 2020). However, although the COVID-19 pandemic has disrupted all forms of travel (Eisenmann et al., 2021), trip reductions have not been the same for all transport modes. The existing evidence indicates a significant shift of commuters from public transport to individual modes such as private car and non-motorised modes (Abdullah et al., 2020). For example, Bucsky (2020) reported that the modal split of public transport decreased from 42% to 18% in Budapest, while private car usage increased from 43% to 65%. Kolarova et al. (2021), using an online survey applied in Germany in April 2020, also reported a significant shift from public transport to private modes. The evidence shows that despite the lifting of mobility restrictions and the success of several vaccination campaigns worldwide, passengers have remained reluctant to use public transport services again (Almlöf et al., 2021). Some of the causes which have been associated with this behaviour have been the perceived contagion risk (Przybylowski et al., 2021), the fear of the virus's consequences (Abdullah et al., 2020), and the changes in people's time use due to the pandemic adaptations related for example to teleworking and online shopping (Bin et al., 2021; Zannat et al., 2021).

Although many studies to date have investigated the impact of COVID-19 on travel behaviour, most of them have been conducted using online surveys, either cross-sectional (Bucsky, 2020; Kolarova et al.,

2021) or considering a limited number of waves (Beck and Hensher, 2021; Molloy et al., 2021). Such online surveys typically have small sample sizes, have a limited capability to capture the day-to-day variability in people's mobility, have not been particularly focused on public transport and rely on respondents' memories to reconstruct pre-pandemic travel patterns. On the other hand, passive data sources such as smart cards, GPS traces and mobile phone records, which have digital mobility footprints of many people over time, can help overcome those limitations (Zannat and Choudhury, 2019), complementing the analyses of people's mobility adaptation through the COVID-19 outbreak. In particular, several studies have implemented smart card data to analyse at an aggregate level (system level, by area or station) the change in the public transport demand caused by the pandemic (Fernández Pozo et al., 2022; Jenelius and Cebecauer, 2020; Rodriguez Gonzalez et al., 2021; Zhang et al., 2021). In comparison, only a few attempts to study the impact of COVID-19 at an individual level considering smart cards have been carried out. Two exceptions are Almlöf et al. (2021), who studied the propensity to stop travelling during the pandemic in Stockholm, and Carney et al. (2022), who focused on accessibility issues on senior cardholders of the West Midlands, England, between 2019 and 2020. Then, the characterization of the recovery in mobility patterns of public transport users that continued travelling after lockdowns remains limited.

2.2. Passenger profiling using smart card data

Smart card data has become a reliable and extensive data source to analyse travellers' travel behaviour and improve public transport planning (Pelletier et al., 2011). Many large-medium cities in the world have implemented Automatic Fare Collection systems (AFC) to collect public transport payments, but also to analyse the public transport travel demand (Kusakabe and Asakura, 2014). Smart cards automatically and continuously store each fare payment of transit users and associate it with an ID card. IDs are unique numbers given to smart cards that allow the study of travel habits, trip sequencing and route preferences, among other characteristics (Pelletier et al., 2011). Each fare payment usually saves information about the card ID, timestamp, service number, card type and fare. In this way, it is possible to use smart card data to study travel demand changes, and to identify anonymously public transport users in different periods, which is a significant advantage compared with traditional data sources (Zannat and Choudhury, 2019).

In previous studies, user profiling has been carried out with smart card data considering passengers' interpersonal and intrapersonal travel behaviour to reveal unseen patterns (He et al., 2018). That exploration has usually been implemented with non-traditional transport models, such as machine learning techniques to classify users depending on their public transport frequency use (Briand et al., 2017). The literature shows that methods such as hierarchical clustering analysis and K-means have been widely implemented on smart card data to group cardholders based on their trip regularity. For instance, He et al. (2018) and El Mahrsi et al. (2017) used smart card data to classify public transport users depending on their trip frequency. Clustering techniques can also be implemented to group cardholders regarding their spatial–temporal trip patterns (Egu and Bonnel, 2020).

3. Data

3.1. Case study

Smart card data from Santiago de Chile at the individual level were available for this study. Santiago's public transport involves a complex system that integrates urban buses, the underground (that is called Metro) and an inter-urban rail. The system serves a population of around seven million inhabitants, with 4.5 million transactions daily before the outbreak of COVID-19. The system consists of around 7.000 buses, more than 10,000 bus stations, 379 bus routes and seven metro lines with 136
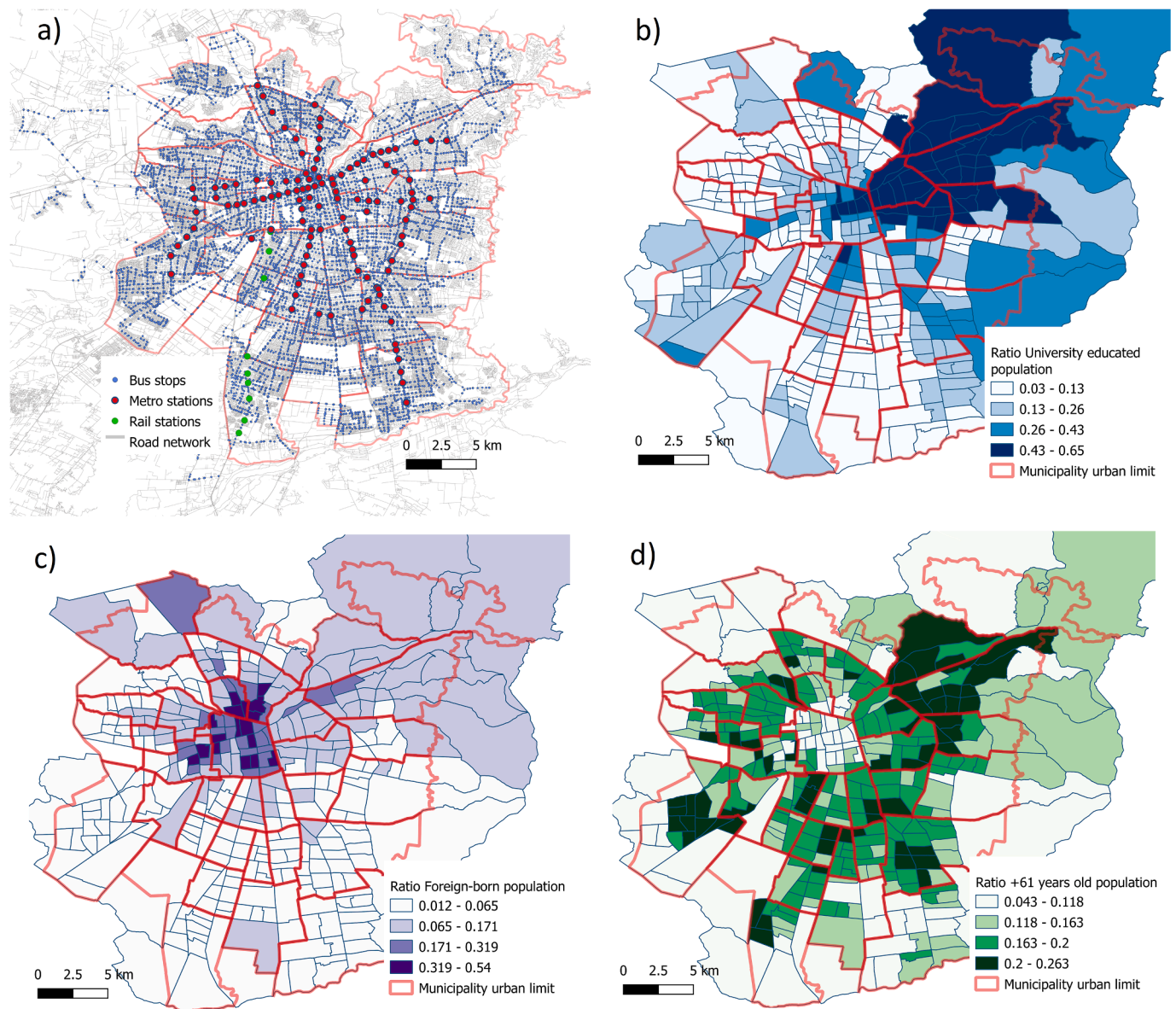
Fig. 1. Spatial distribution in Santiago's Metropolitan area by census district zone of a) Public transport stops/stations, b) Ratio of university-educated population, c) Ratio of foreign-born population, and d) Ratio of the elderly population.

stations and a length of 140 km. Fig. 1 shows the spatial distribution of bus stops and metro/rail stations and three sociodemographic characteristics of the population in the metropolitan area of Santiago across 352 census district areas considering 34 municipalities. A smart card (called bip!) is the only payment method accepted in Santiago's public transport system. Transaction information is recorded and associated with a unique anonymous ID card. Tapping in the card is requested only to board public transport modes, at which time passive data are recorded, such as the card ID, timestamp, and bus service/metro station. The smart card system of Santiago does not gather information about the alighting stops. Instead, the methodology developed by Munizaga and Palma (2012) is applied to infer alighting information. That method identifies alighting locations following the trip chain of an ID card during the day and examining the position and time of the boarding. Then the alighting stop of a trip is estimated considering the boarding position of the next transaction through the minimization of a generalized travel time function. Adult cards are not customized. Hence, they

may eventually be shared among multiple users. It may be noted that bus fare evasion has been recognized as an issue for Santiago's authorities. Therefore, the smart card data may provide a conservative estimate of the ridership in the Santiago public transport system.

3.2. The COVID-19 pandemic in Chile

The first case of COVID-19 in Chile was confirmed on 3 March 2020, and the Chilean government applied the first measures to face the spread of the virus on 16 March 2020. The first lockdown was implemented in Chile on 26 March in seven municipalities of Santiago, and during the entire pandemic, this measure was applied at a municipality level, avoiding implementing a national lockdown. Under this strategy, each municipality could enter or exit a lockdown depending on the number of new cases confirmed and the availability of critical care beds (Bennett, 2021). Even with the implementation of this tactical strategy to tackle the spread of the virus, the number of new cases and deaths increased
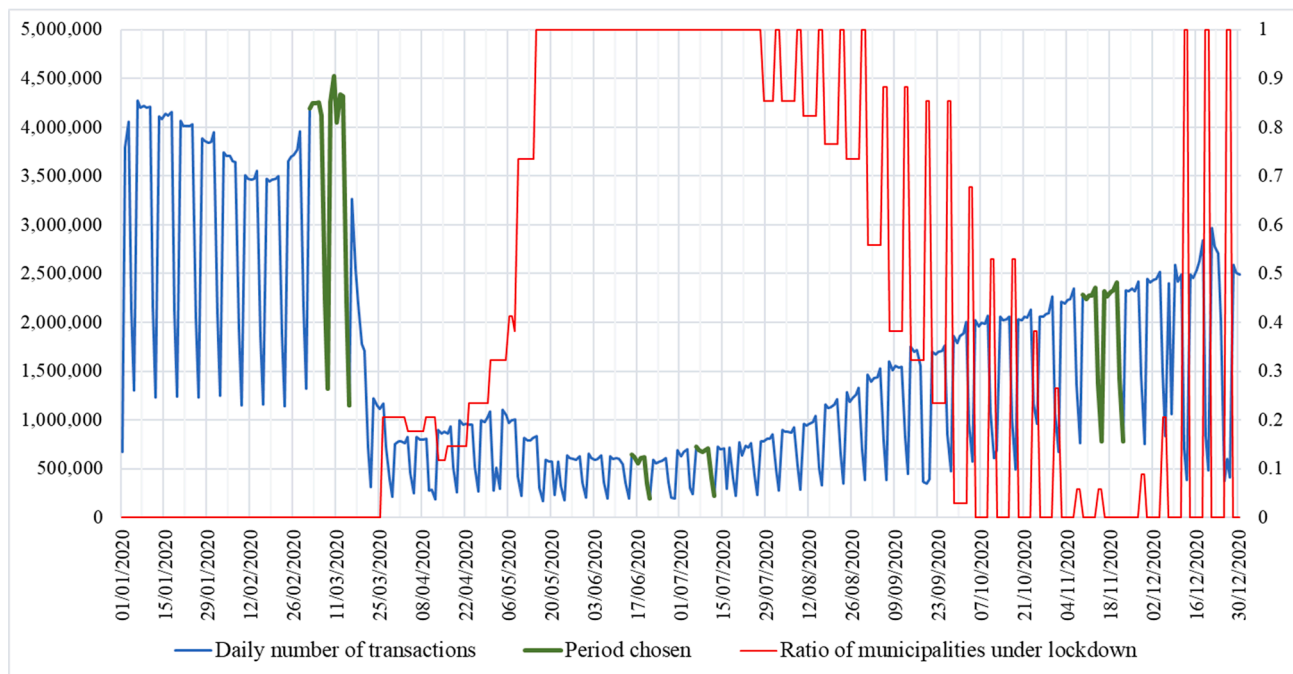
Fig. 2. Daily variability of public transport demand and lockdowns in the metropolitan area of Santiago during 2020.
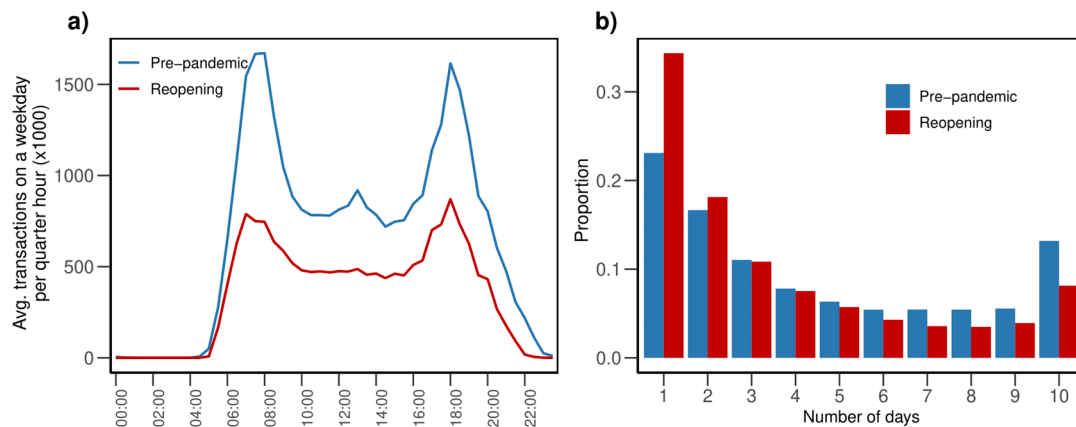

Fig. 3. a) Ridership distribution on weekdays, per period analysed (average values every 30 minutes). b) Proportion of cardholders regarding the number of weekdays travelled.

sharply. Then, the authorities decreed a total lockdown for Santiago on 15 May 2020; this unified lockdown lasted until 27 July, when the first municipalities were released (see Fig. 2, red line, for lockdown progression in Santiago's municipalities). The same month the government announced the "step-by-step" strategy, establishing five possible phases for municipalities depending on the outbreak's severity. Phase 1 meant total lockdown, Phase 2 lockdown only on weekends, while Phase 3 to Phase 5 meant the end of lockdowns but continuing with restrictions at different levels (Villalobos Dintrans et al., 2021). Thereby, on the last days of July, the first municipalities in the Metropolitan area of Santiago started to transition from Phase 1 to Phase 2. Gradually other municipalities followed the same trend. Therefore, many of Santiago's municipalities were still under lockdown on weekends between August and September. This situation is depicted in Fig. 2, where the share of municipalities under lockdown spiked every weekend during the second half of 2020. Eventually, by 5 October, all of Santiago's metropolitan area had been lifted from Phase 1, being municipalities in Phases 2 and 3. From 16 November to 27 of the same month, no lockdowns were in place; however, substantial restrictions were still present (a curfew,

face-to-face classes were still not allowed, gyms and events were not permitted to open yet, mandatory use of face-mask and social distancing protocols were active, among others). Chile's mass COVID-19 vaccination campaign would start only in February 2021, and Santiago's Metropolitan area would enter new full lockdowns during 2021.

3.3. Study period

Following the aim of this study, homogeneous periods were identified during 2020 to characterise passengers' PT usage recovery, in particular of those travellers that were active during the pre-pandemic and after the lockdown. Fig. 2 illustrates the variation of the two factors used to identify the appropriate study period: the share of the municipalities of Santiago's metropolitan area under lockdown and the daily variation of public transport demand. Thus, three key periods of 2020 were chosen: pre-pandemic (PP), lockdown (L) and reopening (O). Regarding the extension of each period, although the literature has considered one week, such as a minimum unit to observe a cycle related to travel behaviour, we decided to use two weeks. Thus, smart card data
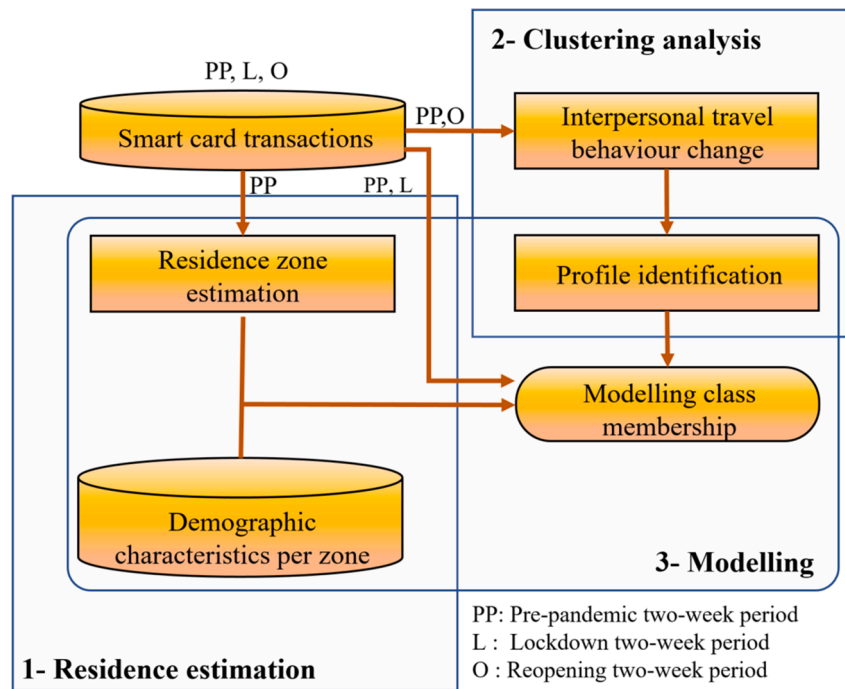
Fig. 4. Flow chart with the three-stage approach implemented in this study.

records of Santiago de Chile's public transport system between March 2–15 were used to illustrate pre-pandemic public transport use, data from June 15–21 and July 6–12 for the lockdown period, and transactions between November 9–20 for the reopening period. For the lockdown, two non-consecutive weeks were chosen to capture any natural between-month variability in this period. The reopening period chosen is still a settling-in time for urban mobility. Mobility and, in particular, public transport ridership continued changing highly during 2021 as a consequence of new waves of the virus that were tackled with new full lockdowns enacted in the metropolitan area. During the reopening, many offices continued teleworking, some called their employees back to face-to-face work, and others adopted a hybrid scheme. This means mobility was significantly lower during this period than in the pre-pandemic. In fact, movement trends provided by Google indicated 41.5% lower activity in workplace locations during the reopening compared with the pre-pandemic weeks.

The progression of the overall public transport demand in Santiago during 2020 is shown in Fig. 2, displaying a massive reduction in the use of the system after the start of the outbreak. In fact, the demand reached an average of 4.3 m transactions on weekdays during the pre-pandemic, but in the total lockdown, a barely daily average of 0.6 m transactions was recorded. As the lockdowns were eased, the public transport demand started to recover, reaching a plateau around the reopening period, with an average of 2.3 m transactions registered on weekdays. On the other hand, most of the services of Santiago's PT system operated in the reopening almost at the same frequencies compared with the pre-pandemic weeks. Minor adjustments were implemented in specific services to strengthen frequencies during peak hours and reduce them in periods of low demand, particularly associated with the metro operation. The recovery of the frequency of services after the reduction implemented during the lockdown was supported by authorities even though the high drop in ridership to ensure social distance protocols and give reliability to users in terms of the level of service of public transport.

Fig. 3a illustrates the differences between the trip distribution during business days (Monday to Friday) for the pre-pandemic and reopening periods for the overall demand. Differences are evident not only in terms of the number of trips but also in terms of their distribution. Morning and evening peaks were displaced (passengers carried out their morning

trips later and the return ones earlier), and the difference in the demand between peak and out-of-peak hours were reduced. Also, the noon peak, a typical characteristic of Chilean cities, almost disappeared. In addition, Fig. 3b shows the proportion of cardholders regarding the number of weekdays travelled by period. The graph displays that during the reopening period, the proportion of passengers that travelled only one or two days in the two-week window increased compared with the pre-pandemic period, while the proportion of cardholders that travelled more than two days declined.

4. Method

4.1. General framework

A three-stage approach was proposed to identify and model profiles of public transport users who continued travelling after the lockdown based on their travel behaviour recovery (Fig. 4). The first stage considered the enrichment of smart card data through the estimation of the residential area of cardholders and the imputation of aggregate demographic characteristics from the Chilean Census, using the pre-pandemic period records. Secondly, seven indicators were proposed to measure the intrapersonal variability of public transport usage between the reopening phase and the pre-pandemic period. Then, the K-means algorithm was applied to identify discrete recovery profiles by splitting cardholders into classes with more homogenous public transport recovery. Finally, Gradient Boosting Decision Tree (GBDT) and logistic regression model (LRM) were applied to relate explanatory variables to the previously-identified clusters. Variables such as individuals' travel history during pre-pandemic and lockdown, card type and aggregate demographic characteristics were used to explain class membership.

4.2. Residential zone estimation and demographic characteristics per zone

As Santiago's public transport system does not collect users' socioeconomic information, we used the socioeconomic characteristics of the predicted home location of the cardholders as a proxy of user characteristics. This information was retrieved from the Chilean Census, that in 2017 gathered sociodemographic data across the country through

Table 1
Explanatory variables used to model cardholders' class membership.

Dimension	Variable	Description
Travel history lockdown (THL)	Lockdown trips	Total number of weekday trips associated to each card during the lockdown period.
Travel history pre-pandemic (THPP)	PP trips - weekdays	Total number of weekday trips associated to each individual card during the pre-pandemic period.
	PP trips - weekend	Total number of weekend trips associated to each individual card during the pre-pandemic period.
	PP days travelled weekdays	Number of different weekdays with trips associated to each card during the pre-pandemic period.
	PP avg. travel time per trip	Average travel time per trip associated to each individual card during the pre-pandemic period.
Demographic characteristics of the traveller (CT)	Senior card	Dummy. 1 if the card is a senior card, 0 if it is an adult card.
Characteristics of the traveller's residential location (CRL)	Share - Women	The ratio between the women population and the total population, per CDZ.
	Share - Age < 13	The ratio between the < 13 years old population and the total population, per CDZ.
	Share - Age + 60	The ratio between the + 60 years old population and the total population, per CDZ.
	Share - Foreign born	The ratio between the foreign-born population and the total population, per CDZ.
	Share - Students	The ratio between the population that declared to be students and the total population, per CDZ.
	Share - University educated	The ratio between the population that have a university degree and the total population, per CDZ.
	Share - Workers	The ratio between the population that declared do paid work and the total population, per CDZ.

household surveys. Information such as gender, age, educational level, employment and migrants can be spatially analysed at three levels of aggregation: blocks, census district zones (CDZ) and municipalities. After analysing the three levels, we chose CDZ, as it offers an intermediate spatial resolution of the population characteristics of the metropolitan area and matches better with the criteria used in the residential location procedure. A total of 352 CDZ for the metropolitan area of Santiago were considered. Table 1 describes the aggregate sociodemographic variables, estimated as the ratio between a target population and the total population for a particular CDZ. These shares should be interpreted as a characterisation of the area where a cardholder lives instead of individual demographic conditions. This approach is particularly appropriate for Santiago's context due to its elevated level of urban and social segregation that causes a high homogeneity in demographic characteristics within neighbourhoods (Gainza and Livert, 2013).

To associate sociodemographic information of the CBZ, the potential residential location of cardholders must be found. We adapted the methodology implemented by Amaya et al. (2018), who proposed to estimate the residential location of a cardholder as the centre of gravity of the coordinates associated with the first transaction of each day, by implementing the DBSCAN algorithm (Ester et al., 1996). DBSCAN is a clustering technique whose advantage on residential estimation is the recognition of outliers. The algorithm was applied over the spatial coordinates of the first trip's boarding coordinate of each day throughout the two pre-pandemic weeks only to those cardholders that carried out trips for at least three days in that period. As parameters, we used 1 km as the maximum distance between two coordinates to be considered part

of the same spatial cluster. This value reflects a walkable distance between cardholders' real residence and their reachable bus stops. At least 40% of the total first boarding coordinates were required to make up a residential cluster. Fig. A1 summarises the steps followed to estimate the residential location of cardholders in this work using smart card transactions. As a final step, the gravity centre of the boarding coordinates of a certain cardholder that only present one residential cluster is assigned to a unique CDZ.

4.3. Clustering analysis

The second stage involved the clustering of cardholders based on the change in their public transport usage between the pre-pandemic and the reopening period. Here, three steps were followed: data processing, estimating intrapersonal travel variability and clustering considering interpersonal differences.

4.3.1. Data processing

Even though disaggregate smart card data is a rich data source to study public transport demand patterns, the literature also recognises the need to include a data processing step to analyse and clean such data (Ordóñez Medina, 2018). In the present study, to obtain suitably cleaned data, a sequence of criteria were considered as determined by the study's goals, the data quality and the pandemic context. The cleaning criteria are listed below; we have also included the number of remaining ID cards after successively applying each criterion.

- Keeping cards that only were active on both pre-pandemic and reopening weekdays (1,385,711 cards).
- Only adults and elderly cards were analysed (1,028,460 cards).
- Removing card IDs with multiple tap-ins (947,800 cards).
- Cardholders at least carried out trips on three different days during the 14-day period in the pre-pandemic weeks (415,762 cards).
- Only cards with an estimated residential location remained (379,115 cards).
- ID cards with no imputed information at all about the alighting stops were removed (360,190 cards).

First, validation records were analysed to identify active cards during the pre-pandemic and the reopening weeks. 3.9 million different cards were active during the pre-pandemic and 2.7 million during the reopening period. However, the analysis found that from the total cards active during the reopening, only 1.38 million cards could be traced to the pre-pandemic period. We hypothesise that the remaining corresponded to travellers that renewed their cards between both periods (possible causes could be the loss or damage of the card) and to the arrival of new travellers. Therefore, for non-traced cards, there was no way to infer whether a user had lost/ changed cards or discontinued using PT. To overcome this data limitation, we focus on those cardholders that are traceable between the pre-pandemic and the reopening period. An example where a comparable approach is considered is Egu and Bonnel (2020), who applied a similarity analysis strictly on traceable public transport users.

Invalid records are recommended to be filtered (Gong et al., 2017). Consequently, cards that were validated more than once in a very short time were removed. The tap-in-only format and the lack of personalisation of the smart cards may induce using one card for multiple validations in a row (usually associated with trips with relatives). An examination showed that a 60-second lapse was an appropriate cut-off point to detect multi-transactions. Then, cards with multi-transactions were not considered to avoid including this noise that may affect the analysis. On the other hand, the analysis of specific public transport users can help to reveal more meaningful findings (Gutiérrez et al., 2020). In particular, student cards were not included because most classes remained online during November 2020. Therefore, more meaningful conclusions for a post-pandemic era may be obtained by

Table 2

Mobility indicators considered to measure interpersonal variability of public transport (PT) usage change between the pre-pandemic and reopening period.

Indicator	Description	Pre-pandemic		Reopening		Variation		
		Mean	Median	Mean	Median	Mean	Median	SD
PT trip intensity								
PT total weekday trips	Total PT trips in the ten weekday	17.69	18.00	9.60	8.00	−8.09	−8.00	8.55
Trip segments per trip	Average PT trip segments per trip	1.4	1.33	1.33	1.16	−0.07	0.00	0.4
Day SI (DSI)	Day-sequence similarity index	–	–	–	–	0.52	0.50	0.29
PT mode use								
Bus usage	Ratio between bus transactions and total PT transactions (%)	50.77	51.28	48.24	50.00	−2.52	0.00	0.29
PT temporal variability								
Time first transaction	Average hour when the first transaction of the day is made	9.73	9.09	11.11	10.53	1.38	0.75	3.64
Temporal SI (TSI)	Temporal similarity index	–	–	–	–	0.43	0.44	0.27
PT spatial variability								
Location SI (LSI)	Boarding location similarity index	–	–	–	–	0.40	0.40	0.28

observing non-student users.

The last criteria were implemented to identify cardholders' residential locations and, in this way, add additional features to the data. Residential location identification allowed the retrieval of aggregate socioeconomic characteristics from census areas and their association with where cardholders lived. It is important to notice that implementing these criteria may lead to the analysis of more habitual travellers. This limitation was not unique to our work, and previous works where smart card data have been used considered criteria that lead to focus the analysis on regular travellers (Caicedo et al., 2021; Espinoza et al., 2018). Considering all these criteria assures the replicability of the analysis carried out in this work in different contexts, facilitating their comparison.

4.3.2. Intrapersonal variability indicators

The second step in the clustering analysis stage, according to Fig. 4, was the estimation of the indicators that describe the change in individuals' public transport use, considering a multidimensional characterization. Then, the intrapersonal pattern comparison was based on seven mobility indicators that describe the change in the public transport use of frequent passengers of the pre-pandemic period (PP) that continued travelling after the lockdown (O). In particular, three similarity indices were adapted from Egu and Bonnel (2020) to measure these changes.

Firstly, a day-sequence similarity index (DSI) was estimated. For each period p (PP and O) and cardholder m, a boolean vector $D_m^p = \left(d_{m,1}^p, d_{m,2}^p, \cdots, d_{m,N}^p \right)$ of length N equal to 10 was defined (representing the 10 business days of the two consecutive weeks considered per period), where d_n takes value one if there was at least one trip during that day, otherwise, the value is 0. Then the similarity measure between D^{PP} and D^O for each cardholder m was calculated considering the simple matching distance, as follows:

$$DSI_m(D_m^{PP}, D_m^O) = \frac{\sum_{n=1}^N \left(1 - d_{m,n}^{PP}\right)\left(1 - d_{m,n}^O\right) + \sum_{n=1}^N d_{m,n}^{PP} d_{m,n}^O}{N} \quad (1)$$

Where $\sum_{n=1}^N \left(1 - d_{m,n}^{PP}\right)\left(1 - d_{m,n}^O\right)$ represents the number of days where D_m^{PP} and D_m^O are zero and $\sum_{n=1}^N d_{m,n}^{PP} d_{m,n}^O$ is the number of days where D_m^{PP} and D_m^O are one. The two vectors are considered similar when there is a mutual absence or presence of trips on the same days between both periods. The DSI also gives values between 0 (a completely different day sequence pattern) and 1 (the same), facilitating its interpretation.

Additionally, two indices were used to measure the similarity of public transport usage at individual level in terms of the temporal and spatial patterns of active passengers between the pre-pandemic and reopening weeks. In terms of similarity, public transport usage between two periods may be considered similar for a particular cardholder if the same proportion of trips is distributed similarly during the day or if they are distributed similarly in terms of the boarding locations. Thus, for the

temporal and spatial intrapersonal variability, a temporal (TSI) and boarding location (LSI) similarity indices are proposed. Let us define T_m^{PP} and T_m^O as the total number of trip registered in the system for a cardholder m during the ten-weekday period during the pre-pandemic (PP) and the reopening (O). Then, for the TSI, h_r^{PP} and h_r^O indicate the number of transactions h registered during the period of the day r, for the PP and the O. While for the LSI, l_z^{PP} and l_z^O are the number of transactions l registered in the location z, also, for both pandemic periods. Then TSI and LSI were estimated as follows:

$$TSI_m = 1 - \frac{1}{2} \sum_{r=1}^R \left| \frac{h_{r,m}^{PP}}{T_m^{PP}} - \frac{h_{r,m}^O}{T_m^O} \right| \quad (2)$$

$$LSI_m = 1 - \frac{1}{2} \sum_{z=1}^Z \left| \frac{l_{z,m}^{PP}}{T_m^{PP}} - \frac{l_{z,m}^O}{T_m^O} \right| \quad (3)$$

where R refers to the total number of periods within a weekday that make up the temporal grid for public transport demand and Z represents the total number of CDZ from where boarding was carried out. Note that by definition $\sum_r h_r / T$ and $\sum_z l_z / T$ for any cardholder and pandemic period are equal to one. In this way, the TSI and LSI measure how different was the distribution of public transport trips in terms of the temporal and spatial variation between the pre-pandemic and reopening period. We decided to use the TSI above other methods, such as Dynamic Time Warping (DTW) or the distance between two empirical Cumulative Distribution Functions (eCDFs), due to the TSI's interpretability advantage over the distance value calculated using these methods. TSI and LSI do not depend on the variation in the number of trips between both periods. If the relative temporal or spatial distribution of the trips is the same between both pandemic periods, the difference estimated is zero, and TSI/LSI are equal to 1. By way of contrast, if the temporal or spatial travel pattern for a specific cardholder has changed completely and there is no match between the two periods, the second term is 1, and the similarity indices take the value of 0. Therefore, independently of whether a cardholder reduced their trip intensity in terms of the number of trips or the days travelled, TSI and LSI analyse only the differences in terms of how the trip distribution has changed temporally (across the day) and spatially (in terms of the areas where a cardholder boards public transport modes). To identify the proper total number of periods of the day R, the criterion of homogenous periods associated with the overall demand in the system and the fare scheme in Santiago's public transport was applied. Then an R equal to eight was used, considering the next time intervals: before 7:00 am, 7 to 9 am, 9 to 12 pm, 12 to 2 pm, 2 to 4 pm, 4 to 6 pm, 6 to 8 pm and after 8 pm. For the LSI estimation, the spatial grid was defined using the 352 Census zones defined in Section 4.2, and required matching them with the location of the boarding of each trip. Therefore, for the TSI, the comparison between trip distributions was made among the eight-time intervals, each of which represents a particular time period during the day, and for the LSI, the analysis was made on the variation in boarding trips among 352
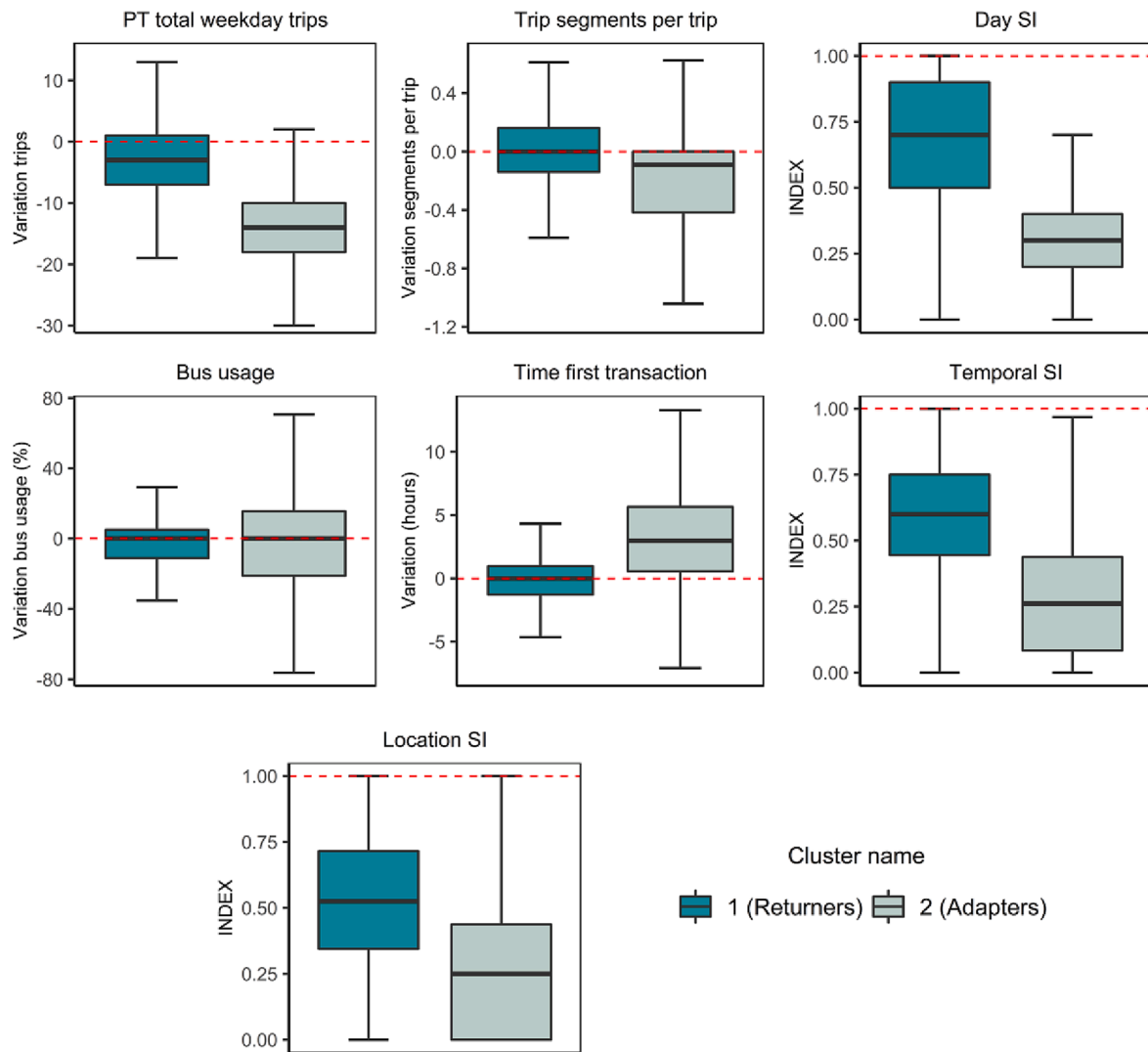
Fig. 5. Cluster profiling by variation in mobility indicators. Red dashed lines indicate the condition of no change between the reopening and the pre-COVID-19 period.

Table 3
Modelling results, GBDT and binomial logistic regression.

Variable	GBDT		Binomial logistic regression			
	Ranking	Relative importance	Odds value	Coefficient	Standard error	Significance
Intercept			2.233	0.803	0.193	<0.001
Travel history lockdown (THL)						
Lockdown trips	1	49.4%	1.097	0.093	0.001	<0.001
Travel history pre-pandemic (THPP)						
PP trips - weekdays	2	28.5%	0.956	−0.045	0.001	<0.001
PP trips - weekend	5	4.0%	1.054	0.053	0.002	<0.001
PP days travelled weekdays	3	6.7%	0.937	−0.065	0.003	<0.001
PP avg. travel time per trip	4	4.8%	1.003	0.003	0.003	<0.001
Demographic characteristics of the traveller (CT)						
Senior card (Adult card ref.)	6	1.4%	1.474	0.388	0.021	<0.001
Demographic characteristics of the travellers' residential location (CRL)						
Share - Women	8	1.1%	0.164	−1.804	0.338	<0.001
Share - Age 00–13	13	0.4%	–	–	–	–
Share - Age + 61	11	0.5%	3.794	0.134	0.266	0.003
Share - Foreign-born	9	0.8%	1.414	0.347	0.079	<0.001
Share - Students	12	0.5%	5.488	1.703	0.297	0.003
Share - University educated	7	1.3%	0.697	−0.361	0.050	<0.001
Share - Workers	10	0.6%	2.497	0.915	0.236	0.002

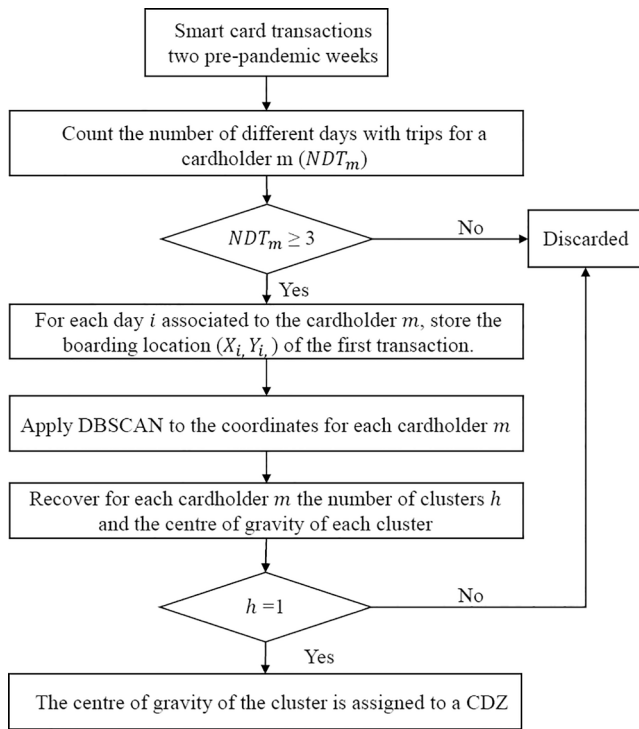The reference class (0) is the adapters' cluster. "-" indicates a non-significant variable.

Figure A1. Framework implemented to identify residential location using smart card data. Adapted from Amaya et al. (2018).

zones distributed across the city.

The remaining five indicators describe the variation between the reopening and pre-pandemic period of variables usually used to characterize public transport usage when smart card data is available. Those variables are the total number of trips, the number of segments per trip, bus usage and the time of the first transaction of the day. All of these were calculated on the ten weekdays of the two periods. Bus usage is estimated as the ratio between the number of validations made on the bus mode and those made in all public transport modes during the ten workdays in each pandemic period. We incorporate this variable to identify whether passengers have systematically reduced or increased the use of the bus mode compared with the metro/rail, as some evidence suggests that metro/rail is more positively rated than buses during the pandemic. The variable time of the first transaction of the day is

estimated by averaging the time of all workday's first transactions, using as a reference midnight (00:00). The definition of trips and trip segments associated with smart card transactions was adopted from Munizaga and Palma (2012). Table 2 presents the seven indicators with their characteristic values per period calculated on the final dataset.

4.3.3. Clustering

Once the indicators that describe intrapersonal public transport usage variability had been estimated, the next step was to use them to identify classes of passengers with similar mobility profiles. K-means, a well-known hard clustering algorithm, was then implemented, aiming to partition the data set into a predefined number of clusters. This technique is considered one of the easiest and fastest clustering algorithms (Viallard et al., 2019) and has demonstrated a high performance due to its capacity to handle big data samples (Ma et al., 2013). As a result of the clustering stage, the optimum number of clusters K was found, and a class membership was assigned to each cardholder depending on the impact of the COVID-19 pandemic on their public transport usage.

4.4. Modelling

Although revealing unseen mobility profiles based on grouping public transport passengers can give a valuable comprehension of the impact of the pandemic on a post-COVID-19 era, understanding the variables that underlie the adoption of a particular profile may drive meaningful insights. Thus, the class membership of each cardholder was studied using the categorical label assigned to each cardholder as a dependent variable and travel history, card type and aggregate demographic characteristics as a set of explanatory variables. Two models were used to complement each other, a Gradient Boosting Decision Tree (GBDT) and a Logistic Regression Model (LRM). GBDT, a supervised machine learning technique, iteratively composes multiple decision trees to find the best results. GBDT provides the relative importance of each explanatory variable used in the classification model, allowing linear and non-linear relationships, with no distributional assumptions, working at high speed with large-size samples. A LRM was also estimated with the aim of complementing the results of GBDT. Following Equation (4), P_k is the probability that a cardholder m belongs to the cluster k, which depends on a linear function V_k (Equation (5)), where α, β, μ and γ are the regression coefficients and x are a set of explanatory variables associated with each observation. If $K + 1$ clusters are considered, only K linear functions are estimated, indexed by k. Therefore, each probability associated to the cluster k will have its own set of
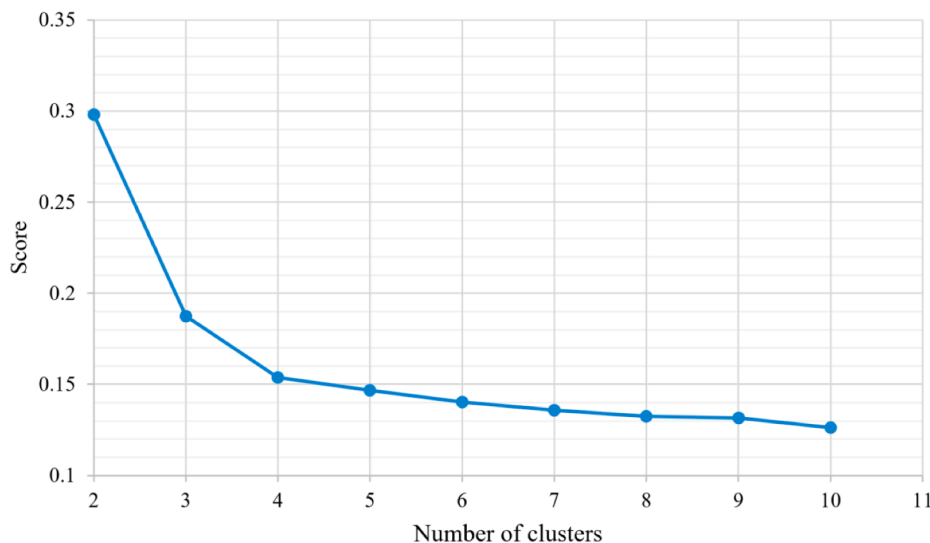

Figure A2. Silhouette scores for the clustering analysis, indicating that the recommended optimal number of cluster is two.

regression coefficients except from the base cluster, which probability is estimated as $1 - \sum_{k=1}^{K} P_k$.

$$P_{k,m} = \frac{e^{V_{k,m}}}{1 + \sum_{l=1}^{K} e^{V_{l,m}}} \quad for\ l = 1, \ldots, K \quad (4)$$

$$V_{k,m} = \sum_r \alpha_{k,r} x_{m,r}^{THPP} + \sum_s \beta_{k,s} x_{m,s}^{THL} + \mu_k x_m^{CT} + \sum_t \gamma_{k,t} x_{m,t}^{CRL} \quad (5)$$

For both models, GBDT and LRM, the set of explanatory variables included travel history during pre-pandemic (THPP), travel history during the lockdown (THL), card type (CT) and aggregate demographic factors associated with the census area where each cardholder resides (CRL). A detailed description of each indicator is given in Table 1.

5. Results

5.1. Public transport user profiles

A summary of the characteristic values of the seven mobility indicators used to capture individuals' public transport usage variability between the pre-pandemic and the reopening period is presented in Table 2. As was expected, an overall comparison between periods indicated a reduction in trip intensity (46% in the number of trips) and a significant adaptation in the temporal and spatial patterns (on average, only 40% of the cardholders' spatial–temporal travel patterns of the pre-pandemic were observed in the reopening).

For clustering cardholders, the K-means algorithm was applied using as observations each of the cardholders of the final dataset and as features the seven indicators that describe the change of cardholders' public transport usage. Therefore, under the K-means approach, cardholders that have similar variations in their public transport usage during PP and reopening were grouped in the same class. The number of clusters was obtained using the silhouette score, which maximized its value when the number of clusters was equal to two (See Fig. A2). This criterion was also confirmed, considering the interpretability of the clustering results and the outcomes obtained in the membership modelling related to other numbers of clusters. Thereby, two well-defined classes of users were detected regarding their public transport usage recovery after the lockdown period. The algorithm classified 47% of the users as members of cluster 1 and 53% as members of cluster 2. The cluster profiling regarding the value distribution of the indicators used for each cardholder class is shown in Fig. 5. Looking at these results, two apparent labels emerge to describe the clusters. Members of cluster 1 were those close to recovering (total or partially) their pre-pandemic mobility patterns in the post-lockdown period; therefore, the name "returner" was given to them. By contrast, cluster 2 was made up of those users whose public transport usage was more highly impacted; thus, they received the label "adapters". To validate the differences in the values of each indicator between the two clusters, the Mann-Whitney U test was conducted, confirming that there was a significant difference in the distributions of the seven indicators between returners and adapters (all had p-values < 0.05).

The differences between the two classes were evident. The returners' cluster showed a median for the variation of total trips of -2.4, which means that 50% of this group almost recovered their trip intensity. By contrast, the same measure was -14.5 trips for the adapters, exhibiting that 75% of their members had a reduction equal to or higher than 10 trips from the pre-pandemic period to the reopening. The distribution of DSI values showed that 75% of the members of returners recovered at least 50% of their trip sequence during the reopening, whilst 75% of the cardholders that belong to the adapters' class showed a much greater change and reproduced less than 40% of their pre-pandemic trip sequence. The average time when the first transaction of a day is made also showed a remarkable difference between the two classes. Returners seem to have maintained the time of their first transaction, showing a

median very close to zero variation. In the adapters' cluster, on the other hand, 75% of their members exhibit a delay in their first trips carried out during the reopening compared with the pre-pandemic of at least 0.5 h, with a median value of around three hours for the class. Regarding TSI and LSI, in the returners' cluster, at least 50% of the cardholders had the temporal and location indices above 0.5, which means that during the reopening period, they carried out a minimum of 50% of their trips in the same time periods and locations that they did during the pre-pandemic. In comparison, 50% of the adapters reached only around 0.25 (25%) similarity with their pre-pandemic behaviour in terms of the period of the day when trips are made and boarding locations in the reopening period.

It is important to note that returners, although belonging to the cluster that recovered most of their pre-pandemic public transport use during the reopening, still exhibited a non-negligible variation in their temporal and spatial trip distributions. We interpret this result as an inherent impact of COVID-19 on people's activities and time use that were still highly present during the post-lockdown period (Molloy et al., 2021). Bus usage did not display an evident dissimilarity between the two user segments if their medians were analysed.

Furthermore, because the literature has exclusively reported the reduction in public transport demand during the pandemic, the expected findings were that all the clusters would show values for trip intensity below the pre-pandemic levels. However, returners illustrated a different situation. The results indicated that around 25% of their members carried out more trips during the reopening than the pre-pandemic. Also, around 50% of their members had an increase in the average number of trips per day and the number of trip segments per trip. Finally, although the clustering analysis indicated that the optimal number of clusters was two, it was evident that the actual number of different strategies that describe all passengers may be as many as the sample size. Therefore, the clusters found were the best aggregation of those adaptations, which inherently limit the visualization of all the strategies related to the changes in public transport usage but help with the interpretation of the main ones.

5.2. Modelling user profiles

This research adopted GBDT and LRM, intending to explore the link between explanatory variables such as travel history, card type and aggregate demographic characteristics with each cluster profile found in the previous section (returners and adapters). GBDT was implemented to provide information about the most important explanatory features associated with class membership. Each indicator mentioned in Table 1 was ranked depending on its relative importance. The relative importance is associated with the number of times a variable is chosen for splitting the sample over all trees. The GBDT model was fitted using a set of parameters, including the number of trees, the learning rate and the maximum tree depth. As the literature suggests, a five-fold cross-validation method was implemented to find the final setting and to control overfitting. The final set of parameters included a shrinkage value of 0.1, 100 trees and a depth equal to 5. On the other hand, the LRM was estimated using Equations (5) and (6), considering the binary nature of the labels found. A Nagelkerke R-value of 0.11 and an acceptable accuracy of 62.1% and 62.5% were obtained for the LRM and GBDT respectively, results in line with those achieved in previous works where comparable data and methods have been implemented (Almlöf et al., 2021). Specific outcomes of GBDT and LRM are presented in Table 3.

First, GBDT identified three types of variables depending on their relative importance (RI). The first category, comprising around 80% of the total RI, includes the variables number of trips during lockdown (49.4%) and the number of trips on weekdays during the pre-pandemic (28.5%). A second group are those characteristics that describe travel history during the pre-pandemic period, such as average travel time per trip (4.8%), weekdays travelled (6.7%), weekend trips (4.0%) and if the card was a senior one (1.4%). Finally, aggregate demographic factors

were the attributes with the lowest RI scores. It is important to note that this does not imply that residential attributes are not relevant, but rather, as is usual in supervised machine learning, variables that result in the most significant set of partitions during the learning process end up showing more relevance (Victoriano et al., 2020). Thus, variables with low importance should be tested using complementary methods such as logistic regression for an appropriate interpretation. Moreover, we hypothesise that the low importance that GBDT assigned to the residential location characteristics is due to the aggregate nature of those variables, gaining more explanation from the variables with a cardholder-level variability. The findings regarding the focus of relative importance in two variables is in line with other studies where GBDT have been implemented with smart card data. In those studies, travel history variables frequently rank first, presenting one or two variables with the greatest relative importance (Tang et al., 2020).

In the LRM, the odds values are estimated as the exponential of the coefficients (see Table 3). A value of 1 indicates that a variable has no influence on the class membership, a value greater than one indicates an increase in the likelihood that an individual is in the returners' cluster, and if the value is smaller than one means a negative effect. Thus, variables of travel history that increase the probability of being part of the returners' class are the number of trips during the lockdown, weekend trips in the pre-pandemic and the average travel time per trip in the pre-pandemic. In particular, the odds value of trips during lockdown was 1.097, which indicates that as trips carried out in this period increase by one, the likelihood of a cardholder being in the returners' cluster will increase by 9.7%. Namely, those who travelled in the most challenging period of the pandemic showed a higher probability of recovering their pre-COVID travel patterns during the post-lockdown stage. The odds that a cardholder belongs to the returners' cluster increases by 5.4% with each trip made on weekends during the pre-COVID period. It may imply that those who carried out more weekend trips (usually associated with non-mandatory activities) probably were more engaged with public transport or had fewer options to choose alternative modes. Moreover, the higher the number of observed pre-pandemic weekday trips, the easier it was for the users to reduce their public transport demand and, consequently, to belong to the adapter cluster. We believe that having a higher pre-pandemic trip intensity can be associated with more flexibility in terms of trip purpose, period of time and mode available, allowing passengers to develop a higher adaptation during the reopening. Additionally, this result suggests that those with more "compact" mobility during weekdays in the pre-pandemic could recover more of their previous mobility patterns than those with higher trip intensity on weekdays. The last travel history indicator, average travel time per trip during the pre-pandemic, showed that the greater its magnitude, the more likely it is that a cardholder belongs to the returners' cluster (3% increase for every 10 min of travel time). The result helps to understand the characteristics of each user segment: higher travel times in Santiago are associated with the municipalities with the lowest income (Gschwender et al., 2016).

Equity aspects were also present in our results. Related to the card type involved, the results indicated that if a senior cardholder was active during the reopening, there was a 47% more chance that the person belonged to the returners' cluster than users with adult cards. This result may initially seem counterintuitive when compared with the existing literature that has found that seniors avoided public transport during the pandemic (Schaefer et al., 2021; Zhao and Gao, 2022). However, given that this study only considered cards that were active in both periods, we hypothesize that most of the senior cardholders that could have had the chance to stop using public transport made that decision at the early start of the pandemic and were already out of the PT system during the reopening. Therefore, we are observing the behaviour of those seniors who likely had no choice rather than to continue using the system during the reopening period, and in that context, the result reveals that if a senior cardholder was active during the reopening, they had more chance to have recovered their pre-pandemic public transport use. This

finding is significant because it provides evidence of heterogeneous responses among the members of the same vulnerable group.

Finally, the effect of the residential area characteristics assigned to cardholders was consistent with the presence of inequality in Santiago's metropolitan area and similar to the one reported in other contexts of the Global South (Caicedo et al., 2021; Vallejo-Borda et al., 2022). In terms of the effect of the home-area demographic factors, results indicated that the higher the share of worker and immigrant population in the areas where cardholders were assigned, the higher the probability they had returned to their pre-COVID public transport use patterns. In fact, as is mentioned by Abduljabbar et al. (2022), public transport is a key mode, especially for specific groups of the population, such as workers and non-nationals, who could face more constraints in deciding freely whether to travel or not. In contrast, cardholders whose residences were located in areas with a higher share of women and university-educated individuals were less likely to be in the returners' cluster. Indeed, gender (female) and higher educational level/income have been associated widely with a higher reduction in public transport use (Abdullah et al., 2020).

6. Conclusion

To our knowledge, the study reported here is the first study where a large passive data source collected during the pandemic of COVID-19 is used to analyse the recovery in public transport demand at a disaggregate level based on a multidimensional approach. This work complements existing literature by analysing the changes in the public transport usage of pre-pandemic users that continued travelling after a long-term lockdown, using smart card data records from the public transport system of Santiago de Chile. The observed results are in reasonable agreement with previous work carried out in the Global South, where sociodemographic disparities have been linked with the change in public transport usage caused by the COVID-19 disease (Caicedo et al., 2021; Vallejo-Borda et al., 2022). However, this study extends existing empirical evidence, demonstrating that the public transport usage recovery among passengers that continued travelling after the lockdown was dissimilar.

Two clusters of public transport users were identified using seven indicators that described the changes in passengers' public transport usage between the pre-pandemic and the reopening. One class of cardholders was named as returners as they showed a pronounced return to their pre-pandemic public transport use during the reopening, whilst the second class was labelled as adapters as they exhibited the greatest changes. Although the class labelled as returners showed a slight change in travel intensity and bus usage between the pre-pandemic and reopening periods, temporal and spatial public transport use patterns showed more strongly evident adaptations, which is in line with previous findings based on ridership analysis during the pandemic (Mützel and Scheiner, 2021). Finally, using disaggregate smart card data it was possible to detect that not all passengers reduced their public transport trip intensity during the reopening. In fact, as many as 25% of the members of the returners' cluster showed an increase in the number of trips during weekdays. This finding is unexpected, and challenges existing literature as, to the best of our knowledge, no evidence of trip intensity increase during the reopening stage that followed the first lockdowns in 2020 has been reported. We speculate that those cardholders could be users that shifted to a type of employment demanding higher mobility due to the pandemic restrictions, likely related to providing services at customers' locations.

The influence of both pre-pandemic and lockdown travel history, demographic characteristics at the residential level and card type were considered as potential variables to explain the membership of each cardholder to each mobility profile using GBDT and logistic regression. The pre-pandemic trip intensity showed a heterogeneous impact on the change in public transport usage between the reopening and pre-pandemic periods. Cardholders that carried out more trips on

weekdays during the pre-pandemic showed a greater likelihood of belonging to the adapters' class. In contrast, those who made more pre-pandemic weekend trips were most likely to have a returner profile. Public transport usage during lockdown was also considered, showing that those who continued travelling during the lockdown exhibited a higher probability of belonging to the returners' cluster.

Equity aspects were also present in results. Our findings confirmed the relationship between the spatial distribution of sociodemographic characteristics across the city and the changes in PT use during the first stage of the pandemic. As Fig. 1 depicted, the highly-educated population, the presence of immigrants and the population's age were characteristics greatly concentrated in specific areas of the city. Inequality issues was also observed on the PT level of service. Indeed, longer PT travel times were related to a lower PT use adaption. Longer PT travel times in Santiago have been historically associated with commuting trips from the city periphery to its northeast area, where the higher income and number of services can be found. In this regard, the lowest capacity of these users to adapt their PT use could be related to the mandatory need for in-person work as soon as the lockdown finished and, secondly, by their strong dependency on specific PT services. This last element would have made them extremely vulnerable to service changes during the opening, which certainly was mitigated by PT authorities' decision to keep PT services and frequencies as close as possible to the pre-pandemic. Therefore, to reduce urban inequalities when future disruptions such as a new pandemic happens, particular emphasis in policy development should be placed on the specific needs of vulnerable and PT-dependent sectors of the population.

Although smart card data is a rich source to explore travel behaviour, individual demographic information of each passenger is typically missing. In fact, individual demographic characteristics, the possibility of teleworking, and an assessment of travellers' risk perception toward public transport could have helped to give a deeper understanding of the profiles found. In our results, the hidden effect of those variables may be represented indirectly by the travel history variables. Therefore, although this work demonstrates the advantages of exploring individual travel behaviour of public transport users, it does not replace the richness provided by traditional surveys in terms of individual explanatory variables. If suitable data is available in future, combining such a survey with passive data will be an interesting direction for future research.

Our finding allows us to conceive three main implications, which expand the current understanding of the changes of COVID-19 on public transport demand and give insights into the post-pandemic scenario but also to eventual new pandemics. Firstly, given that temporal and spatial patterns of public transport passengers have changed considerably, efforts to characterise these adaptations should be made continuously during the pandemic and even in the post-pandemic to propose and adjust services where required. Secondly, as equity disparities are related to a higher recovery of the pre-pandemic public transport use during the reopening, measures that provide benefits to captive cardholders should be considered to support that recovery but also, to mitigate the greater post-lockdown need for mobility found for a considerable proportion of cardholders. For example, as a complement to the pay-as-you-go scheme in Santiago's public transport, travel passes could be a policy in that direction. Finally, our results imply that as an aftermath of the pandemic, public transport systems may experience severe difficulties in recovering their pre-pandemic ridership during the post-COVID-19 period. In fact, even though the return of pre-pandemic users to public transport modes in the reopening, a substantial proportion of them carried out fewer trips than the pre-COVID-19. This suggests that government policies to ensure the sustainability of public transport will be needed for a long-term period. This support will ease the pressure on PT operators to reduce PT supply or increase fares, which may only worsen given the public transport situation. Although this recommendation is theoretically possible in many government-supported public transport systems worldwide, it is a huge challenge for the Global South, where public transport is less regulated, and often

there is no direct subsidy.

CRediT authorship contribution statement

Maximiliano Lizana: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Charisma Choudhury:** Supervision, Writing – review & editing. **David Watling:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix

Fig. A1 and Fig. A2

References

Abduljabbar, R.L., Liyanage, S., Dia, H., 2022. A systematic review of the impacts of the coronavirus crisis on urban transport: Key lessons learned and prospects for future cities. Cities 127, 103770.

Abdullah, M., Dias, C., Muley, D., Shahin, M.d., 2020. Exploring the impacts of COVID-19 on travel behavior and mode preferences. Transp. Res. Interdiscip. Perspect. 8, 100255.

Almlöf, E., Rubensson, I., Cebecauer, M., Jenelius, E., 2021. Who continued travelling by public transport during COVID-19? Socioeconomic factors explaining travel behaviour in Stockholm 2020 based on smart card data. Eur. Transp. Res. Rev. 13 (1) https://doi.org/10.1186/s12544-021-00488-0.

Amaya, M., Cruzat, R., Munizaga, M.A., 2018. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. J. Transp. Geogr. 66, 330–339. https://doi.org/10.1016/j.jtrangeo.2017.10.017.

Beck, M.J., Hensher, D.A., 2021. Australia 6 months after COVID-19 restrictions- part 1: Changes to travel activity and attitude to measures. Transp. Policy 128, 286–298. https://doi.org/10.1016/j.tranpol.2021.06.006.

Bennett, M., 2021. All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in chile. World Dev. 137, 105208.

Bin, E., Andruetto, C., Susilo, Y., Pernestål, A., 2021. The trade-off behaviours between virtual and physical activities during the first wave of the COVID-19 pandemic period. Eur. Transp. Res. Rev. 13 (1) https://doi.org/10.1186/s12544-021-00473-7.

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. https://doi.org/10.1016/j.trc.2017.03.021.

Bucsky, P., 2020. Modal share changes due to COVID-19: The case of Budapest. Transp. Res. Interdiscip. Perspect. 8, 100141.

Caicedo, J.D., Walker, J.L., González, M.C., 2021. Influence of Socioeconomic Factors on Transit Demand During the COVID-19 Pandemic: A Case Study of Bogotá's BRT System. Front. Built Environ. 7, 642344 https://doi.org/10.3389/fbuil.2021.642344.

Carney, F., Long, A., Kandt, J., 2022. Accessibility and Essential Travel: Public Transport Reliance Among Senior Citizens During the COVID-19 Pandemic. Front. Big Data 5, 867085. https://doi.org/10.3389/fdata.2022.867085.

Egu, O., Bonnel, P., 2020. Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. Travel Behav. Soc. 19, 112–123. https://doi.org/10.1016/j.tbs.2019.12.003.

Eisenmann, C., Nobis, C., Kolarova, V., Lenz, B., Winkler, C., 2021. Transport mode use during the COVID-19 lockdown period in Germany: The car became more important, public transport lost ground. Transp. Policy 103, 60–67. https://doi.org/10.1016/j.tranpol.2021.01.012.

El Mahrsi, M.K., Come, E., Oukhellou, L., Verleysen, M., 2017. Clustering Smart Card Data for Urban Mobility Analysis. IEEE Trans. Intell. Transp. Syst. 18 (3), 712–728. https://doi.org/10.1109/tits.2016.2600515.

Espinoza, C., Munizaga, M., Bustos, B., Trépanier, M., 2018. Assessing the public transport travel behavior consistency from smart card data. Transp. Res. Procedia 32, 44–53. https://doi.org/10.1016/j.trpro.2018.10.008.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise* Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR.

Fernández Pozo, R., Wilby, M.R., Vinagre Díaz, J.J., Rodríguez González, A.B., 2022. Data-driven analysis of the impact of COVID-19 on Madrid's public transport during each phase of the pandemic. Cities 127, 103723. https://doi.org/10.1016/j.cities.2022.103723.

Gainza, X., Livert, F., 2013. Urban Form and the Environmental Impact of Commuting in a Segregated City, Santiago de Chile. Environ. Plann. B. Plann. Des. 40 (3), 507–522. https://doi.org/10.1068/b38045.

Gong, Y., Lin, Y., Duan, Z., 2017. Exploring the spatiotemporal structure of dynamic urban space using metro smart card records. Comput. Environ. Urban Syst. 64, 169–183. https://doi.org/10.1016/j.compenvurbsys.2017.02.003.

Google. *COVID-19 Community Mobility Reports*. Retrieved 10 January 2023 from https://www.google.com/covid19/mobility/.

Gramsch, B., Guevara, C.A., Munizaga, M., Schwartz, D., Tirachini, A., 2022. The effect of dynamic lockdowns on public transport demand in times of COVID-19: Evidence from smartcard data. Transp. Policy 126, 136–150. https://doi.org/10.1016/j.tranpol.2022.06.012.

Gschwender, A., Munizaga, M., Simonetti, C., 2016. Using smart card and GPS data for policy and planning: The case of Transantiago. Res. Transp. Econ. 59, 242–249. https://doi.org/10.1016/j.retrec.2016.05.004.

Gutiérrez, A., Domènech, A., Zaragozí, B., Miravet, D., 2020. Profiling tourists' use of public transport through smart travel card data. J. Transp. Geogr. 88, 102820.

He, L., Agard, B., Trépanier, M., 2018. A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. Transport. A: Transp. Sci. 16 (1), 56–75. https://doi.org/10.1080/23249935.2018.1479722.

Jenelius, E., Cebecauer, M., 2020. Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts. Transp. Res. Interdiscip. Perspect. 8, 100242 https://doi.org/10.1016/j.trip.2020.100242.

Kolarova, V., Eisenmann, C., Nobis, C., Winkler, C., Lenz, B., 2021. Analysing the impact of the COVID-19 outbreak on everyday travel behaviour in Germany and potential implications for future travel patterns. Eur. Transp. Res. Rev. 13 (1) https://doi.org/10.1186/s12544-021-00486-2.

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. https://doi.org/10.1016/j.trc.2014.05.012.

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. https://doi.org/10.1016/j.trc.2013.07.010.

Molloy, J., Schatzmann, T., Schoeman, B., Tchervenkov, C., Hintermann, B., Axhausen, K.W., 2021. Observed impacts of the Covid-19 first wave on travel behaviour in Switzerland based on a large GPS panel. Transp. Policy 104, 43–51. https://doi.org/10.1016/j.tranpol.2021.01.009.

Munizaga, M.A., 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. https://doi.org/10.1016/j.trc.2012.01.007.

Mützel, C.M., Scheiner, J., 2021. Investigating spatio-temporal mobility patterns and changes in metro usage under the impact of COVID-19 using Taipei Metro smart card data. Public Transport 14 (2), 343–366.

Ordóñez Medina, S.A., 2018. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. Travel Behav. Soc. 12, 93–101. https://doi.org/10.1016/j.tbs.2016.11.005.

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. https://doi.org/10.1016/j.trc.2010.12.003.

Politis, I., Georgiadis, G., Nikolaidou, A., Kopsacheilis, A., Fyrogenis, I., Sdoukopoulos, A., Verani, E., Papadopoulos, E., 2021. Mapping travel behavior changes during the COVID-19 lock-down: a socioeconomic analysis in Greece. Eur. Transp. Res. Rev. 13 (1) https://doi.org/10.1186/s12544-021-00481-7.

Przybylowski, A., Stelmak, S., Suchanek, M., 2021. Mobility Behaviour in View of the Impact of the COVID-19 Pandemic—Public Transport Users in Gdansk Case Study. Sustainability 13 (1). https://doi.org/10.3390/su13010364.

Rodriguez Gonzalez, A.B., Wilby, M.R., Vinagre Diaz, J.J., Fernandez Pozo, R., 2021. Characterization of COVID-19's Impact on Mobility and Short-Term Prediction of Public Transport Demand in a Mid-Size City in Spain. Sensors (Basel) 21 (19). https://doi.org/10.3390/s21196574.

Schaefer, K.J., Tuitjer, L., Levin-Keitel, M., 2021. Transport disrupted – Substituting public transport by bike or car under Covid 19. Transp. Res. A Policy Pract. 153, 202–217. https://doi.org/10.1016/j.tra.2021.09.002.

Tang, T., Liu, R., Choudhury, C., 2020. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. Sustain. Cities Soc. 53, 101927.

Tori, S., Sejournet, A., Macharis, A., 2023. Reactions of the public transport sector to the COVID-19 pandemic. Insights from Belgium. Travel Behav. Soc. 31, 244–253. https://doi.org/10.1016/j.tbs.2023.01.001.

Vallejo-Borda, J.A., Giesen, R., Basnak, P., Reyes, J.P., Mella Lira, B., Beck, M.J., Hensher, D.A., Ortúzar, J.D.D., 2022. Characterising public transport shifting to active and private modes in South American capitals during the COVID-19 pandemic [Article]. Transp. Res. A Policy Pract. 164, 186–205. https://doi.org/10.1016/j.tra.2022.08.010.

Viallard, A., Trépanier, M., Morency, C., 2019. Assessing the Evolution of Transit User Behavior from Smart Card Data. Transp. Res. Record J. Transp. Res. Board 2673 (4), 184–194. https://doi.org/10.1177/0361198119834561.

Vickerman, R., 2021. Will Covid-19 put the public back in public transport? A UK perspective. Transp. Policy 103, 95–102. https://doi.org/10.1016/j.tranpol.2021.01.005.

Victoriano, R., Paez, A., Carrasco, J.-A., 2020. Time, space, money, and social interaction: Using machine learning to classify people's mobility strategies through four key dimensions. Travel Behav. Soc. 20, 1–11. https://doi.org/10.1016/j.tbs.2020.02.004.

Villalobos Dintrans, P., Castillo, C., de la Fuente, F., Maddaleno, M., Morrissey, K., 2021. COVID-19 incidence and mortality in the Metropolitan Region, Chile: Time, space, and structural factors. PLoS One 16 (5), e0250707.

Wang, H., Noland, R.B., 2021. Bikeshare and subway ridership changes during the COVID-19 pandemic in New York City. Transp. Policy 106, 262–270. https://doi.org/10.1016/j.tranpol.2021.04.004.

Wielechowski, M., Czech, K., Grzęda, Ł., 2020. Decline in Mobility: Public Transport in Poland in the time of the COVID-19 Pandemic. Economies 8 (4). https://doi.org/10.3390/economies8040078.

Zannat, K.E., Bhaduri, E., Goswami, A.K., Choudhury, C.F., 2021. The tale of two countries: modeling the effects of COVID-19 on shopping behavior in Bangladesh and India. Transp. Lett. 13 (5–6), 421–433. https://doi.org/10.1080/19427867.2021.1892939.

Zannat, K.E., Choudhury, C.F., 2019. Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions. J. Indian Inst. Sci. 99 (4), 601–619. https://doi.org/10.1007/s41745-019-00125-9.

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, P., Gao, Y., 2022. Public transit travel choice in the post COVID-19 pandemic era: An application of the extended Theory of Planned behavior. Travel Behav. Soc. 28, 181–195. https://doi.org/10.1016/j.tbs.2022.04.002.