Exploring Travel Pattern Variability of Public Transport Users through Smart Card Data: Role of Gender and Age

Shasha Liu, Toshiyuki Yamamoto, Enjian Yao, and Toshiyuki Nakamura

***Abstract*—To improve passenger experience and service provision, public transport (PT) authorities should gain a better understanding of travel pattern variability. Although studies have extensively examined the travel pattern variability of PT users, these studies are often limited to a short analysis period or to only one dimension of travel behavior. In addition, limited information is available on how the demographic details of PT users are associated with their travel pattern variability. To address these limitations, we develop a novel measure that simultaneously considers multiple dimensions of travel behavior to quantify the intrapersonal variability in weekly PT usage. Moreover, we examine interpersonal variability by identifying clusters of users who share similar weekly profiles. Based on smart card transaction data for 52 weeks and an anonymous cardholder database (including age and gender) from Shizuoka, Japan, we analyze the intrapersonal and interpersonal variability in weekly PT usage as well as the role of gender and age in travel pattern variability. The results indicate that gender and age play an important role in the travel pattern variability of PT users. Female users exhibit higher intrapersonal variability than their male counterparts. Weekly patterns are the most diverse for users aged 70 or over, followed by the users aged 65–69. Regarding interpersonal variability, we identify five clusters of users, each characterized by a distinct weekly profile. Significant associations also exist between the gender and age of users and the identified clusters of weekly patterns.**

***Index Terms*—Age, Gender, Interpersonal variability, Intrapersonal variability, Public transport, Smart card data**

I. Introduction

Travel behavior is dynamic, varying across individuals and varying for the same person over time [1]. Interpersonal variability is reflected in heterogeneous travel patterns across travelers. Intrapersonal variability refers to the longitudinal variability in the travel behavior of the same person over time [1], [2]. Owing to travel pattern variability, a variety of travel needs and habits have arisen in public transport (PT) systems. A better understanding of travel pattern variability will help PT authorities to improve passenger experience and service provision. More specifically, information on the interpersonal variability in PT usage is useful for identifying similarities among users and revealing how different types of users can be grouped according to the way they use PT, thereby providing great potential to realize multiple applications (e.g., fare policy design, customized service provision, targeted travel demand management) [3], [4], [5]. Intrapersonal variability is often measured based on the repetition of the same patterns in a person's behavior [2]. An in-depth understanding of the intrapersonal variability in PT usage can be used in customized journey planning and personalized information provision about disruptions and crowding levels. For example, if we know that a user usually travels between stations A and B on weekday mornings and evenings, an alert can be sent to the user if a service disruption occurs and the user's commute trip is expected to be disrupted [6].

The investigation of intrapersonal and interpersonal variability over multiple days requires the longitudinal observation of PT users. Although conventional multiday travel surveys provide detailed information about travelers and trips [7], [8], [9], these surveys are costly and thus typically limited to short observation periods and small sample sizes [1], [10]. In contrast, smart card data provide long-term and continuous information about PT trips of a large number of users, making it feasible to explore travel pattern variability through longitudinal analyses [2], [3], [11]. Smart card data–based examination of the interpersonal variability in PT usage is typically achieved by grouping users with similar travel patterns using clustering techniques [3], [12], [13], [14]. The scalar or vector aggregation of individual trips is often used as the feature for clustering analysis. However, these studies are often limited to a short analysis period (e.g., 1 month) or to only one dimension of travel behavior (e.g., temporal or spatial pattern).

To measure intrapersonal variability, some methods rely on the relative frequency of trips (e.g., percentage of trips made within certain time periods) [12], [13] and some focus on the variance of trips (e.g., variance in the number of trips per day) [8]. However, these methods usually do not consider how

Manuscript received on March 30, 2020. We would like to thank the Shizutetsu Group for cooperating with us and allowing us to use the smart card data. This work was partly supported by the National Key R&D Program of China (No. 2018YFB1601303).

S. Liu and T. Yamamoto are with the Institute of Materials and Systems for Sustainability, Nagoya University, Nagoya 4648603, Japan (e-mail: ssliubj@bjtu.edu.cn; yamamoto@civil.nagoya-u.ac.jp).

E. Yao is with the Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Beijing Jiaotong University, Beijing 100044, China (e-mail: enjyao@bjtu.edu.cn).

T. Nakamura is with the Institute of Innovation for Future Society, Nagoya University, Nagoya 4648603, Japan (e-mail: tnakamura@mirai.nagoya-u.ac.jp).

multiple trips are combined. Moreover, to develop a female-friendly and age-friendly PT system, it is necessary to examine the association of gender and age with the travel patterns of PT users [15], [16]. However, because demographic characteristics are seldom included in smart card data, information on the association between the demographic characteristics and travel pattern variability of PT users remains limited. Based on smart card transaction data for 1 year and an anonymous cardholder database (including age and gender) from Shizuoka, Japan, we attempt to address these limitations and explore the interpersonal and intrapersonal variability in PT usage as well as the role of gender and age in travel pattern variability. Because the flattening of the profiles of all weekdays into a single daily profile may cause the variations among weekdays to be neglected [14], we examine the travel patterns on a weekly basis rather than on a daily basis.

The contribution of this study is twofold. From a methodological perspective, we propose a novel representation of week sequences that simultaneously considers multiple dimensions of PT usage and develop a measure to quantify the intrapersonal variability in weekly PT usage. In addition, we develop aggregate weekly profiles for users and identify clusters of users who share similar weekly patterns to explore interpersonal variability. From an empirical perspective, we analyze the intrapersonal and interpersonal variability in PT usage in an extensive network through the large-scale application of the proposed methodology to Shizuoka's PT system. We reveal the impact of age and gender on intrapersonal variability and show the demographic association of the identified clusters of weekly patterns.

The remainder of this paper is organized as follows. The next section reviews previous work on travel pattern analysis using smart card data and the sociodemographic association of travel patterns. The third section introduces the smart card transaction data and anonymous cardholder database from Shizuoka, Japan. The fourth section presents the methods used to examine the intrapersonal and interpersonal variability in PT usage. The application of the proposed methodology to Shizuoka's PT users is presented in the subsequent section, revealing the intrapersonal and interpersonal variability as well as the role of age and gender in travel pattern variability. The discussion and conclusion are given in the final section.

II. Literature Review

To identify relevant methods, findings, and gaps in the existing research, we first review the methods that have been employed to explore interpersonal and intrapersonal variability using smart card data. Then, we review the association between sociodemographic characteristics and travel pattern variability. Finally, we present a summary of the research gaps and the objectives of this study.

A. Travel Pattern Analysis using Smart Card Data

Smart card data are widely used to explore the travel patterns of PT users in spatial and temporal dimensions at the aggregate [17], [18], [19] or disaggregate [3], [12], [14], [20] levels. To examine the interpersonal variability in travel patterns, many studies used clustering techniques to group PT users with similar travel patterns. Ma et al. [3] introduced four scalar features—number of travel days, number of similar first boarding times, number of similar route sequences, and number of similar stop ID sequences—to classify PT users into five clusters of varying regularity levels using a k-means++ clustering algorithm. Kieu et al. [13] applied the density-based spatial clustering of applications with noise (DBSCAN) algorithm to recognize regular trips in space and time. Then, based on a subjective segmentation rule, four segments of passengers were identified: irregular passengers, regular origin–destination (OD) passengers, habitual time passengers, and transit commuters. EI Mahrsi et al. [14] built an aggregate weekly temporal profile for each user, describing the trip frequency over each hour on each day of the week. They identified 13 clusters of temporal patterns using a generative model-based clustering approach. In these studies, the features used for clustering or segmentation are derived from a scalar or vector aggregation of individual trips, wherein the organization of multiple trips over time is not considered. In contrast, some studies constructed sequences of activity–travel events to explore travel pattern variability. Goulet-Langlois et al. [4] proposed a longitudinal representation of an individual 4-week activity sequence based on smart card data. They employed principal component (PC) analysis to reduce the dimension of sequences and used the projections of the user sequence onto the first eight PCs as the input to k-means clustering. They identified 11 clusters of users with distinct sequence structures.

Apart from interpersonal variability, the same individual's travel pattern shows variations in terms of the trip, day, and week [1], [2], [10], [21]. Intrapersonal variability is often measured based on the repetition of the same patterns in a person's travel behavior. Deschaintres et al. [2] investigated the cluster membership of weekly profiles over 51 weeks. Each user was then represented by a sequence of week clusters. Two indicators—Shannon entropy and the average Euclidean distance between two successive weeks—were used to measure the intrapersonal variability in weekly PT usage. Egu and Bonnel [10] identified the classes of trips based on the boarding time and stops and then applied a trip-based similarity measure to quantify the day-to-day intrapersonal variability in PT usage. Additionally, intrapersonal variability can be measured by variance indicators (e.g., variance in the number of trips per day) [8]. However, these methods usually do not consider how multiple trips are organized.

B. Sociodemographic Association of Travel Patterns

Both the interpersonal and intrapersonal variability in travel patterns is influenced by sociodemographic characteristics [9]. Because sociodemographic information (e.g., age and gender) is usually not included in smart card data, limited studies [4], [22] have examined the association between the sociodemographic characteristics and travel pattern variability of PT users. However, some studies have investigated how travel patterns vary among passengers with different smart card types [2], [10], [13], [23]. Goulet-Langlois [4] combined household travel survey data with smart card transactions and

obtained a small sample of PT users with detailed sociodemographic information. They then explored the demographic association of the interpersonal variability in multiweek activity–travel sequences using odds ratio analysis and a multinomial logit model. Kandt and Leak [22] connected an anonymous cardholder database (including age, gender, and residential address) to smart card transactions and examined how age, gender, and the residential built environment relate to the interpersonal variability in ordinal boarding sequences over nearly 6 years for the elderly aged 65 or over. However, the sociodemographic impacts on the intrapersonal variability in PT usage have rarely been analyzed.

In addition to the sociodemographic association of the variability in PT usage, many studies have used active travel survey data to explore the impacts of sociodemographic characteristics on activity–travel pattern variability [21], [24], [25], [26]. Raux et al. [8] analyzed the variability in daily activity–travel patterns based on 7-day travel diary data. The results revealed that men have lower intrapersonal variability than women in terms of daily trips, time use, and activity sequence and that older adults show lower variability. Moiseeva et al. [9] used 8 weeks of diary data to examine the variability in activity–travel patterns. They stated that the interpersonal variability between genders is considerably greater than intrapersonal variability. Women tend to have more diverse activity and location sequences than men. However, the travel survey data used in these studies are typically constrained by small sample sizes and short observation periods. Schlich and Axhausen [7] argued that the observation period for measuring intrapersonal variability should not be less than 2 weeks.

In summary, existing studies have highlighted the potential of smart card data to examine the travel pattern variability of PT users. However, these studies are often limited to a short analysis period (e.g., 1 month) or to only one dimension of travel behavior (e.g., temporal or spatial pattern). In addition, the methods used to quantify intrapersonal variability usually do not consider how multiple trips are combined. Moreover, the sociodemographic association of the longitudinal variability in travel patterns is rarely examined using a large sample, and information on the sociodemographic impacts on the variability in PT usage remains limited. To address these limitations, we develop a novel measure that quantifies the intrapersonal variability in PT usage; the measure simultaneously considers multiple dimensions of travel behavior, including the boarding time, OD pairs visited, and organization of multiple trips over time. In addition, we reveal the association of gender and age with the intrapersonal and interpersonal variability in PT usage by observing a large number of PT users over a year.

III. Data

The smart card data were collected from the Automated Fare Collection system of Shizutetsu Bus and Railway from April 2, 2018 to March 31, 2019; this period consists of 52 complete weeks (from Monday to Sunday). Shizutetsu Bus and Railway constitutes the urban PT system of Shizuoka city and its neighboring cities in Shizuoka Prefecture, Japan, and includes one railway line with 15 stations and a bus network with approximately 1600 bus stops.

The full dataset contains approximately 23.51 million transactions, of which 72% are bus transactions and the rest are railway transactions. The data was collected from 194,346 smart cards. Each transaction record contains the card ID, tap-in date and time, tap-out date and time, origin and destination station names, travel cost, bus line, and bus route. Apart from the transaction records, we also collected smart card registration data, including the card ID, home address, age, and gender.

Because the smart card transaction records of PT users who take PT occasionally are not sufficient to reveal travel pattern variability [12], [14], frequent users were selected by grouping users based on their PT usage characteristics. Each user was characterized by the number of active days over the 1-year analysis period and the spread of active days between the first and last days [4]. These two features were then used for k-means++ clustering [27]. Three clusters of users were identified: low-frequency users who traveled for a few days spread over a short period (average of 13 days traveled, spread over 69 days), intermittent users who made trips on a few days spanning most of the analysis period (average of 33 days traveled, spread over 311 days), and frequent users who traveled on many days spread over the analysis period (average of 212 days traveled, spread over 352 days). There were 37,827 frequent users, among which 37,649 users had included personal information (e.g., age and gender) in their registered smart cards. These 37,649 users, accounting for 19% of all users and making over 69% of trips, were the focus of this study.

IV. Methodology

To identify the origins and destinations of trips, we construct trips based on smart card transaction data. If two successive boarding transactions occur within 30 min and are not made on the same route, the two are considered parts of the same trip. The transfer time threshold of 30 min is defined according to the 4th Shizuoka Metropolitan Area Person Trip Survey conducted in 2012, which reveals that more than 94% of transfer activities take less than 30 min in the PT network. Successive trips on the same route in opposite directions usually indicate a return trip, whereas travel in the same direction indicates a nontransfer activity [28].

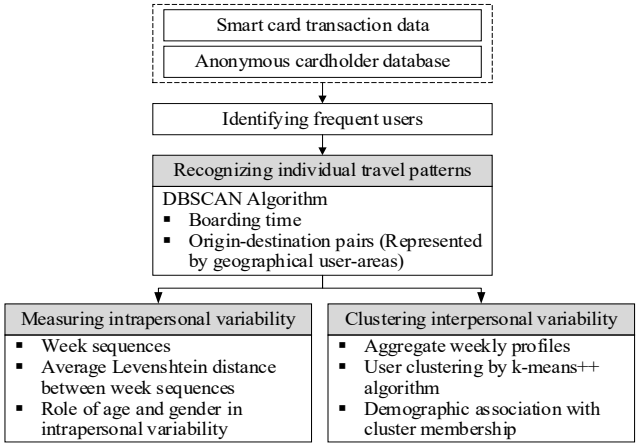


Fig. 1. Analysis framework

Fig. 1 illustrates the analysis framework of this study. We first identify individual travel patterns according to the boarding time and OD pairs of trips. Then, we build consecutive week sequences for each user and use the average Levenshtein distance to measure the intrapersonal variability. We further analyze the role of age and gender in intrapersonal variability. In addition, we construct aggregate weekly profiles for users and employ the profiles for user clustering. We then evaluate the demographic association of the identified clusters of weekly patterns.

A. Defining Geographical User Areas

There is usually more than one PT stop or station around activity locations. A PT user may use different PT stops or stations to access the same activity locations. Hence, stops and stations are grouped into geographical user areas using complete-distance hierarchical clustering. Similar to [4], a separate set of geographical areas $A_u = \{A_1, A_2, \ldots, A_{k_u}\}$ is defined for each user *u*. Let $S_u = \{s_1, s_2, \ldots, s_{n_u}\}$ be the set of all trip origin or destination stops or stations visited by a user *u*. Each s is initialized to a singleton area such that $A_u = \{A_1, A_2, \ldots, A_{n_u}\}$. Next, pairs of areas are successively merged until the smallest distance between two areas is not smaller than a predefined distance threshold D. The threshold D ensures that the distance between any stops and stations in the same area does not exceed a predefined walkable distance. The distance between two areas is measured as follows:

$$d(A_m, A_n) = \max(d'(s_i, s_j)) \quad s_i \in A_m, s_j \in A_n, \tag{1}$$

$$d'(s_i, s_j) = \begin{cases} d(s_i, s_j) & \text{if } N_{ij}/N_u < \lambda \\ \max(d(s_i, s_j), D) & \text{if } N_{ij}/N_u \geq \lambda \end{cases}, \tag{2}$$

where $d(s_i, s_j)$ is the Euclidean distance between s_i and s_j, N_{ij} is the number of trips made between s_i and s_j by user *u*, N_u is the total number of trips made by user *u*, and D and λ are predefined parameters. Because stops or stations between which a user frequently makes trips are likely to be associated with different activities, a parameter λ is introduced to ensure that these pairs of stops and stations are not grouped into the same user area [4]. To determine the appropriate values of D and λ, their impacts are evaluated through sensitivity analysis [6]. For each user, a different set of areas is defined for all 12 combinations of D = {500 m, 800 m, 1000 m} and λ = {10%, 30%, 50%, 100%}.

B. Recognizing Individual Travel Patterns

A PT user may repeatedly visit the same OD pairs during the same time period. To recognize individual spatiotemporal patterns, the boarding time and OD pairs are considered simultaneously for clustering. The OD pairs are represented by the geographical user areas of origin and destination stops or stations. The boarding time is characterized in terms of minutes from midnight. For example, 7:30 is represented as 450 and 18:00 as 1080.

The DBSCAN algorithm is used to recognize the spatiotemporal pattern of each user. This algorithm does not require the definition of the number of clusters and can identify clusters of arbitrary shapes. In addition, it is able to recognize noise and is robust to outliers [3], [13]. The main concept of DBSCAN is to locate dense regions that are separated by regions of lower density. Two key parameters need to be specified: the radius of the neighborhood (*eps*) and the minimum number of points within the neighborhood (*minPts*). For more details on the DBSCAN algorithm, refer to [29]. Considering the spatial and temporal characteristics simultaneously, the distance between trips $d'(t_i, t_j)$ is measured as follows:

$$d'(t_i, t_j) = \begin{cases} d(t_i, t_j) & \text{if } t_i \text{ and } t_j \text{ have the same OD pair} \\ T & \text{else} \end{cases}, \tag{3}$$

where $d(t_i, t_j)$ is the boarding time difference between trips t_i and t_j and T is a predefined threshold, which is greater than *eps* and ensures that trips with different OD pairs are not grouped into the same cluster.

To determine appropriate values of *eps* and *minPts*, their effects are evaluated through sensitivity analysis. For each user, a different set of clusters is defined for all 20 combinations of *eps* = {15 min, 30 min, 45 min, 60 min} and *minPts* = {4, 8, 12, 16, 20}.

C. Measuring Intrapersonal Variability

Based on the identified spatiotemporal patterns, each trip is linked to a cluster ID, which represents the cluster to which the trip belongs. The cluster IDs of trips undertaken on the same day are connected chronologically to derive a daily sequence. The same daily sequences represent an identical pattern of daily PT usage. Then, successive week sequences are built for each user, describing the distribution of all trips in space over time for each day of each week. For a period of N_w weeks, divided into N_d days per week (N_d = 7, from Monday to Sunday), each user's weekly profile is represented by an $N_w \times N_d$ matrix, whose each cell indicates a daily sequence. Table I presents an example of week sequences. "1,2" denotes two trips, with one trip belonging to cluster 1 and the other belonging to cluster 2. In addition, "0" represents the trip that is recognized as noise, and the daily sequence is characterized by "*n*" if no trips are undertaken on the day.

TABLE I
EXAMPLE OF WEEK SEQUENCES

Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
1,2	1,2	1,2	1,2	1,2	*n*	*n*
1,2	1,2	1,2	1,2	1,0,2	*n*	1,0
1	1,2	1,2	1,2	1,3	3	0,3
…	…	…	…	…	…	…

After building weekly profiles, we use the average Levenshtein distance between week sequences to measure the intrapersonal variability, as summarized in (4) and (5). A daily sequence is considered as a string. The Levenshtein distance is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into another and indicates how different two strings are. For

example, the Levenshtein distance between daily sequences "1,2" and "1,3" is 1 (substitute "2" with "3"), and the Levenshtein distance between "1,2" and "1,0,2" is 2 (delete "0" and ","). The higher the number, the more different are the two strings, suggesting that the daily patterns are more different. Further, the Levenshtein distance between week sequences is represented by the sum of the Levenshtein distances between daily sequences of the same day from two weeks:

$$d_u = \frac{2}{N_w \times (N_w - 1)} \sum_{j=i+1}^{N_w} \sum_{i=1}^{N_w - 1} d_u(w_i, w_j), \quad (4)$$

$$d_u(w_i, w_j) = \sum_{k=1}^{N_d} d_u(w_i^k, w_j^k), \quad (5)$$

where d_u is the average Levenshtein distance between week sequences for user u, $d_u(w_i, w_j)$ is the Levenshtein distance between week sequences w_i and w_j, N_w is the number of weeks in the analysis period, $d_u(w_i^k, w_j^k)$ denotes the Levenshtein distance of sequences between day k of week w_i and day k of week w_j, and N_d (= 7) represents the number of days per week. The smaller the average distance, the smaller is the variation in the weekly patterns of PT usage, and so, the more stable and predictable is the user over time.

D. Clustering Interpersonal Variability

To explore the interpersonal variability in weekly PT usage, we construct an aggregate weekly temporal profile for each user, describing the distribution of all trips over time for each day of the week. To reduce the dimension of the time of day, the trip boarding time is grouped into six time slots per day: before 7:00, 7:00 to 10:00, 10:00 to 14:00, 14:00 to 17:00, 17:00 to 21:00, and after 21:00. Each user's aggregate weekly profile is represented by a vector of 6 × 7 elements, each of which indicates the number of trips the user made during a time slot on each day of the week. For example, the first element of the vector denotes the number of trips made by the user before 7:00 on Monday, and the second denotes the number of trips the user made between 7:00 and 10:00 on Monday. It should be noted that the boarding time of a trip is represented by the temporal center of the cluster to which the trip is identified to belong (see subsection *B*). This representation of the boarding time addresses the limitation that the predefined time-window discretization may not be appropriate for everyone and that temporal patterns cannot be recognized well if the boarding time is distributed around the border of two time slots [13]. For the trips identified as noise, the real boarding time is used for representation.

Then, the weekly profiles of users are normalized by min–max feature scaling. The k-means++ algorithm [27] is applied to identify users sharing similar weekly patterns, considering that the k-means++ algorithm is fast and efficient in terms of the computational cost while handling large datasets. The k-means++ algorithm improves the initialization process using a randomized seeding technique and outperforms the standard k-means method in terms of both accuracy and speed.

After determining the clusters of all users, we link the cluster membership of each user to the anonymous cardholder database. Odds ratio analysis is performed to examine the association between the demographic characteristics and cluster membership. As summarized in (6), the odds ratio $OR_{c,k}$ measures how much more (or less) likely a user in cluster k is to have a given characteristic c than a user in other clusters [4].

$$OR_{c,k} = \frac{N_{c,k} / N_{c',k}}{N_{c,k'} / N_{c',k'}}, \quad (6)$$

where $N_{c,k}$ is the number of users having a characteristic c in cluster k and k' refers to the aggregation of all clusters other than k. The demographic characteristics except c are aggregated as c'. A value of $OR_{c,k}$ greater than 1 indicates a positive association and vice versa [10]. The natural logarithm of $OR_{c,k}$ is normally distributed, and whether $OR_{c,k}$ is significantly different from 1 at a given confidence level can be tested using the logit method [30].

V. Results

Based on the sensitivity analysis of all 12 combinations of D = {500 m, 800 m, 1000 m} and λ = {10%, 30%, 50%, 100%}, we obtain the following results: (1) an increase in the distance threshold D from 500 m to 800 m results in a significant decrease in the number of user areas; (2) when the value of D increases from 800 m to 1000 m, the number of users with a single area significantly increases; and (3) the number of trips with the origin and destination in the same area shows an upward trend as λ increases, which suggests that the smaller the value of λ, the better. Therefore, we set D to 800 m and λ to 10%, which indicates that all stops and stations are grouped into the same user areas of less than 800 m in diameter and that the OD pair between which the user made over 10% of trips is not grouped into the same area. Then, these OD areas and the boarding time of trips are used to identify individual travel patterns using the DBSCAN algorithm. The threshold T is set to 90 min. The values of *eps* = 30 min and *minPts* = 8 are selected based on the sensitivity analysis of all 20 combinations of *eps* = {15 min, 30 min, 45 min, 60 min} and *minPts* = {4, 8, 12, 16, 20}. The main reasons for selecting these values are as follows: (1) the number of clusters decreases with respect to both *eps* and *minPts*, and an increase in the value of *minPts* from 4 to 8 or an increase in the value of *eps* from 15 min to 30 min results in a significant decrease in the number of clusters; (2) the number of trips that are recognized as noise decreases significantly as the value of *eps* increases from 15 min to 30 min; and (3) the number of users with a single cluster shows a growing trend with increasing *minPts*.

After identifying individual travel patterns, we build 52 consecutive week sequences and an aggregate weekly profile for each user. The intrapersonal and interpersonal variability in weekly PT usage is further explored.

A. Intrapersonal Variability by Gender and Age

Table II summarizes the PT usage characteristics according to age and gender. In the frequent PT user group, female users

are more than males, especially among users aged 70 or over. On average, male users have a greater number of days traveled on weekdays than their female counterparts, but the males use PT less frequently on weekends. Compared to users under the age of 65, users aged 65 or over make fewer PT trips on weekdays but travel more frequently on weekends. Additionally, female users tend to have more spatiotemporal clusters than males, which implies that the PT usage of female users may be more diverse in space and time. In addition, users aged 65 or over have a higher average number of spatiotemporal clusters than younger users, suggesting that elderly users are more likely to have diffuse spatiotemporal patterns.

TABLE II
DEMOGRAPHIC STATISTICS OF PUBLIC TRANSPORT USAGE

		Age						
		<18	18–24	25–40	41–59	60–64	65–69	70+
V1	M	1655	1708	3230	5482	1211	684	960
	F	2742	4045	4731	6107	1165	1142	2787
V2	M	344	330	379	386	369	340	317
	F	348	321	355	363	315	313	303
V3	M	12	20	22	31	23	43	17
	F	35	13	52	36	77	79	82
V4	M	184.49	177.10	197.31	199.60	192.48	177.67	162.89
	F	183.36	172.77	189.42	189.00	175.04	167.61	157.05
V5	M	18.16	19.22	21.18	26.37	23.81	33.39	20.23
	F	29.76	19.46	36.39	31.08	45.40	42.76	45.68
V6	M	4.06	4.45	3.93	3.97	4.00	4.89	6.60
	F	4.23	4.87	4.24	4.30	4.67	5.99	7.47

Notes: The first column represents variables characterizing PT usage: V1 - Number of users; V2 - Median number of trips (weekdays); V3 - Median number of trips (weekends); V4 - Average number of weekdays traveled; V5 - Average number of weekend days traveled; V6 - Average number of spatiotemporal clusters. The second column denotes gender: M - Male, F - Female.

Based on 52-week sequences of each user, we calculate the average Levenshtein distance between week sequences to measure the intrapersonal variability in weekly PT usage. Analysis of variance (ANOVA) is used to examine the role of gender and age in intrapersonal variability. As indicated in Table III, gender and age have significant influences on intrapersonal variability. The interaction effect of gender and age is also statistically significant. Fig. 2 illustrates the distribution of the average distance between week sequences by age and gender.

TABLE III
ANOVA SUMMARY TABLE

Variable	df	SS	MS	F-value	p-value
Gender	1	9525	9525	638.030	<0.001
Age	6	71479	11913	798.019	<0.001
Gender*Age	6	859	143	9.595	<0.001
Residuals	37635	561834	15		
Total	37648	643697			

For the gender differential in intrapersonal variability, the average distances are greater for female users on average, indicating that female users have higher intrapersonal variability than their male counterparts in weekly PT usage. In particular, gender differentials are greater among users aged 65–69 than among users of other age groups. The reason may be that although the retirement age is typically 65, elderly men may be more likely to continue working until the age of 70 than elderly women, which results in a higher proportion of commuting trips and less diverse spatiotemporal patterns among elderly men aged 65–69 than among their female counterparts. This is demonstrated by the travel survey data. According to the 4th Shizuoka Metropolitan Area Person Trip Survey conducted in 2012, for elderly men and women aged 65–69, 19% and 12% of PT trips were for work purposes, respectively.

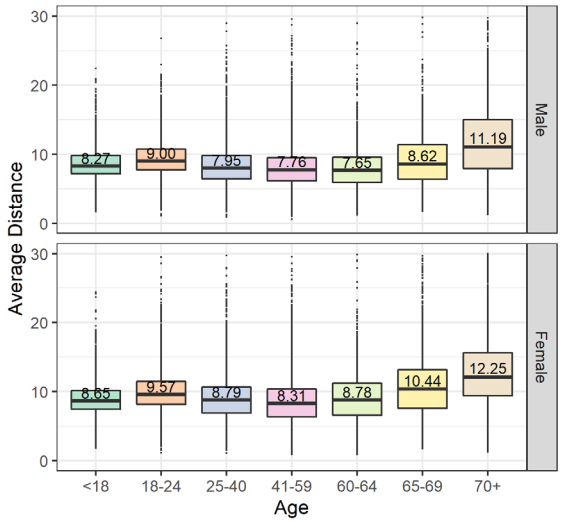


Fig. 2. Intrapersonal variability by age and gender

Regarding the age effect on intrapersonal variability, the weekly patterns are the most diverse among users aged 70 or over, followed by the users aged 65–69. People aged 65 or over are usually retirees and have a flexible schedule, resulting in more diffuse travel patterns. Moreover, some people continue working until 70, which explains the observation that the weekly patterns of users aged 70 or over are the most variable. In addition, users aged 18–24 have the third most diverse weekly patterns because users in this age group are usually undergraduate or graduate students who have a relatively flexible schedule. For users in other age groups, their weekly patterns are less diverse because these users are typically students or workers who have a fixed commuting time and limited activity locations.

B. Interpersonal Variability Analysis

The k-means++ algorithm is used to group PT users with similar weekly profiles. Fig. 3 shows the Davies–Bouldin index (DBI), a measure of internal cluster validation based on the ratio of the within-cluster scatter to the between-cluster separation. A lower DBI value indicates that the clustering results are better. For values of k from 2 to 20, the clustering results are the most compact when $k = 5$. Therefore, PT users are grouped into five clusters, each characterized by a distinct weekly profile. Fig. 4 illustrates the centers of these five clusters. The center of the

cluster is the average of all the elements that belong to the cluster, representing the average trip frequency over time for each day of the week.

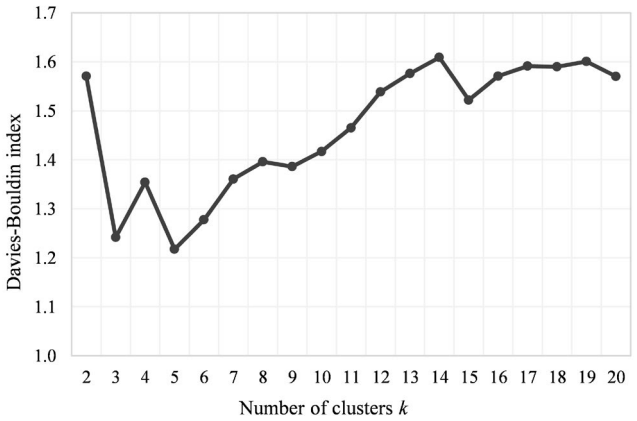


Fig. 3. Davies–Bouldin index for clustering

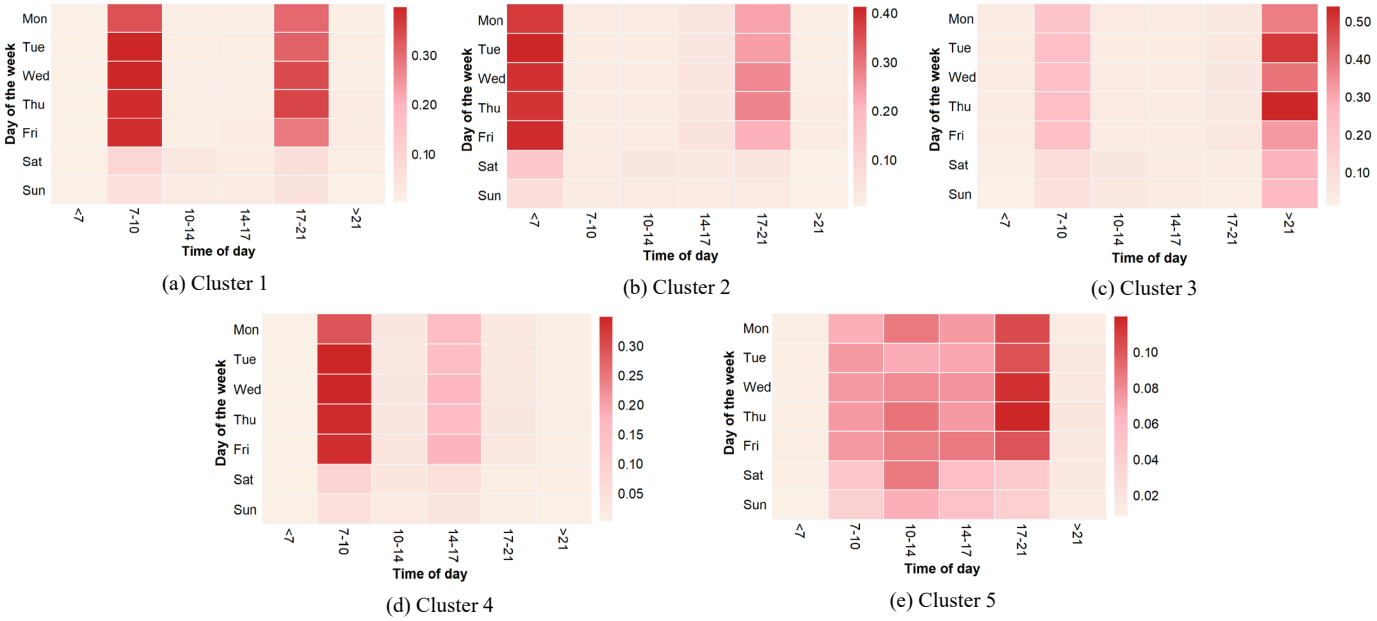
Cluster 1: Users in cluster 1 are characterized by a higher percentage of PT usage during the periods of 7:00–10:00 and 17:00–21:00 on workdays and by reduced travel on weekends. As indicated in Table IV, cluster 1 is the most common, accounting for 40% of the frequent users. On average, users in this cluster visit a smaller number of distinct locations but complete more trips. Odds ratio analysis (Table V) indicates that users aged 25–59 are two times more likely to be assigned to cluster 1. These characteristics suggest that the users in cluster 1 take PT exclusively for work purposes.

Cluster 2: Like users in cluster 1, users in cluster 2 are associated with distinct working days and decreased travel on weekends. However, users in this cluster typically depart earlier in the morning on weekdays, typically before 7:00. Odds ratio analysis reveals that users in cluster 2 are more likely to be younger than 18 or aged 41–64. In addition, male users are 2.50 times more likely to be classified into this cluster. These characteristics imply that cluster 2 is mainly composed of male students and senior male workers, who also use PT as their primary mode of commuting.

Cluster 3: Unlike users in clusters 1 and 2, users in cluster 3 are characterized by distinct working days and frequent travel on weekends. They tend to return later in the evening, typically after 21:00. Based on odds ratio analysis, users aged 25–40 are 2.09 times more likely to be assigned to cluster 4, and male users are 1.99 times more likely to be classified into this cluster. These characteristics imply that cluster 3 is closely associated with young male users who use PT for work and other purposes.

Cluster 4: The major features characterizing cluster 4 are frequent PT usage during the periods of 7:00–10:00 and 14:00–17:00 on weekdays and reduced travel on weekends. The temporal distribution of trips is consistent with school hours in Japan. Odds ratio analysis indicates that users younger than 18 are 4.94 times more likely to be assigned to this cluster. These characteristics suggest that users in cluster 4 are mainly students.

Cluster 5: This cluster accounts for 27% of the frequent users and is characterized by no distinct working days, no distinct departure time, a smaller number of trips and days traveled, and more distinct locations visited. Users in this cluster are 2.60 times more likely to be aged 65–69 and 7.54 times more likely to be older than 70. This suggests that cluster 5 is most strongly associated with retirees, in line with the analysis of the age effect on intrapersonal variability. In addition, female users are 1.98 times more likely to be assigned to cluster 5, implying that elderly women have a high probability of using PT.



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5

Fig. 4. Representation of five centers of weekly profile clusters. The color coding represents the average normalized trip frequency, which is set independently for each cluster to make patterns more apparent.

TABLE IV
DESCRIPTIVE STATISTICS OF CLUSTERS

Cluster ID	1	2	3	4	5
Number of users	14996	3468	1271	7718	10196
Percentage of frequent users	40%	9%	3%	20%	27%
Public transport usage characteristics					
Median number of trips (weekdays)	405	405	360	337	227
Median number of trips (weekends)	28	26	74	17	53
Average number of weekdays traveled	206.76	206.96	190.40	183.46	143.10
Average number of weekend days traveled	26.18	24.81	46.12	20.44	35.85
Average number of distinct locations	7.07	7.39	6.84	7.49	9.73
Demographic attributes					
Age (median)	41	46	39	26	52
Male (%)	44%	14%	5%	18%	19%
Female (%)	37%	6%	2%	22%	32%

The clustering results reveal that the k-means++ algorithm can handle the data well: (1) the clustering results are reasonable, and the travel behavior characteristics of each cluster reflect the sociodemographic characteristics of cluster members and (2) the average silhouette index is 0.42, suggesting that the clustering results are acceptable. The silhouette index is another internal clustering validation measure in addition to DBI. It ranges from -1 to 1, and a value closer to 1 indicates better clustering results.

TABLE V
ODDS RATIO STATISTICS

Cluster ID	1	2	3	4	5
Age					
<18	0.39	1.32	0.18	4.94	0.48
18–24	0.84	0.58	1.09	1.56	0.99
25–40	2.21	0.73	2.09	0.49	0.58
41–59	1.97	1.64	1.29	0.48	0.56
60–64	1.06	1.36	0.51	0.74	1.09
65–69	0.38	0.90	0.75	1.00	2.60
70+	0.09	0.48	0.24	0.89	7.54
Gender					
Male	1.34	2.50	1.99	0.76	0.50
Female	0.74	0.40	0.50	1.32	1.98

Note: All odds ratios are significant at the 95% confidence level.

VI. DISCUSSION AND CONCLUSION

Based on 52-week smart card transactions and an anonymous cardholder database in Shizuoka, Japan, we explored the intrapersonal and interpersonal variability in PT usage as well as the role of gender and age in travel pattern variability.

Our methodological contribution is that we propose a novel representation of week sequences that simultaneously considers multiple dimensions of travel behavior and develop a measure to quantify the intrapersonal variability in weekly PT usage. Existing methods to measure intrapersonal variability remain limited in scope. Some methods focus on the relative frequency of trips [12], [13], and some rely on the variance of trips [8]. However, these methods usually do not consider how multiple trips are combined. In contrast, the proposed measure of intrapersonal variability can consider variability in multiple dimensions of PT usage, including the boarding time and OD pairs of trips as well as the organization of multiple trips over different times of the day. In addition, we develop aggregate weekly profiles for users and examine interpersonal variability by identifying clusters of users sharing similar weekly patterns.

Our empirical contribution pertains to the large-scale application of the proposed methodology to Shizuoka's PT system and the role of age and gender in the intrapersonal and interpersonal variability in PT usage revealed by a large sample over a long period. Both age and gender have a significant impact on the intrapersonal variability in weekly PT usage. Among frequent PT users, female users show higher intrapersonal variability than male users, which is in line with the results of previous studies [8], [9]. This may result from the division of roles in the family and in the labor market [31]. Although women's labor force participation and household roles are changing, women usually undertake the majority of household-serving travel [32]. According to the 4th Shizuoka Metropolitan Area Person Trip Survey, 70% of household-serving trips (e.g., chauffeuring, shopping, and errands) were made by female travelers and only 30% were completed by male travelers. Additionally, the gender differentials in intrapersonal variability seem greater among frequent users aged 65–69 than among other age groups. According to interpersonal variability clustering results, frequent users take PT primarily for work or school purposes apart from those in cluster 5, who are most strongly associated with the elderly aged 65 or over. Further, 19% and 12% of PT trips are for work purposes among elderly men and women aged 65–69, respectively, according to the 4th Shizuoka Metropolitan Area Person Trip Survey. A higher proportion of commuting trips may result in less diverse spatiotemporal patterns among elderly men than among their female counterparts.

Regarding the association between age and intrapersonal variability, the weekly patterns are the most diverse for users aged 70 or over, followed by those aged 65–69; the users aged 18–24 have the third most variable weekly patterns. In contrast, the weekly patterns are less diffuse for users in other age groups. The role of age in intrapersonal variability is consistent with the employment status of PT users. For example, people aged 65 or over are usually retirees who have a flexible schedule and no limitations on activity locations, resulting in more diverse travel patterns. Some people keep working until 70, which explains the observation that the weekly patterns of users aged 70 or over are the most variable. It is worth mentioning that the intrapersonal variability in PT usage may be influenced by the household role, employment status, income, etc. in addition to gender and age. Because these explanatory variables are not available from smart card data and most of them are closely associated with gender and age, we examine the role of age and gender in intrapersonal variability by assuming that age- and gender-based groups are comparable. In the future, the effect of age and gender on the intrapersonal variability in PT usage can

be examined while controlling for other sociodemographic characteristics.

Regarding the interpersonal variability in PT usage, we identified five clusters of PT users, each associated with a distinct weekly profile. Clusters 1, 2, and 4 are characterized by distinct working days and reduced travel on weekends, whereas PT users in cluster 3 travel frequently on both weekdays and weekends. Cluster 5 is most strongly associated with retirees and is characterized by no distinct working days, no distinct departure time, and more distinct locations visited. Overall, commuters (clusters 1–3: primarily for work purposes; cluster 4: primarily for school purposes) and the elderly (cluster 5) comprise the frequent PT user group. Further, odds ratio analysis reveals the association of gender and age with the interpersonal variability in PT usage. Male users are more likely to depart earlier in the morning (cluster 2) and return later in the evening (cluster 3) than their female counterparts. Elderly women have a higher probability of using PT than elderly men (cluster 5). Regarding the age impact on interpersonal variability, the temporal distribution of trips in each cluster is consistent with the most likely employment status for the age group in which the cluster members are most likely to be. For instance, users younger than 18 are most likely to be students. They are 4.94 times more likely to be assigned to cluster 4, whose temporal distribution of trips is in line with the school hours in Japan. This suggests the value of smart card data in forecasting certain demographic characteristics of PT users.

The study findings provide a better understanding of the travel pattern variability of PT users and can be used to improve passenger experience and service provision. For example, users in clusters 1–4 usually travel during fixed time slots. These users with lower intrapersonal variability can be identified as spatiotemporally regular users, and they can be provided personalized information about disruptions and crowding levels. If users are informed in advance that their commute trips are expected to be disrupted or highly crowded, they can reschedule trips. In addition, users in cluster 5 with higher intrapersonal variability have more flexible schedules, and they are most likely to respond to policies that aim to divert users away from highly crowded routes and time periods (e.g., off-peak discounts and volume rebates). To maximize the effectiveness of the policies, they should be targeted through travel demand management campaigns. In addition, certain types of monthly passes can be designed. Users in cluster 5 are closely associated with the elderly aged 65 or over and are characterized by no distinct departure time and more distinct locations visited. Hence, monthly passes with no restrictions on time and space can be issued to the elderly aged 65 or over. Although users in clusters 1–4 are characterized by regular departure times, they have approximately seven distinct activity locations on average. The users in clusters 1–4 with higher intrapersonal variability (that is, probably a greater number of distinct activity locations) may prefer multiple-destination commuter passes [33] over typical commuter passes that are limited to a single spatial interval.

This study indicates that age and gender play an important role in the intrapersonal and interpersonal variability in PT usage. Female users exhibit higher intrapersonal variability than their male counterparts, and the weekly patterns are more diverse over time for the elderly aged 65 or over. Regarding interpersonal variability, five clusters of users are identified, each characterized by a distinct weekly profile and associated with certain age groups and gender. Nevertheless, this study has some limitations that need further investigation. First, we focused on frequent PT users in this study. Although they made over 69% of the trips in Shizuoka's PT system, the analysis of low-frequency users and intermittent users may provide further valuable insight to PT authorities. Second, we analyzed the role of age and gender in the variability in PT usage without considering other sociodemographic characteristics (e.g., household role, employment status, occupation) because of the unavailability of data. A survey can be distributed by mail to registered smart card holders with registered home addresses to collect more sociodemographic characteristics of PT users. Then, the relationship between the variability in PT usage and more sociodemographic characteristics can be evaluated. Third, we do not consider the influence of external factors (e.g., weather, holidays, and big events) on PT usage. The consideration of the impact of external factors may provide a more in-depth understanding of passenger travel patterns. Fourth, we do not distinguish between urban railway users and bus users. Because the travel habits and sociodemographic characteristics of railway users and bus users may be different, the travel patterns of railway users and bus users should be examined separately in future research.

REFERENCES

[1] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao, "Measuring regularity of individual travel patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1583–1592, May 2018, DOI: 10.1109/TITS.2017.2728704.

[2] E. Deschaintres, C. Morency, and M. Trépanier, "Analyzing transit user behavior with 51 weeks of smart card data," *Transp. Res. Rec.*, vol. 2673, no. 6, pp. 33–45, Apr. 2019, DOI: 10.1177/0361198119834917.

[3] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. Part C Emerg. Technol.*, vol. 36, pp. 1–12, Nov. 2013, DOI: 10.1016/j.trc.2013.07.010.

[4] G. Goulet-Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. Part C Emerg. Technol.*, vol. 64, pp. 1–16, Mar. 2016, DOI: 10.1016/j.trc.2015.12.012.

[5] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation (Amst).*, vol. 35, no. 6, pp. 697–722, Aug. 2008, DOI: 10.1007/s11116-008-9178-4.

[6] G. Goulet-Langlois, "Exploring regularity and structure in travel behavior using smartcard data," M.S. thesis, Dept. Civil Environ. Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 2015.

[7] R. Schlich and K. W. Axhausen, "Habitual travel behaviour: Evidence from a six-week travel diary," *Transportation (Amst).*, vol. 30, no. 1, pp. 13–36, Feb. 2003, DOI: 10.1023/A:1021230507071.

[8] C. Raux, T. Y. Ma, and E. Cornelis, "Variability in daily activity-travel patterns: The case of a one-week travel diary," *Eur. Transp. Res. Rev.*, vol. 8, no. 4, Oct. 2016, DOI: 10.1007/s12544-016-0213-9.

[9] A. Moiseeva, H. Timmermans, J. Choi, and C. H. Joh, "Sequence alignment analysis of variability in activity travel patterns through 8 weeks of diary data," *Transp. Res. Rec.*, vol. 2412, no. 1, pp. 49–56, Jan. 2014, DOI: 10.3141/2412-06.

[10] O. Egu and P. Bonnel, "Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon,"

Travel Behav. Soc., vol. 19, pp. 112–123, Apr. 2020, DOI: 10.1016/j.tbs.2019.12.003.

[11] M. P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, Aug. 2011, DOI: 10.1016/j.trc.2010.12.003.

[12] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017, DOI: 10.1109/TITS.2017.2679179.

[13] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015, DOI: 10.1109/TITS.2014.2368998.

[14] M. K. El Mahrsi, E. Come, L. Oukhellou, and M. Verleysen, "Clustering smart card data for urban mobility analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 712–728, Mar. 2017, DOI: 10.1109/TITS.2016.2600515.

[15] K. Hamilton and L. Jenkins, "A gender audit for public transport: A new policy tool in the tackling of social exclusion," *Urban Stud.*, vol. 37, no. 10, pp. 1793–1800, Sep. 2000, DOI: 10.1080/00420980020080411.

[16] F. Shao, Y. Sui, X. Yu, and R. Sun, "Spatio-temporal travel patterns of elderly people – A comparative study based on buses usage in Qingdao, China," *J. Transp. Geogr.*, vol. 76, pp. 178–190, Apr. 2019, DOI: 10.1016/j.jtrangeo.2019.04.001.

[17] H. Nishiuchi, J. King, and T. Todoroki, "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data," *Int. J. Intell. Transp. Syst. Res.*, vol. 11, pp. 1–10, Aug. 2012, DOI: 10.1007/s13177-012-0051-7.

[18] C. Zhong, E. Manley, S. Müller Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *J. Comput. Sci.*, vol. 9, pp. 125–130, Jul. 2015, DOI: 10.1016/j.jocs.2015.04.021.

[19] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *J. Transp. Geogr.*, vol. 41, pp. 21–36, Dec. 2014, DOI: 10.1016/j.jtrangeo.2014.08.006.

[20] M. S. Ghaemi, B. Agard, M. Trépanier, and V. Partovi Nia, "A visual segmentation method for temporal smart card data," *Transp. A Transp. Sci.*, vol. 13, no. 5, pp. 381–404, Jan. 2017, DOI: 10.1080/23249935.2016.1273273.

[21] E. I. Pas and F. S. Koppelman, "An examination of the determinants of day-to-day variability in individuals' urban travel behavior," *Transportation (Amst).*, vol. 13, pp. 183–200, Jun. 1986, DOI: 10.1007/BF00165547.

[22] J. Kandt and A. Leak, "Examining inclusive mobility through smartcard data: What shall we make of senior citizens' declining bus patronage in the West Midlands?," *J. Transp. Geogr.*, vol. 79, p. 102474, Jul. 2019, DOI: 10.1016/j.jtrangeo.2019.102474.

[23] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, May 2007, DOI: 10.1016/j.tranpol.2007.01.001.

[24] W. Y. Szeto, L. Yang, R. C. P. Wong, Y. C. Li, and S. C. Wong, "Spatio-temporal travel characteristics of the elderly in an ageing society," *Travel Behav. Soc.*, vol. 9, pp. 10–20, Oct. 2017, DOI: 10.1016/j.tbs.2017.07.005.

[25] D. Yang, H. Timmermans, and A. Grigolon, "Exploring heterogeneity in travel time expenditure of aging populations in the Netherlands: Results of a CHAID analysis," *J. Transp. Geogr.*, vol. 33, pp. 170–179, Dec. 2013, DOI: 10.1016/j.jtrangeo.2013.10.002.

[26] Y. O. Susilo and R. Kitamura, "Analysis of day-to-day variability in an individual's action space: Exploration of 6-week mobidrive travel diary data," *Transp. Res. Rec.*, vol. 1902, no. 1, pp. 124–133, Jan. 2005, DOI: 10.3141/1902-15.

[27] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discret. Algorithms*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[28] J. B. Gordon, H. N. Koutsopoulos, N. H. M. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in London using fare-transaction and vehicle location data," *Transp. Res. Rec.*, vol. 2343, no. 1, pp. 17–24, Jan. 2013, DOI: 10.3141/2343-03.

[29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Min. (KDD-96)*, Portland, OR, USA, 1996, pp. 226–231.

[30] J. A. Morris and M. J. Gardner, "Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates," *Br. Med. J. (Clin. Res. Ed).*, vol. 296, no. 6632, pp. 1313–1316, May 1988, DOI: 10.1136/bmj.296.6632.1313.

[31] W. Ng and A. Acker, "Understanding urban travel behavior by gender for efficient and equitable transport policies," in *International Transport Forum Discussion Papers*, Paris, France, 2018, DOI: 10.1787/eaf64f94-en.

[32] S. Rosenbloom, "Understanding women's and men's travel patterns: The research challenge," in *Res. Women's Issues Transp., Vol. 1: Conf. Overv. Plenary Pap.*, 2006, pp. 7–24, DOI: 10.17226/23274.

[33] T. Yamamoto, S. Liu, and T. Nakamura, "Variability of passenger travel patterns observed using smart card data in Japan," in *Routledge Companion of Public Transport*, 1st ed., C. Mulley, J. Nelson, and S. Ison, Eds., __: Routledge, 2021.

[1] G. Goulet-Langlois, H. N. Koutsopoulos, Z. Zhao, and J. Zhao, "Measuring Regularity of Individual Travel Patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1583–1592, 2018, doi: 10.1109/TITS.2017.2728704.

[2] E. Deschaintres, C. Morency, and M. Trépanier, "Analyzing Transit User Behavior with 51 Weeks of Smart Card Data," *Transp. Res. Rec.*, vol. 2673, no. 6, pp. 33–45, 2019, doi: 10.1177/0361198119834917.

[3] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. Part C Emerg. Technol.*, vol. 36, pp. 1–12, 2013, doi: 10.1016/j.trc.2013.07.010.

[4] G. Goulet Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transp. Res. Part C Emerg. Technol.*, vol. 64, pp. 1–16, 2016, doi: 10.1016/j.trc.2015.12.012.

[5] R. N. Buliung, M. J. Roorda, and T. K. Remmel, "Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS)," *Transportation (Amst).*, vol. 35, no. 6, pp. 697–722, 2008, doi: 10.1007/s11116-008-9178-4.

[6] G. Goulet-Langlois, "Exploring Regularity and Structure in Travel Behavior Using Smartcard Data," Massachusetts Institute of Technology, 2015.

[7] R. Schlich and K. W. Axhausen, "Habitual travel behaviour: Evidence from a six-week travel diary," *Transportation (Amst).*, vol. 30, no. 1, pp. 13–36, 2003, doi: 10.1023/A:1021230507071.

[8] C. Raux, T. Y. Ma, and E. Cornelis, "Variability in daily activity-travel patterns: the case of a one-week travel diary," *Eur. Transp. Res. Rev.*, vol. 8, no. 4, 2016, doi: 10.1007/s12544-016-0213-9.

[9] A. Moiseeva, H. Timmermans, J. Choi, and C. H. Joh, "Sequence alignment analysis of variability in activity travel patterns through 8 weeks of diary data," *Transp. Res. Rec.*, no. 2412, pp. 49–56, 2014, doi: 10.3141/2412-06.

[10] O. Egu and P. Bonnel, "Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon," *Travel Behav. Soc.*, vol. 19, no. August 2019, pp. 112–123, 2020, doi: 10.1016/j.tbs.2019.12.003.

[11] M. P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011, doi: 10.1016/j.trc.2010.12.003.

[12] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, 2017, doi: 10.1109/TITS.2017.2679179.

[13] L. M. Kieu, A. Bhaskar, and E. Chung, "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, 2015, doi: 10.1109/TITS.2014.2368998.

[14] M. K. El Mahrsi, E. Come, L. Oukhellou, and M. Verleysen, "Clustering Smart Card Data for Urban Mobility Analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 712–728, 2017, doi: 10.1109/TITS.2016.2600515.

[15] K. Hamilton and L. Jenkins, "A Gender Audit for Public Transport: A New Policy Tool in the Tackling of Social Exclusion Tracking and Tackling Social Exclusion: The Contribution of a Gender Audit," *Urban Stud.*, vol. 37, no. 10, pp. 1793–1800, 2000, doi: 10.1080/00420980020080411.

[16] F. Shao, Y. Sui, X. Yu, and R. Sun, "Spatio-temporal travel patterns of elderly people – A comparative study based on buses usage in Qingdao, China," *J. Transp. Geogr.*, vol. 76, no. 308, pp. 178–190, 2019, doi: 10.1016/j.jtrangeo.2019.04.001.

[17] H. Nishiuchi, J. King, and T. Todoroki, "Spatial-Temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data," *Int. J. Intell. Transp. Syst. Res.*, vol. 11, no. 1, pp. 1–10, 2013, doi: 10.1007/s13177-012-0051-7.

[18] C. Zhong, E. Manley, S. Müller Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *J. Comput. Sci.*, vol. 9, pp. 125–130, 2015, doi: 10.1016/j.jocs.2015.04.021.

[19] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *J. Transp. Geogr.*, vol. 41, pp. 21–36, 2014, doi: 10.1016/j.jtrangeo.2014.08.006.

[20] M. S. Ghaemi, B. Agard, M. Trépanier, and V. Partovi Nia, "A visual segmentation method for temporal smart card data," *Transp. A Transp. Sci.*, vol. 13, no. 5, pp. 381–404, 2017, doi: 10.1080/23249935.2016.1273273.

[21] E. I. Pas and F. S. Koppelman, "An examination of the determinants of day-to-day variability in individuals' urban travel behavior," *Transportation (Amst).*, vol. 13, no. 2, pp. 183–200, 1986, doi: 10.1007/BF00165547.

[22] J. Kandt and A. Leak, "Examining inclusive mobility through smartcard data: What shall we make of senior citizens' declining bus patronage in the West Midlands?," *J. Transp. Geogr.*, vol. 79, no. June, p. 102474, 2019, doi: 10.1016/j.jtrangeo.2019.102474.

[23] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, 2007, doi: 10.1016/j.tranpol.2007.01.001.

[24] W. Y. Szeto, L. Yang, R. C. P. Wong, Y. C. Li, and S. C. Wong, "Spatio-temporal travel characteristics of the elderly in an ageing society," *Travel Behav. Soc.*, vol. 9, pp. 10–20, 2017, doi: 10.1016/j.tbs.2017.07.005.

[25] D. Yang, H. Timmermans, and A. Grigolon, "Exploring heterogeneity in travel time expenditure of aging populations in the Netherlands: Results of a CHAID analysis," *J. Transp. Geogr.*, vol. 33, pp. 170–179, 2013, doi: 10.1016/j.jtrangeo.2013.10.002.

[26] Y. O. Susilo and R. Kitamura, "Analysis of day-to-day variability in an individual's action space: Exploration of 6-week mobidrive travel diary data," *Transp. Res. Rec.*, no. 1902, pp. 124–133, 2005, doi: 10.3141/1902-15.

[27] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proc. Annu. ACM-SIAM Symp. Discret. Algorithms*, vol. 07-09-Janu, pp. 1027–1035, 2007.

[28] J. B. Gordon, H. N. Koutsopoulos, N. H. M. Wilson, and J. P. Attanucci, "Automated inference of linked transit journeys in london using fare-transaction and vehicle location data," *Transp. Res. Rec.*, vol. 2343, pp. 17–24, 2013, doi: 10.3141/2343-03.

[29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, vol. 96, pp. 226–231, doi: 10.1016/B978-044452701-1.00067-3.

[30] J. A. Morris and M. J. Gardner, "Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates," *Br. Med. J. (Clin. Res. Ed).*, vol. 296, no. 6632, pp. 1313–1316, 1988, doi: 10.1136/bmj.296.6632.1313.

[31] W. Ng and A. Acker, "Understanding urban travel behavior by gender for efficient and equitable trnasport policies," in *International Transport Forum, Paris, France*, 2018, doi:10.1787/eaf64f94-en.

[32] S. Rosenbloom, "Understanding women's and men's tavel patterns: The research challenge," in *Research on Women's Issues in Transportation, Volume 1: Conference Overview and Plenary Papers*, 2006, vol. 1, pp. 7–24, doi: 10.17226/23274.

[33] T. Yamamoto, S. Liu, and T. Nakamura, "Variability of passenger travel patterns observed using smart card data in Japan," in *Routledge Companion of Public Transport*, 2020.