A Method for Extracting Travel Patterns Using Data Polishing

Mio Hosoe (hosoemio@gmail.com)
Tottori University: Tottori Daigaku https://orcid.org/0000-0002-7728-1010
Masashi Kuwano
Tottori University: Tottori Daigaku
Taku Moriyama
Tottori University: Tottori Daigaku

Research

Keywords: public transport, smart card, high order data

Posted Date: September 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-78074/v1

Version of Record: A version of this preprint was published on January 7th, 2021. See the published version at https://doi.org/10.1186/s40537-020-00402-w.

Title

A Method for Extracting Travel Patterns using Data polishing

Authors

Name: Mio Hosoe

Address: 4-101 Koyama-Minami, Tottori, Japan

E-mail: d19t4003b@edu.tottori-u.ac.jp

Name: Masashi Kuwano

Address: 4-101 Koyama-Minami, Tottori, Japan

E-mail: kuwano@tottori-u.ac.jp

Name: Taku Moriyama

Address: 4-101 Koyama-Minami, Tottori, Japan

E-mail: moriyama@tottori-u.ac.jp

Abstract

With the development of ICT (Information and Communication Technology), interest in using the large amount of accumulated data for traffic policy planning has been increasing. In recent years, data polishing has been proposed as a new methodology for big data analysis. Data polishing is one of the graphical clustering methods. This method can be used to extract patterns that are similar or related to each other by clarifying the cluster structures in the data. The purpose of this study is to reveal travel patterns of railway passengers by applying data polishing to smart card data collected in Kagawa Prefecture, Japan. This study uses 9,008,709 data points collected during the 15 months from December 1st, 2013 to February 28th, 2015. This data set includes such information as trip histories and types of passengers. The study uses the data polishing method to cluster 4,667,520 combinations of information about individual rides: day of the week, time of day, passenger type, origin station, and destination station. As a result, 127 characteristic travel patterns were specified from those combinations.

Keywords

public transport, smart card, high order data

1. Introduction

Various kinds of data have been generated and accumulated in real time with the development of ICT (Information and Communication Technology). The amount and type of data available has increased, and big data is attracting a lot of attention. Big data includes many kinds of data from various fields including social media data, multimedia data, sensor data and log data. Analysis and visualization of these big data is expected to enable recognition of phenomena which could not be observed previously and therefore to make possible the creation of new knowledge.

In the field of transportation research, big data such as GPS data and probe vehicle data have been analyzed to better understand people's travel behaviors [1, 2, 3]. In particular, many researchers have been analyzing smart card data to understand transit users' behaviors [4, 5, 6]. Smart cards were originally developed for fare payment and/or toll collecting. However, smart card data contain information about which people passed the ticket gate at which station at what time of day, as well as at which station and at what time of day they ended their trips. Therefore, they allow us to understand the temporal and spatial travel behaviors of card users.

Analyzing smart card data is important to understand travel behaviors. Also, it is expected that the results may be used as new material for consideration in development of traffic policy. However, most previous studies have focused only on a small number of data items in smart card data: the number of uses by day of the week and time of day, by origin time of day, origin point and destination point, etc [7, 8]. Or there are travel behavior analysis focused on single data items and specific elements

of data items [9, 10, 11]. It cannot be said that these studies have considered multiple data items in smart card data simultaneously.

It is accepted that multiple attributes in smart card data (such as origin station, destination station, time of day, day of the week, and passenger type) affect each other. However, if analysts focus on only specific data items extracted from smart card data, they cannot explore whether there is an effect of the excluded data items on travel behaviors. Moreover, results may differ depending on which data items are analyzed. From the perspective of effective big data analysis, this study considers that as many data items as possible should be analyzed simultaneously. By using smart card data in the provincial city of Japan, this study reveals the travel behavior patterns. This study uses five data items (hereafter, attributes) in the analysis: boarding day of the week (hereafter, day), boarding time of day (hereafter, time), passenger type, origin station, and destination station.

Methodologies using tensor decomposition have been proposed for simultaneously considering multiple attributes [12, 13, 14]. Tensor decomposition can be an effective analysis method for 3rd or higher order data [15]. Moreover, this method enables analysis while maintaining the original data structure [16]. A tensor representation permits us to summarize multivariate data in a multi-dimensional array. The lowest order tensors have specific common names: a 0-order tensor is a scalar, a 1st order tensor is a vector, and a 2nd order tensor is a matrix [14, 17]. Tucker decomposition—one model of tensor decomposition—estimates factor matrices representing characteristics of each attribute in high order data [18, 19, 20]. The characteristics of each attribute are called factors. The number of factors is determined arbitrarily [19]. In addition, a core tensor representing the combination of factors for each attribute is estimated at the same time as the factor matrices [18, 19, 20]. Tucker decomposition enables us to understand interactions between attributes in the original data using the estimated factor matrices and core tensor. However, tensor decomposition, like factor analysis, has greater complexity of results with an increasing number of attributes. Also, as the number of elements for each attribute increases, it becomes more difficult with tensor decomposition to uniquely determine the number of factors and interpret the components of the factors [12]. As a result, it may be difficult to understand and interpret factor matrices and core tensors [12]. Moreover, tensor decomposition cannot extract characteristics of elements having the small number of samples since it trends to depend on the number of samples.

This study tries to consider multiple attributes simultaneously by constructing graph. The more combinations of attributes, that is, the more vertices and edges, the more complex the graph structure. It is difficult to grasp data characteristics from the complex graph. Accordingly, this study extracts groups of more relevant vertices from graph by the similarity of vertices. Several pattern extraction methods using the similarity index have proposed [21, 22, 23]. However, their methods are not suitable for extracting patterns from graphs. The graph is represented by a two-dimensional table, which is the symmetric matrix and the diagonal elements are zero. In this case, the combinations of column information on each row are different. This should be noted for considering the pattern extraction method.

This study focuses data polishing approach to extract travel patterns from the graph. It clarifies group boundaries based on a hypothesis: two vertices have many common neighbors in a graph if they are included in a dense sub-graph of a certain size [24]. In data polishing, all vertex pairs having at least a certain number of common neighbors (i.e., the similarity of neighbors is no less than the given threshold) are identified, and each pair is connected by an edge [24]. On the other hand, all edges whose endpoints do not satisfy the condition are deleted because they are not considered to be in the same cluster [24]. The graph is changed by repeating this operation.

Data polishing makes it possible to modify input data so that groups of related vertices are extracted without the loss of group structures in the data [24]. In addition, it is possible to extract vertex groups without depending on the number of samples by focusing on the similarity of vertices. This advantage is useful for this study. This study uses smart card data included information about elderly people or children with small sample sizes. Therefore, this study can have potential to extract travel behaviors of these passengers by using data polishing. Also, multiple characteristics can be grasped for one vertex since data polishing is one of soft clustering. Thus, it may be possible to extract and understand travel patterns flexibly.

This study proposes the method improved to apply data polishing to smart card data analysis. This study extracts travel patterns of smart card users from five attributes. Specifically, this study groups “usage vertices” composed of three attributes—day, time, and passenger type—by strength of connection with “Origin and Destination (hereafter, OD) vertices” composed of two attributes¬—origin station and destination station. Then, we extract groups of usage vertices with similar connections to OD vertices. In addition, we clarify origin station and destination station combinations with the largest number of users for

each usage group. Using this process, the study provides an understanding of characteristic travel patterns in terms of what kind of people move from which station to which station, on which days of the week, and at what time of day.

2. Data descriptions and aggregate analysis

2.2 Data descriptions

This study uses smart card data collected in Kagawa Prefecture, Japan (Fig. 1). Kagawa is located in the northeast of Shikoku, one of four main islands of Japan. About 980,000 people live there (as of October 1, 2015). Most of them depend on automobiles for transportation. However, people who cannot drive an automobile like the elderly have to rely on public transportation. The aging rate of population in Kagawa is about 30% and it is also expected to keep increasing. Therefore, it is important to maintain public transportation. In addition, in order to improve transit services, understanding the travel behavior of current users is necessary.

The smart card used in Kagawa is called IruCa. IruCa was introduced for use on Kotoden trains or buses operated by the Takamatsu Kotohira Electric Railroad Company and also on buses operated by other bus companies in the prefecture. As of March, 2016, the number of IruCa issued was 341,706. Of these, 75,169 were commuter passes and 266,546 were non-commuter passes. It is thought that most IruCa users are residents of Kagawa Prefecture because IruCa can only be used in the prefecture. So, we consider this study will be able to help us understand the travel behaviors of Kagawa Prefecture residents from IruCa data. Although IruCa can be used for both trains and buses, this study uses only the smart card data related to Kotoden trains.

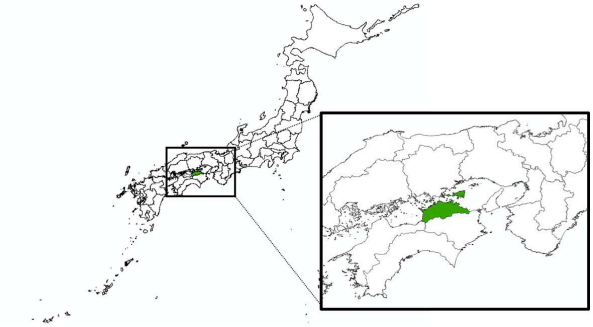

Fig. 1. Kagawa Prefecture

There are 52 stations in total, on three lines: Kotohira line, Nagao line, and Shido line (as shown in Fig. 2). Two stations, Takamatsu-Chikko and Kataharamachi, are connected to both the Kotohira and Nagao lines. And Kawaramachi station is connected to all three lines, so Kotoden passengers can transfer to any line at this station.

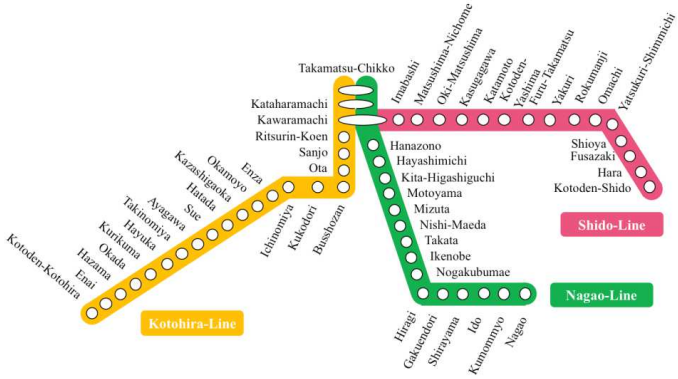

Fig. 2. Kotoden route map

This study uses five attributes related to each trip (day, time, passenger type, origin station, and destination station) that are accumulated in the smart card data as shown in table I. “Passenger type” is explained in detail. Passenger type (1) ~ (5) are non-commuter pass users. On the other hand, (6) ~ (11) are commuter pass users. “Commuter for work” represents a person who use a commuter pass to go to work. “Commuter for school” represents a person who use a commuter pass to go to school

and “Commuter not for school” represents a person who use a commuter pass to travel except for going to school. The smart card data collection period is the 15 months from December 1, 2013 to February 28, 2015. The number of data points collected during this period is 9,033,748. Out of these, this study uses 9,008,709 data points extracted as valid Kotoden trips based on the following criteria: (1) card used by passenger for any train ride during the hours of operation between 5 a.m. and 12 p.m. on any of the three lines, and (2) more than 60 seconds taken to move between stations.

Table I. Five attributes

Attribute	No. of categories	Description
Day	8	(1) Monday, (2) Tuesday, (3) Wednesday, (4) Thursday, (5) Friday, (6) Saturday, (7) Sunday, (8) Public holiday
Time	20	(1) 5:00-5:59, (2) 6:00-6:59, (3) 7:00-7:59, …, (20) 24:00-24:59
Passenger type	11	(1) Adult, (2) Student, (3) Child, (4) Elderly person, (5) Person with disability, (6) Adult commuter for work, (7) Adult commuter for school, (8) Child commuter not for school, (9) Child commuter for school, (10) Disabled commuter for work, (11) Disabled commuter for school
Origin station	52	52 stations
Destination station	52	52 stations

2.2 Aggregate analysis

Figures 3, 4 and 5 show the average daily number of users by day, time, and passenger type, respectively. This helps us understand Kotoden train usage. From the results by day in Fig. 3, it can be seen that the average daily number of users on weekdays from Monday to Friday is almost the same. The average daily number of users on weekdays is 24,794, which is about 2.7 times the average daily number of users on Saturdays, Sundays and public holidays (about 9,026). This is consistent with a large number of weekday trips being taken for commuting and going to school.

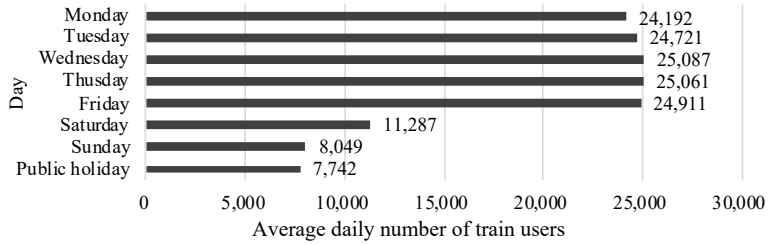

Fig. 3. Average daily number of train users by day

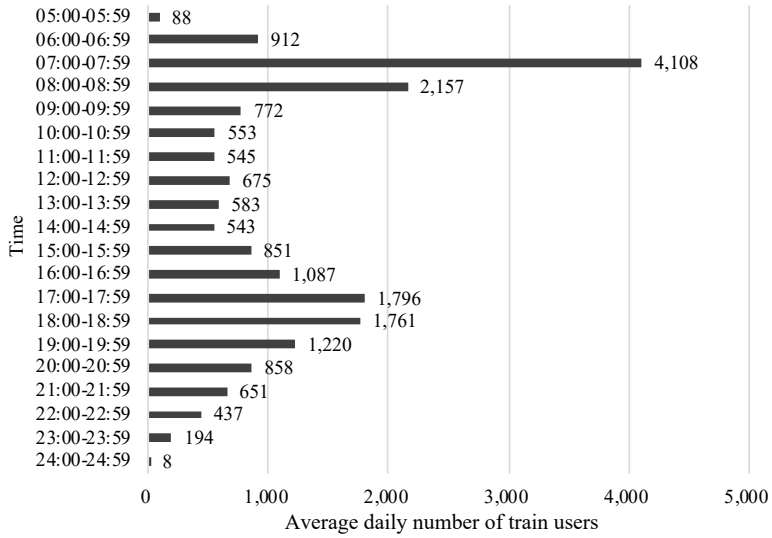

Fig. 4. Average daily number of train users by time

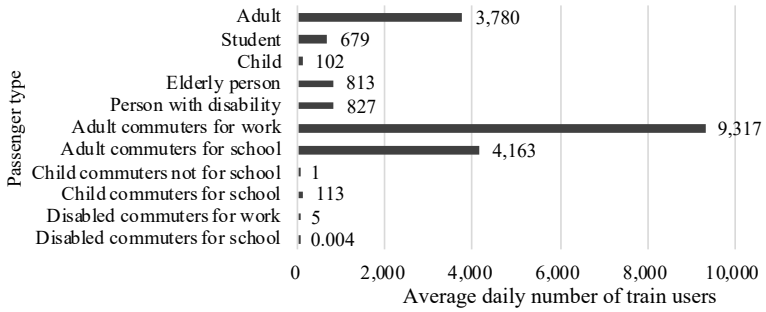

Fig. 5. Average daily number of train users by passenger type

From the results of usage status by time in Fig. 4, it can be seen that the average daily number of users is highest between 7:00 and 7:59. The average daily number of users increases from 5:00 to 7:59, which is understood to be use for commuting to work and to school. The average daily number of users then decreases from 8:00 to 11:59, and thereafter the average daily number of users is almost unchanged until 14:59. After 15:00, the average daily number of users increases again, and is particularly high between 17:00 and 18:59. It is considered that the purpose of use after 17:00 is to return home.

From the results of usage status by passenger type in Fig. 5, it can be seen that the average daily number of users classified as “Adult” is large, and that the average daily number of adult work commuters is as high as 47% (almost half) of the total number of users. On the other hand, it is revealed that the average daily number of child work commuters, and persons with disabilities commuting for work or school are as small as 0.00497%, 0.02696% and 0.00002% respectively.

3. Method

3.1 Preliminary

A graph is consisted with vertex set V and redge set E. All graphs denoted in this paper are undirected graphs.

For any vertex pair containing two vertices u and v, the vertices are called adjacent if the vertex pair is connected by an edge. A vertex u adjacent to v is also called a neighbor of v. The set of neighbors of v is denoted by $N(v)$. A vertex w is a common neighbor of vertices u and v if w is adjacent to both u and v. $N[v]$ denotes $N(v) \cup \{v\}$ and is called the closed neighbor. The number of vertices adjacent to vertex v is denoted by $|N(v)|$.

A vertex set such that every vertex pair is connected by an edge is called a clique C. Although cliques are usually defined by a subgraph, for this study we use this definition, as in Uno et al [24]. A clique included in no other clique is called a maximal clique.

3.2 Extraction of travel patterns by data polishing

This section explains a new methodology applying data polishing as a method for extracting travel patterns from smart card data. In this study, travel patterns are defined by a combination of five attributes: day (8 categories), time (20 categories), passenger type (11 categories), origin station (52 categories) and destination station (52 categories). This study considers these five attributes simultaneously to clarify what kind of people move from which station to which station, on which days of the week, and at what time of day.

Uno et al. [24] focused on the extraction of maximal cliques and proposed a clustering method comprising three procedures: (1) construct the similarity graph, (2) apply data polishing to the similarity graph, and (3) enumerate maximal cliques. This study adds new steps to the proposed clustering method of Uno et al., and proposes a corresponding method for analyzing smart card data. In addition, this study does not focus on maximal cliques alone, but includes all cliques. (See subsection 4) Enumeration of cliques of this section for further detail).

The procedure for extracting travel patterns proposed in this study is comprised of five steps: (1) construct the co-occurrence graph, (2) construct the similarity graph, (3) apply data polishing to the similarity graph, (4) enumerate cliques, and (5) extract the combination of origin station and destination station related to each clique. Each step is explained in order below.

1) Construction of the co-occurrence graph

This study defines the graph constructed from usage vertices and OD vertices having a co-occurrence relationship as the co-occurrence graph G_c. Graph G in Fig. 6 is denoted for constructing the co-occurrence graph G_c. Each vertex and edge in graph G are defined as follows.

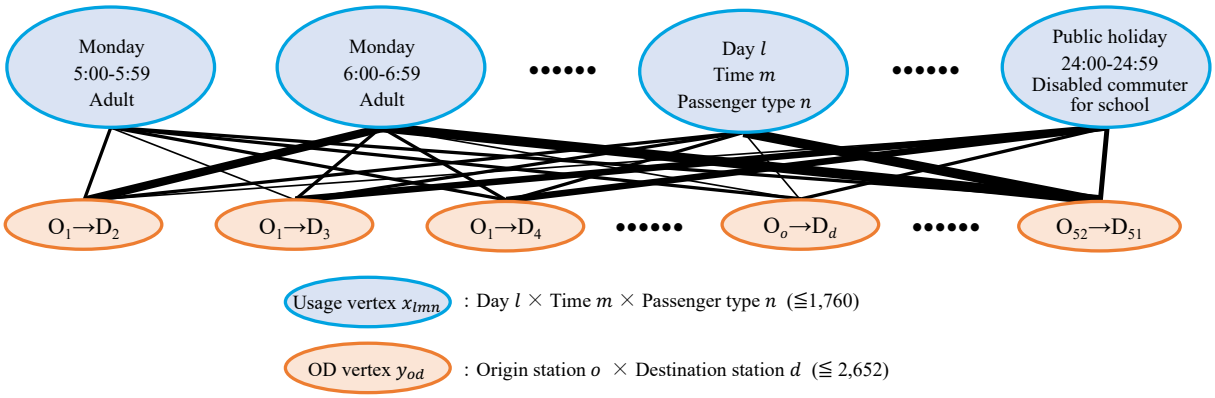

Fig. 6. Graph G representing connection of usage vertices and OD vertices

First, we denote a vertex representing a combination of day $l = 1, \dots, 8$, time $m = 1, \dots, 20$, and passenger type $n = 1, \dots, 11$ by "usage vertex x_{lmn}", and its vertex set is denoted by "usage vertex set $\boldsymbol{X}$". The number of elements in usage vertex set $\boldsymbol{X}$ is 1,760 because it equals the total number of day × time × passenger type combinations. Each usage vertex x_{lmn} has information on the total number of users for each combination of day l, time m and passenger type n. For example, in Fig. 6, the usage vertex ($x_{l=1,m=1,n=1}$) representing the combination of Monday, 5:00-5:59 and Adult has information on the total number of adult users on Mondays at 5:00-5:59. Also, we denote a vertex representing a combination of origin station $o = 1, \dots, 52$ and destination station $d = 1, \dots, 52$ by "OD vertex y_{od}", and its vertex set is denoted by "OD vertex set $\boldsymbol{Y}$". The number of elements in OD vertex set $\boldsymbol{Y}$ is 52×52-52=2,652 because this is the number of all origin station × destination station combinations minus duplicates. Each OD vertex y_{od} has information on the total number of users for each combination of origin station o and destination station d. For example, in Fig. 6, the OD vertex ($y_{o=1,d=2}$) representing O1→D2 has information on the total number of users who move from origin station O1 to destination station D2. Furthermore, each edge connecting each usage vertex x_{lmn} and OD vertex y_{od} has information on the number of users for each combination of day l, time m, passenger type n, origin station o, and destination station d. For example, in Fig. 6, the edge connecting usage vertex ($x_{l=1,m=1,n=1}$) of Monday × 5:00-5:59 × Adults and OD vertex ($y_{o=1,d=2}$) of O1→D2 has information on the number of adult users who move from origin station O1 to destination station D2 on Monday at 5:00-5:59. The number of edges in graph G is 4,667,520 (=1,760×2,652), at maximum.

We construct the co-occurrence graph G_c by extracting combinations having a co-occurrence relationship from all combinations of usage vertices and OD vertices in graph G. In this case, co-occurrence is expressed by the ratio of common users among users included in each usage vertex and each OD vertex. However, a statistical test is performed to determine the significance of co-occurrence because it may occur by chance. This study judges the statistical significance of co-occurrence by a t-test. The t-value used as the test statistic for the t-test is calculated by (1) where W is the total number of users (=9,008,709).

$$t\text{-value} = \frac{\left(|x_{lmn} \cap y_{od}| - \frac{|x_{lmn}| \times |y_{od}|}{W}\right)}{\sqrt{|x_{lmn} \cap y_{od}|}} \tag{1}$$

This study considers that co-occurrence is significant if the absolute value of the t-value is 1.65 or more (significance level 10%). Then, if the combination of usage vertex and OD vertex is a statistically significant co-occurrence relation, the number of users associated with the edge in graph G is replaced with 1, and with 0 otherwise. By this process, the co-occurrence graph G_c is constructed with only the combinations that have significant co-occurrence. In graph G_c, the usage vertex set is denoted by $\boldsymbol{U} = \{u_i | i = 1, \dots, I(I \leq 1{,}760)\}$ and the OD vertex set is denoted by $\boldsymbol{V} = \{v_j | j = 1, \dots, J(J \leq 2{,}652)\}$.

2) Construction of the similarity graph

The similarity graph G_s is constructed on the basis of the co-occurrence graph G_c and expresses the similarity between usage vertices. In this study, we focus on whether the connection of each usage vertex to the OD vertices is same and we construct the similarity graph with the high similarity usage vertices. Although the Simpson coefficient and Dice coefficient, among others, can be considered similarity measures, this study uses the Jaccard coefficient as in Uno et al. [24]. The similarity between any usage vertex u_i and u_i' is calculated by (2).

$$sim(u_i, u_i') = \frac{|N(u_i) \cap N(u_i')|}{|N(u_i) \cup N(u_i')|} \qquad s.t.\ \ u_i, u_i' \in U,\ \ v_j \in V \tag{2}$$

In addition, this study constructs the similarity graph composed of higher similarity usage vertices by setting a threshold value θ_s. We explain how to construct the similarity graph from the co-occurrence graph of Fig. 7 as an example (usage vertex set $\boldsymbol{U} = \{u_i | i = 1, \dots, 6\}$ is a blue circle; OD vertex set $\boldsymbol{V} = \{v_j | j = 1, \dots, 7\}$ is an orange circle). For usage vertices u_1 and u_2, $|N(u_1) \cap N(u_2)| = 1$ and $|N(u_1) \cup N(u_2)| = 4$ since $N(u_1) = \{v_1, v_3\}$ and $N(u_2) = \{v_3, v_4, v_5\}$. Therefore, the similarity between usage vertices u_1 and u_2 is calculated as $1/4 = 0.25$. In this way, the similarity of all usage vertices is calculated and only usage vertex pairs where the similarity equals or exceeds the threshold value are connected by edges. In the example of Fig. 7, when the threshold value θ_s is set at 0.4 we get the similarity graph shown in Fig. 8. In this case, each usage vertex of the co-occurrence graph u_1, u_2, u_3, u_4, u_5, u_6 in Fig. 7 is replaced with s_1, s_2, s_3, s_4, s_5 and s_6 respectively. The similarity graph is constructed by connecting usage vertices in a way similar to how the connections to OD vertices were made, that is, the usage vertices connected by edges represent similar user travel behaviors. For example, the fact that the usage vertices s_1, s_4 and s_5 are connected in Fig. 8 means that the users grouped in these vertices have similar travel behaviors. In this way, the graph representing the similarity relationships between usage vertices is the similarity graph. In the similarity graph G_s, the usage vertex set is denoted by $\boldsymbol{S} = \{s_k | s_k \in \boldsymbol{U}, k = 1, \dots, K(K \leq 1{,}760)\}$.

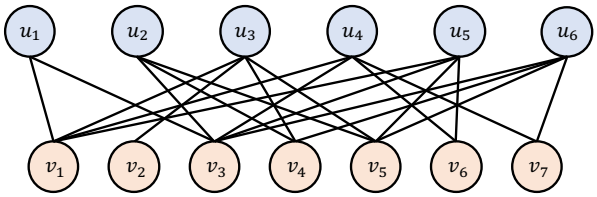

Fig. 7. The co-occurrence graph

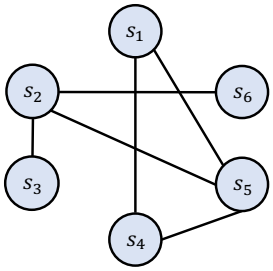

Fig. 8. The similarity graph

3) Application of data polishing to the similarity graph

We next apply data polishing to the similarity graph in order to group usage vertices $s_k \in \boldsymbol{U}$, leaving only the usage vertices with strong connections from the similarity graph. The similarity measure of sets is used to judge whether usage vertex pairs have a strong connection. In this study, the Jaccard coefficient is used as the similarity measure in the same way as for the similarity graph construction. The similarity between any usage vertex s_k and s_k' is calculated by (3).

$$sim(s_k, s_k') = \frac{|N[s_k] \cap N[s_k']|}{|N[s_k] \cup N[s_k']|} \qquad s.t.\ \ s_k, s_k' \in \boldsymbol{S} \tag{3}$$

Equation (3) represents the similarity between the closed neighbors of s_k and s_k'. We explain the data polishing procedure using the similarity graph in Fig. 8. First, the similarity of all usage vertices composing the similarity graph is calculated using (3). The similarity of usage vertex s_1 and s_2 is calculated as an example. $|N[s_1] \cap N[s_2]| = 1$ and $|N[s_1] \cup N[s_2]| = 6$

are found, as the closed neighbor of s_1 is $N[s_1] = \{s_1, s_4, s_5\}$ and the closed neighbor of s_2 is $N[s_2] = \{s_2, s_5, s_6\}$. Therefore, the similarity between usage vertices s_1 and s_2 is calculated as $1/6 = 0.17$. Then, we set the threshold value θ_p, and any vertex pair whose similarity equals or exceeds θ_p is connected by an edge; edges are deleted otherwise. For example, Fig. 8 is replaced with the graph in Fig. 9 if θ_p is set at 0.4. The data polishing is repeated using this newly constructed graph as an input graph and this process is performed until the deformation of the graph converges. Normally, data polishing is applied several times. However, in this example, the shape of the graph does not change from that of Fig. 9 even if data polishing is applied again. Therefore, the data polishing is terminated after only one application. This example requires only one polishing because a simple co-occurrence graph was created in order to simplify the explanation. The final graph constructed by data polishing is called the polishing graph G_p. In G_p, the usage vertex set is denoted by $\boldsymbol{P} = \{p_t | p_t \in \boldsymbol{S}, t = 1, \dots, T(T \leq 1{,}760)\}$.

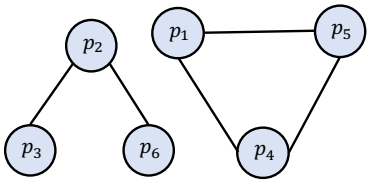

Fig. 9. The polishing graph

4) Enumeration of cliques

This study enumerates all cliques from the polishing graph. As an example, there are three cliques $C_1 = \{p_2, p_3\}$, $C_2 = \{p_2, p_6\}$ and $C_3 = \{p_1, p_4, p_5\}$ in the polishing graph of Fig. 9. In these cliques, there is one maximal clique, $C_3 = \{p_1, p_4, p_5\}$. In this study, we extract all cliques, differeing from the approach of Uno et al. [24] in which they extracted only maximal cliques. Cliques extracted by this proposed method represent groups of usage vertices $p_t \in \boldsymbol{S}$ in which user travel behaviors are similar, as described in 2) of this section. This study allows us to understand users who have similar travel behaviors by considering these groups. If a usage vertex belongs to more than one clique (such as p_2 in Fig. 9), it is suggested that users in this vertex have more than one travel pattern. On the other hand, it is considered that users in maximal cliques have unique travel patterns. Therefore, in this study we consider that various travel patterns of IruCa users can be understood by enumerating all cliques including maximal cliques.

5) Extracting the combination of origin station and destination station related to each clique

For each extracted clique, this study understands that users in that clique frequently move from which origin station to which destination station. As an example, we consider the case of an extracted clique consisting of two usage vertices: $x_{l=1,m=1,n=1}$ (Monday × 5:00-5:59 × Adult) and $x_{l=1,m=2,n=1}$ (Monday × 6:00-6:59 × Adult). First, we extract the OD vertices having co-occurrence with both of these usage vertices on the basis of the co-occurrence graph G_c. Then, we confirm the most frequent OD vertices in co-occurring combinations on the basis of the graph G. By these processes, we clarify what kind of people move from which station to which station, on which days of the week, and at what time of day.

4. Results

4.1 The threshold

In this proposed method, there are two threshold parameters: the threshold value θ_s for constructing the similarity graph G_s and the threshold value θ_p for constructing the polishing graph G_p. As both θ_s and θ_p are criteria for judging similarity, this study considers them on the same value. Therefore, only one parameter is required to extract cliques. Although the threshold value λ $(= \theta_s = \theta_p)$ influences the extraction of cliques, there is no clear criterion for setting it. This study uses the average of the clustering coefficients for determining λ.

The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. It is higher in a graph where vertices adjacent to any other vertex are connected by edges. Here, since all the vertices in the clique are connected by edges, it can be said that the clustering coefficients of the vertices in the polishing graph become higher on average when the clique is generated. Equation (4) shows the calculation for the cluster coefficient for a vertex *i,* where e_i is the number of edges connecting vertices adjacent to *i* and k_i is the number of vertices adjacent to *i*.

$$C_i = \frac{e_i}{k_i(k_i - 1)/2} \tag{4}$$

The average clustering coefficient is the average of the clustering coefficients for all vertices in the graph. It is calculated with (5), where N is the total number of vertices.

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{5}$$

In this study, we focus on the relationship between the average clustering coefficient and the similarity. To construct the cliques, we set the threshold λ to the maximum value of the average clustering coefficient. The procedure for setting λ is as follows. First, we calculate the similarity of all usage vertices in the co-occurrence graph and set λ to 0.01. Next, any edge is deleted if the similarity is less than λ, and the average clustering coefficient is calculated. Usage vertices that are not adjacent to any other usage vertices are removed, λ is raised by 0.01, edges where the similarity is less than λ are removed, and the average clustering coefficient is calculated again. These steps repeat until λ is 1.00.

The average clustering coefficients for each threshold are shown in Fig. 10. Although the decision was made to set the threshold at the maximum of the average clustering coefficients, there are multiple maxima of the average clustering coefficients. Therefore, it is not possible to judge which threshold should be used only by the average clustering coefficients. Accordingly, we also calculate the graph density. The graph density is calculated with (6), where E is the edge set and V is the vertex set.

$$D = \frac{|E|}{|V|(|V| - 1)/2} \tag{6}$$

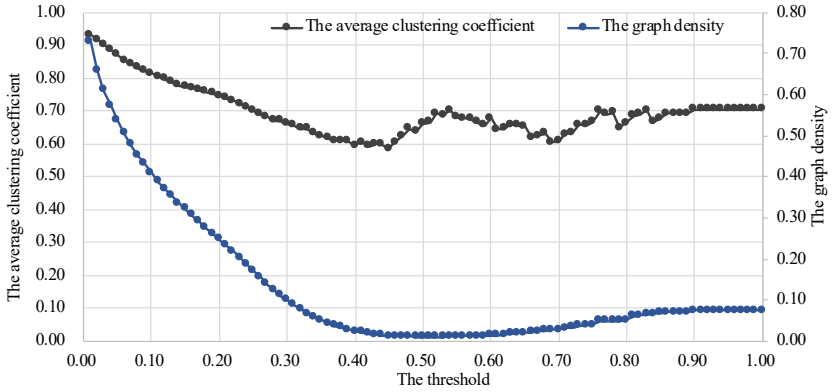

Fig. 10. The average clustering coefficient and the graph density

This value increases as the number of edges gets closer to the maximal number of edges. That is, a dense graph has a high graph density. However, the boundary of the vertex groups (cliques) cannot be judged when the graph is dense. Therefore, this study focuses on the threshold where the graph density is lowest. By using two indices, it is possible to extract cliques such that the boundaries between cliques are clear.

The graph density calculated for each threshold is shown in Fig. 10. In this study, the threshold is set at the point where the average clustering coefficient is at a maximum and the graph density is at a minimum. Thus, from the results of Fig. 10, the threshold is set at 0.54.

4.2 Similarity of usage vertices

As a result of applying the proposed method to the smart card data, the polishing graph as shown in Fig. 11 is constructed. In order to simplify references to each usage vertex, the numbers assigned to them are shown in Fig. 11.

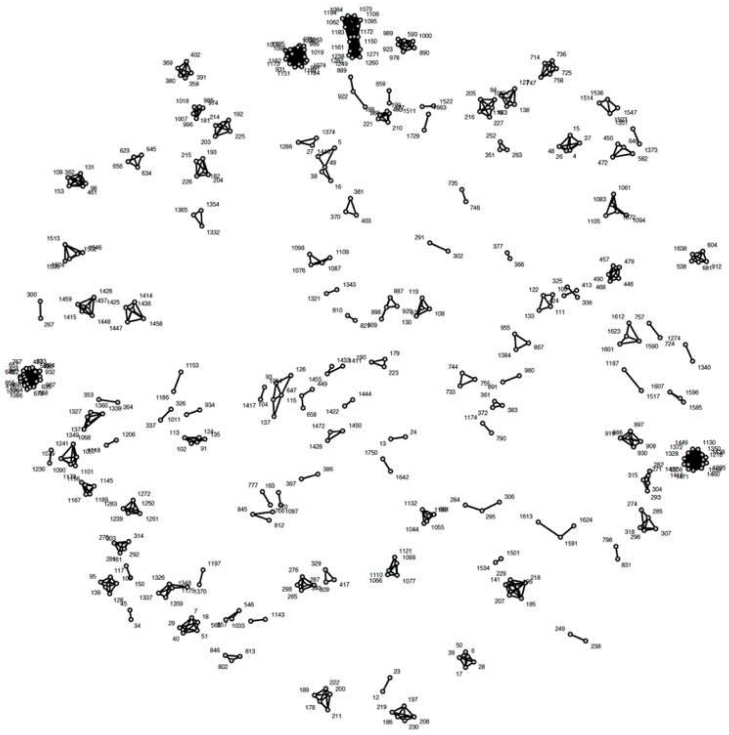

Fig. 11. The polishing graph in this study

From the polishing graph, the total number of cliques extracted is 127. Of these, 52 cliques consist of two usage vertices, 32 of three usage vertices, 5 of four usage vertices, 30 of five usage vertices, 3 of six usage vertices, and 1 each of nine, ten, fourteen, sixteen, and twenty-one usage vertices.

It is not possible to show the composition of all the extracted cliques for want of space. Therefore, this paper focuses on differences in the combinations for day, time, and passenger type. The total number of cliques according to differences in these combinations are shown in tables II to XI, each of which presents the results for the cliques with a given number of usage vertices (as listed in the previous paragraph).

Table II. 52 cliques consisted of two usage vertices

Combination	No. of cliques
Time and passenger type are the same	37
Day and passenger type are the same	2
Only passenger type is the same	10
Day, time, and passenger type are different	3

Table III. 32 cliques consisted of three usage vertices

Combination	No. of cliques
Time and passenger type are the same	28
Day and passenger type are the same	0
Only passenger type is the same	4
Day, time, and passenger type are different	0

Table IV. 5 cliques consisted of four usage vertices

Combination	No. of cliques
Time and passenger type are the same	4
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

Table V. 30 cliques consisted of five usage vertices

Combination	No. of cliques
Time and passenger type are the same	28
Day and passenger type are the same	0
Only passenger type is the same	2
Day, time, and passenger type are different	0

Table VI. 3 cliques consisted of six usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	3
Day, time, and passenger type are different	0

Table VII. 1 clique consisted of nine usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

Table VIII. 1 clique consisted of ten usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

Table IX. 1 clique consisted of fourteen usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

Table X. 1 clique consisted of sixteen usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

Table XI. 1 clique consisted of twenty-one usage vertices

Combination	No. of cliques
Time and passenger type are the same	0
Day and passenger type are the same	0
Only passenger type is the same	1
Day, time, and passenger type are different	0

For example, table III presents the results for cliques consisting of three usage vertices. For the 32 extracted cliques, the number of combinations in which time and passenger type are the same is 28; in the other 4 cliques, only the passenger type is the same.

From tables II to XI, it can be seen that many of the extracted cliques are combinations in which time and passenger type

are the same. Indeed, cliques with this combination make up 76% of the total number of extracted cliques. These cliques represent behaviors of the same type of users at the same time but on different days of the week.

The number of cliques in which only passenger type is the same comprises the second largest proportion of the total (about 19%). For the cliques with six, nine, ten, fourteen, sixteen and twenty-one usage vertices, this is the only combination. This can be understood as behaviors of the same type of users at different times of day on different days of the week.

On the other hand, combinations in which day and passenger type are the same or in which day, passenger type, and time are different, exist only in cliques with two usage vertices. For example, one of the cliques in which day and passenger type are the same is the clique with "Thursday×10:00-10:59×Disabled commuter for work" and "Thursday×12:00-12:59×Disabled commuter for work". This suggests that card users identified as "Disabled commuter for work" have similar travel behaviors between "10:00-10:59" and "12:00-12:59" on "Thursday". The clique with "Saturday×23:00-23:59×Child" and "Public holiday×24:00-24:59×Adult" is a clique in which day, time, and passenger type are different. From this result, it can be said that behaviors of "Child" passengers and "Adult" passengers are similar at midnight on holidays.

Moreover, the extracted cliques are compared by passenger types. For cliques with two, three, four, and five usage vertices, it is found that there are cliques related to various passenger types from Fig. 12. Therefore, it can be said that we can understand several travel patterns of different passenger types from the results of the cliques with small numbers of usage vertices. On the other hand, it is possible to understand travel patterns of specific three passenger types from cliques with large numbers of usage vertices as all cliques with six or more usage vertices relate to adult or child commuter users.

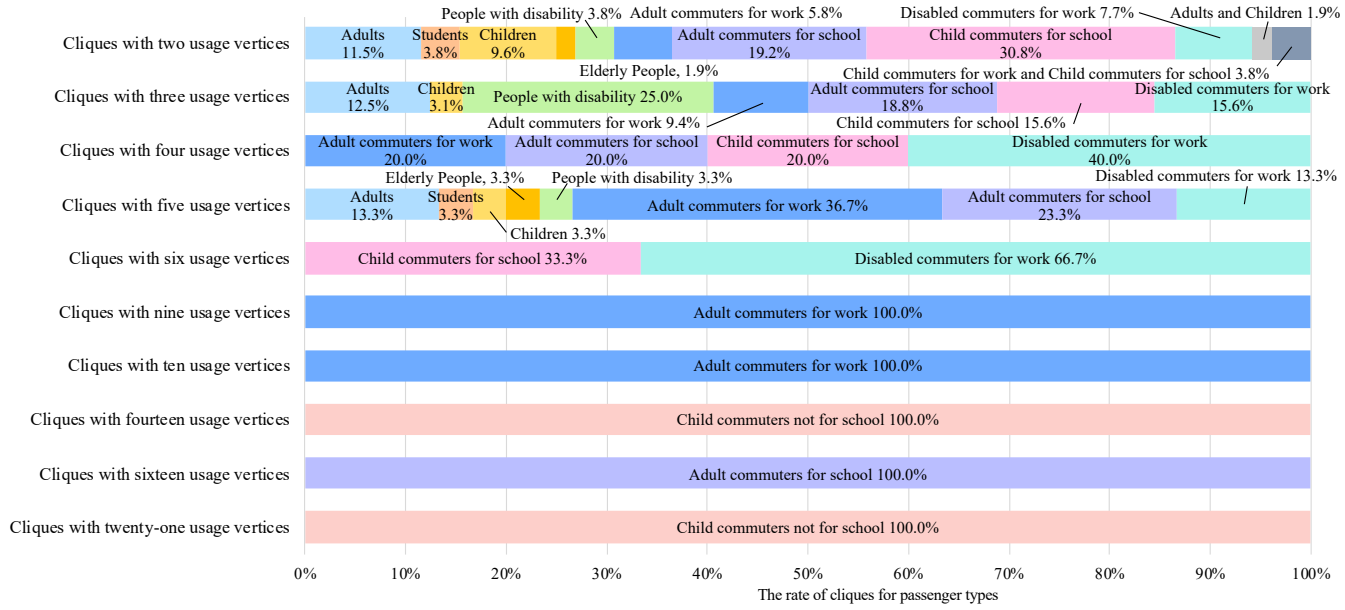

Fig. 12. The rate of cliques for passenger types

4.3 Travel patterns of IruCa users

In this section, we present the characteristic combinations of origin station and destination station for each clique based on the results of the extracted cliques. This corresponds to step (5) of the procedure. Because the number of cliques is large, it is impossible to show the combinations of origin station and destination station for all cliques. Here, we focus on cliques related to children ("Child", "Child commuter not for school", and "Child commuter for school"), elderly people, or people with disabilities ("Person with disability", "Disabled commuter for work", and "Disabled commuter for school"). In the extracted cliques, the number related to children is 35, the number related to elderly people is 2, and the number related to people with disabilities is 28. Although most of these are cliques consisting of the same passenger types, there are 3 cliques that consist of different passenger types. Hereafter, we focus on these latter 3 cliques and clarifies the origin station and destination station combinations with the largest number of users for each clique. The details of the three cliques are as follows;

(C1) "Sunday×6:00-6:59×Child commuter for school" and "Thursday×17:00-17:59×Child commuter not for school".

(C2) “Public holiday×13:00-13:59×Child commuter for school” and “Wednesday×18:00-18:59×Child commuter not for school”.

(C3) “Saturday×23:00-23:59×Child” and “Public holiday×24:00-24:59×Adult”.

For each clique, the origin station and destination station combinations with the largest number of users are shown in table XII. From the table, we can see that the travel behavior of “Child commuter for school” at “6:00-6:59” on “Sunday” and the travel behavior of “Child commuter not for school” at “17:00-17:59” on “Thursday” are similar, and they move from “Ota” to “Shioya”. Moreover, it can be concluded that they move from “Ota” to “Kawaramachi” and then to “Shioya”, based on the Kotoden route map in section II. It is likely that the users in C2 also transfer at Kawaramachi. In C3, we conclude that “Child” at “23:00-23:59” on “Sunday” and “Adult” at “24:00-24:59” on “Public holiday” move from “Kawaramachi” to “Kotoden-Kotohira”. By extracting the OD vertex with the largest number of users for each clique, we can understand what kind of people move from which station to which station, on which days of the week, and at what time of day. In other words, we are able to discover characteristic travel patterns. However, it should be noted that the origin and destination stations are not the actual origins and destinations of the user's trips. Although this study tries to guess the actual place from the land use around the station, they cannot be identified on these cliques. For example, on C1, the actual origin is guessed home since there is a residential area around “Ota”. On the other hand, there are a cultural facility and a beach around “Shioya”. However, it is not clear whether these places are related to children's commuting trips. In the future, we will clarify them.

Table XII. The origin station and destination station combination

Cliques	Origin station	Destination station
(C1) “Sunday×6:00-6:59×Child commuter for school" "Thursday×17:00-17:59×Child commuter not for school”	Ota	Shioya
(C2) “Public holiday×13:00-13:59×Child commuter for school" "Wednesday×18:00-18:59×Child commuter not for school”	Sanjo	Fusazaki
(C3) “Saturday×23:00-23:59×Child" "Public holiday×24:00-24:59×Adult”	Kawaramachi	Kotoden-Kotohira

5. Discussion

From the results of 127 extracted cliques, it was found that many cliques were composed of the same passenger types. On the other hand, it was suggested that the similarity between different passenger types could be quantitatively clarified from cliques in which day, time and passenger type are different.

It may be difficult to predict these usage vertex combinations in advance. For example, in this study, a clique consisting of the “Saturday×23:00-23:59×Child” usage vertex and the “Public holiday×24:00-24:59×Adult” usage vertex was extracted. Basic analysis such as aggregate analysis cannot find this combination. There is also a method to find similar combinations of day and time of use for each passenger type. However, this does not seem to be a realistic method because the number of combinations becomes enormous as the number of categories to be considered increases. Therefore, we conclude that a combination such as “Saturday×23:00-23:59×Child” and “Public holiday× 24:00-24:59×Adult” cannot be extracted without using the proposed method which can consider multiple attributes simultaneously.

This study was also able to provide an understanding of what kind of people move from which station to which station, on which days of the week, and at what time of day. Characteristic travel behaviors of smart card users and origin station and destination station combinations were extracted for each clique.

The results show that the proposed method can extract travel patterns of IruCa users from graphs consisting of day × time × passenger type × origin station × destination station. The travel patterns based on the extracted cliques include behaviors that represent adult commuters and students moving in the morning and returning home in the afternoon. These travel patterns are intuitive results, which support the fact that pattern extraction is performed properly. Moreover, the proposed method can extract travel patterns without depending on the number of samples as about half of all cliques were found to relate to small numbers of smart card users such as children, elderly people and people with disabilities.

When the utilization promotion of the public transportation is examined, it is important not only to raise the utilization

frequency of users who uses it frequently, but also to carry out the measure for raising the utilization frequency of the users who uses it infrequently. However, it is difficult to find their travel patterns in data analysis because the less frequent users have fewer samples. This study shows that data polishing is effective when there are users with different frequency and the distribution of the number of samples is biased.

6. Conclusions

This study proposed a method of extracting travel patterns from smart card data using data polishing. Specifically, we presented a method composed of five steps: (1) construct the co-occurrence graph, (2) construct the similarity graph, (3) apply data polishing to the similarity graph, (4) enumerate cliques, and (5) extract the combination of origin station and destination station related to each clique. We used this method to test the applicability of data polishing to smart card data.

Data from the IruCa smart card, used on the Kotoden rail system in Kagawa Prefecture, Japan, were used in this study. To analyze the data, we constructed a graph representing the relationships between day (8 categories), time (20 categories), passenger type (11 categories), origin station (52 categories) and destination station (52 categories), and applied the proposed method to this graph.

Usage vertices with high similar relationships were grouped by applying data polishing to the similarity graph. The groupings were extracted as cliques, allowing us to understand the similarities between card user behaviors in the extracted cliques and clarifying user groups with similar behaviors. Then, we clarify origin station and destination station combinations with the largest number of users for each usage group. By these processes, this study can understand what kind of people move from which station to which station, on which days of the week, and at what time of day.

In future research, it would be useful to develop an efficient algorithm that eliminates complex calculation procedures and requires fewer processing steps. Also, further study is needed to clarify the occurrence factors and similarity factors on the extracted travel patterns.

Declarations

Availability of data and materials

The data that support the findings of this study are available from the Takamastu-Kotohira Electric Railroad Co. Ltd. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.Data are however available from the authors upon reasonable request and with permission ofthe Takamastu-Kotohira Electric Railroad Co. Ltd.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by JSPS Grants-in-Aid for Scientific(KAKENHI) Grant Numbers 20H02277 and Grant-in-Aid for JSPS Research Fellow Grant Numbers 20J15417.

Authors' contributions

MH**:** Methodology, Software, Formal analysis, Data curation, Writing-Original Draft and Project administration

MK: Conceptualization, Methodology, Writing-Review & Editing and Supervision

TM: Validation, Methodology and Supervision

Acknowledgements

The authors especially thank the Takamastu-Kotohira Electric Railroad Co. Ltd. in Takamatsu City, Kagawa Prefecture, Japan

for providing the smart card data used in this paper.

References

[1] Hofleitner A, Herring R, Bayen A. Arterial travel time forecast with streaming data: a hybrid approach of flow modeling and machine learning. Transport. Res. Part B. 2012;46:1097-1122.

[2] Krause C, Zhang M. L. Short-term travel behavior prediction with GPS, land use, and point of interest data. Transport. Res. Part B. 2019;123:349-361.

[3] Jenelius E, Koutsopoulos H. N. Travel time estimation for urban road networks using low frequency probe vehicle data. Transport. Res. Part B. 2013;53: 64-81.

[4] Espinoza C, Munizaga M, Bustos B, Trepanier M. Assessing the public transport travel behavior consistency from smart card data. Transport. Res. Procedia. 2018;32:44-53.

[5] Morency C, Trepanier M, Agard B. Measuring transit use variability with smart-card data. Transport Policy. 2007;14:193-203.

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

[7] Agard B, Morency C, Trepanier M. Mining public transport user behavior from smart card data. IFAC Proceedings. 2006;39:399-404.

[8] Medina S. A. O, Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. Travel Behaviour and Society. 2018;12:93-101.

[9] Nazem M, Chu A, Spurr T. Analysis of travel pattern changes due to a medium-term disruption on public transit networks using smart card data. Transport. Res. Procedia. 2018;32:585-596.

[10] Li Y. T, Iwamoto T, Schmocker J. D, Nakamura T, Uno N. Analyzing long-term travel behavior: a comparison of smart card data and graphical usage patterns. Transport. Res. Procedia. 2018;32:34-43.

[11] Zhang Y, Martens K, Long Y. Revealing group travel behavior patterns with public transit smart card data. Travel Behaviour and Society. 2018;10:42-52.

[12] Sun L, Axhausen K. W. Understanding urban mobility patterns with a probabilistic tensor factorization framework. Transport. Res. Part B. 2016;91:511-524.

[13] Han Y, Moutarde F. Analysis of large-scale traffic dynamics in an urban transportation network using non-negative tensor factorization. Int. J. Intell. Transport. Syst. Res. 2016;14:36-49.

[14] Vazifehdan M, Moattar M. H, Jalali M. A hybrid bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. J. King Saud Univ. – Comput. Inf. Sci. 2019;31:175-184.

[15] Yao D, Yu C, Jin H, Ding Q. Human mobility synthesis using matrix and tensor factorizations. Inform. Fusion. 2015;23:25-32.

[16] Simsekli U, Virtane T, Cemgil A. T. Non-negative tensor factorization models for Bayesian audio processing. Digit. Signal Process. 2015;47:178-191.

[17] Taneja A, Arora A. Cross domain recommendation using multidimensional tensor factorization. Expert Syst. Appl. 2018;92:304-316.

[18] Wang L, Bai J, Wu J, Jeon G. Hyperspectral image compression based on lapped transform and Tucker decomposition. Signal Process. Image Commun. 2015;36:63-69.

[19] Correa F. E, Oliveira M. D. B, Gama J, Correa P. L. P, Rady J. Analyzing the behavior dynamics of grain price indexes using Tucker tensor decomposition and spatio-temporal trajectories. Comput. Electron. Agr. 2016;120:72-78.

[20] Favier G, Fernandes C. A. R, Almeida A. L. F. D. Nested Tucker tensor decomposition with application to MIMO relay systems using tensor space-time coding (TSTC). Signal Process. 2016;128:318-331.

[21] Briand A. S, Come E, Trepanier M, Oukhellou L. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transport. Res. Part C. 2017;79:274-289.

[22] Ma X, Wu Y. J, Wang Y, Chen F, Liu J. Mining smart card data for transit riders' travel patterns. Transport. Res. Part C. 2013;36:1-12.

[23] Faroqi H, Mesbah M, Kim J, Tavassoli A. A model for measuring activity similarity between public transit passengers using smart card data. Travel Behaviour and Society. 2018;13:11-25.

[24] Uno T, Maegawa H, Nakahara T, Hamuro Y, Yoshinaka R, Tatsuta M. Micro-clustering: finding small clusters in large diversity. 2016;arXiv preprint arXiv:1507.03067v2.

Figures

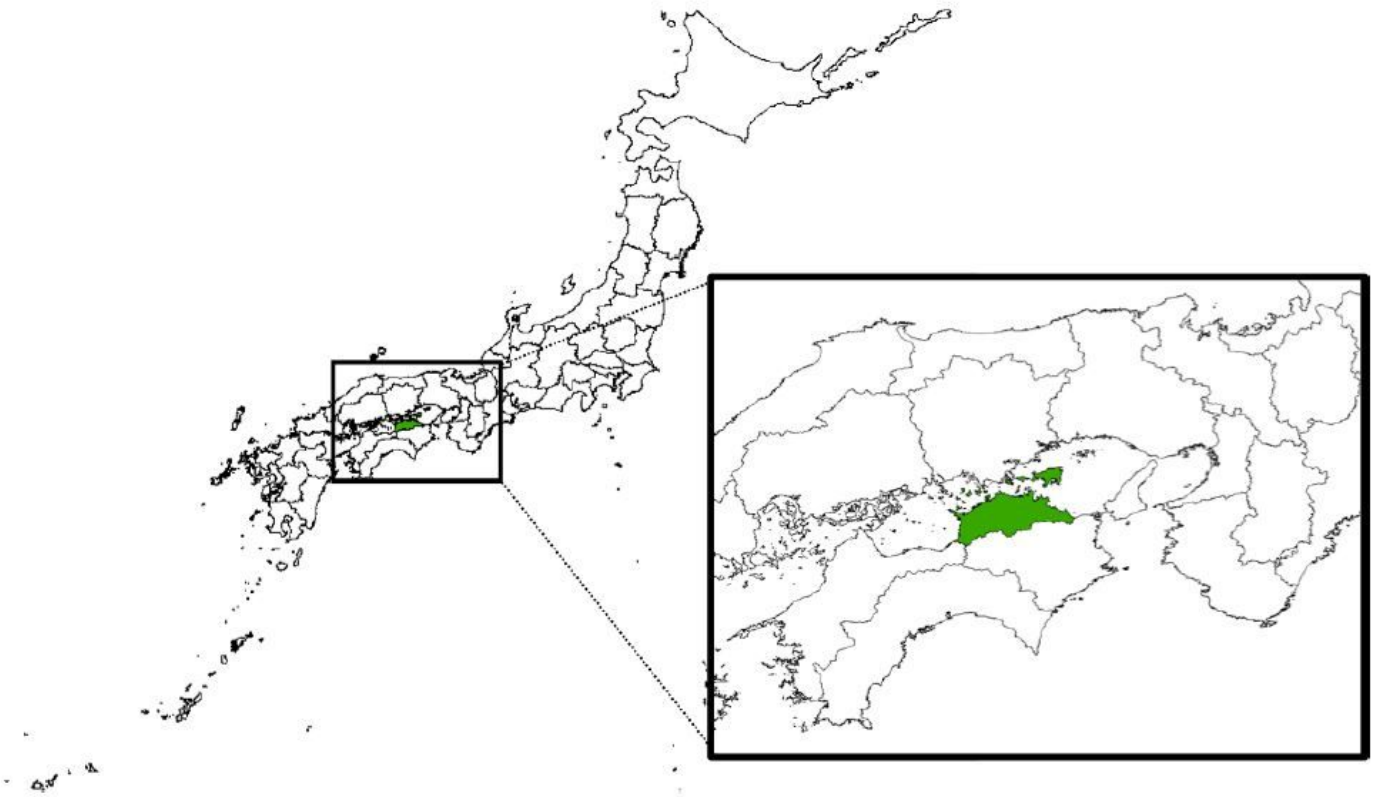

Figure 1

Kagawa Prefecture

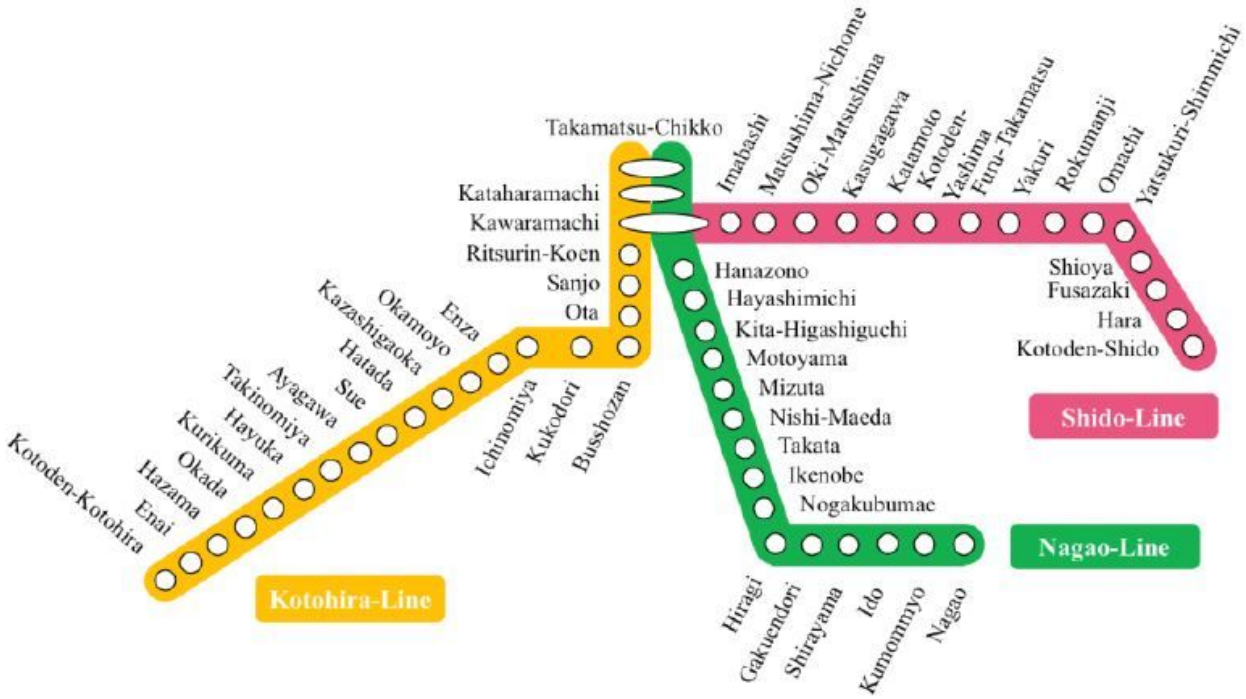

Figure 2

Kotoden route map

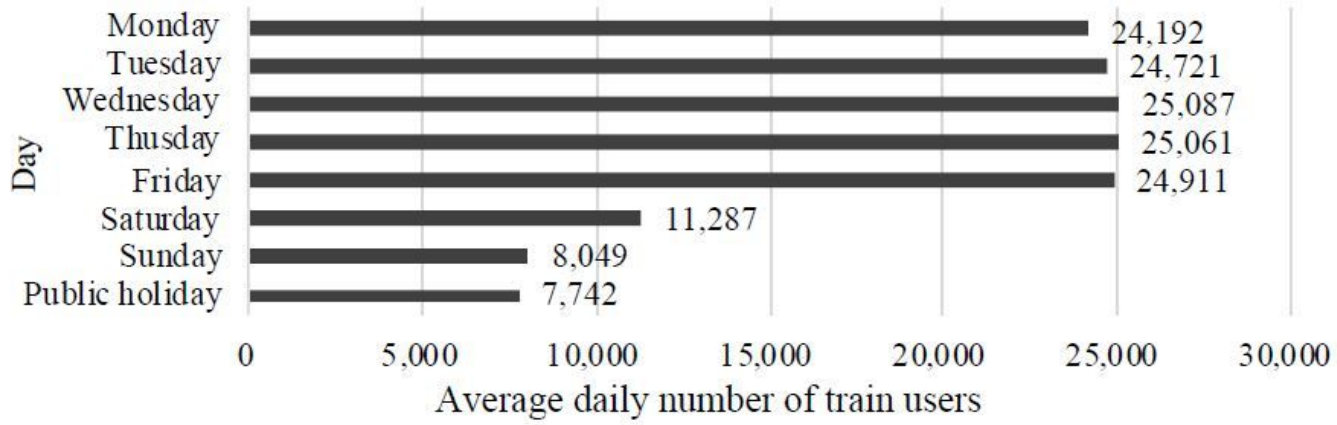

Figure 3

Average daily number of train users by day

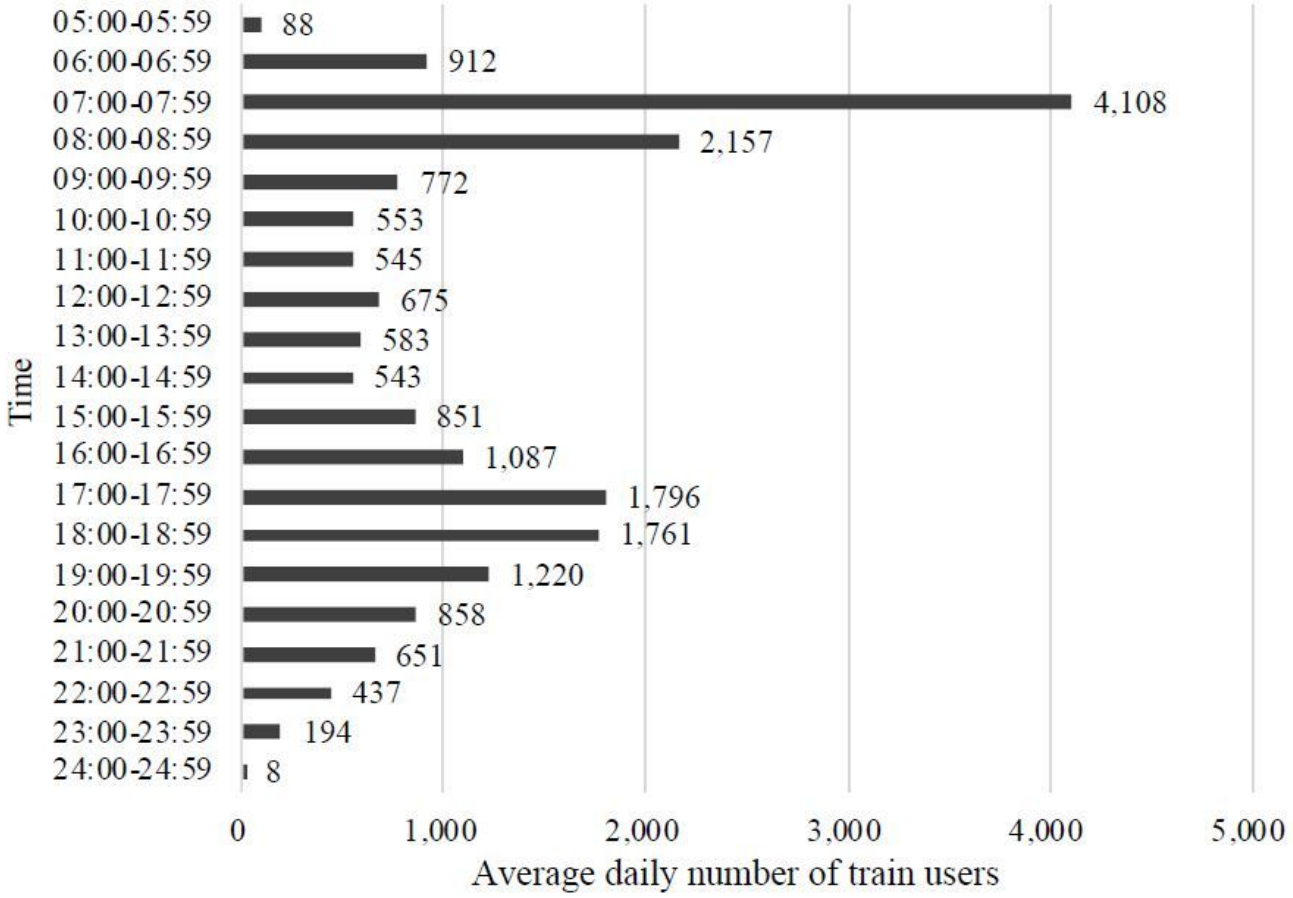

Figure 4

Average daily number of train users by time

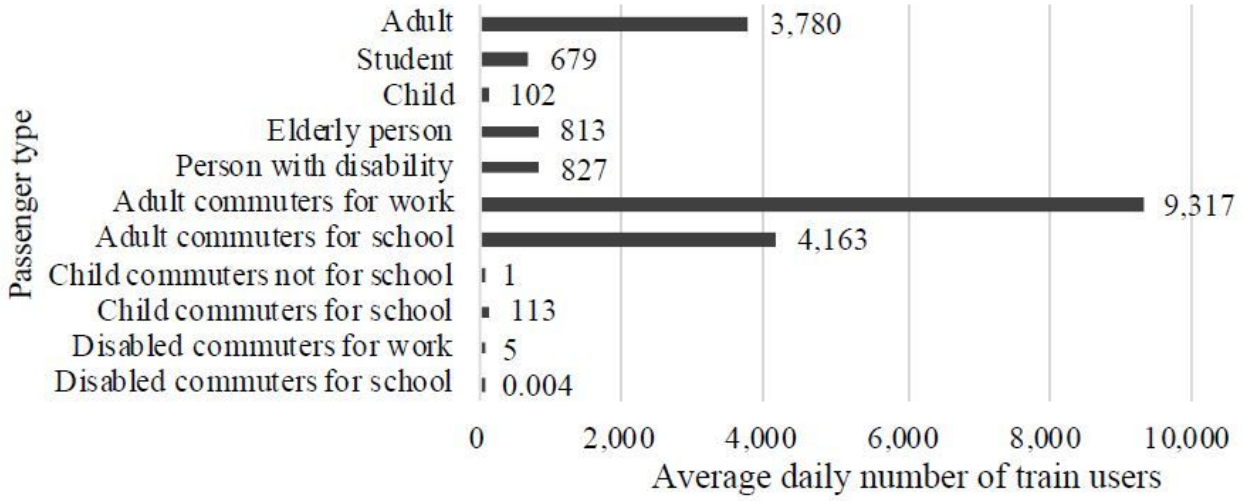

Figure 5

Average daily number of train users by passenger type

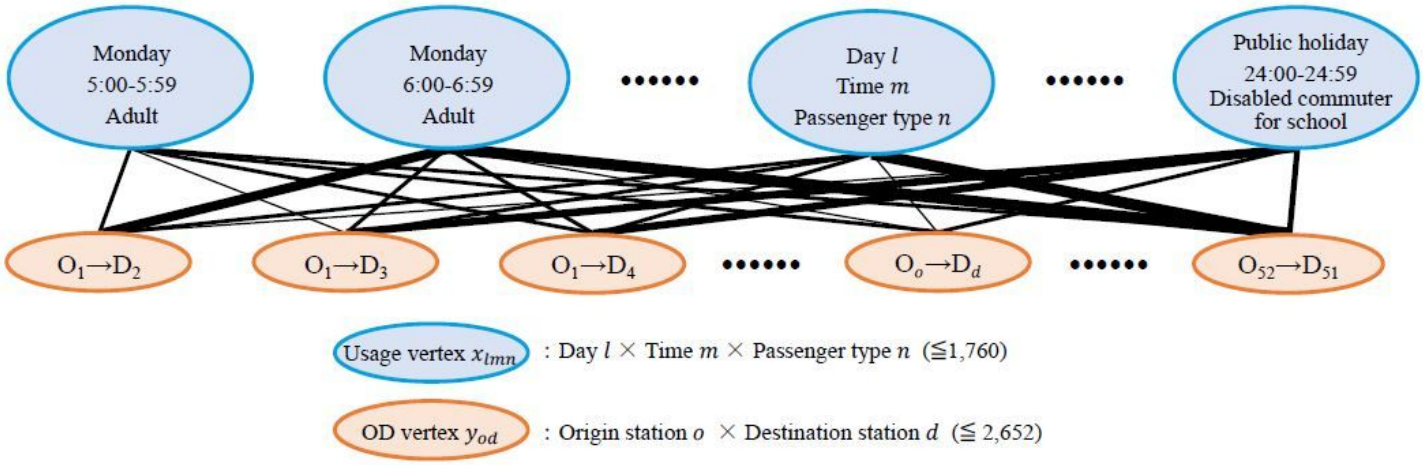

Figure 6

Graph representing connection of usage vertices and OD vertices

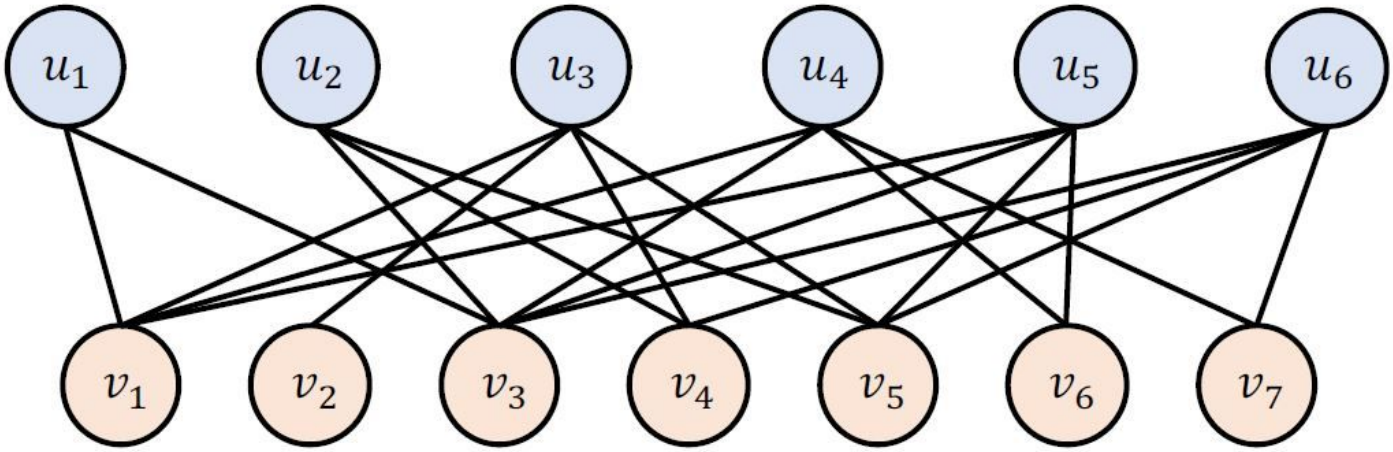

Figure 7

The co-occurrence graph

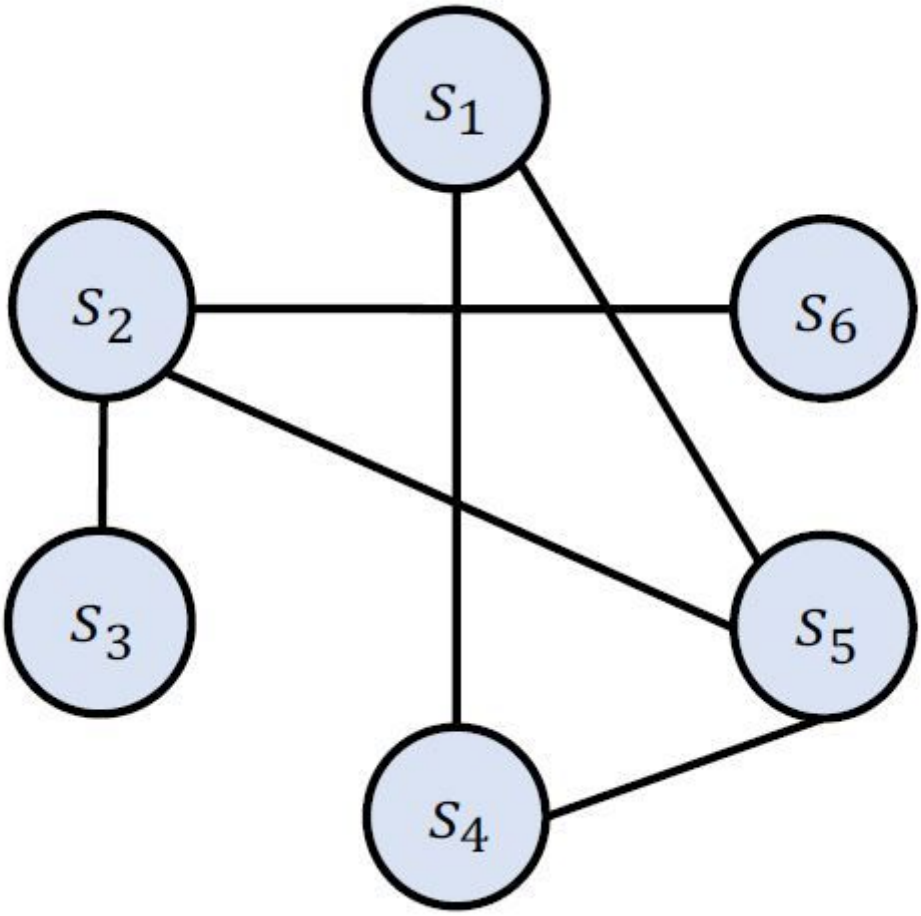

Figure 8

The similarity graph

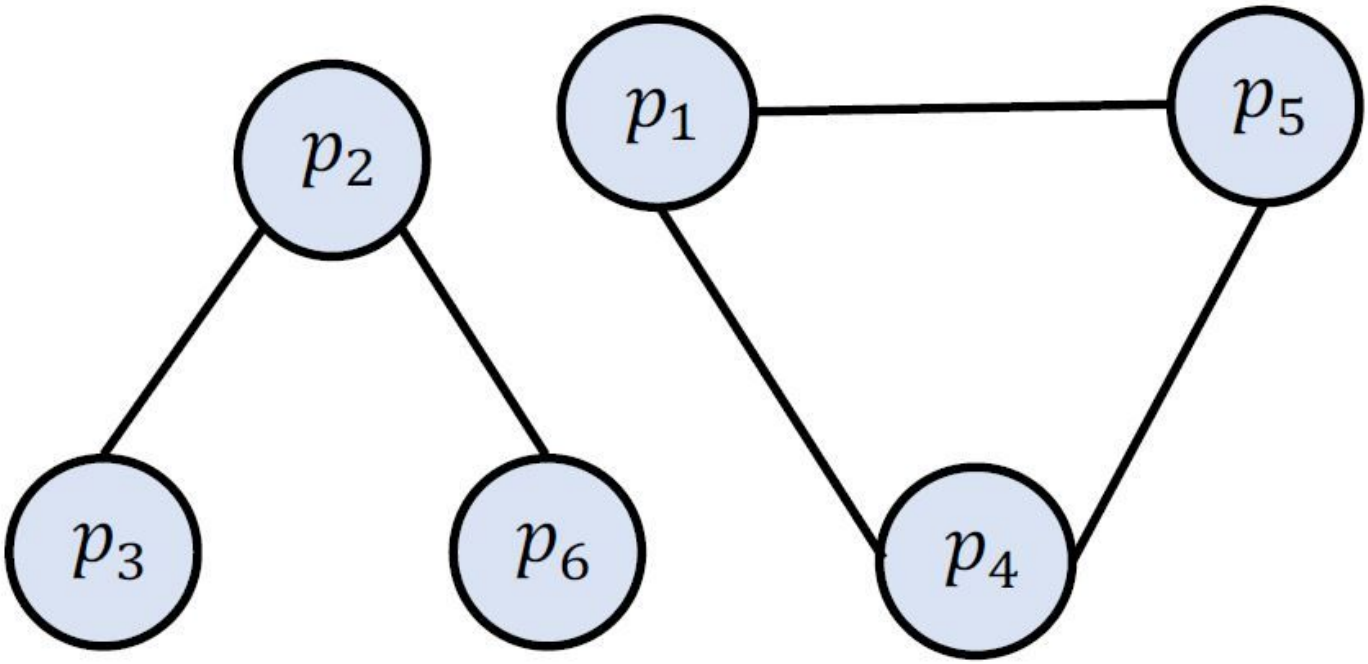

Figure 9

The polishing graph

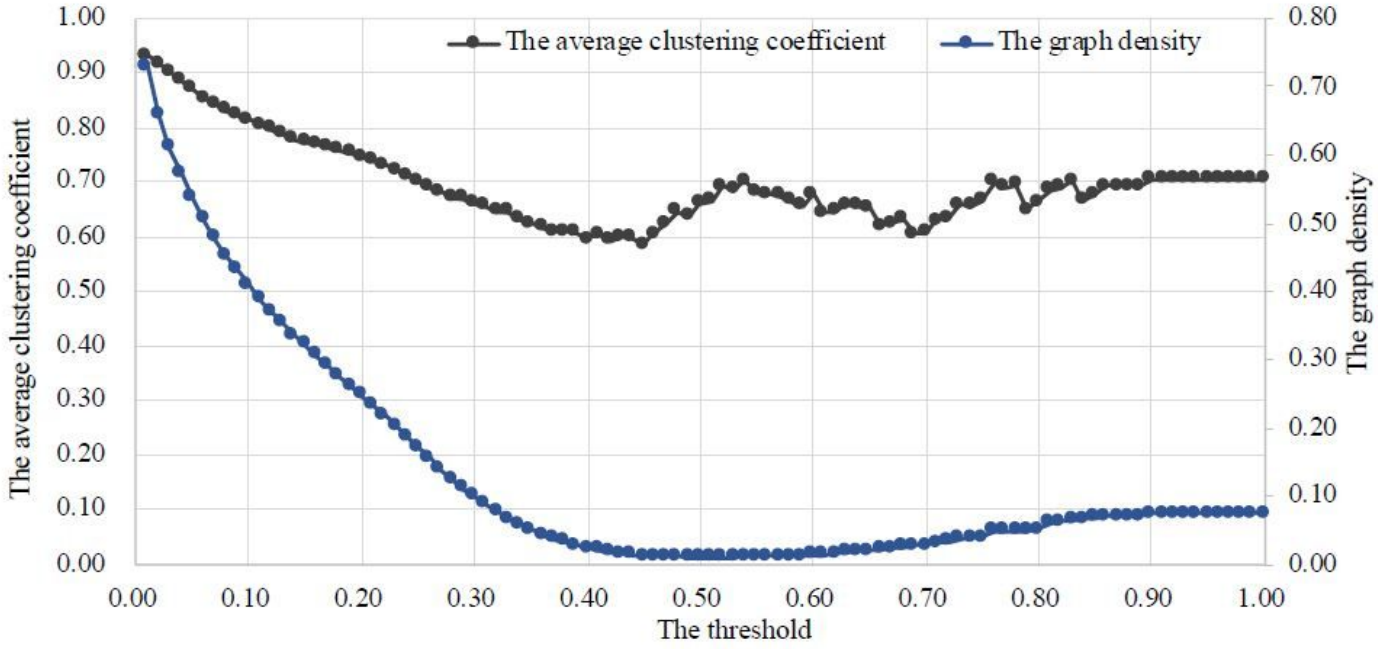

Figure 10

The average clustering coefficient and the graph density

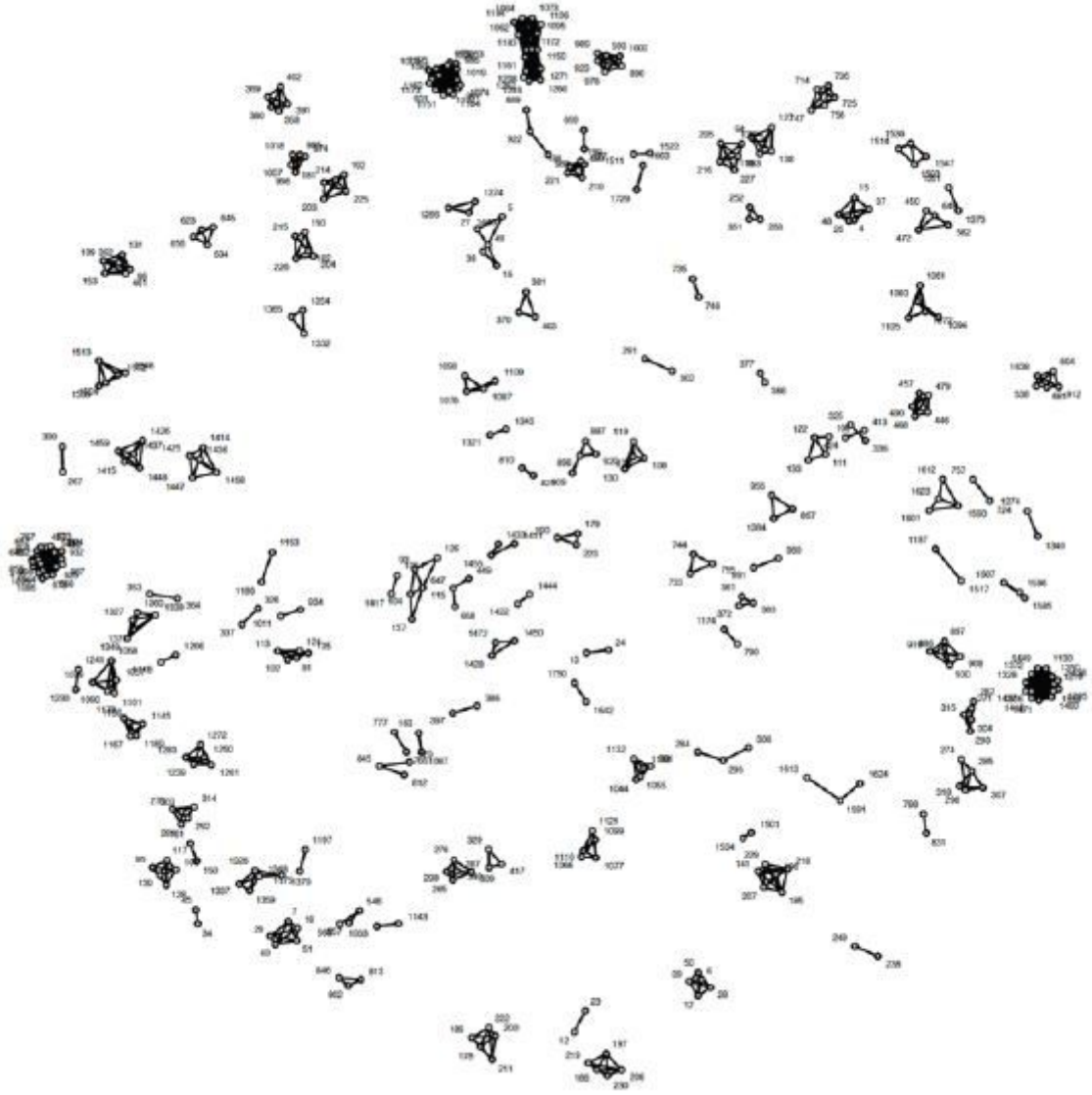

Figure 11

The polishing graph in this study

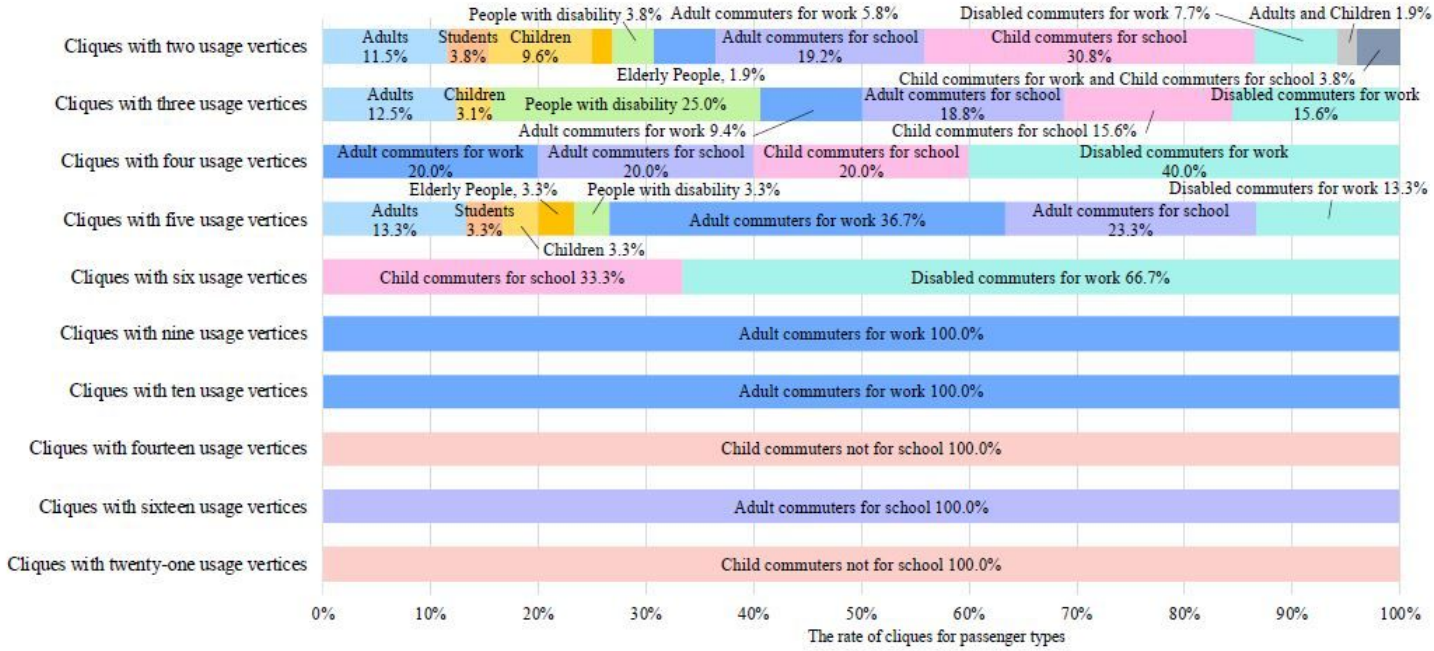

Figure 12

The rate of cliques for passenger types