공학박사학위논문

Estimating of Bus-Trip Destinations Using Temporal Travel Patterns of Smart Card Data

교통카드데이터의 시간적 통행패턴을 활용한 버스 통행 목적지 추정

2019년 8월

서울대학교 대학원

공과대학 건설환경공학부

이 인 묵

Estimating of Bus-Trip Destinations using Temporal Travel Patterns of Smart Card Data

지도교수 고 승 영

이 논문을 공학박사 학위논문으로 제출함

2019년 4월

서울대학교 대학원

공과대학 건설환경공학부

이 인 묵

이인묵의 박사 학위논문을 인준함

2019년 6월

위 원 장 이 정 원 (인)

부위원장 고 승 영 (인)

위 원 차 상 훈 (인)

위 원 김 경 태 (인)

위 원 김 동 규 (인)

Abstract

Estimating of Bus-Trip Destinations Using Temporal Travel Patterns of Smart Card Data

Inmook LEE

Department of Civil and Environmental Engineering

The Graduate School

Seoul National University

Compared to existing transportation data sources (e.g. surveys, travel records, etc) for public transit planning and operation, the smart card data has the following advantages: First, since all trips (transactions) by the cardholder are recorded in the system, extensive data collection is possible. It is in contrast to a small sample by investigation. Second, accurate timestamps and geographic tags (coordinates) are specified for each transaction, so accurate information can be obtained in time and space. Third, since each transaction contains a unique card number (encrypted or anonymized), individual transactions can be tracked, enabling longitudinal studies on long-term travel behavior.

However, it is possible to use the origin information for most of the smart card systems, but the destination information is often omitted. In addition, data on travel purposes are not available. This information is

valuable not only in understanding travel behavior but also in estimating travel demand distribution.

The purpose of this study is to improve the possibility and accuracy of destination (alighting stop) estimation, which is typical missing information of smart card data. In this regard, a number of studies have been conducted to estimate the destination based on the trip chain method. However, there is still a limit to estimates for trip that lacks the reference information required for estimation, such as 'unlinked trip'. To solve this problem, this study used the temporal pattern information extracted from the smart card data from various dates in the past to estimate the destination.

In order to accomplish the purpose of this study, this study proposes a methodology to utilize the historical trip information recorded in the smart card data of the past several days (or more) for the location estimation. In particular, we propose a travel pattern-based alighting location estimation method that generates the travel patterns of public transportation users from the trip records of the past several days and applies the travel patterns to the alighting information estimation.

This study presents an algorithm for estimating the alighting location using historical smart card data of various dates according to the constructed model. For the generation of the first-order clusters, the k-means algorithm suitable for clustering high-dimensional data is applied. For each of the generated clusters, the Gaussian Mixture Model (GMM) was applied. The Expectation-Maximization (EM) algorithm is used as a solution for GMM parameter estimation.

This study validated the validity of the travel pattern model and its suitability for high-capacity smart card data. In particular, by applying the parametric method Gaussian mixed model, the model was developed to suit the classification of travel patterns for high-capacity smart card data.

Historical boarding records were used to generate and utilize temporal pattern information, while stochastic estimates of the destination of individual trip were made by referring to historical boarding location records, which are spatial patterns. Following the method of this study, the probability of estimating the destination of 'unlinked trip' such as 'single trip' was improved.

In addition, a combined algorithm was developed to estimate a destination by combining the conventional trip chain method and the travel pattern method developed in this study. First, after applying the trip chain method, the travel pattern method was applied to trips where the destination could not be estimated by the trip chain method. It was confirmed that the matching rate and the accuracy compared to the conventional trip chain method has been improved.

The developed models and algorithms were analyzed and verified using the smart card data of Daejeon city. In addition, from a practical point of view, the results of categorizing clusters and the sensitivity analysis according to variables and conditions were presented.

Keywords: Smart card data, Automated fare collection data, Destination estimation, OD estimation, Alighting stop, Travel pattern, Passenger clustering, Historical trip record

Student Number: 2009-30943

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Background of this study

The bus system reorganization of Seoul in 2004 was one of the major turning points in the history of public transportation in Korea. First, the bus route system has been completely reorganized. In addition, in exchange for taking the authority to adjust the bus routes, the local government subsidized the profit deficit route, and the installation of the central bus exclusive lane was made to improve the speed of bus travel. It is evaluated as a basis for dramatically improving the public transportation system. As a result of the reorganization of the bus system in Seoul, several large cities are making efforts to improve the public transport system. It is now the central and local government's obligation to efficiently operate public transportation for the convenience of citizens.

Another change due to the reorganization of the bus system in Seoul is the fare system. It was reorganized into a so-called “the Metropolitan Integrated Fare System”, in which a single fare and a distance proportion for each area were mixed to charge a fare based on the total distance traveled by the user regardless of the modes of the bus or the metro. The travel distance calculation is necessary for the implementation of this fare system. It was necessary to get information on boarding as well as alighting. Therefore, the method of using public transportation has been changed so that the smart card is tagged both when boarding and alighting. In addition, the discount rate and transfer discount were provided only to the smart card users, which led to an increase in the smart card usage rate.

As a result, the smart card data included trip information for most public transportation users (passengers).

Studies on the utilization of the smart card data have also been conducted since 2004, when the Seoul Metropolitan Government improved the smart card system in accordance with the reorganization of public transportation mentioned above. As one of the early studies, the Korea Transport Research Institute (KOTI) (2006) verified the availability of the smart card data as a reference material for establishing public transportation policies, and presented the analysis results of public transport status through basic analysis of the smart card data. The Seoul Institute (SI) (2007a) estimates station OD information from the smart card data, summarizes the error type of data and suggests correction method of information using travel history information. The SI (2007b) proposed a basic public transportation service evaluation model using the smart card data and mentioned the possibility of expansion of analysis through linkage with geographic information in the future. Also, they proposed a model of metro travel behavior for estimation of metro users' route choice information, which is one of the limits of the smart card data. The SI (2009) used the smart card data to calculate values for service indicators such as number of service, number of people, and congestion.

Until 2010, researches on basic utilization methods such as analysis of current situation and calculation of statistic amount which can assist policy establishment as the traffic data of public transportation data were conducted mainly. Since then, more research has been carried out on how to use the smart card data for a specific purpose or analysis results. The Gyeonggi Research Institute (GRI) (2011) applied the concept of trip chain, not the simple trip, to calculate the final destination-based travel pattern from the origin place. Using the calculated travel pattern information, they proposed a

method to classify transfer stops. The SI (2011) improved the estimation of the internal route information of the metro by using the dynamic optimal route search algorithm that associates the smart card data with the metro train schedule information, and estimated the congestion of trains and transfer stations using the estimated route information. The SI (2012) compared the level of public transportation supply with the socioeconomic indicators (population, number of employees) and the smart card data.

The government recognized the value of the smart card data and revised the “Act on the Support and Promotion of Utilization of Mass Transit System” in December 2015. The amended law defines the smart card data as one of the transportation data and authorizes the government to collect data from each smart card system operator by standardizing the nationwide smart card data and provides it to public institutions such as local governments and research institutes to enable utilization. They also built a computerized system to support these laws.

The use of smart cards is universal in the world. Bagchi and White (2005) have identified the utility of the smart card data in public transit behavior analysis. Research on the generation of information using smart cards in the field of transportation is being carried out because it is possible to grasp the individual travel of the public transportation users. The most important of these is replacing or supplementing cost-consuming house hold survey transport reports with the smart card data.

In particular, at the level of public transportation operation, the OD matrix of the stop-to-stop is more effective than the OD matrix calculated by the traffic zone. Since the smart card data is generated based on the stop-to-stop basis, it is possible to derive a finer OD matrix than the existing OD matrix based on the survey. However, since the smart card data is generated for the purpose of charging or payment of the public

transportation fare, there is often missing information about alighting the bus other than the boarding information except for special cases (fare charging plan, passenger management, etc). Table 1-1 shows the alighting-tag ratio of bus trips in major cities in Korea. The alighting-tag ratio means the number of trips with the alighting-tag information against the total number of bus trips. In the case of the metropolitan area where the “Metropolitan Integrated Fare System” is applied, most bus trips (97.5%) include alighting information, depending on the characteristics of the fare system. However, cities outside of the Seoul metropolitan area have alighting-tag ratios of around 30% excluding Daejeon city. Even Daejeon city is about 52%. Most of the small and medium cities outside the metropolitan city also do not exceed the rate of 30% alighting-tag ratio.

Table 1- 1 Alighting-tag ratios of bus trips in major cities in Korea (based on the smart card data for 28 May 2018)

City	Total Bus Trip Data (A)	Bus Trip Data with alighting info. (B)	Alighting-tag Ratio (B/A)
Seoul	5,860,119	5,684,969	97.0%
Gyeonggi (Province)	3,391,945	3,344,978	98.6%
Incheon	854,314	821,067	96.1%
Busan	1,535,012	464,151	30.2%
Daegu	754,835	235,082	31.1%
Daejeon	474,432	248,292	52.3%
Gwangju	415,257	121,067	29.2%
Ulsan	334,388	96,566	28.9%

In the estimation of the OD matrix based on the smart card data, estimating the destination (alighting stop) is the most important issue, and

studies mainly applying the trip chain method have been carried out. Especially, since the amount and items of data included in the smart card data are insufficient in many cities, researches have been carried out focusing on the theory related to complementing or correcting information such as the alighting information.

Barry et al. (2002) proposed a deterministic method for estimating the destination by the trip chain. After that, Trépanier et al. (2007) improved the methodology of the trip chain by reflecting the concept of maximum walking distance, Munizaga and Palma (2012) improved estimation accuracy by estimating the generalization time of each trip. After that, Alsger et al. (2016) improved the estimation accuracy by linking with the bus schedule and measured the change of estimation accuracy according to the allowable walking distance and the transfer time. Kim and Lee (2017) applied the generalization distance concept which is easier to apply than generalization time and proposed an optimal allowable walking distance by sensitivity analysis. These are covered in more detail in Chapter 2.

1.2. Objectives of research

Compared to existing transportation data sources (e.g. household survey, travel records, etc.) for public transportation planning and operation, the smart card data has the following advantages: First, since all trips (transactions) by the cardholder are recorded in the system, extensive data collection is possible. It is in contrast to a small sample by investigation. Second, accurate timestamps and geographic tags (coordinates) are specified for each transaction, so accurate information can be obtained in terms of

time and space. Third, since each transaction contains a unique card number (encrypted or anonymized), individual transactions can be tracked, enabling longitudinal studies on long-term travel behavior.

However, it is possible to use the origin information for most of the smart card systems, but the destination information is often missing. In addition, data on travel purposes are not available. This information is valuable not only in understanding travel behavior but also in estimating travel demand distribution.

The main purpose of the smart card data is to facilitate the billing and management, but the large-scale continuous smart card data is used for passenger travel pattern analysis, travel behavior analysis and public transportation system evaluation by researchers, operators and planners. Especially, in the study of smart card data, information about boarding and alighting is very important. The alighting (destination) information is essential data of the route-specific OD matrix and the entire transit service's OD matrix. The OD matrix is essential to effectively implement the public transportation operation plan such as route planning, operation schedule optimization, and in-vehicle congestion analysis.

However, in most city's smart card system, destination information is not available. As a solution to this problem, existing researches have estimated the destination by using the trip chain method, which is an intuitive and efficient method. However, there are limitations to the estimation of trips that still do not constitute chains. As a result of analyzing the smart card data from April to May 2018 in Daejeon city, about 30% of the trips per day was used only once on one date. In the case of existing trip chain method, there is a limitation that the trip is excluded from the object of estimation in the case of a so-called “unlinked trip” in which the trip chain is not linked.

The purpose of this study is to develop a methodology for estimating the alighting location of unlinked trips. It is intended to improve the estimation rate and accuracy of alighting the public transport. In order to achieve the purpose, we utilize historical smart card data and historical travel pattern information.

1.3. Research overview

The purpose of this study is to improve the matching rate and the accuracy of estimating destinations (alighting stops) of bus trips, which is typical missing information of the smart card data. Because metro includes gate-in and gate-out information, metro is excluded from the estimation object. Specifically, the target is to estimate the alighting stop location for each transaction of the smart card data. Although the bus trips except for the metro was set as an object of estimation, the boarding record of the metro was used as the reference information of the alighting.

In order to accomplish the purpose of this study, this study proposes a methodology to utilize the historical travel information recorded in the smart card data of the past several weeks (or more) for estimating the destination. In particular, we propose a travel pattern-based destination estimation method that generates the travel patterns of public transportation users from the travel records of the past several days and applies the travel patterns to the destination estimation.

In Chapter 2, we review the existing research on the destination estimation and the travel pattern analysis using the smart card data, and then draw out the implications and define the directions and problems of

this study.

Chapter 3 summarizes the specifications and descriptions of Daejeon city's smart card data used in this study. The setting of test data and verification data is also described.

In Chapter 4, we propose a model of travel pattern generation and destination estimation using smart card data. The travel pattern generation is divided into a process of generating a travel pattern cluster from the accumulated travel information obtained from various dates and a process of estimating travel pattern parameters of each travel pattern cluster.

Chapter 5 presents an algorithm for estimating the destination using historical smart card data from many dates according to the constructed model. For the generation of the first-order clusters, the k-means algorithm suitable for clustering high-dimensional data is applied. For each of the generated clusters, Gaussian Mixture Model (GMM) was applied.

Chapter 6 describes the result of case study. The performance of the algorithm is evaluated, and the performance is compared with the existing alighting estimation model. In addition, we proposed a method to combine the travel pattern method and the trip chain method.

For insight into the characteristics of the model, Chapter 7 presents the results of analysis by various application methods and conditions. Analysis of categorization of clusters, sensitivity analysis and comparisons with other urban cases are discussed.

Chapter 8 summarizes the contents of this study and describes limitations, additional applications, and future research. The research process is shown in Figure 1-1.

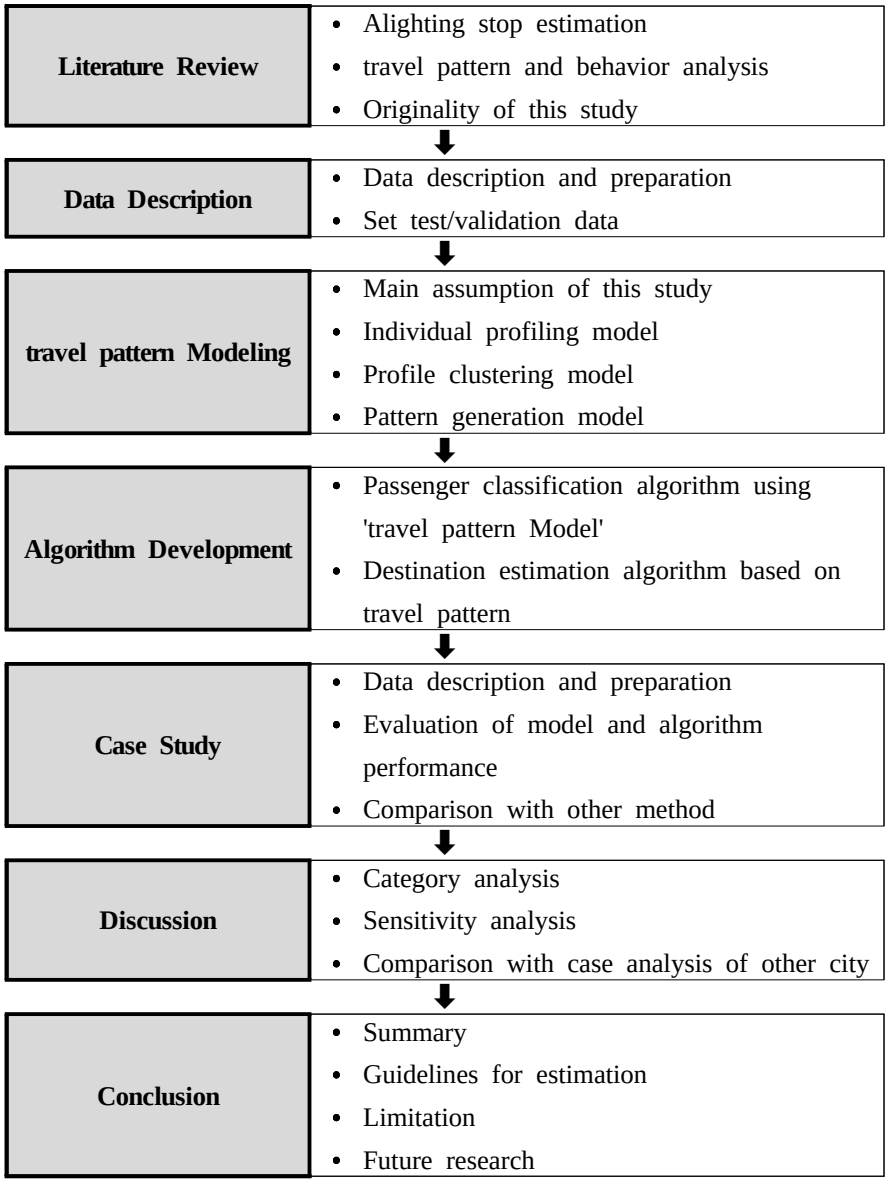


Figure 1- 1 Research Procedure

Chapter 2. Literature Review

In this chapter, we review the methodology of the studies for estimating the destination of smart card data and analysis the main assumptions and features of each study. In addition, from the viewpoint of improving the performance of the destination estimation, the existing study related to the analysis of the travel pattern of smart card data users is reviewed, and the implications and research directions are derived.

2.1. OD estimation using smart card data

2.1.1. Trip chaining model

The early state of the 'trip chain method (or model)', established by Barry et al. (2002), is to structure an individual passenger's trip chain by connecting all sequential trips within a day or a certain time period. There are two basic assumptions on how to estimate the veiled alighting stop: 1) most passengers tend to begin their next trip where their previous trips end, and 2) the very last stop for most passengers is where they began their first trip (home-based trip). These assumptions were built upon a statistical analysis of a travel diary survey in New York, which found 90% of samples adhered to these two assumptions.

With Barry et al. (2002)'s assumptions, Lianfu et al. (2007) estimated alighting stops with a dataset for peak hours, and Farzin et al. (2009) focused on a verification procedure for household survey data. In addition,

Zhao et al. (2007) and Trépanier et al. (2007) extended the algorithm by introducing the concept of the “allowable walking distance” to the next boarding stop. Herein, “maximum allowable walking distance” can be defined as the longest walkable distance between the previous alighting stop and the following boarding stop. Zhao et al. (2007), however, did not investigate how many meters are appropriate for the maximum allowable walking distance, but assumed a 400-meter maximum allowable walking distance for metro smart card data. The alighting stop of the corresponding metro trip is then assigned as the closest stop to the boarding stop of the following bus trip if passengers transfer from the metro to a bus. In Zhao et al. (2007), the estimation success rate with an advanced algorithm, which calculates the ratio of estimated samples to the whole sample and does not include matching accuracy, is 71.2%. Trépanier et al. (2007) compensated for incomplete trip chains with missing links (legs) referring to other days' trip chains. With this assumption, 66% of alighting stops from a sample with trip chains were successfully estimated. Instead of the maximum walking distance, Nassir et al. (2011) introduced a 90-minute maximum transfer time from corresponding boarding time to the next boarding, and the minimum activity duration was assumed to be 30 minutes. With the assumption of Barry et al. (2002), the sample, which consisted of only trip chains, was subject to a validation process, and 92.5% of all alighting stops could be estimated. Based on the results of Zhao et al. (2007), Wang et al. (2011) estimated 68% of alighting stops for bus passengers and validated the results with passengers from five different bus routes. As a result, the total summation of the OD matrix was underestimated due to a sample with unfound alighting stops.

In contrast, Munizaga and Palma (2012) introduced the searching method of minimizing passenger's generalized travel time, which is an

alternative to the “allowable walking distance”. Considering the allowable walking distance as the only factor may lead to estimating unreasonable alighting stops. To overcome such an obstacle, the alighting stop can be determined by minimizing the total transfer cost, including in-vehicle and walking time after the candidates for the alighting stops are selected. With this enhanced algorithm, the estimation rates reached up to 80% of the sample size of 36 million records, and 90% of the matching accuracy (with 300,000 transactions for the trip chain sample) was validated by Munizaga et al. (2014). Alsger et al. (2015) and Nunes et al. (2015) attempted to validate the accuracy with complete OD data from Brisbane (Australia) and Porto (Portugal), respectively. Due to the lack of transfer information in smart card records, a validation process has not been fully conducted.

Alsger et al. (2016) successfully conducted a validation process with the existing algorithm based on the trip chaining principle using a complete dataset, including the exact alighting location and time. To improve the search algorithm, the expected arrival time of passengers was compared to the next departure time of the transit at the target stop. In particular, the same approach was applied to the last trip to improve the estimation. As a result of this improvement, the estimation error (or the gap between the actual and estimated stop) was reduced by 530 meters from the original 806 meters.

Previous research has limited the verification of actual estimation accuracy by verifying the results of a large amount of smart card data as a small amount of survey-based data. Therefore, only the estimation success rate (matching rate) or small accuracy verification by some household survey results are presented. For this same reason, “allowable walking distance”, which is one of the key assumptions in the trip chain method, cannot be readily validated by an accuracy test. Recently, as some of the smart card

data with alighting information becomes available, studies are under way to improve or verify the core assumptions of algorithms such as allowable walking distance.

He at al. (2015) measured the decrease in accuracy and the decrease in the number of opposing samples measured with increasing “tolerance distance” (the same concept as the allowable walking distance). Intuitively, 1,000m was suggested as the most appropriate tolerance distance according to the tendency shown in the results. Alsger et al. (2016) experimentally demonstrated the effect of accuracy on the adjustment of allowable walking distance and transfer time (time boundary for determining transfer trip). Through comparison with actual OD, allowable walking distance of 800 meters or more does not affect algorithm performance. However, the allowable walking distance change was still not covered by the detailed distance, and there was a limit to the accuracy of the change in accuracy.

Due to the well-known problems of trip chain-based algorithms, estimating alighting stops of unlinked trips has essentially been excluded from the onset of arranging input data. To tackle this issue, researchers have attempted to expand the sample size to multiple days, but this is not able to solve the entire problem of estimating unlinked trips. For example, Barry et al. (2009) assumed passengers would share the same travel behavior and OD if they also shared the same stops regardless of unlinked trips and trip chains. A validation process was conducted with records from only 2 routes, which is not sufficient to prove the given assumption.

2.1.2. Probability model and deep learning model

Dou et al. (2007) applied a probabilistic model to estimate the OD

matrix. The main idea of the study is to calculate the probability of alighting at the stop considering the travel distance and the number of passengers. More importantly, this paper found interesting regularity that passengers' travel distance follows the Poisson distribution.

This model has been cited by Zhou et al. (2012) and Yang et al. (2015). Zhang et al. (2014) improved Dou et al. (2007)'s model by adding historical transfer capacity and land use around the station as an element. The transfer capacity of the station was calculated by the number of bus routes. Land use around the station was calculated using the number of passengers.

He and Trépanier (2015) estimate the alighting location according to the priority of the four criteria. Trip sequence, in turn, refers to the next trip's boarding stop if the next trip's boarding stop is within a certain distance from any stop on the previous stop. Last trip of day, in the case of the last trip of the date, it is regarded as a return trip and refer to the boarding stop of the first trip of the date. First trip of next day, if not estimated by the previous two assumptions, refers to the first boarding stop of the next day. The assumptions from the first to the third are the same as those of the previous studies. The fourth assumption was to estimate the alighting location of the 'unlinked trip'. Unlinked trip, if not estimated through the previous three assumptions, the temporal and spatial alighting probability from the historical trip records was calculated. A stop with the highest probability of alighting is estimated as the destination. The temporal and spatial kernel density functions were constructed and the alighting probability of each stop was estimated using the product of the spatial and temporal probabilities of historical departure records related to the estimated stops.

Jung and Sohn (2017) developed a deep learning model to estimate the destination using the smart card data and the land use data. This study

constructed a model architecture with four layers including one input layer, one output layer, and two hidden layers. The total nodes in the input layer are up to 135, including transaction variables (e.g. boarding time, number of transfers, network travel time, generalized travel time) and land use variables (e.g. residential area, commercial area etc).

2.2. Travel pattern analysis using smart card data

Devillaine et al. (2012) developed a methodology that uses smart card data to detect and estimate the place, time, period, and purpose of public transportation users, and then classify them by the trip purpose. Data from the Gatineau (Canada) and Santiago (Chile) were used to detect the activities of public transit passengers. Activity identification modules were used to label each trip with both transfer and non-transfer. Based on the available data sets and estimated activity attributes based on the distinguished criteria, four types of travel purposes were assigned. Based on the number of activity stages in both cities, the average travel time were estimated. The study also identified other behavioral patterns in consideration of sociological, geographical and cultural differences.

Seaborn et al. (2009) conducted basic data analysis based on the smart card data collected from London's buses and subways. They analyzed different transfer behavior groups (subway to bus, subway to bus, and bus to bus). Transfer time was also divided into three (low, middle, high) groups. The results showed that passengers' transfer behavior can be quantified using smart card data.

Ma et al. (2013) developed an effective data mining methodology using

Beijing public transportation data. In their study, bus stop information was estimated using a Bayesian decision tree algorithm based on the Markov chain. The algorithm extracts changes in the amount of boardings in time between two boarding records, and calculates the probabilities for all potential stops using the velocity profile of the GPS data. It is assumed that the stop with the maximum probability becomes the boarding stop. The trip chains were identified based on a fixed time threshold. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was used to analyze historical travel patterns of public transit passengers. They used the k-means clustering algorithm and the rough-set theory to detect the regularity of individual travel patterns. As a result of comparing the two algorithms, they found that the rough set theory is more accurate and efficient. They noted that this approach will help in travel behavior studies, transportation market analysis and OD estimation.

Ma and Wang (2014) also conducted a study to build a platform that combines GPS data and smart card data together. Based on the Bayesian theory of estimating the likelihood of each stop, the boarding stop was inferred. In the inference of the boarding stop, they applied spatial movement activity identification, daily trip chain analysis and historical travel pattern in an integrated manner. A light transit GIS data model was developed using Google's General Transit Feed Specification (GTFS). The platform not only served as a data visualization platform to observe public transport network performance for planning and operation, but also intended to combine information and communication technologies for data-driven transportation research and applications. The main functions of the platform are 'speed map of transit network unit', 'boarding analysis of station unit', 'headway distribution of station unit', and 'travel time reliability of route unit'.

Agard et al. (2007) developed a k-means cluster-based method to analyze public transport passenger behavior. The data includes card numbers, card types, and transaction dates, which are grouped into 20 groups. After applying the method, four passenger pattern groups were created. The results are presented in the form of ratios of each card type category in different travel pattern clusters. This study presented a feasible data mining technique and suggested that public transport operators could help better understand the behavior of passengers.

Kusakabe and Asakura (2014) have developed a data fusion method for estimating travel attributes that can not be recorded in smart card data using survey-based data. Although information on the location and the time of both boarding and alighting can be used in these two data sets, it can not cope with the one-on-one because of the difference in accuracy between smart card data and survey. However, if both the smart card data and the survey data have the same conditional probability distribution ratio, the other attributes of the two data sets can be estimated by the probability model based on the high concentration ratio. Using the Näive Bayes classifier, which applies probability distribution of behavioral properties, they added the attribute of travel (commuting, commuting, leisure, work, and return home) to each record of the smart card data.

Long et al. (2014) used the smart card data to analyze the spatio-temporal travel patterns of public transit passengers. A user who travels noticeably earlier than an average user; A user who travels uncommonly late; A user commuting from an overly long distance; And users with excessive use over the course of the day. In addition, household survey data was used as supplementary data to analyze the socioeconomic background of passengers.

Jánošíková et al. (2014) described how to use smart card data in the

logit model for the route choice. In-vehicle time, walking time for transfer, the number of transfers, and the headway of the route were considered. A case study of Žilina (Slovakia) was conducted and a total of 115,578 trip records were collected. The travel characteristics were estimated by dividing into the morning peak and the non-peak hour. Estimation results show that in-vehicle travel time is the most influential factor.

Kieu et al. (2013) conducted a study on the spatial and temporal characteristics of public transportation users using Queensland (Australia) smart card data. The trip chain method was configured with the time threshold of 60 minutes. DBSCAN approach was used to calculate the trip regularity of passengers who frequently use public transportation. The first step in the method was to analyze the entire passengers' itinerary. The second step was to examine the first boarding location of each cluster in Step 1. Finally, an example of using the classification of regular passengers and the estimation of individual travel time variability is presented.

Chang and Zhao (2016) presented a comprehensive data mining methodology that derives the travel patterns of individual transit passengers from incomplete information in a data set. The trip chain of each transit user was created on the basis of individual trip records. DBSCAN clustering was used to derive travel patterns of public transit users based on past trip chains. The travel pattern clustering was divided into three categories, "space-time rule", "space rule" and "time rule" respectively. This study analyzed the variable travel behavior of passengers.

El Mahrsi et al. (2014) conducted a study to derive a travel pattern from smart card data. They constructed the temporal profile of the passengers by using the boarding information and applied clustering based on the 'topic model' which is a unigram model to search for groups (or clusters) of passengers with similarity in view of the boarding time. Also,

they analyzed the relationship between passengers' travel behavior and socioeconomic characteristics (housing type: collective, individual, income, population density). For this purpose, passengers based on boarding information were grouped by residence area. And how socioeconomic characteristics are distributed among passengers' temporal clusters. The methodology was applied to the Rennes region of France using 4 weeks of smart card data and socioeconomic survey data.

In a subsequent study (El Mahrsi et al., 2017), each station was clustered according to similar usage patterns, and the relationship between time of day, location, and occurrence was analyzed. For this purpose, the boarding frequency profile describing the use of each stop was aggregated and clustered by applying model-based clustering.

Briand et al. (2016) carried out research to cluster the passengers according to the travel time in order to analyze passengers' temporal travel behaviors and patterns. A new method for estimating the Gaussian mixture model (GMM) in the temporal trip profile of passengers was proposed, considering the continuous time instead of the time bin used in most existing methods. They also conducted extensive experimental research using actual data sets and analyzed how this approach relates to various passenger behavior types (irregular passengers, typical commuters, etc.). In the same way as El Mahrsi et al. (2014), the cluster results of each passenger are analyzed in connection with the card type (General, youth, seniors, civil rights, short-term pass, etc) according to the fare of the smart card data.

Table 2- 1 Summary of destination(alighting) estimation researches based on smart card data

Authors	Year	Model	Method	Target Mode	Reference data range (days)	Validation
Barry et al.	2002	Trip Chaining	- Chaining to : (1) Next trip of the day, (2) Daily first trip for Last trip	Metro	1 day	Small samples of survey
Zhao et al.	2007	Trip Chaining	- Chaining to : (1), (2) - Threshold: 400m, 5min	Metro to (Metro,Bus)	1 day	-
Trépanier et al.	2007	Trip Chaining	- Chaining to : (1), (2), (3) First trip of next day for Last trip - Threshold: 2km	Bus	3 days (D-day: -1,0,+1)	-
Farzin	2008	Trip Chaining	- Chaining to : (1), (2)	Bus	1 day	5% samples of survey
Wang	2011	Trip Chaining	- Chaining to : (1), (2) - Threshold: 1km (12min)	Bus	1 day	-
Munizaga and Palma	2012	Trip Chaining	- Chaining to : (1), (2) - Threshold: 1km - Min(Generalized time)	Bus, Metro	1 day	-

Authors	Year	Model	Method	Target Mode	Reference data range (days)	Validation
He and Trépanier	2015	Trip Chaining, Probability	- Chaining to : (1), (2), (3) - Threshold: 2km - Kernel function using historical estimated alight records for Unlinked trip	Bus	Multiple days (no limitation)	-
Alsger et al.	2016	Trip Chaining	- Chaining to : (1), (2) - Threshold: 400-1100m, 30-90min - Scheduled time table constraint	Bus, Metro, Ferry	1 day	O
Kim and Lee	2017	Trip Chaining	- Chaining to : (1), (2) - Threshold: 500m - Min(Generalized distance)	Bus to(Metro,Bus)	1 day	O
Jung and Sohn	2017	Deep Learning	- Input variables :Transaction, Land use - Relaxed accuracy criteria: allow 2nd best candidate	Bus	1 day	O
This Study	**2019**	**Probability, Trip Chain**	**- travel pattern and historical trip records & Combination with existing trip chain (1), (2), (3)**		**Multiple days (no limitation)**	**Validate 50% of the analytical data**

Table 2- 2 Summary of studies on clustering of passenger behavior based on smart card data

Authors	Year	Objective	Method	Research Focus	Data Size
Agard et al.	2006	Clustering temporal passenger pattern	- k-means, Hierarchical Ascendant Classification	Status analysis of passengers and trips	2,147,049 transactions made by 25,452 cardholders
Morency et al.	2007	Measurement of spatial and temporal variability of transit passengers	- k-means	Status analysis of passengers and trips	2,200,000 transactions made by 7,118 cardholders
Ma et al.	2013	Individual travel pattern recognition, Travel regularity mining	- DBSCAN (pattern recognition), k-means++ (travel regularity)	Patterns and behavioral analysis	Transactions made by 3,845,444 cardholders
El Mahrsi et al.	2014, 2017	Clustering temporal passenger pattern	- Unigram Mixture (Topic model)	Development of passenger cluster model	5,404,096 transactions made by 134,979 cardholders

Authors	Year	Objective	Method	Research Focus	Data Size
Briand et al.	2016, 2017	Clustering temporal passenger pattern	- k-means, Gaussian Mixture Model	Development of passenger cluster model	3,492,310 transactions made by 82,223 cardholders
He et al.	2018	Classifying passengers' temporal boarding profile	- Hierarchical clustering (distance: CCD, DTW)	Analysis of traffic cluster methodology	100,000 transactions made by 3,095 cardholders
This Study	**2019**	**Clustering passengers by temporal pattern**	**- k-means, Gaussian Mixture Model**	**Alight estimation using temporal pattern**	**22,376,614 transactions made by 1,336,080 cardholders**

2.3. Direction of this research

As a result of reviewing the related works, many improvements have been made to the method of estimating the destination of the smart card data based on the 'trip chain method' proposed by Barry et al. (2002) for the first time. In particular, the performance of the model has been improved through the study of trip chaining rules, sensitivity analysis for key parameters (transfer threshold etc.), and constraints (vehicle schedule).

However, the method of estimating the alighting location by the trip chain has the limitation of the so-called 'unlinked trip' estimation that can not estimate the alighting location if the trip chain can not be constructed logically. Analysis of the new trip chaining rules is also a process to reduce these unlinked trips.

As an attempt to solve this problem, He and Trépanier (2015) proposed a method of estimating the temporal and spatial expectation values from historical data and estimating the alighting stop according to the alighting probabilities. However, this method has two limitations. First, since the estimated alighting location information is used for the estimation again, there is a limit to a reasonable basis information. Second, there is a limit to the solution of the unlinked trip because there is a problem that the historical trip records which have not succeeded in estimation is not used for the estimation.

The study on the travel pattern generation and clustering using the smart card data mainly applied the clustering methodology of the distribution criterion of travel time. The k-means method and the mixture model, which have the EM algorithm as a solution, maximize the expected value of the distribution are mainly applied as the methodology. Hierarchical clusters and

density-based clusters (DBSCAN, etc.) have been partially applied depending on the data conditions. There are various studies to generate travel pattern from the viewpoint of data mining. However, there is a lack of research on the utilization field and utilization methodology in terms of utilization of travel pattern information.

In this study, we propose an application method of 'historical smart card data' to solve the problem of 'unlinked trip' alighting location estimation of the trip chain method, focusing on the limit of the trip chain method and utilization of travel pattern information. This study presents a model and a solution method for estimating the alighting location of the smart car data that does not consist the trip chain logically.

This study defines travel pattern as distribution of boarding time. In order to apply the travel pattern to the alighting location estimation, all passengers are clustered according to the boarding time and the detailed boarding time distribution of each cluster is estimated. And we develop a method for estimating the destination (alighting location) according to the distribution of the generated boarding time. In addition, this study proceeds with comparative analysis with the existing trip chain method and examines the combined method of estimation with the trip chain method.

Chapter 3. Data Description

This chapter describes the general structure of the smart card data in Korea and describes the smart card data of Daejeon city (Korea) used in this study. In addition, the results of basic statistical analysis on the use of public transportation in Daejeon using smart card data are presented.

3.1. Data overview

3.1.1. Smart card data overview

The smart card data is composed of 'transaction data', which are generally referred to as smart card data, in a narrow sense and 'base data'. The base data is composed of bus route, stop (station), and bus route-stop data though there are some differences by region. Since the information available from the transactional content data without the base data is very restrictive, both the transactional content data and the base data are needed to obtain information on the public transit and operation.

Collection items and types of smart card data by smart card company or region are different, but they are included in the scope of the standard collection items of the Ministry of Land, Infrastructure and Transport. Table 3-1 shows the transaction data of the Ministry of Land, Infrastructure and Transport standards.

Summarizing the contents of the transaction data, it includes the time of using the public transportation, location information, and route information. It is possible to analyze the so-called trip chain by connecting individual

transaction (boarding) information using transfer information.

Table 3- 1 Data fields of general smart card data in Korea

Item	Contents
Serial Number	Given number of each data record
Area Code	Business area code of AFC companies
Virtual Card Number	Encrypted smart card ID
Transaction ID	Given numerical code when smart cards are tagged (assigned the same code with the previous tagging if the corresponding boarding is considered a transfer)
Passenger Type	General, children, student, silver, disabled etc.
Vehicle ID	Given numerical code of each vehicle (bus)
Operation Start Time	The start time of each run count (hh:mm:ss)
Operation End Time	The end time of each run count (hh:mm:ss)
Mode	Metro, bus (trunk, local, feeder), etc.
Route	Actual route number
Operator	Bus or metro operator name (or code)
Boarding Time	Date and time of boarding (hh:mm:ss)
Boarding Stop	ID, name and coordinates of boarding stop
Alighting Time	Date and time of alighting (hh:mm:ss)
Alighting Stop	ID, name and coordinates of alighting stop
Number of Transfers	Number of transfers within a journey (within same transaction)
Fare	Fare determined by the total length of the corresponding journey
Total Distance	Accumulated distance within a journey

The bus route information among the base data includes general information such as the covering region of each bus route, the number of stops, and the operating distance. Route-stop data is constructed to identify the routes through which actual routes operate, including the order (sequence) of stops and the distance between stops. In particular, the bus

route-stop data can be used to identify the moving distance of the actual user or to correlate the coordinates of each stop provided by the bus stop information to data to be.

At the heart of bus stop data is the ability to extend this to spatial analysis by providing coordinates for each stop. Also, since the administrative area of the stop is included, smart card transaction contents data can be constructed from aggregated data. The stop ID is used for linkage. The related metro station data includes information such as station name, route name, and operating agency, but coordinates and station connectivity information (network) for spatial analysis are not included, so it is necessary to construct information separately so that normal smart card data can be utilized.

The core of bus stop data is coordinate information for each stop. Coordinates can be extended to spatial analysis. Also, the administrative area of the stop is included, so that smart card data can be represented as geographically aggregated data.

3.1.2. Data description

This study utilizes the smart card data of Daejeon Metropolitan City for a total of 61 days from April 1, 2015 to May 31, 2018.

The 16 days of data consists of 12 days of holidays (including Sunday), 42 days of Saturdays (excluding Saturdays on holiday), 7 days of weekdays. The data includes trip records of Daejeon city-approved bus routes and Daejeon metro (1 line). Routes authorized by neighboring cities (such as Sejong city) are not included in the data. However, the trip records used Daejeon city's routes in other cities are included in the data. For example,

records using Sejong city' routes in Daejeon are not included in the data, but records using Daejeon city' routes in Sejong are included.

The data consist of a total of 32,119,951 individual boarding records. From this data, 2,977 records without information of stop codes are excluded from the analysis, and finally 32,116,974 records are set as the basic data for analysis.

Daejeon city's public transportation system consists of 114 bus routes and 1 metro line. When the number of data is classified by means, 25,922,133 (80.7%) records the number of bus trips to be estimated, and 6,194,841 (19.3%) in the subway.

There are 2,790 bus stops and 22 metro stations. However, since it is a logical stop recognized by the smart card system rather than a physical stop, it may be indicated on the code as if it is the same as another stop. However, due to the nature of the study model, coordinates are used rather than the individual code (ID) of the stop, so there is no problem in application. The average distance between bus stops was analyzed as 521m except for one Bus Rapid Transit (BRT) line which had a relatively long bus stop distance.

In the study related to estimating the destination using smart card data, it is important how much of the data contains the alighting information. This is because the accuracy of the model can be verified by using this alighting information. In case of Daejeon city, 50.7% of the number of bus trips include alighting information. The metro originally has a check-in and check-out system, so it includes 100% of alighting information. Including the metro trip records, 60.2% of the total boarding records include the alighting information.

This study is based on the estimation of the destination of the smart card data without the alighting information. Therefore, it is only used for

the verification of the estimation. Data used to develop models and algorithms intentionally masked the alighting information. In addition, we remove the external factors that may affect the model and construct the methodology using only the information shown in Table 3-2 below.

Table 3- 2 Data fields utilized in this study

Item	Contents
Virtual Card Number	Encrypted smart card ID
Passenger Type	General, children, student, silver, disabled etc.
Mode	Metro, bus
Route	Actual route number
Boarding Time	Date and time of boarding (hh:mm:ss)
Boarding Stop	ID, name and coordinates of boarding stop
Alighting Stop (For verification)	ID, name and coordinates of alighting stop

3.2. Basic statistics

We analyzed the trips by each day within 61 analysis days. Figure 3-1 and Figure 3-2 show the result of analyzing trips of all modes (bus and metro) and bus by day, respectively.

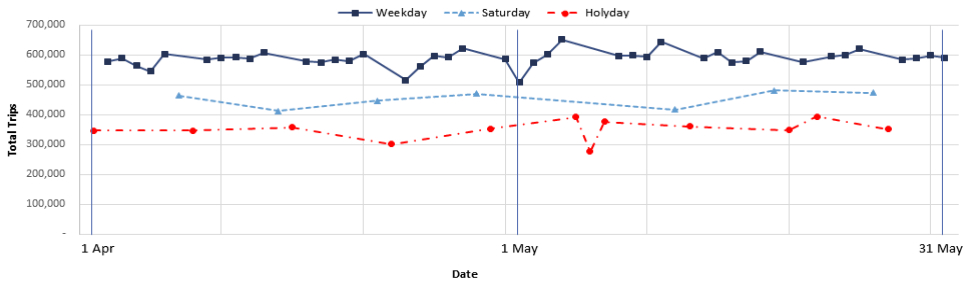


Figure 3- 1 Number of bus and metro trips during April 1 - May 31

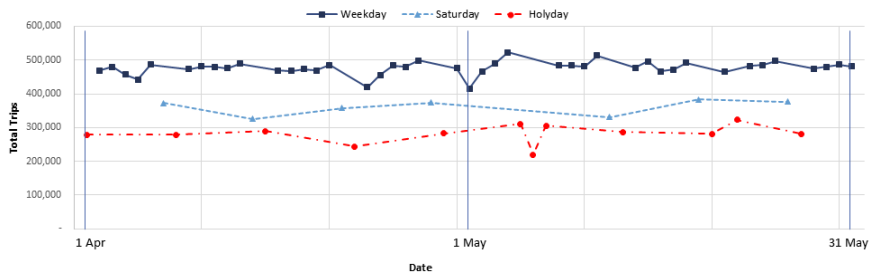


Figure 3- 2 Number of bus trips during April 1 - May 31

Comparing the two graphs, there is no intuitive difference between the overall pattern and the bus pattern. The number of trips per day is high in the order of weekdays, Saturdays and holidays. In the case of holiday on May 6, the date is the middle of the three-day holiday (May 5 to 7). Therefore, it is deduced that the number of trips is low compared to other dates. Table 3-3 summarizes the average number of trips by date types.

Table 3- 3 Average number of buses and metro trips during April 1 - May 31

Mode	Weekday	Saturday	Holidays
Bus + Metro	589,106	452,330	350,685
Bus	476,646	360,105	281,855

A total of 1,666,903 passengers used public transportation during the analysis period, and 1,398,207 (83.9%) used buses. Figure 3-3 shows the distribution of the number of passengers by frequency of use. The number of passengers who use public transportation only once during the analysis period is 300,741, which is 18.0% of all passengers. The trips generated by these passengers can not estimate destinations due to the assumption of the trip chain method. Also, 51.1% of passengers used public transportation less than 5 times in 61 days. Passengers with a low frequency of trip are the subjects to be considered when generating passenger profiles in Chapter 5.

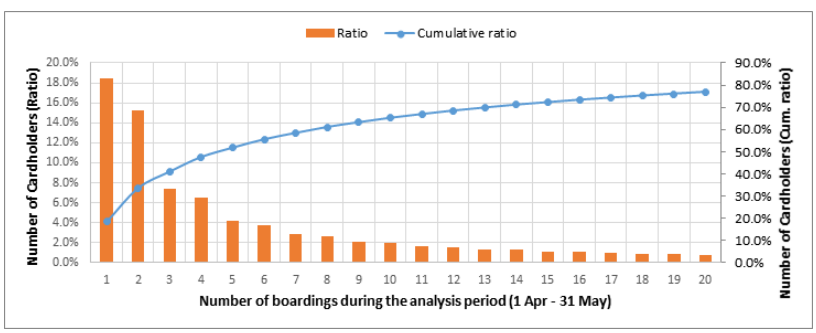


Figure 3- 3 Number of cardholders per number of boardings during the analysis period (1 Apr − 31 May)

Figure 3-4 shows the distribution of the number of passengers by the

number of days of use. During the analysis period, 521,678 people, 31.3% of total passengers, used public transport for only one day. Also, 52.3% of passengers used public transportation less than 3 days. This should also be taken into consideration when generating trip profiles for passengers.

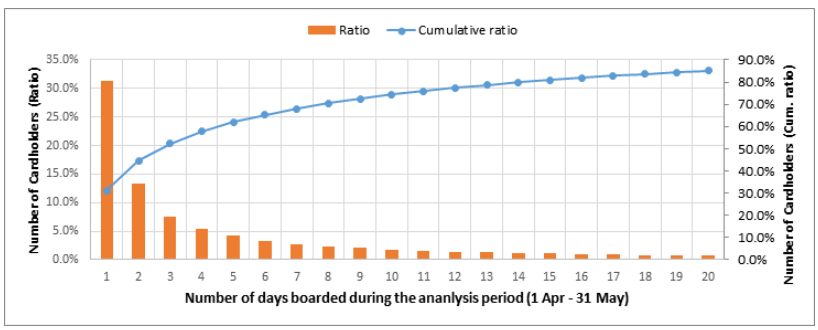


Figure 3- 4 Number of cardholders per number of days boarded during the analysis period (1 Apr − 31 May)

Figure 3-5 shows the distribution of the number of passengers per use frequency by date. It can be seen that passengers using one or two times account for more than 70%.

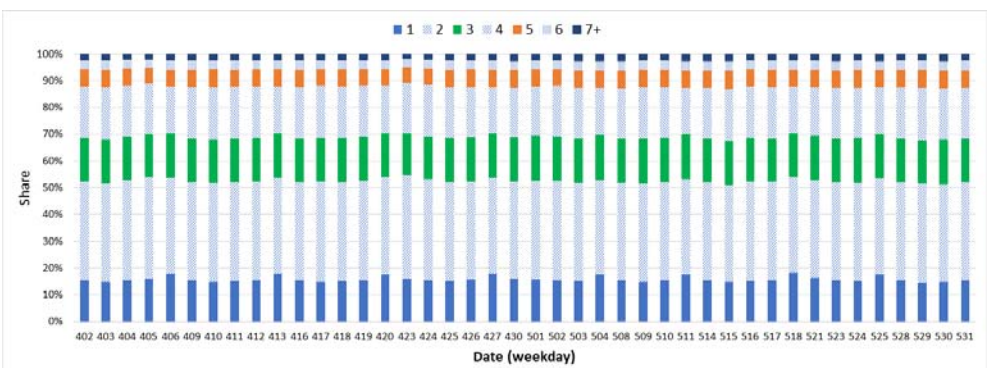


Figure 3- 5 Share of cardholders by number of boardings a day

Chapter 4. Travel Pattern Modeling

In this chapter we summarize the main assumptions of this study for the generation of travel patterns. According to these assumptions, we propose a model of travel pattern generation and destination estimation using smart card data. The travel pattern generation is divided into a process of generating a travel pattern cluster from the accumulated travel information obtained from various dates and a process of estimating travel pattern parameters of each travel pattern cluster.

4.1. Problem definition

The matching rate and the accuracy are improved by using the algorithm based on the trip chain method for estimating destination using smart card data. Nevertheless, 'unlinked trips', which still can not constitute a trip chain within a single trip or temporal and spatial threshold, are still limited in the alighting estimation.

Recently, as the chance of utilization of the smart card data as the public data has increased, the utilization range of the smart card data, which could only utilize the data within the past few days, has been extended to several months or more. This study hypothesizes that the missing information can be supplemented or new information can be derived by expanding the range of data to be used from the viewpoint of data mining.

From the point of reference to other trip, which is the basic assumption of the trip chain method, it is possible to consider a method of extending

the range of reference trips to more than several days of months' trips and utilizing it for the estimation of the destination. Figure 4-1 shows the result of the basic analysis of the method of estimating destinations according to the trip chain method using the smart card data of past several days. The analysis used the smart card data of Daejeon city for two months from April to May 2018. Regardless of the date, each trip was chained to the next boarding record in time. Therefore, the time difference (interval) between the reference data may be several days or more.

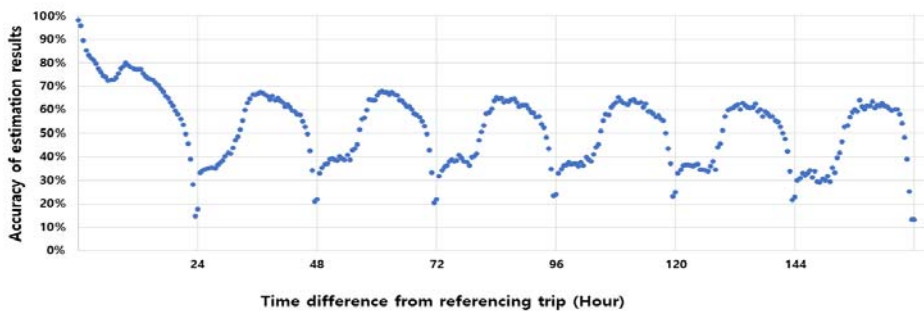


Figure 4- 1 Preliminary analysis: alighting stop estimation

As can be seen in Figure 4-1, there is a 24-hour time-series characteristics on the change in the accuracy of estimation with respect to the time difference between a boarding and the reference boarding (trip in the following order regardless of date). Particularly, when referring to the trip of the same time zone of different dates, a phenomenon that the estimation accuracy becomes lowest is observed. It can be deduced that the accuracy of the cross reference to the same time zone on the other day is low. For example, it is not advisable to refer to trips around 8:00 am in order to estimate the destination of a trip around 8:00 am.

That is, in order to refer to the historical boarding record, it is

reasonable to develop and apply the algorithm considering the temporal travel pattern. In this study, considering the temporal travel behavior, we refer to the appropriate historical trip records for the estimation of the destination. For this purpose, deriving the temporal travel behavior is the first detailed problem, and the second detailed problem is to estimate the destination using the derived travel behavior.

4.2. Modeling overview

This section describes the configuration and application of the major models needed to solve the defined problems. The purpose of this study is to generate a travel pattern from the trip history of public transit passengers and to utilize them to search for historical trips that can be referred to by the estimation object trips. For this purpose, the model was constructed to generate the travel pattern from the smart card data and to estimate the alighting location using the travel patterns.

This study follows the following five assumptions.

Assumptions 1. There are patterns of public transit trip, and passengers follow their own travel pattern.

Assumptions 2. The travel patterns of public transit passengers can be represented by the distribution(s) of the time of boarding.

Assumptions 3. The distribution(s) of the boarding time has a center point (intensive boarding time).

Assumptions 4. The passenger can be clustered by the similarity of the intensive boarding time(s) during the day.

Assumptions 5. The same temporal boarding behavior is followed in the same cluster. (travel pattern: assuming in the form of distribution(s) centered on intensive boarding time(s))

In the travel pattern generation based on the smart card data, a trip profile of each passenger (cardholder) is constructed from the historical data, and clustering into K clusters according to the similarity of profiles. Then, we apply a mixture model, which is a parametric method, to decompose the travel pattern of each cluster into H Gaussians. Through this process, a passenger is given one of K travel patterns, and all trips of the passenger belong to one of H distributions.

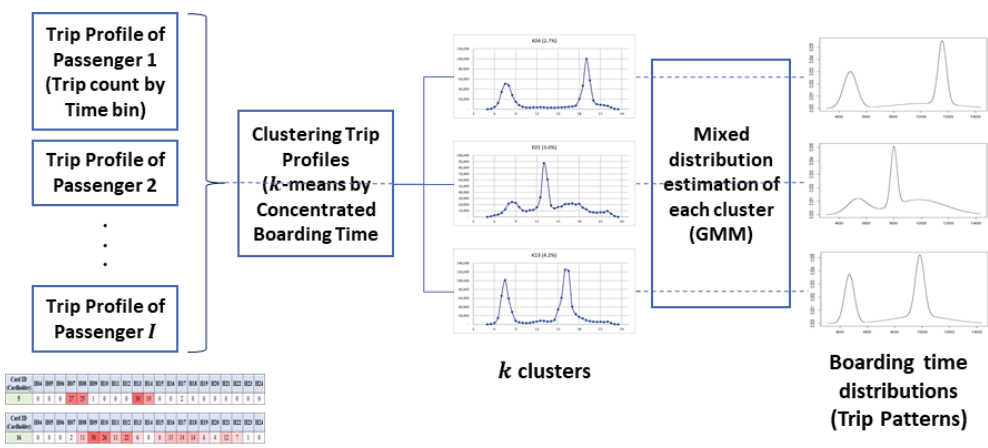


Figure 4- 2 Modeling concept of this study

In the step of estimating the destination based on the travel pattern, first, it is searched whether the passenger of the trip to be estimated has the previously generated travel pattern information or not. When there is his/her travel pattern, it estimates the alighting location probabilistically by using the parameter value of the travel pattern and the historical trip records of the passenger.

4.3. Travel travel profile generation for each passenger

In this study, groups of passengers with similar patterns in terms of time only (i.e. passengers who use public transit at the same time zone without regard to boarding locations) are grouped. Intuitively, the discovery of these groups can help identify frequent patterns of how passengers use public transit and analyze the characteristics of demand accordingly.

Generating a pattern of trip starts with creating a 'personal trip profile' by aggregating records of all the trips in the temporal scope of the analysis (excluding the date for verification). In this regard, El Mahrsi et al. (2014) compiled all trip records for each passenger in the form of counts of hourly (0h - 23h) hours per day (Monday - Sunday). This is in the form of a weekly profile and represented as a vector of 168 variables (24 hours × 7 days) per passenger.

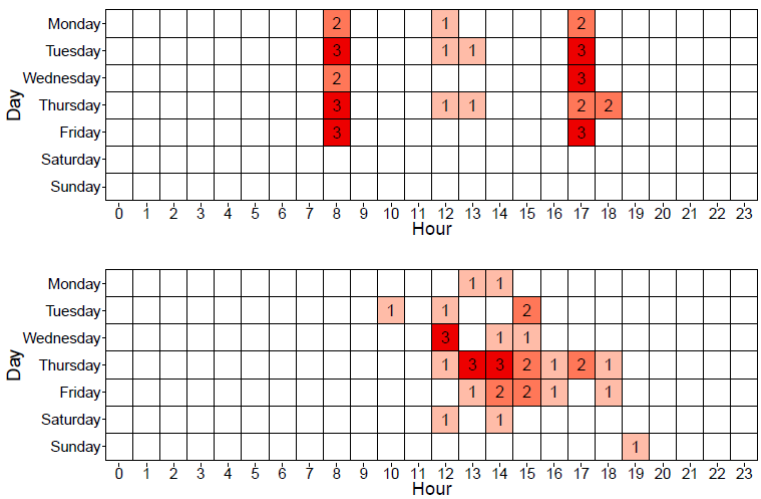


Figure 4- 3 Temporal profiles of two passengers sampled from the smart card data set (El Mahrsi et al., 2014)

This study refers to El Mahrsi et al. (2014)'s method and generates the travel profiles for individual passengers and time zone during the analysis period. When the size of the time bin decreases, the possibility of aggregation by each time unit becomes small. On the contrary, when the size of the time unit becomes large, the probability of aggregation increases, but it becomes difficult to segment the characteristics of the individual trip. Thus, like El Mahrsi et al. (2014), time units are set in one-hour increments. However, in consideration of the actual public transportation time, only 19 time zones (5h - 23h) except for the field are set as variables at dawn time. In consideration of the characteristics of general travel, the day of the week is divided into three at weekdays, Saturdays and weekends to generate a profile. Finally, in this study, the trip profile of each passenger (cardholder) is expressed in the form of a vector (u_i^w) of 19 variable types for each day class as shown in the following formula.

$$u_i^w = (u_i^1, u_i^2, \ ..., u_i^B)$$

- u_i^w : Travel profile vector of passenger i on date type w
- B : Numbers of time bins (In this study, 19)
- u_i^B : Total number of boardings excluding transfer trips during the analysis period by each time bin

Figure 4-4 is an example of a trip profile for 10 passengers configured on a weekday basis. Each column represents a time zone. For example, H08 means the sum of the cars from 4 am to 5 am. On weekends and holidays, the profile is configured in such a way that 44 variables are added horizontally. The number of rows of the trip profile dataset corresponds to the number of passengers included in the data of the trip profile generation. Since the data structure has been converted into a data structure having one record per passenger, clusters can be assigned per passenger rather than individual pass units.

Card ID (Cardholder)	H05	H06	H07	H08	H09	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23
5	0	0	27	25	1	0	0	0	30	19	0	0	2	0	0	0	0	0	0
16	0	0	2	11	30	26	11	22	6	0	8	13	14	14	6	4	12	7	1
17	0	0	1	13	23	2	1	0	0	0	0	0	0	0	0	8	29	2	0
23	0	0	0	32	0	0	0	0	23	0	1	8	7	1	0	0	1	0	0
44	0	0	0	13	14	0	1	5	0	0	0	0	12	16	0	0	0	0	0
47	0	34	56	10	0	2	10	12	12	15	21	13	4	2	2	4	0	0	0
63	0	0	0	16	0	1	0	0	1	0	0	0	6	10	1	0	2	0	0
72	0	0	42	26	4	1	0	0	0	2	1	0	6	9	18	17	9	1	0
77	0	0	1	40	23	0	0	0	1	11	38	14	0	0	0	0	0	0	0
98	0	0	3	16	8	2	2	2	2	2	0	0	0	4	0	0	10	21	0

Figure 4- 4 Individual travel profile (sample)

4.4. Travel profile clustering

In this study, two clustering processes are performed. First clustering is performed to analyze behaviors according to time series of trip distribution, and then clustering is performed to estimate Gaussian mixture distribution for each cluster. In order to prevent duplication of terms, the former is referred to as a travel profile clustering and the latter as a travel pattern generation.

4.4.1. Choosing a clustering method

Concerning the clustering of time series data, Aghabozorgi et al. (2015) analyzed that the time series clustering can be improved in four aspects of time series representation, similarity (distance) measures, prototype and clustering algorithm.

Next, we analyze the clustering algorithm. Hierarchical clustering has the disadvantage that it can not be effectively processed when the number of records (N) of data to be clustering is large because of the complexity of $O(N^2)$ or $O(N^2 \log N)$ in the computational complexity (Wang et al., 2006). It is difficult to apply hierarchical clustering because there are hundreds to several tens of millions of profiles in the characteristics of the large smart card data base. Model-based clustering also limits the application of large amounts of data. In addition, model-based clustering is sensitive to the results based on analyst's assumptions.

Density-based clustering, such as DBSCAN and OPTICS, is known to perform well in low-dimensional areas such as spatial clustering. However,

considering the characteristics of high-dimensional time series data (Fu, 2011), it is difficult to apply it. Aghabozorgi et al. (2015) reported that density-based clustering is not widely used in time-series data clustering due to high complexity in previous studies.

In this study, a partitioning-based clustering model is applied considering characteristics of large-scale smart card data and characteristics such as high-dimensional variables. Particularly, k-means, which is a subdivision cluster methodology suitable for large-volume data, is applied. And this study applies the Euclidean distance, which is most commonly used in clustering time series data as a similarity measure method (Aghabozorgi et al., 2015).

4.5. Pattern generation

4.5.1. Gaussian mixture model

In classifying patterns, it is important to analyze the distribution characteristics of data. In particular, it is necessary to model the data distribution assuming an appropriate probability density function (model). Under the assumption that there is a central point in the boarding time distribution within the cluster (Assumption 3), this study applied the Gaussian probability model as a function of the distribution of the boarding time. The Gaussian model is a probabilistic model that is suitable for representing the distribution form in which the entire observed data set is aggregated around the mean. However, the Gaussian probability distribution has a limitation that only the unimodal form in which the data are grouped

into one group around the mean can be expressed. Since the distribution of boarding time during the day appears to overlap with several distributions, we need a probability model that can represent more general forms. Therefore, in this study, 'Gaussian Mixture Model (GMM)' is applied to estimate the distribution of several Gaussian distributions.

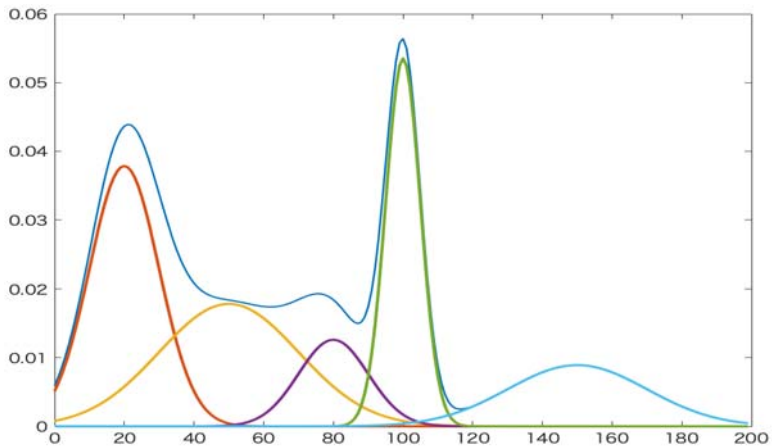


Figure 4- 5 An example of Gaussian Mixture Model
(Source: https://en.wikipedia.org/wiki/Mixture model)

As shown in Figure 4-5, this distribution is difficult to represent as a single distribution, but it can be expressed as a mixture of four Gaussian distributions. By applying the Gaussian mixture model, it is possible to approximate not only a distribution characteristic that one Gaussian distribution function can not show, but also a complex shape distribution as much as desired by using a sufficient number of Gaussian distributions. The total probability density function, defined as a linear combination of H simple probability density functions (or components), is expressed as:

$$p(x)=\sum_{h=1}^{H}\pi_h N(x|\mu_h,\Sigma_h)$$

- $p(x)$: probability density function
- H : numbers of Gaussians
- π_h : parameter (proportion of mixture component h, $0\leq\pi_h\leq 1$)
- N : mixture component (Gaussian distribution)
- μ_h : parameter (mean of Gaussian h)
- Σ_h : parameter (covariance of Gaussian h)

To approximate the Gaussian mixture model, we have to solve with a discrete latent variable. By expressing the distribution with the conditional probability through the latent variable, the formulas that can apply the EM (Expectation Maximization) algorithm, which is a solution of the Gaussian mixture model, are derived. The latent variable Z having the H dimension has a value of only one of H elements of 1 and the remainder of 0.

$$Z_h\in\{0,1\},\qquad \sum_{h=1}^{H}Z_h$$

- Z_h : latent variable

The joint probability distribution with latent variable is as follows.

$$p(x,Z)=p(Z)p(x|Z)$$

The probability of the latent variable is as follows.

$$p(Z) = \prod_{h=1}^{H} \pi_h^{Z_h}$$

The boarding time x follows the Gaussian distribution N when Z_h is satisfied.

$$p(x|Z_h = 1) \sim N(x|\mu_h, \Sigma_h)$$

The conditional probability of x for the latent variable Z is as follows.

$$p(x|Z) = \prod_{h=1}^{H} N(x|\mu_h, \Sigma_h)^{Z_h}$$

Using the above equations, the probability density function can be calculated as follows.

$$p(x) = \sum_{Z_h} p(Z_h) p(x|Z_h) = \sum_{h=1}^{H} \pi_h N(\mu_h, \Sigma_h)$$

As a result, the original probability density function formula is derived. The important thing in this process is that the joint probability distribution $p(x, Z)$ is calculated. The mixture model can be estimated according to the maximum likelihood estimation (MLE). The log-likelihood function of the Gaussian mixture model is as follows.

$$\log p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\log\left\{\sum_{h=1}^{H}\pi_h N(x_n|\mu_h,\Sigma_h)\right\}$$

A suitable means for performing MLE in a model with latent variables is the EM algorithm. The EM algorithm estimates a value using an iterative method. The GMM includes the initialization step, the E (Expectation) step, the M (Maximization) step, and an iteration.

In the initialization step, the average μ_h, covariance Σ_h, and mixture coefficient π_h are initialized to appropriate values.

The E step is a step of obtaining the posterior probability $\gamma(Z_{nh})$ using the value of the current parameter.

$$\gamma(Z_{nh})=\frac{\pi_h N(x_n|\mu_h,\Sigma_h)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}$$

In the M step, each parameter value is calculated using the posterior probability value calculated in the E step.

$$\mu_h^{new}=\frac{1}{N_h}\sum_{n=1}^{N}\gamma(Z_{nh})x_n$$

$$\Sigma_h^{new}=\frac{1}{N}\sum_{n=1}^{N}\gamma(Z_{nh})(x_n-\mu_h)(x_n-\mu_h)^T$$

$$\pi_h^{new}=\frac{N_h}{N}$$

$$N_h = \sum_{n=1}^{N} \gamma(Z_{nh})$$

After performing the M step, an updated new parameter is derived from the previous parameter, and using this, the E step of calculating the expected value or probability of the latent variable Z can be performed again. When the iteration of the M and the E steps is performed, the log likelihood value is increased. As a result, it converges to the local maximum value of the likelihood function through the EM algorithm.

4.5.2. Mixture modeling

This study assumes that the "the travel pattern of each cluster consists of a mixture of boarding time distributions and each distribution follows a Gaussian distribution". Therefore, it is necessary to estimate the distribution according to the two conditions of cluster K and Gaussian H to generate the travel pattern.

As we have seen, the Gaussian mixture model introduces a latent variable Z for model estimation. Since the precondition of the model is the number of clusters and Gaussian distributions, the latent variables Z^1 and Z^2 are set for each. At this time, $Z_k^1 \in \{0,1\}$, $Z_h^2 \in \{0,1\}$ and $\sum_{k=1}^{K} Z_k^1 = 1$, $\sum_{h=1}^{H} Z_h^2 = 1$. First, if the passenger belongs to the cluster k, $k \in \{1,...,K\}$, it belongs to the polynomial distribution M^1. The equation is expressed as follows.

$$Z_k^1 = 1 \sim M^1(1, \pi_k)$$

π_k is the proportion of cluster k and satisfies $0 \leq \pi_k \leq 1$. The probability equation for this is as follows.

$$p(Z_k^1) = \prod_{k=1}^{K} \pi_k^{Z_k^1}$$

Next, when a trip belongs to the distribution h in the cluster k, the trips follows the polynomial distribution M^2.

$$Z_h^2 | Z_k^1 = 1 \sim M^2(1, \tau_{kh})$$

Where τ_{kh} is the proportion of the Gaussian distribution h belonging to cluster k, satisfying $0 \leq \tau_{kh} \leq 1$. The probability of this is the same as the following equation.

$$p(Z_h^2 | Z_k^1) = \prod_{h=1}^{H} \tau_{kh}^{Z_h^2}$$

Therefore, the probability of Z^1 and Z^2 occurring at the same time is as follows.

$$p(Z_k^1 Z_h^2) = p(Z_k^1) p(Z_h^2 | Z_k^1) = \prod_{k=1}^{K} \left(\pi_k \prod_{h=1}^{H} \tau_{kh}^{Z_{kh}^2} \right)^{Z_k^1}$$

Meanwhile, the observation (boarding time) x that satisfy the cluster k and distribution h follow the Gaussian distribution N.

$$p\left(x|Z_k^1 Z_h^2 = 1\right) \sim N\left(x|\mu_{kh}, \sigma_{kh}\right)$$

μ_{kh} represents the mean of the Gaussian distribution h belonging to cluster k, and σ_{kh} means its standard deviation. In order to obtain the probability density function, the above expression is generalized as follows.

$$p\left(x|Z_k^1 Z_h^2\right) = \prod_{k=1}^{K}\left(\prod_{h=1}^{H} N\left(x|\mu_{kh}, \sigma_{kh}\right)^{Z_{kh}^2}\right)^{Z_k^1}$$

Using this formula, the probability density function (marginal probability) can be calculated. The calculation of marginal probability is as follows.

$$p(x) = \sum_{Z_k^1}\sum_{Z_h^2} p\left(Z_k^1 Z_h^2\right) p\left(x|Z_k^1 Z_h^2\right) = \sum_{k=1}^{K} \pi_k \left(\sum_{h=1}^{H} \tau_{kh} N\left(\mu_{kh}, \sigma_{kh}\right)\right)$$

As a result, the parameters needed to be estimated in this study can be summarized as π_k, which is the ratio of each cluster, τ_{kh}, which is the ratio of each Gaussian distribution in each cluster, μ_{kh}, and the standard deviation, σ_{kh}, of each Gaussian distribution.

The likelihood at this time can be summarized as follows.

$$L(\Theta)=\prod_{i=1}^{N^p}\sum_{k=1}^{K}\pi_k\left(\prod_{j=1}^{N^t}\sum_{h=1}^{H}\tau_{kh}N(\mu_{kh},\sigma_{kh})\right)$$

- $\Theta=(\pi,\tau,\mu,\sigma)$: Parameters of the likelihood
- i : ith passenger, j : jth trip
- N^p : Number of passengers
- N^t : Number of trips

The model of this study aims to generate H Gaussian distributions for each cluster after clustering trips with K first. Therefore, the likelihood function constitutes each of the latent variables for clustering. This can be expressed by the following equation.

$$L(\Theta|X,Z^1)=\prod_{i=1}^{N^p}\prod_{k=1}^{K}\left(\pi_k\prod_{j=1}^{N^t}\sum_{h=1}^{H}\tau_{kh}N(x_{ij}|\mu_{kh},\sigma_{kh})\right)^{Z_k^1}$$

Chapter 5. Algorithm

Chapter 5 presents an algorithm for estimating the destination using historical smart card data from many dates according to the constructed model. For the generation of the first-order clusters, the k-means algorithm suitable for clustering high-dimensional data is applied. For each of the generated clusters, the Gaussian Mixture Model (GMM) was applied.

5.1. Algorithm overview

5.1.1. Algorithm concept

The algorithms for solving the problems defined in this study are divided into the following two stages: (1) Algorithm for classifying public transit passengers based on the travel pattern model defined in Chapter 3, (2) Algorithm for estimation the alighting location based on passenger's travel pattern and historical boarding records.

From the viewpoint of solving the problem defined in this study, it is necessary to generate travel patterns to estimate the destination. The travel pattern derivation stage consists of building a trip profile using the number of boarding trips for each passenger by time bin (1 hour), clustering the trip profile, and generating a pattern using Gaussian mixture model for each cluster. Due to the nature of the model, the number of clusters K and the number of Gaussian distributions H should be determined in advance. To solve this problem, we generated travel pattern alternatives by changing the

K and H values and applied Integrated Completed Likelihood (ICL) criterion to search for the best fit among the generated travel pattern alternatives.

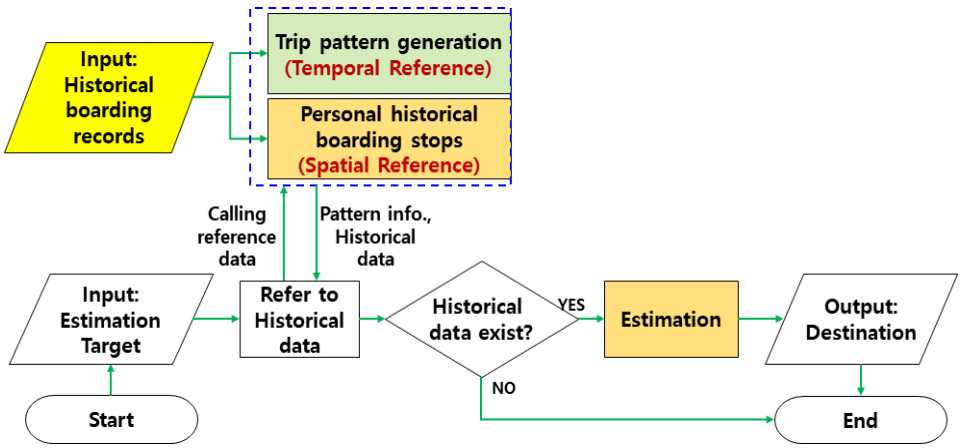


Figure 5- 1 The algorithm concept for this study

Generally, the Bayesian Information Criterion (BIC) standard is used to verify the GMM fitting. However, the BIC criterion tends to overestimate the number of clusters if appropriate models are not considered in the clustering process (Biernacki and Govaert, 1997). To solve this problem, Biernacki et al. (2000) suggested ICL criterion. The BIC is an approximation of the maximum log-likelihood, but the ICL criterion approximates the maximum complete log-likelihood. Indeed, ICL criterion are known to improve the disadvantages of the BIC standard when applied to standard data (Hamdan and Wu, 2013).

5.1.2. Algorithm notations

The notations used in the present algorithm are shown below.

Travel pattern generation

- $T_{i,d,j}$: passenger i's j-th trip on day d
- T_i^P : trip history (past trips) of passenger i
- $O_{i,d,j}$: boarding stop of passenger i's j-th trip on day d
- $x_{i,d,j}$: boarding time of $T_{i,d,j}$
- R^m : boarding transit route with the route ID number of m
- $R_{i,d,j}^m$: boarding transit of passenger i's j-th trip on day d
- K : total number of clusters
- h : Gaussian number (time group)
- u_i : travel profile of passenger i
- u_i^w : travel profile of passenger i on date type w
- b : number of time bin (where, max(b) = B)
- u_i^b : percentage of trip occurrence on time bin b for total trips of passenger i
- c_k : centroid value of cluster k
- E_k : Set of points belonging to cluster k
- π_h : 1st Gaussian parameter (proportion of Gaussian h, $0 \leq \pi_h \leq 1$)
- μ_h : 2nd Gaussian parameter (mean of Gaussian h)
- σ_h : 3rd Gaussian parameter (covariance of Gaussian h)

- τ_{kh}, μ_{kh}, σ_{kh} : Gaussian parameters of cluster k
- $Z^2_{i,j,h}$: latent variable (if $x_{i,d,j}$ belongs to Gaussian h, 1, else 0)
- $h^c_{i,d,j}$: Gaussian number to which $x_{i,d,j}$ belongs
- I : total number of passengers to be analyzed
- H_k : number of Gaussians constituting cluster k
- N_i : total number of trips of passenger i
- $N_{i,d}$: total number of trips of passenger i on day d
- h_k : Gaussian number assigned to $T_{i,d,j}$ belonging to cluster k

Destination estimation

- $D_{i,d,j}$: estimated destination of trip $T_{i,d,j}$
- $T^+_{i,d,j}$: Estimated trip information adding $D_{i,d,j}$ to $T_{i,d,j}$
- S : total number of stops
- $S^s_{i,h}$: stop s where passenger i was boarding in time group h
- $f^s_{i,h}$: boarding frequency of passenger i at S^s in time group h
- M : total number of transit (bus) routes
- Q^m : total number of stops constituting transit route R^m
- S^m_q : q-th stop of transit route R^m
- R^{mP}_i : transit route R^m concerning passenger i's trip history
- $R^{mP}_{i,h}$: transit route R^m concerning passenger i's trip history in time group (Gaussian) h
- $S^{mP}_{q,i}$: historical trip record of passenger i boarding near stop S^m_q
- $S^{mP}_{q,i,h}$: historical trip record of passenger i boarding near stop

S_q^m in time group h

- $S_i^{m,q,s}$: stop s where passenger i was boarding near stop S_q^m
- $S_{i,h}^{m,q,s}$: stop s where passenger i was boarding near stop S_q^m in time group h
- $f_i^{m,q,s}$: boarding frequency of passenger i at $S_i^{m,q,s}$
- $f_{i,h}^{m,q,s}$: boarding frequency of passenger i at $S_i^{m,q,s}$ in time group h
- d^E : Euclidean distance
- W_r : Allowable walking distance with radius r
- z : candidate alighting stops
- h^* : Gaussians other than the target Gaussian h

5.1.3. Algorithm pseudo-code

From here we describe the flow to the algorithms of this study. Algorithms are represented in the form of pseudo-code to aid understanding. First, the travel pattern generation algorithm can be expressed as the following Figure 5-2.

Algorithm 1: Travel Pattern Generation

Input : Boarding record of each trip $T_{i,d,j}^P = \{O_{i,d,j}, x_{i,d,j}, R_{i,d,j}^m\}$

Output : Cluster number of each passenger $k \in \{1,2,...,K\}$
Gaussian label of each trip $h \in \{1,2,... H\}$
Gaussian parameters of each Gaussian distribution τ_{kh}, μ_{kh}, σ_{kh}

1) Trip Profile Generation

for $i = 1 : I$ **do**

for $b = 1 : B$ **do**

Counting trips by time bin $u_i^b = \sum_d \sum_j count\left(T_{i,d,j}^P\right)$

end

end

return $u_i = \left(u_i^1, u_i^2, ..., u_i^B\right)$

2) Trip Profile Clustering

Randomly initialize centroid of each cluster $c^0 = \left\{c_1^0, ..., c_K^0\right\},\ E_k = \{\ \}$

repeat

for $i = 1 : I$ **do**

$j = \underset{j}{argmin} \left| u_i - c_j \right|$ (distance between observation and centroids)

$E_j = E_j \cup \left\{u_i\right\}$ (reassignment of trip profile vectors)

end

for $k = 1 : K$ **do**

$c_k = \frac{1}{\left|E_k\right|} \sum_{u_i \in E_k} u_i$ (recomputation of centroids)

end

until cluster labels of the trip profile does not change any more

return trip profile clusters (cluster label) $u_i = \left(u_i^1, u_i^2, ..., u_i^B, k\right)$

cluster label for each trip $T_{i,d,j}^P = \left\{O_{i,d,j}, x_{i,d,j}, R_{i,d,j}^m, k\right\}$

3) Trip Classification in a Cluster

// GMM Estimation

Randomly initialize parameters for each cluster $\tau_{kh}^0,\ \mu_{kh}^0,\ \sigma_{kh}^0$

for $k = 1 : K$ **do**

repeat $t = 1 : T^{converge}$

// Expectation Step

for $i = 1 : I, j = 1 : N_i, h = 1 : H_k$ **do**

$\gamma\left(Z_{i,j,h}^2\right)$ (calculating probability for Gaussian h of i's j^{th} trip)

end

// Maximization Step

for $h = 1 : H_k$ **do**

τ_{kh}^{t+1}, μ_{kh}^{t+1}, σ_{kh}^{t+1} (updating parameters by probability)

end

until Log-Likelihood converged

end

return Gaussian parameters of each Gaussian distribution τ_{kh}, μ_{kh}, σ_{kh}

// GMM Classification

for $k = 1 : K$ **do**

for $i = 1 : I, d = 1 : D, j = 1 : N_{id}$ **do**

$h_{i,d,j}^c = \underset{h}{argmax}\ \gamma\left(Z_{i,j,h}^2\right)$

end

end

return Gaussian label for each trip $T_{i,d,j}^P = \left\{O_{i,d,j}, x_{i,d,j}, R_{i,d,j}^m, k, h_k\right\}$

Figure 5- 2 Travel pattern generation algorithm

When the process of deriving the travel pattern is completed, the cluster number k (one of K) is given as the information of the travel pattern for each trip (transaction), and a Gaussian number h is assigned to each k. In this case, since h is a value depending on k, even if any two trips have the same h, another pattern is obtained when k (cluster) is different.

Since the Gaussian mixture model is a parametric methodology, the parameters (proportion: τ, mean: μ, standard deviation: σ) of the Gaussian constituting each cluster are calculated. However, in the case of trip that can

not be reasonably deduced from the travel pattern due to insufficient number of trips during the travel pattern derivation process, travel pattern information is not generated and classified as unlinked trips.

When the travel pattern generation process for each trip is completed, the destination estimation process is carried out. The stage of destination estimation utilizes the travel pattern information generated in the preceding stage and the personal boarding records. The detailed algorithm's pseudo-code is shown in Figure 5-3.

Algorithm 2: Destination Estimation

Input : Alighting estimation object (trip) $T_{i,d,j} = \left\{ O_{i,d,j}, x_{i,d,j}, R^m_{i,d,j} \right\}$

Output : Estimated alighting location (stop) $T^+_{i,d,j} = \left\{ O_{i,d,j}, x_{i,d,j}, R^m_{i,d,j}, D_{i,d,j} \right\}$

---

1) Assign travel pattern Parameters

for $i = 1 : I$ **do**

$$T_{i,d,j} = \left\{ O_{i,d,j}, x_{i,d,j}, R^m_{i,d,j}, k \right\}, \quad \forall d, j$$

(assigning cluster label for each trip)

using $u_i = \left(u_i^1, u_i^2, ..., u_i^B, k \right)$ from Algorithm 1

end

for $k = 1 : K$ **do**

for $h = 1 : H_k, i = 1 : I, d = 1 : D, j = d : N_{id}$ **do**

$$h^c_{i,d,j} = \underset{h}{argmax} \frac{1}{\sigma_{kh}\sqrt{2\pi}} e^{\frac{-\left(x_{i,d,j} - \mu_{kh}\right)^2}{2\sigma^2_{kh}}}$$

(assigning Gaussian label for each trip))

end

end

return Gaussian label for each trip $T_{i,d,j} = \left\{ O_{i,d,j}, x_{i,d,j}, R^m_{i,d,j}, k, h^c_{i,d,j} \right\}$

---

2) Alighting Estimation

// Collection of stops with historical boarding records by individual passengers

for $i = 1 : I, s = 1 : S, h = 1 : H_k$ **do**

$$f_{i,h}^{s} = \sum_{h} count\left(S_{i,h}^{s}\right), \quad \forall d, j$$

(Counting boarding records of stop S^s in time group h)

end

for $m = 1 : M, q = 1 : Q^m, h = 1 : H_k$ **do**

$$S_{q,i,h}^{mP} = \left\{S_{q,i,h}^{mP}\right\} \cup \left(S_{i,h}^{m,q,s}, f_{i,h}^{m,q,s}\right) \quad \text{s.t. } d^{E}\left(S_{q}^{m}, S_{i,h}^{m,q,s}\right) \leq W_r, x_{i,d,j} \in h$$

(Collecting boarding records around the current route)

end

return Boarding records around the current route's each stop

$$S_{q,i,h}^{mP} = \left\{S_{q}^{m}, \left(S_{i,h}^{m,q,1}, f_{i,h}^{m,q,S}\right), \ldots, \left(S_{i,h}^{m,q,1}, f_{i,h}^{m,q,S}\right)\right\}$$

// Estimation of Alighting Location

for $i = 1 : I$ **do**

for $m = 1 : M, q = 1 : Q^m, h = 1 : H_k$ **do**

$$p(z) = \frac{\sum_{s=1}^{S} f_{i,h^*}^{m,q=z,s}}{\sum_{q=1}^{Q}\sum_{s=1}^{S} f_{i,h^*}^{m,q,s}} \quad \text{s.t. } h^* \neq h \text{ (Probability of alighting at stop } z\text{)}$$

$$D_{i,d,j} = \underset{z}{argmax}\, p(z), \quad \forall d, j \text{ (Assigning as the alighting stop)}$$

end

end

return Estimated alighting stop $T_{i,d,j}^{+} = \left\{O_{i,d,j}, x_{i,d,j}, R_{i,d,j}^{m}, D_{i,d,j}\right\}$

Figure 5- 3 Destination estimation algorithm

The method of estimating the destination based on the travel pattern is as follows. Estimates the distribution to which the estimated trip belongs by using the Gaussian parameters for each passenger calculated in the travel

pattern generation stage. The value of the probability model belonging to each Gaussian is obtained from the characteristics of the Gaussian mixture model. And the trip is assigned to Gaussian distribution with the highest probability value for the trip.

After estimating the Gaussian to which each trip belongs, we search historical records of boarding around all stops (within the allowed walking distance) via the trip's route. In this study, the allowable walking distance is assumed to be 500m. The corresponding Gaussian numbers are stored together with the historical boarding records. As a result, only the trips which belong to the Gaussians other than the Gaussian of the target trip are utilized for the estimation. Using the historical trip records, it is estimated that a stop with a lot of nearby boarding records is the destination.

5.2. Travel pattern generation

5.2.1. Data preprocessing

In this study, the trip profile is defined as the ratio of trip occurrences (boarding trips) by time. Therefore, the passenger (cardholder), the boarding time and the boarding time zone (bin) information are stored in the database for all trips of the smart card data. To do this, the data of the entire date is sorted by date of boarding for each passenger by dividing date characteristic (weekday, Saturday, holiday). The data of the last several days for verification are separated.

For normal generation of the time-scale trip profile, only the representative trip (the first trip in the detailed trip connected by the

transfer) must be used during the single-purpose trip that is connected to the transfer. For example, if a passenger leaves at 8:10, transfers first at 8:30, and then transfers at 8:50, a tree trips are counted from 8:00 to 9:00. However, when considering a transfer, only one trip can be counted.

Temporal and spatial assumptions are required to distinguish transfer trips. As a temporal condition, the j-th boarding and the following j+1-th boarding should occur within a certain time threshold. As a spatial condition, there should be a stop of the route of the j-th trip within a certain spatial threshold of the boarding location of the j+1-th trip. This study assumes that the temporal threshold is 1 hour and the spatial threshold is 500 meters.

Assume "$T_{i,d,j+1}$" as a transfer trip of "$T_{i,d,j}$"

when $\left(x_{i,d,j+1} - x_{i,d,j}\right) \leq t^F$ and $\left\{z \middle| d^E(z, O_{i,d,j+1}) \leq W_r\right\} \neq \varnothing$

s.t. $\max(j) \geq 2, \quad j < \max(j), \quad z \in R^m_{i,d,j}, \quad q_{O_{i,d,j}} < q_z,$

- t^F : threshold of transfer time between boarding times

After determining transfer trips, utilize only the trips, not the transfer trips, to generate the trip profiles. However, transfer trips are still included in the estimation target of the destination.

5.2.2. Trip profile generation

Define the conditions that can affect the trip profile before generating the passenger trip profile within the analysis period. First, the data is divided into weekdays, Saturdays, and holidays. Second, since the passage pattern is information based on past traffic records, the minimum number of public transportation service use and the number of days of use are imposed. Third, set a time bin size, which is the minimum unit of the traffic profile. As mentioned in Chapter 4, the size of the time bin determines the aggregation possibility and characteristic segmentation.

The trip profile (u_i^w) is generated only when the minimum number of uses or conditions for the number of days of use are met. Therefore, for passengers who are not satisfied with this condition, a trip profile cannot be generated, and the destination cannot be estimated.

$$u_i^w = (u_i^1, u_i^2, u_i^b, \ ..., u_i^B)$$

$$u_i^b = count\left(T_i^b\right) / \sum_{b=1}^{B} count\left(T_i^b\right)$$

s.t. $w \in$ {day type 1, ⋯, day type W}, $\quad b \le x_{i,d,j} < b+1$,

$$count(d_i) \ge d^{\min}, \quad count(T_i) \ge T^{\min}$$

- T_i : passenger i's total trips during the analysis period
- T_i^b : passenger i's total trips on time bin b excluding transfer trips
- d_i : number of days that passenger i used public transit

- d^{min} : minimum number of days of use during the analysis period to generate trip profile
- T^{min} : minimum number of trips during the analysis period to generate trip profile

5.2.3. Trip profile clustering

This process combines the trip profiles generated according to the time of board for each passenger into K clusters by temporal characteristic. Through this process, passengers with similar time and frequency of board are clustered together.

As described in Chapter 4, the partitioning clustering method is applied considering the size of the sample and the number of dimensions. The most common partitioning method, k-means clustering, was applied.

In the k-means clustering, the sum of the square of the distance between each centroid of each cluster and each object in the cluster is set as a cost function, and the cost function is minimized.

$$E_k = \underset{E_k}{argmin} \sum_{k=1}^{K} \sum_{i \in E_k}^{I} |u_i - c_k|^2$$

After setting the initial centroid, the cluster assignment is performed by repeating the cluster assignment and the center update process. In the cluster allocation step, the Euclidean distance from each data to each cluster is calculated, and data is allocated by finding the closest cluster in the data.

$$E_k^{(t)} = \left\{ u_i \middle| \left| u_i - c_k^{(t)} \right|^2 \le \left| u_i - c_{k^*}^{(t)} \right|^2 \forall i,\ 1 \le k \le K,\ k \ne k^* \right\}$$

In the update phase, the centers are updated to the cluster level assigned in the previous step.

$$c_k^{(t+1)} = \frac{1}{\left| E_k^{(t)} \right|} \sum_{u_i \in E_k^{(t)}} u_i$$

And, repeat the above two steps until the results are converged.

5.2.4. Trip classification in a cluster

The key to the trip pattern generation phase is to estimate the Gaussian mixture distribution of time of board for each cluster. The estimated number of Gaussian distributions by cluster and the parameters by Gaussian mean the travel pattern for each cluster.

Since k is estimated in the previous step, the EM (Expectation Maximization) algorithm can be applied to each cluster to estimate the Gaussian parameter.

Initialize the mean μ_{kh}, standard deviation σ_{kh}, and distribution ratio τ_{kh} to a reasonable value. In the next E step, the posterior probability $\gamma\left(Z_{ijh}^2\right)$ is obtained using the value of the initial parameter.

$$\gamma\left(Z_{ijh}^2\right)= \frac{\tau_{kh}^{(t)} N\left(x_{ij}|\mu_{kh}^{(t)},\sigma_{kh}^{(t)}\right)}{\sum_{h=1}^{H} \tau_{kh}^{(t)} N\left(x_{ij}|\mu_{kh}^{(t)},\sigma_{kh}^{(t)}\right)}$$

s.t. $x_{ij} \in$ cluster k

- $\tau_{kh}^{(t)}$, $\mu_{kh}^{(t)}$, $\sigma_{kh}^{(t)}$: initial parameters of Gaussian distribution

Each parameter value is calculated using the probability calculated in the M step.

$$\mu_{kh}^{(t+1)} = \frac{1}{N_h^k} \sum_{j=1}^{N_h^k} \gamma\left(Z_{ijh}^2\right) x_{ij}$$

$$\sigma_{kh}^{(t+1)} = \frac{1}{N_h^k} \sum_{j=1}^{N_h^k} \gamma\left(Z_{ijh}^2\right)\left(x_{ij}-\mu_{kh}^{(t)}\right)\left(x_{ij}-\mu_{kh}^{(t)}\right)^T$$

$$\tau_{kh}^{(t+1)} = \frac{N_h^k}{N^k}$$

$$N_h^k = \sum_{j=1}^{N^k} \gamma\left(Z_{ijh}^2\right)$$

- $\tau_{kh}^{(t+1)}$, $\mu_{kh}^{(t+1)}$, $\sigma_{kh}^{(t+1)}$: updated Gaussian parameters
- N^k : total number of trips belonging to cluster k
- N_h^k : total number of trips belonging to Gaussian h in cluster k

The likelihood is calculated according to the recalculated parameter

value. If it does not converge, go back to the E step to update the parameter. The likelihood function is given by the following equation as derived from the model development in Chapter 4.

$$L(\Theta|X,Z^1)=\prod_{i=1}^{N}\prod_{k=1}^{K}\left(\pi_k\prod_{j=1}^{N_h^k}\sum_{h=1}^{H}\tau_{kh}N(x_{ij}|\mu_{kh},\sigma_{kh})\right)^{Z_k^1}$$

However, in this study, we assume that the number of Gaussian distributions H may be different according to each cluster k, so the likelihood can be summarized as follows.

$$L^k(\Theta|X)=\prod_{j=1}^{N_h^k}\sum_{h=1}^{H}\tau_{kh}N(x_{ij}|\mu_{kh},\sigma_{kh}),\quad \forall k$$

The log-likelihood function at this time is as follows.

$$\ln L^k(\Theta|X)=\sum_{j=1}^{N_h^k}\sum_{h=1}^{H}\gamma(Z_{ijh}^2)\ln(\tau_{kh}N(x_{ij}|\mu_{kh},\sigma_{kh})),\quad \forall k$$

Therefore, we estimate H and parameters that maximize log likelihood for each cluster. If all the calculations converge, the cluster k and the Gaussian h value can be assigned to each trip where a trip profile exists. Further, not only the π value as the clustering parameter but also the parameters τ, μ, and σ by the Gaussian mixture model can be given. This ends the process of generating the travel pattern of each trip.

5.3. Destination estimation

5.3.1. Data preprocessing

The smart card data is updated as shown in Table 5-1 by assigning the cluster number k, Gaussian distribution number h, and Gaussian distribution parameter derived in the previous travel pattern generation stage for each trip (The items added are shown inside the bold lines of Table 5-1). Since the clustering process is performed for each passenger and the application of the Gaussian mixture model is performed for individual trips for each cluster, the cluster is assigned based on the card ID and the Gaussian number and parameters are given on the basis of the passenger and the boarding time.

The estimation target of this study is the alighting location of 'bus' trips. However, in the process of generating the trip chain and the trip pattern, boarding records of 'metro' trips were utilized as the reference information. Therefore, the database also includes metro trips. And, the alighting information contained in some trip data was used only for verification of the estimated results.

Table 5- 1 Data fields of general smart card data in Korea

Item	Contents
Serial Number	Given number of each record (ID of each trip)
Card ID	Encrypted Smart card ID (Passenger)
Passenger Type	General, Children, Student, Silver, Disabled etc.
Mode	Bus (trunk, local, feeder), Metro etc.
Route	Route ID, Route number (name)
Boarding Time	Date and time of Boarding (hh:mm:ss)
Boarding Stop	Stop ID, Sequence and Coordinates of Boarding stop
Cluster Number	Cluster number k to which it belongs
Gaussian Number	Gaussian number h to which it belongs
Gaussian Parameters	π_{kh}, τ_{kh}, μ_{kh}, σ_{kh}
Alighting Stop (For verification)	Stop ID, Sequence and Coordinates of Alighting stop

5.3.2. Destination estimation algorithm

The purpose of this study is to improve the estimation of alighting location, especially of the single trip. For this purpose, we propose a method to utilize the trip history information recorded in the smart card data of past several days for the destination estimation. In particular, we propose to generate travel patterns of public transportation users from past trip records and apply them to the destination estimation.

He and Trépanier (2015) estimated the destination of 'unlinked trip', which is the limitation of the previous trip chain model. The methodology is to estimate the temporal and spatial expectation value from the historical alighting records and to estimate the alighting location with a high alighting probability. The kernel density function of time and space was constructed to calculate the alighting probability of each stop. The destination information used in this study is the result estimated by the trip chain model, not the actual destination information.

However, this method poses two limitations. First, by reusing the estimated values in the estimates, the meaning of the 'data-driven' model is partially tarnished. Second, it is not possible to use the historical trip records which are not estimated by the trip chain model. For example, even if stops have high frequency and repeatability of boards, such as Figure 5-4, and can be inferred that there are relationships between the trip chain, this trip pattern cannot be used for the trip chain model if it does not meet the time constraints of the traffic chaining.

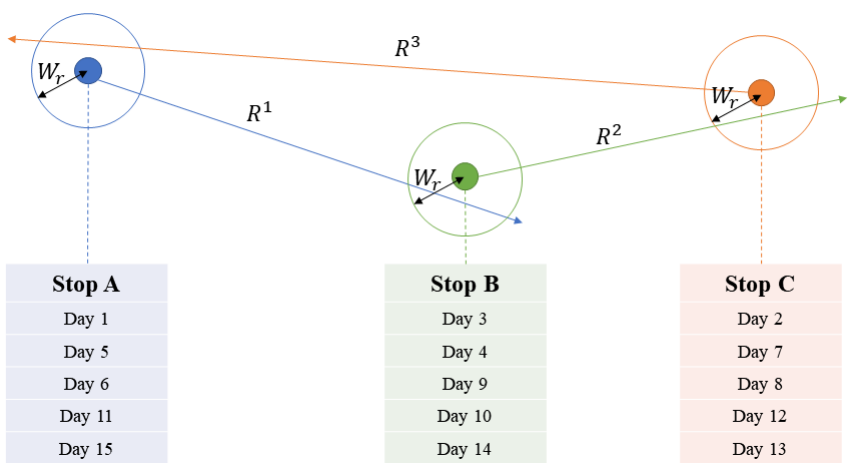


Figure 5- 4 Case that cannot be chained by conventional method despite frequency

In this study, the historical records are used to estimate the destinations of 'the unlinked trips', such as single trips, if the frequency of boarding are high for stops with historical boarding records.

$$T_i^P = \left\{ \left(S_i^1, f_i^1 \right), \ \ldots \ , \left(S_i^s, f_i^s \right) \right\}$$

- S_i^s : Stop where passenger i has ever boarded
- f_i^s : boarding frequency (number of past boarding times) of passenger i at stop S_i^s

For each stop that constitutes the boarding route to the estimated target trip, the corresponding elements of the trip are attached, and the nearby boarding record vector is constructed for each stop passing through. Estimate one of the stops via this vector as the alighting stop. This can be expressed in formulas as follows.

$$R_i^{mP} = \left\{ S_{1,i}^{mP}, S_{2,i}^{mP}, \ldots, S_{q,i}^{mP}, \ldots, S_{Q^m,i}^{mP} \right\}$$

$$S_{q,i}^{mP} = \left\{ S_q^m \right\} \cup \left(S_i^{m,q,s}, f_i^{m,q,s} \right)$$

$$\text{s.t. } d^E \left(S_q^m, S_i^{m,q,s} \right) \leq W_r$$

$S_{q,i}^{mP}$ S can be rearranged by the following formula.:

$$S_{q,i}^{mP} = \left\{ S_q^m, \left(S_i^{m,q,1}, f_i^{m,q,1} \right), , \left(S_i^{m,q,s}, f_i^{m,q,s} \right) \right\}$$

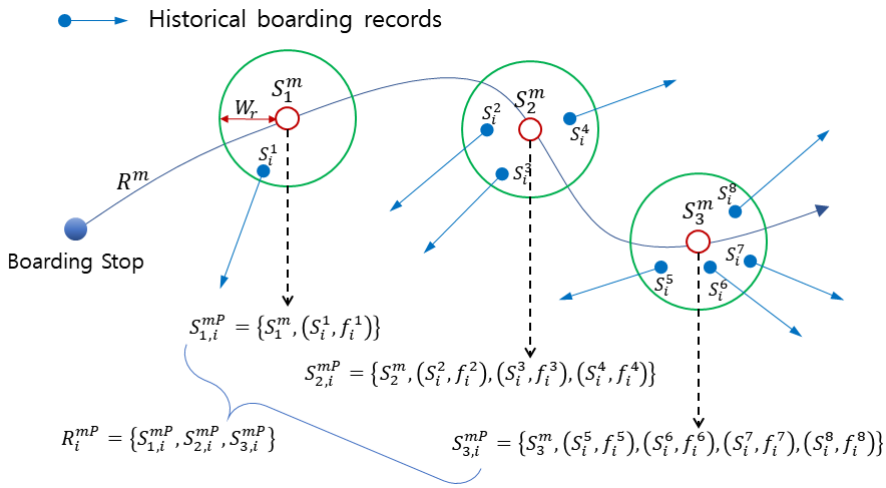


Figure 5- 5 Concept of historical trip record references

However, as discussed in Chapter 4, when estimating a destination by reference to another trip, it was confirmed that the time difference characteristic from the trip referred to was important. The estimation accuracy was low when referring to the trip of time adjacent to the estimated trip, but the estimation accuracy was relatively high when referring to the trip in different time zone. Therefore, the time-related variables should be included in the establishment of the boarding record information. For example, in estimating the alighting stops of the morning rush hour trip, it is reasonable to refer to the trip information in the evening rush hour rather than the trip record in the morning rush hour of the other day. As a result, the formula was supplemented by adding time-related factors (h) to the previously defined R_i^{mP} and $S_{q,i}^{mP}$.

$$R_{i,h}^{mP} = \left\{ S_{1,i,h}^{mP}, S_{2,i,h}^{mP}, \ldots, S_{q,i,h}^{mP}, \ldots, S_{Q^m,i,h}^{mP} \right\}$$

$$S_{q,i,h}^{mP} = \left\{ S_q^m, \left(S_{i,h}^{m,q,1}, f_{i,h}^{m,q,1} \right), , \left(S_{i,h}^{m,q,s}, f_{i,h}^{m,q,s} \right) \right\}$$

$$\text{s.t. } d^E\left(S_q^m, S_{i,h}^{m,q,s} \right) \leq W_r, \; x_{i,d,h} \in h$$

In this study, travel pattern is assumed to be distribution of boarding time of all trips. At this time, it is assumed that the boarding time in the day is a mixed distribution composed of H Gaussian distributions. For example, in the case of a commuting route, the boarding time of the morning (normally work trip) trip and the boarding time of the evening (normally return trip) trip are respectively followed by the Gaussian distribution.

This study assumes that trips capable of estimating exit stops are limited to those involving post-trip activities and that by the time taken for those activities, the following trips take place outside the time range (distribution). This means that the next trip in time distribution h^{α} occurs in distribution h^{β} after its activity time.

$$\text{If } p(x_{i,j}) \sim N(x_{i,j} | \mu_h, \sigma_h), \quad \text{then } p(x_{i,j+1}) \sim N(x_{i,j+1} | \mu_{h^*}, \sigma_{h^*})$$

$$x_{i,j} + t_{i,j}^A = x_{i,j+1}$$

$$\text{s.t. } h \neq h^*$$

- $x_{i,j}$: boarding time of passenger i's j-th trip
- t^A : activity time (including staying time at home)

And uses this information to estimate the departure stop $D_{i,d,j}$ of the

trip $T_{i,d,j}$. First, the time element of the boarding time is used to establish the trip record using the trip information of the rest of the time except the trip information near that time. These trip information become the reference trip information for the estimation.

$$T_{i,d,h}^{Ref} = \left\{ T_{i,d^*,h^*}^{P} \right\}$$

$$\text{s.t. } d \neq d^*,\ h \neq h^*$$

- $T_{i,d,h}^{Ref}$: set of passenger i's reference trips for the estimation
- d^* : dates other than the target date d

$$R_{i,h}^{mRef} = \left\{ S_{1,i,h}^{mRef}, S_{2,i,h}^{mRef}, ..., S_{q,i,h}^{mRef}, ..., S_{Q^m,i,h}^{mRef} \right\}$$

$$S_{q,i,h}^{mRef} = \left\{ S_q^m, \left(S_{i,h^*}^{m,q,1}, f_{i,h^*}^{m,q,1} \right), , \left(S_{i,h^*}^{m,q,s}, f_{i,h^*}^{m,q,s} \right) \right\}$$

$$\text{s.t. } d^E\left(S_q^m, S_{i,h^*}^{m,q,s} \right) \leq W_r,\ x_{i,d,h^*} \notin h$$

- $R_{i,h}^{mRef}$: reference route information for estimating the destination for a trip in the time group h
- $S_{q,i,h}^{mRef}$: reference stop information for estimating the destination for a trip in the time group h

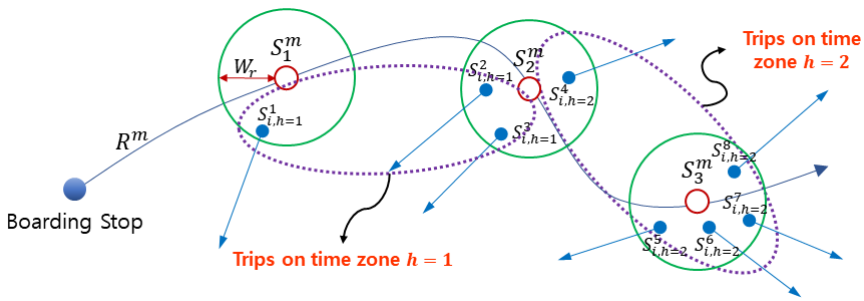


Figure 5- 6 Concept of historical trip records and travel pattern references

Next, the alighting probability according to the boarding frequency of the nearby stop is calculated for the stop after the boarding stop among the stops S_q^m constituting the boarding route $R_{i,d,j}^m$. In this study, the probability of alighting at any stop is proportional to the sum of the frequency of entry at the stop's adjacent stops, and the stop with the highest probability of alighting is estimated as the trip's destination.

$$p(z) = \frac{\sum_{s=1}^{S^s} \left(f_{i,h}^{m,q=z,s} \right)}{\sum_{q=1}^{Q^m} \sum_{s=1}^{S^s} \left(f_{i,h}^{m,q,s} \right)}$$

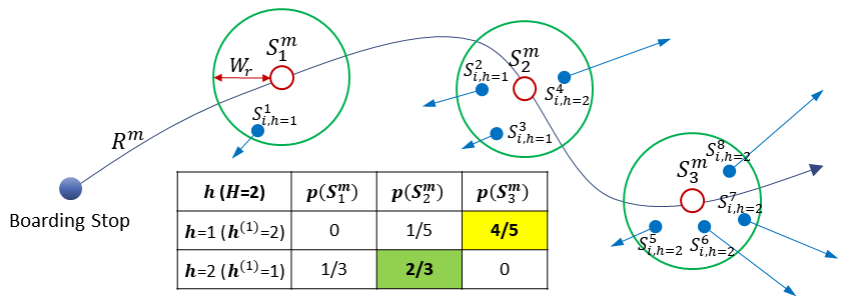
$$\text{s.t.} \quad z \in \left\{ S_1^m, \ldots, S_q^m, \ldots, S_{Q^m}^m \right\}, \qquad q_{i,d,j} < q_z$$

However, if the alighting probability are the same (or within the margin of error), a station with a minimum generalized distance is estimated as the alighting stop. This is to eliminate alternatives with large degree of bypass.

$$D_{i,d,j}^{C} = \underset{z}{argmax}\ p(z)$$

- $D_{i,d,j}^{C}$: alighting stop candidates of trip $T_{i,d,j}$

$$D_{i,d,j} = \underset{z}{argmin}\ d(z)$$



h (H=2)	$p(S_1^m)$	$p(S_2^m)$	$p(S_3^m)$
h=1 ($h^{(1)}$=2)	0	1/5	**4/5**
h=2 ($h^{(1)}$=1)	1/3	**2/3**	0

Figure 5- 7 Conceptual diagram of the alighting estimation based on travel pattern

The key to estimating a destination based on a trip pattern is how to define the trips to be referenced in time.

5.3.3. Selecting priority reference time group

The alighting stop is estimated using the boarding record of the Gaussian other than the Gaussian to which the boarding time belongs based on the boarding time of each trip. In the case of two Gaussian distributions (H = 2), the trips belonging to h = 1 refers to the trips belonging to h =

2, while the trip belonging to h = 2 refers to the trip belonging to h = 1.

However, if H is more than 3, it is necessary to apply the methodology of selecting one of Gaussian other than the corresponding Gaussian or assigning priority of selection. In this study, we utilize the qualitative travel pattern of each Gaussian derived from the travel pattern generation stage.

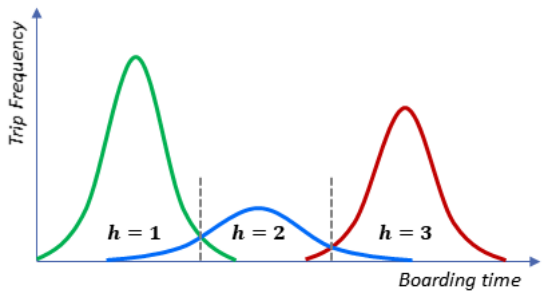


Figure 5- 8 Example of three Gaussians

At this time, according to the assumption of estimating the alighting location with reference to the trip record of the boarding time group h and another boarding time group, it can be expressed as the following expression.

$$T_{i,h}^{Ref} \in T_{i,h^*}^{P}$$

$$\text{s.t.} \quad h \neq h^*$$

- $T_{i,h}^{Ref}$: set of passenger i's reference trips for alighting estimation

In Figure 5-8, when the above expression is applied, the sets of

reference trips can be expressed in the following form:

$$T_{i,h=1}^{Ref} \in \left\{ T_{i,h=2}^{P},\ T_{i,h=3}^{P} \right\}$$

$$T_{i,h=2}^{Ref} \in \left\{ T_{i,h=1}^{P},\ T_{i,h=3}^{P} \right\}$$

$$T_{i,h=3}^{Ref} \in \left\{ T_{i,h=1}^{P},\ T_{i,h=3}^{P} \right\}$$

Among the Gaussian parameters, when the average of Gaussian is $\mu_{h=1}$, it means the distribution of the earliest time. This study assumes home-based trip when $h=1$. Therefore, in the case of a trip which is inferred to correspond to a pattern of returning home, the boarding information with $h=1$ is referred first. In the case of passengers who are deemed to correspond to the pattern of going to work, first refer to the boarding information showing a high peak pattern in the evening.

If the alighting stop can not be estimated in the primary reference in the travel pattern, the information is referenced in descending order of τ_h value. For example, if $H=3$ and $\tau_{h=3} > \tau_{h=1} > \tau_{h=2}$, then the trips corresponding to $h=1$ estimate the alighting stop by referring to the boarding record of the trips corresponding to $h=3$. For trips that can not be estimated, we refer to the next order, $h=2$. When it can not be estimated, the trips corresponding to the Gaussian of the next rank are referred to.

$$h^{ref(1)} = \underset{h^*}{argmax}\, \tau_{kh^*}$$

$$\text{s.t. } h=1,\ h^* \neq h$$

- $h^{ref(1)}$: first reference Gaussian for trips in Gaussian h

$$h^{ref(n)} = \underset{h^*}{argmax}\, \tau_{kh^*}$$

s.t. h=1, $h^* \neq h$, $h^* \notin \{h^{ref(1)}, ..., h^{ref(n-1)}\}$

- $h^{ref(n)}$: n-th reference Gaussian for trips in Gaussian h

$$h^{ref(1)} = 1$$

s.t. $h \neq 1$, $h^* \neq h$

$$h^{ref(n)} = \underset{h^*}{argmax}\, \tau_{kh^*}$$

s.t. $h \neq 1$,, $h^* \neq h$, $h^* \notin \{h^{ref(1)}, ..., h^{ref(n-1)}\}$, $n \geq 2$

In addition, by qualitatively determining the pattern shape of the cluster, a highly relevant distribution can be referred first. For example, in the case of a pattern deduced in the form of commuting, it may be considered that the returning-home trips are first referred to in estimating the destination of commuting trips.

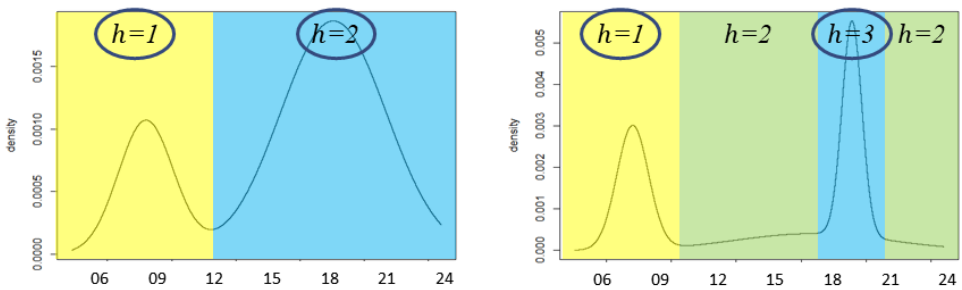


Figure 5- 9 Examples of 2 and 3 distributions

5.4. Combined estimation with the trip chain method

In addition, we analyze combination method of trip chain method and travel pattern method. In the trip chain method, it is estimated by using the next trip of the same date as the priorities of the detailed model application of the existing method, and then estimated by using the first trip of the date when the trip is the last trip of the date.

However, a model estimated using the first trip of the date or next day implies a home-based trip, which corresponds to a kind of travel pattern application. However, the next day can not be concluded that the first trip in time is a real home-based trip. Therefore, in this paper, we propose a combined method of estimating the alighting location by the trip distribution and applying the travel pattern to the trip where the alighting location is not estimated by referring to the next trip. Also, referring to the next trip and the first trip on the same day in stages, a method of applying the travel pattern to a trip where the alighting location is not estimated will be examined.

The first trip chain phase is to refer to the next boarding stop of the trip on the date. The second trip chain phase refers to the first boarding

stop on the date for trips that can not be estimated by the first phase. And, the third phase is to refer to the first boarding stop of the next day for trips that can not be estimated by the previous two phases.

The trip chain method is based on a phased approach, as shown in Figure 5-10. Estimated by the first phase and then estimated by the second phase for trips that can not be estimated by the first phase. Finally, the third phase applies to trips. This study takes this step-by-step approach and presents and evaluates the combined method.

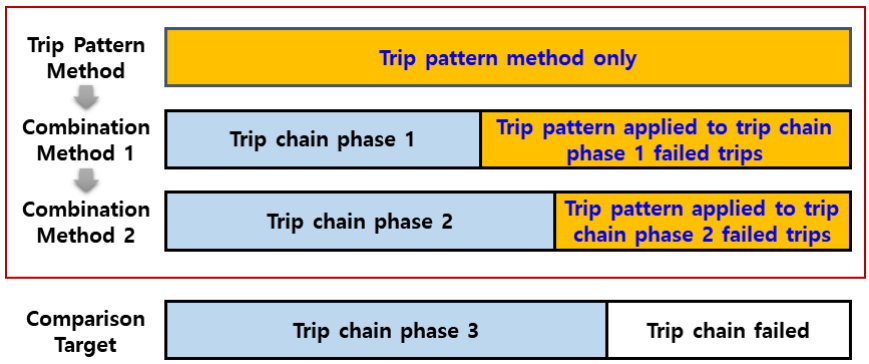


Figure 5- 10 Concept of combination method

Chapter 6. Case Study

Chapter 6 describes the results of case study conducted on the smart card data of Daejeon, Korea. The performance of the algorithm proposed in Chapter 5 is evaluated, and the performance is compared with the existing destination estimation method. The analysis results were presented according to the steps of the model and algorithms.

6.1. Data setting for validation

To verify the model and algorithm, training data and verification data for pattern generation were separated. As described above, the data used in this study is smart card data for 2 months from April to May 2018 of Daejeon Metropolitan City.

In this study, pattern generation data were built on weekdays basis with typical travel patterns. Specifically, there are 38 consecutive weekdays data within the April 2 to May 25 period. Out of the total 32,116,974 trips, 22,376,614 (69.7%,) trips were used to generate the travel pattern. For bus trips, trips data of 18,098,102 (69.8%) of the total 25,922,133 trips were used to generate the travel pattern. Among the pattern generation data, 9,089,230 of 13,148,952 are available for further verification by including the destination information.

Verification data to estimate the destinations was set to the last four weekdays of the collected data. Specifically, there are four consecutive weekday data from May 28 to May 31. Based on the basic statistical

analysis of the verification data, the number of transactions using public transit over a four-day period totaled 2,365,837, which is equivalent to 10.6% of the pattern generation data. Bus trips were observed at 1,921,036 cases, of which a total of 1,023,464 transactions were confirmed to be verifiable as they contain information on the destination.

The cluster number k is assigned to each card number in the pattern generation data, and the Gaussian number H and the parameter τ, μ, σ are assigned to each cluster. Parameters are used to determine the Gaussian distribution to which the verification data belongs.

6.2. Travel pattern generation

6.2.1. Passenger trip profile generation

The first step in generating a travel pattern is to create a trip profile for each passenger. Prior to the creation of a trip profile, transfer trip filtering and frequency-of-use filtering of data are required. Transfer trip filtering is a process for calculating the number of passengers based on single-purpose trip. In this study, transfer trip is assumed to be within 1 hour after the last boarding, and within the allowable transfer walking distance of 500m.

In this study, passengers (cardholders) who used public transportation for more than four days, which is about 10% of the 38 days of analysis, was set as the target of profile generation.

Table 6- 1 Data filtering results for passenger profiling

Contents	Before filtering	After filtering	Ratio
Number of trips	22,376,614	20,158,698	90.1%
Number of passengers	1,336,080	625,760	46.8%

As a result, for the remaining passengers and trips after filtering, it is possible to generate trip patterns and estimate the destinations based on the travel pattern. Trips by passengers without their own trip profiles are excluded from the estimation (analysis scope) of this study.

Trip profiles are generated by counting the number of passes per hour's aggregated time bin. Since zero cell problems can occur, 19 units were set up from 05:00 to 24:00 considering the hours of operating public transportation in Daejeon, not 24 hours. The following is a formula for the generation of a trip profile:

$$u_i^{weekday} = \left(u_i^1, u_i^2, ..., u_i^{21}\right)$$

- u_i^1 : Total number of boardings from 04 to 05 o'clock

The individual trip profile is generated as shown in Figure 6-1. The X-axis is the time bin, and the Y-axis is the individual encrypted card number. Therefore, as the number of passengers to be profiled increases, the data is structured in such a way that records (rows) increase. In this case study, we generated trip profile data composed of 19 items (columns) for 625,760 records.

Card ID	H05	H06	H07	H08	H09	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23
1	0	0	0	0	0	0	0	0	0	0	0	3	1	0	1	1	0	0	0
5	0	0	1	22	0	0	0	0	5	16	0	0	1	0	0	0	0	0	0
10	0	1	0	1	0	0	0	0	1	1	0	0	0	2	2	0	0	1	0
11	0	0	0	3	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0
16	0	0	0	1	7	14	4	4	4	0	2	3	6	4	3	2	7	2	0
17	0	0	1	11	21	2	1	0	0	0	0	0	0	0	0	5	28	2	0
23	0	0	0	29	0	0	0	0	21	0	0	8	5	1	0	0	1	0	0
27	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0
30	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0
31	0	0	0	0	0	2	1	0	2	1	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	1	0	0	2	1	1	8	9	2
40	0	0	0	2	2	0	16	0	1	7	6	5	0	0	0	0	0	0	0
42	0	0	13	7	4	0	0	0	0	0	0	0	6	0	1	6	3	3	0
43	0	0	0	0	1	0	0	0	3	1	7	4	23	0	1	0	0	0	0
44	0	0	0	12	12	0	0	5	0	0	0	0	10	15	0	0	0	0	0
46	0	0	0	0	2	0	1	0	1	0	1	1	3	0	0	0	1	0	0
47	0	4	22	8	0	1	5	6	8	4	8	9	4	2	1	2	0	0	0
49	0	0	0	0	0	0	0	0	3	3	5	6	2	6	8	4	2	0	0
50	0	24	0	0	0	0	1	1	0	1	0	2	22	4	0	1	1	0	0
51	0	0	5	8	0	0	0	1	0	0	1	0	1	11	3	2	1	5	0
52	0	3	0	2	12	9	9	1	10	4	3	2	3	1	0	1	0	0	0
53	0	0	0	1	3	4	6	8	8	5	1	0	1	1	0	0	0	0	0
54	0	0	0	0	4	0	0	0	0	1	1	3	2	2	0	1	0	1	0
55	0	0	0	0	0	0	0	0	1	2	2	2	0	0	0	0	0	0	0
56	0	0	5	1	0	0	0	22	2	3	11	1	1	0	1	0	0	1	0
58	0	0	0	0	0	0	0	1	0	1	0	1	1	1	2	0	0	2	0
60	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0

Figure 6- 1 Individual travel profile (sample)

6.2.2. Trip profile clustering

The following steps is to cluster the generated trip profiles using k -means clustering method. Each element of the trip profile vector corresponds to the dimension. In other words, it is a process of clustering 625,760 samples with 19 dimensions.

The number of clusters must be specified in advance according to the model characteristics of k-means clustering. In this study, ICL (Integrated Completed Likelihood) values are calculated for each cluster after completion of the next travel pattern generation step, and a proper number of clusters is selected. For this purpose, the number of clusters was set from 2 to 9, and a total of 10 repeated experiments were performed.

Table 6-2 shows the clustering results for k = 8 (8 clusters). The number of passengers in each cluster and the number of passengers in the total cluster are shown.

Table 6- 2 Summary of clustering result 1: share by clusters

Cluster	Passengers	Share	Rank
1	5,074,274	25.2%	1
2	837,181	4.2%	8
3	2,772,699	13.8%	3
4	1,927,074	9.6%	5
5	1,833,134	9.1%	7
6	1,898,078	9.4%	6
7	2,433,870	12.1%	4
8	3,382,388	16.8%	2

Figure 6-2 below shows the distribution of trips by time bins for each cluster and identifies the characteristics of each cluster. The x-axis is the time bin, and the y-axis is the total number of trips.

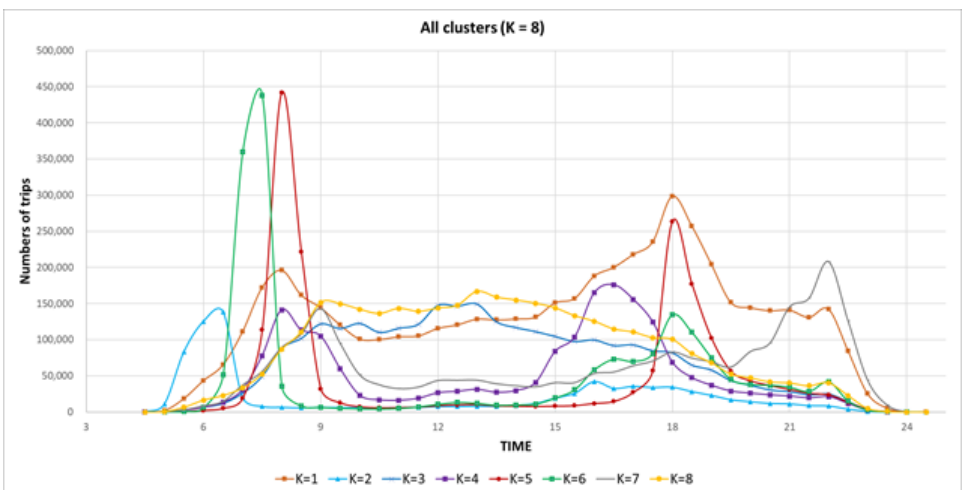


Figure 6- 2 Summary of clustering result 2: distribution diagrams

Because there are several clusters, it may be somewhat difficult to distinguish. Figure 6-3 shows the distribution of the number of trips by time bin for each cluster. The characteristics of temporal trips can be identified for each cluster.

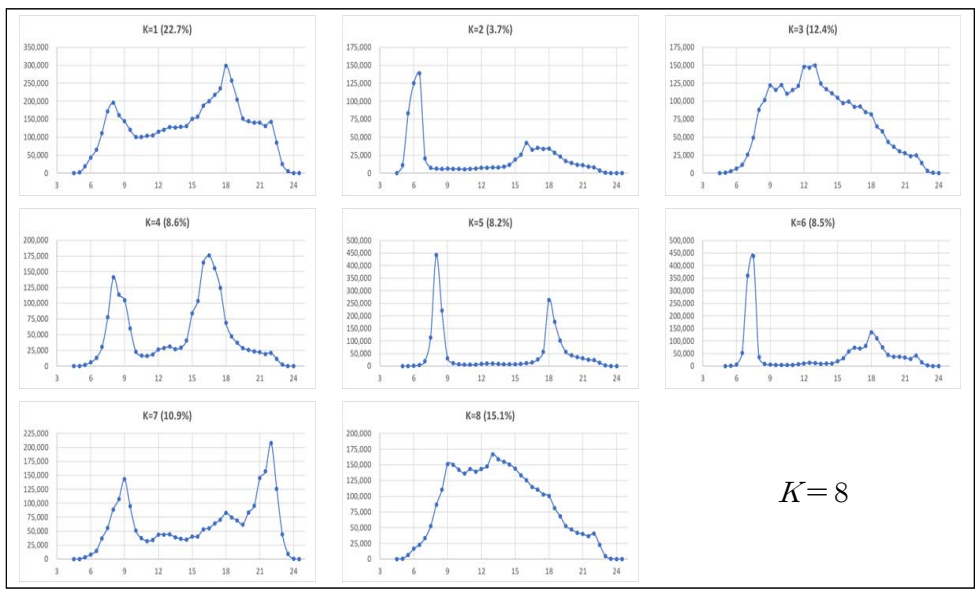
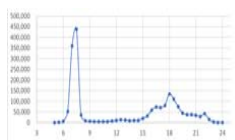
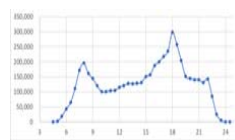
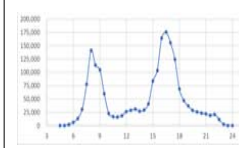
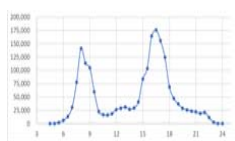
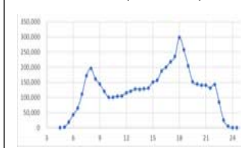
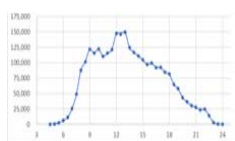
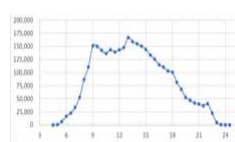
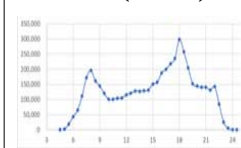
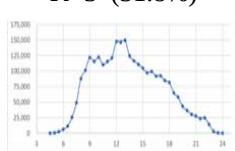
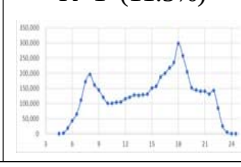


Figure 6- 3 Distribution diagram for each cluster

The smart card data generally has limited information on the characteristics of user types. However, the smart card of Daejeon city includes codes that can distinguish user types such as general, youth, children, the elderly, and the disabled. By this information, we analyzed the relevant cluster by user type. It is represented in a cluster order with a high occupancy rate.

Table 6- 3 Analysis of clustering results by passenger types

Passenger Type	Cluster share ranking		
	Rank 1	**Rank 2**	**Rank 3**
Students	K=6 (24.3%) 	K=1 (20.6%) 	K=4 (16.3%) 
Children	K=4 (24.9%) 	K= N/A (23.0%)	K=1 (22.8%) 
The elderly	K=3 (38.0%) 	K=8 (25.8%) 	K=1 (12.2%) 
The disabled	K=3 (31.8%) 	K=8 (19.3%) 	K=1 (11.5%) 

In the case of students, the K6, K1, and K4 clusters have the highest proportion. It can be seen that the trips are highly distributed in the time zones of going to and from the school. The distribution of children is similar to that of students. Students in this smart card data means middle and high school students. Considering this, it can be seen that the morning peak time is earlier than that of the child. In the case of children, 23% of them can not generate a profile due to frequent use less than 3 days during the analysis period.

The distributions of trips between the elderly and the disabled are similar. Rather than the characteristics of the normal morning and afternoon peaks, it can be seen that more trips are distributed during the daytime outside the peak hours.

6.2.3. Travel pattern generation

The following table shows the number of optimal distributions for each cluster and the parameters (τ_{kh}, μ_{kh}, σ_{kh}) for each Gaussian distribution based on k = 8.

Table 6- 4 Parameter estimation result by GMM

Cluster (k)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Number of Gaussians (h)		2	3	2	2	3	2	3	2
τ_{kh}	**h=1**	29.5%	44.9%	47.7%	30.4%	45.1%	47.4%	27.0%	47.7%
	h=2	70.5%	8.7%	52.3%	69.6%	24.4%	52.6%	44.1%	52.3%
	h=3	-	46.5%	-	-	30.5%	-	28.9%	-
μ_{kh}	**h=1**	554.0	381.9	650.3	517.7	498.9	450.4	541.5	655.0
	h=2	1,059.4	611.0	976.9	1,007.2	1,070.8	1,080.2	1,017.6	990.4
	h=3	-	1,057.1	-	-	1,113.5	-	1,312.1	-
σ_{kh}^2	**h=1**	11,822.1	773.7	15,238.1	3,041.2	538.6	472.7	4,955.8	17,414.7
	h=2	30,151.5	15,697.8	30,696.4	19,587.2	54,850.3	24,528.4	33,898.5	30814.0
	h=3	-	17,955.5	-	-	1,026.6	-	2,197.9	-

The following Figure 6-4 is a graph of the Gaussian mixture distribution for each cluster. It can be seen that it is similar to that shown in Figure 6-3. The traffic distribution was composed of two or three Gaussian distributions for each cluster.

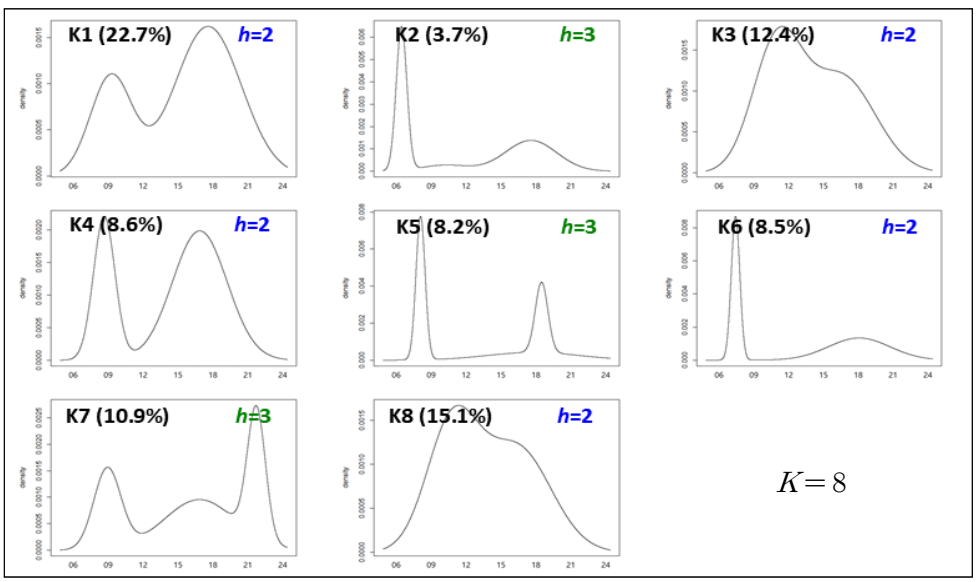


Figure 6- 4 Results of Gaussian mixture

6.3. Estimation results

6.3.1. Result analysis method

In the analysis of the estimation results, the accuracy criteria applied the relaxing criteria that allowed up to one difference in the sequence of the estimated and actual stops. In other words, it is assumed that the next and

previous stops at the actual stop are valid in the determination of accuracy.

The reason for applying the relaxing criteria is that the average distance between stops in the analyzed area (Daejeon city) is about 500 meters, which is similar to the allowable range for the nearby station assumption in this study. In other words, in the case of collecting the historical trips, the enxt and previous stops may also be included in the neighboring station's boarding records. This is a characteristic limitation of this study's model.

Trépanier et al. (2007) method was chosen as the existing trip chain method for comparison. However, instead of the generalization costs function of Trépanier et al. (2007), the generalization distance function proposed by Kim and Lee (2017) for easy application was applied. As a result of the application of the trip chain method, it was possible to estimate the destinations of the trips except 186,208 out of the total 1,023,464 trips, and 73.8% of the total trips can be estimated within one stop error.

Table 6- 5 Estimation result by trip chain method

Method	Trip chain phase 1	Trip chain phase 2	Trip chain phase 3
Sample Size (Estimation Target)	1,023,464	1,023,464	1,023,464
Matched Samples (Matching Rate)	651,944 (63.7%)	817,540 (79.9%)	837,256 (81.8%)
Samples Failed to Match	371,520	205,924	186,208
Accuracy (Tight Criterion)	485,317 (74.4%)	575,313 (70.4%)	582,247 (69.5%)
Accuracy (Relaxing Criterion)	611,202 (93.8%)	744,644 (91.1%)	755,168 (90.2%)
Accuracy to All Samples (Relaxing Criterion)	59.7%	72.8%	73.8%

Matching rate and accuracy (tight criterion and relaxing criterion) were applied as indicators of estimation performance. The matching rate is the ratio of the estimated number of trips to the total samples (trips), and the accuracy is the ratio of the number of trips that have been judged to be correct for the estimated trips.

As a result of the application of the trip chain method, the estimation rate for the total 1,023,464 trips was measured as 81.8%. The relaxed accuracy and the accuracy to all samples were 90.2% and 73.8%, respectively.

Next, Table 6-6 is the estimation result based on the travel pattern of this study. The estimation rate was measured as 71.0%. The relaxed accuracy and the accuracy to all samples were 58.0% and 41.1%, respectively. It can be seen that there is a performance limitation only by the travel pattern method.

Table 6- 6 Estimation result by travel pattern method

Method	Trip chain phase 3	Travel pattern
Sample Size (Estimation Target)	1,023,464	1,023,464
Matched Samples (Matching Rate)	837,256 (81.8%)	726,448 (71.0%)
Samples Failed to Match	186,208	297,016
Accuracy (Tight Criterion)	582,247 (69.5%)	201,685 (27.8%)
Accuracy (Relaxing Criterion)	755,168 (90.2%)	420,981 (58.0%)
Accuracy to All Samples (Relaxing Criterion)	73.8%	41.1%

In the next step, the travel pattern method was applied in combination with the trip chain method. First, we estimated the destinations according to the trip chain method phase 1, and applied a travel pattern to the failed trips of the trip chain model. In this case, the accuracy to all samples was 75.6%. This combination method is slightly superior to the case of using only the trip chain method. (Table 6-7)

Table 6- 7 Estimation result by combination method 1

Method	**Trip chain phase 3**	**Combination method 1 (Chain Phase 1 + Pattern)**
Sample Size (Estimation Target)	1,023,464	1,023,464
Matched Samples (Matching Rate)	837,256 (81.8%)	898,581 (87.8%)
Samples Failed to Match	186,208	124,883
Accuracy (Tight Criterion)	582,247 (69.5%)	565,168 (62.9%)
Accuracy (Relaxing Criterion)	755,168 (90.2%)	773,916 (86.1%)
Accuracy to All Samples (Relaxing Criterion)	73.8%	75.6%

And, we estimated the destinations according to the trip chain method phase 2, and applied a travel pattern to the failed trips of the trip chain model. In this case, the accuracy to all samples was 79.4%. This combination method is superior to the case of using only the trip chain method. (Table 6-8)

Table 6- 8 Estimation result by combination method 2

Method	Trip chain phase 3	Combination method 2 (Chain Phase 2 + Pattern)
Sample Size (Estimation Target)	1,023,464	1,023,464
Matched Samples (Matching Rate)	837,256 (81.8%)	933,616 (91.2%)
Samples Failed to Match	186,208	89,848
Accuracy (Tight Criterion)	582,247 (69.5%)	608,752 (65.2%)
Accuracy (Relaxing Criterion)	755,168 (90.2%)	812,264 (87.0%)
Accuracy to All Samples (Relaxing Criterion)	73.8%	79.4%

In order to accurately determine the effect and performance of the travel pattern method in each combination method, only the samples estimated from the travel pattern in each combination method were analyzed. The results are shown in Table 6-9. Considering the estimated performance index, the combination method 1 (trip chain method phase 1 + travel pattern) has the best performance.

Table 6- 9 Estimation result by travel pattern method in combination method

Method	travel pattern	travel pattern in Combination Method 1	travel pattern in Combination Method 2
Sample Size (Estimation Target)	1,023,464	371,520	205,924
Matched Samples (Matching Rate)	726,448 (71.0%)	246,637 (66.4%)	116,076 (56.4%)
Samples Failed to Match	297,016	124,883	89,848
Accuracy (Tight Criterion)	201,685 (27.8%)	79,851 (32.4%)	67,620 (28.8%)
Accuracy (Relaxing Criterion)	420,981 (58.0%)	162,714 (66.0%)	77,517 (58.3%)
Accuracy to All Samples (Relaxing Criterion)	41.1%	49.5%	37.6%

Chapter 7. Discussion

For insight into the characteristics of the model, Chapter 7 presents the results of analysis by various application methods and conditions. The category analysis of clusters, the sensitivity analysis by parameters and conditions and the comparisons with other city case are discussed in this chapter.

7.1. Travel pattern category analysis

7.1.1. Configuring categories

We conducted a category analysis of the cluster. The reason for applying the Gaussian mixer to the clusters derived from this study is to find the temporal breakpoints that distinguish the trip characteristics under the assumption that they do not refer to the same time group trips.

This breakpoint corresponds to the intersection of the Gaussian distribution in the cluster. In fact, the temporal reference in the trip destination estimate is determined by this breakpoint, so similar clusters can be categorized according to the breakpoint's prevalence. In other words, if the breakpoints are similar, it is to estimate the trip destination based on the simplified cluster, grouped into categories.

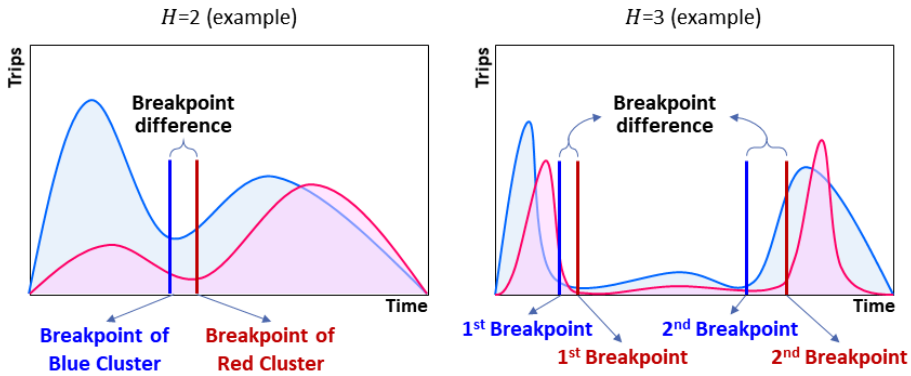


Figure 7- 1 Temporal breakpoint difference between clusters

How to organize the categories is as follows. First, it is divided into two Gaussian clusters and three Gaussian clusters. Then we calculate the overlap between the clusters and then group the clusters in order of the degree of mutual overlap. The degree of overlap is shown as the trips in which the cluster is maintained even when breakpoints of other clusters are applied, as shown in the figure below.

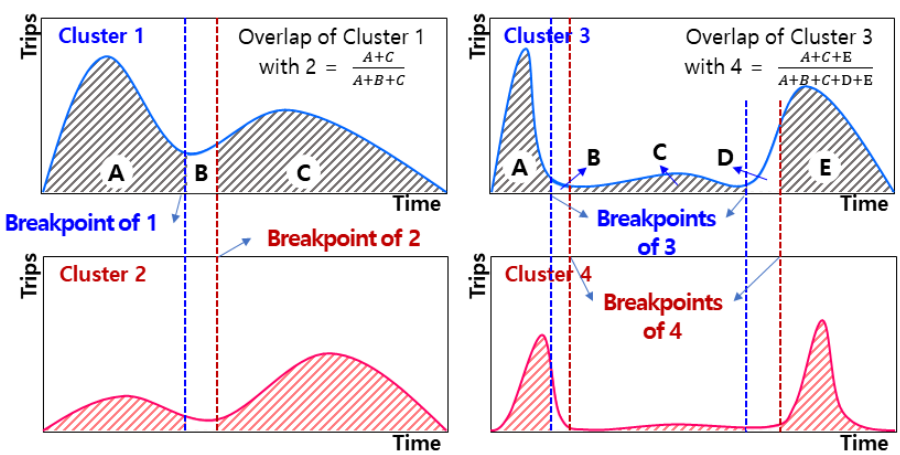


Figure 7- 2 Calculating the degree of mutual overlap

The categories are organized in the following order. First, the degree of overlapping of the comparison clusters for the reference clusters is calculated for every pair of clusters. Even within the same pair, if the reference and the comparison clusters are switched, the degree of overlap will be changed, so both directions will be calculated. When the calculation is completed, pairs with higher values of mutual product of degrees of overlap between the clusters are preferentially categorized. Rerun Gaussian Mixer based on categorizing results. When the execution is complete, calculate the degree of overlap again to tie the next higher overlay cluster. Repeat this process until the entire clusters are merged step by step.

As a result, for the two Gaussian groups, K1 and K3 were the first to be bound, followed by K8 and finally K4. For the group of three Gaussians, K5 and K6 were the first to be bound, followed by K2 and finally K7.

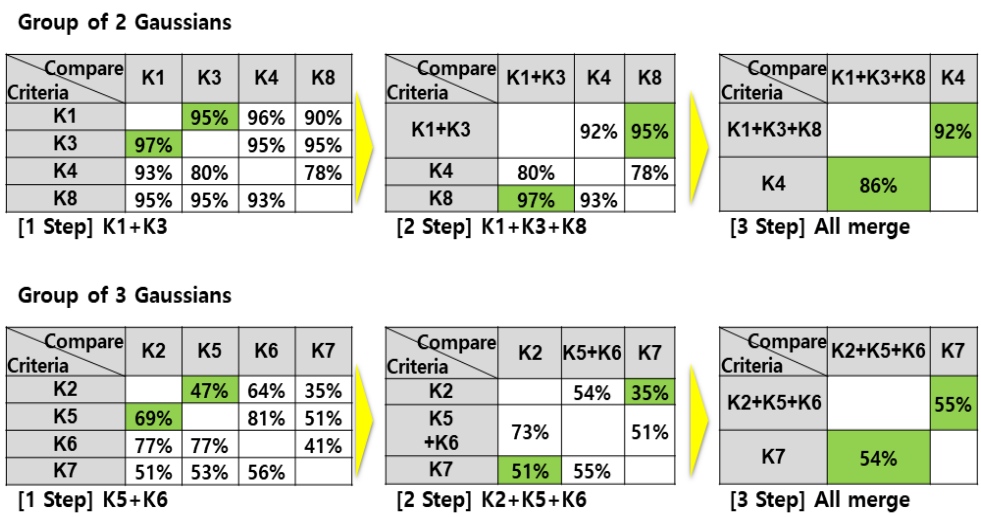


Figure 7- 3 Categorizing steps

We have evaluated the appropriateness of the category step-by-step category structure. Compared with the ratio of Gaussian changes belonging to the passengers before and after the category composition. The smaller the Gaussian variance, the smaller the effect on accuracy of estimation.

As a result, the two Gaussian groups could be grouped into one category in the 1, 3, and 8 clusters. In the case of three Gaussian groups, groups 2, 5 and 6 could be grouped into one category.

Table 7- 1 Evaluating the adequacy of a category configuration (2 Gaussians)

Category	K1+K3		K1+K3+K8		K1+K3+K8+K4	
	Cluster	Change rate	Cluster	Change rate	Cluster	Change rate
Group of 2 Gaussians (K: 1,3,4,8)	K1	8%	K1	9%	K1	3%
	K3	5%	K3	6%	K3	19%
	-	-	K8	5%	K8	18%
	-	-	-	-	K4	1%

Table 7- 2 Evaluating the adequacy of a category configuration (3 Gaussians)

Category	K5+K6		K2+K5+K6		K2+K5+K6+K7	
	Cluster	Change rate	Cluster	Change rate	Cluster	Change rate
Group of 3 Gaussians (K: 2,5,6,7)	K5	15%	K5	15%	K5	15%
	K6	3%	K6	3%	K6	5%
			K2	10%	K2	14%
					K7	29%

As a result of categorization, we could reconstruct a total of four clusters as shown in Figure 7-4. It can be seen that the clusters are divided by the interval between Gaussian and the trips during the daytime.

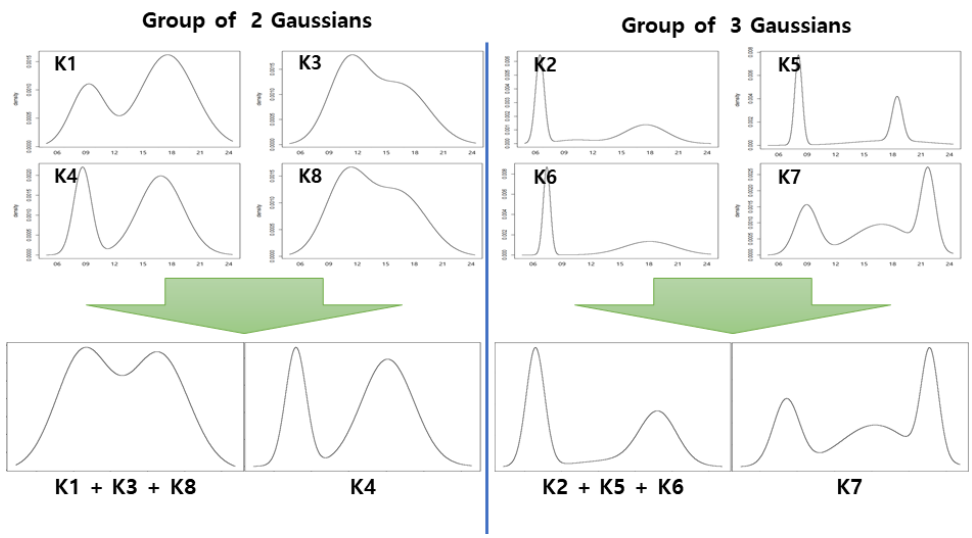


Figure 7- 4 Category configuration results

7.1.2. Results of category analysis

Based on the category analysis, the destination was estimated by the newly constructed cluster. As a result, it has been confirmed that the result of destination estimation is maintained even when the number of clusters is reduced. We have confirmed that the category configuration is valid.

Table 7- 3 Estimation result by cluster categorizing (travel pattern method)

Method	Before Categorizing	After Categorizing
Sample Size (Estimation Target)	1,023,464	1,023,464
Matched Samples (Matching Rate)	731,212 (71.4%)	741,832 (72.5%)
Samples Failed to Match	292,252	281,632
Accuracy (Tight Criterion)	202,807 (27.7%)	203,874 (27.5%)
Accuracy (Relaxing Criterion)	423,506 (57.9%)	425,938 (57.4)
Accuracy to All Samples (Relaxing Criterion)	41.4%	41.6%

Table 7- 4 Estimation result by cluster categorizing (travel pattern method)

Method	Before Categorizing	After Categorizing
Sample Size (Estimation Target)	1,023,464	1,023,464
Matched Samples (Matching Rate)	731,212 (71.4%)	741,832 (72.5%)
Samples Failed to Match	292,252	281,632
Accuracy (Tight Criterion)	202,807 (27.7%)	203,874 (27.5%)
Accuracy (Relaxing Criterion)	423,506 (57.9%)	425,938 (57.4)
Accuracy to All Samples (Relaxing Criterion)	41.4%	41.6%

7.2. Sensitivity analysis

7.2.1. Boarding time zone of trips

An analysis of the estimated performance of the travel pattern method for the morning peak (07:00~10:00) and the evening peak (16:00~19:00) shows that the travel pattern method is more effective for the first of the morning. In other words, it can be seen that the repetitive characteristic of the morning trip is more clear.

Table 7- 5 Estimation result by target trips' boarding time zone

Method	Morning peak		Evening peak	
	Trip Chain Phase 3	Combination Method 1	Trip Chain Phase 3	Combination Method 1
Sample Size (Estimation Target)	220,015	220,015	252,242	252,242
Matched Samples (Matching Rate)	183,100 (83.2%)	207,448 (94.3%)	205,484 (81.5%)	216,496 (85.8%)
Samples Failed to Match	36,915	12,567	46,758	35,746
Accuracy (Tight Criterion)	129,203 (70.6%)	137,250 (66.2%)	139,280 (67.8%)	131,446 (60.7%)
Accuracy (Relaxing Criterion)	166,863 (91.1%)	183,393 (88.4%)	184,603 (89.8%)	186,219 (86.0%)
Accuracy to All Samples (Relaxing Criterion)	75.8%	83.4%	73.2%	73.8%

7.2.2. Number of days referenced

The range of data referenced to the travel pattern was analyzed by expanding to 2 weeks, 4 weeks, 6 weeks, and 8 weeks. It is clear that the estimation rate improves with increasing data. On the other hand, estimation accuracy is constant. Since the accuracy to all samples can be expressed as the product of the estimation ratio and the accuracy, it is natural that the accuracy to all samples also improves according to the reference data range. The reason for the sensitivity analysis was to find out the critical point of the estimated performance according to the range of the reference data. However, due to the limit of the available data, the performance improvement aspect was found, but the finding of the convergent point was limited.

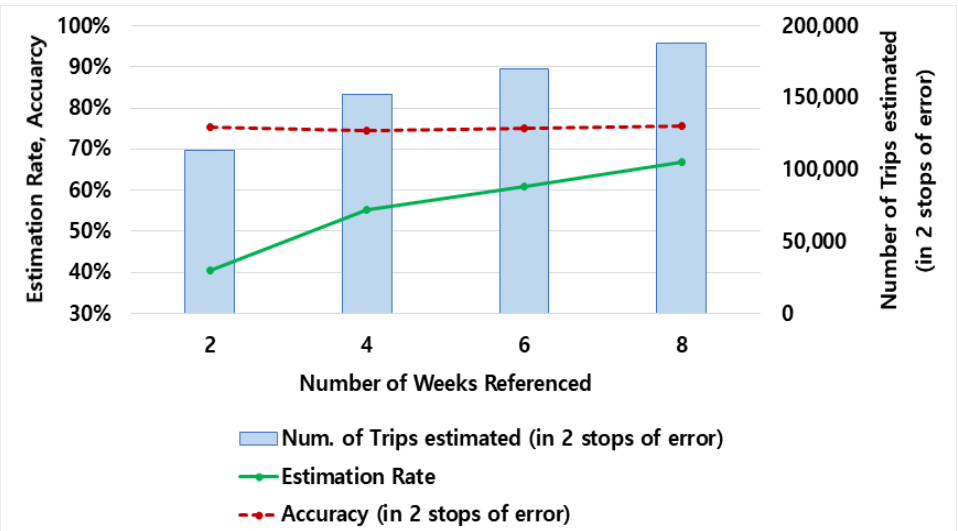


Figure 7- 5 Estimation result by number of days referenced

7.2.3. Repeatability of the stop used

The sensitivity of the estimation performance according to the number of stops used per day was analyzed. The number of stops that have been used for passengers who used public transit for more than 4 days during the analysis period (38 days) are calculated by dividing the number of stops that have been used by the total number of days of use. A small number of average stops used is assumed to be concentrated at a particular stop.

The analysis results are shown in Figure 7-6. It can be seen that the more distributed the stations used, the significantly lower the estimated accuracy by the pattern of traffic. According to the algorithms of this study, it is estimated that the stop with the highest number of past uses among candidate stops is the destination. The higher the number of stops used, the higher the number of stops targeted for candidates. Therefore, the more decentralized the use of the stop, the less accurate the estimate will be. The travel pattern method can be found to be advantageous in applying it when it shows repeated trip characteristics.

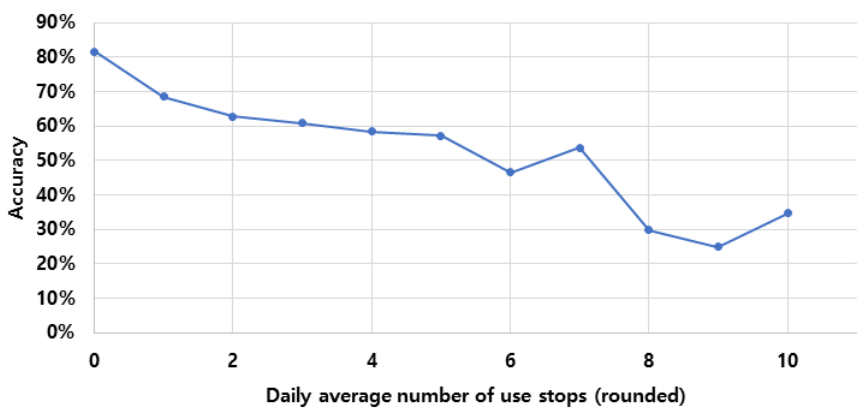


Figure 7- 6 Estimation result by repeatability of the stop used

7.2.4. Number of uses

We analyzed the accuracy performance according to the frequency of use for passengers who used more than 4 days during the analysis period (38 days). If the frequency of use is too large, it is assumed that the case is not general and excluded from the analysis. The results show that the accuracy of estimation by travel pattern method is maximized for the users who used 70 times (about 1.8 times / day). Travel pattern model is effective for repeated trips, including commuting.

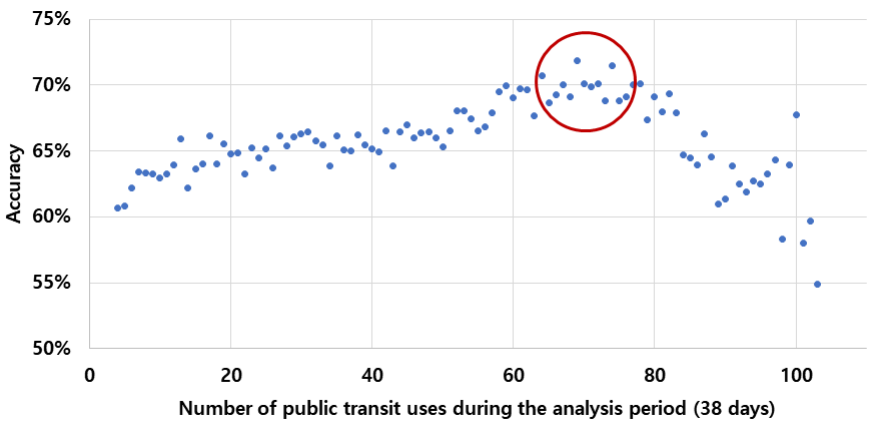


Figure 7- 7 Estimation result by number of uses

7.2.5. Alighting tag ratio

We analyzed the correlation between the alighting tag ratio and the estimation performance by each cluster. In addition, the trips that failed to estimate was correlated with the travel pattern characteristics of the cluster.

Table 7- 6 Estimation result by the alighting tag ratio

Category Criteria	K1	K2	K3	K4	K5	K6	K7	K8	Avg.
Alighting tag ratio	41.7%	41.2%	41.4%	32.7%	30.5%	26.9%	30.1%	44.4%	37.7%
Relaxing accuracy	65.1%	66.7%	64.8%	71.4%	71.4%	71.6%	60.7%	63.2%	66.0%

Correlation between the alighting tag rate and the estimation accuracy was observed. Therefore, it can be deduced that there is a relation between the alighting tag behavior and the travel pattern. If a passenger do not take the alighting tag, it is likely that the passengers are not going to transfer. That is, it can be advantageous for the travel pattern estimation.

7.2.6. Determining the number of clusters

For each cluster set in the trip profile clustering stage, the fitness of the ICL (Integrated Completed Likelihood) criterion with the change of the Gaussian number h was verified. The number of Gaussian was set from 1 to 5. Figure 7-8 shows the result of ICL measurement according to the change of number of clusters. At this time, a graph is created based on h which minimizes the ICL, and the h value is the number of optimal Gaussian of each cluster. The experimental results show that the h value is 2~3 for each cluster.

As shown in the figure, the fitting of the model improves (ICL criterion converges to 0) as the number of clusters increase, but we observe that

convergence speed of ICL criterion slows down based on k = 17. In this study, we set the number of reference clusters as k = 17 considering the sensitivity of the number of clusters and the characteristic of the inflection point. However, if more detailed traffic behavior analysis is required according to the purpose of the analysis, it can be analyzed by increasing the number of clusters (Briand et al, 2017).

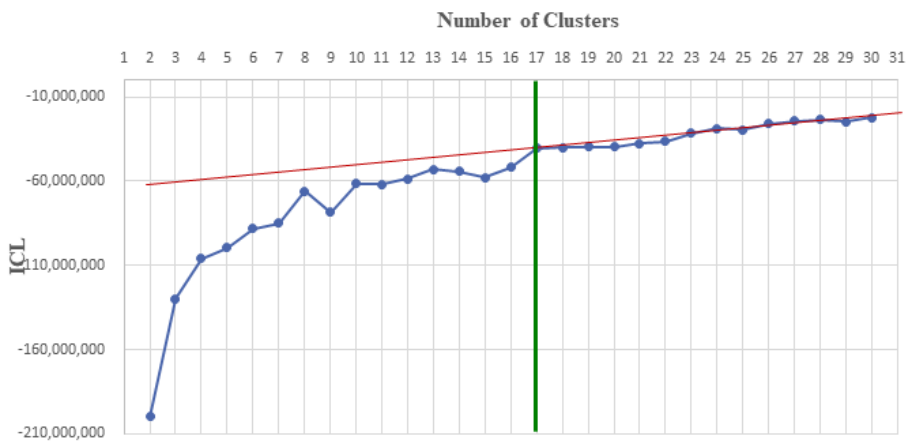


Figure 7- 8 ICL criterion with different numbers of clusters

These results had numerical implications, but did not have a significant impact in terms of destination estimation. Nevertheless, it is expected to be available for more detailed passenger analysis or other trip analyses other than the destination estimation.

7.3. Case study comparison

The methodology of this study was applied to other city other than Daejeon city to confirm the reliability of the methodology and to analyze the application characteristics. The comparative city is Sejong city, Korea, and Table 7-7 compares the main items of Daejeon city and Sejong city.

Table 7- 7 Basic information comparison between Daejeon city and Sejong city

Criteria	Daejeon (A)	Sejong (B)	Ratio (B/A)
Population	1,495,234	297,604	20%
Area	539.98 ㎢	464.90 ㎢	86%
Number of bus routes	114	148	130%
Number of bus stops	2,790	1,170	42%
Analysis period	April-May 2018	Same as left	-
Number of users (bus users)	1,666,903 (1,398,207)	246,496 (〃)	18% (7%)
Numbe of trips (bus trips)	32,116,974 (25,922,133)	2,253,339 (〃)	9% (13%)
Alighting tag rate of bus	50.7%	76.9%	152%

We estimate the estimation of Sejong city's smart card data by the same method applied to Daejeon city. As a result of the comparative analysis, the matching rate is low compared to the Daejeon. This means that the proportion of public transit is low, and that the percentage of continuous trip is low. And, the accuracy of the travel pattern method is higher than

that of Daejeon. The accuracy of the travel pattern method is relatively better. This means that the use of public transit in Sejong city is characterized by repetition. The effect of applying the travel pattern method is great compared to Daejeon city. Especially, when the combined method 2 is applied, the accuracy to all samples is greatly improved from 54.5% to 66.6%.

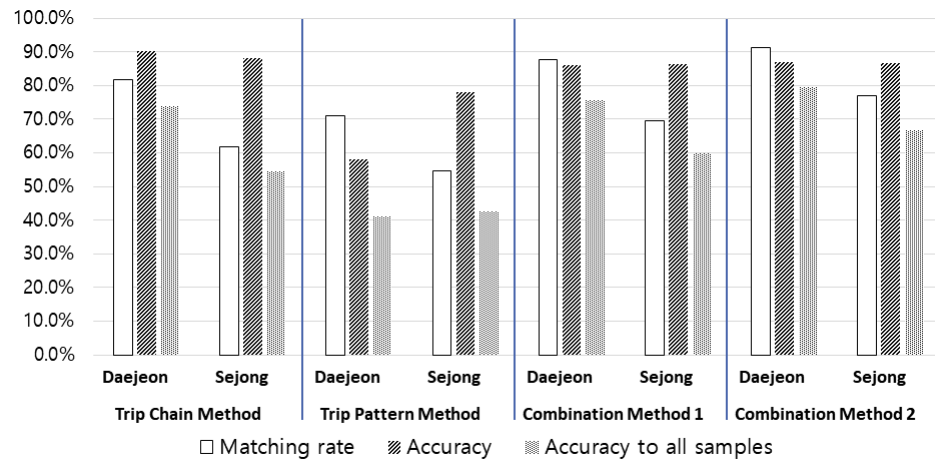


Figure 7- 9 Analysis results of Sejong city compared with Daejeon city

Chapter 8. Conclusion

8.1. Conclusion and contribution

8.1.1. Conclusion

This study verified the validity of the definition of the travel pattern model and the applicability of the model to the large smart card data. We have developed a travel pattern generation model that estimates the Gaussian mixture distribution by user cluster and cluster using the boarding time profile. In particular, it is easy to classify patterns for new observations (trips) according to the characteristics of the parametric method.

Also, we developed a method to estimate the alighting location using historical boarding records. Using the temporal reference information based on the travel pattern model and the spatial reference information from the historical boarding location records, the alighting location of individual trip was estimated. In this way, the possibility of obtaining the alighting location of "Unlinked trip" such as single trip is expanded.

We developed an algorithm to estimate the alighting location by applying a combination of the trip chain method and the travel pattern method. In this study, we applied the trip chain-based estimation first and then suggested the application of the travel pattern-based estimation to the trips that can not be estimated by the trip chain method.

As a result of Daejeon city case study, the matching rate and the accuracy were improved by 11.5% and 7.6%, respectively. In the case of Sejong City, which is the comparative case study, it was found that the

application of this study is more effective. The matching rate was improved by 24.4% and the accuracy was improved by 22.2%. The accuracy of the estimation by the travel pattern method varies depending on the city structural characteristics and repetitive characteristics of trips.

8.1.2. Contribution

The academic contribution of this study is as follows. First, we overcome the limitation of the trip chain methodology. This paper presents a method for estimating the location of a passenger on a pass that can not form a trip chain. In addition, we developed a methodology to utilize historical records that are not temporally adjacent. Also, we propose a combined methodology of existing trip chain method and travel pattern method.

Next, we developed a method for estimating the alighting from the travel pattern model. Unlike the existing rule-based trip chain methodology, the proposed method is based on the probability model. In addition, a combined methodology of rule-based method and probability-based methodology is presented.

Finally, the sensitivity analysis was performed to search the appropriate application conditions of the travel pattern based alighting location estimation. It is confirmed that the performance of the travel pattern-based estimation methodology is high in spite of the time-intensive trip occurrence time and spatially frequent use of the station. Also, it is confirmed that the estimation performance improves as the range of historical data used increases.

8.2. Limitation and future research

8.2.1. Limitation

The limitations of this study are as follows. First of all, it is possible to secure the performance of the alighting estimation only for the trips of the passengers who have sufficient historical trip records. According to the structure of the travel pattern model, it is necessary to secure a profile of the degree of intensive boarding time. Past public transit usage records are essential, and the more records you have, the more likely it is to improve performance. That is, there is a limit that the trip of a new passenger (without a history) is excluded from a travel pattern-based estimation object.

Next, for the trips with low travel pattern characteristics, the estimation performance is somewhat inferior. Experimental results show that the performance is good when the algorithm is applied to the travel pattern with strong trip characteristics, that is, the travel pattern with clear separation of boarding time distribution, while the performance is low when the travel pattern is opposite. There is still a limit to securing the performance of estimating the alighting location because a large percentage of passengers have weak patterns of trip characteristics.

Finally, there was a lack of research and analysis on the temporal transition of the distribution of travel patterns.

8.2.2. Future research

From the perspective of trip behavior, it is necessary to compare the characteristics of trips between passengers with and without alighting information. In this study, since the verification was conducted only for the trips that include the alighting tag, it is necessary to study the verification method considering the characteristics of the two groups in the future and to improve the algorithm accordingly.

Next, research on the temporal expansion of analytical data is needed. In this study, it was confirmed that the estimation performance improves according to the data enlargement within 2 months of data. In other words, the data extension shows that the possibility of unlinked trip can be improved. However, due to lack of data, we have not been able to study the critical point of performance enhancement due to data enlargement and to calculate the appropriate data size using it. We need to deal with this in the future.

There is a need to develop a method for estimating the alighting location for passengers who are new or have low trip frequency. To do this, it is necessary to examine the development of models and algorithms considering spatial characteristics and distribution of other users.

Finally, it is necessary to study the secondary clustering methodology of passengers with weak travel pattern characteristics and the transition between distribution patterns of travel patterns.

Reference

1. Abbas, O. A. (2008). Comparisons Between Data Clustering Algorithms. The International Arab Journal of Information Technology. 5(3), pp. 320-325.
2. Agard, B., Morency, C., Trépanier, M. (2006). Mining public transport user behaviour from smart card data. IFAC Proceedings Volumes. 39(3), pp. 399-404.
3. Aghabozorgi, S., Shirkhorshidi, A. S., Wah, T. Y. (2015). Time-seriesclustering – A decadereview. Information Systems. 53, pp. 16-38.
4. Alsger, A., Assemi, B., Mesbah, M., Ferreira, L. (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. Transportation Research Part C. 68, pp. 490-506.
5. Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. Transportation Research Part C. 87, pp. 123-137.
6. Barry, J. J., Newhouser, R., Rahbee, A., Sayeda, S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. Transportation Research Record. 1817, pp. 183-187.
7. Benmouiza, K., Cheknane, A. (2013). Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. Energy Conversion and Management. 75, pp. 561-569.
8. Biernacki, C., Govaert, G. (1997) Using the classification likelihood to choose the number of clusters. in: proceedings of The second World Conference of the International Association for Statistical Computing, Pasadena, USA.

9. Biernacki, C., Celeux, G., Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22(7), pp. 719 - 725.
10. Briand, A. S., Côme, E., El Mahrsi, M. K., Oukhellou, L. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. in: proceedings of 2015 IEEE International Conference on Data Science and Advanced Analytics. Paris, France. 1, pp. 37-50.
11. Briand, A. S., Côme, E., Trépanier, M., Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C. 79, pp. 274-289.
12. Enders, C. K., Bandalos D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling. 8(3), pp. 430-457.
13. El Mahrsi, M. K., Côme, E., Oukhellou, L., Verleysen, M. (2017). Clustering Smart Card Data for Urban Mobility Analysis. IEEE Transactions on Intelligent Transportation Systems. 18(3), pp. 712-728.
14. El Mahrsi, M. K., Côme, E., Baro, J., Oukhellou, L. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. in: proceedings of ACM SIGKDD Workshop on Urban Computing.
15. Fan, W., Chen, Z. (2018). Estimation of Origin-Destination Matrix and Identification of User Activities Using Public Transit Smart Card Data. Center for Advanced Multimodal Mobility Solutions and Education.
16. Farzin, J. M., (2008). Constructing an Automated Bus Origin - Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil. Transportation research record. 2072, pp. 30-37.

17. Fu T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence. 24, pp. 164-181.
18. Golob T. F. (2000). A simultaneous model of household activity participation and trip chain generation. Transportation Research Part B. 34, pp. 355-376.
19. Goulet-Langlois, G., Koutsopoulos, H. N., Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. Transportation Research Part C. 64, pp. 1-16.
20. Hamdan, H., Wu, J., (2013). Model selection with BIC and ICL criteria for binned data clustering by bin-EM-CEM algorithms. in: proceedings of 2013 IEEE International Conference on Systems, Man, and Cybernetics. Manchester, UK.
21. Han, G., Sohn, K. (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. Transportation Research Part B. 83, pp. 121-135.
22. He, L., Trépanier, M. (2015). Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data. Transportation Research Record. 2535, pp. 97-104.
23. He, L., Nassir, N., Trépanier, M., Hickman, M. (2015). Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems. CIRRELT-2015-52.
24. He, L., Agard, B., Trépanier, M. (2018). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. Transportmetrica A: Transport Science.
25. Hong, S. P., Min, Y. H., Park, M. J., Kim, K. M., Oh, S. M. (2016). Precise estimation of connections of metro passengers from Smart Card data. Transportation. 43(5), pp. 749-769.

26. Hörcher, D., Graham, D. J., Anderson, R. J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. Transportation Research Part B. 95, pp. 105-125.
27. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 31, pp. 651-666.
28. Jung, J. Y., Sohn, K. M. (2017). Deep-learning architecture to forecast destinations of bus passengers from entryonly smart-card data. IET Intelligent Transport Systems. 11(6), pp. 334-339.
29. Kim, K. T., Lee, I. M., (2017). Public Transportation Alighting Estimation Method Using Smart Card Data. Journal of the Korean Society for Railway. 20(5), pp. 692-702.
30. Kinnunen, T., Sidoroff, I., Tuononen, M., Fränti, P. (2011). Comparison of clustering methods: A case study of text-independent speaker modeling. Pattern Recognition Letters. 32, pp. 1604-1617.
31. Kou, G., Peng, Y., Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Information Sciences. 275, pp. 1-12.
32. Kusakabe, T., Asakura, Y., (2014). Behavioural data mining of transit smart card data: A data fusion approach. Transportation Research Part C. 46, pp. 179-191.
33. Lai, H. P., Visani M., Boucher, A., Ogier, J. M. (2012). An experimental comparison of clustering methods for content-based indexing of large image databases. Pattern Anal Applic. 15, pp. 345-366.
34. Li, T., Sun, D., Jing P., Yang, K. (2018). Smart Card Data Mining of Public Transport Destination: A Literature Review. Information. 9(1), pp. 18-21.
35. Ma, X., Wu, YJ., Wang, Y., Chen, F., Liu, J. (2013). Mining smart card data for transit passengers' travel patterns. Transportation Research

Part C. 36, pp. 1-12.

36. Ma, X., Liu, C., Wen, H., Wang, Y., Wu, Y. J. (2017). Understanding commuting patterns using transit smart card data. Journal of Transport Geography. 58, pp. 135-145.
37. Morency, C., Trépanier, M., Agard, B. (2007). Measuring transit use variability with smart-card data. Transport Policy. 14(3), pp. 193-203.
38. Munizaga, M., Devillaine, F., Navarrete, C., Silva, D. (2014). Validating travel behavior estimated from smart card data. Transportation Research Part C: Emerging Technologies. 44, pp. 70-79.
39. Munizaga, M., Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin – Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C. 24, pp. 9-18.
40. Nigam, K., Mccallum, A. K., Thrun, S., Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning. 39, pp. 103-134.
41. Nunes, A. A., Dias, T. G., Cunha, J. F. (2016). Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. IEEE Transactions on Intelligent Transportation Systems. 17(1), pp. 133-142.
42. Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy. 87, pp. 3538-3545.
43. Rish, I., (2001). An empirical study of the naive Bayes classifier. in: proceedings of IJCAI 2001 Workshop on Empirical Methods in AI. Seattle, USA. pp. 41-46.
44. Steele, R. J., Raftery, A. E. (2009). Performance of Bayesian Model

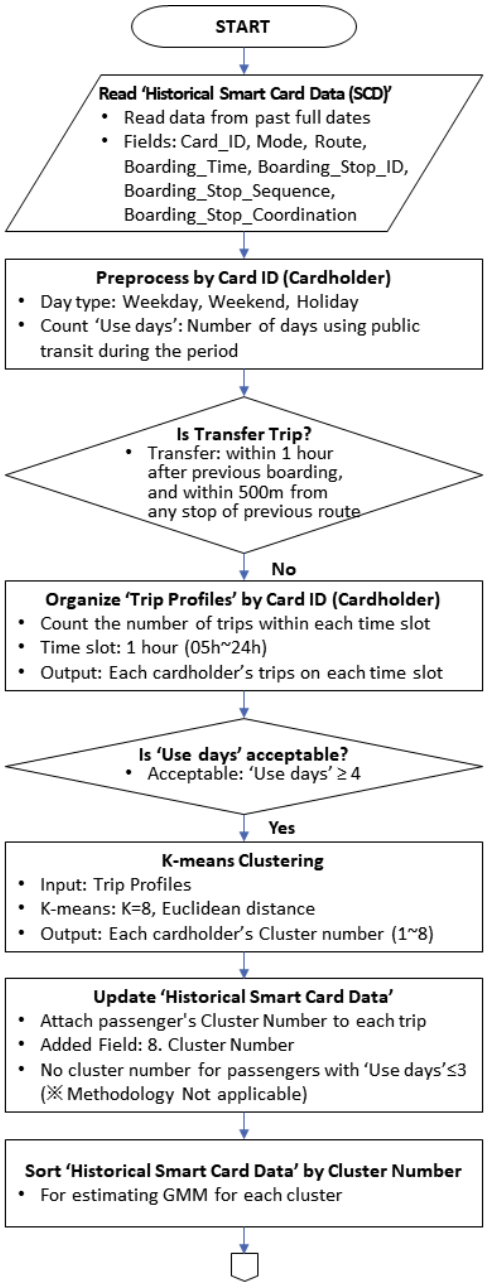
Selection Criteria for Gaussian Mixture Models.

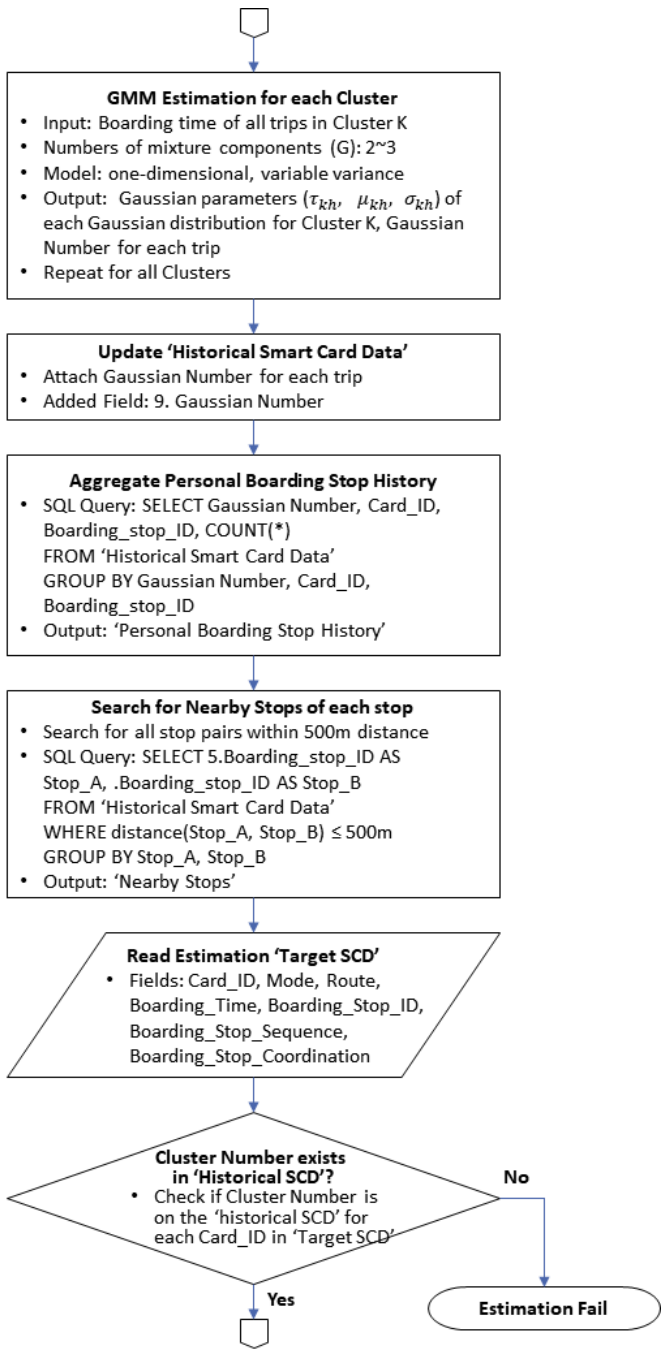
45. Tao, S., Rohde, D., Corcoran, J. (2014). Examining the spatial‐temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. Journal of Transport Geography. 41, pp. 21-36.
46. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. 22(19), pp. 2405-2412.
47. Trépanier, M., Tranchant, N., Chapleau, R., (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. Journal of Intelligent Transportation Systems. 11(1), pp.1-14.
48. Wang, H., Wang, Wei., Yang, J., Yu, P. S. (2002). Clustering by Pattern Similarity in Large Data Sets. in: proceedings of the 2002 ACM SIGMOD international conference on Management of data. Madison, USA. pp. 394-405.
49. Wang, W., Attanucci, J. P., Wilson, N. H. M. (2011). Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems. Journal of Public Transportation. 14(4), pp. 131-150.
50. Wen, L., Zhou, K., Yang, S. (2019). A shape-based clustering method for pattern recognition of residential electricity consumption. Journal of Cleaner Production. 212, pp. 475-488.
51. Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., Sekhar, C., Tham, K. W. (2017). k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings. 146, pp. 27-37.
52. Yuan, N. J., Wang, Y., Zhang, F., Xie, X., Sun, G. (2013).

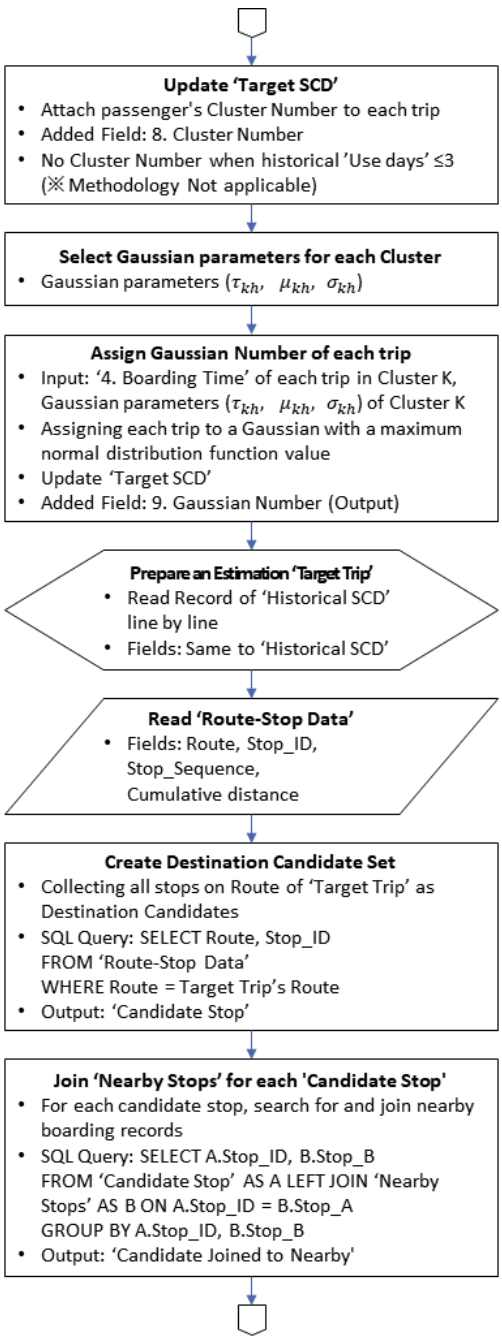
Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach. in: proceedings of 2013 IEEE International Conference on Data Mining. Karlsruhe, Germany. pp. 877-886.

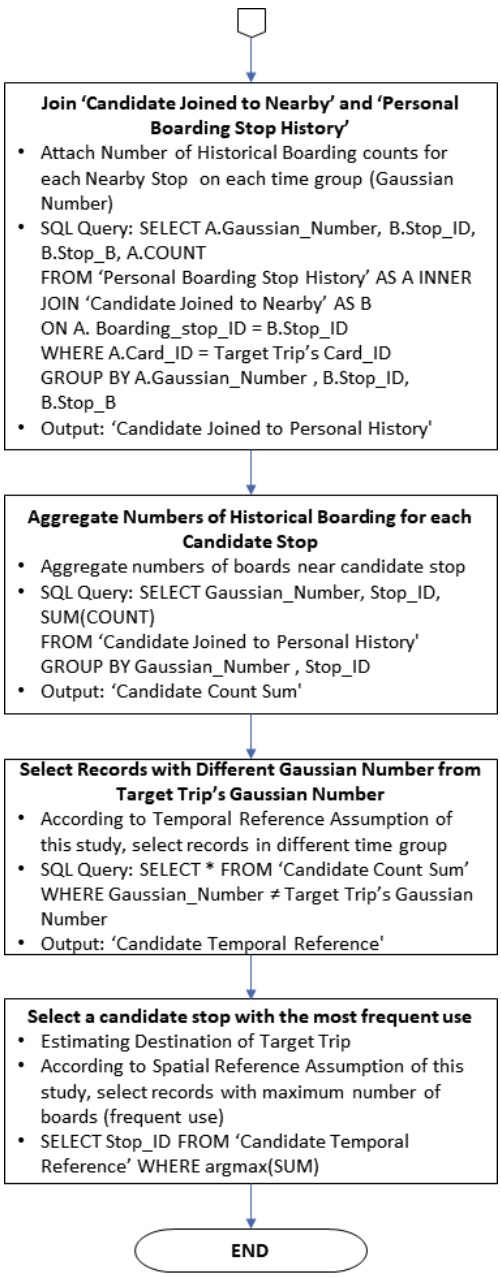
53. Zhang, F., Yuan, N. J., Wang, Y., Xie, X., (2015). Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach. Knowledge and Information Systems. 44, pp. 299-323.
54. Zhao, Z., Koutsopoulos, H. N., Zhao, J. (2018). Individual mobility prediction using transit smart card data. Transportation Research Part C. 89, pp. 19-34.

Appendix 1. Detailed flowchart of the algorithm




GMM Estimation for each Cluster
• Input: Boarding time of all trips in Cluster K
• Numbers of mixture components (G): 2~3
• Model: one-dimensional, variable variance
• Output: Gaussian parameters (τ_{kh}, μ_{kh}, σ_{kh}) of each Gaussian distribution for Cluster K, Gaussian Number for each trip
• Repeat for all Clusters
Update 'Historical Smart Card Data'
• Attach Gaussian Number for each trip
• Added Field: 9. Gaussian Number
Aggregate Personal Boarding Stop History
• SQL Query: SELECT Gaussian Number, Card_ID, Boarding_stop_ID, COUNT(*) FROM 'Historical Smart Card Data' GROUP BY Gaussian Number, Card_ID, Boarding_stop_ID
• Output: 'Personal Boarding Stop History'
Search for Nearby Stops of each stop
• Search for all stop pairs within 500m distance
• SQL Query: SELECT 5.Boarding_stop_ID AS Stop_A, .Boarding_stop_ID AS Stop_B FROM 'Historical Smart Card Data' WHERE distance(Stop_A, Stop_B) ≤ 500m GROUP BY Stop_A, Stop_B
• Output: 'Nearby Stops'
Read Estimation 'Target SCD'
• Fields: Card_ID, Mode, Route, Boarding_Time, Boarding_Stop_ID, Boarding_Stop_Sequence, Boarding_Stop_Coordination
Cluster Number exists in 'Historical SCD'?
• Check if Cluster Number is on the 'historical SCD' for each Card_ID in 'Target SCD'
No
Yes
Estimation Fail


Update 'Target SCD'
• Attach passenger's Cluster Number to each trip
• Added Field: 8. Cluster Number
• No Cluster Number when historical 'Use days' ≤3 (※ Methodology Not applicable)
Select Gaussian parameters for each Cluster
• Gaussian parameters (τ_{kh}, μ_{kh}, σ_{kh})
Assign Gaussian Number of each trip
• Input: '4. Boarding Time' of each trip in Cluster K, Gaussian parameters (τ_{kh}, μ_{kh}, σ_{kh}) of Cluster K
• Assigning each trip to a Gaussian with a maximum normal distribution function value
• Update 'Target SCD'
• Added Field: 9. Gaussian Number (Output)
Prepare an Estimation 'Target Trip'
• Read Record of 'Historical SCD' line by line
• Fields: Same to 'Historical SCD'
Read 'Route-Stop Data'
• Fields: Route, Stop_ID, Stop_Sequence, Cumulative distance
Create Destination Candidate Set
• Collecting all stops on Route of 'Target Trip' as Destination Candidates
• SQL Query: SELECT Route, Stop_ID FROM 'Route-Stop Data' WHERE Route = Target Trip's Route
• Output: 'Candidate Stop'
Join 'Nearby Stops' for each 'Candidate Stop'
• For each candidate stop, search for and join nearby boarding records
• SQL Query: SELECT A.Stop_ID, B.Stop_B FROM 'Candidate Stop' AS A LEFT JOIN 'Nearby Stops' AS B ON A.Stop_ID = B.Stop_A GROUP BY A.Stop_ID, B.Stop_B
• Output: 'Candidate Joined to Nearby'


Join 'Candidate Joined to Nearby' and 'Personal Boarding Stop History'
• Attach Number of Historical Boarding counts for each Nearby Stop on each time group (Gaussian Number)
• SQL Query: SELECT A.Gaussian_Number, B.Stop_ID, B.Stop_B, A.COUNT
FROM 'Personal Boarding Stop History' AS A INNER JOIN 'Candidate Joined to Nearby' AS B
ON A. Boarding_stop_ID = B.Stop_ID
WHERE A.Card_ID = Target Trip's Card_ID
GROUP BY A.Gaussian_Number , B.Stop_ID, B.Stop_B
• Output: 'Candidate Joined to Personal History'
Aggregate Numbers of Historical Boarding for each Candidate Stop
• Aggregate numbers of boards near candidate stop
• SQL Query: SELECT Gaussian_Number, Stop_ID, SUM(COUNT)
FROM 'Candidate Joined to Personal History'
GROUP BY Gaussian_Number , Stop_ID
• Output: 'Candidate Count Sum'
Select Records with Different Gaussian Number from Target Trip's Gaussian Number
• According to Temporal Reference Assumption of this study, select records in different time group
• SQL Query: SELECT * FROM 'Candidate Count Sum' WHERE Gaussian_Number ≠ Target Trip's Gaussian Number
• Output: 'Candidate Temporal Reference'
Select a candidate stop with the most frequent use
• Estimating Destination of Target Trip
• According to Spatial Reference Assumption of this study, select records with maximum number of boards (frequent use)
• SELECT Stop_ID FROM 'Candidate Temporal Reference' WHERE argmax(SUM)
END

Appendix 2. Case study of Sejong city

The results of the estimation of the alighting position of Sejong city are as follows.

Table. Estimation result by trip chain method (Sejong city)

Method	Trip chain phase 1	Trip chain phase 2	Trip chain phase 3
Sample Size (Estimation Target)	119,821	119,821	119,821
Matched Samples (Matching Rate)	45222 (37.7%)	71,164 (59.4%)	74,005 (61.8%)
Samples Failed to Match	74,599	48,657	45,816
Accuracy (Tight Criterion)	32,088 (71.0%)	47,927 (67.3%)	49,107 (66.4%)
Accuracy (Relaxing Criterion)	41,200 (91.9%)	63,527 (89.3%)	65,253 (88.2%)
Accuracy to All Samples (Relaxing Criterion)	34.4%	53.0%	54.5%

As a result of the application of the trip chain method, the estimation rate for the total 119,821 trips was measured as 61.8%. The relaxed accuracy and the effective estimation rate were 88.2% and 54.5%, respectively.

Next, it is the estimation result based on the travel pattern. The estimation rate for the total 119,821 trips was measured as 54.6%. The relaxed accuracy and the effective estimation rate were 78.0% and 42.6%, respectively.

Table. Estimation result by travel pattern method (Sejong city)

Method	**Trip chain phase 3**	**travel pattern**
Sample Size (Estimation Target)	119,821	119,821
Matched Samples (Matching Rate)	74,005 (61.8%)	65,376 (54.6%)
Samples Failed to Match	45,816	54,445
Accuracy (Tight Criterion)	49,107 (66.4%)	32,416 (49.6%)
Accuracy (Relaxing Criterion)	65,253 (88.2%)	51.025 (78.0%)
Accuracy to All Samples (Relaxing Criterion)	54.5%	42.6%

In the next step, the travel pattern method was applied in combination with the trip chain method. First, we estimated the alighting location according to the trip chain method phase 1, and applied a travel pattern to the failed trips. In this case, the effective estimation rate was 60.0%. This combination method is slightly superior to the case of using only the trip chain method.

Table. Estimation result by combination method 1 (Sejong city)

Method	**Trip chain phase 3**	**Combination method 1 (Chain Phase 1 + Pattern)**
Sample Size (Estimation Target)	119,821	119,821
Matched Samples (Matching Rate)	74,005 (61.8%)	83,217 (69.5%)
Samples Failed to Match	45,816	36,604
Accuracy (Tight Criterion)	49,107 (66.4%)	51,106 (61.4%)
Accuracy (Relaxing Criterion)	65,253 (88.2%)	71,840 (86.3%)
Accuracy to All Samples (Relaxing Criterion)	54.5%	60.0%

And, we estimated the alighting location according to the trip chain method phase 2, and applied a travel pattern to the failed trips. In this case, the effective estimation rate was 66.6%. This mixing method is superior to the case of using only the trip chain method.

Table. Estimation result by combination method 2 (Sejong city)

Method	**Trip chain phase 3**	**Combination method 2 (Chain Phase 2 + Pattern)**
Sample Size (Estimation Target)	119,821	119,821
Matched Samples (Matching Rate)	74,005 (61.8%)	92,099 (76.9%)
Samples Failed to Match	45,816	27,722
Accuracy (Tight Criterion)	49,107 (66.4%)	57,983 (63.0%)
Accuracy (Relaxing Criterion)	65,253 (88.2%)	79,760 (86.6%)
Accuracy to All Samples (Relaxing Criterion)	54.5%	66.6%

요약(국문초록)

교통카드데이터의 시간적 통행패턴을 활용한 버스 통행 목적지 추정

교통량 조사, 설문 조사 및 통행기록 등 기존의 교통 데이터와 비교할 때에, 교통카드데이터는 다음의 이점이 있다. 우선, 모든 통행 이력이 통행자(카드소지자)별로 시스템에 기록이 되므로 장기간의 자료수집이 가능하다. 기존의 조사기반 자료의 소규모 표본과 대비된다. 둘째, 각 통행기록에 정확한 시간적, 공간적(좌표 등) 정보가 기록되므로, 시공간적 분석에 유리하다. 셋째, 각 통행기록이 고유의 번호(암호화된 카드번호)를 가지고 있으므로, 이를 이용하여 통행기록을 추적하거나 장기간의 종점인 연구를 수행할 수 있다.

하지만, 대다수 교통카드데이터에는 출발지(승차지점)에 대한 정보는 포함되나, 목적지(하차지점) 정보는 포함되지 않는다. 또한, 대중교통의 과금 목적이 자료이므로, 통행 목적에 대한 정보 역시 포함되지 않는다. 이러한 정보의 제약은 대중교통 통행행태 및 수요 분포의 추정에 교통카드데이터를 활용하는 데에 장애 요소이다.

이 연구는 교통카드데이터의 대표적인 누락 정보인 목적지(하차지점) 정보의 추정 가능성 및 정확성을 향상하는 것을 목적으로 한다. 이와 관련하여, 기존에는 통행사슬 방법을 중심으로 목적지를 추정하는 연구가 다수 수행된 바 있다. 하지만, 여전히 'unlinked trip' 등 추정의 근거가 부족한 통행의 목적지 추정문제가 남아 있다. 이 문제의 해결을 위하여, 본 연구는 과거 여러 날짜의 교통카드데이터로부터 추출한 시간적 통행패턴 정보를 목적지 추정에 활용하였다.

일반적으로 승하차 개찰구 출입 정보를 포함하는 도시철도보다는 목적지 추정의 필요성이 높은 버스의 통행을 대상으로 연구를 수행하였다. 특히, 본 연구는 교통카드데이터의 개별 통행에 대해 목적지 정보를 추정하였다. 추정의 대상이 개별 버스통행의 목적지 정보이지만, 추정에 필요한 참조 정보로서 도시철도 승차기록 자료를 활용하였다.

본 연구의 목적을 달성하기 위하여, 본 연구는 과거 여러 날짜의 교통카드데이터를 이용하여 대상 통행의 목적지를 추정하는 방법론을 제안하였다. 특히, 과거 여러 날짜의 교통카드데이터로부터 통행패턴 정보를 생성하고, 이를 활용하여 목적지(하차지점)를 추정하는 통행패턴 기반의 방법론을 제안하였다.

본 연구는 교통카드데이터의 목적지 추정을 위한 모형을 구축하고, 이에 따른 추정 알고리즘을 개발하였다. 우선, 통행패턴 생성을 위한 모형으로써, 고차원, 고용량 데이터의 군집화에 적합한 k-means 알고리즘 및 혼합분포의 추정을 위한 가우시안 혼합 모형(Gaussian Mixture Model)을 적용하였다. 다음으로, 가우시안 혼합 모형 형태의 통행패턴에 적합한 통행기록 기반의 목적지 추정 알고리즘을 제시하였다.

이 연구는 통행패턴 모델의 타당성과 이 모델이 고용량 교통카드데이터에 적합함을 검증하였다. 특히, 모수적 방법인 가우시안 혼합 모형을 적용함으로써 고용량 교통카드데이터에 대한 통행패턴의 분류에 적합하도록 모형을 개발하였다.

과거의 승차기록을 이용하여 시간적 통행패턴 정보를 생성하여 활용하는 한편, 공간적 패턴인 과거의 승차 위치 기록을 참조하여 개별 통행의 목적지를 확률적으로 추정하였다. 이 연구의 방법에 따라, ‘single trip’과 같은 ‘unlinked trip’의 목적지 추정의 가능성이 향상되었다.

또한, 기존의 통행사슬 방법에 본 연구에서 개발한 통행패턴 방법을 조합하여 목적지를 추정하는 혼합 알고리즘을 개발하였다. 일차적으로 통행사슬 방법을 적용한 후, 통행사슬 방법으로 목적지를 추정할 수 없는 통행에 대하여 통행패턴 방법을 적용하였다. 이에 따라 기존의 통행

사슬 방법 대비 추정 성공률 및 정확성이 개선되었음을 확인하였다.

개발된 모형 및 알고리즘은 대전시의 교통카드데이터를 활용하여 분석하고 검증하였다. 또한, 실용적 관점에서 통행패턴 카테고리 분석, 변수 및 조건에 따른 민감도 분석결과를 제시하였다.

주 요 어: 교통카드데이터, AFC 데이터, 목적지 추정, OD 추정, 하차 정류장, 통행패턴, 통행자 군집화, 과거 통행기록

학 번: 2009-30943