A Survey of Big Data Analytics in ITS: Passenger O-D Matrices Estimation

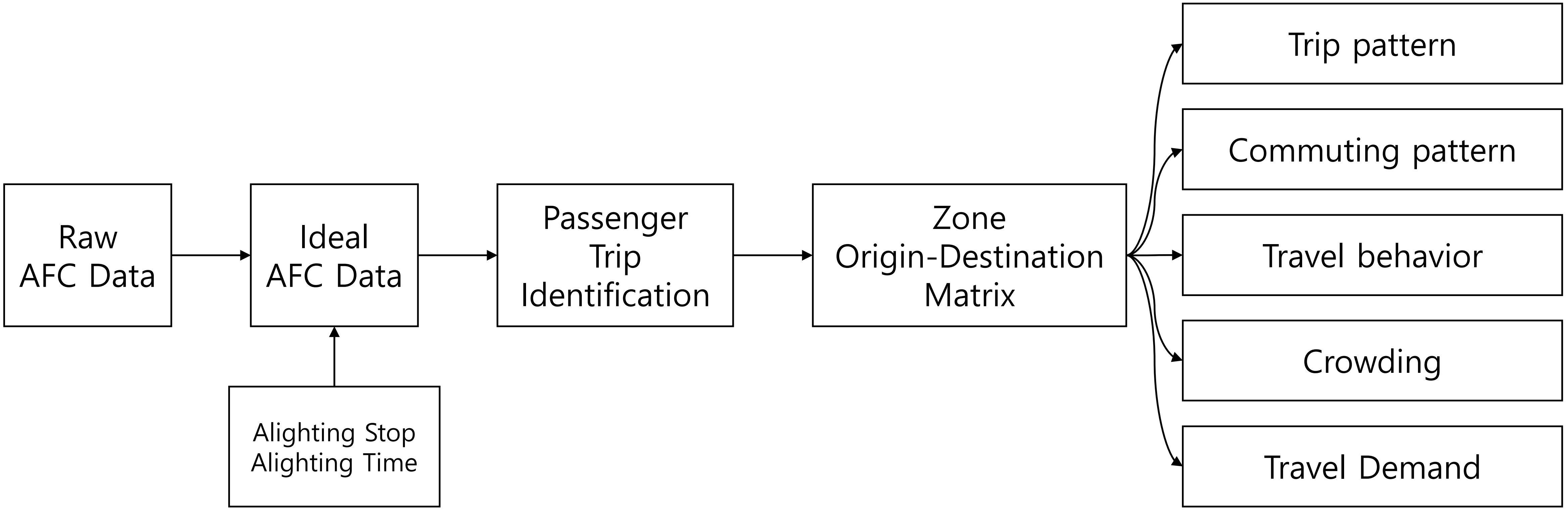
지능형 교통 시스템의 빅데이터 분석 기술 조사: 승객 기종점 행렬 추정

Sanghyuck Lee
tkdgur658@gmail.com
2020. 10. 20.

Introduction

- Importance of Intelligent Transportation Systems (ITS)
  - ITS aims to integrate information, communications, computers and other technologies and apply them to the transportation domain to have a large, full-functioning, accurate and efficient transport management system [1].
- From AFC data To various Application
  - The AFC data can be used in various fields such as analysis of travel patterns, commuting patterns, travel behavior, crowding, travel demand [2].
- Passenger Destination Estimation
  - Passenger destination estimation is the most basic problem to be solved before advancing to various applications based on AFC data [3].

Introduction: ITS Process using AFC data



Introduction: AFC data

- Some studies for available dataset; basically AFC data
    - Trip pattern: understand how passengers' travel patterns vary temporally and spatially[4, 5].
    - Commuting pattern: identifying commuting trips, because commuting significantly impacts urban traffic congestion and emission [6].
    - Travel behavior: optimize transit vehicles, waiting time [7]
    - Passenger OD: estimate public transport origin-destination matrix [8]
    - Crowding: in vehicle, estimate crowding costs [9]
    - Travel demand: e.g. ridership prediction [10]

Dataset	Available
Automated Fare Collection (AFC)	O
Geographic Information System (GIS)	O
Global Positioning System (GPS)	X
Video	X
CAV and VANET	X
Sensor	X

Available data and not available data

Introduction: Passenger destination estimation

- Passenger destination estimation algorithm is a core algorithm for advancing most application based on AFC data
  - Input: AFC data
    - CARD ID
    - BOARDING STOP
    - BOARDING TIME
    - Alighting Stop (Optional)
    - Alighting Time (Optional)

Transaction Number	Boarding Time (yyyy/mm/dd/hh/mm/ss)	ID of Boarding Location	Alighting Time (yyyy/mm/dd/hh/mm/ss)	ID of Alighting Location
1	20150318092455	4196066	20150318092714	4196117
2	20150318214734	4196061	20150318215035	4196065
3	20150318225205	12174	20150318225855	8001017
4	20150318075726	70647	20150318082805	10585
5	20150318070948	216	20150318084718	4101317
...	...	...	...	...

AFC data including alighting information

Card ID	Card style	Route	Stop	Veh. ID	Time	Date	Fare
-1760518818	Monthly card	115	156	2798	06:27:16	2009-07-07	-
-1760518818	Monthly card	115	3	2346	18:47:21	2009-07-07	-
-1760517522	Common card	115	19	2810	08:34:53	2009-07-07	0.9
-1760517522	Common card	115	151	2808	17:47:31	2009-07-07	0.9
-1760515746	Monthly card	115	133	2794	11:56:22	2009-07-07	-
-1760515746	Monthly card	115	133	2814	18:52:29	2009-07-07	-
-1760515714	Monthly card	115	23	2794	17:32:51	2009-07-07	-

AFC data NOT including alighting information

Introduction: Passenger destination estimation

- There are various ITS applications based on the AFC system [4–13, 16].
- However, destination information is missing in many AFC systems.
- Therefore, passenger destination estimation algorithm enables more accurate and efficient AFC system applications [16].

Transaction Number	Boarding Time (yyyy/mm/dd/hh/mm/ss)	ID of Boarding Location	Alighting Time (yyyy/mm/dd/hh/mm/ss)	ID of Alighting Location
1	20150318092455	4196066	20150318092714	4196117
2	20150318214734	4196061	20150318215035	4196065
3	20150318225205	12174	20150318225855	8001017
4	20150318075726	70647	20150318082805	10585
5	20150318070948	216	20150318084718	4101317
…	…	…	…	…

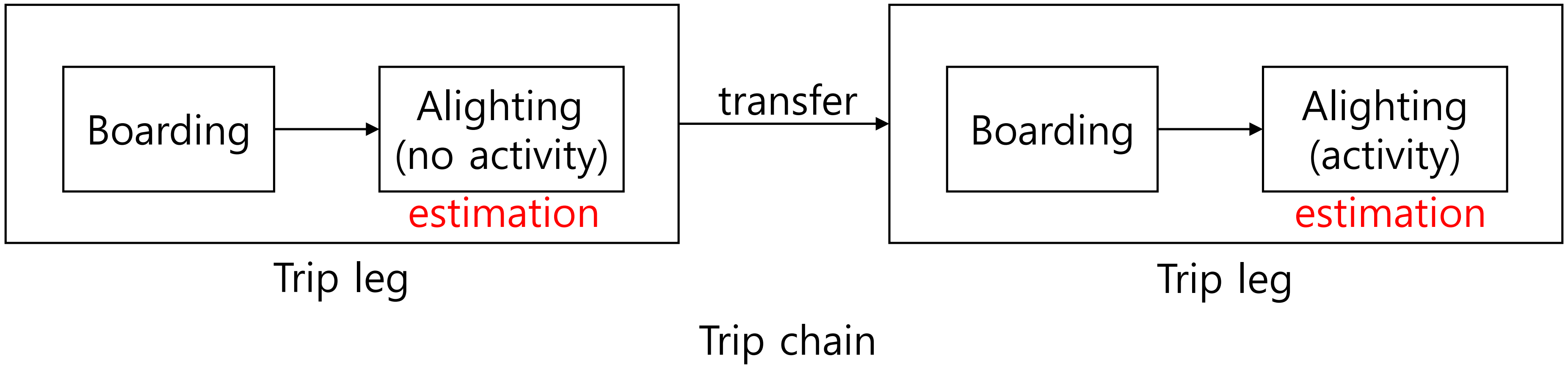
AFC data including alighting information

Card ID	Card style	Route	Stop	Veh. ID	Time	Date	Fare
-1760518818	Monthly card	115	156	2798	06:27:16	2009-07-07	-
-1760518818	Monthly card	115	3	2346	18:47:21	2009-07-07	-
-1760517522	Common card	115	19	2810	08:34:53	2009-07-07	0.9
-1760517522	Common card	115	151	2808	17:47:31	2009-07-07	0.9
-1760515746	Monthly card	115	133	2794	11:56:22	2009-07-07	-
-1760515746	Monthly card	115	133	2814	18:52:29	2009-07-07	-
-1760515714	Monthly card	115	23	2794	17:32:51	2009-07-07	-

AFC data NOT including alighting information

Introduction: Passenger destination estimation

- In many studies related to passenger destination estimation, a "trip" is referred to as either an entire journey of the passenger (i.e. trip chain) to the destination in which he/she wants to perform an activity or a single transaction of a passenger (i.e. trip leg) [8, 12, 80-98].
- The "destination" to be estimated means the alighting point of individual trip leg.



Related work - 1st paragraph (Overall)

Main Category	Sub Category	Keywords
Collection	Smart card	trip pattern [11], O-D matrix [12], transit demand modelling [13]
	GPS	travel mode detection [14], traffic noise mapping [15]
	Video	real-time application [16]
	Sensor	real-time application [17]
	CAV & VANET*	real-time query service [18], scheduling of CAV [19], management in CAV environment [20], anticipation and alert system in VANET [21]
	Passive collection	inferring trip purposes(taxi) [22], predicting stay time of mobile users(contextual information) [23], predicting mobility patterns(social network) [24], sustainable transportation(social network) [25], smart grid [26], Communication Based Train Control(CBTC) [27]
Analytic methods	Regression	short-term forecasting [28], speed estimation [29], trip generation evaluation [30]
	Decision tree	incident detection [31], accident severity analysis [32]
	Neural network	mode choice [33], short-term flow prediction [34], travel time prediction [35], incident detection [36], remaining parking space forecasting [37]
	SVM	travel time prediction [38], arrival time prediction [39], incident detection [40]

* Connected and Autonomous Vehicles & Vehicle Ad Hoc Network

Related work - 1st paragraph (Overall)

Main Category	Sub Category	Keywords
Analytic methods	Unsupervised learning	macroscopic planning [41], travel time prediction [42]
	Reinforcement learning	signal control [43]
	Deep learning	congestion evolution prediction [44], intelligent fault diagnosis of train [45], traffic data imputation [46], flow prediction[47], speed prediction [48]
	Ontology based	automatic traffic light settings [49], driver assistance [50]
Applications	Accidents analysis	freeway weather condition [51], space-air-ground [52], roadside features [53], run-off-roadway [54], rural road way [55], driver, crash and vehicle characteristics [56]
	Flow prediction	traffic flow [57], traffic speed [58], short-term [59], real-time [60], local traffic state [61]
	Public transportation	congested conditions [62], intermodal passenger [63], mobile phone data [64], behavior dynamics [65], O-D matrix [66], demand estimation [67], arrival information [68]
	Personal route	arriving time [69], driving routes [70], mobile data [71], vehicle GPS data [72], train, private cars or bicycles [73]
	Rail transportation	evaluating rail transit timetable [74], precise train stopping [75], stop control [76]
	Asset maintenance	Railways [77], ontology-driven [78], fault diagnosis [79]

Preliminary: from ideal AFC data to generate OD matrix

- Input ideal AFC data:

$CARD\ ID$	$BOARD\ STATION$	$BOARD\ TIME$	$ALIGHTING\ STATION$	$ALIGHTING\ TIME$
1	B_{11}	$T_{B_{11}}$	A_{11}	$T_{A_{11}}$
1	B_{12}	$T_{B_{12}}$	A_{12}	$T_{A_{12}}$
1	B_{13}	$T_{B_{13}}$	A_{13}	$T_{A_{13}}$
2	B_{21}	$T_{B_{21}}$	A_{21}	$T_{A_{21}}$

1) define $\text{Th}reshold\ of\ ITT = 30$
2) Select $CARD\ ID$, $CARD\ ID = 1$

Preliminary: from ideal AFC data to generate OD matrix

- Input data:

	$CARD\ ID$	$BOARD\ STATION$	$BOARD\ TIME$	$ALIGHTING\ STATION$	$ALIGHTING\ TIME$
initial	1	B_{11}	$T_{B_{11}}$	A_{11}	$T_{A_{11}}$
	1	B_{12}	$T_{B_{12}}$	A_{12}	$T_{A_{12}}$
	1	B_{13}	$T_{B_{13}}$	A_{13}	$T_{A_{13}}$
	2	B_{21}	$T_{B_{21}}$	A_{21}	$T_{A_{21}}$

3) Select transaction of current $CARD\ ID$
4) If current $BOARD\ TIME$ < other $BOARD\ TIME$ in same $CARD\ ID$ (i.e. $T_{B_{11}} < T_{B_{12}}$)

   Label the current transaction 'initial', select next transaction and go to 3)

   Else

   Go to 5)

Preliminary: from ideal AFC data to generate OD matrix

- Input data:

	$CARD\ ID$	$BOARD\ STATION$	$BOARD\ TIME$	$ALIGHTING\ STATION$	$ALIGHTING\ TIME$
initial	1	B_{11}	$T_{B_{11}}$	A_{11}	$T_{A_{11}}$
transfer	1	B_{12}	$T_{B_{12}}$	A_{12}	$T_{A_{12}}$
	1	B_{13}	$T_{B_{13}}$	A_{13}	$T_{A_{13}}$
	2	B_{21}	$T_{B_{21}}$	A_{21}	$T_{A_{21}}$

5) If ITT(in case of $T_{B_{12}}$, $T_{B_{13}} - T_{B_{12}}$) $< threshold\ of\ ITT$

    Label the current transaction 'transfer'

6) If there is no $CARD$ transaction that has more $BOARD\ TIME$ than current $BOARD\ TIME$ (in case of $T_{B_{12}}$, maintain 'transfer', but in case of $T_{B_{13}}$, label 'last')

    Label the current transaction 'last'

Preliminary: from ideal AFC data to generate OD matrix

- Input data:

	$CARD\ ID$	$BOARD\ STATION$	$BOARD\ TIME$	$ALIGHTING\ STATION$	$ALIGHTING\ TIME$
initial	1	B_{11}	$T_{B_{11}}$	A_{11}	$T_{A_{11}}$
transfer	1	B_{12}	$T_{B_{12}}$	A_{12}	$T_{A_{12}}$
last	1	B_{13}	$T_{B_{13}}$	A_{13}	$T_{A_{13}}$
initial	2	B_{21}	$T_{B_{21}}$	A_{21}	$T_{A_{21}}$

6) Repeat the same operation for CARD ID of 1, then we can generate trip chain consisting of transactions with a CARD ID of 1

   Trip chain of CARD ID of 1: $(\mathrm{B}_{11} -> \mathrm{A}_{11}) -> (\mathrm{B}_{12} -> \mathrm{A}_{12}) -> (\mathrm{B}_{13} -> \mathrm{A}_{13})$

7) Also, Repeat the same operation for other CARD ID

Preliminary: from ideal AFC data to generate OD matrix

	CARD ID	*BOARD STATION*	*BOARD TIME*	*ALIGHTING STATION*	*ALIGHTING TIME*
initial	1	B_{11}	$T_{B_{11}}$	A_{11}	$T_{A_{11}}$
transfer	1	B_{12}	$T_{B_{12}}$	A_{12}	$T_{A_{12}}$
last	1	B_{13}	$T_{B_{13}}$	A_{13}	$T_{A_{13}}$
initial	2	B_{21}	$T_{B_{21}}$	A_{21}	$T_{A_{21}}$

D \ O	$\cdots$	B_{11}	$\cdots$	B_{21}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_{13}	$\vdots$	1	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_{21}	$\vdots$	$\vdots$	$\vdots$	1	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

8) Then, we can make O-D matrix using not trip legs, but Trip chain; trip legs is movement with just one mode of transport, so it is hard to grasp socioeconomic transportation demand, but trip chain can.

For example, in case of trip chain of CARD ID of 1: $(\mathrm{B}_{11} \rightarrow \mathrm{A}_{11}) \rightarrow (\mathrm{B}_{12} \rightarrow \mathrm{A}_{12}) \rightarrow (\mathrm{B}_{13} \rightarrow \mathrm{A}_{13})$

Count B_{11} to A_{13} as one in O-D matrix.

Preliminary: from ideal AFC data to generate OD matrix

- However, in most cases, there are no alighting information.
- So, we have to estimate alighting station and time to generate a more accurate O-D matrix.

$CARD\ ID$	$BOARD\ STATION$	$BOARD\ TIME$	$ALIGHTING\ STATION$	$ALIGHTING\ TIME$
1	B_{11}	$T_{B_{11}}$	**NaN**	**NaN**
1	B_{12}	$T_{B_{12}}$	**NaN**	**NaN**
1	B_{13}	$T_{B_{13}}$	**NaN**	**NaN**
2	B_{21}	$T_{B_{21}}$	**NaN**	**NaN**

Related works

- There are some related works about passenger destination estimation based on AFC transaction
- Reference paper:
  - Title: Smart Card Data Mining of Public Transport Destination: A Literature Review [3]
  - Authors: T Li, D Sun, P Jing, K Yang
  - Journal: Information (ESCI)

Destination Estimation Model	Advantage	Disadvantage
Trip Chaining Model	• Only need smart card data • Simple algorithm • Infer alight station of each passenger	• Hard to validate the model with sufficient data
Probability Model	• More comprehensive considerations	• Only infer the total on-off passenger number
Deep learning Model	• Very comprehensive considerations • Infer alight station of each passenger • Validated by numerous travel data	• Need abundant data • Only appropriate for entry-exit system • Complex algorithm

Comparison of three models

Related works

1) Trip Chaining Model
   - Barry et al. (2002)'s assumptions:
     1) The alighting station of the current subway trip is the aboard station of next subway trip.
     2) Most riders end their last journey of the day at the station where they begin their first trip of the day.
   - J Zhao (2004)'s assumptions:
     1) There is no other mode of transportation (e.g., car, motorcycle, bicycle, etc.) between two consecutive trips.
     2) Travelers will not walk a long distance when transferring.
     3) Travelers will end their last trip of the day at the station where they begin their first trip of the day.
   - J Zhao (2004)'s assumptions were developed by other researchers.
     1) The first assumption has not yet improved. This is because personal means of transportation cannot be inferred by smart cards.
     2) Various papers have been written using walking time and walking distance as parameters [82-90].
     3) Several researchers have improved that the third assumption cannot infer the destination of a single trip [91, 92].

Related works

2) Probability Model

- Dou Huili et al. (2007) calculated the probability of getting off at each remaining stop by considering the travel distance and the number of passengers.
- This model was improved by adding factors of station transfer capacity and land use level around the station [94-96].

3) Deep Learning Model

- Yu et al. (2006) first used artificial neural network to estimate the OD matrix.
- Jung J and Sohn K (2017) developed a deep-learning model to estimate the alighting number utilizing smart-card data and land-use data.

References

[1] An, Sheng-hai, Byung-Hyug Lee, and Dong-Ryeol Shin. "A survey of intelligent transportation systems." 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks. IEEE, 2011.

[2] Zhu, Li, et al. "Big data analytics in intelligent transportation systems: A survey." IEEE Transactions on Intelligent Transportation Systems 20.1 (2018): 383-398.

[3] Li, Tian, et al. "Smart card data mining of public transport destination: A literature review." Information 9.1 (2018): 18.

[4] Kim, Mi-Kyeong, Sangpil Kim, and Hong-Gyoo Sohn. "Relationship between spatio-temporal travel patterns derived from smart-card data and local environmental characteristics of Seoul, Korea." Sustainability 10.3 (2018): 787.

[5] Ma, Xiaolei, et al. "Mining smart card data for transit riders' travel patterns." Transportation Research Part C: Emerging Technologies 36 (2013): 1-12.

[6] Ma, Xiaolei, et al. "Understanding commuting patterns using transit smart card data." Journal of Transport Geography 58 (2017): 135-145.

[7] Ali, Atizaz, Jooyoung Kim, and Seungjae Lee. "Travel behavior analysis using smart card data." KSCE Journal of Civil Engineering 20.4 (2016): 1532-1539.

[8] Li, Daming, et al. "Estimating a transit passenger trip origin-destination matrix using automatic fare collection system." International Conference on Database Systems for Advanced Applications. Springer, Berlin, Heidelberg, 2011.

[9] Hörcher, Daniel, Daniel J. Graham, and Richard J. Anderson. "Crowding cost estimation with large scale smart card and vehicle location data." Transportation Research Part B: Methodological 95 (2017): 105-125.

[10] Ding, Chuan, et al. "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees." Sustainability 8.11 (2016): 1100.

[11] Nishiuchi, Hiroaki, James King, and Tomoyuki Todoroki. "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data." International Journal of Intelligent Transportation Systems Research 11.1 (2013): 1-10.

[12] Munizaga, Marcela A., and Carolina Palma. "Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile." Transportation Research Part C: Emerging Technologies 24 (2012): 9-18.

References

[13] Chu, Ka Kee Alfred, and Robert Chapleau. "Enriching archived smart card transaction data for transit demand modeling." Transportation research record 2063.1 (2008): 63-72.

[14] Gong, Hongmian, et al. "A GPS/GIS method for travel mode detection in New York City." Computers, Environment and Urban Systems 36.2 (2012): 131-139.

[15] Asensio, C., et al. "GPS-based speed collection method for road traffic noise mapping." Transportation Research Part D: Transport and Environment 14.5 (2009): 360-366.

[16] Courage, K. G., et al. Video image detection for traffic surveillance and control. No. WPI 0510618,. 1996.

[17] Lopes, J., et al. "Traffic and mobility data collection for real-time applications." 13th International IEEE Conference on Intelligent Transportation Systems. IEEE, 2010.

[18] Wang, Xin, et al. "Improved rule installation for real-time query service in software-defined internet of vehicles." IEEE Transactions on Intelligent Transportation Systems 18.2 (2016): 225-235.

[19] Hu, Jiajun, et al. "Scheduling of connected autonomous vehicles on highway lanes." 2012 IEEE Global Communications Conference (GLOBECOM). IEEE, 2012.

[20] King, Ray. "Traffic management in a connected or autonomous vehicle environment." (2015): 6-20.

[21] Al Najada, Hamzah, and Imad Mahgoub. "Anticipation and alert system of congestion and accidents in VANET using Big Data analysis for Intelligent Transportation Systems." 2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2016.

[22] Gong, Li, et al. "Inferring trip purposes and uncovering travel patterns from taxi trajectory data." Cartography and Geographic Information Science 43.2 (2016): 103-114.

[23] Liu, Sen, et al. "Predicting stay time of mobile users with contextual information." IEEE Transactions on Automation Science and Engineering 10.4 (2013): 1026-1036.

[24] Alesiani, Francesco, Konstantinos Gkiotsalitis, and Roberto Baldessari. "A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data." Transportation research board: 93rd annual meeting. 2014.

References

[25] Chen, Ying, Andreas Frei, and Hani S. Mahmassani. "From personal attitudes to public opinion: Information diffusion in social networks toward sustainable transportation." Transportation Research Record 2430.1 (2014): 28-37.

[26] Grob, Gustav R. "Future transportation with smart grids & sustainable energy." 2009 6th International Multi-Conference on Systems, Signals and Devices. IEEE, 2009.

[27] Zhu, Li, et al. "Cross-layer handoff design in MIMO-enabled WLANs for communication-based train control (CBTC) systems." IEEE Journal on Selected Areas in Communications 30.4 (2012): 719-728.

[28] Zhu, Xiaobo, et al. Short Term Forecasting of Remaining Parking Spaces in Parking Guidance Systems. No. 16-5060. 2016.

[29] Shan, Zhenyu, Danna Zhao, and Yingjie Xia. "Urban road traffic speed estimation for missing probe vehicle data based on multiple linear regression model." 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). IEEE, 2013.

[30] Zenina, Nadezda, and Arkady Borisov. "Regression analysis for transport trip generation evaluation." Information Technology and Management Science 16.1 (2013): 89-94.

[31] Payne, Howard J., and Samuel C. Tignor. "Freeway incident-detection algorithms based on decision trees with states." Transportation Research Record 682 (1978).

[32] Abellán, Joaquín, Griselda López, and Juan De OñA. "Analysis of traffic accident severity using decision rules via decision trees." Expert Systems with Applications 40.15 (2013): 6047-6054.

[33] Xie, Chi, Jinyang Lu, and Emily Parkany. "Work travel mode choice modeling with data mining: decision trees and neural networks." Transportation Research Record 1854.1 (2003): 50-61.

[34] Vlahogianni, Eleni I., Matthew G. Karlaftis, and John C. Golias. "Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach." Transportation Research Part C: Emerging Technologies 13.3 (2005): 211-234.

[35] Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies 13.5-6 (2005): 347-369.

[36] Jin, Xin, Ruey Long Cheu, and Dipti Srinivasan. "Development and adaptation of constructive probabilistic neural network in freeway incident detection." Transportation Research Part C: Emerging Technologies 10.2 (2002): 121-147.

References

[37] Zhu, Xiaobo, et al. Short Term Forecasting of Remaining Parking Spaces in Parking Guidance Systems. No. 16-5060. 2016.

[38] Vanajakshi, Lelitha, and Laurence R. Rilett. "Support vector machine technique for the short term prediction of travel time." 2007 IEEE Intelligent Vehicles Symposium. IEEE, 2007.

[39] Bin, Yu, Yang Zhongzhen, and Yao Baozhen. "Bus arrival time prediction using support vector machines." Journal of Intelligent Transportation Systems 10.4 (2006): 151-158.

[40] Xiao, Jianli, and Yuncai Liu. "Traffic incident detection using multiple-kernel support vector machine." Transportation research record 2324.1 (2012): 44-52.

[41] Meng, Yan, and Xiyu Liu. "Application of K-means algorithm based on ant clustering algorithm in macroscopic planning of highway transportation hub." 2007 First IEEE International Symposium on Information Technologies and Applications in Education. IEEE, 2007.

[42] Nath, Rudra Pratap Deb, et al. "Modified K-means clustering for travel time prediction based on historical traffic data." International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer, Berlin, Heidelberg, 2010.

[43] Abdulhai, Baher, Rob Pringle, and Grigoris J. Karakoulas. "Reinforcement learning for true adaptive traffic signal control." Journal of Transportation Engineering 129.3 (2003): 278-285.[44] Ma, Xiaolei, et al. "Large-scale transportation network congestion evolution prediction using deep learning theory." PloS one 10.3 (2015): e0119044.[45] Hu, Hexuan, et al. "Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks." IEEE Transactions on Industrial Informatics 13.4 (2017): 2106-2116.

[44] N. Polson and V. Sokolov, Deep learning for short-term traffic flow prediction, 2016, [online] Available: https://arxiv.org/abs/1604.04527.

[45] H. Hu, B. Tang, X. Gong, W. Wei and H. Wang, "Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks", IEEE Trans. Ind. Informat., vol. 13, no. 4, pp. 2106-2116, Aug. 2017.

[46] Duan, Yanjie, et al. "A deep learning based approach for traffic data imputation." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2014.

[47] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." Transportation Research Part C: Emerging Technologies 79 (2017): 1-17.

[48] Ma, Xiaolei, et al. "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data." Transportation Research Part C: Emerging Technologies 54 (2015): 187-197.

References

[49] Fernandez, Susel, and Takayuki Ito. "Using SSN ontology for automatic traffic light settings on inteligent transportation systems." 2016 IEEE International Conference on Agents (ICA). IEEE, 2016.

[50] Zhao, Lihua, et al. "Ontologies for advanced driver assistance systems." Journal of the Japanese Society for Artificial Intelligence (JSAI), Technical Report, SIG-SWO-035-03. Available online: http://www. ei. sanken. osaka-u. ac. jp/sigswo/papers/SIG-SWO-035/SIG-SWO-035-03. pdf (accessed on 12 August 2016) (2015).

[51] Golob, Thomas F., and Wilfred W. Recker. "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions." Journal of transportation engineering 129.4 (2003): 342-353.

[52] Xiong, Gang, et al. "Novel ITS based on space-air-ground collected big-data." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2014.

[53] Lee, Jinsun, and Fred Mannering. "Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis." Accident Analysis & Prevention 34.2 (2002): 149-161.

[54] Lee, Jinsun, and Fred Mannering. "Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis." Accident Analysis & Prevention 34.2 (2002): 149-161.

[55] Karlaftis, Matthew G., and Ioannis Golias. "Effects of road geometry and traffic volumes on rural roadway accident rates." Accident Analysis & Prevention 34.3 (2002): 357-365.

[56] Bedard, Michel, et al. "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities." Accident Analysis & Prevention 34.6 (2002): 717-727.

[57] Li, Runmei, et al. "Traffic flow data forecasting based on interval type-2 fuzzy sets theory." IEEE/CAA Journal of Automatica Sinica 3.2 (2016): 141-148.

[58] Jeon, Seungwoo, and Bonghee Hong. "Monte Carlo simulation-based traffic speed forecasting using historical big data." Future generation computer systems 65 (2016): 182-195.

[59] Liu, X. L., et al. "Short-term traffic flow forecasting based on multi-dimensional parameters." Journal of Transportation Systems Engineering and Information Technology 11.4 (2011): 140-146.

References

[60] Canaud, Matthieu, et al. "Probabilty hypothesis density filtering for real-time traffic state estimation and prediction." Networks and Heterogeneous Media (NHM) 8.3 (2013): 825-842.

[61] Antoniou, Constantinos, Haris N. Koutsopoulos, and George Yannis. "Dynamic data-driven local traffic state estimation and prediction." Transportation Research Part C: Emerging Technologies 34 (2013): 89-107.

[62] Lu, Chung-Cheng, Xuesong Zhou, and Kuilin Zhang. "Dynamic origin–destination demand flow estimation under congested traffic conditions." Transportation Research Part C: Emerging Technologies 34 (2013): 16-37.

[63] J. B. Gordon, "Intermodal passenger flows London's public transport network: Automated inference full passenger journeys using fare-transaction vehicle-location data", 2012.

[64] L. Alexander, S. Jiang, M. Murga and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data", Transp. Res. C Emerg. Technol., vol. 58, pp. 240-250, Sep. 2015.

[65] S. Tao, "Investigating the travel behaviour dynamics of bus rapid transit passengers", 2015.

[66] J. Chan et al., "Rail transit OD matrix estimation and journey time reliability metrics using automated fare data", 2007.

[67] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff and M. C. González, "The path most traveled: Travel demand estimation using big data resources", Transp. Res. C Emerg. Technol., vol. 58, pp. 162-177, Sep. 2015.

[68] B. Ferris, K. Watkins and A. Borning, "OneBusAway: Results from providing real-time arrival information for public transit", Proc. SIGCHI Conf. Human Factors Comput. Syst., pp. 1807-1816, 2010.

[69] B. Ferris, K. Watkins and A. Borning, "OneBusAway: Results from providing real-time arrival information for public transit", Proc. SIGCHI Conf. Human Factors Comput. Syst., pp. 1807-1816, 2010.

[70] [online] Available: https://www.waze.com/.

[71] [online] Available: http://moovitapp.com/.

[72] [online] Available: http://inrix.com/mobile-apps/.

[73] [online] Available: http://gaode.com/.

References

[74] Z. Jiang, C.-H. Hsu, D. Zhang and X. Zou, "Evaluating rail transit timetable using big passengers' data", J. Comput. Syst. Sci., vol. 82, no. 1, pp. 144-155, 2016.

[75] J. Zhou, "Applications of machine learning methods in problem of precise train stopping", Comput. Eng. Appl., vol. 46, no. 25, pp. 226-230, 2010.

[76] Z. Hou, Y. Wang, C. Yin and T. Tong, "Terminal iterative learning control based station stop control of a train", Int. J. Control, vol. 84, no. 7, pp. 1263-1274, 2011.

[77] A. Núñez, J. Hendriks, Z. Li, B. De Schutter and R. Dollevoet, "Facilitating maintenance decisions on the dutch railways using big data: The aba case study", Proc. IEEE Int. Conf. Big Data, pp. 48-53, Oct. 2014.

[78] J. Tutcher, "Ontology-driven data integration for railway asset monitoring applications", Proc. IEEE Int. Conf. Big Data, pp. 85-95, Oct. 2014.

[79] F. Wang, T. Xu, T. Tang, M. Zhou and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems", IEEE Trans. Intell. Transp. Syst., vol. 18, no. 1, pp. 49-58, Jan. 2017.

[80] Barry, James J., et al. "Origin and destination estimation in New York City with automated fare system data." Transportation Research Record 1817.1 (2002): 183-187.

[81] Zhao, Jinhua. The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples. Diss. Massachusetts Institute of Technology, 2004.

[82] Nunes, António A., Teresa Galvão Dias, and João Falcão e Cunha. "Passenger journey destination estimation from automated fare collection system data using spatial validation." IEEE transactions on intelligent transportation systems 17.1 (2015): 133-142.

[83] Trepanier, Martin, and Robert Chapleau. "Destination estimation from public transport smartcard data." IFAC Proceedings Volumes 39.3 (2006): 393-398.

[84] Zhao, Juanjuan, et al. "Spatio-temporal analysis of passenger travel patterns in massive smart card data." IEEE Transactions on Intelligent Transportation Systems 18.11 (2017): 3135-3146.

[85] Munizaga, Marcela A., and Carolina Palma. "Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile." Transportation Research Part C: Emerging Technologies 24 (2012): 9-18.

References

[86] Munizaga, Marcela, et al. "Validating travel behavior estimated from smartcard data." Transportation Research Part C: Emerging Technologies 44 (2014): 70-79.

[87] Alsger, Azalden, et al. "Validating and improving public transport origin–destination estimation algorithm using smart card fare data." Transportation Research Part C: Emerging Technologies 68 (2016): 490-506.

[88] Cui, Alex. Bus passenger origin-destination matrix estimation using automated data collection systems. Diss. Massachusetts Institute of Technology, 2006.

[89] Alsger, Azalden A., et al. "Use of smart card fare data to estimate public transport origin–destination matrix." Transportation Research Record 2535.1 (2015): 88-96.

[90] Wang, Wei, John P. Attanucci, and Nigel HM Wilson. "Bus passenger origin-destination estimation and related analyses using automated data collection systems." (2011).

[91] Trepanier, Martin, and Robert Chapleau. "Destination estimation from public transport smartcard data." IFAC Proceedings Volumes 39.3 (2006): 393-398.

[92] Munizaga, Marcela, et al. "Validating travel behavior estimated from smartcard data." Transportation Research Part C: Emerging Technologies 44 (2014): 70-79.

[93] Huili, Dou, Liu Haode, and Yang Xiaoguang. "OD matrix estimation method of public transportation flow based on passenger boarding and alighting." Computer and Communications 25.2 (2007): 79.

[94] Zhou, Xuemei, Xiyu Yang, and Xiaofei Wu. "Origin-destination matrix estimation method of public transportaion flow based on data from bus integrated-circuit cards." Journal of Tongji University. Natural Science 40.7 (2012): 1027-1030.

[95] Yang, Wanbo, et al. "OD matrix inference for urban public transportation trip based on GPS and IC card data." J. Chongqing Jiaotong Univ 34 (2015): 117-121.

[96] Zhang M, Guo Y, Ma Y. "A Probability Model of Transit OD Distribution Based on the Allure of Bus Station." J. Transp. Inf. Saf (2014): 57–61.

[97] YU, Jie, and Xiao-guang YANG. "Estimation a Transit Route OD Matrix Using On/Off Data: An Application of Modified BP Artificial Neural Network [J]." Systems Engineering 4 (2006).

[98] Jung, Jaeyoung, and Keemin Sohn. "Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data." IET Intelligent Transport Systems 11.6 (2017): 334-339.

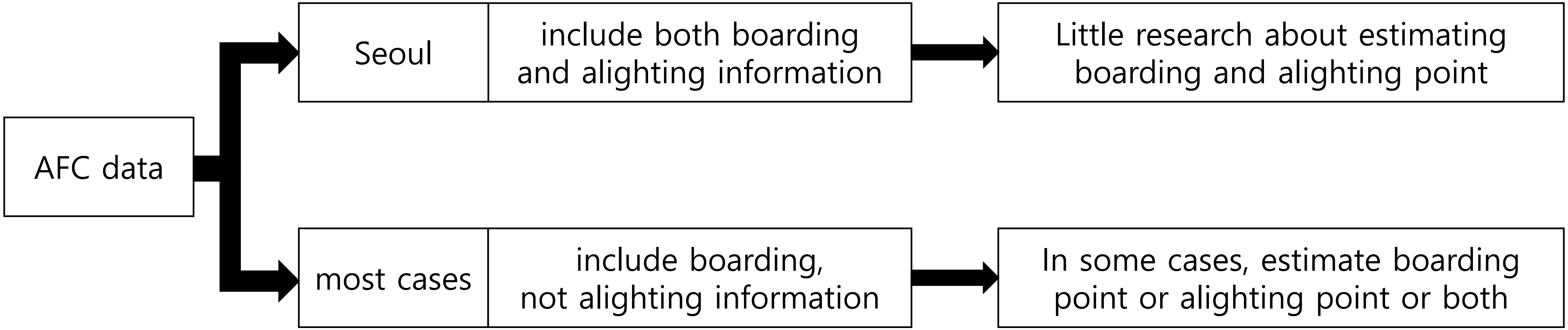
Appendix

Subject selection

1. Compare AFC data of Seoul with that of most cases
2. Survey some studies based on AFC data
3. Methodology analysis

Subject selection

- AFC of Seoul and most cases



Paper survey: 1) Trip pattern

- Title: Relationship between Spatio-Temporal Travel Patterns Derived from Smart-Card Data and Local Environmental Characteristics of Seoul, Korea [4]
- Authors: MK Kim, S Kim, HG Sohn
- Journal: Sustainability (SCIE, IF: 2.576 in 2019)
- Input & Output:

Transaction Number	Boarding Time (yyyy/mm/dd/hh/mm/ss)	ID of Boarding Location	Alighting Time (yyyy/mm/dd/hh/mm/ss)	ID of Alighting Location	Number of Passengers
1	20150318092455	4196066	20150318092714	4196117	1
2	20150318214734	4196061	20150318215035	4196065	1
3	20150318225205	12174	20150318225855	8001017	1
4	20150318075726	70647	20150318082805	10585	1
5	20150318070948	216	20150318084718	4101317	1
...	...	...	...	...	...

The configuration of arranged data from transit records.

=> Clustering => 5 clusters (C1-C5)

Paper survey: 1) Trip pattern

- Input & Output:

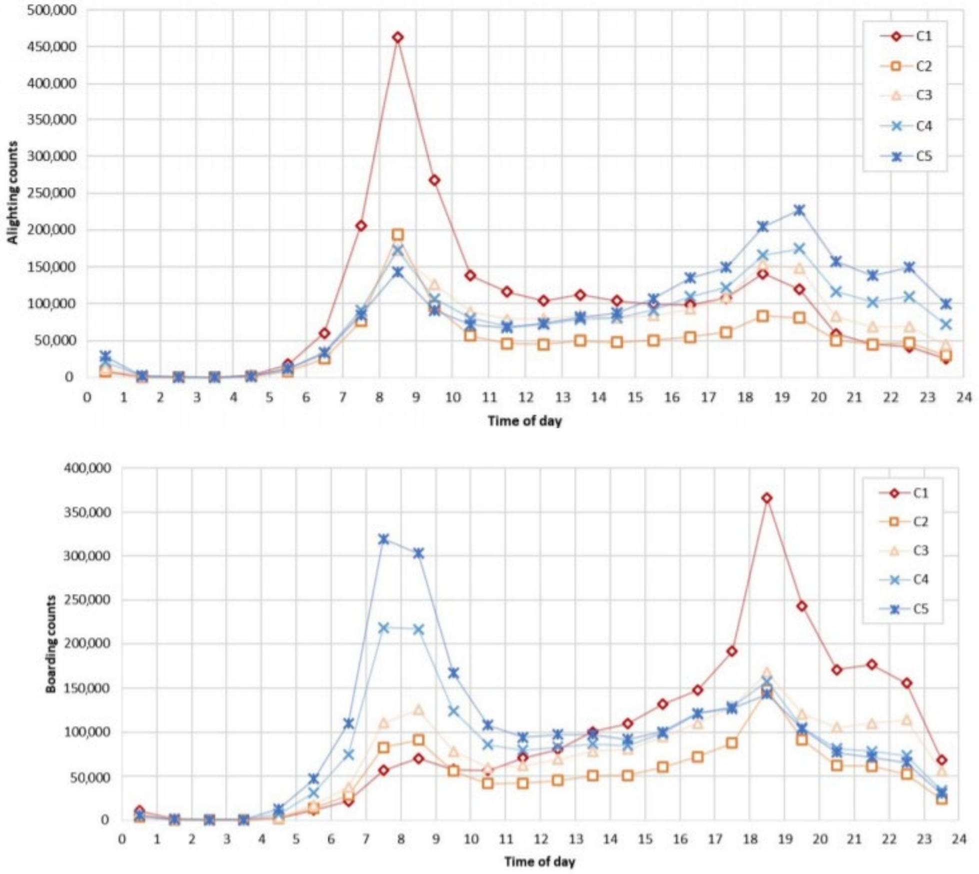
    2) Local environment data provided by Statistics Korea and the Korean Ministry of Environment: 142 variables => Principle component analysis => 3 principle components (3 PC)

Category	Description	Variables
Population	Age/sex distribution Education level (type of last school completed) Marital status	in_age_001~054 in_edu_001~008 in_wed_001~008
Household	Number of rooms Type of occupancy Family type of household	ga_co_001~011 ga_po_001~006 ga_sd_001~006
Housing	Floor space of house Type of housing Construction year	ho_ar_001~009 ho_gb_001~006 ho_yr_001~013
Business	Number of businesses by 9th industrial classification in Korea Number of employees by 9th industrial classification in Korea	cp9_bnu_01~18 cp9_bem_01~18
Land use/Land cover	Residential area Industrial area Commercial area Recreational facilities Roads Public facilities Agricultural area Forest Grass Wetland Bare land Water	lcm_110 lcm_120 lcm_130 lcm_140 lcm_150 lcm_160 lcm_200 lcm_300 lcm_400 lcm_500 lcm_600 lcm_700

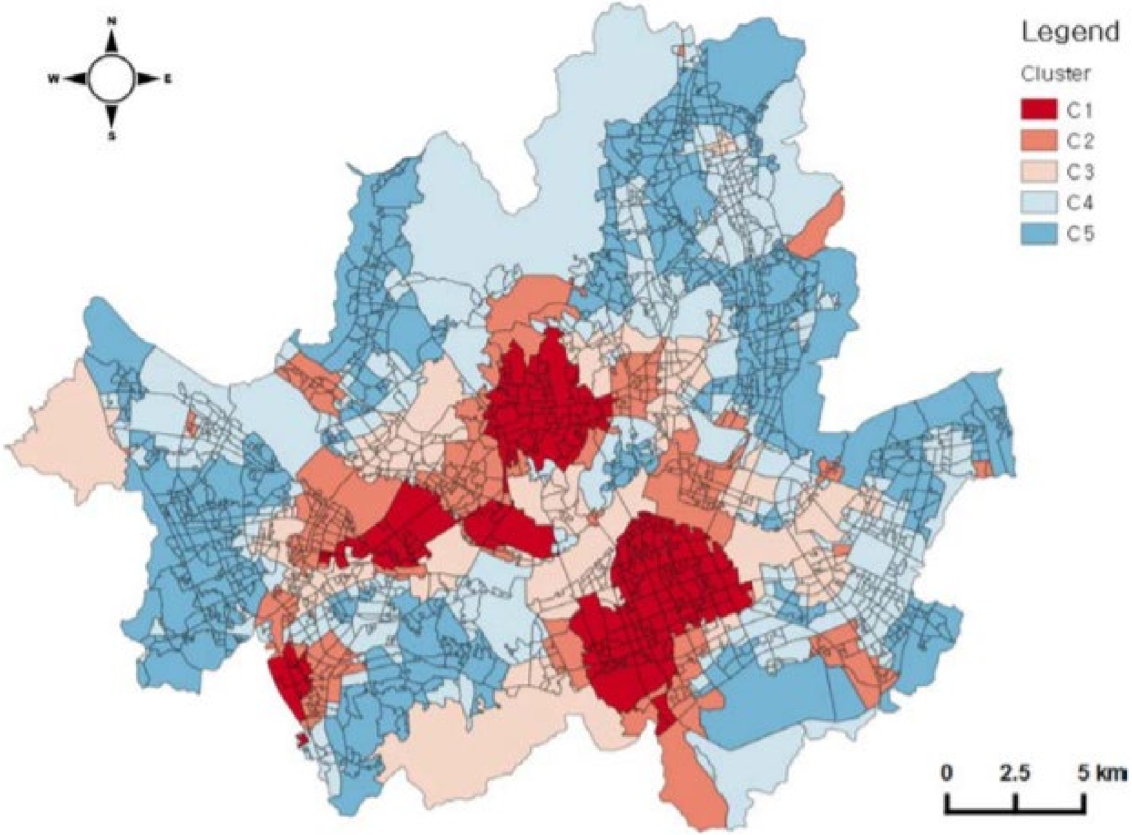
Local environment data

Paper survey: 1) Trip pattern

- Results summary of the paper using 5 clusters and 3 PC:



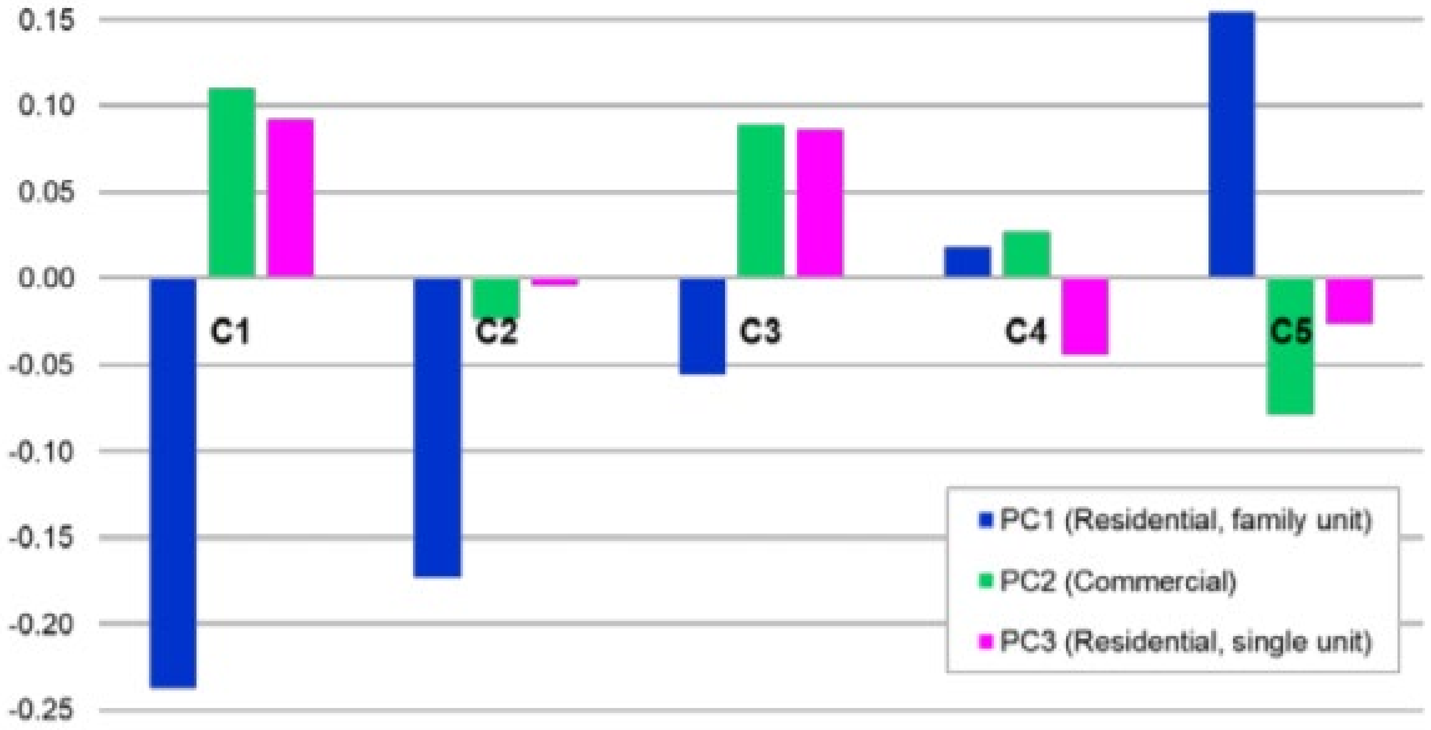
Weekday pattern of boarding and alighting



Spatial distribution of clusters (spatial pattern of mobility)

Paper survey: 1) Trip pattern

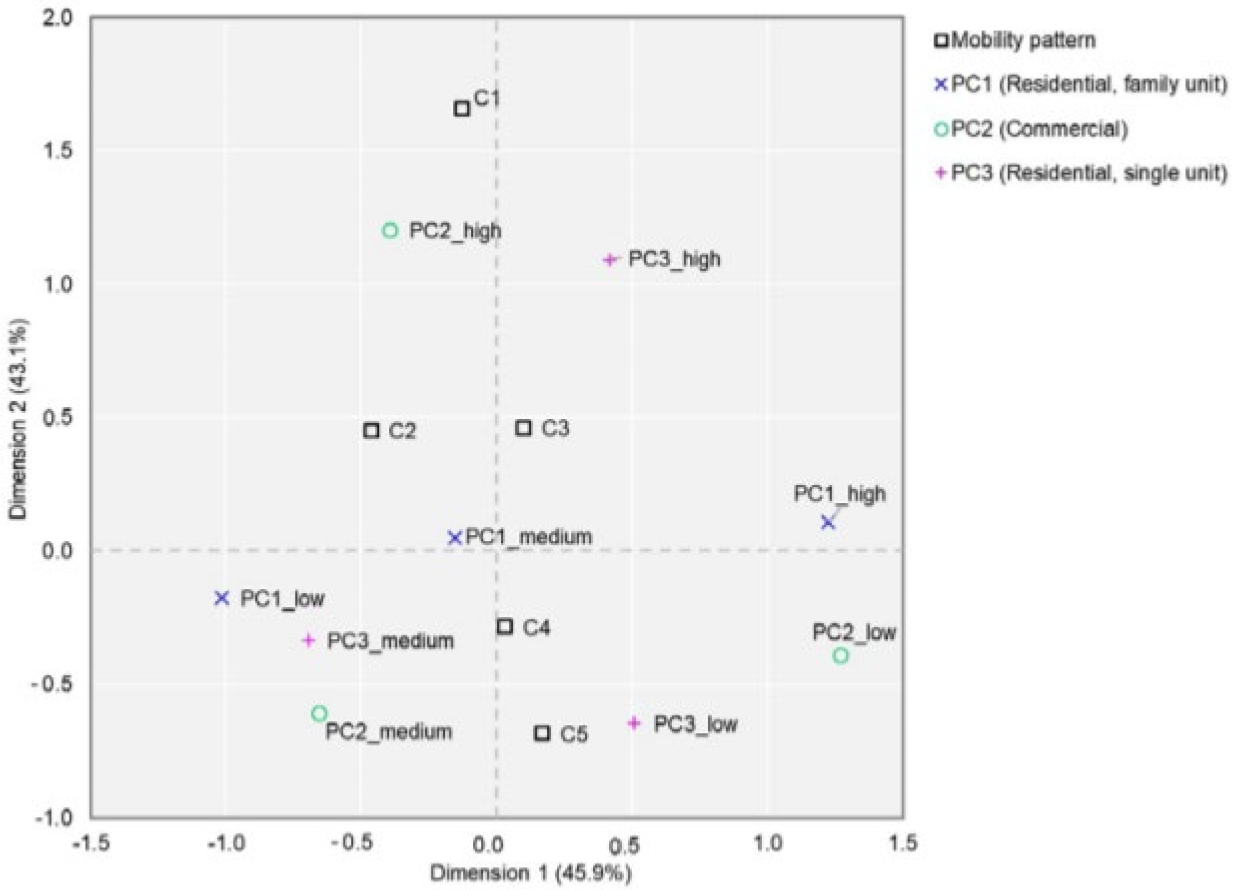
- Results summary of the paper using 5 clusters and 3 PC:



Average of newly constructed variables for five clusters

Paper survey: 1) Trip pattern

- Results summary of the paper using 5 clusters and 3 PC:



Relationship between mobility patterns and environments

Paper survey: 1) Trip pattern

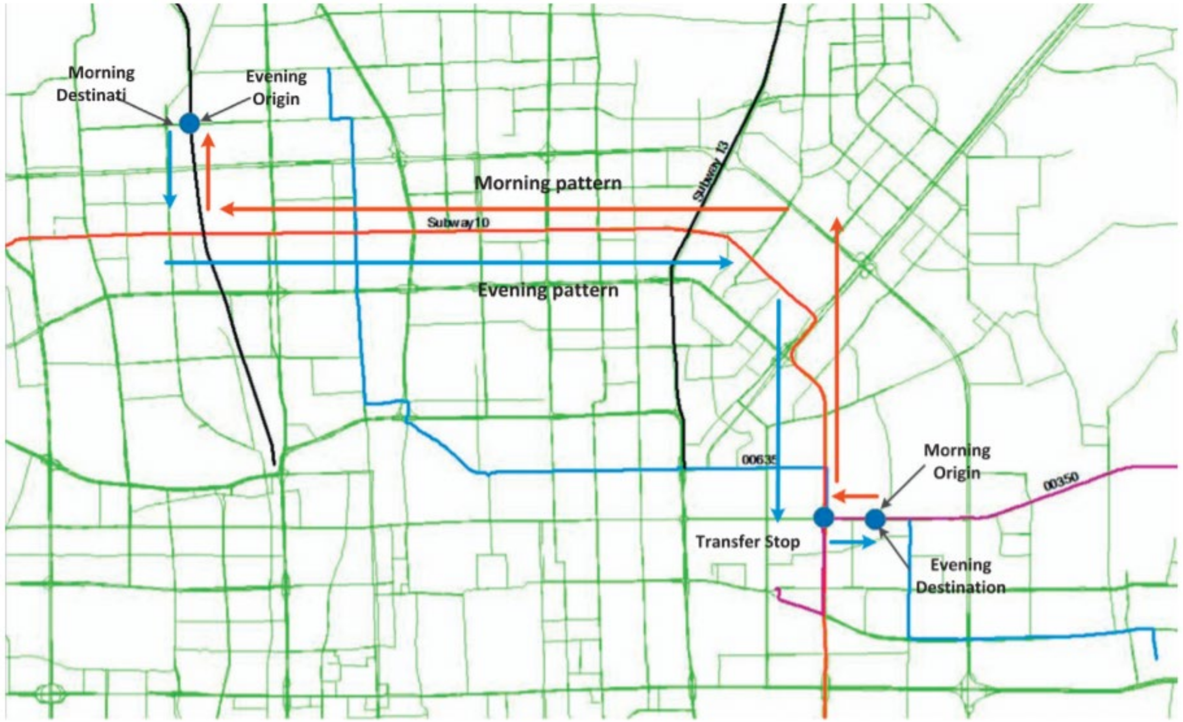
- Title: Mining smart card data for transit riders' travel patterns [5]
- Authors: X Ma, YJ Wu, Y Wang, F Chen, J Liu
- Journal: Transportation Research Part C: Emerging Technologies (SCIE, IF: 6.077 in 2019)
- Flow of a part of the paper: Trip chain generation -> Individual travel pattern recognition -> Regularity clustering
- Input & Output of trip chain generation step:
  - Raw data to trip chain information by estimating transit rider's boarding stops

Chain ID	Card ID	Date	First boarding time	Last alighting time	Route sequence	Stop ID sequence
46388399	1000751018309337	20100705	07:08:45	07:47:28	00635 → 10 → 13	99964,99966 → 50258,50167
46388400	1000751018309337	20100705	18:15:24	18:53:10	13 → 10 → 00635	50192,50245 → 100013,100015
46388401	1000751018309337	20100706	07:19:21	08:01:13	00350 → 10 → 13	91267,91269 → 50258,50167
46388402	1000751018309337	20100706	17:56:08	18:49:50	13 → 10 → 00635	50192,50245 → 100013,100015
46388403	1000751018309337	20100707	07:10:43	07:49:21	00635 → 10 → 13	99964,99966 → 50258,50167
46388404	1000751018309337	20100707	18:29:00	19:06:47	13 → 10 → 00350	50192,50245 → 91276,91278
46388405	1000751018309337	20100708	21:13:58	21:40:10	5 → 10	50125,50246
46388406	1000751018309337	20100709	07:16:24	08:03:46	00635 → 10 → 13	99964,99966 → 50258,50167
46388407	1000751018309337	20100709	17:25:00	18:11:59	13 → 10 → 00635	50192,50245 → 100013,100015
46388408	1000751018309337	20100709	18:30:31	NULL	00031	NULL

Extracted trip chain information for an individual transit rider for the week of 5th–9th July, 2010

Paper survey: 1) Trip pattern

- Input & Output of individual travel pattern recognition:
  - Trip chain information is clustered based on the card ID by using the density-based spatial clustering of application with noise (DBSCAN) algorithm
  - Output: typical trip chains for each individuals
  - Example



Example of a transit rider’s travel pattern

Paper survey: 1) Trip pattern

- Input & Output of regularity clustering:
  - Input: Number of travel days, Number of similar First Boarding Times, Number of similar Route Sequences, Number of similar Stop ID Sequences
  - Use k-means algorithm
  - Output: 5 clusters

Cluster center	C_1	C_2	C_3	C_4	C_5
Regularity	VL	L	M	H	VH
Cluster Center Distance	1.28	5.17	10.42	13.52	19.99
Number of Smart Cards	4809	10330	6483	9502	5877
Percentage of total	13.0%	27.9%	17.5%	25.7%	15.9%

Note: VL = Very Low; L = Low; M = Medium; H = High; VH = Very High.

Summary of five clusters

Paper survey: 2) Commuting pattern

- Title: Understanding commuting patterns using transit smart card data [6]
- Authors: X Ma, C Liu, H Wen, Y Wang, YJ Wu
- Journal: Journal of Transport Geography (SSCI, IF: 3.834 in 2019)
- Flow of some sections of the paper:
  - Trip generation -> Commuting feature extraction and travel pattern recognition -> commuter identification
- Trip generation: estimate O-D matrix using smart card data
- Commuting feature extraction and travel pattern recognition:
  - Grouping temporally pattern into 30-minute unit and clustering spatial pattern using stop locations.
- Commuter identification: using clustering

Paper survey: 3) Travel behavior

- Title: Travel Behavior Analysis using Smart Card Data [7]
- Authors: A Ali, J Kim, S Lee
- Journal: KSCE Journal of Civil Engineering (SCIE, IF: 1.515 in 2019)
- Flow of some sections of the paper:
  - Generating OD matrix -> Travel behavior analysis -> Towards activity identification -> Analysis with Multi-Agent Transport Simulation Toolkit (MATSim)
- Input & Output of Section 3 (i.e. Generating OD matrix):
  - Data: 28 million transactions (provided by Korean Transport Institute for five weekdays from June 11 to June 15, 2012): Unique card ID, boarding time, boarding station ID, alight time, alight station ID, transport method, bus route ID, passenger type, total fare, total distance travelled
  - And some shapefiles for ArcGIS (a software for geographic information system)
  - They was handled in Statistical Package for the Social Sciences (SPSS) and ArcGIS for spatial analysis

Paper survey: 3) Travel behavior

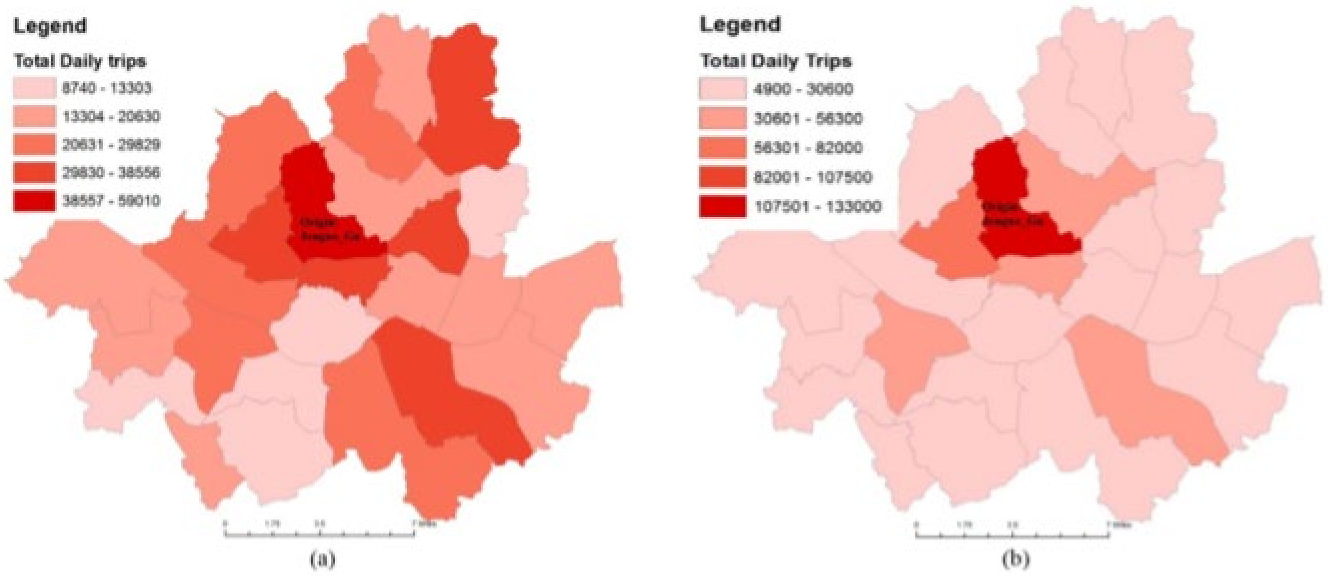
- Input & Output of Section 3 (i.e. Generating OD matrix):
  1) Assigning each transit stop a zone code based on its spatial location
     - Each of 467 sub-districts of Seoul, "dong" is used as a Traffic Analysis Zone (TAZ)
     - Assign the zone codes to the stops using ArcGIS
  2) Joining zonal attributes with database
     - New database: for each trip segment within each trip, xy coordinates and zone code attributes for each boarding and alighting added into the database.
  3) Differentiating trips and trip segments
     - The trips are separated from tip segments the chains are identified and unnecessary data is trimmed off from the database.

Trip Segments				**Individual Trips**		
Card ID	Boarding Time	Alight Time	Description	Card ID	Boarding Time	Alight Time
18558722	08:11:12	08:29:10	$O_1 = B_1$	18558722	08:11:12	08:55:09
18558722	08:33:52	08:55:09	$D_1 = A_2$	18558722	17:20:11	17:32:33
18558722	17:20:11	17:32:33	$O_2 = B_3 \& D_2 = A_3$	18558722	18:10:10	18:42:41
18558722	18:10:10	18:16:50	$O_3 = B_4$			
18558722	18:20:24	18:30:22				
18558722	18:31:30	18:42:41	$D_3 = A_6$			

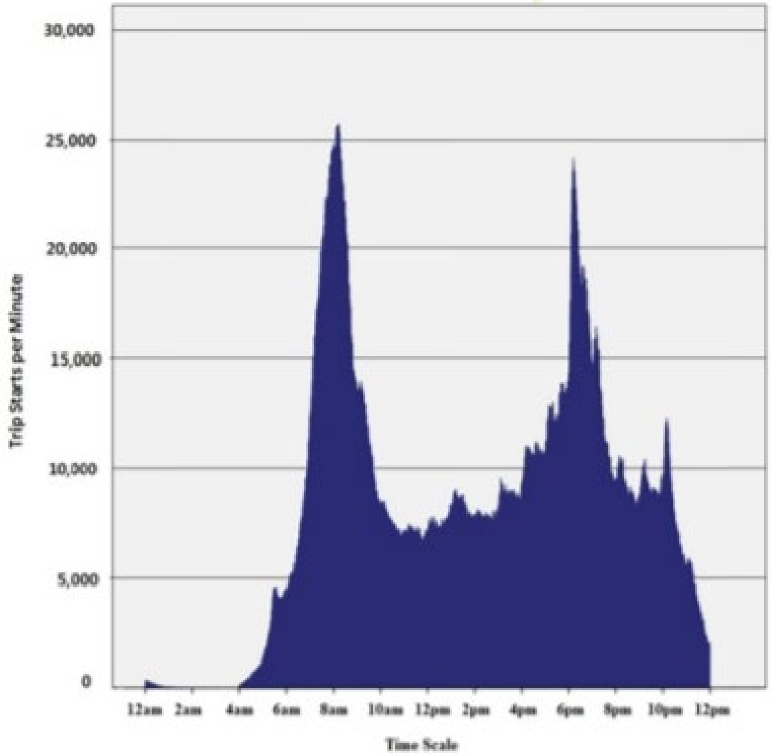
Trip segments and complete trips

Paper survey: 3) Travel behavior

- Section 4 (i.e. Travel behavior analysis):
  - Study in-depth knowledge about spacio-temporal distribution of trips



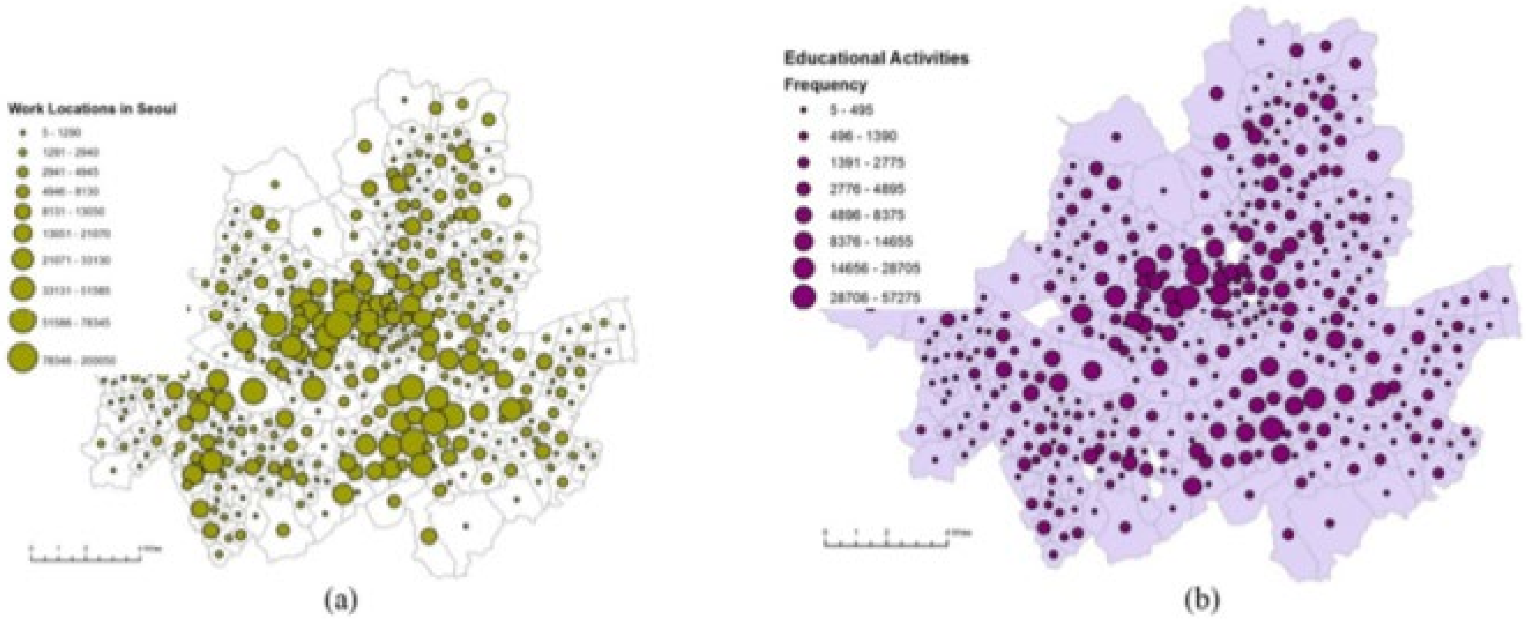
(a) Trip Destination for Jongno-district with main Mode as Subway, (b) Trip Destinations for Jongno-district with main Mode as Bus



A High Temporal Resolution of Trip Start Distribution

Paper survey: 3) Travel behavior

- Section 5 (i.e. Towards activity identification):
  - Set a 30 minutes transfer time as a bench mark for identifying activities.
  - Using some criteria, activities are distinguished as; work, school, academy, shopping.



(a) Work Locations, (b) Educational Activities

Paper survey: 4) Passenger OD

- Title: Estimating a transit passenger trip origin-destination matrix using automatic fare collection system (citied by 22 in Scopus) [8]
- Authors: D Li, Y Lin, X Zhao, H Song, N Zou
- Conference: International Conference on Database Systems for Advanced Applications
- OD estimation model: proposed algorithms (1-3)
- Input data: AFC data in China, there is few alighting information

Card ID	Card style	Route	Stop	Veh. ID	Time	Date	Fare
-1760518818	Monthly card	115	156	2798	06:27:16	2009-07-07	-
-1760518818	Monthly card	115	3	2346	18:47:21	2009-07-07	-
-1760517522	Common card	115	19	2810	08:34:53	2009-07-07	0.9
-1760517522	Common card	115	151	2808	17:47:31	2009-07-07	0.9
-1760515746	Monthly card	115	133	2794	11:56:22	2009-07-07	-
-1760515746	Monthly card	115	133	2814	18:52:29	2009-07-07	-
-1760515714	Monthly card	115	23	2794	17:32:51	2009-07-07	-
-1760515122	Monthly card	115	151	789	18:01:25	2009-07-07	-
-1760514466	Monthly card	115	12	2814	17:38:18	2009-07-07	-
-1760513522	Monthly card	115	31	2802	07:40:42	2009-07-07	-

Partial passenger boarding information from AFC system

Paper survey: 4) Passenger OD

- Algorithm 1: Trajectory search algorithm
  - Input:
    - RP, the set of smart card data from all the passengers in one city
    - $\mathrm{n}_{\mathrm{s}_\mathrm{r}}$, the set of bus stop number for one city
  - Output :
    - Alighting stop of each passenger's every trip. Both boarding stops and estimated alighting ones are combined into the set of trip trajectory PT.
- Algorithm 2: Travel demand matrix
  - Input:
    - PT, the set of trip trajectory
    - SN, the set of bus stop name for one city
  - Output :
    - M, passengers' travel demand matrix for each bus stop; its dimension is equal to $\mathrm{s_n} \times \mathrm{s_n}$, while $\mathrm{s_n}$ represents the number of bus stop name for one city

Paper survey: 4) Passenger OD

- Algorithm 3: Travel demand matrix
  - Input:
    - Passengers' travel demand matrix M
    - The set of all traffic zones $TA = ta_1 \cup ta_2 \cup \cdots \cup ta_{\mathrm{k}}$
  - Output :
    - The travel demand Matrix M of the passengers grouped by traffic zones, and its dimension is equal to $K \times K$, with K representing the number of different bus stop name for one city

O-D matrix estimation algorithm

- Core algorithm of some studies based on AFC data
- Input: AFC data (basically include CARD ID, BOARDING STOP, BOARDING TIME)
- Output: O-D matrix
  - Column and row represent a zone in the study area.
  - T_{ij} shows the number of trips between origin i and destination j.
  - O_i and D_j are total origin and destination respectively.

$Zones$	1	2	$\cdots$	j	$\cdots$	n	O_i
1	T_{11}	T_{12}	$\cdots$	T_{1j}	$\cdots$	T_{1n}	O_1
2	T_{21}	T_{22}	$\cdots$	T_{2j}	$\cdots$	T_{2n}	O_2
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	T_{i1}	T_{i2}	$\cdots$	T_{ij}	$\cdots$	T_{in}	O_{i}
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
n	T_{n1}	T_{n2}	$\cdots$	T_{nj}	$\cdots$	T_{nn}	O_{n}
D_{j}	D_1	D_2		D_j		D_n	

OD matrix illustration [19]

A Simple O-D matrix estimation example

- Simplify the O-D matrix estimation algorithm of Li et al (2011).
- Let us estimate the alighting stop based on the coordinate information of route and stop.
- We have a simplified boarding information table from AFC data, and stop, coordinate, and route information as shown below.

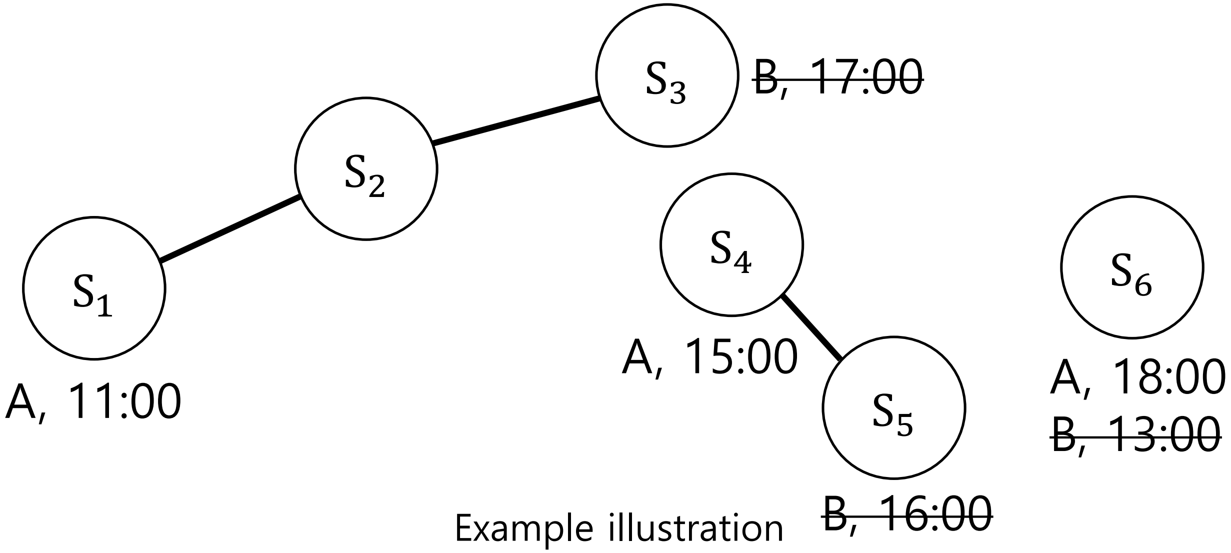
CARD ID	STOP	ROUTE	TIME
A	S_6	R_3	18:00
B	S_6	R_3	13:00
B	S_5	R_2	16:00
A	S_1	R_1	11:00
A	S_4	R_2	15:00
B	S_3	R_1	17:00

Boarding information

STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2
S_5	(x_5, y_5)	R_2
S_6	(x_6, y_6)	R_3

Stop information

A Simple O-D matrix estimation example



Example illustration

STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2
S_5	(x_5, y_5)	R_2
S_6	(x_6, y_6)	R_3

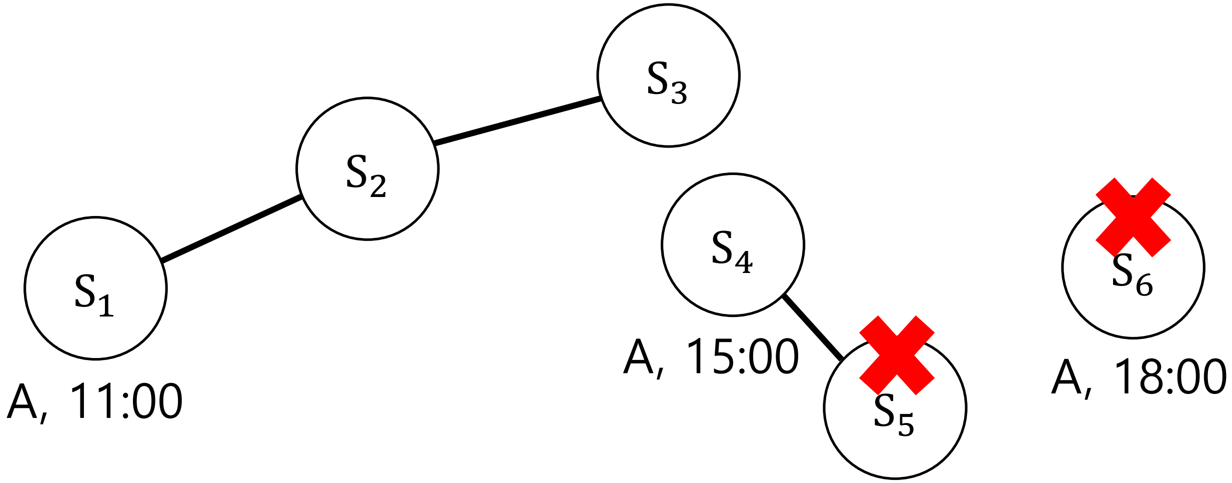
- First, get the information of Card ID A and sort by time.

CARD ID	STOP	ROUTE	TIME
A	S_6	R_3	18:00
A	S_1	R_1	11:00
A	S_4	R_2	15:00

➡

CARD ID	STOP	ROUTE	TIME
A	S_1	R_1	11:00
A	S_4	R_2	15:00
A	S_6	R_3	18:00

A Simple O-D matrix estimation example



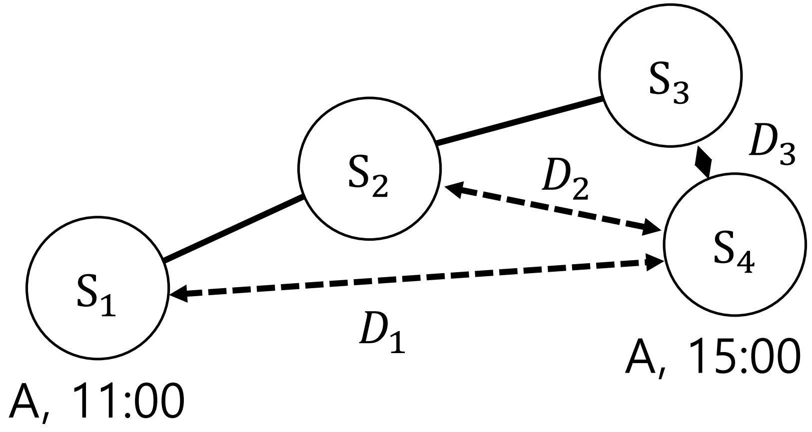
Example illustration

STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2
$\cancel{S_5}$	$\cancel{(x_5, y_5)}$	$\cancel{R_2}$
$\cancel{S_6}$	$\cancel{(x_6, y_6)}$	$\cancel{R_3}$

- Consider ROUTE of the top two transactions, and for first one, get information based on ROUTE and for second one, get information based on STOP.

CARD ID	STOP	ROUTE	TIME
A	S_4	R_1	11:00
A	S_6	R_2	15:00

A Simple O-D matrix estimation example



Example illustration

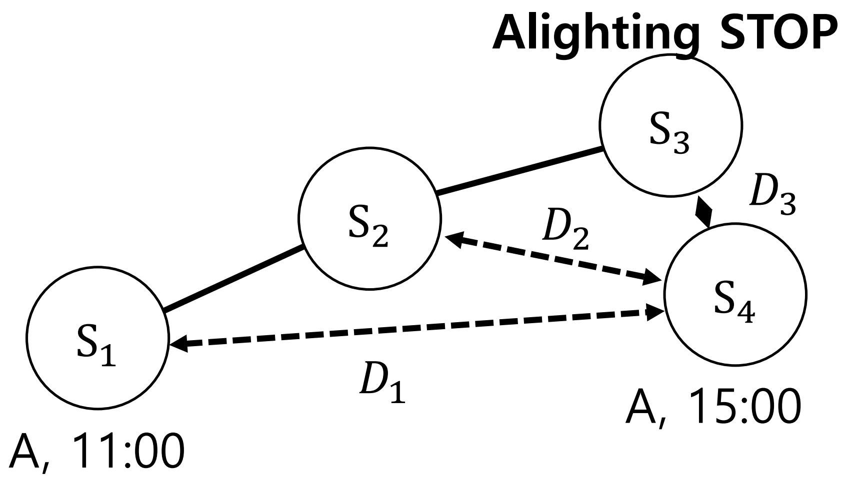
STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2

- Based on STOP table, calculate distance set between every STOPs (i.e. S_1, S_2, S_3) included in ROUTE of first one and STOP (i.e. S_4) of second one.

$$D_1 = \sqrt{(x_4 - x_1)^2 + (y_4 - y_1)^2}$$
$$D_2 = \sqrt{(x_4 - x_2)^2 + (y_4 - y_2)^2}$$
$$D_1 = \sqrt{(x_4 - x_3)^2 + (y_4 - y_3)^2}$$

CARD ID	STOP	ROUTE	TIME
A	S_4	R_1	11:00
A	S_6	R_2	15:00

A Simple O-D matrix estimation example



Example illustration

STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2

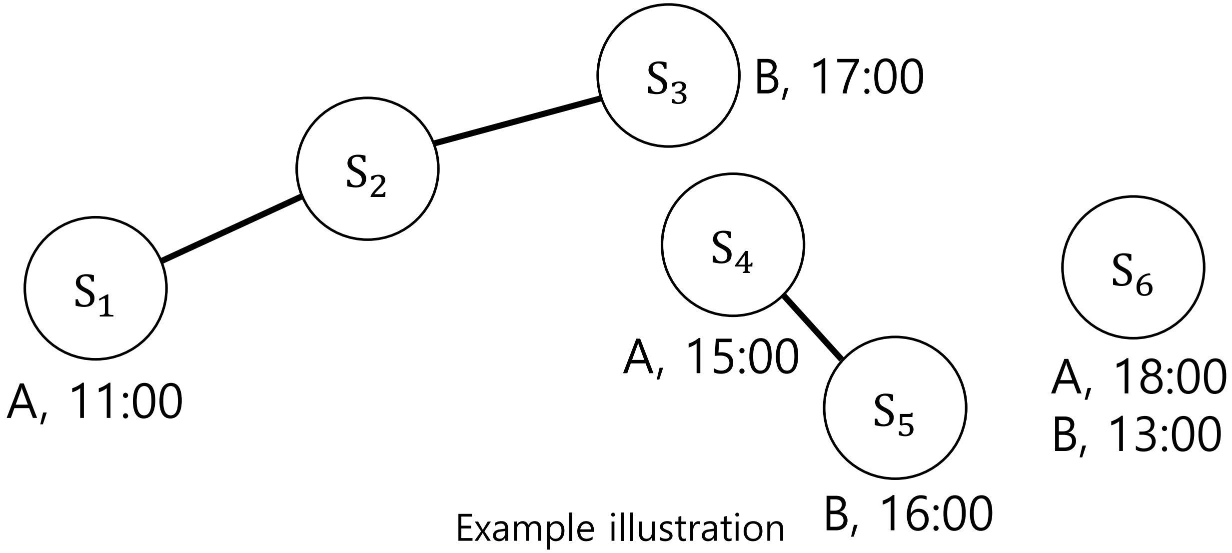
- Assign the minimum value as alighting stop of the transaction.

$$Distance\ of\ the\ transaction = minimum(D_1, D_2, D_3) = D_3$$

$$Alighting\ STOP\ of\ the\ transaction = S_3$$

CARD ID	BOARDING STOP	ROUTE	BOARDING TIME	ALIGHTING STOP
A	S_1	R_1	11:00	S_3

A Simple O-D matrix estimation example



Example illustration

STOP	X-Y coordinate	ROUTE
S_1	(x_1, y_1)	R_1
S_2	(x_2, y_2)	R_1
S_3	(x_3, y_3)	R_1
S_4	(x_4, y_4)	R_2
S_5	(x_5, y_5)	R_2
S_6	(x_6, y_6)	R_3

- Process other transactions in the same way.

CARD ID	BOARDING STOP	ROUTE	BOARDING TIME	ALIGHTING STOP
A	S_1	R_1	11:00	S_3
A	S_6	R_3	18:00	NULL
B	S_6	R_3	13:00	S_6
B	S_5	R_2	16:00	S_4
A	S_4	R_2	15:00	S_5
B	S_3	R_1	17:00	NULL

A Simple O-D matrix estimation example

CARD ID	BOARDING STOP	ROUTE	BOARDING TIME	ALIGHTING STOP
A	S_1	R_1	11:00	S_3
A	S_6	R_3	18:00	NULL->S_6
B	S_6	R_3	13:00	S_6
B	S_5	R_2	16:00	S_4
A	S_4	R_2	15:00	S_5
B	S_3	R_1	17:00	NULL->S_3

- In those NULL cases, we can use some assumptions.
  - For example, we can suppose the last alighting stop is the stop closest to the first boarding stop of the day.
- Using the estimated trajectory, we can continue to generate the OD matrix.

A Simple O-D matrix estimation example

CARD ID	BOARDING STOP	ROUTE	BOARDING TIME	ALIGHTING STOP
A	S_1	R_1	11:00	S_3
A	S_6	R_3	18:00	NULL->S_6
B	S_6	R_3	13:00	S_6
B	S_5	R_2	16:00	S_4
A	S_4	R_2	15:00	S_5
B	S_3	R_1	17:00	NULL->S_3

BOARDING STOP	ALIGHTING STOP
S_1	S_3
S_6	S_6
S_6	S_6
S_5	S_4
S_4	S_5
S_3	S_3

- Delete all features except BOARDING STOP and Alighting STOP.

A Simple O-D matrix estimation example

BOARDING STOP	ALIGHTING STOP
S_1	S_3
S_6	S_6
S_6	S_6
S_5	S_4
S_4	S_5
S_3	S_3

BOARDING STOP	ALIGHTING STOP
S_1	S_3
S_3	S_3
S_4	S_5
S_5	S_4
S_6	S_6
S_6	S_6

- Descend the instances according to BOARDING STOP and ALIGHTING STOP.

A Simple O-D matrix estimation example

BOARDING STOP	ALIGHTING STOP
S_1	S_3
S_3	S_3
S_4	S_5
S_5	S_4
S_6	S_6
S_6	S_6

D \ O	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0	0	0	0	0	0
S_2	0	0	0	0	0	0
S_3	1	0	1	0	0	0
S_4	0	0	0	0	1	0
S_5	0	0	0	1	0	0
S_6	0	0	0	0	0	2

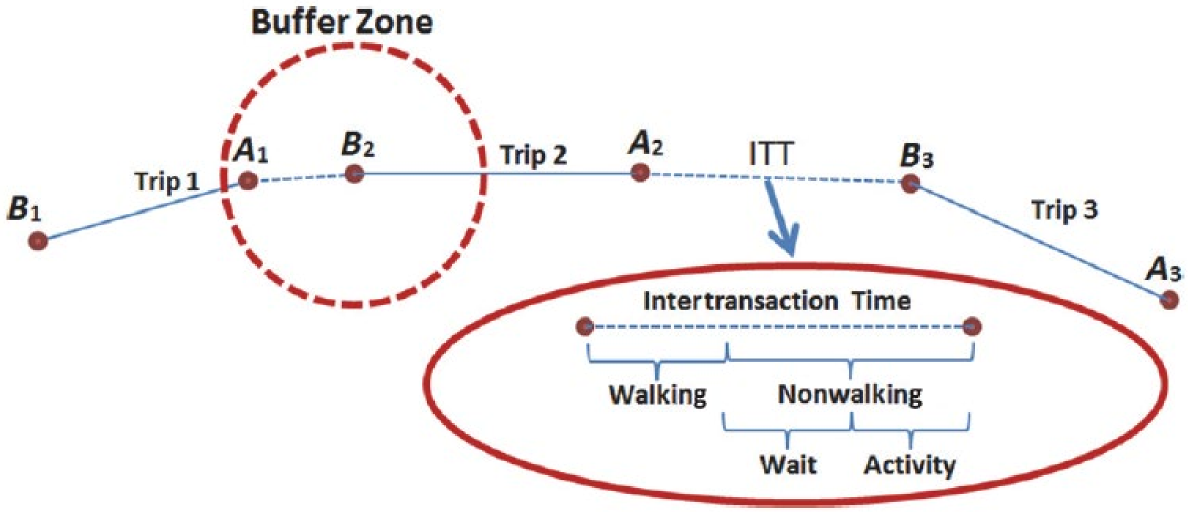
- For each BOARDING STOP, count the number of each ALIGHTING STOP
- Map into O-D matrix.

Trip chain method

- Reference paper:
  - Title: Use of smart card fare data to estimate public transport origin–destination matrix [16]
  - Authors: AA Alsger, M Mesbah, L Ferreira
  - Journal: Transportation research record (SCIE, IF: 1.029 in 2019)
- Most of O-D estimation algorithms are based on trip-chaining method.
- Input: AFC transaction data that has some features: smart card ID, boarding time, alighting time, boarding stop, and alighting stop.

General definition for trip chain method

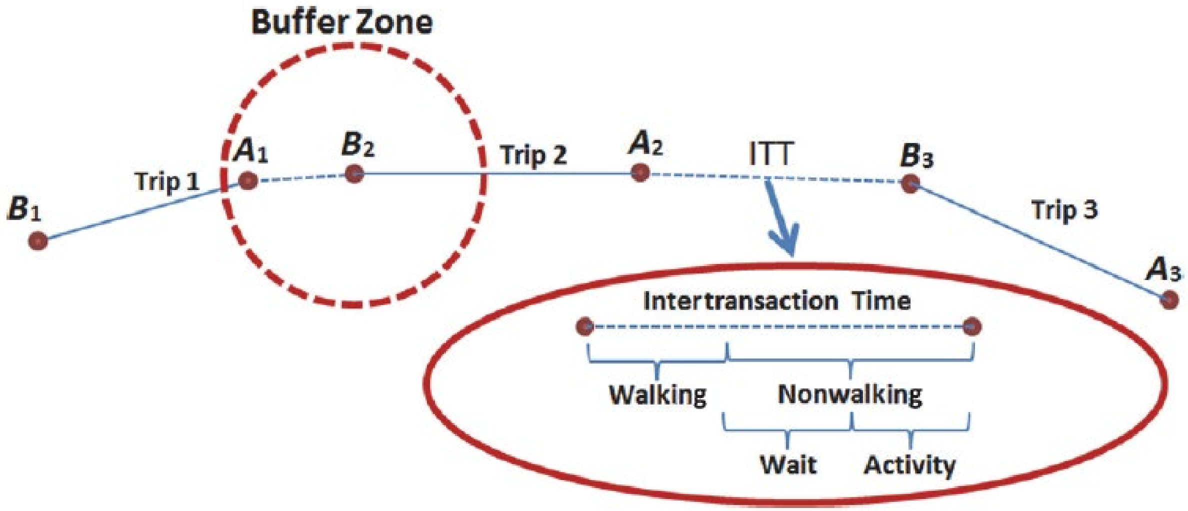
- **Trip leg** is the in-vehicle time or distance (part of a trip) between a passenger's boarding and alighting locations.
- An **O-D trip** describes passenger movement from an origin to a destination by public transport. It may entail one or more trip legs and related transfers.



Example of trip chaining; time between B_1 and A_1 is in-vehicle time.

General definition for trip chain method

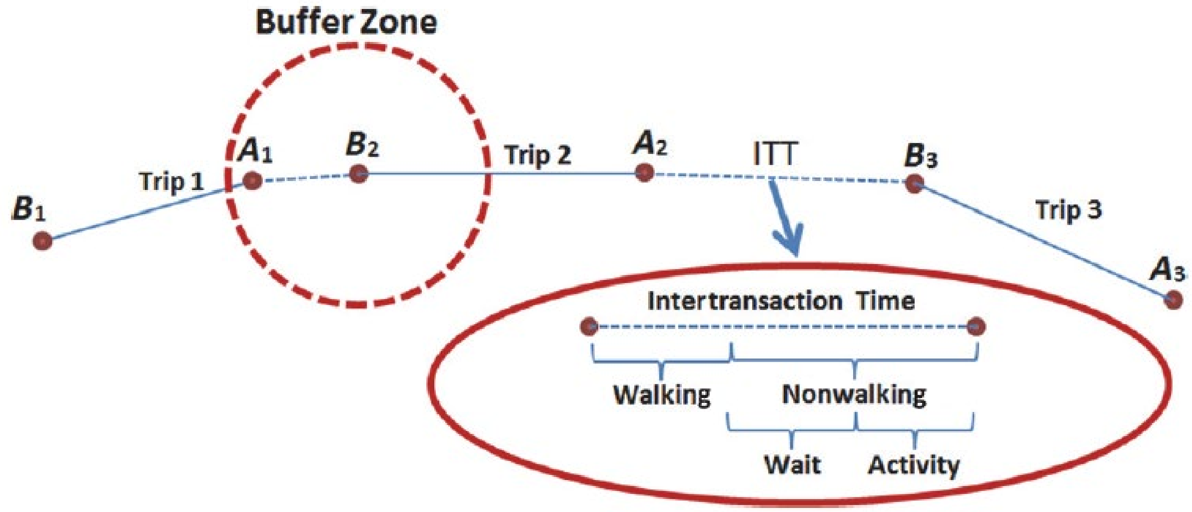
- **Transfer distance** is the walking distance to transfer between subsequent trip legs on public transport (i.e., between a passenger's boarding and alighting locations).
- **Buffer zone** is an assumed area set around a public transport boarding stop that includes any possible public transport alighting stops.



Example of trip chaining; time between B_1 and A_1 is in-vehicle time.

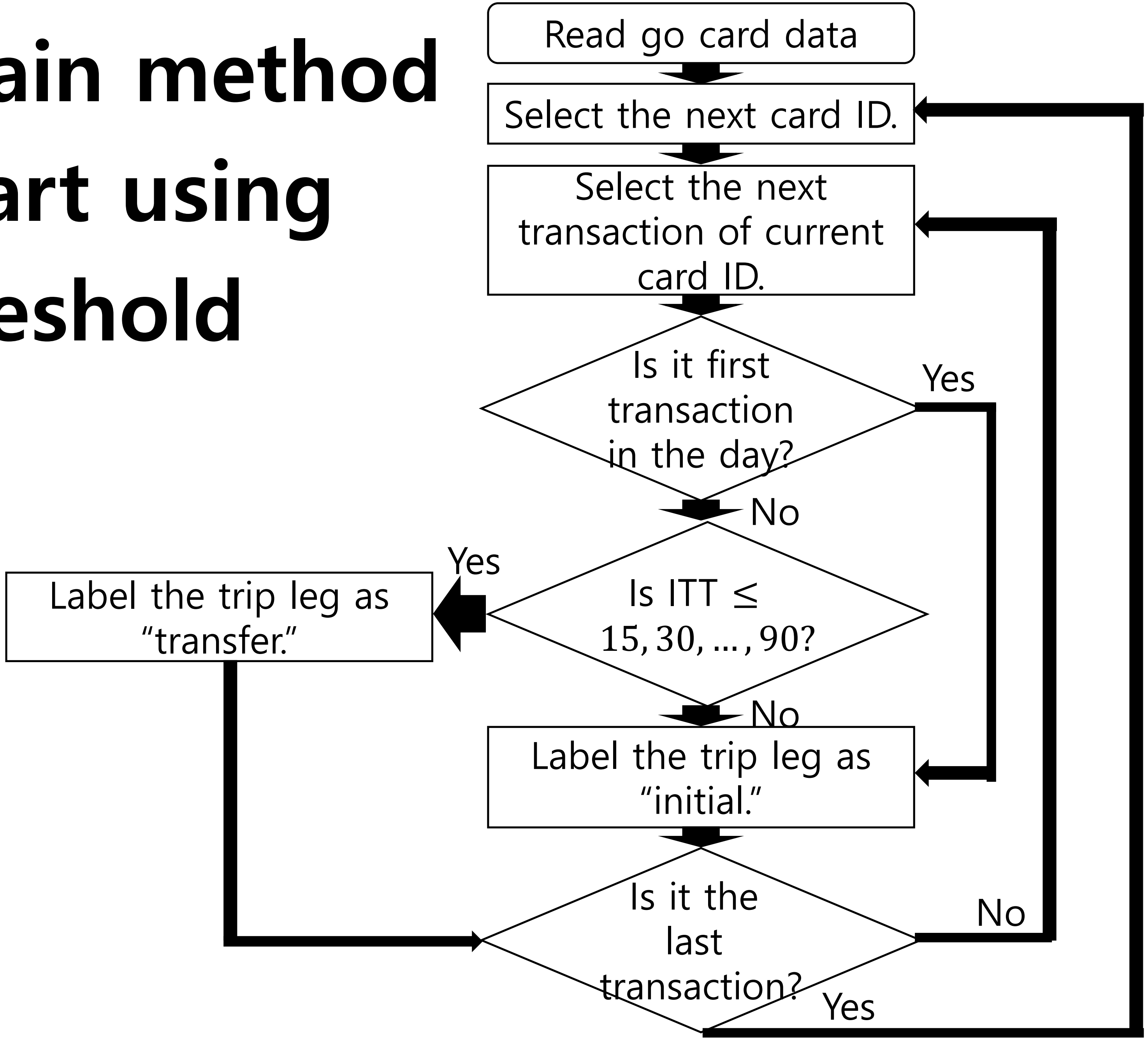
Some assumptions for trip chain method

- A passenger's last destination and first origin locations are the same in a given day
- The major assumption required to connect trip legs and create a passenger trip is a value for the threshold for intertransaction time (ITT): walking time + nonwalking time (waiting time + possible activity time)



Example of trip chaining; time between B_1 and A_1 is in-vehicle time.

Trip chain method flowchart using ITT threshold



Some methods for O-D matrix estimation

- Reference paper:
  - Title: Validating and improving public transport origin–destination estimation algorithm using smart card fare data [17]
  - Authors: A Alsger, B Assemi, M Mesbah, L Ferreira
  - Journal: Transportation Research Part C: Emerging Technologies (SCIE, IF: 6.077 in 2019)
- Estimation attempts:
  1. Walking distances (buffer zones): different walking distances were chosen to infer alighting stops (e.g., 400, 800, 1000 and 1100 m)
  2. Transfer times: different transfer times were chosen to connect trip-legs to infer O–D trips (e.g., 30, 60 and 90 min)
  3. Last destination assumptions: some studies assumed the last destination as the first origin, where others assumed it as the closest stop to the first origin
- Validation attempts
  - Additional data requirement for validation: Additional data (e.g., travel survey, personal interviews, and passenger counting) were used for validation
- Issues of Seoul data
  - Can take validation attempts approach
  - Transfer time of Seoul : fixed at 30 min -> ITT

Some methods for O-D matrix estimation

- Barry et al. (2009) 's assumptions :

  1) A very high percentage of passengers return to their previous alighting station to start their next trip.

  2) A high percentage of passengers finally return to the first station they started their first trip of the day.

- Gordon et al. (2013) 's assumption:

  Last destination is as the closest to the first origin in their methodology.

- Munizaga et al. (2014) with given the lack of alighting information in the main dataset:

  applied exogenous validation (information from travel surveys and personal interviews of a small sample of volunteers).

Some methods for O-D matrix estimation

- Alsger et al. (2017) studied the effect of different data sample sizes on the accuracy level of the generated public transport O-D matrices.
- Prakasa et al. (2017) divided dataset into various time units to create an OD matrix and found that 36.5% transactions the alighting stations still difficult to be inferred because of passenger's certain behavior in their study using trip chain.
- Jung and Sohn (2017) proposed a deep-learning model developed to estimate the destinations of bus passengers based on both entry-only smart-card data and land-use characteristics.

Paper survey: 5) Crowding

- Title: Crowding cost estimation with large scale smart card and vehicle location data [9]
- Authors: D Hörcher, DJ Graham, RJ Anderson
- Journal: Transportation Research Part B: Methodological (SCIE, SSCI, IF: 4.796 in 2019)
- Flow of some sections of the paper:
  - Trip typology -> Assignment methodology -> Modeling
- Trip typology
  - Input: AFC data
  - Output: Type(A-H) of each trip
- Assignment methodology:
  - Input: typed AFC data, Automatic Vehicle Location (AVL) data
  - Output: assignment of each transfer step
- Modeling: travel time multiplier approach

Paper survey: 6) Travel demand

- Title: Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees [10]
- Authors: C Ding, D Wang, X Ma, H Li
- Journal: Sustainability (SCIE, IF: 2.576 in 2019)
- Input:
  - AFC data: 3 subway stations and their nearby bus stops
- Process:
  - Using Gradient Boosting Decision Tree (GBDT)
- Output:
  - Correlation coefficients between subway ridership and transfer volumes from buses
  - Performance of GBDT: R^2, $RMSE$