International Journal of
Information
and
Management
Sciences

http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information and Management Sciences

33 (2022), 271-290. DOI:10.6186/IJIMS.202212_33(4).0001

Artificial Intelligence (AI) Applications Using Big Data and Survey Data for Exploring the Existence of the Potential Users of Public Transportation System

*Hsiang-Chuan Chang**, *Tomohiro Okubo, Akihiro Kobayashi and Akinori Morimoto*
Waseda University/ KDDI CORPORATION/ KDDI Research Inc.

Keywords

Trip generation
cell phone data
artificial intelligence
association analysis
inverse reinforcement learning

Abstract.

The government places emphasis on increasing the usage rate of public transportation nowadays due to public transportation having many benefits for the environment. In order to understand the key factors of trip generation and identify the key trip purposes for selecting transportation modes in a target city, the cell phone data and personal trip survey data were studied by using the machine learning methods of Association Analysis and Inverse Reinforcement Learning. Findings such as hospital, park and elementary school are the most important elements implies that the facilities for mandatory task will attract more people. Also, the elderly age group has very strong tendency to use private vehicle compared to other age groups implies that attracting more young people may be a good strategy. Findings can be a reference for new policy planning, including re-planning the exiting routes of bus systems or integrating different public transportation, by the local government.

1. Introduction

1.1. Background

For developed countries around the world, problems such as budget cut, net outmigration from cities, and aging society, could block the urban development. Public transportation system stands an important role to ease the impacts from the aforenoted problems. How to increase the number of public transportation users becomes an important issue nowadays. The longitudinal data set of passengers can offer useful information for modeling to understand the growth/decade rate of the passengers by public transportation. Bus Rapid Transit (BRT) and Light Rail Transit (LRT) are two major mass

*corresponding author

transit systems. The BRT and LRT are customer-oriented transportation systems which have the merits of delivering fast and comfort. The maintenance costs of the BRT and LRT systems are low to keep urban mobility. Moreover, BRT and LRT provide more flexibility and bus can do service on exclusive lanes along major corridors such as branch out to cover more territory as the destination approaches. The BRT and LRT can provide modern, efficient and comfortable service to public transportation users. BRT is a new solution to improve the efficiency of traditional bus systems. LRT was first introduced in North America by Thompson [19] to describe the new concept of tram transportation. A well-established LRT system can assist city become a TOD (Transit-Oriented Development) City or Compact City through adapting new system (see Takami [16]), and then the city can be more attractive for citizens. It is critical to identify potential public transportation passengers and enhancing their usage rate.

1.2. Objectives

Considering the limitation of acquiring a big data set via using questionnaire survey, the cell phone data can be more competitive for decision making. Nonetheless, traditional survey data can also be applied to Artificial Intelligence (AI) methods to make the calculation faster or obtain more important information. Thanks to the improvement of the software and hardware techniques of computer science, powerful machine learning methods based on AI technologies are emphasized again nowadays. The purpose of this research is using the machine learning method of Association Analysis to identify the key factors of trip generation based on the distance to infrastructures in Utsunomiya City, Japan. Also, another goal is applying the machine learning method of Inverse Reinforcement Learning (IRL) to figure out the tendency of selecting the transportation modes for different trip purposes. The results can be a reference for new policy planning, including re-planning the exiting routes of bus systems or integrating different public transportation systems, by the local government. Furthermore, we expect the proposed methods outperform traditional analysis methods to improve the precision accuracy.

2. Literature Review and Research Flow

2.1 Literature review

The literatures review was conducted for the following three themes. (1) Public transportation system and land use (2) Mobile phone big data, and (3) Machine learning methods related to this study. Numerous studies can be found in literature for analyzing the data sets from public transportation systems. Some recent studies, for example, Zhao et al. [24] quantified the impacts of Urban Rail Transit system on land use change and tied them into the future land maps. Also, Pacheco-Raguz [12] studied the impacts of the development of LRT in Manila on land price, land usage, and population size for the cities along the route. Using correlation and regression methods, the aforementioned variables are analyzed for an accessibility index and network distances obtained from a model built within a Geographic Information System (GIS). In addition, Higgins et al. [7] reviewed the previous literature on LRT and other rail rapid transit systems in

North America, demonstrating that rail transit alone is not a primary driver of land use change. Furthermore, Borchers et al. [4] provide a framework for selecting priority roads to implement bus lanes in medium-sized cities and the solution approach involves consideration of land use. As for the case in Japan, Sakamoto et al. [13] examined the change in urban population size before and after introduction of LRT in 27 cities in Europe. They also analyzed the changes from the population size or areas along the LRT route over time for four cities in Europe. Kriss et al. [9] conducted a study for Toyama City and pointed out that the LRT project in Toyama City got significant success based on the increasing ridership of the LRT. Sato et al. [15] predicted the population distribution in Utsunomiya City, Japan until 2050 by supposing different integration pattern of the LRT and feeder bus systems. Takasugi et al. [17] studied the differences between the BRT and LRT systems and the influence on urban population distribution for Maebashi City. Most of aforementioned researches are mainly concerned about the impacts on the cities after the introduction of public transportation systems.

It is worth to note that the findings of aforementioned researches were highly dependent on using questionnaire survey method. Questionnaire survey is time consuming and asks high cost for acquiring data. Today, cell phone is popular and becomes one of the required devices for people. Cell phone data are more competitive and complete than questionnaire survey data. Using cell phone data for studying the impacts of public transportation systems on related areas have earned more attentions in the near past decade. Widhalm et al. [22] proposed a method to reveal activity patterns that emerge from cell phone data by analyzing relational signatures of activity time, duration, and land use. Also, Wang et al. [21] provided a review of existing travel behavior studies by using mobile phone data. They discussed the potential of mobile phone data in advancing travel behavior research and raised some challenges that needed to be dealt with in the planning of traveling process. There are also papers that using the mobile phone data and taking the consideration of the relation between transportation and land use. Baird et al. [3] investigated how mobile device data can enhance regional and park-level transportation planning efforts by better answering questions about transportation system usage patterns. Also, Wang et al. [20] explored the influencing factors and spatial variations of subway trip origin and destination at the grid level.

Compared with traditional statistical modeling methods, machine learning methods in the AI area can be more powerful for prediction but could be weaker to explain the relationship between the response variable and explanatory variables due to an implicit functional form of hidden layers structure. Moreover, machine learning methods has the potential to automatically learn and update the predicted results through a well design system from updated data sets. Haenlein et al. [6] took a first insight into AI applications by summarizing seven articles published by several of the world's leading experts and specialists in AI that present a wide variety of perspectives on AI. When the relationships among features in data are not clear, the Association Analysis method is one of most competitive machine learning methods to identify the potential relationship between different features with less subjective assumptions. Tan et al. [18] proposed a general idea to set up models by different algorithms. By showing some examples with programming, one can easily understand how to implement Association Analysis method.

The IRL is another machine learning method which is applied in this research. The IRL is about studying from humans. In practice, IRL can be used to study an agent's objectives, rewards and values with the aid of using insights of its behavior. Arora et al. [1] mentioned about advantages and challenges about using IRL. You et al. [23] showed a great example of mobile application via using the IRL for data analysis. Kitani et al. [8] conducted analysis and prediction of walking behavior using IRL model.

2.2. Characterization of research

Most of existing researches are mainly concerned about the impacts on the cities after the introduction of public transportation systems. This research aimed to provide a reference for future planning based on the land use condition nowadays. Also, many existing researches related to the trips condition were based only on the personal trips survey but lack of abundant trial. Here, cell phone data was utilized and conducted the analysis as an estimated trip frequency. A new idea by integrating static and dynamic cell phone datasets from different sources with several analysis steps are proposed in this study for providing to understand the trip generation in the target city. Hereafter, personal trips data was also applied by machine learning methods to make insight analysis and comparison of the results. Therefore, simultaneous utilization of the static data and dynamic data can be a significant contribution of this study. The image of some examples in both the static and the dynamic data sources is reported in Figure 1. The static-type public facilities and boundary data and dynamic-type personal trip survey and cell phone signal data are used in this study. The machine learning methods of Association Analysis and IRL are applied for data analysis to obtain valuable information for making new policy planning by the government of the target city.

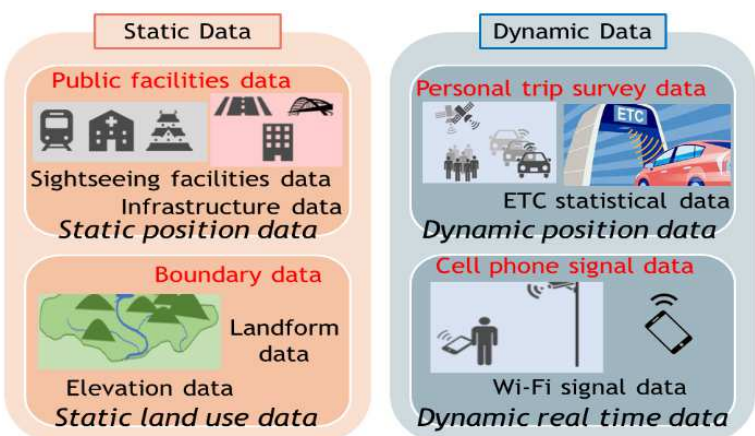


Figure 1: Image of data sources.

2.3. Research framework

Due to the availability of cell phone data from KDDI company, Utsunomiya City in Japan was selected as the target city in this study. Moreover, some open-source datasets

were also downloaded by showing on the map on ArcGIS software. To make the research results reliable, we implemented a data cleaning process to ensure data is correct and consistent. Two steps in Figure 2 are used for data analysis through using Association Analysis in Step 1. In order to make further understanding based on the findings of the Step 1, that is, to confirm the effect of the distance of different infrastructures, IRL was applied to traditional personal trip data in Step 2 to figure out the relation between trip purposes and selection for transportation modes. Finally, results were then shown and concluded. The data preparation and main calculation parts of the research flow are shown in below Figure 2.

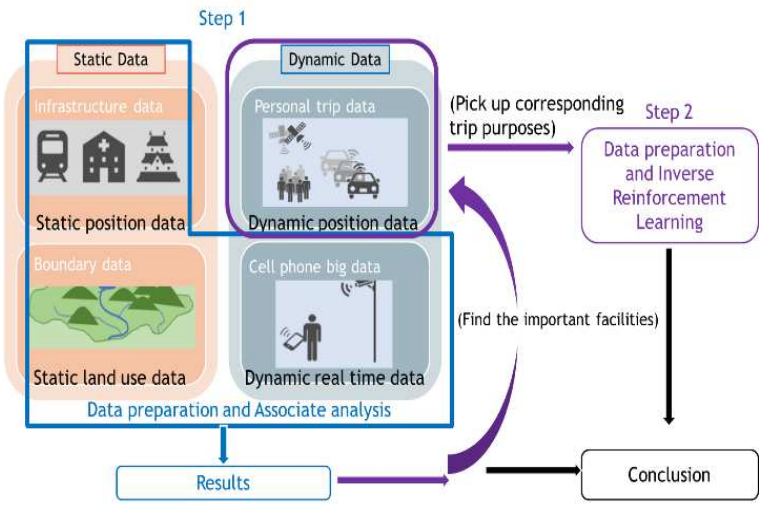


Figure 2: Flow chart of the research.

3. Case Study

3.1. Overview of Utsunomiya city

Utsunomiya City, located in the central part of Japan, is in Tochigi prefecture within the Kanto Region. The prefecture government is located in Utsunomiya City and is located approximately one hour away from Tokyo station. Utsunomiya City is an industrial city and the population size is 26th in Japan within 16 sub districts. Also, it is famous for the good environment for bicycle users. For instance, Japan Cup Cycle Road Race, which is held every October in Utsunomiya City is a well-known event. Nonetheless, by assimilating new challenges, Utsunomiya City is one of the leading suburban cities tackling declining birthrate and aging population (see Nishiyama [10]). The basic geography information of Utsunomiya City is shown in Figure 3.

3.2. The overview of transportation situation in Utsunomiya City

The Utsunomiya City is located in Tochigi Prefecture. Tochigi Prefecture has the second-high private vehicle ownership rate in Japan in 2018. One may find this result

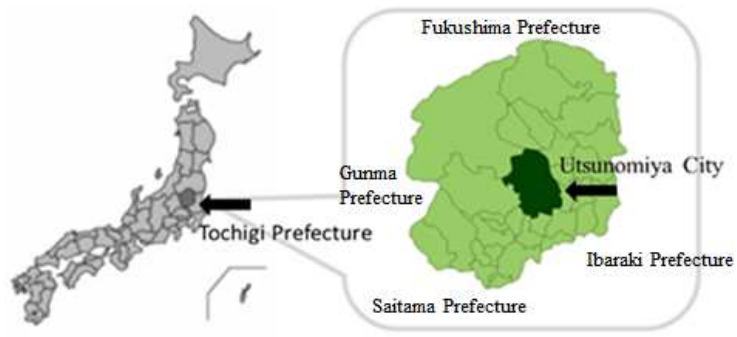
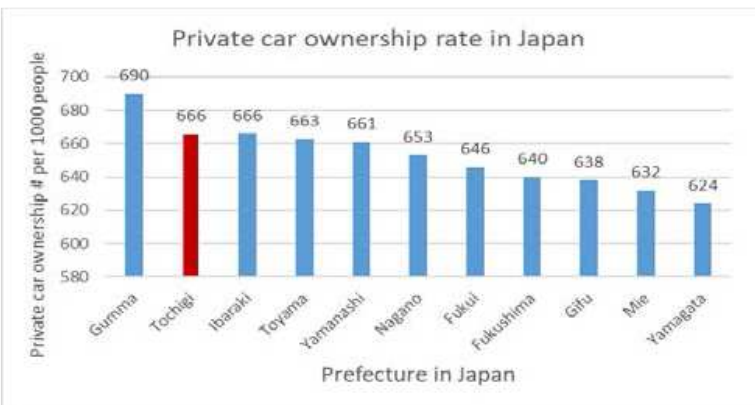


Figure 3: Geography information of Utsunomiya city.

based on Figure 4. Most citizens in Utsunomiya city still preferred to use private car (see Ohmori [11]). The new LRT system in Utsunomiya City is under construction, and is scheduled for operation in 2023. Since there are existing public transportation systems in Utsunomiya City, for example, bus systems and railway systems, the new LRT system in Utsunomiya City will surely serve as a good plan to integrate the public transportation systems. As a result, the potential to attract more citizens or tourists to select public transportation modes seems to be an important issue and therefore Utsunomiya City is chosen for this study.



*Data used: year 2018

Figure 4: Private car ownership in Tochigi Prefecture.

4. Introduction of Data and Data Preparation Process

4.1. Data and software utilized

Table 1 is the summary of source data from KDDI company and National Land Numerical Information download service. One is the cell phone data of Utsunomiya City from KDDI company. In the KDDI cell phone data set, the data was acquired in June

from 2016-2018. Here, de-identification trip data extracted from the GPS logs of cell phones with permission contributes to the first data set. The second data set is about the geographic information from NLNI (National Land Numerical Information). At the same time, personal trip survey data was also prepared for the purpose of broadening the research content.

Table 1: Introduction of two main datasets.

	Cell phone data	Shape-filed data
Data Source	KDDI CORPORATION	National Land Numerical Info.
Data Content	OD Points, Estimated Trip Numbers	Position of Infrastructures
Scale of Data	250m Mesh (JGD2011 Coordinate)	Prefecture, City
File Format	Excel File (.csv)	Point, Line (.shp)

In order to prepare the useful dataset, the original data is needed to be made some calculations. The point data is arranged to find the distance between each infrastructure and the central point of each mesh. The software used here is ArcGIS. ArcGIS is one of the GIS software provided by ESRI, similar to QGIS which can read data, create maps, and output. By the function in ArcGIS, this step can be calculated easily. Another software used in this research is the software related to Python programming. Python is a language that emphasizes code readability. Python has gained a lot of support in the field of deep learning because it is simple code, easy to read, and has abundant libraries that can be used for calculation and statistical processing. Jupiter Notebook is a web-based interactive computing environment that is common use for doing Python programming analysis.

4.2. Steps of data arrangement

After downloading the data from NLNI was put onto the map firstly. Total 10 infrastructures are selected in this study based on the consideration of data availability, see Table 2. The numbers of each facility in the target area are also shown in Table 2. At the same time, the shape files of mesh and the Utsunomiya City were also put onto the map. The central point of each mesh was defined by using the function called “polygon centroids”. Finally, the distance from the central point of each mesh to the nearest 10 infrastructures was calculated by using the function of “Near Tool”.

The estimated trips count from cell phone data was organized in order to know how many trips start or end at one specific mesh. The way of getting the estimated trips count is presented in Figure 5. Here, the estimated trips count was received from KDDI company and the calculation method was conducted by multiplying the expanding factor. For example, if the trips found is 20 and the percentage of users of this cell phone company is 10 percent. The estimated trips count would be 20/0.1. As a result, 200 will be the trips count been used in this situation. Although the judgement based on different distance and time may affect the results, 300m and 15 minutes were set here by the data processing process by KDDI company. Hereafter, since the data was available in

2016-2018, the frequency of trip increasing/decreasing was evaluated. We are interested in evaluate the trip frequency is increased or decreased in 2016-2018 and identifying the causes for the resulting trip frequencies. Finally, the mesh with the symbol of trip increasing was selected, see the flowchart of data processing in Figure 6.

Table 2: 10 infrastructures chosen for this research.

Elementary school (Elem) (*116)	Park (Park) (*884)
Middle school (Midd) (*52)	Attraction facilities** (Attr) (*56)
High school (High) (*24)	Newtown (New) (*27)
Culture facilities*** (Cult) (*254)	University (Univ) (*9)
Police station*** (Poli) (*79)	Hospital (Hosp) (*1074)

* The numbers of this infrastructure in target area
** Attraction facilities: Movie theater, citizen center, etc.
*** Culture facilities: Libraries, art museum, etc.

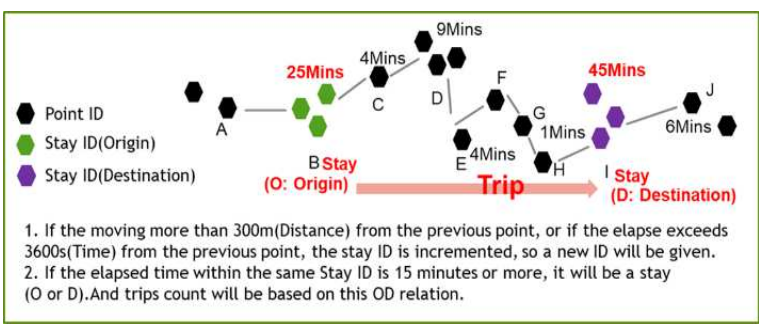


Figure 5: Concept of extracting the trip frequencies.

4.3. Trip generation condition in Utsunomiya City

The mesh areas within Utsunomiya City was focused in the research purpose of this study. The trips changing rate in each mesh in the target areas is calculated. Total 3,328 meshes in the target area with enough data information and the average trip changing rate from year 2016 to 2018 is 0.47. The overall outlook of the trips changing conditions in the target area can be visualized on the GIS map and displays in Figure 7. Although there is no obvious tendency of the trip changing condition in this map, one can find that several regions far from city center have a strong tendency of increasing trips, therefore several corresponding planning should be considered. Hereafter, the meshes were classified into different categories based on different trips changing rate. Figure 8 shows most of the meshes have the increasing rate. Even though the trips count was originally estimated number by multiplying the extracted trips number and the expanding factor so some error may exist, but the trend can be observed through these data. The red highlight

sets represent the target meshes that were selected to utilize for analysis in the Section 5.

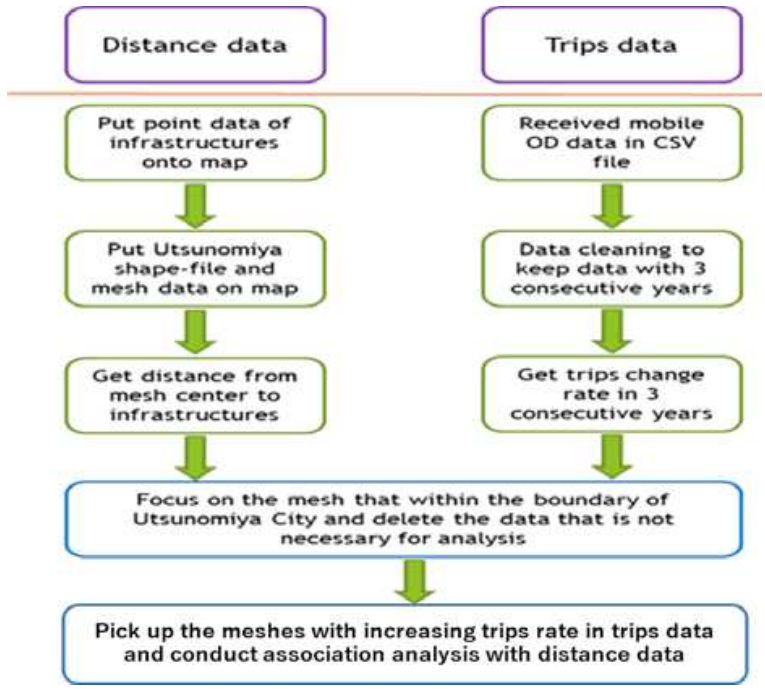


Figure 6: Steps of data processing.

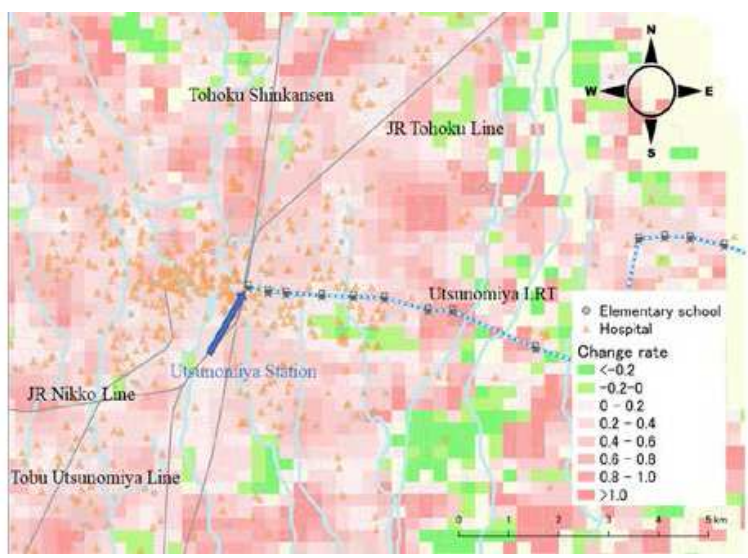


Figure 7: Trips changing condition in Utsunomiya City (2016-2018).

Figure 8: The number of meshes based on different trip changing rates.

5. Apply the Association Analysis

5.1. Concept and principles of Association Analysis

Association Analysis also named market basket analysis is an analytics technique which often conducted by retailers and consultants to understand the purchase behavior of customers. Association Analysis can be used to find interesting relationships in large datasets (see Bai et al. [2]). For market basket analysis, Association Analysis uses the purchase information or transaction data to leverage the effectiveness of sales and marketing. Association Analysis has been actively used in related industries once the immense amount of data has been available due to the creation of an electronic cashier system.

In this research, Association Analysis is conducted for trying to extract key features influencing the trip counts. The purpose is to find the potential relation between the distance from the mesh center to infrastructures and the trip changing condition in each mesh. More precisely, not only the combinations of the facilities with a high number of trips but the potential relations to the distribution of facilities with a high number of trips. Through constructing binary transaction data, it is possible to identify the hidden relationship of each item set within the dataset. Here, Association Analysis provides information on how strong the association rule, or the if-then relationship, is found to be true in the dataset (see Sarker [14]). The following equations are used to provide inferences on the cause-and-effect relationship.

$$\text{Support}(X) = |\{t \in T; X \subseteq t\}|/|T| \tag{5.1}$$

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{supp}\,(X \cap Y)}{\text{supp}\,(X)} \tag{5.2}$$

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{supp}\,(X \cap Y)}{\text{supp}\,(X) \times \text{supp}\,(Y)} \tag{5.3}$$

$$\text{Conviction}(X \Rightarrow Y) = \frac{1 - \text{supp}\,(Y)}{1 - \text{conf}\,(X \Rightarrow Y)} \tag{5.4}$$

where,

X and Y : independent item set

T : set of transactions

t : individual transactions, and

$X \rightarrow Y$: association rule that were under proposing.

In the above equations, the support provides how frequently the item set appears in the data set and is defined as the number of transactions divided by the total number of transactions. The confidence indicates how often the association rule itself was found valid and is defined as the proportion of both item X and Y being purchased to the total purchase of X. The lift value provides the strength of association rule and is defined as Equation (5.3). Lastly, the conviction provides the inference on randomness and is defined as the expected frequency that X occurs without Y.

5.2. Transformation of dataset

Data transformation was made to make the dataset available for implement Association Analysis. For example, the distances are re-scaling to mapping them into 0, 1 feature. The threshold of 1 kilometer was selected, if distance¡1km, it was coded as 1 and distance ¿1km was coded as 0. The mesh with a trip increasing rate¿20% was selected in this study for analysis. Totally 2,817 meshes were chosen for this analysis as shown in the red part of Figure 8.

5.3. Steps and results of Association Analysis

In order to reduce huge rules were generated via using Associate Analysis method for the data set, Figure 9 show several filtered steps are used to extract useful rules. The method of this part is trial and error in order to avoid too many rules but keep the enough rules for discussion. Through several trials, the final criteria that was chosen here for support value 0.1, confidence value 0.8, and lift value 1.0. Pick up the first 10 rules, for example, the part of results can be shown in the following Table 3 for saving space. The results of antecedents and consequents are shown. Overall, there are total 165 rules are obtained in this group.

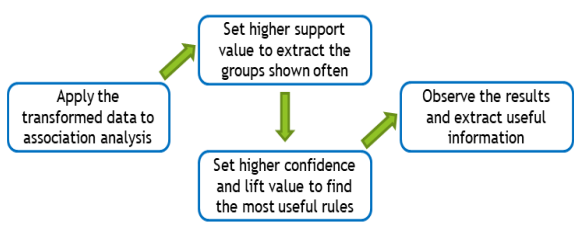


Figure 9: Steps of association analysis.

Through using Association Analysis, the numbers that how many times each infrastructure is shown in antecedents or consequents were known. This numbers in the

Table 3: The parts of filtered results from association analysis.

Rules#	antecedents	consequents	support	confidence	lift
51	({'newt'})	({'hosp'})	0.12	0.95	1.09
54	({'cult'})	({'hosp'})	0.40	0.95	1.10
57	({'poli'})	({'hosp'})	0.38	0.95	1.10
58	({'elem', 'midd'})	({'park'})	0.21	0.94	1.14
64	({'elem', 'attr'})	({'park'})	0.16	0.88	1.07
71	({'elem', 'high'})	({'park'})	0.12	0.97	1.18
76	({'cult', 'elem'})	({'park'})	0.27	0.96	1.17
83	({'elem', 'poli'})	({'park'})	0.29	0.96	1.17
88	({'elem', 'hosp'})	({'park'})	0.48	0.92	1.12
90	({'elem', 'park'})	({'hosp'})	0.48	0.95	1.10

contents of rules imply the trip frequency of each infrastructure. As a result, we find that hospital, elementary school, and park are the major locations where have a higher trip frequency for people, see Figure 10. Short summary here might be that the facilities for mandatory task will attract more people, such as hospital and elementary school.

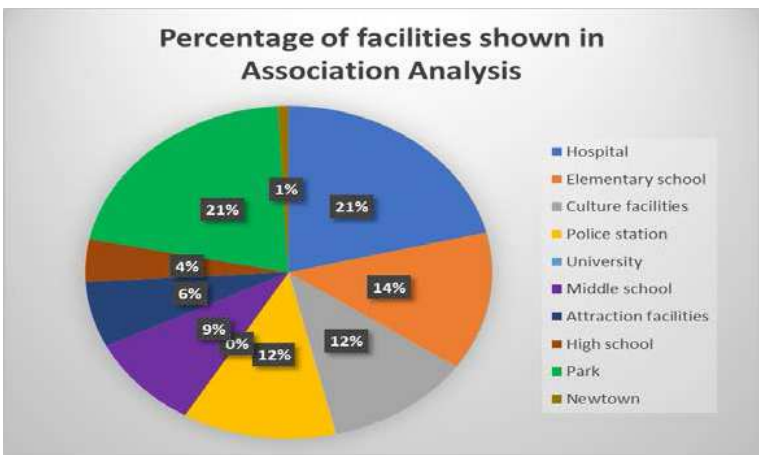


Figure 10: Percentage of each infrastructure shown in above Association Analysis.

6. Apply the Inverse Reinforcement Learning

The relations between trip generation and distance to infrastructures were obtained based on the previous analysis. In order to further understand the relationship between transportation modes and trip purposes, the machine learning method of IRL for traditional questionnaire survey data was also conducted.

6.1. Concept and principles of IRL

IRL method was selected as the analysis method for traditional questionnaire survey data. Reinforcement learning (RL) provides a powerful and general framework for decision making and control, but its application in practice is often hindered by the need for extensive feature and reward engineering. Deep reinforcement learning methods can remove the need for explicit engineering of policy or value features, but still require a manually specified reward function. IRL holds the promise of automatic reward acquisition (see Fu [5]). In the area of behavior analysis, it is difficult to precisely set rewards via using the RL method. The IRL method can be an alternative machine learning method to overcome the drawback of RL method. The decision-making structure of IRL can be flexibly updated. Moreover, IRL can perform iterative calculations based on the data and determine the structural pattern. Hence, IRL is particularly competitive for analyzing behavior for which the decision-making structure is unclear. As the IRL model estimates state value using continuous environment, it can consider the spillover effect. For example, if one state is often observed, IRL estimates the state value of the other states which is closely related to the observed state to be high as well as the observed state. As simple aggregation cannot capture this spillover effect, IRL can estimate more accurately than the simple aggregation.

In this study, the maximum entropy (ME) method was applied to create an IRL model. It is true that various IRL method have been proposed and Bayes method and optimal margin method are often used. However, these IRL method acquire the optimal action in each state. For activity estimation like this research, it is impossible to estimate or define optimal action in each state. On the other hand, ME method can calculate the optimal reward function by inputting the action trajectory even if the optimal action in each state is unknown. Therefore, ME method was used for IRL. The calculation procedure is given in the following: The first step is to make a multidimensional space with the dimensions of the number of elements to be considered. Hereafter, the initial values of the reward function are decided. The state values and measures are therefore optimized from the initial values. Also, based on the optimal state values and measures, the corresponding action trajectory that maximizes the reward is estimated. Finally, the reward function is compared with the observed behavioral trajectory. At the same time, the reward function is updated so that the difference between the two trajectories is minimized.

Let R in Equation (6.1) denote the reward function, which is assumed from the action trajectory ζ and updates the parameter θ. Evaluating R until θ converges to a constant value.

$$R(\zeta \mid \theta) = \theta^T \boldsymbol{f}_\zeta \tag{6.1}$$

where,

$R(\zeta)$: reward function

ζ : action trajectory

θ : parameter

$\boldsymbol{f}_\zeta$: assumed action trajectories.

Numerical calculation could be needed to obtain the value of θ that maximizes the log-likelihood function $L(\theta)$ in Equation (6.2):

$$L(\theta) = -\sum_i \log P(\boldsymbol{f}_{\xi_i} \mid \theta). \tag{6.2}$$

The function $P(\boldsymbol{f}_{\xi_i} \mid \theta)$ represents the probability of selecting a specific locus ξ_i. The gradient of the log-likelihood function is calculated as shown in Equation (6.3).

$$\begin{aligned}
\nabla L(\theta) &= \frac{\partial}{\partial \theta}\Big\{ \frac{1}{M}\sum_{i=1}^{M} \theta^T \boldsymbol{f}_{\xi_i} - \log \sum_{i=1}^{M} \exp \theta^T \boldsymbol{f}_{\xi_i} \Big\} \\
&= \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{f}_{\xi_i} - \frac{1}{\sum_{i=1}^{M} \exp \theta^T \boldsymbol{f}_{\xi_i}} \frac{\partial}{\partial \theta} \sum_{i=1}^{M} \exp \theta^T \boldsymbol{f}_{\xi_i} \\
&= \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{f}_{\xi_i} - \sum_{i=1}^{M} \frac{\exp \theta^T \boldsymbol{f}_{\xi_i}}{\sum_{i=1}^{M} \exp \theta^T \boldsymbol{f}_{\xi_i}} \boldsymbol{f}_{\xi_i} \\
&= \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{f}_{\xi_i} - \sum_{i=1}^{M} P(\xi_i \mid \theta) \boldsymbol{f}_{\xi_i}.
\end{aligned} \tag{6.3}$$

Finally, using this gradient, θ is updated by Equation (6.4) where α is the learning rate given exogenously.

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla L(\theta_{\text{old}}). \tag{6.4}$$

The steps and images of the above part is shown in Figure 11.

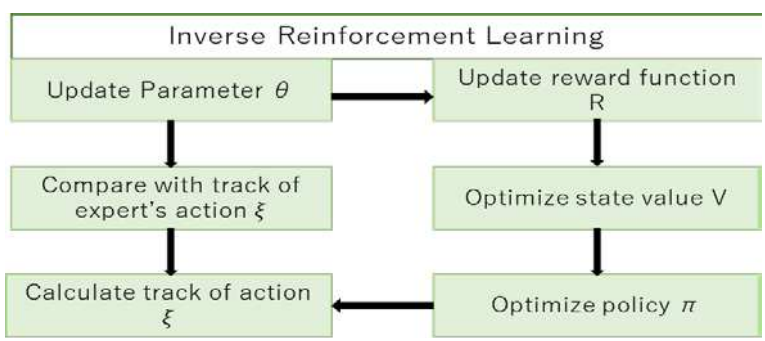


Figure 11: Calculation concept of IRL.

6.2. Data used and steps of IRL

The data of trip survey conducted in Tochigi prefecture in 2014 was used in this research. Focus on the data in Utsunomiya City, the data that origin place is within Utsunomiya City was extracted for the analysis here. The steps of IRL were conducted in the Jupyter environment through Python programming language. The variables of

this part, the trip purposes and transportation modes selected for analysis are displayed in Figure 12. Also, the time interval for the input was set at every 5 minutes as shown in Figure 13. The number of samples of the data applied to the IRL model is almost 800,000. In this research, training data and test data are not completely divided. When training the model, data are extracted in certain sample rate. As the sample rate was 1% or 0.1% in this research, the estimated value of the accuracy is considered to be almost the same as the out-of-sample estimation.

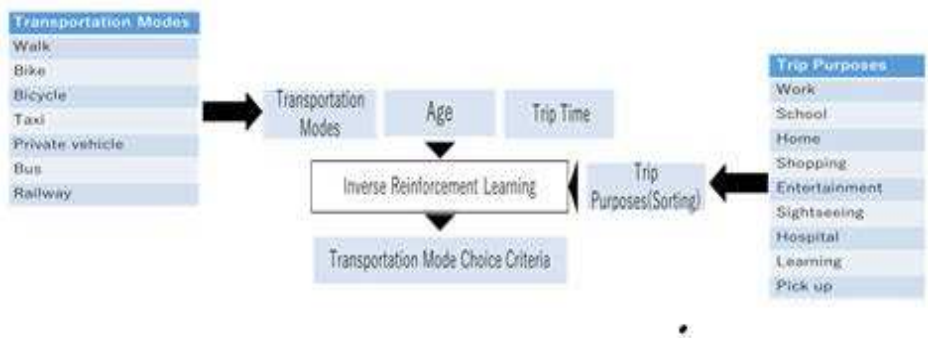


Figure 12: The variables selected for this analysis.

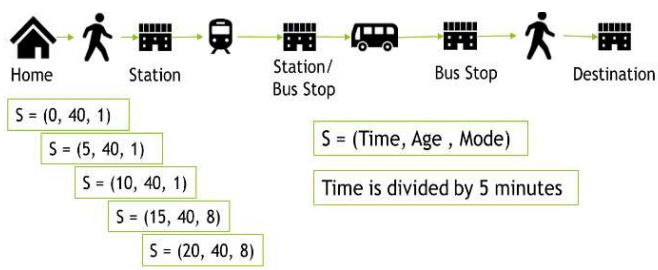


Figure 13: The time interval used here is every 5 minutes.

6.3. Results of IRL model

After analysis by using IRL method, one can get the tendency of many relationships. Figure 14 to Figure 17 display the pattern the transportation modes situation for the purposes of "go entertainment" and "go hospital" vs. different trip time and age, respectively. The higher state value represents that the tendency of selecting this transportation mode is higher. Valuable information can be found from these four figures. For example, Figure 14 and Figure 15 show that private vehicle is often selected when trip time is less than 30 minutes regardless of the purpose is to go entertainment or hospital. However, for longer trips or the trip time longer than 30 minutes, there is no dominate intension of choosing private vehicle, bus or railway to be found. From Figure 16 and Figure 17, we find that the elderly age group over 60 years old has very strong intension to select private vehicle for trips compared to other age groups. Possible reason here may be that middle age group was get used to private vehicle and it is difficult

to change their habits. Attracting more young people under 30 years old may be one possible way to increase the ridership of public transportation system since there were no strong tendency of choosing private vehicle or public transportation. On the other hand, middle age group from 30 to 60 years old will be more important if good design can attract them to change their habits and use the public transportation systems because they used private vehicle more often.

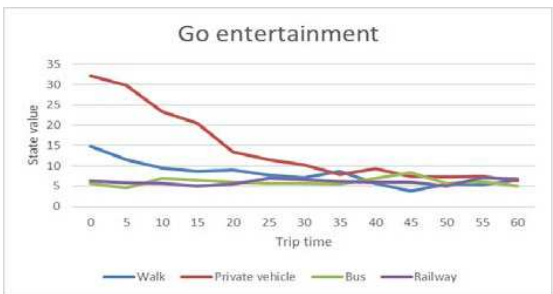


Figure 14: Transportation modes for entertainment purpose based on different trip time.

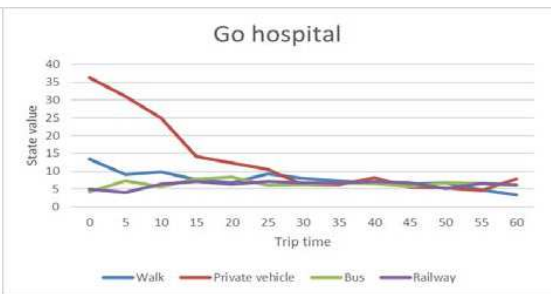


Figure 15: Transportation modes for hospital purpose based on different trip time.

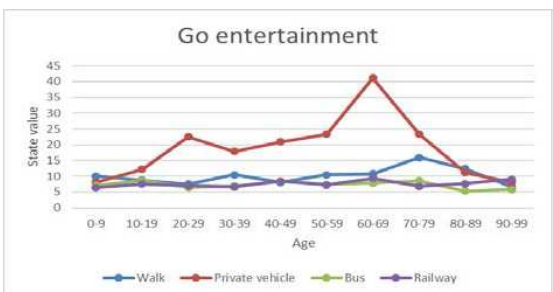


Figure 16: Transportation modes for entertainment purpose based on different age.

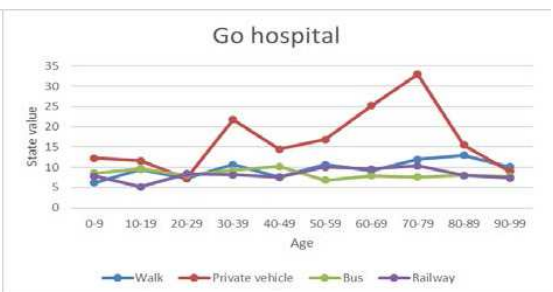


Figure 17: Transportation modes for hospital purpose based on different age.

6.4. Accuracy verification of IRL model

Since the application of IRL in this field is lack of enough information, the accuracy of the model should be checked after the results were acquired in order to verify the model. We compare our proposed model against the linear interpolation model. Table 5 shows how our proposed model outperforms the linear interpolation model. Here, the value represents the NLL (negative log-loss) index of error by the following formula (see Kitani et al. [8]).

$$\mathrm{NLL}(s) = E_{\pi(a|s)}[-\log \prod_t \pi(a_t \mid s_t)] \tag{6.5}$$

where s_t and a_t means a sequence of states and actions, generated by the moving path for a specific configuration of the dataset. After the calculation, IRL represents the value

of IRL model while Linear is the value of linear interpolation model. One can find that the precision accuracy of IRL model outperforms the linear interpolation model with a smaller value of index of error. Also, based on the above Table 4, the t-value test was also conducted for the reason to check the significance probability of this result. The average values of both IRL and Linear models were utilized here. Since the value of significance probability is below 0.05, the results here are good enough for taking reference.

Table 4: The NLL index of error of both model.

	IRL	Linear
Work	0.000	9.15
School	0.705	9.35
Home	0.201	8.78
Shopping	0.000	10.50
Entertainment	3.14	6.94
Sightseeing	5.75	6.20
Hospital	0.000	13.0
Learning	0.000	7.71
Pick up	0.000	8.48
Average	**1.09**	**8.90**

7. Conclusion and Further Implementation

7.1. Conclusion

To conclude, this research proposed a study to utilize both static-type and dynamic-type data for two goals. The first goal is to provide an analysis procedure to find the potential relations between infrastructures and trip generation. By using the cell phone data, the trip purpose in the target city was analyzed. Through association analysis, hospital, park and elementary school are important factors. That is, the locations with the above infrastructures had stronger tendency of increasing trip frequency in this case study in the Utsunomiya City. As a result, the future planning for transportation systems can be taken reference based on this result. The finding in this part that the facilities for mandatory task will attract more people might also be applied to the cities with similar scale. For example, integration of other new transportation systems or rebuilt the existing bus routes should consider this finding. Bike sharing systems and shared automated vehicle fleets in the future might also be possible ways. However, different cities could have different transportation cultures and different city scales. The Associate Analysis method with the proposed steps can be repeated in other target cities to find the important locations to increase trip frequency in the target city and make some comparison studies. Also, other factors such as off-peak hours or well-designed mobility hub should be taken into consideration.

The second goal of this study is to apply IRL method to find that the key factors for selecting transportation modes for different trip purposes using the data obtained from a traditional personal trip survey. In summary, we can find that private vehicle is often selected when the trip time is less than 30 minutes regardless of the purpose is to go entertainment or hospital. However, for longer trips with the trip time longer than 30 minutes, the difference of frequency of choosing private vehicle, bus or railway is not clear. Moreover, the elderly age group over 60 years old has very strong intension to select private vehicle for trips compared to other age groups. As the result, we may conclude that elderly age group was get used to private vehicle and it might be difficult to change their habits. Attracting more young people under 30 years old may be one possible way to increase the ridership of public transportation system. On the other hand, middle age group from 30 to 60 years old will be more important if good design can attract them to change their habits and use the public transportation systems. The government should consider the most suitable policy based on the city culture and population composition in the specific city.

7.2. Limitation and further implementation

The analysis conducted in this research serves as a primary approach to find some useful information behind the cell phone big data set and personal trip survey data by applying to the machine learning method of Associate Analysis and IRL. Nonetheless, due to the different amount and density of different infrastructures, further studies should be conducted in more detail analysis. Also, due to the limitation of datasets, further analysis should be conducted in order to verify the results or broaden the scope of the research if different datasets can be acquired. For example, using different machine learning methods for cell phone data and personal trip survey data can be a further study. Furthermore, because the cost for acquiring survey data is expensive, the years for obtaining the trip survey data and cell phone data are not overlap. Using overlap trip survey data and cell phone data for the proposed method can be another future study.

References

[1] Arora, S., and Prashant D. (2021). *A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress*, Artificial Intelligence, Vol.297, 103500. doi:10.1016/j.artint.2021.103500

[2] Bai, C., Zhou, L., Xia, M. and Feng, C. (2020). *Analysis of the Spatial Association Network Structure of China's Transportation Carbon Emissions and Its Driving Factors*, Journal of Environmental Management, Vol.253, 109765.

[3] Baird, T., Stinger, P., Cole, E. and Collins, R. (2022). *Mobile Device Data for Parks and Public Lands Transportation Planning: A Framework for Evaluation and Applications*, Transportation Research Record, Vol.2676, No.8, 490-500.

[4] Borchers, T. and Ribeiro, R. A. (2022). *A Framework for Selecting Bus Priority System Locations in Medium-Sized Cities: Case Study in Araraquara, Brazil*, Case Studies on Transport Policy, Vol.10, No.4, 2053-2063.

[5] Fu, J., Luo, K. and Levine, S. (2017). *Learning Robust Rewards with Adversarial Inverse Reinforcement Learning*, Conference paper at ICLR 2018. 2018/8/13. doi:10.48550/arXiv.1710.11248.

[6] Haenlein, M. and Kaplan, A. (2019). *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence*, California Management Review, Vol.61, No.4, 5-14. doi:10.1177/0008125619864925.

[7] Higgins, C., Ferguson, M. and Kanaroglou, P. (2014). *Light Rail and Land Use Change: Rail Transit's Role in Reshaping And Revitalizing Cities*, Journal of Public Transportation, Vol.17, No.2, 93-112. doi:10.5038/2375-0901.17.2.5.

[8] Kitani, K. M., Ziebart, B. D., Bagnell, J. A. and Hebert, M. (2012). Activity Forecasting. Computer Vision - ECCV 2012. Springer, Berlin, 201-214.

[9] Kriss, P., Miki-Imoto, H., Nishimaki, H. and Riku, T. (2020). *Toyama City: Compact City Development*, Word Bank, Washington DC. doi:10.1596/34816.

[10] Nishiyama, H. (2019). *The Rise in Vacant Housing in Post-Growth Japan: Housing Market*, Urban Policy, and Revitalizing Aging Cities. Springer, Singapore, Chapter 8. doi:10.1007/978-981-13-7920-8_8.

[11] Ohmori, N. (2017). *Future Perspective of Lifestyles and Mobility in Utsunomiya City*, Proceedings of the 2nd Join Conference of Utsunomiya University andUniversitas Padjadjaran, 280.

[12] Pacheco-Raguz, J. F. (2010). *Assessing the Impacts of Light Rail Transit on Urban Land in Manila*, Journal of Transport and Land Use, Vol.3, No.1, 113-138. doi:10.5198/jtlu.v3i1.13.

[13] Sakamoto, S., Morimoto, A. and Daimon, H. (2015). *A Study on Influence of LRT on Population Shift in European Various Cities*, Journal of the City Planning Institute of Japan, Vol.50, No.3, 774-779. doi:10.11361/journalcpij.50.774.

[14] Sarker, I. H. and Salim, F. D. (2018). *Mining User Behavioral Rules from Smartphone Data Through Association Analysis*, Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 450-461.

[15] Sato, T., Sasaki, T. and Chikuma, M. (2018). *Cost-Benefit Analysis of Developing A Light Rail Transit and Feeder Bus System in Utsunomiya City Considering The Change in Population Distribution*, Asian Transport Studies, Vol.5, No.1, 151-164.

[16] Takami, K. and Hatoyama, K. (2008). Sustainable City Regions. Springer, Tokyo, 183-200. doi:10.1007/978-4-431-78147-9_10.

[17] Takasugi, E., Sato, T., and Chikuma, M. (2018). *Development of The Method of Measuring The Benefits of Developing Light Rail Transit and Bus Rapid Transit Considering Differences of These Systems and Influence on Urban Population Distribution and A Case Study for Maebashi City, Japan*, Journal of the City Planning Institute of Japan, Vol.53, No.3, 1347-1347. doi:10.11361/journalcpij.53.1341.

[18] Tan, P. N., Steinbach, M. and Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley, Boston, Chapter 6.

[19] Thompson, G. L. (2003). Defining an Alternative Future: Birth of The Light Rail Movement in North America. No. E-C058. 2004/4/5. 25-36.

[20] Wang, C., Wang, X., Pan, R. and Yan, Y. (2022). *Influence of Built Environment on Subway Trip Origin and Destination: Insights Based on Mobile Positioning Data*, Transportation Research Record, Vol.2676, No.9, 693-710.

[21] Wang, Z., He, S. Y. and Leung, Y. (2018). *Applying Mobile Phone Data to Travel Behaviour Research: A Literature Review*, Travel Behaviour and Society, Vol.11, 141-155. doi:10.1016/j.tbs.2017.02.005.

[22] Widhalm, P., Yang, Y., Ulm, M., Athavale, S. and Gonzlez, M. C. (2015). *Discovering Urban Activity Patterns in Cell Phone Data*, Transportation, Vol.42, No.4, 597-623. doi:10.1007/s11116-015-9598-x.

[23] You, C., Lu, J., Filev, D. and Tsiotras, P. (2019). *Advanced Planning for Autonomous Vehicles Using Reinforcement Learning and Deep Inverse Reinforcement Learning* Robotics and Autonomous Systems, Vol.114, 1-18. doi:10.1016/j.robot.2019.01.003.

[24] Zhao, L. and Shen, L. (2019). *The Impacts of Rail Transit on Future Urban Land Use Development: A Case Study in Wuhan, China*, Transport Policy, Vol.81, 396-405. doi:10.1016/j.tranpol.2018.05.004.

Doctor's Program in The Dept. of Civil and Environmental Eng., Graduate Schools, Waseda University (3-4-1 Okubo, Shinjuku, Tokyo, 169-8555, Japan)

E-mail: tk16nthu44@aoni.waseda.jp (Corresponding Author)

Major area (s): Light rail transit, machine learning, transportation planning.

Master's Program in The Dept. of Civil and Environmental Eng., Graduate Schools, Waseda University (ditto).

E-mail: t0m4h1r1@moegi.waseda.jp

Major area (s): Light rail transit, machine learning, transportation planning.

KDDI CORPORATION/ KDDI Research Inc. (2-1-15 Ohara, Fujimino-shi, Saitama, Japan).

E-mail: ao-kobayashi@kddi.com

Major area(s): Location big data, deep learning, personality psychology.

Professor, Dept. of Civil and Environmental Eng., Waseda University (3-4-1 Okubo, Shinjuku, Tokyo, 169-8555, Japan).

E-mail: akinori@waseda.jp

Major area(s): Transportation planning, urban planning, traffic safety.

(Received August 2022; accepted October 2022)