Trip Destination Prediction by Cross-City Exploratory Data Analysis Approach in People Flow Data

Ryoichi Kojima
Human-Centered AI Laboratories
KDDI Research, Inc.
Saitama, Japan
ry-kojima@kddi.com

Roberto Legaspi
Human-Centered AI Laboratories
KDDI Research, Inc.
Saitama, Japan
xre-roberuto@kddi.com

Shinya Wada
Human-Centered AI Laboratories
KDDI Research, Inc.
Saitama, Japan
sh-wada@kddi.com

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020611

***Abstract*—Understanding human mobility is mostly based on destination prediction. The reality that training and test data frequently differ makes it challenging to broaden the application of destination prediction, and when a prediction model that was trained on certain areas is then applied to the area of interest but with a different data distribution, the accuracy is suboptimal at best. The objective of the IEEE Big Data Cup 2022 is to solicit a robust and generalizable model that can effectively forecast a person's daily destination based on facts and qualities obtained from four Japanese urban areas, as well as to predict the destination for a new metropolitan area using the models that were trained with the prior data. To address the issue and face the challenge of this Cup, our KDDI Research team has developed a prediction method based on Exploratory Data Analysis (EDA) that does not employ geographical zone information that varies from area to area. Instead, we employ the mobility characteristics of human groups, where each group is categorized according to demographics, which according to our EDA are relatively universal across areas. Our experimental findings show that our method achieves a relatively good prediction accuracy, which is attested by this Cup's leaderboard. Albeit our method is simple and straightforward, we can argue that our approach can be used as good baseline for human mobility destination prediction.**

***Index Terms*—Trip Destination Prediction; Exploratory Data Analysis**

I. Introduction

Research on human mobility has a wide range of potential applications, which include, inter alia, commercial applications such as location-based services [1], traffic congestion prediction in urban planning [2] [3] [4], and evacuation route planning in the event of an emergency or disaster [5] [6] [7]. More importantly, destination prediction is key to understanding human mobility. One of the factors that makes it difficult to expand the scope of the application of destination prediction in human mobility is the fact that training and test data often differ and IEEE's BigData Cup 2022, with the objective *Trip Destination Prediction*, also sheds light on this issue. In general, when a prediction model that is trained on one area is then applied to another area with a different data distribution, the accuracy is often not as good one would expect. Consequently, there are two goals outlined by this Challenge, namely, to build a strong and generalizable model to forecast an individual's daily destination based on easily gathered facts and attributes utilizing data collected from several Japanese urban areas and, secondly, to predict the destination for a new metropolitan region using the models trained on prior data.

To address this problem, we have devised a prediction method that does not use geographical zone information (e.g., population density, number of offices, or longitude and latitude) that differs from area to area. Instead, our method uses mobility characteristics of human groups in which each group is classified based on demographics, such as gender, age, occupation, which are population characteristics that our exploratory data analysis confirmed to be highly universal regardless of the area. For instance, children who are under 10 years of age would need to stay within the same zone when they go out by themselves, housewives return to their own initial, original zones, and individuals who are collectively classified as part of the work force move to many different zones. In other words, while pertinent areas change, the observed behaviors based on human traits and characteristics are highly common regardless of whether, say, it is in Japan's Tokyo or Kinki areas. Our experimental results provide evidence of the good prediction accuracy achieved by our method and as attested by this challenge's leaderboard. We believe that our method, albeit simple, has the potential to be employed as a universal baseline in human mobility prediction.

II. Challenge Dataset

For this challenge, the People Flow Data [8] [9] [10] is given. People Flow Data, which is comprised of spatio-temporal data of entities processed from multiple data sources, is used for monitoring dynamic changes in people's mobile behavior. This utility primarily offers datasets that were processed using the Person Trip Survey Data that was gathered by a nation, which in this case is Japan, or a local government

or municipality in each location. The various area municipalities that collected the Person Trip Survey Data include Tokyo, Chukyo, Kyushu, Higashisurugawan and Kinki.Japan's Ministry of Land, Infrastructure, Transport and Tourism, the Metropolitan Area Traffic Planning Association of each area, and the Japan International Cooperation Agency have all given their approval for the usage of these data sources. Generally, the steps that comprised the data processing include geocoding the start and end points of sub-trips to identify spatio-temporal locations, figuring out the shortest path between two locations, and finally, interpolating minute-by-minute position data based on comprehensive network data.

Each record in the People Flow Data, which represents a daily trip in the metropolitan area, consists of individual ID, demographics (such as age, gender, and occupation), and the origin and destination when moving. This dataset employs zones as the fundamental spatial units to evade reproducing sensitive, albeit significant, locations (i.e., home and office locations) due to privacy issues. In addition to the shapefile that provides the cartographic depiction of the zones, an open data-based corpus that represents each zone's population and number of employees and offices in both secondary and tertiary industries is also provided.

Among the five area datasets, the one from Kinki is used as test data for prediction and the ones from the other four areas consist the training data. As mentioned above, each record includes the representation of an individual's departure time, origin zone ID, and destination zone ID. In this Challenge, the Kinki destination zone ID is the one used as prediction target. Since it is unlikely that the distribution of occupations varies so widely by area, we can surmise that the occupations represented by integer labels differ by area.

III. Exploratory Data Analysis

In statistics, exploratory data analysis (EDA) is a method of examining data sets to highlight their key features, frequently utilizing statistical graphs and other techniques for data visualization. EDA differs from traditional hypothesis testing in that it is primarily used to explore what the data can inform us that is beyond the formal modeling. EDA has been championed since 1970 in order to persuade statisticians to investigate the data and perhaps develop hypotheses that could spawn additional data gathering and experiments. In this paper, however, we leverage EDA to find the rules that can be used to predict trip destinations from analyzing individual demographic attributes.

A. Age and occupation analysis

We labeled the age and occupation attributes using integer values. Fig. 1 shows the the age labels that were plotted according to areas. From the shape of the distributions and the range of labels, we can plausibly infer that age can be labeled with integer values in 5-year increments, e.g., integer 0 for under 5 years old, integer 1 for 5 to 10 years old, and so on. As for the occupation labels, we removed any integer label that is above 17, which appeared to be outliers, and we obtained the histogram shown in Fig. 2. Since it is unlikely that the distribution of occupation would vary so widely according to area, we can infer that the occupations represented by the integer labels differ by area. Hence, referring to the WebAPI specification [11], we can observe a correspondence, albeit incomplete, between the occupation and its integer labels.

We then counted the age and occupation labels and extracted only the resulting rows whose counts are most beneficial when estimating what the occupation label is, i.e., we assume that the row with the largest count, and therefore indicative of the majority for that age group, provides the most viable hypotheses. For instance, if we consider Table I that specifically shows the data counts we gathered from the Tokyo area, we can infer the following: occupation label 11 would refer to junior high school students or lower, label 12 would correspond to high school students, label 13 would be college students, label 14 would be housewives (or househusbands), label 15 would be the unemployed elderly individuals, and the rest would be the working adults. We followed this way of reasoning as analyzed the data counts in the datasets of the other areas (in Tables II, III and IV).

TABLE I
DATA COUNT IN TOKYO

Age	Occupation label	count
5-9	11	38646
10-14	11	37910
15-19	11	4412
15-19	12	17809
15-19	13	7002
20-24	13	13843
50-54(Male)	14	52
50-54(Female)	14	8223
80-84	15	6654

TABLE II
DATA COUNT IN CHUKYO

Age	Occupation label	count
5-9	13	13716
10-14	13	14598
15-19	13	1692
15-19	14	6544
20-24	14	2248
50-54(Male)	22	
50-54(Female)	2033	
80-84	16	3616

TABLE III
DATA COUNT IN HIGASHISURUGAWAN

Age	Occupation label	count
5-9	11	571
5-9	12	1197
10-14	12	1910
15-19	12	298
15-19	13	926
20-24	14	73
50-54(Female) 15	245	
80-84	16	439

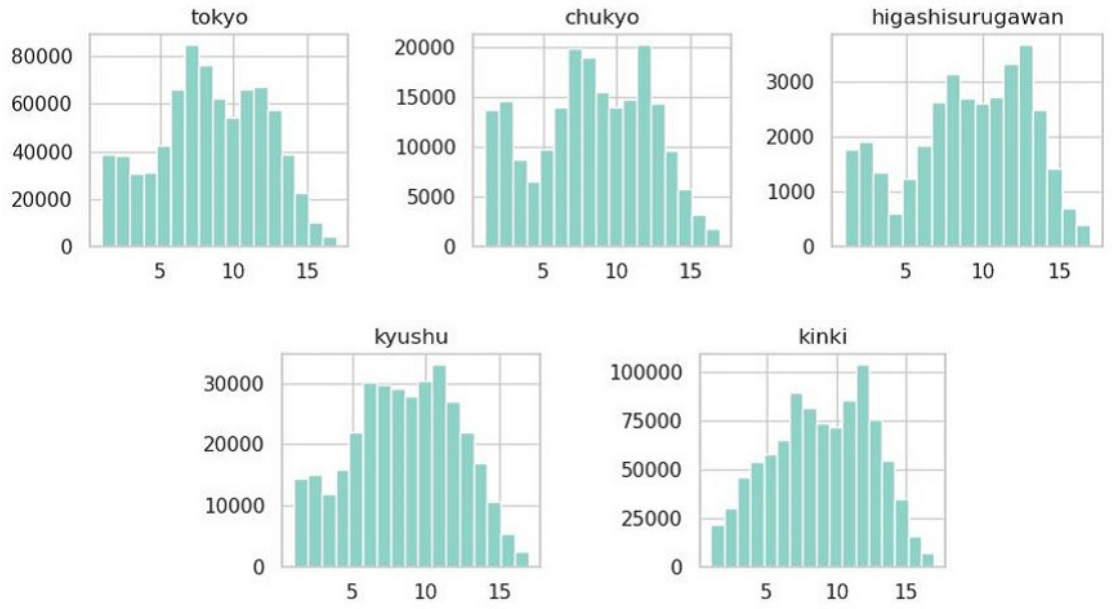

Fig. 1. Age label distribution in each area

TABLE IV
DATA COUNT IN KYUSYU

Age	Occupation label	count
5-9	11	14409
10-14	11	14944
15-19	11	2084
15-19	12	8574
20-24	12	3908
50-54(Male)	13	23
50-54(Female)	13	4827
80-84	14	4136

TABLE V
DATA COUNT IN KINKI

Age	Occupation label	count
5-9	13	21866
10-14	13	29932
15-19	13	3413
15-19	14	40636
20-24	14	25639
50-54(Male)	15	57
50-54(Female)	15	6736
80-84	16	5452

For confirmation, the departure times of the inferred occupations are plotted in Fig. 1. Since the distribution of these departure times is convincingly consistent with common sense departure times for each occupation, the new labels are generally inferred to be reasonable. Thus, counting for the other areas in the same way and inferring the occupation, we assigned the following new labels as common labels for all areas:

0 : Worker
1 : Junior high school or lower
2 : High school or college
3 : Housewife or househusband
4 : Unemployed

and we lay down the specific occupation label conversion rules in Table VI.

TABLE VI
OCCUPATION LABEL CONVERSION RULES

Area	Original label	New label
Tokyo	11	1
	12, 13	2
	14	3
	15	4
	the others	0
Chukyo	13	1
	14	2
	15	3
	16	4
	the others	0
Kyusyu	11	1
	12	2
	13	3
	14	4
	the others	0
Higashisurugawan	11, 12	1
	13, 14	2
	15	3
	16	4
	the others	0
Kinki	13	1
	14	2
	15	3
	16	4
	the others	0

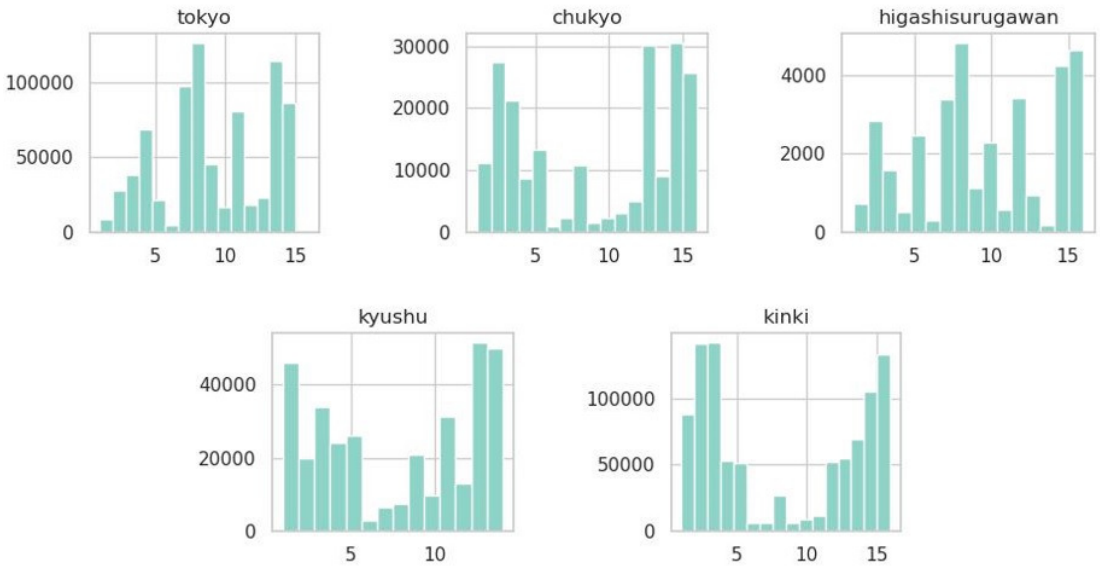

Fig. 2. Occupation label distribution in each area

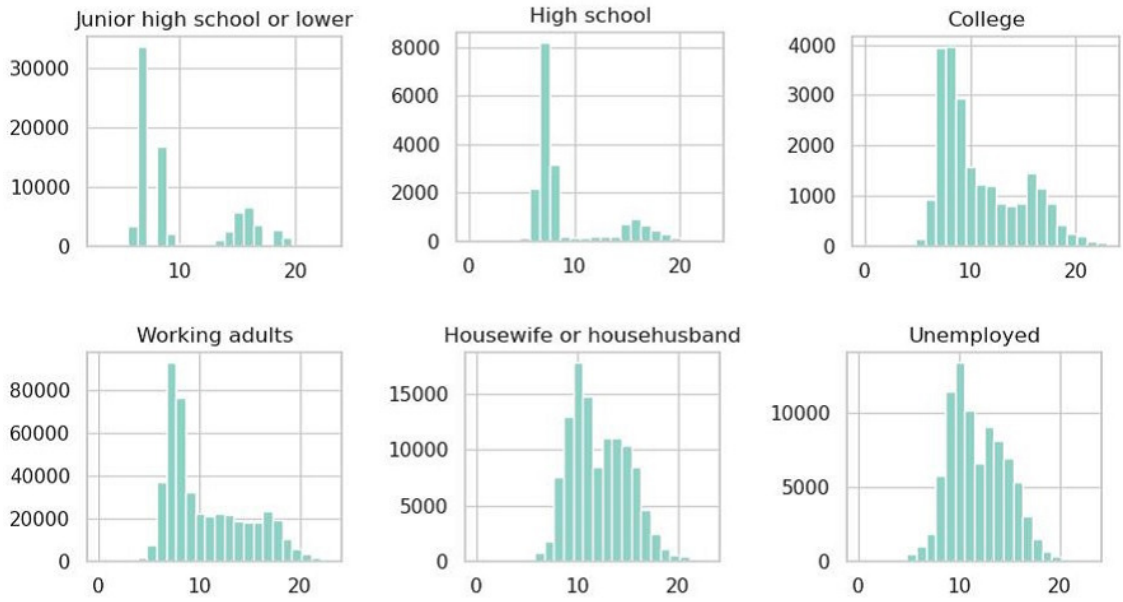

Fig. 3. Departure time distribution in occupation

B. Trip-graph analysis

We visualize a person's trip using a connected directed graph. We call the visualization of the mobility changes in the trip between zones as *Trip-graph*. The node represents the zone, and the number in it is the zone ID. The directed edges represent movements between zones. Fig. 4 shows Trip-graph examples for three individuals distinguished by their corresponding person ID (Pid). Here, Pid 63138 continues to go about within the same zone 4210, Pid 456008 moves across zones from 30221 $\rightarrow$ 30240 $\rightarrow$ 30200 $\rightarrow$ 30200 $\rightarrow$ 30221, and Pid 275715 also moves along three zones. From these people movement flow, we can make the following inferences:

- a trip may be confined within the same zone (self-looping),
- a trip may return to the zone of origin , or
- a trip may return to the previous zone.

Let us take for instance the ratio with which the self-looping trip occurs in the different areas, as shown in Tables VII, VIII and IX. By performing EDA, we observed that about 27% of the 790,613 trips in the Tokyo data are self-looping. Assuming a person simply stays in the same zone, we can still achieve

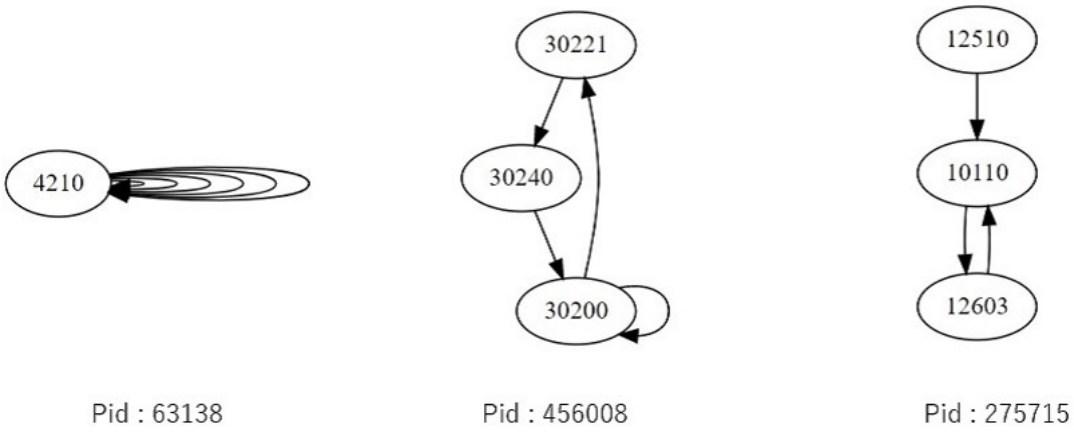

Fig. 4. Trip-Graph examples

the predictive accuracy shown in Table VII. Given for instance that there are actually 1,316 zones in Tokyo, if we predict trip destinations at random, the prediction accuracy would be 1/1316 (about 0.00076). However, if know that there is a 27% probability that the person stays in the same zone, then the prediction accuracy is also around 0.27 by predicting the same zone. We hypothesized then that the great percentage of the trips in the test data (Kinki area) is also self-looping, which means that the ratio is equivalent to the prediction accuracy. We maintained this similar flow of reasoning for when a trip returns to the zone of origin or to the previous zone. The following section shows how to combine several of these high-occurrence trip features to finally achieve high accuracy.

TABLE VII
STAYING IN THE SAME ZONE RATIO

Area	Ratio
Tokyo	0.26753
Chukyo	0.25540
Kyusyu	0.27736
Higashisurugawan	0.30838

TABLE VIII
BACK TO FIRST ORIGIN ZONE RATIO

Area	Ratio
Tokyo	0.25424
Chukyo	0.24887
Kyusyu	0.28375
Higashisurugawan	0.30658

TABLE IX
BACK TO PREVIOUS ORIGIN ZONE RATIO

Area	Ratio
Tokyo	0.31448
Chukyo	0.26983
Kyusyu	0.66990
Higashisurugawan	0.31840

IV. TRIP DESTINATION PREDICTION BY CROSS-CITY EXPLORATORY DATA ANALYSIS

Based on the analysis we elucidated in the previous section, we devised the following simple rule-based prediction method:

- Assuming that the trip graph is not split into sub-trips, the next origin zone in the graph of the same person is the previous destination zone.
- If the occupation is housewife or househusband, or unemployed, the final destination zone of the trip graph returns to the first origin zone of the trip graph; otherwise, it returns to the previous zone.

Using the Kinki area dataset for testing the performance of our EDA, the accuracy with our method is 0.43757. In comparison to the other competitors in this Challenge, the methods of the first and second (top two leaders) placers in the competition that preceded us have reported accuracy of 0.43967 and 0.43838, respectively. Thus, although our results are arguably optimal, it is still relatively effective when contrasted with the other methods that were used by the ones who followed us in this Challenge. Our prediction rules can be viewed as properties that emerged following our EDA, and a possible next work that seems computationally sound is to find a method that can automatically discover the prediction rules that emerged from our EDA.

V. CONCLUSION

Although there have been several studies of the characteristics of a trip [12] [13] [14], we created a simple rule-based prediction model with trip-graphs characterized by people's occupations as discovered by our exploratory data anaysis.

Initially, we also tried to create a distance representation among zones and combined this with other geographical information, such as population, and tried trip destination prediction using nearest neighbor search. However, our use of only a geographical information model led to a very poor prediction accuracy (0.04672).

Moving forward, we aim to investigate the viability of creating a model that leverages the combination of geographical and demographic information to improve accuracy. One idea is to use a graph neural network model to predict links in the trip-graph that was created from the obtained demographic information. For instance, to predict links via transfer learning from the Tokyo, Chukyo, Kyushu and Higashisurugawan areas to the Kinki area.

REFERENCES

[1] "Kddi location analyzer," https://k-locationanalyzer.com/en/.

[2] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. yue Wang, "Traffic flow prediction with big data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 865–873, 2015.

[3] J. S. Naik, N. Kasiviswanath, K. I. Ahamed, and S. R. Reddy, "A survey on traffic flow prediction with deep learning algorithms on big data," *International Journal of Recent Technology and Engineering*, vol. 7, 2019.

[4] W. Jiang and J. Luo, "Big data for traffic estimation and prediction: A survey of data and tools," *Applied System Innovation*, vol. 5, 2022.

[5] T. Horanont, A. Witayangkurn, Y. Sekimoto, and R. Shibasaki, "Large-scale auto-gps analysis for discerning behavior change during crisis," *IEEE Intelligent Systems*, vol. 28, 2013.

[6] X. Song, Q. Zhang, Y. Sekimoto, T. Horanont, S. Ueyama, and R. Shibasaki, "Intelligent system for human behavior analysis and reasoning following large-scale disasters," *IEEE Intelligent Systems*, vol. 28, pp. 35–42, 2013.

[7] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki, "Prediction of human emergency behavior and their mobility following large-scale disaster." Association for Computing Machinery, 2014, pp. 5–14.

[8] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui, and Y. Shimazaki, "Pflow: Reconstructing people flow recycling large-scale social survey data," *IEEE Pervasive Computing*, vol. 10, 2011.

[9] T. Kashiyama, Y. Pang, Y. Sekimoto, and T. Yabe, "Pseudo-pflow: Development of nationwide synthetic open dataset for people movement based on limited travel survey and open statistical data," 2022. [Online]. Available: https://arxiv.org/abs/2205.00657

[10] "People flow project," https://pflow.csis.u-tokyo.ac.jp/data-provision-service/about-people-flow-data/.

[11] "Web api document," https://pflow.csis.u-tokyo.ac.jp/wp-content/uploads/webapi.pdf.

[12] W. Nakanishi, H. Yamaguchi, and D. Fukuda, "Feature extraction of inter-region travel pattern using random matrix theory and mobile phone location data," vol. 34, 2018.

[13] E. Bhaduri, B. S. Manoj, Z. Wadud, A. K. Goswami, and C. F. Choudhury, "Modelling the effects of covid-19 on travel mode choice behaviour in india," *Transportation Research Interdisciplinary Perspectives*, vol. 8, 2020.

[14] Y. Hara and H. Yamaguchi, "Japanese travel behavior trends and change under covid-19 state-of-emergency declaration: Nationwide observation by mobile phone location data," *Transportation Research Interdisciplinary Perspectives*, vol. 9, 2021.