**Midterm Report for Trip Destination Prediction Problem**

*Yicheng Wang 101333295 Wenbin Wang 101191168*

# **1 Introduction**

This project aims to solve the Trip Destination Prediction problem [1]. This problem needs to predict the trip destination in a new area, called Kinki in Japan by using the four other Japanese zone trip trace records. This report outlines our progress and approach.

We divided the problem into two main tasks: (1) classifying whether a trip will be round-trip, and (2) predicting the destination if it’s not. By the midterm, we completed the development of a round-trip classification model. Inspired by the paper, *An Embarrassingly Simple Rule-based Visiting Circulation Approach to Trip Destination Prediction* [2], which assumes that most travellers return to their origin after a trip.

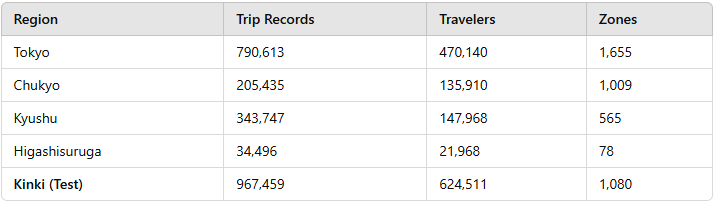
The original method used simple rules based on Pid and Departure\_time to set destinations, achieving 43% accuracy. However, 57% of the trips did not follow this pattern. We improved the approach by incorporating traveller information, such as trip type, to more accurately predict round trips. This provides a strong foundation for further destination predictions.

# **2 Exploratory data visualization**

## **2.1 Dataset Overview**

The dataset includes two parts of data, which is the train data and the test data. The first part is the travel records. Those data contain information about the traveller and the trip proposal. This part of data includes: Pid, Departure\_time, Gender, Age, Occupation, Trip\_type, Origin, and Destination. The target label in the training data is Destination.

The training data comes from four areas in Japan:



Due to geographical characteristics, using an entire area as the validation dataset is preferable. Given the scale and similarity to our test data, Tokyo has been chosen to validate our model.

## **2.2 Feature distribution comparison**

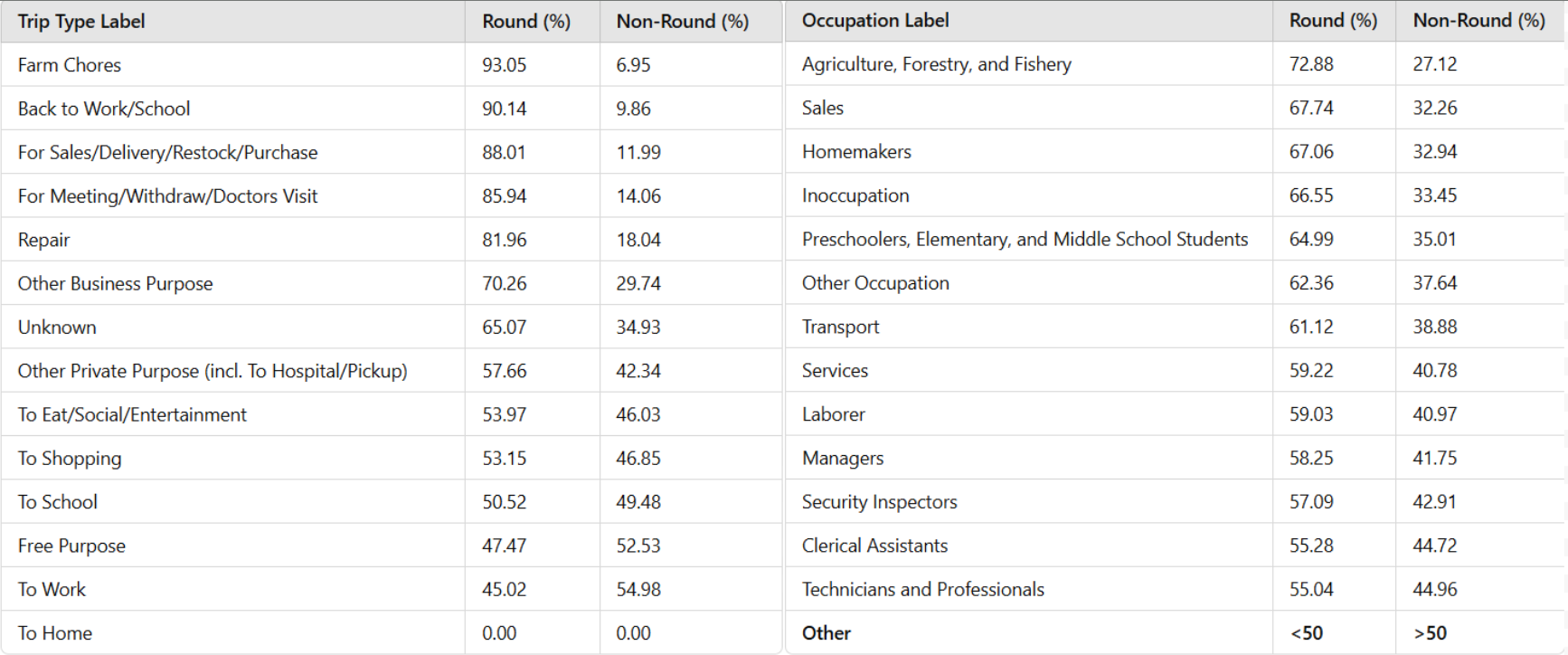
We began by comparing the distribution of features in the training and test datasets, focusing on the following aspects: Gender, Age, Occupation, Trip Type.

To analyze departure times, we divided the departure time into three time periods: morning (0:00 - 7:59), noon (8:00 - 15:59) and evening (16:00 - 23:59), and compared the distributions in both datasets.

Our analysis reveals notable differences in Trip Type and Occupation between the training and test data. The most common trip type in the training data is “To work” (29%), whereas in the test data, “To Eat/Social/Entertainment” is the most frequent (13%). To align the training data more closely with the test data, we plan to increase the representation of “To Eat/Social/Entertainment” labels and decrease the “To work” labels in our training set.

## **2.3 Distribution of round-trip and non-round-trip labels**

We further analyzed the distribution of labels for round-trip and non-round-trip records, finding significant differences in both Trip Type and Occupation between these categories.



For the trip\_type, most trips categorized as “Farm Chores,” “Back to Work/School,” “For Sales/Delivery/Restock/Purchase,” “For Meeting/Withdraw/Doctors Visit,” and “Repair” were round trips. In terms of Occupation, individuals in “Agriculture, Forestry, and Fishery” were more likely to make round trips, while those who labelled as “High School Students and above” tended to take non-round trips.

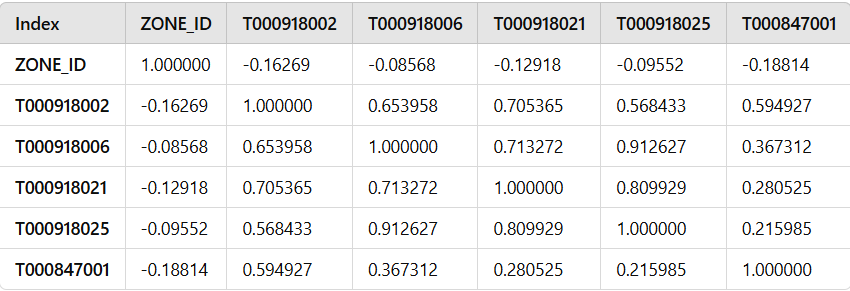
These insights help in understanding travel patterns across different occupations and trip types, which will be beneficial for refining our model.

# **3 Exploration of geographic data**

We conducted a detailed analysis of the geographic data, dividing it into two parts: human geography data and spatial geography data.

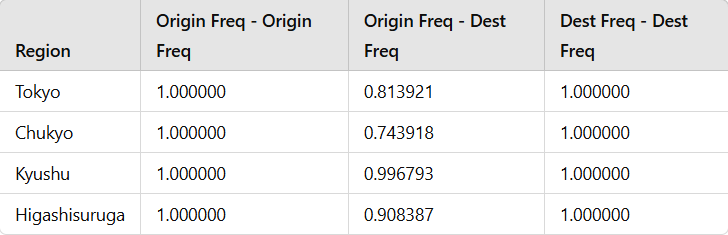
## **3.1 Human geography data**

This including zone number ZONE\_ID, number of commercial facilities, number of employees, and nighttime population. Since most human geography data is continuous, we calculated pairwise correlations to identify relationships between features. For example, we observed a strong correlation between the number of employees in the secondary and tertiary industries (T000918006 and T000918025). Based on this, we consider merging them or selecting one of them as a feature in subsequent classification.

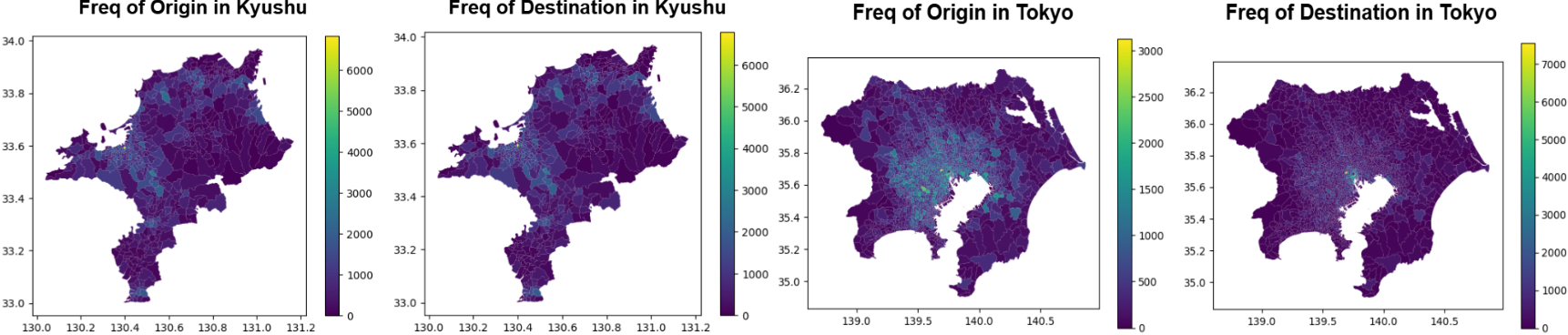


## **3.2 Geospatial data**

This includes the ZONE\_ID and POLYGON of each region. The POLYGON class contains the information about the centre of boundaries of each region. We use these polygons to create visualizations of the region to better display the distribution of data related to geographic location and human geography data.



We also analyzed the frequency distribution between departures and destinations. We found that in areas such as Tokyo, Chukyo, Kyushu, and Higashi-Suroga, popular departures are also common destinations. For example, in Tokyo, the correlation coefficient between departure and destination frequencies is 0.813921, while in Kyushu, this correlation is as high as 0.996793. This shows that most travelers tend to return to their departure points at the end of their trip.



To further improve the prediction ability of the model. We use the departure point in the test data as a potential prediction target area. In addition, according to the analysis of the training data, more than 77.86% of the destinations are within a 0.1 radius of the departure point. Therefore, we can set the 0.1 radius as the candidate area range of the prediction model. This helps to optimize the prediction of the destination.

# **4 Round-Trip classification**

We treated it as a classification problem and used decision trees and logistic regression models. We use one hot encoding to represent trip type, occupation and departure time. We also add three bool attributes(“whether it is a multi-trip”; “whether it is the head of a multi-trip”; “whether it is the end of a multi-trip”). We splitted our train dataset into random train and validate subsets by 0.25. Here is our model performance:

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Data Accuracy** | **Validation Data Accuracy** |
| **Decision Tree** | 95.27% | 79.47% |
| **Logistic Regression** | 77.67% | 77.68% |

Because it is easy to predict the destination for a round-trip, when accuracy is at the same level we hope the [False](https://en.wikipedia.org/wiki/False_positive) negative to be minimum. Here are models’ [False](https://en.wikipedia.org/wiki/False_positive) negatives.

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Data FN** | **Validation Data FN** |
| **Decision Tree** | 32828 | 37140 |
| **Logistic Regression** | 78400 | 26601 |

# **5 Future Work**

Next step, we will proceed with destination prediction for non-round trips. Building on the results from our round-trip classification model. We will incorporate methods such as the gravity model and nearest-neighbour search to enhance destination prediction accuracy. With these steps, we will finalize the destination prediction of our project.

# **6 Conclusion**

In this phase of the project, we successfully implemented and validated a round-trip classification model. We ultimately chose logistic regression because the decision tree performed worse on [False](https://en.wikipedia.org/wiki/False_positive) negative. This two-step modeling process not only helps in accurate trip classification, but also provides a basis for final prediction.

**Reference**

[1] *E.-S. Tu, Y.-H. Chen, E.-C. Liu, H.-Y. Keng, and C.-T. Li, “An Embarrassingly Simple Rule-based Visiting Circulation Approach to Trip Destination Prediction,” in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 6565–6572. doi:* [*10.1109/BigData55660.2022.10020650*](https://doi.org/10.1109/BigData55660.2022.10020650)*.*

[2] *M.-H. Li, B.-Y. Chen, and C.-T. Li, “A Hybird Method with Gravity Model and Nearest-Neighbor Search for Trip Destination Prediction in New Metropolitan Areas,” in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 6553–6560. doi:* [*10.1109/BigData55660.2022.10020439*](https://doi.org/10.1109/BigData55660.2022.10020439)*.*