Presented at 21st ARRB conference in Cairns, Austraria, 18-23 May, 2003

CLEANSING OF PROBE CAR DATA TO DETERMINE TRIP OD

Edward Chung, University of Tokyo, Japan

Majid Sarvi, University of Tokyo, Japan

Yasunori Murakami, University of Tokyo, Japan

Ryota Horiguchi, University of Tokyo, Japan

Masao Kuwahara, University of Tokyo, Japan

ABSTRACT

GPS is increasingly being used to collect travel data as the cost of the equipment is relatively low and it is capable of providing continuous and accurate spatial information and speed in real time. One such example is the Internet Protocol probe car (IPCar) project in Japan which equipped probe cars (consisting of taxis and buses) with GPS. The aim of this project is to explore feasible real time applications of IPCar data either as a stand alone data source or with other data sources such as detector counts and automatic vehicle identification (AVI) travel time. The initial focus of this project is to provide travel time information.

This study focuses on a data cleansing procedure consisting of 6 steps to determine the OD pattern of the probes. The cleansing process addresses data errors and searches for trip ends and the results show that the cleansed data was accurate in terms of trip length distribution, in particular for trips of more than 1km. The probe data also matched 76%-83% of the trips from an independent data from a taxi management system. The OD pattern gives a macroscopic view of the taxi movement and shows that the main generator and attractor of trips are areas around Sakuragi-cho, Yokohama station, Honmoku, Negishi station and Totsuka. The OD pattern and desired line of the probe car also demonstrate that using taxis as probe vehicles only generate intense data for the small areas heavily serviced by taxis and therefore travel time at a higher level of confidence can only be predicted for these small areas. Future full scale implementation of the IPCar project may need to consider private cars and other sources of information such as detectors and AVI data, to supplement area with sparse taxi coverage.

INTRODUCTION

GPS is increasingly being used to collect travel data as the cost of the equipment is relatively low and it is capable of providing continuous accurate spatial information and speed in real time. One such example is the Internet Protocol probe car (IPCar) project in Japan which equipped probe cars (consisting of taxis and buses) with GPS. The aim of this project is to explore feasible real time applications of IPCar data either as a stand alone data source or with other data sources such as detector counts and automatic vehicle identification (AVI) travel time. The initial focus of this project is to provide travel time information and would later extend it to provide other information such as congestion level and emission level.

The quality of travel time information from probe vehicles depends on the frequency of probe vehicles traversing a road link. A large sample of probe vehicles per link per unit time would provide travel time with a higher level of confidence. However, the frequency of probe vehicle is a function of the number of probe vehicles and distribution of probe vehicle trips over the network. In order to address the questions about distribution and frequency of probe vehicles, a detailed understanding of the origin destination (OD) pattern of IPCar is essential for the development of a travel time prediction model so as to meet the specified level of accuracy and

confidence (**see Figure 1**). Only OD pattern of taxis are analysed as buses runs on a fixed route and schedule.

From the OD patterns, the coverage of the probe cars at different times of the day and areas where the probe cars coverage is sparse can be determined. With the travel pattern, it is then possible to see areas where the probe cars coverage is sparse. Even in areas that are well covered, certain routes may be more popular at certain time of day. The next stage is to analyse the link characteristics to determine the frequency of probe car traversing each link per unit time interval. Once accurate OD information of taxis can be determined, it can be decided whether taxis are suitable probe cars to determine accurate travel time.

This study focuses on the data cleansing of probe data to determine the OD pattern of the probes, shown as shaded boxes in Figure 1. The OD pattern gives a macroscopic view of the taxi movement. Detail study of link characteristics such as travel time variance and the development of a travel time prediction model using probe data will be published in the near future. In the following section, an overview of the IPCar system is provided. The steps involved in the cleansing of the probe car data and the trip distribution (OD) of the probe car are presented in detail in the subsequent sections.

IPCAR SYSTEM

The Yokohama IPCar project uses 179 vehicles consisting of 140 taxis and 39 buses. The experiment ran for 11 days in December 2001. The IPCar system is equipped with a GPS and a data logger. GPS collects position data at regular intervals. However, the IPCar system does not store the vehicles position at regular intervals. Instead it logs the state of events as either short stop (SS) or short trip (ST) (**see Figure 2**). The definition of a short stop is when the vehicle speed drops below 3 km/h. When the vehicle speed increases above 3 km/h, the event is considered as a short trip. In other words, instead of a time based data logging, the equipment is an event based data logging. So every time the event changes from SS to ST and vice versa, the GPS position and time stamp are recorded plus the event flag (eg. SS or ST).

This approach reduces the amount of data stored and therefore less data transmission, without sacrificing the quality of the data (Horiguchi, 2002). Note that there is a maximum time limit of 30 seconds for a ST event. If a probe vehicle is moving for 2 minutes, it will be recorded as 4 consecutive ST events. However, there is no time limit for SS event. These event records are temporary buffered in the on-board equipment and transmitted by the polling request from the IPCar data centre. The interval of transmission is normally 1-5 minutes. In addition, the IPCar system records other parallel events information such as the status of the left and right blinkers, the hazard light and the parking brake (**see Figure 2**).

There are instances where data are not recorded ie. contains gap. This could be due to communication or GPS errors. GPS errors might occur when a probe vehicle passes under an infrastructure such as tunnel, or when in the vicinity of elevated structures, the so called urban canyon. Gap could also occur when the engine is switched off because no data will be recorded.

Taxi management system

Three taxi companies took part in the experiment and one of the taxi companies also have their own Taxi Management System installed on the taxi. There are 16 probe vehicles with taxi management system. The taxi management system collects the time, date and position when the fare meter is on or off. In other words, when the taxi goes from in-service to not in-service and vice versa. This data source is used to verify the results from the data cleansing discussed in later section.

DATA CLEANSING

Before the probe data can be used to determine trip OD, the data needs to be cleansed because probe data is a continuous trajectory (**see Figure 3**) and also there are gaps in the data. Therefore, the data cleansing process for the OD analysis is to cut the "continuous" trajectories into trip ends by detecting the following events (**see Figure 4**).

- Gap with parking brake event,
- Long gap,
- Gap with unrealistic speed,
- Long stop,
- Short stop with hazard light, and
- U-turn.

The data cleansing process starts by considering gaps in the data in step 1 to step 3. It then searches for stops which are trip ends in steps 4 to 6. Details of each step are explained below.

Step 1: Gap with parking brake event

Gap in the data could be due to communication error or engine being switched off. However, when there are simultaneous events of a long gap and parking brake event during the gap, it is highly likely that the engine is being switched off. In other words, this occurrence can be considered as a trip end and the trajectory can be cut at this point. Note that parking brake event is checked before and after the gap as no information is obtained during the gap. (**see Figure 5**)shows that most of the gaps with parking brake event occur when the gap is more than 10 minutes, therefore supports the above reasoning. In the data cleansing process, all gaps with parking brake event are considered as trip end.

Step 2: Long gap

There are also instances where a gap occurs without parking brake (see Figure 5).. When a gap is small say 2 minutes and a vehicle is moving, it is fairly safe to bridge the gap by connecting the points before and after the gap with the same travel speed. However, when the gap is large say 15 minutes, numerous combinations of possibilities could occur during this time, such as:

- the vehicle dropping of and picking up passenger,
- the driver waiting at a taxi rank,
- the engine being switched off,
- the driver taking a meal break, and
- the vehicle is on a job.

In this step, 15 minutes is the threshold for gap duration when the gap is considered as trip end. In reality this may not be a true trip end but the lack of further information makes this the best alternative.

Step 3: Gap with unrealistic speed

After removing the long gaps, the remaining gaps are checked for their speed. Since the location and time of the events before and after the gap are known, the speed taken to traverse the gap distance can be computed. From all the data in this experiment, there was no speed greater than 60 km/h. This speed value is used as the upper bound for the speed check and data points above the upper bound are eliminated. For the remaining gaps, if the computed gap speed is greater than 75% of the short travel (ST) speed before the gap, the trajectories before and after the gap will be connected. Otherwise, the gap is considered as a trip end.

Step 4: Long stop

The first 3 steps consider the gaps in the data and steps 4 and 5 search for stops that are trip ends. Stops could happen when a taxi is dropping off or picking up passenger, stopping at an intersection or taxi rank. Obviously picking up and dropping off passenger are considered as

the beginning and end of a trip, respectively. To differentiate between stopping at an intersection and a true trip end can be difficult. Firstly, it takes more than 20 seconds to drop off a passenger ie. the time for a taxi to stop and for the driver to collect the taxi fare. A taxi waiting at a signalised intersection could range from a few seconds to over 100 seconds. It is therefore difficult to distinguish between a genuine trip end and just stopping at intersection. However, from the time distribution of stops with and without parking brake event, 95% or more of the stops are less than 150 seconds. This indicates that it is unlikely for a vehicle to stop at an intersection for more than 150 seconds. In this step, short stop of 180 seconds with parking brake is adopted as the threshold for cutting the trajectories (ie. accepting the long stop as a trip end). From the calibration of maximising the number of correct trip end and minimising the number of false trip end, it was found that cutting a trajectory at short stop greater than 30 seconds without parking brake event gives the best results.

Step 5: Stop with hazard light

The previous step does not recognise stops for dropping of or picking up passengers. In Japan, taxi drivers turn on the hazard light when picking up and dropping off passengers. However, hazard light is also used to acknowledge other drivers for allowing a vehicle to merge or pass, commonly referred as "thank you hazard". Analysis of stops with hazard light when picking up and dropping off shows that the minimum stop time is 20 seconds. In this step, short stop greater than 20 seconds with hazard light more than 10 seconds is used as a cutting point for trip end.

Step 6: U-turn

The last cleansing step looks at the shape of the trajectory that resembles a loop or u-turn. U-turn is often a point close to a trip end for example after dropping off a passenger, the taxi may make a u-turn to go back where it came. Some u-turns are sharp turns (eg. 3 point turn) and others are more gradually. It is also important to note that the geometric configuration of some road network is shape like a loop such as clover interchange, and on and off ramps. Firstly, an exception list of all loops in the road network is created. The list is used to ignore loops detected in the excluded area. Secondly, loops are ignored in the CBD area because there are one-way streets. Excluding the exception list and CBD area, the u-turn algorithm checks the turning angle of all trajectories. If the turning angle of its current position with respect to the last 10 ST trajectories of length more than 20 metres exceeds 170 degree, it is considered as a u-turn (**see Figure 6**).

After the data are cleansed, all the cut points become trip ends. The cleansing process also generates some very short trip ends due to gaps in the data and also due to imprecision in the search for trip ends. It is decided that trip ends less than 500 metres are eliminated as almost all trips are longer than that.

COMPARISON WITH TAXI MANAGEMENT DATA

A secondary data source from the taxi management system is used to verify the results from the data cleansing. The verification uses two measures namely trip length distribution and coverage ratio (ie. the proportion of the probe cars trip ends matching the taxi management trip ends) and false ratio.

Trip length distribution

There are a few fundamental differences between the probe data and taxi management data. Firstly, only straight line distance (as the crow flies) can be calculated from the taxi management data (ie. between the beginning and end of a trip). Whilst it is possible to compute the actual travel distance from the probe data, only straight line distance is used for comparison.

Secondly, the definition of trip needs to be clarified. A trip is define as a travel between origin and destination for a single trip purpose. Whilst the taxi management data may depicts a trip

when the taxi is in-service accurately, it may not do so when the taxi goes from not in- service to in-service. For example a taxi which is not in-service which goes to a petrol station for fuel and then pick up a passenger is considered as 2 trips according to the above definition of trips. However, the taxi management data would consider this as 1 trip, ie. from not in-service to in-service. As the example shows, there would be more trips and also shorter trips from the probe data than the taxi management data. The average taxi management trip length is approximately 3 km.

On the other hand, there would be less discrepancy in the number of trips between the two data sources when the taxi is in-service. For example, it is not frequent for a taxi to pick up 2 passengers wanting to be dropped off at different points. Also, the data cleansing process would only pick this type of trip as 2 separate trips when the taxi has to make a u-turn or when the taxi stops for more than 20 seconds with hazard light switched on for more than 10 seconds. Note that from the probe data it is not possible to tell whether the taxi is in-service or not. Hence, when comparing the trip length distribution between the two data sources, all trips have to be used (**see Figure 7**). Considering the difference in trip definitions discussed above, the result shows that the data cleansing process is proving to be accurate for the trip length distribution. Almost all the discrepancies are of trip length between 500 to 1000 metres.

Coverage and false ratios

The comparison of trip length indicates how well the trip length distributions fit. However, it does not show whether the locations where the trips took place matched. To test what proportion of the taxi management trip ends match the probe trip ends, the coverage ratio and false ratio were used. The coverage ratio measures the proportion of the taxi management trip ends which are matched by the probe trip ends. And the false ratio measures the proportion of probe trip ends which are not matched by the taxi management trip ends.

Under optimum condition, GPS has a horizontal accuracy of 5 metres but could increase to 30-50 metres when conditions are unfavourable. Furthermore, the time when the vehicle positions are taken by each system differs. Hence, when matching trip ends, a radial search of 500 m is used. A temporal constraint is also introduced to ensure that the right trip ends are matched. For example, a taxi may make multiple trips to major train stations such as Yokohama station in one day. Therefore the matching of trip end location alone is not stringent enough to ensure a proper match. A time limit of ± 15 minutes is also used.

The taxi management data can be divided into 2 categories, in-service and not in-service. The probe trip ends are matched against the in-service and not in-service trip ends and the results are shown in Table 1. As expected, the in-service trip ends have a higher coverage ratio (83%) than the not in-service trip ends (76%). This is due to the difference in how trips are measured by the probe and taxi management system which was discussed earlier. The false ratios are 33.7% and 28.8% for not in-service and in-service trips respectively. This is mainly due to gaps in the data and imprecise determination of trip ends. Based on the trip length distribution, and coverage and false ratios, the data cleansing process seems to be performing satisfactorily, in particular if the shorter trips could be reduced.

Table 1 Comparison of trip ends between taxi management and probe data

	Not in-service trip	In-service trip
Coverage ratio (%)	76.3	83.0
False ratio (%)	33.7	28.8

ORIGIN DESTINATION PATTERN

This study divides Yokohama, the study area, into grids of 500m to define the OD zones. From these zones, OD matrices for different times of the day are compiled. A list of the top 20 attractors and generators in terms of trip ends are compiled. The results show that whilst the ranking of areas generating and attracting trips at different times of the day may be different, the main areas however are similar. For example, the Sakuragi-cho area (including Kannai and China-town) and Yokohama station area are consistently amongst the major generators and attractors. Other activity centres include the Honmoku, Negishi station and Totsuka areas. Since taxis are used as probe cars and train stations are often bases for the taxis, the results are not unexpected. In other words, the probe (taxi) OD does not necessary represent the true vehicle OD of Yokohama. This is an important factor to consider when designing a full-scale implementation. Furthermore, the 3 taxi companies that took part in this study used the Yokohama station, Sakuragi-cho and Totsuka areas as their base.

The desired line of trips originating from activity centres such as the Sakuragi-cho area shows that these trips cover approximately a 3 km radius (**see Figure 8**), which is also the average length of a taxi trip. Although the trip rate at midnight is much lower, the trip length is longer and the destinations more scattered (**see Figure 9**). This is partly a characteristic of the travel pattern in major city centres in Japan, where most people commute to the city and take a taxi from the station if their destination is not within walking distance. However, after midnight, train stops running and therefore taxi trips are longer and the destinations more scattered.

CONCLUSION

A data cleansing process consisting of 6 steps was developed to search for trip ends while addressing data errors. The cleansed probe data was compared against the taxi management data which was an independent data source. The results show that the cleansed probe data produced accurate results in terms of trip length distribution, in particular for trips more than 1km. However, for short trips (ie. less than 1 km), there were some discrepancies mainly resulting from gaps in the data and imprecise determination of trip ends. Coverage and false ratios were also used to measure the performance of the probe data, against the taxi management data. The probe data matched between 76% and 83% of the taxi management data. Between 29% and 34% of probe trips were not found in the taxi management data.

OD patterns from the trip ends generated through the data cleansing process were compiled. The results show that the main generator and attractor of trips are areas around Sakuragi-cho, Yokohama station, Honmoku, Negishi station and Totsuka. It is noteworthy that the probe (taxi) OD does not necessarily represent the true vehicle OD of Yokohama. This is an important factor to consider when designing a full-scale implementation.

The desired line of trips originating from activity centres such as the Sakuragi-cho area shows that these trips cover approximately a 3 km radius. This is a travel pattern of people commuting to the city and taking a taxi from the station if their destination is not within walking distance. However trip lengths are longer and destinations more scattered from midnight when train stops running.

The OD pattern and desired line of the IPCar demonstrates that using taxis as probe vehicles only generate intense data for small areas. This implies that travel time information at a higher level of confidence can only be predicted for the areas heavily serviced by taxis. For other areas where taxi coverage is sparse, private cars may be needed along with other sources of information such as detectors and AVI data. It is also important that many different taxi companies should be signed up for the project to give greater coverage as taxi companies in Japan have an unspoken rules about the areas which they are limited to service.

Future research in the data cleansing process would explore the possibility of utilising geographic information and sequence of events to improve the accuracy of determining trip ends. For example, short stop at the vicinity of an intersection and gap in a tunnel section.

ACKNOWLEDGEMENT

We would like to express our sincere appreciation to Association of Electronic Technology for Automobile Traffic & Driving (JSK), who kindly provided the IPCar data. We also thank NEC and NEC Soft, Ltd. for their kindness to give us the details of IPCar data.

REFERENCES

Horiguchi, R. (2002). The Advantage of Event-Periodic Data Recording for Probe Vehicle System, *Proceedings of Infrastructure Planning,* Vol. 26, CD-ROM, November.. 2002 (in Japanese).

AUTHOR BIOGRAPHIES

EDWARD CHUNG is a visiting Professor at the Centre for Collaborative Research, University of Tokyo. He received his Bachelor degree in Civil Engineering and PhD in Traffic Engineering from Monash University. Edward worked as a senior scientist with ARRB Transport Research and as a manager of Infrastructure Analysis and Modelling with the Department of Infrastructure for the state of Victoria. He has also worked in the private sector with R.J. Nairn & Partners. Edward's research interests include transport management, transport emission, transport planning and traffic operations.

MAJID SARVI is a research fellow at the Centre for Collaborative Research, University of Tokyo. He received his Ph.D. in Civil Engineering from University of Tokyo. Majid worked as chief researcher of ITS research group of Social System Research Institute and as a transport analyst with the Hong Kong Transport Department. Majid's research interests include traffic operations, traffic flow theory, transport modelling and highway operations.

YASUNORI MURAKAMI is a researcher at the Centre for Collaborative Research, University of Tokyo. He received his Bachelor's degree in Civil Engineering from Tokyo Institute of Technology and Master's degree in Civil Engineering from University of Tokyo. Murakami's research interests include transport planning and traffic operations.

RYOTA HORIGUCHI is the president of i-Transport Lab. Co. Ltd. He received his MS degree in Computer Science from Hokkaido University and PhD in Civil Engineering from University of Tokyo. Horiguchi's major interests include traffic simulation, traffic operation, and traffic information processing. Horiguchi also has about 10 years experience in consultation as a practitioner.

MASAO KUWAHARA is a Professor at the Centre for Collaborative Research as well as at Institute of Industrial Science, University of Tokyo. He received his Ph.D in Civil Engineering from University of California, Berkeley. Kuwahara has about 20 years experience in teaching and researching on traffic engineering in University of Tokyo. Kuwahara's research interests include dynamic network analysis, traffic simulation, highway capacity, traffic signal control, transport emission, and transport planning.

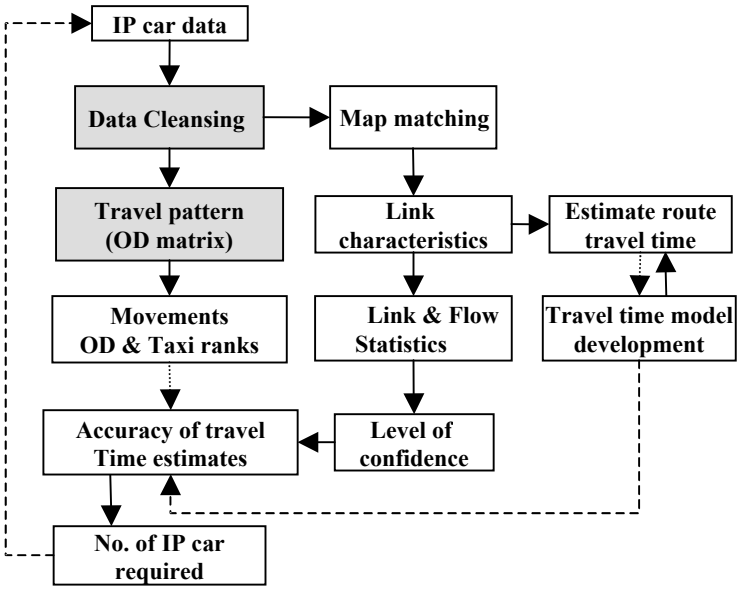

Figure 1 Travel time prediction framework of the IP Car Project

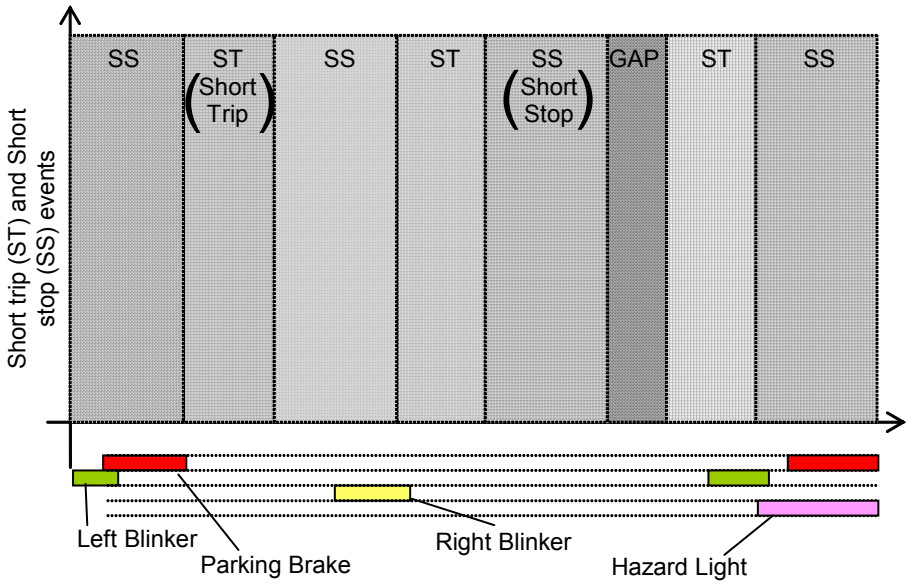

Figure 2 Data collected by the IP Car system

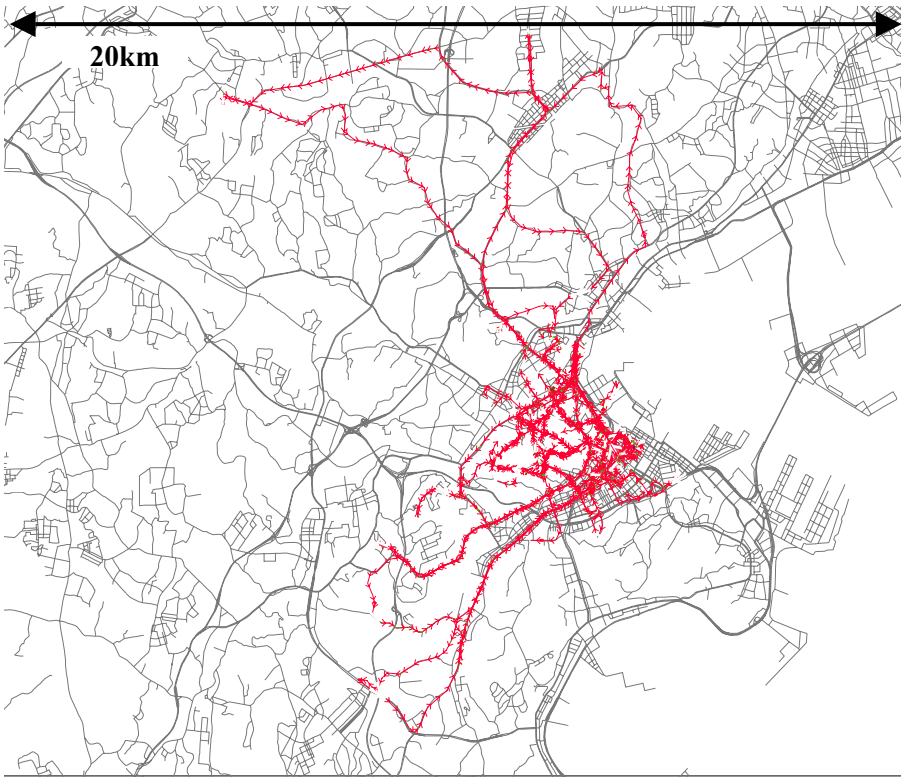

Figure 3 Continuous trajectory of one probe vehicle including gaps

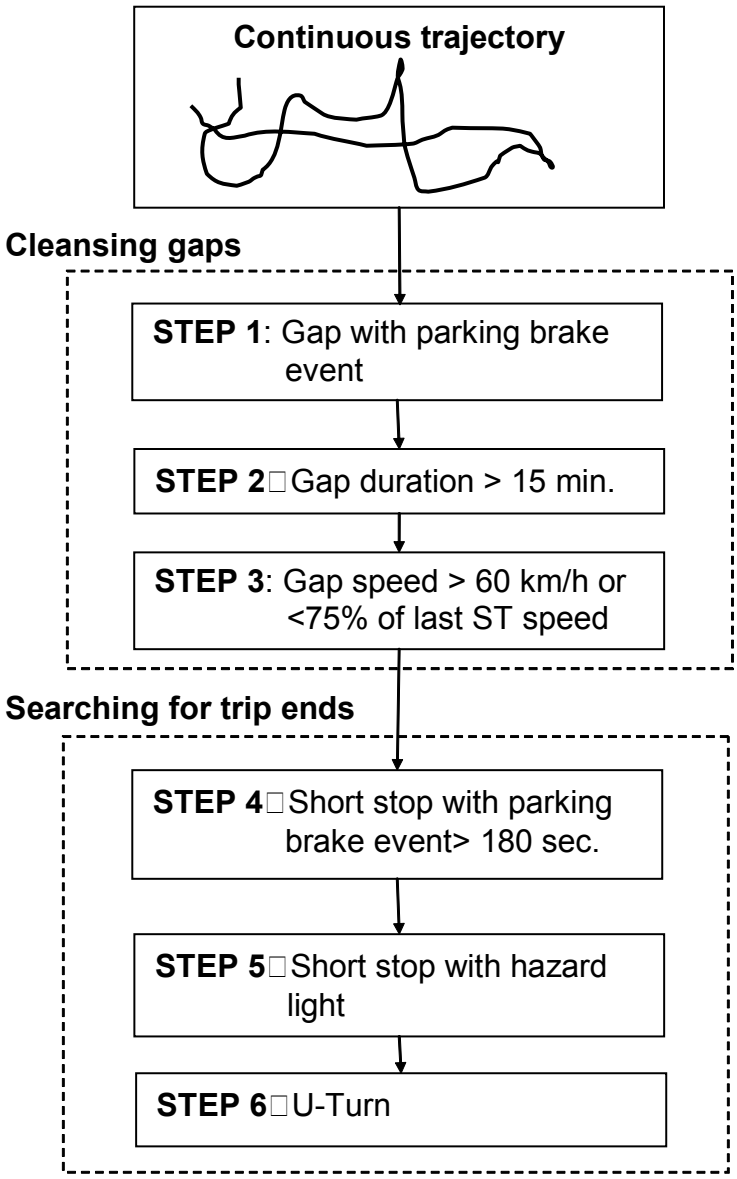

Figure 4 Data cleansing process

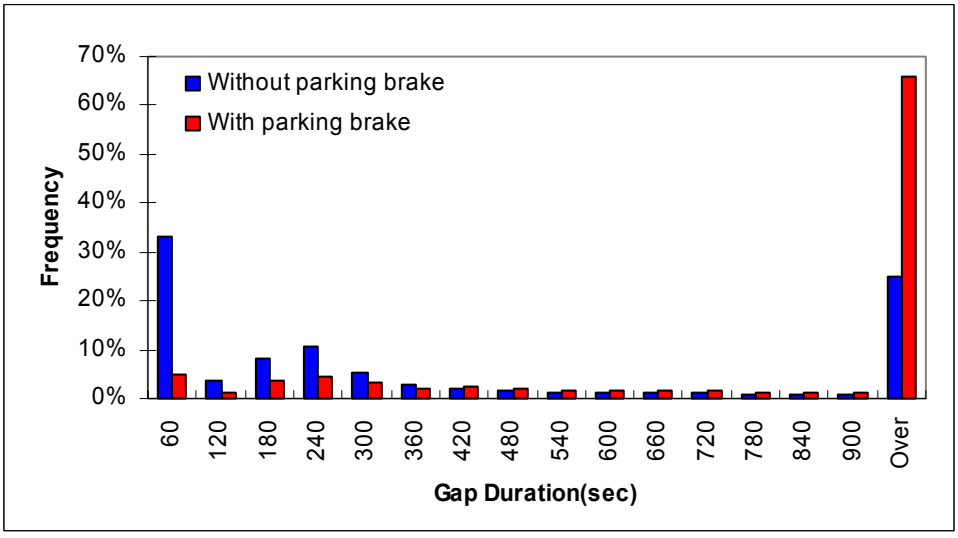

Figure 5 Frequency of gap with parking brakes

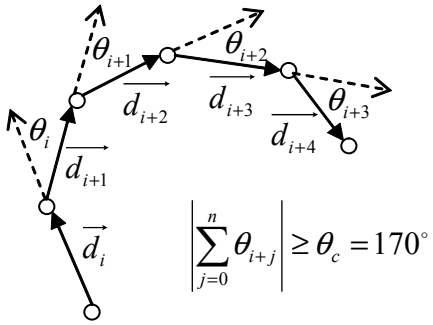

Figure 6 U-turn algorithm

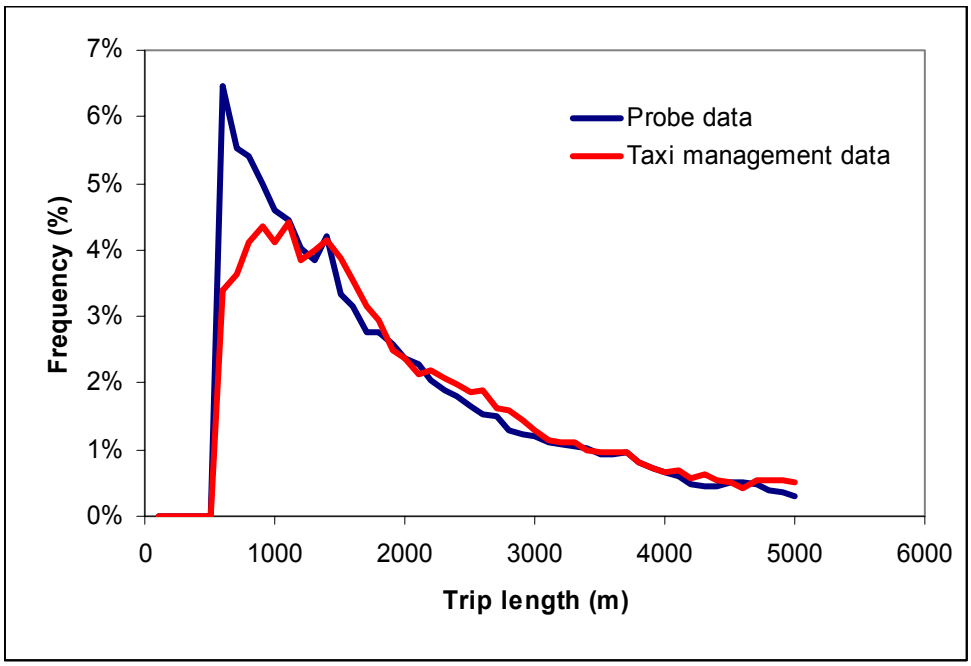

Figure 7 Comparison of trip length distribution

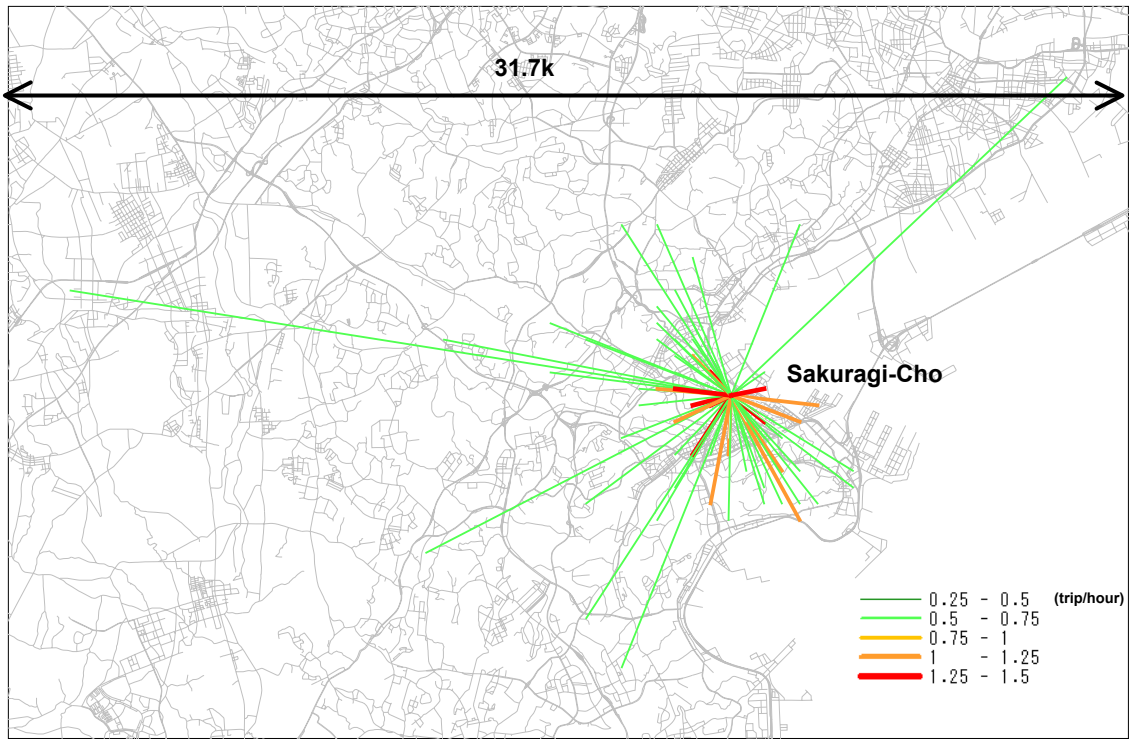

Figure 8 Desired line of trips originated from Sakuragi-cho between 7-9 am.

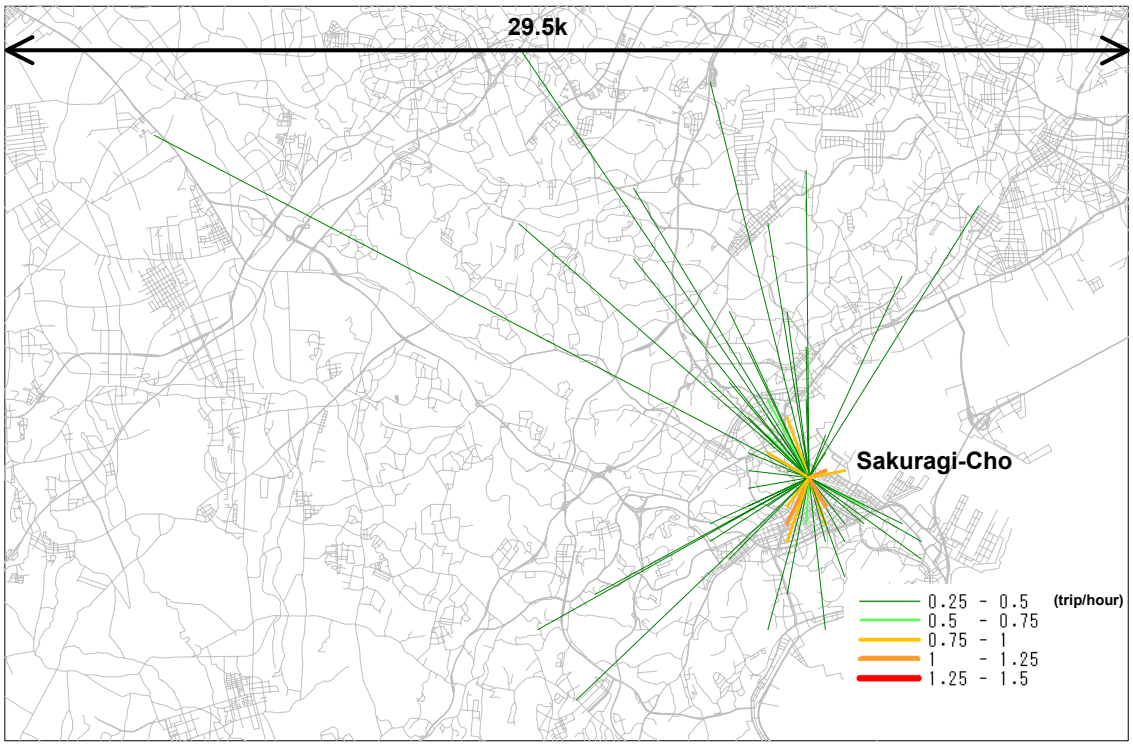

Figure 9 Desired line of trips originated from Sakuragi-cho between 23:00-03:00 am.