# A METHODOLOGY FOR DATA CLEANSING AND TRIP END IDENTIFICATION OF PROBE VEHICLES

**Majid Sarvi [1), Edward Chung [1), Yasunori Murakami [1), Ryota Horiguchi[2),**
**Masao Kuwahara [1)**

## 1. INTRODUCTION

This study firstly proposes a methodology, which estimates link travel time. In addition, potential approaches to obtain probe vehicles traffic demands from intelligent probe car (IP car) is introduced. Then a general methodology to cleanse the GPS data will be discussed in detail. Several intelligent vehicle-highway system demonstration projects are currently assessing the feasibility of using probe vehicles to collect real-time traffic data for advanced traffic management and information systems. However, only a few generalized algorithms have been developed and tested. Travel time, or the time required to traverse a route between any two points of interest, is a fundamental measure in transportation. Travel time is a simple concept understood and communicated to a wide variety of audiences. Various kinds of travel time estimation and OD prediction are being implemented. Several data collection techniques can be used to measure or collect travel times. These techniques are designed to collect travel times and average speeds on designated roadway segments or links. Because these techniques differ from point-based speed measurement, the resulting travel time and speed data are much different than spot speeds. An overview of the techniques relevant to this study is provided in the following paragraphs.

**The test vehicle technique** (often referred to as "floating car") are the most common travel time collection methods and consist of a vehicle(s) that is specifically dispatched to drive with the traffic stream for the data collection purpose.

Key words: Probe vehicle, travel time/ OD estimation, data cleansing

1) University of Tokyo, IIS, Kuwahara Lab., 4-6-1 Komaba, Meguro Ku, Tokyo 153-8505, Japan, Email: msarvi@nishi.iis.u-tokyo.ac.jp Tel: 03-5452-6098, Fax: 03-5452-6420

2) i-Transport Lab. Co., Ltd., 2-12-404, Ageba-cho, Shinjuku-ku, Tokyo, 162-0824, Japan, Email: horiguchi@i-transportlab.jp, Tel/Fax: 03-5261-3077

**The emerging and non-traditional technique** is currently being researched or developed, or may be considered non-traditional when compared to existing methods. These techniques use a variety of methods, such as inductance loops, weigh-in-motion stations, or aerial video to estimate or calculate travel times.

**The ITS probe vehicle technique** utilizes passive instrumented vehicles in the traffic stream and remote sensing devices to collect travel times. The ITS probe vehicles can be personal, public transit, or commercial vehicles. ITS probe vehicles also typically report travel time data to a transportation management center (TMC) in real-time. Probe vehicles may be equipped with several different types of electronic transponders or receivers. The monitoring of cellular telephone activity is also being tested for potential travel time collection applications. GPS receivers use a network of 24 satellites to determine vehicle position and are becoming common for route guidance and "mayday" security applications.

This research firstly addresses the outline of travel time and OD estimation using IP car equipped with GPS. Then the general methodology used to cleanse the GPS data is discussed in detail.

## 2. OUTLINE OF TRAVEL TIME AND OD ESITIMATION

Figure 1 shows the general layout of travel time and OD estimation as well as evaluation of required number of probe cars. First, the data obtained from 179 vehicles including 140 taxies and 39 buses running in Yokohama area are used for the study. This data is collected and accumulated over 15 days.

Due to communication and GPS errors there are instances where data are not recording i.e. contains gap. GPS error might happens where an IP vehicle passes an underground street or in vicinity of elevated structures or due to engine being turned off. During data cleansing process, an attempt has been undertaken to discard those inaccurate data.

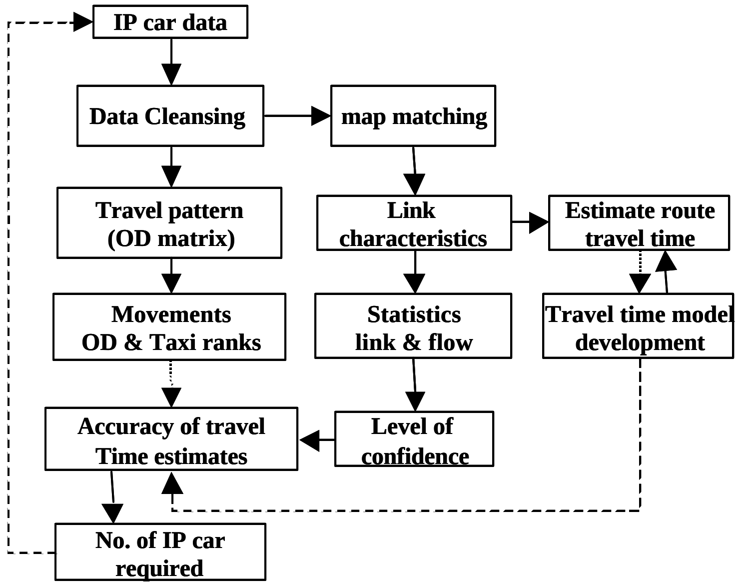



**Figure 1. General layout of travel time and IP vehicles OD estimation.**

Due to the accuracy limit of GPS (about 30 meters), a map matching process is needed to assign a specific route associated with the collected data to the trajectory of each IP vehicle. However, for the OD estimation, map matching is not necessary because OD is area based. In addition to data cleansing the trajectory of each IP vehicle during one-day trip is a continuous line (path), which should be broken up into origin and destination. This process will be discussed later in detail. The obtained OD matrix will demonstrate the movement patterns of IP vehicles within the network. However, in future by equipping more vehicles with GPS such as private cars the network overall OD might be obtained. As well the taxi ranks movement during morning or evening peaks can also be gained which is valuable to the taxi management center. The study continues with estimation of travel time and frequency of IP car over each link. In this step, based on travel time analysis, the characteristics of traffic pattern, congestion and free flow, could additionally be achieved. One of the most important questions has to be addressed yet, is the number of IP vehicles required in the network. Since there is a relationship between the number and movement pattern of IP vehicles, and application of the data, the appropriate number of IP vehicles need to be estimated. This approach is designed to ensure that a specific level of certainty for the entire link is achieved i.e. a predetermined frequency of IP vehicles will be observed.

## 3. DATA CLEANSING PROCESS

The outline of data cleansing process is illustrated in Figure 2-a. As mentioned earlier the trajectory of each IP vehicle during one-day trip is a continuous path, which should be broken up into origin and destination. In addition, there are communications and GPS errors, which required be recognized and cleansed for each IP car data (see Figure 2-b).

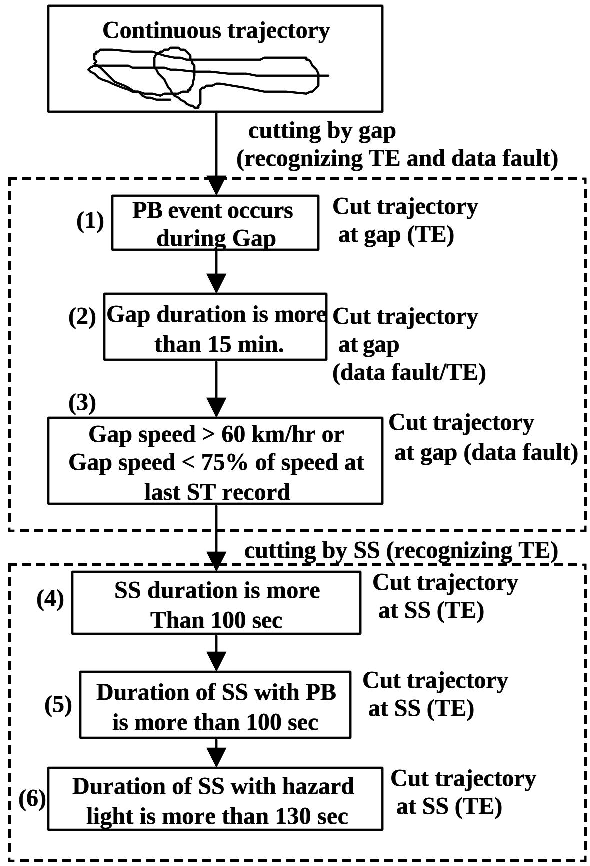
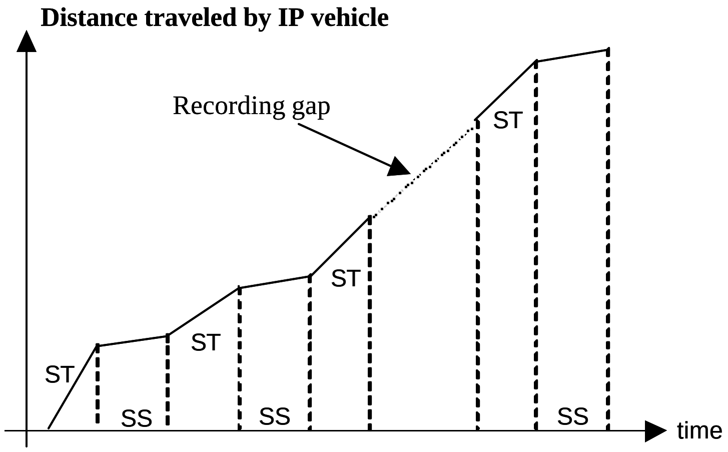



**Figure 2-a. Outline of data cleansing process.**

In Figure 2-a, PB is parking brake, SS is shortstop (when the vehicle speed < 3 km/hr), TE is the trip end, and ST is short trip. Gap speed means D/T where D (gap distance) is straight distance between the two points containing gap (communication error) and T is gap duration.

Figure 2-b demonstrates the trajectory (pattern) of an IP vehicle. It consists of sequential ST and SS events and likely a recording gap due to the engine being turned off or any other previously mentioned causes.

The procedure in Figure 2-a is as follows: (1) Figure 3 demonstrates the relationship between gap duration with parking brake and its frequency (relative ratio means number of observed value/total number of observed values). Based on this figure most of long gaps happen with parking brake. In other words, a long communication gap simultaneous with a parking brake is most likely that an engine has being turned off. Therefore, we can consider this event as a trip end and terminate a trip of the IP car at this point.




**ST: short trip**
**SS: short stop**

**Figure 2-b. ST, SS and recording gap definition**

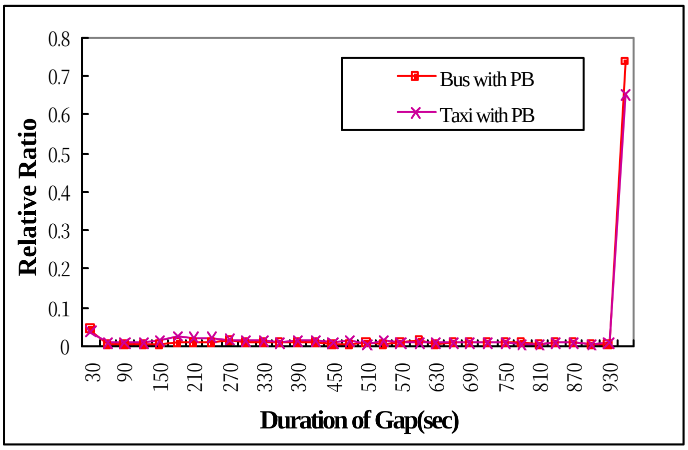



**Figure 3. Duration of Gap with PB and its frequency**

(2) The relationship between gap duration and gap distance is shown in Figure 4-a. Based on our data cleansing methodology all points with longer than 15 minutes gap duration are considered as trip end events (even if they might satisfy the step 3 of Figure 2-a). However, these points may or may not be considered as destination of trips. In other words, when a gap or communication error longer than 15 minutes happens, judging about what happened during that period of time is tentative. Therefore, is prudent to consider this point as a trip end event. On the other hand for the gap shorter than 15 minutes our methodology requires more investigations in order to distinguish the destination of each trip as it is explained in step 3 of Figure 2-a.

(3) Those points without parking brake, which are located above the 60 km/hr line of Figure 4-b (average maximum speed of car in surface street) practically are unfeasible and would be considered as an errors or data faults.

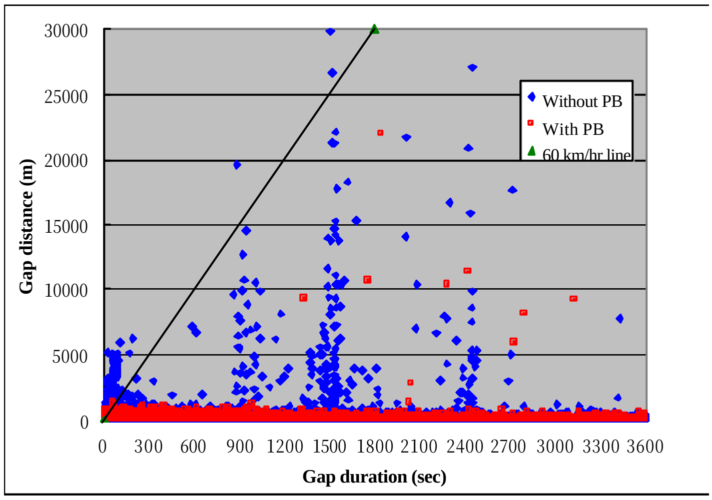



**Figure 4-a. Gap duration and distance relationship**

Consequently, those points would be eliminated from the database. The points below the line however, are feasible and could happened due to passage of an IP car through underground streets or in vicinity of elevated structures. For those points if the gap speed is less than 75% of the speed at the last recorded short trip (ST), it is considered as trip end. Otherwise, the two trajectories before and after gap will be connected with a straight line to produce a complete trajectory (see Figure 2-b).

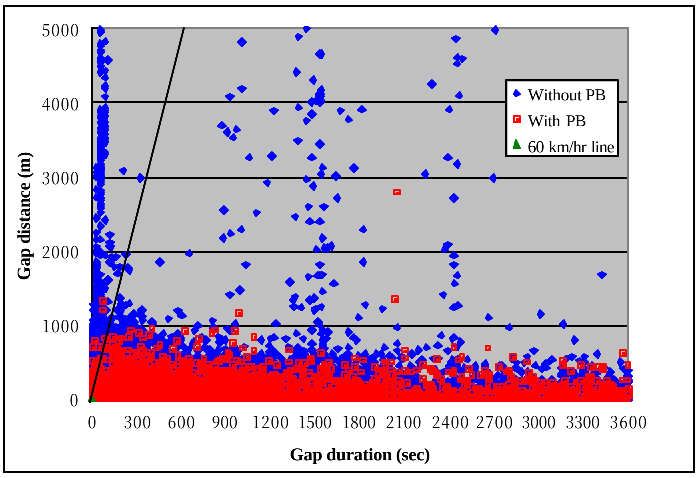



**Figure 4-b. Illustration of data faults**

(4) Having the communication and GPS error cleansed, the next step would be separating the continuous trajectory based on shortstop (SS) events as illustrated in the second portion of Figure 2-a. The relation between SS duration and SS frequency is demonstrated in Figure 5.

The cumulative percentage of SS without parking brake longer than 100 sec is less than 5 percent. Additionally, the average cycle time of signals are about 100 sec. Therefore, it is quite rational to consider a SS event longer than 100 sec. as the destination of trip.

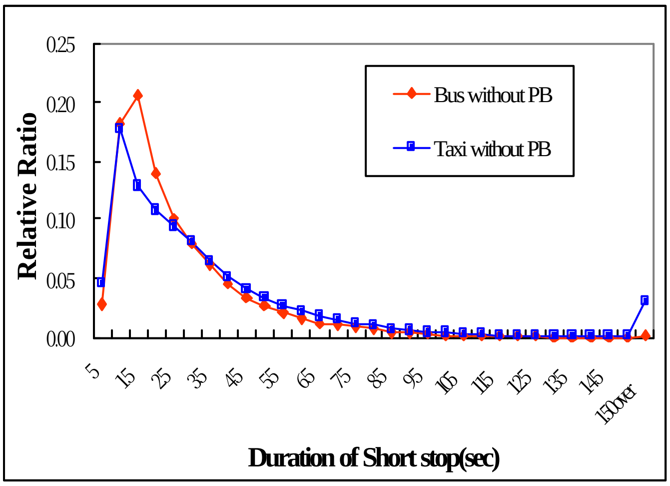



**Figure 5. SS duration without PB and its frequency**

(5) Figure 6 shows the relationship between SS duration with parking brake and its frequency. SS with parking brake could happen due to embarking or disembarking of passengers. According to this figure all SS event with parking brake longer than 100 sec could be considered as TE event.

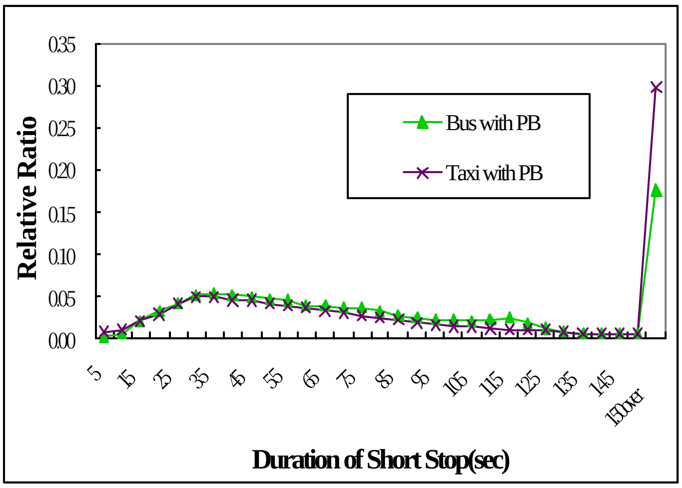



**Figure 6. SS duration with PB and its frequency**

(6) The last step used to decipher the TE event in the methodology could be those SS with hazard lights. It is assumed that some drivers might use hazard light when they are dropping off or picking up passengers. The relation between SS duration with hazard light and SS frequency is demonstrated in Figure 7. According to this Figure 7 all SS events with hazard light longer than 130 sec could be considered as TE event.

However, It is worth mentioning that utilizing the proposed data cleansing methodology might not produce the desired results. In other words, there could be TE events that need to be identified (e.g. those SS with/without parking brake shorter than 100 sec). If that is the case, more cleansing criterions should be taken into consideration. Some potential criterions could be cutting a continuous trajectory at loops, u-turn or self-crosses.

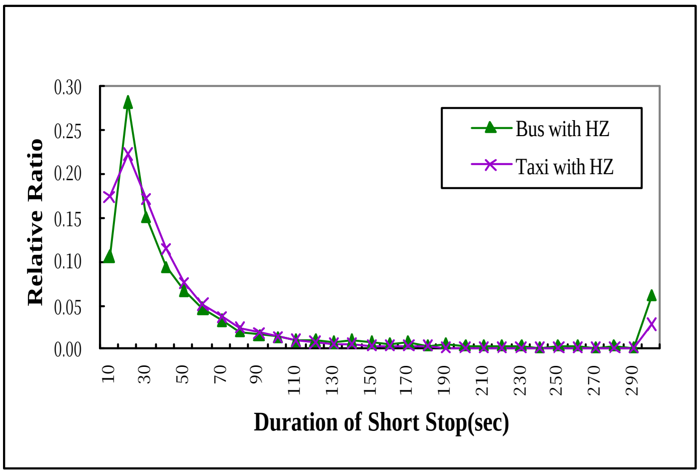



**Figure 7. SS duration with HZ light and its frequency**

## 4. CONCLUSIONS

This study proposed a general methodology, to cleans IP vehicle data and to determine trip ends from a continuous trajectory of each IP vehicle. In addition a framework on an application of the cleansed data to determine travel pattern (OD) of IP vehicles, travel time statistics and ultimately travel time prediction is presented. The application of IP vehicle data is not limited to travel time prediction only. Theses data can also be used in other applications such as travel behavior and vehicle emission.


## ACKNOWLEDGMENT

The authors would like to thank NEC, NEC soft and Japan association of electronic technology for automobile traffic and driving for providing us with the IP vehicles data.


## REFERENCES


1) Chul, G., Jaihak, O., Sigon, K. (1998).”Determination of optimal number of probe vehicle for real-time traffic flow information.” 5th ITS world congress, Seoul, Korea, 1998.

2) Makimura, M., et al. (2002).”Performance indicator measurement using car navigation system.” 81st TRB annual meeting, WD, 2002.