Research Trends and Results

Development of a data cleansing method for probe travel speed data

Toshikazu Matsushima, Guest Research Engineer
Hiroyoshi Hashimoto (Ph.D. (Engineering)), Senior Researcher
Susumu Takamiya (Ph.D.), Head
Road Division, Road Traffic Department

(Key words) probe data, travel speed, data cleansing, ETC2.0 probe data

1. Introduction

A large amount of probe data obtained from automobile navigation systems and other sources are currently being collected and used to analyze traffic conditions. The probe data for travel speeds, which are processed to determine the travel speed of each link, contain peculiar high-speed and low-speed data as a result of individual driving preferences and stopping or parking a vehicle on a roadside. Therefore, to accurately analyze the data, it is necessary to remove such peculiar data.

2. Overview and characteristic of data cleansing method

One characteristic of the data cleansing method that we are developing is that it focuses on the differences in the amounts of time required by individual cars to pass the section. The data obtained stopped or parked vehicles, as shown in Figure 1, are unsuitable for a traffic condition analysis, and we need to remove such data. However, we cannot distinguish such vehicles from low-speed running vehicles caught in a traffic jam, if we just use the travel speeds of individual vehicles. Thus, we focused on the fact that stopped or parked vehicles required significantly larger amounts of time to pass the section than other vehicles, whereas vehicles moving slowly as the result of a traffic jam passed the section in smaller amounts of time that were not very different from those of other vehicles. We considered a vehicle to be peculiar and subject to removal if the difference between the actual amount of time required and the minimum required time during the same time period was larger than a certain threshold time.

3. Discussion of trial result of data cleansing

We used the data cleansing method for the travel speed data classified by DRM links and the upper or lower directions, from ETC2.0 probe data (from April to June in 2015, nationwide data), by setting the threshold time corresponding to the delay from the minimum required time to 600 s. The number of subject vehicles was approximately 274 million vehicle-links in total, and the number of removed vehicles was approximately 0.7 million vehicle-links in total. Thus, the percentage of removed vehicles was approximately 0.32%. We conducted a comparison that focused on individual vehicles in the same time period. As a result, we confirmed that the data that were supposed to be peculiar running data such as for stopping or parking on the roadside were removed.

4. Closing remarks

When using a large amount of data, it is important to perform data cleansing that agrees with the purpose of an analysis. In the future, we will further examine a data cleansing method and consider generalized measures such as creating a manual.

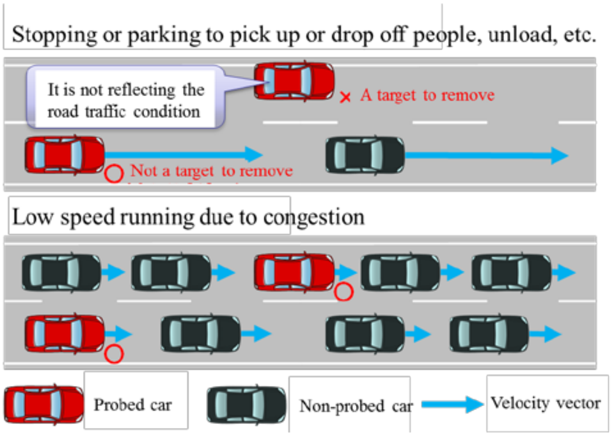
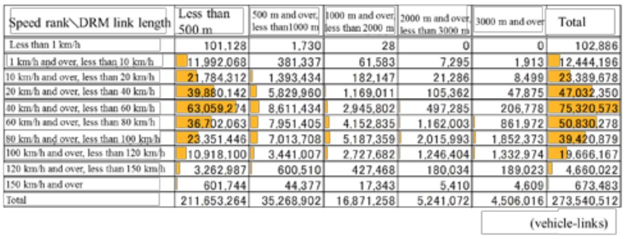

Stopping or parking to pick up or drop off people, unload, etc.
It is not reflecting the road traffic condition
× A target to remove
○ Not a target to remove
Low speed running due to congestion
Probed car
Non-probed car
Velocity vector

Figure 1 Example of running data subject to removal

Table 1 Overview of data cleansing

Item	Content
(1) Removing data based on threshold values	Data are removed when the speed is less than 1 km/h, or the speed is 150 km/h or more.
(2) Removing data based on a road traffic situation in the same time period	Samples are removed when they are 600 s or more later than the minimum required time in the same time period. Note that the lower limit of the minimum required time is determined by the required time for the speed of 80 km/h when a vehicle is running on a highway, and that for the speed of 30 km/h when running on a local road.

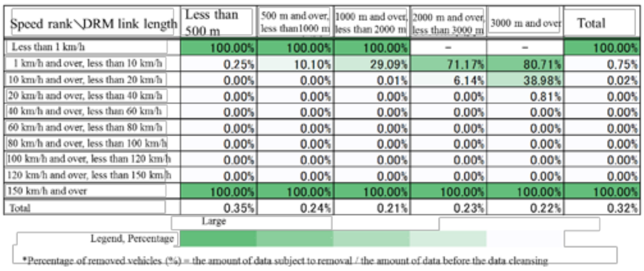
Table 2 Amount of data before data cleansing



Speed rank＼DRM link length	Less than 500 m	500 m and over, less than1000 m	1000 m and over, less than 2000 m	2000 m and over, less than 3000 m	3000 m and over	Total
Less than 1 km/h	101,128	1,730	28	0	0	102,886
1 km/h and over, less than 10 km/h	11,992,068	381,337	61,583	7,295	1,913	12,444,196
10 km/h and over, less than 20 km/h	21,784,312	1,393,434	182,147	21,286	8,499	23,389,678
20 km/h and over, less than 40 km/h	39,880,142	5,829,960	1,169,011	105,362	47,875	47,032,350
40 km/h and over, less than 60 km/h	63,059,274	8,611,434	2,945,802	497,285	206,778	75,320,573
60 km/h and over, less than 80 km/h	36,702,063	7,951,405	4,152,835	1,162,003	861,972	50,830,278
80 km/h and over, less than 100 km/h	23,351,446	7,013,708	5,187,359	2,015,993	1,852,373	39,420,879
100 km/h and over, less than 120 km/h	10,918,100	3,441,007	2,727,682	1,246,404	1,332,974	19,666,167
120 km/h and over, less than 150 km/h	3,262,987	600,510	427,468	180,034	189,023	4,660,022
150 km/h and over	601,744	44,377	17,343	5,410	4,609	673,483
Total	211,653,264	35,268,902	16,871,258	5,241,072	4,506,016	273,540,512

(vehicle-links)

Table 3 Percentage of vehicles removed in data cleansing



Speed rank＼DRM link length	Less than 500 m	500 m and over, less than1000 m	1000 m and over, less than 2000 m	2000 m and over, less than 3000 m	3000 m and over	Total
Less than 1 km/h	100.00%	100.00%	100.00%	–	–	100.00%
1 km/h and over, less than 10 km/h	0.25%	10.10%	29.09%	71.17%	80.71%	0.75%
10 km/h and over, less than 20 km/h	0.00%	0.00%	0.01%	6.14%	38.98%	0.02%
20 km/h and over, less than 40 km/h	0.00%	0.00%	0.00%	0.00%	0.81%	0.00%
40 km/h and over, less than 60 km/h	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
60 km/h and over, less than 80 km/h	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
80 km/h and over, less than 100 km/h	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
100 km/h and over, less than 120 km/h	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
120 km/h and over, less than 150 km/h	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
150 km/h and over	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Total	0.35%	0.24%	0.21%	0.23%	0.22%	0.32%

Large

Legend, Percentage

*Percentage of removed vehicles (%) = the amount of data subject to removal / the amount of data before the data cleansing