APPLICATION OF DATA MINING TECHNIQUES TO CONGESTION DATA ANALYSIS: THE CASE OF SAPPORO URBAN AREA

Mikiharu ARIMURA, Assistant professor, Dr.Eng., Department of Transportation Engineering and Socio-Technology, College of Science and Technology, Nihon University, TEL&FAX:+81-47-469-5355,E-mail: arimura.mikiharu@nihon-u.ac.jp

Toshiyuki NAITO, Senior Engineer, Transportation Department, Docon Co., Ltd., 1-5-4-1 Atsubetu-cho, Atsubetu-ku, Sapporo, 004-8585, Japan, Fax: +81-11-801-1521,E-mail: tn1254@docon.jp

Hironobu HASEGAWA,Research Associate,Dr.Eng.,Dept. of Civil and Environmental Eng., Akita National College of Technology,1-1 Iijimabunkyo-cho, Akita City, Akita, 011-8585,JAPAN,Fax: +81-18-847-6071,E-mail: hasegawa@ipc.akita-nct.ac.jp

Tohru TAMURA, Professor, Dr.Eng., Dept. of Civil Engineering and Architecture Muroran Institute of Technology, 27-1, Mizumoto-chou, Muroran, 050-8585, Japan, Fax: +143-46-5288 E-mail: tamura@mmm-muroran-it.ac.jp

ABSTRACT

This study aims to identify mid- and long-term characteristic congestion trends in the urban area by classifying time-series data collected at sensor-installed points using the k-means method as which a major unsupervised clustering technique, and to support measure planning for each point using the results obtained from the classification.
In this study, temporally and spatially characteristic congestion patterns were extracted from a large amount of congestion data obtained from sensors installed at approximately 2,200 locations across Sapporo urban area where has heavily and snowy weather condition.
The identification of regular congestion patterns that occur at certain locations and hours is expected to facilitate support for the planning of traffic measures that require temporal and spatial consideration.
As the result of this study, congestion trends and congestion-point distributions which are including of summer and winter seasons were then classified into a number of patterns, allowing the selection of effective measures and the identification of targets for countermeasures.

Keywords: Road Performance Measurement, ITSs, Data Mining, Clustering Analysis

1. INTRODUCTION

The spread of intelligent transport systems (ITSs) has led to a significant increase in the amount of data recorded during application processing. In addition, the recent wave of new public management has also created a need for performance indicators and visualization techniques for administrative services.

Large amounts of detailed ITS data can be used to meet these needs, but the data recorded during ITS processing are not intended to be used statistically for measuring road traffic performance. Although congestion data – a valuable resource in terms of traffic management – are commonly used, such data are mainly leveraged to assess projects by comparing pre- and post-implementation situations and to conduct aggregative analyses focusing on the scale of development activities, including support for decision making in project planning (e.g., priority explicit curves for new roads investments).

However, optimal use of infrastructure, including the improvement of operation methods, has recently been debated in Japan. In addition to the need to increase road capacity, the selection of appropriate measures (including provision of services to users, determination of the locations and timing for the implementation of such measures, and the setting of effective target areas) can also involve important assessment items in terms of traffic management. Again, large amounts of detailed ITS data can meet these needs. In analyzing such data, however, along with an approach in which data are collected to test hypotheses and are statistically examined, it is important to use a hypothesis-finding approach in which statistically testable data groups and characteristic patterns are mechanically selected to enable evaluation of the underlying hypotheses.

In this study, temporally and spatially characteristic congestion patterns were extracted from a large amount of congestion data obtained from sensors installed at approximately 2,200 locations across Sapporo City. The identification of regular congestion patterns that occur at certain locations and hours is expected to facilitate support for the planning of traffic measures that require temporal and spatial consideration.

2. THE PURPOSE OF THIS STUDY

The Vehicle Information and Communication System (referred to below as VICS) data used in this study were provided by the Japan Road Traffic Information Center. Through this system, drivers can obtain congestion and traffic regulation information via their car navigation systems in real time.

Many previous studies using VICS data have estimated or predicted travel times in accordance with the main purpose of VICS. Funabashi *et al* (2003) developed a short-term travel-time prediction technique that uses similarities in travel-hour fluctuation patterns for limited spatial and temporal ranges. Yamane (2004) proposed a technique that estimates future congestion trends by processing VICS data, thereby improving information provision to

road users. Tsukahara *et al.* (2005), targeting a wide-area road network, proposed a VICS-information prediction technique based on the nearest-neighbor method and an interpolation method for roads without such information. Ando *et al.* (2006) developed the Vehicle Routing and Scheduling Problems with Time Windows-Probabilistic (VRPTW-P) model, which accumulates travel-time information from VICS data or other data sources and uses travel-time distribution as historical information.

This study aims to identify mid- and long-term characteristic congestion trends of the Sapporo urban area which is the largest city of Japan's cold, snowy region by classifying time-series data collected at sensor-installed points using the k-means method and to support measure planning for each point using the results obtained from the classification. Here, in Sapporo, snowing makes the road condition totally change. Therefore this study paid more attention to the characteristic congestion trends of the variation between summer and winter season.

3. OVERALL DATA TRENDS

3.1 Data outline

Congestion data collected by prefectural traffic and road administrators and summarized by the Japan Road Traffic Information Center were used in this study. These data are released to the public through media such as VICS, TV and radio. Congestion status was monitored at five-minute intervals at 2,200 points in Sapporo between April 1, 2003 and March 31, 2008 as period of five years. The data were recorded only when congestion was observed, and a total of 5,091,077 records were collected during the period. Each of the congestion data sets contains eight fields: date, hour, route, direction, address, congestion length, latitude (degrees, minutes and seconds) and longitude (degrees, minutes and seconds). Figure 1 shows the locations at which sensors are installed.

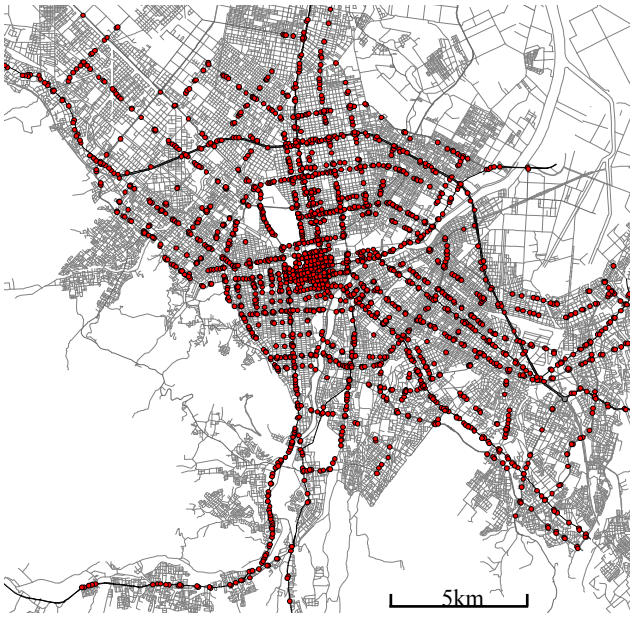

Figure 1 Distribution of sensors in Sapporo

3.2 Data preprocessing

As the data were not suitable for analysis in their raw form, they were preprocessed for each fiscal year through the following steps: first, for each point, 1) the annual hourly congestion frequencies (for zero to 23 hours) and 2) the annual monthly congestion frequencies (for April to March) were summarized; then, for each of the days on which congestion was observed, the total congestion frequency was calculated for all points. Table 1 shows the types of preprocessed data and their purposes in this study.

Table 1 – Types and purposes of preprocessed data

Data		Purpose
By point	Annual hourly congestion frequency	Identification of hourly fluctuation patterns
	Annual day-of-the-week congestion frequency	Identification of day-of-the-week fluctuations
	Annual monthly congestion frequency	Identification of monthly fluctuations
	Annual congestion frequency	Narrowing-down of data to be targeted
By day	Total congestion frequency of all points	Identification of annual trends

3.3 Congestion frequency and distribution

Figure 2 shows the relationships between the daily congestion frequency and the number of points at which congestion occurred between FY 2003 and FY 2007. The horizontal axis indicates the total congestion frequency observed per day, and the vertical axis indicates the total number of points where congestion occurred. For each fiscal year, plots for April to November (summer) and for December to March (winter) are visually differentiated in the graph.

In summer, congestion occurred at around 600 points, often at the same locations. However, in winter, congestion occurred at over 1,200 points on some days, and Figure 2 indicates that both the number and area of congestion points increased with the level of snowfall.

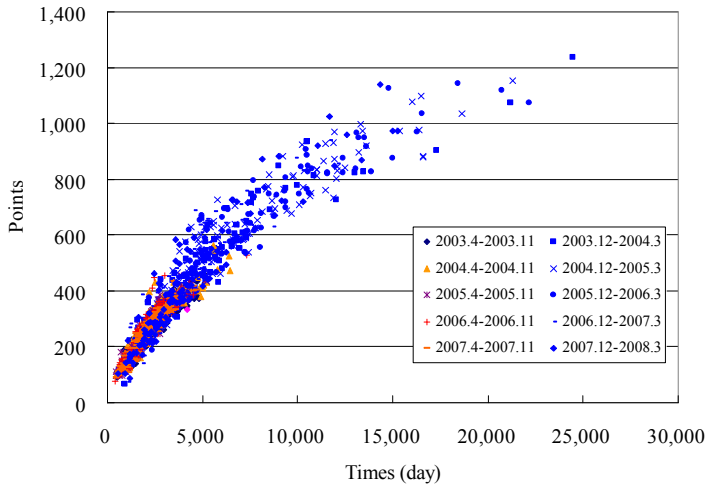

Figure 2 Distribution of the daily congestion frequency and the daily number of congestion points

Figure 3 shows the daily congestion frequency at all points between FY 2003 and FY 2007 arranged in ascending order. The vertical axis indicates congestion frequency, and the horizontal axis indicates date ID. Here, the sumer season is from April to November and the winter season is from December to March. The top 350 dates are concentrated in winter

Compared with summer, winter congestion occurs unexpectedly and expands spatially, and the actual travel time is often different from that perceived by drivers. Although foreseeable fluctuations in travel time can be incorporated into driving plans, unexpected congestion events are likely to greatly compromise the service level of road travel.

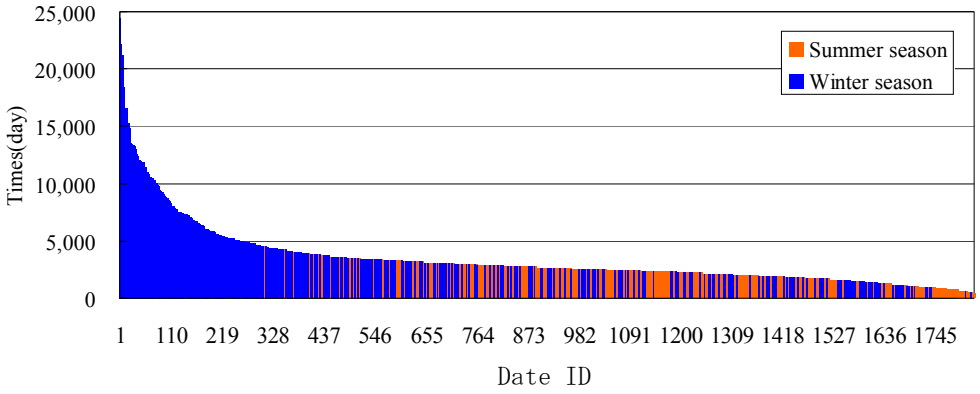

Figure 3 Daily congestion frequencies (All sensor-installed points: FY 2003 – 2007)

3.4 Identification of spatial trends

Here we used the data of four-year period from FY2003 toFY2006 for the follow analysis because of lack of data attribute in FY2007. To extract points with high congestion frequency in summer and winter over, all locations were sorted in decreasing order of congestion frequency for the periods of summer and winter, and those with an accumulated congestion frequency within the top 80% of the total frequency for all points in summer and winter were extracted. Summer and winter were defined as the periods between April and November and between December and March, respectively. By way of example, the distribution of congestion frequency in summer and winter in 2006 is shown in Figure 4. The points plotted to the left of the dotted line in the figure are those below the 80% line.

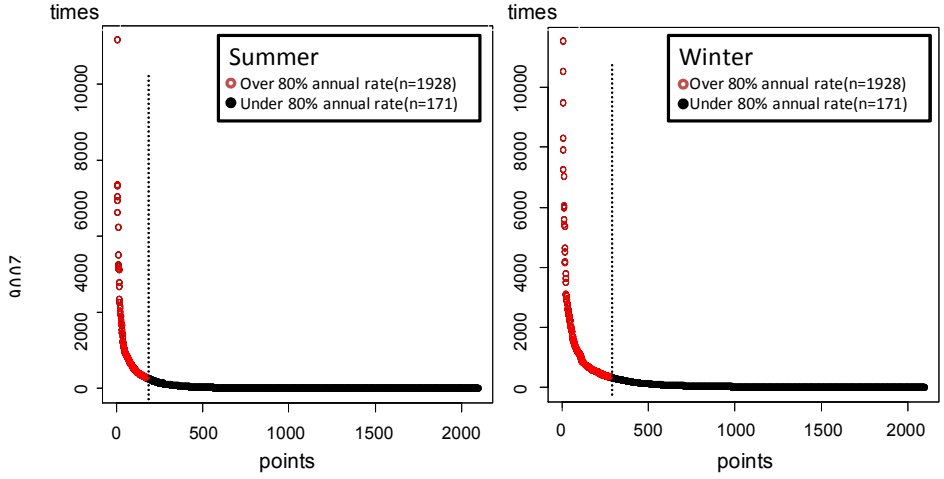

Figure 4 Distribution of annual congestion frequencies in FY 2006

Next, points whose accumulated congestion frequency rate was in the top 80% from FY2003 to FY2006 both for summer and winter were extracted. A conceptual diagram of this is shown in Figure 2. The points whose congestion frequency rate was under the 80% line in the fourth year both for summer and winter are shown in Table 2.

Table 2 – The number of points under the 80% congestion frequency rate

		SUMMER	WINTER	SUMMER ∩WINTER
The number of points	2003	309	213	192
	2004	344	286	253
	2005	338	242	222
	2006	171	272	137

The number of points whose accumulated congestion frequency was below the 80% line throughout the four years for both summer and winter was 99. With regard to points with high congestion frequency regardless of season, countermeasures will be necessary in terms of hardware.

Additionally, for low-congestion points below the 80% line (the tail part in Fig. 1), it is possible that some locations have problems such as the simultaneous occurrence of multiple instances of congestion, and therefore require countermeasures. Detailed analysis of this situation, including the extraction of congestion patterns in consideration of micro-time zones, is a future issue for investigation.

3.5 Extraction of data for cluster analysis

Next, data were extracted to enable cluster analysis for summer and winter. The data for 2006 were used here. Due to the importance of information from points with high congestion frequency in winter in developing countermeasures, 272 such points were selected for analysis. However, 13 of these had zero or a low number (10 or fewer) of instances of congestion in summer, and were excluded from the analysis since they would not produce meaningful time-series patterns for summer. Clustering was therefore performed with the remaining 259 points. Figure 5 shows a conceptual diagram of the analysis data extraction, and Figure 6 shows the locations of the congestion points extracted.

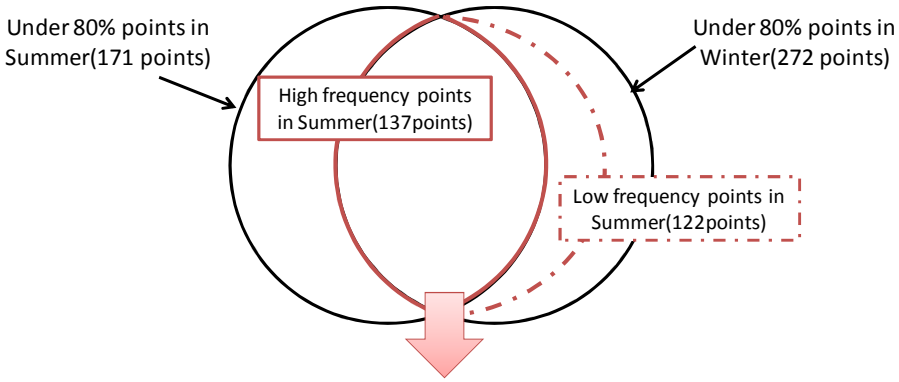

Figure 5 conceptual diagram of the analysis data extraction

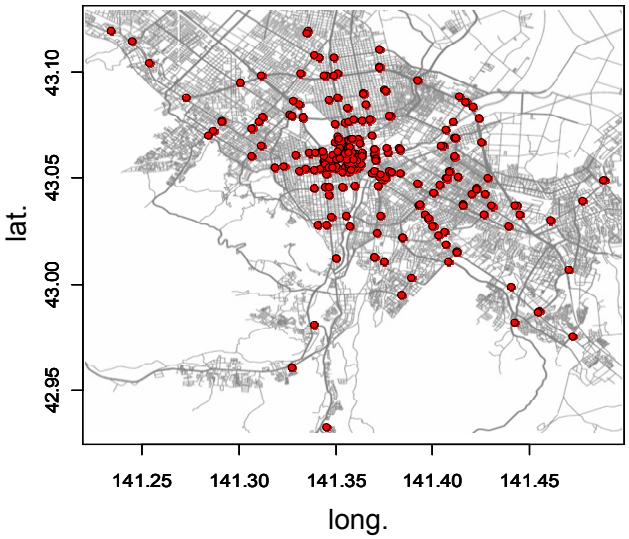

Figure 6 Target for clustering analysis

4. IDENTIFICATION OF PATTERNS FROM CONGESTION DATA

4.1 Temporal fluctuation trends

To determine the temporal fluctuation patterns of congestion in summer and winter at points of high winter congestion where sensors were installed, the figures for congestion frequency in summer and winter at each of the 259 points extracted were normalized using the total congestion frequency in summer and winter, and a graph was created (Fig. 7). Multiple congestion patterns coexisted in the time-series data for each point, making the identification of overall trends difficult. Accordingly, we analyzed fluctuations in the number of congestion hours observed for each point using the k-means method as non-hierarchical clustering approach to identify time-series patterns.

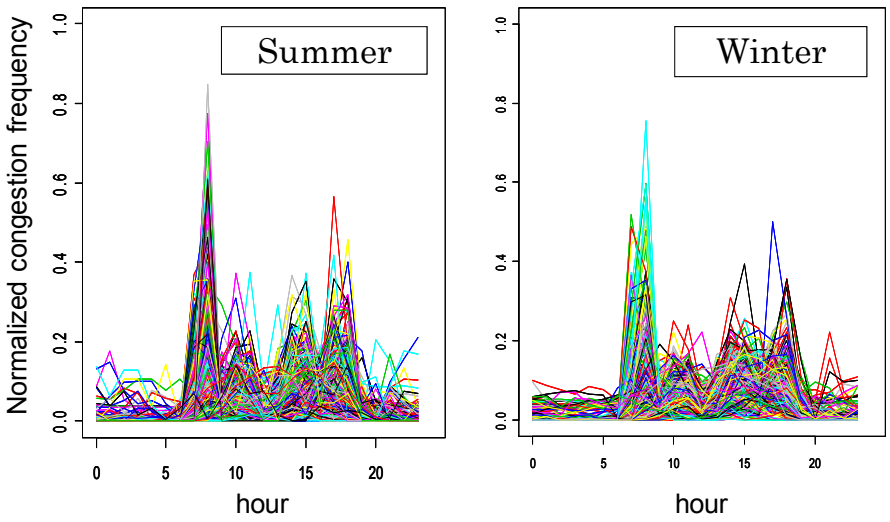

Figure 7 Daily congestion frequencies

4.2 Clustering by the k-means method

Using each time-zone frequency for summer and winter as an input variable, points with similar time-series congestion patterns were clustered using the k-means method. Figure 8 shows the time-series pattern of each cluster. The default number of clusters was set at five based on a study conducted by the author and colleagues4) in Sapporo, in which hypotheses for developing measures were most easily formed when classifying congestion time-series representations into five patterns.

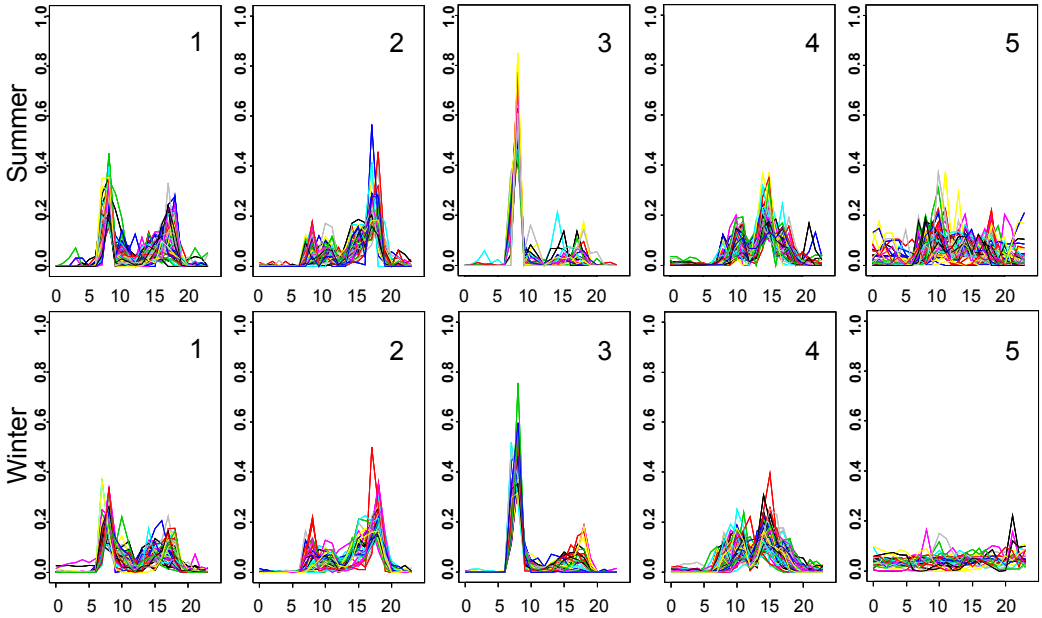

Figure 8 Daily congestion frequencies

4.3 Discussion

The temporal characteristics of each cluster are as outlined below (Table 3).

Table 3 – The number of points under the 80% congestion frequency rate

Cluster	Summer	Winter
1	Peak in morning and evening (Commuting traffic)	
2	Peak in the evening (Commuting traffic)	
3	Peak in the morning (Commuting traffic)	
4	Peak around 10 a.m. and 3 p.m.(Commercial traffic)	
5	Mixied distridution wth peak around 10 a.m. and daytime contiguous	No peak, chronic congestion

Summer clusters 1-4 and winter clusters 1-4 had similar time-series patterns. Congestion frequency showed differences between summer and winter, but the time-series patterns were generally the same between the two seasons.

Based on the results of cluster analysis for summer and winter, the seasonal fluctuation of congestion patterns was examined for each congestion point. Since summer and winter both

showed five time-series patterns, there were 25 congestion-point distributions. In this study, 20 congestion-point distribution patterns were extracted from cluster relocation among two seasons. Figure 9, Figure 10, Figure 11 and Figure 12 show the 4 cases of congestion points of shifting time-series pattern among two seasons.

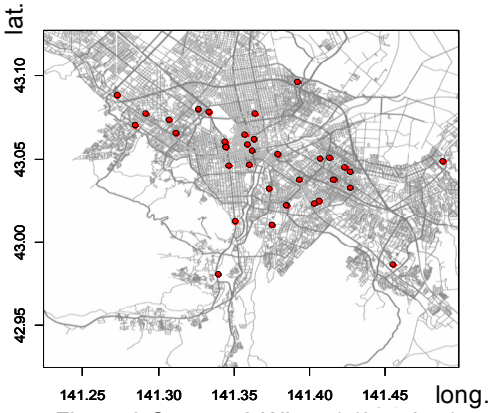

Figure 9 Summer2-Winter1 (37points)

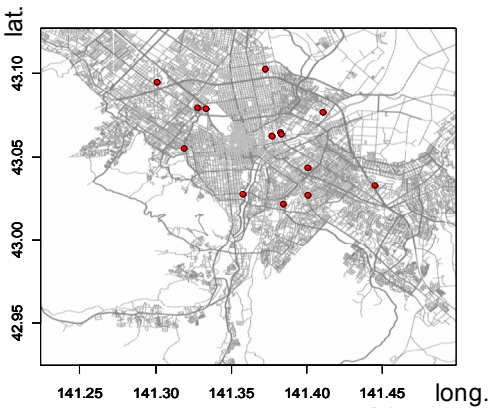

Figure 10 Summer3-Winter3 (14points)

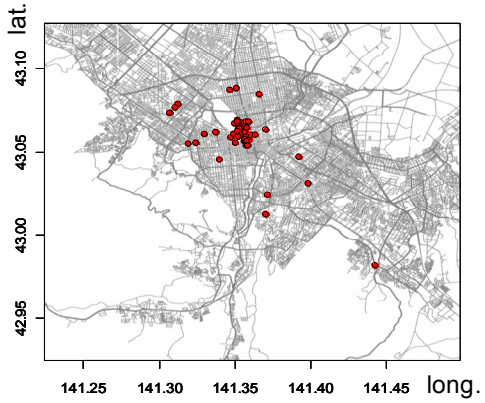

Figure 11 Summer4-Winter5 (53points)

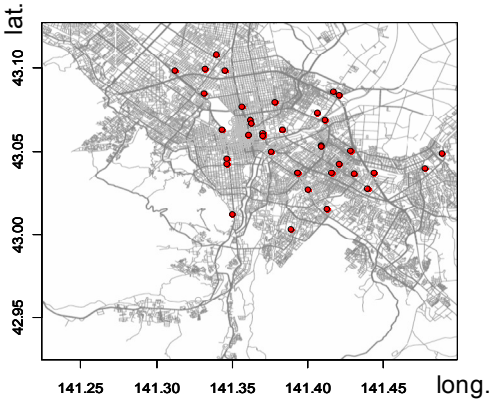

Figure 12 Summer1-Winter4 (37points)

a) Shift to morning-peak type

At the points in summer 1–Winter 3 and Summer 2–Winter 1 (Fig.9), the time-series patterns underwent a shift to a morning-peak type between the seasons. These points were dotted across the city. At these locations, winter-specific measures (such as snow removal before the morning peak hours and prioritized spreading of antifreeze to prevent road-surface freezing in the morning) would be effective.

b) Summer and winter both showing peaks in the morning

At the points in summer 3–Winter 3 (Fig.10), congestion was concentrated in the morning throughout the year. The points were situated on the Kanjo Line and the Toyohira River (bridges). It was suggested that these bridges represent bottlenecks caused by temporal and spatial factors, resulting in a traffic capacity deficiency in the morning. Possible countermeasures include encouragement of off-peak commuting aimed at road users and companies and the promotion of public transportation usage (e.g., introduction of a park-and-ride system).

c) Commercial traffic flow pattern to Constant congestion

At the points in summer 4–Winter 5 (Fig.11), congestion was caused in summer by commercial traffic, but in winter, roads were constantly congested with no peaks. These points were mainly situated in the city center. For summer congestion, urban-type countermeasures (such as the setting of zones and hours in which the concentration of goods handling is required), may be suggested. However, it is difficult to select effective measures against winter congestion, and the provision of information that is useful for selecting routes and predicting arrival times in consideration of expected congestion on bulletin boards and by other means may be suggested.

d) Commuting traffic flow pattern to Commercial traffic flow pattern

At the points in summer 1–Winter 4 (Fig.12), congestion peaks moved from the morning and late afternoon in summer to around 10 a.m. and 3 p.m. in winter. It was suggested that these points had seasonally different congestion factors related to commuting traffic in summer and commercial traffic in winter. These points were dotted across the city, and congestion countermeasures need to be different between summer and winter. In summer, the provision of information on less-congested parallel routes via bulletin boards and other means is suggested. In winter, for the city center, the setting of hours during which illegal parking is rigorously controlled and the concentration of goods handling is suggested, and for arterial roads, snow removal and the provision of information on alternative routes for different time zones is proposed.

Based on these results, it was possible to develop congestion countermeasures with a higher level of detail than those reported in other studies by identifying temporal and spatial congestion trends in consideration of their seasonal fluctuations.

5. CONCLUSION

In this study, a dataset created using congestion data recorded in Sapporo was classified using the k-means method to identify spatial congestion trends and daily fluctuation trends of congestion. Congestion trends and congestion-point distributions in the city were then classified into a number of patterns, allowing the selection of effective measures and the identification of targets for countermeasures.

Future issues to be addressed include: 1) establishment of a congestion prediction model that uses past congestion patterns, 2) refinement of the congestion prediction model through matching with meteorological data and other information, 3) development of outcome indicators that reflect the congestion characteristics of each area and are easier for local residents to understand, and 4) comprehensive understanding of congestion through analysis using data from all congestion points.

Acknowledgement: The authors would like to express their gratitude to the Sapporo Development and Construction Department of the Hokkaido Regional Development Bureau for providing congestion data.

REFERENCES

Funabashi, K., Nishimura, S., Horiguchi, R., Akahane, H., Kuwahara, M., Oneyama, H. (2003): Short Term Travel Time Prediction Using Historical VICS Data, Proceedings of Infrastructure Planning, Vol. 27, CR-ROM. Japan Society of Civil Engineers.

Yamane, K. (2004): Short Term Travel Time Prediction Using Historical VICS Data, Proceedings of Infrastructure Planning, Vol. 29, CD-ROM, Japan Society of Civil Engineers.

Tsukahara, S., Furukawa, T., Hara, K., Karino, H. (2005): Prediction technique of wide area VICS data using the nearest neighbor method based on time similarity, the 67th National Conference of the Information Processing Society of Japan, 6F-5, Information Processing Society of Japan.

Ando, N., Taniguchi, E., Yamada, T. (2006): Advanced vehicle routing and scheduling problems with ITS, Proceedings of Infrastructure Planning, Vol. 33, CD-ROM, Japan Society of Civil Engineers.