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 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. 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 in the city 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, ITS, Data mining, Clustering analysis

1. INTRODUCTION

The spread of intelligent transport systems (ITS) 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 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 road investments).

The 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. TRENDS IN TRAFFIC DATA ANALYSIS AND 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, which targets the urban area of Sapporo (the largest city of Japan’s cold, snowy region), aims to identify mid- and long-term characteristic congestion trends in the area by classifying time-series data collected at sensor-installed points using the k-means method (a major unsupervised clustering technique) and to support measure planning for each point using the results obtained from the classification.

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 (a 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, a) the annual hourly congestion frequencies (for zero to 23 hours) and b) 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>
<thead>
<tr>
<th>Data</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual hourly congestion frequency</td>
<td>Identification of hourly fluctuation patterns</td>
</tr>
<tr>
<td>Annual day-of-the-week congestion frequency</td>
<td>Identification of day-of-the-week fluctuations</td>
</tr>
<tr>
<td>Annual monthly congestion frequency</td>
<td>Identification of monthly fluctuations</td>
</tr>
<tr>
<td>Annual congestion frequency</td>
<td>Narrowing-down of data to be targeted</td>
</tr>
<tr>
<td>Total congestion frequency of all points</td>
<td>Identification of annual trends</td>
</tr>
</tbody>
</table>

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](image)

Figure 3 shows the daily congestion frequency at all points between FY 2003 and FY 2006 arranged in ascending order. The vertical axis indicates congestion frequency, and the horizontal axis indicates date ID. Here, the summer season is from April to November and the winter season is from December to March. The top 350 dates are concentrated in winter.

![Figure 3. Daily congestion frequencies](image)
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.

3.4 Identification of spatial trends

To identify spatial trends of congestion, data for FY 2007 – the most recent full-year set – were examined. Figure 4 shows the annual congestion frequencies at all points for FY 2007 arranged in descending order. The right vertical axis indicates congestion frequency, and the left vertical axis indicates cumulative congestion frequency. The points indicated in red experienced 80% of all congestion events when annual congestion frequencies at all points are accumulated.

The congestion frequencies at all 2,236 points were arranged in ascending order to enable calculation of cumulative percentages. A cumulative percentage of 80% was attained by the top 267 points, and the congestion frequencies recorded at other points were very small. Eighty percent of all congestion events were recorded at approximately 12% of all points. This study attempted to identify spatial trends of congestion targeting these 267 points. It should be pointed out, however, that other points with lower annual congestion frequencies may also require countermeasures for problems such as the frequent occurrence of congestion during snowfall. The extraction of such congestion patterns with seasonal variations is an issue for future study.

4. IDENTIFICATION OF CONGESTION PATTERNS

4.1 Identification of temporal fluctuation trends
To identify trends in temporal fluctuation, hourly congestion frequencies at the 267 points detailed in the previous section are shown in Figure 5. Several fluctuation patterns were confirmed, making it difficult to identify characteristic temporal fluctuation patterns of congestion and the points where such patterns occur.

![Figure 5. Hourly fluctuations in congestion events (FY 2007)](image)

Targeting points with high congestion frequencies, this study performed clustering analysis using the k-mean method to identify temporal and spatial fluctuation patterns of congestion. Additionally, in accordance with the characteristic time-series congestion patterns and the spatial distribution of congestion on the road network for each cluster (determined by clustering analysis) congestion patterns that are useful in terms of management tasks – such as the estimation of congestion factors and the planning of countermeasures – were extracted.

### 4.2 Clustering by the k-means method

#### 4.2.1 Outline of the k-means method

Cluster analysis methods consist of hierarchical and non-hierarchical techniques, and the k-means method is one of the major non-hierarchical ones. With this method, clustering is performed using k cluster centers randomly assigned as initial values. Samples are classified by assigning each one to the nearest cluster center. As a classification standard, the squared Euclidean distance between the cluster center and each sample is used. Once samples have been assigned, new centers are set again. This procedure is repeated until a suitable level of convergence has been reached, thereby producing optimal cluster centers and completing data classification. Unlike hierarchical methods, this technique produces several clusters.

Since the clustering results obtained by the k-means method depend on initial values, clustering was performed by randomly changing the initial values several times, and the
clustering results for which the function below produced the minimum value were used for analysis. For this study, 10,000 trials were performed. The number of clusters was determined by testing several times the number for which analysts can come up with countermeasures, based on the relationship between temporal fluctuations in congestion and congestion points.

\[
obj = \sum_{i=1}^{K} \sum_{x \in C_j} \|x - c_j\|^2
\]

(1)

Input: The number of clusters after division (K), and the data set
Output: Assignment of data to clusters to minimize the evaluation function above

4.2.2 Cluster analysis using the hourly congestion frequency as a feature quantity

Using the hourly congestion frequency as an input variable, samples were classified by the k-means method into clusters with similar congestion time-series patterns. The characteristics of each cluster are shown in Table 2, and the sensor locations and time-series patterns for each cluster are shown in Figure 6. Points were grouped into 47 clusters, and based on the spatial characteristics of and congestion patterns at congestion points, interviews regarding potential countermeasures were conducted with consultants and experts who know the study area well. As a result, the number of clusters for which analysts can comfortably interpret congestion occurrence patterns in the Sapporo urban area was identified as five.

Table 2. Cluster characteristics

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>No. of points</th>
<th>Temporal fluctuation characteristics</th>
<th>Spatial characteristics</th>
<th>Examples of support for measure planning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>192</td>
<td>Many congestion patterns</td>
<td>/Distribution across the urban area /National Route 36 /Near Ishiyama Dori road</td>
<td>Produced too many points to come up with measures for. Clustering was therefore performed again by changing the feature quantity.</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>12</td>
<td>Peaks in the morning and early evening /Multiple congestion patterns, including nighttime congestion at 2 points</td>
<td>/Sapporo Shindo road /Kanjosen road /Ishiyama Dori road in the city center</td>
<td>Provision of information on congestion status and alternative routes to users of Sapporo Shindo. With regard to commercial transport, quantification of delay risks and provision of feedback to companies.</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>27</td>
<td>Congestion from 10 a.m. to 3 p.m. /Relatively constant congestion events at three points during the day and at nighttime</td>
<td>/City center /National Route 5 toward Teine-ku /East side of Kanjosen road</td>
<td>Congestion patterns created in the activity area of commercial transport. Can be used to set access-restriction areas in the city center.</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>32</td>
<td>Congestion events concentrated in the hour from 8 a.m. /Peaks in the morning and early evening</td>
<td>/Kanjosen road toward the eastern and western sides /National Route 12</td>
<td>Promotion of public transport for commuting, and management such as setting of snow-removal hours</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>4</td>
<td>More congestion events than in other clusters /No decrease in congestion events in the hour from 12 p.m.</td>
<td>/Near the south exit of Sapporo Station /Near the north exit of Sapporo /Kita 1 East 1/West 1, Chuo-ku /Yonesato, Shiroishi-ku</td>
<td>With regard to the area near the station: introduction of ITS to taxi bays and exclusion of passing traffic</td>
</tr>
</tbody>
</table>

Figure 6 shows the numbers of congestion points, temporal fluctuation characteristics,
Clustering performed using hourly congestion frequencies showed that two factors – congestion frequency and hourly fluctuation patterns – influenced the classification results. Each cluster was stratified by congestion frequency, and the points in Cluster 1, which had fewer congestion frequencies, were classified into one large group. For Cluster 1, since points with the maximum congestion frequency of 800 or less were spread across the urban area and many congestion patterns existed, effective measures could not be designed. Clustering analysis was therefore performed again using feature quantities in which the congestion frequencies were normalized.

4.2.3 Cluster analysis using normalized hourly congestion frequency
In Cluster 1, the number of annual congestion frequencies recorded at five-minute intervals was lower than 800 – fewer than those in other clusters – whereas several time-series congestion patterns existed visually. Accordingly, as congestion countermeasures, traffic flow management measures – such as changes in road operation methods, provision of congestion-point information and alternative routes, and optimization of snow-removal work – may be more appropriate than a permanent increase in traffic capacity (i.e., infrastructure improvement). For this reason, in performing reclustering, the difference between the average congestion frequency for each hour and the congestion frequency for each point was divided by the standard deviation for normalization, and was used as a feature quantity for the congestion pattern at each congestion point. The numbers of congestion points, temporal fluctuation characteristics, spatial distribution characteristics and potential countermeasures for each cluster are shown in Table 3. In Figure 7, the top graphs show sensor locations, the middle graphs show time-series patterns normalized by the standard deviations, and the bottom graphs show the clustering results converted to hourly congestion frequencies.

4.3 Discussion on ITS-data mining and urban traffic management
In this study, frequent congestion points were extracted by processing and sorting large amounts of data obtained using VICS, and time-series congestion patterns at these points were
Table 3. Cluster characteristics
(hourly congestion frequencies averaged by the standard deviation)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>No. of points</th>
<th>Temporal fluctuation characteristics</th>
<th>Spatial characteristics</th>
<th>Examples of support for measure planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>13</td>
<td>Congestion in the evening after 5 p.m. and during the night</td>
<td>/Sapporo Ekimae Dori road</td>
<td>Thought to be due to taxis in the downtown area and distribution-related transport going into and coming out of the distribution center. With such a small number of congestion points, the priority of this cluster in terms of countermeasure planning is low.</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>36</td>
<td>Congestion in the hour from 8 a.m.</td>
<td>/Kanroson road</td>
<td>A result of regular commuting traffic. For short-distance commuting traffic in inner-city areas, traffic-mode switching measures such as the installation of bicycle lanes may be suggested.</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>60</td>
<td>Congestion peaking at 10 a.m., easing at noon and significantly worsening again in the hour from 3 p.m.</td>
<td>/Kotoni Ekimae Dori road</td>
<td>Congestion patterns created in the activity area of commercial transport. May be used for setting access-restriction areas in the city center.</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>41</td>
<td>Congestion peaking in the hour from 10 a.m. and then gradually easing</td>
<td>/Kanroson road toward Kita-ku</td>
<td>Mainly due to commercial traffic. Charges for traffic passing through the city center may be suggested.</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>42</td>
<td>Congestion peaking in the hour from 5 p.m.</td>
<td>/National Route 12 in the city</td>
<td>A result of after-work commuting traffic. When congestion is predicted, ITS measures, such as the transmission of congestion prediction information to mobile terminals, may be suggested.</td>
</tr>
</tbody>
</table>

Figure 7. Sensors’ locations, Standard normalized hourly congestion frequencies and hourly congestion frequencies
classified by cluster analysis using the k-means method. By extracting stable patterns, the large amounts of data could be converted to information that can be interpreted by analysts.

It should be noted, however, that the findings obtained were the feature quantities of subjectively set samples, and were based on results that were mechanically classified using a clustering algorithm that depends on default values. Data mining is a heuristic approach by which analysts obtain useful information from analysis results. In this sense, for urban traffic management using ITS data, a process in which such relations are coordinated with various restrictions and the empirical information of analysts to obtain new findings is essential, in addition to the analysis of inter-data relations hidden within large amounts of data obtained from traffic sensors installed in the urban road network.

In this study, clustering results were presented to several urban traffic experts for discussion of potential countermeasures based on the spatial distribution of and the time-series congestion patterns at the classified congestion points. Although various constraints must be taken into account in formulating actual measures, the authors believe it is possible to facilitate support for the planning of measures by clearly identifying the locations of congestion events and the congestion patterns found there.

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 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


