

# タグ付き時系列データの可視化と対話的な詳細度制御の一手法

八木佐也香<sup>†</sup> 内田悠美子<sup>†</sup> 伊藤 貴之<sup>†</sup>

<sup>†</sup> お茶の水女子大学大学院人間文化創成科学研究科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: <sup>†</sup>{sayaka,yumi-ko,itot}@itolab.is.ocha.ac.jp

**あらまし** 時系列データの可視化において、折れ線の集合でデータを表示する手法は広く普及しており、データ解析等の目的にも用いられている。しかし、データが大規模になると線分同士の交差が増え、数値の時系列変化を適切に読み取ることが困難になる。また、その多くはグラフの値や形状のみをもとに処理を行っており、天気などのメタ情報と数値の関連は考慮されていない。そこで本論文では、クラスタリングを利用した詳細度制御を行うことで、グラフの読みやすさを向上し、かつ数値とメタ情報の関係性も逃さない表示手法を提案する。これにより、例えば運動・食事と血圧の変化の関係等、メタ情報と数値・時系列との相関性を発見・分析できると考えられる。

**キーワード** 情報可視化, タグ付き時系列データ, 詳細度制御

## A Polyline-Based Visualization Technique for Tagged Time-Varying Data

Sayaka YAGI<sup>†</sup>, Yumiko UCHIDA<sup>†</sup>, and Takayuki ITOH<sup>†</sup>

<sup>†</sup> Graduate School of Humanities and Sciences, Ochanomizu University

Otsuka 2-1-1, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: <sup>†</sup>{sayaka,yumi-ko,itot}@itolab.is.ocha.ac.jp

**Abstract** We have various interesting time-varying data in our daily lives, such as weather data (e.g., temperature and air pressure) and stock prices. Such time-varying data is often associated with other information: for example, temperatures can be associated with weather, and stock prices can be associated with social or economic incidents. Meanwhile, we often draw large-scale time-varying data by multiple polylines in one space to compare the time variation of multiple values. We think it should be interesting if such time-varying data are effectively visualized with their associated information. This paper presents a new time-varying data visualization technique which supposes the associated information is attached as tags of the values of each time step. The technique enables the level-of-detail by means of clustering similarly-shaped and -tagged polylines. It also provides a user interface so that users can intuitively discover interesting local features from large number of polylines.

**Key words** Information visualization, Tagged timevarying data, Level-of-detail control

### 1. Introduction

“Big data” is a common term of the information technology industry in recent years. Various computer systems store huge information from various datasets including computer system logs, sensor measurements, and transactions. These datasets easily get huge amount because they are stored time-by-time. Information visualization is a useful and effective approach for understanding, analysis, and monitoring of such time-varying datasets.

Polyline chart is one of the most popular representation for time-varying datasets. Newspapers and Web sites show various polyline charts for time-varying datasets including stock

prices, foreign currency, sports performance, and weather measurements such as temperatures. We also commonly draw multiple time-varying values in a single polyline chart space so that users can compare the time-varying values. For example, when we want to visually compare the temperatures at multiple locations, we would draw all of them in a single polyline chart space.

On the other hand, we often deal with hundreds or even thousands of time-varying values in the above mentioned various fields. We may want to visually compare the time variation of stock prices of many companies, or temperatures of many locations. It result in severe cluttering, and this may cause difficulty in reading data values and following data

trends, if we draw hundreds or thousands of polylines in a single space. Thus, improved techniques for polyline-based visualization are required.

Meanwhile, such time-varying data is often associated with other information: for example, temperatures can be associated with weather, and stock prices can be associated with social or economic incidents. Such information may be tightly correlated with the time-varying values. For example, minimum temperature will be much lower in the morning of sunny days in winter due to radiative cooling. On the other hand, difference between minimum and maximum temperature will be smaller in the cloudy or rainy days. This kind of correlation between time-varying values and associated information can be useful for various purposes such as retrieval of frequent and outlier patterns. However, it is not always easy to determine the correlation between them. We believe it should be useful if time-varying data visualization techniques simultaneously display such associated information to assist the understanding of the correlation.

This paper presents a new time-varying data visualization technique which supposes tags assigned to the values of each time step. We suppose that tags consist of a set of predefined terms: for example, sunny, cloudy, and rainy for weather data, or exercising, eating, or sleeping for health care data. As a preprocessing, the technique clusters polylines based on their shapes and tags, and then selects representative polylines from the clusters. It realizes smooth level-of-detail control by interactively controlling the number of polylines to be displayed. Also, the technique features click and sketch interfaces so that users can interactively select particular polylines which are tagged with the user-interested terms.

This paper presents the effectiveness of the presented technique with Japanese weather data recorded by AMeDAS (Automated Meteorological Data Acquisition System). The dataset consists of time-varying temperature values with weather tags including “Clear”, “Sunny”, “Cloudy”, “Rainy”, and “Snowy” at 376 observation points. We visually discovered different patterns of temperature variations from the visualization result.

## 2. Related Work

Polyline chart is one of the most popular representation for time-varying data in our daily life, and many recent works have proposed polyline-based techniques to visualize time-varying data. Polyline-based visualization technique has the following features:

- Better to read the finely values than visualization techniques which use colors to denote values.
- Suitable for represent continuous values (e.g. the measurements of science).

Wattenberg et al. presented a sketch-based query interface to search for specific shapes of polylines [1]. Hochheiser et al. pre-sented Timeboxes and TimeSearcher [2], a gradient- and range-based query interface for polyline-based time-varying data visualization. This technique constrains the number of polylines to display by selecting range of rectangles. It is useful to reduce visual clutter and to observe local features of polylines. Though, it is hard to observe overall features of polylines from the constrained polylines, and not always possible to select interesting polylines. Our technique solve these problems by shape- and tag-based clustering, and interactive selection of polylines.

Some works focused on similarity-based pattern and outlier discovery. Buono et al. presented a technique to interactively search for similar pattern [3] as an extension of TimeSearcher, and a similarity-based forecasting technique [4] for the forecasts of the future patterns. Lin et al. presented a technique to discover non-trivial patterns [5], by clustering a set of time-varying values and searching for outliers. Wang et al. presented a technique for important polyline selection [6], by conducting a block-wise analysis of time-varying volume data in the joint feature-temporal space. Recently we presented two time-varying data visualization techniques, featuring sketch query on the clustered view [7], and pattern display on the heatmap [8].

Though, there has been comparatively little research which visualize time and associated information simultaneously.

## 3. Level-of-detail control for time varying data visualization

This section briefly introduces a level-of-detail control and a sketch interface [7] for time-varying data visualization previously we presented.

The technique supposes the following time series data, consisting of a set of values  $\mathbf{P} = (p_1, p_2, \dots, p_n)$  represented as  $n$  polylines. We describe the values of a polyline as  $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{im})$ ;  $p_{ij}$  denotes the value at the  $j$ -th time of the  $i$ -th polyline. We draw the set of values as a polyline chart, while the horizontal axis denotes the 1st to the  $m$ -th time, and the vertical axis denotes the magnitude of the values.

As a preprocessing, the technique temporarily quantizes polylines, generates clusters of them, and selects representative polylines from the clusters. While the quantization step, the technique generates a grid surrounding all polylines, and calculates intersections between the polylines and grid-lines. It then generates rough polylines by connecting the intersections, and uses them for the clustering. Number of clusters can be controlled by the resolution of the grid as well as similarity threshold values, and our implementa-

tion prepares several clustering results so that the number of representative polylines smoothly varies.

The technique then initially displays representative polylines. Smoothly switches clustering results, it can seamlessly change the number of representative polylines to be displayed. The technique also provides a click interface, so that users can specify interesting representatives by directly clicking. It also provides a sketch interface, so that users can specify interesting representatives which have partial shapes similar to the sketched curves.

#### 4. Extension to the tagged time-varying data visualization

This section proposes an extended visualization technique for tagged time-varying data. This paper extends the aforementioned time series data as follows: we describe the tags of the  $i$ -th polyline as  $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{im})$ ;  $w_{ij}$  denotes the tag at the  $j$ -th time of the  $i$ -th polyline, as well as  $p_{ij}$  denotes the value at the  $j$ -th time of the  $i$ -th polyline.

##### 4.1 Clustering and Representative Polyline Selection

The extended technique displays adequate number of representative polylines to reduce the cluttering among the polylines and improve readability. The technique clips the polylines interval by interval, and then generates clusters of clipped polylines for each interval, where clipped polylines in a cluster is similarly shaped and tagged.

The procedure of quantization and clipping step is as follows.

(1) Generate a grid covering the drawing area, and then divides into  $a \times b$  subspaces as shown in Figure 1(a).

Here this thesis formalizes the grid as follows:

- $h_i$  is the  $i$ -th horizontal line of the grid ( $0 \leq i \leq b$ ).
- $v_i$  is the  $i$ -th vertical line of the grid ( $0 \leq i \leq a$ ).
- $t_i$  is the time at  $v_i$ .
- $b_i$  is the value at  $h_i$ .

(2) Sample  $P$  at  $t_0$  to  $t_a$ , and temporarily quantize the sampled values at  $b_0$  to  $b_b$ . Here,  $b_i$  is the value at the  $i$ -th horizontal line of the grid.

(3) Generate groups of polylines, if the polylines have the same quantized values both at  $t_{i-1}$  and  $t_i$ , as shown in Figure 1(b).

(4) Clip polylines of a group by  $t_{i-1}$  and  $t_i$ , as shown in Figure 1(c), and generates clusters of the clipped polylines.

Figure 2 denotes the procedure for clustering of tagged polylines, while the colors of polylines denote their dominant tags. Here, the technique regards the clipped polylines as  $n$ -dimensional vectors, while they contain  $n$  time steps between  $t_{i-1}$  and  $t_i$ .

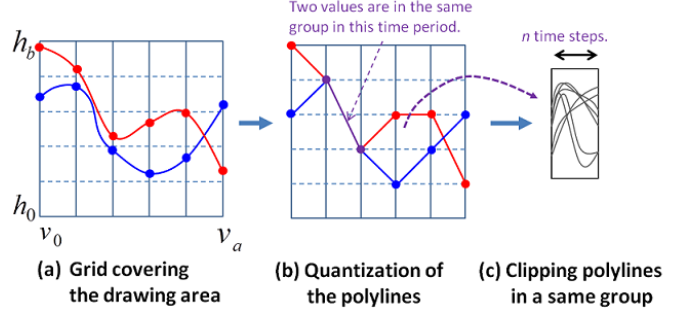


Figure 1 Quantization and clipping of polyline.

(1) Divide the clipped polylines according to tags, as shown in Figure 2(b), using a dendrogram from the polylines constructed according to similarities of their  $k$ -dimensional vectors ( $w_{jt_{i-1}}, \dots, w_{jt_i}$ ).

(2) Apply a non-hierarchical clustering, e.g. k-means, to the polylines in each cluster, using their  $n$ -dimensional vectors ( $p_{jt_{i-1}}, \dots, p_{jt_i}$ ), and generate clusters consisting of similarly tagged and shaped fragments of polylines, as shown in Figure 2(c).

Here, we predesignate the number of clusters as specified by users.

(3) Select representative polylines, for each cluster, as shown in Figure 2(d).

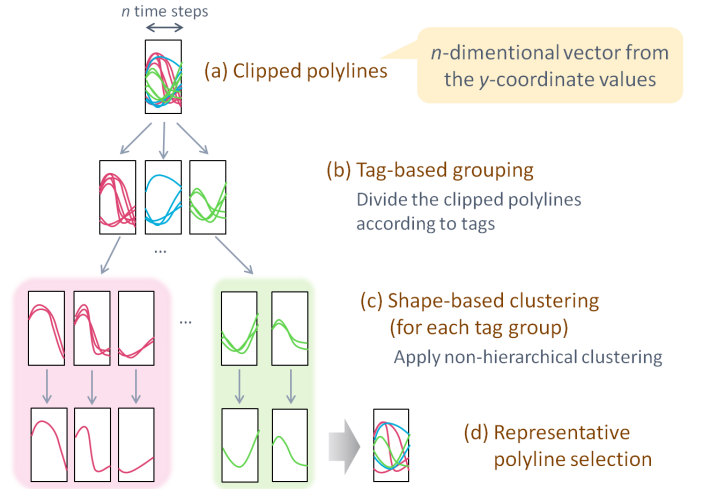


Figure 2 Clustering of polylines.

Our current implementation simply extracts a representative polyline, which is the closest to the center of a cluster in a  $n$ -dimensional vector space. This strategy is basically good because it can select average polylines. Moreover, by using the strategy, we can firstly designate the number of clusters, then designate the number of polylines. It is adequate for our purpose to control the number of displayed polylines. Here, if one or more polylines in a cluster have been already selected as the representative polylines of other

clusters, the technique does not select any new representative polyline from the current cluster, so that we can reduce the total number of representative polylines.

## 4.2 Interactive Visualization

### 4.2.1 Display of Representative polylines

The extended technique represents the time-varying data as colored polylines. It assigns colors to the tags (e.g. red to sunny, and gray to cloudy, to the tags of weather data), and draws the polyline in the assigned colors. If the vertices of a segment of a polyline have different tags, it interpolates the colors along the segments. The technique also features selective polyline display based on the tags. Users can select particular tags so that the technique can draw only corresponding parts of the polylines. They can also select particular tags to be filtered from the display of the polylines.

Initially our technique draws only the representative polylines. Here, number of representative polylines depends on the resolution of the grid and the threshold of the clustering process. Our current implementation generates several clustering results, with several configurations of the grid and the clustering process. Smoothly replacing the clustering results, our technique seamlessly displays several levels of numbers of representatives.

Here, the technique adjusts transparency of representative polylines in each time steps. One drawn representative polyline has time steps which are selected as representative and are not selected as representatives as shown in Figure 3. In time steps which are selected as representative, the technique draws a polyline as low transparency based on the number of original polylines in the cluster. Meanwhile, it draws a polyline as quite high transparency in time steps which are not selected as representative. Through this process, the technique can reduce the visual clutters while displaying consecutive polylines. Users can also observe how many original polylines are represented as the drawn polylines.

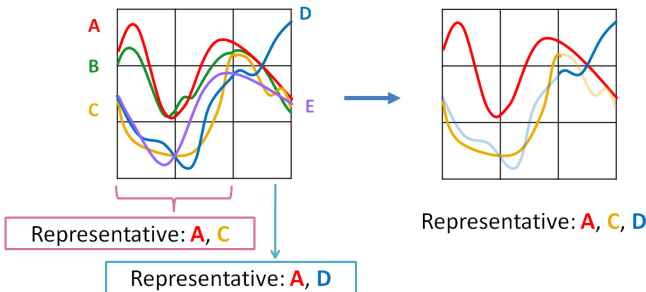


Figure 3 Visualization of representative polylines.

### 4.2.2 Interactive Extraction of Polylines

The technique provides click and sketch interfaces, so that users can specify interesting representatives by directly clicking or sketching particular shapes.

When a user clicks a point on the display, the technique calculates distances between the point and all segments of the drawn polylines. If at least one of the segments of a polyline is enough close to the clicked point, the technique highlights the current polyline. When a user draws a curve on the display, as shown in Figure 4, the technique samples several points on the curve, and calculates distances between y-coordinate values of the sampled points and y-coordinate values of all segments of the representative polylines. Concretely, the technique compares one point on the sketched curve to the points of the representative polylines which have closest x-coordinate value with it. If at least one of the segments of a polyline is enough close to each of the sampled points, the technique highlights the current polyline.

Users may want to specify polylines by logical operation of the polylines specified by multiple click/sketch operations. To satisfy such a requirement, the technique can implement the logical operation. The technique figures out logical AND/OR operation of several times inputs, and highlights the calculated representatives. The technique can also highlight polylines which have similar shape with sketched segment at a different height. It calculates slope between two points instead of comparing y-coordinate values.

Users can specify particular tags to be extracted by the above query operations. It can highlight only the parts of the polylines corresponding to the specified tags while the click or sketch operations. It can also reactivate only the parts of the non-representative polylines corresponding to the specified tags.

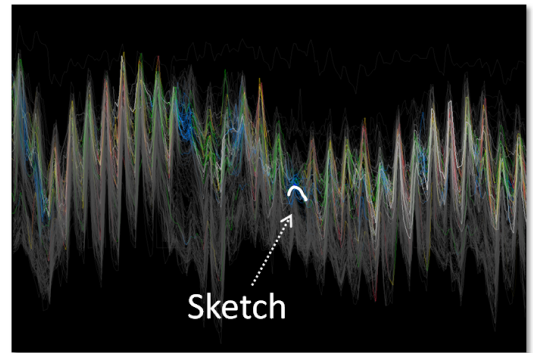


Figure 4 Sketch interface.

## 5. Examples

We developed the presented technique with JDK (Java Development Kit) 1.6, and executed on a personal computer (CPU 2.67GHz Dual Core, RAM 2.0GB) with Windows Vista (32bit). The processing time of the technique is estimated as follows. The time for level-of-detail control is  $O(n(a + s))$ ; for quantization and clipping is  $O(na)$  and for

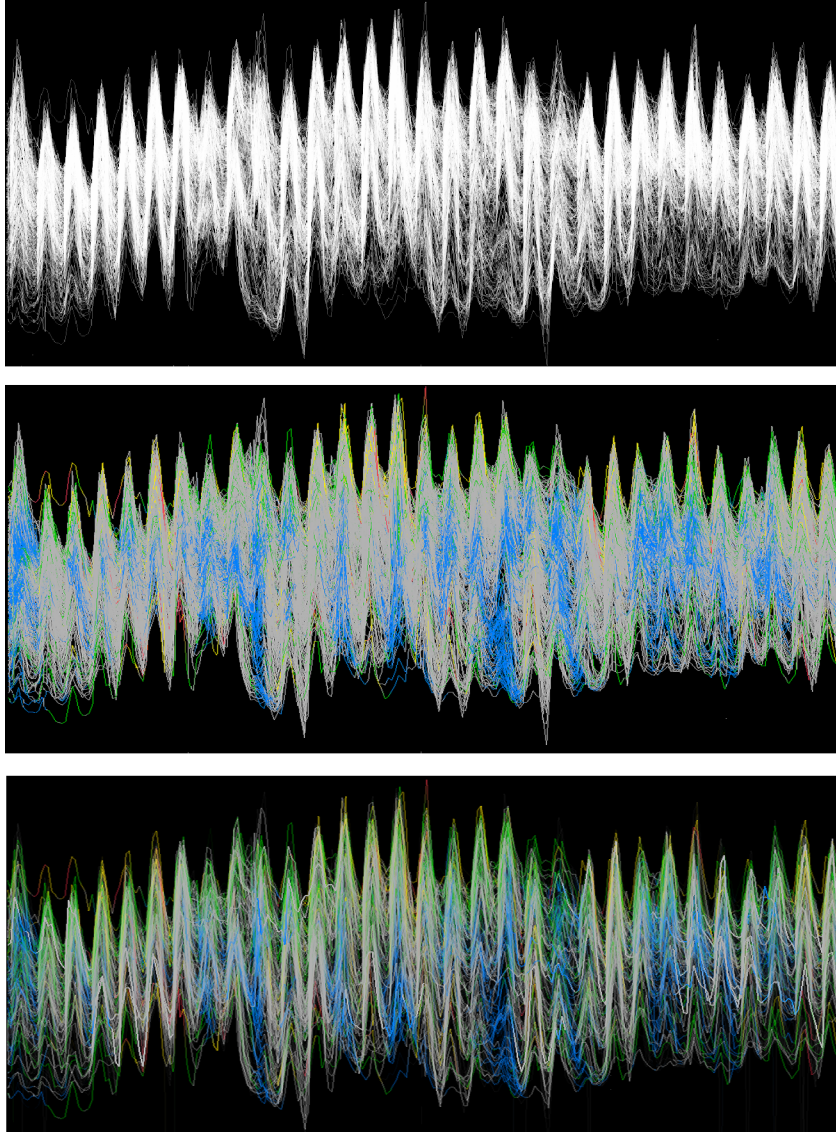


Figure 5 (Upper) Overview without tags. (Center) Overview with tags(before applying level-of-detail control). (Lower) Overview with tags(after applying level-of-detail control).

Table 1 Average number of sunny day in July (1981-2010 normal value)

Sapporo	Sendai	Niigata	Tokyo	Nagoya	Osaka	Hiroshima	Takamatsu	Fukuoka	Kagoshima	Naha
13.0	9.7	14.2	12.1	14.2	16.5	15.9	17.7	15.7	17.9	24.1

clustering is  $O(ns)$ , for sketch interface is  $O(rq)$ .

Here, we formalize the each variable as follows:

- $n$  is the total number of polylines
- $a$  is the number of sampled time steps
- $s$  is the total number of time steps
- $r$  is the number of representative polylines
- $q$  is the number of sketched time steps

In our measurement, the average time for level-of-detail control was 177 milliseconds, and the average of sketch process was 1 milliseconds, which  $n = 376$ ,  $a = 10$ ,  $s = 240$ ,  $r = 249$ , and  $q = 8$ . This result denotes that the technique is enough fast for the interactive visualization.

We applied Japanese weather data recorded by AMeDAS (Automated Meteorological Data Acquisition System) to the presented technique. We extracted time-varying temperature data observed at 376 points around Japan in every 3 hours. Again, each polyline denotes each observation point, the horizontal axis denotes time steps and the vertical axis denotes temperature values.

We then assigned weather tags including “Clear”, “Sunny”, “Cloudy”, “Rainy”, and “Snowy” to temperature value of each time of each point. Here, we assign one tag to a particular time of a particular polyline, and do not cover in the case where multiple tags are simultaneously assigned.



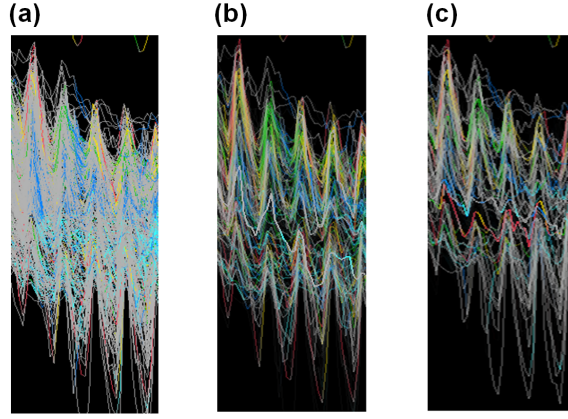


Figure 6 Zoom up view: (a) without level-of-detail control. (b) with level-of-detail control (tag-based clustering). (c) with level-of-detail control (shape-based clustering).

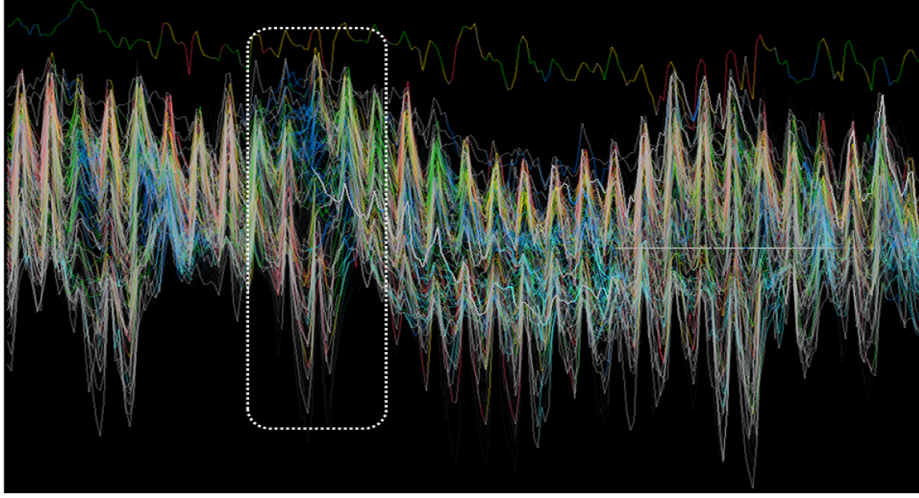


Figure 7 Overview applying level-of-detail control with tag-based clustering.

Our implementation draws the segments of “Clear” polylines in red, “Sunny” in yellow, “Cloudy” in green, “Rainy” in blue, and “Snowy” in cyan, interpolating the colors if tags of two ends of a segment is different. Exceptionally it may draw the segments in gray if we cannot obtain the weather data. While using click or sketch interfaces, our implementation draws selected polylines brightly, and others in gray.

### 5.1 Example (1)

We visualized a dataset during 1 to 31, July, 2009. Figure 5 shows an overview of the temperature data. Figure 5(Upper) shows an overview without tags, and Figure 5(Center) shows an overview with tags. Here, the both results show the results including all observation points before applying level-of-detail control. While Figure 5(Upper) displays just major and outlier variations of temperature, Figure 5(Center) demonstrates much more information. Colors in the dense parts denote major weather of the days. Blue (“Rainy”) is observed in several consecutive days and it appears several times during the period. This result well demonstrates the effectiveness of the visualization of tagged time-varying

data, since it is difficult to obtain such knowledge from Figure 5(Upper).

To observe the detailed variation, we applied level-of-detail control to the original polylines including all observation points. Figure 5(Lower) shows the results. In this figure, severe overlaps of polylines are reduced, especially non-tagged ones colored in gray. As mentioned above, we can observe long rain have continued in various regions. We can also observe only a few red or yellow (“Clear” or “Sunny”) days, while there are about 10 to 20 sunny days in an average year as shown in Table 1.

### 5.2 Example (2)

We visualized a dataset during 1 to 31, December, 2009. Figure 6 shows a zoom up view of temperature variation during five days. Figure 6(a) shows the original view of the tagged temperature data before applying level-of-detail control, and Figure 6(b) shows after applying level-of-detail control with tag-based clustering. Figure 6(a) has many overlaps of polylines, while Figure 6(b) represents the features of the data more clearly. In Figure 6(b), we can find that extremely

lower temperature can be often observed from “Clear” or “Sunny” points, and it shows typical feature of winter temperature due to radiative cooling. It is difficult to obtain such knowledge from the visualization results with shape-based clustering without considering tags, as shown in Figure 6(c). Here, representative polylines are selected based on only their shape. In Figure 6(c), polylines which have extremely lower temperature is drawn as gray, and missed important weather features as mentioned above. The result denotes that we can observe overall features without missing detailed features while reducing cluttering among polylines by the effort of level-of-detail control considering tags on clustering step.

Figure 7 shows the overview applying level-of-detail control with tag-based clustering. In the middle of month indicated by a dotted white rectangle, we can observe that temperatures drawn as red or yellow polylines have been drastically decreased. While temperatures drawn as blue or cyan polylines have been slightly varied.

To observe long-term variation of 2 patterns above, we clicked the polylines. Figure 8(Upper) shows the result which we clicked low-range points while selecting two tags “Clear” and “Sunny”; consequently polylines tagged as “Clear” or “Sunny” at the clicked points were highlighted. It represents that range of temperature between the daytime and the night is relatively large. We can also find that extremely lower temperature can be often observed from “Clear” or “Sunny” points. Figure 8(Lower) shows the result which we clicked high-range points while selecting two tags “Rainy” and “Snowy”; consequently polylines tagged as “Rainy” or “Snowy” at the clicked points were highlighted. It represents that range of temperature between the daytime and the night is relatively small, and the average is a bit higher than temperatures of “Clear” and “Sunny”. We can also observe sunlight is little through the periods.

These results denote that we can discover interesting local features from large number of polylines, and observe the features by the effort of interactive extraction of polylines.

## 6. Conclusion

This paper presented a polyline-based visualization technique for tagged time-varying data. The paper first described the definition of the tagged time-varying data, and presented techniques for level-of-detail control and interactive polyline selection. It also demonstrated effectiveness of the technique by applying temperature data with weather tags.

Our potential future work includes the followings:

**[Many kinds of tags:]** Since our current implementation represents tags as colors, it may be difficult to visually distinguish if we have many kinds (e.g. more than 10) of tags.

We would like to discuss what kinds of visual metaphor can be more effective for the representation of more kinds of tags.

**[Multiple tags at a point:]** It is also difficult for our current implementation to represent if multiple tags are simultaneously assigned to a particular time of a particular polyline. Again, we would like to discuss what kinds of visual metaphor can be more effective for the representation of multiple tags.

**[Observation of tag-change:]** It is interesting for several kinds of data to observe how time series values vary when the assigned tags change. We would like to add features to the technique so that we can focus on time series value variation with particular patterns of tag changes.

**[More applications and tests:]** We would like to apply more various data to the technique, including medical measurement datasets, system measurement datasets, and stock price datasets. We would also like to have experiments for subjective and objective evaluations of the technique.

## Reference

- [1] M. Wattenberg, D. Jones, “Sketching a Graph to Query a Time-Series Database”, *SIGCHI Conference on Human Factors in Computing Systems Extended Abstract (CHI2001)*, 381-382, 2001.
- [2] H. Hochheiser, B. Shneiderman, “Dynamic query tools for time series data sets: Timebox widgets for interactive exploration”, *Information Visualization*, 3(1):1-18, 2004.
- [3] P. Buono, A. Aris, C. Plaisant, A. Khella, B. Shneiderman, “Interactive Pattern Search in Time Series”, *Conference on Visualization and Data Analysis '05*, 175-186, 2005.
- [4] P. Buono, C. Plaisant, A. Simeone, A. Aris, B. Shneiderman, G. Shmueli, W. Jank, “Similarity-based forecasting with simultaneous previews: A river plot interface for time series forecasting”, *11th International Conference on Information Visualisation*, 191-196, 2007.
- [5] J. Lin, E. Keogh, S. Lonardi, “Visualizing and Discovering Non-trivial Patterns in Large Time Series Databases”, *Information Visualization*, 4(2), 61-82, 2005.
- [6] C. Wang, H. Yu, K.-L. Ma, “Importance-driven time-varying data visualization”, *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547-1554, 2008.
- [7] Y. Uchida, T. Itoh, “A visualization and level-of-detail control technique for large scale time series data”, *13th International Conference on Information Visualisation*, 80-85, 2009.
- [8] M. Imoto, T. Itoh, “A 3d visualization technique for large scale time-varying data”, *14th International Conference on Information Visualisation*, 17-22, 2010.

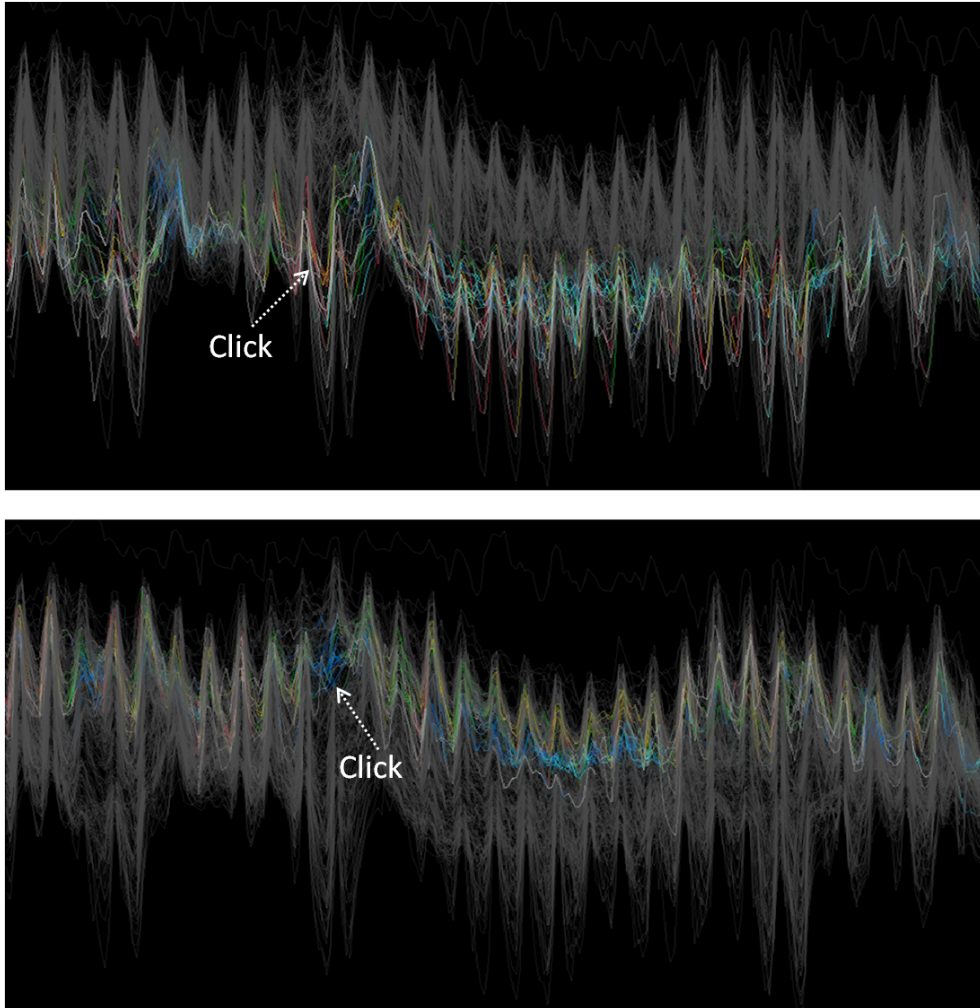


Figure 8 Interactive selection. (Upper) Polyines tagged as “Clear” or “Sunny” at the clicked points are highlighted. (Lower) Polyines tagged as “Rainy” or “Snowy” at the clicked points are highlighted.