Exploratory Visual Analytics for Winter Road Management Using Statistically Preprocessed Probe-Car Data

Yuzuru Tanaka, Hajime Imura, and Jonas Sjöbergh

Abstract Social CPSs (Cyber-Physical Systems) denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure systems. They utilize both cyber data stored in databases and physical data coming from sensor networks in the target physical world for the analysis and optimized control of urban infrastructure systems such as traffic, energy, and water services. This paper focuses on the winter road management in Sapporo where we have the world biggest annual snow fall among the cities with more than 1 million populations. For monitoring the road conditions over the whole city, the use of probe-car data without violating personal data protection is fundamental. This paper first shows that probe car data statistically preprocessed over road links for an urban-scale area still allow us to visualize the dynamic change of the traffic flow in terms of the divergence and flow vector field. These give us sufficient information about the dynamic change of hotspots of traffic, main traffic streams, and route selection preference. The paper also shows more complex and advanced analyses of such data, especially for better winter road management in Sapporo. We extend the well-known multiple coordinated views framework for exploratory visual analytics to multiple coordinated views and analyses by integrating analysis tools with their result visualization views into the same environment. These newly added views may also coordinate with others, and allow users to directly select clusters or mined patterns calculated at runtime to further quantify the underlying database view. Exploratory visual analytics with such an environment enables us to detect road links for effective pinpoint snow removal.

Y. Tanaka (✉)
Graduate School of Information Science and Technology, Hokkaido University,
Kita 13 Nishi 8, Kita-Ku, Sapporo, Hokkaido, Japan
e-mail: tanaka@meme.hokudai.ac.jp

H. Imura • J. Sjöbergh
Meme Media Laboratory, Hokkaido University, Kita 13 Nishi 8, Kita-Ku,
Sapporo, Hokkaido, Japan
e-mail: hajime@meme.hokudai.ac.jp; js@meme.hokudai.ac.jp

Y.-L. Lin et al. (eds.), *Smart Sensors and Systems*,
DOI 10.1007/978-3-319-14711-6_6

1 Introduction

Advances of sensor devices and their networking technologies have enabled the real-time monitoring of physical systems ranging from small-size healthcare devices to large-scale plant systems, or even to urban-scale infrastructure service systems. Social CPSs (Cyber-Physical Systems) denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure systems. A CPS originally denotes an integration of a control system and a computer system that utilizes both cyber data stored in files or databases in the computer system and physical data coming from sensor networks in the target physical system for the sand boxing of embedded medical devices and the optimized operation of large plant systems. A social CPS also utilizes both cyber data stored databases and physical data coming from sensor networks in the target physical world for the analysis and optimized control of urban infrastructure systems such as traffic, energy, and water services. This paper focuses on the winter road management in Sapporo, where we have the world biggest annual snowfall among the cities with more than 1 million populations. Since the targets of Social CPSs are necessarily complex systems of systems, they need to deal with a variety of heterogeneous cyber and physical, real-time and retrospective big data including, for example, probe-car data, weather data, snow removal records, and traffic accident records.

One of the problems we are facing these days in big data R&Ds may be a big gap between the core technology R&Ds and the application R&Ds especially for complex target systems. Through the involvement in a Japanese government-initiative project from 2012 to 2016 on social cyber-physical systems for optimizing social system services such as the snow plowing and removing in Sapporo City, we have been facing difficulties filling the gap between varieties of available data analysis methods and the goal of finding optimized resource schedulings for snow plowing and removing.

For the analysis of dynamically changing traffic and road conditions in an urban-scale area, probe-car data may play the most important role. Inherently they are real time data, and have the potential to cover urban-scale areas. They can tell us not only about dynamically changing traffic and road conditions, but also about people's dynamically changing mobility demands and/or activities.

With the increasing use of car navigation systems networked to service centers, urban-scale or even nationwide scale probe car data are accumulated at service providers and ready for advanced meaningful analyses. However, their advanced utilization is currently facing two major obstacles. First, they are accumulated in different silos belonging to different service providers, and are not open for any analyses by third parties. Second, there are privacy concerns since each trajectory obtained from probe-car data may reveal lots of personal information about its driver, such as his or her home location, office location, frequently visited places like malls, and visits to some acquaintances. The second obstacle is often one of the main reasons for service providers not to open those data silos to public use of their probe-car data. The service providers are generally too cautious in providing their probe-car data to third parties even for nonprofit utilization. On the other hand, probe-car data can inherently give us lots of information about traffic and road conditions, route selection preference, travelling time, and daily or seasonal dynamic change of population mobility in urban-scale, or even nationwide-scale areas. Their utilization is fundamental in the analysis and optimization of the sustainable and safe management of urban infrastructure services such as energy supply and transportation. IT-based management systems for this are sometimes called social cyber-physical systems. IBM and Microsoft, for example, call such systems smarter city systems and urban computing systems, respectively.

There may be two known practical solutions to ensure the protection of personal information when using probe car data. One is to remove the head and tail of each trajectory for some constant time interval or for some constant distance to hide the origin and destination. Such modification can also further fragment the remaining part of each trajectory into segments of some constant time interval or of some constant distance. Another approach is to provide only statistically processed probe car data including for example the average speed, the maximum speed, and the number of cars in each direction of each road link during every time interval of a fixed length, e.g., 5 min. However, it has not been well discussed what kind of statistical data enables what kind of analysis. We need mathematical research on this subject. This paper will first show that we can calculate the divergence and the flow-vector field of the traffic only using the dynamic change of the average speed and the number of cars in each direction of each road link during each time interval of a fixed length. Using statistically preprocessed probe-car data of occupied taxi cars, we can visualize the temporal change of the hotspots for taking taxis and getting-off taxis during a day. We can also observe the temporal change of route preference depending on the traffic jams caused by heavy snowfalls through flow-vector field visualization of the traffic.

In the latter half of this paper, we will deal with more complex analyses for better winter road management in Sapporo. Probe-car data are fundamental to monitor and estimate the urban-scale dynamic change of traffic and road conditions of all the road links, and also to monitor the snow removal operations. We can use probe data from private cars and taxis for the former purpose, and from snow plowing and removing vehicles for the second purpose. Sapporo has 1.9 million citizens and an average annual snowfall of 6 m. It spends more than 150 million US dollars every winter just for snow plowing and removing. Our preparatory study on the influence of snowfall and snow removal operations on traffic revealed that the effect is not uniform even among consecutive road links along the same main route in the central city area. Because of this heterogeneity, we believe that macro analysis applied to the whole urban area cannot give us any meaningful result. One of the possible solutions may be exploratory visual analytics that enables us to freely repeat micro analyses, consisting of hypothesis making and leading to improvisational data segmentation and hypothesis checking through improvisational data analysis and visualization. This paper introduces the Geospatial Digital Dashboard as an open system for such exploratory visual analytics. This system exploits a coordinated multiple views framework, which is a well-known framework for exploratory visual analytics, and extends this framework to integrate analysis tools with their result visualization views as additional coordinated views. During the last 1 year and a half of our project, we have focused on the detection of those directed road links which seem to have caused severe traffic jams in nearby road links. This paper shows that exploratory repetition of micro analyses using the Geospatial Digital Dashboard enables us to detect such hotspot road links.

2 Related Research

Personal information might be handled in many different ways in the probe vehicle system [17]. Even if each trajectory data is anonymized there are many possibilities to identify the same vehicle among a large number of different trajectories through behavioral analysis. ISO/TC204/WG16 published international standard about personal data protection in probe vehicle system as ISO 24100 Intelligent transport systems—Basic principles for personal data protection in probe vehicle information services in 2009 [7]. Even if data cannot identify an individual directly, if it can do so indirectly it should be regarded as personal data, and as a target of protection. While the fragmentation of trajectory data into anonymized pieces of data may protect the privacy, trajectories cannot be reconstructed anymore, i.e., the data lose traceability. Some solutions are proposed ranging from practical ones either to trim both the head and the tail of each trajectory or to provide only statistical data for each road link, to more theoretical ones [1, 6, 13, 21]. The first one [1] protects privacy by shifting trajectory points in space that are already close to each other in time. Clusters of k trajectories are enforced to be close to each other so that they fall in the same area of uncertainty. In the second one [6], privacy is preserved by removing some points such that uncertainty between consecutive points is increased to avoid identification. Both of these assume the allowance to bring some uncertainty into trajectory data. Work in [21] limits the probability of disclosing the tail of the trajectories given the head of the trajectories. The last one [13] extends the concept of k-anonymity to trajectory data.

This paper focuses on one of the practical approaches: giving only statistically preprocessed data. While this lacks traceability, we show that we can still calculate the divergence rate and flow vector field of the traffic to analyze the hotspots and the dynamic changes of the traffic flow and the route preference.

There are many systems for visualization and exploration of data that use coordinated multiple views [16]. However, to the best of our knowledge, there are no other systems that provide an interactive visualization environment with coordinated multiple views and advanced analysis tools. We believe that allowing advanced analysis components in such an environment to create new complex data at runtime, and making such components coordinated with other views like any other visualization view components, are useful extensions of the coordinated multiple views framework. We have not seen any systems that allow interaction with visualization results that then affect the input to such analysis components, and that allow further visualization of and interaction with the results, where changes in any coordinated views affect all the other coordinated views and analyses.

There are systems that use graphical interfaces such as a flow chart style interface to set up what types of preprocessing, analysis, data mining, and visualization etc. should be done. Changes in the flow chart lead to updated visualizations, but interaction with the actual visualization results is very limited or not possible. Examples of such systems include RapidMiner [12] and DEVise [11].

There are systems that use coordinated multiple views of data and allow interaction with the visualization results. Selecting data in one view updates the other views to show only this selection, or clicking on one item in one view shows details about it in another view, etc. Creating new complex data in analysis components is, however, not possible. Examples of such systems include: SpotFire [2] Tioga-2 [3] (now TiogaDataSplash), and Snap-Together Visualization [14].

One example system quite similar to the Geospatial Digital Dashboard described in this paper is KNIME [4], which uses a graphical flow chart to set up processing and visualization of data. It allows adding new user built components and uses multiple linked views for visualization. Selecting subsets of data in one view highlights these in other views, but unlike our system, it does not trigger recalculation of related data mining results.

Another system with many of the features of Geospatial Digital Dashboard is Orange [5]. It sets up data flows in a flow chart, has both visualization and analysis components, allows user built components, and selections in a visualization component can trigger recalculation in data mining components. Unlike our system, two components cannot feed back into each other, so selections in one component can affect the other, but selections in the other component cannot be reflected back to the first.

There are also some precursors to our system. The VERD [18] system is based on IntelligentBox, a 3D version of Meme Media [19]. It visualizes relational databases and allows Web resources to be treated as relational schema. Unlike our system, visualizations are set up in a flow and interaction with visualization results only affects other visualizations in the downstream of the interaction point.

3 Divergence and Flow-Vector Field Analyses of Urban-Scale Traffic

Let us consider a small geographical area S_i. Let $P_i(t)$, $D_i(t)$, $G_i(t)$, $A_i(t)$ respectively denote the population, the divergence rate, the generation rate, and the absorption rate at time t of this area S_i. The population is the number of objects, e.g., cars in this case, in this area at time t. The divergence rate is the difference between the rate of the outbound flow from this area and the rate of the inbound flow to this area at time t. The generation rate and the absorption rate are respectively the increase rate of the newly generated objects and the decrease rate of the newly absorbed objects in this area at time t. For a sufficiently small time interval δ, the following holds:

$$P_i(t+\delta) - P_i(t) = (-D_i(t) + G_i(t) - A_i(t))\delta$$

There are two typical cases of mobility data. In case of mobile phone mobility data, we may assume that mobile phones are usually kept on, which implies that $G_i(t) = A_i(t) = 0$. Therefore, the following holds:

$$D_i(t)\delta = P_i(t) - P_i(t+\delta)$$

This implies that the divergence can be calculated simply from the time variation of the population during δ. In case of cars, we may assume that the flow is rather stable during a sufficiently small time interval, i.e., $P_i(t+\delta) \approx P_i(t)$. This implies the following:

$$D_i(t) \approx G_i(t) - A_i(t)$$

Let us consider the case of taxi cars. Suppose that the probe-car data are available with the distinction of empty cars and occupied cars. For the flow of occupied taxi cars, the rate $G_i(t)$ denotes the rate of taking a taxi in the area S_i, while the rate $A_i(t)$ denotes the rate of getting off a taxi in this area. The assumption $Pi(t+\delta) \approx P_i(t)$ means that the population in each area is stable and does not change during a sufficiently small time interval.

Suppose we have statistically processed probe data of occupied taxi cars showing the number of cars and the average speed in each road link at every 5 min. For a sufficiently small area S_i, every road link may cross the border of this area at one point, two points, or no point. The last case may have two cases, i.e., the whole road link may lie either inside or outside of this area. Figure 1 shows these four cases. It is obvious that the road links a, c, and d in this figure have no contribution to the divergence rate $D_i(t)$. The road links a and c have neither any outbound flow from this area nor any inbound flow to this area. The road link d has the same amounts of outbound flow and inbound flow to cancel each other. We only need to consider those road links such as b that cross the area border at only one point. Let XR_i denote the set of all the road links that cross the border of S_i, only once.

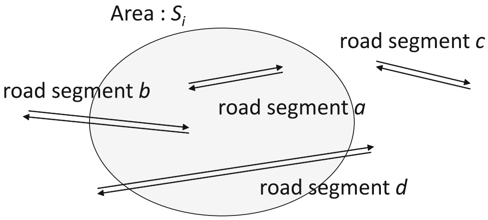


Fig. 1 Four different relationships between a road link and a small geographical area

Fig. 2 Outbound directed road link that crosses the border of S_i only once

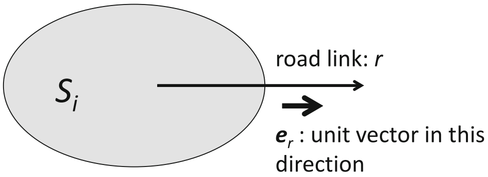


Figure 2 shows an outbound directed road link that crosses the border of S_i only once. Let the length of this road link r be L_r. Let the total number and the average speed of occupied taxi cars in this road link r at time t be respectively $N_r(t)$ and $V_r(t)$. Approximately we may conclude that cars in this area that are within the distance $v_r\delta$ from the border can go out of this area by time $t + \delta$. Approximately, we may assume that cars are uniformly distributed along this road link with the density of $N_r(t)/L_r$. Therefore, the total number of cars that will go out of this area along this outbound directed road link in the time interval δ can be approximated as $(N_r(t)/L_r)v_r(t)\delta$. Actually we have both inbound and outbound traffic on each road link, i.e., $(N_r^+(t), v_r^+(t))$ for inbound traffic, and $(N_r^-(t), v_r^-(t))$ for outbound traffic. If the road link r is a one way road, either of $N_r^+(t)$ or $N_r^-(t)$ always becomes zero. Therefore the following holds:

$$D_i(t)\delta = \Sigma_{r\in XR_i}(N_r^-(t)\cdot v_r^-(t) - N_r^+(t)\cdot v_r^+(t))\delta/L_r$$

From this equation, the divergence rate of the area S_i at time t can be approximately calculated as follows:

$$D_i(t) = \Sigma_{r\in XR_i}(N_r^-(t)\cdot v_r^-(t) - N_r^+(t)\cdot v_r^+(t))/L_r$$

Figure 3 shows the divergence rate of the occupied taxi traffic flow at 8:00 a.m. in the central area of Sapporo City, using gray scale. This was calculated for each cell in a mesh of 250 m by 250 m cells using the average speed and the number of occupied taxi cars in each road link at every 5 min. Around the center of this map is Sapporo Station. This station has two taxi stands, the north one and the south one. The visualized result shows that, in the morning, people get off taxis mainly at the north stand, and take taxis mainly at the south stand. Sapporo has its main business areas in the south of its main station. People like to take taxis at the south stand to go directly to their business. Since this causes the congestion around the south stand, people taking trains from the station like to get off taxis at the north stand.

Using the same data set, we can also calculate the traffic flow vector field. Let $\boldsymbol{F}_i(t)$ denote the traffic flow vector of a sufficiently small area S_i, and $\boldsymbol{e}_r$ denote a unit vector along a directed road link r which crosses the border of this area at least once. If the directed road link r is curved, we approximate the unit vector $\boldsymbol{e}_r$ to be in the direction from the origin to the destination of this directed road link. This time we need to consider the contribution of those directed road links that cross the border twice since the traffic along this directed road link may partially contribute a traffic

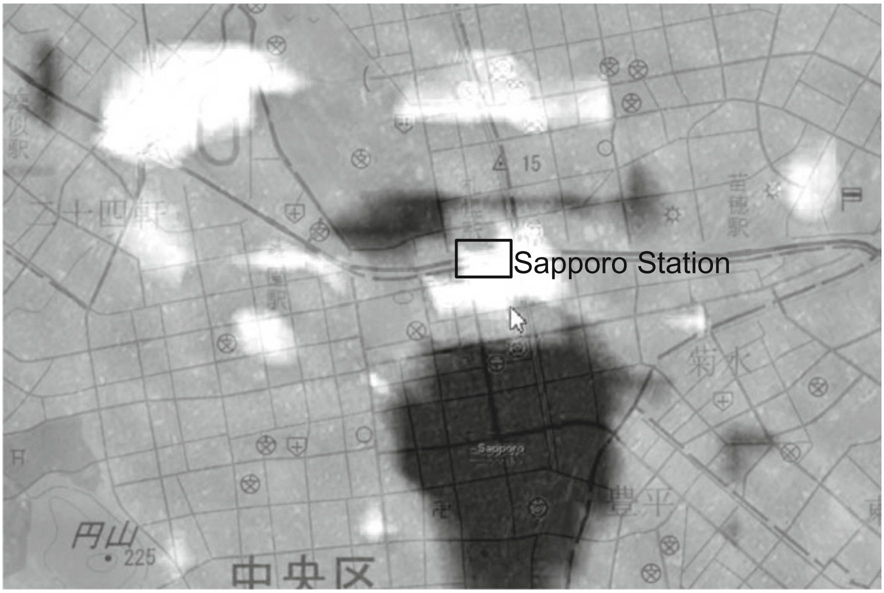


Fig. 3 A divergence rate of the occupied taxi traffic flow at 8:00 a.m.

flow passing through this area. The traffic flow vector $\boldsymbol{F}_{\boldsymbol{i}}\ (t)$ can be approximately calculated as follows:

$$\boldsymbol{F}_{\boldsymbol{i}}(t) = \Sigma_{r\in XR_i^*}(N_r(t)v_r(t)/L_r)\boldsymbol{e}_{\boldsymbol{r}}$$

Here we use XR_i^* instead of XR_i. The set XR_i^* is the set of all the directed road links that crosses the border of S_i. Figure 4a shows the flow vector field of the occupied taxi traffic at 19:00 in the central area of Sapporo City. Sapporo has a nightlife district called “Susukino” around the area 1 km to the south of Sapporo Station. Every evening, you can observe a big inbound flow of occupied taxis toward Susukino. Figure 4b, c show how route preference may change between a day in early December before any serious snowfall and a day in late December after heavy snowfalls. They show that snowfalls in late December cause traffic jams along the biggest southbound stream in early December, and shift the biggest southbound stream to the east.

These analyses show that statistically processed probe car data containing only the number of cars and the average speed in each road link will give us sufficient information about the divergence rate and the vector field of the traffic flow if the time interval is sufficiently small.

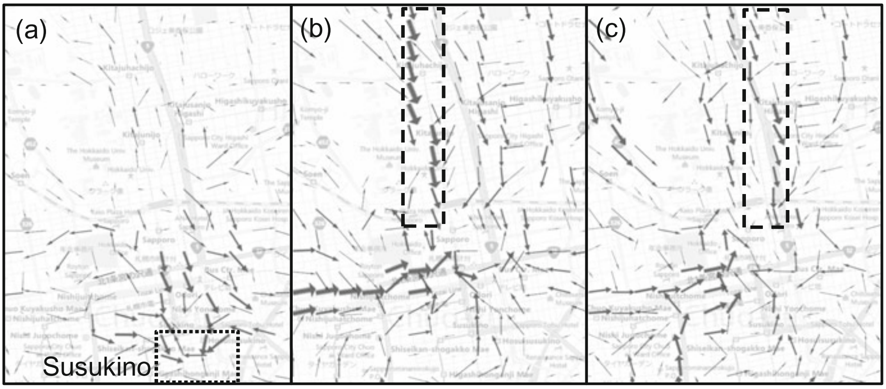


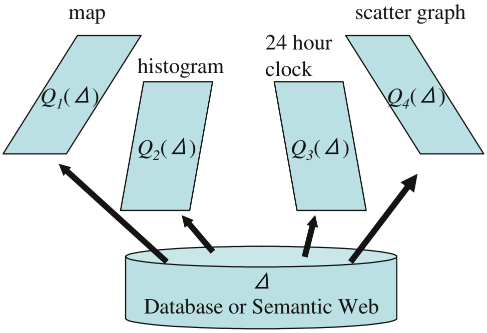
Fig. 4 Flow vector fields of occupied taxi traffic

4 Exploratory Visual Analytics for Pinpoint Snow Removal

Our preparatory study on the big data approach for efficient snow removal in Sapporo showed that macro analyses of probe car data for the whole central city area would not give any meaningful knowledge about the influence of snow fall, snow plowing and removing on traffics [20]. Using the average speed in each road link at every 5 min interval for 24 h, we can characterize each road link as a vector of 288 dimensions. Clustering analyses of road links represented by such vectors told us that even those links along the same route may fall in different clusters on a day with heavy snow fall. In summer time, they almost always fall into the same cluster. This means that the influence of snow on each road link may follow different models even on the same route. Therefore, we may think that influence of snow on traffic in the whole city cannot be modeled by a single monolithic model covering the whole city. It should be considered as a complex system of different models, each of which may be a simple monolithic model. Therefore, macro analyses applied to the whole central city area will not give any meaningful result. In order to obtain meaningful knowledge from such a complex system, we first need to find out a set of subsystems, i.e., subsets of road links in this case, each of which consists of objects, i.e., road links in our case, that can be modeled by the same monolithic model or the same type of models.

One possible solution to this problem may be to exploit an exploratory visual analytics approach. It is well known that exploratory visual analytics [8, 22–24] may use the "coordinated multiple views" visualization framework [16] as its basis. This framework provides more than one view for the visualization of the same database Δ from different aspects. Each view V_i may be a chart view, a map view, a graph representation view, or a calendar view, and shows the evaluation result $Q_i(\Delta)$ of some query Q_i associated with this view using its specific visualization scheme. Each view allows users to select a set of visualized objects by directly specifying each of them or enclosing some of them, which defines a new additional

Fig. 5 A coordinated-multiple-views visualization framework



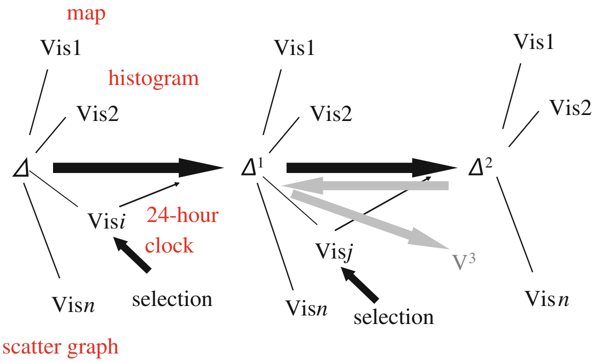


Fig. 6 A process of exploratory visual analytics

quantification condition C to quantify the objects stored in the underlying database Δ. This quantification defines a new database view Δ′ that is obtained by modifying the original view Δ with this additional quantification condition C. Each view V_j including V_i itself then immediately changes its visualization from $Q_j(\Delta)$ to $Q_j(\Delta')$, or just highlights the objects in $Q_j(\Delta')$ in the visualization of $Q_j(\Delta)$, depending on the user specification of its visualization mode. Multiple visualization views are mutually coordinated in this sense (Fig. 5).

In exploratory visual analytics with a coordinated-multiple-views visualization system, each user may start with the original database Δ, and repetitively try different selections of visual objects on different views for the exploration of different quantifications on database objects to find out a meaningful group of database objects. He or she may roll back the preceding visual object selection to try a different quantification through a different visualization view. Figure 6 schematically shows such a process of exploratory visual analytics.

Each view in coordinated-multiple-views visualization is, however, just a database visualization view. No analysis is actually applied in such exploratory visual analytics processes. In order to apply analysis tools to quantified sets of objects, we need to integrate these tools together with their analysis result visualizations into a coordinated multiple view visualization. Many researchers emphasized the importance of integrating various analyses and visualizations [9]. However, to the best of our knowledge, apart from some statistical chart tools to show histograms, correlations, or heatmaps [15], no other analysis tools such as clustering and frequent pattern mining tools have ever been integrated into coordinated multiple views visualizations to allow users to directly select a cluster or some of the mined frequent patterns for further quantifying database objects and for further analyzing those quantified database objects.

Exploratory visual analytics requires an extended coordinated multiple views visualization framework to which we can integrate any analysis tools so that their result visualizations can also be coordinated with other visualization views and analysis result views. It also requires an open library of analysis tools that can be integrated into the framework. These requirements made us to choose our Meme Media technologies [19] as enabling technologies to develop such a framework.

Figure 7 shows a coordinated multiple views visualization system using our Webtop Meme Media technology Webble World [10].

This system, called the Geospatial Digital Dashboard, was developed for exploratory visual analytics of social cyber-physical data related to the winter road management in Sapporo City. It can be set up with many different views. The setup in Fig. 4.2 has several views. The map view shows average taxi speed in each direction of each road link, and the distribution of tweets with geotags. It

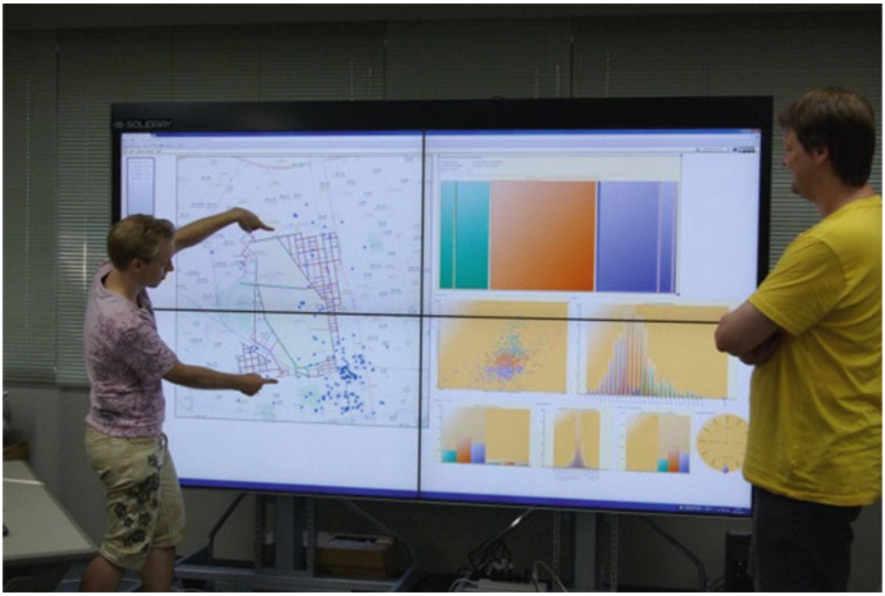


Fig. 7 The Geospatial Digital Dashboard as a coordinated multiple views visualization system

allows us to specify a rectangular area to select only the road links or tweets in this area. The clock view at the bottom right corner shows the population of taxis in each of the 24 h as the area of each circle around the circumference of the clock circle. It allows us to select time intervals for selecting statistical probe car data from only the specified time intervals. Other views include various kinds of charts such as correlation charts and histograms. You may also bring a calendar view with the weather data of each day into this environment. On such a calendar, you may specify only the days with heavy snowfall to focus on only data from heavy snowfall days.

In a coordinated multiple views framework, you may easily connect some analysis tool with its result visualization to any of these multiple views to apply this analysis to the set of objects visualized by this view. However, the visualized analysis result does not normally allow us to select one of the clusters or mined frequent patterns to further quantify those objects in the selected cluster or with the selected pattern. We need to integrate varieties of analysis tools and their result visualizations into our coordinated-multiple-views framework so that these analysis result views may also allow us to directly select some visualized objects, clusters, or mined patterns to further quantify the current database view. We call such an integrated framework a coordinated-multiple-analyses framework.

Let us first consider the integration of statistical analysis tools and their result visualizations into the coordinated-multiple-views framework. Any statistical analysis specifies the group-by attributes and, for each group of records, some aggregate function to calculate the aggregate value. The result can be represented as a relation Stat(GBattributes, Afunction), where the attribute GBattributes is a list of attributes specified as group-by attributes, and the attribute Afunction is a derived attribute whose value is obtained by applying the specified aggregate function such as average, count, minimum, maximum, and correlation to the set of values of the specified attribute in each group. This specified attribute is called the measure attribute. This relation can be visualized in various visualization schemes. Each visualization needs to provide a direct manipulation operation to quantify the GBattributes value and the Afunction value. This quantification further quantifies the values of the database attributes in GBattributes, which modifies the underlying database view from Δ to Δ'. Our taxi probe car data are stored in a relation Taxi(Date, Time, RoadLink, Speed, NumberOfCars, MaxSpeed). For the group-by attributes GBattribute=(Date, RoadLink) and the derived attribute Afunction=average(Speed × NumberOfCars), the attribute AFunction takes the value of the average taxi traffic flow in each day. If we quantify the average taxi traffic flow to be higher than a specified threshold, we will obtain, for each day, those road links satisfying this quantification. This quantification modifies the database view, and changes the other view visualizations and analysis visualizations in the coordinated-multiple-analyses system.

Let us now consider the integration of clustering tools and their result visualizations into the coordinated-multiple-views framework. The result of any clustering applied to objects identified by the values of some attribute A of the underlying database can be considered as a relation Cluster(A, ClusterID), where the values of A work as the object IDs of objects that are clustered, and ClusterID denotes

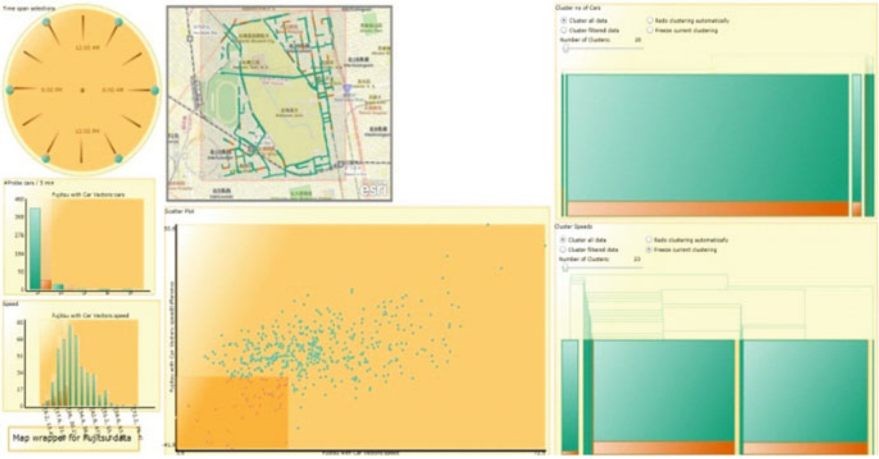
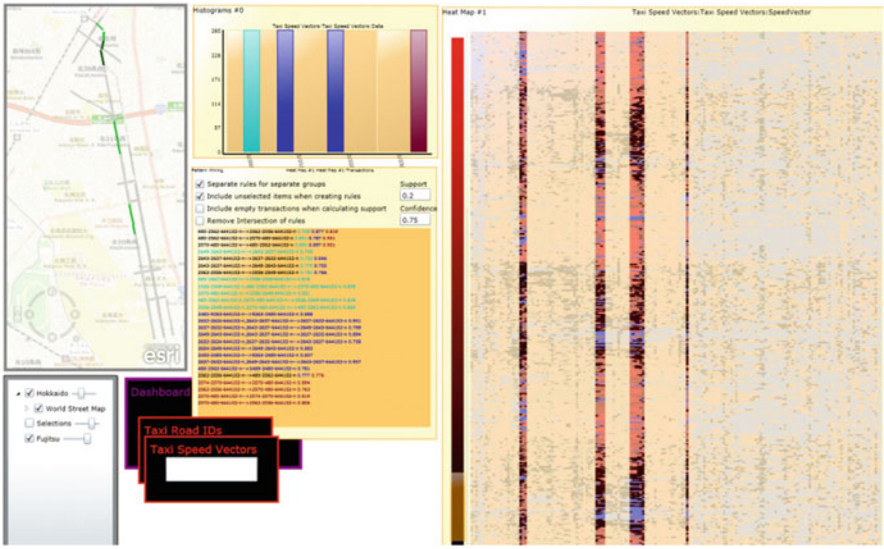


Fig. 8 Extended Geospatial Digital Dashboard with the integration of a clustering tool

the ID of each cluster. This relation Cluster(A, ClusterID) can be visualized in one of various visualization schemes. Each visualization needs to provide a direct manipulation operation for users to select some objects or some clusters. Such a direct selection corresponds to a quantification condition on the attribute A or ClusterID, which further quantifies the underlying database objects. Such a clustering tool with its result visualization can be easily implemented as a Webble, i.e., a visual component of the Webble World system. Figure 8 shows the extended Geospatial Digital Dashboard with the integration of a clustering tool. It has two clustering visualization views in its rightmost area. Each rectangle in each of them represents a cluster. Its size is proportional to the cluster size, i.e., the number of different A attribute values in the cluster. Each of these clustering result views also shows the phylogenetic tree of clusters over these clusters. In this example, road links are clustered in terms of the daily change of the number of taxis and the daily change of the average taxi speed in each road link. These values are available for each 5 min interval. Therefore, the changes of these values characterize each road link as two different vectors of 288 dimensions. For each of these two vector representations of each road link, we clustered the road links based on the similarity of their vector representations. In addition to the initial selection of road links on the map, the lower middle correlation chart showing the correlation between average speed of each road link in summer time and that in winter time is used to further select all the road links in the specified rectangular region. The selected portion of objects is highlighted in each cluster of the two clustering result. You may also choose one of these clusters to highlight only those road links in this cluster on the other visualization views.

Let us now consider the integration of frequent pattern mining tools and their result visualizations into the coordinated-multiple-views framework. Any frequent pattern mining result can be represented by two relations, Mining(Pattern, Supp


Histograms #0
Taxi Speed Vectors:Taxi Speed Vectors:Date
Heat Map #1
Taxi Speed Vectors:Taxi Speed Vectors:SpeedVector
esri
Dashboard
Taxi Road IDs
Taxi Speed Vectors

☑ Separate rules for separate groups
☑ Include unselected items when creating rules
□ Include empty transactions when calculating support
□ Remove Intersection of rules
Support 0.2
Confidence 0.75

☑ Hokkaido
☑ World Street Map
□ Selections
☑ Fujitsu

Fig. 9 Extended Geospatial Digital Dashboard with the integration of a frequent pattern mining tool

(, Conf)) and Include(A, Pattern). The first relation lists up each frequent pattern with its support index, and if necessary also with confidence index. The second relation tells which objects among those identified by the attribute value of A include each of the mined frequent patterns. The first relation can be visualized in various visualization schemes to list up mined frequent patterns together with their support and confidence index values. Each visualization needs to provide a direct manipulation operation for users to change the threshold values of support and confidence indices to list up only those patterns with their support and confidence indices higher than the thresholds, and further to directly select some frequent patterns in the list. Using the second relation, this selection of some patterns is converted to the corresponding quantification condition on the attribute A, which further quantifies the underlying database and changes the other coordinated visualization views. Figure 9 shows an extension of Geospatial Digital Dashboard with the integration of a frequent pattern mining tool as well as a clustering tool.

This figure shows a heatmap on the right side, and the result of association rule mining around the center. Each row of the heatmap represents a 5 min interval in a day. All the rows for 4 days are shown here. Each column represents a road link. The highlighted all the columns are those road links selected on the map view. Here we want to apply a micro analysis to the set of road links colored with light gray on the map. These are the road links along the busiest north-south route and its neighboring ones in the north ward in Sapporo. The intensity of each cell in this heatmap represents the average speed in each road link during each 5 min interval.

In general, such a heatmap can be defined for two nominal value attributes A1 and A2, and one numerical value attribute A3 that are arbitrarily selected out of the underlying database attributes. Its row and column represent the domains of A1 and A2 respectively, and each cell shows the numerical value of A3 as the color intensity. In this example, the attribute A1 is the Time attribute representing each 5 min interval, the attribute A2 is the RoadLinkID represented by a geo-location pair of two end points of each road link, and the attribute A3 is the AverageSpeed of taxis in this road link at this time interval.

On the left side of the heatmap, we have a color gradient bar showing how different intensities are mapped to different colors. This bar allows us to specify more than one intensity interval. If you specify k intervals, the heatmap webble generates k items for each of its columns, and will interpret each row as a transaction, and each column as *k* items. Each item in each transaction has a binary value.

We can apply item set mining and association rule mining to this set of transactions. In Fig. 9, we specified only one intensity interval to focus on the average speed lower than 10 km/h. We wanted to focus on traffic jams in road links. Figure 9 shows the highlighting of those road links with traffic jams in the heatmap, and only the result of applying the association rule mining to the set of transactions, i.e., all the 5 min intervals. This mining tool has two numerical entries for users to specify the threshold values of the support and the confidence. This association rule mining result view allows us to specify some of the mined patterns. This selection quantifies those transactions having one of these selected patterns in the heatmap. Unselected transactions, or rows, will be dimmed in the heatmap. The mining result view Webble also allows us to select each item appearing in each selected pattern. Since each item in this example is a road link, you may pick up a mined association rule to see how the traffic jam in some road links may propagate to other road links on the map view.

Figure 10 shows three analysis results obtained by the same environment as shown in Fig. 9. Here we used an item-set-mining analysis view instead of an association-rule-mining analysis view. The backmost one shows the analysis result for a day before a snowy day. The middle one shows the result for 2 days with very heavy snowfall. The front one shows the analysis result for the day after the snowfall and the immediate snow removal. These show that the traffic jam sections become seriously extended by the snowfall, and that the situation significantly improves after the snow removal.

These traffic jam expansions seem to be caused by areas with traffic jams already before the snowfall, when these areas that already have throughput problems become even worse from snow narrowing the street or making the roads slippery. These traffic jam areas require further analysis to examine if the problems are specific to winter or if there is a more general problem throughout the year. If some of them are specific to winter, they are usually caused by roadside snow piles locally stretching out into the street. Snow removal usually starts with snow plowing to pile snow on the sides of the roads. This is followed by snow removing, when snow is

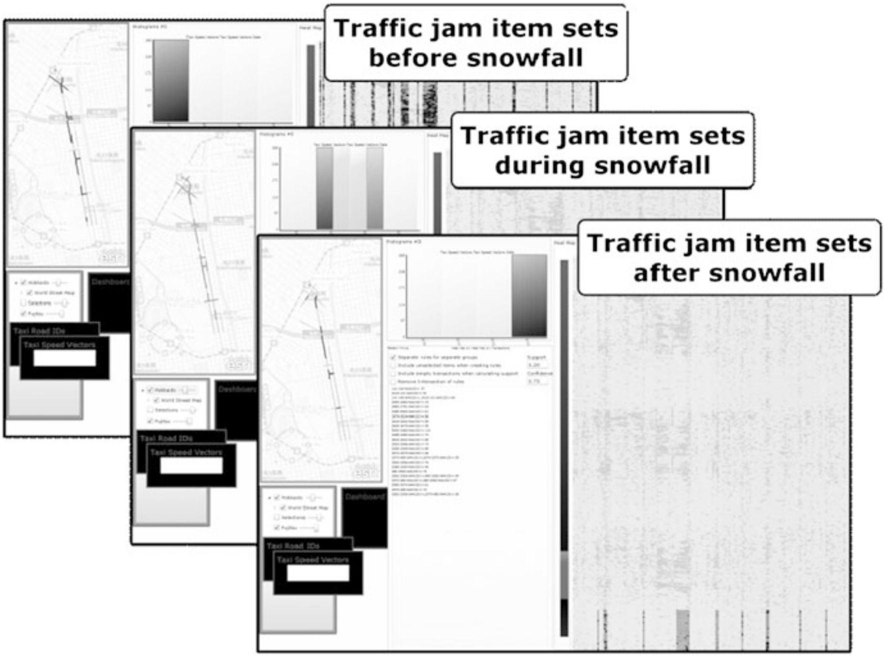

Traffic jam item sets before snowfall
Traffic jam item sets during snowfall
Traffic jam item sets after snowfall
Dashboard
Taxi Road IDs
Taxi Speed Vectors

Fig. 10 Item set mining results on traffic jams on a day before a big snowfall, the 2 days of snowfall, and the day after the snow and the snow removal

cut out of the roadside piles and removed (e.g. by truck) to recover road width. To keep the sidewalks and/or parking spaces clean, people further pile snow on top of the remaining snow piles, which often makes some part fall down into the road. Such snow piles locally stretching out into the streets are the main cause of winter specific local traffic jams. From experience we know that the locations with such local stretch out do not change. Winter-specific traffic jams exacerbated by snowfall are likely to have such local stretch out. Figure 11 shows 3D measurements of road side snow piles and road surface both around some road links with traffic jams that grew after snowfall, and around some road links along the same route but that had no traffic jam. The first one shows lots of local stretch-outs from snow piles, while the second one shows no such stretch out of them. We used a laser range scanner installed on a car to gather the data.

The 3D measurement data was used only as ground truth data to check the validity of our analysis and inference results. This kind of analysis of statistically preprocessed probe car data may give us information about locations of snow stretch outs that cause traffic jams, which may help snow removal operation centers to make decisions on pinpoint snow removal just before the expansion of traffic jams.

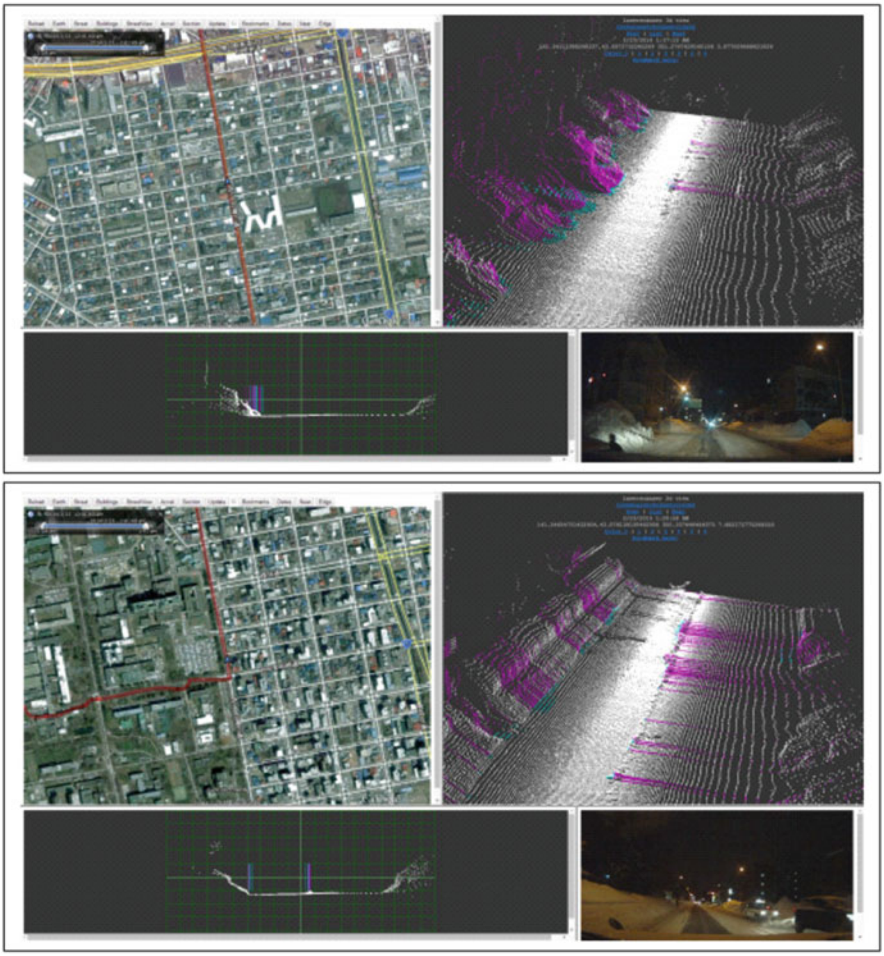


Fig. 11 3D measurement of road surface and road side snow piles both around a traffic jam area that grew during the snowfall, and around a no-traffic-jams area

5 Conclusions

This paper showed that advances of sensor devices and their networking technologies have enabled the real-time monitoring of physical systems ranging from small-size healthcare devices to large-scale plant systems, or even to urban-scale infrastructure service systems. It focused on social CPSs (Cyber-Physical Systems) that denote the extended application of the idea of CPSs to the monitoring and control of urban-scale social infrastructure services such as traffic, energy, and water services, especially on the winter road management in Sapporo, where we have the world biggest annual snowfall among the cities with more than 1 million populations. Since the targets of Social CPSs are necessarily complex systems of systems, they need to deal with a varier of heterogeneous cyber and physical, real-time and retrospective big data including, for example, probe-car data, weather data, snow removal records, and traffic accident records. For the monitoring and analysis of the traffic and road conditions of the whole city, the use of probe-car data may be the only practical solution today. Their advanced usage without violating personal data protection is fundamental.

This paper has shown that statistically preprocessed probe-car data of road links for an urban-scale area can give us lots of information about traffic flow, such as the temporal change of the traffic flow divergence and the traffic flow vector field. These will tell us the hotspots for taking taxis and getting off taxis, main traffic streams at every time, and dynamic change of route selection preference. Such data also allow us to do more advanced complex analyses. We have picked up the optimization of snow plowing and removing in Sapporo as our target of such analyses.

Since our preceding study showed that the influence of snow and/or snow removal on the traffic and road conditions is not homogeneous across even road links along the same main route in the central city area, we proposed the Geospatial Digital Dashboard system for exploratory visual analytics. This system uses the well-known multiple coordinated views framework for exploratory visual analytics, and extends it to integrate analysis tools with their result visualization views in the same environment so that these may also be coordinated with other visualization views. Users can quantify database views through direct selection of visual objects not only in coordinated views but also using analysis result views. Each quantification is immediately reflected not only in other views but also in all other linked analysis result views. Using the extended Geospatial Digital Dashboard, we have analyzed statistically preprocessed probe car data to detect road links with serious snow stretch-outs narrowing the effective road width. This may allow us to advise the snow removal center to conduct pinpoint snow removal to remove such snow stretch-outs, which may effectively prevent serious expansion of traffic jams, and reduce the total snow removal cost.

From the system architecture point of view, exploratory visual analysis in general requires the following features.

1. It should support the repetition of the hypothesis making through data segmentation of a specific set of data and the hypothesis checking through data analysis and visualization of the segmented data set. The analysis result may be also used for further data segmentation.
2. It should provide a large library of analysis and visualization tools open for the future extension. It should be easy to improvisationally wrap external tools and services into components and to register them into the library for their future reuse in the visual analytics environment through the improvisational federation of them with other tools and services.
3. It should allow users to improvisationally bring external data sources provided as web services into the visual analytics environment.

The first requirement made us propose a coordinated-multiple-analyses visualization framework as an extension of the coordinated-multiple-views visualization environment. We used the webtop meme media technology as the enabling technology for these frameworks.

Exploratory visual analytics requires various analysis tools and data sources. Its system should be open for the future integration of new analysis tools and new data sources to itself. Our framework is based on the webtop meme media system Webble World, which enables us to improvisationally wrap both varieties of tools developed in R, Octave, Python, and Ruby, and any analysis and/or data providing web services into webbles. These webbles can be registered into the open library to increase its variety. Users can improvisationally federate any of these wrapped tools and services to work together. For example, in the Geospatial Digital Dashboard System, the map view can also be used to visualize the geographical distribution of tweets. These tweets may be obtained from the twitter service. We can improvisationally wrap the twitter service into a webble, and improvisationally federate this webble with the map view webble to obtain such a visualization result.

These days we have a huge variety of related open data sources over the Web. They can be accessed through web services. It is important for us to be able to improvisationally federate these data sources with our visual analytics environment. We can also find out a huge variety of open tools and services for data analysis and visualization. The improvisational knowledge federation capability of Webble World will allow us to improvisationally wrap a large portion of them into webbles, and to reuse them in cooperation with other data sources, tools and services in our exploratory visual analytics environment.

References

1. Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases. In: Proceedings of the 2008 IEEE 24th international conference on data engineering, ICDE '08; 2008. pp. 376–85.
2. Ahlberg C. Spotfire: an information exploration environment. SIGMOD Rec. 1996;25(4):25–9.
3. Aiken A, Chen J, Stonebraker M, Woodruff A. Tioga-2: a direct manipulation database visualization environment. In: Proceedings of the 12th international conference on data engineering, New Orleans, February 26–March 1, 1996. pp. 208–17.
4. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B. Knime - the konstanz information miner: Version 2.0 and beyond. SIGKDD Explor Newsl. 2009;11(1):26–31.
5. Demšar J, Curk T, Erjavec A, Gorup V, Hočevar T, Milutinovič M, Možina M, Polajnar M, Toplak M, Starič A, Štajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B. Orange: data mining toolbox in python. J Mach Learn Res. 2013;14(1):2349–53.
6. Hoh B, Gruteser M, Xiong H, Alrabady A. Preserving privacy in gps traces via uncertainty-aware path cloaking. In: Proceedings of the 14th ACM conference on computer and communications security, CCS '07; 2007. pp. 161–71.
7. ISO. Intelligent transport systems—basic principles for personal data protection in probe vehicle information services. ISO 24100:2010, International Organization for Standardization, Geneva; 2010.

8. Keim D, Mansmann F, Stoffel A, Ziegler H. Visual analytics. In: Liu L, Özsu MA, editors. Encyclopedia of database systems. New York: Springer; 2009. pp. 3341–6.
9. Keim DA, Mansmann F, Thomas J. Visual analytics: how much visualization and how much analytics? SIGKDD Explor Newsl. 2010;11(2):5–8.
10. Kuwahara M, Tanaka Y. Webble world — a web-based knowledge federation framework for programmable and customizable meme media objects. In: IET international conference on frontier computing. theory, technologies and applications; 2010. pp. 372–7.
11. Livny M, Ramakrishnan R, Beyer K, Chen G, Donjerkovic D, Lawande S, Myllymaki J, Wenger K. Devise: integrated querying and visual exploration of large datasets. In: Proceedings of ACM international conference on management of data, SIGMOD '97; 1997. pp. 301–12.
12. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. Yale: rapid prototyping for complex data mining tasks. In: Proceeding of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06. New York: ACM; 2006. pp. 935–40.
13. Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: A generalization-based approach. Trans Data Privacy. 2009;2(1):47–75.
14. North C, Shneiderman B. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In: Proceedings of the working conference on advanced visual interfaces, AVI '00; 2000. pp. 128–35.
15. Perer A, Shneiderman B. Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. IEEE Comput Graph Appl. 2009;29(3):39–51.
16. Roberts JC. State of the art: coordinated & multiple views in exploratory visualization. In: Proceedings of the 5th international conference on coordinated and multiple views in exploratory visualization, CMV '07. Washington: IEEE Computer Society; 2007. pp. 61–71.
17. Sato M, Izumi M, Sunahara H, Uehara K, Murai J. Threat analysis and protection methods of personal information in vehicle probing system. In: Proceedings of the 3rd international conference on wireless and mobile communications; 2007. p. 58.
18. Sugibuchi T, Tanaka Y. Integrated visualization framework for relational databases and web resources. In: Intuitive human interfaces for organizing and accessing intellectual assets. Lecture notes in computer science, vol. 3359. Berlin Heidelberg: Springer; 2005. pp. 159–74.
19. Tanaka Y. Meme media and meme market architectures: knowledge media for editing, distributing, and managing intellectual resources. New York: Wiley; 2003.
20. Tanaka Y, Sjöbergh J, Moiseets P, Kuwahara M, Imura H, Yoshida T. Geospatial visual analytics of traffic and weather data for better winter road management. In: Cervone G, Lin J, Waters N, editors. Data mining for geoinformatics. New York: Springer; 2014. pp. 105–26.
21. Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories. In: Proceedings of the 9th international conference on mobile data management, MDM '08; 2008. pp. 65–72.
22. Thomas J, Kielman J. Challienges for visual analytics. Inf Visual 2009;8(4):309–14. doi:10.1057/ivs.2009.26. http://dx.doi.org/10.1057/ivs.2009.26.
23. Thomas JJ, Cook KA. Illuminating the path: the research and development agenda for visual analytics. National Visualization and Analytics Ctr (2005). http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0769523234.
24. Tukey JW. Exploratory data analysis. Reading: Addison-Wesley; 1977.