

# Spatio-temporal Aggregation for Visual Analysis of Movements

Gennady Andrienko, Natalia Andrienko

Fraunhofer Institute IAIS (Intelligent Analysis and Information Systems), Sankt Augustin, Germany

## ABSTRACT

Data about movements of various objects are collected in growing amounts by means of current tracking technologies. Traditional approaches to visualization and interactive exploration of movement data cannot cope with data of such sizes. In this research paper we investigate the ways of using aggregation for visual analysis of movement data. We define aggregation methods suitable for movement data and find visualization and interaction techniques to represent results of aggregations and enable comprehensive exploration of the data. We consider two possible views of movement, traffic-oriented and trajectory-oriented. Each view requires different methods of analysis and of data aggregation. We illustrate our argument with example data resulting from tracking multiple cars in Milan and example analysis tasks from the domain of city traffic management.

CR Categories and Subject Descriptors: H.1.2 [User/Machine Systems]: Human information processing – Visual Analytics; I.6.9 [Visualization]: information visualization.

Additional Keywords: Movement data, spatio-temporal data, aggregation, scalable visualization, geovisualization.

## 1 INTRODUCTION

One of the strengths of information visualization as an amplifier of human cognition and ideation lies in supporting abstraction and generalization [14]. Thus, appropriate positioning and/or appearance of graphical elements representing data items can stimulate holistic perception of multiple data items as a unit. However, when the size and complexity of data increases, purely visual approaches become insufficient and need to be combined with computational generalization, which includes, among other techniques (e.g. smoothing, filtering), data aggregation. Aggregation is not only a tool to reduce the size of data but also a way to distill general features out of fine-detail “noise”.

This paper considers the use of aggregation for visual analysis of movement data, more specifically, data about multiple discrete entities changing their spatial positions over time while preserving their integrity and identity (i.e. the entities do not split or merge). In our earlier papers we considered the structure and essential properties of movement data and defined the possible general analysis tasks [1] as well as the types of tools that could support these tasks [3]. Among others, we discussed the use of data aggregation and the possible ways of aggregating movement data. In [2] we described a set of complementary tools for analysis of movement data including database transformations, visualization, interactive dynamic filtering, and clustering. We mentioned one particular aggregation method, which was used for visualization of clustering results. Unlike the previous publications, the current paper primarily focuses on various possible ways of aggregating

movement data. The work has been done within an ongoing EU-funded project GeoPKDD (<http://www.geopkdd.eu>).

In [1] and [3] we introduced a formal model of collective movement of multiple entities as a function  $\mu: E \times T \rightarrow S$  where  $E$  is the set of moving entities,  $T$  (time) is the continuous set of time moments and  $S$  (space) is the set of all possible positions. As a function of two independent variables,  $\mu$  can be viewed in two complementary ways:

- as a set of trajectories of all entities:  $\{\mu_e: T \rightarrow S \mid e \in E\}$ , where the function  $\mu_e: T \rightarrow S$ , called *trajectory*, describes the movement of a single entity;
- as a temporal sequence of traffic situations:  $\{\mu_t: E \rightarrow S \mid t \in T\}$ , where the function  $\mu_t: E \rightarrow S$ , called *traffic situation*, describes the spatial positions of all entities at a time moment  $t$ .

The first way will be further called *trajectory-oriented view* and the second one will be called *traffic-oriented view* (we use the term “traffic” in an abstract sense to denote collective movement of any kind of entities). The view to take depends on the analysis goals, as will be further demonstrated by examples. Each view requires different analysis methods and, in particular, different ways of aggregating movement data. In this paper, we investigate what aggregation methods can be used for each of the views. For the presentation purposes, we use an example dataset and example analysis tasks from the domain of city traffic management. However, our work is not specifically oriented to this domain and these tasks; this is a more general research work on the use of aggregation in analyzing massive movement data.

Before presenting the example dataset and discussing the possible methods of aggregation, we shall briefly overview the relevant works concerning aggregation of movement data.

## 2 RELATED WORK

Most software tools designed to support visual examination of large sets of movement data involve data aggregation. There are three basic types of aggregation, spatial (S), temporal (T), and attributive (A), also called categorical [7]. These basic types are used in various combinations.

Several aggregation techniques are described in a series of papers written by D. Mountain and his colleagues (e.g. [6][10]). T-aggregation appears in the form of temporal histogram where the bars correspond to time intervals and their heights are proportional e.g. to the number of locations visited or the distance traveled. S-aggregation is done by imposing a regular grid over the territory and counting trajectory points fitting in each cell. The resulting density counts are represented by coloring or shading of the grid cells on a map display. In S×T-aggregation the densities are computed for consecutive time intervals and shown on an animated map display. Similar to densities, other aggregated characteristics can be computed and visualized. Thus, in [8] the total number of person/minutes spent in each cell is computed. A sophisticated S×T×A-aggregation is suggested in [17]: position records are grouped spatially by cells of a regular grid and then temporal (e.g. by days of the week) and attributive (e.g. by vehicle types) aggregation is applied to each group. The results are represented by multiple treemaps [13] placed inside each cell.

Essentially, all these aggregations do not differ from what was suggested in [7] for aggregating spatially distributed discrete

<http://geoanalytics.net/and>

events: each record from the movement data is, in fact, treated as an independent event. Hence, these ways of aggregation do not capture the specific nature of movement data. The results of the aggregation show the presence of entities in different places at different times but not the movement of the entities from place to place.  $S \times T$ - and  $S \times T \times A$ -aggregations can be helpful where the traffic-oriented view of movement data is required but do not support the trajectory-oriented view.

A different way of aggregating movement data is counting for each pair of places in space how many entities moved from the first to the second place between two time moments. This kind of aggregation may be represented by the formula  $S \times S \times T \times T$  (start place, end place, start time, and end time). The resulting counts may be visualized as a transition matrix where the rows and columns correspond to the places and symbols in the cells or cell coloring or shading encode the counts [9]. For more than one pair of time moments, one would need to build several transition matrices, which could then be compared. However, the limitations of this approach with respect to the length of the time series of movement data are evident. Another problem is that such visualization lacks the spatial context. Tobler [15][16] visualizes aggregated moves on a map by bands or arrows connecting pairs of locations with the widths proportional to the volumes moved between these locations. Unfortunately, such a map may be illegible because of intersecting and overlapping symbols. Therefore, Tobler suggest a specific method for spatial smoothing of aggregated moves and generation of continuous flow maps. Intersections and overlaps between movement symbols may be reduced by involving the third spatial dimension, as in the visualization of the movement of tourists in New Zealand [5] (discussed in [3]). Irrespective of the visualization,  $S \times S \times T \times T$ -aggregation does not fully support the trajectory-oriented view since it hides essential information about the routes of the entities.

In all aggregations discussed so far the results are numeric values such as counts, sums, statistical means, etc. In [4] a kind of geometric summary of several trajectories is derived. The authors use functions of ArcGIS to build a convex hull containing the trajectories, compute the central tendency and dispersion of the paths, and represent the results on a map as the averaged path. Such geometric summarization works well only when the trajectories are similar in shape and close in space. It can be applied, for example, to groups of similar trajectories resulting from clustering. Grouping of trajectories by similarity and/or closeness of the routes followed by geometric and/or numeric summarization may be called R- (route-based) aggregation.

Our earlier paper [2] contains examples of combining route-based grouping of trajectories with  $S \times S \times T \times T$ -aggregation; all together may be called  $R \times S \times S \times T \times T$ -aggregation. It can support the trajectory-oriented view of movement, as will be shown later.

### 3 EXAMPLE DATA AND ANALYSIS TASKS

To present our work, we shall use an example dataset collected by GPS-tracking of 17,241 cars in Milan (Italy) during one week from Sunday to Saturday. Figure 1 shows the variation of the numbers of simultaneously moving cars from the tracked sample over the period of the observation. The numbers have been counted by hourly intervals and range between 80 and 3173. The vertical lines on the graph correspond to 0 o'clock.

The dataset consists of more than 2 million records each including car identifier, time stamp (date and time of the day), geographical coordinates, and speed. The time intervals between the records of the same car are irregular, mostly ranging from 30 to 45 seconds while there are also larger intervals ranging from several minutes to several days. The data have been kindly provided by Comune di Milano (Municipality of Milan) for the use within the project GeoPKDD.

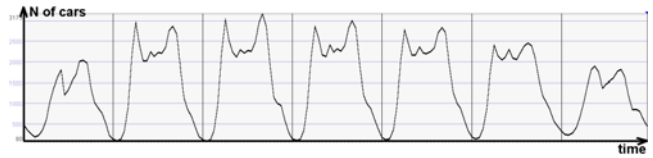


Figure 1. Variation of the number of simultaneously moving cars.

In [2] we have described how we preprocess raw movement data in the database and integrate individual position records into trajectories. There is no unique way of combining position records into trajectories. In [2] we discuss several possible methods. In this paper we shall use trajectories obtained by one of the methods; the details are irrelevant to the topic of the paper. The number of the trajectories is about 176,000.

It should be noted that the whole dataset is too big for loading and processing in the computer's main memory and for interactive exploration with the use of dynamic querying, brushing, and other techniques addressing individual objects, i.e. points or trajectories. Therefore, it is necessary either to aggregate the data inside the database and explore the resulting aggregates or to divide the data into manageable subsets and explore them separately. The results then need to be compared and somehow integrated.

Example analysis tasks related to city traffic management come from the interviews with specialists from mobility agencies and traffic departments of several Italian cities. The interviews have been conducted by our GeoPKDD partners from the Italian telecommunication company WIND and its business school.

According to the interviews, city traffic managers need to cope with the following tasks: (1) estimate the average flows (number of people) between regions of interest and their variation in different time periods and in presence of extraordinary events such as football games, concerts, strikes, etc.; (2) estimate the average travel times between regions and their variation; (3) estimate the "impedance" of a street (obstruction to movement) and its variation; (4) estimate the proportions of the cars leaving a main road on different exits; (5) understand the actual paths used by people to get from one point or region of interest to another.

At the present time, traffic managers do not use data resulting from tracking the movement of vehicles or people. Although such data become widely available, there are no appropriate tools for their analysis. A common practice is to use results of public surveys and traffic monitoring data coming from stationary video cameras or other sensors. Such data are not well suited to the tasks. While methods for reconstructing traffic flows from stationary sensor data are devised in data mining [11], analysis of tracking data could significantly help in coping with the tasks as well as in verifying traffic models built on the basis of data from stationary sensors.

Assuming that the tasks of traffic managers are to be carried out with the use of car tracking data like in our example dataset, we can say that tasks 1, 2, and 5 require the trajectory-oriented view and task 3 requires the traffic-oriented view of the car movement. We shall discuss later which view is more appropriate for task 4.

In the following sections we investigate what aggregation methods can support the two different views of movement data and what visualization techniques are suitable for viewing and exploring the outcomes of the aggregation. We would like to stress that the data and tasks described in this section serve only as examples for illustrating the suggested general framework for analysis of massive movement data with the use of aggregation.

### 4 SUPPORTING THE TRAFFIC-ORIENTED VIEW

We use the term "traffic situation" to denote the spatial positions of all moving entities and the values of the movement-related attributes including speed, direction, acceleration (change of

speed) and turn (change of direction) at some time moment. In the traffic-oriented view, an analyst looks at traffic situations at different time moments and considers the evolution of the traffic situation over time. For practical reasons, the analyst cannot analyze the traffic situation of each second. On the one hand, this would require too much time and effort; on the other hand, the available data may not allow this because of larger time intervals between the measurements. A reasonable approach is to aggregate the data by time intervals of appropriate lengths. Thus, in analyzing city traffic it may be sufficient to use time intervals of the length of one hour or, if this is too coarse, half an hour or quarter of an hour.

S×T-aggregation can adequately support the consideration of aggregate traffic situations on time intervals. Besides dividing the time into intervals, the space (i.e. the territory where the entities move) is divided into appropriate compartments. In our experimental implementation, compartments are defined by building a regular rectangular grid of a desired resolution, but it is possible, in principle, to use other divisions. Then, various aggregates are computed for each pair of space compartment and time interval from the track records fitting in this compartment and this interval: number of different entities, number of visits, total time spent, statistics of the movement-related attributes (minimum, maximum, average, median, etc.). The aggregation can be done in the database. The results are loaded in the main memory and can be visualized in various ways including static and animated maps and non-cartographic displays.

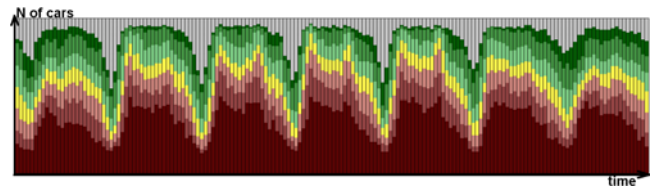


Figure 2. Temporal variation of the median speeds in different places of Milan (grid cells) computed by hourly time intervals.

For example, Figure 2 shows the variation of the frequency distribution of the median speeds throughout the territory of Milan (divided into compartments by a regular grid) over the whole period of the observation from Sunday to Saturday. The data have been aggregated by hourly intervals; the segmented bars represent these intervals. The colors of the bar segments correspond to intervals of the values of the aggregate attribute “median speed”. The breaks are 15, 30, 45, 60, 80, and 100 km/h. Yellow is assigned to the interval from 45 to 60, the shades of red represent median speeds below 45, and the shades of green are used for median speeds over 60 km/h (the color legend can be seen on the left of Figure 3). The heights of the bar segments are proportional to the numbers of the compartments where the median speeds fitted in the respective intervals. Gray segments show the numbers of the compartments with no occurrences of tracked cars during the corresponding time intervals.

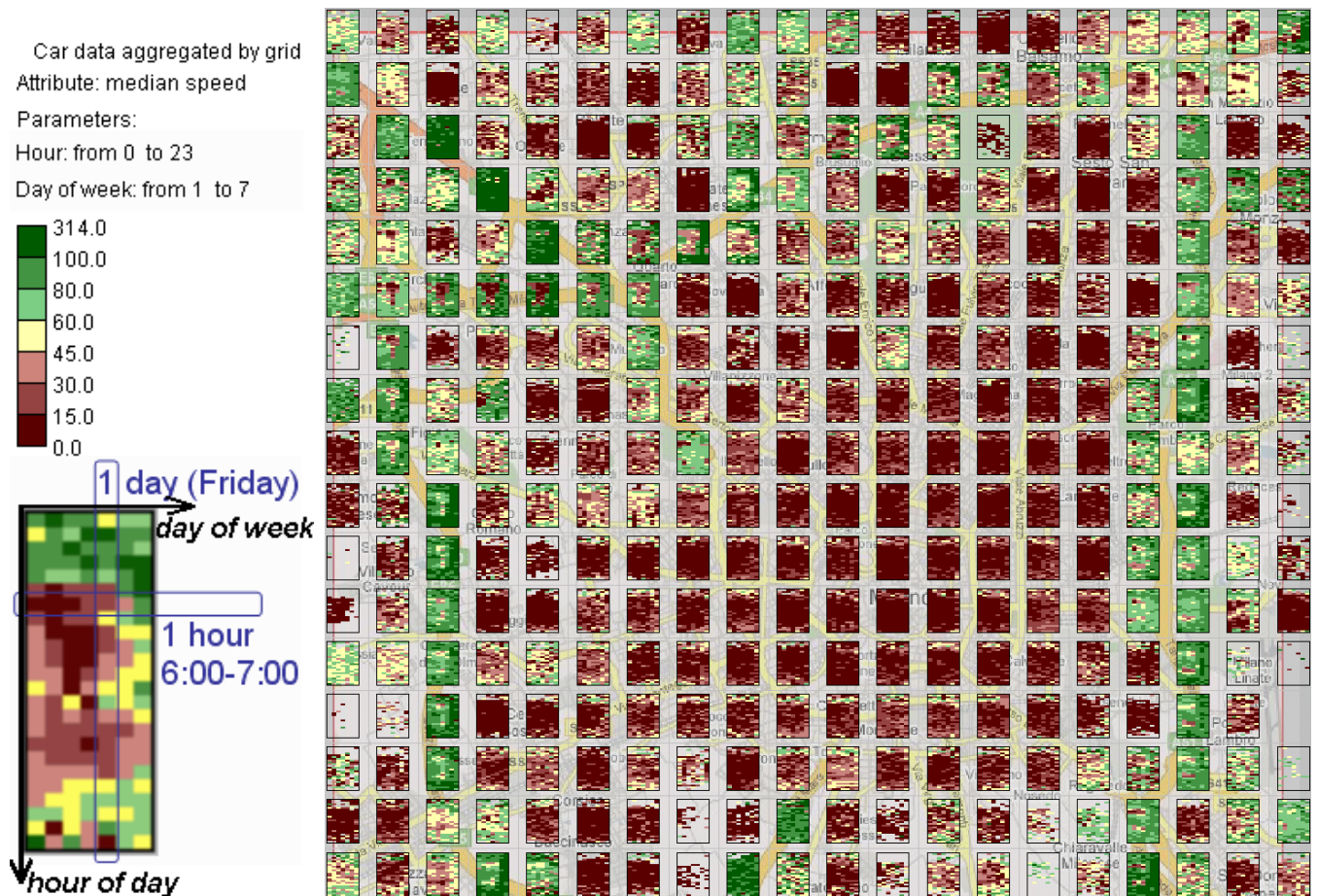


Figure 3. The mosaic diagrams show the variation of the median speeds in spatial compartments by days of the week (columns of the diagrams) and hours of the day (rows of the diagrams). The cells are colored according to the speeds. The breaks and colors for the speed intervals are the same as in Figure 2. Slow speeds are shown in shades of red and fast speeds in shades of green.



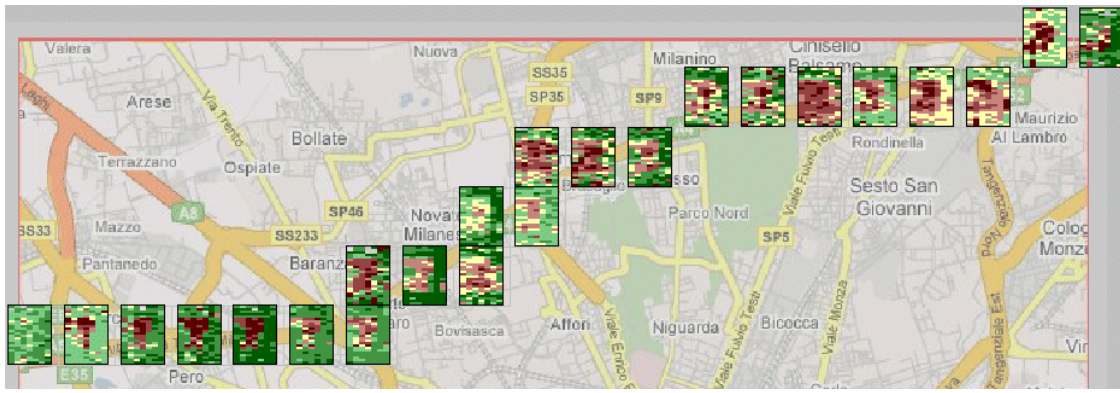


Figure 4. Focusing on selected spatial compartments along a particular road.

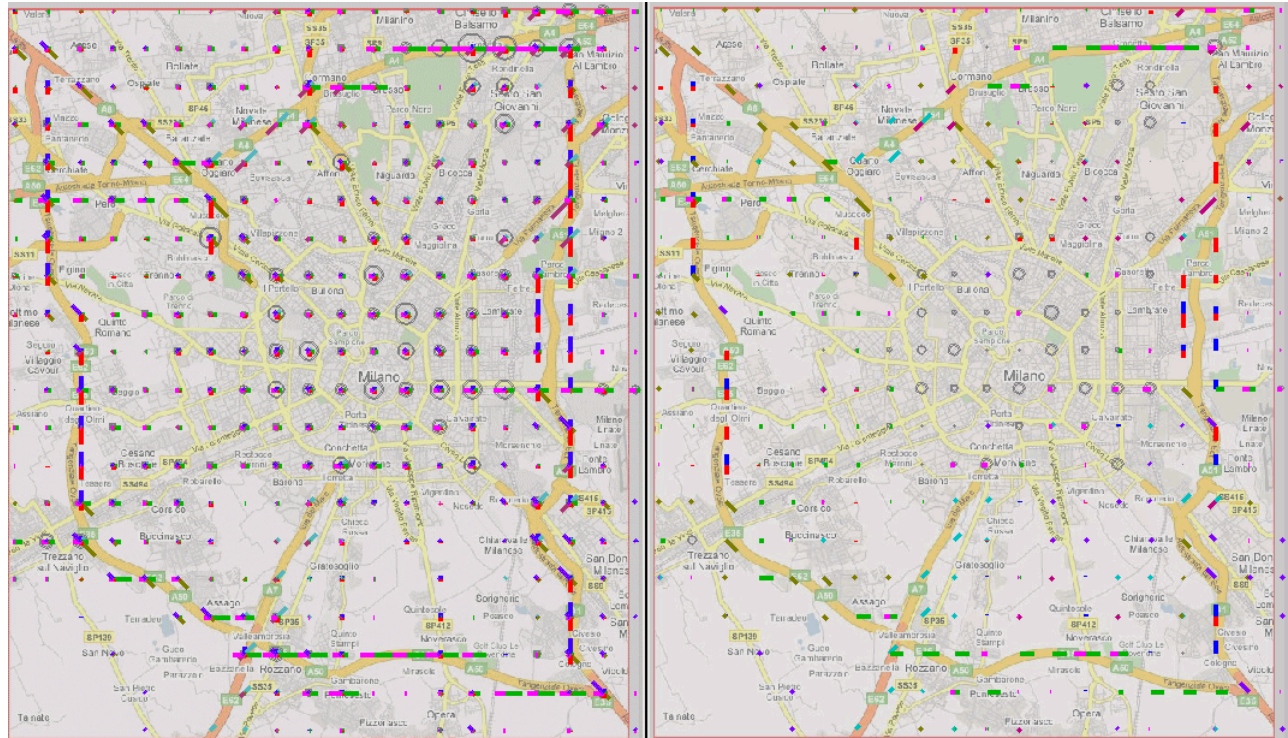


Figure 5. The directional bar diagrams show movement data aggregated by compass directions. The lengths of the bars are proportional to the numbers of the cars that moved in the respective directions during a selected time interval. The radii of the circles are proportional to the numbers of the cars with the speeds below a selected threshold (here 5km/h). On the right, only dominant directions are shown, specifically, where values are at least 25% higher than the next highest value (25% is a selected threshold).

Since time is not only a linearly ordered sequence of moments but also has a cyclical organization, it is possible to aggregate time-related data by dividing their time span according to one or more temporal cycles. Thus, Figure 3 represents aggregates obtained with the use of two temporal divisions: according to the days of the week and according to the hours of the day. The first division groups together data referring to the same day of the week irrespective of the date. The second division groups together data from different days referring to the same hour of the day. As a result, aggregated values have been computed for each combination of space compartment, day of the week, and hour of the day. Each “mosaic” diagram summarizes the daily and weekly patterns of the traffic in a particular place.

A traffic analyst can use this aggregation to explore the impedance of a street (task 3). For this purpose, the analyst can select the space compartments covering the street and look only at the data in these compartments (Figure 4). It should be noted that

regular rectangular compartments may not ideally suit the geometry of a particular street. In this case arbitrarily specified compartments are preferable.

The aggregation discussed so far is not specific to movement but can be applied to other kinds of spatio-temporal data, e.g. point events. In fact, this is the same type of aggregation as used for traffic incidents in [7]. To capture the specifics of movement data, we suggest another aggregation method where the data are aggregated not only by space and time but also by the direction (course) of movement. This aggregation can be denoted by the formula  $S \times T \times D$ , where  $D$  stands for “direction”. Movement directions are often indicated in the original track records. If this is not the case, they can be computed from pairs of consecutive positions of the same entity.

The directions are specified in movement data as numeric values typically representing angular degrees from 0 to 359. For  $S \times T \times D$ -aggregation we suggest to divide this range into intervals

corresponding either to four main compass directions (north, east, south, and west) or to four main and four intermediate directions. Track records fitting in the same spatial compartment and temporal partition are additionally grouped by the movement directions. A separate group is made from records where the speed is below a chosen threshold. This is treated as the absence of movement. Then, various counts and statistics of attribute values are computed for the groups.

To visualize the resulting aggregate data, we suggest a special technique in which the data are represented on a map by directional bar diagrams. Analogously to the *wind rose* used in meteorology, the bars are oriented in four or eight compass directions and their lengths are proportional to the values of the currently selected aggregate attribute corresponding to the respective directions. Thus, the diagrams in Figure 5 (left) portray the numbers of the cars that moved in different directions on Monday between 7 and 8 AM. The bars are colored depending on their orientation; a particular color is assigned to each direction. This helps in gaining an overall view of the prevailing movement directions throughout the whole territory. Besides the directional bars, some diagrams include gray circles representing the groups of records with the speeds below the chosen threshold. The radii of the circles are proportional to the values of the currently selected aggregate attribute computed for these groups of records. The radii can be easily compared with the lengths of the bars. In Figure 5 the circles represent the numbers of the distinct cars that had the speeds below the chosen threshold of 5km/h. Such speeds occur predominantly in the central part of the city but also on the northeast, where the circles located on a segment of a motorway may indicate its congestion.

Visual exploration of traffic with the use of this kind of display can be supported by a number of interactive facilities:

- switch from one aggregate attribute to another, e.g. from the number of entities to the average or median speed;
- select another temporal partition, i.e. another interval, day of the week, time of the day, etc., depending on how the data have been aggregated;
- hide some directions in order to focus on the remaining direction(s), e.g. to see where northward movement occurs;
- choose presenting only the dominant direction(s) in each spatial compartment. A direction is treated as dominant when the corresponding value of the current aggregate attribute exceeds the highest value among the remaining directions by a chosen threshold, which may be either absolute (i.e. minimum difference between the values) or relative (i.e. minimum ratio).

The screenshot on the right of Figure 5 shows the dominant movement directions defined by the relative threshold of 25%. It may be seen that movements towards the center prevail on most radial streets and that movements to the east (green bars) dominate on the motorway on the south. In some compartments there are two or more dominant directions. This means that the respective attribute values differ by less than 25%.

The S×T×D-aggregation together with the visualization can support a more refined exploration of street impedance than it is possible with the S×T-aggregation. An example is shown in Figure 6. To explore the traffic on a particular road, only the space compartments (grid cells) covering this road have been selected. The data have been aggregated according to the four main compass directions. The bar diagrams represent the median speeds in the eastern (green) and western (purple) directions. The diagrams are substantially asymmetric, meaning different speeds of the movement in the eastern and in the western directions. Lower speeds, in turn, signify higher obstruction to the movement. In this way, the impedance of a street to the movement in the different directions can be explored.

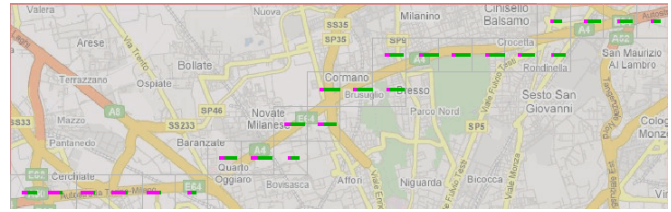


Figure 6. The bars represent the median speeds of the movement toward the east (green) and west (purple) between 11 and 12 AM on Wednesday along a motorway on the north of Milan.

It may seem that the S×T×D-aggregation and directional bar diagrams can also support task 4 – estimation of the proportions of cars leaving a road on its exits. Indeed, some diagrams in Figure 5 (left) show the proportions of the movements in different directions on road exits and crossings. However, these data are not very reliable. The course of the movement in a particular point is determined using the next measured position of the same car. Depending on the temporal spacing between the measurements and the speed of the movement, the next measurement may be taken on another road, somewhere on a curved exit, or on another lane of the same road just in a few meters from the previous measurement. The computed course of a car leaving the road may occasionally coincide with the direction of this road, and on the opposite, the course of a car staying on the road but changing the lane may significantly differ from the road direction. For a more reliable estimation of the proportion of the cars leaving the road, the further routes of the cars need to be taken into account. This means that task 4 requires the trajectory-oriented view of the car movement, like tasks 1, 2, and 5.

## 5 SUPPORTING THE TRAJECTORY-ORIENTED VIEW

In the trajectory-oriented view, collective movement of multiple entities is considered as a set of trajectories of the entities. In practical tasks, the entire trajectory of each entity made during the whole period of the observation is usually divided into parts representing different trips of this entity; the term “trajectory” is also applied to such a part.

In analyzing trajectories, one may be interested in the origins and destinations of the trips, routes, start and end times, durations, distances, variation of the speeds along the routes, intermediate stops, etc. When trajectories are numerous, it is impracticable to examine each of them in detail. They need to be aggregated in such a way that the distribution of the relevant properties over the set of trajectories could be seen. For certain properties, the aggregation may be quite traditional. Thus, a frequency histogram can appropriately represent the distribution of the trip durations or distances. More specific aggregation and visualization techniques are required for the spatial properties (origins, destinations, and routes) and for the spatio-temporal properties (speed variation and intermediate stops).

The general approach is to group the trajectories by similarity in terms of the properties relevant to the goals of the analysis. Then, the groups need to be represented in a summarized way, which appropriately conveys the relevant properties. The easiest case is when the analyst is interested only in the origins and destinations of the trips but not in the routes and spatio-temporal properties. This is the case in tasks 1 and 2. To support such tasks, the trajectories need to be grouped by the origins and destinations.

### 5.1 Aggregation by origins and destinations

In this method, which may be called S×S-aggregation, two approaches are possible. One is to refer the starts and ends of the trajectories to predefined areas of interest, for example, city districts. Then, for each pair of areas, the trajectories starting in



the first area and ending in the second area are grouped together. This applies also to the pairs where the first element coincides with the second one. The other approach is to define areas on the basis of spatial clustering of the start and end points of the trajectories. It is reasonable to assign meaningful names to the resulting clusters so that they could also be used as the names of the origins and destinations of the trips.

For each group of trajectories with a common origin and a common destination, the group size and the statistics (minimum, maximum, mean, median, etc.) of the numeric properties of the trajectories such as trip durations and distances are computed. The results may be displayed in the form of origin-destination matrix where the rows correspond to the origins, columns to the destinations, and the cells contain the values of the computed aggregates. The values in the cells may be represented visually by graduated symbols or diagrams. In our experimental software, the matrix display is linked to a map: clicking on a row, column, or cell highlights the corresponding areas on the map.

Another possibility for the aggregation is to account not only for the areas where a trajectory starts and ends but also for all intermediate areas visited by the trajectory. This means that each trajectory is generalized into a sequence of *moves* between areas. A *move* is a spatio-temporal object defined by the place and time of the start and the place and time of the end. An *aggregate move* combines moves with the same place of the start and the same place of the end. It is characterized by the number of the elementary moves it combines and various statistics of the duration, distance, speed, time, etc. computed from the respective trajectory fragments. These characteristics can be visualized in a matrix display like in the case of complete trajectories (Figure 7 right). The aggregate moves can also be shown on a map display as directed lines (vectors) between areas (Figure 7 left). The widths of the lines may represent the values of a selected aggregate attribute.

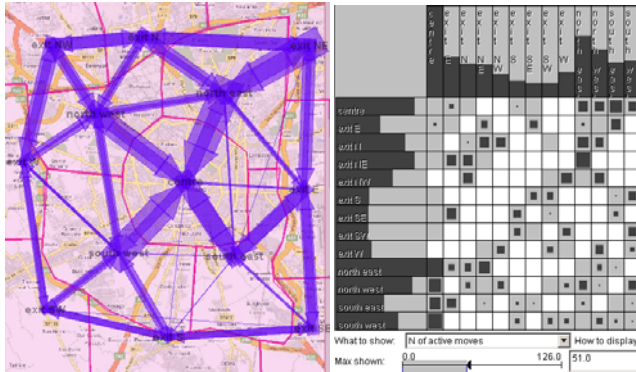


Figure 7. Summarization of trajectories into aggregate moves.

To investigate and compare the trips made in different time periods, the  $S \times S \times T$ -aggregation is used. The time is partitioned into linearly ordered intervals or according to temporal cycles. The  $S \times S$ -aggregation, as described above, is then applied separately to the trajectories or fragments of trajectories made during each of the temporal partitions. Thus, the screenshots in Figure 7 correspond to the time from 5 to 6 AM on Wednesday.

To support task 2 of the traffic managers, the matrix display needs to show the average or median travel times (trip durations) for the pairs of the areas. It may seem that task 1 is adequately supported by the matrix display showing the numbers of the trips between the areas. However, the available dataset does not contain data about *all* cars moving in Milan but only data about a limited number of cars. The true numbers of the trips between the areas cannot be derived from these data. A realistic estimation might be achieved by means of traffic simulation, which takes

into account available movement data about a sample of cars together with measurements from static traffic sensors.

Generally, data about a sample of cars may be sufficient for various analysis tasks where the analyst considers movement speeds, travel times, distances, or routes but not numbers or densities of cars or numbers of trips. The reason is that the cars from the sample mostly use the same roads and streets as the other cars and their speeds cannot differ much from the speeds of their neighbors in the city traffic.

Task 4 (estimate the proportions of the cars leaving a road on its exits) can be supported by the  $S \times S \times T$ -aggregation in a case when a representative set of trajectories going through this road is available (it may result from tracking a sufficiently big number of cars or from a realistic simulation). For this purpose, appropriate areas are built for the road exits and crossings and for the two sides of the road, and the trajectories are summarized into aggregate moves between these areas. Using dynamic filtering, it is possible to focus on outgoing (Figure 8) or ingoing moves, on moves starting or ending on a particular crossing, etc.

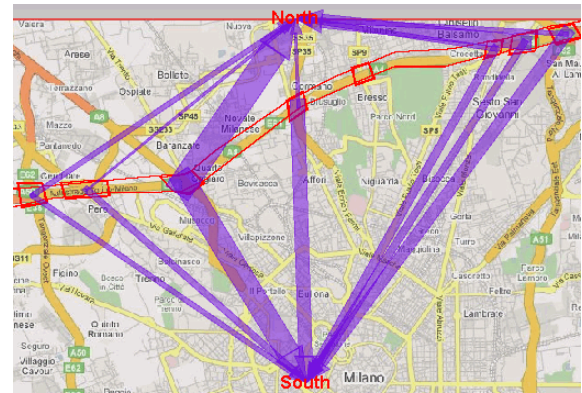


Figure 8. The aggregate moves represent the cars leaving a road on different exits between 7 and 9 AM on Wednesday.

The  $S \times S \times T$ -aggregation described in this subsection does not give an idea about the routes used for getting from place to place. For analysis tasks where routes are relevant, such as task 5 of the traffic managers, it is necessary to have methods for grouping trajectories according to the routes and for presenting the routes in a summarized way.

## 5.2 Aggregation by routes

In all aggregations discussed so far it is possible to specify in advance the groups to be produced in terms of the properties of their members. Thus, in  $S \times T$ -aggregation of track records, the groups are defined in terms of the spatial positions (which must fit in predefined space compartments) and time references (which must fit in predefined temporal partitions). In  $S \times T \times D$ -aggregation, the intervals for the values of movement direction are additionally specified. In  $S \times S$ -aggregation of trajectories, the groups are defined through pairs of areas in which the origins and destinations of the trajectories or their fragments must fit. In  $S \times S \times T \times T$ -aggregation, predefined temporal partitions for the start and end times of the trajectories or fragments are added.

In grouping by routes, it may not be possible to pre-specify a finite number of "model" routes for putting trajectories into groups based on their similarity to this or that route. In such a case, the trajectories may be grouped by means of clustering. In [2] we have described our clustering tool capable of using different measures of similarity between trajectories, also called distance functions. One of the available distance functions, which computes the average distance between corresponding points of two trajectories (the algorithm is given in [2]), can be used for

clustering trajectories by similarity of their routes. The resulting clusters need to be visualized to enable their interpretation.

One of the problems that need to be tackled is that trajectories included in the same cluster are not necessarily very similar and close (Figure 9 left); this depends on the parameter settings for the clustering. In case of high variability, summarizing trajectories by building an envelope around them or deriving an “average trajectory” may yield unclear or misleading results. We suggest representing groups of trajectories by aggregate moves between small areas. This is similar to what is described in the previous subsection except that the areas are not pre-specified but defined automatically using characteristic points of the trajectories, i.e. starts, ends, turns, and stops. The areas are built as circles around clusters of characteristic points from multiple trajectories and around isolated points. The radii vary within a user-specified range. Note that the areas so produced play an auxiliary role and do not need to be visualized (usually they are numerous and clutter the display).

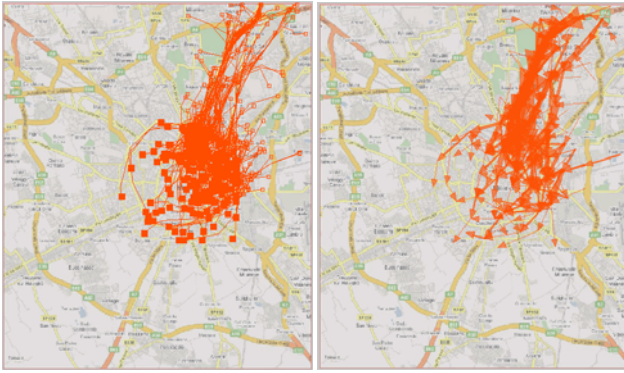


Figure 9. Left: an example of a cluster of trajectories grouped by the routes. Right: Aggregate moves summarizing the cluster.

The immediate result of aggregating a group of trajectories with high internal variability may not look clear enough. Thus, the screenshot on the right of Figure 9 shows all aggregate moves irrespectively of the number of the elementary moves they include. Many of the aggregate moves stand for just one or two elementary moves. By hiding minor aggregate moves through dynamic filtering one can see more clearly what is in common between the trajectories in the group. Furthermore, one can reconstruct one or more prototypical routes followed by the trajectories. This is demonstrated in Figure 10.

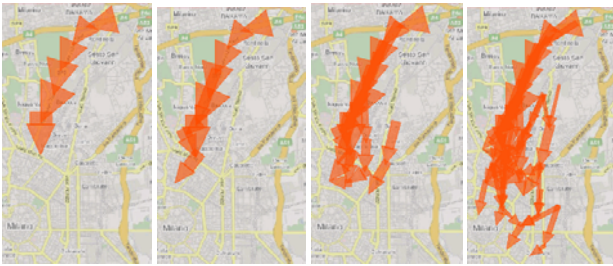


Figure 10. Revealing the commonality between the trajectories in a group and reconstructing the prototypical routes through dynamic filtering of aggregate moves.

The images from left to right correspond to the lower limits 20, 15, 10, and 5, respectively, for showing the aggregate moves on the map. The first two images expose the principal route of the cluster of trajectories. The third image shows branching of the principal route and a shorter secondary route parallel to it. The fourth image reveals further branching and small common segments appearing in at least 5 trajectories.

In an arbitrary set of trajectories, like in the Milan dataset, there may be multitudes of different routes. Clustering by routes will necessarily produce numerous groups of trajectories. There is no possibility to see them all together. In fact, even a few clusters may be hard to explore together if there are intersections and/or overlaps between their trajectories. It does not seem realistic that an analyst considers hundreds of clusters one by one. A more reasonable scenario is that the analyst has a certain focus of interest, for example, typical routes towards the centre of Milan or between two city districts, as in task 5 of the traffic managers. The analyst applies clustering only to the trajectories corresponding to his/her focus and then explores only the biggest clusters.

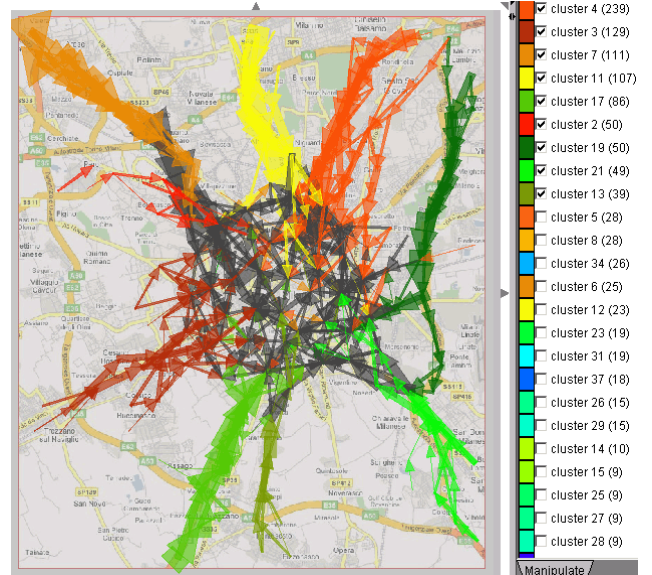


Figure 11. The biggest clusters of trajectories to the center.

For example, in Figure 11 we can see (in an aggregated form) the biggest clusters of trajectories going to the center of Milan. Each cluster is shown in a particular color. We observe the most typical routes towards the center and see how they mix inside the central area of the city; this is indicated by gray-colored lines.

We have demonstrated how groups of trajectories with similar routes can be explored with the help of the S×S-aggregation: the trajectories are transformed into aggregate moves between pairs of automatically defined areas. This can be extended to the S×S×T×T-aggregation, which allows an analyst to explore how the use of the typical routes changes over time. Furthermore, the vectors representing the aggregate moves on a map can be combined with visualization of various statistics related either to the moves or to the areas. For instance, the analyst can explore the temporal variation of the average speeds on different segments of the routes or of the average times spent in different places. The values of selected aggregate attributes can be represented on the map display by graduated symbols or diagrams drawn on top of the vectors. This gives a possibility to explore not only the spatial but also spatio-temporal characteristics of the trajectories.

## 6 POSSIBLE IMPLEMENTATION

The main goal of the paper is to introduce a general framework for analysis of massive movement data with the use of aggregation. Accordingly, we have tried to avoid discussing any implementation specifics. Here we would like to make just a couple of general notes concerning the possible implementation of the suggested methods.

The aggregations supporting the traffic-oriented view (§4) can all be done in a database by means of standard database functions.

Spatial OLAP operations can also be used for this purpose. The  $S \times S \times T$ -aggregation with predefined areas (§5.1) can be fulfilled in a database supporting spatial queries. The route-based aggregation (§5.2) is achieved with the help of clustering. Current clustering techniques work in the main memory. This limits the size of the data that can be analyzed. Our partners in the GeoPKDD project are now developing a scalable clustering tool with a substantial part of analysis performed inside the database.

On the other hand, having (a part of) movement data in the main memory gives very interesting possibilities for interactive analysis. Thus, in our experimental system we have implemented the aggregate moves and generalized places (areas) in the  $S \times S \times T$ -aggregation as *dynamic aggregators*. A dynamic aggregator keeps references to its members (i.e. the objects it aggregates) and reacts to various interactive operations on the set of the objects such as filtering and classification. In response, it adjusts the values of aggregate attributes and, as a consequence, alters its appearance in visualizations. Aggregate moves and places keep references to trajectories or fragments of trajectories. They react to the temporal filter (selection of a time interval), attribute filter (selection of trajectories by attributes such as duration and length), cluster filter (selection of clusters), and assignment of colors to groups of trajectories resulting from clustering or classification. Being represented on a map as vector symbols, aggregate moves can change their thickness or color and hide from the view when all their members are filtered out. Aggregate moves and places also control the appearance of symbols or diagrams representing values of aggregate attributes on a map or in a matrix display. More details about dynamic aggregators are given in [12].

## 7 CONCLUSION

Current positioning and tracking technologies enable collection of huge amounts of movement data. To make sense and use of such data, scalable analysis and visualization tools are very much needed. Visual exploration of massive movement data cannot be done without aggregation and summarization of the data.

We have undertaken an investigation into the aggregation methods suitable for movement data. We have considered known methods, specifically,  $S \times T$ -aggregation (space  $\times$  time) and  $S \times S \times T$ -aggregation (start place  $\times$  end place  $\times$  start time  $\times$  end time), and introduced new methods:  $S \times T \times D$ -aggregation ( $S \times T \times$  direction) and  $R \times S \times S \times T \times T$ -aggregation (route  $\times S \times S \times T \times T$ ). We have systemized these methods according to a framework based on an abstract model of movement data as a function of two variables. This model substantiates the possibility of considering movement data from two different perspectives, which we call traffic-oriented view and trajectory-oriented view. Each view requires different methods for analysis and visualization and, in particular, for data aggregation. The  $S \times T$ - and  $S \times T \times D$ -aggregation support the traffic-oriented view while  $S \times S \times T \times T$ - and  $R \times S \times S \times T \times T$ -aggregation are appropriate for the trajectory-oriented view.

We have also investigated what visualization and interaction techniques can support the exploration of massive movement data in combination with aggregation. We have pointed to known techniques suitable for this purpose and suggested new interactive visual techniques. In particular, the visualization with directional diagrams can be applied to results of the  $S \times T \times D$ -aggregation. The results of the  $R \times S \times S \times T \times T$ -aggregation can be visualized on a map by vectors varying in thickness and color.

In presenting the aggregation and visualization methods, we have used a real example dataset about movement of cars in a big city. We have demonstrated how the methods can support the analysis tasks of traffic managers. However, the methods are not specific for this type of movement data and these tasks. They can

be effective in various domains and for various kinds of data describing both constrained (e.g. by roads) and free movements of vehicles, people, animals, and other entities. Still, we are far from claiming that the aggregations and visualizations we have presented can solve all problems and fully satisfy the needs of all analysts. We are working on combining these techniques with other methods for analysis of movement data.

## REFERENCES

- [1] Andrienko, N., Andrienko, G., Pelekis, N., & Spaccapietra, S.: Basic Concepts of Movement data. In Giannotti, F. & Pedreschi, D. (eds.): *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*, Springer, Berlin, 2008, 15-38
- [2] Andrienko, G., Andrienko, N., & Wrobel, S.: Visual Analytics Tools for Analysis of Movement Data, *ACM SIGKDD Explorations*, 9(2), 2007, 38-46
- [3] Andrienko, N., & Andrienko, G.: Designing visual analytics methods for massive collections of movement data, *Cartographica*, 42(2), 2007, 117-138
- [4] Buliung, R.N. & Kanaroglou, P.S.: An Exploratory Data Analysis (ESDA) toolkit for the analysis of activity/travel data. Proceedings of ICCSA 2004, LNCS 3044, 1016-1025
- [5] Drecki, I., & Forer, P.: *Tourism in New Zealand - International Visitors on the Move (A1 Cartographic Plate)*; Tourism, Recreation Research and Education Centre: Lincoln University, Lincoln, 2000
- [6] Dykes, J. A. & Mountain, D. M.: Seeking structure in records of spatio-temporal behavior: visualization issues, efforts and applications, *Computational Statistics and Data Analysis*, 43 (Data Visualization II Special Edition), 2003, 581-603.
- [7] Fredrikson, A., North, C., Plaisant, C., & Shneiderman, B.: Temporal, geographical and categorical aggregations viewed through coordinated displays: a case study with highway incident data. In *Proc. Workshop on New Paradigms in information Visualization and Manipulation* (Kansas City, Nov. 1999). ACM, NY, 1999, 26-34.
- [8] Forer, P., & Huisman, O.: Space, Time and Sequencing: Substitution at the Physical/Virtual Interface. In Janelle D.G. and Hodge D.C. (eds), *Information, Place and Cyberspace: Issues in Accessibility*, Springer-Verlag, Berlin, 2000, 73-90
- [9] Guo, D., Chen, J., MacEachren, A. M., & Liao, K.: A Visual Inquiry System for Spatio-Temporal and Multivariate Patterns (VIS-STAMP), *IEEE Transactions on Visualization and Computer Graphics*, 12(6), 2006, 1461-1474
- [10] Mountain, D.M.: Visualizing, querying and summarizing individual spatio-temporal behavior. In *Exploring Geovisualization*. (Eds, Dykes, J.A., Kraak, M.J. & MacEachren, A.M.) Elsevier, London, 2005, 181-200
- [11] Ntoutsis, I., Mitsou, N., & Marketos, G.: Traffic mining in a road-network: How does the traffic flow? *Int. J. of Business Intelligence and Data Mining*, 3(1), 2008, 82-98
- [12] Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., & Andrienko, G.: Visually-driven analysis of movement data by progressive clustering, *Information Visualization*, 7(3), 2008 (in press).
- [13] Shneiderman, B.: Tree visualization with treemaps: a 2-D space-filling approach. *ACM Transactions on Graphics* 11(1), 1992, 92-99
- [14] Thomas, J.J., & Cook, K.A., eds.: *Illuminating the Path. The Research and development Agenda for Visual Analytics*, IEEE Computer Society, 2005
- [15] Tobler, W.: Experiments in migration mapping by computer, *The American Cartographer*, 14 (2), 1987, 155-163
- [16] Tobler, W.: Display and Analysis of Migration Tables, 2005, [http://www.geog.ucsb.edu/~tobler/presentations/shows/A\\_Flow\\_talk.htm](http://www.geog.ucsb.edu/~tobler/presentations/shows/A_Flow_talk.htm)
- [17] Wood, J., Slingsby, A., & Dykes, J.: Using Treemaps for Variable Selection in Spatio-Temporal Visualization, *Information Visualization*, 7(3), 2008 (in press).