Extracting Patterns of Individual Movement Behaviour from a Massive Collection of Tracked Positions

Gennady Andrienko and Natalia Andrienko

Fraunhofer Institute IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
gennady.andrienko@iais.fraunhofer.de
http://www.ais.fraunhofer.de/and

Abstract. A EU-funded project GeoPKDD develops methods and tools for analysis of massive collections of movement data, which describe changes of spatial positions of discrete entities. Within this project, we design and develop methods of visual analytics, which combine interactive visual displays with database operations and computational methods of analysis. In this article, we demonstrate by example how visual analytics methods can help in acquiring knowledge about the movement behaviour of an individual from a very large set of movement data.

1 Introduction

A EU-funded project GeoPKDD - Geographic Privacy-aware Knowledge Discovery and Delivery (IST-6FP-014915; see http://www.geopkdd.eu) - aims at developing methods and tools for analysis of massive collections of movement data. Movement data describe changes of spatial positions of discrete entities, which preserve their integrity and identity, i.e. do not split or merge. Within this project, we develop methods for supporting human analysts in visual inspection of movement data and detection of characteristic patterns of movement behaviours.

It is commonly recognised that interactive and dynamic visual representations are essential for gaining understanding of spatial and spatio-temporal data and underlying phenomena. However, visualisations alone are insufficient for exploration and analysis of massive data collections. This is not only the matter of technical limitations such as the screen size and resolution or the speed of rendering but also of the natural perceptual and cognitive limitations of the humans who need to view and interpret the visual displays. Hence, it is necessary to combine visualisation with computational analysis methods, database queries, data transformations, and other computer-based operations.

Recently, we have developed a theoretical basis for the creation of methods for visual analysis of movement data (Andrienko and Andrienko 2007). In particular, we have defined the possible types of behavioural patterns that can be detected by analysing movement data alone and in combination with data about other phenomena. Next, we have envisaged the kinds of data transformations, computations, and visu-

alisations that could enable a human analyst to detect these pattern types in truly massive data, possibly, not fitting in a computer's memory. On the basis of the previous works (Tobler 1987; Dykes and Mountain 2003; Laube, Imfeld, and Weibel 2005; and others), we have suggested a set of techniques where a key role belongs to aggregation and summarisation of the data by means of database operations and/or computational techniques.

For a practical verification of this choice of techniques resulting from a theoretical analysis, we started a prototype implementation of a visual analytics (Thomas and Cook 2005) toolkit for movement data. In this article, we demonstrate by example how visual analytics methods can help in acquiring knowledge about the movement behaviour of an individual from a very large set of movement data.

2 The Example Dataset

The example dataset consists of more than 60,000 records of positions of a car, which has been tracked during 5 months. The data have been recorded only when the car moved, i.e. there are no records for stops and still periods. The temporal spacing of the records is mostly 1 second; however, the records corresponding to periods of uniform movement (i.e. with constant speed and direction) are sparser. The data are stored in a relational database.

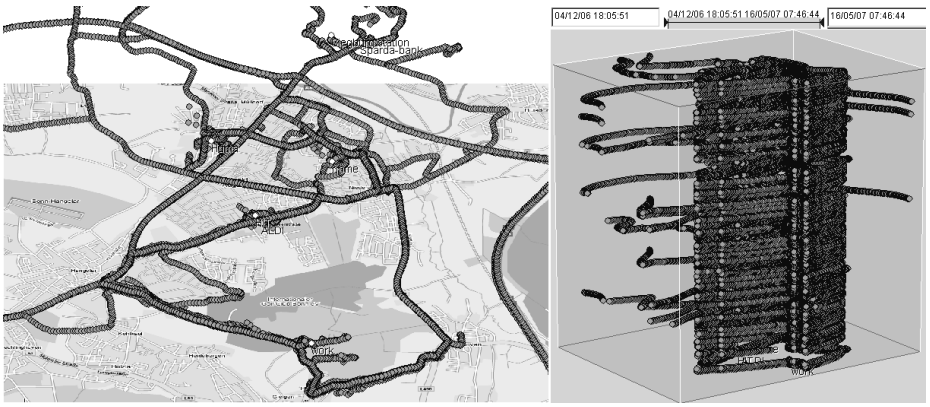

Fig. 1. An attempt to display all individual data items from a large dataset is not productive for data exploration and analysis.

The dataset is too large for a straightforward visualisation of all data items. Fig.1 demonstrates the result of showing all positions of the car on a map (left) and in a space-time cube (right), where the horizontal plane represents the geographical space and the vertical dimension represents the time. The map, in fact, reveals only the street network in the area where the car moves. It is even impossible to see which streets are used frequently and which only occasionally, because the symbols greatly overlap. Temporal filtering and display animation do not help much: the intervals of

movement are very short in relation to the 5-months long time period and therefore hard to extract through interactive filtering and hard to detect by viewing an animated display, where most of the time nothing happens. Hence, to be able to extract useful information from this mass of data, it is necessary to summarise it somehow before trying to visualise.

In our work, we put a particular focus on the use of database operations for data summarisation and other data transformations and computations. On this basis, we strive at developing scalable visual analytics methods, which could be applied even to datasets not fitting in the computer memory. Besides database operations and visualisation, we utilise data mining techniques, as will be seen from the following sections.

3 Detecting important places

A temporally ordered sequence of all positions of an entity is not a meaningful object for analysis since the entity does not necessarily move all the time (thus, in our data, the time of movement is much less than the time of stillness). It is reasonable to divide the sequence into trajectories or into movement episodes. A *trajectory* is a sequence of items corresponding to a trip of an entity from one location (source) to another (destination) where the source and destination are defined semantically (e.g. home, work, shop, etc.) or according to the time the entity spends in a location. *Movement episodes* (Dykes & Mountain 2003) are fragments of trajectories where the movement characteristics (speed, direction, sinuosity, etc.) are relatively constant whereas a significant change indicates the beginning of the next episode.

In our exercise on analysing the car movement data, we assume that we have no background knowledge that would allow us to divide the data into trips using semantic criteria. Moreover, it is one of the tasks of our analysis to extract and interpret the sources and destinations of the trips. Therefore, we need to find the sources and destinations on the basis of the temporal criterion, i.e. according to the time spent in a location.

As we have explained, the records of the car positions have been made only when the car actually moved; hence, the stops and periods of stillness are present in the data implicitly as temporal gaps between successive records. It is easy to find such gaps with the use of database operations; however, it is necessary to specify the minimum temporal distance between records to be treated as a "gap". This threshold can be chosen quite arbitrarily. Interestingly, by setting different temporal thresholds, it is possible to find places of different importance for the moving object, i.e. car user in our case. Thus, setting a threshold of several hours should result in finding places where the person spends much time. These will include person's home and work.

Fig.2 presents the spatial positions of the trip starts and ends, which have been extracted from the database using a temporal threshold of 2 hours. The positions are shown as small circles on a map. The map on the left shows all extracted positions. Surprisingly, there are much more different positions than could be expected. The maps in the middle and on the right show the starts and ends separately. It is easy to notice that the starts are much more dispersed in space than the ends. This looks very strange: the start position of a trip should normally coincide with the end position of

the previous trip. The reason for the observed discrepancy is that the GPS (Global Positioning System) device, which is used for collecting the data, needs some time for warming up, detecting satellites, and establishing connections with them. Therefore, the device starts recording the positions of the car not from the moment when a trip begins but later. Hence, our data are incomplete, and the real times and positions of trip starts are unavailable. This feature needs to be taken into account when analysing the data.

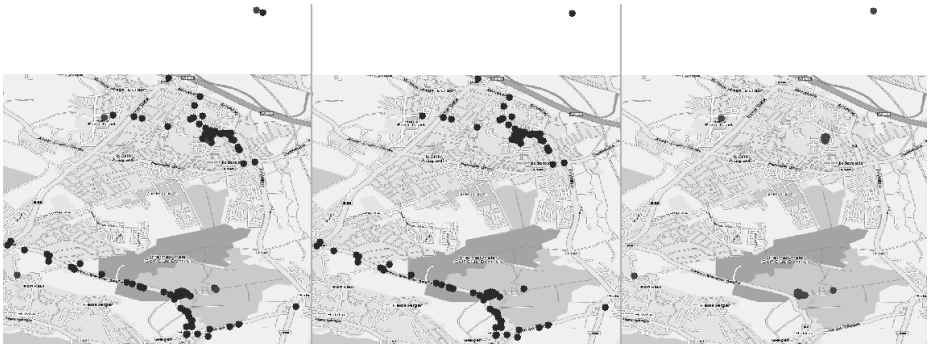

Fig. 2. The positions of the starts and ends of the trips have been extracted by setting a temporal threshold of 2 hours. Left: all extracted positions; middle: starts; right: ends.

It is reasonable to assume that the spatial positions of the trip starts are the same as the destinations of the previous trips. Therefore, we can ignore the extracted starts and look only at the ends, which are more reliable. On the right of Fig.2, we can see 5 different places where the trip destinations are located. Naturally, we are interested first of all in finding frequently visited places, or, in other words, places where the trip destinations are clustered. A map display is not appropriate for this purpose: it is hard to guess how many overlapping circles there are in each place. Instead, we can apply computational techniques for detection of spatial clusters, which are developed in the research area of data mining.

The map on the left of Fig.3 shows the result of applying a clustering tool to the destinations detected with the use of the temporal threshold of 2 hours. The tool has found two clusters; the dark circles mark the corresponding positions. The remaining positions, which are represented by lighter circles, were classified as noise. From the two clusters, the one on the north contains 118 positions and the other, which is on the south, contains 77 positions. It is reasonable to conclude that the larger cluster is located near the home of the car user and the smaller cluster is at the place where the person works.

In the middle of Fig.3, the map presents the results of applying the same clustering tool to the trip destinations extracted using the temporal threshold of 1 hour. Additionally to the two clusters detected before, one more cluster consisting of 11 positions has been found west from the place interpreted as “home”. On the right, the clusters of destinations where the car user spent at least 5 minutes are shown. There

are five clusters, including the three previously detected clusters. Two more clusters have appeared in the centre of the map.

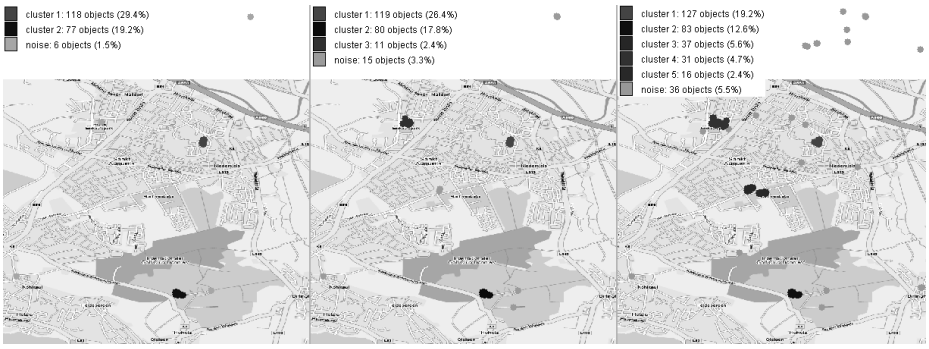

Fig. 3. Results of a clustering method applied to 3 sets of trip destinations extracted using different values for the temporal threshold: 2 hours (left), 1 hour (middle), and 5 minutes (right).

The meaning of the places detected in this way can be established using background knowledge about the territory or information provided by the background map. Unfortunately, the available background map is just an image with a low level of detail. It is only possible to find out that one of the clusters (cluster 3) is located at a shopping centre named "Huma Einkaufspark". However, we are familiar with the territory and can interpret also the clusters 4 and 5: these are located in a shopping area where several stores are separated by a street.

Hence, by extracting and analysing trip destinations, we have found the places where the car user lives, works, and shops. There are two frequently visited shopping areas. The car user spends more time in "Huma Einkaufspark" than in the other shopping area.

Using background knowledge about the territory and/or background map, it is possible to identify also the places visited less frequently, as, for example, post office or bank.

4 Analysing trip directions

After determining the significant places visited by the car user (these places will be henceforth called "places of interest", or POIs), we would like to know how the person moves between them. Thus, we can see that one of the shopping areas is located between the person's home and work. Does the person visit it on the way from the work to home and, if so, how often? Does the car user ever go from the work to the other shopping area? Were there any trips from one shopping area to the other?

In order to answer these and similar questions, it is reasonable to count the number of trips between each pair of POIs. However, each position in the dataset is specified only through a pair of coordinates (latitude and longitude), i.e. as a geographical point

without any semantics. In order to enable a software tool to treat the starting and ending positions of the trajectories as particular places of interest, it is necessary to specify the places explicitly as named areas surrounded by boundaries. For this purpose, one can use computational functions available in geographic information systems (GIS) or spatial DBMS to build buffer zones around the positions of the trip starts and ends and then assign meaningful names to the zones obtained. Another approach is to encircle the areas on the map manually. Here, we shall use manually defined POIs. Besides encircling the trip destinations we could interpret, we also considered the extracted starting positions (Fig.2 middle) and associated some of them with the most probable trip sources. Thus, the start positions of many trajectories lie on the roads passing near the place of the person's work. As we know, these are false starts. It is reasonable to assume that the real source of the corresponding trips is the place of the work. Hence, we have drawn several shapes enclosing the false start positions and named them "(work)*", "(work)**", and "(work)+". Similarly, we have defined an additional POI named "(home)*" by enclosing the false start positions of the trips starting, most probably, at home. Fig.4 shows the POIs we have specified.

Fig. 4. The places of interest defined by encircling areas on a map.

After the places of interest have been defined, a software tool can attach their names to the positions of the trip starts and ends lying within the areas. Then, it becomes possible to count the number of trips between each pair of places. The resulting counts can be visualised, for example, in an interactive matrix display shown in

Fig.5. The rows of the matrix correspond to the trip sources, the columns to the destinations, and the sizes of the rectangles in the cells encode the numbers of the trips. By putting the mouse cursor on a cell we can learn the exact number of trips made from the corresponding source to the corresponding destination.

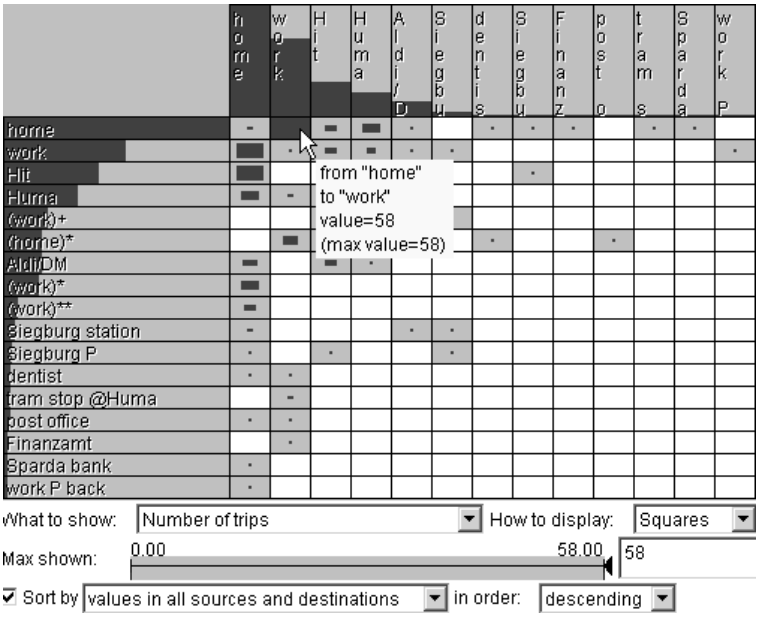

Fig. 5. The numbers of trips between pairs of POIs are represented by proportionally sized rectangles in cells of a matrix where the rows correspond to the trip sources and columns to the destinations.

Thus, during the period under the study, there were 76 trips to work, from which 68 trips were from home (58 from the POI "home" plus 10 from the POI "(home)*" enclosing the false starts of the trips from home). There were 2 trips to work from the shopping area "Huma" and no trips starting in the other shopping area.

Let us now look to which places the person drives from the work. There are several source POIs associated with the place of work. For a more convenient exploration, we can apply an interactive filtering tool to select only the trips starting from any of these POIs. The interactive matrix display reacts to setting the filter by removing irrelevant information (Fig.6). Now, it is more convenient to learn that there were in total 80 trips starting from the work, from which 50 were directly to home, 7 to "Huma" and 19 to the other shopping area (17 to "Hit" plus 2 to "Aldi/DM" on the other side of the street), 2 trips to "Siegburg station" and 2 trips back to work.

As the matrix display lacks the geographical context, it may be useful to complement it with a map display. The map in Fig.7 shows the same summarised information about the trips from the work by vectors (directed lines) connecting the source and destination locations. The widths of the lines are proportional to the numbers of the trips between the respective locations. Unfortunately, the map is not easy to read

because of the overlapping of the vector symbols. Still, the major destinations of the trips from the work can be grasped.

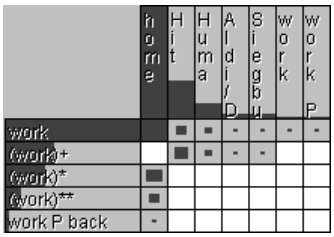

Fig. 6. The matrix display shows only the information about the trips from the work.

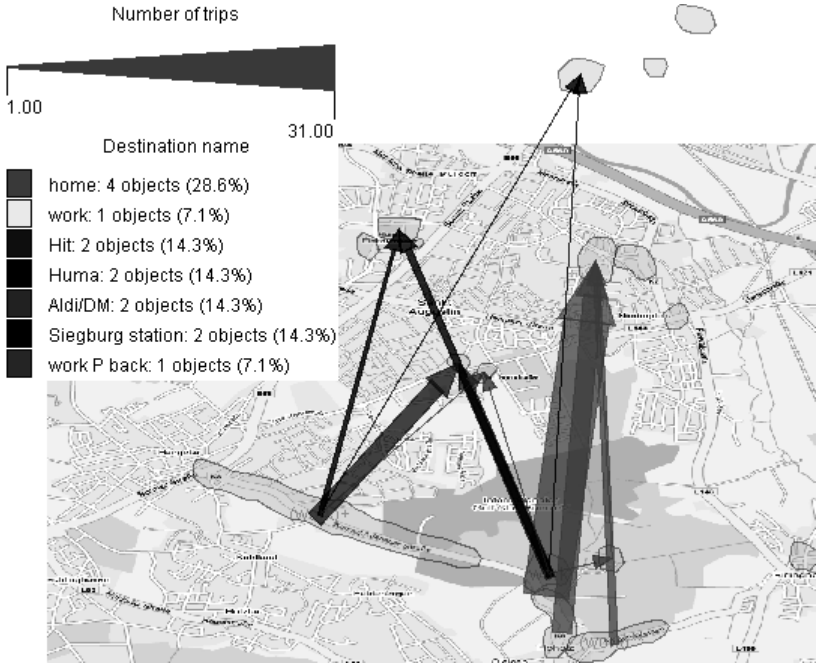

Fig. 7. The same information as in Fig.6 is shown on a map by vectors.

Besides filtering by trip origin, it is possible to set various other filter conditions. For instance, two screenshots of the matrix display presented in Fig.8 show the results of filtering the trips according to the time of the day when a trip begins. On the left, we see the summarised information about the trips starting from 8 to 11 hours, and on the right – from 17 to 20 hours. Most of the morning trips are from home to work, but the evening trips are much more varied. Analogously, it is possible to compare the trips made on working days with the trips made on weekends (Fig.9).

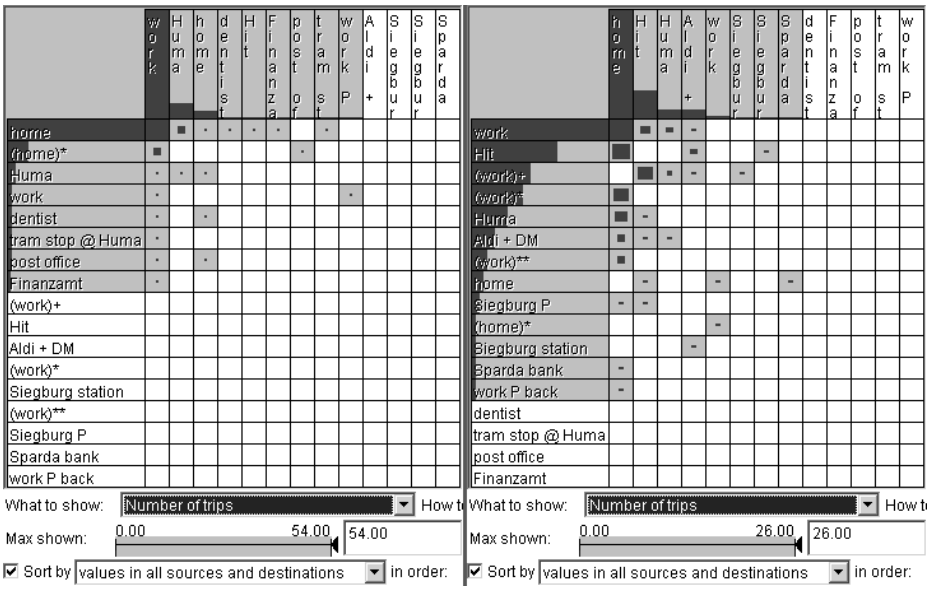

Fig. 8. Filtering of the movement data by the time of the day allows us to compare the major trip directions in the morning (left) and in the evening (right).

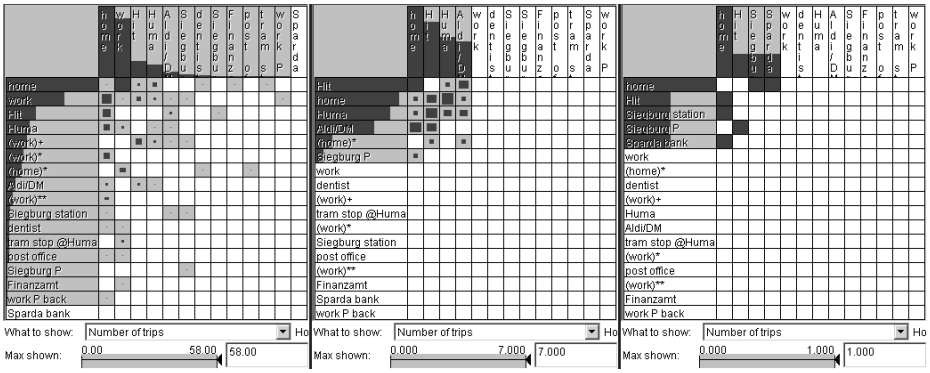

Fig. 9. Left: trips made on the week days from Monday to Friday. Middle: trips made on Saturdays. Right: trips made on Sundays.

5 Analysing trajectories

Both in the matrix display and in the map with vectors (Fig.7) the information about the trips is highly summarised: it is only possible to see the origins and destinations and the numbers of the trips. In studying movement behaviours of individuals, it is

also important to investigate their trajectories, or, in other words, the routes they use. For example, in our case we would like to know what routes the person chooses on the way from home to work and back. If the person uses different routes, it is interesting to find out when which route is preferred and to make plausible guesses about the reasons for choosing this or that route.

For such investigations, we need a detailed representation of the person's trajectories in the geographical context, i.e. on a map or in a space-time cube. However, the representation of all trajectories at once results in an unreadable display similar to what can be seen in Fig.1. A reasonable approach is to explore the trajectories by interpretable portions with the use of the tool for interactive filtering. In particular, it is useful to select subsets of trajectories according to the sources and destinations of the trips. Thus, the map in Fig.10A shows only the trajectories starting from the work and ending at home (the trajectories have been defined using a 2 hour temporal threshold for dividing the sequence of positions). The trajectories are represented as polygonal lines, their starting points are marked by hollow squares (Fig.10B) and end points by filled squares (Fig.10C). It is hard to estimate the number of overlapping lines, but the filtering tool informs us that there are 74 trajectories from work to home among 201 trajectories in total.

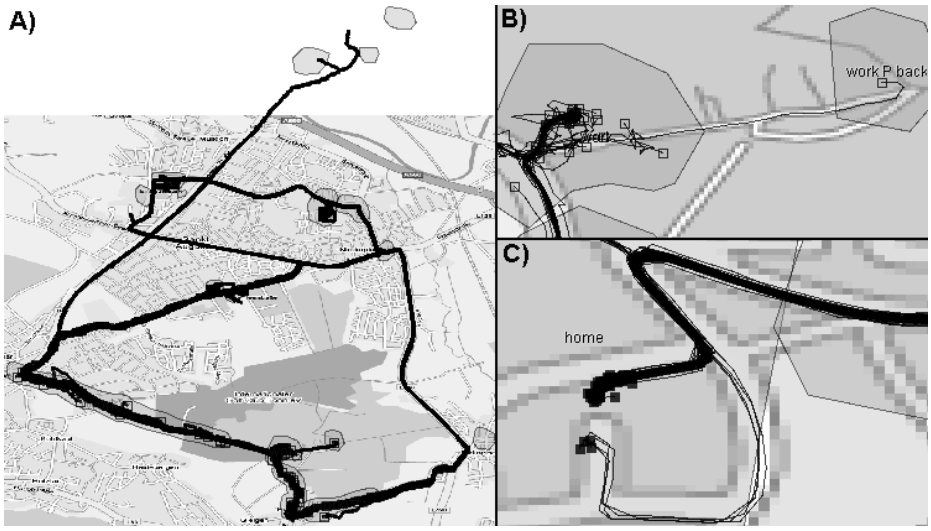

Fig. 10. A) The trajectories from work to home are represented on a map as polygonal lines. B) The start points of the trajectories are marked by hollow squares. C) The end points of the trajectories are marked by filled squares.

As the lines severely overlap, it is hardly possible to understand what routes the person uses for driving from work to home. It is necessary to group trajectories with similar shapes and to look at each group separately. One of the possibilities for groping is by interacting with the map display. Clicking on a position on the map selects all lines passing through this position or close to it. From the lines selected in this way one can make a group, or class. By clicking on different roads, it is possible to

make several selections and to form several classes. Thus, in our case, we have defined 6 classes of trajectories from work to home differing in shape (Fig.11), three of which consist of singular trajectories (classes 4, 5, and 6 in the lower row in Fig.11). The most frequently followed route (upper left of Fig.11) is by the road on the east of the territory; the person used it 43 times. The second frequent route (upper middle), which was used 21 times, passes the shopping area in the middle of the territory.

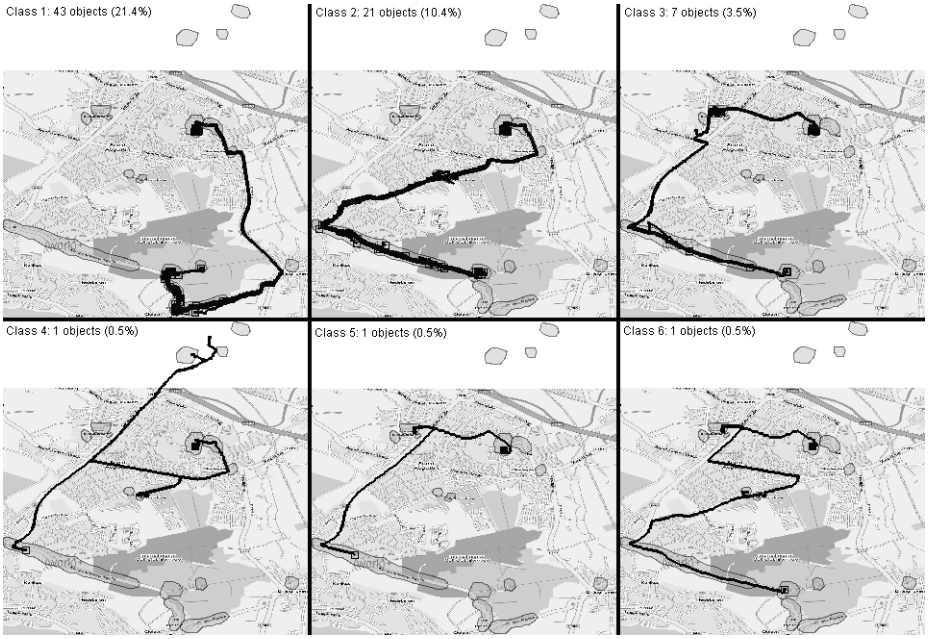

Fig. 11. By interacting with the map display, we have detected 6 different routes from work to home.

	N	N?	min	q1	med	q2	max	ave	stdd
	201	0	390	625	739	2872	14406	2382	2978
■	43	0	435	599	625	662	1028	637	95
■	21	0	641	1658	2745	3192	4501	2607	1163
■	7	0	1234	2018	2832	2971	5765	3239	1256
■	1	0	4306		4306		4306	4306	0
■	1	0	1755		1755		1755	1755	0
■	1	0	4406		4406		4406	4406	0
■	127	0	390	631	828	3358	14406	2863	3513

Fig. 12. Statistics of the trip duration for the different routes shown in Fig.11.

It is useful to look at various statistics about the classes of the trajectories. Thus, Fig.12 shows the statistics of the trip duration, in seconds: minimum, first quartile, median, third quartile, maximum, average, and standard deviation. The upper row of

the table corresponds to the whole set of trajectories, the next 6 rows correspond to the groups of the trips from work to home we have defined, and the last row corresponds to the remaining trajectories. We can see that the route of class 1 takes the least time. This can be explained by the absence of POIs that could be visited on this way. It is highly probable that the routes corresponding to classes 2 and 3 are chosen when the person needs to visit one of the shopping areas. In order to check this, we would need a tool computing the time spent in each POI during each trip; however, such a tool is not available at the moment of writing this paper.

The histogram in Fig.13 shows the distribution of the trips by days of week, from Monday to Sunday. The light grey bars correspond to the entire set of trips. The black, dark grey, and medium grey segments show the proportions of the trips from the classes 1, 2, and 3, respectively. It is notable that the route corresponding to class 2 is most often (7times) chosen on Wednesdays but is used also in other working days of the week. The route corresponding to class 3 was used 3 times on Thursdays, 3 times on Fridays, only once on Monday, and never in the other days of the week.

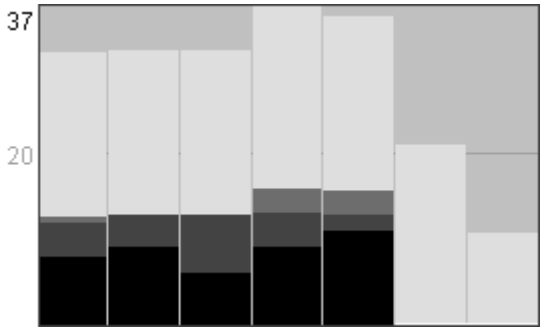

Fig. 13. The histogram shows the distribution of the trips by days of week.

Not only interactive techniques can be used to group trajectories by similarity but also methods for computational clustering, which are preferable in case of great overlaps between trajectories and/or complex shapes with loops and self-crossings. Automatic clustering of any items requires a method that computes the degree of dissimilarity (also called "distance"; this term is used in a wider sense than purely distance in space) between a given pair of items. Such a method is called "distance function". Clustering algorithms and distance functions suitable for trajectories are now under development within the project GeoPKDD. For the car movement data we have, we have implemented a simple distance function that takes into account the incompleteness of the trajectories: it compares trajectories starting from their ends and ignores differences in the lengths. Fig.14 presents a result of automatic clustering of the trajectories from work to home (it should be noted that clustering results may differ depending on the choice of clustering parameters, in our case, the distance threshold – the maximum allowed distance between members of a cluster). The clusters agree very well with the results of our interactive grouping: clusters 1 and 2 are exactly the same as our classes 1 and 2, cluster 3 includes all trajectories from our class 3 plus

the single trajectory we have put in class 5, and the remaining two trajectories are treated as "noise", i.e. as too dissimilar to the other trajectories.

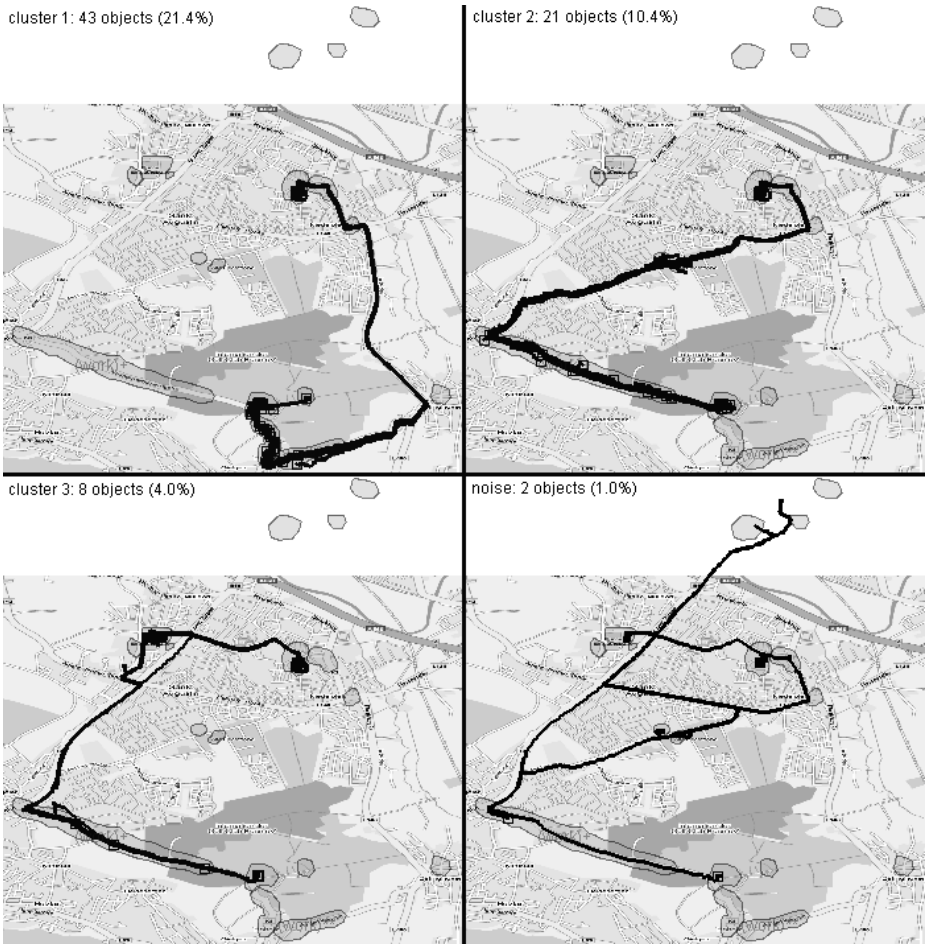

Fig. 14. A result of automatic clustering of the trajectories from work to home. Parameters: the distance threshold is 300 meters and the minimum number of cluster members is 3.

Let us now utilise the clustering tool to investigate how the car user goes from home to work. One of possible results of clustering is presented in Fig.15.

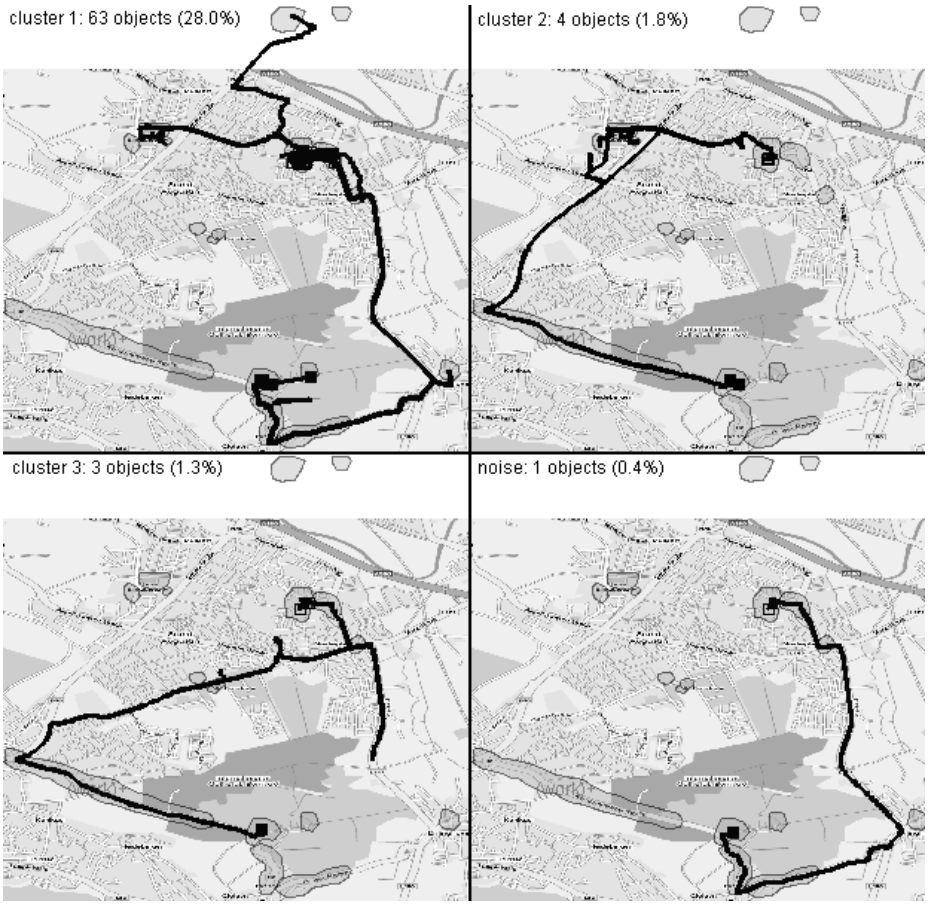

Fig. 15. A result of automatic clustering of the trajectories from home to work. Parameters: the distance threshold is 500 meters and the minimum number of cluster members is 3.

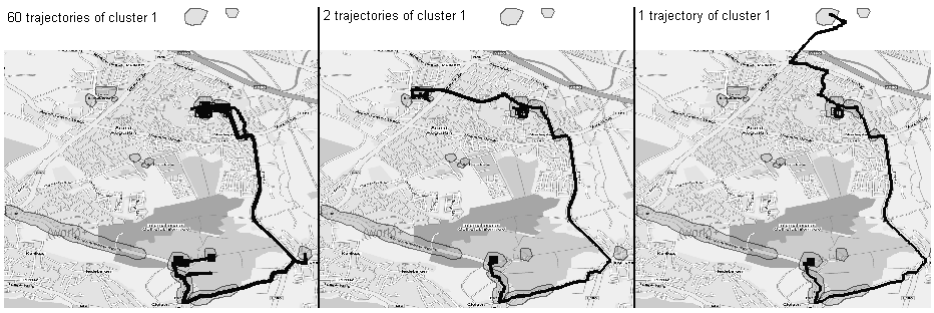

Fig. 16. The composition of cluster 1 from the previous figure.

As can be seen from Fig.15, the person takes almost always the eastern road for driving from home to work. The trajectories following this road are united in cluster 1. Fig.16 shows the composition of this cluster in some more detail. On the left, there are 60 trajectories of very similar shapes. In the middle, there are 2 trajectories where the person first visited the shopping area "Huma", then returned home, and then drove to work. On the right, there is a single trajectory where the person visited the POI "Siegburg station" before going to work.

From the other trajectories from home to work, 4 go along the western road (cluster 2) and 3 trajectories use the diagonal road (cluster 3). Cluster 3 contains one peculiar trajectory: the person first drove half way along the eastern road and then returned back and took the diagonal road. Perhaps, there was some obstacle on the eastern road that day.

It can be noticed that the single trajectory marked as "noise" (bottom right of Fig.15) has the same shape as the standard trajectories in cluster 1 (Fig.16 left). This exposes a weakness of the distance function we use. For some unclear reason, the trajectory marked as "noise" consists of 525 different positions, which is much more than in the standard trajectories of cluster 1 (from 189 to 300). Our distance function cannot properly cope with such a difference and, evidently, requires improvement. At least two important implications can be derived from this observation. First, peculiarities of data to be analysed must be properly taken into account in designing and/or choosing methods for automated analysis. Second, a careful and critical examination of the results of automated methods is absolutely necessary. Interactive visual interfaces are appropriate instruments for this.

6 Conclusion

The main objective of this article was to demonstrate the use of interactive visual tools combined with database processing and computation for the exploration and analysis of large spatio-temporal datasets, more specifically, data about changes of spatial positions of discrete entities. We have shown how patterns of individual movement behaviour can be extracted from a very large number of position records and semantically interpreted. We could continue this investigation and learn much more about the person's life style and habits. Such a possibility raises serious concerns about the privacy of individuals. Therefore, one of the main objectives of the project GeoPKDD is to develop mechanisms for preventing the disclosure of sensitive private information. Such mechanisms need to be incorporated both in computational and in visual tools for analysis.

References

1. Andrienko N and Andrienko G, 2007, Designing Visual Analytics Methods for Massive Collections of Movement Data, Cartographica, 42(2):117-138.

2. Dykes JA and Mountain DM, 2003, Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications, Computational Statistics and Data Analysis, 43 (Data Visualization II Special Edition):581-603
3. Laube P, Imfeld S, and Weibel R, 2005, Discovering relative motion patterns in groups of moving point objects, International Journal of Geographical Information Science, 19(6):639–668
4. Thomas JJ and Cook KA (eds), 2005, Illuminating the Path. The Research and development Agenda for Visual Analytics, IEEE Computer Society, USA
5. Tobler W, 1987, Experiments in migration mapping by computer, The American Cartographer, 14 (2): 155-163