Interactive Cluster Analysis of Diverse Types of Spatiotemporal Data

Gennady Andrienko Natalia Andrienko
Fraunhofer Institute IAIS (Intelligent Analysis and Information Systems)
Schloss Birlinghoven, Sankt Augustin
D-53757, Germany
+49 2241 142486
{gennady.andrienko, natalia.andrienko}@iais.fraunhofer.de

ABSTRACT
We suggest an approach to exploratory analysis of diverse types of spatiotemporal data with the use of clustering and interactive visual displays. We can apply the same generic clustering algorithm to different types of data owing to the separation of the process of grouping objects from the process of computing distances between the objects. In particular, we apply the density-based clustering algorithm OPTICS to events (i.e. objects having spatial and temporal positions), trajectories of moving entities, and spatial distributions of events or moving entities in different time intervals. Distances are computed in a specific way for each type of objects; moreover, it may be useful to have several different distance functions for the same type of objects. Thus, multiple distance functions available for trajectories support different analysis tasks. We demonstrate the use of our approach by example of two datasets from the VAST Challenge 2008: evacuation traces (trajectories of moving entities) and landings and interdictions of migrant boats (events).

1. INTRODUCTION
Clustering, i.e. discovery and interpretation of groups of objects having similar properties and/or behaviors, is one of the most common operations in exploration and analysis of various kinds of data. Clustering is particularly useful in exploring and analyzing large amounts of data since it allows an analyst to consider groups of objects rather than individual objects, which are too numerous. Clustering may also be useful for other purposes, for instance, to detect uncommon objects, which may require special investigation. However, clustering is not a standalone method of analysis whose outcomes can be immediately used for whatever purposes (e.g. decision making). An essential part of the analysis is interpretation of the clusters by a human analyst; only in this way they acquire meaning and value. To enable the interpretation, the results of clustering need to be appropriately presented to the analyst. Visual and interactive techniques play here a key role.

In clustering, objects are often treated as points in multi-dimensional space of properties. However, this approach may be inadequate for structurally complex objects, such as trajectories of moving entities and other kinds of spatiotemporal data. Thus, trajectories are characterized by a number of non-trivial and heterogeneous properties including the geometric shape of the path, its position in space, the life span, and the dynamics, i.e. the way in which the spatial location, speed, direction and other point-related attributes of the movement change over time. Each of these diverse properties needs to be handled in its own way.

There are two main approaches to clustering complex data: (i) defining ad hoc notions of clustering and devising clustering algorithms tailored to the specific data type; and (ii) applying generic notions of clustering and generic clustering algorithms by defining a specific distance function, which measures the degree of dissimilarity between data items. In the second case, the specifics of the data are completely encapsulated in the distance function.

In our research, we pursue the second approach. We use a generic density-based clustering algorithm OPTICS [5], which belongs to the DBSCAN [6] family. Advantages of these methods are tolerance to noise and capability to discover arbitrarily shaped clusters. A brief description of OPTICS is given in [11]. We use an implementation of OPTICS that allows different distance functions to be applied. We have developed a library of distance functions where there are functions suitable for trajectories, events (i.e. objects having spatial and temporal positions), and spatial distributions of events or moving entities in different time intervals.

We demonstrate how the clustering tool can be applied to diverse types of spatiotemporal data and help answer different analytical questions by example of two benchmark datasets from the VAST Challenge 2008 [8]: evacuation traces (trajectories of moving entities) and landings and interdictions of migrant boats (events). Prior to that, we describe the distance functions oriented to different data types: trajectories, events, and spatial distributions of moving entities or events in different time intervals.

2. DISTANCE FUNCTIONS
The clustering tool that we use has three parameters: the spatial distance threshold *maxD*, the minimum number of neighbors of a core object *MinNbs*, and the distance function *F*. The second parameter requires some explanation. Neighbors of an object are such objects whose distances to this object are below the distance threshold *maxD*. A core object is an object located in a dense region, i.e. inside some cluster. The parameter *MinNbs* defines the desired density inside a cluster. Additionally to these, some of the distance functions have their own parameters.

2.1 Distances between trajectories
As we argue in [11], it would not be reasonable to create a single distance function for trajectories that accounts for all their diverse properties. On the one hand, not all characteristics of trajectories may be simultaneously relevant in practical analysis tasks. On the other hand, clusters produced by means of such a universal function would be very difficult to interpret. A more reasonable

approach is to give the analyst a set of relatively simple distance functions dealing with different properties of trajectories and provide the possibility to combine them in the process of analysis.

We suggest and instrumentally support a step-wise analytical procedure called "progressive clustering" [11]. The main idea is that a simple distance function with a clear meaning and principle of work can be applied on each step, which leads to easily interpretable outcomes. However, successive application of several different functions enables sophisticated analyses through gradual refinement of earlier obtained results.

Our distance functions for trajectories are described in [2] and [11]. Here we briefly describe the functions we have used in analyzing the VAST Challenge data [8]. The function *"common destination"* computes the distance in space between the ending points of two trajectories. This is the distance on the Earth surface if the positions are specified in geographical coordinates (latitudes and longitudes) or the Euclidean distance otherwise. The family of functions *"check points"* computes the distances in space between the starting points of two trajectories, between the ending points, and between one or more intermediate check points, and returns the average of the distances. The functions differ in the way of choosing the check points:

- *k points by time*: the user-specified number of intermediate points *k* are selected so as to keep the time intervals between them approximately constant;
- *k points by distance*: *k* points are selected so as to keep the spatial distances between them approximately constant;
- *time steps*: the user specifies the desired temporal distance between the check points;
- *distance steps*: the user specifies the desired spatial distance between the check points.

2.2 Distances between events

We have two distance functions for events. The first one simply returns the distance in space between the spatial positions of the events. The second function, spatiotemporal distance, computes the distance in space and time. For this purpose, it asks the user for an additional parameter: the temporal distance threshold *maxT*, which is assumed to be equivalent to the spatial distance threshold *maxD*. The function finds the spatial distance *d* between the positions of two events and the temporal distance *t* between the times of their occurrence. Then it proportionally transforms *t* into an equivalent spatial distance *d′* and combines *d* and *d′* in a single distance according to the formula of the Euclidean distance.

2.3 Distances between spatial distributions

A spatial distribution is a complex object consisting of spatial positions of several or multiple events or moving entities. Our approach to measuring distances between such objects involves spatial aggregation of the distributions. We divide the underlying territory into spatial compartments using a suitable mesh, which may be regular (rectangular or hexagonal) or irregular. Then, the number of items (i.e. events or entities) in each compartment is computed for every spatial distribution. Hence, each distribution D^i becomes represented by a set of numeric values N_1^i, N_2^i, …, N_M^i, which are the numbers of the items in the compartments C_1, C_2, …, C_M (M is the total number of the compartments).

A straightforward approach to measuring the distance between two spatial distributions D^i and D^j is computing the arithmetic differences between the corresponding counts N_k^i and N_k^j, $1 \le k \le M$, and integrating these differences in a single value by averaging or by the formula for Manhattan or Euclidean distance. However, this approach is not valid. It deals with the spatial compartments as unrelated and does not take into account the spatial proximity between them. Thus, two items from two spatial distributions may be very close in space but occasionally fit in different neighboring compartments, which makes a large contribution to the computed distance between the distributions.

In order to deal with spatial distributions more adequately, our tool counts not only the items inside each compartment but also the items in the neighborhood of each compartment. The presence counts combining the interior and the neighborhood are used in computing the distances between the distributions instead of the simple presence counts N_k^i.

In our implementation, the neighborhood of a compartment C_k is defined as the set of all compartments having a common border segment with C_k. The neighborhood-inclusive presence counts η_k^i are the arithmetic sums of N_k^i and the number of items in the neighborhood of C_k, NN_k^i. Generally, other valid ways of defining the neighborhood and the presence counts are also possible. For instance, the neighborhood of a compartment C_k may be defined as a buffer zone with a chosen width around C_k. A more sophisticated way of counting the items in the neighborhood of C_k is weighting the contribution of each item to the combined count inversely to the spatial distance of the item to C_k, so that closer items contribute more than more distant items. However, the simpler approach works sufficiently well for our purposes.

The neighborhood-inclusive presence counts η_k^i are computed in somewhat different ways in case of events and in case of movement data. In case of events, the counts of items in the neighboring compartments are simply summed. In case of movement data, there is a probability that one and the same item occurs in two or more compartments because it moved during the time interval for which the spatial distribution is defined. In order to avoid counting any item more than once, the tool makes for each compartment and time interval a list of counted items and checks for each item whether it already occurs in the list.

In the following sections, we demonstrate how our generic clustering tool supplemented with the library of different distance functions can be applied to diverse types of data.

3. MINI-CHALLENGE "EVACUATION TRACES"

Clustering is especially helpful in analyzing large datasets. The dataset for the mini-challenge "Evacuation traces" is quite small as it contains only 82 trajectories. Cluster analysis is not really necessary for answering the questions of the mini-challenge. However, it can aptly complement purely visual and interactive techniques, as will be shown below, and the same or similar procedure will be applicable and effective in case of a much larger dataset. According to the focus of this paper, we shall not describe the whole analysis of the dataset but only demonstrate the use of the clustering techniques. A report about a complete analysis (done mostly with the use of other methods) is available

at http://vac.nist.gov/2008/entries/andrienkoevac/index.htm; see also a summary in [3].

3.1 Clustering of traces by "common fate"

The first question we try to answer concerns the fates of the people who were in the building before the explosion and could be affected by the incident: who managed to leave the building and who did not? To answer this question, we cluster the trajectories of the people using the distance function "common destination". After a few experiments with the distance threshold *maxD*, we obtain easily interpretable clusters, which are presented in Figures 1-3. The trajectories are represented by lines; the small hollow squares mark the starting points and the bigger filled squares mark the ending points. In Figure 1, there are four clusters of trajectories that evidently belong to people who managed to leave the building: the ending positions of the trajectories can be interpreted as being at the exits. The two clusters shown in Figure 2 consist of trajectories ending inside the building; hence, the people did not manage to evacuate because they were affected by the incident. In Figure 3, there are five trajectories that do not fit in any cluster. These trajectories need to be considered in detail: the terrorist or terrorists may be among the people who left these traces.

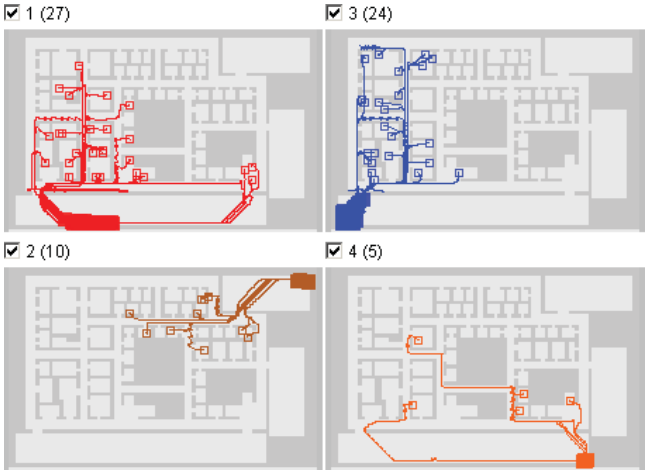

Figure 1. The clusters of the trajectories of the people who evidently managed to leave the building. This and all other figures are available in colors in the online pdf version of the paper.

Figure 2. The clusters of the trajectories of the possible casualties.

The clusters can be very conveniently used for dynamic filtering of the trajectories: the checkboxes above the images of the clusters hide or expose their members. Thus, we can select the clusters corresponding to the possible casualties and find out, with the help of the space-time cube [9][10] (Figure 4), that the people whose trajectories belong to cluster 5 (violet) stopped moving significantly earlier than the people from the second group (cluster 6, green). This means that the former group of people was closer to the place of the explosion than the latter group.

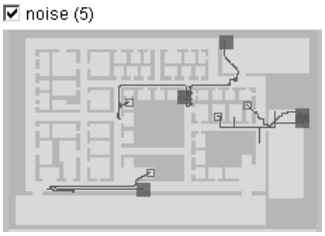

Figure 3. The trajectories that do not belong to any cluster.

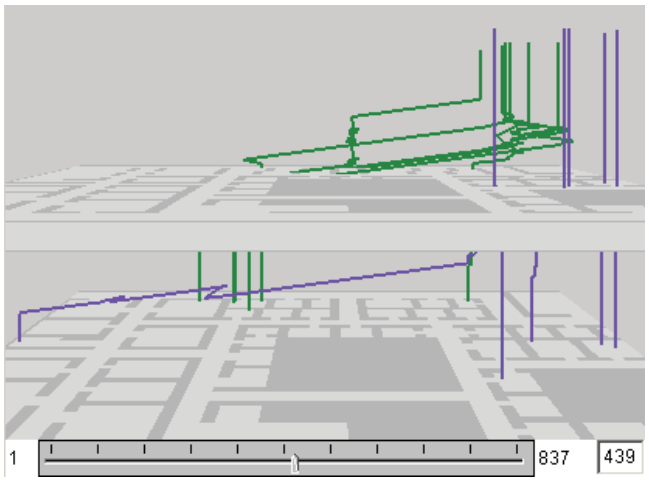

Figure 4. The space-time cube shows the trajectories of the possible casualties. The position of the movable horizontal plane corresponds to the time moment after which there was no movement in cluster 5 (violet).

Now we can select the group of people corresponding to the "noise" (Figure 3) and explore their behaviors looking, in particular, whether they visited the areas where the identified casualties stopped moving. We shall not describe this analysis here. The result is that we identify a person who visited the probable area of the explosion before the explosion occurred, a person who never moved or, possibly, left his RFID tag in his original place, a possible casualty who stopped moving later than the others, and a person who was close to the people from cluster 6 when they stopped moving.

3.2 Clustering of traces by similar routes

Now we want to check whether any of the people who left the building had unusual routes of the movement, which may indicate their possible participation in the incident. As in the previous case, we want to use clustering for the separation of "normal" routes from peculiar ones: the former will be grouped in clusters and the latter will be marked as noise. In our library of distance functions, we have a function "route similarity" [2][11], which measures the correspondence between the geometric shapes of two trajectories and the closeness of their spatial positions. This function appears suitable for our purposes. However, it does not find any clusters in this dataset due to a very high fluctuation of the positions in the trajectories. The fluctuation is visually

signified by zigzag shapes of the trajectory lines, as illustrated in Figure 5. According to the "route similarity" function, the two trajectories shown in Figure 5 are very distant from each other since the zigzags make a large contribution to the computed average distance between the positions. However, the trajectories appear very similar if the fluctuations are ignored. Hence, we need to use a distance function less sensitive to fluctuations.

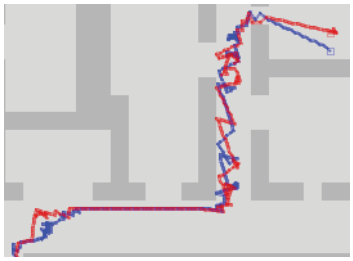

Figure 5. The fluctuations of the positions in the trajectories.

The family of distance functions "check points" can work in this case: if the number of check points is small, the impact of the fluctuations is also small. The functions "*k* points by time" and "time steps" do not suit well to our purposes: they are sensitive to the differences in the starting moments and the velocities of the movement whereas we want to consider only the routes. The function "distance steps" is not a good choice either: it is hard to select a suitable step because of a large variation of the lengths of the trajectories (from 0.5 to 189). The remaining function "*k* points by distance" works adequately. We find out that the results of the clustering do not substantially change when we vary the number of the intermediate check points (parameter *k*) in the range from 5 to 25.

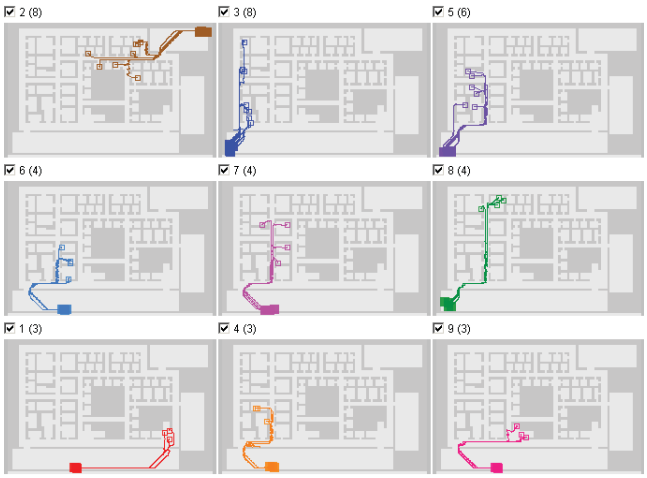

Figure 6. The trajectories of the people who left the building (see Figure 1) have been clustered according to the routes.

Figure 6 presents the clusters discovered among the trajectories of the people who left the building (Figure 1) with the use of the distance function "*k* points by distance" where *k*=15. Figure 7 shows the remaining 23 trajectories, which have not been put in clusters. We can say that the clusters correspond to normal, logical routes of the movement. The remaining trajectories with peculiar routes need to be additionally examined. However, there is no need in a detailed examination of each trajectory. It is sufficient to have a close look at the trajectories of the people who either visited the place of the explosion or interacted with some of the suspects or the victims. As can be seen in Figure 7, none of the uncommon trajectories passes the identified place of the explosion. Hence, we may focus on finding and examining possible interactions between the people who had these trajectories and the possible victims or suspects, whose trajectories are shown in Figures 2 and 3.

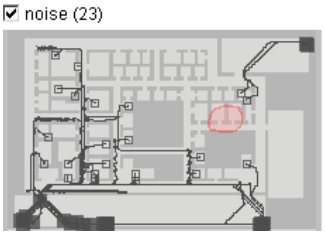

Figure 7. The trajectories not fitting in any cluster. The pink spot marks the identified area of the explosion.

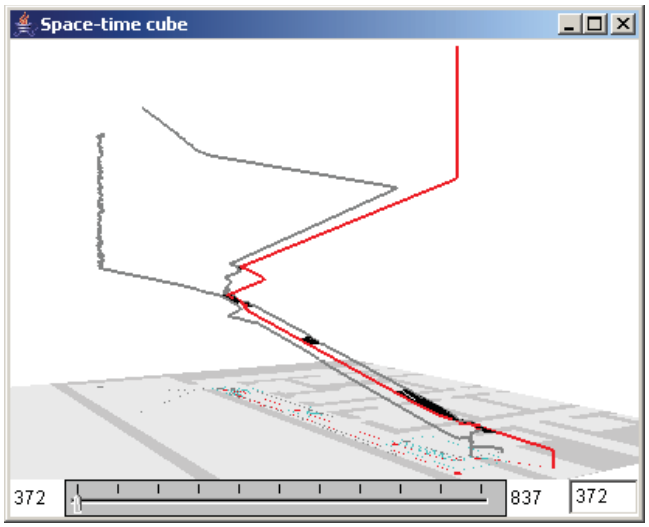

Figure 8. Black marks indicate the probable interactions between one of the possible casualties, whose trajectory is in red, and two other people.

We shall not describe the further analysis in detail. In brief, we applied our computational tool for finding probable interactions, i.e. cases of spatial proximity of moving agents. We found that only three of 23 people might have interactions with some of the victims or suspects. One of them was in the same room as one of the suspects till moment 262, when the latter left the room. The other two people might have interacted with the probable victim who stopped moving later than the other victims (Figure 8). In a case of a real investigation, it would be reasonable to interrogate these three persons.

In this example, we selected the suitable distance function and parameter values empirically, by running the clustering tool several times and looking at the results represented visually. We also use this approach when dealing with much larger datasets. On the first stage, we extract a manageable sample of the dataset and use it for experimenting with the distance functions and parameter values. Then, after finding suitable settings, we check them by applying to another sample. If the results are meaningful, we run the clustering with these settings for the whole dataset.

3.3 Clustering of spatial distributions

One of the questions of the mini-challenge was "Describe the evacuation". The evacuation is a process, i.e. a phenomenon developing in time. In order to describe a process, it is reasonable to try to divide the whole time span in which the process develops into periods of different behaviors such as relative stability (when no or minor changes occur), or some evolution trend, or intensive changes. Clustering can be quite helpful here. It should be applied to data characterizing the states of the process in different time moments or intervals. Clustering will group together similar states of the process and, hence, the corresponding moments or intervals. Successive moments or intervals belonging to the same cluster will signify a period of relative stability or gradual evolution. In periods of intensive changes, the time moments or intervals will be in the "noise" or in different clusters.

The states of the evacuation process can be characterized in terms of the spatial distributions of the people in the building. Hence, the clustering tool should be applied to the spatial distributions. In the experiment described below, we divide the building into compartments by generating a regular rectangular grid. We also divide the time span of the data into intervals of the length 5 units. The last interval starts at moment 836 and includes only two moments, 836 and 837. The counts of presence in the grid cells are computed for these 168 intervals. Four screenshots in Figure 9 show the grid, the positions of the people in the first and the last intervals (blue marks), and the simple and neighborhood-inclusive presence counts (top and bottom, respectively), which are represented by proportionally sized red circles; the largest circle stands for 24 people.

We apply the clustering tool with the distance function computing the average of the absolute differences between the neighborhood-inclusive presence counts. With the distance threshold *maxD*=0.2 and the density threshold *MinNbs*=3, the clustering tool finds four clusters of spatial distributions with the sizes 76, 3, 4, and 24; 61 distributions are treated as "noise".

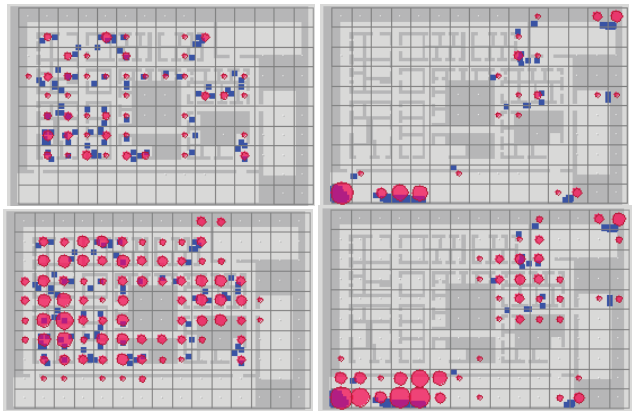

Figure 9. The positions of the people in the time intervals 1-5 (left) and 836-837 (right) and the presence counts for the grid cells, simple (top) and neighborhood-inclusive (bottom).

The table in Figure 10 gives an idea about the composition of the clusters and the relative positions of the corresponding time periods within the whole time span. The vertical dimension represents the time flow, from top to bottom. The first column represents the 168 time intervals, and the following columns correspond to the spatial compartments (i.e. grid cells). The neighborhood-inclusive presence counts for the cells and time intervals are represented by horizontal bars of proportional lengths. The identified clusters are represented by colors of the bars; the dark gray color corresponds to "noise".

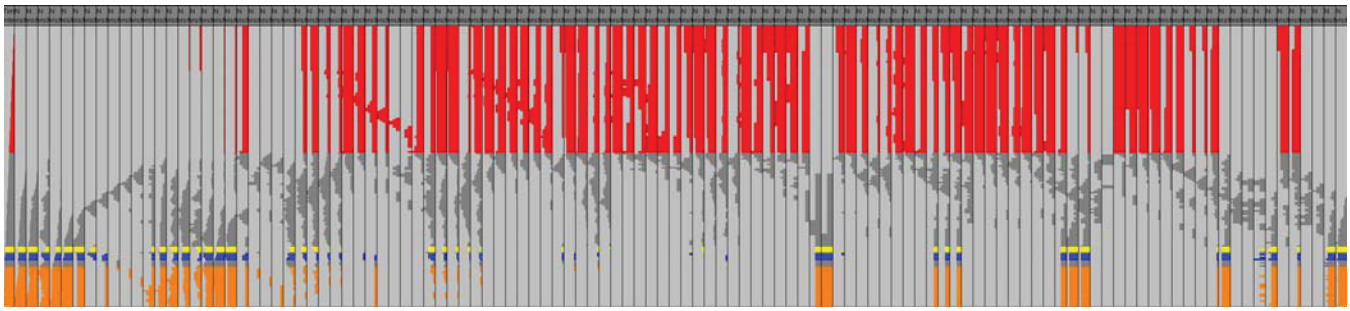

Figure 10. The table display represents in a condensed form the result of clustering the spatial distributions. The rows correspond to the chronologically ordered time intervals, which are represented in the first column. The columns starting from the second correspond to the grid cells. The lengths of the horizontal bars inside the cells are proportional to the neighborhood-inclusive presence counts. The colors of the bars correspond to the identified clusters (dark gray is used for the "noise").

The aggregated spatial distributions at the beginning and end of each of the four periods are shown in Figure 11. The circles represent the simple counts; the maximum circle stands for 23 people. The first cluster corresponds to the time period from interval 1-5 till interval 376-380. In this period most people stayed in the rooms, as can be seen from Figure 11 top. The following period from interval 381-386 till interval 656-660 makes a part of "noise". This is a period of intensive movement of the people towards the exits that began after the explosion. The spatial distributions of the people in the remaining three clusters coincide in the northeastern part, which means that there were no changes in this part after moment 661 except for minor movements at the exit. Some people were staying inside the building without movement. It may be concluded that the explosion occurred in this part of the building and that the people who remained inside the building were killed or injured.

Observable differences among the distributions in the three clusters are in the southwestern part of the building. A clear trend is visible in Figure 11: people move toward the southwestern exit.

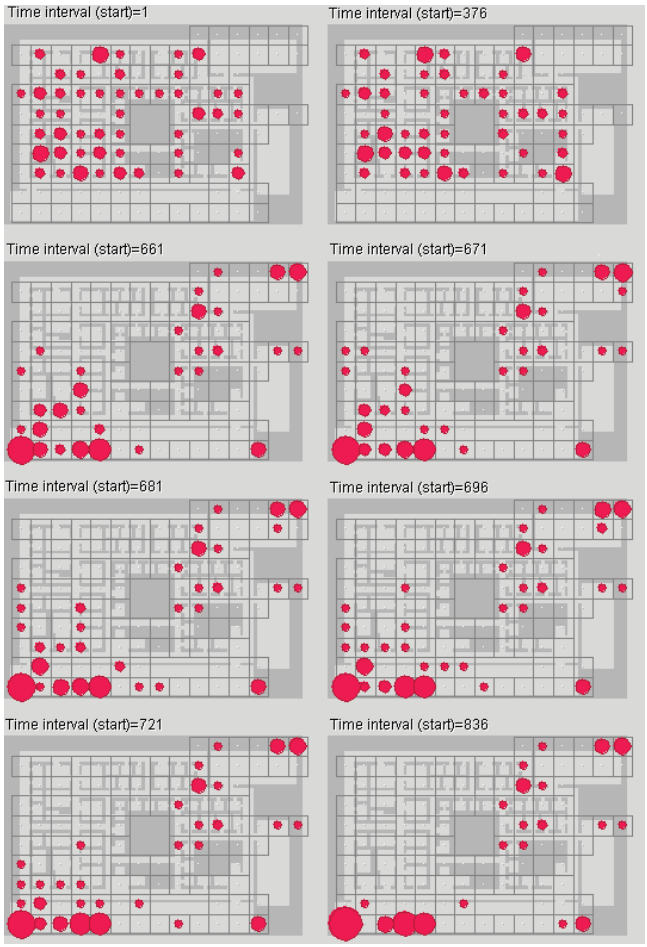

Figure 11. The aggregated spatial distributions of the people in the first and last time intervals of the identified periods of relative stability.

In the second experiment, we use the same spatial division but do not aggregate the data by time intervals. When we apply clustering with the same parameters as before to the counts by time moments, we obtain 16 clusters (most of which are quite small); 28 spatial distributions are treated as "noise".

The first cluster corresponds to the time period 1-382, which is quite consistent with the result of the first experiment. However, the last cluster now corresponds to the period 609-837. This shows that the changes in the spatial distribution from moment to moment are much smaller than from interval to interval, which is natural. Hence, the whole period 609-837 can be regarded as a period of gradual, evolutionary change, specifically, steady movement of the people (except for the casualties) toward the exits. When we gradually lower the distance threshold, the first cluster remains relatively stable while the last cluster significantly reduces. Thus, with the threshold value 1.1, the first cluster reduces to 1-379 while the last one to 679-837. Surprisingly, with the threshold 1.0, the first period is subdivided into two parts 1-76 and 79-379, with a few moments in between. We look at the detailed data and find that only one person was moving till moment 76 and in the moments 77-78 another person started moving. The further lowering of the threshold to 0.08 does not affect the initial period 1-379 (with its two parts) but reduces the last period to 736-837. After further manipulating the distance threshold, we conclude that three periods can be viewed as periods of relative stability, in terms of the spatial distribution: 1-76, 80-378, and 736-837. The initial period of the evacuation (about 379-608) is characterized by higher variability of the distribution than in the following period, when the changes were more coherent. A possible explanation is that the movement of the people was somewhat chaotic at the beginning of the evacuation but then became more coherent.

3.4 Conclusion

In the mini-challenge "Evacuation traces", the density-based clustering of trajectories was useful for several purposes. First, we divided people into groups according to their fates. Two of the groups were interpreted as probable casualties, the others as survivors. Second, we separated normal movement behaviors from peculiar ones. Such separation is possible owing to the specific feature of the density-based clustering, which does not put an object in a cluster if it is not sufficiently similar to others. The flexibility of the clustering tool allows us to choose distance functions according to the goals of the analysis. Third, we used clustering for dividing the time span of the data into periods according to the spatial positions of the people. In this analysis, clustering was applied to spatial distributions, a different type of data derived from the trajectory data. In the next section, clustering is applied to event data, detailed and aggregated.

4. MINI-CHALLENGE "MIGRANT BOATS"

The dataset for this mini-challenge consists of 917 records about landings and interdictions of migrant boats with the spatial positions (geographical coordinates) and times of the landing or interdiction events. The time span of the dataset is three years from the beginning of 2005 till the end of 2007. Among the questions of the mini-challenge, there are questions about the choice of the landing sites over the three years and about the geographic patterns of the interdictions over the three years. These questions may be answered with the help of clustering: using an appropriate distance function, we can discover spatiotemporal clusters of events, in particular, landings or interdictions in the same or close places shortly one after another.

4.1 Spatiotemporal clusters of landings

From the whole set of records, we select only the records about the landings. There are 441 such records. We apply the clustering tool with the distance function "spatiotemporal distance" described in Section 2. With 50 km as the spatial threshold and 21 days as the temporal threshold, we obtain the clusters shown in Figure 12 on a map and in a space-time cube (the use of space-time cube for visual exploration of event data is described in [4] and [7]). The scatterplot in Figure 13 aptly complements these two views. The horizontal and vertical dimensions of the plot represent the time and the latitude of the landings, respectively.

There are two big spatiotemporal clusters of landings located at the coast of Mexico. In the space-time cube, these two clusters appear as vertically aligned dots colored in orange and dark blue. In the scatterplot, the corresponding dots are aligned horizontally. The temporal extent of the orange cluster, which consists of 39 landings, is from April 15 till September 22, 2006. The dark blue

cluster consists of 146 landings, which occurred during the period from February 21 till November 18, 2007. Hence, both the number of landings at the Mexican coast and the duration of the period of active migration significantly increased from 2006 to 2007. As can be seen from the space-time cube and the scatterplot, there were no landings in this area before April 2006.

The spatiotemporal clusters of landings at the coast of Florida and nearby islands are much smaller. In 2005, there were 3 clusters of landings, shown in blue, yellow, and red (5, 9, and 9 landings, respectively); all of them occurred on the islands of the Florida Keys archipelago. In 2006, there were 4 clusters of landings on the Florida Keys islands (light blue, violet, green, and dark red; 26 events in total) and 3 clusters of landings on the western coast of Florida (light cyan, pink, and dark yellow; 16 events in total). In 2007 there was only one spatiotemporal cluster consisting of 6 landings. It is shown in brown; the landings occurred on the western coast of Florida. This may mean that the migrants changed the strategy and avoided repeated landings in the same areas in favor of more distributed targets. This may also mean that repeated attempts to reach the same place were intercepted by the coast guards.

4.2 Spatial clusters of landings

Another kind of analysis can be done by means of spatial clustering of the landing events irrespective of the time. For this purpose, we apply the distance function "spatial distance". With the distance threshold 25km, we obtain the spatial clusters of landings demonstrated in Figure 14 left. The temporal histogram in Figure 14 right shows us how the destinations of the migrants changed over the three years. The bars of the histogram correspond to the years; they are divided into colored segments proportionally to the numbers of landings from the corresponding clusters. We can see that almost all landings in 2005 occurred on the Florida Keys archipelago (red cluster). In 2006, additional destinations appear: at the Mexican coast (orange), on the western coast of Florida (violet, light blue, and dark gray), and at the western end of Florida Keys (pink and yellow). In 2007, the number of landings on Florida Keys significantly decreases while the number of landings in Mexico dramatically increases. Besides, there is an eastern trend: many migrants land on the eastern coast of Florida, which did not occur in the previous years.

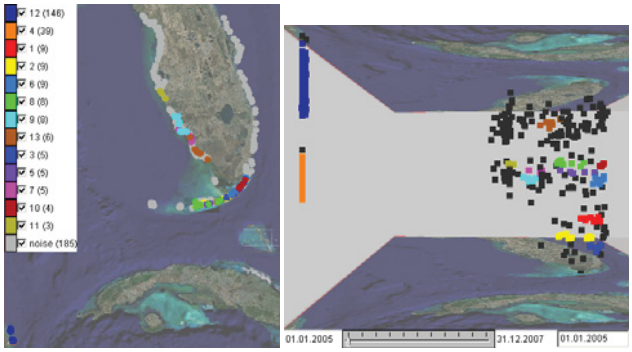

Figure 12. Spatiotemporal clusters of landings on a map (left) and in a space-time cube (right).

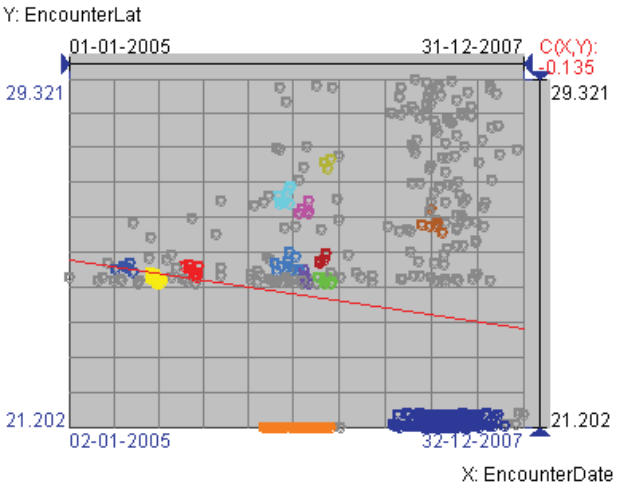

Figure 13. The clusters of landings shown on a scatterplot.

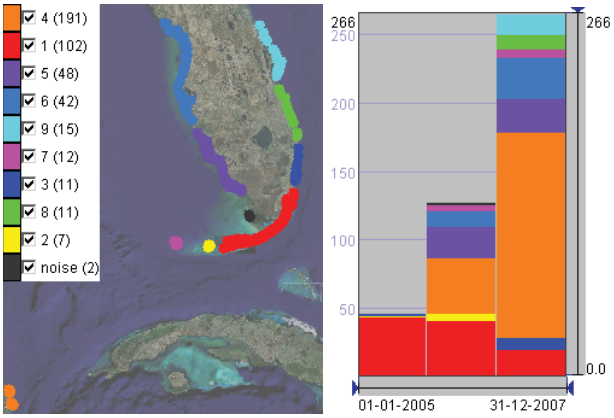

Figure 14. Left: spatial clusters of landings. Right: the distribution of the landings by years.

4.3 Clustering of the interdictions

Now we shall apply clustering to the interdiction events. In Figure 15, we see the spatiotemporal clusters discovered with the use of the distance function "spatiotemporal distance" (*maxD*=50 km; *maxT*=21 days). In Figure 16, we can see how the clusters and the remaining interdiction events ("noise") are distributed over the three years from 2005 to 2007. The temporal histogram in Figure 17 left shows us the sizes of the clusters and "noise" by years.

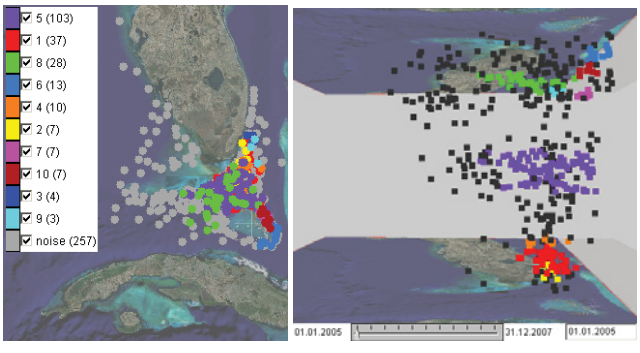

Figure 15. Spatiotemporal clusters of interdictions on a map (left) and in a space-time cube (right).

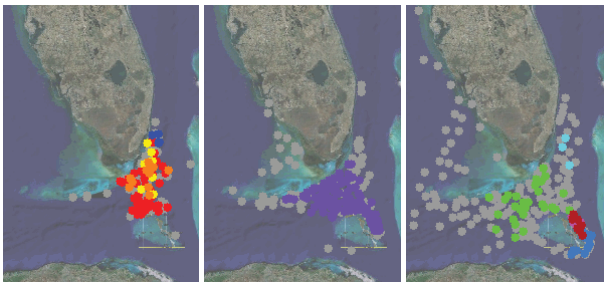

Figure 16. Spatiotemporal clusters of interdictions by years.

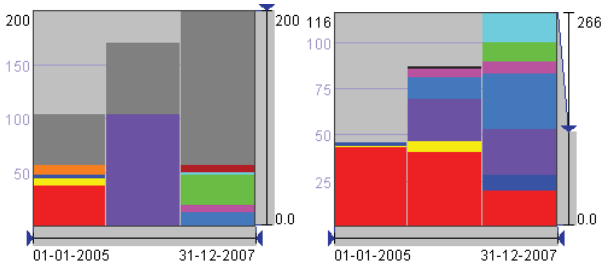

Figure 17. Left: the sizes of the clusters of the interdictions and the "noise" by years. Right: the landings in Florida and on nearby islands in the same years.

The spatiotemporal clusters of interdictions are generally larger than the spatiotemporal clusters of landings (Figure 12), except for the landings in Mexico. This refers not only to the number of events in a cluster but also to its spatial and temporal extent. The larger clusters mean that the interdiction events are spatially and temporally denser than the landing events. The highest spatiotemporal density of the interdictions is reached in 2006, when a single cluster (violet) includes 103 out of 170 events, i.e. over 60%. Like in 2005, the events are concentrated in the area between Florida Keys and Isla Del Sueño, the origin of the migrant trips; however, the spatial extent is larger in 2006. In 2007, the spatial spreading of the interdictions further increases while the spatiotemporal density of the events decreases. This is signified by the larger number of smaller clusters; the largest cluster (light green) is smaller and looser than the largest clusters in the previous years. The ratio between the number of events in the clusters and the size of the "noise" (58 to 142) is much smaller in 2007 than in 2006 (103 to 67) and 2005 (58 to 48).

When we compare these observations with the observations concerning the landings (Sections 4.1 and 4.2), we can conclude that the strategy of the migrants changed over the three years: the migrants diversified their destinations and, evidently, the routes. This, apparently, made the coast guards extend the area of patrolling. Probably, the migrants hoped that the change of the strategy would make them harder to catch and thereby increase the success rate. If we compare the number of landings in Florida and on the nearby islands (visualized on a histogram in Figure 17 right) with the number of interdictions by years, we may conclude that the success rate, indeed, steadily increased over the three years. The ratio between the number of landings and the number of interdictions was 46:106 (0.43) in 2005, 88:170 (0.51) in 2006, and 116:200 (0.58) in 2007. In 2006 and 2007 there were also 41 and 150 landings and no interdictions in Mexico.

For the landing events, we used spatial clustering irrespective of the time, which produced meaningful spatial clusters (Section 4.2). However, this method of clustering does not work well for the interdictions: due to the high spatial density of the events, most of them are united in a single very large cluster. This does not give us new opportunities for the analysis.

4.4 Clustering of spatial distributions

We are now interested to see whether there are repeated patterns of the spatial distribution of the events (landings and interdictions together) during time intervals. For this purpose, we shall apply clustering to spatial distributions of events computed for suitable time intervals. We limit our interest to the events that occurred around Florida, i.e. we exclude the landings in Mexico. We divide the territory into compartments by means of an irregular mesh, which is built according to the overall spatial distribution of the events (Figure 18 left). We divide the time span of the data into time intervals of the length 28 days, or 4 weeks. The suitable length of the intervals was chosen empirically. On the one hand, we believe that the movement of the migrants is very probable to be related to the weekly cycle. On the other hand, when we take weekly or bi-weekly intervals, many of them contain very few events. Figure 18 middle shows the events that occurred during one of the 4-weeks intervals. The map on the right presents the corresponding neighborhood-inclusive presence counts.

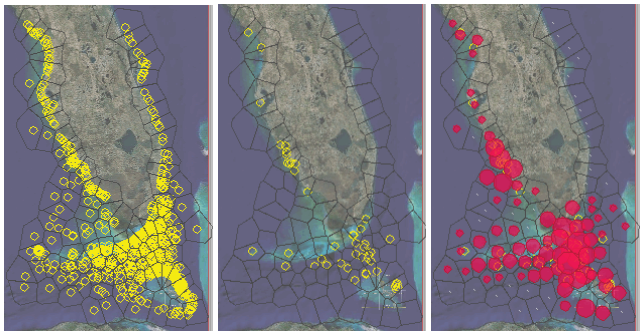

Figure 18. Left: the positions of all events that occurred around Florida (yellow circles) and the spatial compartments (dark grey polygons). Middle: the events that occurred during one of the time intervals. Right: the corresponding neighborhood-inclusive presence counts.

We apply the clustering tool to the resulting 50 spatial distributions using the distance function that combines the differences between the neighborhood-inclusive presence counts in the cells by means of the formula for the Euclidean distance. As usual, we determine the distance threshold experimentally. The clustering tool allows us to identify only one cluster of spatial distributions. With the distance threshold 0.56 we reach a sufficiently low internal variation in the cluster. The cluster consists of 10 spatial distributions with very small numbers of events; the corresponding time periods are 01.01.2005-28.01.2005, 31.12.2005-24.02.2006, 04.11.2006-23.02.2007, and 03.11.2007-31.12.2007, i.e. in winters.

We understand that one and the same distance threshold cannot be used to group both the distributions with few events and the distributions with many events. Therefore, we apply the procedure of "progressive clustering" mentioned in section 2.1. In this case, we apply clustering to the "noise" obtained in the previous step using the same distance function but a different distance threshold. For our set of spatial distributions we made

five steps of progressive clustering and obtained six clusters; two distributions remained in "noise". The results of the five steps are summarized in the table in Figure 19. The rows of the table are chronologically ordered; the colors in the first column, which contains the starting dates of the time intervals, correspond to the identified clusters.

	Clusters (OPTICS; 0.56/3)	Clusters (OPTICS; 0.85/2)	Clusters (OPTICS; 1.3/3)	Clusters (OPTICS; 1.5/3)	Clusters (OPTICS; 1.95/2)	Classes by OPTICS; 5 levels
01.01.2005	1					1.1 (0.56)
29.01.2005	noise	1				2.1 (0.82)
26.02.2005	noise	2				2.2 (0.82)
26.03.2005	noise	2				2.2 (0.82)
23.04.2005	noise	noise	noise	noise	noise	noise
21.05.2005	noise	2				2.2 (0.82)
18.06.2005	noise	2				2.2 (0.82)
16.07.2005	noise	noise	noise	1		4.1 (1.5)
13.08.2005	noise	2				2.2 (0.82)
10.09.2005	noise	1				2.1 (0.82)
08.10.2005	noise	noise	noise	1		4.1 (1.5)
05.11.2005	noise	1				2.1 (0.82)
03.12.2005	noise	1				2.1 (0.82)
31.12.2005	1					1.1 (0.56)
28.01.2006	1					1.1 (0.56)
25.02.2006	noise	noise	1			3.1 (1.3)
25.03.2006	noise	noise	noise	1		4.1 (1.5)
22.04.2006	noise	noise	noise	1		4.1 (1.5)
20.05.2006	noise	noise	noise	noise	noise	noise
17.06.2006	noise	noise	noise	noise	1	5.1 (1.95/2)
15.07.2006	noise	noise	noise	noise	1	5.1 (1.95/2)
12.08.2006	noise	noise	noise	1		4.1 (1.5)
09.09.2006	noise	noise	noise	1		4.1 (1.5)
07.10.2006	noise	1				2.1 (0.82)
04.11.2006	1					1.1 (0.56)
02.12.2006	1					1.1 (0.56)
30.12.2006	1					1.1 (0.56)
27.01.2007	1					1.1 (0.56)
24.02.2007	noise	noise	1			3.1 (1.3)
24.03.2007	noise	noise	1			3.1 (1.3)
21.04.2007	noise	noise	1			3.1 (1.3)
19.05.2007	noise	noise	noise	noise	1	5.1 (1.95/2)
16.06.2007	noise	noise	noise	noise	1	5.1 (1.95/2)
14.07.2007	noise	noise	noise	noise	1	5.1 (1.95/2)
11.08.2007	noise	noise	noise	noise	1	5.1 (1.95/2)
08.09.2007	noise	noise	1			3.1 (1.3)
06.10.2007	noise	noise	1			3.1 (1.3)
03.11.2007	1					1.1 (0.56)
01.12.2007	1					1.1 (0.56)
29.12.2007	1					1.1 (0.56)

Figure 19. Results of five steps of the progressive clustering of the spatial distributions.

Figure 20 gives an example of spatial distributions included in one cluster. This is cluster 5.1 discovered in the fifth step of the progressive clustering. It includes two adjoining time intervals from the summer of 2006 and four consecutive intervals from the late spring and summer of 2007. The distributions are, indeed, quite similar while it is also easy to notice that in 2007 there were additional events on the east and on the northwest.

In Figure 21, we have computed the average presence counts in the spatial compartments (without the neighborhood) for the six discovered clusters. The counts are represented by circles of proportional sizes. Hence, each map gives us a summarized picture of the spatial distributions in the clusters and allows us to understand the main features of the clusters and the differences between them.

4.5 Conclusion

In the mini-challenge "Migrant boats", the density-based clustering helped us to detect compact groups of events in space and time, to asses the spatiotemporal density of the events and its change over time, and to divide events into groups according to their spatial positions in order to examine the changes in the spatial distribution of the events over time. We also explored the variation of the spatial distribution patterns over time in another way, by applying clustering to spatial distributions of events in time intervals.

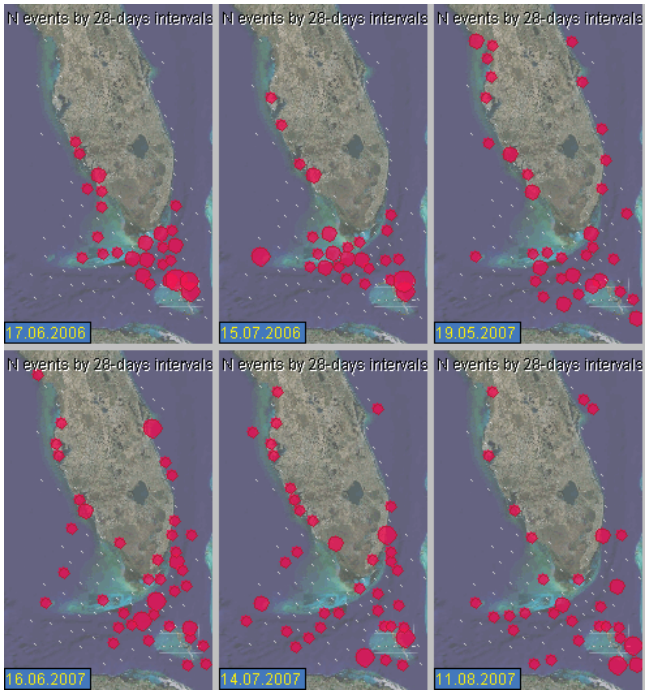

Figure 20. The spatial distributions comprising cluster 5.1 are shown in a summarized form: the circles represent the simple presence counts; the largest circle stands for 4 events.

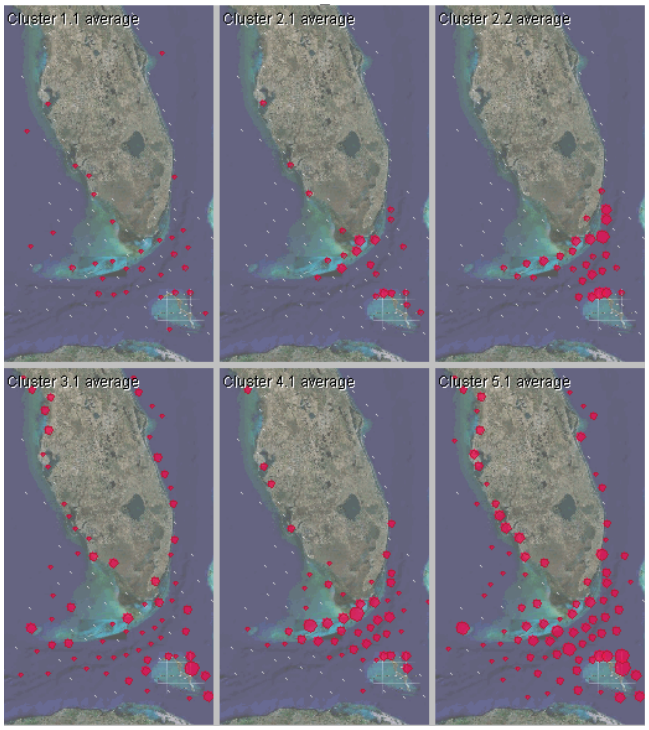

Figure 21. The average presence counts for the six clusters; the largest circle stands for 2 events.

5. GENERAL CONCLUSION

Clustering in combination with interactive visual displays is a powerful instrument of data analysis, in particular, when the data are large and/or complex. Many clustering methods require the data to be represented as points in a multi-dimensional space of properties (in other terms, by feature vectors). However, for complex data with multiple heterogeneous properties there may be no adequate representation by feature vectors. An example of such a complex data type is trajectories of moving objects, characterized by the origin and destination, length, temporal extent, duration, geometrical shape, spatial orientation, dynamics (distribution of the speeds along the way), and, possibly, variation of other attributes during the movement.

A possible approach to the clustering of complex data types is the use of a generic clustering algorithm with a type-specific distance function, which properly accounts for the relevant properties depending on their nature. We have demonstrated this approach by applying the same clustering algorithm to data of different types: trajectories of moving objects, events distributed in space and time, and spatial distributions of moving entities or events. We have also demonstrated that different distance functions oriented to the same type of data may be useful for different analysis tasks.

The clustering tool we use implements a density-based clustering algorithm, which does not strive to put each object in some cluster but finds compact groups of close (similar) objects and leaves the other objects ungrouped. In this way, it not only discovers frequent patterns (combinations of properties) but also enables the analyst to examine the variation of the data density (in terms of close properties) throughout the dataset. In the paper, we have demonstrated how the features of the algorithm are exploited in the analysis.

The VAST Challenge benchmark datasets [8] we have used in this paper are quite small; they could be effectively analyzed without the use of clustering. For larger datasets, clustering gives more significant advantages. Our clustering-based visual analytics tools work well with about 5,000 trajectories, i.e. the reaction time is appropriate for an interactive analysis. Clustering of 10,000 trajectories is possible but requires some patience. The paper [11] describes an analysis scenario with applying the density-based clustering tool to about 4,500 trajectories.

Currently we continue our research related to clustering in two major directions. First, we extend the approach to other types of spatiotemporal data, in particular, interactions between moving objects (mentioned in Section 3.2) and spatially referenced time series data. Second, we look for ways to increase the scalability of clustering with respect to the size of the data. Thus, we have recently devised a visual analytics method for extracting clusters from a dataset not fitting in the computer main memory [1].

6. ACKNOWLEDGMENTS

The work was done partly within the EU-funded research project GeoPKDD – Geographic Privacy-aware Knowledge Discovery and Delivery (IST-6FP-014915; http://www.geopkdd.eu) and partly within the research project ViAMoD – Visual Spatiotemporal Pattern Analysis of Movement and Event Data, which is funded by DFG – Deutsche Forschungsgemeinschaft (German Research Foundation) within the Priority Research Programme "Scalable Visual Analytics" (SPP 1335).

The work on interactive cluster analysis of trajectories was done together with our GeoPKDD partners from the University of Pisa, Italy. We are grateful to them for the cooperation and specially thank Salvatore Rinzivillo for the implementation of the clustering algorithm OPTICS in the way allowing the use of different distance functions.

7. REFERENCES

[1] Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., Giannotti, F. 2009. Interactive Visual Clustering of Large Collections of Trajectories. In *Proc. VAST 2009*, IEEE Computer Society Press, 3-10.

[2] Andrienko, G., Andrienko, N., and Wrobel, S. 2007. Visual Analytics Tools for Analysis of Movement Data. *ACM SIGKDD Explorations*, 9(2): 38-46.

[3] Andrienko, N., and Andrienko, G. 2008. Evacuation Trace Mini Challenge Award: Tool Integration. Analysis of Movements with Geospatial Visual Analytics Toolkit. In *Proc. VAST 2008*, IEEE Computer Society Press, 205-206.

[4] Andrienko, N., Andrienko, G., and Gatalsky, P. 2003. Exploratory Spatiotemporal Visualization: an Analytical Review. *Journal of Visual Languages and Computing*, 14 (6), 503-541

[5] Ankerst, M., Breunig, M., Kriegel, H.-P., and Sander, J. 1999. OPTICS: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD* 1999, 49–60.

[6] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. ACM KDD* 1996, 226-231.

[7] Gatalsky, P., Andrienko, N., and Andrienko, G. 2004. Interactive Analysis of Event Data using Space-Time Cube. In Banissi, E. et al. (Eds.) *Proc. IV 2004 - 8th International Conference on Information Visualization*, July 2004, London, UK, 145-152

[8] Grinstein, G, Plaisant, C., O'connell, T., Laskowski, S. Scholtz, J., Whiting, M. VAST 2008 Challenge: Introducing Mini-Challenges, In *Proc. VAST 2008*

[9] Hägerstrand, T. 1970. What about people in regional science? In: *Papers of the Regional Science Association*, 24, 7-21.

[10] Kraak, M.-J. 2003. The space-time cube revisited from a geovisualization perspective, in: *Proc. 21st International Cartographic Conference*, Durban, South Africa, August 2003, 1988-1995.

[11] Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., and Andrienko, G. 2008. Visually–driven analysis of movement data by progressive clustering, *Information Visualization*, 7(3/4), 2008, 225-239.