Identifying Place Histories from Activity Traces with an Eye to Parameter Impact

Gennady Andrienko, Natalia Andrienko, Martin Mladenov, Michael Mock, Christian Pölitz

Abstract— Events that happened in the past are important for understanding the ongoing processes, predicting future developments, and making informed decisions. Important and/or interesting events tend to attract many people. Some people leave traces of their attendance in the form of computer-processable data, such as records in the databases of mobile phone operators or photos on photo sharing web sites. We developed a suite of visual analytics methods for reconstructing past events from these activity traces. Our tools combine geocomputations, interactive geovisualizations and statistical methods to enable integrated analysis of the spatial, temporal, and thematic components of the data, including numeric attributes and texts. We also support interactive investigation of the sensitivity of the analysis results to the parameters used in the computations. For this purpose, statistical summaries of computation results obtained with different combinations of parameter values are visualized in a way facilitating comparisons. We demonstrate the utility of our approach on two large real data sets, mobile phone calls in Milano during 9 days and flickr photos made on British Isles during 5 years.

Keywords — Event detection, spatio-temporal data, time series analysis, scalable visualization, geovisualization, visual analytics, sensitivity analysis, scale effect

1 INTRODUCTION

In 16th century Francis Bacon wrote that historia (history) is "the knowledge of objects determined by space and time" (http://en.wikipedia.org/wiki/History). An essential part of the history of a place is the events (i.e., noteworthy happenings) that occurred there. Important and/or interesting events usually attract many people, active participants and/or spectators. Nowadays, some of these people may leave traces of their presence in electronic databases. Examples are records about mobile phone calls, georeferenced photos on photo sharing web sites, and georeferenced messages. Many people voluntarily make their data accessible to others via the Web. Goodchild [12] considers citizens as sensors collecting valuable geographical information. By analyzing data related to presence of people in different places, one can discover interesting facts from the modern history of the places. We call such data 'activity traces' or 'activity data', where 'activity' denotes various actions (phone calling, photo taking, message posting, etc.) or movement.

In a general case, activity records include identifiers of people (id), geographic coordinates (x,y), time reference (t) and some attributes. Examples of attributes are the text of a spatially-referenced message, the duration of a mobile phone call and the change of the caller's position during the call, the title and text tags of a photo, etc.

Reconstructing interesting events from activity data is a difficult task. The amount of the data is typically very large; it excludes the possibilities for processing in main memory. The data typically refer to points in space specified by geographic coordinates whereas interesting places are usually areas capable to contain many people rather than just points. Moreover, the definition of a place depends on the intended spatial scale of analysis, e.g. a country, a region, a city, or a building. Similar complexities exist for the temporal dimension. It is necessary to consider the events in the context formed by geography, time, and other events and processes.

We present a visual analytics procedure for detecting and reconstructing events from activity data. It includes division of the territory into areas according to the intended spatial scale of analysis, aggregation of the data into time series, checking temporal correlation within the time series and detecting events in them, and interactive visual analysis and interpretation of the results. We extend our previous work [2] by methods for analyzing the sensitivity of the results to parameters of the procedure.

We illustrate the procedure by applying it to two real datasets. A dataset with the positions of 2,956,739 phone calls made in Milan (Italy) during 9 days 30.10.2008 - 07.11.2008 has been provided by the Italian telecommunication company WIND. A dataset with the positions, temporal references, and titles of 85,041,956 flickr.com photos (mostly taken after January 2005) has been collected by our project partners from University Konstanz. In this paper, we use the subset of the data covering the territory of the UK and Ireland, 8,686,034 records in total.

2 RELATED WORK

A series of works of the MIT group (e.g., [11] and [23]) considers data about mobile phone calls and georeferenced photos as a reflection of people's activity in a city. To demonstrate how known events are reflected in the data, they visualize the spatial distribution of the phone

• *Gennady Andrienko, Natalia Andrienko, Martin Mladenov, Michael Mock and Christian Poelitz are with Fraunhofer IAIS (Intelligent Analysis and Information Systems), Sankt Augustin, Germany. E-mail: gennady.andrienko@iais.fraunhofer.de.*

Manuscript received 15 January 2011

calls or photos during a given time interval by a heat map or density surface. Despite interesting applications, these approaches do not provide sufficient support for finding and interpreting previously unknown events from the past. Of course, an analyst can view an animated density map showing changes over time; however, the effectiveness of animation is limited [25], particularly, due to the perceptual phenomenon known as "change blindness". Hence, it is necessary to combine space-centric approaches with effective techniques for temporal analysis.

Data about flickr photos are used in [20] to present a density-based algorithm for spatial clustering, which does not take into account the temporal references of the photos. There are quite many other papers where data about georeferenced photos are used for presenting particular methods but not for analyzing place histories. Our previous paper [17] explores the potential of such data for gaining information about a region, interests of people, and patterns of their movement. While it uses spatio-temporal aggregation of the data, it does not apply any computational techniques for time series analysis.

Paper [16] considers data about indoor movement collected by motion sensors statically placed inside a building. The data are time series of the sensor activation counts. The authors explore the use of the space over time and detect periods of intensive movement.

Analysis of time series has been in focus of the information visualization researchers for a long time. A calendar display showing similarity of daily profiles of energy consumption is suggested in [26]. A number of papers suggest advanced functionality for a time series graph. TimeSearcher [15] enables interactive querying of time series by their shapes. Paper [3] suggests approaches to representing multiple time series in a summarized form and interactive techniques for detecting sequential increases or decreases of attribute values. A time graph may be combined with a time band where summaries, such as overall averages, or predicted values are represented by coloring or shading [14]. In recent versions of TimeSearcher [7][8], temporal positions of specific features of time series can be marked on the time graph.

Detection of features in time series is a research topic in data mining and statistics. Methods have been proposed for finding specific patterns (motifs) [27], change points [5][18], and periodic patterns [9][24]. In principle, all these methods can be integrated in visual analytics procedures; however, even simpler algorithms combined with interactive visual interfaces may yield high analytical power. This is demonstrated in our current paper, where the focus is not on devising sophisticated algorithms or original visualizations but on combining visual and computational methods so as to benefit from their synergies. The novel contributions of the paper are:

1. A method to transform activity traces (points in space and time) into spatially referenced time series by means of spatio-temporal aggregation based on user-controlled division of the study area into compartments reflecting the spatial density of the data.
2. Synergistic integration of algorithmic methods for detection of peaks/pits and periodicity in time series with interactive visual displays.
3. Interactive visual techniques supporting investigation of the sensitivity of the analysis results to the parameters of the algorithmic methods.
4. Tools for on-demand acquisition of contextual information to enable pattern interpretation.
5. Defining, on the basis of 1-4, a scalable visual analytics procedure for analyzing large sets of point events.

3 Visual analytics procedure

Following the "Visual Analytics Mantra" [19], we start with computational extraction of places (areas) from the data. Using the areas, we aggregate the data into spatial time series. The time series are explored by a combination of interactive visualizations and statistical computational methods. Additional data are acquired on demand for supporting interpretation of the detected features and reconstructing the events that caused them.

All steps of the analytical procedure rely on database processing as the amount of data does not allow complete loading to the main memory. Space tessellation is done on the basis of a data sample. Aggregation can be done either in the database or in the main memory with incremental processing. Additional attributes are loaded from the database on demand. Such architecture allows scaleable processing and analysis of very large data sets.

3.1 Territory tessellation

Space tessellation enables aggregation of point-based data, which is essential for dealing with large datasets. Very often arbitrary territory divisions are used, such as administrative districts or regular grids. Such divisions do not reflect the spatial distribution of the data. It is more appropriate to define space compartments so that they enclose existing spatial clusters of points. However, these clusters may have very different sizes and shapes, which has two disadvantages. First, it is computationally hard to automatically divide a territory into arbitrarily shaped areas enclosing clusters. Second, the areas would differ much in their sizes, and the respective counts would be incomparable. Therefore, we have developed a method that divides a territory into convex polygons of approximately equal sizes on the basis of point distribution [4]. Our algorithm looks for spatial clusters of points that can be enclosed by circles with a user-chosen radius. A concentration of points having a larger size and/or complex shape will be divided into several clusters. The centroids of the clusters are then used as generating points for Voronoi polygons. The centroids are the points with the minimal average distance to the cluster members. They are usually located inside concentrations of points. In data about people's activities, cluster centroids most often indicate the foci of people's attention.

For the tessellation, a sufficiently large sample of the data from the database is loaded in the main memory. To be sure that the spatial distribution properties of the whole dataset are well reflected in the sample, we suggest combining several samples taken from different time intervals. The tessellation method is very efficient:

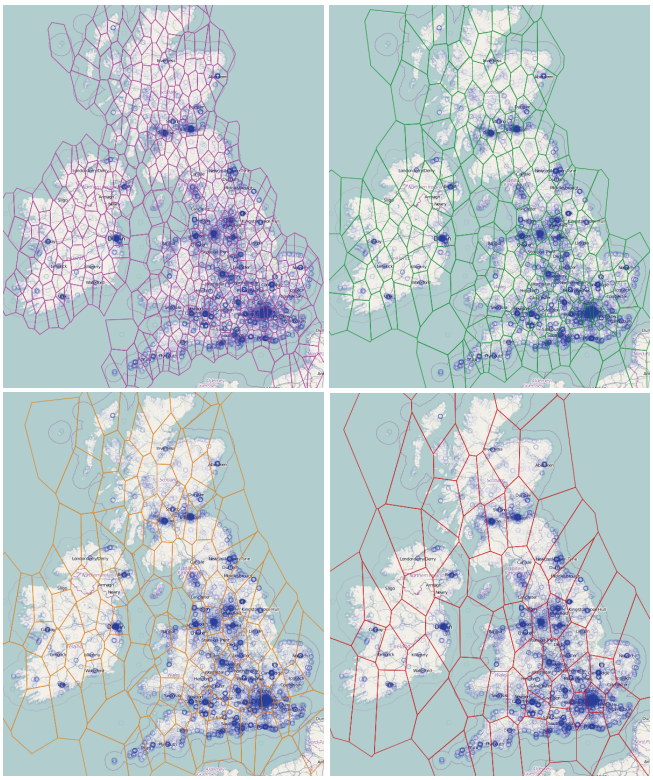

Fig. 1. The territory tessellation has been applied to a sample of the flickr data. Depending on the value of the parameter 'cluster radius' (here 30, 50, 70, and 100km), the divisions differ in the spatial scale.

processing of a 20,000 points sample takes about 3-5 seconds on a standard PC. More strictly, the computational complexity of the point clustering algorithm is $O(n)$, where n is the number of points [4]. The subsequent Voronoi tessellation can be done in $O(k \log k)$ time [10], where k is the number of cluster centroids.

In Figure 1, a random sample of about 23,500 flickr photo records from the territory of UK and Ireland has been used for dividing the territory. The spatial positions of the records are shown on the maps by small circles drawn in dark blue with 5% opacity so that the concentrations of points are easily observable. Four different divisions have been obtained for the values 30, 50, 70, and 100 km of the parameter 'cluster radius'. Owing to the properties of the spatial clustering algorithm, the areas in each division tend to enclose dense concentrations of points.

Our approach may be criticized for using a static (time-invariant) territory division: applying the method to data subsets from different time intervals may result in different tessellations, and static partitioning may hide significant temporal dynamics of clusters. An advantage of a static division is that it enables data aggregation and subsequent analysis of changes over time. An alternative approach to analyzing activity data could be to find and interpret spatio-temporal clusters of points without prior aggregation; however, there are currently no clustering methods that could do this for millions of points in reasonable time. Besides, periodic temporal patterns and events re-occurring in the same places would be more difficult to detect than by our approach.

Another point that may raise questions is the use of a fixed radius for obtaining point clusters. An important advantage of this is that the user can choose a suitable spatial scale depending on the size of the territory and the analysis goals, and also do the analysis at different scales. An apparent disadvantage is that big real clusters can be arbitrarily divided into smaller subclusters. However, preserving the sizes and shapes of real clusters is not essential for the intended way of the further analysis. Moreover, this can be even counterproductive. Thus, almost any large city has a dense cluster of flickr photos covering a large area in the center. Using this area without division would result in too much aggregation and in missing interesting localized events. Our method, which preserves small clusters and divides large ones, is thus adequate to the purpose of the aggregation.

3.2 Spatio-temporal aggregation

Depending on the goals of the analysis, the user selects the time period of interest and divides it into suitable intervals. For the areas of the territory division and the time intervals, the system computes two measures:

1. Number of different people who visited the areas in each interval.
2. Count of the activities that occurred in the areas in each interval (e.g. count of the photos taken).

The first measure indicates the attractiveness of the areas while the second measure represents the activeness of the people in the area. For each measure, the system generates a set of spatial time series $<a_i, t_i, v_i>$, where a_i identifies the area, t_i is the time interval, and v_i is the value of the measure in this interval.

3.3 Time series analysis

Depending on the size of the area under study and the chosen spatial scale, the aggregation procedure may result in hundreds or even thousands of time series, the length of which depends on the length of the time period and the chosen temporal resolution. Such amount of data may be excessive for purely visual analysis. We have devised two computational methods described below.

1. Periodicity (temporal correlation) detection

To test whether specific periods are present in the data, we take the maximum of the (circular) cross-correlation function $\mathrm{ccf}(\tau, x)$ of a time series x and a synthetic test pattern τ generated for a chosen period length T, further referred to as 'target period'. We interpret the value obtained as the periodicity score. The test pattern is a sum of Gaussian functions, which are offset from each other by the target period T:

$$\tau[t] = \sum_{k=0}^{\kappa} \exp\left(\frac{-(t-kT)^2}{\sigma_\tau^2}\right)$$

where $k = |x|/T - 1$ and $|x|$ is the length of the time series x. The ccf can be computed by means of the Fast Fourier Transform in time $O(|x| \log |x|)$

Additionally to $\mathrm{ccf}(\tau, x)$, we compute a normalized periodicity score as $(1/(|x|-1) \cdot \mathrm{ccf}(\tau, x)/\sigma_x\sigma_\tau)$. A high value of the normalized score indicates that the time series is similar to the test pattern regardless of the scale of the data, while the non-normalized score gives more weight

to time series with higher sample variance. Both scores are useful in the analysis process. Time series characterized by extreme values of these measures, either maximal or minimal, deserve special attention.

2. Peak/pit detection

The algorithm, which is a modified version of [6], finds abrupt peaks (increase followed by decrease) or pits (decrease followed by increase) within a given *time window*, i.e., time interval of a given length. The length is measured as the number of time steps in the time series. The algorithm has two parameters: minimum amplitude δ and maximum time window length w (we assume for simplicity that w is even). It identifies a sample x[n] of the time series as a peak if it is a local maximum in the interval w_n:=[n-w/2, n+w/2], and there are samples of value less than or equal to x[n]-δ both before and after x[n] within w_n. A sample is a pit if it is a local minimum in w_n and there are samples of value at least δ+x[n] around it. The algorithm outputs the amplitude of the peak/pit and the sample number n:

$$A^{POS}[n] := x[n] - \min_{k \in w_n}(x[k]) \text{ and } A^{NEG}[n] := \max_{k \in w_n}(x[k]) - x[n]$$

The pseudo code given below includes only the part of the algorithm that detects peaks.

Peak detection algorithm

Given: time series x, minimal amplitude δ, time window length w

Output: {(Apos[n], n) | for all x[n] peaks}

```
1   maxX←∞; minX←∞; maxpos←0; minpos←0;
    lookformax←true
2   for (n in 1 to |x|)
3     current ← x[n]
4     if current > maxX: maxX ← current; maxPos ← n endif
5     if current < minX: minX ← current; minPos ← n endif
6     if (lookformax = true)
7       if (curr < maxX-δ)
8         if ∃ x ∈ {x[maxPos-w/2], ..., x[maxPos]} (x < maxX-δ) ∧
9           ∃ x ∈ {x[maxPos], ..., x[maxPos +w/2]} (x < maxX-δ)
10            output(A^pos [maxPos], maxPos)
11        endif
12        minX ← curr; minPos ← n; lookformax ← false
13      endif
14    endif
15    else if (lookformax = false)
16      if (curr > minX + δ):
17        maxX ← curr; maxpos ← n; lookformax ← true
18      endif
19    endif
20  endfor
```

The algorithm needs only one pass through the time series. However, in order to determine whether a local maximum maxX is a peak, the algorithm goes through all samples in the time window (in lines 8-9) to verify that the definition given above holds. Therefore, the algorithm complexity is O(w | x |)

A modified version of the algorithm allows detection of peaks or pits from normalized values. The values within each time series are transformed to z-scores, i.e. the deviations from the mean of the time series divided by the standard deviation. Respectively, the value of the parameter 'minimal amplitude' is specified as the number of the standard deviations. This modification allows detection of interesting peaks or pits from time series with relatively low values in comparison to other time series; thereby, the scale effect can be explored (see section 6.3).

We shall use the term 'time-series event' or 't-event' to denote peak or pit or, more generally, any kind of abrupt change that may occur in time series. We shall also use the terms 'peak event' and 'pit event' denoting particular types of t-events. For data about presence and/or activities of people, t-events may indicate events that occurred in the real world. Each t-event refers to a certain region and a certain time interval. To find out what real events stand behind the t-events, the user may request *additional contextual information* from the activity records fitting in these regions and intervals. The system computes aggregate values of selected attributes and associates the values with the t-events. For numeric attributes, possible aggregates are, for instance, the average and median values, or the distribution histogram. For texts, the aggregates may represent the most frequent words and word sequences.

3.4 Interactive visual displays

The analytical procedure is supported by a set of interactive visual displays. The time graph (figure 2 top) gives an initial overview of the data[1]. In its usual form, the display suffers from overplotting. We suggest a statistical summary display that shows the average and/or median line, the envelope of all time lines, and the positions of the deciles or other quantiles for all time moments connected by lines (figure 2 bottom, described in [1]). Both variants of the display support interactive operations:

- zooming in the temporal and attribute dimensions;
- brushing that links the graph to other displays such as maps, histograms, scatter plots, and parallel coordinates;
- data transformation by arithmetic functions, normalization, smoothing, calculation of changes, etc.

The peak/pit detection algorithm produces t-events positioned in time and space. The spatial positions are shown on the map and the temporal positions on the time graph. The time graph view additionally contains a linear event bar (figure 3 top, below the plot) – a sequence of rectangles that show the counts of t-events for the time moments by the darkness of shading, darker is more. A display called 'periodicity chart' shows the distribution of the t-events with respect to temporal cycles. Thus, the chart in figure 3 (bottom right) shows the counts of t-events for 24 hours of the day over 9 days. Each row corresponds to one day and each column to one hourly interval of the day. Figure 14 shows the counts of t-events by weekly intervals over 5 years. Each row represents one year. The rows differ in lengths according to the different number of weeks in these years. The vertical bar on the right of the display represents the totals for the rows, i.e. for the days in figure 4 and for the years in figure 14. The horizontal bar in the bottom represents the totals for the columns, i.e. for the same hours of different days in figure 4 and for the same weeks of different years in figure 14.

The space-time cube display (figure 5) shows the positions of t-events in space and in time using two horizontal

[1] The illustrations in this section refer to the Milan calls data.

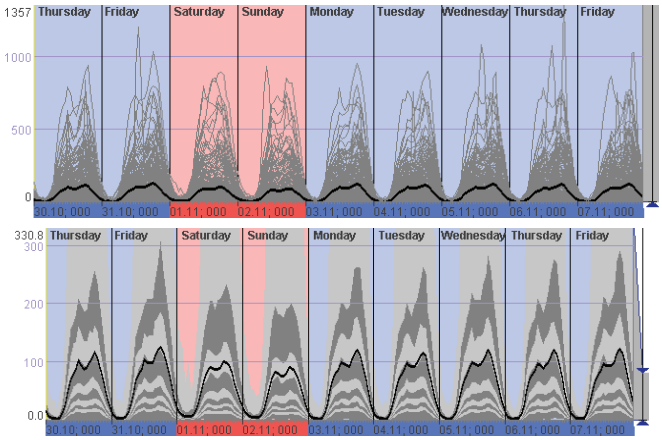

Fig. 2. A traditional time graph display (top) and a statistical summary (bottom). The weekend is marked in light red.

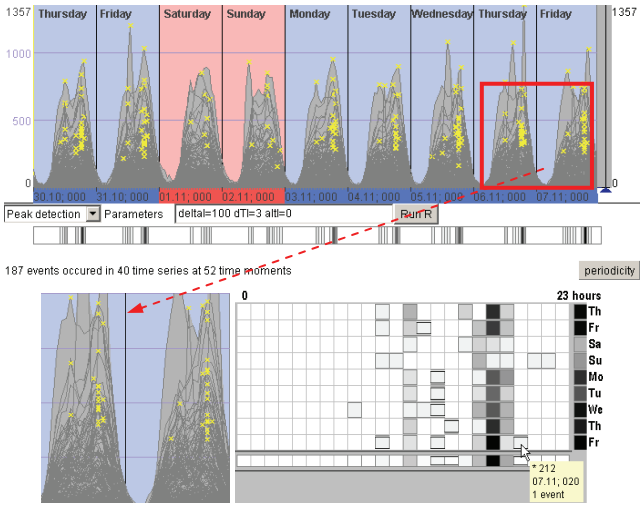

Fig. 3. Positions of peaks are marked on the time graph by yellow crosses. The area within the red rectangle is enlarged in the bottom left. The periodicity chart is in the bottom right.

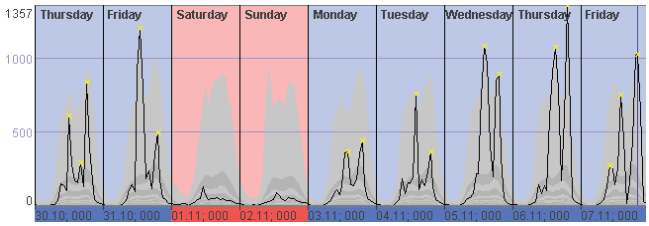

Fig. 4. We used the periodicity chart (figure 3) for selecting the time series event in the evening of Friday, November 7. The corresponding time series is shown on the time graph.

dimensions to represent space and vertical dimension to represent time [13]. It supports visual detection of spatial, temporal, and spatio-temporal clusters of events.

Our implementation enables spatial, temporal, and attribute-based filtering of t-events and time series as well as cross-filtering: after selecting a subset of t-events by any kind of filter or combination of filters, the time series graph can display only the time series containing the selected t-events. Similarly, after selecting a subset of time series, all displays showing t-events (map, space-time cube, event bar, etc.) can represent only the t-events ex-

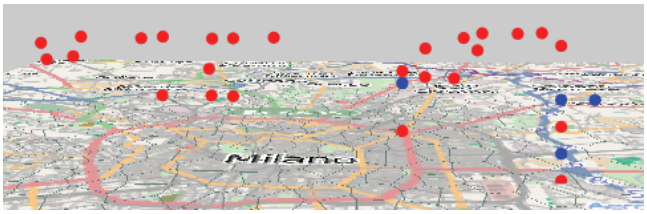

Fig. 5. A fragment of the space-time cube showing the positions of the peaks (red dots) and pits (blue dots). The vertical position of the moveable map plane corresponds to 10:00 on November 7.

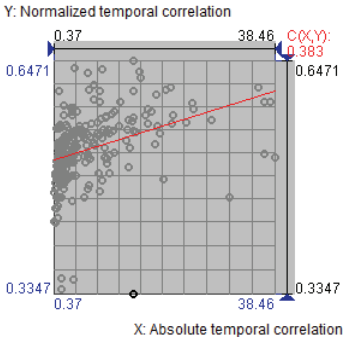

Fig. 6. Absolute and normalized temporal correlations of the Milan calls time series are shown on a scatterplot.

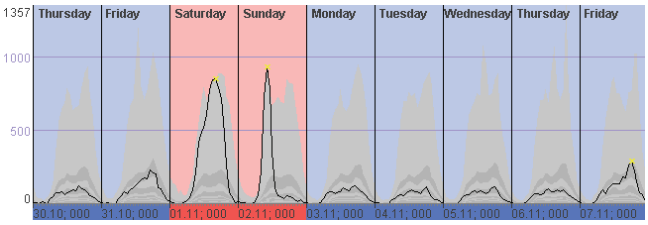

Fig. 7. A non-periodic time series has been selected using the scatterplot (figure 6).

tracted from the selected time series (figure 4).

Results of the periodicity detection algorithm are presented on a scatterplot (figure 6) that shows the absolute correlation scores against the normalized values. This plot is used for selecting time series of interest to be viewed on the time graph (figure 7). The spatial positions of these time series are marked on the map.

4 Parameter Impact

4.1 Division of space and time

A well-known problem in geographic analysis is the modifiable areal unit problem (MAUP) associated with aggregating data by areas: aggregated values vary according to how the area boundaries are defined [21]. MAUP includes two components: scale effect (variation due to the sizes of the areas) and zonation effect (variation due to the delineation of the area boundaries at a given scale).

Generally, a problem similar to MAUP exists for temporal aggregation: the results may depend on the choice of the time intervals. However, the selection of suitable intervals is usually not a difficult problem in analyzing data related to human activities, which are greatly af-

fected by the temporal cycles: daily, weekly, and yearly. Units of these cycles (hours, days, weeks, and months) are typically selected as time intervals for data aggregation since this gives easily interpretable results. Which unit to select, may depend on the semantics of the data, the goals of the analysis, and on the time span of the available data. Thus, we chose hourly intervals for the Milan calls data since the calls are related to the daily activities of people, which depend on the time of the day. For the flickr data, the weekly and yearly cycles appear the most relevant. The suitable units could be days, weeks, or months. With the monthly division, the aggregation is too coarse (many important events fit in the same time bin) and with the daily division too fine (almost in each week there are peaks on the weekend, which do not necessarily indicate interesting events but just the leisure time of many people). Hence, the weekly division is the most appropriate.

For the spatial division, however, there are no justifiable criteria for choosing the units; hence, MAUP is unavoidable. Unfortunately, the literature does not offer generic practical solutions to the problem. A simple strategy that is often used is performing analysis at multiple scales or zones. Although this is not a systematic way to address MAUP, it may give the analyst some idea of the sensitivity of the observed patterns to the space division. It would be beneficial if the analyst could interactively "play" with the tessellation and immediately see how this affects the spatio-temporal distribution of the t-events. However, this scenario is not achievable with the current levels of performance of the commonly used computers and databases. The problem is that the spatio-temporal aggregation of large data takes too much time to allow interactive analysis. Thus, for the data used in this paper, the aggregation time ranges from about 2 to 15 minutes. Hence, interactive visual analysis can start only after the aggregation is done. The effects of the space division can be explored by "playing" with several pre-computed aggregations based on different space divisions.

Figure 1 shows that the analyst can produce multiple divisions of the same territory by varying the parameter 'cluster radius'. Besides, different samples from the database can be used to obtain several divisions at the same spatial scale (i.e., with the same value of 'cluster radius'). The aggregation procedure is implemented so that only one run through the database is needed to obtain the time series for several alternative divisions. The data are loaded from the database to the main memory by portions and aggregated by each of the user-selected sets of areas. After the aggregation is done, the analyst can apply the time series analysis tools to each of the aggregated datasets and use interactive visualizations to compare the results, as will be shown by example in section 6.4.

4.2 Time series analysis

The method for periodicity detection uses one parameter, the expected length of the period. The right value is not difficult to find: depending on the temporal granularity and lengths of the time series, the analyst chooses one of the temporal cycles affecting human activities: daily, weekly, or yearly. The method for pit/peak detection uses two parameters, minimal peak amplitude and time window. It is not always clear what values of these parameters to select. The algorithm, like any computational technique, requires exact numbers whereas the analyst often has only an approximate idea of what values are reasonable. The analyst should be able to see how the variation of the values within reasonable ranges would influence the results of the analysis. We have implemented a set of techniques supporting the analysis of the sensitivity of the method outcomes to the values of the parameters.

The results of the peak/pit detection method monotonously depend on the two parameters in the following way: Increasing the value of the minimal amplitude decreases the amount of detected events; increasing the width of the time window increases the number of events. To test the sensitivity, the method is automatically invoked several times for each value of the time window parameter from the range specified by the user. For the amplitude, the minimal value of interest is used in each run since the algorithm will also find all peaks with higher amplitudes. The results from all method runs are represented in a summarized form on visual displays enabling comparisons[2]. A two-dimensional histogram (figure 8) can show for each pair of the parameters <time window, amplitude> one of the following indicators:

- the number of the extracted t-events;
- the number of places where the t-events occurred;
- the number of time intervals when they occurred.

The indicators can also be displayed in a cumulative mode (figure 8 bottom), thus supporting the assessment of the total numbers of t-events that would be extracted with the different values of the parameter 'minimal amplitude'. When necessary, the user may create two or more displays representing different indicators.

The histogram display can also be used for the investigation of the temporal distribution of the extracted t-events. In figure 9, the horizontal dimension represents the temporal range of the set of the extracted t-events and the vertical dimension, as before, represents the different values of the time window parameter. In the upper image, the heights of the bars show the counts of t-events by the time units of the data. In the bottom, the heights of the bars show the average amplitudes of the events (it is also possible to look at the minimal or maximal amplitudes).

The spatial distribution of the extracted t-events is examined using maps that show event statistics (counts; minimal, maximal, or average amplitudes) by the areas. To investigate the impact of the time window length, two map types are used: a 'small multiple' display with several choropleth maps corresponding to different parameter values (figure 10) and a bar diagram map where each bar within a diagram corresponds to one parameter value (figure 11). The first variant enables overall comparisons of the spatial distributions and the second enables local comparisons of counts or amplitudes.

The user can apply interactive filtering to focus on a

[2] The illustrations refer to the Milan calls data. The t-events are peaks with the minimal amplitude 50 and the time window range from 1 to 5 steps (hours); see section 5.1.

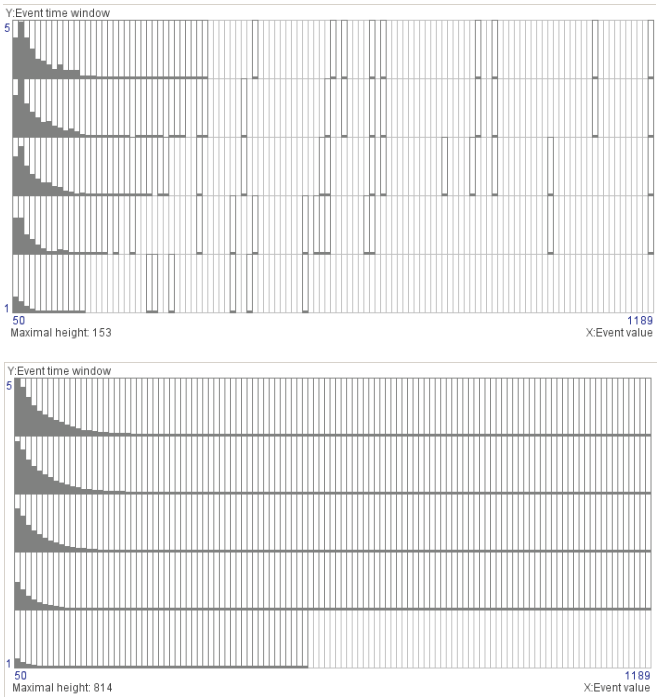

Fig. 8. Top: A two-dimensional histogram shows the counts of extracted t-events depending on the event values (amplitudes) (horizontal axis) and time window (vertical axis). Bottom: the same information is shown in a cumulative mode.

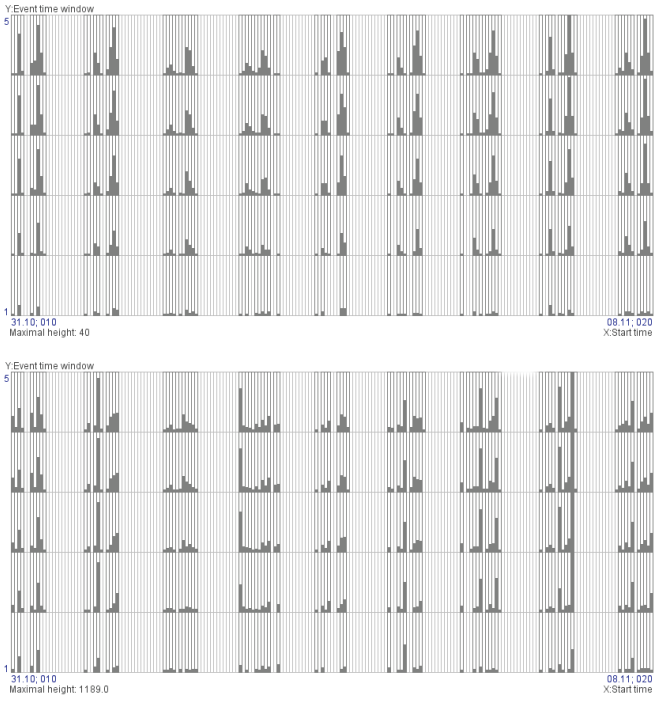

Fig. 9. The temporal distribution of the t-events extracted with different values of the time window (vertical axis). The horizontal axis represents the time range of the data. The upper image shows event counts and the lower image shows their average amplitudes.

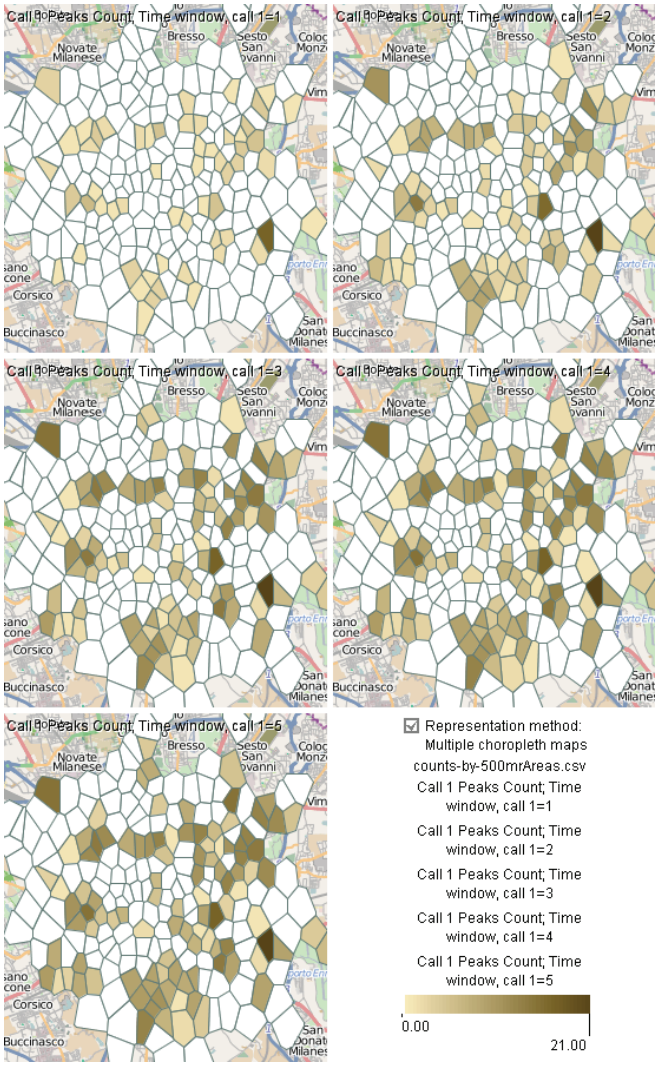

Fig. 10. Small multiples map show counts of t-events in different regions depending on the time window.

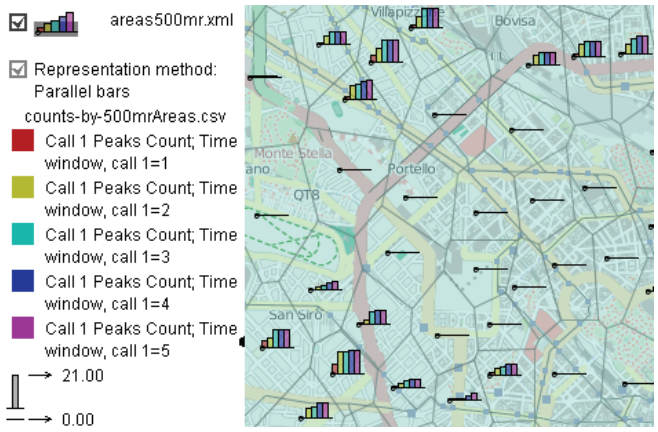

Fig. 11. Bar chart map shows counts of t-events (fragment).

subset of t-events, e.g., within a certain range of amplitudes. The histogram and map displays dynamically react to the filter by re-computing and representing the summaries only for the t-events that passed through the filter.

These interactive visual tools allow the analyst to study the impact of the parameter values on the number of the detected t-events, their spatial and temporal distributions, and the statistical distribution of the amplitudes. On this basis, the analyst can make an informed selection of appropriate values for the parameters.

5 Investigating Mobile Phone Calls in Milan

In this case study, we analyze a set of 2,956,739 call records of 367,730 mobile phone customers in Milan. The data span over 9 days starting from Thursday and ending on Friday next week. We used a sample of about 10,000 calls for a tessellation with the radius 500m, which is roughly a half of the average distance between the antennas of the WIND network. The tessellation resulted in 235 areas, of which 177 include single antennas, 47 include two, 10 include three, and one area includes four antennas. The areas with two or more antennas are located in the centre, where the density of the antennas is high. We could also use Voronoi polygons built around the positions of the antennas; however, in this case some of the areas would be much smaller than the rest. Equalizing the sizes of the areas allows a more valid comparison of the respective counts of people's presence.

For the generated areas, we computed hourly counts of calls for 9 days, which gave us time series of the length 216 hours (figure 2). The available attributes of the calls allow us to classify each call as stationary or mobile based on the customer's displacement during the call. On this basis, we can compute the counts of stationary and mobile calls and their proportion for any combination of area and time interval. This is an example of the acquisition of contextual information, which is mentioned in section 2.

5.1 Peak detection

The time graph and summary statistics display (figure 2) show us the overall character and the ranges of the variation of the calling activities. We shall use the peak detection algorithm for finding sharp rises of the call counts. To choose suitable parameter values, we apply the tools for parameter-sensitivity analysis (section 4.1). We run the algorithm for the range of time window lengths from 1 to 5 and the minimal peak amplitude starting from 50 and inspect the results using two-dimensional histograms (figures 8 and 9) and maps (figures 10 and 11). We see that the time windows 1 and 2 are too restrictive: the counts of the extracted t-events are much smaller than for the other values. The results for the time window lengths 4 and 5 do not differ much from those for the length 3. We choose to focus on the time window length 3, i.e., on the sharper peaks, and disregard the additional gentler peaks extracted for the lengths 4 and 5.

Concerning the minimal amplitude, we see from the histograms (figure 8) that the low values produce very many peaks, e.g., 615 for the value 50 and 251 for the value 90. There are 187 peaks with the amplitudes 100 and higher; we expect them to be more interesting than those with lower amplitudes. So, we choose to focus on this subset of peak events. They occurred in 40 distinct time series at 52 different time moments. Figure 3 shows their temporal positions. The periodicity chart on the bottom right shows that

- the peaks occur more frequently on the working days than during the weekend;
- on the working days, many peaks occur from 12:00 to 13:00 and especially many from 17:00 to 20:00;
- only a few peaks occurred at other times of the working days.

We have discussed the times of the calling peaks with several people living in Italy, who interpreted the evening peaks as a very typical behavior: people call home after the end of the work ("I am coming home, cook pasta!").

We have inspected several peaks that occurred in unusual times. For example, the only peak that occurred on Friday after 20:00 was in the region on the south-east. The time series (figure 4) has a shape typical for office areas, with relatively low values in the weekend but high peaks at lunch time and at the end of the day on the working days. The majority of the calls are stationary in the lunch time peaks and mobile in the evening peaks, which is also typical for office areas. However, the Friday evening peak occurred later than usual. A person knowing Milan informed us that this area contains a large television studio. The unusually late peak of the calls may indicate that the employees had to work longer that day.

In addition to the peaks, we extracted 20 pits that happened in 11 time series at 14 time moments. We plotted the positions of the peaks and pits in the space-time cube. Most pits occur after or before peaks (figure 5). There are also several cases of standalone pits. There are several neighboring areas where the number of calls suddenly dropped from the usual about 200 calls per hour to 0 calls, which lasted for several hours. One more drop occurred nearby on the next day. These t-events may indicate technical problems or maintenance works.

5.2 Periodicity analysis

Now we shall analyze the data by means of the periodicity detection procedure. The nature of the data suggests the target period of 24 hours. We compute the absolute and normalized temporal correlations for this target period and visualize them on a scatter plot (figure 6) linked to the time graph and the map. We find that the highest correlation values are in the residential areas, where the daily profiles for the working days are similar to the weekend profiles. Surprisingly, the residential areas have large proportions of mobile calls at the times of the peak events. Probably, people tend to make many calls during their trips to or from home. Many areas have medium values of the correlation. Temporal profiles for these regions show high similarity among the working days and low activity in the weekend, typical for office areas.

We study in detail the time series with the smallest correlation values. One of them (black dot at the bottom of the scatterplot) refers to an area with a big parking lot near a train station. Figure 7 exhibits very high peaks of calls there on Saturday and Sunday. They are mostly stationary. A person knowing Milan related these peaks to a flea market that sometimes takes place in that area. Other non-periodic time series, shown in figure 12, refer to three neighboring areas containing the stadium Giuseppe Meazza and other sport arenas. Two time series have high peaks on Sunday and Thursday, the third one has a peak on Saturday. The peaks may correspond to sport or cultural events. For a detailed analysis, we aggregate the calls in the stadium area with one minute temporal resolution for the time intervals when the peaks occurred.

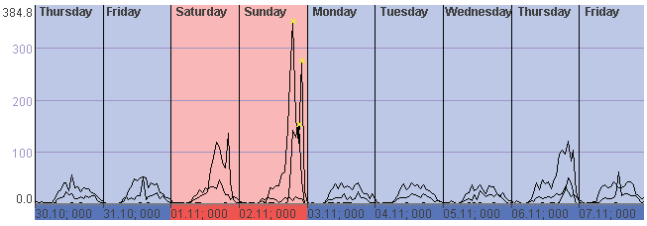

Fig. 12. The 9-days time series for the areas close to the stadium.

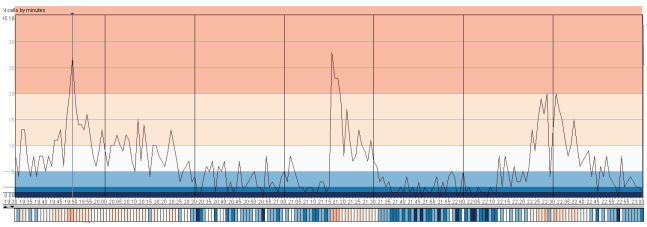

Fig. 13. One-minute aggregates for the area near the stadium. The color band in the bottom shows the dynamics of the calls.

5.3 Detailed analysis in space and time

Figure 13 shows the counts of the calls in the stadium area on Sunday with one minute resolution. The major peaks occurred at 19:50, 21:15, and around 22:30. The calls in the peaks before 22:30 are mostly stationary while after 22:30 they are mostly mobile. Very few calls occurred in the intervals 20:30-21:15 and 21:30-22:20. Very probably, this profile corresponds to a football game. The peak in 40 minutes before the game can be explained by the appearance of the players or by the announcement of the team composition. The periods without calls may correspond to the two halves of the game and the peak between them to the break. The profile of the second half differs from that of the first half, showing a high calling activity at the end. We found that a national championship match[3] attended by about 50,000 spectators started in the stadium on Sunday at 20:30. A single goal was scored at the end of the game, which explains the peak of phone calling.

Another game with about 11,000 attendants took place on Thursday[4]. The profile of the one-minute call counts is similar to that on Sunday. The amplitudes of the peaks are proportional to the attendance of the two games.

5.4 Potential applications

Our techniques allowed us to learn how the presence of people varies in different areas over a day and over a week, identify residential and business areas, detect areas where the presence of people can suddenly increase and areas with unusual temporal patterns of activities. Such findings may be useful for emergency management, transportation planning, and maintenance of public order. We have discussed our analyses with the data providers and made them very enthusiastic about the possibilities for extracting interesting information from their data, which are not seriously analyzed yet.

[3]http://www.fussballdaten.de/italien/2009/10/acmailand-neapel/
[4]http://www.fussballdaten.de/europaleague/2009/zwischenrunde/gruppee/acmailand-braga/

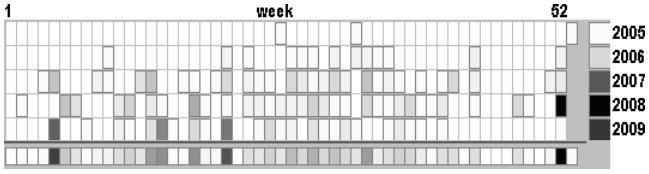

Fig. 14. The periodicity chart shows the counts of the peaks in the number of flickr photographers by weeks (columns) and years (rows). The rightmost column shows the yearly totals, the row in the bottom represents the totals by the corresponding weeks of the different years.

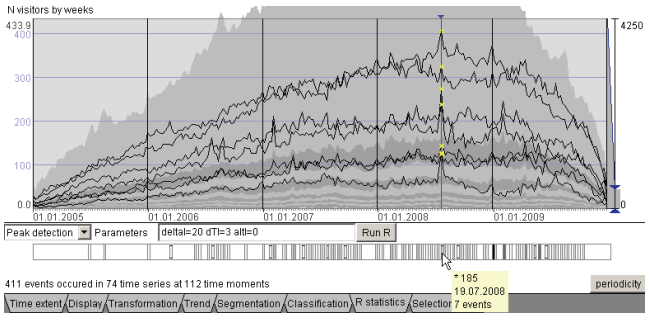

Fig. 15. The time graph highlights the time series that had peaks in the week starting on 19.07.2008.

6 Exploring 5-years history of British Isles

In this study, we analyze 8,686,034 records about the photos made in UK and Ireland by 97,008 flickr.com users from January 2005 till December 2009. We used a sample of about 25,000 photos for producing several tessellations (figure 1). We started our analysis with the division based on the cluster radius 50km, with 164 areas. Later we compared the results with those for other divisions (section 6.4). For the areas, the time series of the weekly attendance by different photographers were computed. The length of the time series is 261 weeks. In the region of London, the values are much higher than elsewhere. We temporarily excluded this region from the analysis and then studied it separately at a finer spatial scale. For the lack of space, we cannot describe our study of the London region here. Interested readers can look at paper [2] and the supplementary video in the IEEE Digital Library.

6.1 Peak detection

Using the peak detection tool, we have extracted 411 peaks with the time window 3 and amplitude 20 or more that occurred in 74 regions in 112 weeks. For interpreting the peaks, we requested the system to extract additional context information from the database, namely, the most frequent words and phrases from the titles of the photos. Hence, each peak event is characterized by a set of frequent words and phrases with their respective frequencies. Pointing on a display element representing a t-event in the time graph, in the map, or in the space-time cube gives access to the attributes of this t-event, including the frequent words and phrases from the photo titles.

The periodicity chart (figure 14) demonstrates that there were only a few peaks in the years 2005 and 2006

(evidently, flickr was not yet very popular). This is indicated by the light shading of the cells in the rightmost column. In the other time dimension, the largest numbers of peaks occurred in the calendar weeks 5, 15, 21, 29, 34, and 52, which is indicated by the dark shading of the respective cells in the bottom row. The elements of the periodicity chart facilitate the access to the corresponding t-events. For example, by clicking on the fifth cell of the bottom row, we select all t-events that occurred in the fifth week of all years (such events were only in 2007 and 2009). We look at the frequent words characterizing the selected peaks. The word "snow" is the most frequent. We conclude that an unusual snowfall attracted people's attention in many regions in the fifths weeks of 2007 and 2009. The regions are highlighted in the map display.

The periodicity chart shows us that the temporal distribution of the peak events is, generally, not periodic. The distribution of the t-events along the time line is shown by the linear event bar below the time graph (figure 15). Like with the periodicity chart, it is possible to select t-events and the time series in which they occurred by clicking on the cells of the linear event bar.

Figure 15 corresponds to the selection of the week of 19.07.2008, in which peak events occurred in six areas. By accessing the frequent words and phrases, we find out that a tall ship race took place in Liverpool, a river festival in Glasgow, a Red Arrows air show in Farnborough, a Ferrari fun day in Newbury, a war peace show in East Sussex, and a Latitude festival and air show in Suffolk. Each of these events attracted from 100 to 400 different photographers above the regular attendance of the areas.

6.2 Detecting periodic patterns

As we have noted, there is no general periodicity in the data. However, periodic patterns of attendance by photographers can be expected in nature areas, which are visited more often in summer than in winter. We run the temporal correlation check with the target period of 52 weeks. The relative periodicity scores are shown on the map in figure 16 (left) by area shading; darker is more. For many areas with high periodicity scores, the time series profiles are similar to the one shown in figure 17 top, which has clear seasonal differences but no large peaks. These areas are located mostly on the coastline and in the rural regions. This corresponds to our expectations.

We also expect that some areas may have regular public events that occur at about the same time every year. Indeed, several areas with high periodicity scores have high peaks in particular weeks of each year, as in figure 17 bottom. The information from the photo titles indicates that the peaks in this time series correspond to the Silverstone Grand Prix annual event. Other regions with similar features are Swindon (Royal International Air Tattoo), Bristol (Glastonbury festival), Eastnor Castle (Big Chill festival), and Littlehampton (Goodwood festival).

Some regions had several peaks with irregular intervals, for example, Nottingham where several festivals took place. Other regions had singular peaks corresponding to occasional events, such as South Devon railway anniversary (April 2009), UEFA Cup Final game in Manchester (May 2008), Edinburgh book festival (August 2007), and veteran car rally London-Brighton (November 2007). Figure 16 summarizes our findings.

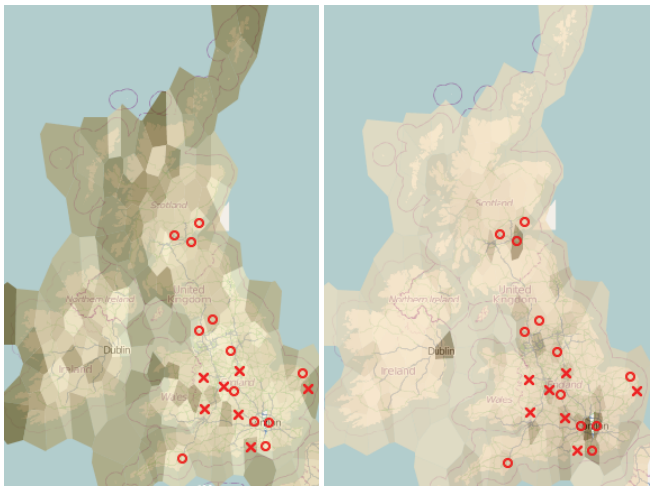

Fig. 16. The shading of the areas shows the relative periodicity scores (left) and the total number of visitors (right). The symbols mark the areas with periodic (crosses) and irregular (circles) peaks.

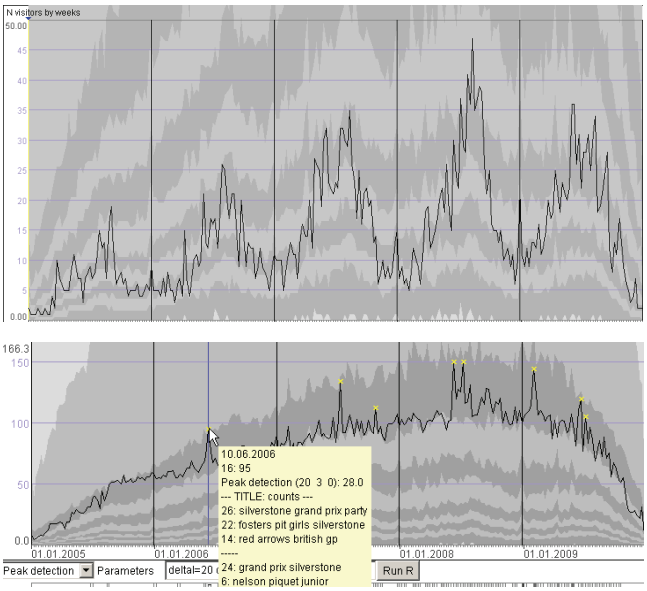

Fig. 17. A periodic time series in Scotland (top) and periodic peaks of the Silverstone Grand Prix (bottom).

6.3 Detecting "small" events

Now we want to find interesting events that might take place in small municipalities and other places where the number of flickr users taking photos is usually small. In such areas, an increase by, e.g., 10 people over the usual number may mean that something interesting happened. To find such happenings, we need to look at the relative increases rather than absolute. Hence, we apply the peak detection method to the visitor counts normalized to z-scores, as mentioned in section 3.3, using the minimal amplitude of 1.5 standard deviations. The method extracts 427 t-events that occurred in 109 regions. The events are visualized in a space-time cube (figure 18 top). Each t-event is represented by a graduated circle with the

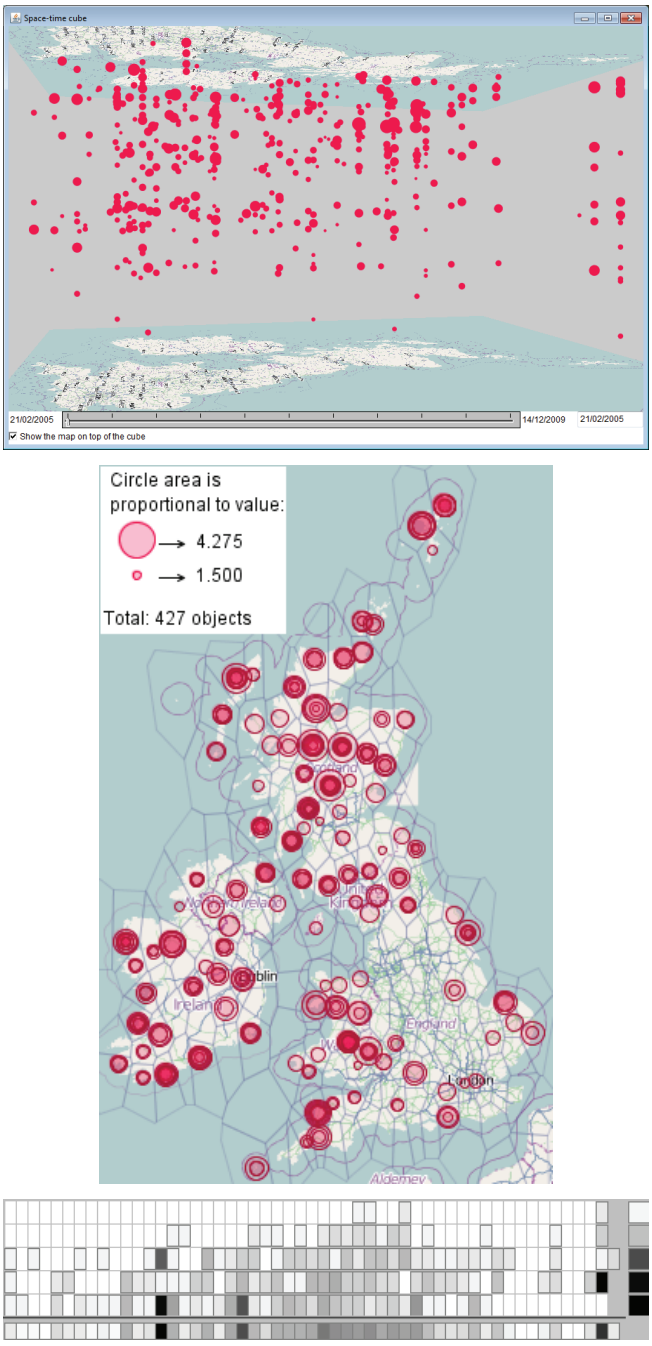

Fig. 18. T-events extracted from normalized time series. Top: the events are represented in a space-time cube by graduated circles with the areas proportional to the event amplitudes. Center: the same circles are shown on a map with 5% opacity. Bottom: the periodicity chart shows the temporal distribution of the events.

area proportional to the event value, i.e., the relative increase of the visitor count. Vertically aligned circles represent t-events that occurred in the same area at different times. The map in the middle of figure 18 can be viewed as a top projection of the cube. The circles are placed on the map according to the spatial positions of the events, irrespective of the time, and filled with 5% opacity. By comparing this map with the one in figure 16 (upper right), it may be seen that the majority of the t-events and the t-events with the largest relative amplitudes occurred in the areas in which the number of flickr photographers is usually low. More precisely, 385 of the 427 events occurred in 84 regions where the median number of photographers does not exceed 25. In the periodicity chart (figure 18 bottom) we see that frequencies of the t-events are generally higher in summer than in the other seasons while the highest frequencies are attained in the times of the Christmas and New Year holidays, on Easter, and at the end of May.

As we did for the absolute peaks, we extract the most frequent words and phrases from the photo titles for the relative peaks. This allows us to find two areas with t-events reflecting the snowfalls of February 2007 and three areas with t-events reflecting the snowfalls of February 2009 additionally to those detected by the absolute amplitudes. We find four t-events in different regions reflecting the floods in UK in June-July 2007. Only one of these t-events (in Oxford) could be detected by the absolute amplitude. The photos also represent various public events, among them 30 festivals, 19 shows, and 8 rallies. Most of these events could not be detected by absolute amplitudes, for example, Welsh food festival, Green Man music festival in Brecon Beacons, Wales, Electric Picnic arts-and-music festival in Stradbally, Ireland, Scottish traditional boat festival in Banff, Ballymena agricultural show in Nothern Ireland, etc. As could be expected, many of the t-events reflect real-world events of merely local importance such as weddings and jubilee parties.

6.4 Exploring the impact of the territory division

As discussed in section 4.1, the results of the analysis may be sensitive to the territory division: sizes of the areas (scale effect) and delineation of the boundaries (zonation effect). As there is no generic solution to the problem, a usual strategy is to do analysis for multiple divisions. To explore the scale dependency, we aggregate data for several divisions of the territory with different area sizes (see figure 1) and apply the peak detection tool to each set of time series. As could be expected, the larger the areas, the more t-events are detected for the same value of the parameter 'minimal amplitude'. Thus, for the value 20, the tool extracts 263, 411, 469, and 497 t-events from the time series related to the divisions with the cluster radii 30, 50, 70, and 100 km, respectively. To diminish the effect of the area sizes and concentrate on comparing the spatial and temporal distributions of the t-events for the different divisions, different values of the parameter 'minimal amplitude' should be used. Suitable values can be selected with the help of the histograms as shown in figure 8.

The spatio-temporal distributions of the t-events are compared with the help of spatial and temporal displays. It is perceptually difficult to observe and compare more than two distributions at a time; therefore, we compare them pair-wise. Here we briefly report only about comparing the results for the 50km and 30km tessellations. Comparable numbers of t-events (411 and 424, respectively) are obtained for the minimal amplitudes 20 and 17, respectively. The temporal distributions are compared using the periodicity charts (figure 19). We observe that the distributions are very similar although not identical.

The overall spatial distributions are compared by

means of juxtaposed maps. Thus, in figure 20, the graduated circles represent the total numbers of t-events extracted from the areas of the 50 km (left) and 30 km (right) divisions. In the regions where the circles are concentrated, the observed differences in their numbers and sizes are likely to be explainable by different grouping of the same t-events by the coarser and finer space divisions: the circles are fewer but larger in the coarser division and more numerous but smaller in the finer division. In both maps there are also isolated circles. For some of them, there are no corresponding circles with approximately the same positions in the other map. There is no simple dependency like one of the divisions just yields additional t-events in comparison to the other. The time graph in figure 21 represents the arithmetic differences between the counts of the t-events in two divisions by the weekly time intervals. We see that there is no consistent prevalence of any of the two divisions.

A detailed exploration of the differences in the spatial distributions by the time intervals is supported by animated maps. Figure 22 shows two screenshots of an animated map corresponding to the time intervals with the highest, by the absolute values, numeric differences (visible on the time graph): +6 and -8. In the map, the areas having t-events in the current time interval are marked by coloring. Violet is used for the areas of the 50km division and red for the areas of the 30km division. The areas are drawn with 50% opacity so that their common parts can be seen. Overlapping areas are very likely to capture the same concentrations of people. In principle, this can be checked by database queries (retrieving the photo records for each area and counting the common records); however, in the scope of this study, we are interested not in the exact counts but in the reasons for the people's concentrations, which can be guessed from the words and phrases frequently occurring in the titles of the photos.

We extract the frequent words and phrases for the t-events detected with the 30km division and compare them with those for the 50km division. Based on a large sample of time intervals and areas chosen with the help of the time graph of the numeric differences (figure 21), we make following observations:

- Events of public interest (shows, festivals, sport events, etc.) detected with the 50km division are also detected with the 30km division.
- Some interesting events have been detected with the 30km division but not with the 50km division, e.g., Bristol kite festival, world fireworks championship in Blackpool, Dublin marathon (all in 2008), and others.
- For the t-events from the 50km areas that have no corresponding t-events in the 30km areas, the frequent words and phrases mostly do not indicate any special happenings. Evidently, the photos come from people that were dispersed in space rather than concentrated in a place of some interesting event.

We can conclude that the results of the t-event extraction do depend on the scale of the territory division used for the data aggregation. While the overall patterns of the spatial and temporal distribution are not very sensitive, interesting events can be missed when the spatial resolution is too coarse. Hence, finer division is preferable; however, the exploration of a large territory may be quite tedious. It may be reasonable to focus on smaller regions.

Similarly to the exploration of the scale effect, we ex-

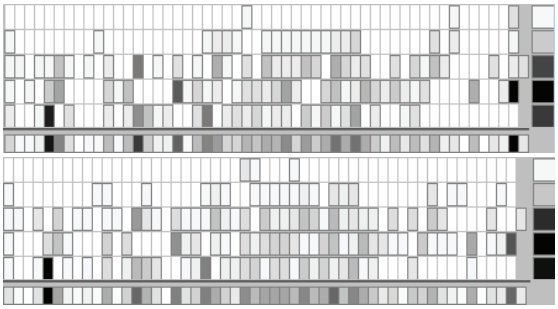

Fig. 19. The periodicity charts show the temporal distributions of the t-events obtained for the 50km (top) and 30km (bottom) divisions.

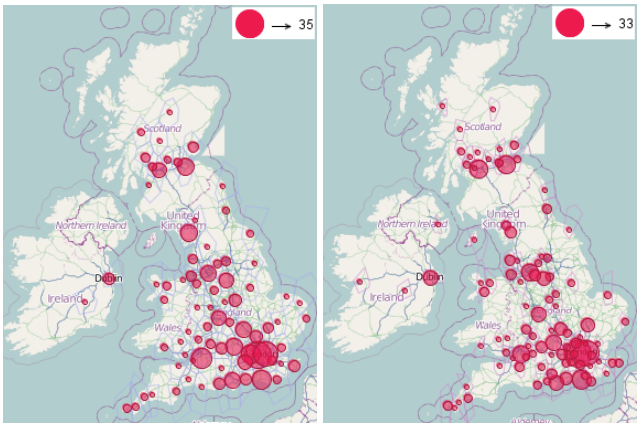

Fig. 20. The maps show the spatial distributions of the t-events obtained for the 50km (left) and 30km (right) tessellations.

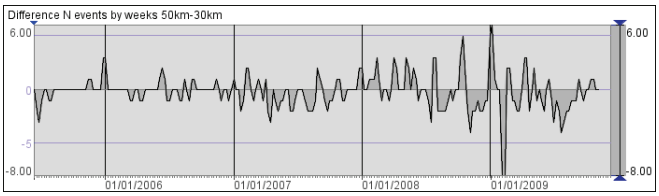

Fig. 21. The time graph shows the differences in the numbers of t-events extracted from different time intervals for the 50km and 30km tessellations.

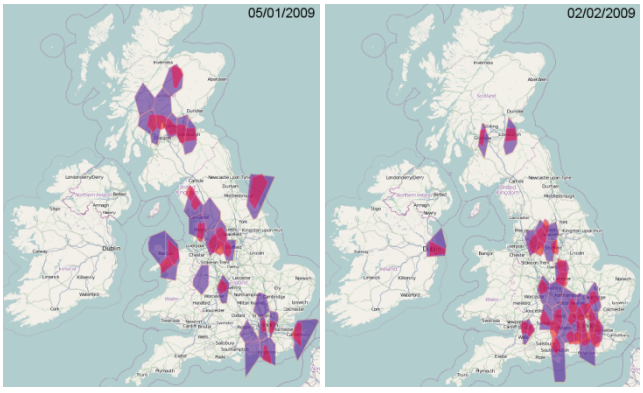

Fig. 22. The screenshots show the areas of the 50km (violet) and 30km (red) divisions having t-events in two selected time intervals.

plore the zonation effect using several territory divisions with 30km cluster radius based on different random samples of points from the database. In this case, the areas of the divisions do not differ much, since the tessellation algorithm takes the generating points for the areas from dense point concentrations, which tend to occur in the same places in different data samples. As could be expected, the overall spatial and temporal distributions of the sets of t-events extracted from different territory divisions are very similar. The numeric differences in the total numbers of the t-events by the time intervals range from -5 to 5. By more detailed pair-wise comparisons involving the frequent words and phrases from the photo titles, we could not find any events of public interest detected from one division and not detected from the other. For the t-events having no counterparts in the other set, the photo titles mostly refer to particular places or landmarks rather than real-world events. Hence, the zonation effect is not as high for this kind of analysis as the scale effect.

6.5 Potential applications

In this case study, we have learned many interesting facts about a foreign country and its capital. The possibility to explore the flickr data in such a way might be useful for tourists, who could in this way learn more about the region or city they are going to visit. Of course, a simple and appealing user interface is required for this purpose. A web service with such an interface is currently being developed at our institute.

7 Discussion and Conclusion

The case studies show that past events of public interest can be detected and interpreted by analyzing data about activities of people. Currently many data sets with people's activity traces are available publicly or can be acquired. We used the public flickr data and a proprietary dataset of a mobile phone company. Other potentially available data sets include Wikipedia articles, twitter messages, and news streams in the public domain as well as data from various stationary and mobile sensors. We suggest a framework for data exploration that takes into account the spatial and temporal distribution of the data records and available numeric and textual attributes.

The idea of reconstructing history by analyzing activity data has its limitations. A major issue is the spatial, temporal, and population coverage of the available data. Important events that are not reflected in the data cannot be detected. For example, the data about mobile phone calls reflect only the events attended by the customers of the phone company and the situations when the use of phones is permitted. The flickr photos data do not reflect the events that were not attended by the flickr users or when taking photos was not permitted or possible. Thus, the event of London bombing of July 7, 2005 was not reflected in the data.

Selection of appropriate scales in space and time is essential for the success of the analysis. Depending on the scale, we can find or miss important events. There is no universal recipe for choosing the most appropriate scale. It is necessary to base the analysis on the domain knowledge, study the sensitivity of the results to the parameters, and perform multi-scale analysis.

An important feature of the suggested framework is the flexibility. To reflect three major components of spatio-temporal data - what, when and where [22] - we implemented several workflows that may be arbitrarily combined:

- what → where + when: for cases of unusual periodicity (either low or high) or unusual number of events, analyze the spatial distribution of the regions and the temporal distribution of the events in these regions;
- when → what + where: for time moments or intervals with unusual numbers of events, find what and where happened;
- where → what + when: for selected regions, investigate what events and when occurred there.

Data analysis according to our framework can be done very efficiently. In our experiments, the time for analyzing a previously unknown dataset was from 30 to 60 minutes.

In our research, we extended the prior works by

- advancing the scalability of the methods by distributed data management and processing;
- supporting the analysis by geocomputations (detection of regions) and statistical methods (peak detection and periodicity testing);
- supporting the investigation of the sensitivity of the analysis results to method parameters;
- enabling flexible workflows for interactive analysis;
- providing novel visualizations (periodic event bar, space-time cube with graduated symbols) and coordination mechanisms (linking elements of different data sets and of different nature, specifically, regions, time series, and events);
- enabling on-demand acquisition of contextual information and fusion of different data types.

In the future, we plan to develop methods for combined analysis of multiple data sets referring to the same territory and time. We are going to extend the library of the time series analysis methods and integrate them in appropriate visual analytics workflows.

Acknowledgment

The work has been supported by the DFG – Deutsche Forschungsgemeinschaft (German Research Foundation) within the research project ViAMoD – Visual Spatiotemporal Pattern Analysis of Movement and Event Data.

References

[1] N.Andrienko, G.Andrienko. Exploratory Analysis of Spatial and Temporal Data. A Systematic Approach. Springer-Verlag, Berlin, 2006

[2] G.Andrienko, N.Andrienko, M.Mladenov, M.Mock, C.Poelitz. Discovering Bits of Place Histories from People's Activity Traces. In IEEE VAST 2010, pp.59-66.

[3] G.Andrienko, N.Andrienko. Visual exploration of the spatial distribution of temporal behaviours. In: IV 2005, pp. 799-806

[4] N.Andrienko, G.Andrienko. Spatial Generalization and Aggre-

gation of Massive Movement Data. IEEE Transactions on Visualization and Computer Graphics, 17(2), 2011 , pp. 205-219

[5] M.Basseville and I.Nikiforov. Detection of Abrupt Changes - Theory and Application. Prentice-Hall, Englewood Cliffs, 1993

[6] E.Billauer. Peakdet: Peak detection using MATLAB. Online, http://www.billauer.co.il/peakdet.htm. Retrieved 26 Feb. 2010

[7] P.Buono, A.Aris, C.Plaisant, A.Khella, B.Shneiderman. Interactive Pattern Search in Time Series. In VDA 2005, SPIE, 175-186

[8] P.Buono, C.Plaisant, A.Simeone, A.Aris, B.Shneiderman, G.Shmueli, W.Jank. Similarity-Based Forecasting with Simultaneous Previews: A River Plot Interface for Time Series Forecasting. In IV 2007, Zurich, Switzerland;

[9] M.Elfeky, W.Aref and A.Elmagarmid, Periodicity Detection in Time Series Databases. IEEE Transactions on Knowledge and Data Engineering Vol. 17, No.7. July 2005

[10] S.Fortune. A sweepline algorithm for Voronoi diagrams. In Proc. Second Annual Symposium on Computational Geometry, Yorktown Heights, New York, United States, 1986, 313–322

[11] F.Girardin, F.Fiore, C.Ratti, J.Blat. Leveraging explicitly disclosed location information to understand tourist dynamics: a case study. Journal Location Based Services, 2008, 2(1), 41-56

[12] M.F.Goodchild. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. International Journal of Spatial Data Infrastructures Research 2: 24-32. 2007

[13] T.Hägerstrand. What about people in regional science? Papers, Regional Science Association, 24, 7-21. 1970

[14] M.Hao, H.Janetzko, P.Sharma, U.Dayal, D.Keim, M.Castellanos. Visual prediction of time series. In IEEE VAST 2009, pp. 229-230

[15] H.Hochheiser and B.Shneiderman, Dynamic query tools for time series data sets: Timebox widgets for interactive exploration, Information Visualization, 2004, 3(1), 1-18

[16] Y.Ivanov, Ch.Wren, A.Sorokin, I.Kaur, Visualizing the History of Living Spaces, IEEE Transactions on Visualization and Computer Graphics, 13(6), 2007, pp.1153-1160

[17] P.Jankowski, N.Andrienko, G.Andrienko, S.Kisilevich. Discovering landmark preferences and movement patterns from photo postings. Transactions in GIS, 4(6), 2010, pp.833-852

[18] Y.Kawahara, M.Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In SIAM Data Mining 2009

[19] D.Keim, G.Andrienko, J.-D.Fekete, C.Görg, J.Kohlhammer, G.Melancon. Visual Analytics: Definition, Process, and Challenges. In Information Visualization – Human-Centered Issues and Perspectives. Vol. 4950 of LNCS, Springer, 2008, pp.154-175

[20] S.Kisilevich, F.Mansmann, D.Keim, P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In COM.Geo 2010, ACM, New York, NY, USA. Article 38, 2010.

[21] S.Openshaw. The Modifiable Areal Unit Problem. Norwich: Geo Books, 1984.

[22] D.Peuquet, Representations of Space and Time. Guilford, 2002

[23] R.Pulselli, P.Romano, C.Ratti, E.Tiezzi. Computing urban mobile landscapes through monitoring population density based on cell-phone chatting. International .Journal of Design & Nature and Ecodynamics, 2008, 3(2), 121-134

[24] M.Small and K.Judd. Detecting periodicity in experimental data using linear modeling techniques. In Physical Review E 59(2), pp. 1379-1385. February, 1999

[25] B.Tversky, J.B.Morrison, M.Bétrancourt. Animation: can it facilitate? International Journal of Human-Computer Studies, 57(4), 2002, pp. 247-262

[26] J.J. van Wijk, E.R. van Selow. Cluster and Calendar Based Visualization of Time Series Data. In InfoVis 1999, pp. 4-9

[27] D. Yankov, E. Keogh, J. Medina, B. Chiu and V. Zordan. Detecting time series motifs under uniform scaling. In ACM KDD 2007

Gennady Andrienko is a lead scientist responsible for the visual analytics research at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). He received his master degree in Computer Science from Kiev State University in 1986 and PhD equivalent from Moscow State University in 1992. He did research on knowledge-based systems at the Institute for Mathematics of Moldavian Academy of Sciences (Kishinev, Moldova), then at the Institute for Mathematical Problems of Biology of Russian Academy of Sciences (Pushchino, Russia). Since 1997, he is working at GMD, now Fraunhofer IAIS. He co-authored the monograph 'Exploratory Analysis of Spatial and Temporal Data' (Springer, 1996), over 20 book chapters and more than 100 conference papers. Since 2007, Gennady Andrienko is chairing the ICA Commission on GeoVisualization. He co-organized scientific events on visual analytics, geovisualization and visual data mining, and co-edited multiple special issues of journals and proceedings volumes.

Natalia Andrienko received her Master degree in Computer Science from Kiev State University in 1985 and PhD equivalent from Moscow State University in 1993. She did research on knowledge-based systems at the Institute for Mathematics of Moldavian Academy of Sciences (Kishinev, Moldova), then at the Institute for Mathematical Problems of Biology of Russian Academy of Sciences (Pushchino, Russia). Since 1997, she has been working at GMD, now Fraunhofer IAIS. Since 2007, she is a lead scientist responsible for the visual analytics research. She co-authored the monograph 'Exploratory Analysis of Spatial and Temporal Data', over 50 peer-reviewed journal papers, over 20 book chapters and more than 100 conference papers.

Martin Mladenov is a Master's student in Computer Science at the University of Bonn and also working at Fraunhofer IAIS. His research interests include efficient inference in probabilistic and logic models and optimization. He is currently working on lifted sovers for linear programs.

Michael Mock is working as a senior scientist at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in the department KD (Knowledge Discovery). He received his diploma degree in Computer Science from the University of Bonn in 1987, doctor degree in 1995 and habilitation degree from the University of Magdeburg in 2004, being a member of the Department of Computer Science since then. His research interests include distributed systems, autonomous systems and real-time systems. He has authored over 50 publications, including textbooks on programming and wireless networks.

Christian Pölitz studied computer science at the Technical University of Ilmenau. 2009 he graduated as Dipl. Inform. (Master) and joint the Fraunhofer Institute for Intelligent Analysis and Information Systems and the University of Bonn. Currently he is a research associate at the Cluster of Excellence at the Saarland University.