Discovering Bits of Place Histories from People's Activity Traces

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

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

ABSTRACT

Events that happened in the past are important for understanding the ongoing processes, predicting future developments, and making informed decisions. Significant 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 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.

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). The history of a place includes as an essential part the events that happened in the place (the term 'event' is used in the sense of 'noteworthy happening', according to the Merriam-Webster's dictionary). Important and/or interesting events usually attract many people, active participants and/or spectators. Nowadays, some of the people attending events may leave traces of their presence in electronic databases. Examples are records about mobile phone calls, georeferenced photos on photo sharing web sites such as flickr, and georeferenced twitter messages. We call such records 'activity traces' or 'activity data', where 'activity' denotes various actions (phone calling, photo taking, message posting, etc.) or movement.

Since many people voluntarily make their data accessible to others via the Web, Goodchild [11] 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.

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 twitter message, the duration of a mobile phone call and the spatial displacement of the caller during the call, the title and text tags of a Flickr photo, etc.

Reconstructing significant events from activity data is a difficult task. The complexity is caused by several factors. 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 meaningful places are usually areas rather than 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.

In this paper, we propose a suite of visual analytics methods for detecting and reconstructing events. The methods support division of the territory into suitable areas according to the intended spatial scale of analysis, aggregation of data into time series associated with the areas, detecting events in time series and checking temporal correlation within time series, and interactive visual analysis and interpretation of the results. Our special focus is analysis of event periodicity. Especially interesting are cases of periodicity in data sets that are mostly not periodic and cases of non-periodicity in mostly periodic data sets.

We illustrate the approach on two real datasets that are kindly provided by our project partners. The first dataset, provided by the Italian telecommunication company WIND, contains positions of 2,956,739 phone calls made in Milan (Italy) during 9 days 30.10.2008 - 07.11.2008. We expect high periodicity in these data according to the daily temporal cycle. The second dataset consisting of the positions, temporal references, and titles of flickr.com photos, has been collected by our partners from University Konstanz within the DFG Priority Research Programme "Scalable Visual Analytics". The data set contains records about 85,041,956 photos worldwide, mostly captured after January 2005. In this paper, we use the subset of the database covering the territory of the UK and Ireland, 8,686,034 in total. We expect this data set to have little periodicity.

2 RELATED WORK

One of the key works in the area is [15]. Indoor movement data have been collected by motion sensors statically placed inside a building. The time series of the sensor activation counts were analyzed to describe the use of the space over time and to detect periods of intensive movement.

Several papers of the MIT group [10][20] considered data about mobile phone calls and georeferenced photos as a reflection of people's activity in a city in a selected period. To demonstrate how known events are reflected in the data, they visualize the spatial distribution of the phone 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. For this task, space-centric approaches should be combined with techniques for temporal analysis.

Analysis of time series data has been in focus of the information visualization community for a long time. Van Wijk [23] proposed a calendar display representing similarity of daily profiles of energy consumption. A large group of papers suggests advanced functionality for a time series graph. TimeSearcher [14] enables interactive querying of time series by their shapes. Paper [2] suggests approaches to representing multiple time series in a summarized form and describes interactive manipulations of the time series display for detecting patterns of sequential increase or

http://geoanalytics.net/and andrienko@geoanalytics.net

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 28, Salt Lake City, Utah, USA
978-1-4244-9486-6/10/$26.00 ©2010 IEEE

decrease of attribute values. The time graph display may be combined with a time band where time-dependent values such as overall averages or predicted values are represented by coloring or shading [13]. In recent versions of TimeSearcher [6][7], temporal positions of specific features of time series can be marked on the time graph.

Detection of features in time series attracts attention of the data mining and statistics communities. Methods have been proposed for finding specific patterns (motifs) in data [22], detecting change points [4][17], and finding periodic patterns [21][8].

Our previous paper [1] describes a basic infrastructure for finding specific patterns in spatial time series by means of statistical techniques and exploring the findings by means of visualization and interaction. Paper [16] explores the potential of the flickr photos data for gaining information about an area, interests of people, and patterns of their movement. The novel contributions of the current paper are:

1. Transformation of activity traces (points in space and time) into spatially referenced time series by means of spatio-temporal aggregation. This is based on user-controlled division of the study area into regions reflecting the spatial density of the activity traces.
2. Algorithmic methods for detection of peaks/pits and periodicity in time series, which are tightly integrated with visual displays focused on the analysis of repeated and periodic patterns.
3. Tools for on-demand acquisition of contextual information to enable pattern interpretation.

3 Visual Analytics Approach

In our approach, we follow the "Visual Analytics Mantra" [18]. We start with computational extraction of places (areas) from the data. Using the areas, we aggregate the data into spatial time series. The set of time series is explored by a combination of interactive visualizations and statistical methods for detection of peaks/pits and periodicity. Additional data are acquired on demand for supporting interpretation of the detected peaks or pits and reconstructing the events that caused them.

Algorithm 1: Event reconstruction

Given: Data records describing activities of objects
$<\{id_i,\}\ x_i, y_i, t_i\ \{, attr_i\}>$
where x_i and y_i are spatial coordinates, t_i is a time reference, id_i is an object identifier, and $attr_i$ is a set of attributes of activities.

Description of the algorithm:

1. Divide the space $\{x_{min}, y_{min}, x_{max}, y_{max}\}$ into non-overlapping polygons of the size suitable for the intended spatial scale of analysis in a way reflecting the distribution of activities.
2. Divide the time $\{t_{min}, t_{max}\}$ into non-overlapping intervals and calculate the amount of activities for each polygon and time interval.
3. Discover significant events and places with periodically varying activities by means of interactive analysis using visualizations and computational methods. Describe the discovered events and places using the attributes associated with the activity records.
4. If some areas exhibit non-standard patterns and are interesting from the domain perspective, repeat the procedure for these areas with refined spatial and/or temporal resolution.

All steps of the algorithm 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, while it is more convenient to use space compartments with approximately equal sizes and convex shapes. We have developed a method that divides a territory into convex polygons of desired size on the basis of point distribution [3]. We use a special algorithm for spatial clustering of points that finds round-shaped clusters with the desired radius. When a real point cluster has a larger size and/or elongated or non-convex shape, the algorithm divides it into several round clusters. The centroids of the clusters so obtained are used as generating points for Voronoi polygons. The centroids are the points with the minimal average distance to the cluster members. They are 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 radius is selected depending on the size of the studied territory and the goals of the analysis. The tessellation method is very efficient: processing of a 20,000 points sample takes about 3-5 seconds on a standard PC. If the analyst finds that the results do not reflect important geographical features, he/she may run the method several times with different parameters until the results are satisfactory. More strictly, the computational complexity of the point clustering algorithm is $O(n)$, where n is the number of points [3]. The subsequent Voronoi tessellation can be done in $O(k \log k)$ time [9], where k is the number of cluster centroids [9].

Our approach may be criticized for using a static (time-invariant) territory division. It is true that applying the method to data subsets from different time intervals may result in different tessellations and that any static partitioning may hide significant temporal dynamics of clusters. However, an advantage of static division is that it enables data aggregation and subsequent use of statistical analysis techniques devised for time series. An alternative approach would 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. Another advantage of static division is that it facilitates the detection of events re-occurring in the same places and of periodic patterns of activities.

Another point that may raise questions is the use of a fixed radius for obtaining point clusters. An important advantage of this approach is that the territory division can be conveniently adjusted to the desired spatial scale of analysis. 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 data 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 aggregation.

3.2 Spatio-temporal aggregation

Depending on the goals of analysis, the user selects the time period of interest and divides it into suitable temporal intervals. For the areas resulting from 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 desired spatial resolution, 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 desired temporal resolution. Such amount of data may be excessive for purely visual analysis. We apply two computational methods for finding time series and time moments of interest:

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 $ccf(\tau, x)$ 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 $ccf(\tau, x)$, we compute a normalized periodicity score as $(1/(|x|-1) \cdot ccf(\tau, x)/\sigma_x\sigma_\tau)$. 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 identifies abrupt peaks (increase followed by decrease) or pits (decrease followed by increase) within the given time window. This is a modified version of [5]. It has two parameters: minimum amplitude δ and maximum width w (we assume for simplicity that w is even). The algorithm will identify 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.

Algorithm 2: Peak detection

Given: time series x, amplitude δ, width w
Output: {(Apos[n], n) | for all x[n] peaks}

Description of the algorithm:

```
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(Apos [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 window (in lines 8-9) to verify that the definition given above holds. Therefore, the algorithm complexity is O(w|x|).

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, time series events may indicate events that occurred in the real world. Each t-event detected in a spatial time series refers to a certain region and a certain time interval.

To find out what real events stand behind t-events, the user may request additional 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 sequences.

3.4 Interactive visual displays

For obtaining an initial overview of the data, the time graph display (figure 2 top) is used. In its usual form, this display suffers from overplotting. To overcome this problem, we use 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 [2]). Both variants of the time graph support interactive data processing functionality:

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

Results of the periodicity detection algorithm are presented on a scatter plot that shows the absolute correlation scores against the normalized values. This plot is used for selecting time series of interest to be presented on the time graph. The spatial positions of these time series are indicated on the map.

The peak/pit detection procedure produces t-events positioned in time and space. The spatial positions are indicated on the map and the temporal positions on the time graph display. The space-

time cube display shows the positions in space and time simultaneously, using two horizontal dimensions to represent space and vertical dimension to represent time [12].

The time graph display 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. Additionally, a display called periodicity chart reflects the cyclic structure of the time. Thus, the periodicity 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. The periodicity chart in figure 11 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 3 and for the years in figure 11. The horizontal bar in the bottom represents the totals for the columns, i.e. for the same hours of different days in figure 3 and for the same weeks of different years in figure 11.

4 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 time span of the data is 9 days starting from Thursday and ending on Friday next week. We used a sample of about 10,000 calls for producing the tessellation with the radius 500m. This resulted in 238 areas. For these areas, we computed hourly counts of calls for 9 days, which gave us time series of the length 216 hours (figure 2). In the course of the analysis, we refined the temporal resolution to 1 minute for selected areas and time intervals, see section 4.3.

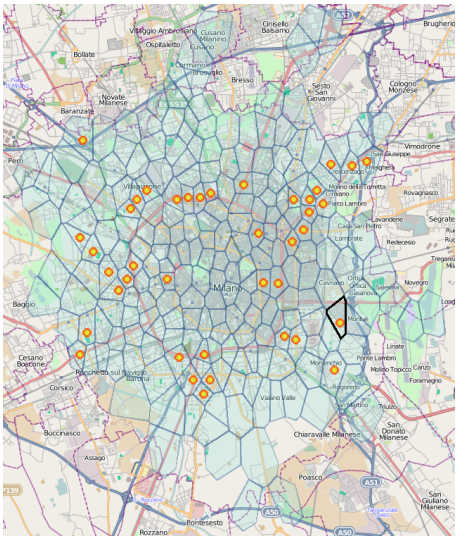

Figure 1. Tessellation of Milan with the radius 500m. A polygon of interest is highlighted. Circles indicate locations of peaks.

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.

4.1 Initial overview

The time graph and summary statistics display (figure 2) indicate that the calling behavior on Saturday and Sunday differs from that in the working days. In all working days, there are two sharp increases of calls, one at lunch time and the other in the evening. The magnitudes slightly differ in different days.

Figures 1 and 3 show the spatial and temporal positions, respectively, of the peaks with the magnitude of at least 100 calls over 3 hours. These 187 peaks occurred in 40 distinct time series at 52 different time moments; hence, the peak events are quite frequent. The periodicity chart in figure 3 shows that

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

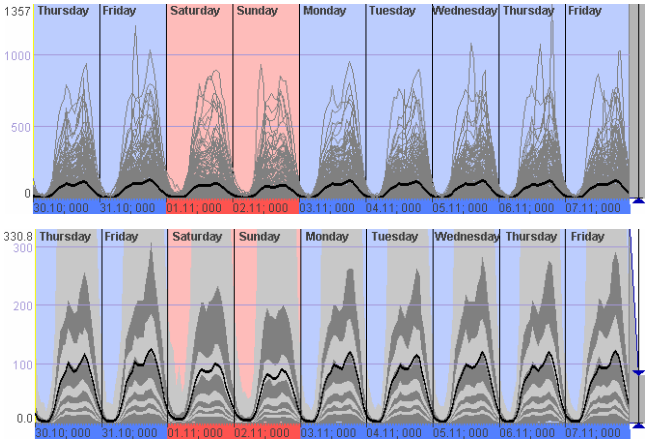

Figure 2. A traditional time graph display (top) and a statistical summary (bottom). The weekend is marked in light red.

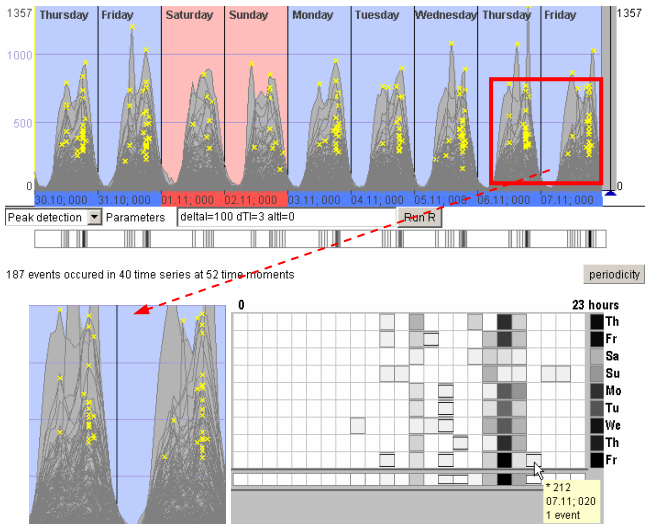

Figure 3. Positions of peaks are marked on the time graph by yellow crosses. The area within the red rectangle is enlarged in the bottom left. The periodicity chart is in the bottom right.

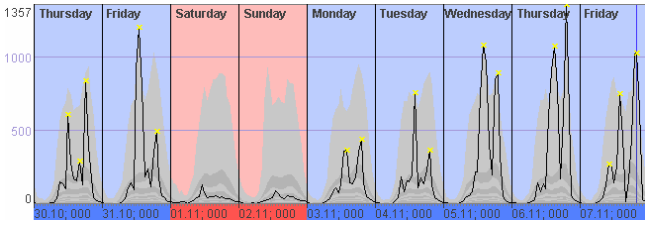

Figure 4. We used the periodicity chart (figure 3) for selecting the time series event in the evening of Friday, November 7. The corresponding time series is shown on the time graph.

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 highlighted in figure 1. The time series for this region (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. By inspecting the map, we found that a major television studio is located in this area. The unusually late peak of the phone calls may indicate that the employees had to work longer that day.

4.2 Detailed analysis in space

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 of before peaks (figure 5). There are also several cases of standalone pits. One example is shown in figure 6. The number of calls suddenly dropped from the usual about 200 calls per hour to 0 calls, which lasted for several hours. A similar drop occurred simultaneously in one of the neighboring areas. One more drop occurred nearby on the next day. These t-events may indicate technical problems or maintenance works in the real world, or, perhaps, missing data.

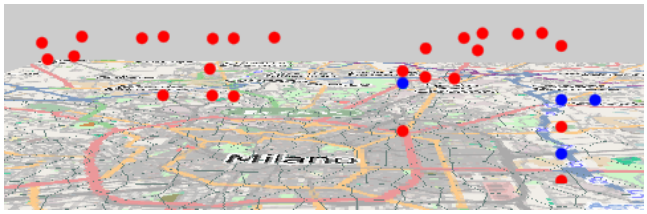

Figure 5. A fragment of the space-time cube showing the positions of the peaks (red dots) and pits (blue dots). The position of the movable map plane corresponds to 10:00 on November 7.

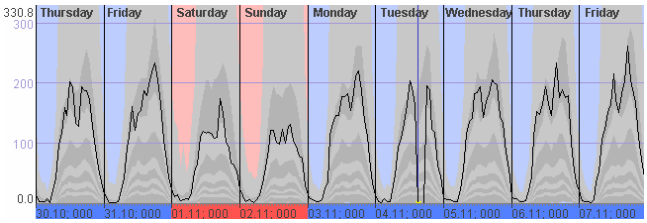

Figure 6. A pit has happened on Tuesday.

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 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 calls in move 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 studied in detail the time series with the smallest correlation values. One of them corresponds to an area on the north-east near a train station, where a big parking lot is located. Figure 7 exhibits very high peaks of calls on Saturday and Sunday. The majority of them are stationary. Possibly, an event like a flea market took place there. Other non-periodic time series, shown in figure 8, refer to three neighboring areas containing the stadium Giuseppe Meazza and a few other sport arenas. Two time series have high peaks on Sunday and Thursday, the third one has a peak on Saturday. Probably, the peaks correspond to sport or cultural events. For a detailed analysis, we aggregated the calls in the stadium area with one minute temporal resolution for the time intervals when the peaks occurred.

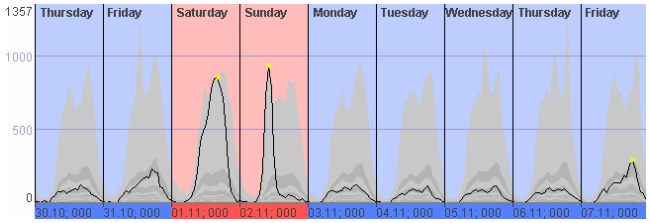

Figure 7. Non-periodic example: parking on North-East.

4.3 Detailed analysis in space and time

Figure 9 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 (both before and after that). The calls in the peaks before 22:30 are mostly stationary while after 22:30 they are mostly on the move. There are very few calls 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 whereas the peak between them occurred in the break. The profile of the second half differs from that of the first half, showing a high calling activity at the end of the game. We found that a national championship match[1] started in the stadium on Sunday at 20:30. It was attended by about 50,000 spectators. A single goal was scored at the end of the game, which explains the peak of phone calling.

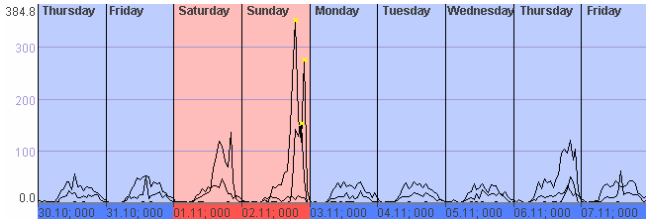

Figure 8. The 9-days time series for the areas close to the stadium.

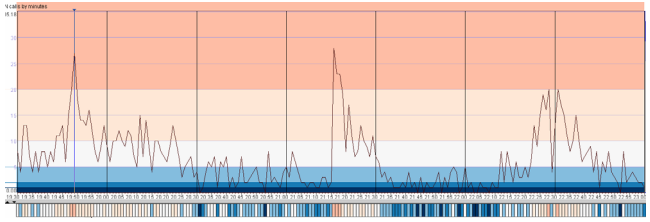

Figure 9. One-minute aggregates for the area near the stadium. The color band in the bottom shows the dynamics of the calls.

Another game (international, about 11,000 attendants) took place on Thursday[2]. 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.

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

4.4 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 (like in figure 7). Such findings may be useful for emergency management, transportation planning, and maintenance of public order.

5 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 20,000 photos for producing a tessellation with the radius 50km. This resulted in 164 regions (figure 10). For the regions, the time series of the weekly attendance by different photographers were computed. The length of the time series is 261 weeks. In the region around London, the values are much higher than elsewhere. We analyzed this area separately at a finer spatial scale (section 5.3).

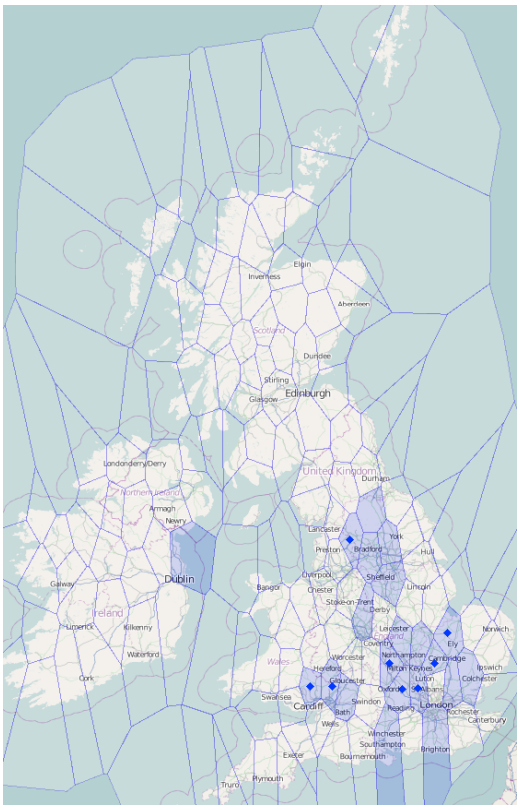

Figure 10. The study area divided into 164 regions. The shading marks 21 regions with snow-related peaks in February 2009; the dots mark 8 regions with snow in February 2007.

Like in the first case study, we applied two computational methods: temporal correlation check with the target period of 52 weeks and peak detection with the peak amplitude of at least 20 different people. 411 peaks were found in 74 regions in 112 weeks (figure 11). For interpreting the peaks, we requested the system to extract the most frequent words and phrases from the titles of the respective photos stored in the database. 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 provides access to the attributes of this t-event, in particular, to the frequent words and phrases extracted from the photo titles.

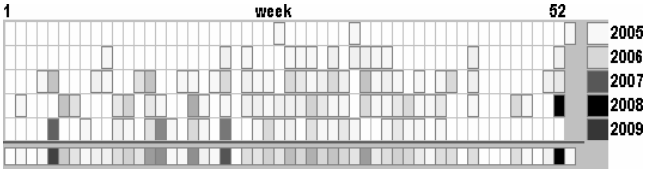

Figure 11. The periodicity chart shows the counts of the peaks in the number of flickr photographers by weeks (columns) and years (rows). The rightmost column shows the yearly totals, the row in the bottom represents the totals by the corresponding weeks of the different years.

5.1 Initial overview

The periodicity chart (figure 11) 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 annual event counts in the rightmost column. In the other dimension of time, 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 especially 2009. The regions are shown in figure 10. The blue shading marks the regions with peaks in 2009 and the dark blue dots mark the regions that had peaks in 2007.

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 12). Like with the periodicity chart, it is possible to select t-events and the time series in which they occurred by clicking on the elements of the linear event bar.

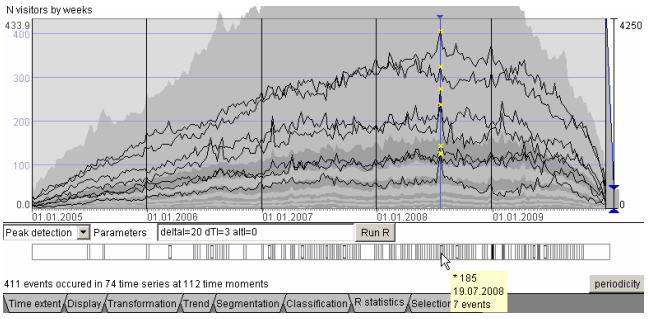

Figure 12. The time graph highlights the time series that had peaks in the week starting on 19.07.2008.

Figure 12 corresponds to the selection of the week of 19.07.2008, in which peak events occurred in six areas. By accessing the peak attributes (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 regions.

5.2 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. For many areas with high periodicity scores, the time series profiles are similar to the one shown in figure 13 top, which has clear seasonal differences but no significant peaks. These areas are located mostly on the coastline and in the rural regions. This corresponds to our expectations.

We also expect that some regions have regular public events that occur at about the same time every year. Indeed, several regions with high periodicity scores have high peaks in particular weeks of each year and low variation in the remaining time, as in figure 13 bottom. The statistical summaries of the photo titles suggest 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).

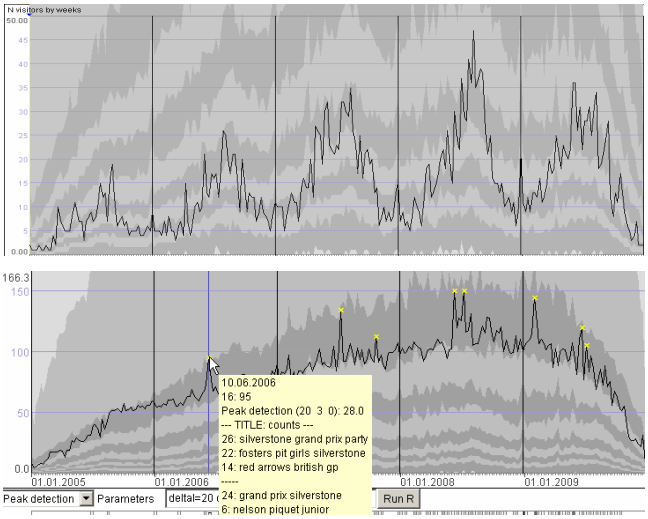

Figure 13. A periodic time series in Scotland (top) and periodic peaks of the Silverstone Grand Prix event (bottom).

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 14 summarizes our findings.

5.3 Refining analysis for a selected region

Now we focus on the photos taken in the London area. We use a sample of about 20,000 photo positions (out of about 2 million) for tessellation with the radius of 1km, resulting in 1,236 polygons. For the polygons, we again compute the weekly time series of the counts of different photographers. The peak detection procedure with the threshold of 20 people above the regular values extracted 158 peak events referring to 42 regions and 77 distinct weeks. By means of the correlation detection procedure, we found a number of areas with highly periodic behaviors of the time series. Some of them are parks and open spaces with clear seasonal patterns. Other places are the locations of annual public events or shows such as Chinese New Year (China town, February), Chelsea (flower show, May), Wimbledon (tennis competitions, June), Hide Park (festival, June), Biggin Hill (air show, June), Canary Wharf (motor show, July), Notting Hill (carnival, August), Arsenal stadium (Emirates Cup, football, August), and others. The events were identified by accessing the frequent words and phrases from the titles of the photos.

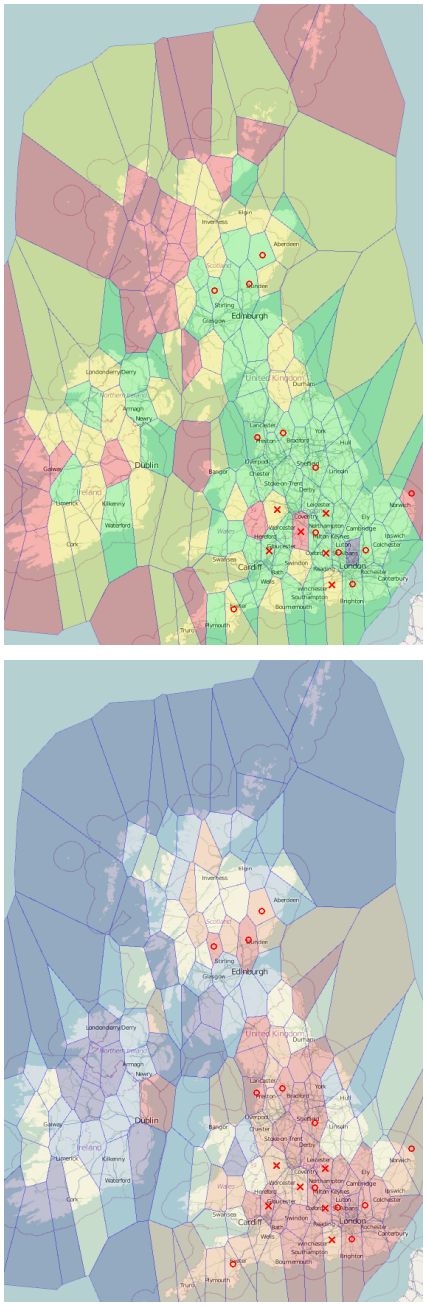

Figure 14. Top: the regions classified according to the periodicity of the time series: low (green), medium (yellow) and high (red). Bottom: the counts of photos in the regions are shown by colors, from blue (few) to red (many). The symbols mark the regions with periodic (crosses) and irregular (circles) peaks.

We found also irregularly re-occurring events at the Wembley stadium and several occasional events such as Yahoo hack day (Alexandra Palace, June 2007) or London Marathon (in several places, April 2009). In July 2007, the Tour de France race resulted in simultaneous activity peaks in several regions. The exceptional snowfall in February 2009 was reflected by numerous peaks in London like in other regions.

5.4 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 needs to be designed for this purpose, which is quite feasible.

6 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 [19] - we implemented several workflows that may be arbitrary 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);
- enabling flexible workflows for interactive analysis;
- providing novel visualizations (periodic event bar) and coordination mechanisms (linking elements of different data sets and of different nature, specifically, regions 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. Another direction of our work is making the tools accessible to general public in the Web 2.0 environment and integration with search engines for acquiring relevant data and interpreting discovered facts. We consider involving in the analysis landmark data bases and geocoding services.

References

[1] G.Andrienko, N.Andrienko, M.Mladenov, M.Mock, Ch.Poelitz. Extracting events from spatial time series. In: IV 2010

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

[3] N.Andrienko, G.Andrienko Spatial Generalization and Aggregation of Massive Movement Data. IEEE Transactions on Visualization and Computer Graphics, 2010

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

[5] E.Billauer. peakdet: Peak detection using MATLAB. *Online, http://www.billauer.co.il/peakdet.htm*. Retrieved 26 February, 2010

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

[7] 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;

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

[9] 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

[10] 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

[11] 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

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

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

[14] 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

[15] 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

[16] P.Jankowski, N.Andrienko, G.Andrienko, S.Kisilevich. Discovering landmark preferences and movement patterns from photo postings. Transactions in GIS, under review

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

[18] 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. Volume 4950 of LNCS, Springer, 2008, pp.154-175

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

[20] 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

[21] 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

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

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