A Framework for Using Self-Organizing Maps to Analyze Spatio-Temporal Patterns, Exemplified by Analysis of Mobile Phone Usage

Gennady Andrienko[1], Natalia Andrienko[1], Peter Bak[4], Sebastian Bremm[2], Daniel Keim[4], Tatiana von Landesberger[2,3], Christian Pölitz[1], Tobias Schreck[2]

[1] University of Bonn & Fraunhofer IAIS, Germany
[2] Technische Universität Darmstadt, Germany
[3] Fraunhofer IGD, Germany
[4] University of Konstanz, Germany

Give full correspondence details here

(Received 28 April 2010; final version received 21 September 2010)

We suggest a visual analytics framework for the exploration and analysis of spatially and temporally referenced values of numeric attributes. The framework supports two complementary perspectives on spatio-temporal data: as a temporal sequence of spatial distributions of attribute values (called spatial situations) and as a set of spatially referenced time series of attribute values representing local temporal variations. To handle large amounts of data, we use the Self-Organizing Map (SOM) method, which groups objects and arranges them according to similarity of relevant data features. We apply the SOM approach to spatial situations and to local temporal variations and obtain two types of SOM outcomes, called space-in-time SOM and time-in-space SOM, respectively. The examination and interpretation of both types of SOM outcomes are supported by appropriate visualization and interaction techniques. The paper describes the use of the framework by an example scenario of data analysis. We also discuss how the framework can be extended from supporting explorative analysis to building predictive models of the spatio-temporal variation of attribute values. We apply our approach to phone call data showing its usefulness in real-world analytic scenarios.

Keywords: geovisualization; spatio-temporal data, visual cluster analysis

Introduction

Large amounts of data associated with positions in space and time are difficult to visualize and explore. Existing visualization techniques such as animated map, space-time cube, and time graph dynamically linked to map [Andrienko et al. 2003] become ineffective when the number of different places and times (i.e. moments or intervals) increases. Clustering is one of the standard approaches to dealing with large datasets: the amount of the data is reduced by means of uniting similar data items into groups (clusters) such that the internal data variability is significantly lower than inter-cluster differences. This approach requires solving several problems: first, developing an appropriate measure of similarity depending on the data structure and the analysis goals; second, choosing or devising a suitable clustering algorithm; third, representing the results of the algorithm to a human analyst in a way enabling interpretation and exploration, that is, in a visual way. Obviously, the visualization also depends on the structure of the data and the analytical task.

In the research reported in this paper, we deal with spatially- and temporally-referenced data whose structure can be formally represented as $S \times T \rightarrow A$, where S

stands for space (set of places), T stands for time (set of moments or intervals, jointly called time units), and A is the thematic, or attributive, component of the data, possibly, multi-dimensional (Andrienko and Andrienko 2006). Note that S and T in this formula are independent components whereas A depends on S and T. This can formally represented as $a=f(s,t)$, where *a* is a variable representing various attribute values or combinations (in case of multi-dimensional attributes), *s* represents various places, *t* represents time units, and *f* symbolizes a function $f: S \times T \rightarrow A$. For any chosen place $s \in S$, i.e. value of variable *s*, there is a respective $T \rightarrow A$, i.e. the series of values of variable *a* corresponding to the possible values of variable *t*. We shall call this time series local temporal variation in place s. Similarly, for any chosen time unit $t \in T$, i.e. value of variable *t*, there is a respective $S \rightarrow A$, i.e. the set of values of variable *a* corresponding to the possible values of variable *s*. We shall call this spatial distribution of attribute values spatial situation in time unit t.

Accordingly, a spatiotemporal dataset can be considered from *two complementary perspectives*: as a set of local temporal variations ($T \rightarrow A$) distributed over space and as a temporal sequence of spatial situations ($S \rightarrow A$). Which perspective to take, depends on the analysis goals. For a comprehensive exploration, it may be necessary to consider the data from both perspectives.

We have developed a visual analytics framework for the exploration of large datasets consisting of spatially and temporally referenced values of numeric attributes with the use of clustering. We have chosen the Self-Organizing Map (SOM) (Kohonen 2001) as the clustering algorithm. Depending on the taken perspective, the SOM is applied either to local temporal variations or to spatial situations. The similarity of objects is measured in terms of the Euclidean distance (or, more generally, Minkowski distance) in a multi-dimensional space of feature vectors. Hence, we represent local temporal variations and spatial situations by suitable feature vectors. We have designed visualization techniques to display the outcomes of the SOM in such a way that the spatial, temporal, and attributive components of the data can be seen and explored by an analyst.

The main features of the software tools have been earlier described by Andrienko et al. (2010). The focus of this paper is the general framework for the analysis of spatio-temporal data using self-organizing maps. The framework is exemplified by exploring a large dataset about mobile phone calls in Milan (Italy) aggregated by 238 spatial compartments and 216 hourly time intervals (the data have been provided by the Italian telecommunication company WIND).

The remainder of the paper is structured as follows. The next section gives an overview of the related works. After that, we briefly introduce the tools and then describe the framework by example of exploration of the phone calls data. Then we summarize the suggested framework in the form of flow chart. This is followed by a discussion of the possible use of the exploration results in spatiotemporal modelling.

Related works

The SOM methodology is discussed in depth in Kohonen's monograph (2001); a brief introduction is given by Skupin and Agarwal (2008). Vesanto (1999) describes the basic analytic tasks that can be addressed and possible visualizations supporting these tasks. The tasks include analysis of cluster structure, of prototype vectors, and of overall data distribution.

The SOM method is applicable to any data type that can be represented by feature vectors. Specifically, complex and multimedia data can be addressed by the SOM if they are represented by appropriate feature vectors. Thus, the SOM to date has been successfully applied, for example, to financial data (Deboeck and Kohonen 1998), texts (Nuernberger and Detyniecki 2006), images (Barthel 2008), and time-dependent scatter data represented as trajectories in two-dimensional abstract space of attribute values (Schreck et al. 2009). There are also numerous applications of the SOM to geospatial data; many of them are described in (Agarwal and Skupin 2008). Colour-coding is used to link the locations of data elements in the SOM space with the respective positions on a cartographical map (Koua and Kraak 2008, Spielman and Thill 2008) and elements of other displays such as lines in a parallel coordinates plot and cells in a reorderable matrix (Guo et al. 2006). Besides using simple linearly scaled two-dimensional colour maps, approaches exist for advanced colour mappings adjusting for non-uniform distributions of the SOM distances (Kaski et al. 2000) and considering perceptual issues (Guo et al. 2005).

Skupin (2008) describes an interesting way of linking the SOM space to geographical space and time: a trajectory made by a person in the geographical space is projected onto the SOM space where geographical places are arranged according to their similarity in terms of multiple attributes. Variants of the SOM algorithm exist that include also geospatial coordinates in the SOM training process, allowing a trade-off of multivariate and geospatial data properties in the obtained SOM (Bação et al. 2003).

The SOM has been also applied to spatiotemporal data with the structure $S \times T \rightarrow A$, which is in our focus. Guo et al. (2006) apply the SOM to combinations of values of multiple attributes characterizing pairs <place, time unit>. Assessing similarities and differences among spatial situations is done by visual inspection of multiple maps (one map per time unit) where each place has the colour of the SOM cell containing the particular combination of this place and the time unit. Assessing similarities and differences among local temporal variations is done using a reorderable matrix where the rows correspond to the places, columns to the time units, and cells have the colours of the SOM nodes. Hierarchical clustering groups the rows by similarity.

Hewitson (2008) applies the SOM to time series of spatial distributions of air pressure values in order to find the archetypal distributions for a region and then looks for certain temporal patterns such as frequencies of the archetypes in dry and wet years. Hewitson does not consider the complementary analytic task, analysis of the spatial distribution of the local temporal variations.

Presentation of the tools

Here we briefly introduce our software tools; a more detailed description is given in (Andrienko et al. 2010). We use the implementation of the SOM algorithm available in the SOM_PAK package (Kohonen et al. 1996). The input of the algorithm is a set of feature vectors, i.e. combinations of values of multiple attributes. The output is a two-dimensional layout consisting of rectangular or hexagonal cells. The hexagonal topology is recommended for an improved quality (isotropy) of the display when the purpose is visualization of data spaces (Kohonen 2010). Our main purpose is visualization of clusters. We use the rectangular grid topology, which is more convenient for our display design.

Each cell of the SOM grid corresponds to a certain vector, called prototype vector. Prototype vectors are not elements of the input data but are constructed during the unsupervised learning process according to the SOM method. The input data vectors are assigned to the cells with the closest prototype vectors. Therefore, the cells can be interpreted as clusters. As a key property provided by the SOM method, the output cells themselves are positioned in the two-dimensional layout (in our case matrix) according to the similarity of their prototype vectors. Hence, the SOM not only groups data but also produces a layout that is very convenient for visualization.

The desired numbers of rows and columns in the grid can be chosen by the user. As explained by Bação et al. (2008), larger grids should be used for the purpose of exploring the data distribution and small grids are used when the user is interested in clustering, as in our case.

We apply the SOM method to two types of complex objects: (1) spatial situations $S \rightarrow A$ that occurred in different time units $t \in T$, and (2) local temporal variations $T \rightarrow A$ that occurred in different places $s \in S$. The feature vector representing a spatial situation in one time unit consists of the concatenated attribute values corresponding to this time unit and all places. The feature vector representing a local temporal variation in one place consists of the attribute values corresponding to this place and all time units. The input data are normalized by the SOM tool.

The primary visualization of the SOM output in our system is the SOM matrix display. When the SOM is applied to spatial situations, the resulting matrix is called ‘space-in-time SOM’. The result of applying the SOM to local temporal variations is called ‘time-in-space SOM’. Colouring of the cells in the SOM matrix display is used to link the matrix with additional data views for supporting multi-perspective data analysis. This will be demonstrated in the example analysis scenario.

The cells in the matrix display may include two types of images: feature images and index images. Feature images represent the objects to which the SOM tool has been applied, i.e. spatial situations in a space-in-time SOM and local temporal variations in a time-in-space SOM. Spatial situations are represented by maps (Fig.1) and local temporal variations by diagrams called ’temporal mosaics’ (Fig.2). A map image portrays the attribute values attained in all places in one time unit. A temporal mosaic portrays the attribute values attained in one place in all time units. Each pixel corresponds to one time unit; the pixels are arranged in rows of user-specified length. Values of space and time-dependent numeric attributes are represented by colour coding. In all illustrations given in this paper, one of the scales from Color Brewer (Harrower and Brewer 2003) is used. The colour scale and the division of the attribute value range into intervals can be explicitly specified by the user. If this is not done, the system automatically divides the value range into a default number of equal intervals and chooses a diverging colour scale with the midpoint corresponding to the central interval. Diverging colour scales are used for increasing the visual salience of the feature images. The feature images are not meant to support accurate decoding of the attribute values (which can be better done using other visualization techniques available in the system) but for a comparative overall assessment.

Index images show the temporal or spatial positions of the objects included in the SOM matrix cells. In a space-in-time SOM, index images show the temporal positions of the spatial situations (Fig.1). An image consists of small squares representing the time units, which are temporally ordered and arranged in rows of user-chosen length. The squares representing the objects included in the respective

SOM cell are filled in black. In a time-in-space SOM, index images show the spatial positions of the local temporal variations. Each image is a map where the spatial positions are marked by black filling of the corresponding territory compartments (Fig.2). The combination of feature images and index images provides a combined representation of the space, time, and attribute values. The user may arbitrarily switch on and off the drawing of the feature images and the index images.

The number of objects included in a SOM cell is represented graphically by the proportional length of the filled segment in the bar drawn on top of the cell. The colour of the bar, white or grey, depends on the darkness of the background colour of the cell. We remind that the purpose of the colouring of the SOM cells is to support visually cross-referencing data elements over multiple complementary views.

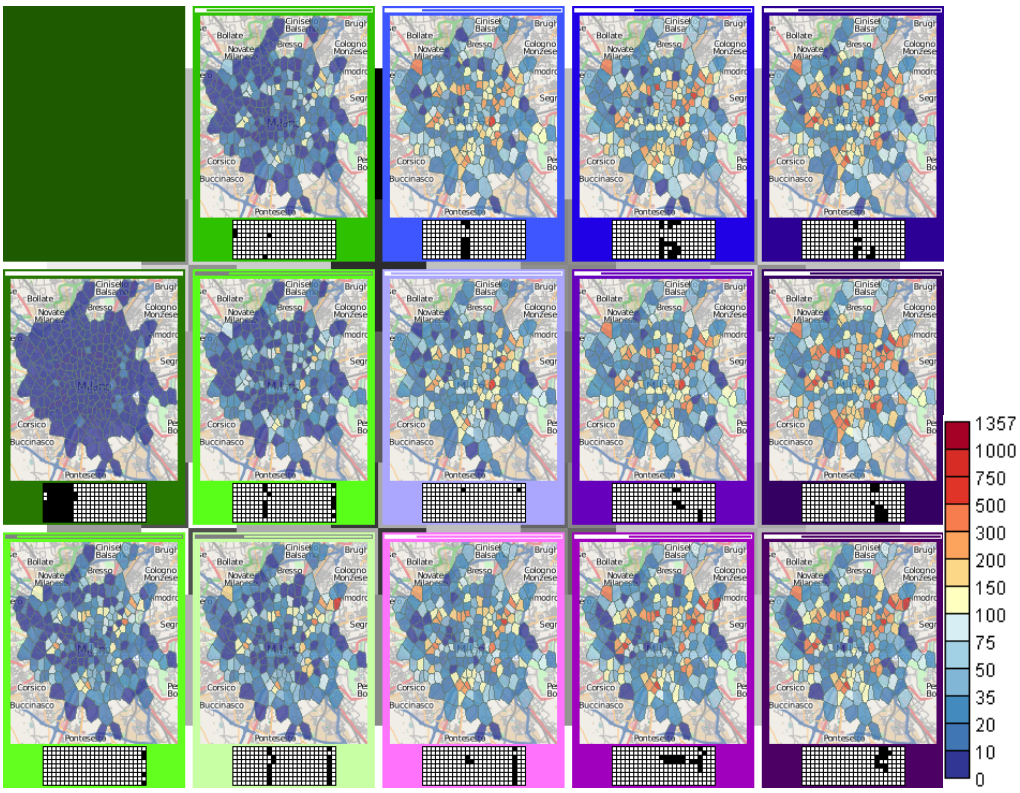

Figure 1: Space-in time SOM matrix. The maps displayed in the cells are feature images representing spatial situations. Below the maps are the index images. Each index image is a matrix with 24 columns representing hours of a day and 9 rows representing consecutive days. The filled squares show the temporal positions of the spatial situations included in the cells.

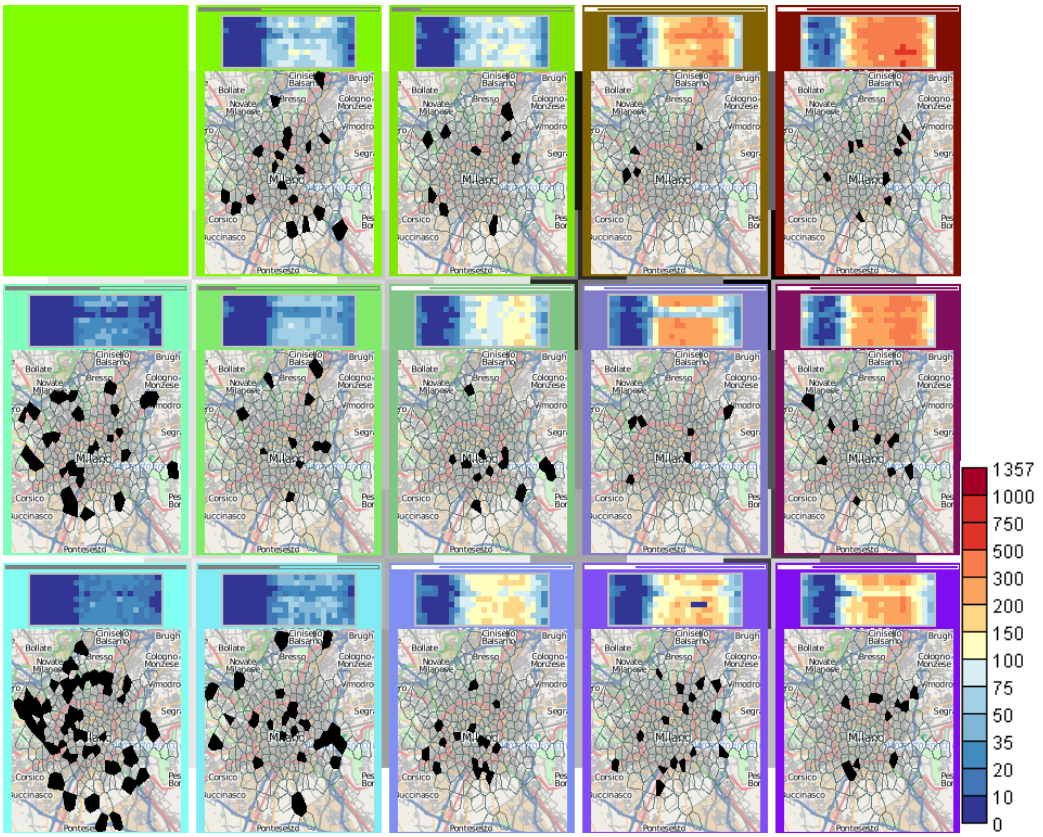

Figure 2: Time-in-space SOM matrix. Inside each cell at the top is a ‘temporal mosaic’, which is a feature image showing the temporal variation of the attribute values. The columns correspond to 24 hours of a day and the rows to 9 days. The maps serve as index images: the spatial positions of the temporal variations included in the SOM cells are shown by black filling of the respective areas.

Figs.1 and 2 demonstrate the use of two different two-dimensional colour scales for the colouring of the SOM matrix cells. The scale in Fig.1 has been obtained by re-projecting the prototype vectors onto an abstract two-dimensional space according to the distances among them by means of the Sammon’s projection algorithm (Sammon 1969) and applying polar colour mapping. This means that the positions in the projection space are expressed in polar coordinates relative to the centre of the bounding rectangle. The angle determines the hue and the distance determines the saturation and brightness. The scale in Fig.2 has been obtained by creating a colour matrix with 10 times more rows and 10 times more columns than in the SOM matrix. The colours of the cells are generated by means of rectangular colour mapping: four distinct colours are put in the corners of the matrix, and the colours for the remaining cells are obtained by mixing the primary colours proportionally to the distances from the cells to the corners. The prototype vectors from the SOM matrix are iteratively placed within this larger colour matrix according to the distances to their previously placed neighbours.

Optionally, the SOM matrix display can convey additional information about the pair-wise distances between neighbouring SOM cells in the attribute space, known as U-matrix (Ultsch 1999). The distances are represented by shading of the cell borders (Figs.1 and 2). The border of a cell is divided into 8 segments corresponding

to the 8 neighbours of this cell. The degree of darkness of each segment between white and black is proportional to the Euclidean distance to the respective neighbour in terms of the prototype vectors. Note that neighbouring cells with small distances between them have similar colours while the colours of the neighbouring cells with large distances between them differ quite much. The visualization of the U-matrix and the colouring of the grid cells support the analyst in assessing the possibly non-linear distance relationships between neighbouring regions in the SOM. To understand the U-matrix, the user needs to have some knowledge about the SOM method; otherwise, the visualization of the inter-cell distances may be confusing. Therefore, it is optional and appears only upon user's request.

A SOM matrix display with feature images in the cells was previously suggested by Schreck et al. (2009). The feature images were built by overlaying semi-transparent lines representing individual objects included in the SOM cells. In our case, overlaying is not meaningful. The feature images shown in the cells correspond to the objects whose feature vectors are the closest to the respective prototype vectors. Another option would be to represent averaged attribute values. Images of all objects included in a cell can be seen in an additional detail-on-demand window (e.g. Fig.6) which appears after the analyst clicks on the respective cell. Index images have not been used in the display of SOM outputs before.

The following section describes an example scenario of phone call data analysis with the use of the SOM-based framework. When necessary, we shall provide further explanations of the tools and techniques used.

Data analysis with the SOM

The SOM matrix displays demonstrated in Figs.1 and 2 have been obtained by applying the SOM tool to the mobile phone calls data from Milan. Our framework is applicable to spatio-temporal data having the structure $S \times T \rightarrow A$, where S is a finite set of places, T is a finite set of time units, and each pair of place and time unit is characterized by value(s) of one or more numeric attributes. The original data consist of records about individual phone calls characterized by the call time and position (plus some other attributes, such as call duration, which we do not use in our analysis). The structure of these data can be described by the formula $C \rightarrow S \times T$, where C is the set of calls, or $C \rightarrow S \times T \times A$ (A stands for the additional attributes of the calls). Note that S and T in this formula are dependent components. This data structure can be transformed to the structure $S \times T \rightarrow A$ by means of spatio-temporal aggregation.

We divided the territory of Milan into polygonal compartments. The time extent of the data is 9 consecutive days starting from Thursday and ending on Friday next week. This time was divided into hourly intervals, which results in 216 time units. For each compartment and time interval, we obtained a count of mobile phone calls, which gave us the required structure $S \times T \rightarrow A$. The counts range from 0 to 1357; however, very high values occur in only a few compartments. Therefore, we encode count values by colours in a way that emphasizes differences between small values. The colour legend is shown in the bottom right corner of Figs.1 and 2. The elements of the index images in Fig.1 and feature images (temporal mosaics) in Fig.2 are arranged in rows of the length 24 so that each row represents 24 hourly intervals of one day.

In aggregating data, there are many possibilities for dividing the space and time. The analyst should be aware of the possible scale sensitivity of the analysis, in

particular, of the modifiable areal unit problem (Openshaw 1984). We suggest that the analyst should check whether the same or similar patterns are observed after changing the partitioning of the space and of the time. Our software allows the user to re-aggregate data quickly and easily. However, a detailed discussion of the issues of choosing appropriate spatial and temporal scales of analysis and delineation of the spatial and temporal units is out of the scope of this paper.

For a comprehensive analysis, it may also be appropriate to apply various transformations to the values of the attributes (Andrienko and Andrienko 2006). In the example described in this paper, we deal with the absolute counts of the phone calls. The analysis should be continued by applying the SOM also to relative values, in particular, normalized deviations from the averages over time by place and averages over space by time. These and other transformations of attribute values can be easily done in our system. However, we shall not describe the transformation tools and the application of the SOM to the transformed data to avoid increasing the size of the paper and blurring its focus.

Exploring the changes of the spatial situation over time

A spatial situation consists of the counts of the mobile phone calls made in all compartments during one hour. We apply the SOM to the spatial situations corresponding to all hourly intervals. As a result, the intervals are grouped and positioned in the SOM matrix according to the similarity (i.e. closeness in the attribute space) of the respective spatial situations. The matrix is shown in Fig.1. The feature images, in the form of maps, show representative spatial situations from the cells. The index images show the days and hours in which these and similar situations occurred.

As can be seen from the feature images in Fig.1, the cells painted in shades of green in the left part of the matrix include spatial situations with lower calling activity throughout the whole territory than in the middle and on the right of the matrix. The situations with the lowest activity are included in the dark green cell on the middle left; the neighbouring cells painted in light green shades include situations with somewhat higher call counts. The situations with the highest calling activity are on the right of the matrix, where the cells are pained in very dark shades of blue, violet, and purple. These observations are consistent with the information obtained from the time graph display with the background painted in the colours of the SOM matrix cells (Fig.3). Note that the plot area is painted in a semi-transparent mode, which makes the colours look paler than in the matrix. The true colours are used in painting the bar below the plot area. The variation of the number of calls in each area is represented in the time graph by a line. We can see that the values in the “green” intervals are much lower than those in the “blue”, “violet”, and “purple” intervals.

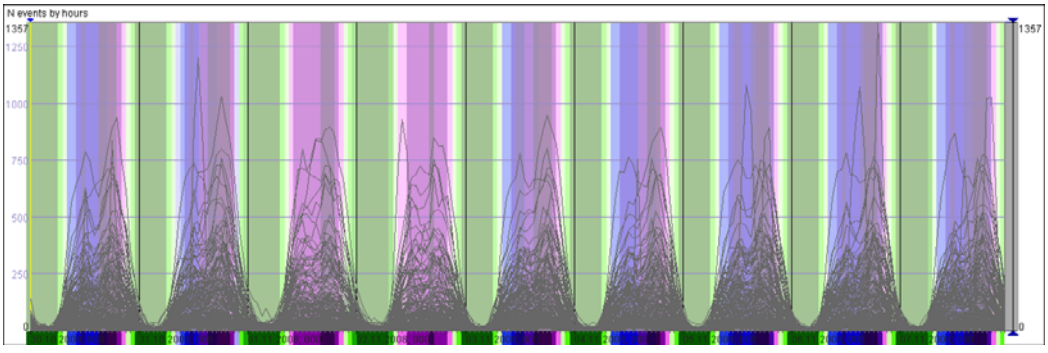

Figure 3: Time graph with the background painted according to the cell colours in the SOM matrix in Fig.1.

Hence, the horizontal dimension of the SOM matrix corresponds to numeric differences among the spatial situations. The differences among the situations included in the same column but different rows in the middle and right parts of the matrix are not immediately obvious. However, after a closer look at the feature images, we understand that the situations in the upper row (shades of blue) have higher call counts in some areas in the city centre than the situations in the lower row (shades of pink and purple). The situations in the middle row seem to be intermediate in this respect.

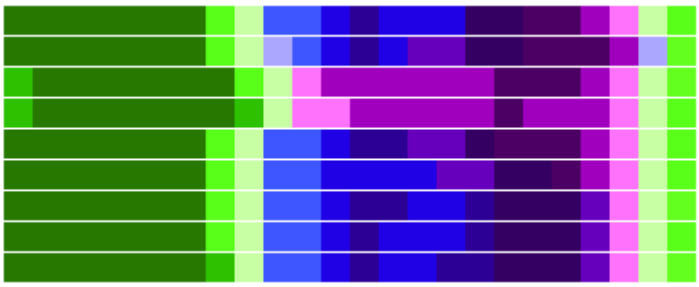

Figure 4: Time arranger corresponding to the space-in-time SOM shown in Fig.1.

The index images in the SOM matrix and the time graph demonstrate that the calling activity varies periodically according to the daily cycle while Saturday and Sunday (days 3 and 4) differ from the working days. An additional tool to explore the temporal variation is the time arranger, shown in Fig.4. As in the index images, the rectangular elements (pixels) represent time intervals and are arranged in rows. In our case, the rows have the length 24, corresponding to 24 hours of the day. Like in the time graph, the intervals are painted in the colours of the respective SOM cells (generally, a time arranger can represent any classes or clusters of time units). By mouse-pointing on a pixel, the user receives the information about the respective time unit in a popup window.

The periodic character of the temporal variation is manifested through the vertical arrangements of identically or similarly coloured pixels in the time arranger. The differences between the working days (rows 1-2 and 5-9) and the weekend (rows 3-4) are well visible. Knowing the meanings of the cell colours in the SOM matrix, we can describe the temporal variation of the phone calling activity as follows. On the working days, the calling activity is very low in hours from 0 to 6. It gradually increases in hours 7 and 8, then increases much more in hours 9-10 and 11 and remains high until hour 19; however, the calling activity in the centre decreases towards afternoon and evening. After 19, the number of calls gradually decreases. In the weekend, the calling activity in hour 0 is higher than on the working days. The period of low activity lasts from hour 1 to hour 7, i.e. it is shifted by one hour forward compared to the working days. The activity gradually increases from hour 8 to midday, remains quite high till hour 20, and then decreases. The maximal activity is in hours 17-19 on Saturday and hour 17 on Sunday. Shades of blue are absent in rows 3-4, which means that the increase of the calling activity in particular areas in the centre does not occur in the weekend.

The cyclic character of the temporal variation of the calling activity and the distinctions between the temporal patterns of the working days and the weekend could be expected. The consistence of the SOM outcome with these expectations demonstrates the validity of the approach. Besides exhibiting the expected general patterns, the SOM and related visual tools allowed us to do a detailed investigation of the temporal variation of the spatial situations.

Exploring the spatial distribution of the local temporal variations

The spatial distribution of the local temporal variations is studied with the help of a time-in-space SOM matrix, which is obtained by applying the SOM tool to the time series of the mobile phone call counts in all areas. The SOM groups and arranges areas by closeness of the respective time series. The example time-in-space SOM with 5 columns and 3 rows in Fig.2 is suitable for illustrative purposes since the cells are large enough. However, when we look at the statistics of the distances of the objects in the cells to the respective prototypes (Fig.5), we find out that a finer resolution of the SOM matrix would be more appropriate for the exploration. The statistics can be seen when the display of the feature and index images is switched off. Fig.5 shows the same SOM matrix as in Fig.2. The numbers in each cell represent, from top to bottom: the number of objects included in the cell, the maximum of the distances from the prototype vector of this cell to the prototype vectors of the neighbouring cells, the maximum of the distances from the prototype vector to the feature vectors of the member objects, and the average of the distances. When the maximum distance from the cell prototype to the cell members exceeds the maximum distance to the neighbouring cells, this is an indication that the cell contains at least one outlier, i.e. an object significantly dissimilar to the other objects in this cell. Thus, in the middle right cell, the maximum distance to the prototype is 2.6 times bigger than the maximum distance to a neighbouring cell. The cells above and on the left of this cell also contain outliers.

As mentioned earlier, clicking on a SOM matrix cell opens a window exhibiting the feature images of all objects included in the cell. Fig.6 shows the content of the middle right cell of the matrix presented in Figs. 2 and 5. The feature image corresponding to area 152 is very distinct from the others, which means that this area is an outlier. In the map in Fig.8, this area is marked by a thick black border. We inspect in the same way the contents of the other cells where the maximum distances to the prototypes are bigger than the distances to the neighbouring cells and easily detect the outliers among the temporal variations represented by the mosaic diagrams.

Hence, because of the low resolution chosen, the SOM algorithm could not properly place objects that differ much from the others, which means that the resolution of the matrix should be increased. We did not do this for the space-in-time SOM considered in the previous section: although there were outliers in a few cells, they did not differ so dramatically from the other cell members as in the time-in-space SOM matrix.

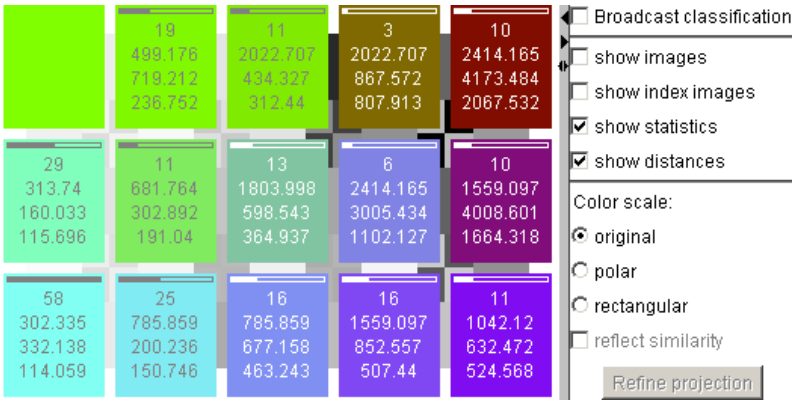

Figure 5: The time-in-space SOM matrix from Fig.2 in the mode of showing the statistics of inter-cell and intra-cell distances. The numbers in each cell represent, from top to bottom: the number of objects included in the cell, the maximum of the distances from the prototype vector of this cell to the prototype vectors of the neighbouring cells, the maximum of the distances from the prototype vector to the feature vectors of the member objects, and the average of the distances.

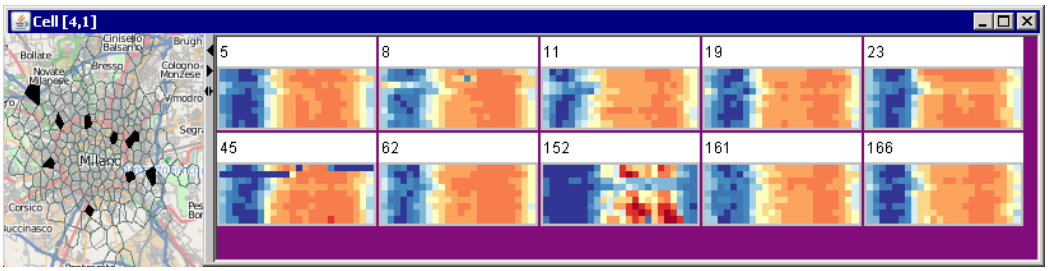

Figure 6: A window showing the content of the middle right cell of the SOM matrix presented in Figs.2 and 5.

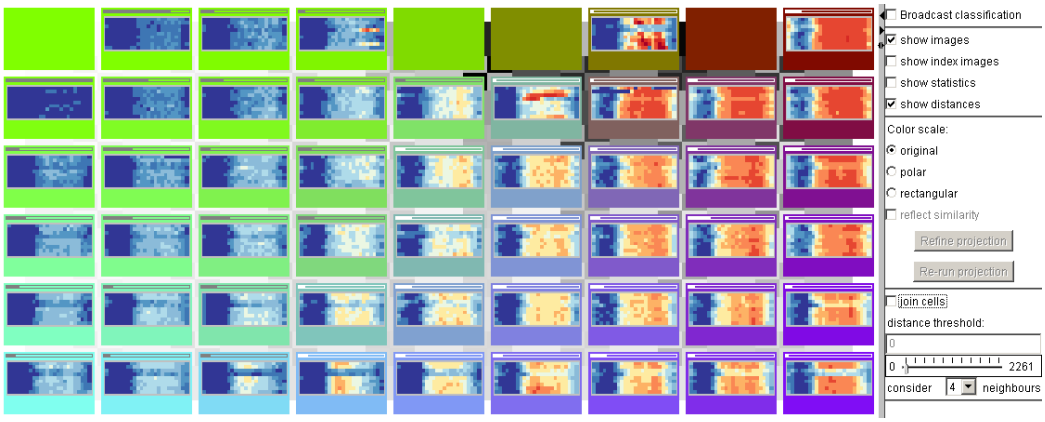

Figure 7: Time-in-space SOM matrix with 9 columns and 6 rows for the same data as in Fig.2.

Fig.7 presents the time-in-space SOM matrix with the resolution of 9 columns per 6 rows produced for the same data as the matrix in Fig.2 and Fig.5. The algorithm has placed the most distinctive outliers in separate cells. In particular, the feature image of area 152 can be seen in column 7, row 1. The colour of this cell differs significantly from the colours of the neighbours, which means that the temporal variation of the call counts in area 152 is very unusual. Unfortunately, we have no local knowledge to explain this and other outliers.

The feature images (temporal mosaics) in the matrix in Fig.7 show how the calling activity varies over time in the areas included in the cells. In the left part of the matrix, the call counts are quite low all the time, and in the right part, the counts are much higher, except for the night and early morning hours. There is a tendency that in the lower part of the matrix the values on the weekend (rows 3 and 4 of the mosaics) are notably lower than on the working days. In the upper part of the matrix, the working days and weekends have close values. In Fig.8, the colours of the matrix cells are used for painting the areas on the map of Milan. We can observe that there is no smooth spatial pattern in the distribution of the local temporal variations. Apparently, the temporal patterns depend on the major people's activities in the areas, i.e. whether the areas are residential, industrial, commercial, etc. We can guess that in the areas where the call counts significantly decrease on the weekend there are industrial enterprises or offices.

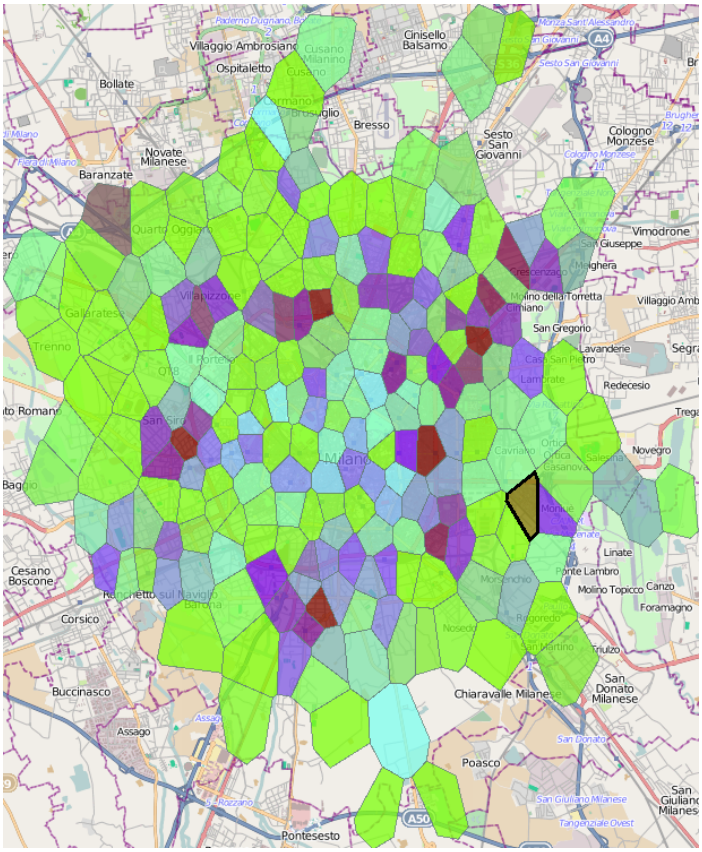

Figure 8: Map with areas coloured according to the positions in the time-in-space SOM matrix shown in Fig.7.

Returning back to the matrix in Fig.7, we can observe that the inter-cell distances, which are conveyed through the shading of the cell borders and through the degree of similarity of the cell colours, are much lower in the left part of the matrix than on the right. Hence, the objects in the left part are much closer to each other in the attribute space than those in the right part. A projection algorithm that is not constrained by a predefined grid layout would put more similar objects closer in two-dimensional space than less similar objects. For example, Fig.9 shows the time series of the call counts projected onto continuous two-dimensional space by means of the Sammon's projection algorithm (Sammon 1969). Each area is represented by a dot (circle) having the colour of the SOM cell in which this area was placed, like in Fig.8.

It can be seen that light green and light cyan dots representing areas with low counts are put very close together in the Sammon's projection. Hence, the SOM can better reveal distinctions among objects with low values than the Sammon's projection. Another advantage of the SOM over continuous projection techniques, such as Sammon's projection, MDS, and PCA, is a better use of the display space.

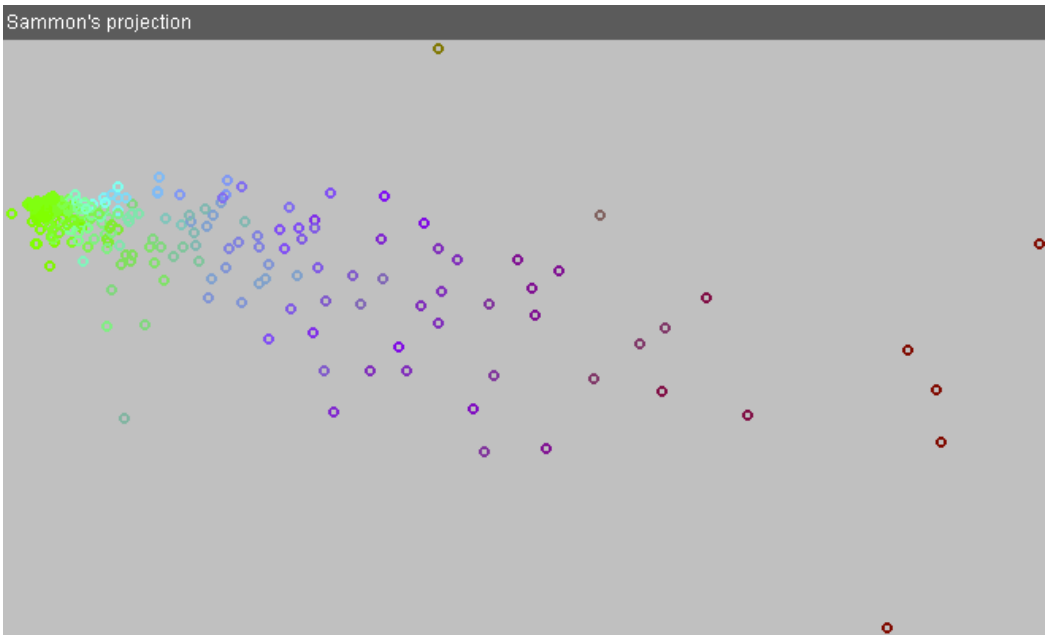

Figure 9: Sammon's projection of the same data as in the time-in-space SOM (Fig.7).

Building clusters with the help of SOM

In discussing the choice of parameter values for SOM, Bação et al. (2008) explain that the resolution of the SOM grid (in other terms, size of the SOM map) is chosen depending on the analysis task. The task of exploring the data distribution requires a large SOM map (i.e. high resolution grid). When the user is interested in clustering data, a small map (i.e. low resolution grid) is used. In our case, the large amount of the data makes us interested in clustering. However, as we have demonstrated in the previous section, when the chosen grid resolution is too low, some of the clusters produced by the SOM can be spoiled with outliers. Hence, it is reasonable to choose a higher resolution. On the other hand, when the grid resolution is high, the distinctions among objects included in different SOM cells may be excessively fine. It may be appropriate to put them in the same cluster. An optimal grid resolution that would result in pure while clearly distinct clusters hardly exists because, as can be seen from the illustrations, the differences (i.e. distances in the attribute space) between neighbouring SOM cells vary significantly throughout the matrix. Therefore, the following strategy is reasonable: first, choose such a grid resolution that results in sufficiently pure cell contents, i.e. without outliers; second, unite cells that are close in the attribute space into larger clusters.

Yan and Thill (2008) use the k-means clustering algorithm to build clusters on the basis of SOM output. They apply k-means to the prototype vectors of the SOM grid cells. The contents of the cells whose prototype vectors are put in the same k-means cluster are united. The k-means algorithm requires the desired number of

clusters (k) to be specified as the parameter. Since the true or optimal number of clusters is not known in advance, Yan and Thill run the k-means method many times with different values of k and then choose the clustering variant with the best Davies-Bouldin index, which is the measure of cluster separation (Davies and Bouldin 1979). We suggest that uniting SOM cells into larger clusters can be done in a more direct and user-controlled way by interactive joining of neighbouring cells having similar contents.

Another possible strategy is, on the opposite, to use a coarse grid resolution for the SOM and then refine the results by applying another clustering method (e.g. hierarchical clustering, as suggested by John et al. 2008) to the contents of the SOM cells. For our purpose, the first strategy is preferable since it requires less effort from the user's side.

In our implementation, the user specifies a distance threshold, and the system joins neighbouring SOM cells such that the distance between them in the attribute space is below the threshold. The threshold can be specified and modified by moving a slider (bottom right of Fig.10). The system dynamically reacts to the changes: it displays cluster labels in the SOM cells and changes the colouring of the cells that are joined with their neighbours to the "average" colour of the joined cells. For example, in Fig.10 cluster 1 unites the lower three cells in the first column and the cell in column 2, row 4. Cluster 2 unites three cells in the upper left corner of the SOM matrix. The colours of the cells that have been joined with others are very close to the original colours visible in Fig. 7. This is due to the similarity of the original colours from which the "average" colour is derived (we remind that similarity of cell colours means closeness in terms of the attributes). When the distance threshold is increased, more cells are joined together, and the colours change more noticeably.

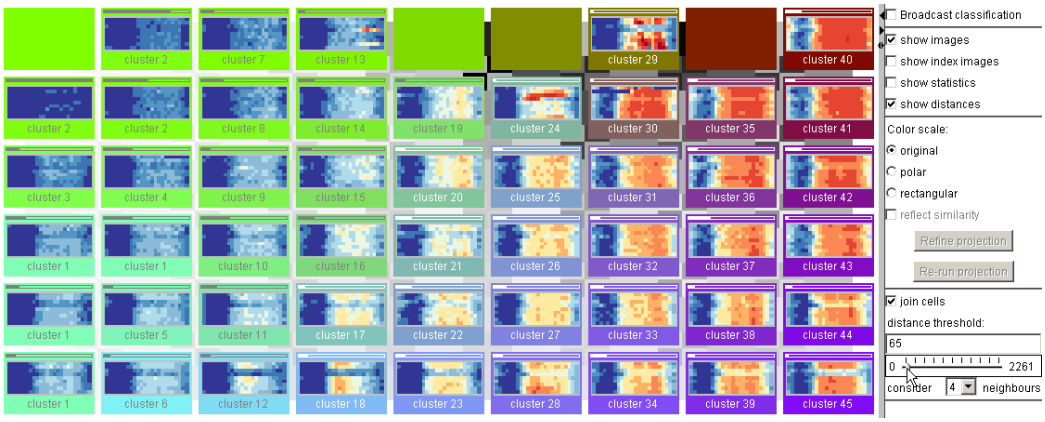

Figure 10: SOM matrix 9x6 after cell joining.

An important question in joining cells is when to stop. A significant change of cell colour is an indication that the cluster unites quite dissimilar cells, which is not desirable. To inspect the content of a cluster, the user clicks on any of the cells making the cluster and gets an additional window like in Fig.5 with the feature images of all cluster members. This allows the user to visually estimate the degree of similarity and variability within the cluster and decide whether to keep it or to decrease the distance threshold. When a more rigorous way for controlling the "purity" of the clusters is desired, a display like in Fig.11 can be used. The display was produced for cluster 2, which combines the cells (1,2), (2,1), and (2,2) (the first number is the column and the second is the row). For each original cell and for the

resulting cluster, the system computed the time series of the mean values and time series of the standard deviations from the original data. These time series are shown on two time graphs, which allow the user to evaluate the result of cell joining more accurately.

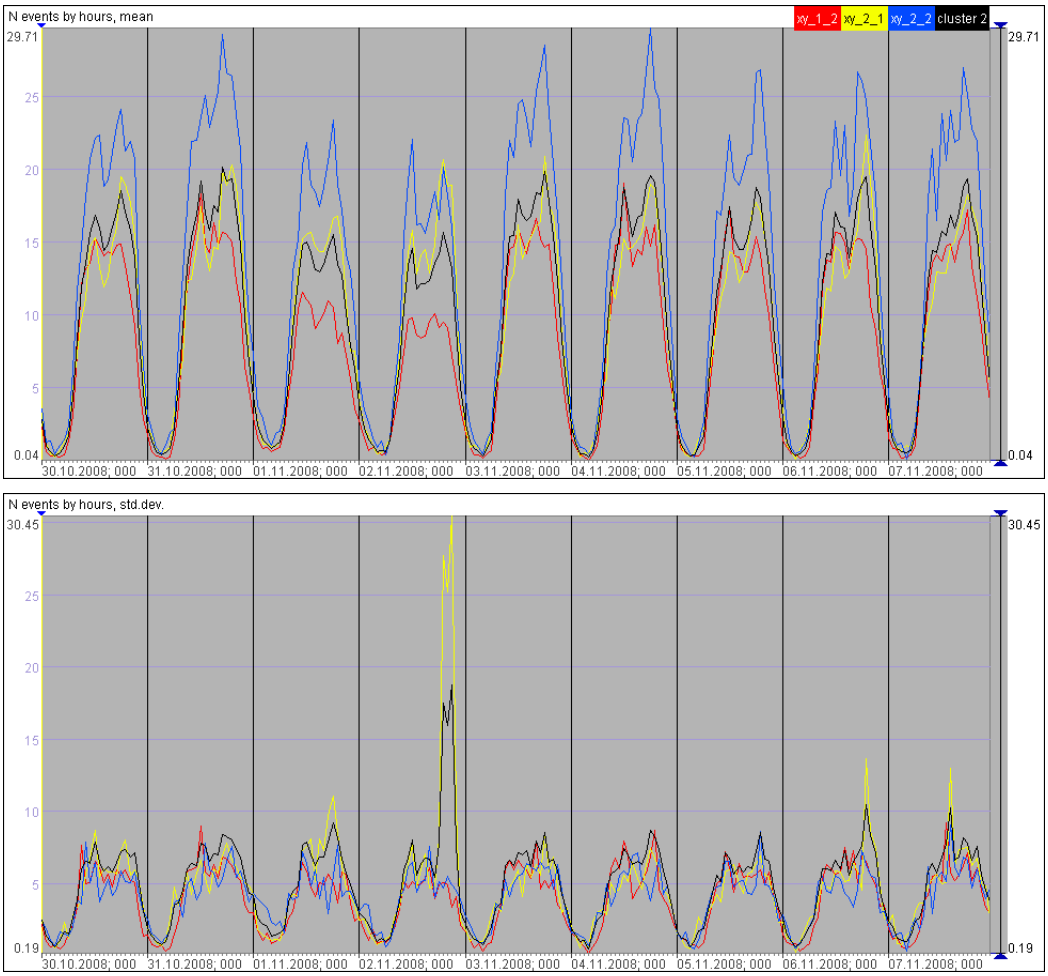

Figure 11: The time series of the mean values (top) and standard deviations (bottom) in the cells joined in cluster 2 are shown in red, yellow, and blue and the time series of cluster 2 as a whole in black.

In Fig.11 bottom we can see that the standard deviations in the original three cells and in cluster 2 are very close in all but a few time units. Exceptionally high values in the cell (2,1) (yellow line) in some time units mean that the cell includes an outlier that could not be separated from the other data by the SOM method. Accordingly, the standard deviations in cluster 2 in these time units are also high. However, if we disregard the outlier, we can say that the internal variance in cluster 2 is not higher than in the original cells. From the top time graph in Figure 11 we can see that the values in the cell (1,2) (red) in the weekend are slightly lower than in the other two cells and that the values in the cell (2,2) (blue) are, on the average, slightly higher than in the other two cells. It is up to the user to decide whether the cells are similar enough to be joined in one cluster.

Concerning the outlier contained in the cell (2,1), we could find out that this is the area of the stadium San Siro (also known as Meazza), where two football games

took place on Sunday and Thursday. The call counts were exceptionally high in this area in the time intervals before and after the games, which explains the high values of the standard deviation in these intervals (Fig.11 bottom). The peak of the call counts on Sunday was much higher than on Thursday, which can be explained by the difference in the attendance of the games: 50,000 and 11,000 spectators, respectively. Apart from these particular time intervals, the call counts in the area were close to the values in the other areas included in the cell (2,1).

Summary of the framework

The flow chart in Fig.12 summarises the general framework for using the SOM for the analysis of spatio-temporal data that have the structure $S \times T \rightarrow A$. This is not a very severe limitation to the use of the framework since very often spatio-temporal data that originally have a different structure can be converted to $S \times T \rightarrow A$. Thus, we had the phone calls data with the structure $C \rightarrow S \times T$, but we transformed them to $S \times T \rightarrow A$ by means of spatio-temporal aggregation. A similar transformation can be applied to movement data having the structure $O \times T \rightarrow S$, where O is a set of moving objects; see Andrienko and Andrienko (2010) for a detailed discussion. Of course, aggregation is appropriate only when the analyst is interested in general patterns rather than details.

Depending on the analysis goals, the analyst may wish to know how the patterns of the spatial situations change over time or how the patterns of the temporal variations are distributed over space, In the first case, the analyst applies SOM to the spatial situations (i.e. $S \rightarrow A$) in the different time units $t_i \in T$ and obtains a space-in-time SOM. This case is reflected on the left of the flow chart. In the second case, the analyst applies SOM to the temporal variations (i.e. $T \rightarrow A$) in the different places $s_i \in S$ and obtains a time-in-space SOM. This case is reflected on the right of the chart. In both cases, the analysis is done according to the common procedure. The analyst chooses a suitable SOM grid resolution, runs the SOM algorithm, and checks whether the resulting SOM cells contain significant outliers. If so, the resolution of the grid should be increased, and the algorithm should be re-run. When the contents of the cells are sufficiently pure, the analyst may (optionally) build bigger clusters by interactively joining close cells. The resulting patterns are observed on additional displays, to which the colours of the SOM cells are transmitted. The temporal patterns of the changes of the spatial situations are observed on temporal displays, such as time graph (Fig.3) and time arranger (Fig.4). The spatial patterns of the distribution of the temporal variations are observed on spatial displays, first of all, on a map (Fig.8).

As we mentioned previously, various transformations of the attribute values may be appropriate for a comprehensive analysis. In particular, it may be useful to transform absolute attribute values into relative ones, such as deviations from the average over time in each time series $T \rightarrow A$ or deviations from the average over space in each spatial situation $S \rightarrow A$. The SOM-based analytical procedure is applied to the transformed attribute values in the same way as to the original values.

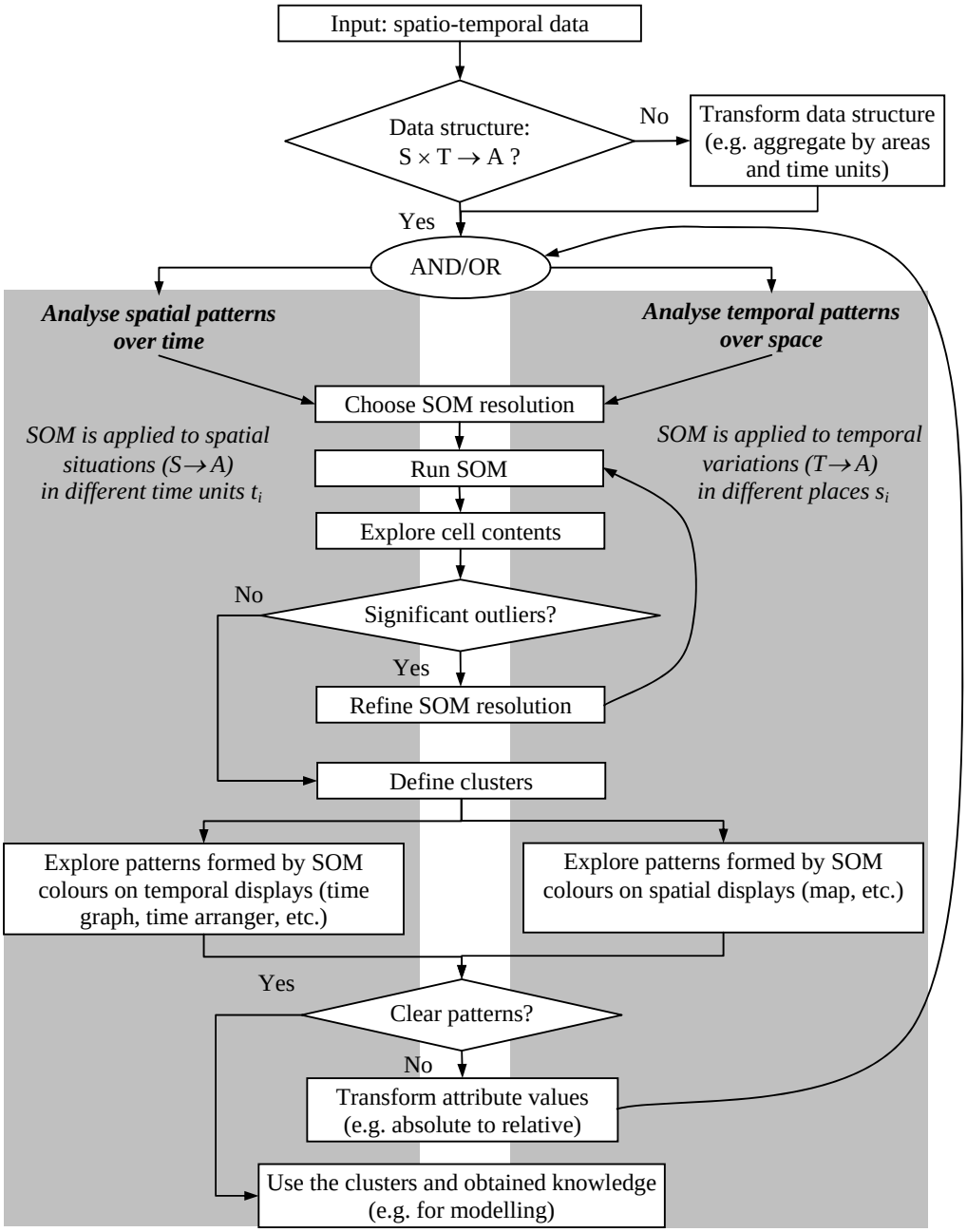

Figure 12: The general framework for using self-organizing maps to discover and analyze spatio-temporal patterns.

The direct outcome of the analysis is the clusters of the spatial situations and/or of the temporal variations. However, this is not the only outcome and not the most valuable one. Perhaps, the most valuable is the knowledge the analyst obtains about the properties of the spatial and temporal distribution of the phenomenon under investigation. In the following section, we briefly outline how the results of the analysis, i.e. the knowledge as well as the clusters, can be used for predictive spatio-temporal modelling.

Predictive modelling based on SOM analysis results

One of our current research interests is extending visual analytic approaches to produce explicit and practically utilizable results, such as predictive models. In particular, we work on extending the SOM-based analysis framework to the generation of models that can predict attribute values in different places and time units. The approach we currently investigate is combining existing methods for time series prediction with techniques and/or ideas from geographic analysis so that the spatial variation of the data and spatial dependencies among places could be taken into account. The rationale for using time series-based predictive modelling is that a number of established methods together with the corresponding body of knowledge exist in statistics and that the methods are widely available in various statistical analysis packages such as R or SAGE.

Since it is necessary to account for the spatial variation, creating one global temporal prediction model for the whole territory is not reasonable. On the other hand, creating a separate local model for each place on the basis of only its individual time series is also a poor choice due to the risk of overfitting, i.e. capturing occasional fluctuations rather than the general pattern of change. A suitable strategy may be to use the idea on which the geographically weighted regression technique (GWR) is based (Fotheringham et al. 2002). A traditional regression model is an equation with parameters, which are assumed to be constant. In GWR, the parameters are allowed to vary over the study area. In a similar manner, the temporal variation of a space- and time-dependent attribute may be described by an equation with parameters that vary over the space.

In applying this idea to spatio-temporal modelling, the analyst needs to make two choices: a method for the time series-based prediction and a method to account for the spatial variation of the model parameters. Both choices can be made on the basis of the knowledge obtained by the analysis according to our framework. The method for the time series prediction is chosen according to the properties of the temporal variation, which are conveyed by the temporal mosaic images in the time-in-space SOM (Fig.2) and the temporal displays incorporating the colours of the cells of the space-in-time SOM, in particular, time graph (Fig.3) and time arranger (Fig.4). The method to account for the spatial variation is chosen on the basis of the properties of the spatial distribution, which are conveyed by the map images in the space-in-time SOM (Fig.1) and the map display incorporating the colours of the cells of the time-in-space SOM (Fig.8).

The most important characteristics of the temporal variation that determine the choice of the prediction method are the presence of a trend and the presence of periodicity. Thus, simple exponential smoothing techniques can be used when there is no trend and no periodicity, double exponential smoothing can account for a trend in the data and triple exponential smoothing (Holt-Winters' method) is used when there is both trend and periodicity (NIST/SEMATECH 2010, Gelper et al. 2010). For example, for our phone call data, which vary in a periodic manner, the Holt-Winters' method is appropriate. The methods differ in the number of parameters: one in the simple exponential smoothing, two in the double exponential smoothing, and three in the triple exponential smoothing. In the case of spatio-temporal data, these parameters vary over the space.

The GWR approach usually assumes that the values of each parameter form a continuous surface in the geographical space. Data observed near a place are assumed

to have more influence on the prediction for this place than data located farther from it. This is reflected in the weights assigned to the observations according to their spatial distances to the place. Usually the weights are specified by a continuous function of the distance. For example, the weight may decrease with the distance according to a Gaussian curve. This approach can be taken for the spatio-temporal prediction in a case when the spatial variation of the phenomenon is smooth, i.e. there are no big differences between neighbouring places. However, this is not always the case. Thus, our phone calls data are spatially abrupt rather than smooth (MacEachren 1995), as can be seen from Fig.8.

In a case of a spatially abrupt phenomenon, the spatial variation of the parameters for the temporal prediction needs to be represented by a discontinuous function. A straightforward idea is to use the clusters of the places obtained from the time-in-space SOM. Each cluster unites places with similar temporal variations. Hence, the temporal prediction parameters for a place can be defined using the time series in this place as well as the time series of the other members of this cluster. The simplest approach is to weight all cluster members equally. This means practically that a common temporal prediction model is created for each cluster. A possible modification is to weight cluster members on the basis of their spatial distances to the place for which the prediction is made.

As a proof of concept, we have tried to build a predictive model for our example data about phone calls. Since the temporal variation is periodic, we have chosen the Holt-Winters' method for temporal prediction. To account for the differences between the working days and the weekends, we created distinct sub-models for the working days and the weekends. For this purpose, we produced two time series from the original 9-days time series associated with each area: one consisting of the working days and the other consisting of Saturday and Sunday.

A specific problem of this dataset is that the time series are quite short: they include seven working days and only one weekend. To deal with this problem, we randomly divided the areas contained in the cluster into two subsets: one subset to be used for deriving the model and the other for testing, since we do not have additional data for another time period to check the predictions. The time series of the areas from the subset used for deriving the model were concatenated into one long time series. This was done separately for the working days time series and for the weekend time series. We applied the Holt-Winters smoothing procedure that is available in the statistical package R. We set the expected periodicity to 5 for the working days time series and 2 for the weekend time series. The R procedure gave us the models for working days and for weekends. We evaluated the models by applying them to the subset that has not been used for the model derivation and computing the mean squared errors. Generally, the evaluation results are quite good, indicating the validity of our approach.

Hence, both the implicit outcomes of the SOM analysis (knowledge of the spatial and temporal distribution patterns) and the explicit ones (clusters of places) can be utilized for building predictive models. We do not claim that modelling on the basis of SOM results should necessarily be done in the way just described. We plan to continue our research in this direction by considering data with other characteristics of spatial and temporal distribution. We shall also investigate the applicability of other types of statistical and geostatistical models.

Conclusion

We have developed a visual analytics framework in which computational and visual techniques are combined to enable interactive exploration and analysis of large spatio-temporal datasets. The framework supports two complementary perspectives on spatio-temporal data: as a temporal sequence of spatial situations and as a set of spatially distributed time series representing local temporal variations. We have described the use of the framework by applying it to a real dataset and discussed a possible way towards producing tangible, practically usable results in the form of predictive models. Currently we are making only the first steps in this direction but we consider it as one of the strategic directions in our future research. This does not apply solely to the utilization of SOM outcomes but also to results of other visual analytics tools.

Short biographical notes on all contributors

References

Agarwal, P., and Skupin, A., editors (2008): Self-Organising Maps: Applications in Geographic Information Science. Wiley

Andrienko, G., and Andrienko, N. (2010): A General Framework for Using Aggregation in Visual Exploration of Movement Data. The Cartographic Journal, 2010, v.47 (1), pp. 22-40

Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., von Landesberger, T., Bak, P., and Keim, D. (2010): Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. Computer Graphics Forum, Vol.29(3), pp. 913-922.

Andrienko, N., Andrienko, G., and Gatalsky, P. (2003): Exploratory Spatio-Temporal Visualization: an Analytical Review. Journal of Visual Languages and Computing, Vol. 14 No.6, December 2003, pp. 503–541

Andrienko, N and Andrienko, G. (2006): Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Berlin: Springer

Bação, F., Lobo, V., and Painho, M. (2003): The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Computers & Geosciences, 31(2):155 – 163, 2005. Geospatial Research in Europe: AGILE 2003.

Bação, F., Lobo, V., and Painho, M. (2008): Applications of Different Self-Organizing Map Variants to Geographical Information Science Problems. In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp.21–44, Wiley.

Barthel, K. U. (2008): Improved image retrieval using automatic image sorting and semi-automatic generation of image semantics. In Proc. Int. Workshop Image Analysis for Multimedia Interactive Services, pp. 227–230.

Davies, D.L., Bouldin, D.W. (1979): A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 2, pp.224-227.

Deboeck, G. and Kohonen, T., editors (1998): Visual Explorations in Finance with Self-Organizing Maps. Springer

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2002): Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley: Chichester.

Gelper, S., Fried, R.; and Croux, C. (2010): Robust forecasting with exponential and Holt-Winters smoothing. Journal of Forecasting, Vol.29, no.3, pp.285-300

Guo, D., Chen, J., MacEachren, A., and Liao, K. (2006): A visualization system for space-time and multivariate patterns (VIS-STAMP). IEEE Transactions on Visualization and Computer Graphics, 12(6):1461–1474

Guo, D., Gahegan, M., MacEachren, A., and Zhou, B. (2005): Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. Cartography and Geographic Information Science, 32(2):113–132

Harrower, M., and Brewer, C.A. (2003): Colorbrewer.org: An online tool for selecting colour schemes for maps. The Cartographic Journal 40(1):27–37.

Hewitson, B. C. (2008): Climate analysis, modelling, and regional downscaling using self-organizing maps. In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp.137–163, Wiley

John, M., Tominski, C., and Schumann, H. (2008): Visual and Analytical Extensions for the Table Lens. In Proceedings IS&T/SPIE Annual Symposium Electronic Imaging - Visualization and Data Analysis (VDA), San Jose, USA, 2008.

Kaski, S., Venna, J., and Kohonen, T. (2000): Coloring that reveals cluster structures in multivariate data. In Australian Journal of Intelligent Information Processing Systems, pp.6–82

Kohonen, T. (2001): Self-Organizing Maps. Springer

Kohonen, T. (2010): SOM Toolbox: Intro to SOM by Teuvo Kohonen. http://www.cis.hut.fi/somtoolbox/theory/somalgorithm.shtml. Retrieved 23 July 2010.

Koua, E., and Kraak, M.-J. (2008): An integrated exploratory geovisualization environment based on self-organizing map. In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp. 45–66, Wiley

MacEachren, A.M. (1995): How Maps Work: Representation, Visualization, and Design. Guilford, New York

NIST/SEMATECH (2010): NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, accessed 27 July 2010.

Nuernberger, A., and Detyniecki, M. (2006): Externally growing self-organizing map and its application to e-mail database visualization and exploration. Applied Soft Computing, 6(4):357–371

Openshaw, S. (1984): The Modifiable Areal Unit Problem. Norwich: Geo Books.

Sammon, J. W. (1969): A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, 18:401–409

Schreck, T., Bernard, J., von Landesberger, T., and Kohlhammer, J. (2009): Visual cluster analysis of trajectory data with interactive Kohonen maps. Information Visualization, 8(1):14–29

Skupin, A. (2008): Visualizing human movement in attribute space. In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp.121–135, Wiley

Skupin, A., and Agarwal, P. (2008): Introduction: What is a Self-Organizing Map? In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp.1-20, Wiley

Spielman, S. E., and Thill, J.-C. (2008): Social area analysis, data mining, and GIS. Computers, Environment and Urban Systems, 32(2):110–122

Ultsch, A. (1999): Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In Kohonen Maps (1999), Elsevier, pp. 33–46.

Vesanto, J. (1999): SOM-based data visualization methods. Intelligent Data Analysis, 3(2):111–126

Yan, J., and Thill, J.-C. (2008). Visual Exploration of Spatial Interaction Data with Self-Organizing Maps. In Agarwal, P., and Skupin. A., editors, Self-Organising Maps: Applications in Geographic Information Science, pp. 67-88, Wiley