Self-Organising Maps for exploration of spatio-temporal emergency response data

Olga Špatenková[1], Urška Demšar[2], Jukka M. Krisp[1]

[1]Laboratory of Geoinformation and Positioning Technology, Helsinki University of Technology,
P. O. Box 1200, FI-02015 TKK, Finland
Email:olga.spatenkova@tkk.fi, jukka.krisp@tkk.fi

[2]National Centre for Geocomputation,
National University of Ireland, Maynooth, Ireland
Email: urska.demsar@nuim.ie

1. Introduction

The motivation for this paper is the problem of civil protection agencies to develop a risk model for their service planning. Within the strategic levels of planning and management, improvements in the identification of high or low risk areas can assist the emergency preparedness setting and resource evaluation (Krisp et al. 2005). The core problem is in gaining an insight into complex relationships of the underlying phenomena and consequently in the selection of meaningful variables for the model. One of the possible solutions is to use exploratory data analysis and data mining methods to attempt to identify the potentially relevant hidden relationships from the data supplied by the emergency response services.

Due to the progress in data acquisition and data processing technologies, real-world data have been recently collected and stored in large amounts. Various datasets can therefore be combined for the analysis. However, to find relevant patterns from the diverse types of data is a challenging task. Knowledge discovery from databases (KDD) and data mining in particular offer methods for discovery of yet unknown, but potentially useful, information from the data (Fayyad et al. 2002, Han et al. 2002).

The special nature of spatial data, in particular spatial autocorrelation, heterogeneity, non-stationarity and complex geometrical and topological relationships between spatial and spatio-temporal objects, poses more challenges to the data mining process (Miller and Han 2001). There are two approaches to handle these special issues. The first option is to apply the conventional data mining methods to the pre-processed data, where the spatial relationships have been considered, while the other option consists of the development of new spatial data mining algorithms. Spatial data mining is computationally a very demanding process and it is therefore often complemented with methods for visual exploration (Keim et al. 2004, Ahonen-Rainio 2005, Klein 2005, Demšar et al. 2006). Visual data mining avails of the ability of the human brain to recognise interesting patterns in the data more efficiently and faster than any computer.

Space and time are intrinsically interrelated in geography. One of the fundamental problems in GIScience has therefore been the support for spatio-temporal analysis. Because of the complexity of the problem, no fully adequate temporal GIS exists at the present time (Yuan 2007). With regard to knowledge discovery, spatio-temporal data mining has been identified as a research area, where knowledge is discovered from datasets that contain explicit or implicit temporal or spatio-temporal information

(Hamilton et al. 2006). Recently, a number of studies focused on developing either computational or combined visual-computational methods that could be used to identify patterns that are simultaneously spatial and temporal (Andrienko et al. 2003, Kwan 2004, Laube and Purves 2006, Compieta et al. 2007).

In this paper we focus on the improvement of the existing risk model (Ihamäki 1997) for the Fire and Rescue Organisation in the Helsinki Metropolitan Area, Finland. There is a need to explore the relations between the occurrence of recorded emergency response incidents and the characteristics of the surrounding areas of these incidents. Previous work focused on the investigation of the incident location and the surrounding spatial objects (Karasová et al. 2005) and exploration of the socio-economic aspects of the phenomenon (Špatenková and Krisp 2007). However, while the temporal component is of a special interest in this case (Ahola et al. 2007), it has not yet been investigated properly.

This paper describes an approach to identify spatio-temporal relationships in emergency response data by combining spatio-temporal data pre-processing, application of a traditional computational data mining algorithm and finally visualisation and visual exploration of the result. The paper is structured as follows: the methodology for the data-pre-processing and exploration is presented in section 2. Section 3 lists a selection of the more interesting results and section 4 contains conclusions about the suitability of the chosen data mining approach as well as plans for further research.

2. Methodology

A Self-Organising Map has been chosen for this study to discover patterns in spatio-temporal data. The data were pre-processed so that the relevant spatial and temporal information could be handled by the conventional SOM algorithm. The computational step was followed by visualisation and visual exploration of the results. All SOM calculations were done and visualisations produced using the SOM toolbox for Matlab.

2.1 Self-Organising Map (SOM)

The Self-Organising Map (SOM) is an unsupervised neural network that maps multidimensional data onto a two-dimensional lattice of cells. It preserves the probability density and the topology of the input data and thereby also the similarity patterns that exist in the higher dimensional space (Kohonen 1997, Silipo 2003).

The SOM defines a mapping from the n-dimensional input data space onto a two-dimensional array of neuron cells in a rectangular, hexagonal or irregular lattice. In the beginning, every cell is assigned a vector of weights, m_i. During the training phase, the SOM finds the location of the neuron that is most similar (i.e. the best match) to the input data vector. After each input the weight vectors m_k of each neuron in a neighbourhood of the best match neuron m_c are recalculated according to the distance to the best match neuron and the neighbourhood kernel function $h_{ck}(t)$. This function reaches the highest value at the best matched neuron m_c and monotonically decreases towards 0 with distance from the central neuron (fig. 1). This means that nearby cells activate each other up to a certain distance to learn from the same input data vector. Similar data vectors are therefore mapped to cells that are close to each other in the output space (Kohonen 1997).

The two-dimensional result space of the SOM makes it a very nice method to represent visually. For the work described here, the relevant visualisations are the distance matrix (D-matrix) and the component planes, while other visualisations can be found in Vesanto (1999).

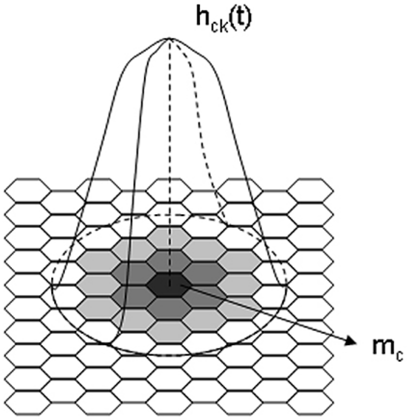


Figure 1. The Self-Organising Map with a hexagonal neuron lattice. The neighbourhood function $h_{ck}(t)$ is centred over the best matched neuron m_c, which is shown as a black cell. The neighbouring neurons that have their weights recalculated by this best match are shown in grey. Other neurons are not affected.

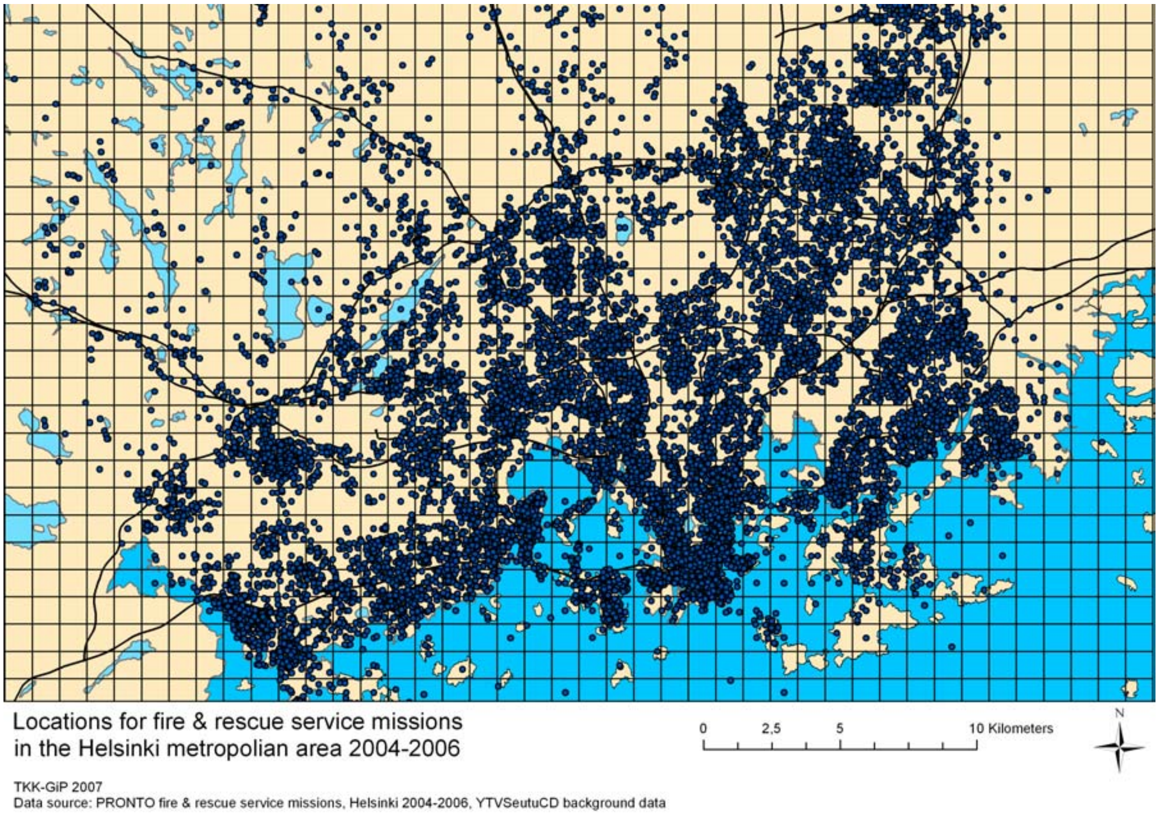
In a D-matrix a grey shade or a colour is assigned to each lattice cell according to the distance to its immediate neighbours. Light areas in such a map indicate groups of similar cells, i.e. clusters. Dark areas contain less similar neurons and mark borders between clusters (Kohonen 1997, Vesanto 1999). In the SOM toolbox used in our study, the distance values in the D-matrix are represented on the blue-green-yellow-orange-red colour scale, where blue corresponds to low and red to high values. The clusters in our D-matrix can therefore be identified as contiguous bluish areas, while orange to red patches indicate areas of greater dissimilarity and as such represent borders between clusters.

The result of the SOM can also be visualised using the component planes, with one lattice shown for each attribute. Each lattice is coloured according to the values of a particular attribute. In the SOM toolbox, the colours for the planes range from blue through green, yellow and orange to red on a similar scale as the one for the D-matrix. Again, blue represents low values and red high values of each attribute. The planes can be compared to each other to find correlations between attributes – these are revealed as similar patterns at identical locations in different component planes (Vesanto 1999, Koua and Kraak 2004).

The SOM is useful as a knowledge discovery tool for spatial datasets because it preserves both the topology and the distribution of data vectors in the input space. It has been used for knowledge discovery in a number of spatial applications (Takatsuka 2001, Gahegan et al. 2002, Jiang and Harrie 2004, Koua and Kraak 2004, Guo et al. 2005, Demšar 2007), but spatio-temporal applications are fewer, for example Skupin and Hagelman (2005).

2.2 Discovering spatio-temporal patterns with help of a SOM

Apart from the spatial distribution, the emergency response data used in our study have distinct temporal characteristics. The incidents are recorded over the time of the year, the day of the week and the hour of the day. These characteristics add an additional temporal value to the already multidimensional spatial dataset. In our study, the SOM was used to project this dataset to a two-dimensional geometric space, which can be considered as a spatial metaphor or a spatialisation (Skupin and Fabrikant 2007) of the original dataset. The aim is to reduce the complexity of the data while at the same time preserving the ability to discover patterns that are simultaneously spatial and temporal. Since SOM preserves the topology of the original multidimensional space, it is a suitable clustering method for this purpose. Because of the SOM's topology preserving property, spatial patterns discovered by visual exploration of the SOM space correspond to spatio-temporal patterns in the original dataset.


Locations for fire & rescue service missions
in the Helsinki metropolian area 2004-2006

TKK-GiP 2007
Data source: PRONTO fire & rescue service missions, Helsinki 2004-2006, YTVSeutuCD background data

Figure 2: Map of the study area.

The data used in this study consist of the PRONTO records of fire and rescue incidents in the Helsinki Metropolitan Area (fig. 2) for the years 2004-2006. This is a point dataset which was for this experiment combined with background information about the population and the infrastructure. We have extracted 11 spatio-temporal attributes from the incidents dataset, describing the incident occurrence: the day in the year (1-365), the day of the week (1-7), the hour (0-23), the incident type, X-coordinate, Y-coordinate, and the types of the five nearest incidents. The background information comes from "SeutuCD", a data collection of population, business, topology, and infrastructure information published by the regional council of Helsinki in 2003. From this data source, we have selected four additional attributes describing the surroundings of the incidents:

the distance to the nearest building, the type of the nearest building, population density and age density (calculated based on average age of people living in each building, e.g. areas with houses for elderly people have higher values of age density than areas with young families).

3. Results

This section demonstrates the process of discovering the patterns in the data by using the SOM visualisations and presents a selection of interesting results. It should be noted that these results describe observations which implicitly describe the relationships in the studied dataset and can serve as the basis for hypothesis formation. However, to avoid the interpretation fallacy based on coincidental patterns in this particular dataset, the results should be confirmed and validated by the domain experts before they can be used. This should be done to ensure that the coincidental patterns existing in the data without causality relationships are excluded from further consideration before the actual risk modelling takes place.

Clusters represent similarities in the original spatio-temporal attribute space. They can be identified from the D-matrix as areas of predominantly blue colour, which indicates that cells in these areas are similar to their neighbours. Cells in red areas of the D-matrix are different from their respective neighbours and as such represent borders between clusters. On inspection, the D-matrix can be subdivided into seven clusters (fig. 3). In fig. 3, these clusters have been superimposed on the component planes for each attribute. Looking at these other planes, we can observe that the clusters match the patterns in the component planes to some degree. For example the triangular cluster in the bottom left corner represents incidents occurring in the built-up area of the highest age density, which is not densely populated. The incidents in this cluster occurred mainly in the summer and during the daytime, the actual location varies, but similar incidents occur in their neighbourhood (as indicated in the planes for attributes NEAR_TYPE1 to NEAR_TYPE5, where the area of this cluster has approximately the same colour in all these five planes). This cluster also represents a group of similar incident types located in the higher values of the incidents' classification - the exact codes should be further investigated.

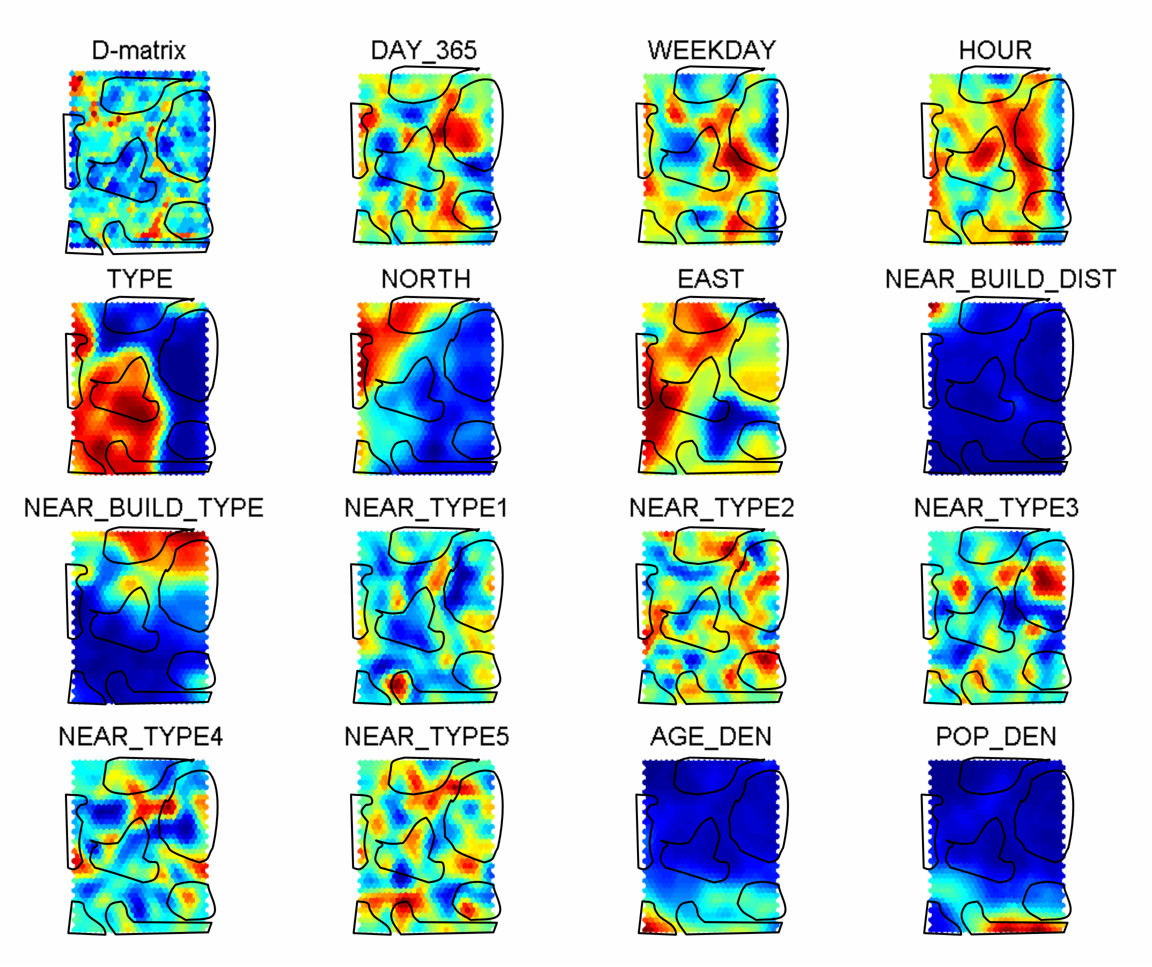


Figure 3. Identification of clusters from the D-matrix and their location in the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

By comparing different areas of the SOM in the component planes, we can discover relationships between the attributes. For example, from fig. 4 we can observe that the weekend incidents (which are located in the dark red areas of the relevant component plane – WEEKDAY) do not seem to occur during the early morning, but instead during the afternoon and in the evening. They occur in the southern part of the study area where population and age density are low and in the proximity of buildings classified by low values (building types should be further investigated). Similarly, we can say that winter incidents occur in the north-east in the afternoon, but in the south in the morning. Summer incidents, on the other hand, occur in areas of high population density and are of different types during the mornings and during the afternoons. During the weekends, incidents occur mainly in the south and south-west. Early morning incidents in general do not occur during the weekends and they are located either in the south-west, or north-east.

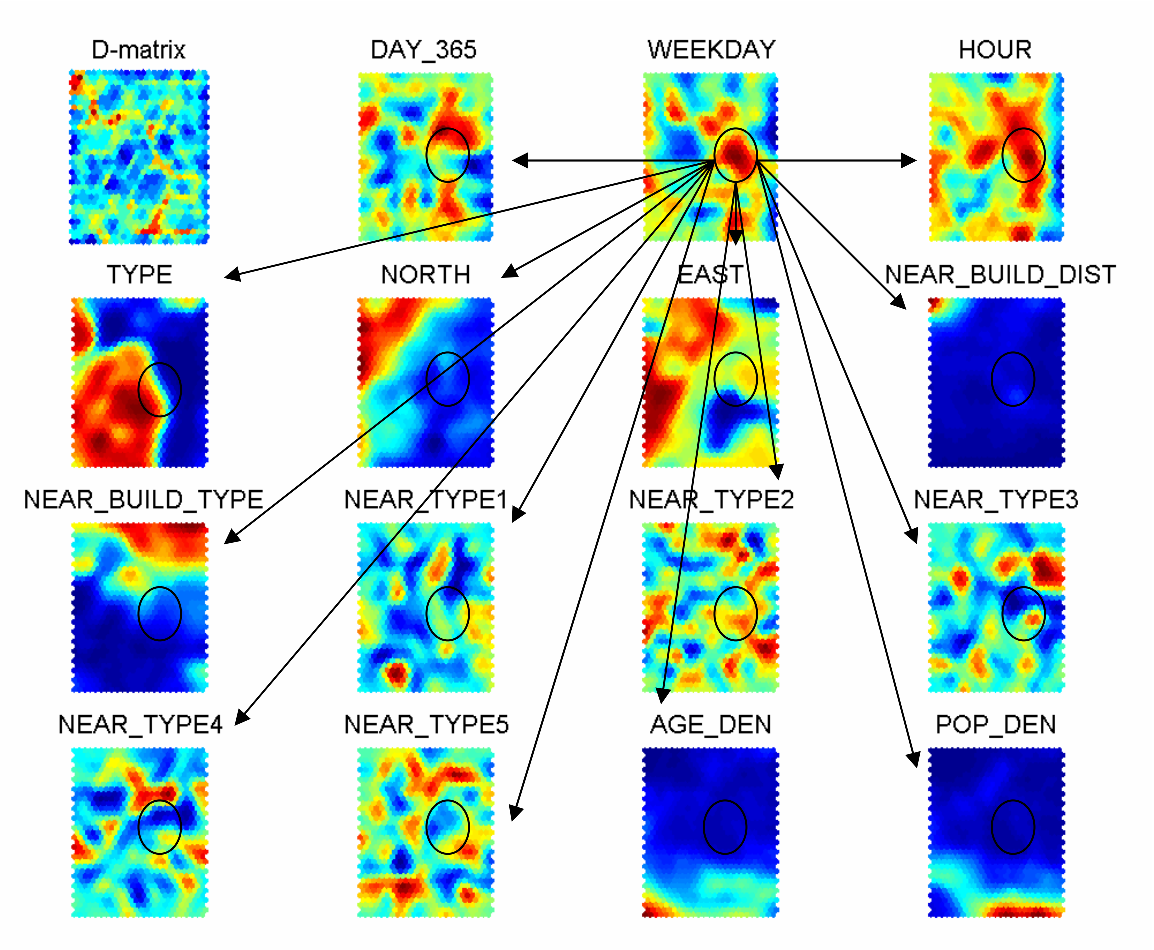


Figure 4. Discovering relationships between the attributes from the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

Comparing the patterns in the component planes for attributes NEAR_TYPE1 to NEAR_TYPE5, two areas with similar colours in all these five planes can be identified. These two areas are roughly indicated with a rectangle and a circle in fig. 5 and represent areas of similar incidents. It seems that incidents in both these areas occur on Saturday evenings, but they differ considerably in location and in the building types in the neighbourhood.

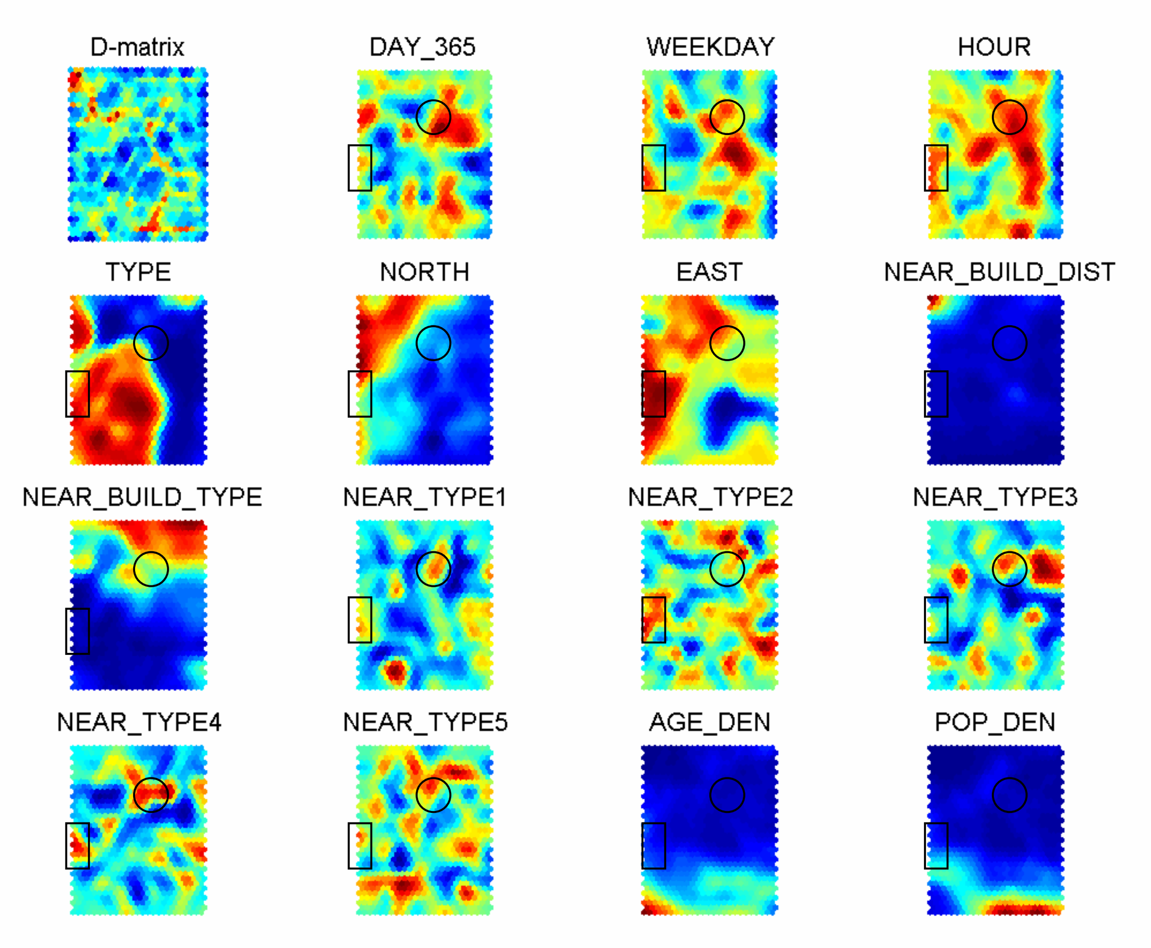


Figure 5. Areas of similar neighbouring incidents, as indicated by similar colours in the component planes NEAR_TYPE1 – NEAR_TYPE5. Other characteristics of these incidents can be found from the remaining component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

4. Conclusions

This paper describes how a computational data mining method - the SOM - can be effectively used to provide an insight into the relationships in the spatio-temporal data and thus support the modelling of the studied phenomenon. The observations as presented above can serve for hypothesis formation describing the relationships in the studied dataset and after consultation with the domain experts be used for emergency risk modelling. Since knowledge discovery is an iterative process, more expert opinion could be included in the exploration process, to explain and validate some of the structures and patterns revealed by the use of the SOM.

Combining the computational method with the visual exploration of the result helped to make fast and easy-to-interpret observations about the patterns in the dataset. On the other hand, because of the "throwing many things in one cell" property of the SOM, the details were obscured. For example, in this study it was difficult to distinguish between the particular incident types or building types in question, we could only estimate the time of day in general terms such as morning/afternoon/evening and we could not properly differentiate between the days of the week. Further investigation of these patterns with other methods is therefore necessary to gain real meaning from the results.

Another issue is that since the SOM visualisations are a spatialisation in an abstract two-dimensional space which has nothing to do with the original geographic space, the connection with the real-world locations can be difficult to infer. Integrating the component planes with a real spatial representation of the input geographic space, i.e. a map, would make the interpretation easier to follow, but the SOM toolbox does not provide this option. One of the possible goals for future research is to repeat and improve the exploration procedure using an appropriate software package that provides the integration of the SOM with other visualisations, such as for example the GeoVista Studio (Gahegan et al. 2002). However, a disadvantage of such applications is the non-flexibility regarding the size and the training method of the SOM, since these are usually predetermined and can not be changed. Another possibility would therefore be to consider developing bespoke software components for coupling the SOM with an existing GIS software, i.e. something similar as done by Skupin and Hagelman (2005).

5. Acknowledgements

The authors would like to thank the Finnish Fire & Rescue Services for cooperation and supporting the research.

6. References

Ahola T, Virrantaus K, Krisp JM and Hunter G, 2007, A spatio-temporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science (IJGIS)*, accepted for special issue.

Ahonen-Rainio P, 2005, *Visualization of geospatial metadata for selecting geographic datasets.* PhD Thesis, Helsinki University of Technology, Helsinki.

Andrienko N, Andrienko G and Gatalsky P, 2003, Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing*, 14:503–541.

Compieta P, Di Martino S, Bertolotto M, Ferrucci F and Kechadi T, 2007, Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages and Computing*, accepted manuscript.

Demšar U, 2007, Knowledge discovery in environmental sciences: visual and automatic data mining for radon problems in groundwater. *Transactions in GIS*, 11:255-281.

Demšar U, Krisp JM and Křemenová O, 2006, Exploring geographical data with spatio-visual data mining. In: G. Elmes (eds), *Proceedings of the 12th International Symposium on Spatial Data Handling*, Vienna, Austria, 12-14 July, 149-166

Gahegan M, Takatsuka M, Wheeler M and Hardisty F, 2002, Introducing Geo-VISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems*, 26:267-292.

Guo D, Gahegan M, MacEachren AM and Zhou B, 2005, Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*, 32(2):113-132.

Fayyad U, Grinstein GG and Wierse A, 2002, *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, San Diego.

Hamilton HJ, Geng L, Findlater L and Randall DJ, 2006, Efficient spatio-temporal data mining with GenSpace graphs. *Journal of Applied Logic*, 4:192–214.

Han J, Altman RB, Kumar V, Mannila H and Pregibon D, 2002, Emerging scientific applications in data mining. *Communications of the ACM*, 45(8): 54-58.

Ihamäki V-P, 1997, *Paikkatietojärjestelmien (GIS) käyttö palo- ja pelastustoimen yhteistoiminnan suunnittelussa (Geographic information systems in planning fire and rescue services cooperation)*. ProGradu Thesis, Helsinki University, Helsinki.

Jiang B and Harrie L, 2004, Selection of streets from a network using self-organizing maps. *Transactions in GIS*, 8:335–350.

Karasová V, Krisp J and Virrantaus K, 2005, Application of spatial association rules for development of a risk model for fire and rescue services. In: Hauska H and Tveite H (eds), *Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS)*, Stockholm, Sweden.

Keim DA, Panse C, Sips M and North SC, 2004, Visual Data Mining in Large Geospatial Point Sets. *Computer Graphics and Applications*, 24(5): 36-44.

Klein P, 2005, TheCircleSegmentView: A User Centered, Meta-data Driven Approach for Visual Query and Filtering. PhD Thesis, Universität Konstanz, Konstanz.

Kohonen T, 1997, *Self-Organizing Maps*. 2nd edition, Springer Verlag, Berlin-Heidelberg.

Koua E L and Kraak M-J, 2004, Alternative visualization of large geospatial datasets. *The Cartographic Journal*, 41:217–228.

Krisp JM, Jolma A and Virrantaus K, 2005, Using explorative spatial analysis to improve fire and rescue services in Helsinki, Finland. In: P. Oosterom, S. Zlatanova and E. Fendel (eds), *Geo-information for Disaster Management*, Springer, Delft, The Netherlands, pp. 1282-1296.

Kwan MP, 2004, GIS Methods in Time-Geographic Research: Geocomputation and Geovisualization of Human Activity Patterns. *Geografiska Annaller B*, 86:267–280.

Laube P and Purves RS, 2006, An approach to evaluating motion pattern detection techniques in spatio-temporal data. *Computers, Environment and Urban Systems*, 30:347–374.

Miller H and Han J, 2001, *Geographic data mining and knowledge discovery*. Taylor & Francis, London.

Silipo R, 2003, Neural Networks. In: Berthold M and Hand DJ (eds), *Intelligent Data Analysis*, 2nd edition. Springer Verlag, Berlin-Heidelberg, 269-320.

Skupin A and Fabrikant SI, 2007, Spatialization. In: Wilson JP and Fotheringham SA (eds), *The Handbook of Geographic Information Science*, in press, Blackwell Publishing, 61-79.

Skupin A and Hagelman R, 2005, Visualizing Demographic Trajectories with Self-Organizing Maps. *Geoinformatica*, 9(2):159-179.

Špatenková O and Krisp JM, 2007, The Use of Contingency Tables to Value Variables for Spatial Models. In: *Proceedings of the 5th International Symposium on Spatial Data Quality*, Enschede, the Netherlands, in press.

Takatsuka M, 2001, An application of the Self-Organizing Map and interactive 3-D visualization to geospatial data. In: *Proceedings of the 6th International Conference on Geocomputation*, Brisbane, Australia.

Vesanto J, 1999, SOM-based data visualization methods. *Intelligent Data Analysis*, 3:111-126.

Yuan M, 2007, Adding Time into GIS Databases. In: Wilson JP and Fotheringham SA (eds), *The Handbook of Geographic Information Science*, in press, Blackwell Publishing, 169-184.