Data Mining of Collaboratively Collected Geographic Crime Information Using an Unsupervised Neural Network Approach

Julian Hagenauer, Marco Helbich (corresponding author),

Michael Leitner, Jerry Ratcliffe, Spencer Chainey, and Ricky Edwards

ABSTRACT: Crime intelligence analysis and criminal investigations are increasingly making use of geospatial technology and geographic profiling methodologies to improve tactical and strategic decision-making. However, its full potential has yet to be exploited. In particular, current geospatial technology is limited in handling the increasing volume of police recorded data and can be weak in considering information that is voluntarily provided and/or sourced from eyewitness accounts and other public sources. Thus, the objective of this research is to promote data mining methods, particularly the self-organizing map algorithm and its visualization capabilities, to explore the value of 'hidden' information in such data sources and to gain insight into the complex behavior of the geography of crime. The approach is applied to a high-profile and unsolved murder series in the city of Jennings, Louisiana. In a collaborative effort with the Jennings Police Task Force, the analysis uses a range of information sources, including email correspondence, transcribed face-to-face interviews, and phone calls that have been stored in Orion, an FBI database of "Information Packages (IPs)". In this research, close to 200 IPs relating to Necole Guillory, the eighth and last victim (whose body was discovered in August 2009), are analyzed and resulted in new geographic patterns and relationships previously unknown to the Task Force.

KEYWORDS: Crime mapping, text mining, self-organizing map, information packages

Introduction

Besides the variety and large amounts of geospatial data recently available, criminology profits from the technological progress in data analysis to counteract the rise of crime throughout society. Profiling methodologies and geographic information systems (GIS) based methods, as discussed in Chainey and Ratcliffe (2005), are successfully applied in day-to-day operations of police and governmental agencies. Nevertheless, these methods are not capable to explore large amounts of high dimensional data and have only limited applicability to unstructured text documents such as crime protocols. This requires data mining techniques (see Han & Kamber 2006) to discover previously unknown information hidden in data which would not be readily apparent when sifting through the data manually (Fayyad et al. 1996).

Therefore, the objective of this research is to present first results of using data mining approaches in criminology. In particular, a thitherto neglected subsidiary data source, namely the information and hints of eyewitnesses, the general public, or other relevant persons, will be explored by means of document mining techniques and self-organizing maps (SOM; Kohonen 2001), which represent an unsupervised neural network. To the best knowledge of the authors such approach to investigative documents collected by the police and stored in the form of textual information, has never been done before.

The real-world case study is associated with a much publicized, still unsolved, serial killer case in the city of Jennings, Jefferson Davis Parish (JDP), LA. Between May 2005 and August 2009 eight women were killed and dumped in rural areas just outside of the city limits of Jennings. The age of the women ranged from 17 to 30, six of the women were whites the other two were blacks. All eight women were residents of Jennings. They were drug-addicts and made their living mostly from prostitution, making them highly vulnerable and relative easy targets for the serial killer. About one hundred similar unsolved serial killer cases involving mostly young female victims that are drug-addicts and prostitutes exist currently in the US. The Jennings Police Task Force uses Orion, an FBI database of "Information Packages (IPs)" to store email correspondence, transcribed face-to-face interviews, and phone calls associated with this crime series. For this study only the 172 IPs related to the last of the eight victims, Necole Guillory, were extracted from Orion and subsequently analyzed. If the results from this research turn out to be useful to the task force, then it is planned to rerun the analysis with the entire set of all IP's included in the Orion database. Obviously, such vast amount of data records can only be analyzed with a data mining approach in a meaningful way.

Materials and Methods

Study Site and Data

All but one victim's body dump site is located in Jefferson Davis Parish (JDP), LA outside of the city of Jennings. The last victim's body was found in Acadia Parish, which neighbors JDP immediately to the east (Figure 1). JDP is a poor and mostly rural parish with a median family income of US$30,783 and dominated by agricultural products, such as sugar cane, rice, cotton, sweet potato, etc. The city of Jennings is the parish seat with a population of 10,986 according to the 2010 US census. The ethnic composition of Jennings is about 70% white and 28% black, with the majority of the black population living south and the majority of the white population north of a railroad track that runs through the city from northwest to southeast. Interstate I-10 crosses Jennings in the north.

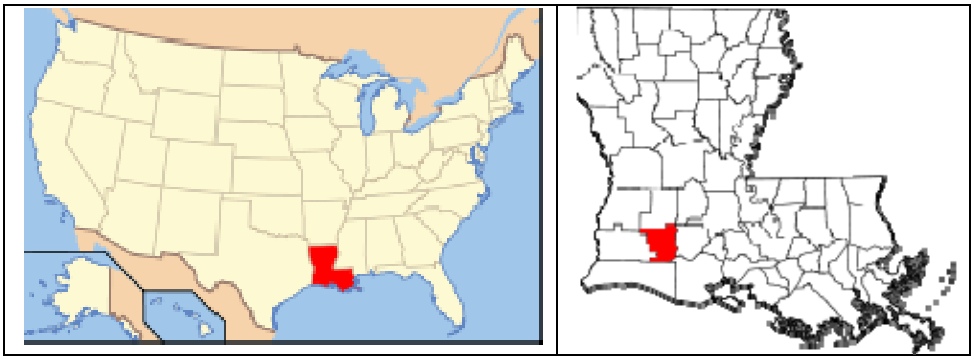

Figure 1: The location of Louisiana within the 48 contiguous states of the US (left) and the location of Jefferson Davis in Louisiana (right).

The eight body dump sites (numbered from 1 through 8) and the date of recovery, in parenthesis, are shown in Figure 2. In the middle top of this image, the city of Jennings, LA can be seen. Also the I-10 corridor and the railroad tracks are slightly visible.

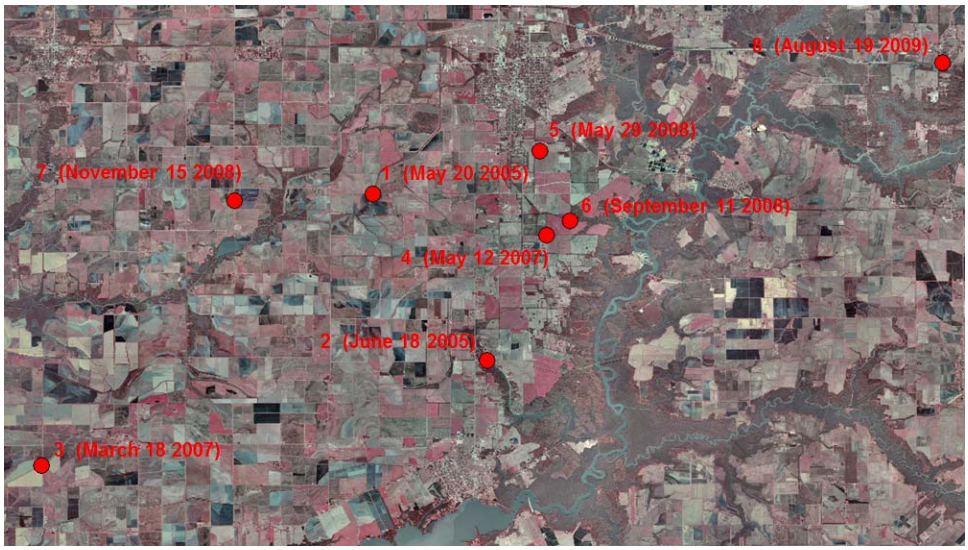

Figure 2: The location of all eight body dump sites superimposed over a digital orthophoto.

Data preprocessing

In order to analyze the IP's, basically textual documents, it is necessary to transform them into a data representation, which enables computational processing. In order to achieve this task, this study employs information retrieval methods (e.g. Manning et al. 2008).

At the beginning, all words are extracted from the IP documents. For computational reasons, the order of words is neglected. All words that are not text (e.g. headings), shorter than 3 characters, occur less than 5 times, or more than 100 times per IP, are discarded. These parameter settings were empirically determined from preliminary experiments. Note that the document's metadata were not removed, because they can also be valuable sources of information. After that, the remaining words are set to lower case, stop words (e.g. "and", "which", "that") are removed, and word stems are extracted (Porter 1980). A word stem represents the part of the word that is common to all its inflected variants. The stemming algorithm alters sometimes the orthography of words, e.g. the letter "y" is often substituted by "i". Therefore, it is often not possible to directly infer the original word from a word stem. Furthermore, their meaning is also often unclear. Table 1 supports a better understanding and lists the ambiguous or unclear word stems that are relevant for this publication, there corresponding words, and provides a brief explanation of their meaning, if applicable. Due to privacy reasons, word stems of first and last names of persons other than the victims' are made anonym: E.g. "fn1" refers to the first name of the person with the number 1 assigned, and "ln3" to her/his last name, which has number 3 assigned. Finally, vectors are formed by calculating the inverse document-frequency (IDF) of the word stems for each IP. The IDF measures the

importance of a term in a document collection by relating the occurrence of a term in that document to the total occurrence of the word in all documents (Harman 1992).

Table 1: Ambiguous or unclear word stems and their meanings.

Word stem	*Word/Explanation*
deputi	deputy
unusu	unusual
calcasieu	Calcasieu Parish
det	abbreviation of detective (det.)
viewip	part of an referred URL
goe	goes
impala	Chevrolet Impala
roug	(Baton) Rouge
lpr	Mait/lpr
boi	boy
detect	detective
mailto	reference to an Email address
lspcl	Louisiana State Police Crime Laboratory
gmc	truck manufacturer
piec	pieces
leo	part of an referred URL

Most of the IP's contain geographic information in the form of coordinates related to people's residences. These coordinates are also extracted from the IP's and are assigned to the corresponding vector to enable geographic mapping of the results. To ensure privacy, some random noise is added to the coordinates.

Self-organizing maps

The self-organizing map (SOM; Kohonen 2002) is an unsupervised artificial neural network approach. The SOM takes a high-dimensional input vector and maps it two a low-dimensional output map. Thereby, the SOM creates models of different types of data

in the dataset, and organizes theses models in an ordered fashion in a map. Therefore, the SOM can also be interpreted as an adaptive display method, which is particular suitable for the representation of complex data and large data sets (Kaski and Kohonen, 1996). These properties have made the SOM especially popular in GIScience (Skupin and Agarwal 2008).

The SOM consists of an arbitrary number of neurons. Associated with each of these neurons is a prototype vector of the same dimension as the input space. Additionally, neighboring neurons are connected with each other. These connections reflect the topology of the map. In principle, the dimension of a SOM is arbitrary, but in practice mostly two-dimensional SOMs are used for visualization purposes. A set of input vectors is used to train the SOM. For each input vector the neuron with the shortest distance of its prototype vector to this input vector is determined. This neuron is commonly termed the best matching unit (BMU). Then, the BMU's prototype vector and the prototype vectors of the neurons within a certain vicinity of the BMU are moved into the direction of the presented input vector. The strength of adaption depends on the distance of the neuron to the BMU and on the actual learning rate. Both, the size of the vicinity as well as the learning rate decrease monotonically in the course of the learning process. The rationale is that in the beginning of the learning phase, the arrangement of neurons on the map can alter distinctly, while at the end of the training phase only small changes are made to fine tune the map. After training, the SOM represents a low-dimensional map of the input space, where each neuron represents some portion of the input space. Furthermore, the distance relationships of the input space are mostly preserved in the map. For an in-depth discussion of the SOM algorithm the reader is referred to Kohonen (2001).

To visualize SOMs in order to allow in-depth analysis of the map's structure, U-Matrices (Ultsch & Siemon 1990) are convenient for interpretation and analysis. The U-matrix plots the differences of neighboring neurons' prototype vectors within the map by means of a color scale. Clusters become visible in the U-Matrix by distinct outlines of the cluster boundaries. If no crisp outlines are visible, then it means that clusters in the input space are less distinct. Thus, the U-matrix shows both the present cluster structure and the quality of the clustering.

Results

An 8×6 SOM is trained with 10,000 iterations. For ease of visualization, the neurons are arranged in a hexagonal grid. The learning rate decreases linearly from 0.5 to 0. The kernel function for adapting the neurons is the Gaussian function, initially covering the entire map in order to coarsely arrange the neurons in the beginning of the training phase (Agarwal & Skupin 2008).

Figure 3 shows the results of the SOM in many different ways. The base of the map is a U-Matrix. The U-Matrix shows a complex structure with lots of different areas, which represent distinct areas of the input space. In particular, a few light-colored regions at the left and right, as well at the bottom of the SOM are noticeable. In order to more emphasize these light-colored regions, their outline is drawn with different color hues.

These outlined regions are distinct from other areas of the map in that their color values of the U-Matrix are below a certain threshold. In this way, the coloring can be interpreted similar to the clustering of different watersheds (Vincent & Soille 1991). Additionally, Figure 3 shows the prototype vectors' words with high IDF values. The label sizes correspond the IDF values, thus word stems with a high IDF have also large labels. Word stems with IDF values below 0.2 are not shown.

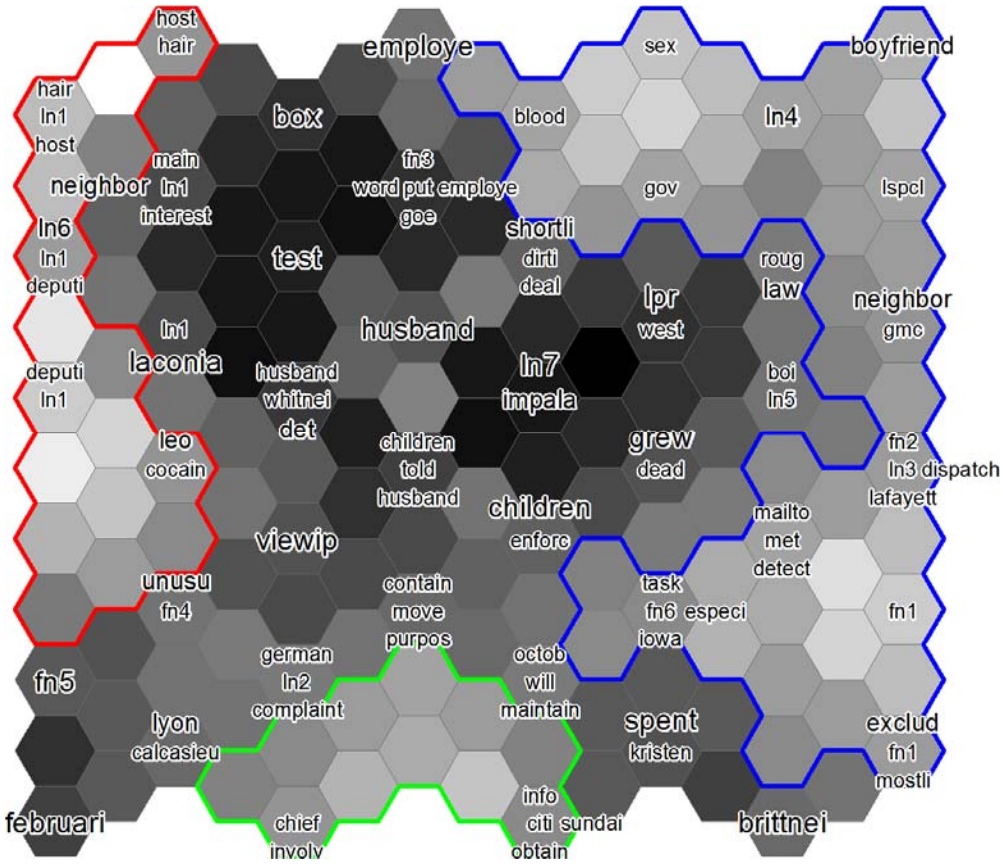

Figure 3: Cluster sizes and strengths represented in the form of an U-Matrix

The comprehensive visualization of the dataset in Figure 3 reveals interesting insights into the input dataset. At first it is noticeable, that some metadata word stems like "viewip" or "mailto" appear in the map, which indicates that they exhibit high IDF values. Further, some relationships between different terms are apparent. The words "sex", "boyfriend", and "blood" are all located in the same cluster, indicating that documents with these words share many similarities. Also the murder victims "Kristen" (Kristen Lopez) and "Brittnei" (Brittney Gary) are mapped to "regions" in the map that are close to each other. It thus can be concluded that the criminal investigations regarding both victims tended to be closely related. On the other side, the word "laconia" referring to another victim (Laconia Brown), is located far away from the two previously mentioned victims. This indicates that for some reason there is some difference in the conducted criminal investigations between this last and the two previously mentioned victims. It is interesting to note, that some murder victims are not represented in Figure 3. The reason for this is that the records that have high IDF values for such victims' name

stems are mostly dissimilar to each other, so that the emergence of high values for the victims' name stems is prevented by the training algorithm of the SOM.

In order to characterize the outlined clusters in a more general fashion, the mean of the cluster's prototype vectors are calculated. The five highest IDF values and the corresponding word stems of the resulting mean vectors for the different clusters are shown in Table 2.

Table 2 The five highest IDF values and the corresponding words of clusters' mean prototype vectors

	1st	*2nd*	*3rd*	*4th*	*5th*
Cluster 1 (Red)	ln1,	deputi,	host,	hair,	remain,
	0.447	0.305	0.197	0.197	0.182
Cluster 2 (Green)	sundai,	chief,	obtain,	involve,	info,
	0.121	0.119	0.118	0.076	0.075
Cluster 3 (Blue)	fn1,	piec,	sex,	prepar,	ln3,
	0.319	0.312	0.268	0.266	0.258

Cluster 1 shows a notably high value for "ln1", which refers to the word stem of a common last name. In fact, a lot of interrogations of different people with that last name have been made in the course of the criminal investigation of the last victim, Necole Guillory. It is noticeable, that the word stem of "ln1" is especially closely related to "laconia", which is mapped nearby cluster 1. In general, cluster 2 has low IDF values and does not seem to exhibit a noticeable pattern. Cluster 3 has a high IDF value for the word stem of the first name "fn1". Furthermore, cluster 3 has also a high IDF value for the word stem of last name "ln3". It should be noted that the criminal investigations refer to people with that last name quite often. The word stem of last name "ln3" is mapped to the same neuron as the words "lafayette" and the stem of first name "fn2", suggesting that there is a close relation between these three names with "lafayette", referring to "Lafayette", which is both the name of a city and a parish in Louisiana. Lafayette Parish is very close to Jefferson Davis Parish and both the city of Lafayette and Jennings are located on the interstate I-10. Furthermore, the three clusters have high IDF values for general word stems that seem to appear often when transcribing police recorded data and information that is voluntarily provided and/or sourced from eyewitness accounts and other public sources (e.g. chief, obtain, involve, prepare, info). Moreover, the IDF values of these word stems notably differ for the different clusters. Thus, it can be concluded that in particular the word stems that result from the process of reporting describe distinct properties of the clusters.

Figure 4 shows the coordinates of the reports of each identified cluster in a geographic map. Several important remarks can be drawn from this map. First, it is notable that most of the reports of cluster 1 and cluster 3 are centrally located in Jennings, whereas cluster 3 is sparsely distributed in the whole area. Moreover, a notable portion of cluster 1's

reports are located in the surroundings of Lafayette, whereas only few reports are located nearby Lake Charles. The opposite is true for the reports of cluster 3: Only a few reports of cluster 3 are located close to Lafayette, whereas a notable number of reports are located in the surroundings of Lake Charles. This geographic pattern indicates, that the clusters that were outlined by the inspection of the SOM, exhibit unique spatial properties which complement their distinct textual characteristics.

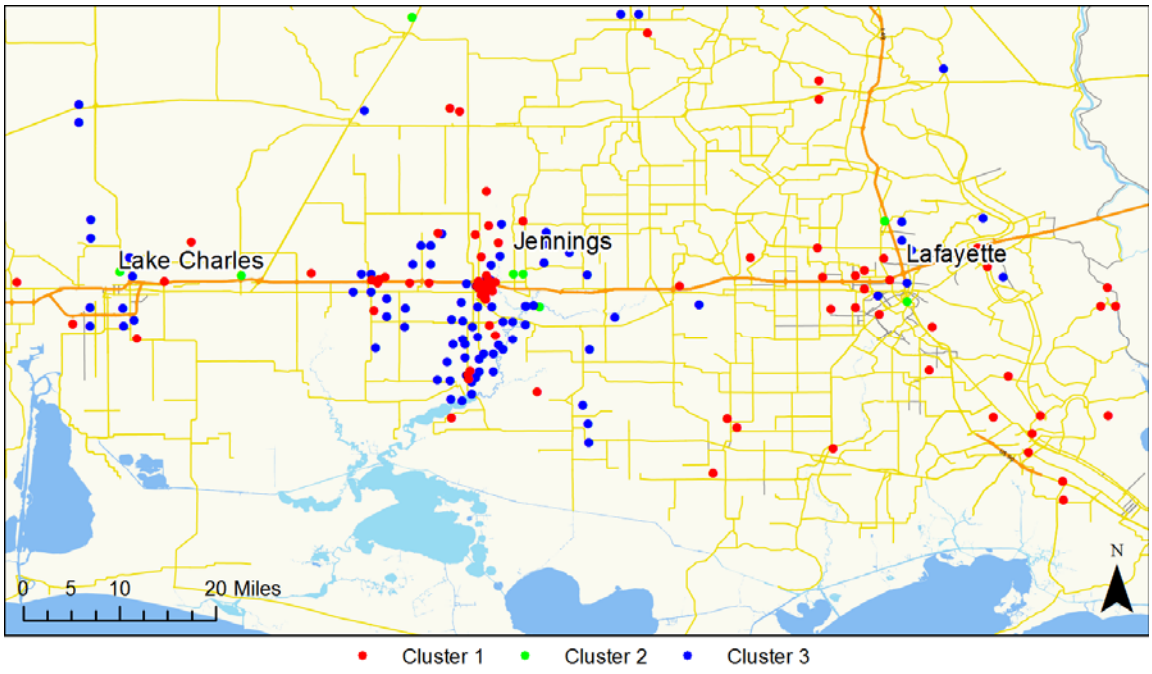

Figure 4 Coordinates of the reports

Conclusions

Results from the data mining exercise will be presented and shared with the Jennings task force. However, as a yet unsolved serial killer case, results of this research may be difficult to evaluate, unless they would lead directly to the apprehension of the serial offender. However, solved criminal investigations would allow the evaluation of the data mining results by comparing trends, relationships, novel information, etc. derived from the data mining results with the specific information that led to the arrest of the offender. As a matter of fact, the authors have already started to data mine crimes occurring at or near the Louisiana State University (LSU) Campus that fall into the jurisdiction of the LSU police. It is planned to only select those criminal investigations that are also broadcasted to the LSU community (students, faculty, and staff) by the LSU police. It is proposed to data mine the information that the LSU police receives from the LSU community to those broadcasted crime investigations. We will select both solved and unsolved cases in this ongoing research project. Another future research endeavor is the data mining of all IP's related to all victims in the Jennings serial homicide case and to further investigate the cluster's coordinates map by means of point pattern techniques.

Acknowledgements

This project is funded by the DFG's Initiative of Excellence, Global Networks, University of Heidelberg. Marco Helbich is funded by the Alexander von Humboldt Foundation.

References

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

Chainey, S. and Ratcliffe, J. H. (2005) *GIS and Crime Mapping*. Chichester: Wiley.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) Advances in Knowledge Discovery and Data Mining. In Fayyad, U. M. et al. (eds.) *Advances in knowledge discovery and data mining*. Cambridge: MIT Press, pp. 1-34.

Han, J. and Kamber, M. (2006) *Data Mining. Concepts & Techniques*. Amsterdam: MK.

Harman, D. (1992) Ranking algorithms. In Frakes, W. B. and Baeza-Yates, R. (eds.) *Information Retrieval. Data Structures & Algorithms*. Upper Saddle River: Prentice Hall, pp. 363-392.

Kaski, S. and Kohonen, T. (1996) Exploratory Data Analysis By The Self-Organizing Map: Structures Of Welfare And Poverty. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, pp. 498-507.

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

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Porter, M. F. (1980) An Algorithm for Suffix Stripping. *Program*, 14, pp. 130-137.

Vesanto, J. (1999) SOM-based Data Visualization Methods. *Intelligent Data Analysis*, 3, pp. 111-126.

Vesanto, J. and Alhoniemi, E. (2000) Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11, pp. 586-600.

Ultsch, A. and Siemon, H. P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proceedings of International Neural Networks Conference, pp. 305-308.

Vincent, L. and Soille, P. (1991) Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, pp. 583-598.

Marco Helbich, Institute of Geography, University of Heidelberg, Berliner Straße 48, D-69120 Heidelberg, Germany. Email <helbich@uni-heidelberg.de>