Cartography and Geographic Information Science, 2013
http://dx.doi.org/10.1080/15230406.2013.779780

Taylor & Francis
Taylor & Francis Group

Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach

Marco Helbich[a]*, Julian Hagenauer[a], Michael Leitner[b] and Ricky Edwards[c]

[a]*Institute of Geography, University of Heidelberg, Berliner Straße 48, Heidelberg D-69120, Germany;* [b]*Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA;* [c]*Louisiana Sheriffs Association, 1175 Nicholson Drive, Baton Rouge, LA 70802, USA*

(*Received 10 November 2012; accepted 9 February 2013*)

Crime intelligence analysis and criminal investigations are increasingly making use of geospatial methodologies to improve tactical and strategic decision-making. However, the full potential of geospatial technologies is yet to be exploited. In particular, geospatial technology currently applied by law enforcement is somewhat limited in handling the increasing volume of police recorded and relatively unstructured narrative crime reports, such as observations and interviews of eyewitnesses, the general public, or other relevant persons. The main objective of this research is to promote text mining, particularly the self-organizing map algorithm and its visualization capabilities, in combination with point pattern analysis, to explore the value of otherwise hidden information in a geographical context and to gain further insight into the complex behavior of the geography of crime. This methodological approach is applied to a high-profile and still unsolved homicide series in the city of Jennings, Louisiana. In a collaborative effort with the Jennings Police Task Force, the analysis is based upon a range of information sources, including email correspondence, transcribed face-to-face interviews, and phone calls that have been stored as "Information Packages" in the Orion database, which is maintained by the Federal Bureau of Investigation. Close to 200 individual information packages related to Necole Guillory, the eighth and last victim whose dead and dumped 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; information packages; text mining; word cloud; self-organizing map; point pattern analysis

Background

Crime in all its facets including robberies, cybercrime, and terrorism is an integral part of our daily life and affects society as a whole (Costa 2010). Because crime impacts our sense of security, affects our quality of life, and has far reaching economic consequences, crime protection and crime combat has gained significant importance in the general public, police agencies, politics, and science. To counteract the impact of crime throughout society, crime agencies, among other things, have reverted to the application of modern intelligence and geospatial technologies, which have quickly become an emerging scientific research field for tackling such security needs. The development of such technologies is often founded on a solid theoretical basis that includes such well-known theories as routine activities (Cohen and Felson 1979), rational choice (Clarke and Cornish 1985), and environmental criminology (Brantingham and Brantingham 1981).

One of the main challenges that law enforcement is increasingly facing is the proliferation of crime and crime-related data (Chen et al. 2004; Chen and Wang 2005; Hagenauer, Helbich, and Leitner 2011). Profiling methodologies and geographic information systems (GIS)-based methods (Chainey and Ratcliffe 2005) have already been successfully applied retrospectively and prospectively in day-to-day operations by law enforcement agencies (e.g., Leitner and Helbich 2011; Helbich and Leitner 2012). However, these methods are not capable of exploring large amounts of unstructured narrative crime reports or protocols that are increasingly stored in criminal justice databases. A large part of these data is provided voluntarily to the police by eyewitnesses, the general public, or other relevant persons, who report hints and narrative descriptions of observations associated with particular crime events. As such, this corresponds to Goodchild's (2007) "human as sensors paradigm", where people voluntarily collect and share geographic and textual information with the police. Major serial crimes, especially serial homicides or rapes, usually receive much media coverage and the number of such protocols can easily exceed hundreds to thousands of unstructured items in the form of emails, written statements, and transcribed telephone recordings. In contrast to standard structured data, which can be explored by basic queries, such crime reports are stored in the form of plain text documents, which require alternative and novel methods for their analysis (Jones and Purves 2008). Chen et al. (2004) also note that such unstructured data are further affected

*Corresponding author. Email: helbich@uni-heidelberg.de by noisy content, resulting from spelling mistakes in addition to typographical and grammatical errors.

Although, such unstructured documents are often very important for crime analysis and are useful in providing valuable insights into the complex behavior of crime, they have rarely been investigated comprehensively in empirical analysis, and thus have played a marginal and less significant role in criminal investigations. The exploration of unstructured documents requires text data mining techniques (Delen and Crossland 2008; Manning, Raghavan, and Schütze 2008) to extract and discover previously unknown, potentially useful, and hidden information, which would not be readily apparent when sifting through the data manually (Han and Kamber 2011).

At present, only a few studies exist that focus on the mining of narrative crime reports and more research is needed to evaluate the capability of text mining approaches for law enforcement. For instance, Chau, Xu, and Chen (2002) automatically extracted entities, including names and addresses, from a sample of 36 narrative reports from the Phoenix police department using a neural network-based approach. The authors concluded that the accuracy of the empirical results varied depending on the kind of entity mined. For example, in contrast to names, which can be precisely extracted, addresses show a lower accuracy rate. Similarly, Ku, Iriberri, and Leroy (2008) analyzed police and witness reports with an information extraction system, combining several algorithms from natural language processing. They reported a high proportion of correctly extracted features. More relevant for this research is the work by Chen et al. (2003) and Alruily, Ayesh, and Al-Marghilani (2010), who used self-organizing maps (SOMs) for clustering and visualization of crime data from public media. The latter study successfully extracted crime phrases, such as keywords, from Arabic news articles in order to characterize diverse crime types. Common to these studies is that they are solely based on text exploration and neglect the ability to link the mining output to geographical space. This is clearly an important limitation, since "space" is seen as an inherent property of crime (see Leitner and Helbich 2011). Therefore, the present research applies a multistage methodology which projects the mining output to geographical space and then analyzes this output with spatial statistical methods. To summarize, while the mining of (unstructured) crime reports is still in its infancy, it represents a vibrant and emerging research field that has not yet been linked to a GIS and its analysis and visualization capabilities.

Hence, the main objective of this study is to propose a text mining approach to expand upon the spatial analytical capabilities of the investigative effort carried out by law enforcement personnel. In particular, this research mines an already available subsidiary textual data source, namely "Information Packages (IPs)" stored in the Orion database which is operated by the US Federal Bureau of Investigation (FBI). All IP's are related to a much publicized, yet unsolved, serial homicide case in the city of Jennings, located in Jefferson Davis Parish (JDP), LA. This novel approach allows uncovering certain aspects of the information content and possible relationships between IPs, previously unknown to the Jennings Police Task Force. Additionally, both the content and relationships can be linked to their geographic context. To the best knowledge of the authors, using such a data mining approach to investigative an important on-going crime series has never been done before. The remainder of this article is organized as follows: In the "Study area and data" section the study area and the data set are introduced. In the "Methodology" section, the methodology is reviewed. Important results are presented in the "Results" section. Finally, the "Conclusions" section draws conclusions and suggests valuable directions for future research.

Study area and data

Between May 2005 and August 2009 eight women were killed and dumped in rural areas just outside of Jennings, the largest city and seat of JDP, LA. All but the last victim's body dump site are located in JDP. The last victim's body was found in Acadia Parish, which neighbors JDP immediately to the east. The Task Force being assigned to this crime series assumes that all homicides are linked to the same perpetrator. All body dump sites (numbered from 1 through 8) and the date of recovery (in parenthesis) are shown in Figure 1. In the middle top part of Figure 1 the city of Jennings can be seen. The age of the women ranged from 17 to 30 years, six of the women were whites and 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. Unfortunately, the Jennings homicide series is not unique and a large number of very similar drug-involved prostitute crime series exist around the US (Fox and Levin 2010; Quinet 2011).

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 has 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.

The Jennings Police Task Force uses the Orion database of "Information Packages (IPs)" to store email correspondence, transcribed face-to-face interviews, and phone calls associated with its homicide series. For this study only the 172 IPs related to the last of the eight victims,

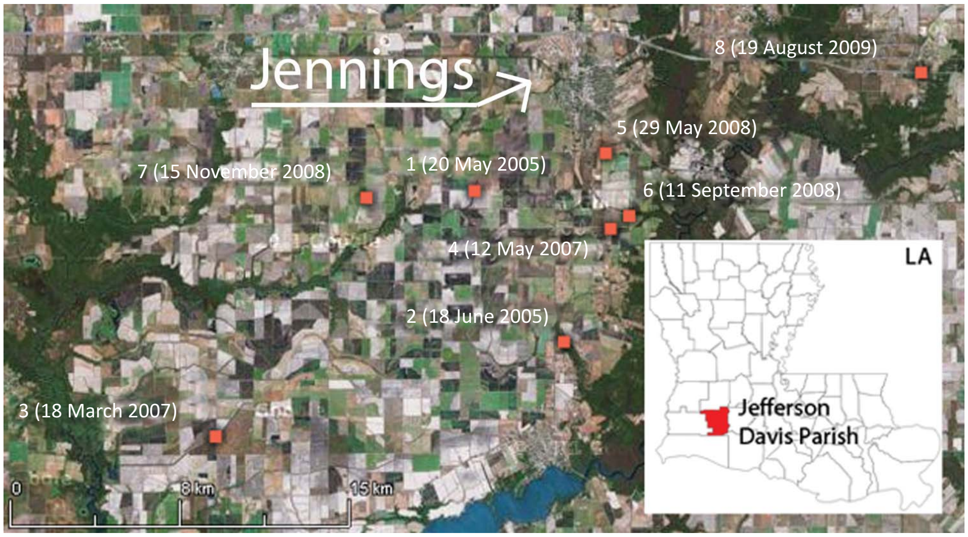


Figure 1. The location of Jefferson Davis Parish in the State of Louisiana and the location of all eight body dump sites superimposed over a satellite image by Google.

Necole Guillory, were extracted from Orion and subsequently analyzed. One example of a rather short IP used in this research is shown in the appendix. For reasons of confidentiality, personal information, such as names and addresses, are masked. Obviously, such a vast number of unstructured data records can only be efficiently analyzed with a text data mining approach.

Methodology

This section illustrates our methodological workflow. Following necessary pre-processing tasks of the IPs ("Data pre-processing" section), the SOM algorithm is applied to derive meaningful clusters and for visualizing the results ("Self-organizing maps" section). Finally, the clusters are linked back to the geographic space, which permits the exploration of the spatial distribution of the clusters using point pattern analysis methods ("Bivariate *K* (*d*)-function" section).

Data pre-processing

In order to analyze the IPs, basically text documents, it is necessary to transform them into a data representation, which enables computational processing. To achieve this task, this study employs information retrieval methods (e.g., Manning, Raghavan, and Schütze 2008). At the beginning, all words are extracted from the IP documents and stored as plain text. For computational reasons, the order of words is neglected. All words that are not part of the main text (e.g., headings) are discarded. Additionally, very frequently and rarely occurring words are removed because they do not affect the clustering of the reports, but reduce the dimensionality of the data set considerably. Testing several cut-off values, words shorter than three characters and that occur less than 5 times or more than 100 times per IP are removed. Furthermore, commas, article prepositions, conjunctions, and non-informative stop words (e.g., "and", "which", "that") are eliminated. The remaining words are set to lower case. 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 this filtering process, a stemming algorithm is employed (Porter 1980). A word stem represents the part of the word that is common to all its inflected variants. By removing the affixes of the words, the terms are transformed to their roots, which carry the crucial aspects of the semantic content (Singhal 2001). As an example, the three words "maps", "mapped", and "mapping" are reduced to their common root "map". The stemming algorithm sometimes alters the orthography of words, e.g., the letter "y" is often substituted by "i". This makes it sometimes difficult to directly infer the original word from a word stem with the consequence that the actual meaning of such a word stem may become unclear. Table 1 provides a list of ambiguous or unclear word stems that are relevant for this research, their corresponding words, and a brief explanation of their meaning, if necessary. Due to privacy reasons, word stems of first and last names of persons other than the victims' are anonymized and coded. For example, the code "fn1" refers to the first name of a person with the number 1 assigned to it and "ln3" to a person's last name with the number 3 assigned to it.

In order to represent word stems as an algebraic model, the data are transformed in accordance with the

Table 1. Ambiguous or unclear word stems and their meaning.

Word stem	Explanation	Word stem	Explanation
Deputi	Deputy	Roug	Baton Rouge
Unusu	Unusual	Lpr	License plate recognition
Calcasieu	Calcasieu Parish	Boi	Boy
Det	Abbreviation of detective	Detect	Detective
Viewip	Part of a referred URL	Mailto	Reference to an email address
Goe	Goes	Lspcl	Louisiana State Police Crime Laboratory
Impala	Chevrolet Impala	Gmc	General Motors Company
Piec	Pieces	Leo	Part of a referred URL

vector space model (Singhal 2001). Thereby, the reports are represented as vectors, where words are the columns and documents represent the rows. Finally, the term frequency–inverse document frequencies (TFIDF) of all word stems for each IP are calculated. The TFIDF measures the importance of a term in a document collection by relating the occurrence of a term in that document to the total number of occurrences of that term in all documents (Harman 1992). Most IPs contain geographic information, for instance, in the form of coordinates related to people's residences. These coordinates are also extracted from the IPs and are assigned to the corresponding vector to enable geographic mapping of the results.

Self-organizing maps

Several algorithms have been proposed for data clustering, including the *k*-means algorithm and SOMs (Kohonen 2001). The latter is an unsupervised artificial neural network. The main reason for using SOMs in this research is that several performance studies have verified that SOMs are superior compared to other alternative algorithms (e.g., Watts and Worner 2009). Additionally, several authors, including Chen et al. (2003) as well as Alruily, Ayesh, and Al-Marghilani (2010) have promoted SOMs specifically for text mining. Because SOMs reduce the high-dimensional input vector to a low-dimensional output map, they have gained popularity in geographic information science (see Agarwal and Skupin 2008; Hagenauer, Helbich, and Leitner 2011). The SOM creates models of different types of data in the data set and organizes these models in an ordered fashion in a map. Kaski and Kohonen (1996) point out that the SOM can also be interpreted as an adaptive display method, which is particularly suitable for the representation of complex and large data sets.

The SOM consists of an arbitrary number of neurons, determining the dimension of a SOM. In practice, only two-dimensional SOMs are used for visualization purposes (Vesanto and Alhoniemi 2000). Associated with each neuron is a prototype vector – in our case a 1197 dimensional vector – of the same dimension as the input space. Additionally, neighboring neurons are connected with each other, reflecting topological relationships in the map. A set of input vectors is used to train the SOM. For each input vector the neuron with the shortest Euclidean distance of its prototype vector is determined, which is also referred to as the best matching unit (BMU). Then, the BMU's prototype vector and the prototype vectors in 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. Thus, at the end of the training phase only small changes are made to fine tune the map. After training, the SOM represents a two-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 a more detailed discussion of SOMs refer to Hagenauer and Helbich (2013).

In order to analyze and interpret the map's structure, U-matrices (Ultsch and Siemon 1990) are a convenient method to visualize SOMs (Vesanto 1999). 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 clusters in the input space are less distinct. Thus, the U-matrix shows both present cluster structures and the quality of the clustering. To avoid visual cluster delimitations which are subjective to the analyst, a watershed algorithm (Vincent and Soille 1991) is often applied to segment the U-matrix computationally.

Bivariate K(d)-function

To explore the mapped coordinates of the IPs of each identified SOM cluster, the bivariate extension of the Ripley's $K(d)$-function is employed. This function is a widely used second-order statistic (e.g., Helbich 2012) and allows determining the strength and type of the spatial association between different spatial point patterns (Dixon

2002). In this case, it is tested whether the IPs which are separated by the SOM in distinct clusters, relate to each other in geographic space. Following Rowlingson and Diggle (1993), the bivariate $K(d)$-function is formally defined as the expected number of points of pattern A, i.e., points of SOM cluster 1, within a distance d of an arbitrary point of pattern B, i.e., points of SOM cluster 2, divided by the overall density of the points in pattern A. In accordance to Helbich (2012) and for ease of interpretation and practicality, the function is often transformed to the so-called $L(d)$-function. Thus, positive $L(d)$-function values represent attraction between the two point patterns, while $L(d)$-function values below 0 indicate repulsion of the two point distributions at a given distance d. If $L(d)$ equals 0, the null hypothesis of non-spatial interaction must be accepted. The significance level represented by confidence envelopes is determined through simulations of toroidal shifts (Dixon 2002).

Results

As an initial step, all IPs are converted to plain text, which allows a simple visualization by means of a word cloud (Figure 2; Feinberg 2010). Such word clouds provide an impression of common words covered by the IPs and show the number of times a certain word appears in all IPs. This is expressed by varying font sizes that is, larger font sizes represent words that appear more often than words represented by smaller font sizes (Cidell 2010). It should be noted that the gray tones used to display the words in Figure 2 do not have a specific meaning and should simply help to distinguish the words. Except for removing commonly used and non-informative words (e.g., “a”, “the”), so-called stop words, the word cloud computation does not perform additional stemming. Figure 2 depicts the word cloud for the 172 IPs included in this research. It can be seen that the words “person”, “jennings”, “joc” (Jennings Operation Center), and “investigations”, among others, frequently occur in the IPs. Although, the interpretation is somewhat vague and subjective, the resulting word cloud suggests that *something* (“female”, “victim”, “investigation”) has happened *somewhere* (“lat/long”, “jennings”) at a *certain time* (“pm”).

While constructing a word cloud is a first step in the analysis, it is certainly limited in its contribution to uncover hidden information and relationships between IPs. For this reason, the second step in the analysis involved training a SOM. Following recent empirical SOM applications (e.g., Hagenauer and Helbich 2013; Kourtit, Arribas-Bel, and Nijkamp 2012), a SOM dimension of 8 × 6 neurons with 10,000 learning iterations is selected. 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 and Skupin 2008).

Figure 3 shows the results of the trained SOM. The U-matrix depicts a complex structure, which represents distinct areas of the input space. In particular, a few light-gray regions at the left and right, as well at the bottom of the SOM are noticeable, which result from the U-matrix segmentation by means of the watershed algorithm. Compared to the dark-gray regions located in the center of the SOM, the light-gray regions contain IPs which are similar and homogeneous to each other. In order to emphasize these light-gray regions, their outline is drawn with different colors (red, blue, and green). In this way, the three outlined regions are explored and can be interpreted similar to three clusters. Additionally, Figure 3 depicts the neuron’s word stems that have a high TFIDF value, representing important terms. The label sizes correspond to the TFIDF values, thus word stems with a high TFIDF have also large labels. To avoid overcrowding of the display, only important word stems (measured on the basis of the TFIDF values) are shown in Figure 3.

The visualization of the data set in Figure 3 reveals interesting insights into the input data set, which are not

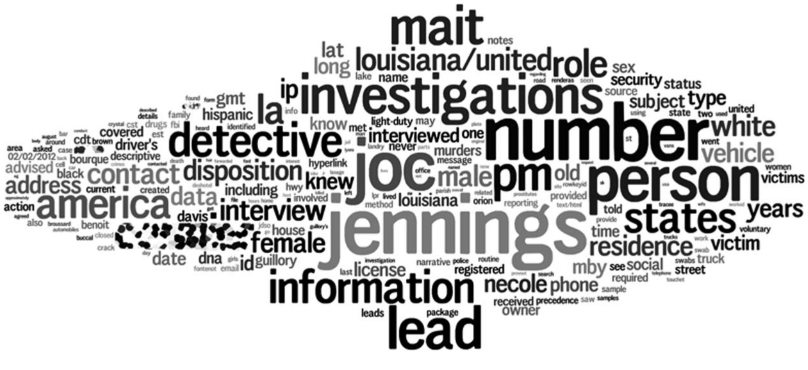


Figure 2. Word cloud of the IPs (names of people are masked for privacy reasons).

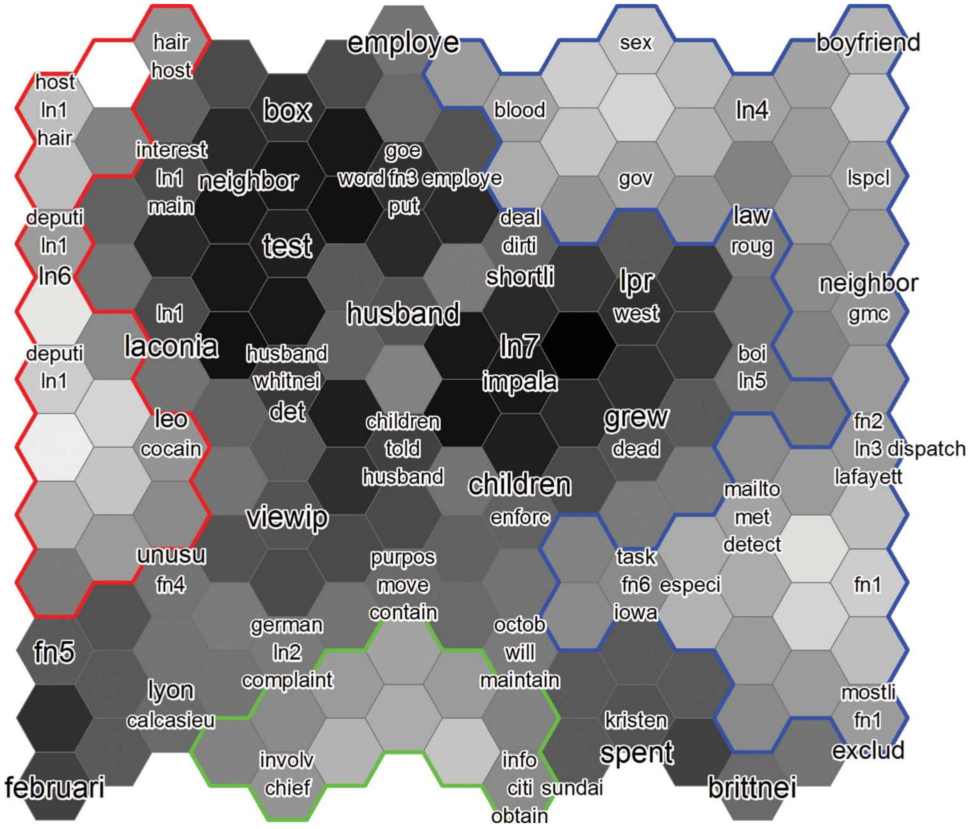


Figure 3. U-matrix. The borders of three clusters are shown in different colors (cluster 1 = red, cluster 2 = green, cluster 3 = blue).

apparent when simply using the word cloud. As far as the three revealed clusters are concerned, 17 IPs are assigned to cluster 1, 11 IPs to cluster 2, and 44 IPs to cluster 3. All other IPs cannot be mapped to a distinct cluster, such as the central area of the SOM. At first it is noticeable that some metadata word stems like "viewip" appear in the map, which indicates that they exhibit high TFIDF values. Further, some relationships between different terms are apparent. Words, such as "sex", "boyfriend", and "blood" are all located in the same cluster, indicating that IPs that include these words share many similarities. Also the murder victims "Kristen" (Kristen Lopez) and "Brittnei" (Brittney Gary) are mapped to regions in the map (in the lower right) that are close to each other. It thus can be concluded that the criminal investigations regarding both victims tended to be closely related. In contrast, 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 not all murder victims are represented in Figure 3. The reason for this is that the records that have high TFIDF 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 TFIDF values and the corresponding word stems of the resulting mean vectors for the different clusters are shown in Table 2 (higher mean TFIDF values refer to more important terms).

Cluster 1 shows a notable high value for "ln1", which refers to the word stem of a common last name. In fact, many 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 closely related to "laconia", which is mapped just outside the boundaries of cluster 1. Compared to cluster 1 and 3, cluster 2 has medium TFIDF values. Cluster 3 has the highest mean TFIDF value for the word stem of the first name "fn1". Furthermore, cluster 3 has also a high mean TFIDF value for the word stem of the last name "ln3". It should be noted that the criminal investigations refer to people with that particular last name quite often. The word stem of the last name "ln3" is mapped to the same neuron as the words "lafayette" and

Table 2. The five highest mean TFIDF values and the corresponding word stems of the three cluster means' prototype vectors.

	1st	2nd	3rd	4th	5th
Cluster 1 (Red)	Ln1, 0.223	Deputi, 0.152	Host, 0.098	Hair, 0.098	Remain, 0.091
Cluster 2 (Green)	Sundai, 0.121	Chief, 0.119	Obtain, 0.118	Involve, 0.076	Info, 0.075
Cluster 3 (Blue)	Fn1, 0.074	Piec, 0.072	Sex, 0.062	Prepar, 0.061	Ln3, 0.060

the stem of the first name "fn2", suggesting that there is a close relationship between these three words. "Lafayette" refers to the name of a city and a parish in Louisiana. Lafayette Parish is very close to JDP and both the city of Lafayette and Jennings are located on interstate I-10. Furthermore, the three clusters listed in Table 2 have high mean TFIDF values for general word stems that are likely 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 mean TFIDF values of these word stems notably differ between the three clusters. Thus, it can be summarized that distinctive patterns appear in the IPs related to the homicide.

In the next step of the analysis, the three SOM clusters and the geographic context of each IP are projected to geographical space and the resulting map being displayed in Figure 4. Several observations can be drawn from this map. Most reports associated with cluster 1 and cluster 3 are located in and around Jennings, whereas cluster 2 is sparsely distributed across the entire map. Moreover, a significant portion of the IPs from cluster 1 are located close to the city of Lafayette, whereas only a few reports are located nearby the city of Lake Charles to the west of Jennings. In contrast, only a few reports included in cluster 3 are located close to Lafayette, whereas a notable number of reports for cluster 3 are located around 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.

Nevertheless, based on the distribution of IPs in Figure 4, it is difficult to ascertain whether a spatial relationship exists between the three clusters. For this reason, bivariate $L(d)$-functions are computed between each possible pair of the three point patterns making up the three clusters. To determine statistical significance

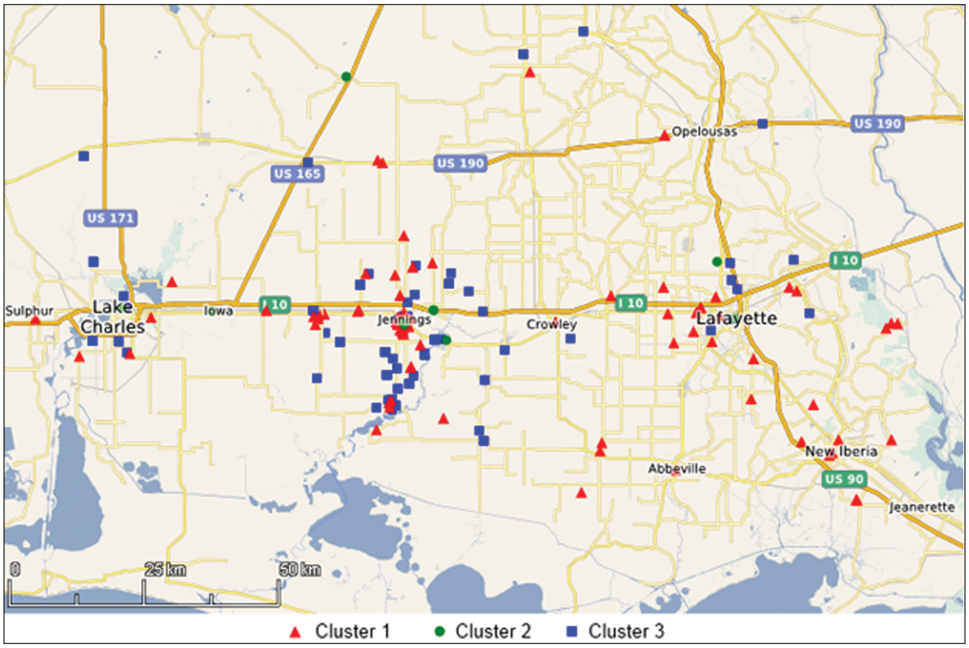


Figure 4. Geographic distribution of the IPs belonging to the three SOM clusters (Note: One IP which is located in Seattle, WA is not included).

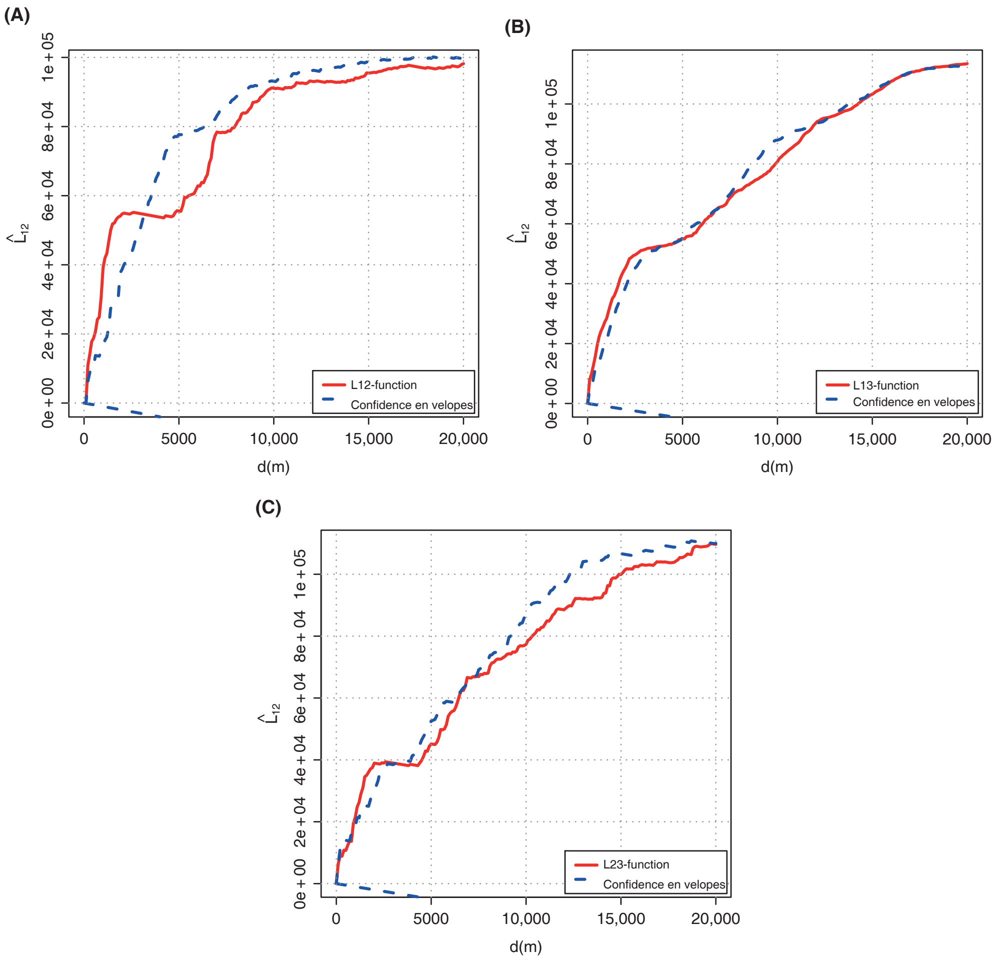


Figure 5. Bivariate *L*(*d*)-functions of mapped SOM clusters. Top left panel cluster 1 vs. cluster 2, top right panel cluster 1 vs. cluster 3, and lower panel cluster 2 vs. cluster 3. The distances *d* on the x-axes are given in meters (m).

9999 toroidal shifts are conducted. Figure 5 shows the results of these three pair-wise comparisons. In general, the three empirical functions (red solid curves in Figure 5) show rather similar patterns. Since function values are all above 0, there is some evidence of attraction in geographic space between the IPs mapped in Figure 4. Within a distance interval of 0 and 3 km, the empirical functions of cluster 1 versus cluster 2 as well as cluster 1 versus cluster 3 are above the upper confidence envelops (represented by the blue dotted curves in Figure 5). Thus, for this distance interval evidence of significant spatial attraction of the points between clusters 1 and 2 as well as clusters 1 and 3 can be observed. In contrast, the empirical function comparing clusters 2 with 3 is only significant within a small interval from 2 to 3 km. For all other distance intervals, the three empirical functions consistently fall between the upper and the lower confidence envelopes. This means that, with the exception of two small distance intervals, the three SOM clusters in a pair-wise comparison do not show significant spatial interaction (i.e., spatial attraction) in geographical space. Overall, the three SOM clusters are not spatially co-located, meaning that attribute similarity is statistically independent of the locational similarity.

Conclusions

This article analyzes whether crime reports or protocols of eyewitnesses and/or the general public can serve as valuable subsidiary information source for law enforcement in their criminal investigations. Because such reports/protocols are only available in the form of unstructured text documents, a text data mining approach is needed to extract possible hidden relationships and information. This research is based on 172 text documents, provided as IPs by the Task Force investigating a serial homicide case in the city of Jennings, LA. The methodological approach analyzing the content of these text documents focuses on the visualization and clustering capabilities of an unsupervised neural network approach, specifically the SOM algorithm.

Based on the word content of each text document, the SOM analysis resulted in three distinct clusters of all 172 IPs. The mapping of the IPs from each cluster as point patterns in geographical space allows the spatial arrangement of such IPs to be explored further. The results provide evidence that the three attribute-based IP clusters that were originally extracted by the SOM show only a partially significant spatial association in geographic space. The results from this data mining exercise have already been presented to and shared with the Jennings Task Force, which confirmed that this information was previously unknown and may provide new and important clues in this criminal investigation. However, due to confidentiality reasons and this being still an open criminal investigation, the authors of this research cannot go into more detail as far as the specifics of this "previously unknown information" and "new and important clues" are concerned.

This research demonstrates that such "collective surveillance" by individuals who serve as "human crime sensors" providing voluntary information have the potential to become a powerful data source to possibly revolutionize crime protection and crime combat methods. Besides the interpretation of the results, it should be noted that the effectiveness and the impact of the proposed multistep approach to eventually solving a crime is partly affected by the subjective interpretation of the analyst and difficult to evaluate, unless it would directly lead to the apprehension of the serial offender. However, criminal investigations that have already been solved would allow the evaluation of the multistep approach applied in this research. This can be done 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. Although the evaluation of exploratory spatial data analysis results is inherently challenging and remains an active research topic, the results from this research recommend law enforcement agencies to undertake an *ex post* evaluation as soon as the case is solved. As a matter of fact, one future research direction that the authors of this article take is to evaluate the proposed multistep approach using "collective surveillance" information from the community (students, faculty, and staff) of the Louisiana State University (LSU) in Baton Rouge with solved crime cases that have been collected and investigated by the LSU police.

Acknowledgements

Marco Helbich and Julian Hagenauer have been funded by the German Research Foundation's (DFG) Global Networks Initiative of Excellence of the University of Heidelberg. Marco Helbich also acknowledges the Alexander von Humboldt-Foundation for funding. We gratefully acknowledge support for this research from Jerry Ratcliffe and Spencer Chainey. We also thank Maxim Rylov, who helped us with the visualization of the content of the information packages using the MapSurfer.NET software and the AutoCarto 2012 program committee for selecting this article for possible inclusion in this Special Issue of *CaGIS*.

References

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

Alruily, M., A. Ayesh, and A. Al-Marghilani. 2010. "Using Self Organizing Map to Cluster Arabic Crime Documents." In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 357–363. Los Alamitos, CA: IEEE Computer Society Press.

Brantingham, P. J., and P. L. Brantingham, eds. 1981. *Environmental Criminology*. Beverly Hills, CA: Sage.

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

Chau, M., J. J. Xu, and H. Chen. 2002. "Extracting Meaningful Entities from Police Narrative Reports." In *Proceedings of the National Conference for Digital Government Research*, 271–275 Los Angeles, CA: Digital Government Society of North America.

Chen, H., H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, et al. 2003. "COPLINK: Visualization for Crime Analysis." In *Proceedings of the 2003 Annual National Conference on Digital Government Research*, 1–6. Boston, MA: Digital Government Society of North America.

Chen, H., W. Chung, J. J. Xu, G. W. Y. Qin, and M. Chau. 2004. "Crime Data Mining: A General Framework and Some Examples." *Computer* 37: 50–56.

Chen, H., and F.-Y. Wang. 2005. "Artificial Intelligence for Homeland Security." *IEEE Intelligent Systems* 20: 12–16.

Cidell, J. 2010. "Content Clouds as Exploratory Qualitative Data Analysis." *Area* 42: 514–523.

Clarke, R., and D. Cornish. 1985. "Modeling Offenders' Decisions: A Framework for Research and Policy." *Crime Justice* 6: 147–185.

Cohen, L., and M. Felson. 1979. "Social Change and Crime Rate Trends: A Routine Activity Approach." *American Sociological Review* 44: 588–608.

Costa, A. M. 2010. "The Economics of Crime: A Discipline to be Invented and a Nobel Prize to be Awarded." *Journal of Policy Modeling* 32: 648–661.

Delen, D., and M. D. Crossland. 2008. "Seeding the Survey and Analysis of Research Literature with Text Mining." *Expert Systems and Applications* 34: 1707–1720.

Dixon, P. 2002. "Ripley's K Function." In *Encyclopedia of Environmetrics*, edited by A. El-Shaarawi, and W. Piegorsch, Vol. 3, 1796–1803. Chichester: Wiley.

Feinberg, J. 2010. "Wordle." In *Beautiful Visualization. Looking at Data through the Eyes of Experts*, edited by J. Steele, and N. Iliinsky, 37–58. Sebastopol: O'Reilly.

Fox, J. A., and J. Levin. 2010. *Extreme Killing: Understanding Serial and Mass Murder*. 2nd ed. Thousand Oaks: Sage.

Goodchild, M. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69: 211–221.

Hagenauer, J., and M. Helbich. 2013. "Contextual Neural Gas for Spatial Clustering and Analysis." *International Journal of Geographical Information Science* 27: 251–266.

Hagenauer, J., M. Helbich, and M. Leitner. 2011. "Visualization of Crime Trajectories with Self-Organizing Maps: A Case Study on Evaluating the Impact of Hurricanes on Spatiotemporal Crime Hotspots." 25th. International cartographic conference, Paris, July 3–8.

Han, J., and M. Kamber. 2011. *Data Mining. Concepts & Techniques*. Amsterdam: MK.

Harman, D. 1992. "Ranking Algorithms." In *Information Retrieval. Data Structures & Algorithms*, edited by W. B. Frakes, and R. Baeza-Yates, 363–392. Upper Saddle River, NJ: Prentice Hall.

Helbich, M. 2012. "Beyond Postsuburbia? Multifunctional Service Agglomeration in Vienna's Urban Fringe." *Journal of Economic & Social Geography* 103: 39–52.

Helbich, M., and M. Leitner. 2012. "Evaluation of Spatial Cluster Detection Algorithms for Crime Locations." In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organization*, edited by W. A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, and J. Kunze 193–201. Heidelberg: Springer.

Jones, C. B., and R. S. Purves. 2008. "Geographical Information Retrieval." *International Journal of Geographical Information Science* 22: 219–228.

Kaski, S., and T. Kohonen. 1996. "Exploratory Data Analysis by The Self-Organizing Map: Structures of Welfare and Poverty." In: *Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, 498–507. Singapore: World Scientific.

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

Kourtit, K., D. Arribas-Bel, and P. Nijkamp. 2012. "High Performers in Complex Spatial Systems: A Self-Organizing Mapping Approach with Reference to the Netherlands." *Annals of Regional Science* 48: 501–527.

Ku, C. H., A. Iriberri, and G. Leroy. 2008. "Crime Information Extraction from Police and Witness Narrative Reports." IEEE international conference on technologies for homeland security, May 12–13.

Leitner, M., and M. Helbich. 2011. "The Impact of Hurricanes on Crime: A Spatio-Temporal Analysis in the City of Houston, Texas." *Cartography and Geographic Information Science* 38: 214–222.

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

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14: 130–137.

Quinet, K. 2011. "Prostitutes as Victims of Serial Homicide: Trends and Case Characteristics, 1970-2009." *Homicide Studies* 15: 74–100.

Rowlingson, B., and P. Diggle. 1993. "SPLANCS: Spatial Point Pattern Analysis Code in S-Plus." *Computers and Geosciences* 19: 627–655.

Singhal, A. 2001. "Modern Information Retrieval: A Brief Overview." *IEEE Data Engineering Bulletin* 24: 35–43.

Ultsch, A., and H. P. Siemon. 1990. "Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis." In: *Proceedings of International Neural Networks Conference*, 305–308. Paris: Kluwer Academic Press.

Vesanto, J. 1999. "SOM-Based Data Visualization Methods." *Intelligent Data Analysis* 3: 111–126.

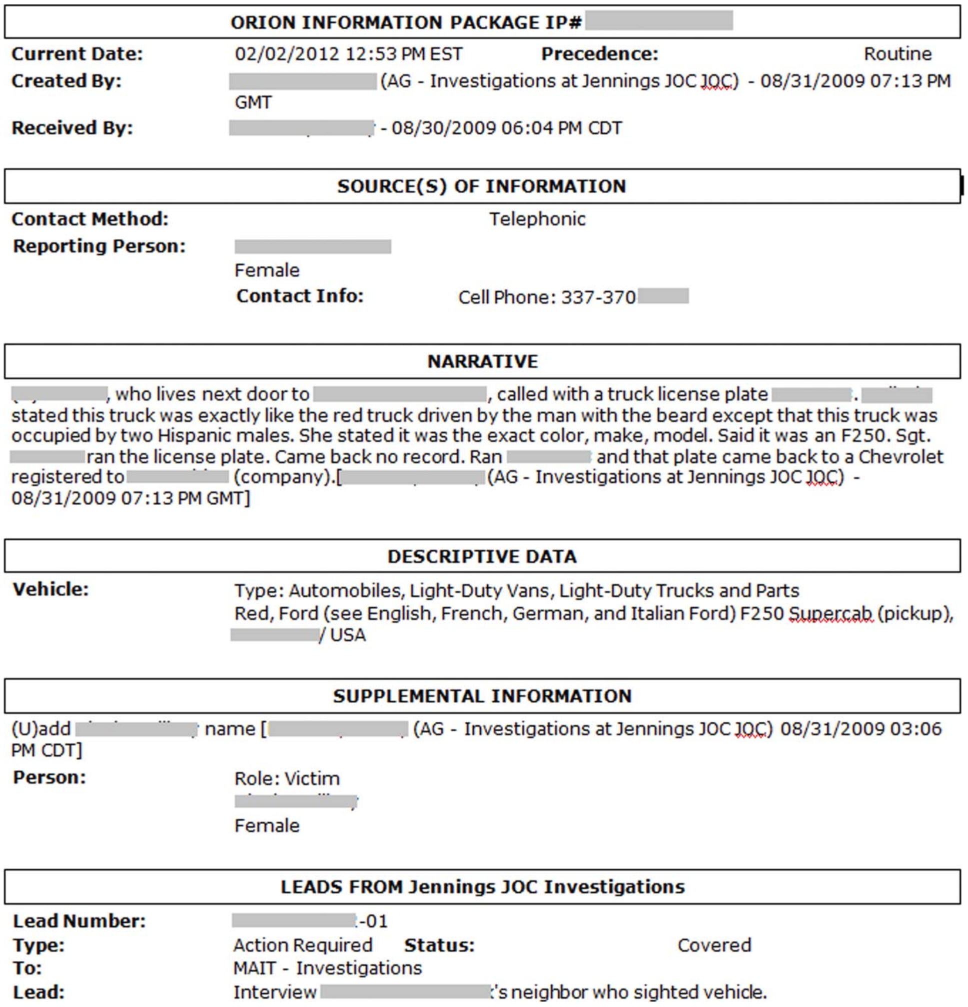
Vesanto, J., and E. Alhoniemi. 2000. "Clustering of the Self-Organizing Map." *IEEE Transactions on Neural Networks* 11: 586–600.

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

Watts, M. J., and S. Worner. 2009. "Estimating the Risk of Insect Species Invasion: Kohonen Self-Organising Maps Versus k-Means Clustering." *Ecological Modeling* 220: 821–829.

Appendix

As an example, the below document shows one anonymized information package extracted from the Orion database.



ORION INFORMATION PACKAGE IP# [redacted]

Current Date: 02/02/2012 12:53 PM EST **Precedence:** Routine
Created By: [redacted] (AG - Investigations at Jennings JOC JOC) - 08/31/2009 07:13 PM GMT
Received By: [redacted] - 08/30/2009 06:04 PM CDT

SOURCE(S) OF INFORMATION

Contact Method: Telephonic
Reporting Person: [redacted]
Female
Contact Info: Cell Phone: 337-370[redacted]

NARRATIVE

[redacted], who lives next door to [redacted], called with a truck license plate [redacted]. [redacted] stated this truck was exactly like the red truck driven by the man with the beard except that this truck was occupied by two Hispanic males. She stated it was the exact color, make, model. Said it was an F250. Sgt. [redacted] ran the license plate. Came back no record. Ran [redacted] and that plate came back to a Chevrolet registered to [redacted] (company).[[redacted] (AG - Investigations at Jennings JOC JOC) - 08/31/2009 07:13 PM GMT]

DESCRIPTIVE DATA

Vehicle: Type: Automobiles, Light-Duty Vans, Light-Duty Trucks and Parts
Red, Ford (see English, French, German, and Italian Ford) F250 Supercab (pickup), [redacted] / USA

SUPPLEMENTAL INFORMATION

(U)add [redacted] name [[redacted] (AG - Investigations at Jennings JOC JOC) 08/31/2009 03:06 PM CDT]

Person: Role: Victim
[redacted]
Female

LEADS FROM Jennings JOC Investigations

Lead Number: [redacted]-01
Type: Action Required **Status:** Covered
To: MAIT - Investigations
Lead: Interview [redacted]'s neighbor who sighted vehicle.

Figure 6.