Driving Forces of Non-Violent Crime in Houston, TX: A Spatially Filtered Negative Binomial Model

Marco HELBICH[1], Jamal JOKAR ARSANJANI[1] and Michael LEITNER[2]

[1]Institute of Geography, University of Heidelberg/Germany · helbich@uni-heidelberg.de
[2] Louisiana State University, Baton Rouge/USA

This contribution was double-blind reviewed as full paper.

Abstract

The analysis and understanding of spatial crime patterns is crucial for law enforcements to improve strategic and tactical decision-making. In this context, generalized linear models, such as count regressions, are commonly applied. These non-spatial models are challenged by spatial autocorrelation effects, contradicting fundamental model assumptions. Therefore, the purpose of this research is to present a spatially explicit approach, which combines a negative binomial model and spatial filtering to explain the spatial distribution of non-violent offences in Houston, TX, for the year 2010. The results provide evidence that the non-spatial negative binomial model is biased while the supplementary consideration of a spatial filter is capable to absorb these undesirable spatial effects and results in a well-specified regression model. Moreover, besides the significant importance of space in the explanation of the non-violent crime patterns, only the percentage of renter-occupied housing units and the percentage of Asian population are significantly related to the crime. The former covariate has a stimulating effect while the latter has an inhibiting effect.

1 Introduction

Since the last two decades geographic information system and spatial analysis are effectively used in daily operations of governmental agencies involved in public security and safety (WANG 2012; LEITNER 2013). Their applications have grown tremendously, comprising, for instance, of hotspot analysis (e.g. HELBICH & LEITNER 2012), crime forecasting on the basis of time series analysis (e.g. BERK 2008), and data mining applications (e.g. HELBICH et al. 2013b). In particular, regression models are of great importance to law enforcement agencies and academic researchers (e.g. OSGOOD 2000). These models support the understanding of the underlying spatial and social processes affecting and contributing to the presence or absence of criminal offenses. But mostly, traditional regression models are challenged by the fact that criminal activities are not spread evenly across space and tend to clump in certain areas. For instance, LEITNER and HELBICH (2011) have shown that socially disadvantaged areas in Houston, TX, exhibit higher crime rates. This has serious consequences for most statistical analysis conducted with area-based crime data. Such a coincidence of locational and attribute similarity is referred to as spatial autocorrelation, a well-

Jekel, T., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2013): GI_Forum 2013. Creating the GISociety.
 ISBN 978-3-87907-532-4.
 eISBN 978-3-7001-7438-7, doi:10.1553/giscience2013s117.

known concept in geography (e.g. HELBICH et al. 2013a). To receive unbiased and correct inference results, spatial autocorrelation must be explicitly considered in statistical analysis (TITA & RADIL 2011). This is non-trivial in count regressions, where the response variable used is the number of crimes within a spatial unit (e.g. OSGOOD 2000; GRIFFITH & HAINING 2006). Bayesian spatial modeling provides an alternative but such models are highly complex and computationally intensive and thus of limited use for non-statisticians and larger datasets.

Therefore, the present research makes the following two contributions: First, since count regression models are misspecified whenever spatial autocorrelation is present in the residuals, spatial eigenvector filtering is introduced and applied to crime analysis and modeling. Thus, the paper responds to a recent call by BERNASCO and ELFFERS (2011) that spatial filtering might be a relevant technique for quantitative criminology. Second, by analyzing the distribution of non-violent crime for the year 2010 in Houston, TX, a deeper understanding of the major driving forces is gained which allows formulating more appropriate and situational policies and actions from law enforcements.

The paper is structured as follows: Section 2 presents the study area and data, while section 3 briefly introduces the main aspects of count regression, specifically the negative binomial model, and spatial filtering. Main results are discussed in section 4. The paper concludes through a discussion of major findings.

2 Study Area and Data

The study area is the metropolitan area of Houston, TX, which is located in Harris County. To avoid enclaves within the study site, a slightly modified version of the original metropolitan area is considered to secure a coherent area. The spatial units for subsequent statistical analysis comprise of 467 census tracts for the year 2010, which represent an official administrative unit commonly used by the U.S. Census Bureau. This area is "controlled" by the Houston Police Department, one of the ten largest police departments in the U.S., serving a population of over two million residents, and covering an area of approximately 600 square miles.

Crime data were received from the police department by an official data request in 2012. Approximately 120,200 offences occurred within the study area in 2010. The police agencies classify each offence in accordance to the classification scheme provided by the Federal Bureau of Investigation. This study is exclusively based on 99,600 non-violent crimes, while the remaining 20,500 violent offenses are not further considered. Following the Uniform Crime Reporting classification scheme, non-violent crimes include the following offense types: burglaries, larceny, auto theft, and arson. After pre-processing (e.g. data cleaning) all non-violent crime data, the crime incident locations are geocoded by means of the TIGER street network provided by the U.S. Census Bureau. With a geocoding accuracy of about 92%, the results show a high completion rate and guarantee a high correspondence with the actual crime distribution. Subsequently, these geocoded crime data are aggregated to each corresponding census tract and henceforth reflect the corresponding variable in the regression analysis. Figure 1 provides an overview of the spatial crime distribution for the year 2010.

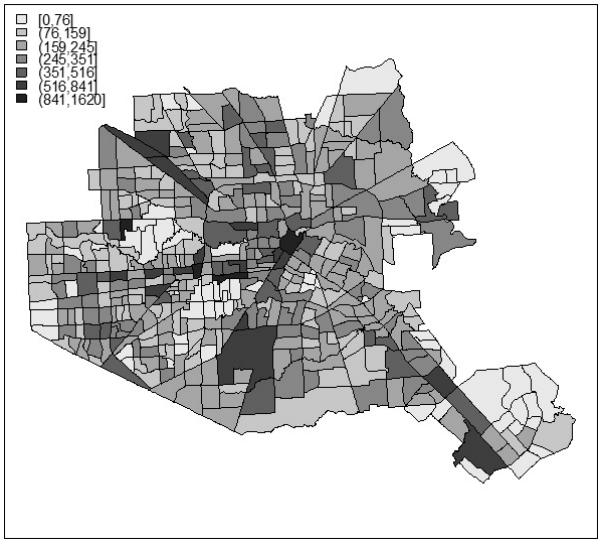

Fig. 1: Spatial distribution of non-violent crimes in Houston (TX) in 2010

Figure 1 makes this fact evident that these patterns reflect some spatial variation in non-violent crimes. The majority of the census tracts depict less than 400 crimes during the study period (figure 2), while others show an exceptional high amount of crime (dark areas in figure 1).

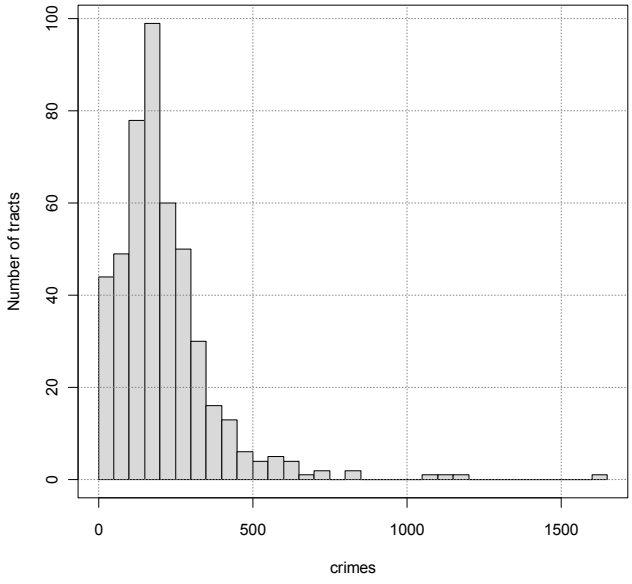

Fig. 2: Number of non-violent crimes per census tract in Houston (TX) in 2010

The utilization of census tracts allows to easily link the crime data with several underlying socio-economic and ethnic characteristics, which may serve as driving forces of crime. Based on LEITNER and HELBICH (2011), seven variables are considered as independent variables. These variables are published by the U.S. Census Bureau and cover the same time period as the crime data, namely 2010. The following variables are considered within the spatial analysis:

1) percentage of white population (WHPOP),
2) percentage of black population (BLPOP),
3) percentage of Asian population (ASPOP)
4) percentage of renter-occupied housing units (ROH)
5) homeowner vacancy rate (%; HVR)
6) rental vacancy rate (%; RVR), and
7) Euclidian distance to the nearest police station (meters; DIST).

3 Methods

This section describes the main modeling steps. The applied analysis framework consists of two parts: First, a count regression model is introduced to determine the relevant driving forces, explaining the amount of crime that occurred at a statistically significant level (section 3.1). Second, due to expected existing spatial autocorrelation effects, a spatial filter is derived and integrated into the count regression model, which aims to absorb biasing spatial autocorrelation effects (section 3.2).

3.1 Negative Binomial Regression Model

In general, count regression analysis relates the response variable to several independent variables, which might be theoretically relevant to explain a given response (CAMERON & TRIVEDI 1998). Because count data are formed by non-negative integers, among other particularities, basic ordinary least square models are not appropriate and generalized models, in particular Poisson regressions, are required (CAMERON & TRIVEDI 1998; COXE et al. 2009). Empirical criminological studies (e.g. OSGOOD 2000) clearly note that Poisson regression is based on model assumptions that are too strict, i.e. that the mean is equivalent to the variance, which is rarely the case in reality and erroneously results in overdispersed models.

Thus, the negative binomial model is a suitable extension, which relaxes this assumption about the mean-variance equidispersion through an additional variance specification explicitly controlling for overdispersion (COXE et al. 2009). In particular, OSGOOD (2000) promotes the negative binomial model for crime analysis. Nevertheless, the spatial independence of the residuals is a mandatory requirement to receive unbiased estimates (GRIFFITH & HAINING 2006). However, the application of a spatial filter (see section 3.2 for details) should fulfill this prerequisite and enable residuals not to be affected by spatial autocorrelation. This, of course, requires a strict testing of model assumptions.

The following negative binomial model, in which spatial autocorrelation is modeled by means of spatial filtering is empirically tested in this study:

$$CRIME_i = \beta_0 + \beta_1 \mathrm{WHPOP}_i + \beta_2 \mathrm{BLPOP}_i + \beta_3 \mathrm{ASPOP}_i + \beta_4 \mathrm{ROH}_i + \beta_5 \mathrm{HVR}_i + \beta_6 \mathrm{RVR}_i + \beta_7 \mathrm{DIST}_i + \sum_{l=1}^{l} \beta_l EV_{l,i} + \varepsilon_i$$

Where *CRIME* is the absolute number of non-violent crimes in census tract *i*, the βs are the coefficients to be estimated, the independent variables are those introduced in section 2, and the EV_l are a subset of selected spatial filters which turn the non-spatial negative binomial model in a spatially explicit model. ε represents the error term.

3.2 Spatial Filtering

The absence of spatial autocorrelation is a fundamental regression model assumption. A widely used approach to investigate spatial autocorrelation is the Moran's *I* statistic, on which the recently introduced spatial eigenvector filtering approach is based upon (GRIFFITH 2000; TIEFELSDORF & GRIFFITH 2007). Essentially, spatial eigenvector filtering extracts eigenvectors (EVs) from an exogenously defined spatial contiguity matrix, which represents the spatial arrangement of the census tracts. Possibilities of how to define such a spatial contiguity matrix are discussed, for instance, in PATUELLI et al. (2011). For simplicity, this study only focuses on one contiguity definition, specifically the first-order queen contiguity with a globally standardized coding scheme. EV decomposition is used to extract orthogonal and uncorrelated EVs from the following matrix:

$$\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^T}{N}\right)\boldsymbol{C}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{N}\right)$$

where $\boldsymbol{I}$ represents the $N \times N$ identity matrix having 1s in the main diagonal and 0s elsewhere, $\mathbf{1}$ is a $N \times 1$ vectors of 1s, $\boldsymbol{C}$ gives the spatial arrangement of *N* spatial units, and *T* denotes the matrix transpose. The first EV has the largest achievable Moran's *I* value for a given contiguity definition. The second EV contains the set of numerical values that has the largest achievable Moran's *I* by any set of numerical values uncorrelated with the first EV, and so on. As outlined by TIEFELSDORF and GRIFFITH (2007), each of these *N* EVs represents a distinct map pattern, ranging from global patterns to local ones. In general, the global patterns follow the main cardinal directions while higher EVs show some regional or local map characteristics. Within the next step, these *N* EVs are reduced to a set of candidate EVs. As a rule-of-thumb criterion, TIEFELSDORF and GRIFFITH (2007) recommend EVs having a certain degree of autocorrelation (approximately 0.25). The application in this research follows a recent study by PATUELLI et al. (2011) and investigates positive autocorrelation effects, usually present in geospatial data. Subsequently, to receive the final set of EVs, the response variable is regressed on the set of candidate EV and stepwise variable selection is utilized to determine only the EVs significantly related to the response variable. Finally, a spatial filter is constructed by constituting a linear combination of the significant EVs. This filter can now be included as an additional independent variable in regression models and serves as a surrogate for possible missing predictors and absorbs undesirable spatial residual autocorrelation.

4 Results

Descriptive statistics show that the mean number of crimes across all census tracts is 213. To further investigate the spatial crime distribution statistically and to evaluate whether spatial autocorrelation may affect regression analysis, the global Moran's *I* statistic is computed. As outlined above, a globally standardized first-order queen contiguity matrix is used. With a value of 0.206, the Moran's *I* statistic confirms significant spatial autocorrelation (p-value < 0.001), meaning that similar values are located nearby in space. Therefore, the spatial filtering approach seems to be a rational choice to avoid complications associated with spatial autocorrelation. Following section 3.2, 467 EVs are extracted for the given contiguity matrix. As expected, the first EVs comprise of the highest positive Moran's *I* values, while the last EVs show the lowest negative Moran's *I* values. The transition of the Moran's *I* values from positive to negative autocorrelation is obvious in figure 3.

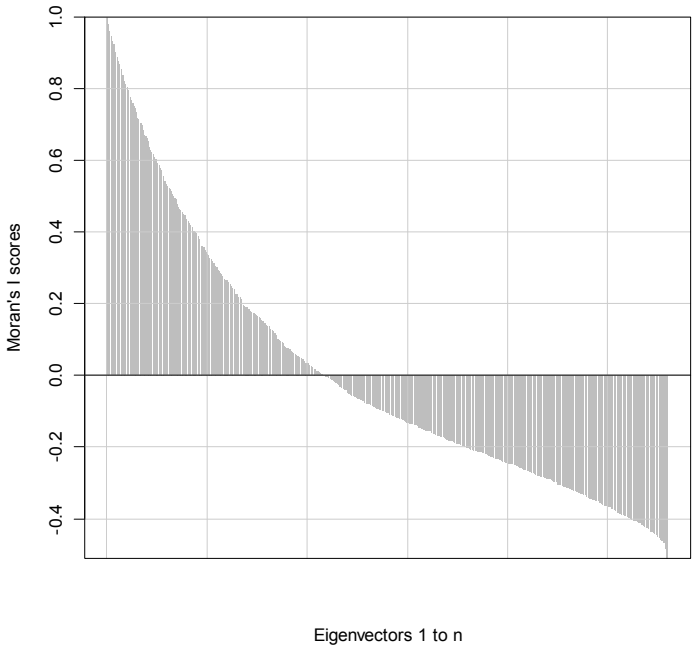

Fig. 3: Moran's *I* values derived for each EV

To reduce these 467 EVs to a set of candidate EVs, a Moran's *I* threshold value of 0.3 is applied which results in 92 candidates. Four examples of the extracted EV are visualized in figure 4. While EV3 (e3) depicts large-scale trends, the lower ranked EVs, e.g. EV15 (e15), reflects more regional patterns. In contrast, EV40 (e40) and EV78 (e78) show local patterns. Because not all EVs are significantly related to the response variable, a stepwise variable selection algorithm which minimizes the Akaike information criterion (AIC) is applied for a negative binomial model. Choosing this model is necessary because both the basic Poisson and the Quasi-Poisson models reveal considerable overdispersion, contradicting an

important model assumption. For example, the overdispersion test statistic for the basic Poisson model clearly rejects equidispersion (p-value < 2.2e-16). A re-estimation using a Quasi-Poisson model does not show any improvement. The final negative binomial model incorporates 43 EVs significantly related to the crime pattern. In contrast to the basic and Quasi-Poisson model, the final negative binomial model shows negligible overdispersion of 1.259 that is close to the ideal value of 1. Lastly, the 43 EVs are combined by means of a linear combination and constitute the spatial filter.

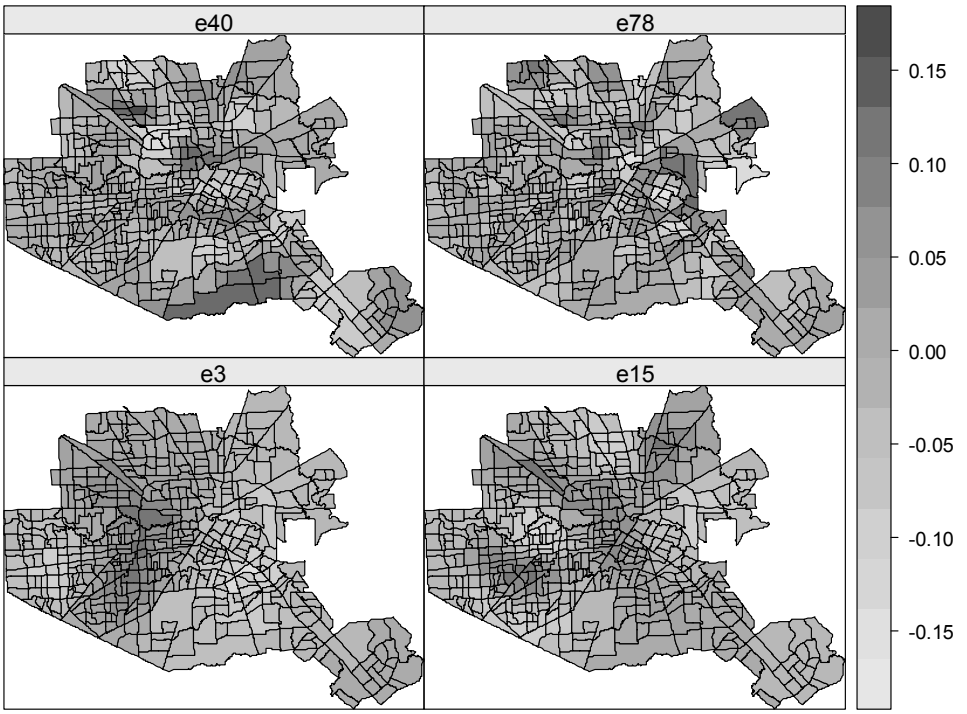

Fig. 4: Spatial patterns of four EVs representing global (e3), regional (e15, e40), and local spatial autocorrelation patterns (e78)

To investigate the relationship between the crime pattern and the selected socio-economic and ethnical variables, count regressions are estimated. As before, only the negative binomial model is not affected by overdispersion. Significant residual autocorrelation (Moran's I = 0.125; p-value < 0.001) of the non-spatial model justifies the application of the spatially filtered negative binomial model. Testing the independence assumption for the spatially filtered model confirms that the residuals are randomly distributed across space (Moran's I = -0.031; p-value = 0.265) and that spatial autocorrelation is effectively modeled. Overall, the model's pseudo-R^2 indicates that approximately 43% of the variation is explained. The AIC provides clear evidence that the spatial NBM (AIC = 5,851) is more appropriate compared to the non-spatial model (AIC = 5,871). Detailed estimation results of the final model are provided below in Table 1.

Table 1: Estimation results of the spatially filtered negative binomial model

	Estimate	Std. Error	*z*-value	Pr(>|*z*|)	
Intercept	4.9370	0.2860	17.263	< 2e-16	***
WHPOP	-0.0003	0.0033	-0.115	0.908	
BLPOP	0.0001	0.0031	0.036	0.971	
ASPOP	-0.0096	0.0045	-2.140	0.032	*
ROH	0.0059	0.0015	4.032	5.53e-05	***
HVR	0.0172	0.012	1.385	0.166	
RVR	0.0012	0.0046	0.271	0.787	
DIST	0.0000	0.0000	0.280	0.779	
Spatial filter	0.8903	0.0631	14.108	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

The model estimation reveals that two out of the seven predictors are significantly associated with the non-violent crime pattern. The covariate percentage of renter-occupied housing units (ROH) shows a significant positive relationship ($p < 0.001$) to the number of non-violent crimes and the covariate percentage of Asian population (ASPOP) shows a significant negative one ($p < 0.050$). As such, both variables can be interpreted as criminogenic factors. The former stimulates crime while the latter inhibits crime. Furthermore, the spatial filter is highly significant and provides empirical evidence that location matters (see also LEITNER & HELBICH 2011). This confirms a previous study by MORENOFF et al. (2001), which concludes that spatial effects are of utmost importance in analyzing homicide rates.

5 Summary and Conclusions

Area-based crime analysis is often affected by spatial autocorrelation, having serious consequences for aspatial regression models. While for Gaussian models a rich set of possible alternatives exist to account for such patterns (e.g. autoregressive models), little research has been conducted on generalized linear models, including count regressions.

The main objective of this paper was to demonstrate the usefulness of a relatively novel approach, namely spatial eigenvector filtering, to model spatial autocorrelation explicitly in count regression models. Using non-violent crime as well as selected socio-economic and ethnical data for the year 2010 for the city of Houston, TX, the usefulness and effectiveness of this approach is demonstrated. It is shown that the derived spatial filter absorbs the existing residual autocorrelation effects through a negative binomial model yielding a well-specified model. Neglecting these residual autocorrelation effects would have unequivocally resulted in a misspecified model and resulted in wrong conclusions drawn by law enforcement agencies. Moreover, the model confirms that the percentage of renter-occupied housing units as well as the percentage of Asian population are significant criminogenic factors explaining Houston's non-violent crime pattern. The significance of the spatial filter also points to the relevance of "space" being an important variable to explain the "spatial" distribution of crime patterns.

To conclude, correctly specified models are not only of particular importance in terms of scarce monetary resources for law enforcement policies and safety, but also to get a deeper

understanding of criminological processes. In this sense, spatial eigenvector filtering provides a cutting-edge spatial statistical technique which turns basic non-spatial models in spatially-explicit and flexible models. To fully explore the method's capacities and prevailing criminal conditions in Houston, future research must deal with a more comprehensive dataset describing the neighborhood conditions more broadly. Moreover, future models must account for the population specific effects within the census tracts via an additional offset term. Thus, crime analysis remains a vital on-going research area.

Acknowledgements

Marco Helbich and Jamal Jokar Arsanjani gratefully acknowledge the financial support from the Alexander von Humboldt Foundation. Additionally, this research has been funded by the by the German Research Foundation's Global Networks Initiative of Excellence of the University of Heidelberg.

References

BERK, R. (2008), Forecasting Methods in Crime and Justice. Annual Review of Law and Social Science, 4, 219-238.

BERNASCO, W. & ELFFERS, H. (2011), Statistical Analysis of Spatial Crime Data. In: PIQUERO, A. R. & WEISBURDM D. (Eds.), Handbook of Quantitative Criminology. Springer, Heidelberg, 699-724.

CAMERON, C. & TRIVEDI, P. (1998), Regression Analysis of Count Data, Cambridge University Press, Cambridge.

COXE, S., WEST, S. G. & AIKEN, L. S. (2009), The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. Journal of Personality Assessment, 91, 121-136.

GRIFFITH, D. (2000), A Linear Regression Solution to the Spatial Autocorrelation Problem. Journal of Geographical Systems 2, 141-156.

GRIFFITH, D. & HAINING, R. (2006), Beyond Mule Kicks: The Poisson Distribution in Geographical Analysis. Geographical Analysis, 38, 123-139.

HELBICH, M. & LEITNER, M. (2012), Evaluation of Spatial Cluster Detection Algorithms for Crime Locations. In: GAUL, GEYER-SCHULZ, W. A., SCHMIDT-THIEME, L. & KUNZE, J. (Eds.), Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, 193-201.

HELBICH, M., BRUNAUER, W., HAGENAUER, J. & LEITNER, M. (2013a), Data-Driven Regionalization of Housing Markets. Annals of the Association of American Geographers (online first, DOI: 10.1080/00045608.2012.707587).

HELBICH, M, HAGENAUER, J., LEITNER, M. & EDWARDS, R. (2013b), Exploration of Unstructured Narrative Crime Reports: An Unsupervised Neural Network and Point Pattern Analysis Approach. Cartography and Geographic Information Science (under revision).

LEITNER, M. (2013), Crime Modeling and Mapping Using Geospatial Technologies. Springer, Heidelberg.

LEITNER, M. & HELBICH, M. (2011), The Impact of Hurricanes on Crime: A Spatio-Temporal Analysis in the City of Houston, TX. Cartography and Geographic Information Science, 38, 214-222.

MORENOFF, J., SAMPSON, R. J. & RAUDENBUSH, S. (2001), Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence. Criminology, 39, 517-560.

OSGOOD, W. (2000), Poisson-Based Regression Analysis of Aggregate Crime Rates. Journal of Quantitative Criminology, 16, 21-43.

PATUELLI, R., GRIFFITH, D., TIEFELSDORF, M. & NIJKAMP, P. (2011), Spatial Filtering and Eigenvector Stability: Space-Time Models for German Unemployment Data. International Regional Science Review, 34, 253-280.

TIEFELSDORF, M. & GRIFFITH, D. (2007), Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach. Environment and Planning A, 39, 1193-1221.

TITA, G. & RADIL, S. (2011), Spatial Regression Models in Criminology: Modeling Social Processes in the Spatial Weights Matrix. In: PIQUERO, A. R. & WEISBURD, D. (Eds.), Handbook of Quantitative Criminology. Springer, Heidelberg, 101-121.

WANG, F. (2012), Why police and policing need GIS: an overview. Annals of GIS, 18, 159-171.