Master Thesis

submitted within the UNIGIS M.Sc. programme

Interfaculty Department of Geoinformatics - Z_GIS

University of Salzburg

„Prediction of Assaults with a Combination of Retrospective and Prospective Analyses on the City of Graz“

by

B.Sc. Sebastian Rath

103522, UNIGIS MSc Jahrgang 2014

A thesis submitted in partial fulfillment of the requirements of the degree of Master of Science (Geographical Information Science & Systems) – M.Sc. (GIS)

Advisor:

Dr. Michael Leitner, Louisiana State University

Nuremberg, 29.12.2016

Acknowledgments

I hereby want to give my thanks to everyone who supported me throughout the thesis and the master's study. I especially would like to thank the "Bundesministerium für Inneres" of Austria for providing me with the necessary crime data, and my family, friends and colleagues who gave me moral support.

Furthermore, I would like to thank the University of Salzburg UNIGIS team for supporting me through the master's program. I would also like to thank Dr. Michael Leitner for guiding me through the master thesis. Special thanks go to Mathis Dollinger and CeCelia Hubbartt for their counselling throughout the thesis and the master's program.

Declaration of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Nuremberg, 29/12/2016

Abstract

Police forces around the globe are under constant pressure to improve their work with fewer resources. Preventing crimes even before they occur is the latest approach to distribute the police resources more efficiently. This research focuses on the enhancement of current crime analyses with modern crime prediction methods. More specifically, this research highlights a crime prediction model that combines retrospective and prospective analyses in order to measure the prediction accuracy of assaults in the city of Graz, Austria.

The main objective of this research is to test the prediction rate of this model and compare it to the prediction of the common retrospective approach. For the input of retrospective analyses, assault incident data from the years 2009 to 2013 were provided from the "Sicherheitsmonitor" of Austria. The prospective analysis used Open Street Map data to estimate the risk of an occurring crime based on the proximity to certain locations. The accuracy of both prediction approaches was evaluated with the assault incidents of 2014.

The result of the Hotspot analysis predicted the locations of 26.7% of the assault incidents of 2014, whereas the final model reached 19.0%. The prediction model of this research could not enhance the prediction accuracy of the retrospective approach, but it did predict emerging hotspots which were not included in the Hotspot analysis. The lack of research on prospective crime prediction and the imperfection of the Open Street Map data revealed the obstacles which need to be overcome for a successful improvement of the retrospective crime prediction.

Keywords: crime prediction, GIS, retrospective analysis, prospective analysis, OSM, python

Table of Contents

*** RESTRICTION NOTE ***

(SPERRVERMERK)

List of Figures

List of Tables

List of Formulas

List of Acronyms & Abbreviations

BMI	Bundesministerium für Inneres
GIS	Geographical Information Systems
IACA	International Association of Crime Analysts
KDE	Kernel Density Estimation
NNH	Nearest Neighbour Hierarchical Clustering
OSM	Open Street Map
PAI	Prediction Accuracy Index
RTM	Risk Terrain Modelling
SIMO	Sicherheitsmonitor
STAC	Spatial and Temporal Analysis of Crime
WGS84	World Geodetic System 1984

1. Introduction

1.1. Motivation and Background

Police forces around the globe are under constant pressure to improve their work with fewer resources. Therefore, an efficient reallocation of resources is a crucial element of today's police work. To accomplish improvement goals, police forces invested in technical equipment and administrative infrastructure, specifically through the introduction of Geographic Information Systems (GIS). This innovative system had a powerful impact on police institutions. Using geographic data required the implementation of physical infrastructure. Therefore, police forces had to adapt to digital recording of incidents and their handling in databases. However, changing the technical component was not the only burden to overcome. Legal boundaries had to be discussed in order to establish efficient police work between jurisdictions and administrative guidance on a larger scale. After overcoming these obstacles, digital crime analyses were able to become established in the daily routine of police forces (Levine, 2006).

With GIS implemented, police forces were able to use digital analyses for patrol organization, automated recording of incidents and crime prediction. Most prediction analyses however, were retrospective, meaning crimes were analysed on the basis of incidents and their relating factors that had already occurred (Porteous, 2003). This commonly used approach lacks to identify crime hotspots where no incidents have previously occurred. which is a debilitating downside for identifying crime clusters. Prospective analyses use crime related information to estimate the crime risk of an area. This approach has "greater accuracy than either retrospective hotspot mapping or simple target repeats" (Mazerolle & Wortley, 2008), resulting in an increased ability to identify hotspot areas before they evolve.

Early efforts in predictive crime analyses were conducted by Liu and Brown in 2003. They expanded the existing crime clustering methods by the offenders' preferences in selecting crime sites. The idea was to predict locations which match the same type and spatial attributes of the victimized crime sites. The algorithm, based on a point pattern density model, highlights the areas which have an increased likelihood of victimization.

Bowers (2004) analysed the risk of victimization in a nearby area after a burglary. He used a "Search Efficiency Rate," which is based on the density of incidents per area.

It was concluded that the properties within 400 meters of the burgled location are at a significantly higher risk to be victimized in the following two months. Both studies, however, are not widely accepted due to the subjectivity of the hotspot methods.

The Predictive Accuracy Index (PAI) was the first objective metric to evaluate the effectiveness of different hotspotting methods in different study areas. Chainey and Ratcliffe (2008) introduced the PAI to compare the hit rate to the area size of the predicted hotspots in relation to the whole study area. The calculation is easy to use and independent of the size of the study area and approach used for the analysis.

Although prospective analyses have been found to aid crime predictions, they are not widely used in today's police work. At the 2016 International Association of Crime Analysts (IACA), experts in the field began presenting their research on the comparison of retrospective and prospective analyses. Since the promotion of this research is still new for many experts, there is currently no widely used study on the influence of prospective analyses on the accuracy of retrospective crime prediction.

The new research examines police records and free accessible data related to assaults to create a model combining retrospective and prospective analyses, while statistically testing the model parameters for an optimal crime prediction.

1.2. Research Statement & Objectives

The purpose of this thesis is to assess the question: Can the combination of retrospective and prospective analyses improve the prediction accuracy of a solely retrospective approach? To best determine the answer to this question, the following queries must be examined:

- Is the improvement of the crime prediction statistically significant?
- Is a prospective crime prediction as accurate as a retrospective prediction?

1.3. Expected Results

Current research suggests prospective crime prediction is useful to overall retrospective analyses. But due to the lack of research on prospective analyses and the questionable accuracy and actuality of the OSM data, retrospective crime prediction should prove to be more accurate than prospective analysis alone. The combination of prospective and retrospective analyses is expected to minimize the disadvantages of both approaches and combine their benefits. Therefore, the prediction model of this analysis will highlight the existing hotspots and identify emerging hotspots as well.

1.4. Target Groups

This project pioneers the field of prospective crime prediction. The methods used and results achieved will prove useful to police institutions and researchers who will implement or further develop their crime prediction expertise in police work and the academic field.

1.5. Thesis Structure

The research is divided into six chapters. Chapter 1 introduces the thesis, which includes the thesis abstract. Chapter 2 focuses on the literature review and theories behind retrospective and prospective analyses. The methods, the selection of the parameters for the analyses and the structure of the Python model are outlined in Chapter 3. The fourth chapter presents the results, as well as their evaluation and discussion. Limitations of the applied methods and the interpretation of the results are discussed in Chapter 5. Finally, Chapter 6 concludes the research findings and provides further recommendations.

2. Literature Review

The connection between crime and geography has been a focus of study for nearly two centuries. In the early 19th century Guerry (1833) identified a link between higher property crime rates and affluent locations. By analysing data on crimes, suicides, literacy and on other demographic statistics, he conducted the first comprehensive study relating crime to social demographics. Almost simultaneously, Quetelet (1831, 1835) led the foundation of criminology and sociology in Europe with his study of "moral statistics". Their works are considered to be the foundation of modern social science (Friendly, 2007).

The studies of Quetelet and Guerry in addition to those conducted by the British government afterward were conducted on a large administrative scale due to the lack of crime data on smaller geographical levels (Piquero & Weisburd, 2010). To combat this issue, Shaw and McKay (1942) manually mapped nearly 6000 juvenile delinquency cases in an effort to find patterns between delinquency rates and community characteristics. Their objective was to determine the economic and social characteristics of local areas which led to a variation in the male juvenile delinquency rate. They concluded that without changes in a community the configuration of delinquency "changes very slowly, if at all" (Shaw & McKay, 1942:222).

The pioneering works before the 1970s were hampered by the inability to collect, store and process geographic and crime information (Weisburd, 1998). With a combination of technological developments, the ability for digital processing in institutions and by overcoming organizational hurdles, crime analyses has grown more appealing to police agencies (Weisburd & Lum, 2005).

The Illinois Criminal Justice Information Authority developed one of the first software in 1989, Spatial and Temporal Analysis of Crime (STAC), to analyse crime. The police department of Chicago, Illinois, was the first police agency to request this program to help identify Hotspot areas. STAC requires a user-defined search radius, as well as a minimum cluster and cell sizes. The application of these defined parameters creates a grid over the study area and places a circle with the user-defined radius over each grid cell intersection. The number of aggregated incidents in each circle determines their rank. Since no point should belong to multiple clusters, the software combines the incidents until there are no overlapping circles. This method results in a standard deviational ellipse or its convex hull (Levine, 2015).

STAC is one of four approaches for analysing patterns in point data. The "Spatial Fuzzy Mode" is the simplest type of a hotspot analysis. Its algorithm requires a user-defined search radius to cluster the crime events that fall within each radius. The Fuzzy Mode calculates a radius where most incidents fall into and outputs a table with the ranks, total number of included points, and the X and Y coordinates of the centre (Levine, 2015). A third option for the hotspot analysis is the "Nearest Neighbour Hierarchical Clustering" (NNH). The NNH, based on the hierarchical clustering of King (1967), requires the user to define a minimum cluster size and a threshold value. Points that are located closer than expected under spatial randomness are grouped together if the minimum cluster size is reached. The smallest clusters are defined as first-order clusters. If the new clusters are spatially closer to each other than the defined threshold value, these clusters are merged again. Like a pyramid scheme, this process is repeated until the lowest order clusters are achieved (Eck, Chainey, Cameron, Leitner & Wilson, 2005).The final point-pattern hotspot analysis is the "Kernel Density Estimation" (KDE). This method contains a weighted raster surface across the study areas, based on the proximity of the point locations (Chainey and Ratcliffe, 2005). In the first step, a grid is created with a user defined gird size that covers all events. Afterward, a three-dimensional Kernel function calculates the distance to each surrounding point with the user-defined search radius. The final cell value results from the sum of the distances to each point. This occurrence is due to the Kernel function, which decreases with growing distances. The result of the KDE is a weighted grid.

The detection of crime clusters and emerging hotspots is, as every kind of analysis, sensitive to its use. Having different options to analyse crimes requires tools to identify the best option for the study. Especially by comparing different types of analyses, an objective measure of the accuracy of each analysis is needed. For example, calculating the crime density of resident burglary with the total population may not be as appropriate as using the number of occupied households (Harries, 1991).

Bowers (2004) was one of the first researchers to find a more suitable measure than just a hit rate. The Search Efficiency Rate calculates the incident number per square kilometre of the prediction area. The application of this index is restricted by the study area.

If the Search Efficiency Rate is applied to areas with different sizes, the results will become biased (Bowers, 2004). A measure that takes the size of the study area into account is the Predication Accuracy Index (PAI). The PAI calculates the accuracy by comparing the hit rate to the percentage of the prediction area in relation to the total study area. This accuracy index is quite easy to perform and universally applicable (Chainey, Tompson & Uhlig, 2008).

With modern GIS and today's wealth of tools and models, a Hotspot analysis is quite easy to perform, making it the most commonly used form of crime analysis in today's police work. While the analyses have made a strong impact on police work, researchers and officials still deal with the problem they faced 30 years ago: Aggregating spatial data lead to the so called "Modifiable Areal Unit Problem" (MAUP).

"When thematically mapped, different boundaries may generate different visual representations of where the hotspots may exist" (Chainey and Ratcliffe, 2005: 152).As long as the focus of the analyst is on a certain area, MAUP is not a problem. However, if a study area is subdivided, the aggregation of incidents within each of the boundaries will result in a distortion of the result (Piquero & Weisburd, 2010). To overcome this issue, Chainey and his colleagues (2008) changed the spatial level. Hotspot analyses conducted with STAC resulted in a spatial ellipse that was fitted through the areas with the highest crime density (Craglia, Haining & Wiles, 2000). For an improved calculation of crime density, Chainey used a continuous surface to identify the hotspots, resulting in the Kernel Density Estimation (KDE) (Chainey et al., 2008).

The enhancement of the accuracy of hotspot analyses is quite challenging after conquering the MAUP and digital information quality. However, researchers are still putting big effort in fine-tuning the methodology for the detection of hotspots. Without changing the scale of the analysis, the improvement of the results will become stagnate.

In the last few decades, crime prediction measurements are more commonly accepted by police institutions. The analyses and methods stated above are just some examples of crime predictions. But they all are retrospective and therefore focused on the incidents which have already occurred in the study area.

Like the connection between a place and an incident, there is a second important aspect for understanding crime: time. Almost 50 years ago, Hagerstand (1970) pointed out that crimes require the opportunity to be committed. Daily routines and behaviours of the offender and victim therefore limit these opportunities and are changing over time. This understanding led to the “Temporal Constraint Theory”. By linking the opportunity to commit a crime with the temporal restriction of the daily life, preventive measures can be determined (Ratcliffe, 2006). This kind of crime prediction is called prospective analysis. Analysing current data to determine areas at risk becomes increasingly relevant with near real time recording of data and universal access. "The true promise of crime mapping lies in its ability to identify early warning signs across time and space" (Shiode & Shiode, 2014: 438).

Knox and Bartlett (1964) had one of the first prospective elements in their analysis of the time and space restrictions of repeat victimization. The so called "Knox-type method" indicates that if an incident occurs at one location, the nearby area has an increased risk of being victimized again in the near future (Johnson, Lab & Bowers, 2008). This theory, now called "Near Repeat Victimization", is very promising for improving crime predictions (Johnson et al., 2007), but it has limited use. The chance for a near repeat is highly restricted by its spatial and temporal proximity to the original incident. The spatial extent for near repeat victimizations of domestic robbery in Merseyside, England reached 300 to 400 meters, since the parameters of a suitable target are most similar in the same environment (Johnson and Bowers, 2004). These parameters are subject to change over time. Haberman and Ratcliffe (2012) determined the time span for near repeats of street robberies in Philadelphia up to seven days. Therefore, "Near Repeat" analyses are beneficial for short term predictions, but cannot be applied to a longer time span.

A prospective approach for long term predictions is the Risk Terrain Modelling (RTM). RTM is an approach to assess the risk of an area by combining separate map layers which represent the spatial influence and intensity of a crime risk factor. The idea is to produce grids which are weighted according to the influence of a crime risk factor. Overlaying the grids results in a so called "risk terrain map", which allows crime prone areas to be identified. This approach offers a "statistically valid way to articulate and communicate crime-prone areas at the micro-level according to the spatial influence of criminogenic features" (Caplan and Kennedy, 2011: 7).

The first study site where Caplan and Kennedy (2011) used RTM was Irvington, New Jersey. The township of Irvington experienced an increase in crime rates during the start of the twenty-first century. With a murder rate of 38.7 per 100,000 inhabitants, Irvington was well above the national average of 4.9 for cities of similar size. The increase in violent crimes was partly due to a vibrant drug market and high gang activity. To conquer this trend, state police executives were looking to implement an accurate crime prediction. Caplan and Kennedy based their analysis on the prominent risk factors, such as known gang residences, drug arrest locations, infrastructure and shooting sites. The Risk Terrain Modelling was applied to each crime factor and their results aggregated (Figure 1).

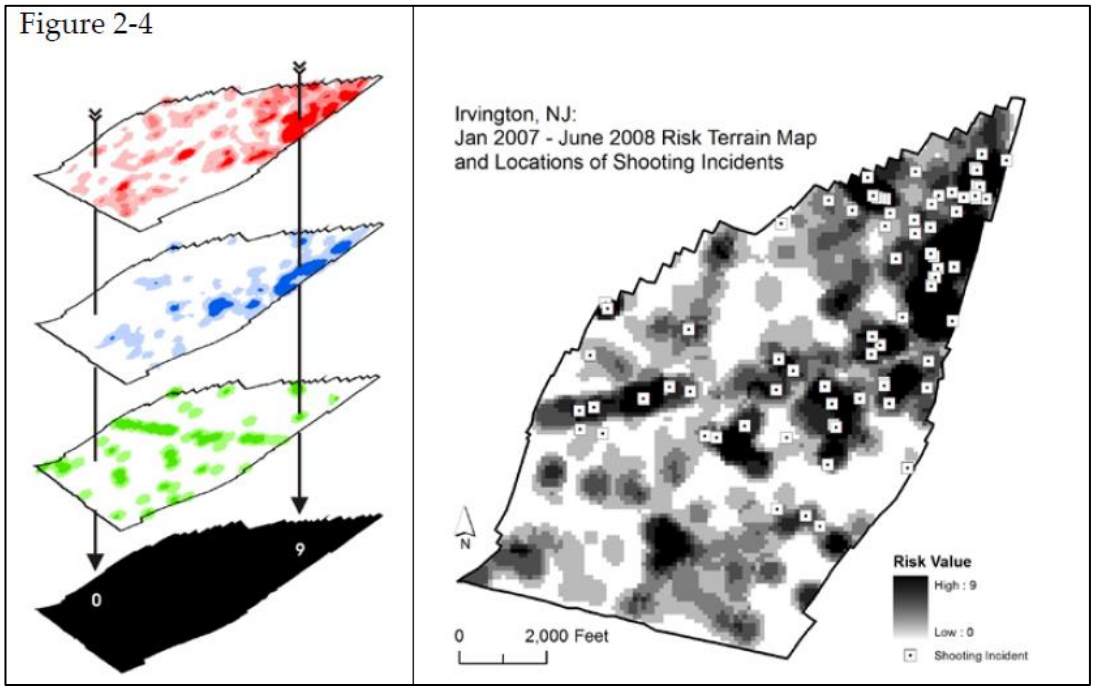

Figure 1: Risk Terrain Map of Shooting Incidents in Irvington, NJ; Source: Caplan & Kennedy 2011

The parameters used for the RTM were a bandwidth of 1,000 feet and a cell size of 100 feet. The bandwidth was chosen in relation to the average block face of 350 feet. The 100x100 foot cell size was the smallest level of detail the computers were able to process in a reasonable matter of time. They added in their conclusion that a smaller cell size "would provide the most utility for operational policing compared to larger, less specific, units of analysis" (Caplan & Kennedy, 2011: 19). Caplan and Kennedy also concluded that the prospective analysis of criminogenic factors was an effective tool to predict the shootings.

The findings of Caplan and Kennedy and those of the researchers mentioned before, show the potential of retrospective and prospective analyses for predicting crime. Since the adoption of technological development and eradication of legal restrictions, fighting crime was elevated to a new level. But the possibility for further improvements is rising with every new method. Therefore, this research attempts to reveal a new scope of predicting crime by combining the benefits and challenging the weaknesses of retrospective and prospective analyses when implemented on their own.

3. Methods

The following chapter presents the parameters and methodology of the research. It starts with the introduction of the study area and preparation of the input data. The third part gives an overview about the method of solution. Section 3.4 focuses on the examination of the assaults analysed in this research. Next, the retrospective and prospective analyses are assessed and the model parameters are chosen. In section 3.6, a short introduction is given to the programming language Python, which is used for the prediction model. The last chapter sums up the methodology of the prediction model.

3.1. Defining the Study Area

This research is a project in the scope of the UNIGIS Master on the University of Salzburg. The area studied is the city of Graz, Austria. Graz is about 900 years old and divided into 17 districts (Figure 2). The total city area covers 127.57 km² and had 282.479 inhabitants as of January 1, 2016 (Referat für Statistik, 2016).

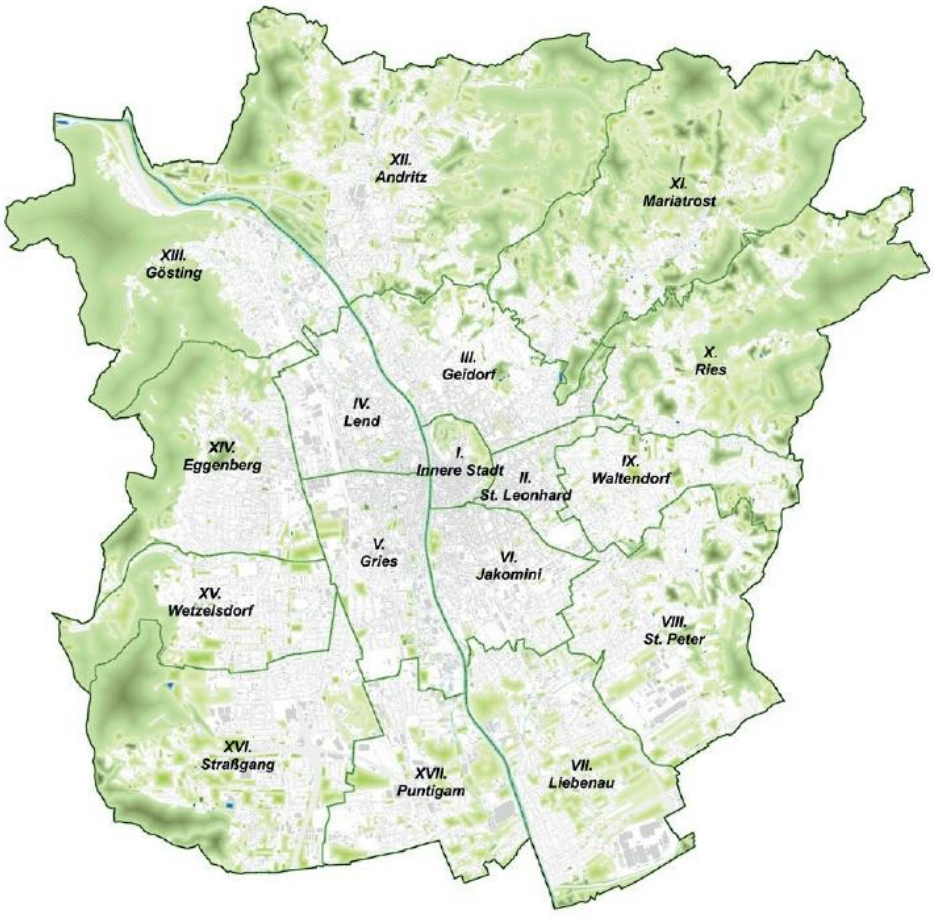

Figure 2: Districts of Graz, Source: Referat für Statistik, 2016

3.2. Preparing the Input Data

The data used in this research were provided by the “Bundesministerium für Inneres” (BMI) and by Open Street Map. In the following two sections, an overview is given how these data were prepared for their use in the retrospective and prospective analyses.

3.2.1. Retrospective Analysis

The incident data used are provided by the BMI of Austria. To receive the data, an application must be sent and approved by the BMI. Because this process takes time, the data had to be requested before the research focus was fully determined. Therefore the input from the "Bundesministerium für Inneres" also includes crimes of aggravated assault, burglary, theft and murder.

The input data were provided in an Excel format. The file included the assault, aggravated assault, burglary, theft and murder incidents from January 2009 to December 2014. Also included were

- An unique identification key
- The district, postcode and city of the incidents
- The coordinates of the crimes as referenced by the World Geodetic System 1984 (WGS84)
- The time of the crime
- The classification and the place of the crime
- The goods involved
- A binary classification if the crime is solved and if it was committed or just attempted; and
- The date and time when the incident was recorded in the Sicherheitsmonitor (SIMO) database.

There were several unusable entries in the database that had to be removed from the dataset during preparation for the research. In some entries there were no coordinates given or, due to a spelling error, were unusable. After cleaning up the input data, it was prepared for ArcGIS. This study focuses on assault, which is defined in §83 StGB (Austria). The assaults were identified by code classification.

The data were geolocated by their WGS84 X- and Y- coordinates. Some incidents' coordinates were lying outside the place of study. These incidents cannot be used for the analysis and had to be identified. The area of Graz was determined using digital boundaries for the municipality of Graz, provided by from Michael Bauer Research GmbH. All incidents outside the municipality were removed from the dataset within ArcGIS. The ArcGIS version used for the research was 10.3.1.

Although the data were already geographically referenced by the WGS84 coordinates, they were not projected. This step is actually not necessary for the analyses, but for easier handling of the determination of the parameters, the incident data were projected to Universal Transverse Mercator. The advantage of projection is to have the linear unit of one meter. Therefore the testing parameter could be chosen intuitively.

3.2.2. Prospective Analysis

The prospective analysis was based on data from Open Street Map. The OSM data were assessed via the Open Street Map website. The bulk download of the OSM data includes different shapefiles with different geometries. Based on that theme, there were point, line and polygon shapefiles. The polygon data of Graz were selected by using the same digital boundaries of Graz as described before.

Creating the risk terrain map requires several weighted grids. To create these grids, the OSM data were utilized. In the "Risk Terrain Modelling Compendium", Caplan and his colleagues at Rutgers University identified several risk factors by identifying their correlation to different crime types. The risk factors for assault as defined in §83 StGB (Austria) were:

- Restaurant, Bars, and Nightclubs;
- School Buildings and School Property; and
- Entertainment Venues.

(Caplan & Kennedy, 2011)

The polygon data that met these risk factors for assaults were selected and exported from the OSM dataset. The problem of the freely accessible OSM data is that there is no guarantee of the integrity and accuracy. In the process of conducting the RTM analysis, it became evident that the risk weights were distorted. As shown in Figure 3, all of the buildings in the "buildings" shapefile of the Open Street Map data were recorded separately. Therefore, even if the buildings belong thematically together, like universities and schools, there was a separate polygon for each of them.

Figure 3: Distortion of the RTM result by OSM duplicates, Source: Own representation based on BMI and Google Earth

In order to prevent this distortion, the OSM data had to be visually examined and thematically merged. In order to do so, the attributes of the polygons were searched for repeatable use of the same category and name, as well as visual analysis of buildings with a high spatial proximity. The merging process was conducted in ArcGIS.

3.3. Presenting the Method of Solution

This chapter provides an overview of the methodology of the research (Figure 4). The assault incidents of the Sicherheitsmonitor of Graz were used as the input for the Hotspot analysis. As described in chapter 3.2.1, the data were cleaned in Excel and projected in ArcGIS. The preparation of the Open Street Map data was conducted in ArcGIS. Both resulting point shape files were imported into the research's geodatabase. This preparation had to be performed manually and could not be automated. The researcher must visually evaluate the accuracy, integrity and validity of the input data.

The Python model was utilized after the preparation of the data. The Hotspot Analysis section started with dividing the SIMO dataset for each crime year. The next steps were performed for every crime year via a loop. The locations were placed on the same geographical coordinates and aggregated to show the real influence of the location. The Kernel Density Estimation was executed with the weighted point data set. The grid values of the resulting five risk layers were aggregated and weighted. To secure the comparability of the Hotspot analysis and the Risk Terrain Modelling results, the newly created grid was normalized with a fuzzy algorithm. The method used for the fuzzyfication is "MSLarge". This method weights higher values stronger than lower values. Therefore, an area with high values is given a higher value in the fuzzyfication than a group of low values. This is the most appropriate option for identifying hotspots.

The RTM started with the buffering of OSM locations with three different buffer radii. The value of each resulting buffer was calculated in relation to its proximity to the incident location. The closest buffer had the highest weight; the farthest buffer had the lowest value. These weighted buffers were intersected with a newly created grid. The grid was designed to have had the same cell size as the KDE rasters. Using the minimum boundary geometry of all OSM incidents, the area size of the grid was calculated. The weights of each grid cell were aggregated for each intersected cell. The same fuzzy algorithm as in the Hotspot analysis was used to normalize the RTM risk map.

The Hotspot and RTM grids were overlayed and the value of each cell was aggregated. The result was the final risk terrain map of the crime prediction. To evaluate the accuracy of the prediction and to answer the research objectives, hotspot polygons were created out of the risk terrain map. The incidents of the projection year within these hotspots were counted and exported. To evaluate the accuracy of each separate approach, the same procedure was utilized to count the number of the future incidents. The resulting statistics are used to answer the research objectives and show the final projection accuracy of the crime prediction.

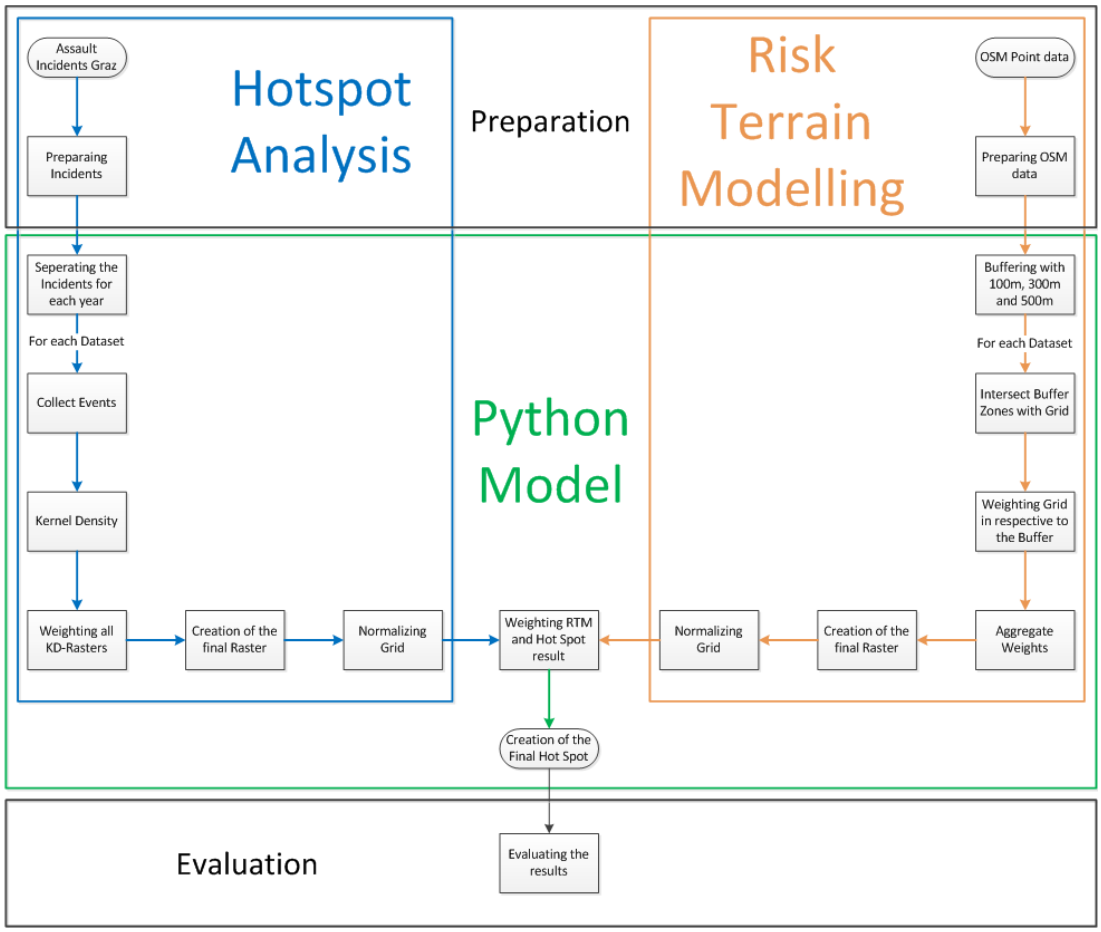

Figure 4: Workflow of the crime prediction model, Source: Own representation

3.4. Examining Assaults

Assaults defined in §83 StGB (Austria) describes the action of "a physical act that results in harmful or offensive contact with another's person without that person's consent" (Cornell University Law School). The German word "Körperverletzung" is not the exact translation of the word "assault". Legally, assault describes the intention to harm another person, whereas battery is defined as the actual act of harming (Cornell University Law School). However, the term "assault" is used in common language to define the act of harming a person as well. This thesis will use the word "assault" for the act of intentionally harming a person.

3.5. Assessing the Prediction Analyses

The importance of predicting crime is becoming increasingly prevalent in police work. Forecasting future crime events aid tactical and strategic planning to deploy police department's efforts, such as allocating patrols more efficiently and to detect and intervene in troubled areas. The following two chapters introduce the two crime prediction analyses used in this research.

3.5.1. Hotspot Analysis Parameters

Crime is not evenly distributed, therefore it is important to detect the crime prone areas to fight criminality (Eck et al., 2005). The most common crime prediction tool in today's police work is the Hotspot Analysis. Depending on the data available for the study area, there are different types of Hotspot analyses. The two commonly used hotspot analyses the point locations of the input directly assessed and the data aggregated into other geographic boundaries. The input data for this analysis was the locations of incidents recorded in the SIMO of Austria. Therefore a point-pattern hotspot analysis was chosen.

This research combines retrospective and prospective analyses to analyse the enhancement of the prediction rate compared to a solely retrospective approach. Critical to this task is the ability to have comparable results of each method in order to combine them. The ability of grids to perform calculation on the same geographical level makes them the best choice.

The resulting risk terrain of the KDE makes it the best hotspot analysis option for this research. The crucial parameter of the KDE is the search radius.

A series of tests were conducted to determine this parameter. The idea for the evaluation of the best search radius is to count the significant clustered incidents in each hotspot (Figure 5).

The input for the KDE contains the assault incidents from 2009 to 2013. The data were separated into a shape file for each year. In these files, each incident is located by their WGS84 coordinates, but one location can have multiple offenses. Therefore crimes were aggregated for each location.

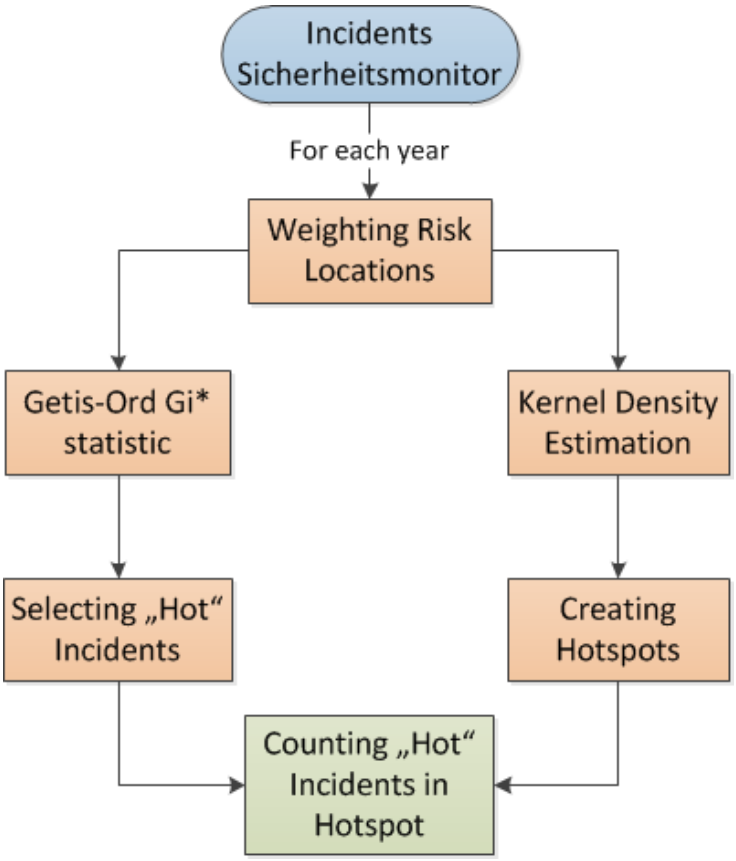

Figure 5: Evaluation workflow hotspot search radius, Source: Own representation

The resulting shape file was used for the Kernel Density Estimation. The KDE was performed with eleven different search radii, ranging from 100 meters to 350 meters by a 25 meter interval. The resulting weighted grid was classified by the "Natural Breaks (Jenks)" method. This slicing method results in a grid that is classified by natural groupings. The "Natural Breaks (Jenks)" method "specifies that the classes will be based on natural groupings inherent in the data" (ArcGIS Help 1). The grid values will be aggregated in classes. The classes are produced to contain the best group of similar values and to maximize the difference between the classes (Figure 6+7).

This step is necessary for the evaluation of the prediction model, because the final hotspot polygons are created by selecting the highest class of the reclassified values and converting it into a polygon shapefile.

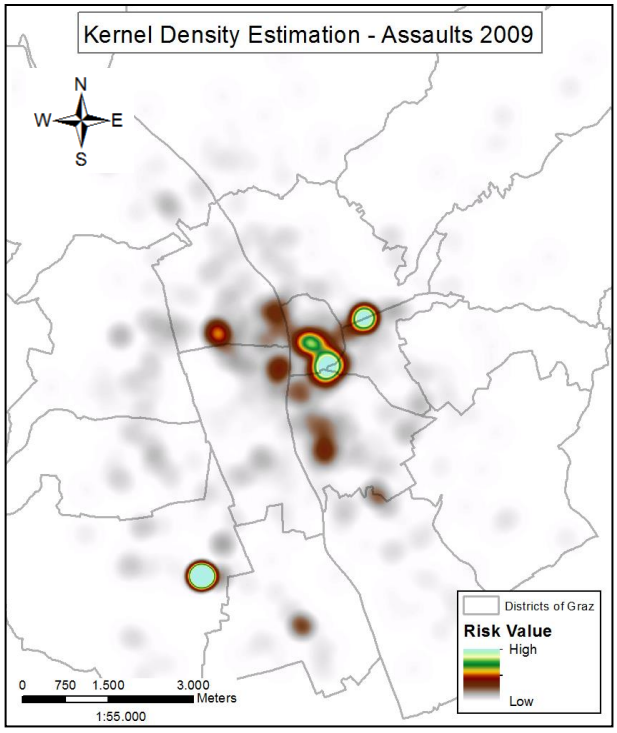

Figure 6: KDE of the Assaults of 2009 – Cyan represents high risk, white represents low risk, Source: Own representation based on BMI

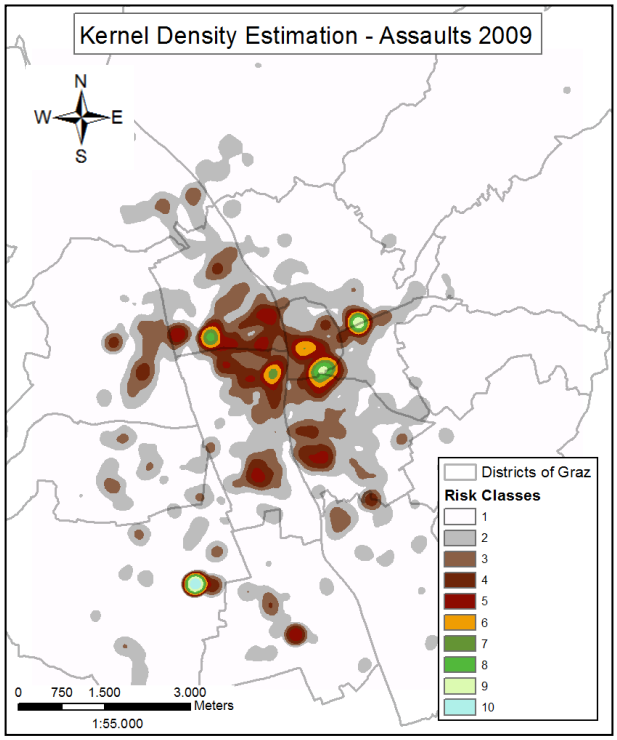

Figure 7: KDE of the Assaults of 2009 reclassified with Natural Breaks - 10 represents the areas with the highest risk, 1 represents the areas with the lowest risk, Source: Own representation based on BMI

The second part of the search radius evaluation is the statistical testing of the clustering significance. The "Getis-Ord Gi*" test identifies statistically significant spatial clusters. The clustering of high values is defined as a hot spot; the clustering of low values creates a cold spot.

As shown in Formula 1, the Gi* score is derived by comparing the local weight of a feature and its neighbours to all features. A feature with a high value, meaning a high crime count, is not seen as a hotspot if the neighbouring features do not have a higher value as well.

The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \quad (1)$$

where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X}\right)^2} \quad (3)$$

The G_i^* statistic is a z-score so no further calculations are required.

Formula 1: Getis-Ord Gi* staistical testing, Source: ArcGIS Help 2

The resulting z-score indicates the presence of a hot or cold spot. The p-value represents the level of statistical significance (ArcGIS Help 2). The incidents that had the highest z-scores while being highly statistical significant are shown in Figure 8.

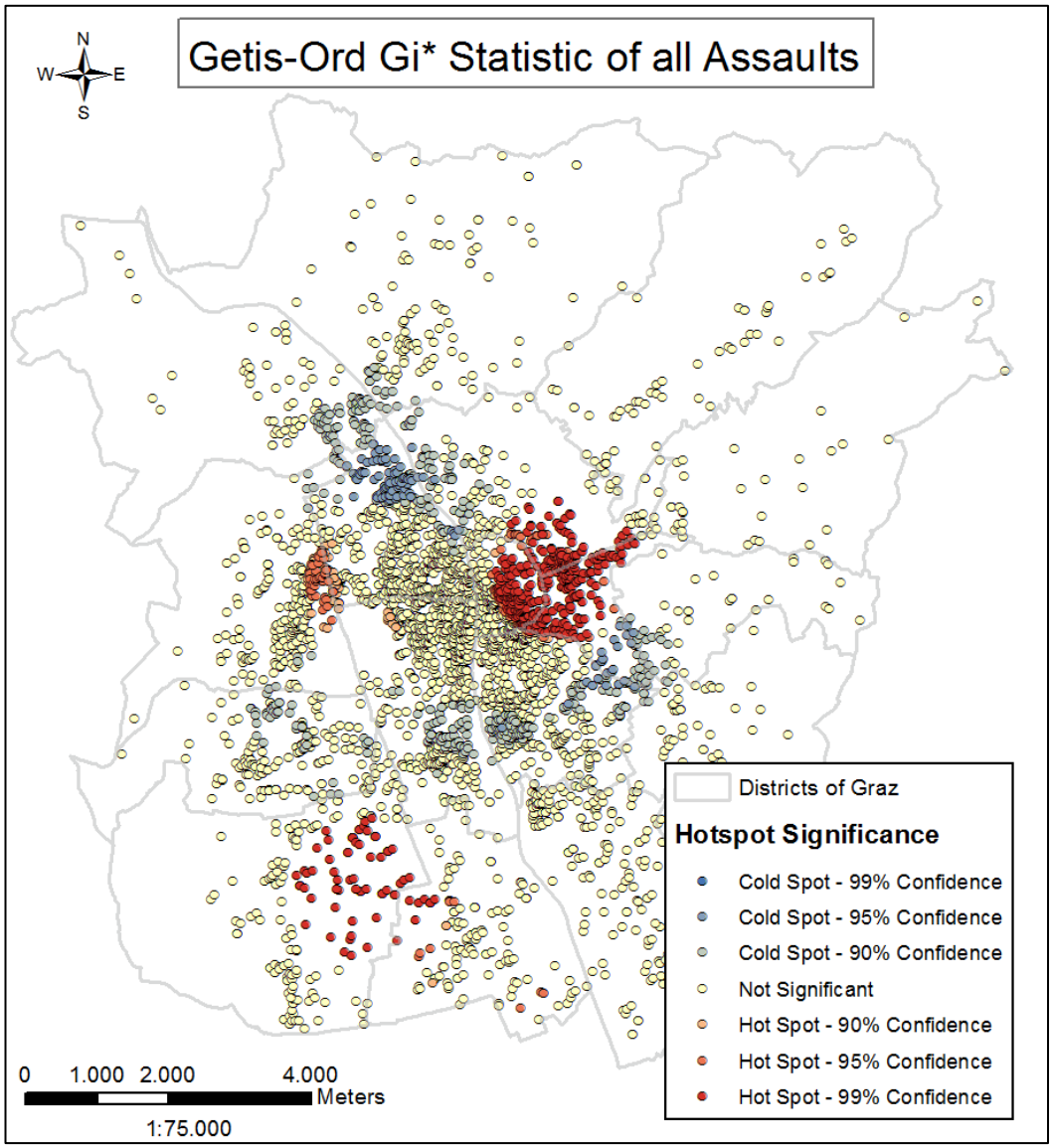

Figure 8: Getis-Ord Gi* - Statistic showing the clustering significance of all incidents, Source: Own representation based on BMI

Another tool for evaluating the optimal search radius for the Hotspot analysis is the Prediction Accuracy Index. In comparison to the hit rate, the PAI also considers the size of the hotspots. The PAI is defined as:

$$\frac{\left(\frac{n}{N}\right)*100}{\left(\frac{a}{A}\right)*100} = \frac{HitRate}{AreaPercentage} = \text{Prediction Accuracy Index} \quad (1)$$

Formula 2: Calculation of the PAI, Source: Chainey et al., 2008: 14

Where *n* is the number of future crimes within the hot spots from the retrospective data, *N* is the number of new crimes in the whole study area. *A* is the size of the whole study area and *a* is the sum of the size of the generated hot spot area. The PAI finds higher values if the area percentage is lower. Therefore the PAI is sensitive to the percentage of area covered by the hotspots (Chainey et al., 2008).

3.5.2. Risk Terrain Modelling Parameters

Retrospective analyses are restricted by their scope. Hotspot or Near Repeat analyses can only analyse patterns and clustering where crimes were already committed. To expand the focus of a crime prediction to the whole study area, another data base must be included. Perry and his colleagues (2013) introduced environmental and temporal features to expand their research. They analysed time routines of a day or week, the influence of seasons and the weather, type and proximity of locations, and demographic and economic data of the crime area.

This research assesses the influence of a prospective approach based on the type of buildings in Graz. A free and easily accessible way to get a hold of building data for the study area is Open Street Map. But, as described in chapter 3.2.2, the OSM data needed to be prepared for the analysis. After cleaning and thematically merging the OSM data, the input for the Risk Terrain Modelling was ready. The RTM makes "predictions about crime risk based on how close given locations are to theses risk-inducing features" (Perry, McInnis, Price, Smith & Hollywood, 2013: 50). The "risk-inducing features" are defined by Caplan and Kennedy (2011) in their "Risk Terrain Modelling Compendium". Based on their findings, the buildings were sorted by their OSM tags.

The crucial parameter of the RTM is the buffer distance, which determines the weight of the risk factor for the analysis. Around every risk factor or selected OSM location, a buffer is created. This buffer determines the influence of this risk factor. The influence of a risk to an area decays over space (Perry et al., 2013). The literature provides no distinct buffer distance for the influence of the risk factors on assault. Using one buffer is quite difficult, due to the lack of information regarding the spatial influence of the risk locations. The buffer distance was chosen in order to strike a balance between the number of locations included and a reasonable spatial influence.

The opposite of choosing one buffer distance would be a grid where cell values decline over the distance. In this case, the weight of the grid value would be different for every grid cell according to its spatial distance to the risk location. This could cause issues when two locations have the same distance to a third location. In this scenario, their influence would be different if the quadratic cell's boundary just does not include the location's coordinate.

Because of missing information in the literature and the technological problems of calculating the exact risk value of a location, different buffer zones will be created. By using this method, the closest grid values will have the highest influence on the incidents, but more distant areas still have an influence on the analysis.

3.5.3. Risk Terrain Map

The result of the prediction model is a Risk Terrain Map. It is created by combining the results of the Hotspot analysis and the RTM and aggregating the cell values. The final cell value represents the risk level of the specific area.

In order to evaluate the accuracy of the prediction model, the hotspot locations must be determined. Therefore, the continuous grid needed to be converted into polygon boundaries. Defining the cut-off value for the slicing is critical in this process. ArcGIS offers three option s for slicing grid data. Figure9 shows the slicing process using the method "Equal Intervals". This method divides the maximum range of input values by the number of output classes. For this research, ten classes are created to analyse the study site.

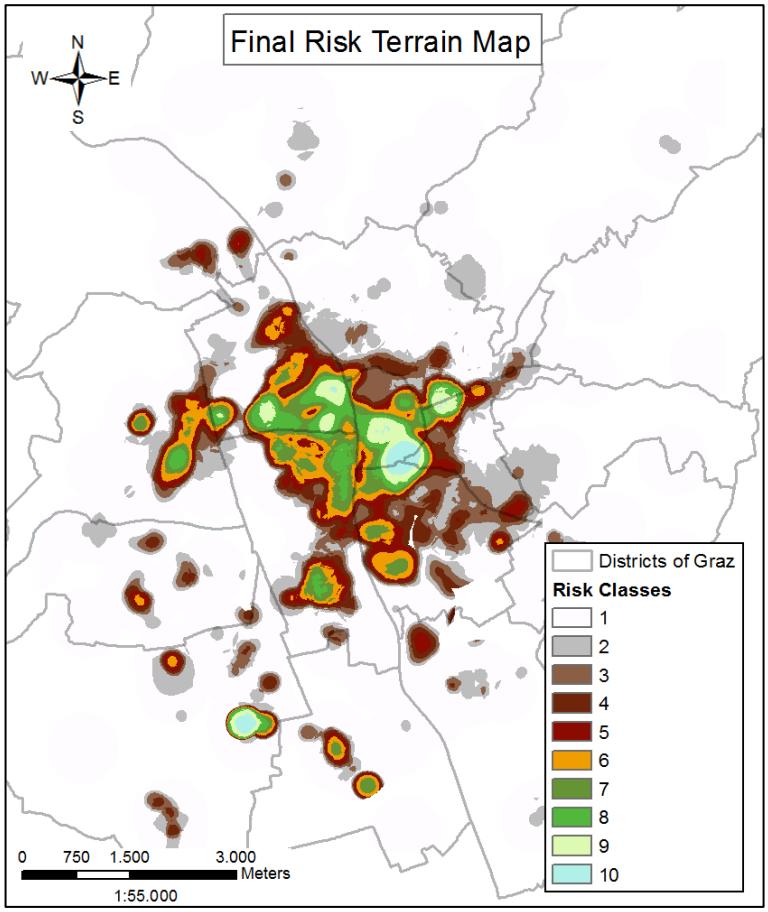

Figure 9: Final Risk Terrain Map sliced using the classification method "Equal Intervals", Source: Own representation based on BMI

Figure 10 shows the slicing method "Equal Areas". This reclassification creates output classes with a similar amount of cell values for each class. It becomes obvious that the result could not be divided into ten classes, as required.

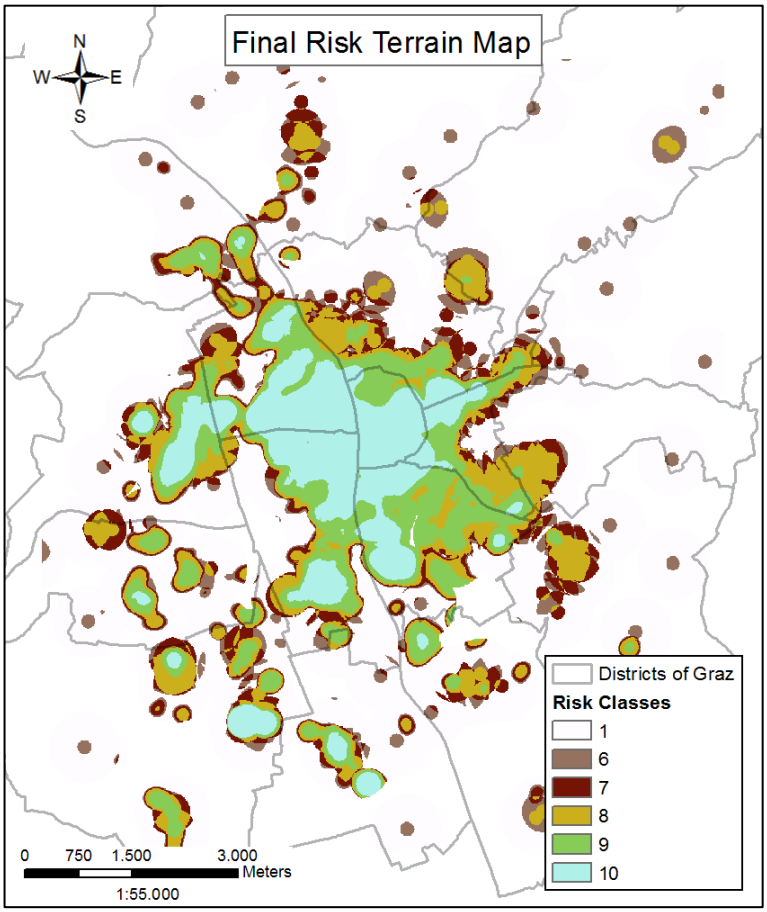

Figure 10: Risk Terrain Map slice using the classification method "Equal Areas", Source: Own representation based on BMI

The last slicing method available in ArcGIS is the "Natural Breaks (Jenks)" method (Figure 11). The algorithm searches for natural groups in the input data and creates the specified amount of classes by grouping similar values. Simultaneously, the method tries to maximize the difference between the classes.

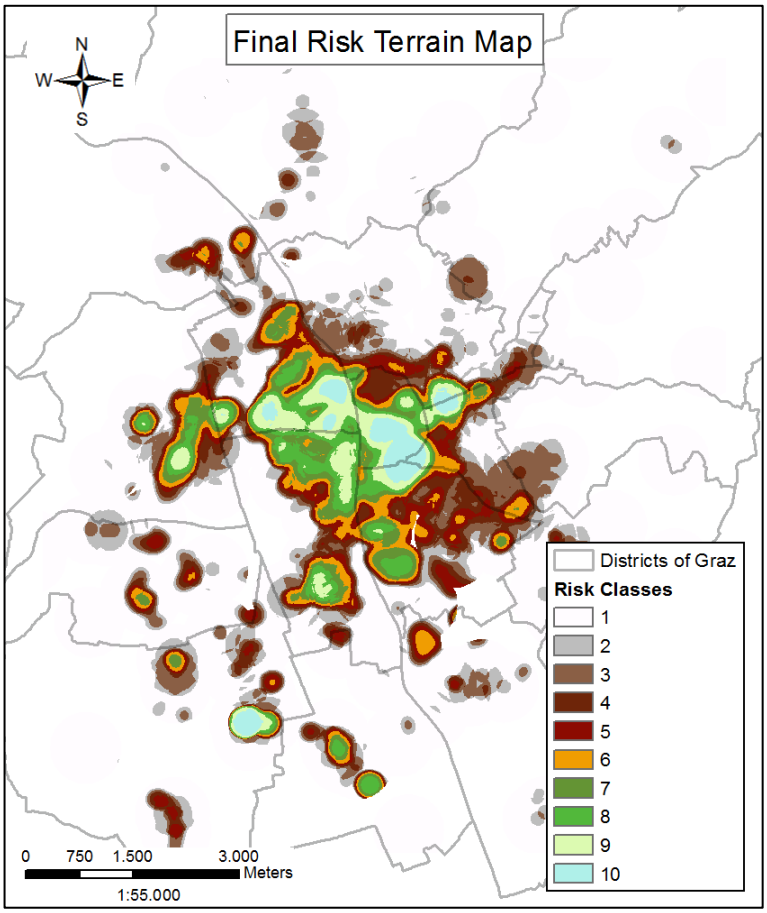

Figure 11: Risk Terrain Map sliced using the classification method "Natural Breaks (Jenks)", Source: Own representation based on BMI

As shown in the figures above, using different cut-off values massively influences the resulting hotspot areas. The aim of the prediction analysis is to identify the areas at the highest risk of being victimized. Therefore, the slicing method "Natural Breaks (Jenks)" serves as the best option for the categorization of an area as a hotspot.

3.6. Introducing Python

Python is a free and universally applicable programming language invented by Guido van Rossum in 1990s. The interpreted language contains a vast standard library that allows it to be widely applicable. Since the first full version of Python 1.0 in 1994, there have been several updates that have added new modules and resources with each update. Because of its expressive syntax and availability, it granted many advantages in comparison to other programming languages of this time. Where other languages like C or C++ are quite difficult to learn and must be compiled, the interactive interpreter of Python allows a real-time code development (Dubois, 2007).

The programming language Python was chosen for this research, because of its implementation in the ArcGIS software. The module "ArcPy" allows easy access and use of geoprocessing tools in ArcGIS. Pieces of a script can be easily tested by entering the information in the built-in Python window in ArcMap. Next to the vast community, which grants help to almost any problem with the use and structure of ArcGIS tools, the implementation of Python in ArcGIS has another major benefit, which influenced the choice of the programming language. It is possible to export a model created in ArcGIS "Model Builder" into Python. ArcGIS Model Builder allows the user to create a model by visually connecting tools and data. With Drag-and-Drop functionality, tools can be efficiently added to a model. This allows a relatively easy development of complex scripts in Python (Toms, 2015).

3.7. Conducting the Crime Prediction

The theoretical background and the methodological approach were defined in previous chapters. This chapter presents the final model with the selected parameters as it is implemented in Python.

The Python model starts with the definition of paths and variables, which are steadily used throughout the model. Examples for these global variables are the paths for the platform, the project, the geodatabase and its feature datasets. Setting the global environment options allows the user to define the spatial reference system and the option if the script should overwrite the output files with each run.

The preparation for the Hotspot analysis was initially described in chapter 3.2. After cleaning and projecting the incident data, a shapefile is created with a projected geographic coordinate system which has a linear unit of one meter. The resulting feature class is divided by the year of the crime. For each new shape file, the Kernel Density Estimation is conducted. The parameters used are as determined in chapter 3.4.1. In order to create the hotspots of all five years, the grid cells of each KDE are aggregated. Research suggests that the influence of past incidents on future events decays over time. Therefore, each KDE result is weighted according to the time difference to the prediction year. This means, that the year 2013 has the highest influence on the aggregate hotspot, and 2009 the lowest. After conducting this weighted aggregation of the KDE results, the Hotspot analysis is finished.

Performing the Risk Terrain Modelling required the preparation of the Open Street Map data. As described in chapter 3.2.2, the locations were selected, which are specified by Caplan and Kennedy (2011) to induce a risk correlated to assault. To avoid a distortion in the risk terrain map, locations that are thematically combined were merged in the OSM dataset. The resulting shape file was the basis for the RTM. In chapter 3.4.2, the cause was described for weighting the area around the risk locations. Buffer zones needed to be placed around each location to determine the influence of the nearby area. The OSM dataset used was a polygon file, meaning it had to be converted into a point layer. This shapefile is used to conduct the buffering. It is important for the buffer to have the ability to overlap to ensure the correct RTM result.

The option to dissolve the overlapping parts would lead to a distortion of the risk terrain map, since the areas that are lying in multiple buffer zones would not get the aggregated value of each buffer.

The weighting process itself is performed using a newly created grid with the same cell size as the Hotspot analysis. This grid is converted into a point layer using the centroid of each grid cell. This step monitors the buffer zones and determines if they were intersected with the original grid, since some cells would belong to multiple buffer zones. This is an important step in the process, since each cell should only belong to one buffer zone of each location. After intersecting the buffer zones and the point layer, the weights of each centroid is aggregated. After joining the centroid weights to the original grid, the Risk Terrain Modelling is completed.

The last step of the model is to combine the results of the Hotspot analysis and the RTM. Since both rasters have different ranges in their values, they must be normalized. On each grid, a fuzzy function is applied leading the grids to range from 0 to 1 with 0 representing the lowest values of the original grid and 1 representing the highest. The normalized grids were combined and their values summed up. The result is the risk terrain map of the analysis.

4. Results

Chapter 4 is presenting the results of the analyses and is organised into three sections. The first part presents the descriptive statistics of the study area and the assault data. The evaluation of the prediction parameters for both approaches is shown in section 4.2. Chapter 4.3 closes with the discussion of the final results and the comparison of the prediction accuracy.

4.1. Describing the Spatial Statistics

The number of reported crimes in Austria in 2015 has been 517,870, of which 55,491 (Bundeskriminalamt Wien, 2015) were in the Steiermark and 24,291 in the city of Graz in 2015 (Landesstatistik Steiermark, 2015). Of the total 24,291 crimes in Graz, 13,523 offenses were categorized either as assault, aggravated assault, murder, theft and burglary. 79.183 crimes of the specified crime types were committed in the city of Graz from 2009-2014. 12.7 % of these were located in the central district (Figure 12).

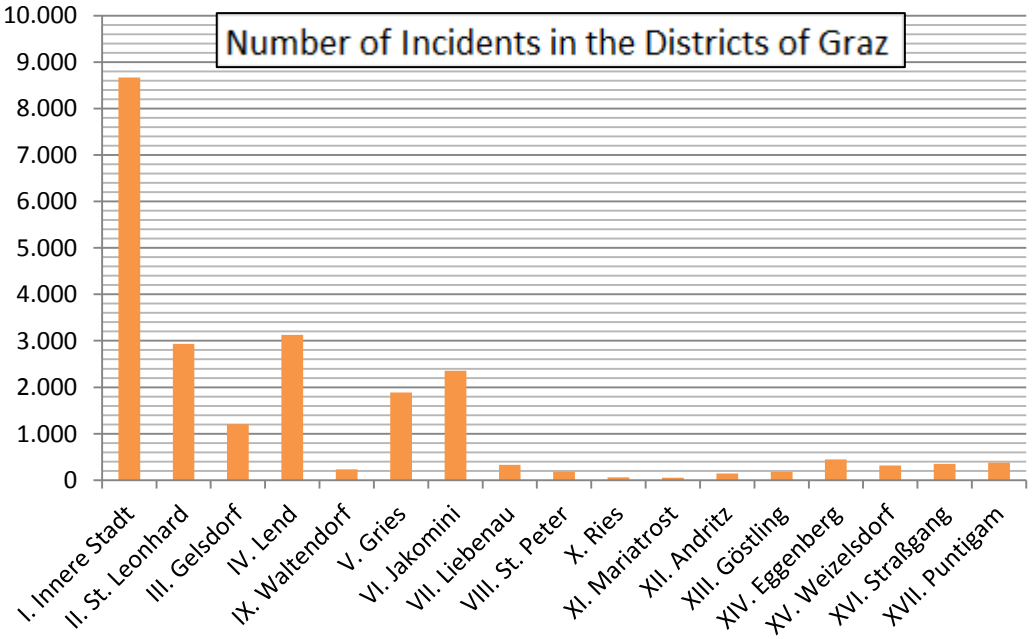

Figure 12: Number of Incidents for each District of Graz from 2009-2014, Source: Own representation based on BMI

The crimes committed in Graz are, not equally distributed across the districts. Figure 13 shows the spatial distribution of the incidents per area for each district in Graz. A darker colour implies a higher incident rate.

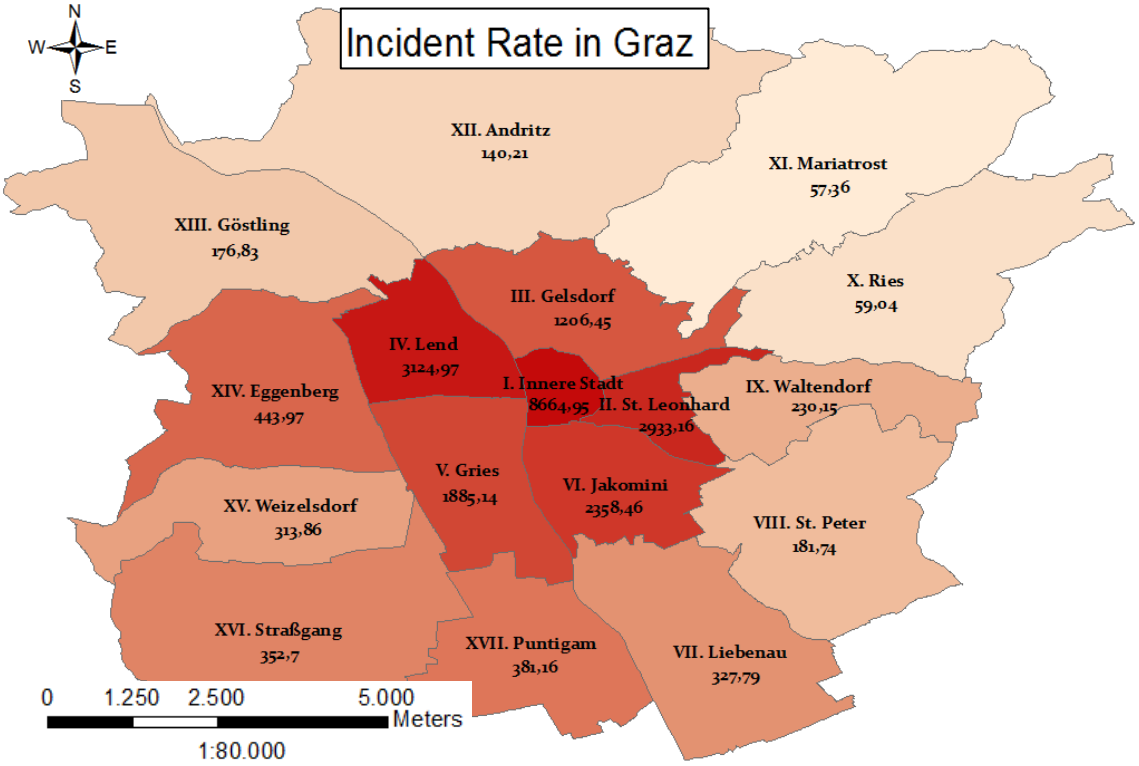

Figure 13: Incidents per Area for each District in Graz from 2009-2014, Source: Own representation based on BMI

Figure 14 and 15 show the distribution of all assaults reported in the time span of 2009-2014. The pattern is similar to the distribution of all crimes types within the districts. A vast majority of the assaults was committed in the inner city districts.

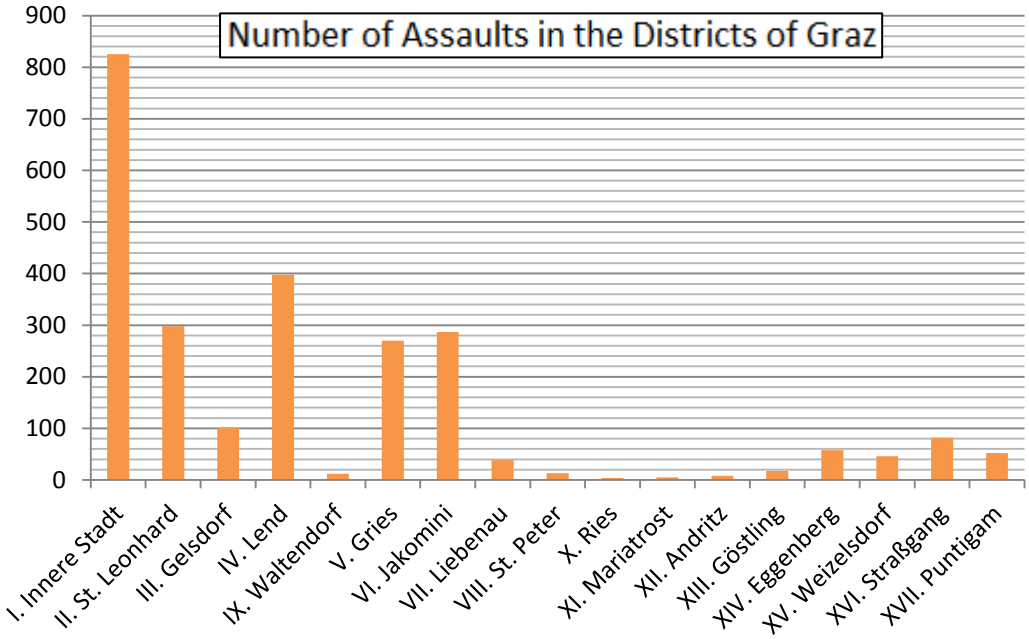

Figure 14: Number of Assaults for each District of Graz from 2009-2014, Source: Own representation based on BMI

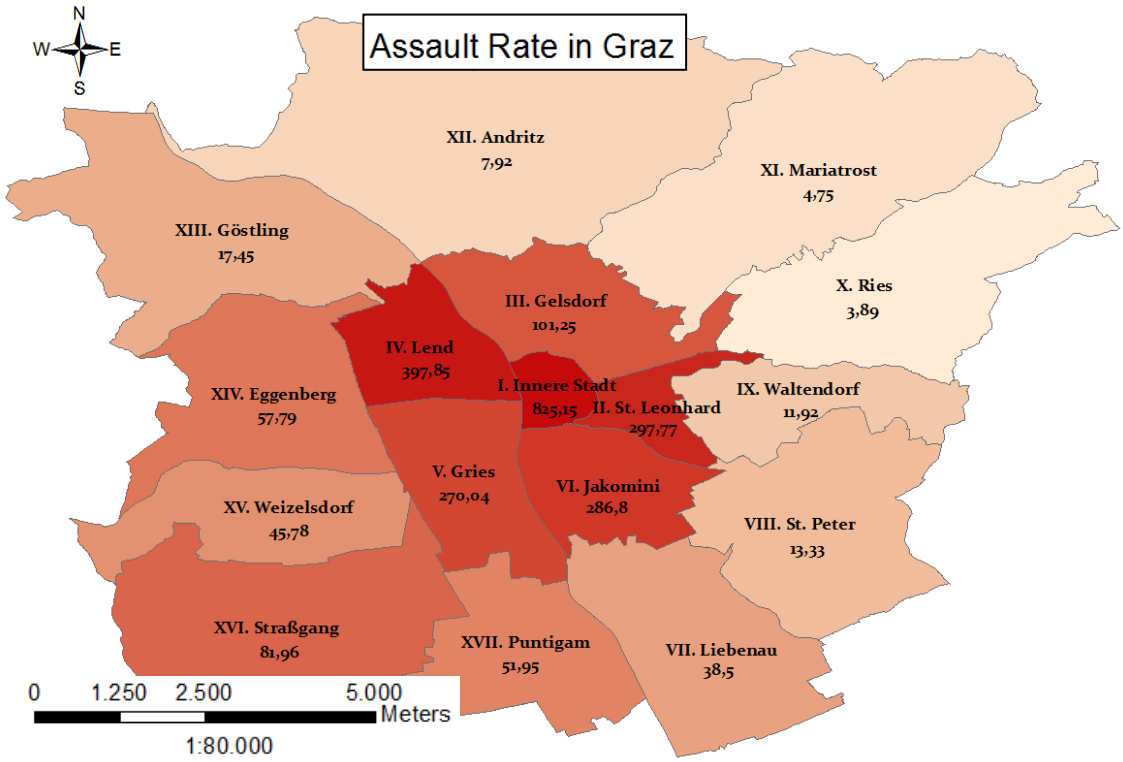

Figure 15: Assaults per Area for each District of Graz from 2009-2014, Source: Own representation based on BMI

The rate of incidents per area is sensitive to the total area of each district. But the amount of assaults can also be observed visually using a pinpoint map of the incidents (Figure 16 + 17).

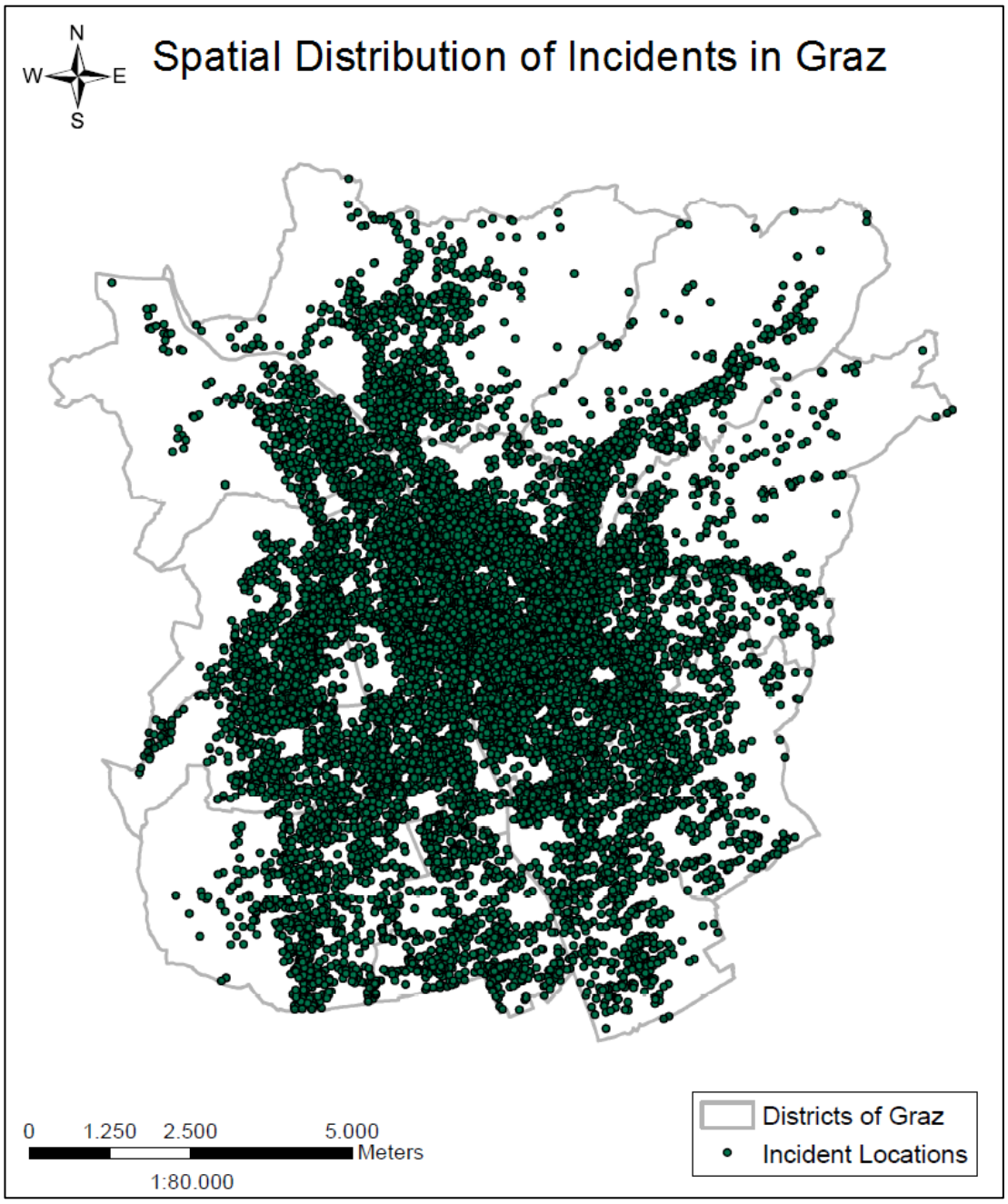

Figure 16: Spatial Distribution of all Incidents, Source: Own representation based on BMI

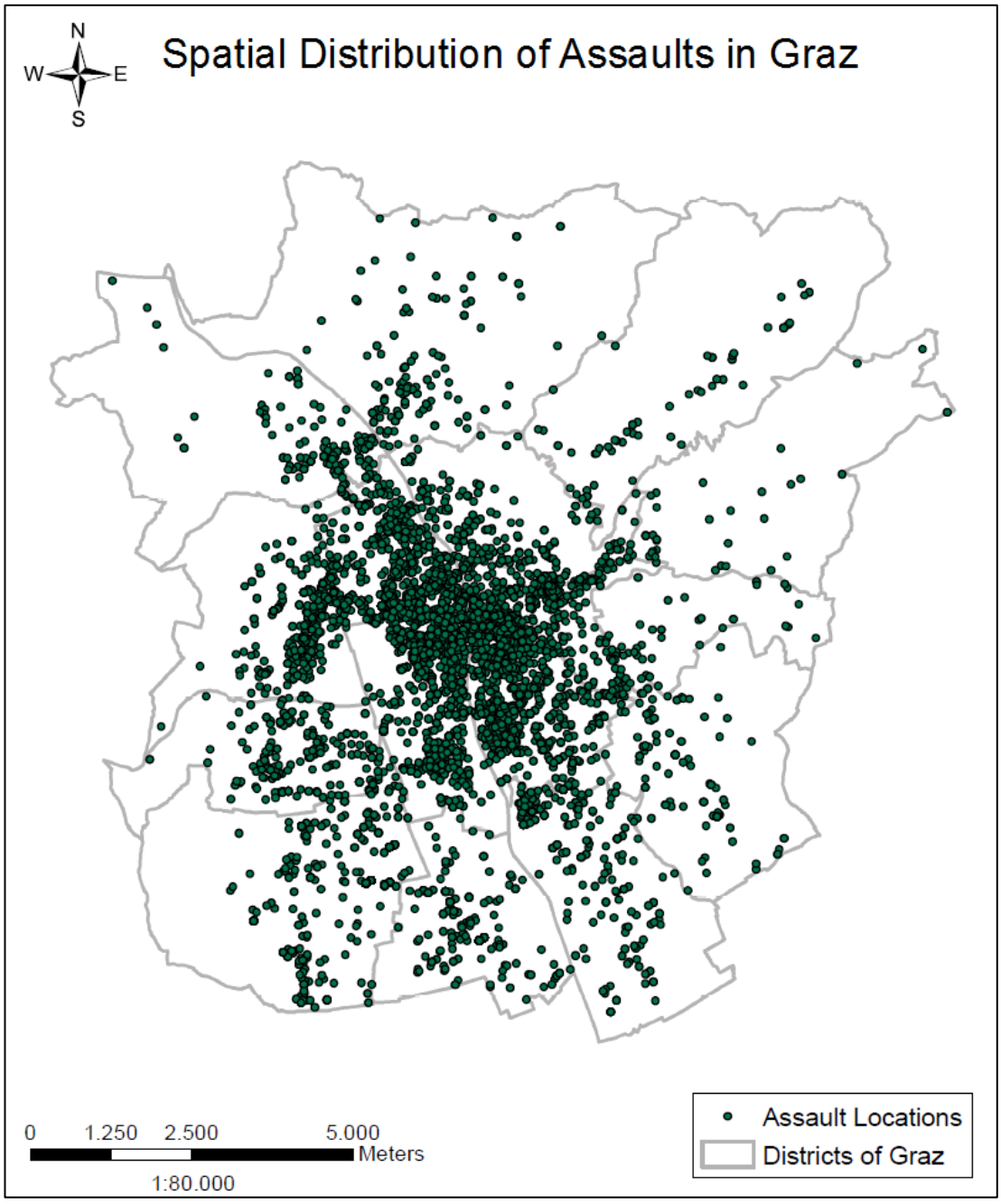

Figure 17: Spatial Distribution of all Assaults, Source: Own representation based on BMI

Figures 16 and 17 portray the spatial distribution of the incident locations, but not the individual weight of each location. By comparing the weighted locations with the simple locations, the prevalence of certain locations is highlighted (Figure 18).

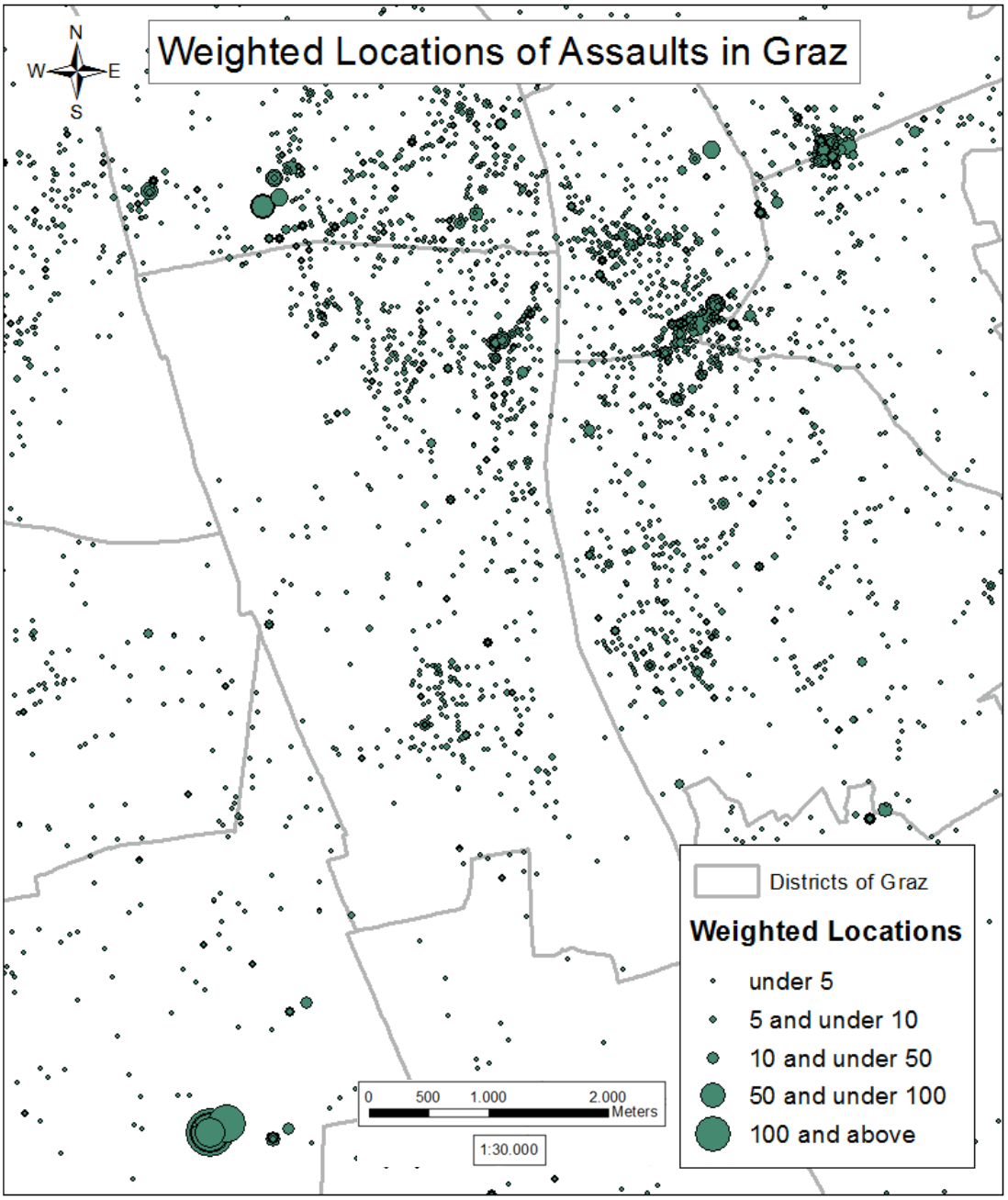

Figure 18: Central Extent of Weighted Assault Locations - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

To get an overview of the shifting and the prevalence of the hotspots over time, the weighted locations for each crime year of the analysis are portrayed (Figures 19 to 23).

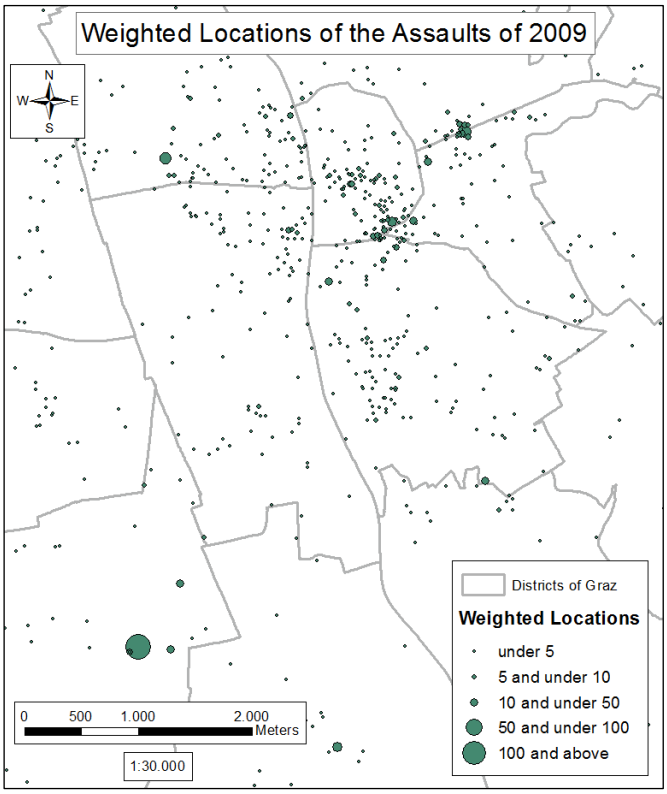

Figure 19: Central Extent of Weighted Assault Locations from 2009 - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

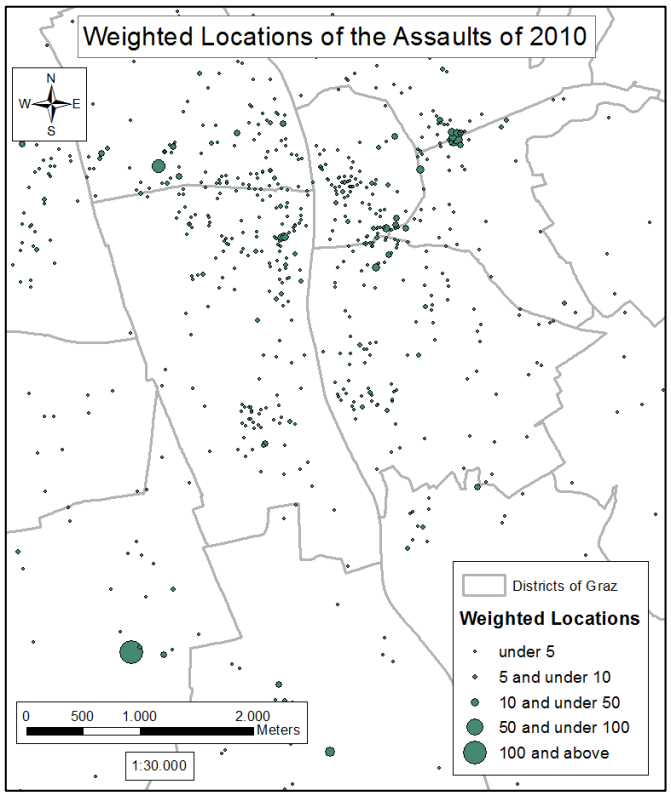

Figure 20: Central Extent of Weighted Assault Locations from 2010 - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

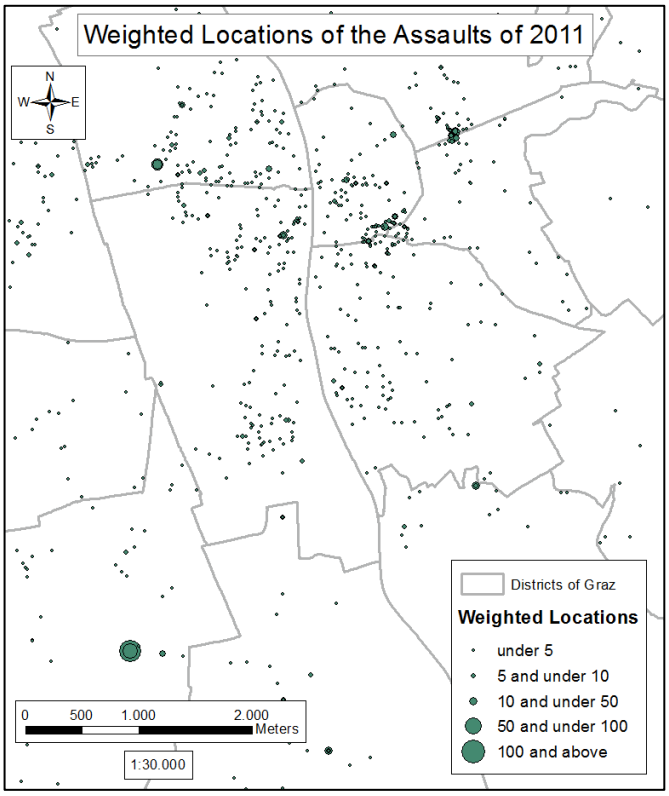

Figure 21: Central Extent of Weighted Assault Locations from 2011 - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

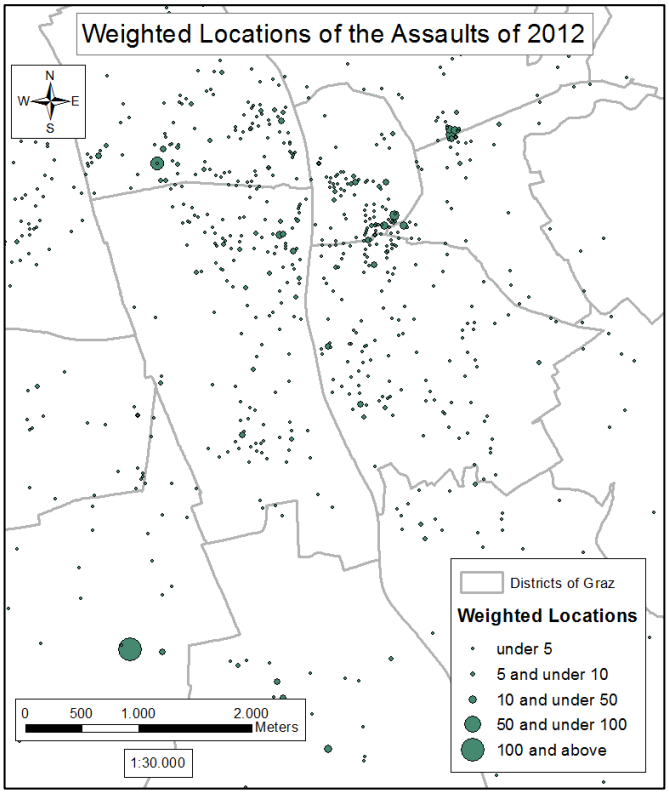

Figure 22: Central Extent of Weighted Assault Locations 2012 - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

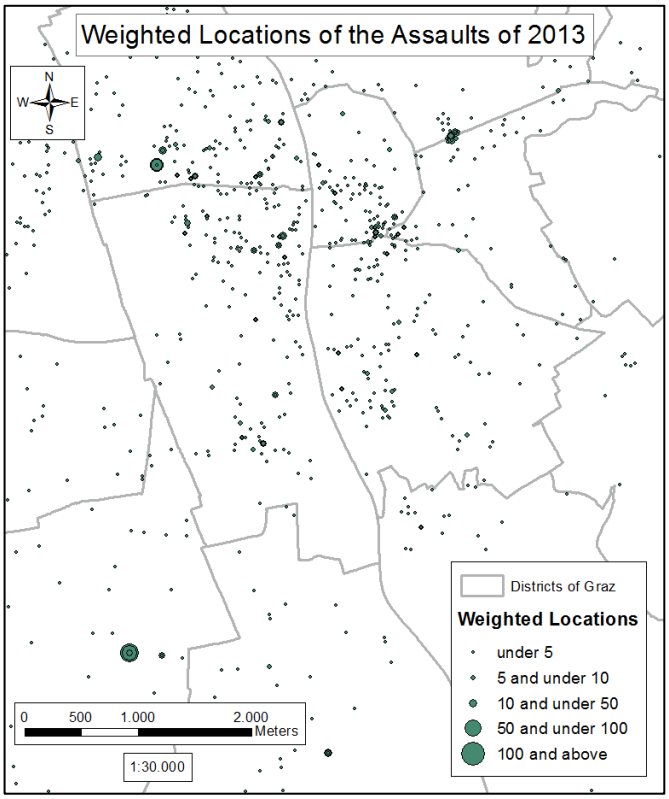

Figure 23: Central Extent of Weighted Assault Locations from 2013 - Classified by Natural Breaks (Jenks), Source: Own representation based on BMI

As expected, the locations and the individual weights shift throughout the years of the analysis. Interestingly, the areas with the highest density of weighted locations appear similar throughout the time span of the analysis.

4.2. Evaluating the Prediction Parameters

Pioneering works are dependent on experimenting with the model parameter to create the best possible outcome of the analysis with the given data. The next three sections describe the process how the model's parameters were evaluated.

4.2.1. Parameters of the Hotspot Analysis

The critical parameter for the Hotspot analysis is the search radius. As described in chapter 3.5.1, the Getis-Ord Gi* statistic and the Prediction Accuracy Index were used to evaluate this parameter.

The incidents that had the highest z-scores while being highly statistical significant were aggregated into the hotspots from the KDE (Figure 24). The KDE was performed for the incidents of every crime year with 11 distances, ranging from 100 to 350 meters.

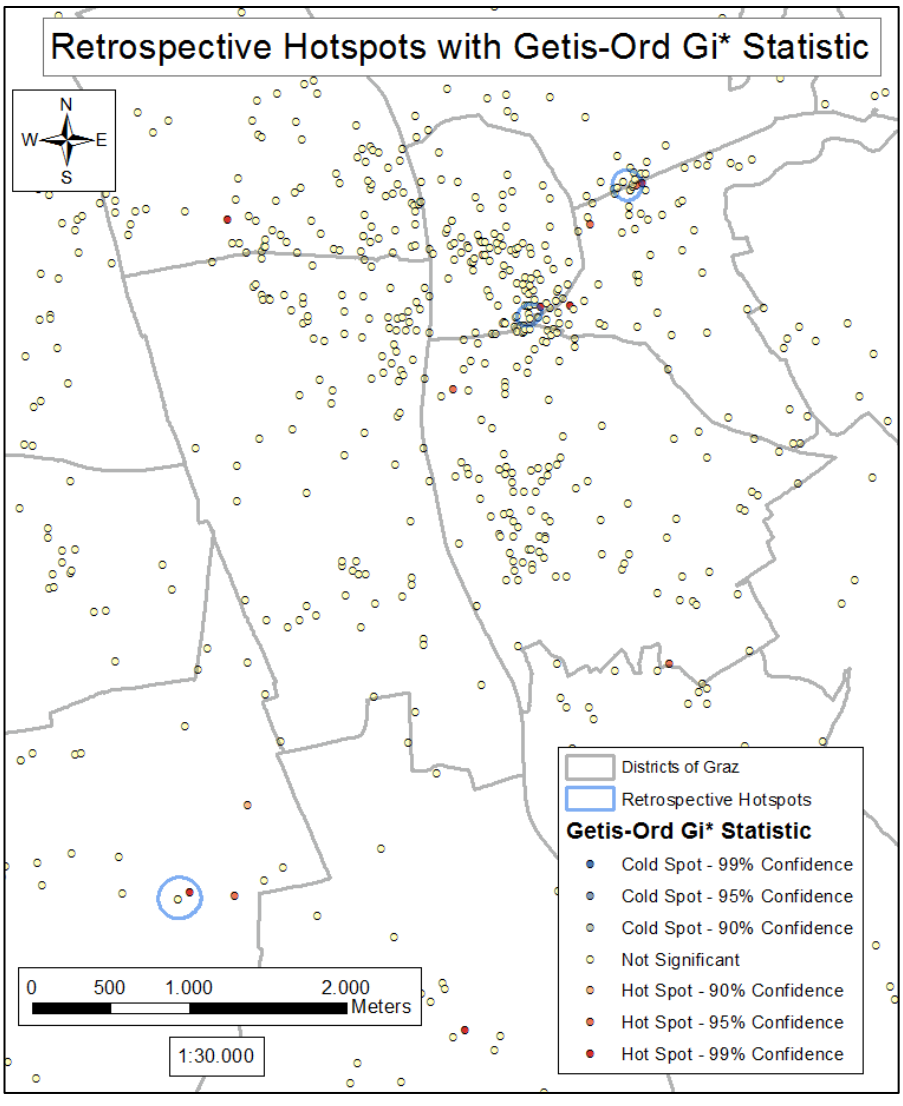

Figure 24: "Hot" Incidents of 2009 within the Hotspots created by a KDE with 250m search radius, Source: Own representation based on BMI

For the final evaluation of the search radius, three values were calculated. Table 1 shows the results for each search radius with their respective evaluation parameters.

Search radius [m]	Hot / Total Hot	Sum / Total Incidents	Prediction Accuracy Index
100	67%	10%	392
125	71%	10%	284
150	70%	10%	131
175	70%	10%	160
200	73%	11%	84
225	75%	11%	93
250	**81%**	**12%**	**76**
275	78%	11%	69
300	80%	12%	61
325	80%	12%	56
350	86%	13%	45

Table 1: Results of the Evaluation of the Hotspot search radius, Source: Own representation based on BMI

The first column lists the different search radii. The hit rate of the highly significant incidents, which are lying in the resulting hotspots from the KDE compared to all highly significant incidents in the study area, are shown in the second column. The third column displays the share of all incidents of all years lying in the hotspots compared to all incidents in the study area. The PAI is displayed in the last column.

The search distance of 250 meters is chosen to be the best option for the hotspot analysis. This distance strikes the best balance between the highest rate of hot incidents compared to all hot incidents and the highest rates of coverage for all incidents and the least total area of the study site.

4.2.2. Parameters of the Risk Terrain Modelling

The RTM was based on Open Street Map data. The buildings used for the analysis were selected by Caplan and Kennedy's (2011) classification of risk factors. Critical for the identification of the risk factors was the correlation of a location to assault incidents.

After preparing the input for the analysis, the weight of the building locations had to be set. There is no recommendation in the literature to weight one of these factors higher than the other. It is not the subject of this research to determine the weight of the risk factors. Therefore, each risk factor's influence on assault was treated equally. As discussed in section 4.2.2, buffer zones will be created around the OSM building centroids in order to assess the spatial influence of the risk factors. The buffer zones are weighted according to their spatial distance from the risk location. There is no guideline to determine an appropriate weight. The reasoning was to give areas that are located between several risk locations a higher risk value than those with no risk location nearby. The distance for each buffer zone must fit to the study area and the spatial distribution of the risk locations. The search radius of the Hotspot analysis, which is 250 meters, was used as a starting point for the determination of the optimal search distance. By visually analysing the OSM locations and their distances to each other, it was chosen to use three buffer zones with the distances of 100, 300 and 500 meter. This way, visual clusters of risk locations are targeted and the weight used for the RTM is comparable to the search distance of the hotspot analysis. The weight of the grid cells in each of the buffer distances was set from 0.5 to 1.5, giving the grid cells with the highest proximity the highest risk value (Figure 25).

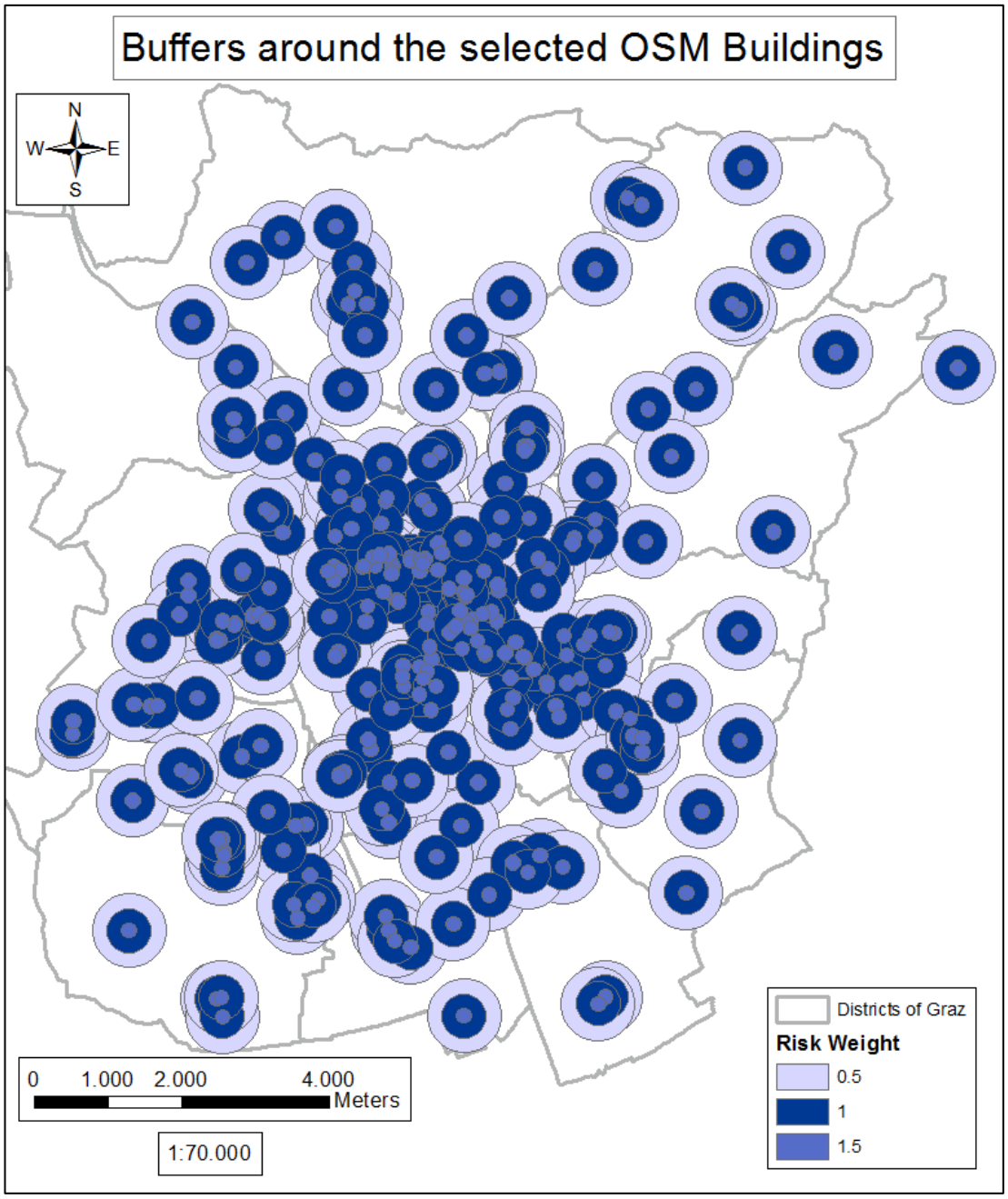

Figure 25: Risk Values for the OSM Risk Locations, Source: Own representation based on BMI

4.2.3. Combining the results of the Hotspot Analysis and the RTM

The goal of this research is to improve the accuracy of crime predictions by combining retrospective and prospective analyses. This endeavour requires the comparability between the results of the Hotspot analysis and the Risk Terrain Modelling. Each approach results in a grid. The final risk terrain map is acquired by combining the grids and aggregating their grid values. Therefore, the cell size and grid orientation for each analysis was used with special care.

Even after taking these precautions, there is still an unseen risk by overlaying the resulting grids. Figures 26 and 27 show the results for the Hotspot analysis and the result of the RTM, respectively.

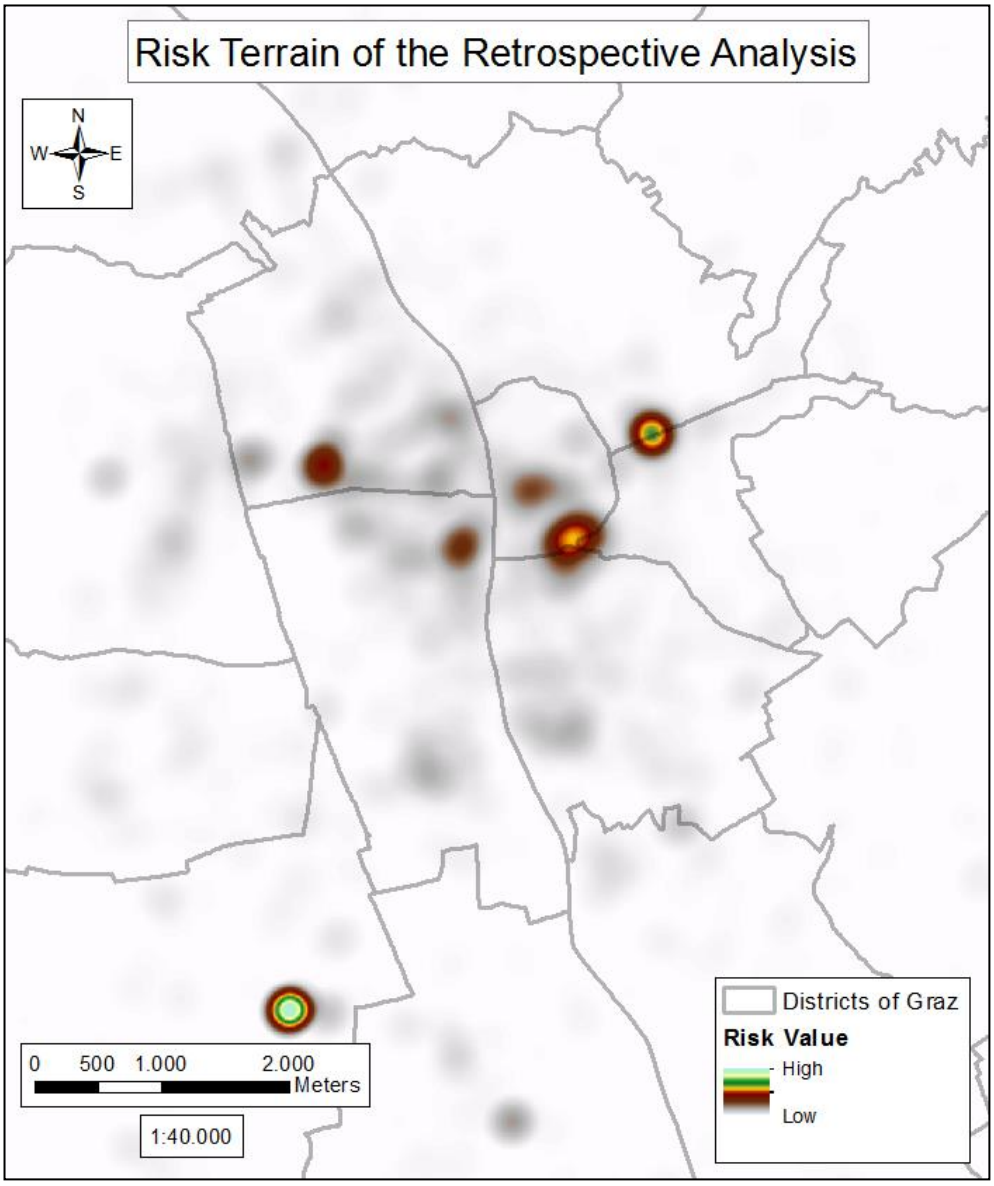

Figure 26: Result of the Hotspot Analysis, showing the risk values for each grid cell, Source: Own representation based on BMI

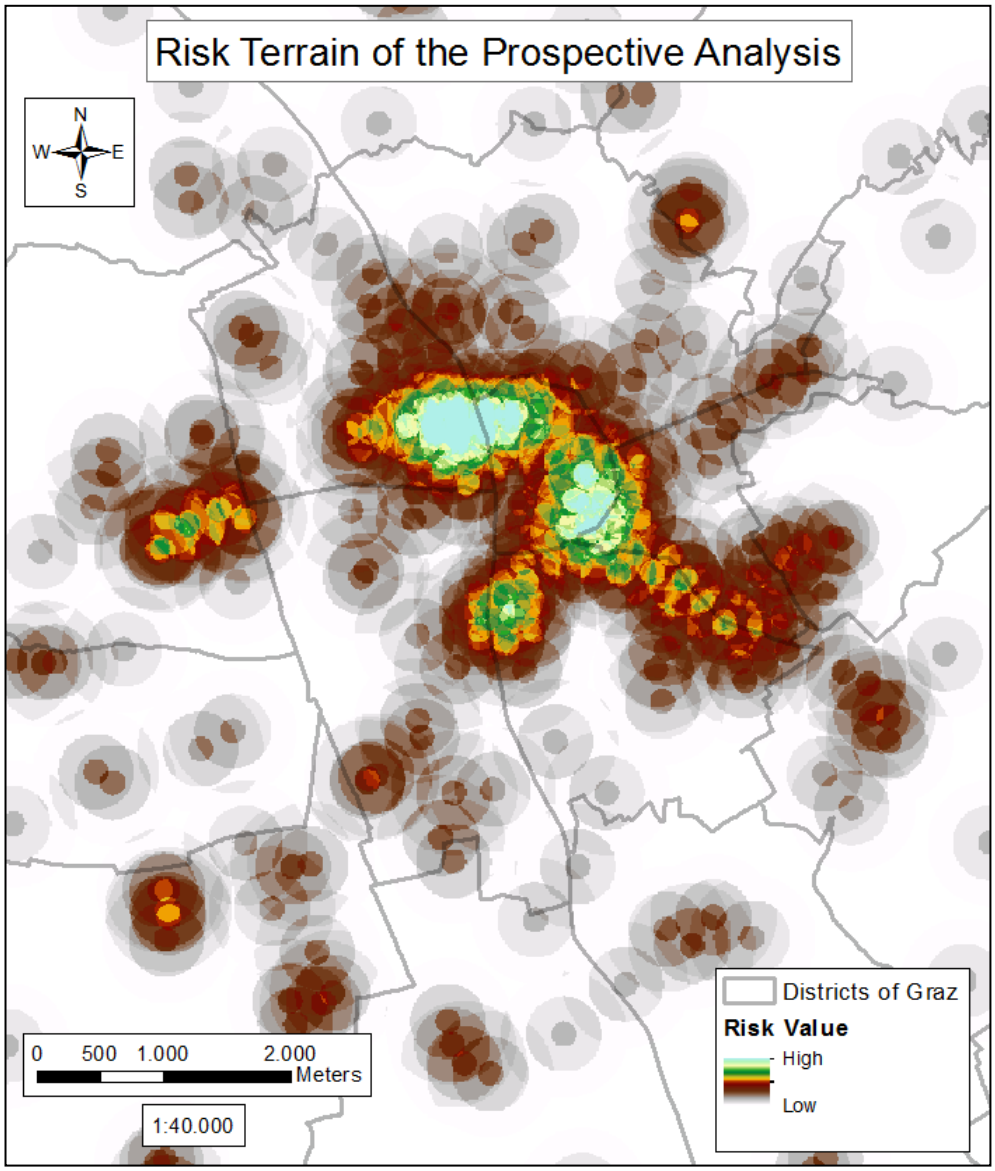

Figure 27: Result of the RTM, showing the risk values for each grid cell, Source: Own representation based on BMI

The visual interpretation of the two risk maps already indicates why the results are still incomparable. Due to the different calculations of the grid values, the approaches differ in their range of cell values. The RTM achieves higher risk values than the Hotspot analysis, but this is not related to the prediction accuracy of the approaches.

The different ranges need to be standardised in order to lead to an undistorted result. A fuzzy function is used to normalize the grid values. After the fuzzyfication, the resulting grids range from 0 to 1, meaning they are comparable. The result of the fuzzyfication of the Hotspot Analysis and the RTM is shown in Figure 28 and 29.

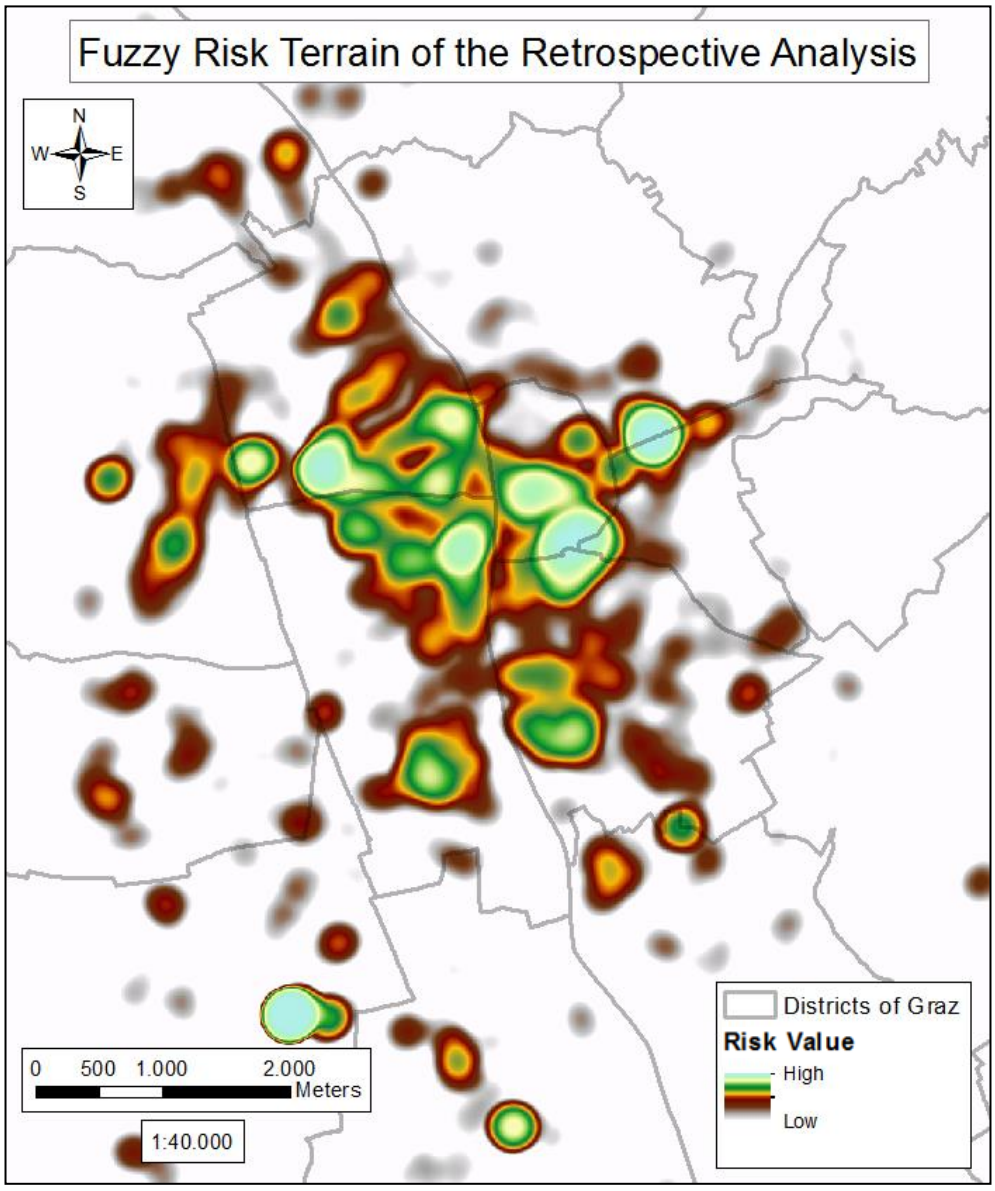

Figure 28: Fuzzyficated results of the Hotspot Analysis, showing the risk values for each grid cell, Source: Own representation based on BMI

It can be observed, that the main hotspot areas highlighted in the original analysis result are the same as in the fuzzyficated result of the Hotspot analysis, but the areas with medium risk are now stressed more than before. The same behaviour can be observed in the Risk Terrain Modelling. Former medium risk areas are now included in the high risk areas.

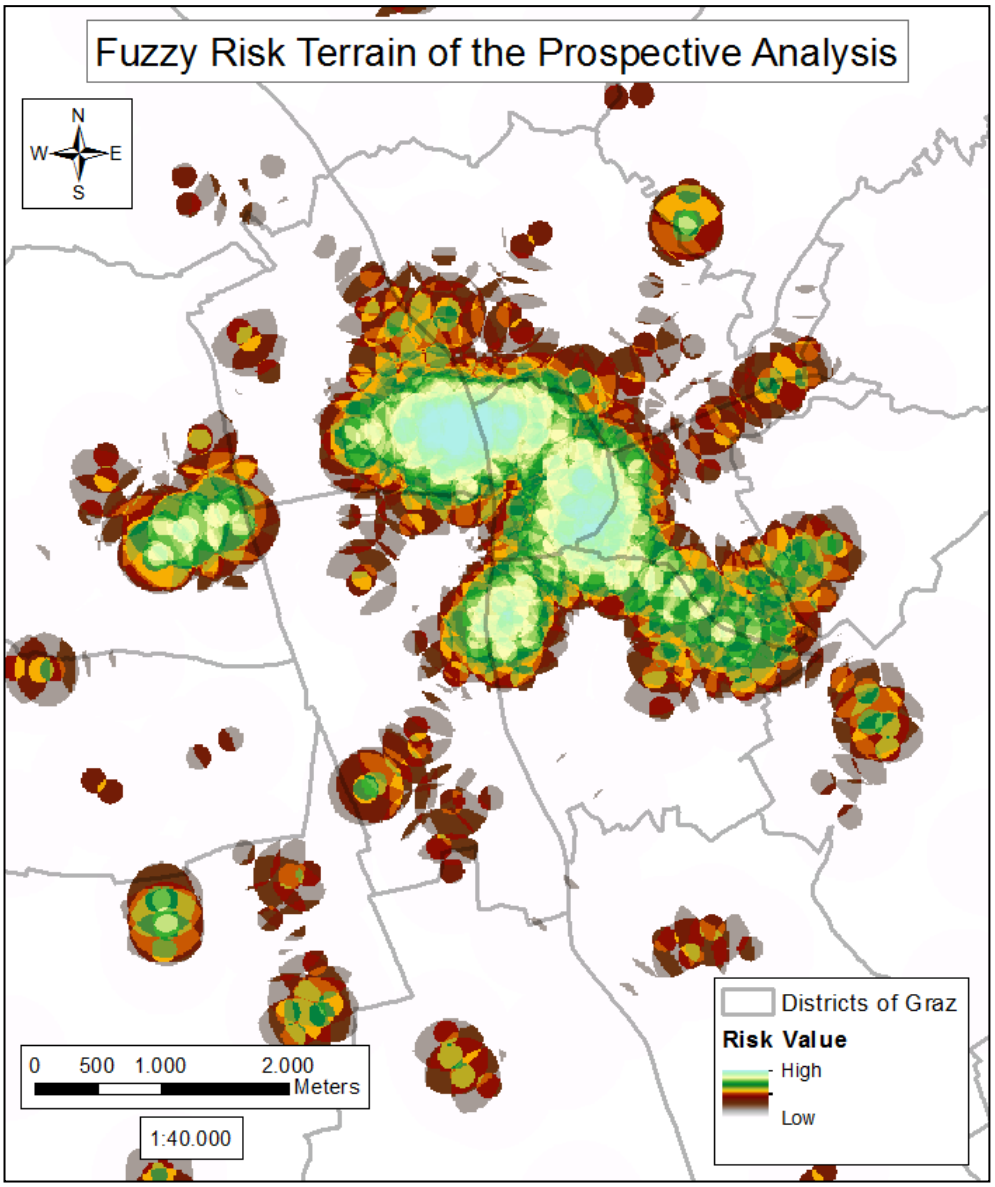

Figure 29: Fuzzyficated results of the RTM, showing the risk values for each grid cell, Source: Own representation based on BMI

The following figure shows the risk terrain map of the analysis. It is the result of combining the grids of the Hotspots Analysis and the RTM. The results of the retrospective and the prospective approach were weighted equally.

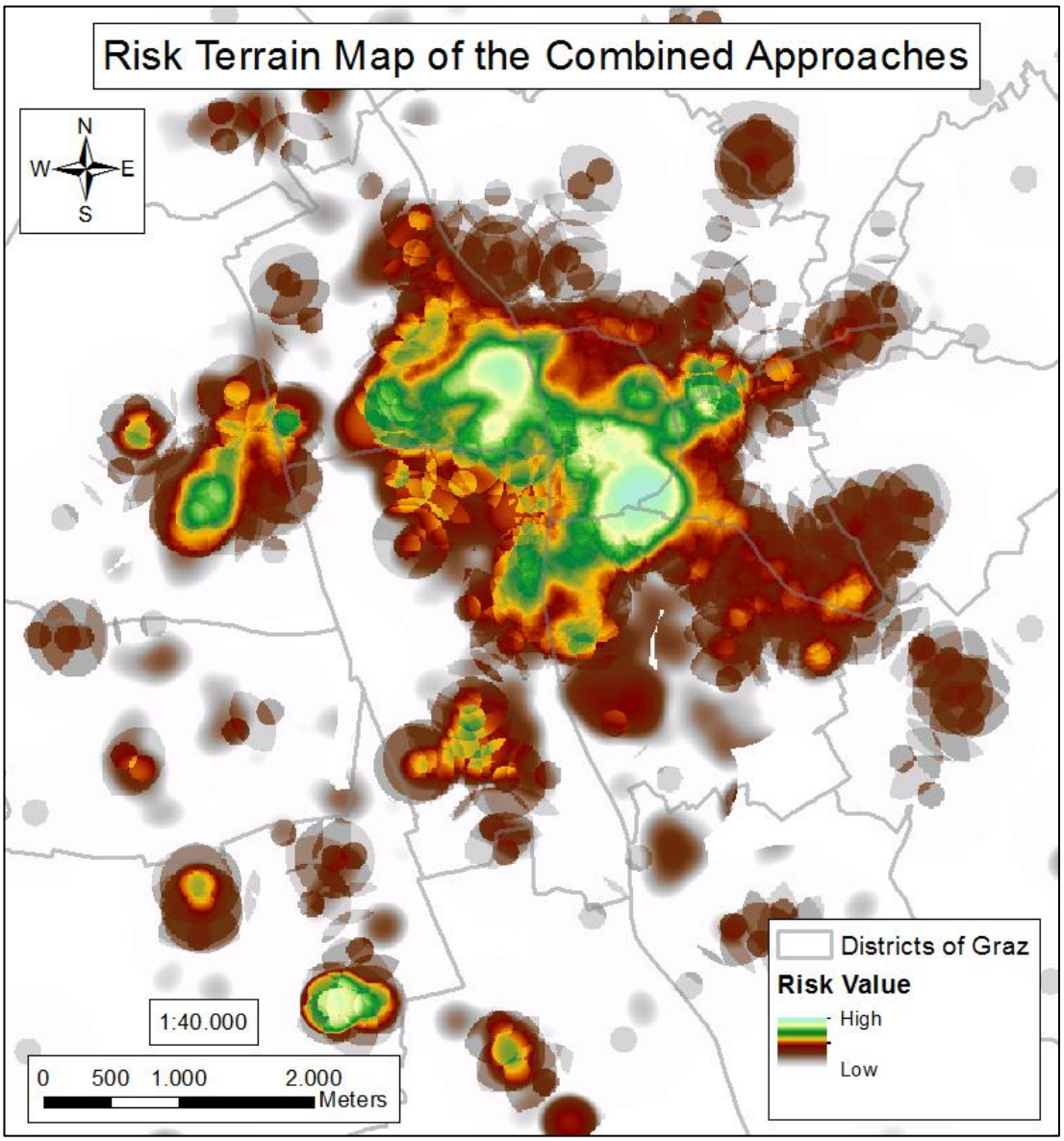

Figure 30: Risk Terrain Map of the Prediction Model, showing the risk values for each grid cell, Source: Own representation based on BMI

4.3. Discussing the Results

In this research, a model was created that combines analyses from a retrospective and a prospective prediction approach. The resulting risk terrain map is a continuous raster layer. This chapter contains the discussion of the applied prediction analysis and its prediction accuracy.

4.3.1. Problems in the Prediction Method

To evaluate the accuracy of the prediction model, the final grids must be converted into polygons. As described in chapter 3.5.3, the risk terrain map was sliced, using a cut-off value that is the highest of the ten reclassified classes. The slicing was conducted using the Natural Breaks (Jenks) method. With this method, the final hotspots are created in the same way as in the Hotspot analysis, making the results comparable.

After converting the grid cell with the highest risk value into a polygon file, the final hotspot area can be analysed. Figure 31 shows the areas where assaults are assumed to happen in the prediction year.

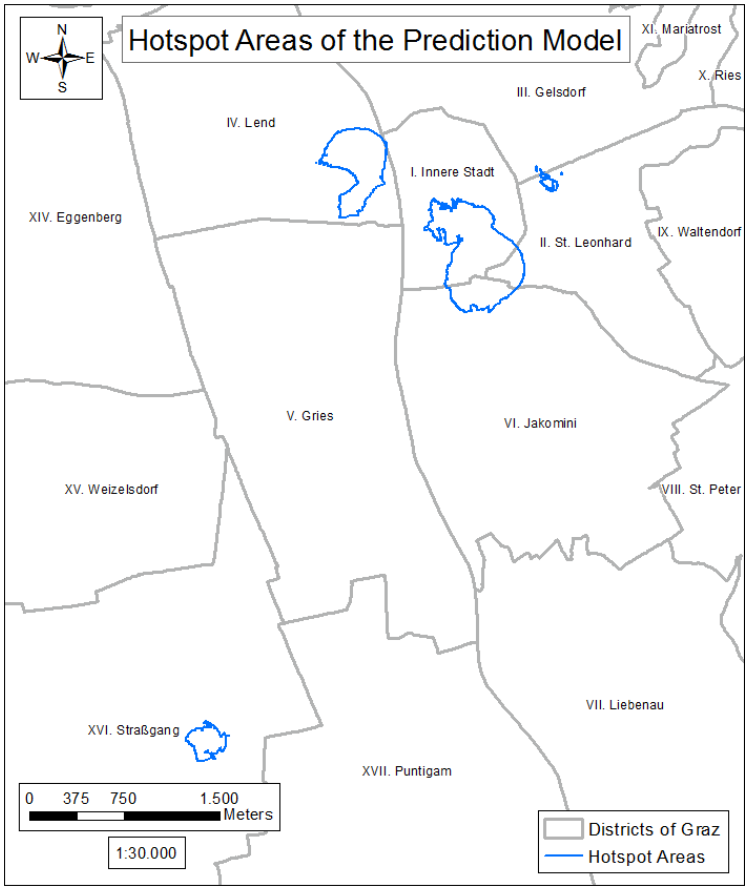

Figure 31: Hotspot Areas of the Prediction Model, Source: Own representation based on BMI

It can be observed that there are three major hotspot areas. Two of which are located in just one district. The biggest hotspot area is divided by three districts. These districts are "Innere Stadt", "St. Leonhard" and "Jakomini". Each of these districts is located in the inner city. “Lend”, the second biggest hotspot, is also located in the inner city.

Before the two hotspots are examined any further, attention should be put on the sliver polygons. The shapefile, which resulted from the conversion of the grid into a polygon, has a total of 25 entries (Table 2).

Feature	AREA [m²]
1	18
2	58.964
3	3
4	18
5	43
6	240
7	61
8	146
9	419.913
10	22
11	7
12	2
13	158
14	1
15	240
16	1
17	171
18	1
19	7.777
20	9
21	122
22	16
23	205
24	2
25	209.068

Table 2: Size of the Final Hotspot Areas, Source: Own representation based on BMI

Next to the three major hotspot areas, there were several other "hotspots" created. This can be explained by recalling the method to create the final hotspots. The risk terrain map was reclassified into ten classes. These classes are constructed by grouping the most similar risk values together while having the biggest difference within all classes. The reclassification does not consider the spatial location of each value. Therefore, every grid value located in the highest class is converted to a polygon, which leads to the creation of very small polygons. These by-products of the analysis will be called "sliver polygons" from now on.

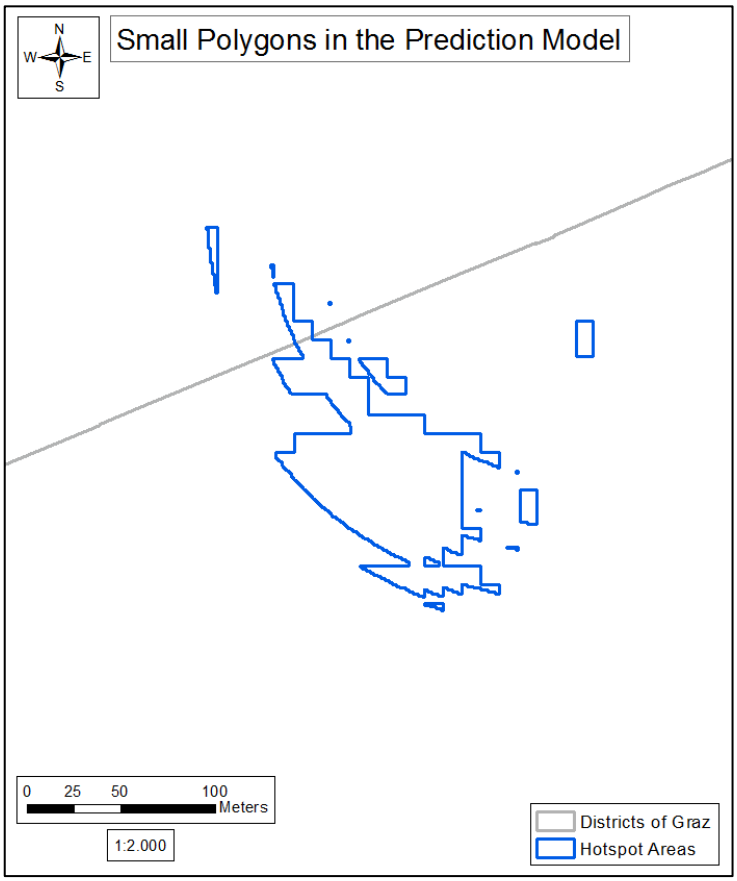

Figure 32: Very Small Polygons in the Result of the Prediction Model, Source: Own representation based on BMI

The classification of sliver polygons is subjective and has to be adjusted for a new study site. The threshold for categorizing very small polygons in this analysis was 10.000 m^2, making all polygons portrayed in Figure 32 sliver polygons. For the evaluation of the prediction accuracy, the sliver polygons are not taken into account. There is no incident of the prediction year located in these areas, however since they show where the Hotspot analysis and the RTM are working in opposite directions, their appearance is helpful for further analyses.

Figure 33 shows an extent of the Hotspots analysis result for the largest sliver polygons. To show the values with which the values of the risk terrain map are calculated, the fuzzyficated grid layers are shown.

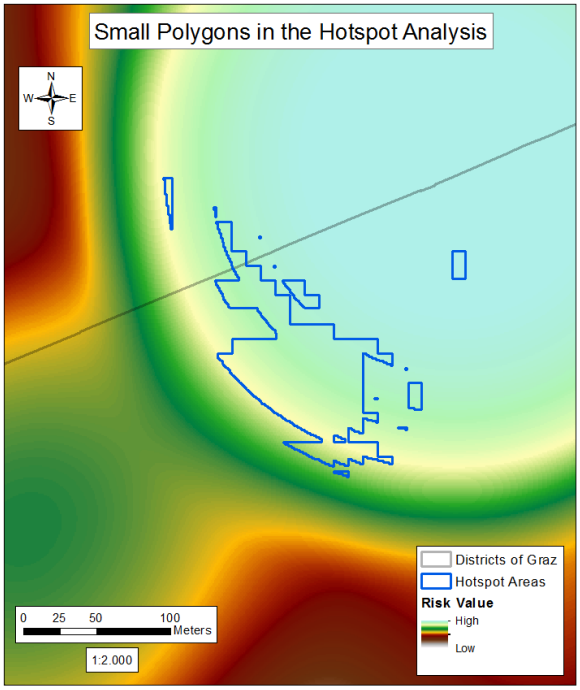

Figure 33: Hotspot Analysis Result zoomed onto the biggest Sliver polygons, Source: Own representation based on BMI

Most sliver polygons are located on the outside, but are still in the high risk area. This indicates that there are relatively high grid values used for the calculation of the final value.

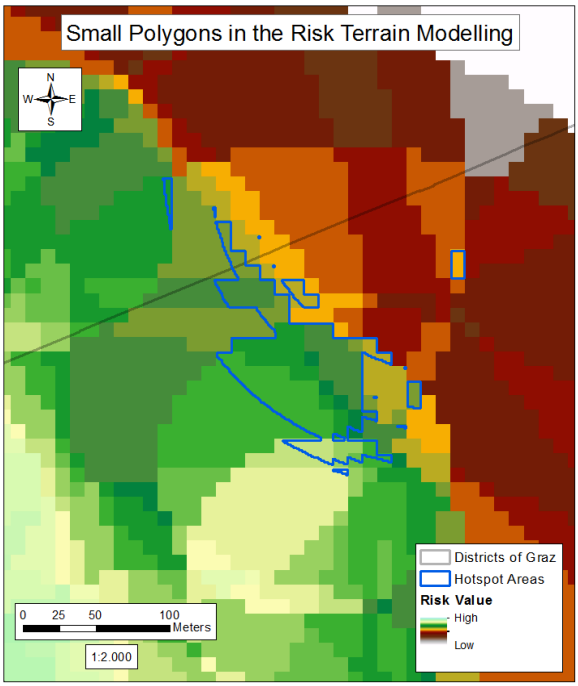

Figure 34: RTM Result zoomed onto the biggest Sliver Polygons, Source: Own representation based on BMI

In the RTM result (Figure 34), the sliver polygons are located on the outside of a high risk area, that is located on the opposite site as the one of the Hotspot analysis. Only the cells that achieved an average or above average value were used for the calculation of the final risk value.

The sliver polygon, located in the light brown area, is showing the compromise between the two hotspot areas derived from the Hotspot analysis and the RTM.

The creation of sliver polygons shows evidence of where the two approaches are working in opposite directions. These areas can give crucial information regarding the determination of the model parameters. Balancing the search distance of the Hotspot analysis and the buffer distance of the RTM reduces the appearance of sliver polygons.

The sliver polygons are not the only feature of the prediction result, which might confuse the user. Because of the use of raster data, the boundaries of the hotspot areas are not smooth, but rather edged (Figure 35). Especially on a closer view, a grid cell might or might not cover a street intersection or a house.

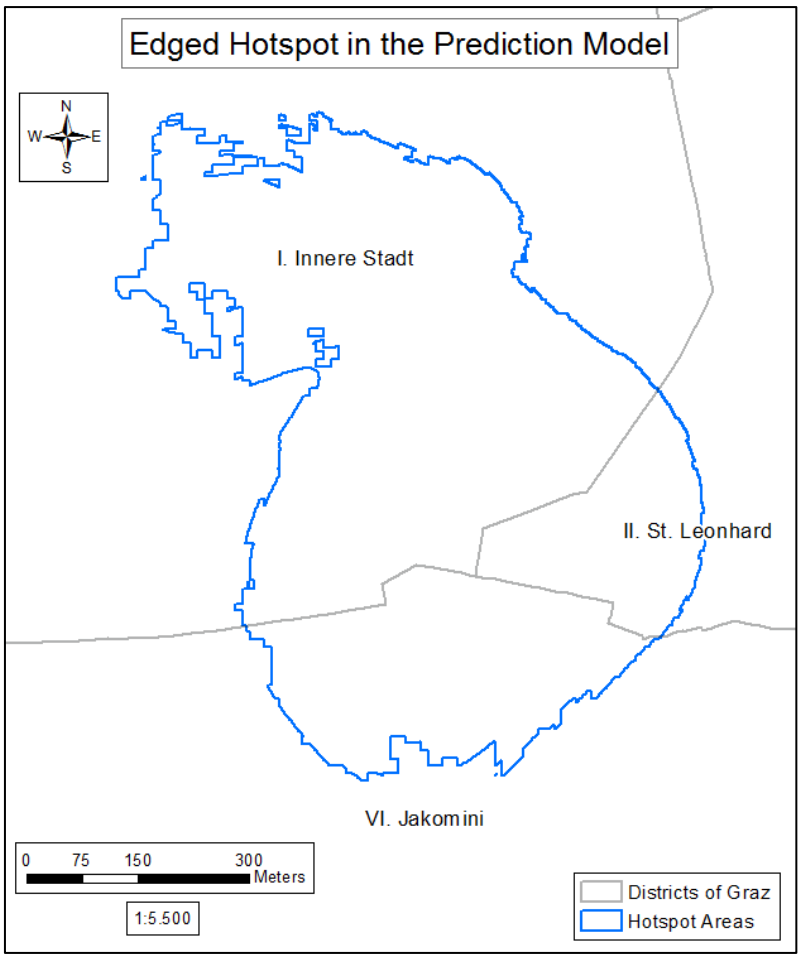

Figure 35: Edged Appearance of the Hotspot Boundaries, Source: Own representation based on BMI

The focus of this research was not to produce a risk terrain map for studying the area covered by the final hotspots. If the user needs to create hotspots with smooth boundaries, the hotspots can be smoothed. Generalizing the hotspots' boundaries creates a more appealing visual representation of the areas at risk (Figure 36).

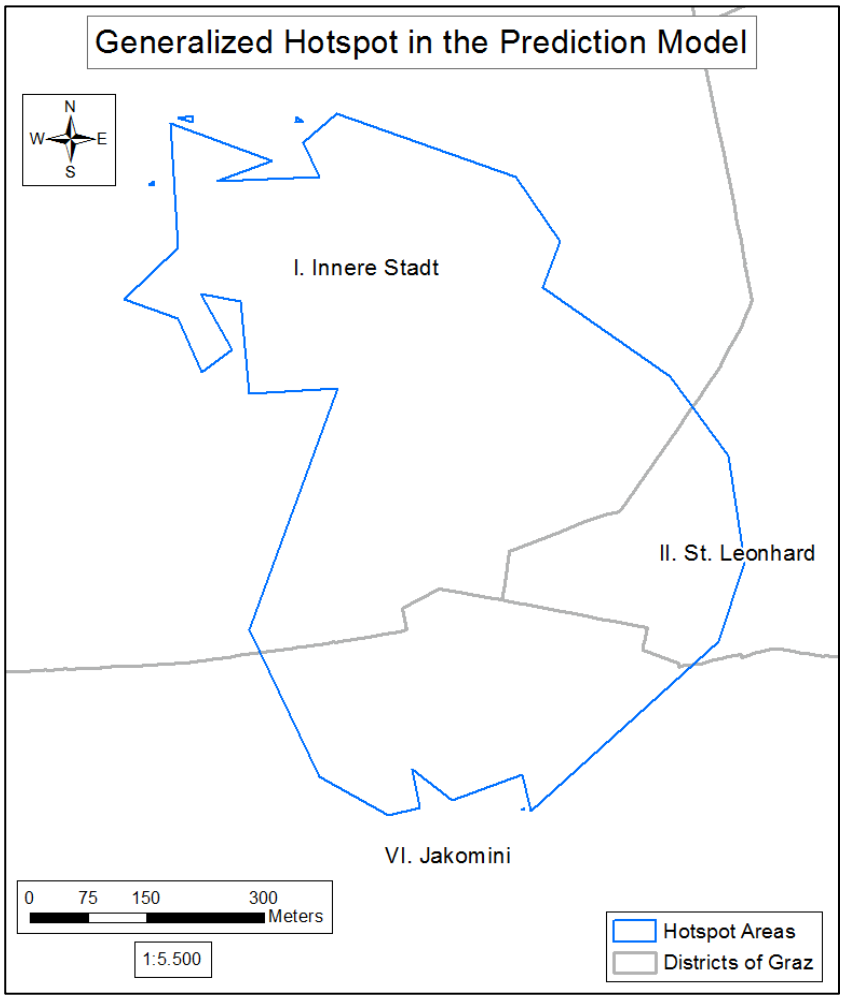

Figure 36: Generalized Hotspot Areas with 25 meter tolerance, Source: Own representation based on BMI

The generalization might be more visually appealing, but it is not a simple task to perform. Smoothing boundaries might lead to a loss of information, as well as a creation of unwanted multiparts. If the research purpose requires such a treatment of the results, special care must be taken not to avoid distorting the new result.

4.3.2. Evaluating the Prediction Accuracy

There were 1,544 assault incidents reported in Graz throughout 2014. 149 or 9.7% of them were located inside the hotspot area identified by the Hotspot analysis (Figure 37).

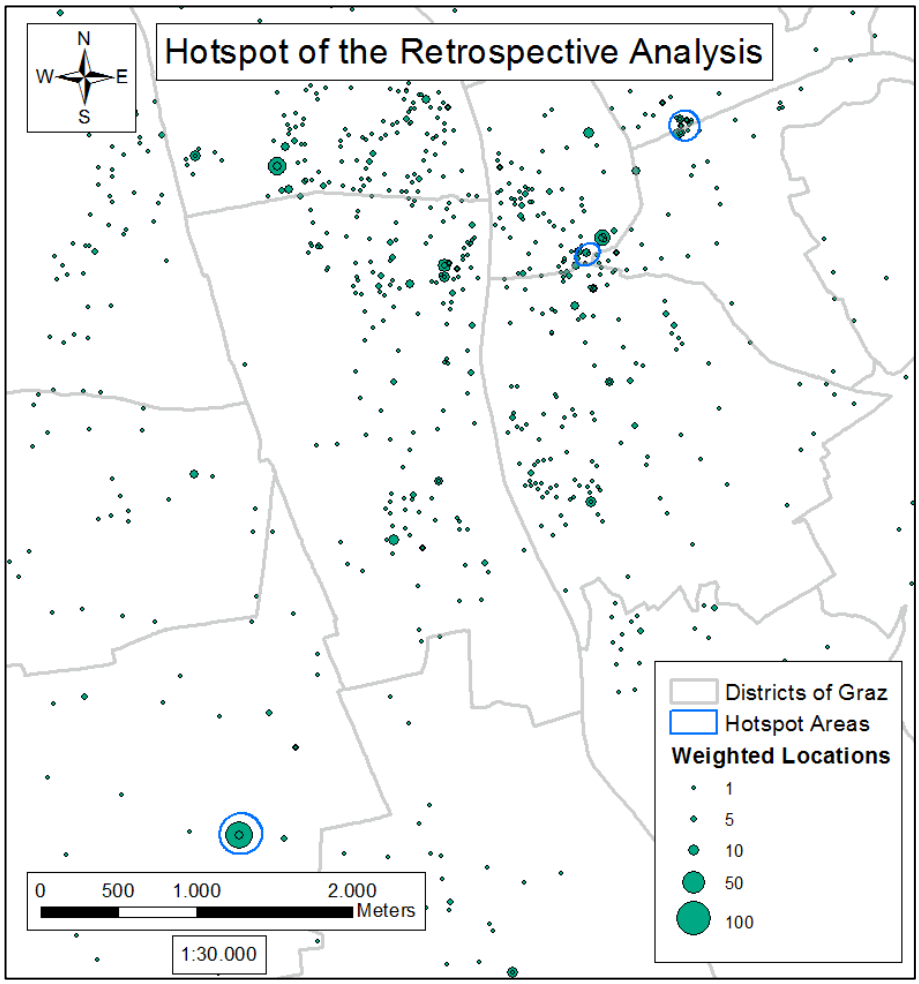

Figure 37: Hotspot Analysis Result with agglomerated Incidents of 2014, Source: Own representation based on BMI

Research suggests the fineness of this approach for predicting crime incidents, but it implies the restrictions of the method as well. Retrospective analyses are solely able to predict hotspot areas which are located within the geographic extent of the input data. Rising crime prone areas located outside of this extent cannot be detected using a retrospective approach.

The only way to identify hotspot areas in new locations is to analyse its risks of becoming victimized. The prospective approach used in this research is represented by the Risk Terrain Modelling. The RTM used Open Street Map data to calculate the identification of areas that possess a suspicious amount of aggravating factors. The hotspot areas predicted by the RTM are created by the same methodology as used in the Hotspot analysis. The resulting hotspots cover 30 or 1.9% of the 1.544 assault incidents of 2014 (Figure 38).

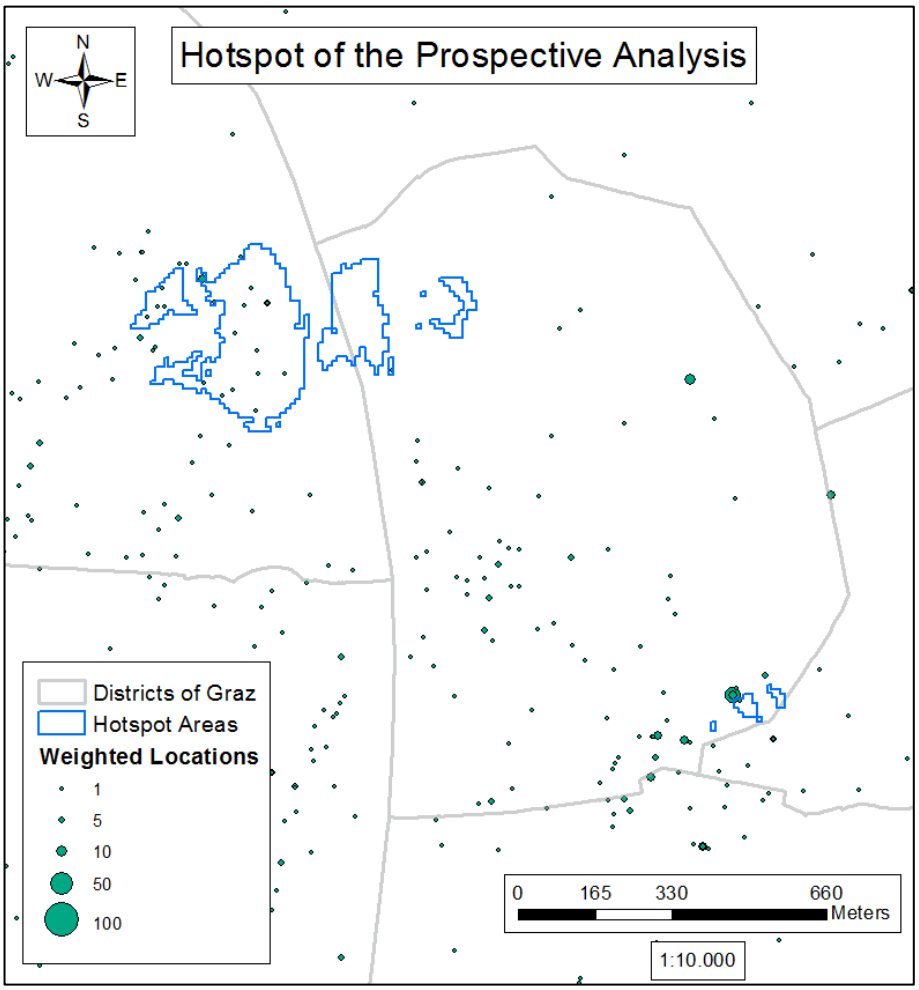

Figure 38: RTM Result with agglomerated incidents of 2014, Source: Own representation based on BMI

Comparing the prediction accuracy of both approaches leads to an evident result. The hotspot areas of the Hotspot analysis cover five times more incidents of the prediction year as the ones of the RTM. The comparison of the resulting shape of both analyses raises an interesting point. While the hotspots of the Hotspot analysis are represented by ovals, the result of the RTM shows several edged areas. Although these areas are quite close to each other, the grid cells do not form a cohesive shape.

The reason for this dispersed appearance is easily explained by recalling the methodology of the RTM. The grid cells achieve a higher value when they are targeted by the spatial influence of the risk location. Due to the distribution of the risk locations and their spatial influence, an irritating picture results by colourizing the grid cells by their risk value (Figure 39).

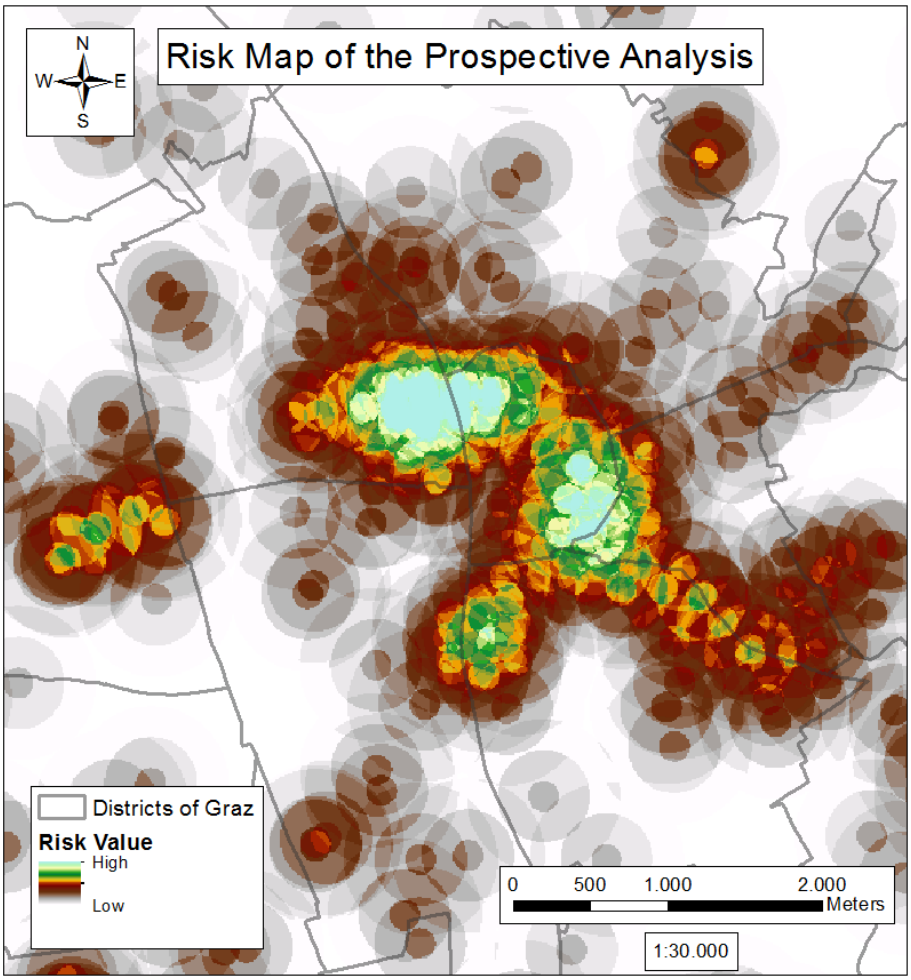

Figure 39: Central Extent of the RTM Risk Map, Source: Own representation based on BMI

The two major hotspot areas detected by the RTM are jagged, which is problematic for defining hotspot areas. However, the prospective approach gives valuable information to new areas at risk and supports the prevalence of retrospective hotspots (Figure 40).

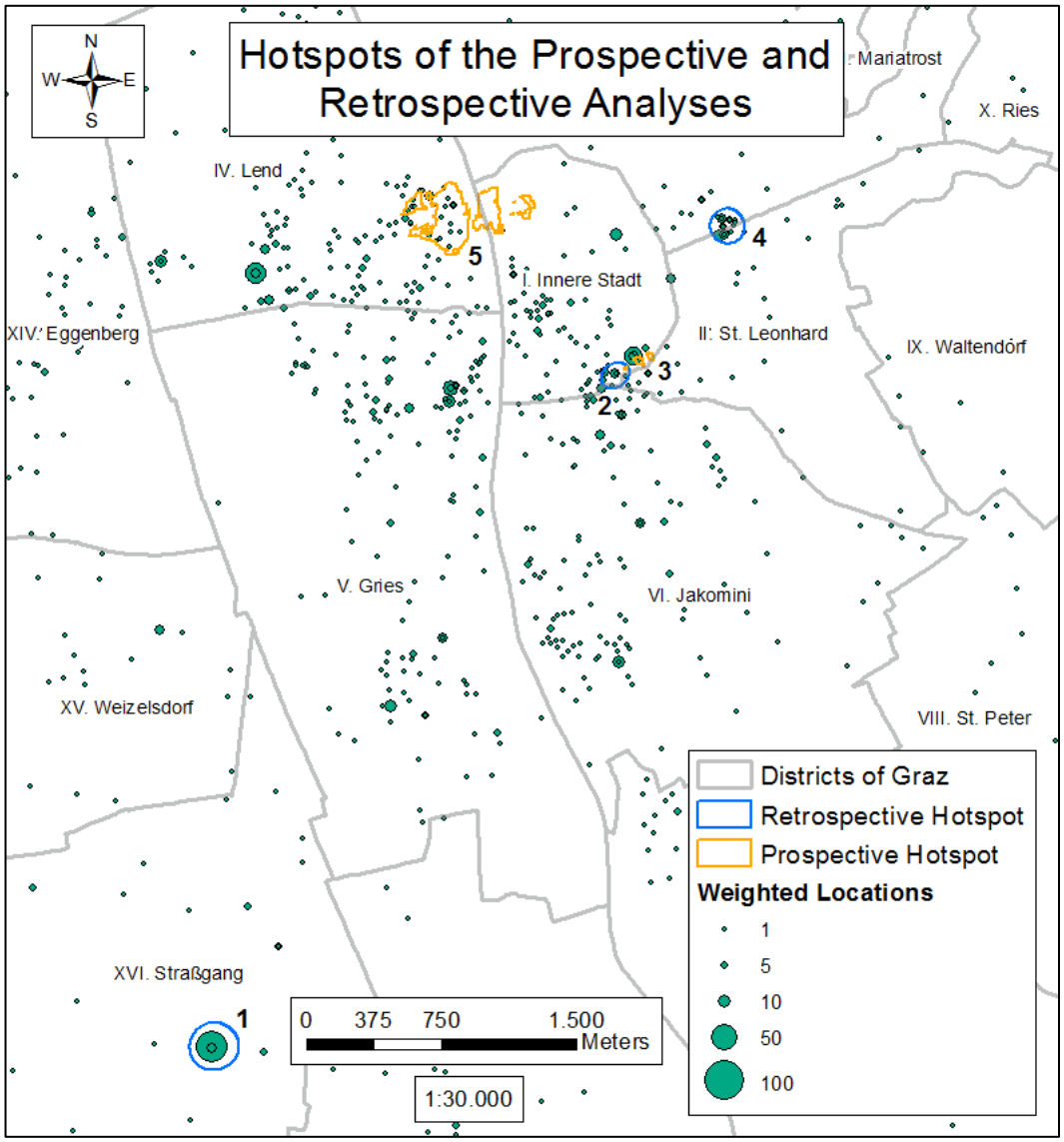

Figure 40: Hotspots of the RTM and the Hotspot Analysis with the agglomerated Incidents of 2014, Source: Own representation based on BMI

One of the hotspot areas of the RTM (number 3) is not within the retrospective hotspot (number 2), but it identifies that area to be at risk as well. Although the RTM fails to predict the hotspot areas 2 and 3, it manages to highlight a completely new area at risk (number 5). There are 28 incidents located in this hotspot area, which cannot be predicted by the retrospective approach.

Essentially, the prediction accuracy of the hotspot areas identified by the methodology presented in chapter 3.5 is higher using the retrospective approach. The prospective analysis is able to predict a new hotspot area, but fails to predict two major hotspot of the retrospective analysis (number 1 and 4).

In order to create the final hotspot of the analysis, the risk maps of the RTM and the Hotspot analysis were combined. As described in chapter 3.5.3, the grid values of the results for each approach varied in their value range. Therefore a fuzzy algorithm was used to normalize the grids. By combining the normalized results of the retrospective and the prospective analysis, 293 or 19.0% of the assaults of 2014 can be predicted (Figure 41).

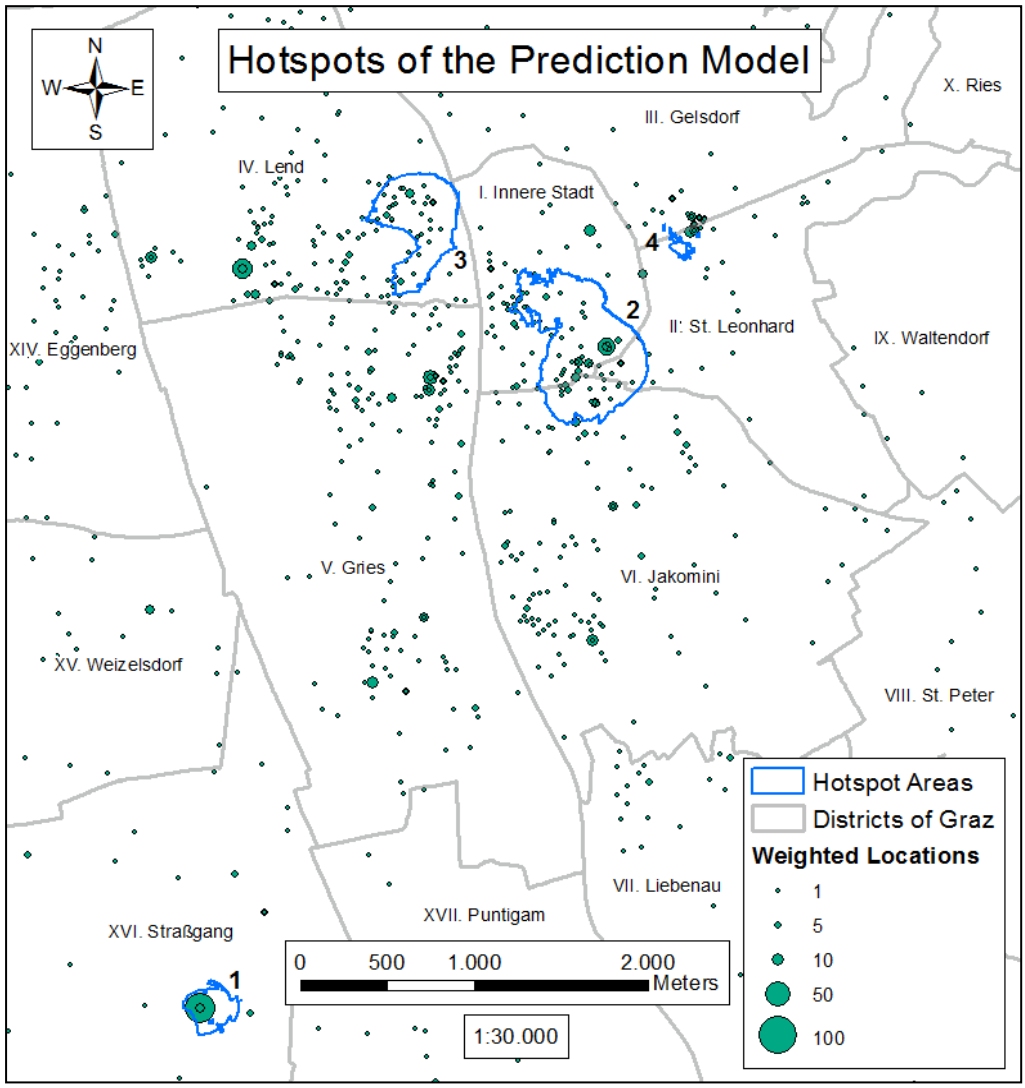

Figure 41: Result of the Prediction Model with the agglomerated Incidents of 2014, Source: Own representation based on BMI

The improvement of the prediction accuracy from 9.7% to 19.0% is astonishing, but biased as well. For the evaluation, the final grid was reclassified into 10 classes by the Natural Breaks (Jenks) method. The cell values of the highest class were used as the final hotspots. This approach was used for the results of the three risk maps, created in the analysis. The evaluation results of the Hotspot analysis and the RTM are comparable, but the comparison of the Hotspot analysis and the final risk terrain map is distorted.

As seen in Figure 42, the fuzzyfication increases the area covered by the hotspot areas.

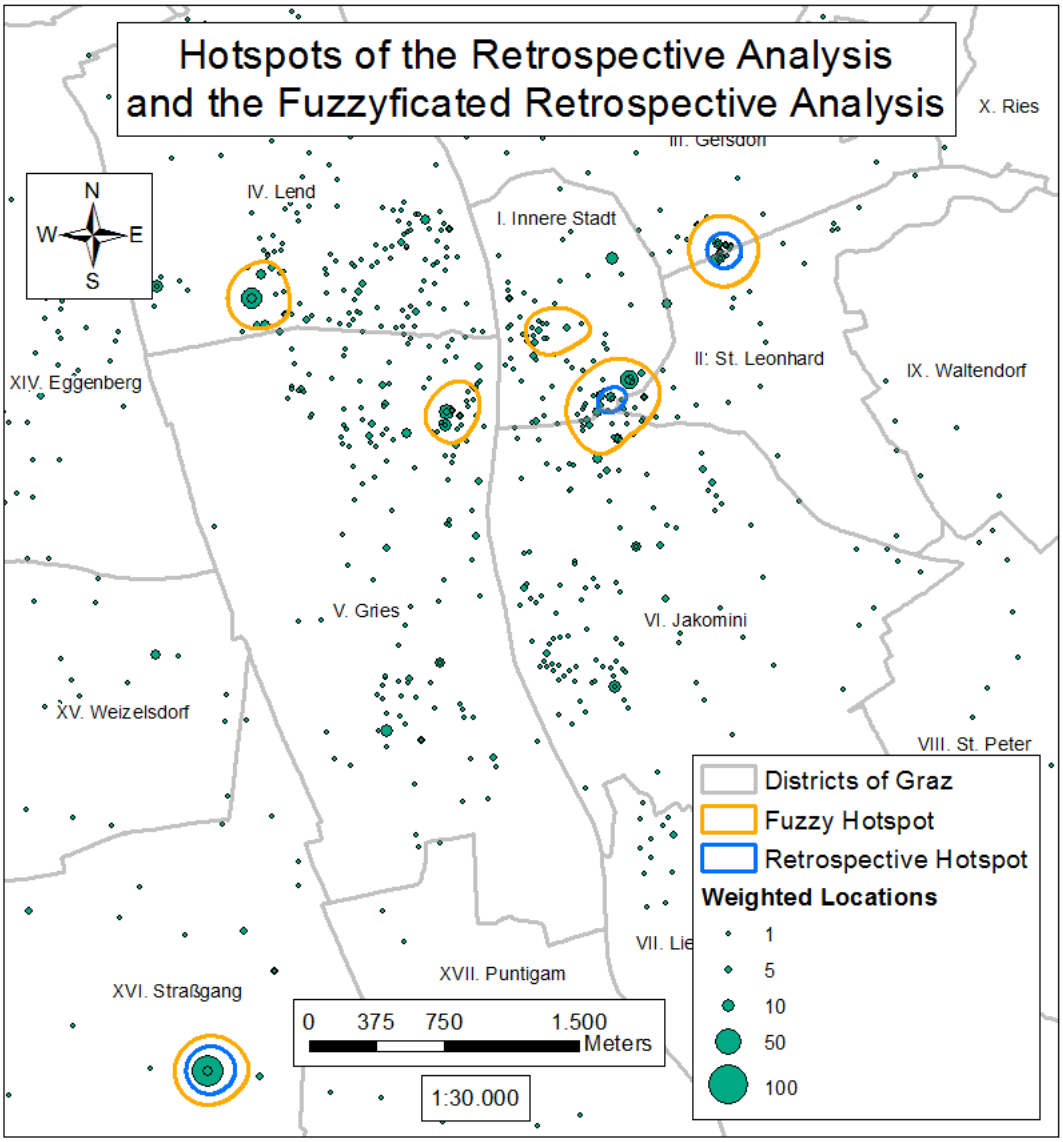

Figure 42: Fuzzyficated Result of the Hotspot Analysis in comparison to the original result, Source: Own representation based on BMI

Due to the normalization of the grid, more grid values fall into the highest category. Therefore, the size of the fuzzyficated hotspots is around 6 times larger than the Hotspot results. This results in a massive increase in the prediction accuracy.

To perform an unbiased comparison of the prediction accuracy of the Hotspot analysis result and the result of the prediction model, the hotspots of the normalized Hotspot analysis must be used. As shown in Table 3 the retrospective approach exceeds the result of the prediction model in the prediction accuracy rate by 7.7%.

Method	Incidents of 2014 in the Hotspots	Incidents of the Prediction Year 2014	Prediction Accuracy	Prediction Accuracy Index
Fuzzy Hotspot Analysis	412	1,544	26.7 %	49.36
Prediction Model	293	1,544	19.0 %	34.70

Table 3: Prediction Accuracy Results of the Final Hotspots and the Fuzzyficated Hotspots of the Retrospective Approach, Source: Own representation based on BMI

Figure 43 shows both hotspot areas with the weighted point locations of the assault incidents that occurred throughout 2014.

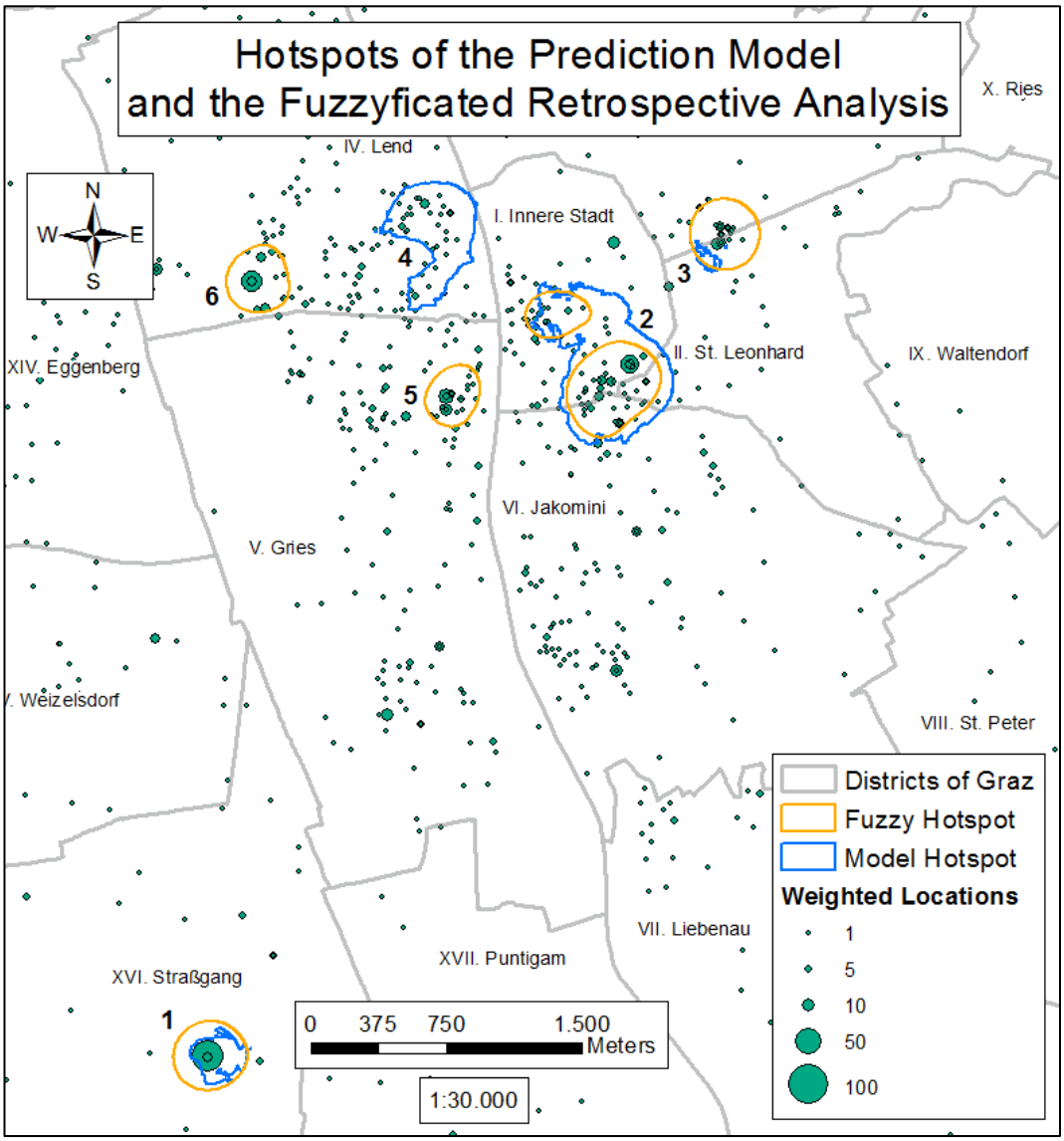

Figure 43: Result of the Prediction Model and Fuzzyficated Hotspot Analysis Result with the Agglomerated Incidents of 2014, Source: Own representation based on BMI

Two hotspot areas are highlighted by both results, including the most important hotspot in the southwest (1) and the largest hotspot areas in the north east (2). The approaches conflict with each other in the four hotspots in the north and north-western corner of the map extent (Figure 44).

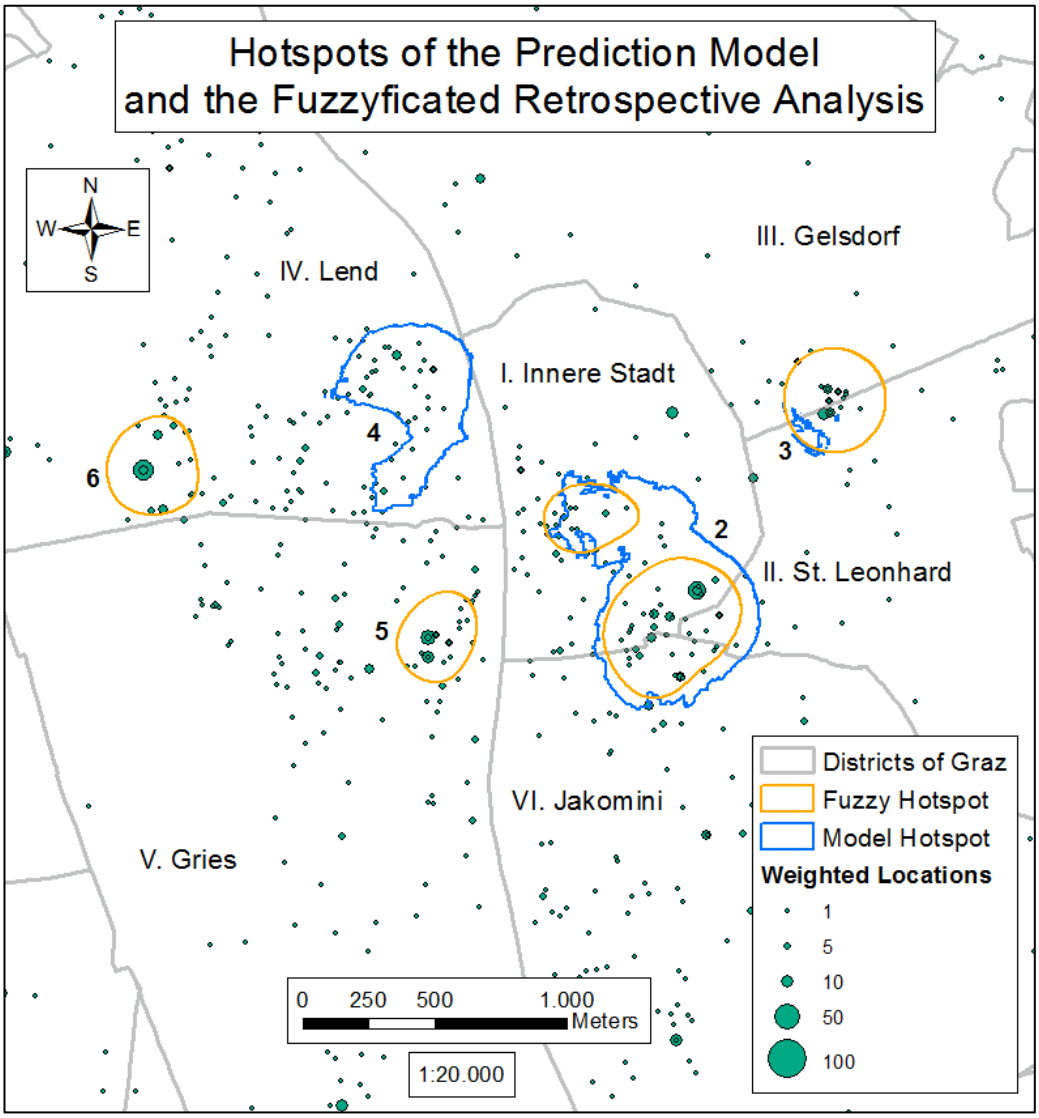

Figure 44: Central Extend of the Hotspots - Result of the Prediction Model and fuzzyficated Hotspot Analysis Result with the Agglomerated Incidents of 2014, Source: Own representation based on BMI

While the prediction model predicts the hotspots 1 and 2, it fails to predict the hotspot 3, 5 and 6. It barely misses the locations of 47 assault incidents of 2014 located in hotspot 3. The 57 incidents in hotspot 5and the 71 incidents in hotspot 6 are not included in the final analysis at all. But the 50 incidents located in hotspot 4 can only be predicted by the final model. The reduction of the prediction accuracy results from the opposing risk values of the Hotspot analysis and of the RTM. The appearance of sliver polygons shows the opposing influence on the risk values. But the prediction accuracy is also lowered by the increase in the size of the predicted hotspot area.

The fuzzyficated results cover an area of about 69 ha, whereas the result of the prediction model covers 80 ha. The PAI of the fuzzyficated Hotspot analysis surpasses the prediction model by 14.66 points (Table 3). Several tests were performed on the search distance of the Hotspot analysis and the buffer distance of the RTM to balance the opposing risk weights, but with the available data, it decreases the total accuracy of the prediction model. There were only 245 building locations available for the RTM, while there were 7,445 incident locations used for the Hotspot analysis. This implies that a higher amount and more accurate building data would greatly influence the result of the prediction accuracy and thus the final model.

Overall, the final prediction model achieved a lower prediction accuracy and PAI as the normalized Hotspot analysis, but the prospective approach was able to identify emerging hotspot areas, which were unable to be identified by the retrospective approach. Therefore, the prospective approach is promising in enhancing common crime predictions, but it is strongly dependent on the quality of the input data. Using high-quality building data would greatly increase the Risk Terrain Modelling and hence the prediction model.

5. Summary & Outlook

This research was designed to evaluate if a combination of a retrospective and prospective approach can improve the prediction accuracy of a solely retrospective analysis. A model, created in Python combined both approaches. The final risk terrain map is the result of an overlaying of the grids from the retrospective Hotspot analysis and the prospective Risk Terrain Modelling. The aggregated grid cell value represents the risk value of this area. The hotspot areas derived from the risk terrain map were used to aggregate the assault incident locations of 2014 to determine the final accuracy of the prediction model. The same method was used to identify the prediction accuracy of the solely retrospective analysis.

The retrospective Hotspot analysis is the most common method used for identifying and predicting hotspots. Based on the locations of previous incidents, future events are predicted. In this research, the assault incidents of Graz from 2009 to 2013 are used for the crime prediction. The influence of crime location on future crimes is decaying over time. For this purpose, the risk maps for each crime year are weighted based on the temporal proximity to the prediction year. This methodology is well researched and presumed to achieve a good basis for a crime prediction. Literature implies that crime predictions using retrospective analyses are restricted to areas where previous events occurred. New crime prone areas cannot be detected using incidents of the past. These regions can only be detected by analysing risk factors that indicate the presence of aggravating factors of a crime. This prospective approach was implemented using the Risk Terrain Modelling. Open Street Map data on buildings were used as basis for the analysis. The risk factor of a building was derived from its aggravating influence on assaults, identified by Rutgers' Centre of Public Security.

By combining the results of the Hotspot analysis and the Risk Terrain Modelling, the benefits of the retrospective approach were enhanced and the disadvantages minimized. 293 or 19.0% of the assaults of 2014 can be predicted using this methodology. The prediction accuracy of the solely retrospective Hotspot analysis only reached 9.7%. Therefore, the final model was able to increase the crime prediction, but it increased the size of the prediction area as well. In order to combine the Hotspot analysis and the RTM, a normalization had to be performed, resulting in an increase in area covered by the hotspots.

If the final model is compared to the normalized result of the Hotspot analysis, the prediction accuracy is 7.7% lower than in the retrospective approach. The reason why the final model cannot achieve an improvement in prediction accuracy compared to the normalized result of the Hotspot analysis, is the unsatisfying performance of the prospective approach. The Risk Terrain Modelling can only predict 4.7% of the incidents of 2014 due to the low quality of the OSM data, but it still manages to predict an emerging hotspot that is not covered by the retrospective analysis.

In the process of conducting this research, problems as well as improvements were identified that could not be considered in the analysis. The prospective approach used in this analysis is suitable for predicting crime and can aid current analyses, but it is highly dependent on the quality of the input data. It is therefore recommended for further research to acquire high-quality data on buildings with accurate and complete descriptions of the buildings' type of use. Having a higher amount of risk locations can lead to the detection of more emerging hotspot areas and may fit better with the retrospective hotspots, resulting in an increase in accuracy of the prediction model.

Computer-aided analyses had a great impact on predicting crime and elevated the efficiency of fighting crime. But the expertise of police officials or researchers is necessary to bring the information gained from these analyses to its best use. This research combined common retrospective analyses with a prospective approach to test the improvement of current prediction methods. The model created was not able to exceed the traditional approach, but gave valuable insight why the prediction model failed. Furthermore, the new approach of combining retrospective and prospective analyses can be the basis for further developments of analyses in other fields of application as well. Prospective parameters such as seasonality, daily routines and regional events might be able to improve short-term crime predictions. Combining these indicators with retrospective methods like the Hotspot analysis or the Near-Repeat Victimization may open up new possibilities on predicting crime.

6. References

ArcGIS Help 1. *How Hot Spot Analysis (Getis-Ord Gi*) works.* Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/hot-spot-analysis.htm

ArcGIS Help 2. *Slice.* Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/slice.htm

Bowers, K. J. (2004). *Prospective Hot-Spotting: The Future of Crime Mapping?* British Journal of Criminology, 44(5), 641–658.

Bundeskriminalamt Wien. (2015). Sicherheit 2015.

Caplan, J. M., & Kennedy, L. W. (2011). *Risk Terrain Modeling Compendium.* Newark, NJ.

Chainey, S., &Ratcliffe, J. (2005). *GIS and crime mapping. Mastering GIS.* Chichester, West Sussex, England, Hoboken, NJ: Wiley. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10307946

Chainey, S., Tompson, L., & Uhlig, S. (2008). *The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime.* Security Journal, 21(1-2), 4–28.

Cornell University Law School. *Assault.* Retrieved from https://www.law.cornell.edu/wex/assault

Craglia, M., Haining, R., & Wiles, P. (2000). *A Comparative Evaluation of Approaches To Urban Crime Pattern Analysis.* Urban Studies, 37(4), 711–729.

Dubois, P. F. (2007). *Python: Batteries Included.* Computing in Science Engineering, Volume 9(3).

Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., & Wilson, R. E. (2005). *Mapping Crime: Understanding Hot Spots.*

Friendly, M. (2007). A.-M. *Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis.* Statistical Science, 22(3), 368–399.

Guerry, A. M. (1833). *Essai sur la statistique morale de la France.* Crochard, Paris.

Haberman, C. P., &Ratcliffe, J. H. (2012).*The Predictive Policing Challenges of Near Repeat Armed Street Robberies.* Policing, 6(2), 151–166.

Harries, K. D. (1991). *Alternative denominators in conventional crime rates.* Environmental Criminology, 147–165.

Johnson, S. D., & Bowers, K. J. (2004).*The Burglary as Clue to the Future: The Beginnings of Prospective Hot-Spotting.* European Journal of Criminology, 1(2), 237–255.

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., & Townsley, M. (2007). *Space–Time Patterns of Risk: A Cross National Assessment of Residential Burglary Victimization.* Journal of Quantitative Criminology, 23(3), 201–219.

Johnson, S. D., Lab, S. P., & Bowers, K. J. (2008).*Stable and Fluid Hotspots of Crime: Differentiation and Identification.* Built Environment, 34(1), 32–45.

King, B. (1967). *Step-Wise Clustering Procedures.* Journal of the American Statistical Association, 62(317), 86.

Landesstatistik Steiermark. (2015). *Angezeigte Fälle der Behörden in der Steiermark.* Retrieved from http://www.statistik.steiermark.at/cms/beitrag/10888582/109801486/

Levine, N. (2006). *Crime Mapping and the Crimestat Program.* Geographical Analysis, 38(1), 41–56.

Levine, N. (2015). *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 4.02).* Washington, D.C:.

Liu, H., & Brown, D. E. (2003). *Criminal incident prediction using a point-pattern-based density model.* International Journal of Forecasting, 19(4), 603–622.

Mazerolle, L. G., & Wortley, R. K. (Eds.) (2008). *Crime science series. Environmental criminology and crime analysis.* Cullompton, UK, Portland, Or: Willan. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10294453

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations.* Santa Monica: RAND Corporation. Retrieved from http://gbv.eblib.com/patron/FullRecord.aspx?p=1437438

Piquero, A. R., &Weisburd, D. (2010). *Handbook of Quantitative Criminology.* New York, NY: Springer New York.

Porteous, D. (2003). *Crime Prevention Studies, Volume 13: Analysis for Crime Prevention.* Safer Communities, 2(1), 29–57.

Quetelet, A. (1831). *Recherches sur le penchant au crime aux differentsages.* Hayez, Brussels.

Quetelet, A. (1835). *Sur l'hommeet le d´eveloppement de sesfacult´es, ouEssaid'une physique sociale.* Bachelier, Paris.

Ratcliffe, J. H. (2006). *A Temporal Constraint Theory to Explain Opportunity-Based Spatial Offending Patterns.* Journal of Research in Crime and Delinquency, 43(3), 261–291.

Magistrat Graz - Präsidialabteilung. (2016). *Graz in Zahlen 2016.* Graz. Retrieved from http://www1.graz.at/statistik/Graz_in_Zahlen/GIZ_2016.pdf

Shaw, C., & McKay, H. (1942*). Juvenile delinquency and urban areas: A study of rates of delinquents in relation to differential characteristics of local communities in American cities.* Chicago, Ill.: The University of Chicago Press.

Shiode, S., &Shiode, N. (2014). *Microscale Prediction of Near-Future Crime Concentrations with Street-Level Geosurveillance.* Geographical Analysis, 46(4), 435–455.

Toms, S. (2015). *ArcPy and ArcGIS - geospatial analysis with Python: Use the ArcPy module to automate the analysis and mapping of geospatial data in ArcGIS.* Open source - community experience destilled. Birmingham, Mumbai: Packt Publishing.

Weisburd, D. (1998). *Crime mapping and crime prevention. Crime prevention studies: Vol. 8*. Monsey, NY: Criminal Justice Press.

Weisburd, D., & Lum, C. (2005).*The Diffusion of Computerized Crime Mapping in Policing: Linking Research and Practice.* Police Practice and Research, 6(5), 419–434.

Appendix

The supplementary material of this research is available in a Dropbox. Included are the assault incidents and the OSM building data as an ArcGIS shapefile. The link will be available until March 31, 2017.

https://www.dropbox.com/sh/gaxjt3002wv7ep1/AAB5kuZ0jRN-bN-uUbeBei53a?dl=0