Methods of segmenting and analyzing of road accident data

A dissertation submitted by:

Ma’en Qaseem Ghadi

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Supervisor: Dr. Árpád Torok

Budapest, 2019-2020
Contents

1. Introduction ................................................................................................................................. 3
   1.1 Background ............................................................................................................................ 3
   1.2 Literature review ...................................................................................................................... 4
   1.3 Motivation and novelty of the research topic .......................................................................... 10
   1.4 Objectives ............................................................................................................................... 11
   1.5 Research Methodology .......................................................................................................... 11
   1.6 Hypotheses ................................................................................................................................ 12
2. Comparison of different black spot identification methods on different road categories ........................................................... 13
   2.1 Short summary ....................................................................................................................... 13
   2.2 Introduction ........................................................................................................................... 13
   2.3 Sliding moving window with accident frequency ..................................................................... 16
   2.4 Spatial autocorrelation .......................................................................................................... 17
   2.5 Consistency test for different segmentation criteria ............................................................... 19
   2.6 Data Description ................................................................................................................... 20
   2.7 Results ..................................................................................................................................... 21
      2.7.1 Sliding moving window results ....................................................................................... 22
      2.7.2 Spatial autocorrelation results ....................................................................................... 23
   2.8 Discussion and conclusion ..................................................................................................... 26
   2.9 Thesis ..................................................................................................................................... 28
3. Developing a new spatial clustering method for road accidents ................................................................................................. 29
   3.1 Short summary ....................................................................................................................... 29
   3.2 Introduction ........................................................................................................................... 29
   3.3 Data description ..................................................................................................................... 31
   3.4 Developing of a new spatial clustering segmentation method ................................................. 31
      3.4.1 K-means clustering ........................................................................................................ 31
      3.4.2 The proposed spatial clustering segmentation method .................................................. 33
      3.4.3 Measuring the optimal number of cluster ....................................................................... 34
      3.4.4 Further segmentation (optional) ..................................................................................... 34
   3.5 Performance evaluation ......................................................................................................... 35
      3.5.1 Description of the other segmentation approaches ......................................................... 35
      3.5.2 Performance criteria (development of accident prediction models) ........................... 38
   3.6 Results and discussions ......................................................................................................... 40
   3.7 Conclusion ............................................................................................................................. 45
   3.8 Thesis ..................................................................................................................................... 46
4. A comparative analysis of black spot identification methods and road accident segmentation methods ................................................................. 47
   4.1 Short summary ....................................................................................................................... 47
   4.2 Introduction ........................................................................................................................... 47
   4.3 Data description ..................................................................................................................... 48
   4.4 Description of segmentation methods .................................................................................... 48
   4.5 Description of BSID methods ................................................................................................. 50
      4.5.1 Empirical Bayesian ......................................................................................................... 50
      4.5.2 The other BSID methods ............................................................................................... 51
   4.6 Performance evaluation tests ................................................................................................ 52
      4.6.1 Site consistency test ....................................................................................................... 52
      4.6.2 Method consistency test ................................................................................................ 52
4.6.3 Total rank differences test ............................................. 53
4.6.4 Total score test.............................................................. 53
4.7 Results ........................................................................... 54
4.7.1 Result of the SPF development process ............................ 55
4.7.2 Consistency tests results ................................................. 56
4.8 Discussion of the results .................................................. 60
4.9 Conclusion ................................................................. 62
4.10 Thesis ........................................................................... 63
5. Spatial clustering segmentation approach: practical applications ... 64
5.1 Short summary ............................................................. 64
5.2 Introduction .................................................................. 65
Part one: ............................................................................... 66
Integration of probability and clustering based approaches in the field of black
spot identification .................................................................. 66
5.3 Data description ............................................................ 66
5.4 Methodology .................................................................. 67
5.5 Results and discussion .................................................... 67
5.6 Summary of part one ....................................................... 69
Part Two: ............................................................................... 70
Introducing a new qualitative-spatial approach to explain road accidents ...... 70
5.7 Decision analysis ........................................................... 70
5.8 Data description ............................................................ 71
5.9 Methodology .................................................................. 71
5.10 Results and discussions .................................................. 73
5.10.1 Road segmentation results ........................................... 73
5.10.2 Black spots identification results ................................... 73
5.10.3 Extracting decision rules from decision tree ................. 74
5.11 Summary of part Two ...................................................... 78
5.12 Thesis ........................................................................... 79
6. Multilevel analysis of road accident frequency: the impact of the road category 80
6.1 Short summary ............................................................. 80
6.2 Introduction .................................................................. 81
6.3 Data Description ............................................................ 82
6.4 Methodology .................................................................. 84
6.1.1 Level-one modelling ................................................ 85
6.1.2 Level-two modelling (Developing multilevel model) ......... 86
6.5 Results ........................................................................... 87
6.6 Conclusion .................................................................... 91
6.7 Thesis ........................................................................... 92
7. Overall conclusions and scopes for the future studies .................... 93
7.1 Short summary ............................................................. 93
7.2 New scientific results ...................................................... 94
7.3 Application of the scientific results ................................... 101
7.4 Scope for future work ..................................................... 103
List of Figures

Figure 2.1 Sliding moving window mechanisms .................................................. 16
Figure 2.2 Different indications of spatial autocorrelation values .......................... 18
Figure 2.3 Illustration of the consistency measurement for three different segment criteria .......................................................... 20
Figure 2.4 BS density per section-km of the (a) Motorway (b) Urban roads, measured for W1, W2, and W3 .................................................. 22
Figure 2.5 Global Spatial Autocorrelation Report (by the ArcGIS) resulted by each polygon length (W1, W2, and W3) for the (a) Motorway (b) Urban roads .......................................................... 24
Figure 2.6 BS density per section-km of the (a) Motorway (b) Urban roads, measured for W1, W2, and W3 .................................................. 25

Figure 3.1 The main process of the k-means clustering algorithm ....................... 33
Figure 3.2 A visual illustration of the developed spatial clustering methodology .... 34
Figure 3.3 Graph the linear correlation between the observed accidents (x-axis) and the predicted accidents (y-axis) of the year 2016 ......................... 44

Figure 5.1 Defining the value of the optimal number of clusters based on the changes in the SSE (variance) .............................................................................. 68
Figure 5.2 Proposed framework for analysing the spatial and causal pattern of road accidents. L is the length of a segment n. N is the number of accident for each road segments RSG .......................................................... 72
Figure 5.3 The output of CART tree ..................................................................... 76

Figure 6.1 Multilevel hierarchical structure of road accident data ..................... 85
Figure 6.2 Estimated intercept for different roadways from the multilevel analysis. Where, LR= local road, SR= Secondary road, AR= Arterial, ER= Expressway, and MR= Motorway .......................................................... 88

Figure 7.1 A Illustration of the consistency measurement for three different segment criteria (W1, W2, W3) .............................................................................. 95
Figure 7.2 A visual illustration of the developed spatial clustering methodology. .... 96
Figure 7.3 Linear correlation between the observed accidents (x-axis) and the predicted accidents (y-axis) of the year 2016 .................................................. 97
Figure 7.4 The proposed framework for analyzing accident spatial distribution and patterns ................................................................................. 99
Figure 7.5 Estimated intercept for different roadways from the multilevel analysis. Where, LR= local road, SR= Secondary road, AR= Arterial, ER= Expressway, and MR= Motorway .................................................. 100
List of Tables

Table 2. 1 A review of some popular BSID techniques .......................................................... 14
Table 2. 2 Descriptive statistics of the segmentation characteristics of the motorway and urban roads ........................................................................................................ 21
Table 2. 3 Sliding window consistency test measured for different window lengths .. 23
Table 2. 4 Global Moran's I Summary ....................................................................................... 24
Table 2. 5 Spatial autocorrelation consistenct test measured for different polygon lengths ...................................................................................................................... 26

Table 3. 1 Description of the data used in developing the models ........................................ 31
Table 3. 2 Descriptive statistics of length and accident numbers per segment ................. 37
Table 3. 3 Correlations coefficient parameter of cluster segments .................................. 39
Table 3. 4 Summary of k-means segmentation results by road reference ..................... 41
Table 3. 5 Values of the model parameters, (p-value), QIC, PCC and over-dispersion (k) for: (a) model 1 with one explanatory variable; AADT, b) Model 2 with three explanatory variables; AADT, speed, HDA ........................................ 41

Table 4. 1 Descriptive statistics of segmentation results ..................................................... 54
Table 4. 2 Estimated values of SPF parameters and goodness-of-fit. ............................... 55
Table 4. 3 Site consistency test results ............................................................................... 57
Table 4. 4 Method consistency test results ....................................................................... 58
Table 4. 5 Total differences test results .............................................................................. 59
Table 4. 6 Total score test results ....................................................................................... 60

Table 5. 1 Cluster information and the final BS identification process ............................. 69
Table 5. 2 Descriptive statistics of the resulted spatial segmentation processes .......... 73
Table 5. 3 Model parameters and goodness-of-fit results ................................................ 74
Table 5. 4 Description of accident attributes per segment’s groups risk level ............... 74

Table 6. 1 Descriptive statistics of the resulted segments per road category ............... 83
Table 6. 2 Variance values of the random effect components ....................................... 88
Table 6. 3 Comparing the goodness-of-fit of a single level and multilevel models ... 89
Table 6. 4 Resulted fixed effect coefficient values ......................................................... 89

Table 7. 1 Resulted parameters and goodness-of-fit values of the five road segmentation methods .................................................................................................................. 96
Table 7. 2 Total score test results for top 5% and 10% BS ............................................. 98
Table 7. 3 Resulted fixed effect coefficient values ......................................................... 101
### List of abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>Average Annual Daily Traffic</td>
</tr>
<tr>
<td>AF</td>
<td>Accident frequency</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>AR</td>
<td>Accident rate</td>
</tr>
<tr>
<td>AR</td>
<td>Arterial Road</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BS</td>
<td>Black Spot</td>
</tr>
<tr>
<td>BS1, BS2, BS3, BS4</td>
<td>Black Spots (ranked from higher risk (BS1) to lower risk (BS4) in ascending order)</td>
</tr>
<tr>
<td>BSID</td>
<td>Black Spot Identification</td>
</tr>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>DR</td>
<td>Decision Rule</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>EB</td>
<td>Empirical Bayes</td>
</tr>
<tr>
<td>EEB</td>
<td>Excess Empirical Bayesian method</td>
</tr>
<tr>
<td>ER</td>
<td>Expressway</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross Domestic Products</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographical Information Systems</td>
</tr>
<tr>
<td>HDA</td>
<td>Horizontal deflection angle</td>
</tr>
<tr>
<td>HSM</td>
<td>Highway Safety Manual</td>
</tr>
<tr>
<td>I</td>
<td>Global Moran's Index</td>
</tr>
<tr>
<td>I_i</td>
<td>Local Moran's Index</td>
</tr>
<tr>
<td>K</td>
<td>The number of clusters (generated by the K-means method)</td>
</tr>
<tr>
<td>LR</td>
<td>Local road</td>
</tr>
<tr>
<td>M3</td>
<td>Hungarian Motorway number 3</td>
</tr>
<tr>
<td>MCT</td>
<td>Method Consistency Test</td>
</tr>
<tr>
<td>MR</td>
<td>Motorway</td>
</tr>
<tr>
<td>N_E</td>
<td>Expected Number of Accidents</td>
</tr>
<tr>
<td>N_o</td>
<td>Observed Number of Accidents</td>
</tr>
<tr>
<td>N_P</td>
<td>Predicted Number of Accidents</td>
</tr>
<tr>
<td>PCC</td>
<td>Pearson Correlation Coefficient</td>
</tr>
<tr>
<td>QIC</td>
<td>Quasi-likelihood under Independence Model Criterion</td>
</tr>
<tr>
<td>RP</td>
<td>Reference Point</td>
</tr>
<tr>
<td>RSG</td>
<td>Road Segment</td>
</tr>
<tr>
<td>RSI</td>
<td>Road Safety Inspection</td>
</tr>
<tr>
<td>RTM</td>
<td>Regression-to-the-Mean</td>
</tr>
<tr>
<td>SC1, SC2, SC3</td>
<td>The identified spatial cluster number 1, 2, or 3 (Figure 3.2)</td>
</tr>
<tr>
<td>SCL</td>
<td>Spatial Clustering</td>
</tr>
<tr>
<td>SCT</td>
<td>Site Consistency Test</td>
</tr>
<tr>
<td>Seg-1</td>
<td>Spatial Clustering Segmentation Method</td>
</tr>
<tr>
<td>Seg-2</td>
<td>HSM based Segmentation Method</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Seg-3</td>
<td>Constant AADT Segmentation Method</td>
</tr>
<tr>
<td>Seg-4</td>
<td>Constant length Segmentation Method</td>
</tr>
<tr>
<td>Seg-5</td>
<td>Curve based Segmentation Method</td>
</tr>
<tr>
<td>SMW</td>
<td>Sliding Moving Window</td>
</tr>
<tr>
<td>SPA</td>
<td>Spatial Autocorrelation</td>
</tr>
<tr>
<td>SPF</td>
<td>Safety Performance Functions</td>
</tr>
<tr>
<td>SR</td>
<td>Secondary Road</td>
</tr>
<tr>
<td>SSE</td>
<td>Sum of Squared Error</td>
</tr>
<tr>
<td>TRDT</td>
<td>Total Rank Differences Test</td>
</tr>
<tr>
<td>TST</td>
<td>Total Score Test</td>
</tr>
<tr>
<td>W1, W2, W3</td>
<td>Road segmentation methods with different criteria (Figure 2.3)</td>
</tr>
<tr>
<td>W12,W13,W23,W123</td>
<td>Intersection areas between W1, W2, and W3 (Figure 2.3)</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

Today the most negative effects of building and developing transportation systems are related to road accidents. According to the definition of the Hungarian Central Statistical Office [1], a traffic accident is, an unexpected, unintentionally occurring traffic accident that has resulted in death or personal injury or property damage. Accidents can be classified by the outcome as follows:

- Fatal accident where at least one person died at the scene of the accident or within 30 days;
- Serious injury means an accident which has resulted in at least one person being seriously injured within 8 days;
- Minor injury is an accident that has resulted in at least one person being healed within eight days.

Implementation of safety measures are costly and restricted funds set limitations on the number of operative interventions. Therefore, it is necessary to define a priority based hierarchy of the high-risk sites and their related safety measures in order to utilize the limited resources as effectively as possible.

The identification of hazardous road sections is one of the main parts of any successful road safety management process. The identification of hazardous road sections, high-risk accident locations or black spots (BS) receives a great interest from road agencies and safety specialists. Black spot identification (BSID) can be defined as the process of searching for locations in the transportation network which can be characterized by higher accident risk than other similar locations. Beside this, in many cases the increased accident risk is caused by local risk factors [2], [3]. Errors in the BSID process can result in false positive (a site is involved in safety investigation while it is not needed) and false negative (a site is not involved in safety investigation while it is needed) cases. In other words, accidents can occur in both safe and unsafe locations, and one of the most challenging task of the BSID process is to avoid false positive and negative cases in identifying the most critical locations. The differentiation of locally occurring black spots and safety issues (e.g. risky sections) occurring on a network level is a very important scientific objective, which is still
needed to be clarified by the scientific society. However, the investigation of the mentioned field is out of the scope the current research. Accordingly, irrespectively of the length of the investigated risky sections, they have been identified as black spots.

Considering that traffic accident data is heterogeneous, in general, data segmentation is considered to be the first and most important step of the BSID process. The success of any method applied in identifying BS on roads should depend fundamentally on how the data is organized into specific homogeneous segments. In road safety, road accident segmentation is the process of organizing road infrastructure network data into homogenous entities. Data segmentation can also be defined as a process of dividing and classifying a large and complex dataset into small and simple homogeneous groups or entities in which data within a group shall be similar but data between groups shall rather be dissimilar. The more accurate the accident data is segmented, the more accurate the accident prediction model will be, and this will consequently affect the performance of the applied BSID methods that rely on accident prediction models in their criteria.

The effect of road segmentation on BSID methods is still an inadequately explored area. This has led to unobserved deficiencies in many of the developed BSID methods, demonstrated by the referred variation in their performance for different applied case studies. Moreover, the previously performed studies have also differed in the applied methods related to the identification of the homogeneous road segments based on different applied variables. Segmentation variables mostly include traffic features (traffic volume, light vehicle and truck volume, speed) and/or road geometrical characteristics (number of lanes, horizontal curves, vertical slope, roadside hazard). Segmentation methods have just rarely considered the spatial factor of accidents distribution. Neighboring accidents occurred relatively close to each other spatially and temporally are more likely to share similar properties. Clustering of spatially related accidents can be a useful technique to find hidden relationships and patterns related to road accidents and divide them into small groups. Generally, the appropriate combination of black spot identification (BSID) method and segmentation method can greatly contribute to the reduction in false positive and false negative cases in identifying BS segments.

This dissertation aims to suggest a new methodological framework for BSID; based on the development and evaluation of new segmentation method focusing on road accident data. The dissertation also aims to develop a new comprehensive multilevel model for predicting accident frequency for different road categories related to different road segment characteristics.

### 1.2 Literature review

Traffic accidents cause large social and economic losses to countries, families and individuals. Every year about 1.35 million people dies by road accidents around the world, and the external annual cost of the accidents is about US $ 518 billion (1-3% of
GDP) [4], [5]. In Europe alone, road accidents cost approximately US $220 billion million in 2017 [6]. The identification of hazardous road sections is one of the main parts of any successful road safety management process. Over the years researches have been widely performed in the field of identifying “black spot”, “site with promise”, “high risk” or “hotspot” identification [7]. The spread and multiplication of BSID methods has resulted in different BS definitions. Different applied definitions can significantly affect the result, in some cases even the accuracy and the performance of the applied method. Elvik [8] used the expected number of accidents concept and defined four categories of sites as follows (ref):

Correct positive: \( \text{if } E \geq c \text{ and } R \geq c \)

False positive: \( \text{if } E < c \text{ and } R \geq c \)

Correct negative: \( \text{if } E < c \text{ and } R < c \)

False negative: \( \text{if } E \geq c \text{ and } R < c \)

Where: \( E \) is the expected number of accidents, \( R \) is the observed accidents, and \( c \) denotes a selected critical value.

Correct and false are defined based on the comparison of the expected number of accidents and selected critical value. While Positive and negative are defined based on the comparison between the observe number of accidents and the critical value.

Generally, the accuracy of any BSID method depends on three main factors: the applied methodology (i.e. statistical modeling), segmentation method, and the applied predictor variables. Accordingly, it is necessary to establish an integrated methodological approach which can facilitate segmenting and identifying BS in a highly efficient way. This section chronologically reviews the evolution of BSID methods and all related issues regarding road segmentation, statistical modeling, and applications.

In literature, diversity of definitions has explored and described the meaning of road accident BS. Hakkert and Mahalel [9] defined BS as a site that has higher accident frequency than a specific prescribed expected level. Mahalel et al. [10] assumed BS as a site which has a maximum expected reduction in its total accidents by treatment compared to other similar sites. Universally, there is no one accepted BS definition, but generally BSID has a main objective of searching for specific locations of the transportation network which can be characterized by higher accident numbers than other similar locations mainly caused by local risk factors [2], [3]. In accordance with this, accidents can occur in both safe and unsafe locations, and the challenge is to avoid false positive and negative cases in identifying the most dangerous locations. Therefore, the correct identification of accident BS is essential to provide efficient allocation of the available sources and maintain a sustainable road safety system [11], [12].
Elvik (2008a) [8] investigated and compared those BSID methods which are applied the most frequently in practice in Europe. Furthermore he has defined the following list of the five most commonly used ones:

1. Number of accident count (accidents per year or accident per km),
2. Accident rate (accidents per vehicle per kilometers or per entering vehicles),
3. Number of accidents and high accident rate,
4. Expected number of accident, and
5. Empirical Bayes (EB) dispersion criterion.

As it has been presented many approaches have been suggested to analyze road traffic safety level based on accident rates and frequency. For instance, Virtisen [13] has ranked BS locations according to their reported number of accidents. Jorgensen [14] has selected those accident locations which have higher accident frequency than a certain threshold value. Foldvary [15] used simple models based on the mean and variance values to study variations in accident rates.

As it can be observed some of the applied methods and other similar researches [16]–[18] are not necessarily based on statistical models, while other investigations have been purely based on safety performance functions (SPF) [19]–[23] to predict accident frequency. As it has been proved, this way of prediction reflects usually to an unrealistic assumption that factors affecting safety can remain the same by the passage of time. On the other hand, the result of the SPF might be affected by the natural fluctuation in accident frequencies over time for any given location, such phenomena are known as regression-to-the-mean (RTM) bias (see Davis (2008) [24]). RTM bias occurs when sites are selected for treatment based on long-term trends of observed crash frequency, which can result in a biased selection of sites that need treatment. For instance, accidents may be affected due to the novelty of certain road improvements but after a while they get used to the intervention and the effect size will set to a normal level. To overcome the regression-to-the-mean effect, Jorgensen introduced a new method for measuring the expected number of accident calculated by combining the benefit of the observed and predicted accident numbers. As a result, the expected number of accidents using the Empirical Bayesian (EB) has been applied as a new parameter in defining BS.

The EB method combines the benefit of observed and predicted accident frequencies. The two values are weighted in a statistical model based on the reliability level of the predicted value which is derived from an SPF. Higle and Witkowski [25] used the EB approach and have focused on the identification of road sections with unusually large accident rates. These authors have proposed that a road section can be identified hazardous if its accident rate exceeds a certain value. Hauer et al. [26] also developed an EB technique to evaluate safety performance of road infrastructure components, however they have focused on signalized intersections.
A minor improvement has been suggested for the EB method to give it the ability to identify and rank BS according to the potential for accident reduction. McGuigan [27], [28] suggested BS ranking according to their accident reduction potential, which can be defined based on the difference between the observed number of accidents and the expected value of accident counts at sites described by the same investigation parameters value. Nguyen et al. [29] applied an analytical framework for identifying BS based on potential crash cost saving. The method searches for BS where safety improvement measures are expected to have the greatest economic effectiveness. Generally, the applied analytical models aim to represent the relationship of the known information and characteristics of accidents to understand how accidents happen.

In most cases, researchers focus only on the method of BSID itself, neglecting the effect of data organization and preparation, as an integral part of BSID process. Considering that traffic accident data are heterogeneous, in general, data segmentation is considered to be the first and outstandingly important step of BSID process. Thomas [30] has argued that applying different lengths for segmenting road network can result in different definitions of hazardous locations which, in turn, affects the reliability of results. Koorey [31] has discussed the benefit of applying variable length segments and their effect on BS determination. It has been confirmed as well that selecting the appropriate segmentation method can have a significant effect on reducing false positive and negative cases during the determination of BS segments. Moreover, road segmentation has a significant effect on the validity of the developed SPF. The more accurately the road network data is segmented, the more accurate the SPF will be [32], and according to our assumptions this consequently affects the performance of BSID methods that directly use the results of accident prediction models.

Most often, road segmentation methods are based on researchers’ experiences, methodological decisions or intuitions. Cafiso et al. [32] evaluated the change of SPF effectiveness based on five different segmentation methods considering geometric and/or traffic-related attributes. The findings have revealed that SPF effectiveness can be characterized by the highest value in case of design parameters (i.e. curvature characteristics) based road segmentation since the resulting set of high-risk sections have seemed to be well-correlated with the set of locations characterized by high accident density. Cafiso et al. [33] suggested that segment lengths should be related to the Annual Average Daily Traffic (AADT) to increase the performance of BSID methods, where segments with lower AADT values should be longer.

According to the Highway Safety Manual (HSM) [20], roadway segments should be characterized by homogenous cross-sections not smaller than 0.1 mile, and their endpoints should be identified based on the changes in AADT values or in other roadway features. In practice, it is not always easy to implement this type of segmentation as not all the variables are available [34]. Other researchers [35] used constant length segments. Constant length segmentation can result in a different BS
set depending on the selected length and starting point of a segment [30]. Generally, in case of such a segmentation method, which aims to generate homogenous sections, too many segmentation variables can result in very short average segment lengths. This can be disadvantageous due to the too many zero-accident sections, which can consequently affect the accuracy of the accident prediction models [36]. In contrast, increasing segment length can sacrifice the homogeneity.

On the other hand, a few studies have considered spatial distribution of accidents as an appropriate indicator for homogeneity. Spatial distribution based road accident segmentation can provide a better understanding of the patterns and processes that cause the accidents [37], [38]. Benoit et al. [39] tried to understand empirically the spatial occurrence of road accidents by comparing the two well-known methods in this area; kernel and spatial autocorrelation methods. Finally the outcomes of the research have provided almost the same efficiency in both cases for the two analysed methods. Anderson [40] used Geographical Information Systems (GIS) and Kernel Density Estimation to study the spatial patterns of injury-related accidents in London. This author has classified individual accidents into a smaller number of more extended groups in a hierarchical procedure. He has justified; this framework could lead road safety professionals to a better understanding, not only the types of BS but their spatial patterns as well. The recent development of the GIS tools has allowed more extensive spatial analysis for a larger dataset of accidents [41]–[44]. Long Tien and Sekhar [45] applied GIS software to identify pedestrian-vehicle crashes near bus stops in a spatial model. Generally, based on the investigated studies it can be assumed that accidents located closer to each other in space and time are more homogeneous than single accidents located separately [39], [46]. Since accidents located closer to each other are more likely to occur under similar circumstances.

Usually, in case of those scientific fields where "spatial segmentation" is applied, the concept of “clustering” appears. Clustering techniques have begun to attract more and more attention in the field of accident data segmentation [47]–[49]. Clustering can be a useful technique to find hidden relationships and patterns for large datasets of accidents. Karlaftis and Tarko [50] applied cluster analysis to categorize accidents into different groups and in the next step they have analyzed the results of the clustering process by assuming a negative binomial distribution to study the impact of driver age on road accidents. Jianming Ma and Kockelman [51] used clustering to group accident data into different segments and finally have identified relationships between different accident characteristics. Ng et al. (2002) [52] used cluster analysis to group homogeneous accident data aimed to develop an algorithm to estimate accident frequency on roads and to investigate the risk distribution of accidents. Yannis G. et al [53] applied clustering to explore the change of accident risk related to driving under influence of alcohol. A latent cluster analysis has been performed by Depaire et al. [47].

Information generated by the segmentation process can assist road safety professionals to enhance estimation accuracy of locations characterized by high
accident density and also to understand accident patterns along the road. In this regard, data mining techniques can be useful tools to investigate road accident patterns [54]–[57]. Data mining is a mostly heuristic process aiming to discover hidden connections from a large dataset. Its main goal is to collect useful information from a large heterogeneous dataset after being classified into homogeneous structures. Accordingly, data mining techniques can be used to segment and investigate road accident data. Decision analysis is one of the popular supervised data mining techniques that has been widely applied in the field of road safety [58]–[60]. For example, Kumar and Toshniwal [61] applied K-means clustering and the association rule of mining to analyse patterns of accidents that occurred under different circumstances in India. The findings have revealed that the combination of k mode clustering and association rule mining can be very useful in road safety investigations. Abellán et al. [58] applied a methodology to define decision rule (hereinafter DR) from more decision trees (hereinafter DT) to enable more effective interaction between accident attributes, focusing on accident severity as a label variable. Kashani et al. [62] applied the classification DT to identify the important factors influencing injury severity of drivers involved in traffic accidents in Iran.

From a methodological aspect, a wide range of models has been applied to investigate accident patterns and their main contributing factors. For example, Guerrero Barbosa et al. [63] studied the influence of factors related to roadway geometry, traffic volumes and speeds on accident frequency in case of an urban road network. Joly et al. [64] investigated the influence of geographical and socio-ecological variations on pedestrian and cyclist accidents for different spatial locations. Generally, the previous studies have explored traffic safety at either micro- or macro-level, separately.

At micro-level, road accident analysis has included detailed factors related to one or a small group of accidents, such as; road geometric features [65], [66], traffic characteristics [67], or weather and environment conditions [68], [69]. Most of these studies have been performed for small spots or road segments. Contrary to this, the factors applied at macro-level have related to a large group of accidents usually on a regional level, such as; road category [70], socio-economic characteristics including population, buildings density [71], or travelled vehicle mile [72]. However, the selection of a suitable analysis level depends on the purpose of the investigation and the type of available data.

On the other hand, most of the reviewed accident prediction models in have been limited to a certain road category or intersection. However Castro et al. [73] applied a latent variable generalized ordered response framework for modelling accident count at urban intersections. The results have revealed some important factors influencing accident propensity at intersections, (e.g. road parameters, functional type of the connecting roads, and total daily entering traffic volumes). Persaud & Mucsi [74] used hourly traffic volumes in a regression model for estimating accident potential locations on two-lane rural roads. B. Persaud & Dzbik [75] developed a generalized linear model to study the change of accident frequency and severity on freeway. Most
of these studies have been limited and dedicated for a specific road type or intersection. Differences in the characteristics of different roads make it difficult to construct a general model for all road categories. Ghadi & Török [76] compared the performance of some black spot identification methods between highway and secondary roads. They found that the performance of the applied methods can significantly be changed by the road category, and speed factor can have a significant impact on the results. However, failure to account for the hierarchical nature of accident data can result in underestimating of accident prediction model parameters.

Multilevel models are specifically developed to analyse data with hierarchical structure. Road safety related multilevel analysis aims to model the relationship between different accidents groups by identifying the hierarchical system of the data, which can take the advantages of the clustered dataset [77], [78]. This means that the outcome of the model is affected by a nested relationship between the lower level characteristics (level-one) of individual accidents and higher level groups characteristics (level-two). Lenguerrand et al. [79] investigated the advantages of applying a multilevel logistic model to analyse the hierarchical structure of accident data compared to traditional generalized estimating model and logistic model. They have found that the multilevel analysis provides a more efficient model than the traditional models. Haghighi et al. [66] explored the nested relationship between individual crash characteristics and environment and roadway features. They have applied a multilevel ordinal logistic regression to analyse the hierarchical structure of accident data and its impact on accident severity outcome. Only a few researches have investigated the applicability of multilevel analysis in the field of accident analysis. Cai et al. [80] developed a Bayesian integrated spatial model, to analyse accident frequency at macro- and micro-levels also considering districts and road components (i.e. segments and intersections), simultaneously. The results have indicated that the model can simultaneously identify both micro- and macro-level factors contributing to accident occurrence.

At this point, it is preferable to mention that “accident” and “crash” expressions are used to refer to the same meaning in this dissertation.

1.3 Motivation and novelty of the research topic

In the reviewed literature a wide range of road safety strategies has focused on BSID methods [3], [7], [81]. The diversity of BSID methods has resulted in different definitions of BS which influences the accuracy and performance of the applied methods. In most cases, researchers focus only on the BSID model itself, neglecting the effect of data organization and preparation, as an integral part of BSID process. Since traffic accident data are considered to be heterogeneous in general, data segmentation is the first and most important step of the BSID.

Accordingly, it seems to be reasonable to establish an overall methodological approach which can facilitate segmenting and identifying BS in an integrated and
consistent way. The appropriate combination of BSID method and segmentation method can significantly contribute to the reduction in the number of false positive and false negative cases related to the identified BS segments.

Until now the field of accident segmentation has gained little attention from road safety experts. Furthermore, the effect of segmentation on the performance of BSID methods and SPFs is still under researched. This dissertation introduces a new methodology for BSID based on development and evaluation of novel network data segmentation especially considering the spatial characteristic of traffic accident data. Furthermore, the dissertation also presents a new comprehensive multilevel model for predicting accident frequency in case of different road categories related to the variation of road segment characteristics. The aim of the newly developed methodology is to reduce the number of false positive and negative cases related to BSID method.

1.4 Objectives

The main objectives of the dissertation are summarized as follows.

- Reviewing the relevant BSID methods and identify key factors that can lead to variation in their performance by different road categories and different segmentation characteristics.
- Development and evaluation of a novel network segmentation method based on road accident data to assist in reducing the number of false positive and negative cases during BSID process.
- Studying and comparing the effect of the methodological diversity of road network segmentation on the performance of different BSID methods.
- Identifying the most efficient combinations of segmentation methods and BSID methods.
- Investigating the practical application of combining the newly developed segmentation method with the EB method to determine the spatial locations of BSs.
- Investigating the practical application of combining the newly developed segmentation method with the data mining techniques to determine the patterns of the spatially distributed road accidents.
- Modelling risk variability between different road segments on different road categories at micro- and macro-level.

1.5 Research Methodology

A variety of statistical tools and software have been used for different purposes.
Spatial statistics tools of the ArcGIS have been used to find the spatial autocorrelation of accidents.

ArcGIS toolboxes, including Geo-processing and System tools, have been used to support the spatial segmentation and arrangement of accident datasets.

K-means clustering and linear referencing techniques have been used to contribute to spatial segmentation of accident dataset.

Safety performance function has been used to predict accident frequency.

Quasi-likelihood, Pearson Correlation, Akiake’s information criterion, and other statistics have been used to examine and compare the goodness-of-fit of the developed models.

Decision tree and decision rules have been used to interpret accident patterns.

R software package, IBM SPSS, Microsoft Excel, Microsoft Access and Weka software have been used to support the statistical analysis.

1.6 Hypotheses

The application of the same BSID method and/or segmentation method for different road categories can result in significantly different outcomes.

The efficiency of BSID process depends fundamentally on how data being organized into homogeneous segments.

The application of a suitable combination of BSID methods and segmentation methods can result in the reduction of wrongly identified BS locations.

The homogeneity of the generated spatial accident groups can strongly influence the reliability of the developed SPF.

Clustering accident data based on the spatial factor can reduce the heterogeneity of accident dataset.

A hierarchically structured prediction model can reveal the hidden multilayer relationships between accident data, road segments and higher level entities like road categories or regions through the variation in accident predictors’ characteristics.
Chapter 2

Comparison of different black spot identification methods on different road categories

2.1 Short summary

The identification of road sections characterized by high-risk accidents is the first step for any successful road safety management process, considering the limited available resources. Comparisons of different BSID methods are still not sufficiently explored. This chapter aims to measure and compare the performance of two popular BSID methods; Sliding Moving Window and Spatial Autocorrelation, when applied on two different road categories; Motorway and Urban roads. The chapter also aimed to highlight the effect of data organization and segmentation on the performance of different BSID methods. In the evaluation process, a consistency test has been used in order to measure and compare the performance of the applied BSID method in identifying the same BS with each different applied conditions. The result reveals that the performance of the two applied BSID methods is significantly affected by two factors: the characteristics of the roadway where the analysis is performed (i.e. road category) and the characteristics of the method which is applied to divide the road network into segments and to classify accidents into small groups (i.e. segmentation method). Based on the results it found that the outcome of the applied BSID is significantly affected by both the segmentation criteria and the road category. This may put some question marks about the reliability in applying the best BSID method that suits enough the conditions of the case study in which it is applied.

2.2 Introduction

In road safety, searching for high-risk sites or BS is a well-known topic. Contrary to the achieved results, important parts of this field still need to be revealed. The identification of road sections characterized by high accident risk is the first step for any successful road safety management process, considering the limited available resources.
In the identification process of BS three main methods can be used: screening methods, clustering methods and crash prediction methods. Many literatures and case studies have been written describing the pros or cons of each method [7], [8], [82]. Table 2.1 reviews the methodologies, advantages, and disadvantages of some methods applied in this field.

<table>
<thead>
<tr>
<th>BSID Method</th>
<th>Methodology</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident count</td>
<td>Sites are ranked at first according to their reported number of accidents</td>
<td>• Simplicity</td>
<td>• This method is very sensitive to random variation in accident counts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Does not account for regression to the mean bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Does not account for traffic volume</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Does not estimate a threshold</td>
</tr>
<tr>
<td>Accident rate</td>
<td>Similar to the accident count method but considers traffic volume or entering vehicle volume as well</td>
<td>• Could be modified to account for severity</td>
<td>• Does not account for regression to the mean bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Consider the AADT in the analysis</td>
<td>• Does not identify a threshold</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Will mistakenly prioritize low volume, low collision sites</td>
</tr>
<tr>
<td>Empirical Bayes</td>
<td>Both the observed and predicted accident frequencies are weighted to calculate an expected average crash frequency</td>
<td>• Account for regression to the mean</td>
<td>• Requires SPFs calibrated to local conditions</td>
</tr>
<tr>
<td>Excess Empirical Bayes (safety potential)</td>
<td>calculate the difference between the estimated EB value and the predicted accident frequency to rank BS sites according to their potential for safety improvements.</td>
<td>• Account for regression to the mean</td>
<td>• Requires SPFs calibrated to local conditions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Identifies a threshold</td>
<td>• largely affected by the validity of the developed SPF</td>
</tr>
</tbody>
</table>
Sliding window

- A window with fixed length moves along the roadway with small increments to satisfy a pre-specified hot spot selection criterion (threshold)
- Easy to understand
- Easy to develop
- The selection of BS is based on first come first serve without overlapping rule
- Identify black spots without further classification
- Each black spot has an equal length

Spatial Autocorrelation

- Measuring the degree of association (co-variation) between approximate spatial units (crashes) using the Maron's index I to determine accident clusters
- Flexible length black spots
- General road design and traffic volume are not taken into consideration
- The efficiency of the method depends on choosing the right scale

Preliminary Road Safety Inspection (RSI)

- A risk index is calculated by a group of road experts using driving vehicle method to assess the road conditions (pavement condition, road side, horizontal and vertical curve, etc).
- Flexible black spot length.
- Proactive accident prevention method
- Can be done without road and traffic data
- The method is highly depends on inspectors judge (different assessment from different inspectors)
- Only focus on site without real accidents data

The methods described in Table 2.1 can be divided into performance measure (i.e. accident count, accident rate, Empirical Bayes, safety potential), screening methods (i.e. sliding window), clustering method (i.e. spatial autocorrelation), and proactive method (i.e. RSI).

Generally, the literatures are mostly focused on the methodology of the BS identification itself, disregarding the other related techniques focusing on data organization and preparation. Moreover, different BSID methods with different road categories or with other different influencing parameters can result in different outcomes especially considering the success and precision of the applied method. Therefore, firstly the most important question has to be answered is that which method should be applied for which road. This question can be answered by comparing different applied methods for different road categories. However the comparison of different BSID methods is still not an adequately explored area.

In the reviewed literatures it has been claimed that changing segment lengths applied by the sliding moving window (SMW) method can influence significantly the BS searching process, and consequently could result in many false positive and false negative. Therefore, different clustering techniques have been developed (i.e. the
spatial autocorrelation - SPA) in order to tackle the problem and to apply an adaptable BS length. However, using clustering techniques can be more useful in adapting the length of the investigated BS area to the certain accidents. Contrary to this, the use of SMW with a fixed window length, for a low-speed road type could have some deficiencies. For instance, in such a case, where a relevant part of the identified BS does not include any collision and the analysis still result in partially false positive. On the other hand, changing the window characteristics (i.e. window length, sliding distance) can give different outcomes [83].

In accordance with this, the chapter investigates the pros and cons of two BSID methods (i.e. SMW and SPA) on two different road categories using a consistency test method. The investigation is based on the assumption that the differences in road characteristics may affect the spatial form of accident distributions along the road. This could affect the reliability of the applied BSID technique.

2.3 Sliding moving window with accident frequency

Network screening is a well-popular segmentation approach. According to the HSM [20] there are three main methods of network screening: sliding moving window, peak searching, and simple ranking methods. The sliding moving window (SMW) is a commonly used dynamic segmentation method which is designed to slide with a fixed length window on a homogeneous road segment, in terms of design parameters and built-in environment attributes, aiming to identify the appropriate start and end points of BS locations along a roadway and ranking them by severity.

In the general methodology, the SMW method requires the user to input the window length and one of the BSID method (i.e. accident frequency, accident rate, EB, etc.) to identify critical road sections. Then, the window with the specific length moves incrementally along the road, starting from a predefined point to the end of a road. The performance measure for selecting critical road sections is applied to each window after each step (sliding), and the results of the analysis are recorded for each window. When the criteria are met for any sliding window, a BS location is identified, and the search for another BS is continued from the next segment without overlapping, as described in Figure 2.1.

![Figure 2.1 Sliding moving window mechanisms](image-url)
In this section, the accident frequency method has been investigated to evaluate the safety characteristics of each window. To define the threshold value, a fixed length scanned window is moved along every category of road segments that have similar environment properties, separately. The number of accidents in every SMW is tabulated. Then, the window is moved again along the same length of the road considering any location \((i)\) to be unsafe if the number of accidents exceeds the threshold value, as represented in the following Equation (2.1):

\[
x_i > \bar{x} + \frac{CI \times SD}{n}
\]

(2.1)

Where: \(x_i\) is the number of accidents at location \((i)\). \(\bar{x}\) is the average number of accidents for all locations. \(CI\) is a confidence interval. \(SD\) is the standard deviation. And, \(n\) is the total number of all SMW. The HSM recommends an average of about 500 meter SMW length, however, for simplifying the process, a length of 1000 meter has been chosen, here.

### 2.4 Spatial autocorrelation

Autocorrelation literally means that if pairs of objects are close to each other they are more likely to have similar characteristics. Spatial autocorrelation (SPA) measures how much the values of nearby objects are similar. Identifying BS in SPA is based on locating and measuring the spatial aggregation of contiguous spatial units (crashes) that are geographically approximate [84].

The assessment of SPA is based on measuring the degree of co-variation between spatial units’ values at each location and the nearby locations using the Global index. Global spatial autocorrelation is a measure of the overall correlation of the data. One of the statistics used to measure the global spatial autocorrelation is the Moran's I. The global index returns three important values: the Moran's I Index, z-score, and p-value. Moran's I Index is ranged between -1 and +1. Positive I index value with significant p-value or z-score indicates a positive association between variables and the tendency of the investigated parameter to be clustered spatially. While a negative Moran's I index indicates a tendency toward dispersion, and (0) indicates a random distribution of the variables with no relationship, as described in Figure 2.2.
Global spatial autocorrelation yields only one statistic to summarize the whole problem without the indication of the exact cluster locations. Moreover, sometimes the global index fails in detecting the presence of clusters, for example when an equal amount of positive and negative clustering cases exists. Therefore, a local autocorrelation index is recommended to be used to investigate the spatial variation and spatial associations between approximate spatial units. Spatial units could indicate the accidents themselves or road sections (polygons) with different accident frequency, as in our case. In the case of polygons, a contiguity edge method is applied to measure distances between the spatial units in which polygon features that share a boundary and similar values are more likely to be correlated. The value of the Morans’ I index can depend quite a bit on the assumptions of the spatial weights matrix $w_{ij}$ describing the neighboring relationships between all considered $i$ and $j$ locations. The idea is to construct a matrix that reflects the assumptions about a spatial phenomenon.

During the implementation of SPA the ArcGIS has been applied to identify BS based on the observed number of accidents per a spatial unit (polygon) and spatial contiguity matrix, as follows:

- Divide the road into small spatial units (polygons) and count the number of accidents $x_i$ for each $i$.
- Calculate the Global index to find out if a correlation exists.
- Calculate the local index $I_i$ for each $i$ location considering $j$ values for all other locations, as follows.

$$ I_i = z_i \sum_j w_{ij} z_j $$ \hspace{1cm} (2.2)

Where:

$z_i = x_i - \bar{x}$, for all $i$,

$z_j = x_j - \bar{x}$ for all $j$,

$\bar{x}$: the critical number of accidents,
$w_{ij}$: is the neighboring relationship between all considered $i$ and $j$ and $\sum_j w_{ij} = 1$ (*the weight here is row standardized*).

The $z$-score can give a sense of the intensity of spatial clustering or dispersion.

### 2.5 Consistency test for different segmentation criteria

The consistency test method is proposed to evaluate the consistency of the analysed BSID methods in case of different segmentation characteristics. The test measures the consistency of identifying road segment with high crash-risk by applying different segmentation characteristics (i.e. different segment length and sliding distance). The higher the proportion of shared BS sites is, the more consistent the applied BSID method is. Thus, if the consistency of a BSID method is 1 then all the BS sites identified by the given BSID method are included by the output of the other applied segmentation models.

The developed consistency test is similar to the method consistency test [85], [86]. However in this case the consistency is measured by the proportion of the shared sections related to the outputs of (i.e. $n1$, $n2$, ..., $n$) the same BSID method with different segmentation models.

\[
Consistency = \{n_1, n_2, ..., n\}_{(BSID1)} \cap \{n_1, n_2, ..., n\}_{(BSID2)} \cap ... \cap \{n_1, n_2, ..., n\}_{(BSIDn)} \quad (2.3)
\]

Where $n$ is the index of BS segments identified by the applied BSID method.

Figure 2.3 illustrates the method of performing the consistency test for a given BSID model in case of three different segmentation criteria. Circles represent the length of all sections identified as BS (such as top 5%, 10%...etc. of BS areas) by the different segmentation criteria (i.e. W1, W2, W3). The intersections between the circles represent the sections, which are identified as BS in case of all the related segmentation models. For instance, W12 represents the set of sections identified as BS in case of both segmentation criteria; W1 and W2. Similarly, W123 covers the sections which are indicated as BS sites in case of all the tested segmentation criteria W1, W2, and W3. Thus, the larger the shared section length is, the more consistent the applied BSID method is. In accordance with this, if all circles are concentric ($W1 \cap W2 \cap W3 = W1 = W2 = W3$) then this is the perfect consistency in case of the applied segmentation criteria.
2.6 Data Description

The used data is from Hungary and included accident information during the years 2013-2015 for two different roadway categories. The first road is the motorway M3 that has a total length of about 279 km and connects the capital Budapest with Nyiregyhaza city in the northeast of Hungary. The M3 is characterized by its high speed-limits ranged between 90-130 km/h. The second road category is an urban road. Two urban roads (numbers 1119, 1111) have been selected form the Hungarian road network. The two urban roads length are about 68 km with an average speed limit between 50-70 km/hour. These types of roads were selected to show the impact of the difference in their characteristics on road safety operations. The motorway (M3) and the urban road (1111, 1119) are randomly selected as a representative to high-speed motorways and low-speed urban roads, respectively.

The dataset for both roads includes accident information (i.e. locations, year), casualty characteristics (i.e. frequency, severity), and road characteristics (i.e. category, AADT, speed-limits). Additional roadway geometric (i.e. horizontal curves and lane characteristics) have been identified with the help of the Google Earth and ArcGIS software. It is worth to mention that the accidents geographic locations have been identified by the X and Y coordinate system using the ArcGIS.

During the analysis period (2013-2015), 262 accidents occurred on the motorway and 130 accidents occurred on the urban roads included both; fatal and injury cases. The official national accident database that been applied has been received from the Hungarian Institute of Transport Sciences Ltd (KTI). Such data was originally...
collected by the police. Therefore, it has been assumed that the data are valid and accurate.

### 2.7 Results

In the case study, the consistency of SMW and SPA methods is investigated in identifying BS sites in case of two different road categories. The assessment is based on the consistency analysis of the applied BSID method in case of different segmentation lengths (i.e. different constant lengths).

In order to demonstrate the effect of road characteristics on the applied BSID methods, two different road categories have been considered (i.e. motorway, urban road). Moreover, in order to demonstrate the effect of segmentation criteria; three constant length segmentation approach have been applied (i.e. 200 m, 400m, and 800m) in case of each BSID method for the two road categories. The selected lengths are assumed to be short enough to match the spatial distribution pattern of accidents in case of the urban road and long enough to match the spatial distribution pattern of accidents in case of the motorway. Table 2.2 represents the descriptive statistics of the applied segment lengths in case of the two considered roadway categories.

#### Table 2.2 Descriptive statistics of the segmentation characteristics of the motorway and urban roads

<table>
<thead>
<tr>
<th>Road type</th>
<th>Segment length (m)</th>
<th>Sliding distance (in case of SMW)</th>
<th>Average number of accidents (2013-2015)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorway</td>
<td>200</td>
<td>100</td>
<td>0.19</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>200</td>
<td>0.34</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>400</td>
<td>0.78</td>
<td>1.3</td>
</tr>
<tr>
<td>Urban road</td>
<td>200</td>
<td>100</td>
<td>0.38</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>200</td>
<td>0.78</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>800</td>
<td>400</td>
<td>1.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

The table (2.2) shows the average accident counts related to the investigated segment length values. It is clear that accident densities in case of urban road segments are higher than in case of the motorway. Based on the available accident data, it becomes possible to investigate the consistency of SMW and SPA methods in case of the different segment length and road types.

For the purpose of simplification, the cases of different segment lengths will be symbolized by W1, W2, and W3 (respectively 200m, 400m, and 800m).
2.7.1 Sliding moving window results

Three different window lengths (i.e. 200m, 400m, 800m) have been chosen to measure the accident frequencies in every sliding distance (i.e. 100, 200, 400 respectively), in case of the two selected roads. It is important to emphasize again that when a window satisfies the critical threshold value the next window continues sliding without overlapping, according to the standard SMW method. Figure 2.4 shows the result of SMW for W1, W2, and W3 in case of the motorway and urban roads.

![Figure 2.4](image)

(a) BS density per section-km of the (a) Motorway (b) Urban roads, measured for W1, W2, and W3
The bar diagram of Figure 2.4 represents the density of accidents (i.e. accidents per unit length) identified for the top 10% BS (y-axis) aggregated per every 25 km along the motorway and every 10 km along the urban roads (x-axis). The variation of the bar lengths indicates that if an identified section perfectly represents a BS, then the cause of the safety problem is strictly limited on the identified section. In light of this, any additional extension of the section would lead to the involvement of sections, which are not directly influenced by the certain safety problem so it would result in false positive cases (i.e. section identified as BS while it is not). Practically, this assumption could significantly be affected by the performance of the method itself. This might be examined by measuring the consistency of the method in identifying the same BS for each window length. Table 2.3 shows the resulted consistency values in case of the different window lengths.

Table 2.3 Sliding window consistency test measured for different window lengths.

<table>
<thead>
<tr>
<th>Shared windows</th>
<th>Equation (refer to Figure 2.3)</th>
<th>Motorway % shared area</th>
<th>Urban road % shared area</th>
</tr>
</thead>
<tbody>
<tr>
<td>W12</td>
<td>W12/(W1+W2-W12)</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>W13</td>
<td>W13/(W1+W3-W13)</td>
<td>0.16</td>
<td>0.01</td>
</tr>
<tr>
<td>W23</td>
<td>W23/(W2+W3-W23)</td>
<td>0.35</td>
<td>0.00</td>
</tr>
<tr>
<td>W123</td>
<td>W123/(W1+W2+W3-W123)</td>
<td>0.19</td>
<td>0.00</td>
</tr>
</tbody>
</table>

In Table 2.3, the consistency has been measured by identifying the shared BS sections along a roadway with the help of the ArcMap GIS. The consistency values are identified based on Figure 2.3. Accordingly, the consistency between any two windows is represented by the length of the shared BS sections divided by the sum of all BS in case of all window lengths (without repetition of the shared sections). For example, 27% of the top 10% of BS sites have been identified by both W1 and W2 together in case of the motorway, while the percentage is just 16% in case of the urban road for the same windows. Generally, the consistency of the SMW is apparently better for the motorway in comparison with the urban road. This can be seen noticeably through the zero shared BS area obtained by the urban roads for the windows W1, W2, and W3.

2.7.2 Spatial autocorrelation results
The same segments (i.e. W1, W2, and W3) have also been applied as input polygons in case of the SPA process in order to examine the variation of accident clustering scale for the different roadways. The SPA method starts by applying the Global Moran's index that gives a summary about whether or not there is any kind of a spatial correlation among the dataset. This can be obtained by applying the "spatial..."
autocorrelation” option of the ArcGIS Toolbox. The resulted reports of the Global Index are summarized in Table 2.4 and Figure 2.5.

Table 2.4 Global Moran's I Summary

<table>
<thead>
<tr>
<th>Description</th>
<th>Motorway</th>
<th>Urban road</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W1</td>
<td>W2</td>
</tr>
<tr>
<td>Moran's Index</td>
<td>0.04</td>
<td>0.09</td>
</tr>
<tr>
<td>z-score</td>
<td>1.66</td>
<td>2.38</td>
</tr>
<tr>
<td>p-value</td>
<td>0.09</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Figure 2.5 Global Spatial Autocorrelation Report (by the ArcGIS) resulted by each polygon length (W1, W2, and W3) for the (a) Motorway (b) Urban roads

The interpretation of the resulted global index is interesting, as the segment length of the applied polygon increases the global index increases in case of the motorway while it decreases in the case of the urban road. For instance, Table 2.3 and Figure 2.5

24
show that the identified clustering, in case of the motorway, moves from a low recorded value (I= 0.04, z-score= 1.66), near the random area, to a higher value (I= 0.16, z-score= 2.96). While the accidents of the urban roads show a higher tendency to clustering at lower segment length W1 (I= 0.69, z-score= 12.64) but reach the random area with longer segment length W2 (I= 0.05, z-score= 0.59). Moreover, Figure 2.5 proves the superiority of the long lengths segment (W3) in identifying accident aggregations for the motorway compared to the urban roadways.

Considering that the global index is only an indicator of the spatial correlation; a local Morons' index has been applied to find the exact high-density accident locations. The local index has been calculated for both of the investigated roadways. The resulted accident densities per the top 10% of the BS identified by the SPA method are presented in Figure 2.6.

Figure 2. 6 BS density per section-km of the (a) Motorway (b) Urban roads, measured for W1, W2, and W3
Figure 2.6 can be explained in the same way as Figure 2.4, except that the average density of the urban road segments is significantly higher than that of the motorway.

The consistency of the applied SPA has been identified based on similar rules as the previously mentioned SMW. Every time roadways are divided into different polygons with different length (i.e. W1, W2, W3) and then SPA is calculated based on accident content of each polygon. Table 2.5 shows the resulted consistency values in case of the different applied polygon lengths.

<table>
<thead>
<tr>
<th>Shared BS windows</th>
<th>Equation (refer to Figure 2.3)</th>
<th>% shared area (Motorway M3)</th>
<th>% shared area (Urban road)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W12</td>
<td>W12/(W1+W2-W12)</td>
<td>0.33</td>
<td>0.44</td>
</tr>
<tr>
<td>W13</td>
<td>W13/(W1+W3-W13)</td>
<td>0.20</td>
<td>0.22</td>
</tr>
<tr>
<td>W23</td>
<td>W23/(W2+W3-W23)</td>
<td>0.28</td>
<td>0.15</td>
</tr>
<tr>
<td>W123</td>
<td>W123/(W1+W2+W3-(W12+W13+W23+2*W123))</td>
<td>0.27</td>
<td>0.10</td>
</tr>
</tbody>
</table>

The consistency analysis has resulted in similar values for the two roads. The SPA method appears to be more consistent between the lower length segments in case of the urban road (0.44), but still consistent for longer lengths in case of the motorway. In general, it is clear that the consistency of the SPA is significantly influenced by the segmentation characteristics and the types of roadways in which it is applied.

### 2.8 Discussion and conclusion

This research is aimed to study the pros and cons of the applied black spot identification (BSID) methods by comparing the variation in their outcomes in case of two different road categories. The investigated road categories have been the Hungarian motorway and urban roads. The research has also aimed to reveal the variation among the outcomes of different BSID methods in light of the applied road network segmentation characteristics. The BSID methods include the sliding moving window (SMW) screening method and the spatial autocorrelation (SPA) method which is represented by the spatial correlation among accident datasets. A consistency test method is proposed to evaluate different segmentation characteristics for the same BSID method. The test measures the consistency of high risk road segment identification processes in case of different segmentation characteristics.

The top 10% of the riskiest BS have been examined. The results have shown that the use of different segment lengths (and sliding distance) seems to have a significant impact on the performance of the SMW method. This can be seen from the difference
in the identified BS along the road in case of the different applied segmentation criteria, as presented by the consistency test. Generally, the SMW has shown a comparatively higher consistency in case of the motorway compared with the urban roads. This can primarily be explained by the rule of "first come first serve without overlapping" applied by the standard SMW method and the applied constant window length.

The application of the SPA method has also demonstrated a variation in performance with different segmentation criteria and different road categories. In the low-speed urban roads, accidents showed a higher spatial autocorrelation in case of small segment lengths in comparison with the motorway accidents. This can be explained by the nature of the accident distributions along the roadway. Relatively high spatial autocorrelation of longer segments means accidents are more scattered than in case of high autocorrelation value of shorter segments. The SPA method has shown higher consistency in case of the urban road for lower segment lengths. Generally, urban accidents seem to have a higher spatial dependency within a lower scale in comparison with motorway accidents. However, the most relevant finding of this study can be summarized as follows:

- In case of the analyzed roads which represent different road categories, the different BSID methods have been performed in a significantly different way.
- Every BSID method could lead to different outcomes with different segment length, starting point, and different process and criteria for segmentation.

As a result of this chapter, a new discussion has been opened about the impact of the applied segmentation method and the analysed road category. This is the field, what I try to reveal in the following chapters.

**Limitations and scope for future works**

The limitations of this research relate to the limited number of BSID methods and roadway categories that have been used during the investigation. In this research, only two BSID methods have been analysed, with only two roadway categories. The sliding window is considered to be a representative method for screening models, and the spatial autocorrelation is also considered to be a good representative method for the clustering technique. However it has to be emphasized that, the results can not be generalized for all BSID methods but gives enough indication of some limitation of the applied BSID approaches. However, the same methodology can be applied and developed for further investigation between more BSID methods.
2.9 Thesis

- Thesis (1)

I have introduced new evaluation criteria of two different BSID methodologies applied with different segmentation characteristics and road categories. Based on the result of the investigation, it can be concluded that different road categories and different BSID methods can be characterized by significantly different consistency.

Related publications to Thesis 1: [76]

Chapter 3

Developing a new spatial clustering method for road accidents

3.1 Short summary

In road safety, the process of organizing road infrastructure network data into homogenous entities is called 'segmentation'. The segmenting of the road network is considered the first and most important step in developing a safety performance functions (SPF). The objective of this chapter is to study the benefit of a newly developed network segmentation method which is based on the generation of accident groups, applying K-means clustering approach. K-means algorithm has been used to identify the structure of homogeneous accident groups. According to the main assumption of the proposed clustering method, accident risk is strongly influenced by the spatial interdependence of accidents. The performance of K-means clustering has been compared with four other segmentation methods applying constant average annual daily traffic based homogenous segments, constant length segments, related curvature characteristics and a multivariable method suggested by the Highway Safety Manual (HSM). SPF has been used to evaluate the performance of the five segmentation methods in predicting accident frequency. K-means clustering based segmentation method has been proved to be more flexible and accurate than the other models in identifying homogeneous infrastructure segments with similar safety characteristics.

3.2 Introduction

An appropriate SPF is considered to be one of the basic methods of road safety analysis. SPF represents a mathematical relationship between accident frequency and other related explanatory variables. In most of the cases, the reliability of a SPF depends fundamentally on the validity of the applied statistical methods and the way how data is organized, - with other words - clustered into specific homogeneous sets or groups of similar entities. Most often, road segmentation is based on researchers’ experiences, methodological decisions or objectives. A variety of popular segmentation methods exist [2].
Some scholars apply a constant length segments while keeping other variables constant [88]. Kononov and Allery [89] divided the road into 2-mile segments to identify high accident road segments. Others divide the road based on certain traffic characteristics, road geometrical design features, or both. Abdel-Aty and Radwan [67] divided the road network into homogeneous segments in terms of geometry (horizontal alignment, shoulder and median characteristics, lane width, etc.) and traffic flow features. Sadeghi et al. [90] have suggested that a new homogeneous road sections can be defined when one of these factors change: AADT, speed-limit, curvature rate, and lane width. According to the HSM for secondary roads [20], a new segment is created by changing one of the following parameters: AADT, roadside hazards, or road curvature. However, applying different lengths and start points for segmenting road network can result in different definitions of hazardous locations [5,6] which in turn affect the accuracy and stability of results. Generally, if the road segmentation process targeting the generation of homogeneous groups is based on too many variables, it can result in very short average segment lengths [9] which can lead to the overrepresentation of zero-accident sections. In contrast, increasing segment length would violate the requirements related to homogeneity. Besides this most of the segmentation approaches apply only traffic conditions and road geometrics related attributes in identifying homogeneous segments and dispense with the consideration of accident data, which can in some cases significantly improve the reliability of the model. Boroujerdian et al. [93] claimed that the length of high accident road segments can be identified based on accident arrangement, and he proposed a dynamic segmentation method for that.

Considering that traffic accident data is heterogeneous, in general, identifying accident clusters considering also spatial references can significantly contribute to the outcomes of the applied segmentation method. In other words, traffic accident analysis using a clustering approach can be a useful technique to find hidden relationships and patterns for a large number of accidents. Moreover, what distinguishes this technique as well is its ability to classify accidents into small coherent groups considering their spatial and temporal dependencies.

In this chapter, an accident clustering based road network segmentation method is introduced, which classifies accidents and road dataset into homogeneous cluster segments. The segmentation process is based on the assumption that accidents which occur within close distances, spatially and temporally, are more likely to have similar characteristics. The chapter also introduces criteria for comparing the introduced spatial clustering method with some popular segmentation methods (i.e. based on the HSM specifications, constant AADT segments, constant length segments, and segments characterized by curvature). The SPF has been used to evaluate the performance of the five segmentation methods in predicting accident frequency.
3.3 Data description

The used data is from Hungary and related to the secondary main roads number 25, 35, 36, 49 and 82. The investigated data has been generated from 2013 to 2016, and it has included accident data, traffic characteristics, and road design parameters. Only fatal and injury accidents have been included in this study since they have the greatest impact on safety. The total length of the investigated network is about 359 km (in one direction). Only the accidents of road segments have been considered in the analysis excluding the accidents of intersections. During the analysis period, 870 fatal and injury accidents occurred. The data has been divided into two parts; the data of the first three years (2013-2015) has been applied for developing the models, while the data of the last year (2016) has been used for checking the performance of the developed models. Table 3.1 presents the basic characteristics of the used dataset.

Table 3.1 Description of the data used in developing the models

<table>
<thead>
<tr>
<th>Road Ref.</th>
<th>Accident counts (fatal and injury)</th>
<th>AADT (min)</th>
<th>AADT (max)</th>
<th>Length (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2013</td>
<td>2014</td>
<td>2015</td>
<td>2016</td>
</tr>
<tr>
<td>25</td>
<td>57</td>
<td>56</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>35</td>
<td>30</td>
<td>43</td>
<td>38</td>
<td>33</td>
</tr>
<tr>
<td>36</td>
<td>26</td>
<td>39</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>49</td>
<td>24</td>
<td>37</td>
<td>39</td>
<td>35</td>
</tr>
<tr>
<td>82</td>
<td>59</td>
<td>60</td>
<td>68</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>870</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4 Developing of a new spatial clustering segmentation method

3.4.1 K-means clustering

Classification models usually aim to organize a large dataset into a small predefined number of homogeneous groups using different analytical methods called “clustering”. Cluster analysis can have several definitions depending on the field of application, but generally, it has a major objective of organizing a large data set (or objects) into a number of small homogeneous groups in which the degree of
association between the objects within the same group is maximal, so the generated clusters become more easily and efficiently understandable. Many cluster algorithms have been developed \[94\], \[95\]. K-means clustering is one of the widely used data partitioning techniques designed to classify \( n \) objects into \( K \) clusters \((k_1, k_2, \ldots, K)\). Simplicity of application, even in case of a large dataset, makes this algorithm very attractive (i.e. compared to hierarchical methods and density-based methods).

K-means clustering is generally based on the definition of a cluster structure that minimizes a specific error criterion. The main process of the k-means clustering approach is presented in Figure 3.1. Every object (e.g. accident) in this method is represented by a point in a geographical location and every point has different attributes and coordinates. A good way to measure the affinity between any two points is the definition of their distance. K-Means clustering algorithm starts with an initial set of cluster centres chosen randomly or according to some heuristic methods. During the iteration process each point is assigned to its nearest centre according to the Euclidean Distance method (other distance similarity methods can also be applied). In the next step, cluster centres are re-calculated again, considering the new members, according to Equation (3.1). The iteration stops when no more cluster centres need to be relocated.

\[
S_k = \frac{1}{N_k^2} \sum |x_i - x_j|
\]  

(3.1)

Where \(S_k\) is the measured dissimilarity value for a cluster \(k\). \(N_k\) is the number of objects (accidents) belonging to cluster \(k\). \(\sum |x_i - x_j|^2\) is the sum of squared difference in distances between each object (accident) \(j\) and its nearest centre \(i\) for the same cluster \(k\).

For a given data set, the method of sum of squared error (SSE) is often used to measure the compactness degree of all clusters. It simply measures the sum of square difference between the objects and their nearest centre. The algebraic formula of the \(SSE\) can be represented in the following Equation (3.2).

\[
SSE = \frac{1}{2} \sum_{k=1}^{K} N_k S_k
\]  

(3.2)

Where: \(K\) is the total number of clusters for a given data set.
In case of road accidents, a variety of attributes (coordinates) can be applied to measure and classify accidents into homogeneous groups. The greater the number of attributes used, the more homogeneous the accidents of one group is. However, increasing the number of attributes may decrease the average size of segment length and its accident content while increasing the complexity of the analysis process.

### 3.4.2 The proposed spatial clustering segmentation method

The selection of a proper attribute is crucial in the accident classification process. The developed methodology tries to find out the most homogeneous road segments with lower complexity. The methodology suggests that the spatial location is a key classification variable for road accidents since it has a significantly higher probability that neighboring accidents have strongly related causes than in case of distantly located accidents, because they are more likely to occur within similar environment conditions [39].

Considering that the accidents are represented by two coordinates (the longitude and latitude) on the road network, the k-means algorithm could misunderstand the spatial distribution of accidents in case of considering each coordinate as a separate variable. To reduce the demanded calculation capacity of the problem, instead of a bi-variable representation of accidents’ spatial location in the clustering algorithm, a linear referencing model [96] seems to be appropriate. Linear referencing is used to locate objects along a line from a reference point (RP). The linear referencing method has been applied to locate accidents along every single road where the road has been represented as a one-dimensional line, with a zero-distance starting point as a RP [97], as presented in Figure 3.2. All accidents have been located along the road by measuring their distances from the RP. For instance, in Figure 3.2, the first accident (represented by a point counted from the left) is located 0.6 distance unit away from the RP. This approach results in a one-variable representation for the K-means algorithm contrary to the two-variable geographical coordinates. It also allows
avoiding classifying accidents in the same cluster that are closely located but occurred on different roads.

The application of K-mean clustering and linear referencing can be very attractive; since the road accident based segmentation method makes it possible to relate the length of the identified segments to the clusters’ lengths. Each cluster length is determined by the number of accidents involved and their spatial distribution. The length of a segment equals the distance between the first and last accident within the cluster. For instance, referring to Figure 3.2, the spatial cluster (SC1) has a segment length equals to 0.6 (i.e. \( L_1 = 1.2 - 0.6 \)) and accidents number equals to 3 (i.e. \( N_1 = 3 \)). Moreover, in the developed method at least two accidents are needed to form a segment. Accordingly, empty sections between clusters with no accident history can be easily separated from the other clusters.

3.4.3 Measuring the optimal number of cluster

One of the crucial issues related to K-means clustering is that the number of clusters (K) that has to be determined in advance, which can cause uncertainty without any prior knowledge. The selection of the optimal K is usually affected by the type of study and identified based on different statistical methods and related experiences. For determining the optimal estimated number of clusters, it is required to test different numbers of K and analyze the changes of the variances (or SSE) compared to cluster centers. A method of minimum variance partition has been used [98]. This method searches for a point, which splits the SSE function domain into two parts. Until this point, each added cluster results in a significant reduction in the value of variance, and after the given point, any increase in \( k \) results in less-and-less reduction in the value of variance. This approach represents actually the second derivative of the variance as a function of the number of clusters.

3.4.4 Further segmentation (optional)

Applying the spatial factor to segment road accident networks is very useful in revealing the heterogeneity of accident data. Although the spatial factor seems to be
quite relevant, in some cases it is possible to insert additional factors in the segmentation method. According to literature [20], [99], the traffic volume (i.e. AADT) plays an important role in classifying and predicting accident data. So, it is possible to divide the resulted segments, by the spatial clustering method, for further segments, taking into account the homogeneity of AADT. Other variables can also be used based on the purpose of research. However, sometimes the spatial convergence of accident can be enough to give homogeneous segments (as will be discussed later). However, as mentioned above, increasing variables numbers can result in a large number of small segments with the chance of losing some important information referring to the interrelationships characterizing the analyzed data.

3.5 Performance evaluation

In order to examine the effectiveness and applicability of the proposed segmentation method, it is better to measure and compare its performance with some popular segmentation methods applied in the same field. To do so, the performance of the spatial clustering method has been compared with four other well-known segmentation methods that apply constant AADT segments, constant length segments, related curvature characteristics, and a multivariable method suggested by the HSM. The SPF has been used to evaluate the performance of the five segmentation methods in predicting accident frequency. During the evaluation the accident prediction model (i.e. SPF) has been developed from the dataset resulted from the segments of each method. And, the goodness-of-fit of the resulted models has been compared. The description of the other segmentation methods and the evaluation criteria are in the following sections.

3.5.1 Description of the other segmentation approaches

3.5.1.1 Segmentation variables

According to literature, a variety of variables have been used to reveal the heterogeneity of road accident data. The variables used in the segmentation vary according to the applied method. The main variables included in this study, apart from the spatial factor, are described as follows:

- The average annual daily traffic (AADT): AADT has been considered as a major variable in road segmentation and its value plays an important role in predicting accident frequency [100].

- Speed limit: Speed limit can affect directly the accident occurrence, where since low-speed vehicles can have more perception-reaction time to avoid an accident than high-speed vehicles. Also, the speed limit can affect accidents severity.
• Percentage of trucks and percentage of light vehicles: In this study, light vehicles include passenger car, small van and light truck (under 3.5 ton); while truck includes medium and large vehicles and trailers (over 3.5 ton).

• Horizontal curve: Instead of straight roads with long stopping sight distance, horizontal curves can have an unfavorable effect on stopping sight distance, which can be considered as an important risk factor influencing accident probability. Besides this, the speed limits of infrastructure elements characterized by outstanding curvature values could also have a negative effect on the probability of run-of-road accident.

### 3.5.1.2 The other segmentation methods

Each of the other investigated methods differs in the criteria applied to determine homogeneous road segments. The other segmentation methods investigated in this study represent the most used approaches, which are described as follows:

• HSM method (Seg-2): The generated segments are homogenous in AADT, roadside hazards (RSH), and presence of curves, as recommended by the HSM specifications for highway roads. The identification of RSH should be based on different factors [101], [102]. In this research 4 types of RSH have been differentiated (trench, embankment, trees and bridges) with 4 possible intensity scores (from 1 to 4, in increasing order of potential hazard). The weighted average of RSH for any single SCL base segment \( n \) is calculated by Equation 4, including both road directions.

\[
RSH_n = \max(IN \times TW_j)
\]  

(3.3)

Where \( IN \) is the intensity of the RSH in segment \( n \), and \( TW \) is the relative weight of jth type of RSH based on AASHTO weight index [103].

• Constant AADT (Seg-3): The generated segments are homogenous in AADT while other variables do not affect the segmentation structure.

• Constant length (Seg-4): The road network is split into equidistant intervals, while other variables do not affect the segmentation structure. The segmentation length has been chosen to be 750 meters. It is an average length that is longer than the recommended minimum value of the HSM but short enough to provide homogeneity, so the characteristics of the road do not change too much within the length.

• Curves (Seg-5): Segments are generated considering the endpoints of the horizontal curves and straight lines. Each different curve and straight line is ordered to a different segment. The data organized to have within each segment 2 curves and 2 tangents, avoiding having short segments when using
a single curve [32]. According to the HSM specification [20], the curve is identified from the point where the tangent ends and the curve begins. If the curve has spiral transitions, then it is measured from the effective begin and endpoint of the curve.

The below presented descriptive statistics table (3.2) represents the five segmentation methods describing the minimum, maximum, mean and standard deviation values related to the different approaches.

Table 3.2 Descriptive statistics of length and accident numbers per segment

<table>
<thead>
<tr>
<th>Parameters description</th>
<th>Seg-1</th>
<th>Seg-2</th>
<th>Seg-3</th>
<th>Seg-4</th>
<th>Seg-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment Length</td>
<td>Min length (km)</td>
<td>0.04</td>
<td>0.02</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Max length (km)</td>
<td>10.59</td>
<td>2.61</td>
<td>16.34</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Average length (km)</td>
<td>3.68</td>
<td>0.47</td>
<td>4.01</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>3.35</td>
<td>1.63</td>
<td>2.17</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of segments</td>
<td>42</td>
<td>462</td>
<td>63</td>
<td>465</td>
</tr>
<tr>
<td>Accident frequency</td>
<td>Min</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>18</td>
<td>6</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>4.86</td>
<td>0.43</td>
<td>3.29</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>3.23</td>
<td>0.81</td>
<td>3.19</td>
<td>0.87</td>
</tr>
</tbody>
</table>

The total number of segments, in Table 3.2, represents the number of generated sections. In the modelling part (the following section), accidents number per segment has been averaged for the three study years (2013-2015).

The table (3.2) also shows that all of the segmentation methods count zero-accident segments, except for the proposed K-means clustering method (Seg-1), which needs at least two accidents to form a segment. Its basic concept assumes that single or distinct accidents (far from other accidents spatially and temporally) accidents, which do not have any common characteristics with their neighbouring accidents is considered as a noise which are randomly occurred. Therefore, they are not considered in a further analysis. It is obvious that this consideration is a heavy simplification however on the other hand it can help to reduce negative effects of over fragmentation. However, it has not yet been researched in details that the road should be investigated with or without zero-accident segments, when building prediction models. To investigate the introduced issue, both cases are examined in the following sections.
3.5.2 Performance criteria (development of accident prediction models)

3.5.2.1 Model variables

As it has been introduced, the selected segmentation methodology can significantly determine the output of the following analytical steps (e.g. the results of SPF). Since the main objective of this research is to evaluate safety related characteristics of the infrastructure network, it is important to choose such segmentation variables that significantly affect safety level of an infrastructure component. According to literatures [104], [105], the following are the most applicable explanatory variables for accident prediction.

- AADT: AADT is considered as a major factor in predicting the number of accidents.
- Speed limit: speed is also an important factor related to accident risk.
- Percentage of trucks and percentage of light vehicles: The percentage is measured per total traffic in case of each specific segment.
- Horizontal deflection angle (HDA): In a mathematical sense, the curvature is the reciprocal of the radius. A small curve can easily be characterized by using the radius. But, if the radius is large as a mile or a km, HDA is more convenient for describing the horizontal curve. HDA is defined as the amount of change in the curvature rate of a straight line horizontally. It is measured from a central angle to the ends of a chord of agreed length, and it is mathematically calculated as follows:

\[
HDA \left( \frac{\text{degree per unit length (feet)}}{\text{Radius of curve (feet)}} \right) = \frac{5279}{\text{Radius of curve (feet)}}
\]  

(3.4)

For each specific segment, curves have been determined by their transition points, their dimensions have been measured and the total HDA in degrees has been calculated per unit length of the segment.

When a segment has not got a constant AADT, speed or traffic value, its length value (as a linear length-weight) has been used to estimate average values.

However those variables - which are used to determine the segment structure - should be carefully chosen to ensure their correlation with the dependent variable (i.e. accident frequency) and independence among themselves. For that purpose correlation analysis and stepwise method have been applied. Table 5.3 contains the resulted correlation coefficients for Seg-1.
Table 3. Correlations coefficient parameter of cluster segments.

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Per cent of truck</th>
<th>Speed</th>
<th>HDA</th>
<th>AADT</th>
<th>Per cent of light vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per cent of truck</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speed</td>
<td>-0.251</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDA</td>
<td>-0.082</td>
<td>0.002</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AADT</td>
<td>0.466</td>
<td>0.076</td>
<td>0.120</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Per cent of light vehicles</td>
<td>0.805</td>
<td>-0.153</td>
<td>-0.170</td>
<td>-0.314</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The highest correlation has been detected between three explanatory variables: AADT, percentage of trucks, and percentage of light vehicles. Due to this reason, only one variable related to traffic flow is selected. The same situation has been found with the other segmentation methods (their correlation data are not described in details). The stepwise forward approach has also been used to check the significance of inserting or removing different explanatory variables for the tested SPFs. Finally, AADT, speed, and HDA have been selected to be the input variables of the segmentation models.

3.5.2.2 Model description

The model development process has been implemented in R software package and IBM SPSS Statistics. The maximum likelihood method has been used to estimate the model parameters.

Since accident frequency along road segments is a count data that could include zero values, the negative binomial regression has been used to predict accident frequencies [106]. The negative binomial can be distinguished from the Poisson regression through solving the over-dispersion problem when the observed variance exceeds the mean value. Two types of SPF models have been developed. The first type includes only one explanatory variable: AADT, while the other includes; AADT, speed, and HDA. Accidents have been predicted per unit length, by considering the varying segment lengths as an offset variable.. The general form of the applied SPF is represented by the following equation (3.5) [19]:

\[
SPF = \exp[\alpha + \sum_n \beta_n \times \ln(X_n)] + \ln(length)]
\]  

(3.5)

Where \(\alpha\) is the intercept of the ordinate axis, and \(\beta_n\) is a regression coefficient of the corresponding explanatory variable \(X_n\) (i.e. \(\beta_1 = \text{AADT}, \beta_2 = \text{speed},\) and \(\beta_3 = \text{HDA}\)).

To evaluate and compare goodness-of-fit between the different developed SPFs, two different statistical methods have been applied: the Quasi-likelihood under Independence Model Criterion (QIC) [107], [108] and the Pearson Correlation Coefficient (PCC) [109], [110]. The QIC is one of the most widespread well-
established goodness-of-fit statistics for the Generalized Estimating Equation (GEE), which is an extension of Akaike’s information criterion (AIC) [111]. QIC can be used to select the most applicable estimation model structure in the Generalized Estimating Equation (GEE) analyses. Like AIC, it balances the model fit with model complexity to select the most effective model. The models being compared do not need to be nested, which means that the parameters of the different evaluated models do not need to be the subsets of each other. Therefore, QIC can be used to compare and rank different models with different number of parameters. When QIC is applied to model evaluation, the model with the smallest QIC is preferred. However, QIC does not provide any test regarding a null hypothesis whether the investigated data fits the model results or not.

Thus, Pearson's correlation coefficient (PCC) test is used to measure the efficiency of the developed models in predicting accident data. PCC measures the linear correlation between any two variables X and Y. It can be defined also as a covariance of the two variables divided by the product of their standard deviations. PCC has a value between +1 and −1. A value between 0 and 1 implies a positive linear correlation between X and Y. A value between 0 and -1 implies a negative linear correlation between X and Y. Correlations equal to 1 or −1 correspond to a perfect correlation. PCC has been represented by the following equation (3.6):

\[
PCC = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum x_i^2 - n \bar{x}^2} \sqrt{\sum y_i^2 - n \bar{y}^2}}
\]  

(3.6)

Where: \( x \) is the observed number of accidents that occurred in a segment, \( y \) is the predicted number of accidents, \( n \) is the sample size; and

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i (\text{can be expressed similarly})
\]

The two methods form an integrated statistical methodology for selecting the best fit least complexity model.

### 3.6 Results and discussions

The proposed K-means clustering segmentation has resulted in 126 segments in case of the investigated roads regarding the analysed time period (2013-2015). All zero-accident road segments have been excluded from the analysis, in case of the proposed segmentation method, in order to examine the validity of the model in this case (the
whole road including all the zero-accident segments will be analysed and compared in the next chapter). Table 3.4 summarizes the resulted statistics of the K-means clustering for each roadway in this study.

Table 3.4 Summary of k-means segmentation results by road reference

<table>
<thead>
<tr>
<th>Road Ref.</th>
<th>Total number of segments (total for 3 years)</th>
<th>Average segment length (km)</th>
<th>Average accident frequency (per segment per year)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2013</td>
<td>2014</td>
</tr>
<tr>
<td>25</td>
<td>27</td>
<td>4.41</td>
<td>6.5</td>
</tr>
<tr>
<td>35</td>
<td>27</td>
<td>2.83</td>
<td>4.1</td>
</tr>
<tr>
<td>36</td>
<td>17</td>
<td>4.72</td>
<td>6.3</td>
</tr>
<tr>
<td>49</td>
<td>20</td>
<td>4.13</td>
<td>3.7</td>
</tr>
<tr>
<td>82</td>
<td>35</td>
<td>2.76</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td><strong>Total number of segments = 126</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The resulted data by the other segmentation methods have also been used to develop new SPF models. The model calibration results of different explanatory variables are shown in Tables 3.5a-b.

Table 3.5 Values of the model parameters, (p-value), QIC, PCC and over-dispersion (k) for: (a) model 1 with one explanatory variable; AADT, b) Model 2 with three explanatory variables; AADT, speed, HDA.

(a) Model 1

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Methods of segmentation</th>
<th>Seg-1 (K-means clustering)</th>
<th>Seg-2 (HSM)</th>
<th>Seg-3 (Constant AADT)</th>
<th>Seg-4 (Constant length)</th>
<th>Seg-5 (Curvature)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α (Intercept) [p-value]</td>
<td>-6.811 [&lt;0.001]</td>
<td>-11.710 [&lt;0.001]</td>
<td>-8.866 [&lt;0.001]</td>
<td>-18.616 [&lt;0.001]</td>
<td>-10.220 [&lt;0.001]</td>
<td></td>
</tr>
<tr>
<td>β1 (AADT) [p-value]</td>
<td>0.795 [&lt;0.001]</td>
<td>1.209 [&lt;0.001]</td>
<td>0.957 [&lt;0.001]</td>
<td>1.982 [&lt;0.001]</td>
<td>1.099 [&lt;0.001]</td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>1.132</td>
<td>2.334</td>
<td>1.029</td>
<td>1.090</td>
<td>1.064</td>
<td></td>
</tr>
<tr>
<td>QIC</td>
<td>20</td>
<td>164</td>
<td>66</td>
<td>150</td>
<td>297</td>
<td></td>
</tr>
<tr>
<td>PCC</td>
<td>0.786</td>
<td>0.321</td>
<td>0.738</td>
<td>0.447</td>
<td>0.663</td>
<td></td>
</tr>
</tbody>
</table>
Most of the variables (i.e. $\alpha$, $\beta_1$, $\beta_2$, and $\beta_3$) considered in the stepwise procedure in Table 3.5 are statistically significant to 0.05 significance level (95% confidence level). Speed and HAD are statistically significant to 90% confidence level, for some of the segmentation approaches. This can be yield due to the variation of their values for a single segment. The intercept coefficient $\alpha$ of Seg-1 (Table 3.5b) is significant to 0.1 levels. This can be explained by the relatively small sample size used in the cluster-based segmentation model (Seg-1). Generally, the role of $\alpha$ is not crucial, so a small deficiency does not affect critically the efficiency of the whole model since it rather acts as a calibration factor for the model. Nevertheless, the whole model is statistically significant for all of its parameters.

For example, due to the small constant length segments, in the case of Seg-4, the effect of curves decreases while the effect of AADT and speed increase. According to the model results speed seems to have an inverse effect on accidents, however this phenomenon shall primarily be interpreted in light of the service level improvement including safety related attributes. Since, usually road specifications are improved as
speed-limits increase, especially considering physical separation of the traffic directions. This is also supported partly by Garber & Gadiraju (1989) [112] who have found that drivers’ speed tend to increase as road geometrical characteristics are improved, and that accident rates do not necessarily increase in case of an increase in the average speed but do rather increase in case of an increase in speed variance.

It can be concluded by evaluating Table 3.5a that the models are slightly improved by included the new variables, described in Table 3.5b. Table 3.5 also shows that, the lower best QIC values have been given by the developed segmentation method (Seg-1), (QIC1(a)= 20 and QIC1(b)= 22). While the model developed based on curvature segmentation method shows the worst QIC results; QIC5(a)= 297 and QIC5(b)= 284, followed by the HSM segmentation based method (QIC2(a)=164, QIC2(b)=167).

Pearson Correlation Coefficient (PCC) has been applied to evaluate and compare the developed models by measuring their efficiency in predicting accident data for another year in the future (i.e. year 2016). PCC has been used to describe the strength of the relationship between the predicted and observed values (Tables 3.5). In accordance with this Figure 3.3 presents a scatterplots of relationships between the observed accidents (x-axis) and the predicted accidents (y-axis) for the 2016 accident data, including all the segmentation models presented in Table 3.5. Figures 3.3a-e represent the models of Table 3.5a, while Figures 3.3f-j represents the models of Table 3.5b. Figures 3.3a-j also presents linear regression lines (solid lines) fit the data and their R-square values. The dashed lines indicate a perfect prediction of the accident data. Solid regression lines above and below the dashed lines indicate that model prediction is overestimated or underestimated compared to the actual number of accidents.

It is obvious from Figures 3.3 that all models tend to underestimate the accident numbers at lower frequencies and overestimate it at higher frequencies, except the cluster regression model (Seg-1) (Figure 3.3a). In case of all segmentation methods, except for Seg-2, the quality of prediction is slightly improved by increasing the number of explanatory variables (Table 3.5a and b: regarding PCC and R-squared values). However, the differences in the PCC values between model 1 and model 2 (Tables 3.5a and 4b) do not seem to be decisive. Generally, based on the results of Tables 3.5 it can be concluded that the capability of the models to predict accident frequencies is between weak and moderate (PCC: 0.300 - 0.794, R-squared: 0.08 - 0.63). This can be explained by the relatively small dataset analyzed in this evaluation. However, this study aims to give more concentration on the difference in performance between the models. The poorly fitted regression lines of Seg-2 to Seg-5 (Figures 3.3b-e and g-j) are caused by the strictly ordered discrete data which makes it difficult to provide a well-fitted regression line. In addition, zero-accident segments can also weaken a model in some cases, and this is obvious in case of Seg-2 which gives the lowest PCC around 0.3 and it is almost failed in describing the linear regression line with R-squared values around 10%. Even if the empty segments are eliminated from the model the results do not show much improvement compared to
the proposed model. Despite this, the Seg-3 provides an efficient prediction model with relatively reasonable PCC; $\text{PCC}_{3(a)} = 0.738$ and $\text{PCC}_{3(b)} = 0.760$. In general, the developed cluster based segmentation models (Seg-1) provide the best prediction efficiency, with the highest PCC and R-square values ($\text{PCC}_{1(a)} = 0.786$, $R^2 = 0.62$ and $\text{PCC}_{1(b)} = 0.794$, $R^2 = 0.63$) as well as the lowest QIC values ($\text{QIC}_{1(a)} = 20$ and $\text{QIC}_{1(b)} = 22$).

Figure 3. 3 Graph the linear correlation between the observed accidents (x-axis) and the predicted accidents (y-axis) of the year 2016
3.7 Conclusion

In road safety, the success of any method applied in identifying high-risk locations should depend fundamentally on how data being organized into specific homogeneous segments. In this chapter, a new method for classifying and segmenting road accident network has been introduced. The method is based on applying K-means clustering for segmenting accident data into small homogeneous segments especially considering their spatial dependency. The methodology assumes that accidents which occur within close distances spatially and temporally are more likely to share the same features. Accidents distribution has been measured using a linear referencing method. The linear referencing method has been used to count accident locations along one-dimensional roads, measured from a reference point, in a single attribute.

In order to evaluate the performance of the proposed spatial clustering segmentation method, it has been compared with another four well-known segmentation methods. The second segmentation method is based on the Highway Safety Manual (HSM) specifications, using curvature and average annual daily traffic (AADT). The third method is based on constant length segments, whilst the fourth method is based on constant AADT segments. And in case of the fifth method curvature characteristic has been applied to separate segments. The performance of the five segmentation methods has been examined by comparing their performance in developing new safety performance functions (SPFs), in case of a Hungarian secondary road accident database. Two models with different explanatory variables have been developed for every segmentation method. To develop the SPFs, a negative binomial model has been used. The goodness of fit of the models has been evaluated using two statistical methods: the Quasi-likelihood under Independence Model Criterion (QIC) and the Pearson Correlation Coefficient (PCC). The two methods form an integrated statistical methodology for selecting the best fit least complexity model.

In the end, K-means clustering based segmentation models have given the best goodness-of-fit statistics. This meets my expectations, because the segment length, in this method, is flexibly influenced by data availability, quality, and other variables to optimize the SPF calibration. In contrast, unexpectedly, the curvature and HSM based segmentation methods have been the least, of the investigated cases. This can be explained, to some extent, due to applying several variables in the segmentation process which could separate similar accidents in some locations and produce many zero-segments. However, segmentation approach based on spatial clustering seems to be promising. This can be an important milestone for the further the applicability of this method.

Possibility for implementation in practice

The developed spatial clustering segmentation method has numerous application possibilities in the field of road traffic safety. In road safety, it can be used to improve the reliability of SPF which consequently affects the performance of the relevant BSID methods that rely on accident prediction models in their criterion. In the field of
BS analysis, the developed segmentation method can contribute to the identification of proper segment lengths, which can help in fitting the BS area to the characteristic of the accident distribution. Thesis-2 provides also a methodology to assess and compare data segmentation methods.

3.8 Thesis

- Thesis (2)

I have developed a new road network segmentation method based on the spatial distribution of road accidents by identifying homogeneous accident clusters in terms of their spatial and temporal properties.

Related publications to Thesis 2: [113]–[115]


Chapter 4

A comparative analysis of black spot identification methods and road accident segmentation methods

4.1 Short summary

The success of any method applied in identifying high-risk locations or black spots (BS) on the road should depend fundamentally on the way, how data is organized into specific homogeneous segments. The appropriate combination of black spot identification (BSID) method and segmentation method contributes significantly to the reduction in false positive (a site is involved in safety investigation while it is not needed) and false negative (a site is not involved in safety investigation while it is needed) cases in identifying BS segments. The purpose of this research is to study and compare the effect of methodological diversity of road network segmentation approaches on the performance of different BSID methods. To do this, four commonly applied BSID methods (Empirical Bayesian (EB), Excess EB, accident frequency, and accident ratio) have been evaluated against four different segmentation methods (spatial clustering, constant length, constant traffic volume, and the standard Highway Safety Manual segmentation method). The evaluation has applied consistency tests to compare the joint performances of the BS methods and segmentation methods. In conclusion, BSID methods have showed a significant change in their performance depending on the different segmentation method applied. In general, the EB method has surpassed the other BSID methods in case of all segmentation approaches.

4.2 Introduction

The identification of high-risk accident locations or black spots (BS) receives great interest from road agencies and safety specialists. Compared to the large number of studies focusing on the development of various BSID methods and various segmentation methods, considerably fewer researches have been dedicated to investigate their performance. In the previous chapter, a new road network segmentation method has been developed for classifying a large accident data set into
a small number of clusters based on their spatial distribution. Based on the performed overall literature review, it can be concluded that the research field focusing on the combined impact of certain BSID and network segmentation methods on performance is reasonable under researched.

The main assumption of this thesis is that the performance of each BSID method varies in case of different segmentation methods. To examine this, four commonly applied BSID methods (Empirical Bayesian (EB), Excess EB, accident frequency, and accident ratio) have been evaluated against four different segmentation methods (spatial clustering, constant length, constant traffic volume, and the standard Highway Safety Manual segmentation method). Two evaluations have been used to compare the performance of the methods. The approach first evaluates the segmentation methods based on the accuracy of the developed safety performance function (SPF), as described in the previous chapter. The second evaluation applies consistency tests to compare the joint performances of the BS methods and segmentation methods. It is better to point out that the applied spatial clustering segmentation method has some minor changes, as it is going to be described.

### 4.3 Data description

In this part of the research, traffic accident data of different Hungarian Secondary main road sections has been used. The length of the sample network is 1965 kilometers (in one direction) and the data timeframe of the analysis covers the period from 2013 to 2015. The investigated data includes the main parameters of the road cross-section, accident data, road design parameters, and traffic characteristics. In general, the cross-section of the investigated road category can be characterized by 2 lanes in each direction, divided by a median with barriers. The analysis focuses only on road segments; accordingly, intersections data is out of the scope of the investigation (functional area of the intersection which is assumed to be 50 meters long from the intersection center [32]). The investigated dataset includes 2155 fatal and injury accidents related to the given period and the sample network.

### 4.4 Description of segmentation methods

Four different segmentation methods have been used in this chapter. All the used models are based on different classification variables to define homogeneous network sections. The first segmentation method (Seg-1) is the proposed spatial clustering, which includes some minor improvements. The second segmentation method (Seg-2) is based on constant length segments (750 meters). The third method (Seg-3) applies constant AADT segments, whilst the fourth method (Seg-4) is based on the Highway Safety Manual (HSM) specifications, using curvature and AADT. A description of those segmentation methods can be found in "Chapter 3".
The spatial clustering method applies a K-means clustering technique in order to find spatially and temporally coherent accidents. The spatial location is a key classification variable in this method since it has a significantly higher probability that neighboring accidents have strongly related causes than in case of distantly located accidents [39]. The method uses a linear referencing technique to locate accident along a one-dimensional road network measured from a reference point, so dual representation of accidents’ spatial location in the clustering algorithm can be avoided.

The process of segmenting accidents into spatially identified segments is summarized as follows:

- Localizing accident points on a map using GIS.
- Converting geographically referred roads into linearly referred one-dimensional objects.
- Localizing accident objects along linearly referred one-dimensional roads, measured from the RP.
- Applying the measured distances (in the previous step) as a single input for K-means algorithm.
- Identifying the optimal number of clusters K.
- Applying the K-means algorithm to locate all accident clusters
- The result is segments with different lengths and accident contents.

The performed methodological improvements:

- The method applies only the spatial factor, excluded AADT, for the segmentation process. The goal is to keep the focus on the spatial factor that is related directly to accident distribution, without the need for further segmentation that could break the homogeneity of clustered segments and result in numerous zero-sections.

- The method includes zero-accident segments in the analysis. As mentioned in the previous chapter, the proposed segmentation method assumes that zero-segments without any accident history are not expected to be the location of an additional crash, and so the zero-segments can be ignored in the BSID analysis. From the viewpoint of model performance, the involvement of zero segments increases the computational efforts without any significant advantageous impact on model efficiency since the applied segmentation method always results in a relatively fewer number of zeros compared to other approaches. In order to validate the introduced consideration, in contrast to the previous chapter, zero-segments are included in this analysis, and finally the results are compared.
4.5 Description of BSID methods

Four commonly applied BSID methods have been tested and compared. Each method varies in the criteria and variables used to find BS sites, as described below.

4.5.1 Empirical Bayesian

Accidents are random events; their frequencies fluctuate naturally over time and space. This can result in a biased selection of sites that need treatment, known as regression to the mean (RTM). RTM bias occurs when sites are selected for treatment accident frequencies tend to regress in a long-term. In other words, when a comparatively high number of accidents are observed for a given period, it is more likely that fewer accidents will occur in the following period.

The EB method takes into consideration the RTM phenomenon by applying both the observed and predicted accidents' frequencies, for a specific road network element, in one statistical model (Equation 4.1).

\[
N_E = w\times N_P + (1 - w)\times N_O
\]  

(4.1)

The expected number of crashes \((N_E)\) can be used to estimate the expected average crash frequency for both future and past periods, if both observed \((N_O)\) and predicted \((N_P)\) number of accidents are available. The weight factor \((w)\) in the Equation (4.1) represents the degree of reliability in obtaining \(N_P\), and it is inversely proportional to the over-dispersion parameter. The over-dispersion parameter measures the degree of disperse in obtaining \(N_P\) between the given study years. Therefore, if the resulted value of the \(N_P\) is more dispersed it will have a lower weight in the EB Equation (4.1), and vice-versa. However, it can be noted that the crucial parameter in the last equation (Equation 4.1), to some extent, is \(N_P\). The greater the reliability in the development of \(N_P\) models (for the study years) in predicting accidents, the more weight it has in the equation. The predicted average crash frequency \((N_P)\) can be defined using the Safety Performance Function (SPF) for the study period under the given conditions. SPF is a regression equation that estimates the average crash frequency for a given location. The HSM has developed a number of SPFs for three different types of roads: rural two way two lane, rural multilane highway and urban and suburban arterial.

When calculating SPF the segmentation is also another crucial issue. The recommended HSM segmentation method for each specific road category is applied. The generated segments are homogenous in AADT, number of lanes, median type, and presence of curves, as recommended by HSM specifications. The data sets of the resulted segments, including segment length, AADT, accident frequency, and other parameters if necessary can be used for predicting the EB value (Equation 4.1) after the predicted value \((N_P)\) is measured using the standard SPF (Equation 4.2) [20].
\[ N_p = \exp(a + b \times \ln(AADT) + \ln(L)) \] (4.2)

Where: \( a \) and \( b \) are regression parameters, their coefficient values are affected by the type of road (i.e. number of lane, median type) and type of collision.

4.5.2 The other BSID methods

- Excess Empirical Bayesian method (EEB): The expected accident frequency value resulted by the EB can be used to calculate the EEB, which is the difference between the estimated EB value and the predicted accident frequency value given by the SPF for a given site as shown in Figure 4.1. The same figure (4.1) shows the graphical distribution of the observed accident frequency, predicted accident frequency by the SPF, and the expected number of accidents with respect to the AADT.

![EEB estimate](image)

Figure 4.1 EEB estimate [91]

- The magnitude of the excess EB gives an idea about the potential safety savings at particular locations, therefore it is used for preparing ranked lists of sites for safety interventions. When the EEB value is greater than zero, a site can be characterized by more accidents than expected. When the EEB value is less than zero, a site can be characterized by fewer accidents than expected.

- Accident frequency (AF): AF is a straightforward method. This method ranks BS sites in a descending order based on the observed number of accident per unit length during a given period (e.g. 1 year). However, this method only considers the number of accidents in defining the risk level of a given infrastructure component without taking into account other important safety related variables.

- Accident rate (AR): AR is a complementary method for AF that considers traffic volume or entering vehicle volume as well. The result of this method is the descending order of the evaluated infrastructure components based on their
estimated accident rank values. This method enjoys broad applications by traffic agencies and road safety researchers. The output of this method still refers to historical data, so its applicability in prediction is frequently disputed.

4.6 Performance evaluation tests

The performance of the four BSID methods (EB, EEB, AF, and AR) in identifying BS sites has been tested and compared based on different segmentation methods by applying the method of consistency tests. The consistency test method includes four quantitative evaluation tests [85], [86]: the site consistency test, the method consistency test, the total rank differences test, and the total score test.

4.6.1 Site consistency test

The site consistency test is applied to measure the ability of a BSID method to identify consistently high-risk sites over the successive observation periods. The test is based on the premise that a site which is identified as a high-risk site in the first period should also possess lower safety performance at a later period if the given infrastructure component has remained more or less unchanged compared to its baseline state. In a statistical perspective, the site consistency test assumes that the more efficient a BSID method is, the more the estimated future accident numbers (period 2) of the high-risk sites correlate to the current data (period 1). The site consistency test (SCT) can be presented in the following equation.

\[
SCT = \sum_{n} C_{k,i(i)} > \sum_{n} C_{k,i(i+1)}
\]

Where:

n= number of sites being investigated (i.e. top 5%, 10% of BS).
C= total number of accidents for n sites.
k= the applied BSID method (i.e. EB, EEB, AF, AR).
i= observed period, while i+1 represents the next period.

4.6.2 Method consistency test

The method consistency test is designed to evaluate BS methods by measuring the number of observed road segments that have been identified as high-risk sites in successive periods (i.e. period i and period i+1). The higher the rate (compared to the total number of segments considered) of sites identified as high-risk sites in both periods, the more consistent the BSID method performed. Thus, a good BSID method is the one that identifies the same high-risk sites over the two periods. The two
periods are assumed to be close in time to avoid any changes in the traffic or operational characteristics of the road.

Analytically, the method consistency test (MCT) is simply the intersection of all ranked sites \( n \) (i.e. \( n1, n2, .., n \)) identified in the subsequent periods \( i \) and \( i+1 \), or,

\[
MCT = \{n_i, n_{i+1}, ..., n_i\} \cap \{n_i, n_{i+1}, ..., n_{i+1}\}
\]

(4.4)

Only investigated sites (i.e. top 5%, 10% of BS) are compared.

### 4.6.3 Total rank differences test

The total rank differences test is primarily based on the method consistency test. It takes into account the rank of the high-risk sites in both periods. The total rank differences test calculates the sum of rank total differences of the high-risk sites identified over the two periods. Therefore, contrary to the two previous tests, the higher the test value is, the less efficient the BSID method is.

Accordingly, the total rank differences test (TRDT) equation is given as

\[
\sum_{n=1}^{n} (\mathcal{R}_n)_{k(i)} - (\mathcal{R}_n)_{k(i+1)}
\]

(4.5)

The equation gives the total differences in rank \( \mathcal{R}_n \) for all of the investigated sites between periods \( i+1 \) and \( i \) for a given BSID method.

### 4.6.4 Total score test

The total score test (TST) aims to represent the aggregated result of the previous three tests in a single value, assuming that all tests have the same weight. Accordingly, the test can be represented by the following formula:

\[
TST = \frac{100}{3} \times \left[ \frac{SCT}{\max(SCT)} + \frac{MCT}{\max(MCT)} + \left(1 - \frac{TRDT - \min(TRDT)}{\max(TRDT)} \right) \right]
\]

(4.6)

The total score test facilitates making a comprehensive comparison of the BSID methods, considering all types of the mentioned consistency test approaches. The closer the total score test value of a BSID method is to 100%, the more reliable the given analyzed BSID method is.
During the consistency calculation method, the values of relative accident indicator have been defined proportionately to unit length and to a year-long period. The values of the consistency tests have been expressed proportionately to the number of segments. This can ensure an equivalent comparison between BSID methods applied to any segmentation method, as will be shown later in the “Results” section.

### 4.7 Results

The proposed K-means clustering technique (Seg-1) and the other segmentation methods; constant length (Seg-2), constant AADT (Seg-3) and the standard HSM approach (Seg-4), have been applied for segmenting road accident networks. The clustering method has resulted in a significantly lower number of segments (i.e. 304) with a higher average segment length (i.e. 6.76 km) in comparison with the other segmentation methods in case of the investigated roads regarding the analyzed period (2013–2015). Table 4.1 gives descriptive statistics of the applied segmentation methods.

<table>
<thead>
<tr>
<th>Parameters description</th>
<th>Seg-1 (K-means clustering)</th>
<th>Seg-2 (Constant length)</th>
<th>Seg-3 (Constant AADT)</th>
<th>Seg-4 (HSM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min length (km)</td>
<td>0.22</td>
<td>0.75</td>
<td>0.20</td>
<td>0.42</td>
</tr>
<tr>
<td>Max length (km)</td>
<td>19.32</td>
<td>0.75</td>
<td>18.70</td>
<td>15.53</td>
</tr>
<tr>
<td>Average length (km)</td>
<td>5.12</td>
<td>0.75</td>
<td>5.39</td>
<td>2.37</td>
</tr>
<tr>
<td>Standard deviation (km)</td>
<td>4.08</td>
<td>0.00</td>
<td>4.33</td>
<td>1.49</td>
</tr>
<tr>
<td>Total length (km)</td>
<td>1965</td>
<td>1965</td>
<td>1965</td>
<td>1965</td>
</tr>
<tr>
<td>Number of segments</td>
<td>304</td>
<td>2620</td>
<td>362</td>
<td>829</td>
</tr>
<tr>
<td>Accident frequency (fatality+injury)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>13</td>
<td>6</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>Mean</td>
<td>2.6</td>
<td>0.3</td>
<td>2.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Standard deviation (fatality+injury)</td>
<td>2.6</td>
<td>0.6</td>
<td>2.1</td>
<td>1.1</td>
</tr>
<tr>
<td>Total</td>
<td>2155</td>
<td>2155</td>
<td>2155</td>
<td>2155</td>
</tr>
<tr>
<td>AADT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>867</td>
<td>827</td>
<td>882</td>
<td>827</td>
</tr>
<tr>
<td>Max</td>
<td>22163</td>
<td>56140</td>
<td>55677</td>
<td>56176</td>
</tr>
<tr>
<td>Mean</td>
<td>6724</td>
<td>6473</td>
<td>7885</td>
<td>6651</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3499</td>
<td>4373</td>
<td>5680</td>
<td>4742</td>
</tr>
<tr>
<td>Horizontal Deflection Angle (HDA) (per km)</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Mean</td>
<td>89</td>
<td>106</td>
<td>102</td>
<td>124</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>72</td>
<td>148</td>
<td>96</td>
<td>146</td>
</tr>
<tr>
<td>Speed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>40</td>
<td>30</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>Max</td>
<td>90</td>
<td>100</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>Mean</td>
<td>76</td>
<td>74</td>
<td>72</td>
<td>74</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>13</td>
<td>18</td>
<td>15</td>
<td>18</td>
</tr>
</tbody>
</table>
After segmenting the road network with four different models, it is required to compare the efficiency of each model and to evaluate their impact on BSID methods. Beside the SPF, another evaluation test has been used. Both tests results are presented below.

4.7.1 Result of the SPF development process

The four segmentation approaches have been applied to develop a new process for the input of SPF. The stepwise approach has been used to check the significance of inserting or removing different explanatory variables in case of each SPF. Finally, AADT, speed, and HDA have been selected to be the input variables of the SPF's. Accident data from the first years (2013-2014) of the study sample have been used to develop the models, while data of the year 2015 has been used for checking the performance of the developed models.

According to Table 4.1, accident frequencies in all of the applied methods, show higher variance values (squared standard-deviation) compared with their mean. This suggests some possible presence of over-dispersion. Therefore, the negative binomial distribution has been used for modeling accident frequency. The model calibration results for all segmentation methods are presented in Table 4.2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Method of segmentation</th>
<th>Seg-1 (K-means clustering)</th>
<th>Seg-2 (Constant length)</th>
<th>Seg-3 (Constant AADT)</th>
<th>Seg-4 (HSM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α (Intercept)</td>
<td>[p-value]</td>
<td>-4.721 [0.056]</td>
<td>-6.215 [&gt;0.001]</td>
<td>-3.579 [0.010]</td>
<td>-6.865 [&gt;0.001]</td>
</tr>
<tr>
<td>β1 (AADT)</td>
<td>[p-value]</td>
<td>0.625 [0.086]</td>
<td>0.553 [&gt;0.001]</td>
<td>0.670 [&gt;0.001]</td>
<td>0.675 [&gt;0.001]</td>
</tr>
<tr>
<td>β2 (Speed)</td>
<td>[p-value]</td>
<td>-0.625 [0.086]</td>
<td>-0.269 [0.076]</td>
<td>-0.871 [&gt;0.001]</td>
<td>-0.288 [0.036]</td>
</tr>
<tr>
<td>β3 (DOA)</td>
<td>[p-value]</td>
<td>0.184 [0.007]</td>
<td>0.148 [&gt;0.001]</td>
<td>0.058 [0.023]</td>
<td>0.216 [0.001]</td>
</tr>
<tr>
<td>Over-dispersion</td>
<td></td>
<td>1.387</td>
<td>1.717</td>
<td>1.108</td>
<td>1.096</td>
</tr>
<tr>
<td>QIC</td>
<td></td>
<td>324</td>
<td>1721</td>
<td>390</td>
<td>967</td>
</tr>
<tr>
<td>PCC</td>
<td></td>
<td>0.692</td>
<td>0.198</td>
<td>0.561</td>
<td>0.462</td>
</tr>
</tbody>
</table>

55
Most of the estimated parameters (i.e. $\alpha$, $\beta_1$, $\beta_2$ and $\beta_3$) of the SPFs presented in Table 4.2 are statistically significant to 95% confidence level and the rest to 90% confidence level. By evaluating the SPFs in terms of data fitting, it is evident from Table 4.2 that the Seg-1-based method shows greater potential for fitting accident data; this is obvious from its significantly lower QIC value ($QIC_1=324$). In contrast, Seg-2-based method is rather over-fitting the developed model, with the highest QIC value ($QIC_2=1721$).

The Pearson Correlation Coefficient (PCC) method has also been used to evaluate and compare the developed SPFs based on measuring their efficiency in predicting accident data for another year (2015). The values of the PCC are shown in Table 4.2. It is noted that the SPF developed by the Seg-2 method shows weakness in predicting accident frequency with the lowest PCC value (i.e. 0.20). Contrary to this, the proposed, clustering segmentation (Seg-1)-based model gives the best correlation between the predicted and observed data. This is described by its highest PCC value which is close to 0.7.

The comparison of the changes applied to the proposed spatial segmentation method in this and the previous chapter can lead to some important remarks. In the case of the proposed spatial clustering (Seg-1) method, applying zero-accident segments (Table 4.2) has not affected its general performance, compared to the results presented in Table 3.5. In both cases, the model still presents the best performance among the other segmentation-based models. Moreover, the same can be concluded by ignoring the AADT from the segmentation process.

### 4.7.2 Consistency tests results

Consistency tests have been fed by relative accident data referring to the unit long section of the segments and of a one-year long period, as described previously, to achieve a comprehensive comparison between any BSID method and other segmentation methods. Consistency test results for the top 5% and 10% BS sites are described in the next section.

#### 4.7.2.1 Site consistency test

It is shown in Table 4.3, that from the point of view of the site consistency test, the best-performing models have been the EB and AF methods with reasonably close results. In the case of Seg-1, the AF method performs the best in identifying the top 5% and 10% of BS sites. Accordingly, the AF method has produced the highest equivalent accident frequencies for the same BS sites in case of the top 5% and 10% of BS sites (1.41, 1.14 respectively) between the compared baseline period (2013) and predicted period (2014–2015). On the other hand, the EB method has also been able to excel in identifying the top 5% of BS sites, with a value of 1.24 for the spatial segmentation method (i.e. Seg-1), and also the top 10% of BS sites in case of Seg-1 and Seg-3, with 1.03 and 1.10.

In general, the highest single value of the site consistency test for all segmentation methods and all BSID methods has been provided by the common application of the
AF method and the Seg-1 method in both the top 5% and 10% of high-risk sites, with 1.41 and 1.14 equivalent accident frequencies. Contrary to this, the AR method performs worst in identifying the BS sites from all segmentation methods, especially for Seg-4 method, resulting in the lowest accident frequencies among all BSID methods, with 0.6 accidents for both the top 5% and 10% BS sites.

Table 4. 3 Site consistency test results

<table>
<thead>
<tr>
<th>BS method</th>
<th>Segmentation method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seg-1</td>
</tr>
<tr>
<td>5% risk</td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>1.24</td>
</tr>
<tr>
<td>EEB</td>
<td>1.20</td>
</tr>
<tr>
<td>AF</td>
<td>1.41</td>
</tr>
<tr>
<td>AR</td>
<td>0.93</td>
</tr>
<tr>
<td>10% risk</td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>1.03</td>
</tr>
<tr>
<td>EEB</td>
<td>1.02</td>
</tr>
<tr>
<td>AF</td>
<td>1.14</td>
</tr>
<tr>
<td>AR</td>
<td>0.88</td>
</tr>
</tbody>
</table>

4.7.2.2 Method consistency test

In the method consistency test, the EB method has outperformed the other BSID methods in most cases for the two study periods, based on the number of similar BS sites as shown in Table 4.4. In the case of the constant length segmentation method (Seg-2), the EB model has been the superior in both the 5% and the 10% risk cases, by identifying 35% and 39% of correlated BS sites between the two periods. Similarly, the highest test values in case of the Seg-4 method have also been provided by the EB method for the top 5% and 10% high-risk sites (34% and 43%). In the case of the Seg-1 method, the AF method has achieved the maximum values. This practically means 53% and 54% of sites characterized by similar accident frequency in case of the top 5% and 10% of BS sites, in case of the AF method.
The best results of the method consistency test have been produced by the Seg-1 method and have been achieved by the EB and AF methods, as presented in Table 6.4. The lower test results have been provided by the other BS methods strongly depending on the segmentation method applied. In the case of Seg-2 and Seg-4 methods, AR has performed the worst, whereas the EEB method has been the last BSID method for Seg-3 and Seg-4.

Table 4. Method consistency test results

<table>
<thead>
<tr>
<th>BS method</th>
<th>Segmentation method</th>
<th>Seg-1</th>
<th>Seg-2</th>
<th>Seg-3</th>
<th>Seg-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% risk</td>
<td>EB</td>
<td>0.47</td>
<td>0.35</td>
<td>0.33</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>EEB</td>
<td>0.41</td>
<td>0.28</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>0.53</td>
<td>0.23</td>
<td>0.33</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>0.29</td>
<td>0.23</td>
<td>0.33</td>
<td>0.27</td>
</tr>
<tr>
<td>10% risk</td>
<td>EB</td>
<td>0.45</td>
<td>0.39</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>EEB</td>
<td>0.45</td>
<td>0.33</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>0.54</td>
<td>0.30</td>
<td>0.39</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>0.37</td>
<td>0.24</td>
<td>0.36</td>
<td>0.34</td>
</tr>
</tbody>
</table>

4.7.2.3 Total rank differences test

Table 4.5 illustrates that the EB method has significantly smaller total differences for all applied segmentation methods, which makes it the superior method in this test also followed directly by the AF methods. The lowest values of the EB method have resulted in the case of Seg-1 and Seg-3 methods for both the top 5% and 10% high-risk sites, respectively, with relatively close values, whereas, the poorest test values have been provided by the combined application of the Seg-2 method and the EEB model.
4.7.2.4 Total score test

The total score test combines the results of the previous three consistency tests (site consistency test, method consistency test, and total differences test).

Table 4.6 illustrates the values of the total score test that describe the total efficiency of each BSID method with respect to every segmentation method applied, in case of the two investigated cases (5% and 10% BS sites). The column titled “Total”, presented in Table 4.6, represents the aggregated average performance of the BSID methods considering all the investigated segmentation models. The row titled “Total”, presented in Table 4.6, represents the aggregated average performance of the segmentation models considering all the BS methods investigated.

In the analysed two cases (5% and 10% BS sites), the EB and AF methods have performed better than the other BSID methods. Generally, the best total score test values are included by Seg-1 column (see Table 4.6), except the AR model which has performed better with the constant AADT and HSM segmentation methods (Seg-3 and Seg-4 respectively). Contrary to this, most of the BSID processes which have been based on Seg-3 or Seg-4 methods, are performed reasonably worse. This is obvious, as the lowest total average performance of all BSID methods has been caused by Seg-4 (72.2) for the 5% high-risk sites and by Seg-3 (72.1) in case of the
top 10% of the high-risk sites. Besides this, the AR method has provided the lowest average performance between the BSID methods considering all segmentation models.

<table>
<thead>
<tr>
<th>BS method</th>
<th>Segmentation method</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seg-1</td>
<td>Seg-2</td>
</tr>
<tr>
<td>5% risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>92.2</td>
<td>84.4</td>
</tr>
<tr>
<td>EEB</td>
<td>77.4</td>
<td>77.8</td>
</tr>
<tr>
<td>AF</td>
<td>93.1</td>
<td>72.8</td>
</tr>
<tr>
<td>AR</td>
<td>57.8</td>
<td>72.7</td>
</tr>
<tr>
<td>Total</td>
<td>80.1</td>
<td>76.9</td>
</tr>
<tr>
<td>10% risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>96.8</td>
<td>81.9</td>
</tr>
<tr>
<td>EEB</td>
<td>82.0</td>
<td>77.0</td>
</tr>
<tr>
<td>AF</td>
<td>96.3</td>
<td>74.1</td>
</tr>
<tr>
<td>AR</td>
<td>72.6</td>
<td>69.5</td>
</tr>
<tr>
<td>Total</td>
<td>86.9</td>
<td>75.6</td>
</tr>
</tbody>
</table>

4.8 Discussion of the results

The previously presented results evaluate the performance of the introduced four commonly applied BS identification methods in case of the four analyzed road network segmentation methods, using accident data from the Hungarian secondary main road network.

The outputs of the applied segmentation methods have been fed into different accident prediction models (i.e. SPFs). Also, consistency tests have been used to evaluate and compare the performances of the four BSID methods in the localization of high-risk
accident sites. Based on the prepared analysis, the following conclusions can be drawn:

- The performance of SPFs differs, depending on the applied segmentation method. This result is supported also by the work of Ghadi & Török [113] revealing different goodness-of-fit in case of the analysed SPFs related to five different segmentation methods.

- Therefore the performance of some BSID methods, e.g. EB and EEB, which apply SPFs, is also significantly affected by the segmentation method applied in the model, as also approved by Kwon et al. (2013) [91].

- The performance of each BSID method varies in case of the different segmentation methods.

- The performance of the BSID methods, has been the best in the consistency tests when the spatial segmentation (Seg-1) method has been applied to road accident data (see Table 4.6) [116], except the case of AR. This result is also supported by the SPF test (Table 4.2), where the datasets classified by the Seg-1 method has produced the best fitting SPF [113]. This is logical since the spatial segmentation models classify closely located accidents in the same groups which have higher probability to have similar causes than in case of the distantly located accidents, as analysed by Flahaut et al. [39].

- The EB methods have performed better than the other BSID methods in most tests. This is supported also by theoretical basis, since EB method uses the observed and the predicted values in its statistics, which in turn increases the reliability of its results. In practice, the EB method has proved its efficiency in many studies [2], [82], [86], [117].

- The second best BSID method has been the AF, especially when it is applied with the spatial-based segmentation method (Seg-1). The performance of the AF method has been fairly close to the EB method and slightly outperformed it in some cases. Although, its performance has been poor in case of the Seg-4 method. Similar results have been obtained by Cheng and Washington (2008) [85].

- The EEB and AR methods seem to be reasonably inconsistent in most of the cases. The EEB method is largely affected by the predicted value and consequently the validity of the developed SPF. As the predicted value of accident increases, the likelihood of a site being selected as a BS increases. The main drawback of the AR method is that it incorrectly assumes a linear relationship between accident frequency and traffic volume, since accident intensity can be strongly influenced by other factors, like speed [76]. This fact must be taken into consideration in case of using AR in road safety investigations. These results are also consistent with the results of other researches [86], [118].
4.9 Conclusion

This chapter investigates and compares the influence of applying different roadway segmentation methods on the performance of different black spot identification (BSID) methods. Beside this the research has examined the applicability of the newly developed segmentation method and its contribution to the performance of other BSID methods.

Four different BSID methods have been evaluated against four different segmentation methods. The first, proposed by the author, segmentation method (Seg-1) applies clustering techniques to segment accident data based on their spatial convergence. The second and third methods (Seg-2 and Seg-3, respectively) apply a constant length and constant average annual daily traffic (AADT) segments, respectively, while the last method (Seg-4) is based on the specifications of the Highway Safety Manual (HSM), using curvature and AADT data. In the case of BSID methods, the Empirical Bayesian (EB), Excess EB (EEB), accident frequency (AF), and accident rate (AR) methods have been applied. Two tests have been used to compare the performance of the methods in the case of Hungarian Secondary roads. In the first test, the performance of safety performance functions (SPFs) has been statistically compared in light of the applied segmentation models. In the second part of the investigation, four consistency tests (site consistency test, method consistency test, total rank differences test, and total score test) have been used to compare the joint performances of the BSID methods and segmentation models.

In the first test, which uses the same strategy as the previous chapter, the performance of the segmentation methods has been compared. The best goodness-of-fit value has been obtained by the proposed spatial clustering segmentation model (Seg-1). The obtained result is consistent with the result of the previous chapter. This is rational since in this case numerous factors have been helped to optimize SPF calibration, like the consistency of the segmented datasets, the flexibility of segment length, quality, and other variables. This result has been supported by the second test, which has confirmed that all BSID methods, except AR method, have reached their best performance in the case of the spatial clustering segmentation (Seg-1) method. Beside this, the consistency evaluation tests have also shown that the EB method performed better than the other BSID methods in most of the cases. The test results highlight that the EB method is the most consistent method for identifying priority investigation sites. The AF has directly followed the EB, although it slightly outperformed in a few cases.

Beside this, another interesting result of the consistency tests is the performance variation of BSID methods in case of the different applied segmentation methods. For instance, the AF method has recorded its best performance, with 93.1% consistency, in case of the Seg-1 method, while it has achieved only 66.9% consistency, as a third performance, in the case of the Seg-4 methods, for the top 5% of the studied black spot sites (as presented in Table 4.6). In general, the performance of the EEB and AR
methods has been the weakest in most segmentation cases. This is quite alarming, as many highway agencies and researchers use these methods.

In general, the results of this research try to highlight that data segmentation is an important and fundamental step in road accident data classification and configuration for subsequent operations. This has been demonstrated, in this chapter, by the different performance obtained by the BSID methods with different applied segmentation methods. However, testing other segmentation and BSID methods and comparing them with more reliable approaches will have an increasing relevancy in future research.

**Possibility for implementation in practice**

This thesis proposes a new methodology that can assist experts in selecting the proper BSID and segmentation method combined with the best joint performance. The thesis also calls the attention of the experts that the application of some BSID (e.g. accident rate) and/or segmentation methods (e.g. constant length) can result in misleading conclusions.

4.10 Thesis

- **Thesis (3)**

  I have revealed the effect of methodological diversity of road network segmentation on the performance of different black spot identification methods by adapting a methodology that is able to compare the performance of road segmentation methods, black spot identification methods, and the joint performance of both methods.

  I have demonstrated that the developed spatial clustering segmentation method can improve the performance of different black spot identification methods. The best result was yielded using the spatial clustering segmentation technique in combination with the Empirical Bayes method.

**Related publications to Thesis 3:** [116], [119]


Chapter 5

Spatial clustering segmentation approach: practical applications

5.1 Short summary

Previous chapters of the dissertation have introduced and examined the benefit of applying a new spatial clustering approach for segmenting road accident networks. The proposed spatial segmentation method can be used for classifying a large and complex accident dataset along a road network into smaller homogeneous groups based on their spatial distribution. The new segmentation model maximizes the similarity of accidents within the same group. The advantage of this method enables it to get more reliable results when applied with BSID methods.

This chapter proposes two innovative approaches to the application of the spatial segmentation model. In the first application a complex methodology is applied to analyze BS locations by combining the benefit of both; Empirical Bayes method and the spatial clustering approach. In the first step, the developed spatial clustering segmentation method has been used to define homogeneous accident clusters. Then, the Empirical Bayesian method has been applied to define BS locations based on the determined clusters. Due to the combination of the introduced methods, a powerful technique is provided that is able to identify high-risk locations and cluster the related sections along the road as the output of the model.

The second application aimed to examine the main traffic accident factors that affect the severity of road segments. The practical objective is to assist specialists in identifying risk patterns both from a spatial and casualty point of view. To achieve the desired goals, accidents of the investigated road network have been analyzed through three major steps: segmentation, black spot identification, and decision analysis. The first two steps follow similar procedures in identifying clusters characterized by high risk and homogeneity. Beside this, in the methodology, the technique of decision rules has been applied to identify the main contributors of accidents in certain segments. The results show that there is a significant relationship between road geometric and traffic characteristics (i.e. speed limits, average daily traffic, road alignment) of road segments and their impact on road safety. The results have also revealed that the higher the road segment design level is, the lower risk the road segment has.
5.2 Introduction

Reducing traffic accidents on roads is the final objective of any road safety strategy. The main difficulty of analyzing accident data is its heterogeneity. Typically, road network segmentation is used to reduce heterogeneity and reveal hidden relationships between accident data. The more accurate the segmentation and classification of accidents are, the better safety strategies can be built and this can significantly improve the road safety.

There is a fairly extensive literature focused on developing or applying different black spot (BS) identification methods. Elvik [120] has evaluated and compared the applied BS methods in a number of European countries. He has proved the weakness of most of the applied BS methods in these countries and has recommended using Empirical Bayesian (EB) method, as applied in numerous researches [86], [117], [119]. The EB is a state of the art BS method that combines the use of both observed and predicted accident frequencies. Now it has also proved that the efficient application of EB method makes it reasonable to couple it with other approach that improves the process of generating homogeneous segments.

Clustering techniques are among those most important data-mining approaches that can be used in finding hidden relationships and patterns of a large number of accidents [121], [122], and in classifying them according to the similarities in their attributes and spatial distributions. Depaire et al. [47] have used a latent class generation method to identify homogeneous traffic accident types. They suggest clustering as a preliminary analysis, which can reveal hidden relationships and also can help in generating homogeneous road segments. Practically, homogeneity of accident classes could have many definitions considering the type of accidents, severity, spatial distribution, type and number of involved vehicles, driver characteristics and many other issues. The results of the clustering can vary significantly depending on the applied classification method [105]. Based on the performed evaluations it has been proved that those road network segmentation models which are fitted to the characteristics of input variables can result in more efficient BSID methods. Most of the current studies focus only on the identification of factors related to accident circumstance as input variables for the segmentation process. Only few of them classify the accidents based on their spatial dependence. In both cases, a data mining framework has been applied as a tool to explore the required information.

Data Mining is a process for discovering hidden patterns from a large data set. Its main goal is to collect useful information from a heterogeneous data set after classifying it into homogeneous structures. Accordingly, data mining techniques can be used to segment and investigate road accident data [54]. Decision tree (DT) is a popular supervised data mining technique to analyse and classify accident related factors. This technique is very attractive, since it does not need any certain assumption or predefined hypothesis regarding the relationship between dependent and explanatory variables, which are required in case of the traditional statistical
techniques. DT is based on decision rules (DR) to map the most important influencing factors related to the investigated process. Abellán et al. [58] have applied a methodology to define DR from more DTs to enable more effective interaction between accident attributes, focusing on accident severity as a label variable. Kashani et al. [62] have applied the classification DT to identify the important factors influencing injury severity of drivers involved in traffic accidents in Iran. These researchers usually combine the supervised and unsupervised data mining techniques for a complete data analysis.

The spatial segmentation method is a proposed methodology for classifying a large and complex accident dataset along a road network into smaller homogeneous groups based on their spatial distribution in which the similarity between accidents of the same group is maximum. Accordingly, this chapter proposes two practical applications for the proposed spatial clustering segmentation method. The first part of this chapter examines the practical utility of applying the developed spatial clustering segmentation model as a well applicable complementary approach to EB method. The hypothesis assumes that the practical application of both methods together could improve the process of BSID. The second part of the chapter tries to link the spatial factor and other related attributes during the classification process of road accident data. According to the hypothesis road segments with similar risk properties can be accurately identified and discovered, depending on the characteristics of their accident contents. The outcome can strongly support specialists in identifying risk patterns along the road network.

**Part one:**

Integration of probability and clustering based approaches in the field of black spot identification

### 5.3 Data description

In the case study, the motorway road M-3 in Hungary has been examined. In accordance with the Hungarian specifications, the motorway is a high-speed road with 130 km/hour speed limit, and has two lanes in each direction with an extra emergency lane. This road has a total length of 281 km and it connects the capital Budapest with Nyíregyháza city located in the northeastern part of the country. Each accident occurred on this road from the year 2013-2015 is recorded regarding its geographical coordinates, type and severity (excluding ramp and intersection accidents). Traffic volume (i.e. AADT) for different sections of the road is also included in the dataset for the same period. The data includes fatal, serious, and slight casualties' accidents.
5.4 **Methodology**

The objective of this part is to define a complex methodology to analyze black spot locations of road infrastructure network combining the benefit of both; Empirical Bayesian method and K-means clustering approach (described in chapter 4). In the first step, the developed spatial clustering segmentation method has been used to define homogeneous accident clusters (refer to section 3.4). After applying the K-mean clustering to divide the road into segments, it is required to evaluate the risk level of the resulted segments in order to identify BS sections on the road. The EB method is considered as a state-of-the-art technique for doing that.

The process of identifying BS locations using EB method is started by dividing the road into segments, and followed by calculating the predicted number of accidents by using the formula of Equation 4.1 and 4.2. Each segment is characterized by the number and type of accidents, AADT and length of the given section. Finally, the excess risk derived by the EB method is calculated as the difference between the numbers of expected and predicted accidents. If the expected number of accidents is larger than the predicted number of accidents then the given location can be indicated as a BS [20]. This technique also allows us to arrange the BS segments according to the degree of dangerousness in which higher excess value indicate higher severity.

5.5 **Results and discussion**

A spatial cluster analysis has been applied initially in order to divide road accident data into homogeneous groups based on their spatial inter-correlation assuming homogenous traffic and geometric conditions. To do so, the linear referencing method has been applied by defining the distances of the geographically distributed accidents from the starting point along the M-3. In the next step it has been required to identify the input number of cluster for the K-means algorithm. In order to find out the optimal number of clusters an initial value is required. In our study, it is assumed that the average segment length is 10 km, in this case the number of clusters along the M-3 (281 km length) is about 30 (281/10≈30). Accordingly, the initial number of clusters has been selected to be 30. Starting with this value, the marginal change in the SSE (Equations 3.1 and 3.2) has been examined in light of the unit increase or decrease in the number of clusters (referring to section 3.4). Then, the optimal number of clusters has been identified (See Figure 5.1).
Figure 5.1 Defining the value of the optimal number of clusters based on the changes in the SSE (variance)

It can be noted from the figure (5.1) that as the number of clusters \( k \) is increasing the variance is decreasing until it achieves the zero when every accident has its own cluster. So, the selected point is there, where the difference quotient of function values and domain values changes an order of magnitude (becomes small enough - e.g. when its absolute value is less than 1). Up to this point each added cluster results in a substantial reduction in the value of variance but after that point, increasing \( k \) will result marginal reduction in the value of variance. As mentioned above, this consideration is similar to evaluation methods of the second derivative of the variance as a function of the number of clusters (indicating how much the difference changes as another cluster is added). In accordance with the result of the analysis, the value of 40 appears to be the proper cluster number.

The resulted number of accidents within each cluster and its corresponding segment length are presented in Table 5.1. In this study, only clusters with accidents have been included in the analysis. In contrast, empty segments (between the resulted clusters) with no accident history have been assumed to be safe, to some extent, so they are excluded from any further analysis. The next step is to evaluate the resulted segments, using the Excess EB approach to calculate the excess number of accidents for a given section. Because the size of the available data is small, the accident prediction models recommended by the HSM [20], have been used to identify the coefficients of the SPF (referring to Equation (4.2): \( a=-12.34, b=1.36, \) and \( k(\text{overdispersion})=1.32 \)) and values of weight factors, for each specific road category. The positive excess risk segments are indicated in Table 5.1.
Table 5. 1 Cluster information and the final BS identification process

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Observed number of accidents</th>
<th>Location Section-Km (Start point)</th>
<th>Length (km)</th>
<th>Excess EB</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>3.6</td>
<td>2.78</td>
<td>-29.10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>211.9</td>
<td>9.27</td>
<td>-14.28</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>89.8</td>
<td>6.92</td>
<td>-17.16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>51.2</td>
<td>5.58</td>
<td>-24.84</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>15.4</td>
<td>10.44</td>
<td>-42.42</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>46</td>
<td>171.2</td>
<td>6.94</td>
<td>-7.20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td>248.0</td>
<td>4.36</td>
<td>3.71</td>
<td>BS</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>101.9</td>
<td>0.06</td>
<td>0.87</td>
<td>BS</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>81.8</td>
<td>4.36</td>
<td>-6.30</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>34</td>
<td>34.0</td>
<td>6.05</td>
<td>-29.28</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>34</td>
<td>162.3</td>
<td>8.67</td>
<td>-16.78</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>274.5</td>
<td>11.41</td>
<td>-7.12</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>32</td>
<td>198.4</td>
<td>4.31</td>
<td>0.81</td>
<td>BS</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>179.2</td>
<td>5.29</td>
<td>-9.09</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>148.3</td>
<td>3.09</td>
<td>-2.01</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>42.9</td>
<td>4.39</td>
<td>-18.75</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>14</td>
<td>130.2</td>
<td>4.41</td>
<td>-10.97</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>59.7</td>
<td>3.65</td>
<td>-19.80</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>26</td>
<td>97.5</td>
<td>6.05</td>
<td>-14.02</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>14</td>
<td>107.4</td>
<td>3.17</td>
<td>-7.02</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>44</td>
<td>68.4</td>
<td>6.76</td>
<td>-16.06</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>28</td>
<td>21.1</td>
<td>4.52</td>
<td>-12.12</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>140.8</td>
<td>0.40</td>
<td>0.04</td>
<td>BS</td>
</tr>
<tr>
<td>28</td>
<td>14</td>
<td>54.8</td>
<td>2.91</td>
<td>-12.15</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>119.1</td>
<td>3.88</td>
<td>-7.64</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>22</td>
<td>222.2</td>
<td>8.53</td>
<td>-10.05</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>22</td>
<td>112.4</td>
<td>2.69</td>
<td>-2.61</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>8</td>
<td>26.6</td>
<td>4.15</td>
<td>-25.09</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>24</td>
<td>76.1</td>
<td>5.73</td>
<td>-17.98</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>18</td>
<td>258.6</td>
<td>5.58</td>
<td>0.77</td>
<td>BS</td>
</tr>
<tr>
<td>39</td>
<td>40</td>
<td>283.0</td>
<td>6.58</td>
<td>7.70</td>
<td>BS</td>
</tr>
<tr>
<td>40</td>
<td>24</td>
<td>185.7</td>
<td>5.15</td>
<td>-3.89</td>
<td></td>
</tr>
</tbody>
</table>

Note: cluster 8, 12, 26, 27, 30, 34, 35, 37 are excluded since it represent a single point (crash) cluster

5.6 Summary of part one

Previous researches have demonstrated that K-means algorithm can be a useful tool to classify large data sets in many scientific fields. This study illustrates the benefits of the developed clustering technique in segmenting roads. According to the obtained results, clustering can be a well applicable complementary approach of EB method to
provide an effective and simple BSID technique. Since the proposed spatial clustering method can over bridge some of the deficiencies related to the traditional segmentation methods, it seems to be proper initial methodology component before the risk calculation process. The main advantage of the integrated technique is the ability to narrow the identified segments to those sections, which are really affected by the accidents, through defining network segments based on the spatially coherent crashes. This concept makes it possible to ignore the further analysis of empty and less risky road sections between these clusters. This consequently results in re-sectioning the whole road network according to the spatial distribution of accidents, forming the basis of examining and classifying the risk level of the potential segments.

**Part Two:**

**Introducing a new qualitative-spatial approach to explain road accidents**

### 5.7 Decision analysis

Decision rule (DR) is a very popular data mining technique used in the field of supervised learning. DR helps in recognizing useful information and relations between various attributes in a large dataset. Safety engineers usually apply DR to discover hidden characteristics of road accident datasets [123]. An especially useful DR tool is the Decision tree (DT) method. DR and DT have been widely used for analyzing road accident data [58], [124], [125]. DT includes a set of useful methods which are applicable to uncertain decision processes. DT can be used in the field of classification and regression without any requirement on setting parameters or a prior relationship between dependent and independent variables. These properties make it a useful tool to classify accident data based on the most influencing factors.

The process of extracting DR's from DT is constrained by the structure of the tree. In case of road accidents, classification and regression tree (CART) and C4.5 methods are usually applied to construct the DT [62], [126]. The main differences between the two methods are originated from the structure of the tree and the splitting criteria. CART method uses the Gini index to measure the dispersion among the variables, and it results in a binary tree, while the C4.5 is an algorithm developed by Ross Quinlan used to generate a decision tree based on the Gain ratio [127] based on the entropy measure of probability.

The tree construction [125] started by recursively partitioning the target variable to minimize “impurity” in the terminal nodes using the Gini index criterion. The partitioning is done by searching all possible threshold values for all input variables (splitters) to find the threshold that leads to the greatest improvement in the purity
score of the resultant nodes. At the end of the tree growing, a saturated tree is obtained. The next step is tree ‘‘pruning.’’ Pruning is a mechanism to create a series of simpler trees, through cutting off least important nodes. The last step is to select a tree of the right size from the pruned trees. In this paper, the CART based DT method is used to define DR's from road accident data.

5.8 Data description

The data of the Hungarian network have been analysed in this study; the analysis has covered approximately 1965 km long of Hungarian Secondary Main Roads. During the investigated interval, 2155 fatal, serious and light injury accidents occurred on roadway segments. The investigated data describes main traffic, road geometric and accident parameters from the year 2013 to 2015. Individual accident data includes information on severity level, time of accident (e.g. month, day, hour), type of accident (e.g. pedestrian accident, run off road, etc.), weather condition (e.g. clear, foggy, rainy, snowy), visibility (e.g. daylight, restricted visibility, night with or without public light), and location of accident (e.g. x-y coordinates). Roadway data includes information on road alignment (e.g. vertical and horizontal curvature), roadside hazard, median type, and pavement conditions. Traffic data includes information on speed limit, AADT, and percentage of trucks. The analysis focuses only on the road segments’ accidents, and similarly to the previous chapters, intersections’ accidents are not considered in the evaluation.

5.9 Methodology

Road accidents usually occur as a result of integrating human, traffic and road geometric factors (hereafter referred to accident environment factors) in a particular road segment. The methodology followed, in this study, is divided into three main processes, as shown in Figure 5.2. The first method has been used to segment the road network based on homogeneous clusters of accidents using the spatial clustering technique. The second part has been applied to rank and classify each segment by risk level using the EB method. The last part tries to identify the main contributors of accidents for different segment risk level.
Figure 5.2 Proposed framework for analysing the spatial and causal pattern of road accidents. L is the length of a segment n. N is the number of accident for each road segments RSG

The methodological steps presented in Figure 5.2 can be summarised as follows:

- Localizing of accident points using GIS.
- Localizing of accident objects along linearly referred one-dimensional road objects, measured from the reference point (RP).
- Applying the K-means algorithm to locate all clusters of spatially aggregated accidents and their corresponding segment lengths. The next step is to apply the EB method to rank the resulted segments by risk level, as presented in the previous part of this chapter.
- The dataset of the environment characteristics for each spatially identified road segment has been added, to be used to study the accident patterns and circumstances.
- To better understand the difference in severity, the road segments are divided into four groups of equal size but different risk level. Each group took a particular code; BS1, BS2, BS3 and BS4, and risk level ranked from higher risk (BS1) to lower risk (BS4) in an ascending order.
- DT is then applied to derive the DR from the resulted groups. To verify the quality of the classification process, the data are divided into two sets: training set and test set. The training set has used the 2/3 of the cases and the test set has used the rest of the cases. Since the accident data have been collected sequentially in time, R software has been used to shuffle the data to ensure random selection of the training and testing sets. The training set is used to construct the classification model and test set to assure the accuracy of the model. The CART DT has been formed by some important variables related to individual accident attributes. DR’s have been derived from the constructed CART tree.
5.10 Results and discussions

5.10.1 Road segmentation results

The proposed spatial segmentation method has resulted in 181 segments in case of the investigated road network regarding the analysed time period (2013-2015). The description of the resulted spatial clustering process per segment is summarized in Table 5.2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Total m</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment length (km)</td>
<td>1225</td>
<td>0.23</td>
<td>12.32</td>
<td>6.76</td>
<td>4.57</td>
</tr>
<tr>
<td>Accident frequency</td>
<td>1926</td>
<td>2</td>
<td>13*</td>
<td>4*</td>
<td>2.54*</td>
</tr>
</tbody>
</table>

*The values are averaged per the three study years (2013-2015)

It can also be noted from Table 5.2 that at least two accidents are required to form a cluster. Since it is assumed that a cluster of a single accident that is far enough from any neighbouring accident, spatially and temporally, is more likely to occur randomly and should not be considered in any further analysis. This explains why lower total accident frequency (1926) included in the analysis than the real total number (2155).

5.10.2 Black spots identification results

After segmenting the road based on the spatially referred accident data using the clustering method, the EB method has been applied to identify and rank BS segments. As EB method considers both the observed and predicted accidents frequencies, the predicted values have been estimated using a SPF. Accident data of the years 2013-2014 have been used to develop the SPF for the case study dataset. The stepwise forward approach has been used to check the significance of inserting or removing different explanatory variables of the SPF. Finally, AADT, speed, and degree of horizontal curve have been selected to be the input variables of the SPF. The data of the 181 road segment has been applied to develop the SPF. The model calibration results are presented in Table 5.3.

The goodness-of-fit of the model has been evaluated in term of the Quasi-likelihood under Independence Model Criterion (QIC) and the Pearson Correlation Coefficient (PCC). The model results fits well to the road accident dataset with a low acceptable error as presented in the QIC value (324). Moreover, examining the correlation between the observed and predicted accident data of the year (2015) has resulted in a good correlation characterized by a relatively high positive PCC value (0.692), as shown in Table 5.3.
Table 5.3 Model parameters and goodness-of-fit results.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Frequency</th>
<th>BS1 (29)</th>
<th>BS2 (33)</th>
<th>BS3 (21)</th>
<th>BS4 (17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident Severity</td>
<td>Fatal</td>
<td>143</td>
<td>0.29</td>
<td>0.33</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>622</td>
<td>0.32</td>
<td>0.30</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Slight</td>
<td>1161</td>
<td>0.32</td>
<td>0.29</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>AADT</td>
<td>&lt;3000 (AADT1)</td>
<td>140</td>
<td>0.32</td>
<td>0.11</td>
<td>0.16</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>3000-10000 (AADT2)</td>
<td>1296</td>
<td>0.26</td>
<td>0.29</td>
<td>0.24</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>&gt;100000</td>
<td>490</td>
<td>0.47</td>
<td>0.37</td>
<td>0.15</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The developed SPF has been applied in the EB method to rank the 181 road segments by severity based on the contained accidents. The resulting number of segments has been divided into four equal quarters, ranked from the highest risk segments (BS1), to the least risk segments (BS4) (i.e. evenly distributing segments between groups).

5.10.3 Extracting decision rules from decision tree

Twelve predictor variables have been used with the four qualitative target variables (i.e. BS1, BS2, BS3, and BS4) to detect different accident patterns and contributing factors along the road. The twelve predictor variables include environment characteristics (e.g. AADT, trucks percentage, speed limit, road alignment, vertical alignments, road surface conditions), temporal characteristics of accidents (e.g. time, weekday, month), and accident variables (e.g. severity, visibility, weather condition). Descriptive statistics of segment severity related to the predicted variables is presented in Table 5.4 differentiating the data according to BS risk levels (BS1, BS2, BS3, and BS4).

Table 5.4 Description of accident attributes per segment’s groups risk level

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Frequency</th>
<th>% of accidents per BS risk level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accident Severity</td>
<td>Fatal</td>
<td>143</td>
<td>0.29 0.33 0.21 0.17</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>622</td>
<td>0.32 0.30 0.20 0.18</td>
</tr>
<tr>
<td></td>
<td>Slight</td>
<td>1161</td>
<td>0.32 0.29 0.22 0.16</td>
</tr>
<tr>
<td>AADT</td>
<td>&lt;3000 (AADT1)</td>
<td>140</td>
<td>0.32 0.11 0.16 0.40</td>
</tr>
<tr>
<td></td>
<td>3000-10000 (AADT2)</td>
<td>1296</td>
<td>0.26 0.29 0.24 0.20</td>
</tr>
<tr>
<td></td>
<td>&gt;100000</td>
<td>490</td>
<td>0.47 0.37 0.15 0.02</td>
</tr>
<tr>
<td>(AADT3)</td>
<td>Speed</td>
<td>&lt;=60</td>
<td>0.31</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>Limit</td>
<td>&gt;60</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Horizontal alignment</td>
<td>Straight</td>
<td>1237</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Curve</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consecutive-curves</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Others</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Type of accident</td>
<td>Head-on collision</td>
<td>306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rear-end collision</td>
<td>353</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standing vehicle, solid object collision</td>
<td>451</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slipping, overturning</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Run Off the Road</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hitting a pedestrian</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Others</td>
<td>415</td>
</tr>
<tr>
<td></td>
<td>Weather conditions</td>
<td>Clear</td>
<td>1167</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cloudy</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Foggy</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rainy</td>
<td>184</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Snow</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Visibility</td>
<td>Daylight, natural light</td>
<td>1396</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Night, with active lighting</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Night without public lighting</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Pavement condition</td>
<td>Flawless</td>
<td>1252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fractured, uneven, wavy pits</td>
<td>518</td>
</tr>
<tr>
<td></td>
<td>Vertical alignment</td>
<td>Flat</td>
<td>1741</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slope</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ascending</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Time of day</td>
<td>Mid-night (0-6)</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Morning (6-12)</td>
<td>599</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Afternoon (12-18)</td>
<td>770</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Night (18-0)</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>Day of week</td>
<td>Work day</td>
<td>1337</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Weekend</td>
<td>589</td>
</tr>
<tr>
<td></td>
<td>Month</td>
<td>Winter (12-2)</td>
<td>382</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring (3-5)</td>
<td>392</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Summer (6-8)</td>
<td>633</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Autumn (9-11)</td>
<td>519</td>
</tr>
</tbody>
</table>
During the training process, CART decision tree and the Gini splitting criterion have been used to construct the tree and select the best attributes. R package has been used to construct the tree. Figure 5.3 shows the produced classification tree. The tree has five terminal nodes with five DRR's. It can be easily recognized that AADT, horizontal alignment and speed limit are the primary selected splitter attributes for given phenomenon described by the tree. As the first splitter, AADT has appeared at the top of the tree as the most influential variable in identifying BS accidents by their spatial locations. This indicates that these variables are the most critical in classifying accident groups of the BS locations by severity. All of the selected variables are related to environment factors and all accidents within a certain cluster can take the same value of these environment attributes. In other words, accident density of certain segments and the associated environment characteristics, have an outstandingly important role in classifying BS segments by risk level. This is rational since most of the well-known BS identification methods use these environment factors to identify BS sites. For instance, the EB method relies mainly on the AADT to predict future changes in accident frequency [49], [120]. Similarly, accident ratio method identifies BS segments based on the value of accidents per segment length or accidents per AADT [16].

![Figure 5.3 The output of CART tree](image)

To explain figure 5.3 it has to be mentioned that the values in the output boxes of the branches contain the risk character of the given group referring to the majority of the given rule in the top row. The following line contains the number of accidents classified into BS1, BS2, BS3 or BS4 respectively from left to right. The last row describes the distribution of the accidents in light of the total number of accident. The interpretation of the results is straightforward. Segments of BS1 are ranked to be the riskiest BS sites, according to the applied EB method. These segments include about one-third of the total sample of accidents and have the highest accidents density per
unit length (2.98 accidents per km). It has to be emphasized that about 40% of all BS1 accidents occurred under high traffic volume conditions (AADT >10000). This type of accidents constitutes 26% of total accidents included in the training CART tree, as shown in the terminal node 1 (from the left) of Figure 5.3. This can be explained due to the strong correlation between the AADT and accident frequency. Benedek et al. (2016) [128] has proved that AADT is the most influential variable in accident prediction models, and the number of accidents is directly proportional to the AADT and the intensity of the linear dependency between the two factors (accidents and traffic) is strong.

At lower traffic volume (AADT < 10000) the CART tree is branched again but in this case by the horizontal road alignment. At the first branch, the rule describing accidents occurred at a curved shape section (horizontal curve) is ended in terminal node 2, which is also dominated by accidents of BS1 segments. Generally, accidents occurred on curved segments with low AADT (AADT<10000) constitute about 26% of all accidents analysed in the research, and included 20% of fatal accidents. This is supported by previous studies which have showed that accident rate of horizontal curves is from 1.5 to 4 times higher than straight roads [129].

At the second tree branch of the horizontal road alignment related node, road segments become straighter and less dangerous from the viewpoint of the horizontal curve. In this case high-speed limit (speed≥60 km/hour) is the distinctive factor in the rule related to the terminal node 3 (i.e. rule 3). This is not surprising as driver speed and speed limits have been recognised as one of the main contributing factors of road accidents. Speed limits can affect accident likelihood, severity, and density per road segment [76]. This rule is dominated by accidents ordered to the second road segment risk category (i.e. BS2). However, a high proportion of BS2 accidents follow the introduced BS1 rules (rule 1 and rule 2, as described in Figure 5.3).

The last less hazardous terminal nodes; node 4 and node 5, are dominated by accidents of BS3 and BS4 respectively. The typical characteristics of these nodes can be described by average environment factors and low speed limits or low traffic volume. For instance, accidents of node 4 mainly occurred in case of lower speed limits (<60 km/hour), moderate traffic volume (10000<AADT<3000), and straight road section. While accidents of node 5 mostly occurred under low traffic volume conditions (AADT <3000) and straight road section. In case of BS3 and BS4 segments bad weather conditions have also a strong influence on the occurrence of the accidents (see Table 5.4).

It can also be concluded from the tree that as the rules of the nodes are more dominated by lower severity segments (BS1, BS2, BS3, and BS4 respectively), accident contributing factors (AADT, speed, horizontal alignment) fit better to the safe and non-hazardous road scheme conditions. This is logical since accidents within the low severity BS segments are more likely to be caused by human errors or special events rather than specific environment factors like in the case of BS1 and BS2 segments.
5.11 Summary of part Two

Classification of road accident data still remains an important road safety issue. This chapter has presented a new methodology to examine the main contributing factors affecting the severity of the accidents included by the investigated road segments. The idea behind this research is to link the spatial factor and other accident attributes during the classification process of road accident data, since this kind of novel approach can strongly support the related decision making processes.

To achieve the desired goals, the accident data of the analysed Hungarian secondary main road, has been manipulated in three major steps (see Figure 5.2): spatial segmentation, black spots identification, and decision analysis. Firstly, the developed spatial segmentation approach has been used to classify all accidents in the road network into segments based on their spatial distribution. This has resulted in homogeneous segments with more flexible length pattern. In the second step, the Empirical Bayesian (EB) method has been used to classify the resulted segments into four groups according to risk level of the accidents. In the next step the decision tree has been applied to identify the main accident contributors for each segment group.

Spatial clustering method has resulted in 181 segments with 3.5 accidents per segment on average. The strength of this method has been demonstrated from the robustness of the developed accident prediction model, which is used by the EB method in order to rank road segments by severity. The resulted road segments have been divided into four groups with different risk level. The application of decision tree on the resulted groups has revealed a significant relationship between three environment factors (traffic volume, speed limits, and horizontal road alignment) and the most typical accident severity of the analysed segments.

In this research, the key factors characterizing road environment have been identified for spatially coherent accident groups instead of individual accidents to avoid errors originating from randomly occurred accidents resulted by events independent of infrastructure (e.g. individual human errors). Accordingly, the proposed methodology can help in deriving useful information from similar road segments in terms of their accident characteristics as well as their location on the network. The information can help road safety experts in understanding the accident profiles of the clusters and their pattern along the road network and proposing uniform safety improvements that could fit better to the defined segment classes. In the next step of the research the effect of the most relevant accident factors should be analysed in case of different data samples.

Possibility for implementation in practice

Previously literature was primarily focused on either the identification of BS locations without explaining the causes [58], [123] or classifying accidents into groups based on their causes without considering their spatial distribution [48], [130] (each group could include accidents from totally different locations).
In the case of this newly developed model, the two concepts are merged by identifying similarities in accident characteristics (e.g. causes) for the spatially aggregated accident groups. The spatial aggregation is performed by applying K-means clustering method, while the accident patterns are described with the help of the decision tree. Thus the analysis is performed on the segment level instead of the accident level.

The newly developed methodology can efficiently support the spatial identification of road segments that can be characterized by certain accident risk levels in light of specific road-safety related factors.

5.12 Thesis

• Thesis (4)

I have revealed the practical utility of applying the developed spatial clustering segmentation method as a well applicable complementary approach to the EB method in providing the best-integrated methodology for BSID.

Related publications to Thesis 4: [49]


• Thesis (5)

I have developed a complex methodology to identify accident risk patterns along the road network from a spatial and causal perspective, including the application of supervised and unsupervised data mining techniques.

Related publications to Thesis 5: [131], [132]


Chapter 6

Multilevel analysis of road accident frequency: the impact of the road category

6.1 Short summary

According to the first chapter of the dissertation; a successful road safety investigation strategy should take into account the segmentation processes, the type of predictor variables, and the type of road in order to have a reliable and comprehensive BSID process. The first two mentioned factors have already been discussed in the dissertation, therefore this chapter concentrates on the analysis of accident risk related to different road categories. More specifically, this research suggests a new, general model framework for predicting accident frequency at micro- and macro-level. The investigated hypothesis assumes that the number of accidents may vary on the one hand according to the roadway category and on the other hand according to the characteristics of the smaller road segments. Every individual road segment has different geometric and traffic properties. Contrary to this, every group of road segments within a single roadway can have similar characteristics that may differ from other road categories. To describe the nested relationship between individual road segments and different road categories a multilevel model has been developed. According to the newly developed framework, the multilevel analysis provides opportunity to study the hierarchical nature of road accident properties on micro- and macro-level, attempting to understand the risk of individual road segments within different roadway categories. To do so, fifty-seven roadways have been selected from the Hungarian’s map including five main categories (according to the Hungarian specifications): motorway, expressway, primary arterial, secondary main road, and local road. Furthermore, every roadway has been divided into flexible length segments, where sections are defined to be homogenous with regard to their traffic and geometric properties. The results verify that each different road category has different accident risk patterns. Moreover, road categories with low design standards, with lower traffic volumes, are less safe than well-designed roadways.
6.2 Introduction

The performance of BSID and segmentation methods can vary by road category, as suggested in Chapter 1 of this dissertation [76]. Accident prediction models are mostly limited to certain road categories or intersections. Castro et al. [73] has applied a latent variable generalized ordered response framework to describe the number of accidents at urban intersections. The result has revealed some important unobserved components influencing accident propensity at intersections, just like road parameters, road categories, and daily traffic. B. Persaud & Dzbik [75] have developed a generalized linear model to study the change of accident frequency and severity on the freeway. Most of these studies have been limited and dedicated for a specific road type or intersection. The different characteristics of the certain road categories make it difficult to construct a uniform model.

In general, a road accident is an unexpected event that can occur on different roads under different conditions. Accident density is highly affected by the environment and geometrical characteristics of the certain road segment. Accordingly, accident risk data of individual road segments can be characterized by a nested nature with regard to different road properties (e.g. road categories) and segment specific parameters. Since these datasets can be well-represented by a hierarchical structure, accident risk within road segments (i.e. individuals) of connecting roadways (i.e. groups) may have similar properties which should be considered during the analysis. The multilevel analysis provides opportunity to represent the hierarchical nature of road accident properties, aiming to characterize the risk of individual road segments and considering specific properties of the investigated groups. For example, on the one hand, small segments of a road may vary in some parameters, such as; traffic characteristics (i.e. Average Annual Daily Traffic (AADT), speed-limit, trucks volume) and geometric properties (i.e. roadside hazard, median type, curves, surface condition). On the other hand, accident frequency can be defined based on group characteristics aggregated from lower level characteristics, such as; category or level of service including road design standard, average speed-limit, and average AADT. Few studies have investigated the influence of the multilevel analysis on accident frequency. Cai et al. (2018) [80] have developed a Bayesian integrated spatial model, to analyze accident frequency at the macro- and micro-levels between districts and road entities (i.e. segments and intersections), simultaneously. Based on the results it can be concluded that the model could simultaneously identify both micro- and macro-level factors contributing to the accident occurrence and with higher performance.

Currently, the research field of combined micro- and macro-level model based investigation of the variance of road accident risk for different road categories is an under-researched area. This hypothesis of this thesis assumes that the predicted number of accidents is highly influenced by roadway categories and also the characteristics of their small segments. Every individual road segment has distinguished geometrical and traffic features. In contrast, every group of road segments within a single roadway can share similar characteristics that may differ
from other road categories. In the study the investigated roads are further divided into smaller homogenous segments with regard to environment and geometrical design parameters. Considering the assumed nested relationship between roads and road segments it is reasonable to construct a multilevel model, which is able to predict the number of accidents related to certain road segments in case of more road categories.

6.3 Data Description

In this study fifty-seven roadways have been chosen from the Hungarian network included five different categories: Motorway, Expressway, Arterial (primary main road), Secondary main road, and Local road. According to the Hungarian specifications, the Motorway is defined as two-lane roads in each direction, with an emergency lane and 130 km/h speed limit. The expressway has also two-lane in each direction but without an emergency lane and speed limit around 110 km/hour. Both roads, motorway and expressway, can also be classified as a highway. The arterial road is similar to the expressway, in terms of the number of lanes and speed limits, but with lower standards. The secondary main road is categorized as a state road. The last is a minor or local access road.

The used dataset covers the five road categories with a total of 5074 km length, where the number of investigated accidents is 6025 which contains three years data (2013-2015). The data includes accident information (i.e. geographical location, date, severity, parameters of involved vehicles), road characteristics (category, number of lanes, built-in/ non built-in areas), and traffic characteristics (i.e. traffic volume, speed-limits, trucks volume). Additional roadway characteristics (i.e. horizontal curves, lane characteristics, roadside hazard) have been identified with the help of the ArcGIS software and Google Earth. Only the data of continuous homogeneous road segments have been considered in the analysis, while the effect of intersection have not been taken into account during the investigation.

The research investigates the impact of nested relationship between overall road categories and specific road segments on accident frequency. To represent the effect of different road types on accident frequency, each road has been divided into smaller homogeneous segments with distinguished environment characteristics. The segmentation process has been based on the following factors: average annual daily traffic (AADT), speed limit, roadside hazard, and horizontal curves, as recommended by the Highway Safety Manual (HSM) specifications [20]. Table 6.1 presents descriptive statistics of the resulted segmentation process for each road category.

The segmentation process has resulted in 1590 segments with regard to the whole network. Each segment has different environment characteristics. Considering the parameters of Table 6.1, it can be also noted that each road category has different characteristics.
Table 6.1 Descriptive statistics of the resulted segments per road category

<table>
<thead>
<tr>
<th>Parameters description</th>
<th>Motorway</th>
<th>Expressway</th>
<th>Arterial</th>
<th>Secondary</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of roads</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>Total lengths (km)</td>
<td>993</td>
<td>236</td>
<td>1608</td>
<td>1796</td>
<td>441</td>
</tr>
<tr>
<td>Road segment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min length (km)</td>
<td>0.78</td>
<td>0.4</td>
<td>0.32</td>
<td>0.42</td>
<td>0.45</td>
</tr>
<tr>
<td>Max length (km)</td>
<td>17.29</td>
<td>8.8</td>
<td>14.91</td>
<td>15.53</td>
<td>11.83</td>
</tr>
<tr>
<td>Average length (km)</td>
<td>6.45</td>
<td>4.07</td>
<td>3.54</td>
<td>2.34</td>
<td>2.81</td>
</tr>
<tr>
<td>Standard deviation of length</td>
<td>3.28</td>
<td>1.78</td>
<td>2.52</td>
<td>1.49</td>
<td>1.91</td>
</tr>
<tr>
<td>Number of segments</td>
<td>154</td>
<td>58</td>
<td>454</td>
<td>767</td>
<td>157</td>
</tr>
<tr>
<td>Observed Accidents (per year)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>13</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Mean</td>
<td>2.31</td>
<td>1.08</td>
<td>1.65</td>
<td>0.86</td>
<td>1.12</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.46</td>
<td>1.40</td>
<td>1.49</td>
<td>0.96</td>
<td>1.52</td>
</tr>
<tr>
<td>AADT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>1388</td>
<td>3304</td>
<td>836</td>
<td>827</td>
<td>164</td>
</tr>
<tr>
<td>Max</td>
<td>107962</td>
<td>105023</td>
<td>34768</td>
<td>56176</td>
<td>17134</td>
</tr>
<tr>
<td>Mean</td>
<td>33291</td>
<td>34150</td>
<td>10029</td>
<td>6678</td>
<td>3865</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>21765</td>
<td>30074</td>
<td>6299</td>
<td>4878</td>
<td>3364</td>
</tr>
<tr>
<td>Truck volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>94</td>
<td>487</td>
<td>33</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Max</td>
<td>12748</td>
<td>19476</td>
<td>3489</td>
<td>2973</td>
<td>1213</td>
</tr>
<tr>
<td>Mean</td>
<td>4312</td>
<td>5055</td>
<td>756</td>
<td>634</td>
<td>146</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3224</td>
<td>5420</td>
<td>643</td>
<td>589</td>
<td>158</td>
</tr>
<tr>
<td>Speed (km/h)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Max</td>
<td>130</td>
<td>110</td>
<td>110</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Mean</td>
<td>122</td>
<td>90</td>
<td>75</td>
<td>74</td>
<td>71</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>13</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Horizontal Curves (%)</td>
<td>Straight path</td>
<td>0.73</td>
<td>0.57</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Single curve</td>
<td>0.23</td>
<td>0.33</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Multiple curves</td>
<td>0.05</td>
<td>0.10</td>
<td>0.20</td>
<td>0.12</td>
</tr>
<tr>
<td>Residential area</td>
<td>No</td>
<td>0.70</td>
<td>0.74</td>
<td>0.77</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.57</td>
</tr>
</tbody>
</table>

In general, motorway can be characterized by the highest average speed-limit (i.e. 122 km/hour). Naturally, the lower the level of road category is, also the lower the average
speed-limit is (in Table 6.1). AADT, truck volumes and, horizontal curves can be characterized by similar pattern, however the effect of road category does not seem to be so clear in the case of those properties. Beside this, the relationship of road category and the proportion of segments located in a built-in area seem to be even weaker. In general, according to Table 6.1, some convergence can be observed between the data of motorway and expressway segments, in contrast, the arterial, secondary, and local road segments show another convergence. Accordingly, to reveal the nested relationship between road categories and certain segments and their influence on traffic safety the chapter focuses on the multi-level evaluation of the combined effect of environment and geometrical characteristic and the road categories on accident risk.

## 6.4 Methodology

Accident prediction is considered to be one of the basic methods of road safety analysis. The analysis of accident frequencies can be adapted for both micro- and macro-levels. The characteristics of the different road categories can significantly influence safety risk of the road segments. At the same time, it can also be concluded that beyond the investigated properties of the road categories, accident frequency patterns are also influenced by the characteristics of the certain road segments.

Accident frequency values of the road segments involve zeros and positive integers. If accident locations are relatively rare, their distribution may be fit to Poisson distribution. Poisson distribution requires that its mean is equal to its variance. The observed accident variance is frequently greater than the mean. Frequently, the reason for this is the omission of relevant explanatory variables or dependent observations. The problem of over-dispersion can be solved by using a negative binomial regression.

To analyse the accident frequency and reveal the relationship between explanatory variables at both micro- and macro-levels, it is reasonable to investigate the hidden effects of the hierarchical relationship. The different characteristics of accident frequency can be analyzed at different information levels (see Figure 6.1), including regional level (i.e. population, residents), road level (i.e. category, level of service, average speed-limit), or small road segment level (i.e. AADT, speed limit, roadside hazard, curvature). For instance, possible relationship may exist among the road segments of the same road due to unobserved safety related characteristics. Similarly, roads in the same district can also have similar characteristics. Accordingly, the application of a multilevel model may increase the accuracy and help to produce a more general model. Multilevel analysis includes modeling the relationship between different groups of accidents by identifying a hierarchical system of the data that takes the advantages of the clustered dataset. This means that the outcome of the model is affected by a nested relationship between the lower level characteristics (level-one) of
individual road segments and higher level group characteristics (level-two) related to specific road parameters, as presented in Figure 6.1.

According to the newly developed methodology, this paper tries to analyze accident frequency by revealing the hidden influences of a hierarchical relationship between explanatory variables. To do so, a two-level approach has been applied to examine the nested relationship between a number of roads with five different categories and accident frequency per road segments. Figure 6.1 includes the three introduced major levels and from these levels the developed model focuses on Level-One and Level-Two. The description of the two-level models is given in the following.

![Figure 6.1 Multilevel hierarchical structure of road accident data](image)

### 6.1.1 Level-one modelling

Level-one model treats all road segments in the same way, and at this level the effect of the category, the location or the network role of the roads is not analysed only the certain geometric and environment parameters of the specific road segments. The general single-level model tries to estimate the expected accident rate ($\lambda$) per segment length (offset variable) in a similar model to the SPF, which can be written as follows [133].

$$
\eta_{ij} = \log (\lambda_{ij}) = \beta_0 + \sum_{q=1}^{Q} \beta_q X_q + \log (L_{ij}) \tag{6.1}
$$

Where, $\eta_{ij}$ is equal to the natural logarithm of the expected number of accident per segment length ($\lambda_{ij}$) for segment $i$ at road $j$. $\beta_0$ is the intercept, $X_q$ is the predictor $q$. $\beta_q$ is the fixed effects coefficient related to predictor $q$. An offset variable ($\log(L_{ij})$) has been added to the model to account for the variation of the road segment length ($L$).
The odd ratio can be calculated from raising the \( e \) number (Euler’s number) to power \( \eta_{ij} \) (log-odd coefficients - \( \exp(\eta_{ij}) \)). When the odd ratio is less than one, the probability of an event occurring is less than the probability that the same event will not be occurring. On the other hand it becomes higher than the probability of the complementary event if odd ratio is above one.

### 6.1.2 Level-two modelling (Developing multilevel model)

The difference between the multilevel model and the single-level model is the purpose of characterizing the nestedness of individual observations within the higher-level groups. The multilevel model treats individual road segments as cluster members. Each group belongs to a specific road with distinguished characteristics (i.e. road category). Since individual segments of the same road are likely to share similar characteristics, it seems to be reasonable to expect closer connection between their safety related properties than individuals of other roads, which in turn differs from the concept of single models.

Instead of the fixed-effect of slopes and intercept of the single-level models, multilevel models can be characterized by random variation. The variation in level-one intercept values (\( \beta_{0j} \)) between groups distinguished by \( j \) index (i.e. roads) indicates the influence of group-level characteristics on the outcome. The level-one intercept values represent a group mean when all explanatory variables or their average value are zero [134]. The general form of the level-two model is presented in the following equation [133].

\[
\beta_{0j} = \gamma_{00} + \left( \sum_{j} \gamma_{0j} W_j \right) + u_{0j} \tag{6.2}
\]

At the level-two, more level-one intercepts (\( \beta_{0j} \)) can be modeled as a function of level-two random effect variance (\( u_{0j} \)) and a fixed level-two intercept (\( \gamma_{00} \)). Random change in slope coefficient (\( \gamma_{0j} \)) between groups can be considered as level-two parameter (\( W_j \)) of the model. However, a general multilevel model can be formed by combining both levels (Equations 1+2) in a single formula (Equation 6.3) [133].

\[
\eta_{ij} = \gamma_{00} + \left( \sum_{i} \gamma_{i0} X_i \right) + \left( \sum_{j} \gamma_{0j} X_j \right) + u_{0j} + \log(L_j) \tag{6.3}
\]

Where: \( X_i \) and \( X_j \) are fixed-effect predictors of accident frequency at level-one (individual-level) and level-two (group-level), respectively. \( \gamma_{i0} \) is the coefficient of \( X_i \) and \( \gamma_{0j} \) is the coefficient of \( X_j \).
For multilevel model outcome, there is no separate variance (error) term at level-one, the estimated value of the level-two variance \((u_{ij})\) can describe whether groups are significantly different from each other.

### 6.5 Results

In order to model accident risk on different road categories, several explanatory variables including road segment characteristics (i.e. geometrical features, traffic characteristics) and the characteristics of complete roads (i.e. category, level of service, speed-limit) have been tested. During the preparation process insignificant variables have been excluded. According to Table 6.1, in most of the cases, the squared standard-deviation value of accident counts per road segment is higher than their mean values. This implies the possibility of over-dispersion. Therefore, accident counts have assumed to follow a negative binomial distribution.

To reveal the potentially unobserved correlation effects within the certain roads and address the influence of different road category related characteristics, a two-level negative binomial regression model has been developed for the case study. The level-two module of the model includes attributes related to the complete roads primarily differentiated by categories, while the specific road segment characteristics constitute the level-one module. Beside this accident frequency of road segments is considered as the target variable. Figure 6.2 presents the level-one estimated intercept values for the 57 roads of the case study. Level-one intercepts describe the unit changes in the predicted accident counts assuming the other variables at their base value (average or zero). The changing of intercept values (Figure 6.2) indicates that the predicted accidents are highly associated with different road features.

Figure 6.2 also shows the distribution of roadway intercepts ranked by road category. The figure depicts an increasing pattern representing the variation of the intercept values from the negative to the positive in an ascending order for each road category, assuming the other variables remain at their base values. The intercept describes the unit changes in the predicted accident counts assuming the other variables at their base value; average or zero (i.e. AADT= 10873, truck volumes=1127, speed= 79, road alignment= straight, and built-in area = no, according to all data of the case study). As the value of intercept increases, the accident density increases as well.

The degree of variability of the intercepts is presented in Table 6.2. Beside this, Table 6.2 includes the variance component of the intercept. The ratio of the intercept variance (0.084, \(p=0.01\)) to its standard error (0.033) justifies the use of the multilevel model. These findings support the applicability of the developed multi-level model in explaining hierarchical interrelationship among the identified variables. For example, multi-level models can reveal such correspondences where a certain prediction variable influences accident risk differently in case of different road categories. Since also the \(p\)-value is greater than 0.05, the residual variance value is reasonable enough to indicate the multilevel [133], [135].
Figure 6.2 Estimated intercept for different roadways from the multilevel analysis. Where, LR= local road, SR= Secondary road, AR= Arterial, ER= Expressway, and MR= Motorway.

Table 6.2 also shows the variability of the level-two road category variable. The variance of road category value (0.07, p=0.01) underpins that accident frequency of road segments is more likely to vary among roads characterized by different categories. In other words, the developed multilevel model can explain the effect of the hierarchical relationship between the individual road segments and groups of roads with different categories on accidents frequency.

Contrary to this, applying only a single-level analysis could reduce the performance of the model. Table 6.3 compares the model fitting (goodness-of-fit) of the single-level and multilevel models, developed from the current case study dataset. According to the results (Table 6.3), the multilevel model has significantly lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values (4377, 4388 respectively) compared to the single level model (6682, 6746 respectively). This indicates that a relevant proportion of the variation in the predicted accident counts between different roads with different categories is explained by the hierarchical structure.

<table>
<thead>
<tr>
<th>Variance</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Z-test</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.084</td>
<td>0.033</td>
<td>2.569</td>
<td>0.010</td>
</tr>
<tr>
<td>Road category</td>
<td>0.071</td>
<td>0.028</td>
<td>2.555</td>
<td>0.011</td>
</tr>
</tbody>
</table>
Table 6.3 Comparing the goodness-of-fit of a single level and multilevel models

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Akaike Corrected (AIC)</th>
<th>Bayesian (BIC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-level Model</td>
<td>6682</td>
<td>6746</td>
</tr>
<tr>
<td>Multilevel Model</td>
<td>4377</td>
<td>4388</td>
</tr>
</tbody>
</table>

Table 6.4 summarizes the estimated fixed effect results of applying the multilevel model, considering the length of the road segment as an offset variable.

Table 6.4 Resulted fixed effect coefficient values

<table>
<thead>
<tr>
<th>Model Term</th>
<th>Odd ratio (Coefficient)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.47</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Level-one road segment variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AADT</td>
<td>2.70</td>
<td>0.00</td>
</tr>
<tr>
<td>Speed-limit</td>
<td>0.69</td>
<td>0.00</td>
</tr>
<tr>
<td>Truck volumes</td>
<td>0.68</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment = Consecutive-curves</td>
<td>1.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment = One-curve</td>
<td>1.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment = Straight</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Resident= yes</td>
<td>1.90</td>
<td>0.00</td>
</tr>
<tr>
<td>Resident= No</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>Level-two full road variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road Category = Local road</td>
<td>1.78</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category = Secondary main</td>
<td>1.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Road Category = Arterial</td>
<td>1.93</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category = Expressway</td>
<td>0.20</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category = Motorway</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Over-dispersion</td>
<td>0.53</td>
<td></td>
</tr>
</tbody>
</table>

* Reference variable

Table 6.4 contains the variables of the two hierarchical levels. The level-one micro-level variables (road segments) include AADT, truck volumes, speed-limits, road alignment, and built-in area. The only level-two macro-level (roadway) variable is the road category. Most of the variables, in Table 6.4, are statistically significant to a 0.05 significance level (95% confidence level). The estimated coefficient represents the
odd ratio of accident counts. When the odd ratio is less than 1, the probability of the outcome is less than the reference level while if the odd ratio is above one the probability of the outcome is higher than the reference level. The intercept value (1.47) can be interpreted as the odd ratio that describes increased risk of accidents, assuming the predictors to be zero.

On the other hand, the predictor slope can be interpreted as the reciprocal of the amount required to increase or decrease the odd value of the expected accident rate by one unit, holding the other predictors at their base values. For example, road segments with higher AADT have a higher probability to be involved in accidents with a factor of 2.70. More specifically, each additional increase in AADT by 1/2.7 is expected to increase the odd ratio of the accident rate by one unit if truck volumes=1127, speed=79, road alignment= straight and residential area = no, and road category= motorway.

The correlation between accidents and AADT has been highlighted by many researchers. Cadar et al. (2017) [136] has proved that accident risk is significantly proportional with traffic volume up to a certain level of congestion. The proportional relationship has also been proved by the HSM [20].

Beside this, roads with higher design standards are more likely to have safer road segments when the other predictors are held at their base values. The interpretation of this result can be contradictory since usually increasing transportation speed is assumed to unequivocally increase accident risk and especially severity. In accordance with this, usually road specifications are improved as speed-limits increase, especially considering physical separation of the traffic directions. This is supported by Garber & Gadiraju (1989) [112] who have found that drivers’ speed tend to increase as road geometrical characteristics are improved, and that accident rates do not necessarily increase in case of an increase in the average speed but do rather increase in case of an increase in speed variance. From a statistical perspective, roads with high design standards (e.g. motorway, expressway) have lower accident densities even compared to other low speed-limit roads, as shown in Table 6.1. On the other hand it is still necessary to emphasize that in most cases the reason behind the better safety quality of the roads with higher speed limits can rather be derived from their higher technical specification level (e.g. physical separation of the opposite directions, larger curve radius, better visibility parameters, etc.). However, similar interpretation can describe the odd ratio of truck volumes that showed an inverse relationship with accident risks by 31%. In general, the increasing presence of heavy good vehicles on a road has negative effect on road safety. On the other hand, due to the service level improvement processes, nowadays those roads – which can be characterized by higher heavy good vehicle ration – are usually well developed and equipped with the necessary safety facilities. Contrary to this, a single-curve or a consecutive-curves road segment would increase the odd ratio of the predicted accident rates by 33% and 49% compared to a straight road alignment. Similarly, it is estimated that the residential location of a road segment can increase the odd ration of the accident rate by 90%.
The only level-two variable is the road category. Interpretation of its results is not so simple since it is affected by the pre-determined base values of level-one variables. However, the Motorway (the highest speed with the best design standards road) has been chosen as a reference for the other road categories. Table 6.4 reveals that the accident probability value belongs to the Expressway category decreases by 80% compared to the Motorways (reference variable). In contrast, accident probability value belongs to the Arterial category increases by 93% compared to the Motorways. Therefore, in our case study, the safest road is the Expressway while the riskiest is the Arterial.

In the case of the Expressway, referring to the above-raised discussion, the reason could be associated with the design standards of the road which already designed to cope with high-speed limits (e.g. median separation, large curve radius, roadside railing, better visibility). This can also be influenced by data shortages, since most of the expressways, of the case study, are newly constructed. In the case of the riskiest road category (i.e. Arterial), the explanation could also be related to the design standards, since it is still a high-speed road with the least specifications comes directly before moving to state and local roads. Generally, roads characterized by higher service level and at the same time by higher speed limits (motorway and expressway) seem to have more favorable effects on the number of accidents than the roads characterized by a lower level of service and speed limits.

6.6 Conclusion

Based on the performed analysis it becomes clear that road safety is strongly affected by micro- and macro-level factors, especially in the case of accident frequencies. In this study, a multilevel negative binomial regression model has been developed in order to predict accident frequency at different road segments for different road categories. The multilevel analysis has modelled the relationship between the different groups of accidents on different roads by identifying a hierarchical data structure that can utilize the advantages of the clustered dataset. The negative binomial distribution has been applied to deal with the possible over-dispersion of a count data.

The model includes two levels. In level-two, fifty-seven roads were chosen from the Hungarian road network, included five different categories (i.e. motorway, expressway, arterial, secondary road, and local road). Every road was divided into a group of level-one segments in which every segment has distinguished environment and geometric features. Accident frequency of road segment has been chosen as a target variable. A number of explanatory variables related to road segment characteristics (i.e. geometrical features, traffic characteristics) and full-road characteristics (i.e. category) have been applied in the model.

The resulted level-two variance components of the intercept verify that hierarchical structure is obviously embedded. In other words, it has a higher probability that
accident frequency values are more similar if they belong to road segments included by similar roads. This outcome has proved the applicability of multilevel models in improving the efficiency of accident frequency estimation methods. On the other hand, examination of the random effect of road categories has indicated that slopes between roads of different categories within a group-level are significantly different. This indicates a reasonably strong correlation between road categories and accidents risk that could be difficultly revealed by a one-level model. The result has also indicated that high-speed roads (motorway and expressway) have lower accident rates compared to other roads with lower service level (arterial, secondary road, and local road). This can be explained primarily by the difference in the design standards for the different road categories.

Predicting accident frequencies considering the nested relationships between different micro- and macro-levels can enhance the accuracy of accident estimation models and black spot analysis. However, extending the model by additional levels can also contribute to the improvement of the model performance.

Possibility for implementation in practice

The proposed multilevel analysis of Thesis-6 provides a general model that is able to identify accident frequency at any road segment, with different characteristics, and for any road category. Furthermore, the new model contributes to the improvement of accident prediction accuracy. Beside this, the developed model can open up new horizons in road safety research, especially in connection with BSID.

6.7 Thesis

- Thesis (6)

I have developed a new multilevel negative binomial regression model to predict accident frequency at different road segments for different road categories. The multilevel analysis has modeled the relationship between the different groups of accidents on different roads at micro- and macro-level by identifying a hierarchical data structure that can utilize the advantages of the clustered dataset. In other words, the model can define the accident rate of the investigated road segments considering their environment features and road categories.

Related publications to Thesis 6: [137]

M. Ghadi, “Multilevel Analysis of Road Accident Frequency: The Impact of the Road Category,” in TRB 2020 - 99th Annual Meeting. Accepted.
Chapter 7

Overall conclusions and scopes for the future studies

7.1 Short summary

Indicating road safety-related aspects in the phase of infrastructure planning and operating is always a challenging task. The identification of hazardous road sections or black spots (BS) receives a great interest from road agencies and safety specialists. Black spot identification (BSID) can be defined as a process of searching for locations in transportation systems with a higher number of accidents than other similar locations that is mainly caused by local risk factors [2], [3]. Errors in BSID can result in many false positive and negative cases.

Considering that traffic accident data is heterogeneous, in general, data segmentation is considered the first and most important step of the BSID process. In other words, the success of any method applied in identifying BS on roads should depend fundamentally on how the data is organized into specific homogeneous segments. In road safety, road network segmentation is the process of organizing road infrastructure network data into homogenous entities. Data segmentation can also be defined as a process of dividing and classifying a large and complex dataset into small and simple homogeneous groups or entities in which data within a group are very similar but data between groups are dissimilar.

Most often, road segmentation methods are based on researchers’ experiences, methodological decisions or objectives. Thomas [30] argued that applying different lengths for segmenting road network can result in different definitions of the hazardous locations which, in turn, affects the reliability of the results. Koorey [31] discussed the benefit of applying variable segment length and their effect on BS determination. It has been confirmed as well that the more accurate the network segmentation is, the better performance the accident prediction model has, and this consequently affects the performance of the BSID methods that rely on accident prediction models in their criterion (i.e. EB method). On the other hand, the segmentation process plays also an important role in the mechanism of locating BS segments and identifying their proper lengths.
The main concept of the dissertation is based on the development of some novel statistical tools for clustering and segmenting large accident datasets, based on their spatial aggregations, into small and homogeneous groups. The idea is based on the assumption that spatially aggregated accidents are expected to be more homogeneous than distant accidents. Since, they are more likely to occur under similar environment conditions [39], [46].

The main purpose of the developed segmentation method has been to improve the accuracy of road network segmentation process. Thus, accident prediction and BSID reliability can also be improved. On the other hand, the dissertation has investigated the change in accident risk between different road segments, with different characteristics and road categories by developing a novel multilevel evaluation model.

7.2 New scientific results

(1) New scientific results have been achieved by the evaluation of pros and cons regarding the applied BSID methods. During the investigation, the performance of the following two popular methods have been measured and compared: sliding moving window and spatial autocorrelation methods. The evaluation of the methods has been implemented on two different road categories applying three different segmentation criteria.

- In the evaluation, a consistency test method is proposed in order to evaluate the consistency of the analyzed BSID methods in case of different segmentation characteristics. The test measures the consistency of identifying road segments with high crash-risk by applying different segmentation characteristics (i.e. different segment length and sliding distance). The higher the proportion of shared BS sites is, the more consistent the applied BSID method is. The consistency is measured by the proportion of the shared sections related to the outputs of (i.e. n1, n2, .., n) the same BSID method with different segmentation models.

\[
\text{Consistency} = \frac{\left| n_1 \cap n_2 \cap \ldots \cap n_n \right|_{\text{BSID}_1}}{\left| n_1 \cup n_2 \cup \ldots \cup n_n \right|_{\text{BSID}_1}} \quad (7.1)
\]

Where \( n \) is the index of BS segments identified by the applied BSID method.

- The use of different segment lengths (and sliding distance) seems to have a significant impact on the performance of the SMW method. This can be seen from the difference in the identified BS along the road in case of the different applied segmentation criteria, as presented by the consistency test results. Generally, the SMW has shown a comparatively higher consistency in the case of the motorway compared with the urban roads.
The application of the SPA method has also demonstrated a variation in performance with different segmentation criteria and different road categories. In the low-speed urban roads, accidents showed a higher spatial autocorrelation in the case of small segment lengths in comparison with motorway accidents. This can be explained by the nature of the accident distributions along the roadway.

In the case of the analyzed roads which represent different road categories, the different BSID methods have been performed in a significantly different way. Every BSID method could lead to different outcomes with different segment lengths, starting points, and different processes and criteria for segmentation.

**Related publications to Thesis 1:** [76]

(2) My research work targeting the development of a novel accident dataset based network segmentation method based on the combination of K-mean clustering and linear referencing techniques has achieved scientifically new results in identifying homogeneous accidents clusters based on their spatial distribution (Figure 7.2). The developed method assumes that spatially and temporally coherent are more likely to occur within the same circumstances. In accordance with this, the developed method can divide road networks into flexible length road segments depending on the spatial distribution of accident clusters. Due to the favourable characteristics of the method, which allows us to adapt the segment lengths to the spatial distribution of the accidents, the identification and exclusion of empty road segments can efficiently support the reduction of computational demand of the BSID process.
A complex statistical technique has been used to evaluate the performance of the developed method compared to four other well-known segmentation models, including constant AADT segments, constant length segments, related curvature characteristics and a multivariable method suggested by the HSM. At the end, the developed clustering based segmentation models have achieved the best goodness-of-fit statistics. Table 7.1 presents and compares the resulted statistics of the implemented models by each segmentation method. Figure 7.3 shows graphically the goodness-of-fit of the evaluated methods based on the resulted Pearson correlation test.

### Table 7.1 Resulted parameters and goodness-of-fit values of the five road segmentation methods.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Methods of segmentation</th>
<th>Seg-1 (K-means clustering)</th>
<th>Seg-2 (HSM)</th>
<th>Seg-3 (Constant AADT)</th>
<th>Seg-4 (Constant length)</th>
<th>Seg-5 (Curvature)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>α (Intercept) [p-value]</td>
<td>β1 (AADT) [p-value]</td>
<td>β2 (Speed) [p-value]</td>
<td>β3 (HDA) [p-value]</td>
<td>k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 5.934 [0.062]</td>
<td>0.784 [&lt;0.001]</td>
<td>- 0.531 [0.080]</td>
<td>1.186 [0.007]</td>
<td>1.070</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 5.664 [0.013]</td>
<td>1.107 [&lt;0.001]</td>
<td>- 1.264 [0.018]</td>
<td>2.147 [0.007]</td>
<td>1.909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 7.664 [&lt;0.001]</td>
<td>0.961 [&lt;0.001]</td>
<td>- 0.320 [0.10]</td>
<td>3.414 [0.063]</td>
<td>1.017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 14.269 [&lt;0.001]</td>
<td>1.872 [&lt;0.001]</td>
<td>- 0.846 [0.019]</td>
<td>2.582 [0.057]</td>
<td>1.117</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- 6.785 [&lt;0.001]</td>
<td>1.009 [&lt;0.001]</td>
<td>- 0.638 [0.003]</td>
<td>1.084 [0.067]</td>
<td>1.024</td>
</tr>
<tr>
<td></td>
<td></td>
<td>β3 (HDA) [p-value]</td>
<td>k</td>
<td>QIC</td>
<td>PCC [R-square]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.070</td>
<td>22</td>
<td>0.794 [0.63]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.909</td>
<td>167</td>
<td>0.300 [0.08]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.017</td>
<td>64</td>
<td>0.760 [0.57]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.117</td>
<td>140</td>
<td>0.474 [0.22]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.024</td>
<td>284</td>
<td>0.684 [0.47]</td>
<td></td>
</tr>
</tbody>
</table>
Figure 7. 3 Linear correlation between the observed accidents (x-axis) and the predicted accidents (y-axis) of the year 2016

**Related publications to Thesis 2:** [113]–[115]

(3) As a new scientific result, I have revealed the effect of methodological diversity of road network segmentation methods on the performance of different black spot identification methods, and have performed a comprehensive comparison regarding their joint performance (Table 7.2).
### Table 7.2 Total score test results for top 5% and 10% BS

<table>
<thead>
<tr>
<th>BS method</th>
<th>Segmentation method</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seg-1 (K-means clustering)</td>
<td>Seg-2 (Constant length)</td>
</tr>
<tr>
<td>5% risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>92.2</td>
<td>84.4</td>
</tr>
<tr>
<td>EEB</td>
<td>77.4</td>
<td>77.8</td>
</tr>
<tr>
<td>AF</td>
<td>93.1</td>
<td>72.8</td>
</tr>
<tr>
<td>AR</td>
<td>57.8</td>
<td>72.7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>80.1</strong></td>
<td><strong>76.9</strong></td>
</tr>
<tr>
<td>10% risk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>96.8</td>
<td>81.9</td>
</tr>
<tr>
<td>EEB</td>
<td>82.0</td>
<td>77.0</td>
</tr>
<tr>
<td>AF</td>
<td>96.3</td>
<td>74.1</td>
</tr>
<tr>
<td>AR</td>
<td>72.6</td>
<td>69.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>86.9</strong></td>
<td><strong>75.6</strong></td>
</tr>
</tbody>
</table>

The main achieved results are summarized below:

- The developed methodology makes it possible to evaluate
  - the joint performance of the investigated BSID and segmentation methods.
  - the performance of the different segmentation and BSID methods,
    independently in an integrated model environment.
- The results have also demonstrated how the developed spatial clustering segmentation method can improve the performance of the different BSID methods.
- Practically, the applied methodology can significantly help to choose the best integrated BSID methodology.

**Related publications to Thesis 3:** [116], [119]

(4) During the research a complex practical application has been implemented to examine the practical utility of the developed spatial clustering segmentation method. Based on the achieved outcomes it has become clear that the new approach can efficiently complement the EB model and the combined model framework can provide an outstandingly well performing BSID method. Beside this, the practical application has presented an addition key advantage of the developed integrated technique, which is the ability of grouping spatially coherent accidents in small segments, and so separating and excluding if required the
empty and less risky road sections. After this it becomes possible to identify and rank all road segments which are affected by traffic accidents.

**Related publications to Thesis 4:** [49]

(5) The new scientific achievements of the research provide an integrated methodology that can assist road safety analysis in identifying the patterns of accidents and BS locations along the road. The comprehensive methodology, presented in Figure 7.4, includes the application of the proposed segmentation method with other data mining techniques (i.e. decision tree).

![Figure 7.4 The proposed framework for analyzing accident spatial distribution and patterns](image)

Such methodology can significantly improve the efficiency of the decision making processes and help road safety experts in identifying the location of road segments with similar accident patterns, and hence in applying comprehensive and quality assured safety measures.

**Related publications to Thesis 5:** [131], [132]

(6) As a scientific novelty a multilevel negative binomial regression model has been built up for predicting accident frequency at the micro- and macro-level for different road categories.

\[
\eta_j = \gamma_{00} + \left( \sum_i \gamma_{0i} X_i \right) + \left( \sum_j \gamma_{0j} X_j \right) + \mu_0 + \log(L_j) \quad (7.2)
\]
Where: $\beta_{0j}$ is level-one intercepts. $X_i$ and $X_j$ are fixed-effect predictors of accident frequency at level-one (individual-level) and level-two (group-level), respectively. $\gamma_{i0}$ is the coefficient of $X_i$ and $\gamma_{0j}$ is the coefficient of $X_j$. $u_{0j}$ is level-two random effect variance.

The study has been based on the assumption that the number of accidents may vary according to the roadway category and the characteristics of their smaller segments. Every individual road segment has different geometrical and traffic features. In contrast, every group of road segments within a single roadway can share similar characteristics that may differ from other road categories. The resulted level-two variance components of the intercept verify that the analysed data carries inside a latent hierarchical structure. Accordingly, the reached outcomes strongly support the assumption, that the road segment related accident frequency values are more likely to vary across different roadways. This result is underpinned by the variability of the intercept values for the different roads categories (Figure 7.5).

![Figure 7.5 Estimated intercept for different roadways from the multilevel analysis. Where, LR= local road, SR= Secondary road, AR= Arterial, ER= Expressway, and MR= Motorway.](image)

The developed multilevel model can contribute to the methodology of the road safety analysis with regard to the following aspects.

- Understanding the effect of risk change of accident existed on different road categories, also considering the effect of the environment characteristics of the different road segments.
• Predicting accident frequency with a higher accuracy taking also into account the nested relationships between different micro- and macro-level factors.

Table 7.3 summarizes the estimated fixed effect results of applying the multilevel model, considering that the length of the road segment as an offset variable.

<table>
<thead>
<tr>
<th>Model Term</th>
<th>Odd ratio (Coefficient)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.47</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Level-one road segment variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AADT</td>
<td>2.70</td>
<td>0.00</td>
</tr>
<tr>
<td>Speed-limit</td>
<td>0.69</td>
<td>0.00</td>
</tr>
<tr>
<td>Truck volumes</td>
<td>0.68</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment= Consecutive-curves</td>
<td>1.49</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment= One-curve</td>
<td>1.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Road alignment= Straight</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Resident= yes</td>
<td>1.90</td>
<td>0.00</td>
</tr>
<tr>
<td>Resident= No</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>Level-two full road variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road Category= Local road</td>
<td>1.78</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category= Secondary main</td>
<td>1.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Road Category= Arterial</td>
<td>1.93</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category= Expressway</td>
<td>0.20</td>
<td>0.00</td>
</tr>
<tr>
<td>Road Category= Motorway</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Over-dispersion</td>
<td>0.53</td>
<td></td>
</tr>
</tbody>
</table>

* Reference variable

Related publications to Thesis 6: [137]

7.3 Application of the scientific results

(1) The main idea of Thesis-1 can be concluded as follows: “The success of any method applied in identifying BS locations of the road network should depend fundamentally on how the data is organized into segments”. In other words, the thesis has recognized and underpinned that the initial stage of BSID, represented by road network segmentation, should gain more attention in order to reduce the deficiencies of BSID. Furthermore, the thesis proposed a new consistency test to measure and compare the performance of different BSID method by measuring the consistency in identifying the same BS with each different segmentation
characteristics. This achievement has underpinned the recognition that the following tasks should be completed during research:

a. Developing a comprehensive evaluation methodology applicable to investigate the interaction of segmentation and BSID methods also considering the effect of the integrated model environment.

b. Identifying a general concept combining both segmentation and BSID models would be necessary to improve the performance of the overall methodology.

This finding would be the basis for any related work, regarding the application and development of BSID methods.

(2) The developed road network segmentation method has a number of applications in the field of road traffic safety. In road safety, it can be used to improve the reliability of SPF which consequently affects the performance of the relevant BSID methods that rely on accident prediction models in their criterion. Thesis-2 provides also a methodology to assess and compare data segmentation methods. In the field of BS analysis, the developed segmentation method can contribute to the identification of proper segment lengths, which can help in fitting BS area to the characteristic of the accident distribution.

(3) The third thesis proposes a new methodology that can assist experts in selecting the proper BSID and segmentation method combination that have the best joint performance. The thesis also call the attention of the experts that the application of some BSID (e.g. accident rate) and/or segmentation methods (e.g. constant AADT) can result in misleading conclusions.

(4) The fourth thesis introduces a complex practical methodology to investigate road safety more effectively. This methodology suggests the application of the developed segmentation method along with the EB method. Hence, it can be considered as a good integrated methodology in the area of BSID.

(5) The results of the data mining process can help road safety experts to make a comprehensive assessment of road accident patterns along a road by identifying road segments with similar risk features, so proper and comprehensive safety measures can be applied.

(6) The proposed multilevel analysis of Thesis-6 provides a general model that is able to identify accident frequency at any road segment, with different characteristics, and for any road category. Furthermore, the new model contributes to the improvement of accident prediction accuracy. Beside this, the developed model can open up new horizons in road safety research, especially in connection with BSID.
7.4 Scope for future work

(1) The developed spatial segmentation method can be further improved in term of the applied clustering method and the criteria applied for identifying the number of clusters. On the other hand, the criteria applied in the comparison can include other statistical techniques with more predictor variables related to road segment characteristics.

(2) The other newly developed segmentation and BSID techniques of the future should also be investigated by the prepared methodology.

(3) It is reasonable to develop a software package related to the methodology proposed by Thesis-5 that can achieve the identified objective of accident pattern detection in a more simple way. On the other hand, a variety of variables can also be examined in the decision process.

(4) Related to Thesis-6, future work could be able to propose a more complex hierarchical model with an additional number of levels that could include more comprehensive variables to describe road accidents at the micro- or macro-level more accurately.
List of the author publications (related to this dissertation)


Bibliography


