UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

RIGORÓZNÍ PRÁCE

Spatial analysis of traffic crashes by the use of kernel density estimation

Katedra matematické analýzy a aplikací matematiky
Vedoucí rigorózní práce: **doc. RNDr. Eva Fišerová, Ph.D.**
Vypracoval(a): **Richard Andrášik**
Studijní program: Matematika
Studijní obor: Matematika a její aplikace
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Richard Andrášik

Název práce: Prostorová analýza dopravních nehod pomocí jádrového odhadu hustoty

Typ práce: Rigorózní práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2017

Abstrakt: Spolehlivá identifikace nebezpečných lokalit (hotspotů) dopravních nehod je záležitostí, kterou je potřeba se zabývat. Poněvadž nebezpečná místa na silnicích jsou lokalitami s významně vysokým počtem dopravních nehod způsobených lokálními faktory dané lokality, k nalezení míst, kde se dopravní nehody dějí častěji než se očekává, lze použít shlukovou analýzu. Představujeme metodu KDE+, která překonává známé nevýhody již existujících postupů. KDE+ rozšiřuje standardní metodu jádrového odhadu hustoty o statistický test významnosti a je schopna seřadit výsledné shluky podle jejich nebezpečnosti. Metodou KDE+ jsme analyzovali data o dopravních nehodách v období 2009 – 2013, které se udály na silničních komunikacích v České republice. Efektivita metody KDE+ závisí na počtu simulací, přičemž vyšší počet simulací vede k přesnějším výsledkům. Zkoumali jsme přesnost stanovení prahu pro různé počty simulací a došli jsme k vyváženému nastavení metody s ohledem na časovou náročnost a přesnost. V České republice jsou od roku 2007 všechny dopravní nehody lokalizovány pomocí GPS. Podobná databáze ovšem není v ostatních zemích pravidlem. Z tohoto důvodu bylo použití metody KDE+ rozšířeno i na nepřesně lokalizovaná data (např. zaokrouhlení vzhledem ke staničení pozemních komunikací). Získané výsledky pomáhají administrátorům silnic v efektivní lokalizaci nejnebezpečnějších míst v rámci silniční sítě. Výsledky byly rovněž vizualizovány ve webovémapové aplikaci www.kdebourame.cz.

Klíčová slova: Prostorová analýza, Jádrový odhad hustoty, Monte Carlo, Dopravní nehody, Nebezpečné lokality

Počet stran: 48

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Richard Andrášik

Title: Spatial analysis of traffic crashes by the use of kernel density estimation

Type of thesis: Rigorous thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2017

Abstract: Reliable identification of hazardous locations (hotspots) of traffic crashes is a safety issue which has to be addressed. Since hazardous road locations are places with a significantly high number of traffic crashes due to local factors connected to the location, cluster analysis can be applied to find locations where traffic crashes occur more frequently than expected. We introduce the KDE+ method which overcomes the well-known drawbacks of the existing approaches. The KDE+ method extends the standard kernel density estimation by statistical significance testing and allows for the ranking of the resulting significant clusters. We applied the KDE+ approach to the data on traffic crashes which occurred in 2009 – 2013 on the Czech road network. The effectiveness of the KDE+ method depends a great deal on the number of simulations with a higher number of simulations leading to more precise results. The precision of establishing a threshold was measured for varying numbers of simulations. We found a suitable balance between the time consumption and the accuracy of the Monte Carlo method embedded in the KDE+ approach. Since 2007, all traffic crashes have been precisely localized by GPS in the Czech Republic. Similar databases in other countries do not contain, however, information on the exact positions (GPS) of the accidents. Although the KDE+ method was primarily introduced for data with GPS coordinates, we extended the framework to cover also the non-precisely located data (e. g. rounded to a relative stationing system). The presented results allowed the road administrators to effectively localize the most dangerous places within the road network. The results were visualized in our web-map application www.kdebourame.cz.

Key words: Spatial analysis, Kernel density estimation, Monte Carlo, Traffic crashes, Hazardous road locations

Number of pages: 48

Number of appendices: 0

Language: english

Prohlášení

Prohlašuji, že jsem vytvořil tuto rigorózní práci samostatně za vedení a pomoci doc. RNDr. Evy Fišerové, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne ....

podpis

Contents

Poděkování

Na tomto místě bych rád poděkoval svým kolegům z Centra dopravního výzkumu, zejména doc. RNDr. Michalu Bílovi, Ph.D. a Mgr. Jiřímu Sedoníkovi, za možnost pracovat na tomto zajímavém tématu a spolupráci. Rovněž děkuji Mgr. Tomáši Svobodovi za softwarovou realizaci metody popsané v této práci. Dále bych rád poděkoval doc. RNDr. Evě Fišerové, Ph.D. za čas, který mi věnovala při konzultacích. V neposlední řadě děkuji svým rodičům za podporu v mé výzkumné práci.

Introduction

The aim of any society should be to prevent traffic crashes (TCs) and reduce the severity of their consequences. From the point of view of a road administrator, precise identification of hazardous places (hotspots) within a road network is an essential tool for applying mitigation measures. However, standard methods of hazardous places identification only take into account aggregated data. They evaluate the safety of a road section as a whole [17, 21] or test the general tendency to form clusters on a particular road section [24, 32].

The identification of hazardous locations on roads has substantially progressed during the last years. It has been facilitated by both the application of geographic information systems (GIS) into transportation research and by the possibility of precise localization of TCs through the use of GPS devices. Nowadays, many traffic-crash databases contain the precise locations of the TCs and therefore it is no longer necessary to detect hazardous road locations from aggregated data [14, 16, 20, 32]. Having these accurate positions of the TCs, we are able to focus on the precise identification of spatial patterns of TCs.

In general, there are three types of methods for hotspots identification. The most straightforward approach is based on aggregated counts of records. The sums are either used directly to rank segments of roads, or the local spatial autocorrelation statistic (local Getis Ord statistic) is computed. The latter option seems better because it allows for setting an objective threshold for distinguishing significantly dangerous locations. However, these methods have several drawbacks: segmentation of roads, not considering the regression to the mean and aggregation when exact positions of TCs are known.

Second, various regression models are often built to analyze crash-frequency data. They express the number of TCs by the use of explanatory variables [15]. However, there are many methodological issues which have to be addressed prior to the application of this approach (e. g. time-varying explanatory variables, under-reporting, low sample mean and sample size, omitted variables, disunity in the choice of a functional form, segmentation of roads). Hence, regression analyzes are time-consuming in the case of crash-frequency data and often produce biased results [21]. The empirical Bayes method [10] uses the results from a regression model as the prior estimate of expected crash-frequency counts. Afterwards, the prior information is combined with the real data and the posterior estimate is produced. Although this is a brilliant idea, the accuracy of the empirical Bayes method depends on the prior estimates produced by a regression model. Furthermore, this approach gives no objective threshold for distinguishing significantly hazardous locations.

Since hazardous road locations are places with a significantly high number of TCs due to local factors connected to the location [11], also clustering analysis can be used to find locations where TCs occur more frequently than expected. Clustering methods can

- either testify a general tendency of clustering on a road section; for instance K-function method [24, 32] and nearest-neighbor methods [22, 29]

- or identify exact positions of hotspots within a road section; for example the Kernel density estimation (KDE) method [27], Clumping method [22].

The methods from the first group do not contribute to the localization of clusters within sections. The latter methods are more efficient as they provide the information of this type.

We focused on the KDE method and its application to the spatial analysis of TCs. One advantage of the KDE method compared to other clustering methods is that the uncertainty about the exact position of the TCs is expressed by the bandwidth of the kernel – this means something like spreading the risk of a traffic crash [1].

Since the estimated probability density function is a multimodal function in general, it has more local maximums where clusters can be found. Therefore, it is necessary to determine which ones are statistically significant. Currently, a comprehensive investigation of statistical significance of clusters identified by the KDE method is missing in the literature and the KDE suits better for visualization purposes than for identification of hotspots [26]. The same problem was noticed also by Xie and Yan [31]. They wrote that "... the absence of the significance testing is a drawback of the KDE". Anderson [1] mentioned the same problem: "However one main drawback (of the KDE method) reoccurs, which relates to determining the statistical significance of the resulting clusters. This is an area of research which is something to investigate in further studies".

We aim on presenting an improved procedure of cluster detection, based on the standard KDE method, suitable to identify the most hazardous road locations by testing the significance of the clusters followed by the ordering of the most hazardous places. This procedure is described in chapter 2. We present also further developments of the KDE+ method in chapter 3, particularly the applicability to non-precisely located data.

In addition, TCs on the Czech roads were analyzed with the use of the novel KDE+ method. We performed the analysis of four databases: TCs without distinction (all TCs), single-vehicle TCs, two-vehicles TCs, TCs with severe injury or death. The obtained results are described in chapter 4. Analyzes of other databases (e. g. wet-road collisions, animal-vehicle collisions) are visualized in the web-map applications www.kdebourame.cz and www.srazenazver.cz.

This rigorous thesis is based on the following published papers:

- Bíl, M., **Andrášik, R.**, Janoška, Z.: Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation, Accident Analysis & Prevention 55, 265 – 273 (2013).

- **Andrášik, R.**, Bíl, M.: Traffic accidents: Random or pattern occurence?, Safety and Reliability of Complex Engineered Systems – Podofillini et al. (Eds), 3 – 6, Taylor & Francis Group, London, ISBN 978-1-138-02879-1 (2015).

- Bíl, M., **Andrášik, R.**, Sedoník, J., Svoboda, T.: The KDE+ software: a tool for effective identification and ranking of animal-vehicle collision hotspots along networks, Landscape Ecology 31(2), 231 – 237 (2016).

- **Andrášik, R.**, Bíl, M.: Traffic accident hotspots: Identifying the boundary between the signal and the noise. Risk, Reliability and Safety: Innovating Theory and Practice – Walls et al. (Eds), 1634 – 1637, Taylor & Francis Group, London, ISBN 978-1-138-02997-2 (2016).

Chapter 1

Data

The Czech road network is approximately 37,469 km in length excluding the urban roads. The data on TCs comes from the Czech Police database. This database consists of 90,418 entries which were recorded over the period 2009 – 2013. Such a time span is sufficient because there are enough records and the characteristics of the traffic remain relatively unchanged [8, 11]. We excluded TCs which occurred at intersections because they could hide the existence of a dangerous location within a section [7]. These TCs did not have to be excluded in order to perform the analysis. However, intersections are typically dangerous places by definition. Therefore, we focused on finding dangerous locations within road sections, i. e. between intersections.

In our case, each TC was localized by a police officer by the use of a GPS device. The Police used the Garmin Geko 201 device which is able to localize a point (without additional EGNOS/WAAS corrections) with a maximum error up to 25 m [9]. If the measurement of a TC position is not taken at the exact place where the TC occurred, another error may arise. The police officer sometimes gathered the locations of the TCs, for the sake of their personal safety, at the side of the road. The maximum width of a road in available database was 25 m. Therefore the maximum expected error which may arise from the process of traffic-crash data capturing accounts for 50 m. Hence, the minimum kernel bandwidth, in our case, should not be shorter than the above specified length.

Our proposed approach can be applied to any point (or interval) data situated on a network. We can consider TCs in general, animal-vehicle collisions in particular or even records on roadkills. To demonstrate the performance of our method, we analyzed four databases. Initially, we applied the proposed method to the database of TCs without distinction. Consequently, we performed the analysis in three specific groups of TCs: single-vehicle TCs, two-vehicles TCs and TCs with severe injury or death (see Figure 1.1).

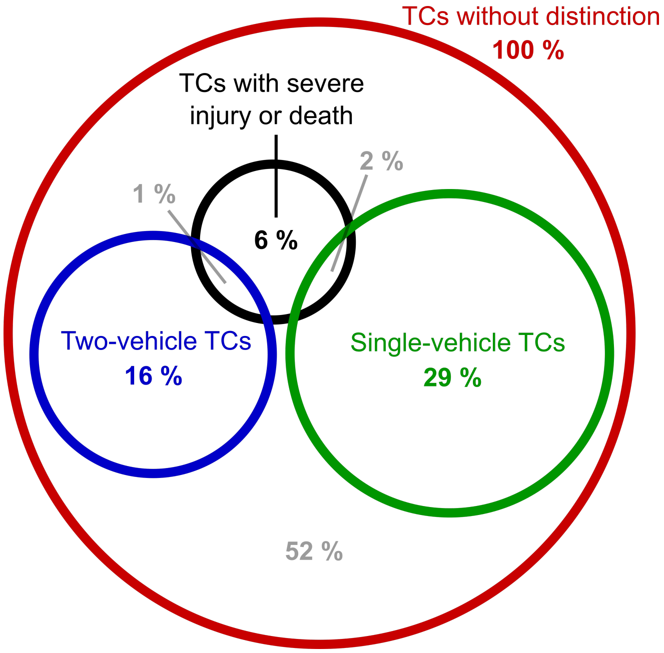


Figure 1.1: TCs in the Czech Republic in the period 2009 – 2013.

TCs with severe injury or death are naturally of a special concern. Although the number of TCs slightly increased from 2011 to 2013, the proportion of TCs with severe injury or death in relation to all TCs fell from 7.2% to 5.2% over this period (see Figure 1.2).

The road network data were obtained from the Road and Motorway Directorate (RMD). The analyzes were performed on primary roads excluding the highways. We also omitted the urban areas, because the RMD data does not contain the complete urban network. The road network was separated into road sections, which did not contain any intersections. We define a road section as a segment of a road network between two intersections and this definition is used throughout this text.

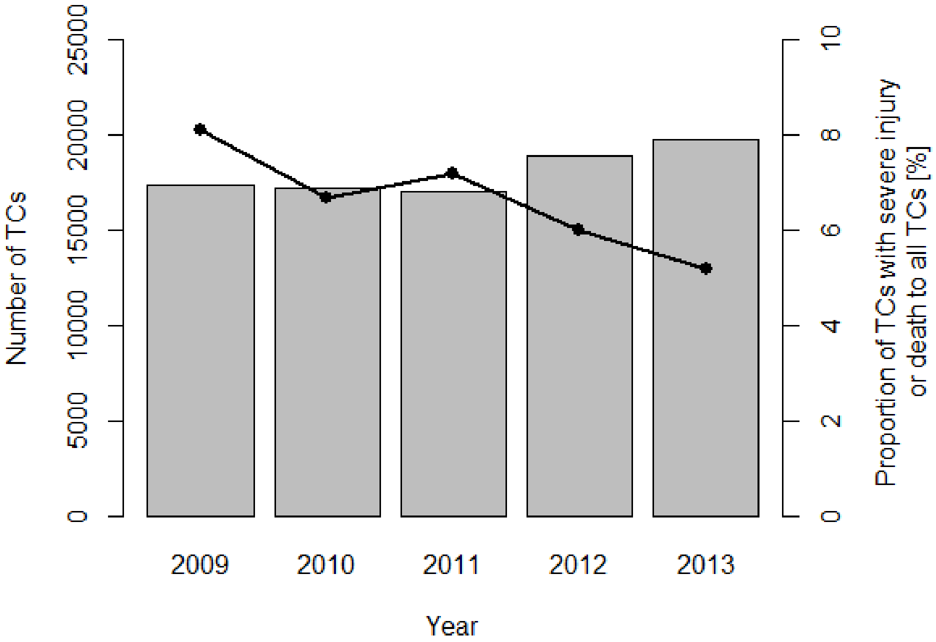


Figure 1.2: Number of TCs over the period 2009 – 2013 (excluding the urban network and intersections) and the proportion of TCs with severe injury or death in relation to all TCs.

Chapter 2

Kernel density estimation

To identify the hazardous locations, we used the KDE which is enriched by a statistical testing procedure to find the significant clusters of TCs and by a cluster ranking procedure. These additional steps were developed to overcome the drawbacks of the KDE method (see Introduction).

Methods like the KDE are usually defined as planar methods, while TCs are bound to the network, which is not a two-dimensional space. Some authors, however, ignored this fact and their results are therefore biased [12, 27].

This limitation can be overcome by using the network variant of the KDE [23, 31]. This approach is, however, not suitable in traffic-crash analyzes, because the use of the network variant of the KDE (or any other method) needs to reflect additional data like annual average daily traffic (or other estimates of the traffic flow). The reason comes from the fact, that the number of TCs depends on the intensity of the traffic and therefore they have to be weighted by a factor to get comparable results. The task is even more complicated, because the relation among the TCs and the daily traffic is not linear [18, 27].

Therefore, the preferred approach, when working with TCs, is to separate the road network into road sections and use the KDE in the one-dimensional space. Using the road sections, we do not require consideration of the effect of the annual average daily traffic, because the traffic remains constant within each section.

Let us briefly describe the KDE method. The KDE method depends on two parameters: the type of a kernel function and the bandwidth. First, the type of a kernel function is selected. A univariate kernel function, denoted by $K(x)$, is a real-valued integrable function satisfying:

- $K(x) \geq 0,\ \forall x \in \mathbb{R}$,
- $\int\limits_{-\infty}^{+\infty} K(x)\mathrm{d}x = 1$,
- $\int\limits_{-\infty}^{+\infty} xK(x)\mathrm{d}x = 0$.

Usually, the kernel function is an even probability density function [13]. In our research we used the Epanechnikov kernel (Figure 2.1). Many other shapes of the kernel could be selected such as rectangular, triangle or Gaussian.

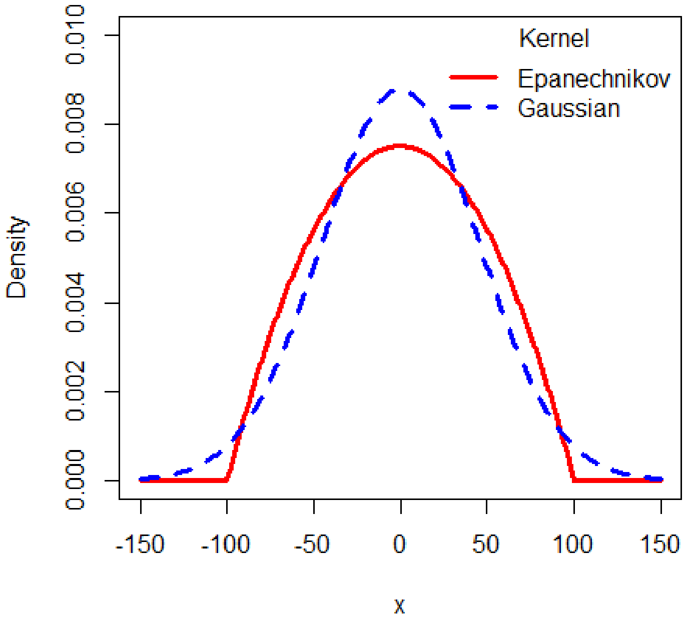


Figure 2.1: A comparison of the Epanechnikov and Gaussian kernels. The Epanechnikov kernel has bounded support. On the other hand, the Gaussian kernel tends asymptotically to zero for x going to $\pm\infty$ and its support is unbounded.

The kernel shape itself does not have a substantial impact on the results when compared with the bandwidth [4, 31]. However, the shape of the kernel should reflect the range of the uncertainty of the real position of a TC. In the case of the Gaussian kernel, which has, from its definition, unbounded support, the uncertainty expands to the whole section and beyond. This is not realistic and the uncertainty extent should be narrowed. Furthermore, the Epanechnikov kernel minimizes the asymptotic mean integrated squared error [19] defined as:

$$AMISE(\hat{f}) = \frac{1}{nd}\int\limits_{-\infty}^{+\infty} K(x)^2 \,\mathrm{d}x + \frac{1}{4}d^4 \int\limits_{-\infty}^{+\infty} x^2 K(x)\,\mathrm{d}x \int\limits_{-\infty}^{+\infty} [f''(x)]^2 \,\mathrm{d}x, \tag{2.1}$$

where $\hat{f}$ is the estimated probability density function, f stands for the underlying original probability density function, $d > 0$ is the bandwidth and $n \in \mathbb{N}$ denotes the number of TCs within the road section. The Epanechnikov kernel is defined as follows:

$$K_d(x) = \frac{3}{4d}\left(1 - \left(\frac{x}{d}\right)^2\right) \mathrm{I}_{(-d,d)}(x), \tag{2.2}$$

where $\mathrm{I}_{(-d,d)}(x)$ is the indicator function on the interval $(-d, d)$.

Secondly, we choose the bandwidth of the kernel. In general, the bandwidth is chosen in order to minimize (2.1) with respect to d [19]. Concerning the KDE method for TCs, the bandwidth choice is dependent on the character of the traffic (maximum speed of vehicles and visibility range). Commonly used bandwidths start at 50 m when applied in urban areas [30] and go up to 500 m in highway segments [12]. In our case we chose a 100 m long bandwidth. This value is reasonable for rural roads with respect to breaking distance and visibility range.

After setting the parameters, the KDE can be computed as a sum of the kernel functions (Figure 2.2), where modal points are locations of TCs. The KDE is defined as:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_d(x - X_i), \tag{2.3}$$

where $X_i \in (0, L)$, $i = 1, \ldots, n$, are the locations of the TCs, $n \in \mathbb{N}$ is the number of TCs within the road section and $L > 0$ is the length of the road section. Obviously, the area below the resulting density curve is equal to one, because it is a density function.

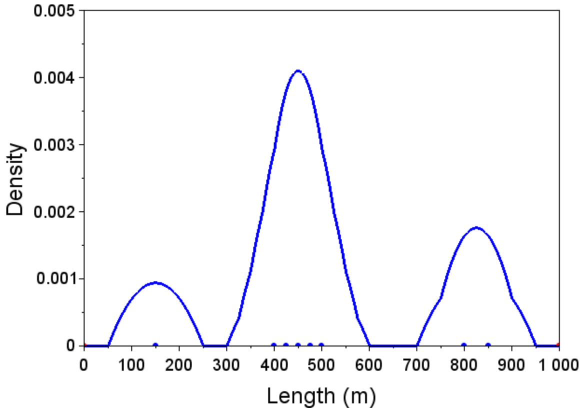


Figure 2.2: An example of the KDE (blue curve). Blue dots stand for the locations of the TCs. The threshold for significant clusters is unknown.

Naturally, the KDE method results in many clusters located at points of local maximums of the KDE. Therefore, the following steps are necessary:

1. Objectively determine the threshold for statistically significant clusters.

2. Select the significant clusters.

3. Order the significant clusters.

We improved the procedure of cluster detection, based on the standard KDE method, by testing the significance of the clusters (see section 2.1) followed by the ordering of the clusters (see section 2.2). Furthermore, we tested accuracy and stability of our approach (see sections 2.3 and 2.4) and we compared our method to other methods which were recently used in cluster analyzes of TCs (see section 2.5). Finally, we prepared a stand-alone application and toolbox for ArcGIS to allow the use of our approach also to other researchers (see section 2.6).

2.1. Testing the significance

At this point the common application of the KDE usually ends. Clusters are identified at places of local maximums of the kernel function. Sometimes an arbitrary threshold is determined [12]. However, we aim on determining the significance of a cluster more objectively.

If there is a spatial pattern of TCs on a particular road section, the distribution of TCs would have to be distinguishable from the uniform distribution. Therefore, our procedure follows by stating the null hypothesis:

$\mathbf{H}_0$: "The TCs are distributed randomly along the road section according to the uniform distribution on the interval $(0, L)$",

where $L > 0$ is the length of the road section. If $\mathbf{H}_0$ is rejected, we found a spatial pattern of TCs and hotspots can be identified. Otherwise, we cannot distinguish, whether there is any spatial pattern of TCs or not, and TCs were likely caused by a spatially random process. Statistical test of the null hypothesis is based on the Monte Carlo (MC) method, which uses repeated random simulations to determine properties of a problem in question.

Let us make the following notation:

h, H	–	thresholds
$\hat{f}(x)$	–	the probability density function of TCs estimated by the use of the KDE method,
L	–	the length of a road section,
n	–	the number of TCs on the road section,
M	–	the number of MC simulations.

Hotspots are located at places where function $\hat{f}(x)$ is significantly greater than the probability density function of the uniform distribution. In other words, we have to objectively determine the threshold expressing the significant difference between $\hat{f}(x)$ and the probability density function of the uniform distribution. Statistical test can be performed either from a local perspective ("local" test) or from a global perspective ("global" test). We apply the both tests simultaneously taking the local test as the primary test. The global test is performed for its informative value. The details concerning these two tests are given in sections 2.1.1 and 2.1.2.

2.1.1. Local test

The local test examines the values of $\hat{f}(x)$, $x \in (0, L)$, pointwise and compares them to the values of the uniform probability density function $1/L$. We set the significance level as $\alpha = 5\%$. The local test can be carried out by the use of the MC method in the following way (see Figure 2.3):

1. We choose n points by random form the uniform distribution on the interval $(0, L)$. The probability density function of this distribution is $1/L$ on the interval $(0, L)$ and zero otherwise.

2. We calculate the KDE of the randomly chosen points in step 1.

We perform a sufficient number of repetitions of steps $1 - 2$. Consequently, we arrive at M probability density functions estimated by the KDE method. Let us denote them as $g_1(x)$, $g_2(x)$, ..., $g_M(x)$.

3. The 95% quantile of $g_1(x)$, $g_2(x)$, ..., $g_M(x)$ is calculated at each point $x_0 \in (0, L)$. We denote this quantile as $q(x_0)$.

4. The threshold is determined as the mean value of $q(x)$, $x \in (0, L)$. More specifically,

$$h = \frac{1}{L} \int_0^L q(x) \, \mathrm{d}x$$

5. Finally, we arrive at significant clusters which are located at places with $\hat{f}(x) > h$. If there exists at least one significant cluster, we reject the null hypothesis.

The main advantage of this local approach is the simple localization of the clusters. In addition, we are able to measure the degree of violation of the null

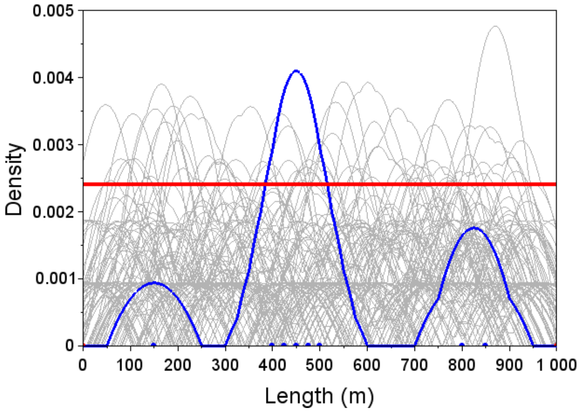


Figure 2.3: An example of the local test. The blue line shows the estimated probability density function of the underlying TCs. The gray lines represent KDEs of uniformly distributed data (the Monte Carlo method). The horizontal red line is the threshold (95th percentile level). In places, where the blue line is above the threshold, a significant cluster is identified.

hypothesis at each point within the road section. This feature facilitates the ordering of the significant clusters (see section 2.2). Due to these reasons, we further use the local test as the primary significance test for identifying and localizing the significant clusters.

The existence of at least one cluster within a road section is determined by the existence of $x_0 \in (0, L)$ such that $\hat{f}(x_0) > h$. Since the significance level α is set for each particular point $x_0 \in (0, L)$, which means that

$$P(\hat{f}(x_0) > h|\mathbf{H}_0) = \alpha,$$

we get

$$P(\exists x_0 \in (0, L) : \hat{f}(x_0) > h|\mathbf{H}_0) > \alpha.$$

Hence, the probability of type I error is likely higher than the predefined significance level α and we can expect a higher false alarm rate.

2.1.2. Global test

The global test examines, whether there is a clustering on the particular road section. It compares the maximum value of $\hat{f}(x)$ to the maximum value of the uniform probability density function, i. e. to the $1/L$. Again, we set the significance level $\alpha = 5\%$. The first two steps of the global test are the same as for the local test. The whole procedure is performed as follows (see Figure 2.4):

1. We choose n points by random from the uniform distribution on the interval $(0, L)$. The probability density function of this distribution is $1/L$ on the interval $(0, L)$ and zero otherwise.

2. We calculate the KDE of the randomly chosen points in step 1 and remember the maximum value.

We perform a sufficient number of repetitions of steps 1 – 2. Consequently, we arrive at M maximums. Let us denote them as m_1, m_2, $\ldots$, m_M.

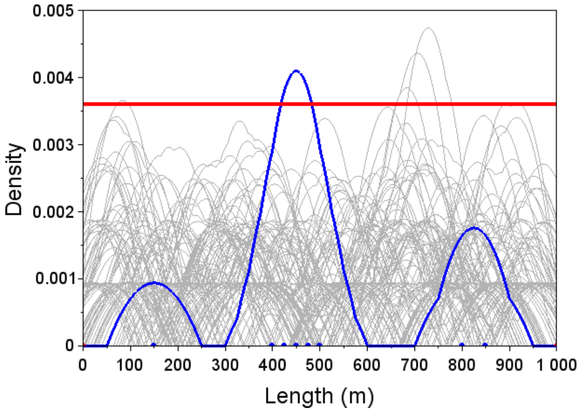


Figure 2.4: An example of the global test. The blue line shows the estimated probability density function of the underlying TCs. The gray lines represent KDEs of uniformly distributed data (the Monte Carlo method). The horizontal red line is the threshold (95th percentile level). In places, where the blue line is above the threshold, a significant cluster is identified.

3. The threshold H is determined as the 95th percentile of maximums m_1, m_2, $\ldots$, m_M.

4. If there exists $x_0 \in (0, L)$ such that $\hat{f}(x_0) > H$, we reject the null hypothesis. In other words, the null hypothesis is rejected if $\max_{x \in (0,L)} \hat{f}(x) > H$.

Significant clusters are located at places where $\hat{f}(x) > H$, $x \in (0, L)$, and in their neighborhood.

Since

$$P(\max_{x\in(0,L)} \hat{f}(x) > H|\mathbf{H}_0) = \alpha,$$

we conclude that the global test has lower false alarm rate than the local test. On the other hand, the miss rate of the global test is higher than in the local test, because $H > h$. In addition, the exact localization of the significant clusters is uncertain when using only the global test. Therefore, we apply the global test only for its informative value. The information obtained from the global test can be used to reduce the false alarm rate in the local test.

2.2. Cluster strength

Applying the proposed method, it is possible to determine which clusters within a road section are statistically significant (see Figure 2.3). Furthermore, we can determine the degree of significance for each cluster which allows us to compare the clusters among themselves. We call it a cluster strength. It is defined as a ratio between the maximum of $\hat{f}(x)$ within the cluster (denoted as $\hat{f}_{max}$) and the value of that maximum (see Figure 2.5), more specifically $(\hat{f}_{max} - h)/\hat{f}_{max}$. According to this definition, it holds that the cluster strength is a positive number less than one.

The cluster strength quantifies the degree of violation of the null hypothesis. The ranking of the clusters makes it possible for a user to sort all the significant clusters from the most hazardous to the least. The cluster strength is a function of the following four factors:

- number of TCs within the cluster and their mutual position,
- the length of the cluster,
- the total number of TCs within the road section and

- the length of the section.

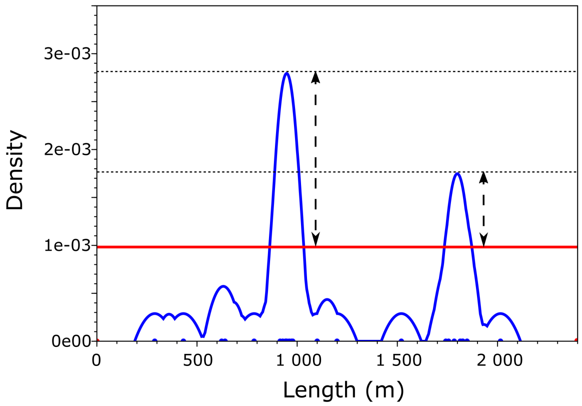


Figure 2.5: Cluster strength is the relative height of the density function (blue curve) above the threshold (red horizontal line). Cluster on the left is stronger than cluster on the right.

In order to show how the cluster strength is influenced by the four mentioned factors, we prepared a table (see Table 2.1), where four rows show clusters of the same strength (0.72), but one of the factors has been changed in every row. It is clear from Table 2.3 that if number of TCs within a cluster or the length of the section increases, the cluster strength also rises. On the other hand, when the length of the cluster or the number of the TCs out of the cluster grows, the cluster strength decreases.

Table 2.1: Each row represents a cluster with the cluster strength of 0.72. Three of the four settings are kept constant.

TCs/cluster	TCs/section	Cluster length [m]	Section length [m]
10	15	100	2200
10	15	170	3000
10	21	100	3000
8	15	100	3000

When the length of a section rises, the probability that TCs occur within a predefined distance drops. For example, if there are two TCs which are located elsewhere in a section but no more than 100 m apart, the probability that these two TCs lie within the selected distance accounts for 19% for a 500 m long section, but only 9.75% for a 1000 m long section. Therefore, the cluster strength is lower (0.16 – 0.22) for the 500 m long road section than for the 1000 m long road section (0.47 – 0.50).

The cluster strength shows, in descending order, the clusters from the most significant to the least. This measure is suitable for ranking the clusters from the view of a driver. It is something like a hazard for an individual driver. On the other hand, we are aware that decision-makers are usually more interested in mitigating the cumulative danger for all drivers. Therefore, the density of TCs within a cluster combined with the cluster strength can be in the interest of road administrators as well.

2.3. Accuracy of the Monte Carlo method

It is apparent that the threshold, and therefore also the cluster strength, varies when performing repeated runs of the KDE+ method for the same data. This is caused by the random character of the MC method. The accuracy of the MC method increases with the number of simulations. In contrast, the number of simulations determines time-consumption of the analysis. Therefore, we needed to find a balance between the accuracy of the MC method and its time-consumption by setting the appropriate number of simulations.

In order to evaluate the accuracy of the MC method, we used confidence intervals of cluster strength. These confidence intervals can be easily derived from the confidence intervals of the threshold following the same approach as during the calculation of the cluster strength. It should be recalled that the threshold is the $100(1-\alpha)\%$ quantile (or its mean value). Thus, the $100(1-\beta)\%$ confidence interval of the threshold (see Figures 2.6 and 2.7) can be calculated by the use of the binomial distribution in the following way.

For a fixed $x_0 \in (0, L)$, let us suppose that M simulations were performed and probability density functions estimated by the use of the KDE method during the simulations, denoted by $g_1(x_0)$, $g_2(x_0)$, …, $g_M(x_0)$, are ordered increasingly. We calculate indices

$$i = \max\left\{ l \in \{0, 1, \ldots, M+1\}; \sum_{k=0}^{l-1} \binom{M}{k} (1-\alpha)^k \alpha^{M-k} \leq \frac{\beta}{2} \right\},$$

$$I = \min\left\{ u \in \{0, 1, \ldots, M+1\}; \sum_{k=u}^{M} \binom{M}{k} (1-\alpha)^k \alpha^{M-k} \leq \frac{\beta}{2} \right\}$$

and denote $q_{low}(x_0) = g_i(x_0)$ and $q_{up}(x_0) = g_I(x_0)$. This process is performed for all $x_0 \in (0, L)$. See [25] for the details on computing indices i and I. Finally, we arrive at the $100(1-\beta)\%$ confidence interval (h_{up}, h_{low}) of h by calculating

$$h_{low} = \frac{1}{L} \int_0^L q_{low}(x) \ \mathrm{d}x, \ \ h_{up} = \frac{1}{L} \int_0^L q_{up}(x) \ \mathrm{d}x.$$

Similarly, the $100(1-\beta)\%$ confidence interval of H accounts directly for (m_i, m_I), assuming that maximums calculated during the simulations, denoted by m_1, m_2, …, m_M, are ordered increasingly.

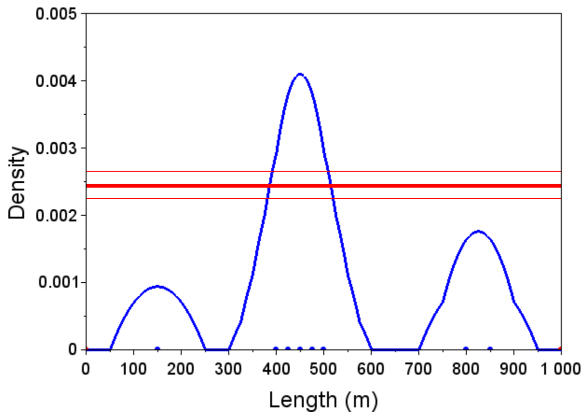


Figure 2.6: The estimated probability density function of the underlying TCs (blue curve), the threshold in the local test (thick red line) and its 99% confidence interval (thin red lines).

The $100(1-\beta)\%$ confidence interval of a cluster strength can be computed as follows:

$$\left(\frac{\hat{f}_{max} - h_{up}}{\hat{f}_{max}}, \frac{\hat{f}_{max} - h_{low}}{\hat{f}_{max}} \right) \subset (0, 1),$$

where $\hat{f}_{max}$ is the maximum of $\hat{f}(x)$ within the particular cluster. Although we order the clusters according to their strength, we can conclude that two clusters differ significantly with respect to the cluster strength only if the confidence intervals of their cluster strengths do not overlap.

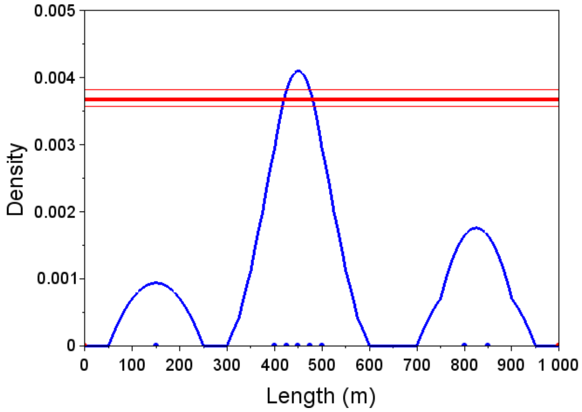


Figure 2.7: The estimated probability density function of the underlying TCs (blue curve), the threshold in the global test (thick red line) and its 99% confidence interval (thin red lines).

2.4. Stability of significant clusters

Stability in general means that a small change in input data leads to a small change in the result. Regarding clusters, two types of stability can be considered: temporal stability and stability related to the database of TCs. We focused on the later type of stability.

Bíl et al. [7] introduced a simple test for stability of a cluster. With the use of the stability test we can focus on the most important clusters. Furthermore, the stability test eliminates possible mistakes in the database (e. g. a TC can be snapped to a wrong road section or the location of a TC can be recorded incorrectly).

Stability of a cluster shows to what extend the cluster depends on the accuracy of the input data. Stable clusters are not affected by a slight change in the number of TCs or in their positions along the road section. On the other hand, an unstable cluster disappears or its cluster strength varies significantly after deleting or adding a single TC. Since underreporting is a frequent problem, the stability of any method is an important feature. We focused on the stability of the KDE+ method with respect to the input data in the following examples.

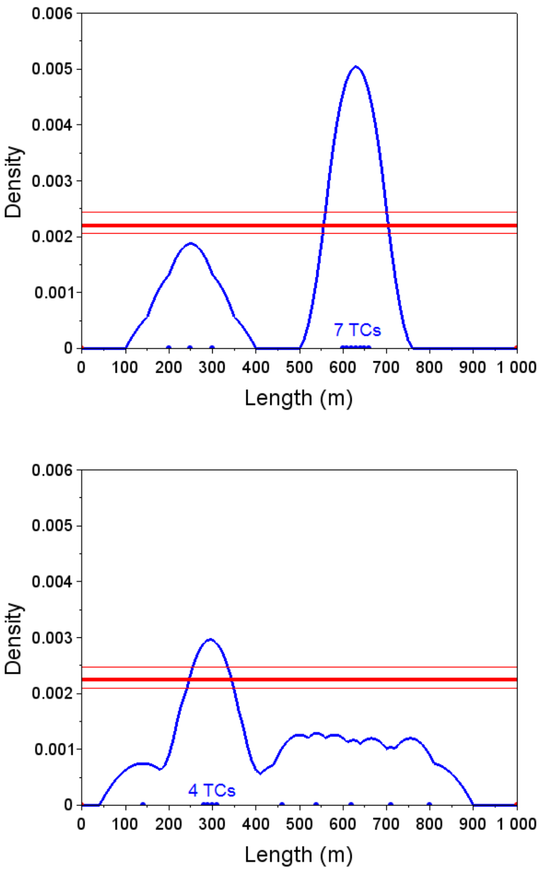


Figure 2.8: Graphical representation of the KDE+ applied to two road sections 1000 m long. TCs occurred at the places represented by blue dots. Thick red horizontal line stands for the threshold and thin red horizontal lines for its 99% confidence interval.

Imagine two road sections, each 1000 m long. Within each road section, ten TCs occurred and one significant cluster was identified by the use of the KDE+ method (see Figure 2.8). The cluster within the first road section is more hazardous (the cluster strength accounts for 0.56) than the cluster on the second road section (the cluster strength is 0.23).

We calculated the relative frequency of finding a significant cluster in the same place as the original cluster after removing TCs on a road section. All possible combinations of removed TCs were taken into account. This calculation was conducted for the two road sections depicted in Figure 2.8. The obtained results (see Table 2.2) demonstrate that the KDE+ is stable even if only a limited number of crashes are available. In addition, the cluster with a greater strength is more stable. It means that more important clusters are less prone to disappear when some crashes are missing in the database. This feature is of exceptional importance because crash database usually contains only a limited number of data. The issue of stability of the significant clusters is further studied on the actual data in section 4.1.

Table 2.2: Relative frequencies of a cluster identification at the same place as the original cluster after removing a given number of TCs on a road section (see Figure 2.8). All possible combinations of removed TCs were taken into account.

Removed crashes [%]	Cluster identification rate [%]	
	Section 1	Section 2
10	100	100
20	100	87
30	100	67
40	100	45
50	99	74
60	97	55
70	82	33

2.5. Comparisons with other clustering methods

To show the contribution of the proposed approach, we compared the KDE+ approach with other methods which were recently used in cluster analyzes of TCs. We used:

- the K-function [32],
- hierarchical clustering [22],
- the dangerousness index (DI) method [30] and
- the clumping method [22].

It is clear that the K-function is able only to specify whether there is clustering on a section without specifying where on the section it occurred.

Hierarchical clustering does not have any tool to evaluate statistical significance of clusters. It was only able to identify clusters of TCs.

The DI method is in fact a special case of the KDE. It is based on the "points of measurements". We think, however, that the positions of crashes should be used instead of "points of measurements". When there is not a substantially dense network of "points of measurements" some localities can be underreported. Two exact situations can be evaluated differently in relation to positions of "points of measurements" and TCs. When the network of "points of measurements" is dense enough, the method converges to the standard KDE.

The clumping method is able to detect the cluster positions, but it is too sensitive. It means that a small change of location of TCs outside a cluster can affect the significance of the cluster itself.

Our approach improves the standard KDE by statistical testing and the cluster strength computation which serves as a tool for direct comparison of the significant clusters (see Table 2.3).

Table 2.3: Comparison of clustering methods used for hotspots detection. Symbols ”+” and ”−” represents YES and NO, respectively.

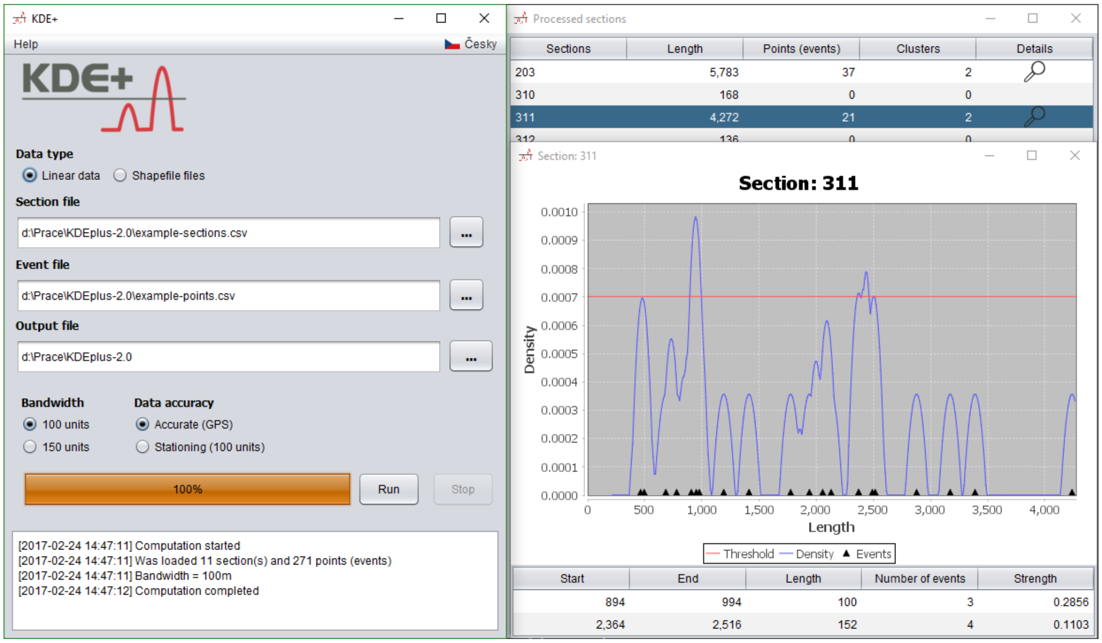
	Cluster localization	Significance test	Ordering	Stability
KDE	+	−	−	+
Dangerousness index	+	+	−	+
K-function	−	−	−	+
Hierarchical clustering	+	−	−	+
Clumping method	+	+	−	−
KDE+	+	+	+	+

2.6. Software realization

The KDE+ method has been already developed in a software package and as a toolbox for ArcGIS. Both the software package and the toolbox can be downloaded as a freeware (only registration of an e-mail address is needed to download the application) from www.kdeplus.cz website as a single compressed zip file. This folder contains application KDEplus.jar. Java Runtime Environment (from version 7) is needed to run this application. It can be downloaded from http://java.com/download. The folder with the application also includes two demonstration files, one for the road section and the second for TCs.

The KDE+ is a desktop application with windows. The main window serves for file import and allows for the running of the computing. Important reports are written in the text box at the bottom. An additional window appears if the user is interested in examining the graphics representation of a particular road section (see Figure 2.9). There are visualized graphs of the estimated probability density function and the threshold. The interface is currently written in Czech and English.

Several computation threats can be run in parallel fashion in order to take advantage of multiple processor cores. This feature shortens the computing time significantly and the computer is used effectively since contemporary computers are equipped with more than one core. The application can also be used without the window environment, i.e. in a command line. This would be useful in grid computing in the case of processing a large amount of data.



Data type
● Linear data ○ Shapefile files

Section file
d:\Prace\KDEplus-2.0\example-sections.csv

Event file
d:\Prace\KDEplus-2.0\example-points.csv

Output file
d:\Prace\KDEplus-2.0

Bandwidth
● 100 units
○ 150 units

Data accuracy
● Accurate (GPS)
○ Stationing (100 units)

100%

```
[2017-02-24 14:47:11] Computation started
[2017-02-24 14:47:11] Was loaded 11 section(s) and 271 points (events)
[2017-02-24 14:47:11] Bandwidth = 100m
[2017-02-24 14:47:12] Computation completed
```

Sections	Length	Points (events)	Clusters	Details
203	5,783	37	2	🔍
310	168	0	0	
311	4,272	21	2	🔍
312	136	0	0	

Start	End	Length	Number of events	Strength
894	994	100	3	0.2856
2,364	2,516	152	4	0.1103

Figure 2.9: Working environment of the KDE+ software.

Chapter 3

The KDE+ method for interval data

The Czech Police adds a GPS location to all TCs since 2007. However, we were aware that certain countries do not use precise GPS positions to georeference TCs. We discovered that the most widely used system for referencing the TCs, different from GPS localization, uses the linear referencing system (LRS) with 100 m accuracy. This means that TCs are located with 100 m precision from the beginning of a road segment to its end. In such a setting, where the data is not precise due to a systematic error (e.g. rounding error), the KDE+ method would not work properly. We took this fact into consideration and extended the KDE+ method to be applicable to interval data. In both the software package and the toolbox for ArcGIS, a user can select the input data type (GPS or LRS).

The modification of the KDE+ method affects only the choice of the kernel function. Originally, we used the Epanechnikov kernel to the exact GPS positions of the TCs. The new kernel $\varphi_{d,v}(x)$ with a bandwidth $d > 0$ and $v > 0$ quantifying the inaccuracy, is derived from the Epanechnikov kernel and reflects the uncertainty of the position of a TC.

The GPS position of a TC belongs to an interval determined by the uncertainty of the LRS. Thus, the error of the LRS for a particular TC can be considered as a random variable and its probability density function is defined on the interval $(-v, v)$, where $v > 0$ is the maximal rounding error (e. g. uncertainty of the LRS). Let us denote this random variable as Z. Since we do not have any further information on the exact position of the TC, we can assume that Z has the uniform distribution with support $(-v, v)$. Hence, the probability density function of Z is

$$g(z) = \begin{cases} \frac{1}{2v}, & |z| \leq v, \\ 0, & |z| > v. \end{cases} \tag{3.1}$$

Let us think symbolically for a while. In the original setting, the risk of a TC is spread by applying the probability density function $K_d(x - X_0)$, where X_0 is the exact position of the TC given by the GPS coordinates. If the GPS position of a TC is uncertain, we have to consider all its possible values. Thus, $\varphi_{d,v}(x) = \sum_z f_d(x|Z = z)P(Z = z)$ according to the formula of the total probability. The conditioned probability density function has the form $f_d(x|Z = z) = K_d(x - z)$, because if the exact position of the TC is z, then we spread the risk of the TC by the use of the Epanechnikov kernel around this point. In order to perform the derivation properly, we calculate the probability density function of $\varphi_{d,v}(x)$ as follows

$$\begin{aligned} \varphi_{d,v}(x) &= \int\limits_{-\infty}^{+\infty} f_d(x|z)g(z)\ \mathrm{d}z = \int\limits_{-\infty}^{+\infty} K_d(x - z)g(z)\ \mathrm{d}z = \\ &= (K_d * g)(x), \end{aligned} \tag{3.2}$$

which means that the new kernel is a convolution of the Epanechnikov kernel and function $g(z)$ (the uniform probability density function in our case). This result does not depend on the assumed shape of $g(z)$. Hence, any other probability density function can be used, if it is reasonable.

The explicit formula of the resulting convolution (3.2) with $g(z)$ given by (3.1) depends on the values d and v. First, we define four auxiliary functions:

$$\begin{aligned} F_1(x) &= \frac{-3vd^2 + (x+v)^3}{8vd^3}, \\ F_2(x) &= \frac{3vd^2 + (x-v)^3}{8vd^3}, \\ F_3(x) &= \frac{3x - 2d}{8vd}, \\ F_4(x) &= \frac{3x + 2d}{8vd}. \end{aligned} \tag{3.3}$$

If $d \geq v$, then

$$\varphi_{d,v}(x) = \begin{cases} F_4(x) - F_1(x), & |x+d| < v, \\ F_2(x) - F_3(x), & |x-d| < v, \\ F_2(x) - F_1(x), & |x| \leq d - v, \\ 0, & |x| > d + v, \end{cases}$$

otherwise ($d < v$), it holds that

$$\varphi_{d,v}(x) = \begin{cases} F_4(x) - F_1(x), & |x+v| < d, \\ F_2(x) - F_3(x), & |x-v| < d, \\ F_4(x) - F_3(x), & |x| \leq v - d, \\ 0, & |x| > v + d, \end{cases}$$

As expected, $\varphi_{d,v}(x)$ has wider support than $K_d(x)$ due to the uncertainty in the data (see Figure 3.1).

Finally, the kernel density estimation is provided by the formula

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{d,v}(x - Y_i), \tag{3.4}$$

where $Y_i \in (0, L)$, $i = 1, \ldots, n$, are the LRS positions of TCs, $n \in \mathbb{N}$ is the number of TCs within the particular road section and $L > 0$ denotes the length of the road section.

The application of kernel function $\varphi_{d,v}(x)$ is a better option than the use of $K_d(x)$ in the case of the LRS data, because:

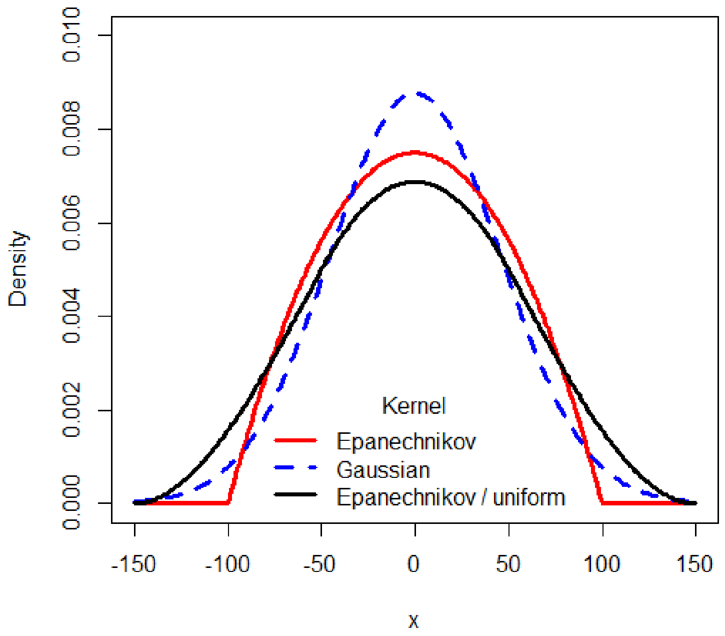


Figure 3.1: A comparison of the Epanechnikov kernel and the combined Epanechnikov/uniform kernel ($d = 100$, $v = 50$).

- $\varphi_{d,v}(x)$ is correct while $K_d(x)$ is incorrect from the theoretical point of view,
- $K_d(x)$ leads to only one possible outcome hidden behind the LRS data, while the use of $\varphi_{d,v}(x)$ takes into account all possible outcomes (see Figure 3.2),
- considering the case study depicted in Figure 3.3, $K_d(x)$ can result in false clusters (although there are three significant clusters determined by the use of $K_d(x)$, there should be only one significant cluster).

We remark that formula (3.4) can be also used to estimate a probability density function from interval data with unequal lengths of intervals. It can be done in the following way:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{d,v_i}(x - Y_i), \tag{3.5}$$

where Y_i, $i = 1, \ldots, n$, are centers of the intervals and v_i, $i = 1, \ldots, n$, are half-lengths of the intervals.

Concerning the spatial analysis of TCs, it stems from the nature of the LRS that the uncertainty in the exact position of a TC is the same for all TCs in a particular database. Thus, we use formula (3.4) and not its more general version (3.5) for estimating a probability density function of TCs given with some systematic error.

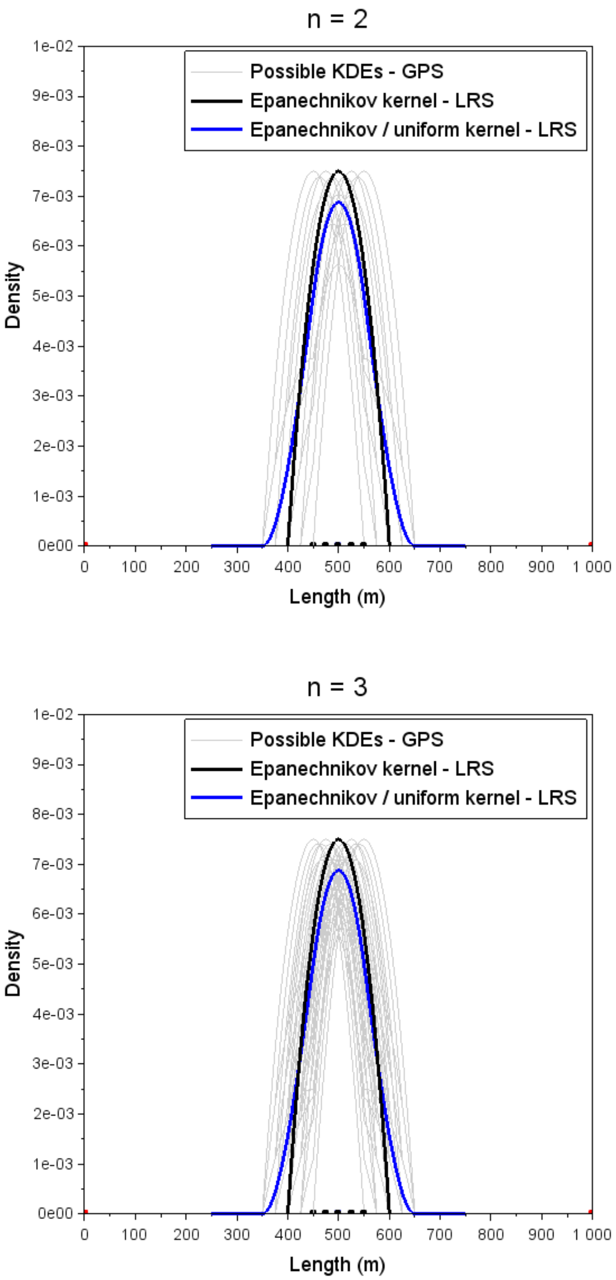


Figure 3.2: A comparison of possible KDEs of GPS data with the classical Epanechnikov kernel and combined Epanechnikov/uniform kernel applied to LRS data for two TCs (top) and three TCs (bottom) which are located in the interval $\langle 450, 550 \rangle$ within a kilometre-long road section.

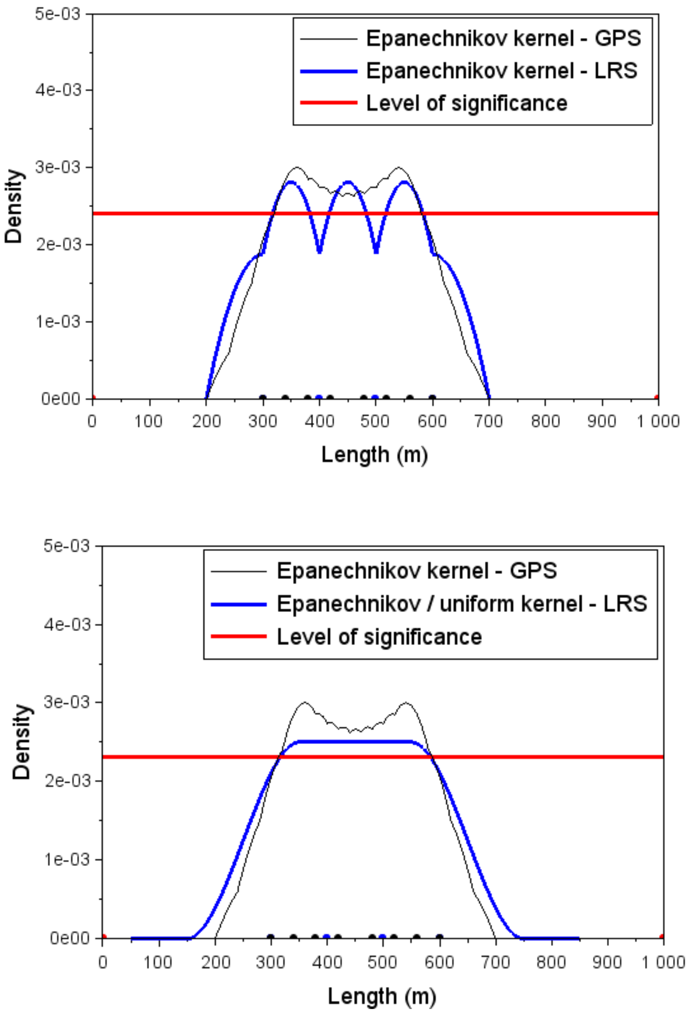


Figure 3.3: Performance of Epanechnikov kernel (top) and Epanechnikov/uniform kernel (bottom) applied to LRS data (8 TCs).

Chapter 4

Results

The KDE+ method was applied to the Czech road network. TCs without distinction were analyzed at first. We identified 3.12% of the entire road network length as hazardous. It consists of 8,739 significant clusters containing 37,885 (41.9%) TCs.

In practice, it is usually not possible to apply mitigation measures at all hazardous places due to a limited budget allocated for this purpose. Therefore, the ordering of the hotspots is necessary. The KDE+ method enabled us to rank the significant clusters according to their strength. Subsequently, road administrators can select the strongest clusters for applying mitigation measures or focus only on the clusters which are significant also according to the global test. For instance, there were 86 clusters with cluster strength greater than 0.7 (see Figure 4.1) covering 22.3 km (0.06% of the entire road network). The most hazardous location was 225 m long and contained 63 TCs. Its cluster strength was 0.88. These very strong clusters are likely not false alarms produced by the method.

We used the KDE+ method to examine clustering of specific types of TCs, namely single-vehicle TCs, two-vehicles TCs and TCs with severe injury or death. Table 4.1 shows the outcomes of the performed analysis. Clusters of TCs with severe injury or death were the shortest on average. TCs with severe injury or death have the lowest tendency to form patterns (only 15.3%). The most hazardous places were depicted in a map [6].

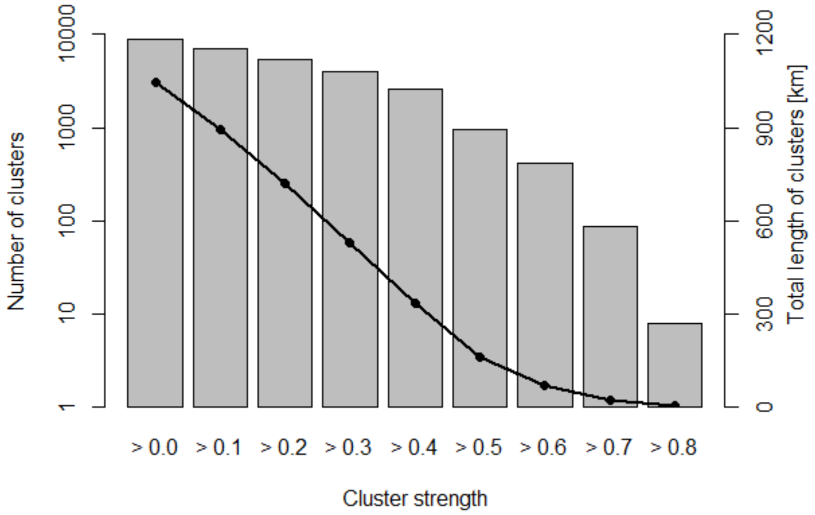


Figure 4.1: Number of clusters of TCs on the Czech road network (bars) and the total length of clusters (dots) with respect to the cluster strength.

Table 4.1: The results of cluster analysis performed by the use of the KDE+ method on the Czech road network. The data on TCs were recorded over the period 2009 – 2013.

Group of TCs	Without distinction	Single-vehicle	Two-vehicles	With severe injury or death
Number of TCs	90,418	59,811	26,512	5,953
Number of clusters	8,910	6,555	2,657	406
Number of TCs in clusters [%]	41.9	39.9	31.8	15.3
Total length of clusters [%]	3.12	1.98	0.71	0.08
Mean length of clusters [m]	120	113	101	70

4.1. Performance of the Monte Carlo method on actual data

We analyzed the TCs without distinction by the use of the KDE+ method with varying number of MC simulations ranging from 200 to 1200. Computations were performed in Scilab 5.5.0 [28] on PC Intel Core i7 (2.7 GHz) with 8 GB RAM. We set $\beta = 0.01$ in our computations.

The results varied very little for the number of simulations greater than 400 and varied insignificantly for the number of simulations greater than 800 (see Table 4.2). Therefore, we concluded that 800 simulations represent a sufficiently good balance between accuracy and time-consumption of the KDE+ method. This balance is primarily needed when performing the analysis multiple times (e. g. for various time periods, for various types of TCs).

Table 4.2: Results from the KDE+ analysis of the Czech road network with varying number of MC simulations.

Number of simulations	Number of clusters	TCs within clusters [%]	Length of clusters [%]	Time [h]
200	8,253	39.7	3.01	1.05
400	8,759	41.4	3.10	2.13
600	8,835	41.7	3.11	3.25
800	8,910	41.9	3.12	4.27
1000	8,953	42.0	3.12	5.17
1200	8,955	42.0	3.12	6.40

We observed that stronger clusters have narrower confidence intervals in their strengths (Figure 4.2). This can be explained in the following way. The cluster strength is defined as the maximum relative height of $\hat{f}(x)$, $x \in (0, L)$, within the cluster above the threshold. Let us denote

$\hat{f}_{max}$	–	the maximum of $\hat{f}(x)$ within the particular cluster,
r	–	the relative error of the MC method,
h	–	the exact threshold (unknown),
$h \pm rh$	–	the threshold calculated by the use of the MC method (known),
s	–	the exact cluster strength (unknown),
$\hat{s}$	–	the estimated cluster strength (known).

From the definition of the cluster strength, we have

$$s = \frac{\hat{f}_{max} - h}{\hat{f}_{max}}, \ \hat{s} = \frac{\hat{f}_{max} - (1 \pm r)h}{\hat{f}_{max}}.$$

Practically, we can calculate only $\hat{s}$. Therefore, we are interested in evaluating the difference between the estimated and exact cluster strengths. It holds that

$$|s - \hat{s}| = \left| \frac{\hat{f}_{max} - h + \hat{f}_{max} + (1 \pm r)h}{\hat{f}_{max}} \right| = \frac{rh}{\hat{f}_{max}} = r(1 - s).$$

Hence, the greater strength of a cluster leads to a more precise estimate. Therefore, we are more certain in terms of more hazardous locations.

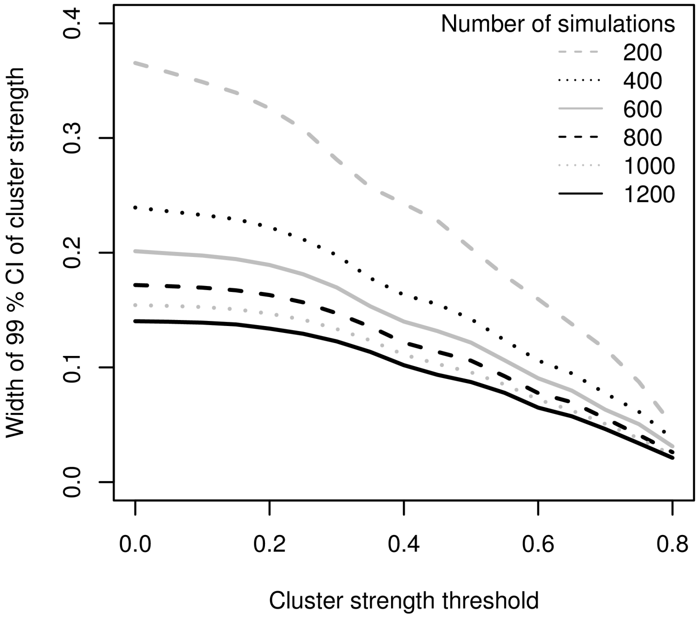


Figure 4.2: The average widths of confidence intervals of cluster strengths for various numbers of simulations and varying minimum cluster strengths.

4.2. Stability of the significant clusters within the Czech road network

We testified significant clusters with the cluster strength of more than 0.5 for their spatial stability with respect to the missing data. Since the number of possible combinations of removed crashes is extremely large for road sections with many TCs, we randomly chose a thousand of combinations for each road section and each particular percentage of removed crashes as an estimate of the relative frequency of cluster identification.

Table 4.3 shows that underreporting of 20% is not a problem in localization of almost all significant clusters with cluster strength greater than 0.5. Furthermore, majority (95%) of the clusters is more likely identified than not when even 50% of data is missing. In average, it is more than three times more likely to find a significant cluster than not when up to 60% data is missing in the database.

Table 4.3: Relative frequencies of cluster identification at the same place as the original cluster after removing a given percentage of TCs on a road section. Actual data – significant clusters within the Czech road network with cluster strength above 0.5 were considered.

Removed crashes [%]	Cluster identification rate [%]	
	In average	5th percentile
10	100	100
20	99	96
30	98	88
40	94	70
50	84	50
60	76	33
70	60	0

Conclusions

We introduced a method which should assist road administrators in the quick identification of the most hazardous places within a transportation network. All the data which are needed for the analysis performed by the KDE+ method are just:

- the positions of TCs on a road section and
- the length of the road section.

This feature is of merit, because any other characteristics about the traffic and the infrastructure are not needed for the identification of hazardous locations.

The method produces a dimensionless number, cluster strength, by which it is possible to order the previously identified hotspots. The presented results allowed the road administrators to effectively localize the most dangerous places within the road network.

A comparison of the KDE+ method with other methods for the identification of hazardous locations was published in [7]. The main advantage of the KDE+ method is its stability and objectivity. In addition, the strength of a cluster is a measure which enables the ordering of clusters. This unique feature of the method helps road administrators apply mitigation measures in the most effective way.

We had the GPS locations of all TCs from 2009 to 2013. This is not, however, the case in many European countries. Therefore, we extended the framework of the KDE+ method to also be applicable for LRS data. A new kernel function was derived and tested. Our results demonstrate that the new kernel function is appropriate for LRS data from both theoretical and practical view.

In order to enable the application of our KDE+ method also to other researches, we developed the KDE+ software [5]. It can be downloaded as freeware from the www.kdeplus.cz website. The KDE+ software can benefit from multi-core computers, because it allows for parallel computing in several threads. This feature significantly shortens the time needed for computation. Therefore, it can be used when processing a large amount of data.

The KDE+ method was applied to the entire Czech road network to obtain a list of significantly hazardous locations (clusters). The presence of clusters indicates the unlikely arrangement of TCs within a road section. TCs inside clusters follow a local spatial pattern. This means that the majority of TCs inside clusters were induced by local factors which should be consequently determined and investigated in the next step of the safety analysis.

Bibliography

[1] Anderson, T. K.: Kernel density estimation and K-means clustering to profile road accidents hotspots, Accident Analysis & Prevention 41, 359 – 364 (2009).

[2] Andrášik, R., Bíl, M.: Traffic accident hotspots: Identifying the boundary between the signal and the noise. Risk, Reliability and Safety: Innovating Theory and Practice – Walls et al. (Eds), 1634 – 1637, Taylor & Francis Group, London, ISBN 978-1-138-02997-2 (2016).

[3] Andrášik, R., Bíl, M.: Traffic accidents: Random or pattern occurence?, Safety and Reliability of Complex Engineered Systems – Podofillini et al. (Eds), 3 – 6, Taylor & Francis Group, London, ISBN 978-1-138-02879-1 (2015).

[4] Bailey, T. C., Gatrell, A. C.: Interactive Spatial Data Analysis, Longman, Essex, UK (1995).

[5] Bíl, M., Andrášik, R., Sedoník, J., Svoboda, T.: The KDE+ software: a tool for effective identification and ranking of animal-vehicle collision hotspots along networks, Landscape Ecology 31(2), 231 – 237 (2016).

[6] Bíl, M., R. Andrášik, Sedoník, J.: Clusters of traffic accidents on the road and motorway network in the Czech Republic over the period 2009 – 2013, map 1:520 000, ISBN 978-80-88074-02-1 (2014).

[7] Bíl, M., Andrášik, R., Janoška, Z.: Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation, Accident Analysis & Prevention 55, 265 – 273 (2013).

[8] Cheng, W., Washington, S. P.: Experimental evaluation of hotspot identification methods, Accident Analysis & Prevention 37, 870 – 881 (2005).

[9] Clegg, P., Bruciatelli, L., Domingos, F., Jones, R. R., De Donatis, M., Wilson, R. W.: Digital geological mapping with tablet PC and PDA: a comparison, Computers & Geosciences 32, 1682 – 1698 (2006).

[10] Elvik, R.: The predictive validity of empirical Bayes estimates of road safety, Accident Analysis & Prevention 40, 1964 – 1969 (2008).

[11] Elvik, R.: A survey of operational definitions of hazardous road locations in some European countries, Accident Analysis & Prevention 40, 1830 – 1835 (2008).

[12] Erdogan, S., Yilmaz, I., Baybura, T., Gullu, M.: Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar, Accidents Analysis & Prevention 40, 174 – 181 (2008).

[13] Everitt, B., Hothorn, T.: An Introduction to Applied Multivariate Analysis with R, Springer-Verlag, New York (2011).

[14] Flahaut, B., Mouchart, M., San Martin, E., Thomas, I.: The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach, Accident Analysis & Prevention 35, 991 – 1004 (2003).

[15] Geurts, K., Wets, G.: Black Spot Analysis Methods: Literature Review. Flemish Research Center for Traffic Safety, Belgium (2003).

[16] Gundoglu, I. B.: Applying linear analysis methods to GIS-supported procedures for preventing traffic accidents: Case study of Konya, Safety Science 48, 763 – 769 (2010).

[17] Hauer, E.: Observational Before-After Studies in Road Safety, Pergamon Press, Oxford (1997).

[18] Hiselius, L. W.: Estimating the relationship between accident frequency and homogenous and inhomogenous traffic flows, Accident Analysis & Prevention 36, 985 – 992 (2004).

[19] Ledl, T.: Kernel density estimation: theory and application in discriminant analysis. Australian Journal of Statistics 33(3), 267 – 279 (2004).

[20] Li, L., Zhu, L., Sui, D. Z.: A GIS-based Bayesian approach for analyzing spatial-temporal patters of intra-city motor vehicle crashes, Journal of Transport Geography 15, 274 – 285 (2007).

[21] Lord, D., Mannering, F.: The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives, Transportation Research A 44(5), 291 – 305 (2010).

[22] Okabe, A., Sugihara, K.: Spatial Analysis Along Networks: Statistical and Computational Methods (Statistics in Practice), Wiley-Blackwell (2012).

[23] Okabe, A., Satoh, T., Sugihara, K.: A kernel density estimation method for networks, its computational method and a GIS-based tool, Journal of Geographical Information Science 23 (1), 7 – 32 (2009).

[24] Okabe, A., Yamada, I.: The K-function method on a network and its computational implementation, Geographical Analysis 33(3), 152 – 175 (2001).

[25] Owen, A. B.: Monte Carlo theory, methods and examples (2013). Available from: http://statweb.stanford.edu/∼owen/mc/

[26] Plug, C., Xia, J., Caulfield, C.: Spatial and temporal visualization techniques for crash analysis. Accident Analysis and Prevention 43, 1937 – 1946 (2011).

[27] Sabel, C. E., Kingham, S., Nicholson, A., Bartie, P.: Road Traffic Accident Simulation Modelling – A Kernel Estimation Approach, Presented at SIRC 2005 – The 17th Annual Colloquium of the Spatial Information Research Centre, University of Otago, Dunedin, New Zealand (2005).

[28] Scilab Enterprises 2012: Scilab. Free and Open Source soft-ware for numerical computation (Version 5.5.0), Available from: http://www.scilab.org.

[29] Stark, B. L., Young, D. L.: Linear nearest neighbor analysis, American Antiquity 46 (2), 284 – 300 (1981).

[30] Steenberghen, T., Aerts, K., Thomas, I.: Spatial clustering of events on a network, Journal of Transport Geography 18, 411 – 418 (2010).

[31] Xie, Z., Yan, J.: Kernel density estimation of traffic accidents in a network space, Computers, Environment and Urban Systems 32, 396 – 406 (2008).

[32] Yamada, I., Thill, J. C.: Comparison of planar and network K-functions in traffic accident analysis, Journal of Transport Geography 12, 149 – 158 (2004).