

Richard Andrášik\*, Michal Bíl

Transport Research Centre, Líšeňská 33a, 636 00 Brno, Czech Republic

\*e-mail: andrasik.richard@gmail.com

# Clustering of traffic accidents with the use of the KDE+ method

## TABLE OF CONTENTS

1 Introduction .....	2
2 Data .....	4
3 The KDE+ method .....	6
Modification of the KDE+ method.....	6
Stability of a cluster .....	10
Stability related to the database of traffic accidents .....	10
KDE+ software .....	10
4 Results .....	12
5 Discussions and conclusions .....	14
References .....	15

# 1 INTRODUCTION

The aim of any developed society should be to prevent traffic accidents (TA) and reduce the severity of their consequences. From the point of view of a road administrator, precise identification of dangerous places within a road network is an essential tool for applying mitigation measures. However, standard methods of dangerous places identification only take into account aggregated data. They evaluate the safety of a road section as a whole (Hauer, 1997; Lord and Mannering, 2010) or test the general tendency to form clusters on a particular road section (Okabe and Yamada, 2001; Yamada and Thill, 2004).

Traffic-accident data collected by the Czech Police have their GPS localizations as of 2007. This feature makes the Czech road accident database unique even within developed countries. Spatial analyses are therefore possible in high detail.

Kernel density estimation (KDE; Sabel et al., 2005; Erdogan et al., 2008; Chung et al., 2011; Plug et al., 2011) can be used to identify the position of a cluster within a road section. However, the results obtained by the KDE method are strongly influenced by the subjective setting of a threshold for selecting the most dangerous locations. In addition, there is no opportunity to order clusters as the main drawback of the method (Xie and Yan, 2008). The ordering of clusters would help road administrators with the decision as to where it is most important to apply mitigation measures. In our previous research (Bíl et al., 2013), we introduced the new two-step KDE+ method which objectively determines the significance of clusters by the use of the Monte Carlo method and allows the ordering of the clusters.

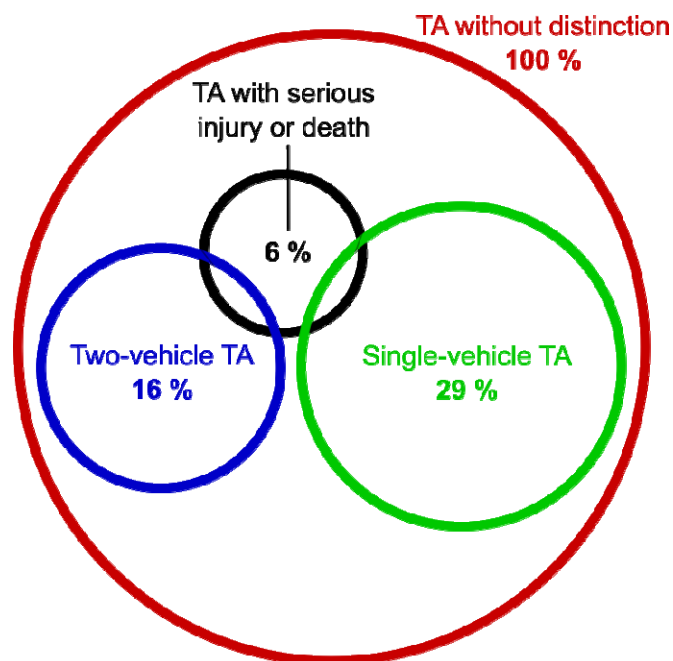
Our aim is to present further developments of the KDE+ method, particularly the applicability to non-precisely located data. In addition, TA on the Czech roads were analysed with the use of the novel KDE+ method. This analysis was performed in four groups of TA: single-vehicle TA, two-vehicles TA, TA with severe injury or death and TA without distinction (all TA).



## 2 DATA

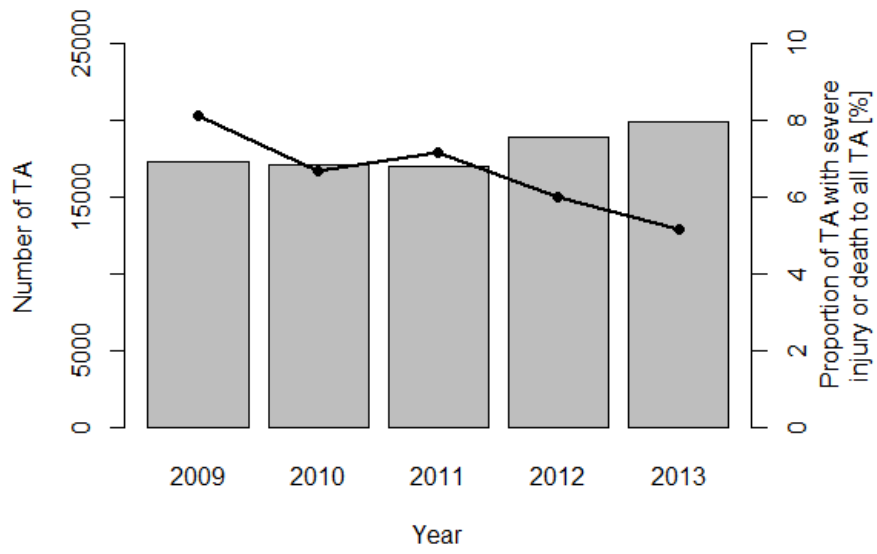
The Czech road network is approximately 37,469 km in length excluding the urban roads. The data on TA come from the Czech Police database. This database consists of 90,418 entries which were recorded over the period 2009 – 2013. We excluded TA which occurred at intersections because they could hide the existence of a dangerous location within a section (Bíl et al., 2013). TA which occurred at intersections 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 – between intersections.

We initially analysed the database for clustering of TA without distinction. Consequently, we performed the analysis in three specific groups of TA: single-vehicle TA, two-vehicles TA and TA with severe injury or death (see Figure 1).



*Figure 1: TA in the Czech Republic.*

TA with severe injury or death are naturally of special concern. Although the number of TA slightly increased from 2011 to 2013, the proportion of TA with severe injury or death in relation to all TA fell from 7.2 % to 5.2 % over this period (see Figure 2).



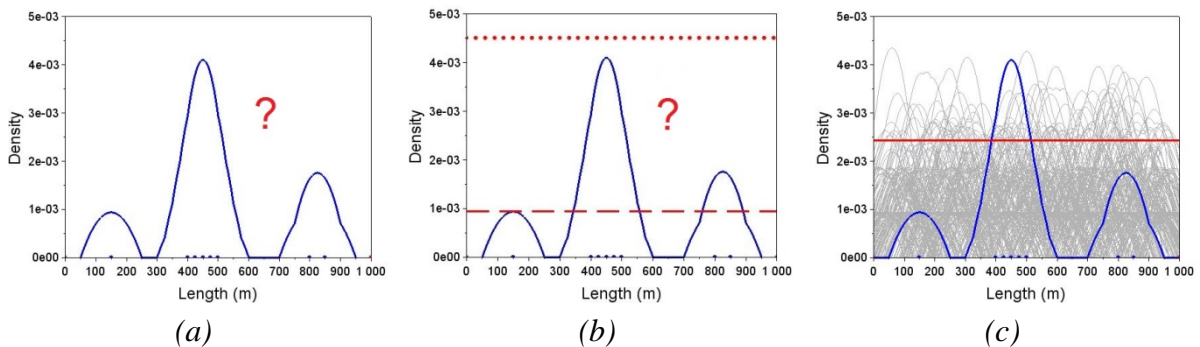
*Figure 2: Number of TA (excluding the urban network and intersections) and the proportion of TA with severe injury or death in relation to all TA over the period 2009 – 2013.*

### 3 THE KDE+ METHOD

The KDE+ technique is based on the KDE method which estimates the probability density function of the underlying data (Figure 3a). However, it is not clear how to set a threshold. Hence, the KDE method can result in a number of clusters located in the neighbourhood of the local maxima of the estimated probability function or significantly dangerous locations can be omitted by setting the threshold excessively high (Figure 3b). Therefore, we improved the KDE approach by adding Monte Carlo simulations to calculate the threshold (Figure 3c).

The resulting significant clusters can be sorted according to the cluster strength (Bíl et al., 2013). The cluster strength is a relative measure of how much the null hypothesis “TA are uniformly distributed along a road section” was violated. The cluster strength depends on the total number of traffic accidents on a particular road and on their mutual position.

The KDE+ method is described in detail in Bíl et al. (2013). Our approach is objective and allows for focusing only on the significant clusters. Furthermore, the KDE+ method is stable and significant clusters can be ordered according to the cluster strength which helps road administrators effectively mitigate dangerous locations.



*Figure 3: KDE with an unknown threshold (a), KDE with two subjectively chosen thresholds (dashed and dotted lines) which significantly influence the results (b) and the KDE+ method (c). The blue line shows the estimated probability density function of the underlying TA. 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 dangerous location is identified.*

#### MODIFICATION OF THE KDE+ METHOD

The Czech Police adds a GPS location to all TA since 2007. This is not, however, the case for all of Europe. When the data on TA are imprecise, the KDE+ method does not perform well

and the significance test is wrong. The most widely used system of TA referencing (apart from GPS localization) is the linear referencing system (LRS) with inaccuracy of 100 m. Therefore, we modified the KDE+ method to also be applicable for this type of data.

The modification of the KDE+ method affects the choice of the kernel function. In the original paper (Bíl et al., 2013), we used the Epanechnikov kernel to the exact GPS positions of the TA. The Epanechnikov kernel (Silverman, 1986) is defined as follows

$$K_d(x) = \begin{cases} \frac{3}{4d} \left(1 - \left(\frac{x}{d}\right)^2\right), & |x| \leq d, \\ 0, & |x| > d, \end{cases}$$

where  $d > 0$  is the bandwidth of the kernel. The new kernel  $\varphi_d(x)$  was derived from the Epanechnikov kernel and reflects the uncertainty of TA.

If there is the GPS position of a traffic accident in the database, it had to belong to an interval determined by the LRS. The GPS position is thus a random variable and has a uniform distribution on the interval  $(x_0 - v; x_0 + v)$ , where  $x_0$  is the location in LRS and  $v > 0$  is the uncertainty of the stationing. Let us denote this random variable as  $\mathbf{X}$ . Obviously, the probability density function of  $\mathbf{X}$  is

$$g(x) = \begin{cases} \frac{1}{2v}, & |x - x_0| \leq v, \\ 0, & |x - x_0| > v. \end{cases}$$

We denoted the exact position of the place which influenced a traffic accident as  $\mathbf{Y}$ . In the original setting, when the GPS coordinates are known, the probability density function of  $\mathbf{Y}$  is  $K_d(y - x)$ , where  $x$  is the GPS position of the traffic accident (Bíl et al., 2013).

Let us think symbolically for a while. If we know the GPS position, we would be able to calculate  $P(\mathbf{Y}) = P(\mathbf{Y}/\mathbf{X} = x)$ . However, we do not know this, thus  $P(\mathbf{Y}) = \sum_x P(\mathbf{Y}/\mathbf{X} = x)P(\mathbf{X} = x)$ . The conditioned variable  $\mathbf{Z} = \mathbf{Y}/(\mathbf{X} = x)$  has the probability distribution function of the form  $f_d(y/x) = K_d(y - x)$ .

In order to perform the derivation properly, we calculated the probability density function of  $\mathbf{Y}$  as follows

$$\begin{aligned}
\varphi_d(y) &= \int_{-\infty}^{+\infty} f_d(y/x)g(x)dx = \\
&= \int_{-\infty}^{+\infty} K_d(y-x)g(x)dx = \\
&= (K_d * g)(y),
\end{aligned}$$

which means that the new kernel is a convolution of the Epanechnikov kernel and the uniform probability distribution function.

In our case, we set  $v = 50$  m because of 100 m inaccuracy in LRS. Regarding the bandwidth, we used  $d = 100$  m for roads and  $d = 150$  m for highways as in our former research (Bíl et al., 2013). As expected,  $\varphi_d(x)$  has wider support than  $K_d(x)$  due to the uncertainty in LRS (see Figure 4).

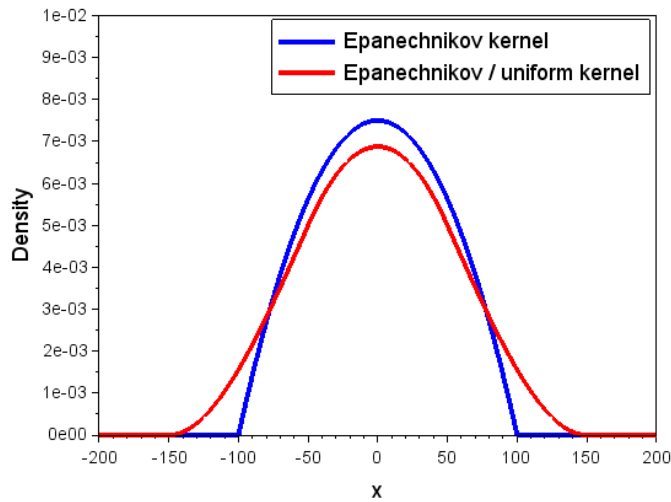


Figure 4: A comparison of the Epanechnikov kernel and the Epanechnikov / uniform kernel ( $d = 100$ ,  $v = 50$ ).

Finally, the kernel density estimation is provided by the formula

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n \varphi_d(x - X_i),$$

where  $X_1, X_2, \dots, X_n$  are the LRS attributes of traffic accidents and  $n$  is the number of them within the particular road section.

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

- $\varphi_d(x)$  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  $\varphi_d(x)$  takes into account all possible outcomes (see Figure 5),
- from the practical point of view,  $K_d(x)$  can result in false clusters (although there are three significant clusters determined by the use of  $K_d(x)$  in Figure 6, there should be only one significant cluster).

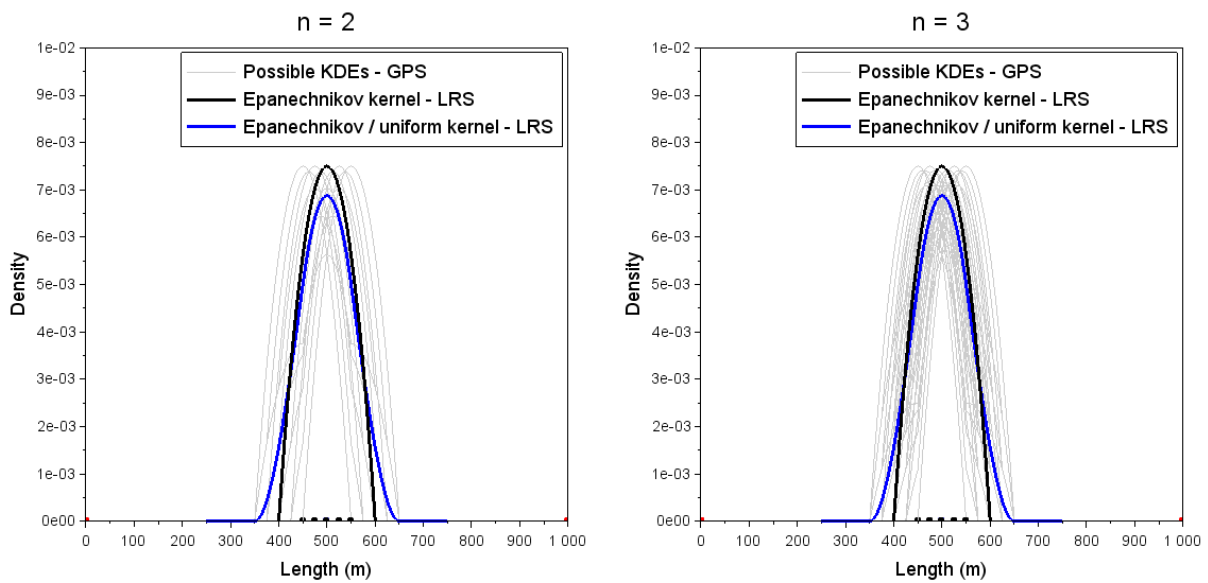


Figure 5: A comparison of possible KDEs of GPS data with classical Epanechnikov kernel and Epanechnikov / uniform kernel applied to LRS data for two TA (left) and three TA (right) which are located in interval  $\{450; 550\}$  within a kilometre-long road section.

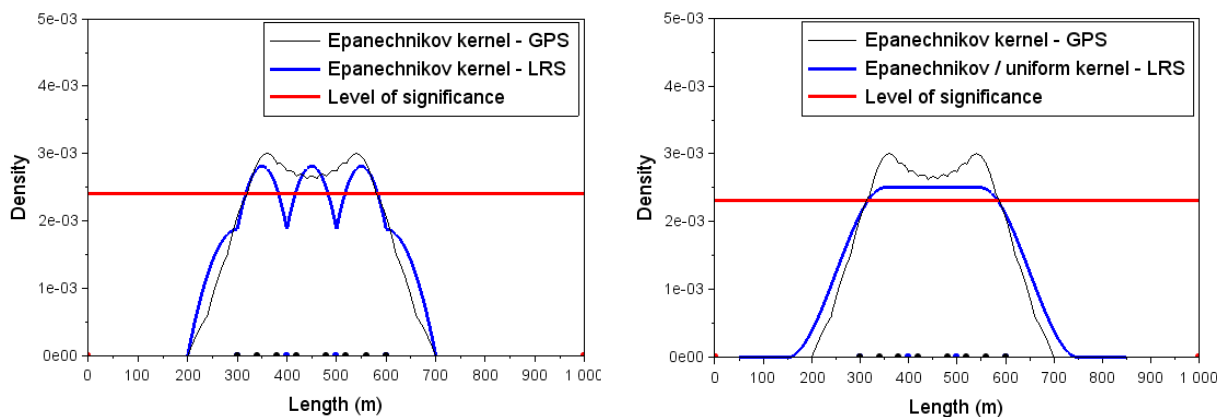


Figure 6: Performance of Epanechnikov kernel (left) and Epanechnikov / uniform kernel (right) applied to LRS data (8 TA).

## STABILITY OF A CLUSTER

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

### Stability related to the database of traffic accidents

Bíl et al. (2013) introduced a simple test for cluster stability. 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. TA was snapped to a wrong road section or the location of TA was recorded incorrectly).

Figure 7 demonstrates the stability of the KDE+ method. It returns almost the same results even if a significant proportion of data is missing. This is important when the lack of data is a common feature. Furthermore, inaccurate data can be excluded from the analysis. The strength of a resulting cluster in Figure 7 is naturally greater in the case of a real dataset, because the clustering is better supported.

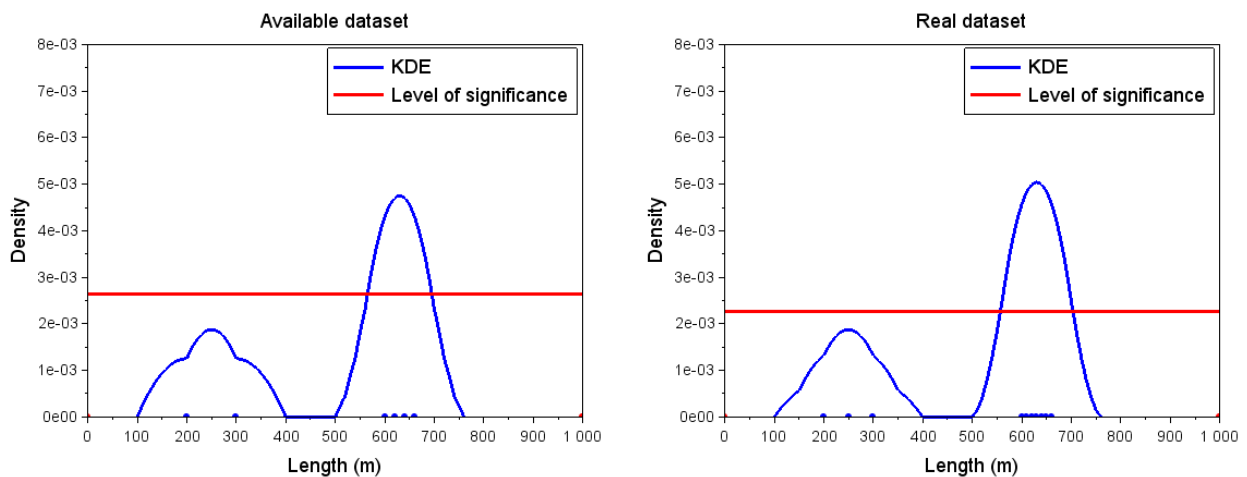


Figure 7: A comparison of resulting KDE for a different number of TA (six on the left and ten on the right).

## THE KDE+ SOFTWARE

A programmed version of the KDE+ method can be downloaded as freeware from the [www.kdeplus.cz](http://www.kdeplus.cz) website. Our KDE+ software is a desktop application. The main window serves for files import and allows for running computation. The important reports are written

in the text box at the bottom of the main window. A graphical representation of a particular road section can also be visualized (Figure 8) by showing the estimated density function and the level of significance.

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 concerning accidents.

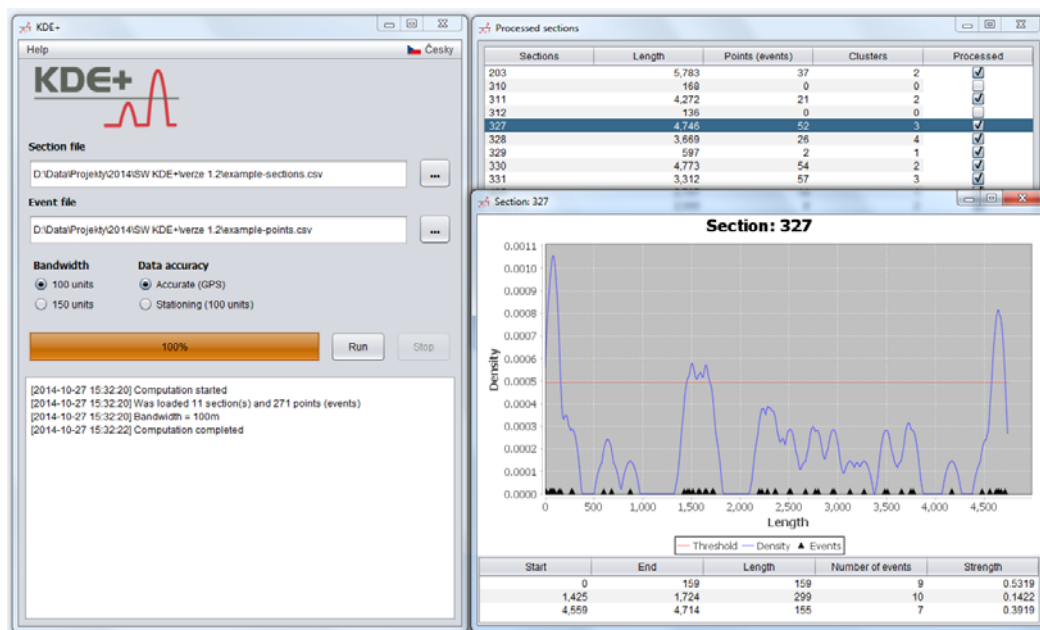
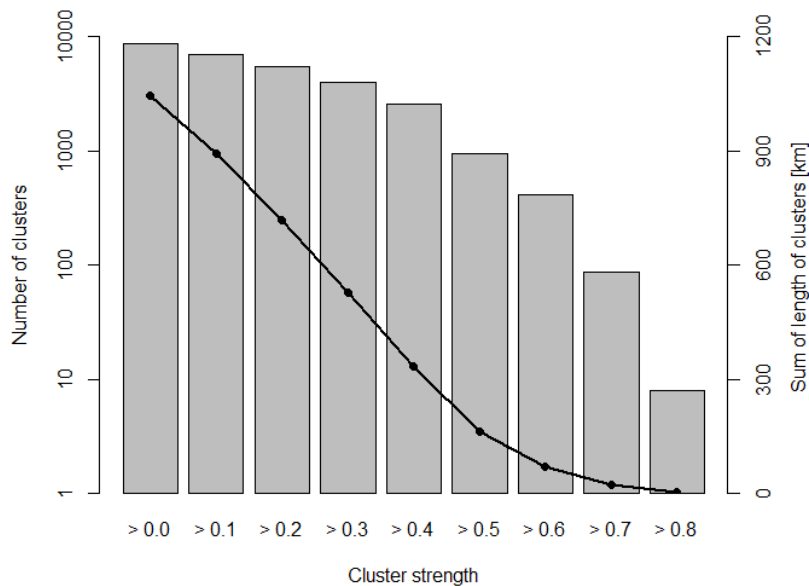


Figure 8: Running the KDE+ software.

## 4 RESULTS

The KDE+ method was applied to the Czech road network. TA without distinction were analysed first. We identified 2.787 % of the entire road network length as dangerous. It consists of 8,739 significant clusters containing 37,585 (41.6 %) TA. This means that more than two fifths of TA form patterns. The most dangerous location was 225 m long and contained 63 TA. Its cluster strength was 0.88.

The KDE+ method enabled us to classify the significant clusters according to their strength. There were 86 clusters with cluster strength greater than 0.7 covering only 22.3 km (0.06 % of the entire road network).



*Figure 9: Number of clusters of TA without distinction (bars) and the total length of clusters (line).*

We used the KDE+ method to examine clustering of specific types of TA, namely single-vehicle TA, two-vehicles TA and TA with severe injury or death. Table 1 shows the outcome of the performed analysis. Clusters of TA with severe injury or death were the shortest on average. This type of TA has, however, the lowest tendency to form patterns (only 15.3 %).

The detailed results, including the attributes of the clusters (e. g. cluster strength and its stability), were visualized in our web-map application [www.kdebourame.cz](http://www.kdebourame.cz). The most dangerous places were depicted on a map (Bíl et al., 2014).

Table 1. The results of a performed clustering analysis with the use of the KDE+ method on the Czech road network. The data on TA were recorded over the period 2009 – 2013.

Group of TA	Without distinction	Single-vehicle	Two-vehicles	With severe injury or death
Number of TA	90,418	59,811	26,512	5,953
Number of clusters	8,739	6,555	2,657	406
Number of TA in clusters [%]	41.6	39.9	31.8	15.3
Total length of clusters [km]	1,044	740	268	29
Total length of clusters [%]	2.79	1.98	0.71	0.08
Mean length of clusters [m]	120	113	101	70

## 5 DISCUSSIONS AND CONCLUSIONS

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

The presented results allowed the road administrators to effectively localize the most dangerous places within the road network. In addition, if road administrators are interested in determining the worst places within the road network, they only need to inspect a short part of the network.

We had the GPS locations of all traffic accidents 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.

A comparison of the KDE+ method with other methods for the identification of dangerous locations was published in Bíl et al. (2013). 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.

The option of the use of the KDE+ method to LRS data was implemented in the KDE+ software. Thus, there are two options in the data accuracy setting: GPS and LRS with 100 m precision. The KDE+ software can be used by any user with an interest in identifying the most dangerous locations of TA. Mitigation measures can be applied to clusters with the highest strength.

## REFERENCES

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

Bíl, M., R. Andrášik and J. Sedoník (2014). 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.

Chung, K., K. Jang, S. Madanat and S. Washington (2011). Proactive detection of high collision concentration locations on highways. *Transport. Res. A-Pol.*, **45**, 927 – 934.

Erdogan, S., I. Yilmaz, T. Baybura and M. Gullu (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Anal. Prev.*, **40**, 174 – 181.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety*. Pergamon Press, Oxford.

Lord, D. and F. Mannering (2010). The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transport. Res. A-Pol.*, **44**(5), 291 – 305.

Okabe, A. and I. Yamada (2001). The K-function method on a network and its computational implementation. *Geogr. Anal.*, **33**(3), 152 – 175.

Plug, C., J. Xia and C. Caulfield (2011). Spatial and temporal visualization techniques for crash analysis. *Accident Anal. Prev.*, **43**, 1937 – 1946.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Xei, Z. and J. Yan (2008). Kernel Density Estimation of traffic accidents in a network space. *Comput. Environ. Urban*, **32**, 396 – 406.

Yamada, I. and J. C. THILL (2004). Comparison of planar and network K-functions in traffic accident analysis. *J. Transp. Geogr.*, **12**, 149 – 158.