PSO based Optimization of DBSCAN Algorithm Parameters for Road Accident Blackspot Localization

Sándor Szénási*, Miklós Sipos† and Péter Mogyorósi‡
John von Neumann Faculty of Informatics, Óbuda University
Budapest, Hungary
Email: *szenasi.sandor@nik.uni-obuda.hu, †sipos.miklos@nik.uni-obuda.hu

***Abstract*—The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a well-known data-mining method capable of localization of accident blackspots of the road network based on the already existing road accident database records. However, its parameterization raises many problems, as its operation is significantly different from the traditional Sliding Window (SW) method. This paper presents a Particle Swarm Optimization (PSO) based method to find a base parameter set for the DBSCAN method which gives similar results to the already existing SW. The fitness function of the PSO algorithm is based on the similarity of accident blackspots, which needs a definition of a novel metric. The evaluation results show that the DBSCAN method used with the recommended parameter set is capable to give similar results to the SW method used by road safety experts.**

***Index Terms*—DBSCAN, Particle Swarm Optimization, Accident Blackspot, Sliding Window**

I. Introduction

The basic objective of road safety management is the protection of all participants and decreasing the property damage caused by accidents [1]. There are several (mostly financial) limitations of preventive actions; therefore, it is essential to find the appropriate locations and operations to maximize the expected benefits. For this reason, it is important to have appropriate methods for finding accident hotspots.

These accident hotspots (also known as blackspots) are some hazardous locations of the public road network where the number of accidents is higher than expected (Fig. 1). blackspot management consists of three consecutive steps: identification, analysis, and treatment of these areas. This paper focuses only on the first step, which is the most crucial part of the process. In this phase, the input is given by several databases (road network, weather, historical accident data), and the expected output is a list of potential blackspots (these can be considered as blackspot candidates until further analysis can prove that these are real hazardous areas).

However, the scientific work in this area has a long tradition; interestingly, there is no generally accepted definition of road accident blackspots. There are several variations and the official definition used by engineers also varies by country. This paper will use the definition of the Hungarian government: outside built-up areas, blackspots are road sections no longer than 1000 meters where the number of accidents during the last 3 years is more than 3.

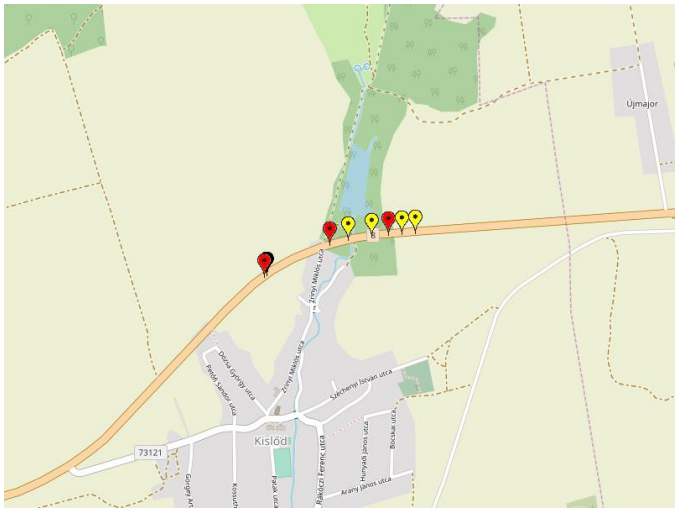

Fig. 1. blackspot example (*image generated by Google Maps*.)

There are various algorithms for blackspot localization, usually according to the specific definitions [2]–[6]. In the Hungarian case, the traditional sliding window method fulfills the requirements. However, this is a somewhat outdated algorithm based on road numbers and section numbers. There are more accurate data-mining-based techniques using the GPS coordinates of accidents like the DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise). But the output of these planar methods significantly differs from the results of the sliding window algorithm, and these require different parameters. This raises several issues for road safety engineers and regulators to adopt these new methods to the already existing road safety methodology.

This paper presents a heuristics-based method to find the appropriate parameter set of the DBSCAN algorithm than can be used to achieve similar results that the output of the sliding window method. The rest of the paper is structured as follows: the next section presents the already existing results in the field of blackspot management and parameter optimization.

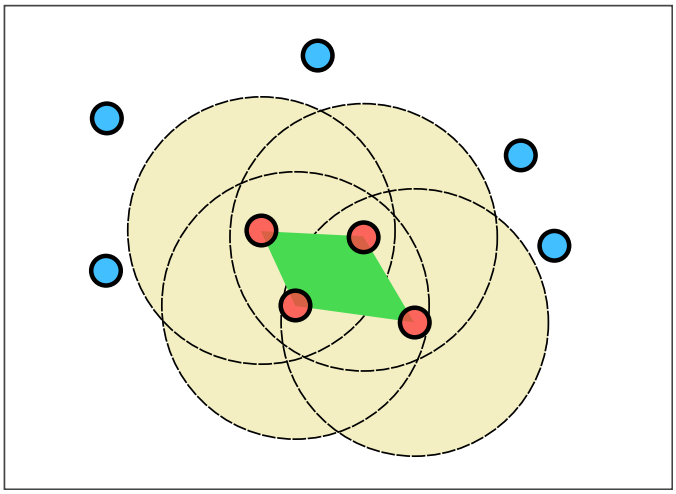

Fig. 2. Example for the DBSCAN algorithm

The third section presents the used methodology and the next contains the evaluation of the results. Finally, the last section contains the conclusions and limitations.

II. Related Work

In the field of accident hot spot identification, one of the most traditional methods is the well-known sliding window technique [7], [8]. It has two input parameters: section length (l) and a threshold value for the minimum number of accidents (a_{min}). The main steps of the algorithm are a) select one road from the network and split it into equally sized sectors (using parameter l) b) select the accidents that occurred in the given section c) if the number of these is higher than the threshold (a_{min}), mark the section as a blackspot candidate.

As a one-dimensional technique, it has several limitations. It is not possible to use the GPS-based accident coordinates directly it is necessary to map these to the road network. Furthermore, since it depends on the accidents on one specific road, it is not possible to find the blackspots at intersections. That was the reason for the emergence of spatial data analysis techniques like the KDE (Kernel Density Estimation) method which gives the accident density estimation at a given reference point [7], [9], [10]. This value is based on the search radius distance, and it is also possible to define several kernel functions.

As an alternative, it is also possible to use the DBSCAN data-mining algorithm for blackspot identification. As a density-based clustering method, its objective is to group items where the elements of the same group are similar to each other; meanwhile, the elements of different groups are not. In the field of road safety engineering, the elements are the accidents in the public road network identified by planar GPS coordinates. Clusters are the potential blackspots (accidents similar to each other) and the outliers are the random accidents not belonging to any hot spot. As a small modification, it is also possible to set a lower limit to the density of the cluster to decrease the number of false-positive results.

The objective of the research work is to optimize the parameters of the DBSCAN algorithm [11]. As there are two floating-point variables (distance, density limit) and one integer (minimum accident count) it is hard to manually find the appropriate values. There are several heuristics applicable for this task, like Hill Climbing, Genetic Algorithm [12], or Particle Swarm Optimization [13].

III. Methodology

A. Dataset

The official road accident database of Hungary was used in the experiments. This dataset contains all accidents with personal injury collected by the police and handled by the Central Statics Department of Hungary. The completeness of the data is ensured by legislation because the participants of any road accident with personal injury are obliged to report it to the police.

The official blackspot definition is based on a 3-year long interval therefore accidents from the years 2017 to 2019 were used. The sliding window method can only examine data for one single road; therefore, an additional filter was set to use only the accidents that occurred on road number 1.

B. Clustering methods

The parameters of the sliding window methods are according to the official blackspot definition: the sliding window length (l) is 1000m, and the minimum number of accidents (a_{min}) is 4.

The objective of the optimization is to find the optimal parameters of the DBSCAN method. These parameters are:

- ϵ – growing distance (float value between 1m and 1000m);
- a_{min} - minimum number of accidents (integer between 2 accident and 10 accident);
- λ - limit for accident density (float value between 0 accident/m^2 and 1 accident/m^2).

The density of a cluster is calculated by the number of accidents divided by the area of the given cluster. Where the area of the cluster is calculated by the Gauss' area formula:

$$A = \frac{\left|\sum_{i=1}^{n-1} x_i y_{i+1} + x_n y_1 - \sum_{i=1}^{n-1} x_{i+1} y_i - x_1 y_n\right|}{2} \quad (1)$$

Where

- A - the area of the cluster;
- n - number of accidents in the cluster;
- (x_i, y_i) - two-dimensional planar coordinates of the i-th accident of the cluster (where $i \in \{1, 2, ..., n\}$).

C. Fitness function

The result of the fitness function is the similarity of the results given by the sliding window and the DBSCAN algorithm. Both methods result in a list of blackspot candidates; therefore, further calculations are required to determine the similarity.

For this purpose, the following definitions have been defined:

- The similarity between two blackspots (BS_1 and BS_2) is
$$s(BS_1, BS_2) = \frac{|BS_1 \cap BS_2|}{|BS_1 \cup BS_2|}. \quad (2)$$
Where $BS_1 \cap BS_2$ is the intersection of the blackspots; therefore, $|BS_1 \cap BS_2|$ shows the number of accidents found at both blackspots. $|BS_1 \cup BS_2|$ is the number of accidents in both blackspots without duplications.
As visible, this value is 1.0 if the two blackspots are the same (both contain the same accidents). It is 0.0 if there is not any intersection between the two clusters. The value is between 0.0 and 1.0 if there is some partial equation between the two blackspots.
- The similarity between two blackspot list is calculated by the following steps:
  1) Two lists (L_1 and L_2) are created based on the blackspots given by the two methods.
  2) Let (BS_1, BS_2) is a pair of blackspots from L_1 and L_2 where the similarity of BS_1 and BS_2 is maximal over any other pairings, based on (2)
  3) Remove BS_1 from L_1 and BS_2 from L_2. And increase an S value with the similarity score of BS_1 and BS_2.
  4) Repeat steps 2-3. until there are no possible pairings
  5) The similarity of two blackspot lists is calculated by
$$f(L_1, L_2) = \frac{S}{|L_1| + |L_2|} \quad (3)$$
Where $|L_1|$ and $|L_2|$ are the initial sizes of L_1 and L_2 lists. f is the similarity between the two blackspot lists, which is the fitness function of the heuristics.

D. Optimization method

The Particle Swarm Optimization method was used to find the appropriate parameter set. The main parameters of the experiment are the following:

- Population size: 10
- Number of iterations: 200
- C_1 parameter: 1.49445
- C_2 parameter: 1.49445
- w parameter: 0.729

The velocity at time $t+1$ of a given P participant is calculated by the following formula:

$$P_{t+1}^{velo} = w*P_t^{velo} + C_1*r_1*(P_t^{opt} - P_t^{pos}) + C_2*r_2*(B_t + P_t^{pos}). \quad (4)$$

Where

- P_t^{velo} - the velocity of the given participant at time t
- P_t^{pos} - the position of the given participant at time t
- P_t^{otp} - the local optimal position of the given participant at time t
- B_t - the global optimal position of the swarm at time t
- w - Inertia weight
- C_1 - Velocity coefficient affected by personal best
- C_2 - Velocity coefficient affected by global best
- r_1, r_2 - random variables between 0..1

The PSO method was started with a random initial population of 10 elements. The initial velocity was 0 for all particles. After that, the algorithm calculates the velocity and the new position of all elements of the population. This process is repeated 300 times, continuously logging the parameters and fitness values of all elements and the best/average fitness of the population by iteration.

IV. Evaluation

As the first step, the sliding window method was executed on the dataset. The result of this procedure is a list of 31 potential blackspots.

The objective of the next step is to find the parameter set able to give similar results using the DBSCAN method. The PSO algorithm was used to find this parameter set.

Fig. 3 shows the fitness values by the iterations of the PSO method.

As visible, the optimization was successful, the final fitness was significantly better than the initial values. The best fitness value was reached at a relatively early stage at iteration 19.

The element with the best fitness value represents the following parameter set:

- $\epsilon = 343.2503m$
- $\alpha = 4$
- $\lambda = 0.0002$ accident/m^2

In the case of the Sliding Window method, the 1000m distance is the length of the window, the ideal ϵ value in the case of DBSCAN is significantly less (343.250m). That makes sense, because in the case of DBSCAN, this is a growing distance, not an absolut length. As also expected, the α value is the same as the $a_m in$ value.

Detailed comparison results are shown in Table I. As visible, two of the original blackspots have no pairs in the DBSCAN results. In the opposite direction, two of the DBSCAN blackspot candidates are not paired to any Sliding Window spots. The remaining pairs show significant similarities. Most of them are almost identical and there are some pairs with minor overlapping accidents.

V. Conclusions

The objective of the paper was to determine a parameter set for the DBSCAN algorithm to give similar results to the traditional Sliding Window method. A novel blackspot pairing and evaluation method were developed to assign a fitness function to the potential parameter sets. A Particle Swarm Optimization was used to iteratively approximate the optimal value. The results of the experiment were evaluated, and these are acceptable.

As a limitation, it is worth noting that the given parameter set is not sure the best parameter set for the DBSCAN method. It is just as similar to the Sliding Window as possible, to help the work of road safety engineers during the transition period of adopting the new method. There may be better parameter

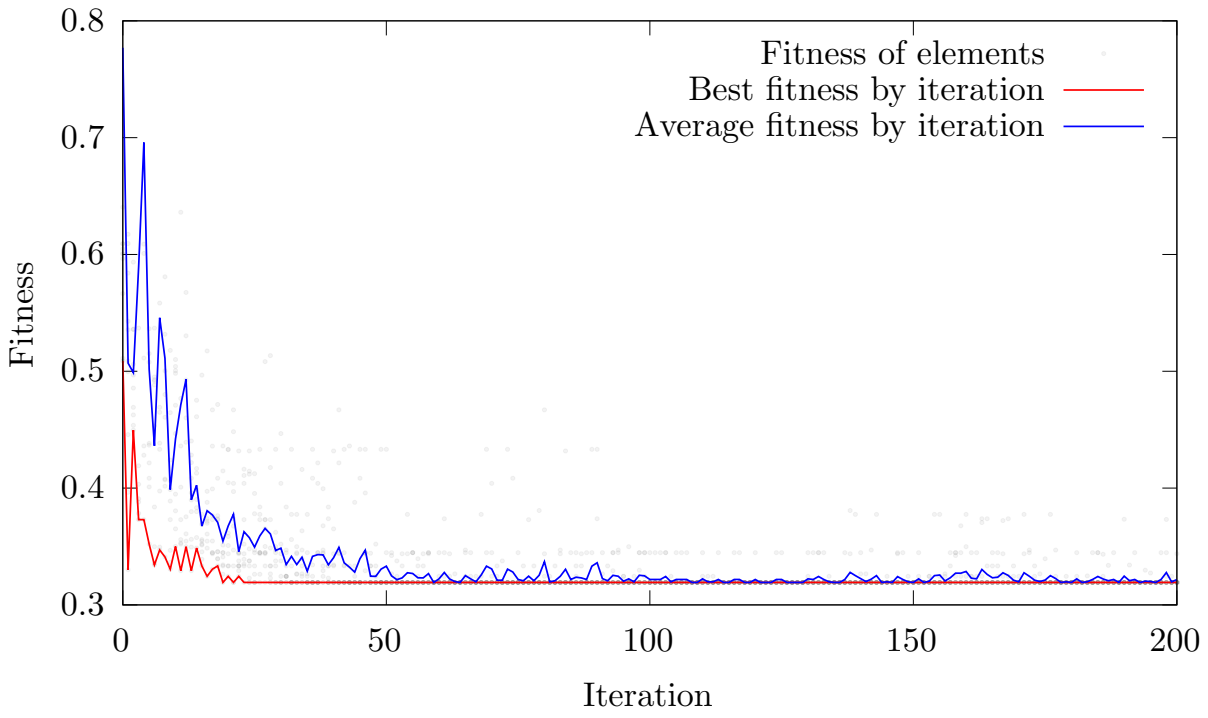

Fig. 3. Results of the PSO algorithm

sets that can make better use of the advantages of the planar method.

ACKNOWLEDGMENT

The authors would like to thank the GPGPU Programming Research Group of Óbuda University for its valuable support. The authors would like to thank NVIDIA Corporation for providing graphics hardware for the experiments. This work was supported by the Hungarian National Research Development and Innovation Office PIACI KFI grant (2020-1.1.2-PIACI-KFI-2020-00003).

REFERENCES

[1] G. Kertesz and I. Felde, "One-Shot Re-identification using Image Projections in Deep Triplet Convolutional Network," in *SOSE 2020 - IEEE 15th International Conference of System of Systems Engineering, Proceedings*. Institute of Electrical and Electronics Engineers Inc., jun 2020, pp. 597–601.

[2] R. Elvik, "A survey of operational definitions of hazardous road locations in some European countries," *Accident Analysis & Prevention*, vol. 40, no. 6, pp. 1830–1835, 2008.

[3] R. Delorme and S. Lassarre, "A new theory of complexity for safety research. The case of the long-lasting gap in road safety outcomes between France and Great Britain," *Safety Science*, vol. 70, no. 0, pp. 488–503, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753514001593

[4] W. Murray, J. White, and S. Ison, "Work-related road safety: A case study of Roche Australia," *Safety Science*, vol. 50, no. 1, pp. 129–137, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925753511001597

[5] A. Montella, D. Andreassen, A. P. Tarko, S. Turner, F. Mauriello, L. L. Imbriani, and M. A. Romero, "Crash Databases in Australasia, the European Union, and the United States," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2386, no. 1, pp. 128–136, 2013.

[6] P. Hegyi, A. Borsos, and C. Koren, "Searching possible accident black spot locations with accident analysis and GIS software based on GPS coordinates," *Pollack Periodica*, vol. 12, no. 3, pp. 129–140, 2017. [Online]. Available: http://www.akademiai.com/doi/abs/10.1556/606.2017.12.3.12

[7] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[8] S. Szénási and D. Jankó, "Internet-based decision-support system in the field of traffic safety on public road networks," in *6th European Transport Conference*, Budapest, 2007, pp. 131–136.

[9] M. Bíl, R. Andrášik, and Z. Janoška, "Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation," *Accident Analysis & Prevention*, vol. 55, no. 0, pp. 265–273, 2013.

[10] B. Flahaut, M. Mouchart, E. S. Martin, and I. Thomas, "The local spatial autocorrelation and the kernel method for identifying black zones," *Accident Analysis & Prevention*, vol. 35, no. 6, pp. 991–1004, 2003.

[11] P. Rosenberger and J. Tick, "Multivariate optimization of pmbok, version 6 project process relevance," *Acta Polytechnica Hungarica*, vol. 18, no. 11, 2021.

[12] I. Lovas, "Fixed point, iteration-based, adaptive controller tuning, using a genetic algorithm," *Acta Polytechnica Hungarica*, vol. 19, no. 2, pp. 59–77, 2022.

[13] I. Felde, "Simplified computation of the heat transfer co-efficient in quenching," *Acta Polytechnica Hungarica*, vol. 18, no. 10, 2021.

TABLE I
COMPARISON OF THE RESULTS OF THE SW AND DBSCAN METHODS.

Sliding Window		DBSCAN		Similarity
Blackspot ID	Accident count	Blackspot ID	Accident count	
8	6	18	6	1.00
23	5	6	5	1.00
25	11	22	11	1.00
26	9	25	9	1.00
27	8	20	8	1.00
16	14	16	15	0.97
13	10	28	11	0.95
9	8	0	7	0.93
3	6	24	7	0.92
14	7	3	6	0.92
12	6	19	5	0.91
10	9	27	11	0.90
5	5	10	4	0.89
6	10	7	8	0.89
15	5	8	4	0.89
17	5	1	4	0.89
4	8	26	9	0.82
30	6	14	7	0.77
22	6	12	5	0.73
20	12	17	25	0.65
24	9	2	4	0.62
0	8	23	16	0.58
21	8	13	6	0.57
18	5	21	5	0.40
7	5	5	4	0.22
11	6	-	0	0.00
28	6	-	0	0.00
-	0	9	4	0.00
-	0	15	4	0.00