Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 187–194

Procedia Computer Science

www.elsevier.com/locate/procedia

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6 - 9, 2020, Warsaw, Poland

On the use of clustering analysis for identification of unsafe places in an urban traffic network

Johan Holmgren [a,*], Luk Knapen[b,c], Viktor Olsson[a], Alexander Persson Masud[a]

[a] *Internet of Things and People Research Center & Dept. of Computer Science and Media Technology, Malmö University, Malmö 205 06, Sweden*
[b] *Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium*
[c] *VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*

Abstract

As an alternative to the car, the bicycle is considered important for obtaining more sustainable urban transport. The bicycle has many positive effects; however, bicyclists are more vulnerable than users of other transport modes, and the number of bicycle related injuries and fatalities are too high. We present a clustering analysis aiming to support the identification of the locations of bicyclists' perceived unsafety in an urban traffic network, so-called bicycle impediments. In particular, we used an iterative k-means clustering approach, which is a contribution of the current paper, and DBSCAN. In contrast to standard k-means clustering, our iterative k-means clustering approach enables to remove outliers from the data set. In our study, we used data collected by bicyclists travelling in the city of Lund, Sweden, where each data point defines a location and time of a bicyclist's perceived unsafety. The results of our study show that 1) clustering is a useful approach in order to support the identification of perceived unsafe locations for bicyclists in an urban traffic network and 2) it might be beneficial to combine different types of clustering to support the identification process.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.

Keywords: Cluster analysis; k-means; iterative k-means; DBSCAN; Click-point data; bicycle impediment

1. Introduction

The bicycle is in general considered to be a sustainable, fast, cost efficient, and healthy alternative to the car in urban environments. The list of positive effects of the bicycle can be made long; however, there are also negative aspects related to bicycling. The perhaps most important negative aspect is that bicyclists are unprotected, hence more vulnerable than users of other transport modes, in particular car and public transport. However, it has been shown that the positive health effects of bicycling significantly overshadow the risk of getting injured or dying from an accident.

* Corresponding author. Tel.: +46-40-665 76 88
E-mail address: johan.holmgren@mau.se

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2020.03.024

See, for example, Andersen et al. [1], who show that bicyclists are expected to live longer even though they are exposed to an increased risk of being injured or dying in an accident.

Still, there are too many accidents involving bicyclists, which is shown in available accident statistics. For example, approximately 2000 bicyclists died in road accidents in the European Union countries (EU) during 2016 [5]. The total number of road fatalities in EU has significantly decreased from approximately 43.000 to approximately 26,000 from 2007 to 2016. The number of bicycle fatalities dropped from approximately 2500 to approximately 2000 from 2007 to 2010; however, since 2010, the number of bicycle fatalities has remained on the same level. It is, therefore, obvious that there is a need for improvements to provide a safer traffic environment for bicyclists, hence contributing to make the bicycle more attractive. This is important as the bicycle play an important role in the strive towards a sustainable society, where urban transport to a larger extent than today is carried out by green transport modes.

We use the term *bicycle impediment* to refer to a location in an urban traffic network where bicyclists have a tendency to feel unsafe. By identifying impediments, it is possible for the responsible authorities to focus their safety improving investments on the locations that are perceived unsafe by bicyclists. However, it is not always straightforward to identify the impediments in an urban transport network. Historic accident statistics is an important source to support the identification of the impediments. However, we argue that the accident statistics does not give a complete view on the locations of impediments as it does not take into consideration the bicyclist's own perception of unsafety.

We present a clustering analysis, aiming to identify bicycle impediments in the city of Lund, Sweden. In particular, we used an iterative k-means clustering approach, which is a contribution of the current paper, and DBSCAN. In contrast to standard k-means clustering, the iterative k-means clustering approach enables to remove outliers from the data set. In our study, we used a set of *click-point* data, which was collected by bicyclists, who were instructed to push a button mounted on their bicycle handlebars, each time they experienced unsafety in the traffic situation.

The identification of accident locations, often referred to as accident hotspots, is widely discussed in the literature, see, for example, Cheng & Washington [3] and Montella [8] for overviews. Several studies use clustering in order to identify accident hotspots, for example, Anderson [2] use k-means clustering and kernel density estimation in order to identify accident hotspots and typical groups of road users involved in accidents, and Xu & Tao [12] use ensemble clustering in order to identify accident hotspots. The existing studies differ from our work in that they mainly focus on historic data on accidents, whereas we focus on perceived unsafety of bicyclists. One exception is the study of Persson Masud & Olsson [9], which makes use of k-means clustering for the same data set as we used in the current study. The focus of their work is to compare different approaches for controlling the size of the generated clusters.

By supporting the identification of bicycle impediments, our study aims to contribute to improved traffic safety for bicyclists in urban environments. In particular, the result of the study is intended to help the municipality of Lund to identify and take action on bicycle impediments. In turn, this may contribute to increased popularity of the bicycle.

2. Click-point data

In our study, we used a, so-called, click-point data set, which was collected by 78 bicyclists traveling in Lund, Sweden, in the autumn of 2018. The bicyclists' where instructed to click a button, mounted on their handlebars, when they felt unsafe in the traffic situation. Each of the clicks, i.e., a data point, contains the unique button identifier, the GPS coordinate (latitude and longitude), the received GPS accuracy, and a time stamp of the click. The buttons were connected via Bluetooth to the bicyclist's mobile phone, which in turn communicated each button click to a server.

As the data set contained several repeated (most likely unintended) clicks, duplicates, and geographic outliers, that is, clicks with GPS coordinates outside Lund, we filtered the data set prior to conducting our cluster analysis. Repeated clicks were those that were provided so fast that we considered it to be very unlikely that they could be provided by a bicyclist pushing the button several times. Even if the clicks that we considered to be repeated would be intended, a sequence of repeated clicks most likely refers to the same impediment, still making it safe to filter them out. Duplicates are sequences of data points provided by the same button, but with the same latitude and longitude.

In the filtering process we conducted the following sequential steps:

1. **Remove repeated clicks.** For each individual, we considered a click point with timestamp t to be an intended click point) if and only if it has no predecessor in the period $t - \Delta t$, where $\Delta t = 1[s]$. This reduced our data set from 3101 (the collected number of data points) to 2142 data points.

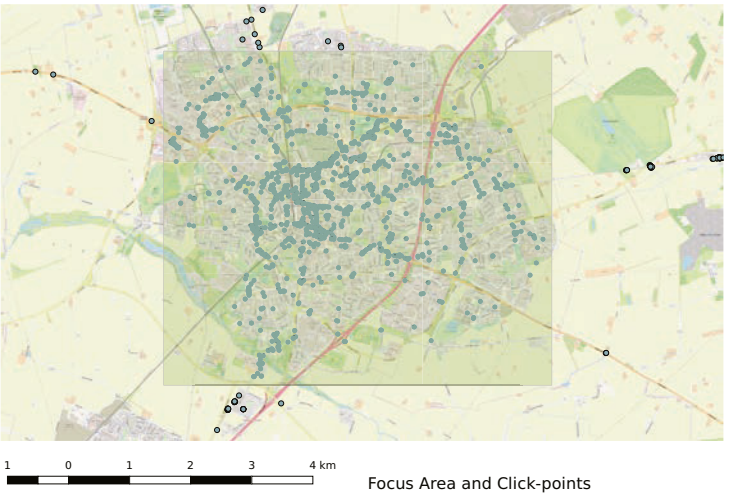

Fig. 1. The click-point data set and focus area covering the city of Lund.

2. **Remove duplicate clicks.** This further reduced our data set from 2142 to 1914 data points.
3. **Remove clicks outside focus area.** We filtered all data points outside the area defined by the longitude interval [55.68, 55.729] and the latitude interval [13.153, 13.254]. This reduced our data set from 1914 to 1774 points.
4. **Remove inaccurate clicks.** We filtered all data points with GPS accuracy larger than 50 meters. This further reduced our data set from 1774 to 1622 data points.

3. Cluster analysis

Cluster analysis is a type of unsupervised learning, where data points are grouped into clusters such that the data points assigned to the same cluster are more similar to each other than to the data points in other clusters. In our cluster analysis focusing on bicycle impediments identification, we used a derivative of k-means clustering [7] and DBSCAN [4]. We implemented our cluster analysis using Python 2.7.17 and the machine learning package scikit-learn 20.3.

It is well known that k-means clustering is mainly suitable for identification of convex clusters, which in our application corresponds to "round" clusters representing impediments that appear, for example, in intersections. On the other hand, DBSCAN supports the identification of elongated clusters, which, correspond to one-dimensional stretched impediments that may appear along roads. As there is no prior evidence about the user's perception; in practice both of the two cases may co-exist, we considered both k-means clustering and DBSCAN in our study.

We also faced the problem that k-means clustering has no "noise concept" similar to, for example, DBSCAN. We therefore developed an iterative k-means approach, which is described in detail below, where outliers are iteratively removed from the data set. In our application, the Euclidian distance between a cluster member and the centroid is limited by an upper bound D, meaning that the distance between two members in a cluster is at most $2D$. The value of D can be determined by reasoning at the application level (maximum distance between a click point and the unknown position of the impediment) or by investigating the clusters (number and size) for different plausible values of D. The value of D is related to the thresholds used in [10, 11].

3.1. K-means clustering

In k-means clustering [7], the number of clusters that should be produced is required as input, and the algorithm operates by iteratively assigning points to clusters represented by cluster centroids, which are updated in each iteration. A cluster centroid is calculated by taking the average in each dimension of all data points included in the cluster. The algorithm is initiated by assigning k randomly chosen points in the data set as centroids, and it iterates between two steps until the clusters have stabilized:

1. Assign each data point to the closest centroid.
2. Calculate new positions of the centroid of each cluster by taking the mean value of all data points assigned to the cluster.

As mentioned above, k-means clustering requires that the number of clusters (k) is provided as input; however, it is typically not straightforward to identify a suitable value of k. Different approaches have been proposed for identifying an approximate optimal value of k, including *rule-of-thumb* and the *elbow method* [6]. In the elbow method, a decreasing cost function $c(k)$ is analyzed for $k = 1, \ldots, n$. The $c(k)$ can, for example, be defined as the sum or average distance from all points in the data set to its centroid. When increasing the value of k, the cost function will typically drop faster for lower values of k and slower for larger values of k. For some applications, the decrease rate of the cost function will, for some value of k, distinctly slow down, and this point (i.e., a k value) is referred to as the elbow point.

For our click-point data set, we illustrate in Fig. 2 the application of the elbow method, with the average distance from any point to its centroid as cost function. By visual inspection, the elbow appears to be in the interval $[35, 40]$.

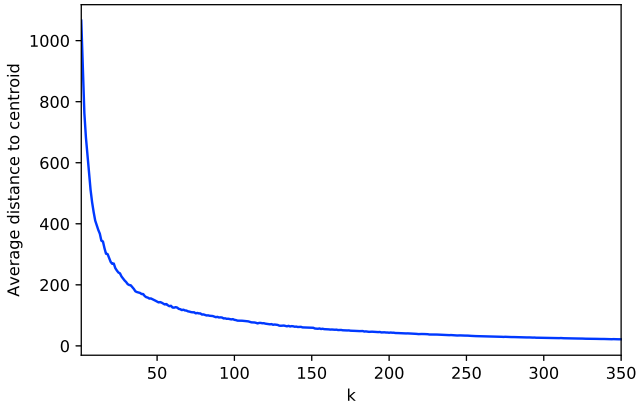

Fig. 2. The average distance from any point to its centroid for different values of k. The elbow appears to be in the interval $[35, 40]$.

However, a closer look at the clusters produced using $k = 40$ reveals that the types of clusters produced is not suitable for identification of bicycle impediments, as the clusters vary significantly in terms of their spatial extension. This is illustrated in Fig. 3, where we plot the average distance from any point to its centroid for each cluster (to the left) and the largest distance from any point to its centroid for each cluster (to the right). In general, the clusters are spatially too big for k:s around 40, and to obtain clusters that are small enough to be connected to bicycle impediments, a significantly higher k-value is required. However, this would instead give us many small clusters with outliers, and some clusters that are still larger than a typical bicycle impediment. For $k = 200$, we illustrate this in Fig. 4.

From Fig. 3 and Fig. 4, we conclude that it is mainly the outliers that prevents us from finding an appropriate value of k. Indeed, k-means clustering does not identify outliers, and it does not allow setting a limit on the minimum number of points to include in a cluster, or a maximum spatial cluster size. These aspects are important in our application, where a minimum number of clicks within a rather small area is required in order to define a bicycle impediment.

Due to the abovementioned limitations, we defined the *iterative k-means clustering with outlier detection* approach, which we used in order to 1) identify an appropriate value of k and 2) remove outliers from our data set. The approach, which is specified below (in **Algorithm 1**), makes use of iterative updates of k and removal of outliers, and it requires the following input parameters:

- *min_samples* - The minimum number of points to allow in any cluster.
- *max_dist_to_centroid* - The maximum distance a point is allowed to be from the centroid of its cluster.
- *outlier_threshold* - A point is considered to be an outlier if and only if the distance to its centroid is larger than *outlier_threshold* times the median of the distance to the centroid for all points in the same cluster.
- $k^{init} \in \mathbb{Z}^+$ - The initial k-value used in the approach.

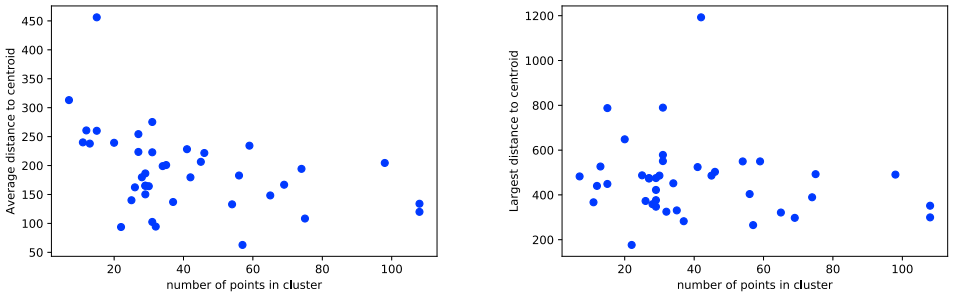

Fig. 3. For each k-means cluster, where $k = 40$, the average distance from any point to its centroid (to the left) and the largest distance from any point to its centroid (to the right).

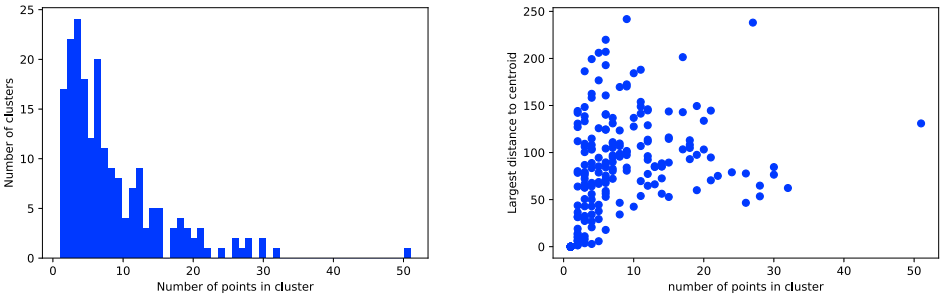

Fig. 4. For each k-means cluster, where $k = 200$, the number of clusters with $1, \ldots, 51$ number of points (to the left) and the largest distance from any point to its centroid (to the right).

Furthermore, we let P denote the input set of click points.

Algorithm 1 Iterative k-means clustering with outlier detection.

Step 0: Set $P^{cur} = P$ and $k^{cur} = k^{init}$.
Step 1: Generate k-means clustering C for $k = k^{cur}$ and data point set P^{cur}.
Step 2: Identify all clusters $C' \subseteq C$ with less than *min_samples* data points. If $|C'| = 0$: Go to Step 3. Otherwise: Go to Step 4.
Step 3: For all clusters $c' \in C'$, remove the point p' with longest distance to the cluster centroid, i.e., set $P^{cur} = P^{cur} \setminus \{p'\}$. Set $k^{cur} = \max(1, k^{cur} - 1)$ and go to Step 1.
Step 4: Identify the point $p' \in P^{cur}$, with distance $d(p')$ to its cluster centroid, with largest value $q = \frac{d(p')}{median_{p \in C'} d(p)}$. If $q >$ *outlier_threshold*: Go to Step 5. Otherwise: Go to Step 6.
Step 5: Remove p', i.e., set $P^{cur} = P^{cur} \setminus p'$. Set $k^{cur} = \max(1, k^{cur} - 1)$ and go to Step 1.
Step 6: If $\max_{P^{cur}} d(p) >$ *max_dist_to_centroid*: Set $k^{cur} = k^{cur} + 1$ and go to Step 1. Otherwise: Terminate with approximate optimal $k^* = k^{cur}$, filtered data point set P^{cur}, and clustering C.

We applied our algorithm with *min_samples* = 8, *max_dist_to_centroid* = 50, *outlier_threshold* = 2.0, and $k^{init} = 30$. This resulted in $k^* = 41$, where 1053 (i.e., all but 569) of our 1622 data points were discarded as outliers. See Fig. 5 for an illustration of the generated clusters.

3.2. DBSCAN

DBSCAN [4] operates by grouping dense groups of points into clusters. It makes use of two input parameters: *min_samples* and ϵ. A core point is a data point with at least *min_samples* − 1 other points within a distance of ϵ. A

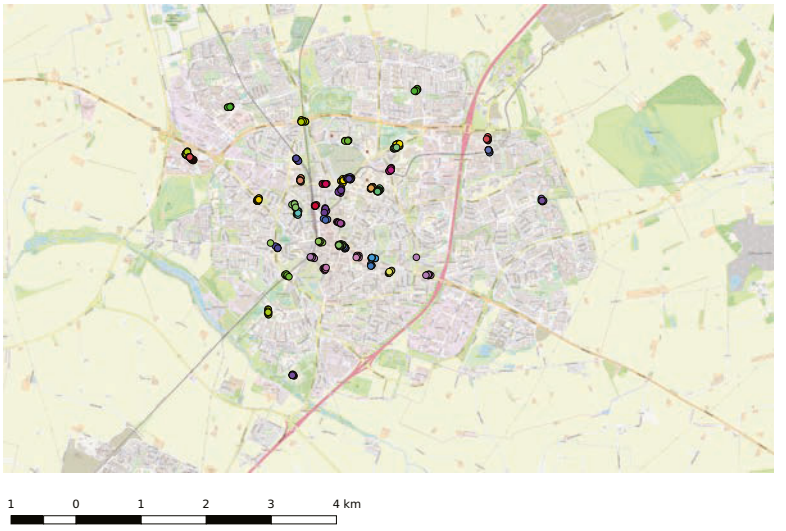

Fig. 5. Generated clusters using our iterative k-means clustering with outlier detection approach (with $k = 41$ and 1053 outliers removed).

cluster contains 1) all core points that are reachable from each other using a sequence of steps, where each step is not larger than ϵ, and 2) all other points whose distance to any core point in the cluster is not larger than ϵ. Outliers are all points whose distance to any core point is larger than ϵ.

As DBSCAN includes functionality to eliminate outliers, it can be argued that it is more straightforward to apply than standard k-means clustering. However, it is far from trivial to choose an appropriate value of ϵ. To get an idea about what ϵ is appropriate for our dataset, we compared, for *min_samples* = 8, the number of clusters generated for ϵ values ranging from 1 to 100. In Fig. 6, we present the results of this comparison and a third degree polynomial $p(x)$, which we fitted to the number of clusters for different values of ϵ. The local maximum of $p(x)$ is $p(63.1) = 46.1$.

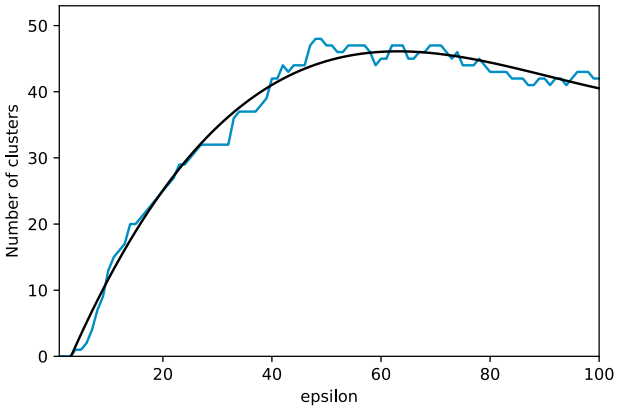

Fig. 6. The number of clusters generated using DBSCAN for ϵ values ranging from 1 to 100, and a third-degree polynomial $p(x)$, which we fitted to the number of clusters for different values of ϵ.

Since we found it reasonable to generate as many clusters as possible, we studied the DBSCAN clustering for $\epsilon = 63$. However, since the appropriateness of this choice was not obvious, we also studied the DBSCAN clusterings for $\epsilon = 53$ and $\epsilon = 73$. In Fig. 7, we present the clusters generated using DBCAN for $\epsilon \in \{53, 63, 73\}$. The number

of points included in the clusters are 855 for $\epsilon = 53$, 930 for $\epsilon = 63$, and 986 for $\epsilon = 73$. However, since there is no larger differences in the generated clusters, it seems to be reasonable to use any of the considered ϵ values.

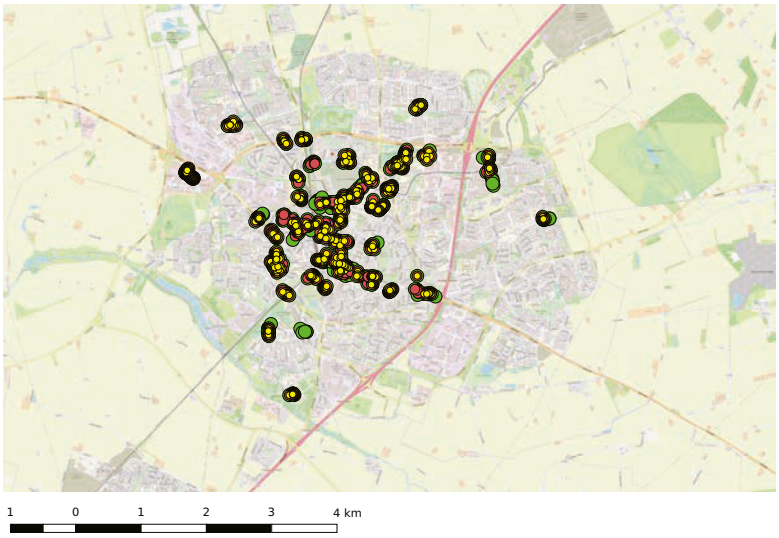

Fig. 7. DBSCAN clusters using $\epsilon = 53$ (yellow), $\epsilon = 63$ (red), and $\epsilon = 73$ (green). We used polynomial fitting in order to estimate the ϵ value that maximizes the number of clusters.

3.3. Discussion on the clustering analysis

From the visualizations of our k-means and DBSCAN cluster analyses, it is possible to make some interesting observations. In Fig. 5, it can be seen that our iterative k-means clustering with outlier detection approach generates both round and elongated clusters. However, the elongated clusters are limited in their extension due to the *max_dist_to_centroid* parameter applied in the approach. DBSCAN, which is visualized in Fig. 7, manages quite well to generate elongated clusters for all of the considered ϵ values. This is an important factor in order to identify those bicycle impediments that, for example, occur along roads.

In Fig. 8 we present the clusters generated using both our iterative k-means approach and DBSCAN with $\epsilon = 63$. The figure clearly shows that the two clustering techniques complement each other for the considered application. From the figure it can be also seen that each k-means cluster (in black) is covered by a DBSCAN cluster (in red) but the opposite is not the case. Furthermore, DBSCAN produces clusters that are larger than the *max_dist_to_centroid* criteria, which we used in our iterative k-means approach to limit the spatial size of the generated clusters. However, this is not negative as we applied DBSCAN cluster analysis in order to complement the clusters generated by k-means by identifying those extended clusters that could not be identified using k-means clustering.

4. Concluding remarks

We have presented a clustering analysis study, where we applied k-means clustering and DBSCAN in order to identify the locations of bicyclists' perceived unsafety in an urban traffic network, so-called bicycle impediments. In particular, we defined an iterative k-means with outlier detection approach that enabled us to identify an appropriate *k*-value and identify the outliers in the data set. Our analysis made use of data collected by bicyclists travelling in the city of Lund, Sweden, where each data point defines a location and time when a bicyclist reported to feel unsafe.

The generated clusters were evaluated by presenting them to traffic management practitioners, of the Lund municipality, who confirmed that the identified bicycle impediments, represented by clusters, appear to be reasonable. In particular, they were able to explain several of the impediments, whereas some of them were previously unknown.

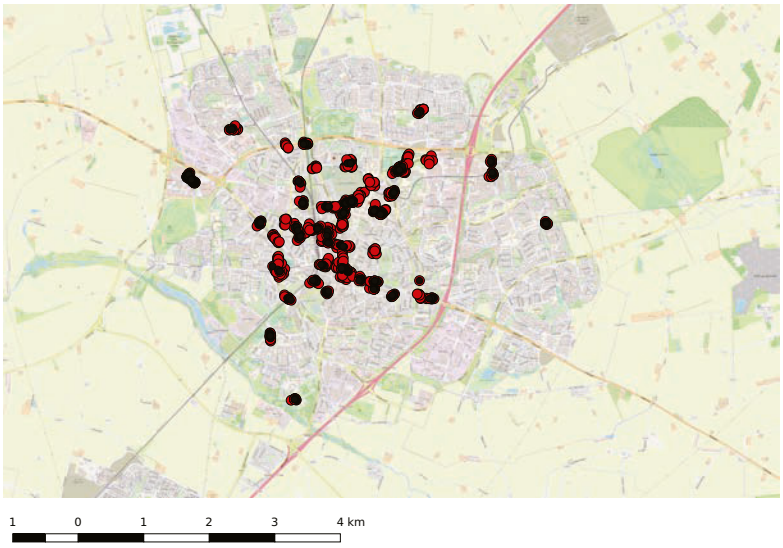

Fig. 8. Visualization of our iterative k-means clustering with outlier detection approach (black) and DBSCAN with $\epsilon = 63$ (red).

From the results of our study, we conclude that clustering is a useful approach in order to support the identification of perceived unsafe location for bicyclists in an urban traffic network. Furthermore, our results show that it is beneficial to combine different types of clustering algorithms for the considered problem.

5. Acknowledgements

The presented study was conducted as part of the project *Smart public environments II*, which was funded by Sweden's Innovation Agency (Vinnova). We acknowledge Trivector for providing the data used in the study.

References

[1] Andersen, L., Riiser, A., Rutter, H., Goenka, S., Nordengen, S., Solbraa, A., 2018. Trends in cycling and cycle related injuries and a calculation of prevented morbidity and mortality. Journal of Transport & Health 9, 217–225.
[2] Anderson, T.K., 2009. Kernel density estimation and k-means clustering to profile road accident hotspots. Accident Analysis & Prevention 41, 359–364.
[3] Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis & Prevention 37, 870–881.
[4] Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.
[5] European Union, 2018. Traffic safety basic facts 2018. EU report.
[6] Kodinariya, T., Makwana, P., 2013. Review on determining of cluster in k-means clustering. International Journal of Advance Research in Computer Science and Management Studies 1, 90–95.
[7] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, California. pp. 281–297.
[8] Montella, A., 2010. A comparative analysis of hotspot identification methods. Accident Analysis & Prevention 42, 571–581.
[9] Persson Masud, A., Olsson, V., 2019. Cyclists' perceived insecurity in urban environment - An unsupervised machine learning study. Bachelor's thesis. Malmö University. Sweden.
[10] Reddy, D., Jana, P.K., 2012. Initialization for K-means Clustering using Voronoi Diagram. Procedia Technology 4, 395–400.
[11] Singh, M., Rani, A., Ritu, S., 2014. An efficient approach - (KCVD) K-means clustering algorithm with Voronoi diagram. International Journal of Advance Computational Engineering and Networking 2, 1–4.
[12] Xu, Q., Tao, G., 2018. Traffic accident hotspots identification based on clustering ensemble model, in: 2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), pp. 1–4.