SPATIAL ANALYSIS OF EXTERNAL INFLUENCES ON TRAFFIC ACCIDENTS USING OPEN DATA

J. Golze[1]*, U. Feuerhake[1], M. Sester[1]

[1] Institute of Cartography and Geoinformatics, Leibniz University Hannover, Germany
(golze, feuerhake, sester)@ikg.uni-hannover.de

Commission IV, WG 3

KEY WORDS: traffic accidents, external influences, environmental context, open data, machine learning

ABSTRACT:

In this article traffic accidents in Germany are investigated based on different environmental impacting factors. These factors are related to the accident surrounding like road characteristics, weather information, population density as well as POIs. For this purpose different open data source are used, like OpenStreetMap and an open weather API. These input data sources are processed in order to generate different categories of features, describing the on-site situation of an accident. Using the k-means clustering, six clusters were identified. In a consecutive investigation, each cluster is not only described with respect to the feature space but also in regards to their spatial distribution. Based on these findings, a regional comparison is done across the state boarders of Germany.

1. INTRODUCTION

In our modern world with large public traffic infrastructures and the aim for autonomous and/or semi-autonomous vehicles, traffic accidents play an important role when it comes to security aspects and the safety of human traffic participants. Especially pedestrians and cyclists are the most vulnerable participants in the road traffic scenario. The greatest risk in any traffic environment originate in traffic accidents.

According to the World Health Organisation (WHO) each year 1.35 million death are caused by traffic accidents, while the majority (93%) goes back to developing countries (World Health Organization, 2019). Compared to European countries, like e.g. Germany, the number of death caused by traffic accidents is relatively small with 2.219 in 2020, when considering nearly 2 million reported accidents with property damage (Statistische Ämter, 2021).
In general, everyone can be affected by traffic accidents, either directly (involved) or indirectly (by consequences). When affected by accident consequences, for example, could result in a delay of travel time or a change in the average speed on a road. These consequences will last for a certain amount of time until the accident site is cleaned up and the traffic is back to normal. The cause(s) of traffic accidents are manifold and often do not only rely on a single aspect. Nevertheless, environmental aspects in the near surrounding of an accident could negatively influence the drivers' attention. This distraction could play a major role of the accident.

This paper investigates traffic accidents in Germany using open data like OpenStreetMap (OSM), accident data as well as weather and population information. The goal is to find influencing environmental factors for traffic accidents. Due to the lack of ground truth data, in a first step, the traffic accidents are categorized using an unsupervised machine learning approach to describe different groups of accidents. In a following step, different regions and cities are compared to get a better understanding of the most dominant traffic accident circumstances. Especially, the potential impact of structural factors is of special interest (e.g. road layout, buildings, streetlights and trees), as these could be changed in future, decreasing their influence.

The outline of this work is structured as follows. At first, the related work is reviewed in Section 2. The different data sources are described in Section 3. Afterwards, Section 4 introduces the methodology and processing pipeline used in this work. The results of our approach are presented in Section 5 and discussed in Section 6. Finally, the conclusion and future work is highlighted in Section 7.

2. RELATED WORK

The investigation of traffic accidents is an active field of research and emerged with the mass production and growing usage of motorized vehicles. In order to analyze the causes of accidents, they can be inspected with various aspects in mind and thus could show different dependencies.
One aspect to investigate traffic accidents could be the adjustment of car accident risk assessment models based on driving behaviour information. With this aspect in mind, various research had inspected different influencing factors of traffic accidents, such as the weather, humans, material or electrical failure of one or multiple cars (especially in the field of autonomous driving). Major influencing factor for traffic accidents is identified as the driver or the human aspect in general. The human aspects in the role of traffic accidents is hard to predict due to many facets of the human nature and various fields of possible experience and how these could be measured. A couple of these experience-related human factors are identified as the knowledge about the route, long-distance driving, reaction capability and others which are proofed to have an impact on the accident risk by the research of Guillen et al. (2019).
Additionally, the GNSS technology enables to create categories based on the temporal aspect of driving such as night driving and peak-hours driving (Ayuso et al., 2019). Furthermore, GNSS based localization can be combined with information on the road network infrastructure. This way geographical aspects can be considered as well as temporal aspects and provide a

* Corresponding author

more fine-grained investigation. Such aspect could be for example the road type typically driven on. Nevertheless, the risk level also depends on other environmental aspects or the urban context a driver usually passes through. Thus, it is suggested to include environmental aspects in the car accident risk assessment modelling (Husnjak et al., 2015). This environmental aspects could be split into different categories e.g. weather, POI and land use.
Many studies have already shown that specific weather conditions do have an increasing effect on the frequencies of traffic accidents. These conditions could be, for example, (heavy) rainfall (Bergel-Hayat et al., 2013; Andrey and Yagar, 1993), fog (Eisenberg and Warner, 2005) and winter precipitation (Black and Mote, 2015).
Two other aspects are the type of POIs and the land use type. Both could attract the attention of the traffic participants, and could lead to distraction and thus could increase the accident risk. While the types of POIs offer a more specific analysis in this context, the land use types indicate a more general description of the overall environmental area. Different researchers have found that POIs of the following categories show an increased accident frequency in their near surrounding: hospitals, markets, restaurants, retails, bank clusters and transportation stations (Jia et al., 2018; Lee et al., 2018). Additionally, it has been shown that the accident frequency is higher in commercial and/or mixed with residential land use while it is less frequent in areas with rural and agricultural land use (Lym and Chen, 2020; Alkahtani et al., 2019; Kim and Yamashita, 2002; Ng et al., 2002).

When considering the aforementioned influencing aspects in the field of traffic accidents, the question of accident risk prediction arise. Moosavi et al. (2019b) have built a neural network for real-time accident prediction based on sparse data. In order to proof their concept they have acquired a three years accident dataset for the US, which, additionally, include environmental attributes such as weather data, POIs and time with respect to traffic events (compare Moosavi et al. (2019a)). In their research they have found that the most significant factors for the accident prediction process are the time and nearby POIs.
Recently, (Dadwal et al., 2021) proposed an adaptive clustering approach for the prediction of traffic accidents based on temporal and regional features. Moreover, they tested different clustering approaches for the spatial aggregation as well as additional prediction methods, which were applied to three German cities. The results show that the feature groups of regional and temporal features are of higher importance.

The research of Yassin and Pooja (2020) makes use of machine learning algorithms to identify contributing factors for road accident severity. They have found that especially the environmental aspects day and light condition (as well as the driver experience and age) strongly contribute to the state of injury caused by traffic accidents in underdeveloped countries. Similar findings regarding the weather and lighting conditions are also concluded for the region of Scotland and the United Kingdom by the work of Fountas et al. (2020).

Similar to our previous work (Golze et al., 2021) on the approximation of the accident impact on the traffic flow and the estimation of the impact duration, Wong and Wong (2016) have also investigated the impact of traffic incidents based on vehicle trajectory data.

3. DATASETS

In this work, different open data sources are combined, in order to derive various descriptive attributes.

3.1 Accident Atlas

The Accident Atlas is a dataset within the framework of open data published by the Federal Statistical Offices[1] on a yearly basis, started in 2016. The dataset contains reported and investigated traffic accidents in Germany. The numbers of yearly accidents are listed in Table 1. For some states the report of accidents started in 2019.

Table 1. Overview of reported accidents provided by the Federal Statistical Offices[1]

Year	Accidents
2016	138.380
2017	195.229
2018	211.868
2019	268.370
2020	237.994
Total	**1.051.841**

The dataset provides precise location information, hour of day in UTC+1, day of the week and the month for the recorded accidents. Furthermore, a set of detailed information is encoded in the category, kind and type of accidents (see Table 2). Moreover, the travel mode of the involved accident participants is recorded (e.g. motorized 4-wheels, motorized 2-wheels, bicycle or others). While the accidents are precise in the spatial domain, they are lacking in the temporal domain. They provide only a rounded hour for the accident time and also no accurate date to identify the day of the month. Thus, multiple days per month (3 - 5) could be the possible accident day, depending on the number of weeks in the respective month. Overall, the time of an accident is defined by combining the year, month, day of the week and the rounded full hour.

3.2 OpenStreetMap

OSM has become a well-known source for freely available data and is possibly the most well-known example of Volunteered Geographic Information (VGI) systems (Jokar Arsanjani et al., 2015). In this research, the different layers of polygons (land uses), polylines (roads) and points are used to derive contextual structural information for the accident sites in Germany (OpenStreetMap contributors, 2017).

3.3 Meteorological Data

Meteorological data is gathered by the meteorological service Germany[2] (DWD), who provides historical meteorological data of each weather station in Germany in the context of their open data program. They contain directly observed but also calculated meteorological information such as temperature, humidity, wind speed, cloud cover percentage, weather condition, dew point, precipitation, pressure and visibility.
In order to receive the weather information for a specific spatio-temporal point, the data of the closest weather station is used. An effective way to query the data can be realized by using the open source project "Bright Sky" which provides a free-to-use JSON API de Maeyer and Pape (2020) for data retrieval.

[1] Statistische Ämter des Bundes und der Länder (www.statistikportal.de)
[2] Deutscher Wetterdienst - Leistungen Open Data (www.dwd.de)

Table 2. Overview of the accident characteristics category, kind and type.

Characteristic		Description
category	1	killed participant
	2	seriously injured participant
	3	slightly injured participant
kind	0	Collision with other vehicle starting, stopping or stationary traffic
	1	Collision with other vehicle driving ahead or waiting
	2	Collision with other vehicle driving sideways in the same direction
	3	Collision with another vehicle coming from the opposite direction
	4	Collision with another vehicle turning or crossing the road
	5	Collision between vehicle and pedestrian
	6	Collision with an obstacle on the roadway
	7	Leaving the carriageway to the left
	8	Leaving the carriageway to the right
	9	Accident of other type
type	1	Driving accident
	2	Turning accident
	3	Crossing accident
	4	Exceeding accident
	5	Accident on stationary traffic
	6	Accident on longitudinal
	7	other accident

3.4 Population data

Population (density) data is based on the open German census data, which was enhanced using machine vision AI using satellite images in the work of Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University (2016). The more precise resulting data are provided in a freely available repository[3].

4. METHODOLOGY

Our approach to analyze the influences on traffic accidents consists of three consecutive steps. In the first step, the accident features are determined and pre-processed. In the second step, these features are handed over to the clustering process to find meaningful accident groups. Therefore, the *k-means* algorithm is chosen in order to reduce the amount of clusters to make them more interpretable compared to other methods, like DBSCAN. Finally, in the third step, the resulting accident clusters are explored and their features are investigated to identify their characteristics. The workflow is depicted in Figure 1.

4.1 Step 1: Calculations of Features

The features are calculated using the data from the given sources to describe the environmental circumstances of the accident location and time. They can be divided into three sets of features, each with another focus of the circumstances.

The first set contains the **accident features**, which are directly retrieved from the Accident Atlas. This kind of features is derived from the accident protocols of the police, investigating the accident site, and does not need further processing. They describe the concrete situation and in which way an accident happened (e.g. how vehicle crashed and during which process),

[3] Global High Resolution Population Density Maps (www.unspider.org)

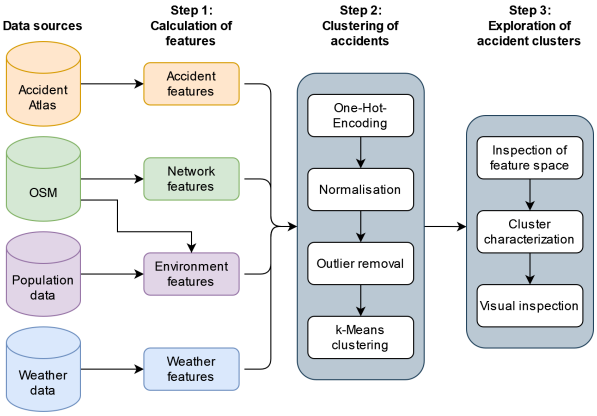


Figure 1. Workflow of the approach. The features are calculated from the different data sources in the first step. Secondly, in step 2, the k-means clustering is executed on the basis of these feature vectors. In step 3, the resulting clusters are investigated.

the type of vehicle (car, bike, etc.) participating and the general on-road situation (e.g. light situation and surface condition). Due to the nature of the documentation of the accidents (see Section 3.1), the available temporal aspects (year, day of week and hour of day) are included in the set of accident features as well. All the features are either of numeric or categorical type. All accident features are listed in Table 3.

Table 3. Overview of the **accident features**.

Accident Feature	Description
Hour	Hour of the day
Weekday	Day of the week
Category	see Table 2
Kind	see Table 2
Type	see Table 2
Participant	Travel mode of the accident participants
Light	Illumination condition [Daylight, Dawn, Darkness]
Road condition	Surface condition of the road

Secondly, there are the **network features**. They reflect the overall road network environment of an accident with regards to the corresponding roads and their related attributes, like road type, speed limit, number of road lanes. They are derived from the OSM polyline layer, which is filtered to only receive the relevant road information. Additionally, the road curvature of the road is calculated based on the approach by Franco (2016), which is the mean of the radii of every sequentially ordered points of a line segment (here: road segment). The radius for a point of the line segment is estimated based on its predecessor and successor point of the same road segment. Thus, the circular radius is calculated based on the subset of three points (see Equation 1).

$$curvature = \frac{1}{n-2} \sum_{i=2}^{n-1} radius(p_{i-1}, p_i, p_{i+1}), \quad (1)$$

where n = number of segment points
$radius(p_{i-1}, p_i, p_{i+1})$ = radius of circum-circle
p_i = i-th point p

The road length is aggregated as additional feature attribute based on the length of the consecutive road segments. Furthermore, the road network complexity is approximated to get

meaning-full insights on the granularity of the road network structure. The approximation is kept on a simple level by counting the number of individual roads in the near vicinity (50 m) of an accident location. Besides the categorical attribute road type, all other network features contain numeric values. All network features are given in Table 4.

Table 4. Overview of the **network features**.

Network feature	Description
Road type	Hierarchical level of the road
Road surface	Material of the road surface
Road length	Total length of the road
Road lanes	Number of road lanes
Road curvature	Curvature of the full road
Speed limit	Maximum allowed speed
Network complexity	Complexity of the road network

Another set of features is also derived from the OSM data. The **environmental features** are acquired from the point and the polygon layer of OSM. The polygon layer is used to extract the land use information, which is the basis for the calculation of the dominant land use type (DLT) within a buffer region of an accident location. The most dominant land use type is identified by the maximum area within this buffer region (see Equation 2).

$$DLT = argmax(lut_1, lut_2, ..., lut_i), \tag{2}$$

where lut_i = area of i-th land use type

Additionally, the land use diversity is calculated, based on the same buffer region, as the Simpson Diversity Index SDI (see Equation 3), which is known from the biological field of research Simpson (1949).

$$SDI = \sum (\frac{n_i}{N})^2, \tag{3}$$

where n_i = total area of i-th land use type
N = total area of inspected region

The land use diversity SDI provides insights on how uniform (SDI = 1) or diverse (SDI = 0) a land use is in a certain region. While a more diverse region could lead to increased traversal traffic due to mixture of available land use types. Furthermore, the points-of-interests (POIs) within each accident location are investigated using the OSM point layer. In order to have a functional separation of the different POI categories, they are split into five categorical subsets (see Table 5). These subsets distinguish the POIs by their usage (touristic, transportation, nightlife, public and commercial). Within each subset, the POIs are aggregated based on their different types (e.g. total number of all commercial POIs like shops, stores, markets, offices, etc.). Besides the dominant land use feature, all other environmental features are of numerical type.

The number of trees and streetlights are counted within the same region as the previous POIs. Both of them could have an effect on the visibility for the traffic participants and thus can be considered as important features for this analysis.

Lastly, the population information is conducted to receive the average population count in the same region as the POIs. The population feature is chosen to support the (sometimes) incomplete land use type information provided by OSM and offer a general idea of how populated a certain area is. All environmental features are listed in Table 6.

Table 5. Assignment of POI categories to the different subsets.

Subset	OSM POI types
Transportation	Bus & Tram stops, Crossings, Rental/Sharing stations, Parking, Taxis
Public	Leisure, Sports, College, Schools, Universities, Libraries, Doctor, Hospital, Police, Post
Touristic	Historic & touristic places, Hotels
Nightlife	Bars, Cafes, Restaurants, Cinemas, Theatres, Nightclubs
Commercial	Shops, Stores, Markets, Offices

Table 6. Overview of the **environmental features**.

Environmental feature	Description
Dominant land use	Major land use type
Land use diversity	Land use diversity based on Simpson Diversity Index
Number of trees	Total number of trees
Number of streetlights	Total number of streetlights
POI transportation POI commercial POI public POI touristic POI nightlife	Total number of POI of the corresponding type (see Table 5)
Population	Approximate population density for the area

Due to the restricted information available regarding the actual timestamp of an accident, the **weather features** are averaged values of all possible accident days in the range of plus/minus one hour of the accident time. In order to prohibit distortion of the values, the averaged weather features are only considered if the values of each possible accident day does not deviate from the derived average value (up-to a certain threshold) and thus have comparable weather conditions on all possible accident days. This is possible, as the weather does not dependent on a certain day of the week. All weather features are summarized in Table 7. Overall, weather features could be calculated for a total of 942.749 accident points (89%).

Table 7. Overview of the **weather features**.

Weather feature	Description	Unit
Cloud coverage	Cloud coverage	%
Dew point	Dew point 2 m above the ground	°C
Precipitation	Total precipitation during previous 60 minutes	mm
Pressure	Atmospheric pressure, reduced to mean sea level	hPa
Temperature	Air temperature 2 m above ground	°C
Visibility	Viewing distance	m
Wind gust speed	Maximum wind gust speed during previous hour, 10 m above the ground	km/h
Wind speed	Mean wind speed during previous hour, 10 m above the ground	km/h

4.2 Step 2: Clustering of Accidents

The second step in the methodology is the clustering procedure of the accident points. The clustering is based on the features described in the previous Section 4.1. Due to the need of standardized features and to handle categorical features properly, one-hot-encoding and feature normalization is applied.

At first, the outlier points are excluded. Therefore, the *DBSCAN* clustering algorithm Ester et al. (1996) is used in order to remove the outlier points from the whole set of samples. The execution of the *DBSCAN* clustering is done using approximations for the two parameters eps (0.8) and min_points (41). The eps parameter is approximated using the elbow-method with a 30% subset of the whole data samples to significantly reduce the calculation time. With the rule of thumbs, the min_points are equal to the number of features (after feature normalization) in the feature vector plus one. In the following calculations, the outlier accident points are not considered anymore.

Afterwards, the accident points are clustered using the wellknwon *k-means* algorithm MacQueen (1967). The Euclidean distance is typically utilized as distance metric for the *k-means* clustering, but is also limiting the feature vector to only contain numeric values. Thus, the non-numeric feature values are transformed using the one-hot-encoding approach, which creates one additional feature column for each characteristic. However, all features are normalized before handed over to the clustering process to have a comparable meaning of distance across the different feature attributes. In order to find the number of clusters k, multiple iterations are carried out and has been inspected. For the inspection, again, the elbow-method was used, as well as indices like the silhouette score Rousseeuw (1987) and Davies-Bouldin index Davies and Bouldin (1979). This way the number of clusters k is fixed.

4.3 Step 3: Exploration of Accident Clusters

The evaluation of the clustering results is done by, first, inspecting the feature space. Therefore, a one vs. all classification of the output clusters is done using a default Random Forest (Breiman, 2001). This way the feature importance of each cluster is investigated and compared among the clusters. Due to the large number of features, only the 10 most important features are considered in this step. The resulting distribution represents the features used to separate each cluster of accidents.

Secondly, the feature distribution within each cluster is used to characterize the contained accidents based on their similarities. Therefore, the clusters are investigated, individually, focusing on the most dominant feature characteristics.

Finally, the accidents are put in relation to their spatial distribution within the study area. This way, the overall tendencies of each cluster are highlighted using a heat map distribution.

5. RESULTS

The created results of the approach presented in Section 4 are highlighted in the following paragraphs.
Using *DBSCAN*, 47.107 accident points (4.49%) are classified as outliers and removed from the further clustering process. The remaining accident samples are handed over to the *k-means* clustering algorithm. Using the approach described in Section 4.2, the number of clusters k is set to $k = 6$. The sample distribution can be seen in Table 8.

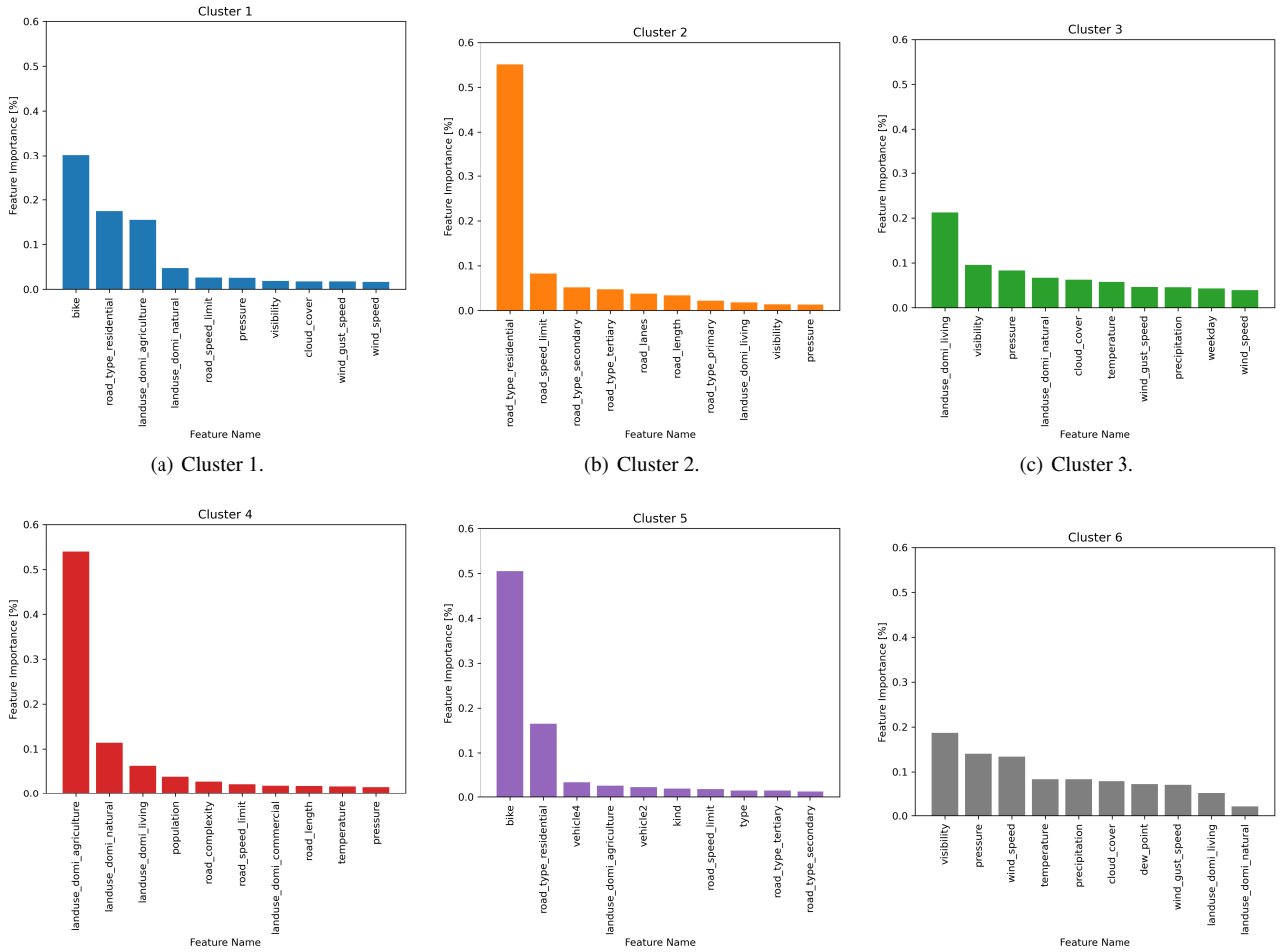
Applying the one versus all classification approach based on a Random Forest classifier and investigating the feature importance, reveals the corresponding distributions presented in Figure 2. It is worth mentioning that clusters cannot be characterized by the feature importance only. Therefore, the characteristics of the most important features have to be analyzed.

Table 8. Overview of accident samples per resulting cluster.

Cluster	Samples	Percentage
1	441.516	43.94
2	189.387	18.85
3	53.026	5.28
4	113.030	11.25
5	151.759	15.10
6	56.016	5.58
Total	**1.004.734**	**100**

It can be seen that $Cluster$ 1 (blue) is mostly represented on higher order road types and thus is distinguished from the other clusters mainly using the features $bike$ (0%), $residential\_road\_type$ (0%). Overall, accidents in this cluster can be described as accidents in natural (31.3%) and residential (52.1%) areas with a very small amount of touristic, public or nightlife related POIs close by the accident site. Accidents mostly happened on primary (22.6%), secondary (36.6%) and tertiary (20.6%) roads. Most parts consist of vehicle collisions with vehicles in front (28.5%) or turn in vehicles (20.8%). $Cluster$ 1 accidents can be described as accidents on major roads, which connect city centers.
$Cluster$ 2 (orange) is entirely described by the feature $road\_type\_residential$, which corresponds to urban areas. All accidents of this cluster happened on residential roads (100%) with a speed limit of 30 to 50 km/h (63.3%). Participants of these accidents are for most cyclists (49.5%). The dominant land use is residential (87.1%) corresponding to the road type. Further, each kind of POIs is equally represented with 10% to 16%. Accidents in $Cluster$ 2 are inner city accidents with a high chance of bicycle involvement.
The $Cluster$ 3 (green) are mostly located in natural (54.5%) or agriculture (33.9%) areas on roads like motorways (16.0%), primary (20.4%) or secondary (32.7%). Special to this cluster is the fact that 27.1% of the accidents happened during darkness and that 24.2% of the accidents occurred because of falsely leaving the road to the right. This cluster, $Cluster$ 6, contains accidents during darkness and accidents drifting off the road primarily in a rural area and along major highways.
The samples of $Cluster$ 4 (red) are located on higher order road types, like primary (19.8%) and secondary (31.6%) but also contain motorways (16.6%) and thus are pre-dominant represented on higher order roads. The fact that all accidents happened in an area with dominant land use agriculture (100%), this cluster is distinct from $Cluster$ 1. In addition, the severity of the accident is higher, heavy accidents increase from 72.9% to 81.5%. Accidents in $Cluster$ 4 are mostly located on intercity roads, surrounded by agricultural land and have a higher risk of serious injury.
$Cluster$ 5 (purple) contains solely bicycle accidents (100%) with mostly slightly injured participants (82.2%) in residential areas (72.6%). These accidents happened mostly on secondary (37.5%) and tertiary (29.4%) roads and contain a corresponding appearance of POIs of all categories, varying from 16% to 28%. $Cluster$ 5 solely contains accidents with bicycle participation in mostly residential areas and city centers.
The accidents of $Cluster$ 6 (gray) happened mostly on residential areas (99.5%), while foremost being located on roads of residential (33.2%) and secondary (27.5%) road type. Special to this cluster is the fact, that one third of the accidents happened during darkness (29.4%). It is very similar to $Cluster$ 5, but instead of bicycle accidents, it contains car accidents.

The resulting six clusters show a clear tendency regarding the spatial distribution in the visual inspection (compare Figure 3).



(a) Cluster 1. (b) Cluster 2. (c) Cluster 3.

(d) Cluster 4. (e) Cluster 5. (f) Cluster 6.

Figure 2. The 10 most important features describing the accident $Clusters$ 1-6, respectively.

The majority of the accidents appear to happen in cities and urban areas around, while also the major roads, like motorways and primaries, and contain a respective amount of accidents. Furthermore, it can be seen that the clusters are mixed in certain areas like cities, as they can be considered as the agglomeration of accidents. In order to distinguish, especially, these mixed areas, the feature attributes need to be looked at.
This aspect is also represented in the heat map visualization in Figure 4. It can be seen that all, but $Cluster$ 4, have the highest concentration of samples in and around the example city (here: Hannover) and smaller cluster centers in other residential areas. The situation is quite different for $Cluster$ 4, which following its characterization, shows the hot spots all over the major primary and motorway roads.

Considering the whole study area, the heat map visualization generally reveals an expected concentration in the largest cities as well as the major motorways.
Nevertheless, the south part is much more affected by accidents of cluster type $Cluster$ 3 and 4, while in the major northern cities (Berlin and Hamburg) accidents of type $Cluster$ 5 are more frequent. As one of the largest cities, Berlin shows only a small concentration of accidents of type $Cluster$ 3.
Cluster type $Cluster$ 1, 2 and 6 show a similar distribution in major cities. Especially, $Cluster$ 6 is very dominant solely in the major cities of south and north Germany.
While bot being part of the data collection since the beginning, the states of Lower Saxony, Saxony Anhalt, Saxony and North Rhine-Westphalia do not contain significant concentrations of any cluster type compared to north and south Germany.

6. DISCUSSION

The presented approach outputs a reasonable amount and interpretative clusters of traffic accidents. Certain features used in the clustering appear to be much more dominant than other features, which could be explained by the feature definition and their relation to the accidents.

However, due to the lack of temporal information of the accidents, the weather information is only queried for a subset of the accidents. This way, no relevance of the impact factor weather could be found in the results, in contradiction to the findings in Section 2. In order use the weather information, they need to be available for all accidents. Here it was only used for those accidents, where approximately the same weather occurred on each possible candidate day of the accident.

Moreover, the selection of POIs could be adapted based on different aspects, like the maximum viewing distance indicated by e.g. location and weather constraints, which could influence the visual range of an accident participant.

Another aspect is the usage of two different clustering algorithms (*DBSCAN* and *k-means*). Although,). Although, the

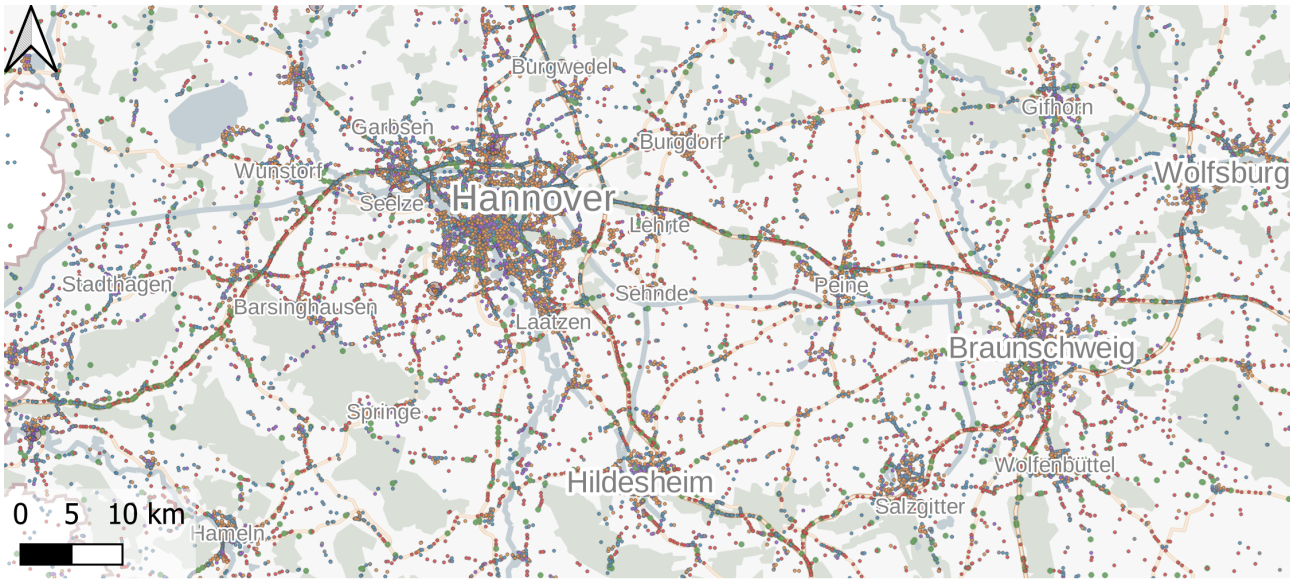
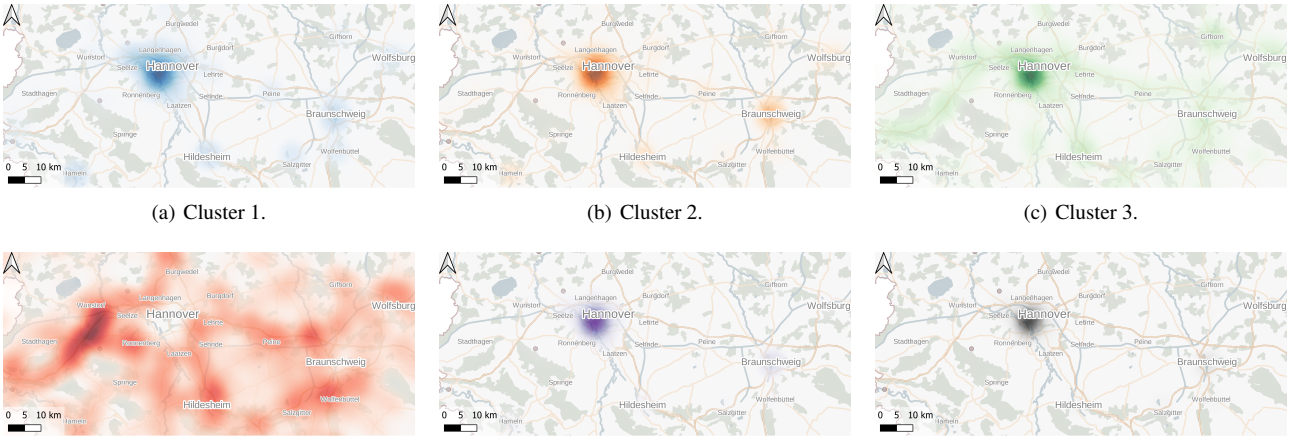


Figure 3. Accident clustering result: it can be seen that certain clusters (encoded by color) are predominantly seen in the city centers and another one mainly seen in the surrounding urban area (basemap: © LGLN — © GeoBasis-DE / BKG).



(a) Cluster 1. (b) Cluster 2. (c) Cluster 3.

(d) Cluster 4. (e) Cluster 5. (f) Cluster 6.

Figure 4. Heat map representation for each cluster for Lower Saxony (basemap: © LGLN — © GeoBasis-DE / BKG).

varying density of the sample points affects *DBSCAN*, it could be used to remove the outlier points of the samples. This way *k-means* could be applied on an outlier-free set of samples to overcome its weakness. Nevertheless, one could consider a different approach to receive a reasonable amount of clusters.

Many features appear to be not important as they have a varying broad range of values making the interpretation complex and the resulting impact very small compared to much more dominant features.
As a dominant feature the road type is not only correlated with other features like the maximum speed and number of lanes, but it is also associated with certain areas, like motorways most of the time pass through natural or agricultural land rather than city districts. Therefore, it could be left out of the analysis to empower the more basic features like maximum speed, land use types and different POIs.

The uneven start of data collection may have resulted in certain regions recording fewer accidents overall and thus fading into the background (especially in the heat map representation).

7. CONCLUSION

In this paper, we presented an approach to investigate accidents and their surrounding environmental factors using a clustering approach and the inspection of the corresponding feature spaces. This leads to a characterization of accident types. In addition, the spatial distribution of these types has been examined. It revealed that accidents could be split into urban and rural accidents, but also that accidents within cities could further be distinguished. A small tendency in different accident types could be seen when north and south Germany are compared.

Further aspects, which could be improved in future work, are to tune the clustering results by refining the set of used features. This could be implemented by a simplification of the different weather features into weather categories. Nevertheless, using weather features is questionable, when using an incomplete timestamp. Moreover, additional features could be included (from other data sources) like the general traffic flow on road segments and similarly the traffic density and local events during the accident time. Another aspect is to simplify the fea-

ture value ranges (e.g. for temperature, number of trees, etc.) to certain ranges to receive more interpretative results. Finally, the (most important) features of each cluster could be used to generate heat map predictions of possible accident regions with respect to the underlying features.

REFERENCES

Alkahtani, K. F., Abdel-Aty, M., Lee, J., 2019. A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. *International Journal of Sustainable Transportation*, 13(4), 255-267.

Andrey, J., Yagar, S., 1993. A temporal analysis of rain-related crash risk. *Accident Analysis & Prevention*, 25(4), 465-472.

Ayuso, M., Guillen, M., Nielsen, J. P., 2019. Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46(3), 735–752.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: Weather effects. *Accident Analysis & Prevention*, 60, 456-465.

Black, A., Mote, T., 2015. Effects of winter precipitation on automobile collisions, injuries, and fatalities in the United States. *Journal of Transport Geography*, 48, 165-175.

Breiman, L., 2001. Random Forests. *Machine Learning*, 5–32.

Dadwal, R., Funke, T., Demidova, E., 2021. An adaptive clustering approach for accident prediction.

Davies, D. L., Bouldin, D. W., 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 224-227.

de Maeyer, J., Pape, G., 2020. Bright Sky - JSON API for DWD's open weather data. `https://brightsky.dev/`. Accessed: 2021-11-16.

Eisenberg, D., Warner, K., 2005. Effects of Snowfalls on Motor Vehicle Collisions, Injuries, and Fatalities. *American journal of public health*, 95, 120-4.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, AAAI Press, 226–231.

Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University, 2016. High Resolution Settlement Layer (HRSL). Source imagery for HRSL © 2016 DigitalGlobe. Accessed: 2021-10-06.

Fountas, G., Fonzone, A., Gharavi, N., Rye, T., 2020. The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. *Analytic Methods in Accident Research*, 27, 100124.

Franco, A., 2016. Curvature - Find twisty roads. `https://roadcurvature.com/`. Accessed 2021-11-16.

Golze, J., Feuerhake, U., Koetsier, C., Sester, M., 2021. Impact Analysis of Accidents on the Traffic Flow based on Massive Floating Car Data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2021, 95–102.

Guillen, M., Nielsen, J. P., Ayuso, M., Pérez-Marín, A. M., 2019. The Use of Telematics Devices to Improve Automobile Insurance Rates. *Risk Analysis*, 39(3), 662-672.

Husnjak, S., Peraković, D., Forenbacher, I., Mumdziev, M., 2015. Telematics System in Usage Based Motor Insurance. *Procedia Engineering*, 100, 816–825.

Jia, R., Khadka, A., Kim, I., 2018. Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis & Prevention*, 121, 223-230.

Jokar Arsanjani, J., Mooney, P., Zipf, A., Helbich, M., 2015. *An introduction to OpenStreetMap in GIScience: Experiences, Research, Applications*.

Kim, K., Yamashita, E., 2002. Motor Vehicle Crashes and Land Use: Empirical Analysis from Hawaii. *Transportation Research Record*, 1784(1), 73-79.

Lee, J., Chae, J., Yoon, T., Yang, H., 2018. Traffic accident severity analysis with rain-related factors using structural equation modeling – A case study of Seoul City. *Accident Analysis & Prevention*, 112, 1-10.

Lym, Y., Chen, Z., 2020. Does space influence on the frequency and severity of the distraction-affected vehicle crashes? An empirical evidence from the Central Ohio. *Accident Analysis & Prevention*, 144, 105606.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1, Oakland, CA, USA., 281–297.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Ramnath, R., 2019a. A countrywide traffic accident dataset.

Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., Ramnath, R., 2019b. Accident Risk Prediction based on Heterogeneous Sparse Data. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Ng, K., Hung, W., Wong, W., 2002. An algorithm for assessing the risk of traffic accident. *Journal of Safety Research*, 33(3), 387-410.

OpenStreetMap contributors, 2017. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Simpson, E. H., 1949. Measurement of Diversity. *Nature*, 163(4148), 688.

Statistische Ämter, 2021. Verkehrsunfälle 2020. *Fachserie. 8, Verkehr. 7*.

Wong, W., Wong, S. C., 2016. Evaluation of the impact of traffic incidents using GPS data. *Proceedings of the Institution of Civil Engineers - Transport*, 169(3), 148–162.

World Health Organization, 2019. *Global Status Report on Road Safety 2018*. World Health Organization, Geneva.

Yassin, S. S., Pooja, 2020. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, 2(9), 1576.