Abstract

The GIDAS (German In-Depth Accident Study) data set includes more than 7,000 cyclists involved in an accident in the German Hannover and Dresden areas – the majority of these accidents were between a bicycle and a car, followed by objects, other bicycles, commercial vehicles etc. Of these cyclists, more than 300 sustained severe injuries (i.e., MAIS 3+1 injuries). The relationship between accident risk, injury risk and other factors in different situations is complex and difficult to understand by standard analysis tools.

The results of the GIDAS data bank analysis form the basis for the study of new, nontrivial relationships within the data set. New subsections of automated cluster and correlation analysis will be used for the first time for the analysis of accident data, with the help of tools such as Dynamic Semantic Data Mining (DSDM). This new method is well-suited to supplement pre-classified data bases by data-driven categories that will be derived from the specific context and need to be detected automatically.

The aim of this study is to arrive at a better understanding of the nontrivial and multiple relationships of accidents with bicycle involvement. Only then can bicycle traffic safety be increased in the future - for example, through new advice and recommendations regarding non-motorised bicycle traffic.

Keywords: GIDAS data base, bicycle traffic, accident analysis, AcubeS, Big Data, Smart Data.
1 INTRODUCTION AND CONTENT DESCRIPTION

In order to further improve cycling safety a deep understanding of accident data is crucial. However, due to the complexity of accident situations it is difficult to understand the relationships between accident risk, injury risk and other factors. The objective of this paper is to demonstrate a new method for a nontrivial multi-relationship accident analysis.

Two accident data sources were used in this study. To analyse the most important accident situations in Germany, the German national accident statistics of the year 2013 was used [1]. The national statistics includes all police reported road accidents with a limited depths. For example the injury severity in the national statistics is described by fatally injured (road user died within 30 days after the accident as a direct result of the accident), severely injured (road user was treated in a hospital for more than 24 hours) directly after the accident and slightly injured (all other injured road users).

For a more detailed analysis of the relevant bicycle accidents the GIDAS (German In-Depth Accident Study) data set of the years 2005 – 2013 was also used. GIDAS (German In-Depth Accident Study) is the largest and most comprehensive in-depth road accident study in Germany. Since mid-1999, the GIDAS project investigates about 2000 accidents in the areas of Hannover and Dresden per year and records up to 3000 variables per crash. The project is supported by the Federal Highway Research Institute (BASt) and the German Association for Research in Automobile Technology (FAT) [2]. The sponsors and the investigation teams have access to the data. In GIDAS, road traffic accidents involving personal injury are investigated according to a statistical sampling process using the “on the scene” approach. That means, teams are called promptly after the occurrence of any kind of road traffic accidents with at least one injured person which occurred in determined time shifts. Along with this method, severe accidents are recorded slightly more frequently than accidents with lower injury severities and this is mainly caused by a lower notification rate or late information. In order to avoid such biases in the database and to approach regional and national representativeness, comparisons are made regularly with the official accident statistics and weighing factors are applied. In addition, the investigation areas were chosen accordingly to the national road network and built-up areas. The detailed documentation of the accidents is performed by survey teams consisting of specially trained students, technical and medical staff. The data scope includes technical vehicle data, crash information, road design, active and passive safety systems, accident scene details and cause of the accident. Surveyed factors include impact contact points of passengers or vulnerable road users, environmental conditions, information on traffic control and other parties (road users) involved. Additionally, vehicles are measured more in detail, further medical information is gathered and an extensive crash reconstruction is performed. In the GIDAS data set the injury severity is described following the national statistics metrics but also using the AIS (Abbreviated Injury Scale describing the mortality risk in an ordinal scale ranging from AIS 0 - uninjured to AIS 6 – no medical treatment possible) code for every individual injury. Because of the more detailed nature of the AIS scale this scale is used as metrics for the analysis of the GIDAS data.

Finally, a selection of the reported bicycle to motorised vehicle accidents in the GIDAS data base was used for the new analysis method.

The German national statistics show a slightly increasing trend for slightly injured bicyclists for all accident configurations, i.e. not only for bicycle-to-motorised vehicle accidents (Figure 1). It is expected that this trend follows an increased share of bicycle trips. For seriously injured bicyclists there is a slightly decreasing trend, and for fatally injured cyclists there is a strongly decreasing trend that follows the general trends in the German road accident statistics.
Figure 1. Development of cyclist accidents and injury severity in Germany.

Most bicycle accidents happen in collision with passenger cars. However, the risk for fatal injuries compared to the accident risk is especially high for single bicycle accidents and for bicycle to-HGV (Heavy Goods Vehicle) accidents (Figure 2).

Following the analysis of the German national statistics car-to-bicycle accidents account for the majority of cyclists injured or killed. In order to analyse this accident situation in more detail, GIDAS car-to-bicycle accident of the years 2005 to 2013 was used.

![Graph showing the development of cyclist accidents and injury severity in Germany.]

Figure 2. Injury severity levels depending on accident opponent (note all injured cyclists include slightly injured, severely injured and fatally injured cyclists).

Here especially the type of accident appears to be a good starting point for the understanding of the individual accident situation. The type of accident describes the conflict situation which resulted in the accident, i.e. a phase in the traffic situation where the further course of events could no longer be controlled because of improper action or some other cause. The type of accident does not describe the actual collision but indicates how the conflict was triggered before this possible collision. The determination of the type of accident also plays an important role in local accident analysis since the type of accident is marked by coloured pins on the maps of the local police authorities. The following seven types of accidents are distinguished (Figure 3):
1) Driving accident: The accident was caused by the driver’s losing control of his vehicle (e.g., due to inappropriate speed or misjudging the course or condition of the road), without other road users having contributed to this. As a result of uncontrolled vehicle movements, however, a collision with other road users may have happened. A driving accident however does not include accidents in which the driver lost control of his vehicle due to a conflict with another road user, an animal or an obstacle on the carriageway, or because of a sudden physical incapacity or a sudden defect of the vehicle. In the course of the driving accident, this vehicle may collide with other road users, so that this is not necessarily a single vehicle accident. For this study the loss of control could be caused by either of the two opponents (here bicycle or HGV).

2) Accident caused by turning off the road: The accident was caused by a conflict between a vehicle turning off and another road user approaching from the same or opposite direction (including pedestrians) at crossings, junctions and entries to premises or car parks. Whoever follows the priority turn of a main road is not considered as turning off.

3) Accident caused by turning into a road or by crossing it: The accident was caused by a conflict between a road user turning into a road or crossing it when he actually should have yielded and a vehicle having the right of way at crossings, junctions, or exits from premises and car parks.

4) Accident caused by crossing the road: The accident was caused by a conflict between a vehicle and a pedestrian on the carriageway, unless the pedestrian walked along the carriageway and unless the vehicle turned off the road. This applies also where the pedestrian was not hit by the vehicle. Even if the pedestrian who caused the accident was not hit, the accident is classified as caused by crossing the road. A collision with a pedestrian walking along the carriageway is recorded as a no. 6 type of accident.

5) Accident involving stationary vehicles: The accident was caused by a conflict between a moving vehicle and a parked/stopping vehicle or a vehicle manoeuvred in connection with parking/ stopping. Accidents with vehicles waiting just because of the traffic situation are not included. For this study this applies most often to stationary trucks and cyclists, for example, an accident with an opening door.

6) Accident between vehicles moving along the carriageway: The accident was caused by a conflict between road users moving in the same or opposite direction, unless this conflict caused by a different type of accident.

7) Other accident: This includes all accidents that cannot be allocated to any other type of accident. Examples: U-turning, reversing, accidents between parked vehicles, obstacles or animals on the carriageway, sudden failure of the vehicle (brake failure, defective tyre, etc.).

The majority of accidents are of the type “turning into a road or crossing it” or “turning off the road”, see Figure 3.
In order to be able to understand the conflict in more detail, GDV\(^2\) developed a three digit code for the type of accident with the first digit being identical to the standard type of accident. For the subsequent analysis detail types of accident exceeding 5% of the accidents were used. In total these accidents account for approx. 45% of the recorded bicycle-to-car accidents in the GIDAS data base.

Table 1. Selected bicycle-to-car accidents for detailed analysis according to their type of accident\(^3\)

<table>
<thead>
<tr>
<th></th>
<th>342</th>
<th>341</th>
<th>321</th>
<th>243</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>W</td>
<td>R</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24%</td>
<td>8%</td>
<td>6.8%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

2 GENERAL INSTRUCTIONS AND ANALYSIS METHOD

Cyclists are not specifically protected in the road environment. At present it is largely unknown how accidents with serious consequences for a certain part of cyclists can be prevented. Therefore, we developed a method for an automated analysis using the latest findings from the Dynamic Semantic Data Mining (DSDM). Free of a

---

\(^2\) GDV: Gesamtverband der Deutschen Versicherungswirtschaft (German Insurance Association)

\(^3\) R matches the cyclist; W matches the direction which has to give way (yield)
corresponding bias the new way should now found divergences. The approach has the objective to develop automatically question identifies by use of latest technology.

### 2.1 The Approach

The scientific approach to Accident Analysis 2.0 deals with applying new methodologies involving pattern recognition to support visual data exploration. It is of paramount importance that domain experts, without the help of technical experts, can sift through densely packed blocks of information with an appropriate visualization tool.

To extract unknown facts from data, new methodologies are necessary, which, in contrast to conventional approaches, reverse the analytical procedure. Up to this point, data analysis has largely involved searching a sea of data for answers to given questions. These questions must be formulated in advance by humans. Without these questions there are no answers. This would usually lead to one of the following two (less than ideal) situations, where personal biases of the domain expert could influence the results:

- The domain expert excludes certain situations or issues, because he or she personally considers them unlikely.
- The domain expert has a certain question and analyses the data based only on facts regarding this issue. Other interesting information will remain hidden from him.

In contrast to conventional methods is the data-driven approach to data analysis. The data driven approach prevents the bias of the domain expert from having an impact on the evaluation of the data. There is no premise that restricts the analysis of the data. Instead of merely looking for answers, new questions from the data are derived. Once the questions have been generated from the data, the next step involves putting a domain expert in a position where he can explore the data in search for answers to these questions. The domain expert is supported by exploratory data analysis tools.

For analysing the data, algorithms used in data mining and the technology of Big Data are combined with classical methods of mathematical analysis. Both the amount of data to be analysed and the number of computations per record are highly scalable. Any mathematical calculations are assembled into an analysis pipeline. Although the computational effort for the analysis of a single record rises with it, such pipelining allows for a better automated interpretation of the results. The resulting data reduction and filtering is presenting domain experts the data with the highest information content. The practical implementation is described in the chapters 2.2 and 2.3.

With this new application logic and the related study of this new topic, a symbiotic collaboration between domain experts and data scientists can be realized as shown in Figure 4.

### 2.2 The analysis Process and Proceedings

The analysis process is divided into several sub-steps that take place in the processes described below.

First, the overall GIDAS data set is clustered using density-based clustering and pattern recognition techniques to identify pertinent relationships within the data. Through various combinations of data evaluation methods, less interesting data is separated out from data promising a high information value. The resulting relevant clusters are then differentiated and visualised according to various attributes, for example, according to correlations of the cluster attributes or according to cluster sizes (see Figure 5).

Big Data plays a major role for the described analysis process, although the data set used for the examples in this paper is quite small. However, if high-quality forecasting and evaluation models are to be trained, the computational effort increases exponentially. If only 10 different parameters are evaluated with 5-fold cross-validation, the computational time is 50 times as long. The amount of data also increases virtually in accordance with the same ratio.
For the project at hand, small amounts of data have been evaluated and re-evaluated using a number of different analytical processes and the results compared to each other. Big Data also plays a central role when evaluating the quality of the clustering in general and the analysis of variance in particular. Density-based clustering is achieved in this study by evaluating the distance between points on an attribute by attribute basis. Furthermore, distributed computing provides the power needed for the extraction of patterns hidden in the data.

---

4 Density based clustering, cluster validation, calculation of correlation coefficients and Spanning Trees
The next step consists of computing the correlation between attributes for each cluster to generate a graph of the reduced set of data containing the most pertinent information for each cluster. Using this overview of cluster graphs, a domain expert would then be in a position to decide which clusters should be explored in more depth (Figure 6).

### 2.3 The Evaluation Options

These cluster graphs, which show not only pairwise relationships between attributes but also chains of relationships across a number of attributes, allow the domain expert to answer a number of the questions arising during the analysis process. In the Discussion section of this publication, two examples will be used to demonstrate how useful information can be gleaned from data using the analysis process presented here. The AcubeS (Analytics as a Service) system developed at the Fraunhofer Institute for Transportation and Infrastructure Systems was used for visual data exploration and for preparing the examples. AcubeS allows comprehensible visualizations, which facilitate the interpretation of the results.

The AcubeS system is based on open source software and frameworks. AcubeS consists of 4 main parts. The first part is the computation clusters based on Linux clusters running on standard hardware. The second part is the Big Data computation software based on Apache Hadoop and Apache Spark. The third part is the Web Portal for user management and storage of computation results. The fourth part is the Web based visualisation of the computation results for the particular user. Each part is based on Open Source software. Each part is substitutable and customizable on every level as necessary. These properties are needs for the research based nature of our work. No proprietary software meets these needs for our work.

### 2.4 Domain Interpretation

As described in the previous section, correlation matrices and spanning trees provide a useful means to demonstrate the correlations between various elements within a cluster. One such example is shown in Figure 7.

---

5 http://dateneinblick.de/index.php/en/
This spanning tree presents correlations between the collision angle (KWINK) of the bike and the other vehicle and the variables VDI1, accident type (UTYP), bike path used (RADWBEN), existing bike path (RADW), gender of driver (GESCHL) and the vehicle class (FZGKLASS). This therefore gives the authors good reason for going deeper into the analysis of the correlations of the collision angle.

The collision angle is defined as the spread of the different directions of speed vectors of both collision participants. It is always given in mathematical positive direction. An example of measuring the collision angle is shown in Figure 8.

A very interesting correlation is of course between the collision angle and the type of accident, which will now investigate this correlation more closely. Figure 9 shows for each case of the basis dataset the link between type of accident on the left and the collision angle on the right.
Figure 9. Accident type - collision angle.

For the accident type 243 (purple) the collision angle of most accidents is in the range of -20° to -90°.

Figure 10. Accident type 243.

The cases with accident type 321 (green) have two ranges. The first range is near 90° and the second range is near -90°. Whether the angle is positive or negative is determined by the accident type and whether or not the bicycle is the vehicle at fault.

Figure 11. Accident type 321.

For the accident type 341 (blue) the range of the collision angle is between 30° and 120°.
Figure 12. Accident type 341.

For the accident type 342 the range is between -70° and -130°.

Figure 13. Accident type 342.

Each accident type considered here is characterized by a relatively specific collision angle between the bicycle and the other vehicle.

3 DISCUSSION

From the technical point of view complex data analytical topics are combined with domain analysis. The analytical topics include:

- The pre-qualification of data and density-based pattern recognition through clustering
- The data reduction through cluster validation and calculation of correlation coefficients
- The automatable data exploration by calculating Spanning Trees

Our continuing research will build on the findings in this paper. We want to extent the described analysis method with the operation of non-numerical data. Furthermore we want to improve the current limits of separation of spurious correlations and serious correlations.

A very plastic example was given by the extraction of the collision angle out of the number of the type of accident. Without asking the Question: “is there a correlation between the type of accident and the collision angle?” the system shows that it is. Regarding the pictures of the type of accident (see figures 10 – 13) a domain expert can validate the detected fact from his knowledge and visual representation. The applied analysis process detected this fact without domain knowledge and without a visual representation. The base is only the cluster analysis and the correlation of all given accident data.

Possibilities of the new method are for example:

- Get unknown connections inside the accident dataset (plausibility, filling up missing or corrupt data points)
- Get connections between the accident data and external data to find new Questions

Especially the second possibility will give future research tasks and some solutions for Questions of safer road traffic, which are not part of any accident database or analytic-method now. The further development of the shown methodology and the ACUBES-system in combination with accident databases will help to get ideas for reducing the number of accidents, injuries and fatalities in traffic.
4 CONCLUSIONS

The methodologies of automated clustering [7] and correlation analysis have been combined for the first time in the analysis of manually compiled accident data [9]. This new method is well-suited to supplement preclassified data sets by data-driven approaches for categorising data, which will result from the specific context and need to be detected automatically.

The study has shown that this methodology can aid to the better understanding of nontrivial and multifaceted relationships of accidents including those where bicycles were involved. Only then can bicycle safety be increased in the future - for example, through new guidelines and recommendations regarding non-motorized bicycle traffic.

In this practical example of the first time application of the new accident data analysis methods presented above – such as Big Data cluster analysis [6], [8] and its correlation dependencies as well as the ways in which domain experts and data scientists can support each other in their work - two general conclusions can tentatively be put forward:

1. The automated analysis of big datasets of accident data leads to real relations and research topics without the need to determine an assumption beforehand.
2. The automatically generated results of the accident analysis are plausible and can be used for numerous validations and quality management processes.

The methodologies and approach presented above have great potential in further enhanced, automated accident analysis.

5 REFERENCES

[1] DESTATIS (Statistisches Bundesamt). Fachserie 8 Reihe 7 Verkehrsunfälle 2013, Juli 2014 (in German)