

EXTRAPOLATION OF GIDAS ACCIDENT DATA TO EUROPE

Jens-Peter Kreiss

Gang Feng

Jonas Krampe

Marco Meyer

Tobias Niebuhr

Technische Universität Braunschweig

Germany

Claus Pastor

Jan Dobberstein

Bundesanstalt für Straßenwesen

Paper Number 15-0372

ABSTRACT

In the paper it is investigated to what extent one can extrapolate the detailed accident database GIDAS (German In-Depth Accident Study), with survey area Hanover and Dresden region, to accident behavior in other regions and countries within Europe and how such an extrapolation can be implemented and evaluated. Moreover, it is explored what extent of accident data for the target country is necessary for such an extrapolation and what can be done in situations with sparse and low accident information in a target region.

It will be shown that a direct transfer of GIDAS injury outcomes to other regions does not lead to satisfactory results. But based on GIDAS and using statistical decision tree methods, an extrapolation methodology will be presented which allows for an adequate prediction of the distribution of injury severity in severe traffic accidents for European countries. The method consists essentially of a separation of accidents into well-described subgroups of accidents within which the accident severity distribution does not vary much over different regions. In contrast the distribution over the various subgroups of accidents typically is rather different between GIDAS and the target. For the separation into the subgroups meaningful accident parameters (like accident type, traffic environment, type of road etc.) have been selected. The developed methodology is applied to GIDAS data for the years 1999-2012 and is evaluated with police accident data for Sweden (2002 to 2012) and the United Kingdom (2004 to 2010). It is obtained that the extrapolation proposal has good to very good predictive power in the category of severe traffic accidents.

Moreover, it is shown that iterative proportional fitting enables the developed extrapolation method to lead to a satisfactory extrapolation of accident outcomes even to target regions with sparse accident information. As an important potential application of the developed methodology the a priori extrapolation of effects of (future) safety systems, the operation of which can only be well assessed on the basis of very detailed GIDAS accident data, is presented.

Based on the evaluation of the presented extrapolation method it will be shown that GIDAS very well represents severe accidents, i.e. accidents with at least one severely or fatally injured person involved, for other countries in Europe. The developed extrapolation method reaches its limits in cases for which only very little accident information is available for the target region.

INTRODUCTION

In this paper we present a methodology which allows for an extrapolation of the detailed German accident database GIDAS (German In-Depth Accident Study) to other regions or countries within Europe. The great advantage of GIDAS is that due to (in the regions of the German cities Hanover and Dresden) accident parameters are recorded in fine detail. However, the accident information within GIDAS is regionally restricted and cannot be transferred in a one-to-one manner directly to other regions or countries (even not in Germany). We suggest to apply well-developed statistical decision tree methods to achieve this goal. In [8] a different so-called weighting methodology has been developed and applied to GIDAS in order to extrapolate GIDAS to other regions in Germany.

To evaluate the proposed methodology we obtained rather detailed police recorded accident material for Sweden and the United Kingdom (UK) from the Swedish Transport Agency and the Bundesanstalt für Straßenwesen.

In the next section we briefly describe the accident data we used in this study. Then we describe in full detail the suggested decision tree method and the way of obtaining extrapolations for other regions or countries (target regions) within Europe. In order to see how the proposed methodology works for real accident data we report on applications to Swedish and British accident data. Since we have detailed accident data for these two countries at

hand we can even evaluate the proposed methodology. It will be seen that the proposed methodology works well for injured people in severe accidents, i.e. not taking uninjured participants into account.

As an example for application of such an extrapolation we present how the effectiveness of a fictional future safety system in vehicles can be predicted for regions or countries in Europe. The term “future safety system” refers to a system with no or low market penetration. For such systems an evaluation of the effectiveness cannot be carried through on the basis of recorded accident data. For complex systems even police recorded data may not be detailed enough to be able to measure effectiveness. Furthermore, we consider in a further section the situation when only low and aggregated accident information in the target country is available. Here we suggest to combine the proposed extrapolation method with the so-called iterative proportional fitting method (IPF). Finally, we will also discuss extrapolation possibilities for countries for which we only have number of fatalities broken down to very few accident parameters like type of vehicle and location of accident (rural versus urban). As an example in this section we have taken Austria. The paper is concluded by a summary of the obtained results.

DATA SETS USED

For the evaluation and application of the proposed extrapolation of injury outcomes in accidents we make use of accident data from GIDAS the database for the period 1999 until 2012. These data contain detailed accident information of 24,341 injury accidents with 32,312 at least slightly injured people. As is discussed later we restricted ourselves to *severe accidents* (accidents with at least one severely or fatally injured person). This restriction leads within GIDAS to 7,474 severe accidents with 10,982 at least slightly injured people.

Based on this accident data we exemplarily extrapolate to severe accidents in Sweden and United Kingdom. To evaluate the obtained extrapolations we make use of Swedish accident data, which we have received from STRADA for the period 2002 until 2012 and make use of STATS19 accident data for UK for the period 2004 until 2010. The STRADA data set for Sweden contains 36,320 severe accidents with 60,999 injured people and the STATS19 data set for UK contains 184,263 severe accidents with 283,201 injured people.

Finally we present an application making use of accident data from CARE for the period 2008 until 2013. A special feature of this accident data set is, that it contains numbers of fatalities in road accidents for European countries, only, but not numbers of injured/uninjured persons involved.

DECISION TREE BASED EXTRAPOLATION OF ACCIDENT BEHAVIOR

The main idea for accident extrapolation of GIDAS injury outcomes in road accidents to various target regions is to apply the well-established statistical method of decision trees (cf. [2], [9] and tailored for accident data [10]). The target variable, which we intend to predict for different countries or areas, is injury severity of people involved in road accidents. In order to obtain a widely applicable procedure we choose a rather simple categorization of injury severity, namely only the four categories *not injured*, *slightly injured*, *severely injured* and *fatally injured*. In GIDAS a person is referred to *severely injured*, if it has been hospitalized. In most European countries police recorded accidents contain some information about the injury severity of people involved in the accidents. Many countries also make use of the above categorization. However, the distinction especially between *slightly* and *severely injured* varies with different countries. Concerning Sweden and United Kingdom and also Austria, to which we exemplarily apply our method, the definition of injury severity is quite similar to that of Germany.

Proper application of decision tree methodology moreover needs decision variables in order to create a set of disjoint accident categories with their own and specific injury severity distributions. We applied a detection of relevant variables assisted by so-called iterated decision trees or bootstrap aggregation (cf. [3], [4], [5] and [14]). It is obtained that essential variables for this accident segmentation are

- ACCIDENT TYPE (UART in GIDAS)
- TRAFFIC ENVIRONMENT/LOCATION OF ACCIDENT (ORTSL in GIDAS)
- VEHICLE TYPE (FART in GIDAS)
- TYPE OF ROAD (STRART in GIDAS)
- NUMBER OF INVOLVED PARTIES (ANZBET in GIDAS)

- SPEED LIMIT (VZUL in GIDAS)
- LIGHTCONDITIONS (TZEIT in GIDAS)
- SEX (GESCHL in GIDAS)
- AGE OF CASUALTY (ALTER1 in GIDAS)
- PEDESTRIAN
- DIRECTION OF FIRST IMPACT (VDI1 in GIDAS).

For practical application it is of course most important to know which accident variables are available for the target region or country. We consider as a case a person involved in an accident. Persons involved in one and the same accident can be grouped into different categories depending on their accident parameters. So the method can take into account that passengers of a vehicle typically suffer different types of injuries as pedestrians or bicycle riders in accidents with each other.

Extensive investigations of extrapolation of GIDAS accidents to other European regions or countries clearly showed that reliable results can only be obtained if the consideration is restricted to at least slightly injured people in *severe accidents*, where the definition of a severe accident is, that at least one severely or fatally injured person is involved in that accident. However, within the category of *severe accidents* all involved injured people are taken into account and not only severely and fatally injured people.

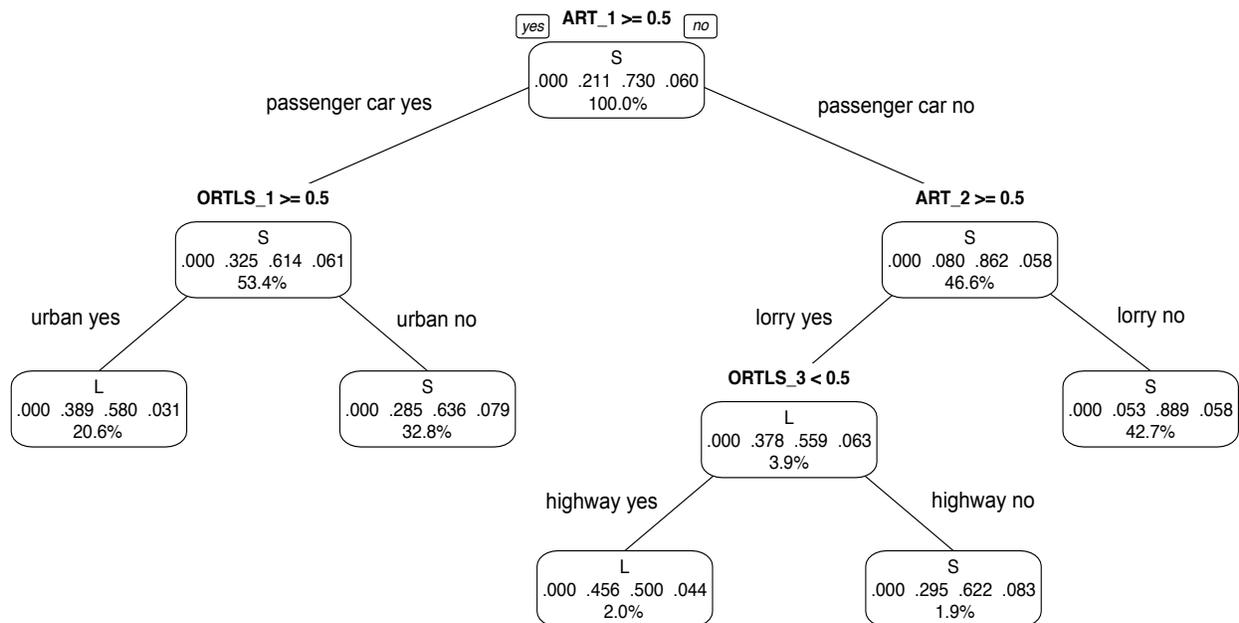


Figure 1: Example of a decision tree based on GIDAS (1999-2012), which can only make use of a very limited number of variables related to traffic participation and accident location.

Based on the selected and available variables a segmentation of injured people involved in severe accidents from GIDAS based on a decision tree method is constructed in a first step. Decision trees have the advantage that the segmentation results are very well visualized. For the application to Swedish accident data a part of such a decision tree is displayed in Figure 2. A decision tree is a kind of flowchart-like structure in which each path from the root to the various leaf nodes is represented by a sequence of classification rules. Each leaf node represents a selected category of injured people in accidents.

A resulting decision tree for accident data is explained in the following for a clear and limited example where only two variables (type of traffic participation and accident location) for each injured person in a severe accident are available (cf. Figure 1). According to the possible expressions of these two characteristics, eight

0-1-variables, namely

- Passenger car (ART_1)
- Lorry or Truck (ART_2)
- Motorized two-wheelers (ART_3)
- Pedal cycle (ART_4)
- Pedestrian (ART_5)

and

- Urban (ORTLS_1)
- Rural (ORTLS_2)
- Highway (ORTLS_3),

may be used in order to set up a decision tree. Put simply, the routine `rpart` now searches for the one of these variables which (according to their values yes/no) best splits the set of all injured people involved in severe accidents within GIDAS into two subgroups (internal nodes). More precisely `rpart` minimizes a specific measure of impurity or diversity (Gini index or information index) of a node (subgroup) in order to obtain a best split. A node is called completely pure if all persons within this node fall into the same injury category, i.e. the injury distribution is degenerate. In the considered example `rpart` selected the variable ART_1, which means “people seated in a passenger car: yes or no”. The internal node *people seated in a passenger car ‘yes’* in a second step is separated according to the variable ORTLS_1, which means “urban accident location: yes or no”. After this step no further separation is suggested and two leaf nodes of the decision tree are obtained. The internal node *people seated in a passenger car ‘no’* is separated according to the variable ART_2, which means “people seated in a lorry or truck: yes or no”. If ‘yes’ a further separation is done depending on whether the accident happened on a highway or not. At the end five leaf nodes (endpoints of the branching) are obtained. Each leaf node contains the injury distribution (uninjured, slightly injured, severely injured and fatally injured). Since we have not included uninjured people the corresponding share always is zero. Above the injury distribution a letter indicates the predicted injury category for this specific leaf node. *L* stands for *slightly injured* and *S* stands for *severely injured*. Below the various injury distributions the share of all injured people in severe accidents who fall into the corresponding leaf node are displayed.

The decision tree, which is completely based on GIDAS data, provides a complete segmentation of all injured persons in severe accidents into so-called leaf nodes. From the injury outcomes of all cases within a leaf node we calculate an injury distribution, which typically varies substantially over the various leaf nodes. As a further and more realistic example consider in Figure 2 (decision tree making use of variables available for Sweden) the leaf node labeled number 6. This leaf node contains all injured people in accidents for which FART_3=0 and FART_5=1. The coding FART_3=0 and FART_5=1 means that this leaf node contains all injured persons which have been involved in severe accidents, have not been passengers of a passenger car (FART_3=0) but have been passengers of a bus (FART_5=1). Within this subgroup of injured people the injury severity distribution is as follows: 67.1% of the involved injured people are slightly injured, 30.3% are severely injured and 2.6% are killed. As a further example consider leaf node number 5. The coding FART_3=1 and ANZBET=1 means that within this subgroup injured people sitting in a passenger car and involved in a single vehicle accident are collected. The injury severity distribution for this subgroup reads rather different, namely: 15.1% of the involved people are slightly injured, 74.5% are severely injured and 10.4% are killed.

In order to apply a decision tree method to a target region (specific European region or country) an essential point to be clarified is, which accident variables for this region are available and whether they can be harmonized with accident information within GIDAS. As an example assume that for the target region we only know for each injured person in a severe accident the ACCIDENT TYPE, TRAFFIC ENVIRONMENT and VEHICLE TYPE and that the values each of the three variables can take are similarly defined in the target region compared to GIDAS. Then the calculation of a decision tree has to be based on these variables.

For the target region or country it is assumed that we do not have any information on the injury outcomes of people involved in the accidents.

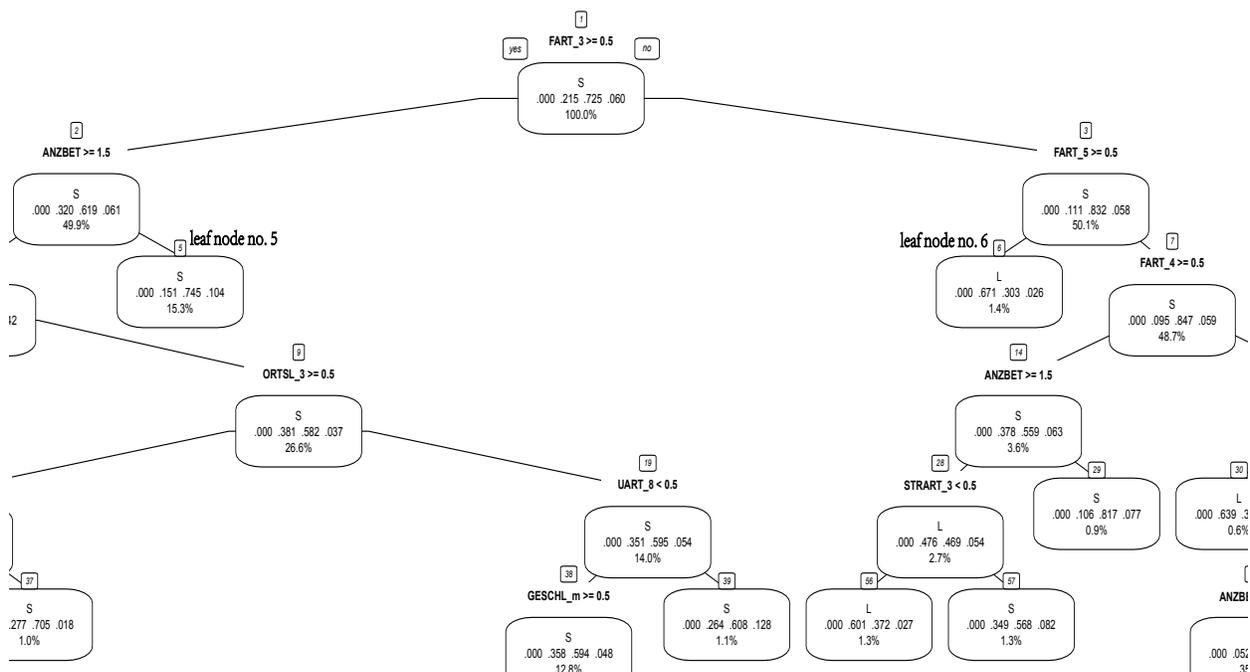


Figure 2. Decision Tree (part) based on GIDAS (1999-2012) and based on variables available for Swedish accidents.

From our experience it is advantageous to *binarize* nominal variables like **VEHICLE TYPE** (**FART** in GIDAS) or **ACCIDENT TYPE** (**UART** in GIDAS), i.e. to create for any value the variable can take a derivative variable, which only takes the values yes (1) or no (0). For **FART** this would lead to binary variables **FART_1**, **FART_2**, ..., **FART_K**, indicating for example by **FART_3=1** that the vehicle of the injured person is a passenger car. A freely available software to create such a decision tree is the **rpart** package in R (cf. [11], [12] and [14]).

Having such a decision tree, based on GIDAS accident data, at hand, the extrapolation now is as follows. Any person involved in a severe accident in the target region will be sorted according to the classification rules of the decision tree to a specific leaf node. Since we do not know the injury outcome of the cases for the target region we replace the injury outcome for the target by the injury distribution of the corresponding leaf node obtained from GIDAS accidents. Having done so we finally obtain the overall injury distribution for the target region as a weighted sum over the injury distributions within all leaf nodes. The distribution within a leaf node stems completely from GIDAS accidents, but the weights over the set of various leaf nodes stem from the target region. In order to obtain sound results it has to be assumed, that the injury distributions within a leaf node (for example single vehicle accidents in a rural area by night happened not on a motorway) does not vary much between GIDAS and the target region. In contrast, the relative frequencies of accidents belonging to the various leaf nodes (relative frequency of leaf nodes) may vary substantially between GIDAS and the target region. As already stated it has for example to be assumed, that the injury outcome for a person seated in a passenger car, which leaves the road by night in a rural area is comparable between GIDAS accidents and accidents in the target region, but the share of such accidents may be much higher or lower in the target region or country compared to GIDAS by what reason ever.

In the following we describe the decision tree methodology a bit more formally. As already said our investigations have shown that a reliable extrapolation via decision trees is only possible if we restrict to severe accidents and exclude non-injured people. For the extrapolation we therefore are interested to predict,

for a given target region or country, the conditional probability that a person in an accident suffers a specific injury severity given that the person is hurt at all and is involved in a severe accident. This probability is stated in equation (1).

$$P(\{PVERL = x\}|\{PVERL \geq 1, UKAT \geq 2\}) \quad , \quad x = 1,2,3 \quad (1).$$

Here and in the following we follow the German accident coding. PVERL stands for the injury severity of a person involved in an accident and PVERL can take the values $0=not\ injured$, $1=slightly\ injured$, $2=severely\ injured$ and $3=fatally\ injured$. UKAT stands for the accident category. $UKAT \geq 2$ indicates that the accident the person is involved in is *severe*, meaning that at least one person in this accident suffered a severe or fatal injury. Decision tree methods now lead to a decomposition of the conditional probability given in equation (1). Let us denote the leaf nodes of a decision tree by B_i , $i=1, \dots, I$, in which B_i stands for a specific subgroup of injured persons (e.g. injured persons in passenger cars involved in a severe single vehicle accident by night in a rural environment). Then we consider on the one hand the injury severity distributions within each B_i (each subgroup) $P(\{PVERL = x\}|\{PVERL \geq 1, UKAT \geq 2\} \text{ and } B_i)$ and on the other hand the distribution of injured people in severe accidents over the B_i , namely $P(B_i|\{PVERL \geq 1, UKAT \geq 2\})$. This immediately leads to the decomposition of the aforementioned conditional probability given in equation (2).

$$P(\{PVERL = x\}|\{PVERL \geq 1, UKAT \geq 2\}) = \sum_{i=1}^I P(\{PVERL = x\}|\{PVERL \geq 1, UKAT \geq 2\} \text{ and } B_i) \cdot P(B_i|\{PVERL \geq 1, UKAT \geq 2\}) \quad (2).$$

The first factor in equation (2), namely $P(\{PVERL = x\}|\{PVERL \geq 1, UKAT \geq 2\} \text{ and } B_i)$, is computed from the decision tree using GIDAS data only. The second factor, namely $P(B_i|\{PVERL \geq 1, UKAT \geq 2\})$, which is the distribution over the various leaf nodes of the decision tree, has to be derived from accident information of the target region or country. An essential assumption for the proposed method to work is that the injury severity distributions within the leaf nodes (subgroups) B_i do not differ much between GIDAS and the target region or country. However, the distributions over the set of leaf nodes (subgroups) may differ substantially between GIDAS and the target.

Now we are ready to apply and to evaluate the proposed extrapolation method for two European countries.

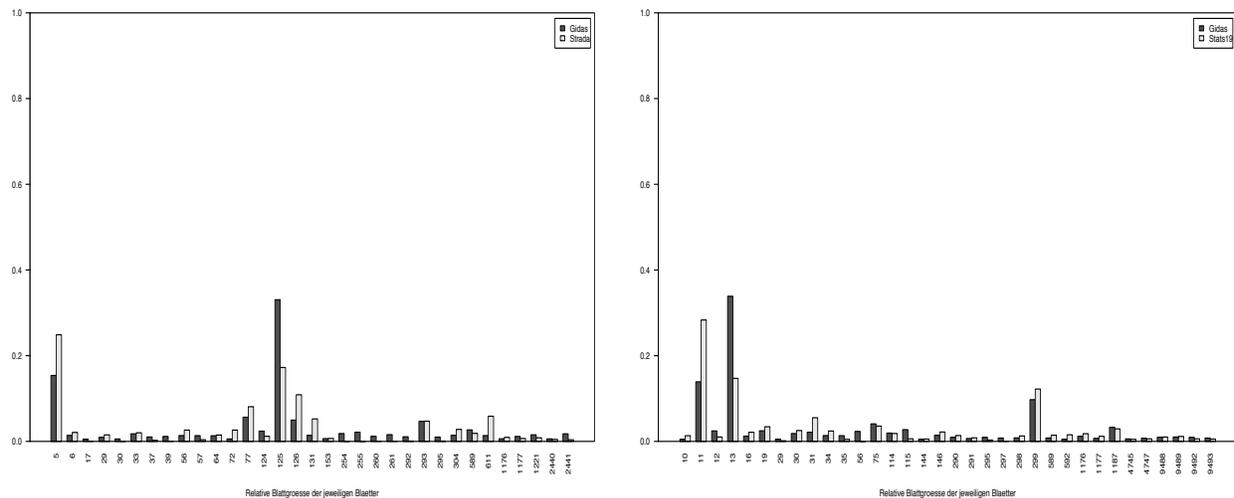


Figure 3: Relative Frequencies of the leaf nodes of the decision tree for GIDAS (1999-2012) (dark) and Sweden (2002-2012) (light) on the left-hand chart and for GIDAS (1999-2012) (dark) and UK (2004-2010) (light) on the right-hand chart.

EXTRAPOLATION RESULTS FOR SWEDEN

For the extrapolation to Swedish accident behavior we have made use of the following variables, which are provided by STRADA and successfully have been harmonized with variables within GIDAS.

- ACCIDENT TYPE (UART)
- TRAFFIC ENVIRONMENT (ORTSL)
- TYPE OF ROAD (STRART)
- TYPE OF VEHICLE (FART)
- NUMBER OF INVOLVED PARTIES (ANZBET)
- LIGHT CONDITIONS (TZEIT)
- SPEED RESTRICTIONS (VZUL)
- SEX (GESCHL)

A part of the obtained decision tree is given in Figure 2. A comparison of the distribution over the set of leaf nodes of the decision tree is displayed in Figure 3. It can be seen that for some leaf nodes the differences are quite substantial. Examples of leaf nodes with big differences are the leaf nodes with number 5, 125 and 126. Leaf node number 5 represents injured passengers in single vehicle accidents of passenger cars, while leaf node number 125 represents people not in passenger cars, lorries, busses or streetcars and injured in accidents that happened in urban areas with one or two parties involved in the accident. Finally leaf node number 126 represents injured people not in passenger cars, lorries, busses, streetcars or motorbikes and injured in accidents that happened not in urban areas. A typical case in the last two leaf nodes could be an injured pedestrian or bicycle rider in urban areas (leaf node 125) or in rural areas (leaf node 126). From Figure 3 it can be seen that in Sweden a significantly higher share of people is involved in single vehicle accidents of passenger cars. In contrast to these substantial differences a closer look at the injury distributions within the leaf nodes shows very slight differences between GIDAS and Sweden for leaf node number 5 and 125 and still acceptable differences for leaf node number 126.

The final extrapolation to Sweden can be seen in the left plot in Figure 4. The plot shows for each of the injury categories *slight*, *severe* and *fatal* three column-charts. The left-hand column represents the relative frequencies of GIDAS to the injury categories *slight*, *severe* and *fatal*. The differently colored segments of the columns show the corresponding contributions of the various leaf nodes. It can be easily seen for example that within GIDAS the largest leaf node, which has number 125 (cf. Figure 3) contributes much more to the category of severely injured people compared to the other two columns. The columns on the right-hand side display the obtained extrapolation for Sweden and the columns in the middle show for evaluation the true injury distribution in Sweden for the categories *slight*, *severe* and *fatal*. It can be seen that the obtained extrapolation matches the true situation in Sweden to a large extent and performs much better than a one-to-one extrapolation directly from GIDAS (left-hand columns) without adaptation to the different distribution over the set of leaf nodes in the decision tree.

It is emphasized that the proposed extrapolation not only performs well in representing the overall distribution over the injury categories *slight*, *severe* and *fatal* in Sweden, which corresponds to the total height of the columns in Figure 4, but that also the breakdown to the leaf nodes (subgroups) within the three injury categories, for example severe injuries, is close to the reality. Of course from this result it cannot be concluded that the same holds true for a further and deeper breakdown of the leaf nodes.

EXTRAPOLATION RESULTS FOR UNITED KINGDOM

For the extrapolation to accident behavior in UK we have made use of the following variables, which successfully have been harmonized with variables within GIDAS.

- TRAFFIC ENVIRONMENT (ORTSL)
- TYPE OF ROAD (STRART)
- TYPE OF VEHICLE (FART)
- NUMBER OF INVOLVED PARTIES (ANZBET)
- DIRECTION OF FIRST IMPACT (VDI 1)

- PEDESTRIAN
- SPEED RESTRICTIONS (VZUL)
- SEX (GESCHL)
- AGE OF CASUALTY (ALTER1)

A comparison of the distribution over the set of leaf nodes of the decision tree is displayed in Figure 3 (right-hand plot). It can be seen again that for some leaf nodes substantial differences occur. A look at the corresponding injury distributions within these leaf nodes does not show any relevant difference.

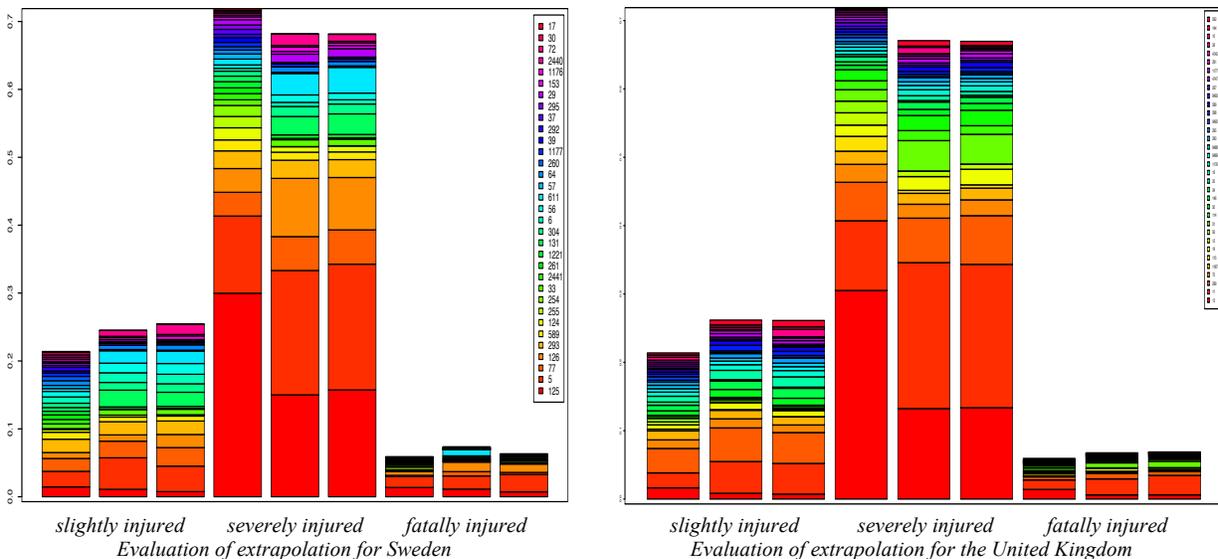


Figure 4: Column-chart of obtained extrapolation for Sweden (left) and for UK (right). Only injured people in severe accidents are considered. The three columns on the left in each chart refer to slightly injured people, the three columns in the middle refer to severely injured people and the three columns on the right refer to fatalities. Within each group of three columns the leftmost column represents the situation within GIDAS (1999-2012) (slightly, severely and fatally injured people in severe accidents within GIDAS), the column in the middle represents the actual situation within Sweden (2002-2012) (left-hand plot) and UK (2004-2010) (right-hand-plot), while the rightmost column represents the obtained extrapolation from GIDAS to Sweden or UK. The breakdown in each of the columns indicates the shares of the various leaf nodes of the decision tree to the specific injury severity. For example, the lowest segment in each of the columns for Sweden (left-hand plot) represents the contribution of the aforementioned leaf node 125 to the injury categories (slight, severe and fatal).

The final extrapolation to UK can be seen in the right-hand plot in Figure 4. The rationale behind the plot is the same as for Sweden. It again can be seen that the obtained extrapolation matches the true situation to UK to a large extent and again performs much better than a one-to-one extrapolation directly from GIDAS (left-hand columns) without adaptation to the different distribution over the set of leaf nodes in the decision tree. It is worth mentioning that again the break down to the leaf nodes (subgroups) within the three injury categories, for example severe injuries, is very close to the reality in UK. Of course from this result it cannot be concluded that the same holds true for a further and deeper breakdown of the leaf nodes.

EXTRAPOLATION OF EFFECTS OF FUTURE SAFETY SYSTEMS

This section is devoted to the application of the developed extrapolation methodology in order to get estimates for the effectiveness of future safety systems in vehicles. These are safety systems, which are either not yet on

the market or only have a small market penetration that effectiveness cannot be quantified from police recorded accidents. The specific question might be to estimate the number of fatalities, which possibly could be avoided if a particular safety system is on board of all vehicles on the roads. The idea is to take advantage out of the fact that for the detailed recorded accidents within the GIDAS database the effects of the safety system of interest can be quantified very well. This means, having the abilities of a safety system in mind, we evaluate all accidents within GIDAS once again and with respect to the injury outcomes for the involved people. In a next step we sort the newly evaluated accidents to the decision tree, which has been set up with the original GIDAS accident data. The following consequences may arise

1. An accident may completely drop out of the category of severe accidents because severe injuries for all involved people would have been prevented by the new safety system. All involved people drop out.
2. An accident stays within the category of severe accidents but part of the involved persons drop out because injuries would have been prevented by the safety system. This might change the injury distribution within some leaf nodes.
3. Injury severities of persons involved in an accident mitigate between the categories slight, severe and fatal.
4. All injury severities of all people involved in an accident stay unchanged.

Because of this, both the injury severity distributions within some leaf nodes of the decision tree as well as the distribution over the set of leaf nodes may change as a result of the safety system.

Based on a precise calculation of the effectiveness of a safety system for GIDAS accidents the extrapolation of the effectiveness to a target region is as follows:

1. For all leaf nodes, new injury distributions obtained from a safety system investigation in GIDAS have to be calculated.
2. The accidents from the target region will be sorted into the decision tree.
3. The change in the absolute case numbers over the set of leaf nodes in GIDAS caused by the safety systems will be extrapolated to the target and lead to adapted absolute case numbers over the set of leaf nodes.
4. As before, for each leaf node, we obtain changed absolute case numbers of slight, severe and fatally injured people. Accumulation leads to an overall extrapolation for the target.

A fictional example of extrapolation

Let us consider a single leaf node obtained from a GIDAS decision tree for which the absolute and relative injury severity distribution is given in the first two rows of Table 1. Let us fictitiously assume that a new safety function reduces the absolute number of injured people in this subgroup by 273 and that the reduction breaks down to the injury categories as given in the third row of Table 1. It is worth mentioning that on the one hand we observe a reduction of fatalities by 13 and on the second hand an increase of the relative frequency of fatalities among the group of injured people from 8.7% to 9.6%. Finally assume that for the target country we have 350 injured people within the considered leaf node (subgroup). As described before the extrapolation to the target is done by transferring the injury distribution within the considered leaf node from GIDAS to the target. This is of course done for the GIDAS injury distribution without and with the new safety system.

It is seen from Table 1 that we extrapolate that the future safety system reduces the number of fatalities within the considered subgroup (leaf node) by three. However, as in GIDAS, the relative frequency of fatalities increases. This phenomenon makes clear that in order to assess a safety function correctly absolute figures are vital.

If one carries through such an extrapolation for all leaf nodes of the decision tree one will receive an extrapolation of the effectiveness of the new safety system for the entire target region or country.

Table 1.
Injury distribution for a single leaf node of a GIDAS decision tree and
an extrapolation of a fictional future safety system.

	Slightly Injured	Severely Injured	Fatally Injured	Sum
Injury distribution GIDAS without new safety system	95	1,147	118	1,363
	7.0%	84.2%	8.7%	100%
Injury distribution GIDAS with new safety system	135	850	105	1,090
	12.4%	78.0%	9.6%	100%
<hr/>				
Extrapolation of injury distribution without new safety system to target	25	295	30	350
	7.0%	84.2%	8.7%	100%
Extrapolation of injury distribution with new safety system to target	35	218	27	280
	12.4%	78.0%	9.6%	100%

ITERATIVE PROPORTIONAL FITTING

For target countries with only sparse accident information the method of iterative proportion fitting (IPF) successfully can be used. Assume that for the target region only marginal distributions from accident databases are available. In contrast, for GIDAS, the full cross tables are available. The goal of IPF is to transfer the cross tables from GIDAS to the target taking into account all available accident information from the target country. For the validity of such an approach it is vital that the inner structure of GIDAS cross tables to a sufficient extent is close to the corresponding structure of the target country. The following subsection illustrates the situation for German accident data.

IPF: An example

Assume that only accident information for Germany as given in the margins (blue background) in Table 2 is available. The variable ANZBET denotes the number of involved parties (ANZBET=1, 2, 3 stands for one, two or three involved parties and ANZBET=4 stands for four or more involved parties in the corresponding accident). The variable ORTSL denotes the location of the accident (ORTSL=3 stands for urban areas and ORTSL=4 stands for rural areas). Finally UART denotes the accident type according to the German coding (e.g. UART=8 denotes an accident where the vehicle has left the road to the right hand side).

Table 2.
Accident information available from target (margins, blue)
and extrapolated via IPF from GIDAS (yellow).

ORTSL, ANZBET	3,1	3,2	3,3	3,4	4,1	4,2	4,3	4,4	
UART									
0	4.26%	1.77%	0.22%	0.06%	1.58%	1.45%	0.57%	0.27%	10.19%
1	0.04%	1.70%	0.44%	0.26%	0.05%	0.53%	0.06%	0.13%	3.21%
2	0.01%	1.23%	0.44%	0.18%	0.05%	3.69%	1.54%	1.47%	8.62%
3	0.02%	0.92%	0.19%	0.02%	0.02%	1.69%	0.38%	0.27%	3.50%
4	0.00%	3.60%	0.60%	0.12%	0.02%	9.77%	1.60%	0.45%	16.16%
5	0.11%	15.68%	1.00%	0.32%	0.02%	6.42%	0.89%	0.04%	24.49%
6	0.01%	8.54%	0.87%	0.09%	0.02%	0.70%	0.14%	0.03%	10.39%
7	0.29%	0.04%	0.00%	0.00%	0.27%	0.03%	0.03%	0.00%	0.66%
8	2.74%	0.28%	0.01%	0.06%	9.36%	0.91%	0.12%	0.06%	13.53%
9	1.63%	0.19%	0.01%	0.03%	5.83%	0.98%	0.47%	0.12%	9.24%
	9.10%	33.95%	3.79%	1.14%	17.22%	26.17%	5.81%	2.82%	100.00%

The method IPF allows to transfer the inner structure from the GIDAS cross-table, which is completely available, to the target region, taking into account the accident information contained in the margins of Table 2. The margins of GIDAS and the target may and typically will be different. The IPF algorithm is described in detail in [1], [6], [7] and [13]. For the validity of IPF it is vital that the interaction-effects or, equivalently, the cross product ratios (cf. [1]) between GIDAS and the target are similar.

Having the full Table 2 at hand a decision tree method based on the accident variables accident location, number of involved parties and accident type can be set up. It has been obtained that the extrapolation for Germany (for which all injury information is available) in the described example works very well and leads to reliable extrapolation.

EXTRAPOLATION TO EU COUNTRIES IN CASE OF LOW ACCIDENT INFORMATION

We present in this section an extrapolation from CARE accident records to the target country Austria for the year of 2008. Note that in this example only fatalities (broken down to accident location and vehicle type) for Austria are available. The final goal is to set up an accident decision tree which allows for extrapolation of GIDAS to Austria and applications as described in the section on extrapolation of the effects of future safety systems.

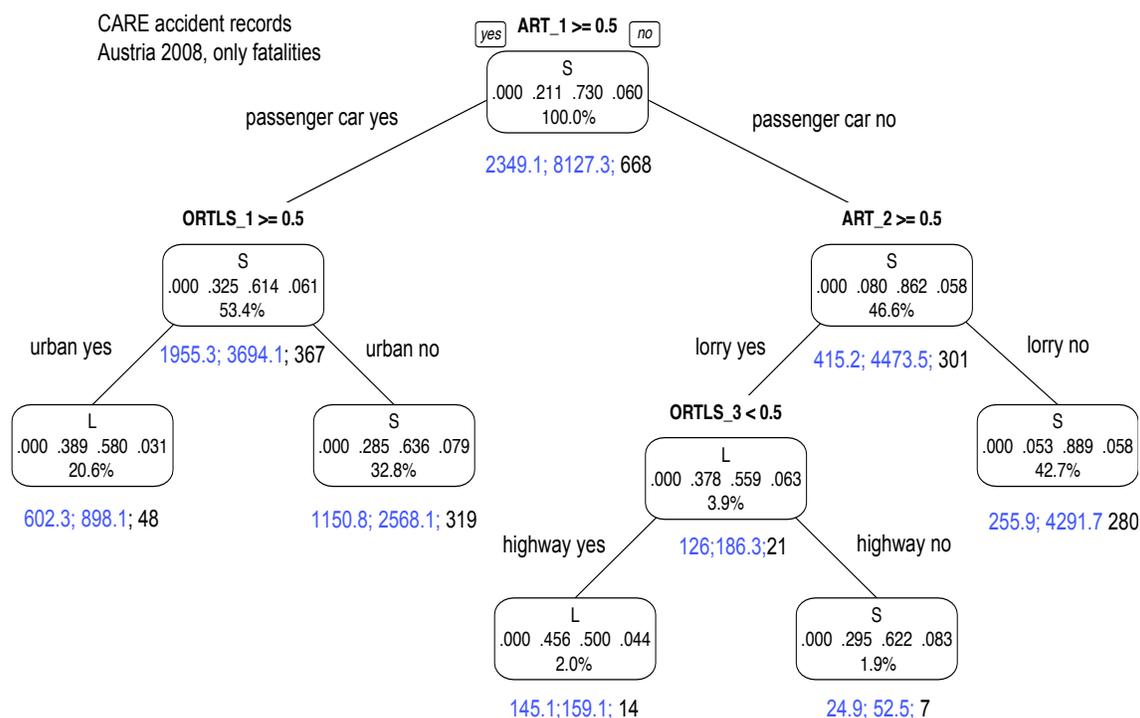


Figure 5: Accident decision tree based on GIDAS (1999-2012) with adaptation to Austria and based on fatalities, only. Below every box the fatalities for Austria (2008) (from CARE database (in black) and the extrapolated numbers of slightly and severely injured people (in blue) are displayed.

Based on GIDAS and the two variables (accident location and vehicle type) we can set up a decision tree, which is given in Figure 5. Below every box the known fatalities for Austria (in black) are given. According to the relative frequencies of fatalities (calculated from GIDAS) within each box one can extrapolate the absolute numbers of slightly and severely injured people involved in the accidents of each group. These extrapolations are given in blue color below each box. From this we obtain an estimate of the unknown distribution of injured people over the set of

leaf nodes for Austria. This information is necessary in order to extrapolate the distribution of injury severity within Austria. Recall that for such an extrapolation we make use of the injury distributions within the leaf nodes (they can be calculated from GIDAS only) and the distribution over the set of leaf nodes, which necessarily has to be calculated from accident data of the target country. Of course the low depth of data affects the extrapolation methodology in the form that the resulting decision tree is less complex than for more detailed accident information within the target country. Compare with the extrapolations for Sweden and UK described above.

CONCLUSIONS

Based on the extensive and detailed GIDAS accident database we have discussed in detail the decision tree method, a method for extrapolation from the GIDAS accident database to European regions or countries. It has been shown that in particular for severe accidents, that is, accidents with at least one severely or fatally injured person, GIDAS has good predictive power. The proposed methodology is able to take into account differences in the accidents between the survey area of GIDAS (Hanover region and Dresden region) and the target region, caused by different distributions of accidents over the various leaf nodes. For not severe accidents, i.e. accidents with uninjured or slightly injured persons, only, the proposed decision tree method reaches not the same quality as it does for severe accidents. This is to a substantial degree due to the fact that the category *slightly injured* in Germany, and probably in other countries as well, has a very soft and partially even diffuse definition. However, evaluation of the developed methodology with police recorded accident data for Sweden and the United Kingdom yields that the method successfully can be applied as long as only severe accidents are taken into account.

The key assumption for validity of the developed extrapolation methodology is that, with sufficient breakdown of accident scenarios, the injury severity distribution within each scenario no longer depends very strongly on the region, i.e. the injury severity distributions within a scenario for GIDAS and the target region are similar.

If a specific target region, for which one wants to extrapolate the injury outcomes from GIDAS, only a low data-depth of accident data is available, the proposed extrapolation methodology typically cannot be directly carried out. This is particularly the case, if for the target region only so-called marginal distributions from accident data are available. For this situation, a combination of the suggested methodology and the so-called iterative proportional fitting (IPF) leads to a reasonable extrapolation concept.

An important application of the proposed extrapolation methodology is, that an a priori assessment of effects of future safety systems can be carried out. Effects of future safety systems often can reliably be quantified for accidents reported in detail in GIDAS. The methodology developed then allows extrapolation of such a quantification of effects of future safety systems to target regions with possibly different accident behavior.

Finally, it is discussed to what extent an extrapolation is possible to target regions or countries with (very) low accident information. As an example this refers to regions or countries in which only fatalities broken down according to very few accident parameters (e.g. only accident location (rural/urban) and type of vehicle (passenger car, ...)) are present. It is shown that even in such a situation the decision tree method can be applied to estimate the effectiveness of (future) safety systems.

Summarizing, it has been shown that advantage can be taken out of a very detailed in-depth accident database within a limited survey area (like GIDAS) for extrapolating accident outcomes to other regions and countries. Thus GIDAS in a sense successfully can be enabled to describe severe accidents in other regions of Europe. Further it is shown that the developed extrapolation method based on field accident data leads to benefits in assessment methods for safety systems in vehicles.

ACKNOWLEDGEMENTS

We are grateful to Swedish Transport Agency for providing a very extensive database of police recorded accidents in Sweden. From the Federal Highway Research Institute we have received an accident database for the United Kingdom and access to CARE accident data.

We benefitted a lot from numerous discussions and suggestions from various colleagues obtained on a sequence of extrapolation workshops to be held at various places in Germany.

REFERENCES

- [1] Agresti (2013). „Categorical Data Analysis.“ 3rd edition. John Wiley & Sons, NJ: Hoboken.
- [2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). „Classification and Regression Trees.“ Belmont, CA: Wadsworth.
- [3] Breiman, L. (1996). „Bagging predictors.“ *Machine Learning* 24, 123–140.
- [4] Breiman, L. (2001). „Random Forests.“ *Machine Learning* 45, 5–32.
- [5] Bühlmann, P. (2004). „Bagging, Boosting and Ensemble Methods.“ In *Handbook of Computational Statistics: Concepts and Methods* (Eds.: Gentle, J., Härdle, W. and Mori, Y.), 877–907. Springer, New York.
- [6] Deming, W.E. and Stephan, F.F. (1940). „On a least squares adjustment of a sampled frequency table when the expected marginal total are known.“ *The Annals of Mathematical Statistics*. 11, No. 4, 427–444.
- [7] Fienberg, S.E. (1970). „An iterative procedure for estimation in contingency tables.“ *The Annals of Mathematical Statistics*. 41, No. 3, 907–917.
- [8] Hautzinger, H., Pfeiffer, M. und Schmidt, J. (2006). „Hochrechnungen von Daten aus Erhebungen am Unfallort. Berichte der Bundesanstalt für Straßenwesen.“ *Fahrzeugtechnik*, Heft F 59.
- [9] Hothorn, T., Hornik, K. and Zeileis, A. (2006). „Unbiased Recursive Partitioning: A Conditional Inference Framework.“ *Journal of Computational and Graphical Statistics* 15, 651–674.
- [10] Kreiss, J.-P., Feng, G., Krampe, J. Meyer, M. and Niebuhr, T. (2014). „Methodik zur Hochrechnung von GIDAS-Daten auf Europa“, Bericht Bundesanstalt für Straßenwesen.
- [11] Milborrow, S. (2014). rpart.plot: Plot rpart models. An enhanced version of plot.rpart. R package version 1.4-4. <http://CRAN.R-project.org/package=rpart.plot>.
- [12] R Core Team (2014). „R: A language and environment for statistical computing.“ R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [13] Stephan, F.F. (1942). „An iterative method of adjusting sample frequency tables when expected marginal totals are know.“ *The Annals of Mathematical Statistics* 13, 166–178.
- [14] Therneau, T., Atkinson, B. and Ripley, B. (2014). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. <http://CRAN.R-project.org/package=rpart>.