An extrapolation method on European accident data based on weighting and data harmonization

Albine Chanove*[1], Maria Pohle[1], Martin Urban[1], Jorge Lorente Mallada[2]

[1]Fraunhofer Institute for Transportation and Infrastructure Systems,Zeunerstr. 38,01067 Dresden, Germany, albine.chanove@ivi.fraunhofer.de
[2]Toyota Motor Europe, Research & Development,Hoge Wei 33,1930 Zaventem, Belgium

Abstract

The validation of the safety performance of Advanced Driver Assistance Systems (ADAS) and highly automated driving functions (AD) is a main objective for their introduction. On this topic, a methodology is used to create simulation files of the pre-crash phase of accidents from police-recorded accident data: the resulting dataset includes various information (e.g., participant types, participant trajectories and speed profiles). These simulation files allow the reconstruction of the crash scene and pre-crash phase as well as the assessment of the effectiveness of ADAS. However, this dataset is only based on police-recorded accidents from Saxony, Germany. Therefore, this paper focuses on developing an extrapolation method, in order to transform the database to the characteristic accident situation on a macroscopic scale. For instance, one aspect may be to assess the effectiveness of a newly-developed safety system at a European level, based on pre-crash simulation files. The methodology starts with a data review to link the simulation files with European accident data, then the extrapolation based on weighting factors is explained. It requires to find common variables between the two datasets, group the data by these variables and calculate the weighting factors. Due to the data difference and data categorization in countries' accident statistics, the grouping and consequently a direct extrapolation are not possible: an in-between database must be created, which contains harmonized data of police-recorded accidents. This allows to group the data with homogenous variables and especially identical accident constellation categories. In addition, the process enables to extrapolate a small dataset of police-recorded accidents to the European level, or certain countries. It provides a method for the calculation of weighting factors, by defining reproducible requirements for the input data and for the variable groups to determine the weighting factors between each dataset.

Keywords: extrapolation, harmonization, European data, police data, weighting factor, accident data

[1] * Corresponding author. Tel.: +49-351-4640-8757;
E-mail address: albine.chanove@ivi.fraunhofer.de

1. Introduction

The validation of safety performance is a main objective to introduce Advanced Driver Assistance Systems (ADAS) and highly automated driving functions (AD). On this topic, a newly-developed methodology creates simulation files of the pre-crash phase of accidents from police-recorded accident data [1, 2]. Participant trajectories are generated by using the information recorded by the police, the accident description, and the aerial image of the accident site. Speed profiles and breaking distances are added through an in-depth analysis of homogenous accident constellations, which define the mean values and standard deviation of different dynamic parameters such as initial speeds or braking decelerations. The resulting dataset called TASC (Traffic Accident Scenario Community) includes information on the accident, the participants, the injuries, and spatio-dynamic variables such as the trajectory or speed profiles. These data can be used to reconstruct the accident scene and pre-crash phase, to carry out an effectiveness assessment of ADAS, could have been beneficial to the participants.

Since the TASC dataset is only based on police-recorded accidents from Saxony, Germany, one important aspect is the transferability of the data to other regions or countries and thereby the extrapolation of the dataset for the purpose of evaluating ADAS or AD on different macroscopic scales with their characteristic accident situation [3]. For instance, the effectiveness of a newly-developed safety system may differ at an European level compared to a certain nation level.

The paper presents the different necessary steps to build an extrapolation method:

- Select the data source, on the level of which the extrapolation will be done. Requirements for these data sources are also developed here.
- Determine the mathematical method for the extrapolation. Following this method leads to the requirements on data cleaning.
    - Find common variables between both data sources. The selection of the variables must be a compromise between the available variables on both sides and their relevance for traffic safety research.
    - Group both data by these common variables. Limitations can appear, given the data structure and data availability of the data sources: a solution by using a harmonization method is given.
- Calculate the weighting factors, which allow to extrapolate the figures.

These steps are applied to the concrete case of the TASC data, with the aim of extrapolating the data on the European level, as set as initial expectation. Following them, the preliminary work consists of a review of the different European data sources, to select the one that best fits the TASC data. Then, following the mathematical process by choosing and grouping the TASC variables reveals limitations due to the existing data: the paper describes how to use a harmonization method to create an in-between database (shadow database), which role aims at having corresponding variables with both databases. Finally, a numerical application is performed on a practical case.

2. Methodology

2.1. Selection of the European data source

With the purpose of extrapolating a small data set to the European level, it is necessary to select the corresponding European data. The selection of the European source strongly relies on the initial data. The TASC dataset contains detailed data with two types of variables: the police recorded ones (such as location, injury severity, participant type, accident type), and the estimated ones (speed, point of collision, and trajectory). All accidents involve two participants, with at least one car, and at least one injured person. To propose a precise extrapolation, it is therefore necessary to find a database containing similar information, or at least the possibility of filtering according to the same conditions. At the European level, there are two major databases of police-recorded road accidents: the CARE database and the IRTAD database. The following Table 1 summarizes the main properties of these sources.

Table 1: CARE and IRTAD meta-data

	CARE [4]	**IRTAD [5]**
Geographical range	33 countries (EU)	33 countries (worldwide)
Number of accidents (per year)	1 M.	
Variables	~70 variables	~30 variables

Comments	- Only accidents with injured persons - Access to aggregated data only

After reviewing both codebooks and the meta-data, it appears that due to its wider range of variables, the CARE database offers better possibilities to group the data with precision with regards to the TASC data. The IRTAD database has for example no accident constellation variable. Moreover, its data are already aggregated in fixed tables: this is not necessarily a disadvantage, but it would require more work to prepare the data in the same way. The choice of variables in the CARE database is in this respect more flexible. Consequently, the CARE database is selected.

2.2. Mathematical process

The extrapolation method is based on weighting factors [6]. Once the level on which the extrapolation will be carried out is chosen, the next step is the explanation of the mathematical process for the extrapolation. In this case, the extrapolation method is based on weighting factors, since the distribution of accidents at the local level and at the macroscopic level is different. These factors are based on certain variables, which must be present in both databases. In other words, both databases are aggregated and grouped by these common variables. Then, the corresponding weighting factors are calculated for each group, following the equation (1).

$$wf = \left(\frac{Accidents_{group}}{Accidents_{group\_total}}\right) \Big/ \left(\frac{Accidents_{group\_EU}}{Accidents_{group\_EU\_total}}\right) \quad (1)$$

With:

$Accidents_{group\_EU}$ *Accidents per harmonized group in Europe (location, injury severity, accident constellation, etc.)*
$Accidents_{group\_EU\_total}$ *Accidents in total Europe*
$Accidents_{group}$ *Number of accidents per harmonized group in the local data source (location, injury severity, accident constellation, etc.)*
$Accidents_{group\_total}$ *Number of accidents in total in the local data source*

Once the factors are calculated, and for a selected harmonized group (a combination of location, injury severity, accident constellation, etc), each accident number from the local source is multiplied by the corresponding factor (2). The result is the expected extrapolation.

$$n_{extrapolated} = n_{original}.wf \quad (2)$$

With:

$n_{extrapolated}$ *Extrapolated number of accidents at the European level*
$n_{original}$ *Number of accidents in the local source*
wf *Corresponding weighting factor*

The next working steps are then determined by this mathematical process, which will be applied to the common variables, that have been identified in the databases. The data will be grouped consequently. A compromise has to be found between accurate extrapolation (i.e. a large number of variables), the availability of information and the allocated time. This work is detailed in the following parts.

2.3. Common variables and clustered information

Information is considered firstly only at the level of available variables. The data review at both accident and participant level allows finding common information and common variables. Even if TASC has more variables than CARE, some information is still on both side available, such as:

- Location type (urban or rural area),
- Road class (primary road, secondary road, etc.),
- Junction type (4-arms junction, T-shape junction, etc.),
- Speed limit of the road (in miles per hour or in kilometre per hour),
- Surface conditions (if the road was dry, slippery, wet, etc.),
- Light conditions (the daylight, the dawn or darkness of the night),
- Hit object (any “object” like animals, lost loading, an element of the road infrastructure, etc.),

- Special infrastructure (accident site located on a bridge, in a tunnel, etc),
- Traffic type (vehicle type),
- Participant type (role of the people involved like pedestrian, driver or bicycle),
- Injury severity (police-injury definition also known as time-injury definition),
- Accident constellation (defined here as the accident type or the participant manoeuvre, according to what is available in the database).

The selected variables are now reviewed at the level of their sub-variables and categories, respectively. Most of them have sub-variables with quite similar definitions. For example, the presence of a motorway is given in CARE as a separated variable. In TASC, the information for the motorway is contained in the variable "road class". However, the information differs for the accident constellation. TASC contains a very detailed accident type catalogue with 297 types of accident (following the Unfallforschung der Versicherer definition [7]), whereas CARE has only 61 different accident types (at the accident level), and about 20 participant manoeuvres (at the participant level). In addition, since CARE relies on the original country data definition, the data is not entirely provided: not all the countries register an accident type or a participant manoeuvre. Only 15 countries record an accident type, and 14 countries record a participant manoeuvre. In addition, CARE is only available in aggregated form, so detailed information per accident and per participant is not available. Working with only half of the countries is not a satisfying solution since it can taint the plausibility of the results. It is therefore not acceptable to extrapolate TASC directly to CARE.

Thereby, it becomes necessary to integrate a third source into the process, whose role is to match data between TASC and CARE. This source is called SHADOW and its range and construction is explained in the following.

2.4. The SHADOW data

A methodology [8] was previously developed and used here to create harmonized data from police sources. It consists of five main steps. Firstly, a state of the art on existing projects dealing with data harmonization in Europe is made, to have a look at the existing. The idea is to highlight the limits of actual processes and develop a method, which would overpass these limits. Then the data review allows to select some variables, on which a harmonization can be made. To do so, European police recorded accident databases are gathered, translated into English and their content is analysed. The next step is the development of the harmonization: by comparing identical and/or similar variables within the different data sources, new meta-variables can be clustered and defined, which keep the highest possible level of detail. Then the newly developed meta-variables are mapped with the original data. Finally, these meta-variables form the basis for a harmonized meta-database.

The SHADOW database is built following this method [8] and contains harmonized data of raw police-recorded road accidents from several European countries, as listed in the following Table 2. The countries are selected based on the availability and accessibility of their data: France, Germany, Great Britain and Spain. Working on police data allows access to the primary source of road death analysis, and has the advantage, in Europe, of presenting a common mode of data collection, and a general injury definition [9]. To be harmonized according to [8], the original data should be available under a non-aggregated form, with information at all levels: accident level, participant level, as well as participant serial number information. For each country, the data is translated by a native speaker into the English language, then is reviewed and harmonized.

Table 2: SHADOW content

Country	Source	Year for study	Number of accidents	Spatial range
Germany	Fraunhofer IVI [10]	2016	13.800	Saxony
France	Observatoire National Interministériel de la Sécurité Routière [11]	2016	59.600	France
Great Britain	Department for Transport [12]	2016	136.500	Great Britain
Spain	General direction of traffic [13]	2017	221	Spain

As a result, all data have identical variables and most of all, identical accident constellation categories: the method delivers a common harmonized manoeuvre classification, to which the TASC accident types can be matched. The harmonisation can be done only on some of the variables listed in 2.3, because of the variable availability. Now, SHADOW contains six harmonised variables: the location type, the road class, the junction type, the participant manoeuvre, the vehicle type and the injury severity. These harmonized variables especially match on one side

some of the TASC variables (e.g. the participant manoeuvre, the vehicle type and the injury severity) and on the other side, some of the CARE variables (e.g. location-type, road class and junction).

2.5. Calculating weighting factors

The weighting method allows calculating two ranges of factors: first, between TASC and SHADOW (e.g. the accident constellation, the vehicle type and the injury severity) at the participant level, and second, between SHADOW and CARE (e.g. location-type, road class and junction) at the accident level, as represented in the following Figure 1. The data is consequently grouped under these variable combinations. To ensure mathematical plausibility, accident data in SHADOW and CARE is filtered to address the same accidents as in TASC, which contains only two-participant accidents, involving at least one car and necessarily resulting in injuries. This is also represented on the following Figure 1 with the smaller boxes entitled “SHADOW 2” and “CARE 2”, respectively in the boxes SHADOW and CARE. The number of accidents for each of these accident collectives are respectively 3.926 accidents in TASC, 51.767 accidents in SHADOW 2 and 575.711 accidents in CARE 2.

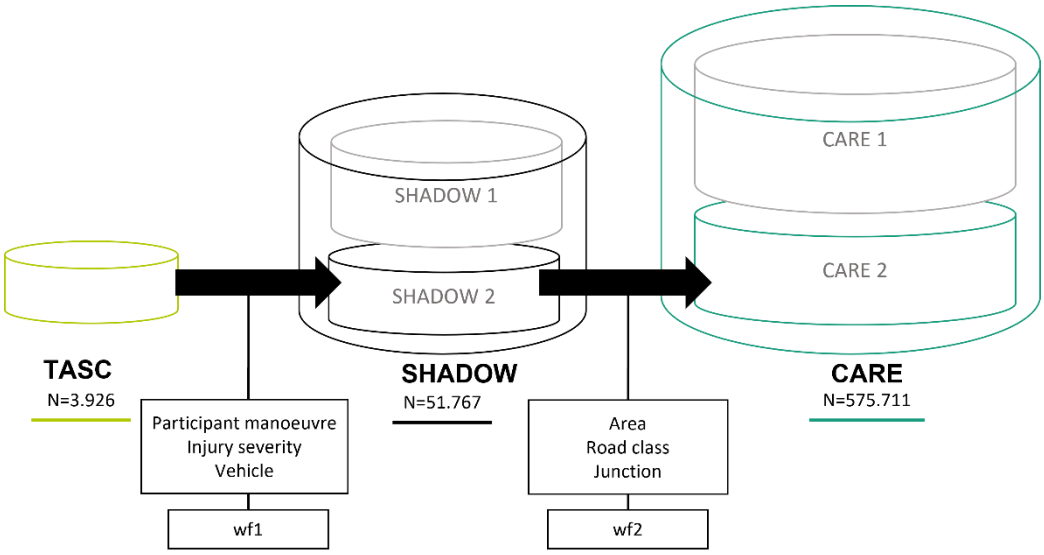

Figure 1: Two ranges of weighting factors between TASC, SHADOW and CARE

Having two ranges of weighting factors solve the issue of different information level, mentioned in 2.3, and still allows the compromises between detailed extrapolation and information availability. The mentioned weighting factor in the equation (2) is in this case the multiplication of the two weighting factors in Figure 1, following equation (3):

$$wf = wf1.wf2 \qquad (3)$$

With:

wf	*Overall weighting factor*
$wf1$	*Weighting factor between TASC and SHADOW*
$wf2$	*Weighting factor between SHADOW and CARE*

The next step is the numerical determination of the weighting factors: one for each variable group. The following Table 3 shows an example of six weighting factors for three different data groups. The accident constellation represents the combination of harmonized manoeuvres performed by the two participants. The severity of the accident is also noted in the same way. The first number is assigned to the first participant, the one who caused the accident, as defined by the codebooks. This notation makes it possible to work at the level of the participants and the level of the accident without losing information. The data in SHADOW are grouped per accident constellation, participant and accident severity, and the corresponding weighting factors are calculated. These factors are designated as the first range of factors. The data in CARE are grouped per location type, junction and road class, and similarly, the corresponding weighting factors are calculated. These factors are designated as the second range of weighting factors.

Table 3: Extract of the weighting factors

Participant manoeuvre	Vehicles	Accident severity	Weighting factor 1	Location type	Junction	Road class	Weighting factor 2
Range wf1				Range wf2			
Going straight VS going straight	Car VS bicycle	Not injured VS severely injured	5.43	Urban	On junction	Secondary road	4.26
Going straight VS going straight	Car VS bicycle	Not injured VS slightly injured	1.93	Urban	On junction	Secondary road	4.26
Going straight VS turning left	Car VS truck	severely injured VS Not injured	37.54	Rural	On junction	Secondary road	3.29

The factors are calculated in this case for the special case of the TASC data extrapolated on the CARE data. They can be used for different applications. In the following is an exemplary application presented.

3. Results

It is assumed that a new ADAS safety system is developed with the aim to avoid collisions between pedestrians and cars. Its effectiveness can be assessed with the TASC simulation files: after evaluation, it prevents around 400 car-to-pedestrian accidents in urban areas. The constructor of the ADAS system wants to know what share of accidents its system would prevent at the European level, since he wants to develop it on the European market. The developed extrapolation method can be here applied. The previously calculated weighting factors are filtered by vehicle type (car-to-pedestrian or pedestrian-to-car) and by location type (in urban area) to match the analysis. The pairs of weighting factors are multiplied to the TASC accident numbers (Equation 2). The sum of each combination (by considering the different combinations of manoeuvres, injury severity, junction types and road classes) is then the wished extrapolation result. It allows forecasting 3,500 possible prevented accidents on a European scale, from the 400 prevented accidents in the TASC database.

4. Discussion

This paper presents a method to extrapolate smaller data source to the European level. One advantage of the method is that the method is applicable and reproducible to any smaller data: the so-called TASC data in the paper could be any other accident source, like German federal police source also an in-depth data base like GIDAS, as long as it follows the requirements of the necessary variables. It solves the challenge of having different data sources and different variables. On this topic, the example on the TASC source showed how the method covers the definition issue between the TASC and the CARE accident types, thanks to the harmonization and the SHADOW source.

The extrapolation method can also be used on other research topics and thereby enable data projections on special questions. For example, a particular accident data source focusing on accidents with e-scooters in a given city could also be used as entry data source (replacing in the paper the TASC data), so that a first extrapolation could be made at the EU-level.

The method also supports accident analysis and accident comparison on a European scale by being able to compare data from different countries. In doing, it supports the effort in Vision-Zero.

To be applicable, the data must however follow the first requirements: both the one to extrapolate and the one to which the extrapolation aims to be performed must have enough equivalent information between both databases. In the case detailed in the paper, as CARE is built on the databases from police sources, the database to be

extrapolated has to preferably follow this structure. In the case of a more "in-depth" database, it will be necessary to ensure that the level of information in CARE can be used again. In any case, some adaptation work will be necessary.

Secondly, the results depend on the harmonization method. The core of the harmonization method is the way in which the harmonized variables are created, which also influences the results. It provides a standard that reduces differences in information, depending on the availability of variables in each country: this implies that the amount of data to be harmonized and therefore extrapolated is limited.

Finally, the results depend on the consistency of the databases used to build the SHADOW. The more representative data it contains, the better the results. Currently, the SHADOW database is composed of 4 countries, including only one region for Germany and only a random extract of 200 accidents for Spain. A future perspective is to work on consolidating the content of SHADOW and use the same years for each dataset.

5. Conclusions

The development of this method by using weighting factors allows the extrapolation on macroscopic scale of smaller datasets like TASC, used for the assessment of safety performance for different ADAS and AD. It uses reproducible requirements for the input data and for the variable groups to determine the weighting factors between each dataset. Such a method supports data difference, by considering the data and the meta-data. It can extend the use of the already-existing European databases, such as CARE and IRTAD. Finally, it enables data projection on special question and supports accident analysis on a European scale towards Vision-Zero.

6. Acknowledgment

Special thanks to the experts of the CARE and IRTAD groups, who helped to understand the data with which this work was carried out.

7. References

1. M. Urban, et al., A methodology for building simulation files from police recorded accident data (for ADAS effectiveness assessment). FISITA Conference, Prague, 2020.
2. C. Erbsmehl, TASC-Scenarios, SafetyUpdate Conference, Würzburg, 2020.
3. S. Alvarez, et al., Prospective effectiveness assessment of adas and active safety systems via virtual simulation: a review of the current practices, ESV Conference, 2017.
4. European Commission, CADAS glossary, 2018 edition ,CARE database, 2018.
5. IRTAD Group, The International Traffic Safety Data and Analysis Group. OECD, 2020.
6. AUDI AG, BOSCH GmBH, DAIMLER AG, VOLKSWAGEN AG, Unfalldatenanalyse, GIDAS-Wirkfeldanalyse ausgewählter sim-Anwendungsfälle zur Darstellung eines maximal anzunehmenden Wirkfeldes. 2009.
7. J. Ortlepp, P. Butterwegge, Unfalltypen-Katalog: Leitfaden zur Bestimmung des Unfalltyps. Unfallforschung der Versicherer, Berlin, 1998.
8. Chanove, A.; Erbsmehl, C.T.; Landgraf, T.; Urban, M.: A Method to Harmonize Accident Databases Between Different Countries. 9th International Expert Symposium on Accident Research ESAR 2021, virtuelles Event, 23.-24. März 2021, Vortrag: A. Chanove
9. D. Adminaite, G. Jpst, H. Stipdonk, H. Ward, An overview of road death data collection in the EU: PIN Flash report 35. European Transport Safety Council, 2018.
10. PHK Laskosky, Datensatz Beschreibung. Statistisches Bundesamt, Polizei Sachsen, 2017.
11. ONISR, Guide de redaction du Bulletin d'Analyse des Accidents Corporels de la Circulation. Ministère de l'intérieur, Paris, 2017.
12. Department of Transport, Instructions for the Completion of Road Accident Reports from non-CRASH Sources: STATS20. Department of Transport, 2011.
13. Ministerio del Interior, Boletín official del estado: Disposición 12411 del BOE núm. 289. Ministerio del Interior, 2014.