H. Hautzinger, M. Pfeiffer, J. Schmidt
Institut für angewandte Verkehrs- und
Tourismusforschung e. V., Mannheim, Germany

Expansion of GIDAS Sample Data to the Regional Level: Statistical Methodology and Practical Experiences

1 Introduction

Data concerning accidents involving personal injury which have been collected in the context of in-depth investigations on scene in the Hannover area since 1973 and in the Dresden area since 1999 represent an important basis for empirical traffic safety research. At national and international level various analyses and comparisons are carried out on the basis of "in-depth data" from the above mentioned investigations. In-depth data play a decisive role e.g. within the validation of EuroNCAP results on secondary safety (crashworthiness) of individual passenger car models. Thus, statistically sound methods of data analysis and population parameter estimation are of high importance.

Since the 1st of August 1984 the "in-depth investigations on scene" in the Hannover area have been carried out according to a sampling plan developed by HAUTZINGER in the context of a research project on behalf of BASt. In the meantime a second region of in-depth investigation on scene was added with surveys in Dresden and the surrounding area. Internationally, the acronym GIDAS (German In-Depth Accident Study) is commonly used for the two above mentioned surveys.

The objective of a current research project (topic of this contribution) is, among other things, to examine and adjust the previous weighting and expansion method for the two regional accident investigations to the current general conditions.

2 The Project GIDAS: In-depth Investigation on Scene in the Hannover and Dresden Areas

2.1 Investigation Methodology

One of the main characteristics of in-depth accident investigations is that the research team

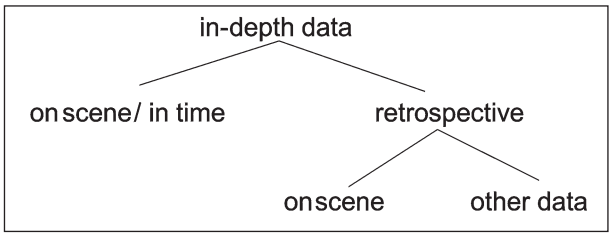

Fig. 1: Categorisation of in-depth data on road traffic accidents

arrives on scene and starts collecting the accident data immediately after having been alarmed by the police, rescue services, or fire department headquarters ("on scene" and "in time"). Apart from the acquisition of accident data on site (gathering information, taking pictures, etc.), the process of data collection also covers additional phases like the interrogation of witnesses or the collection of data at different places (e.g. hospitals, scrap yards). Basically, in-depth data can also be collected exclusively in retrospect by ex post examination of the spot of accident ("on scene") or by gathering relevant data solely at other places.

Accident investigation in the Hannover and Dresden areas takes place daily during two six-hour time intervals (so-called shifts) following a 2-week cycle. During the first week data collection is carried out from 12:00 a.m. to 06:00 a.m. and from 12:00 p.m. to 06:00 p.m. and within the second week those accidents are documented that occur during the other two intervals (06:00 a.m. to 12:00 p.m. and 06:00 p.m. to 12:00 a.m.). Thus, the premise for the acquisition of accident data is that the accident occurs within the respective time interval and within the demarcated investigation area. In any case, however, only accidents involving personal injury are taken into consideration.

Within the shifts, the first reported accident involving personal injury is recorded by the team and subsequently all other accidents. Due to the fact that data acquisition on scene takes about one hour per accident, overlapping of accidents is possible. In this case, the most current accident after reestablishment of the operational readiness is registered.

2.2 GIDAS Survey Plan from a Sampling Theory Point of View

2.2.1 Target Population and Sample

From a sampling theoretical point of view the target population consists of all police-recorded accidents involving personal injuries which occur in

the Hannover and Dresden areas. Accidents which are reported neither to police nor to the rescue services do, strictly speaking, not belong to the target population, since they are not included in the official accident statistics and, therefore, cannot be considered in the expansion factor.

The sampling units (i.e. accidents) can be seen as "events" occurring in time and space. Therefore, at the beginning of the survey period there is no list containing all the elements of the target population which could serve as a sampling frame. This kind of target population can also be referred to as a "bulk of events". Furthermore, neither the annual sample size nor the size of the target population are known in advance.

2.2.2 Selection of Time Clusters as Primary Units

The GIDAS sampling plan for the acquisition of accident data corresponds to a two-stage sampling procedure. The first stage is to randomly select time intervals as primary units (primary selection). With respect to the in-depth investigations, the primary units correspond to time clusters of accidents which are defined as follows:

For each calendar week exist – due to organisational reasons – the following two basic types of survey intervals (each of length 7x12=84 hours):

Type A: daily between 12:00 a.m. and 06:00 a.m. and between 12:00 p.m. and 06:00 p.m.

Type B: daily between 06:00 a.m. and 12:00 p.m. and between 06:00 p.m. and 12:00 a.m.

Consequently, based on one legal year, there exist 104 primary units.

Over the year, the time clusters according to the two basic types are being selected alternately, i.e. 52 out of 104 primary units are chosen. Thus, the selection of primary units can be regarded as a systematic sample with sampling interval 2. Due to this procedure all parts of the year are equally covered by the sample. For this reason – i.e. in view of the way the primary units are selected – systematic random sampling is superior to simple random sampling of time intervals.

Assuming perfect preconditions, that means that within each selected survey time interval (i) all police-recorded accidents are being reported to the investigation team and (ii) all reported accidents are being registered by the investigation team, the GIDAS survey method corresponds to one stage systematic cluster sampling with sampling interval 2. These ideal preconditions are, however, not given in practice: on the one hand not all police-recorded accidents are being reported to the investigation team and on the other hand not every reported accident can be registered by the investigation team. Thus, a sampling procedure for the second stage, i.e. for selection of accidents within the selected time intervals (shifts) is needed.

2.2.3 Selection of Accidents as Secondary Units

With regard to the selection at the second stage, special emphasis has to be given to the documentation of as many accidents as possible. For this reason the first reported accident (involving personal injuries) of a selected time cluster has to be documented and after that all other reported accidents if the team is ready for operation. Due to the fact that the target population generates itself in the context of a randomised process, this selection method can also be referred to as random. The current alarming system, however, does not have an absolute random character in the sense that all survey units (accidents) have the same selection probability: by classifying the survey units according to accident severity (accident with slightly injured persons, with seriously injured persons, with persons killed) it becomes evident, that not all accidents have the same probability of being included in the sample, which essentially can be attributed to the alarming system. From a statistical point of view, inclusion in the sample depends on the results of two subsequent random experiments.

On the basis of a first random experiment it is determined whether or not a police-recorded accident is reported to the survey team. In case of an incoming report a second random experiment determines whether or not an accident will be registered by the team. The accident will be documented either if at the corresponding point in time the team is ready for operation or if the reported accident is the most recent reported accident after reestablishment of operational readiness of the team:

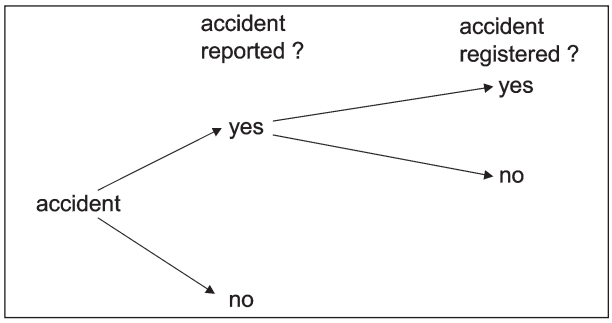

Figure 2

Presuming for a first analysis of the Hannover data from the year 2000, that the severity of accident consequences is the only determining factor for the reporting or non-reporting of an accident, the following estimates of the reporting probability are obtained:

- accident with slightly injured persons: 34.3%
- accident with seriously injured persons: 66.3%
- accident with persons killed: 75.0%

In contrast to accidents with slightly injured persons, accidents with persons killed or seriously injured are significantly more frequently reported to the investigation team. Certainly, a reporting rate of 100% would be ideal.

The following proportions show to which extent the reported accidents are actually documented:

- accident with slightly injured persons: 84.5%
- accident with seriously injured persons: 88.4%
- accident with persons killed: 96.3%.

The ideal case would be given if the documentation rates were identical. That means, even on the assumption of a perfectly working alarming system simple random sampling of secondary units (i.e. accidents on the second stage of the sampling procedure) can not be assumed on the basis of the empirical data from the year 2000.

All in all the following estimates for the selection probabilities can be obtained:

- accident with slightly injured persons: 29.0%
- accident with seriously injured persons: 58.6%
- accident with persons killed: 72.2%

While an accident with persons killed is selected with a probability of over 70%, the selection probability for an accident with slightly injured persons is only about 30%.

3 Weighting Procedure

The previous analyses of selection probabilities show that the raw GIDAS sample is biased at least with respect to severity of accidents. Therefore, a weighting procedure is needed in order to adjust or correct for this bias. The variables used for this fitting process should be those which correlate

Kind of accident	2000			2001		
	GIDAS unweighted	GIDAS weighted	official statistics	GIDAS unweighted	GIDAS weighted	official statistics
	accidents in %					
Collision with another vehicle which starts, stops or is stationary	5.3	5.9	9.8	6.6	6.8	9.4
Collision with another vehicle moving ahead or waiting	13.2	13.6	19.2	12.6	13.4	18.8
Collision with another vehicle moving laterally in the same direction	3.7	3.8	4.4	5.3	4.6	4.3
Collision with another oncoming vehicle	6.3	5.9	5.4	6.9	6.9	6.0
Collision with another vehicle which turns into or crosses a road	31.7	33.2	27.2	31.0	33.9	28.2
Collision between vehicle and pedestrian	12.6	11.7	9.5	10.8	10.6	9.4
Collision with an obstacle in the carriageway	0.2	0.2	0.4	1.0	0.8	0.5
Leaving the carriageway to the right	9.6	8.5	6.0	10.1	8.9	6.8
Leaving the carriageway to the left	7.1	6.6	4.8	7.3	5.7	4.7
Accident of another kind	10.3	10.4	13.4	8.4	8.5	12.0
Total	100	100	100	100	100	100
Chi^2 goodness-of-fit	106.1	79.8	-	79.6	52.7	-

Tab. 1: Accidents involving personal injuries in the Hannover area by kind of accident (year 2000 and 2001)

highly with as many as possible other accident characteristics. At present, the GIDAS weighting procedure is based on three characteristics which cover factual and spatial as well as temporal aspects of accidents:

- severity of accident (accident with slightly injured persons, with seriously injured persons, with persons killed)
- locality of accident (within built-up area yes/no)
- time interval of accident occurrence (12:00 a.m. - 06:00 a.m./06:00 a.m. - 12:00 p.m./12:00 p.m. - 06:00 p.m./06:00 p.m. - 12:00 a.m.)

The weighting procedure consists of a simple adjustment of the above three-dimensional contingency table to the corresponding table from official accident statistics of the respective survey area.

Although a three-dimensional distribution is used for weighting, it was found that the accuracy of some of the estimates is not satisfactory. As an example, in table 1 the weighted and unweighted GIDAS estimates for the Hanover area are compared to the official statistics with respect to the variable “kind of accident”.

The table shows that in each of the two years under investigation the weighted distribution is very similar to the unweighted one, although the fit of the weighted distribution to the official statistics is slightly better. Nevertheless, the effect of the weighting procedure is relatively small. The two distributions of “kind of accident” obtained from GIDAS data, however, differ considerably from the official accident data. In all cases the null hypothesis of equality of GIDAS and official distribution can be rejected since the empirical Chi^2-values are far beyond the critical value of 16,9 (level of significance: 5%).

Due to this results and the above mentioned bias with respect to accident severity it was decided to develop and test an alternative weighting procedure based on the following two variables:

- severity of accident (accident with slightly injured persons, with seriously injured persons, with persons killed) and
- kind of accident (10 categories).

It was assumed that locality of accident is strongly correlated to accident severity and, therefore, using one of them in the weighting procedure might be sufficient. Moreover, it was hoped that considering the variable “kind of accident” in the

Locality of accident	2000			2001		
	GIDAS unweighted	GIDAS weighted	official statistics	GIDAS unweighted	GIDAS weighted	official statistics
	accidents in %					
Within built-up area	72.8	75.5	75.7	69.7	73.2	75.8
Outside built-up area	27.2	24.5	24.3	30.3	26.8	24.2
Total	100	100	100	100	100	100
Chi^2 goodness-of-fit[1]	4.8	0.02	-	19.0	3.3	-
[1] critical value 3.84 (level of significance: 5%)						

Tab. 2: Accidents involving personal injuries in the Hannover area by locality of accident (year 2000 and 2001)

Time interval	2000			2001		
	GIDAS unweighted	GIDAS weighted	official statistics	GIDAS unweighted	GIDAS weighted	official statistics
	accidents in %					
12:00 a.m. - 06:00 a.m.	3.3	2.8	6.1	4.1	3.3	5.0
06:00 a.m. - 12:00 p.m.	34.6	35.1	29.0	28.9	29.5	29.4
12:00 p.m. - 06:00 p.m.	40.4	41.1	43.8	41.1	42.4	43.5
06:00 p.m. - 12:00 a.m.	21.6	21.0	21.1	26.0	24.8	22.1
Total	100	100	100	100	100	100
Chi^2 goodness-of-fit[1]	27.6	33.7	-	9.4	8.0	-
[1] critical value 7.81 (level of significance: 5%)						

Tab. 3: Accidents involving personal injuries in the Hannover area by time interval of accident (year 2000 and 2001)

Light conditions	2000			2001		
	GIDAS unweighted	GIDAS weighted	official statistics	GIDAS unweighted	GIDAS weighted	official statistics
	accidents in %					
daylight	74.1	75.3	73.3	72.5	72.7	73.8
dawn	3.9	3.4	4.6	4.6	5.3	5.0
darkness	22.0	21.2	22.2	22.9	22.0	21.2
Total	100	100	100	100	100	100
Chi^2 goodness-of-fit[1]	1.2	3.9	-	1.7	0.6	-
[1] critical value 5.99 (level of significance: 5%)						

Tab. 4: Accidents involving personal injuries in the Hannover area by light conditions (year 2000 and 2001)

Kind of accident site	2000			2001		
	GIDAS unweighted	GIDAS weighted	official statistics	GIDAS unweighted	GIDAS weighted	official statistics
	accidents in %					
not stated	79.1	79.4	79.1	77.3	77.8	78.3
road crossing	7.0	7.3	6.9	7.2	7.3	6.8
junction	7.8	7.8	8.2	7.6	7.9	8.3
property gateway	2.1	2.1	2.2	2.3	2.3	2.5
gradient	0.8	0.7	0.9	1.3	1.2	1.1
curve	3.2	2.7	2.7	4.3	3.5	2.9
Total	100	100	100	100	100	100
Chi^2 goodness-of-fit[1]	4.1	2.1	-	22.4	5.2	-
[1] critical value 11.1 (level of significance: 5%)						

Tab. 5: Accidents involving personal injuries in the Hannover area by kind of accident site (year 2000 and 2001)

weighting process will compensate possible biases with regard to the temporal distribution of accidents. In tables 2 to 5 it is shown whether these expectations proved true for the GIDAS 2000 and 2001 data. Variables to be analysed are:

- locality of accident
- time interval
- light conditions and
- kind of the accident site.

Concerning locality of accidents, the new weighting procedure (by severity and kind of accident) is substantially improving the fit of the distribution. In both years the null hypothesis can be rejected for the unweighted data whereas a good representation of the target population can be obtained by using the weighted distributions. This is due to the fact that locality is closely linked to accident severity since the probability for a severe accident is lower within built-up areas.

Regarding the variable “time interval of accident”, the new weighting scheme yields no correction of the bias. In 2000, the fit is worse than in 2001. This is mainly caused by the failure of the alarming system. Especially in the stratum 12:00 a.m. – 06:00 a.m. where only a few accidents are occurring anyway, the reporting rate is lower than in the other strata.

In table 4 and 5 the results for the accident characteristics “light conditions” and “kind of accident site” are depicted. Concerning light conditions, the fit of both the weighted and unweighted sample distributions can be regarded as good. One can see that there is a slight coherence between light conditions and time interval since for both cases the fit of the weighted distribution is better in 2001 whereas the opposite is true in 2000.

Finally, the weighted distributions of the variable “kind of accident site” are again closer to the official statistics than the unweighted ones. This holds particularly for the 2001 data where the null hypothesis of equality can be rejected for the unweighted distribution but not for the weighted one.

4 Concluding Remarks

The results described above may be summarised as follows:

- The figures presented in tables 2 - 5 show that appropriate expansion and weighting procedures can substantially improve the accuracy of the data from in-depth accident investigations. Of course, the main objective of expanding in-depth data to the target population is to expand variables which are not included in the official accident statistics (e.g. AIS, EES, etc.). In the present paper variables which are contained both in the sample and in the official statistics have been analysed in order to check the goodness-of-fit of the weighted sample distributions.

- Usually, accident characteristics which are recorded by police are also collected by the in-depth investigation team. However, it might well be that these two different measurements do not yield the same results, e.g. if police assigns an accident to another kind of accident than the research team does. It is important to note that in any case the police recorded data (standard traffic accident reports) of the accidents in the sample have to be used for expansion purposes. Even if the data from the in-depth investigation team are more precise it would be incorrect to base the weighting factors on them because in this case some of the accidents in the sample would be assigned to the wrong stratum (according to the target population).

- As described above, the in-depth investigations in the Hannover and Dresden areas are based on a two-stage sampling process. Therefore, according to the principles of sampling theory it would be most natural to use a two-stage expansion methodology. That is to say, in a first step the secondary units (accidents) would be expanded to the primary units (parts of calendar weeks) and after that the expansion of the primary units would take place. Within the scope of the research project on which the present contribution is based such a methodology has been carefully developed and tested. It was found, however, that the theoretical advantages of this method compared to the simple weighting procedure described above are relatively small, especially, if one takes into account the complexity of the calculation process necessary to obtain the corresponding expansion factors.