Y.W.R. de Vries
TNO-Automotive Safety, Delft, The Netherlands

N.A. Jongerius
DAF Trucks, Eindhofen, The Netherlands

Method For In-Depth Traffic Accident Case-Control Studies

Abstract

Internationally, the need is expressed for harmonized traffic accident data collection (PSN, PENDANT, etc.). Together with this effort of harmonization, traffic accident investigation moves more and more in the direction of accident causation. As current methods only partly address these needs, a new method was set up. The main characteristics of this method are:

- Accident/injury causation (associated) factors can objectively be identified and quantified, by comparison with exposure information from a normal population.
- All relevant accident and exposure data can be included: human-, vehicle-, and environmental related data for the pre-crash, crash and post-crash situation (the so-called Haddon matrix). The level of detail can be chosen depending on interest and/or budget, which makes the method very flexible.

In this paper the accident collection and control group method are presented, including some of the achieved results from a pilot study on 30 truck accidents and 30 control locations. The data were analyzed by using cross-tabulations and classification-tree analysis. The method proved useful for the identification of statistically significant causational aspects.

Notation

N Number of virtual accidents

n_m Number of vehicles in main direction

n_o Number of vehicles in other directions

γ Environmental impact

M Maximum number of virtual impacts (set at 10.000)

A_m Allowed number of vehicles in the main direction

A_o Allowed number of vehicles in the other directions

N_i^A Number of vehicles of type i in the maximum allowed sample

N_i^V Number of vehicles of type i in the video sample of duration D

D Video duration in minutes

W_i Weight factor for vehicles of type i

W_{ij}^I Weight factor for the impact between vehicles of type i and j

L_k Available percentage of road type i

N_k^s Number of locations sampled for road type i

W_k^L Weight factor for the location i

Introduction

Traffic accident investigation is used more and more to address all cells of the Haddon-matrix (see Figure 1). The information from this matrix is used to deploy new activities in relevant areas.

Knowledge on primary safety and pre-crash aspects (i.e. avoiding accidents) requires information about accident causation. Data bases and methods that have been developed for accident causation studies (e.g. EACS, ETAC) up till now, belong to the case-series studies: cases are investigated and frequency counts give information on occurrence of possible risk factors in these accidents. The impossibility to relate these occurrences to reference data is a large drawback as will be shown later.

A new method has been set up, with the main aim to address the afore mentioned limitation. The main objectives of this paper are to discuss the principles of this newly developed method for an epidemiological study into accident causation, and to show the results of a first pilot analysis on truck accidents.

The work has been carried out by TNO, with the support from DAF Trucks, Scania Trucks, and the Dutch Ministry of Transport, Public Works and Water Management and the Dutch Ministry of Economic Affairs.

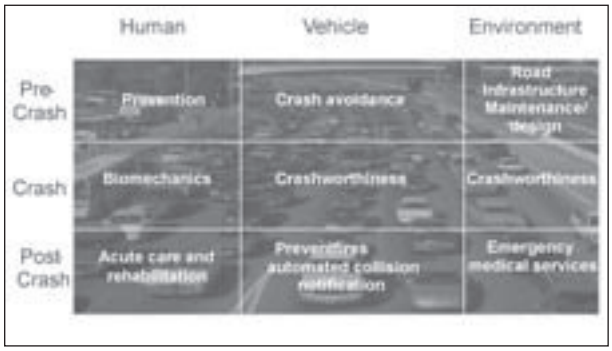

Figure 1: Example of the Haddon matrix

Exposure

Correct information about the cause(s) of accidents is relevant for policy makers and vehicle manufacturers. Identification of accident causes requires the acquisition of large amounts of data per accident, and a correct interpretation of this data (accident reconstruction, etc.).

Incorrect knowledge on accident causes may lead to the implementation of non-effective countermeasures. Such incorrect knowledge may be due to exposure effects or subjective assessment by investigators. High frequencies do not necessarily indicate a risk, but also show the amount of exposure. The investigator's subjectivity can originate from pre-determined ideas and feelings about what the normal operation should be under certain circumstances. These ideas are not necessarily correct and may lead to misjudgments. Therefore some kind of reference data is needed to correctly identify risk factors. A literature survey was carried out to find previously used study setups which made use of reference data. Possibilities for obtaining such observational data are the following:

Internal control groups
Internal control groups are groups of accidents for which a specific parameter is assumed not to have any influence. Differences in the presence or absence of the parameter in the studied group and the control group can indicate a relationship with accident occurrence. Two main problems exist: many cases have to be present and it has to be sure that the parameter under study has no effect in the control group.

Global indicators
Sometimes global indicators (kilometers driven, etc.) are used as reference data to indicate potential problems. However, the results are very dependent on the indicator that is used [ELSEVIER, 1997] and can be tuned with the use of an indicator which provides the results that are wanted. Furthermore, global indicators can not go into the detail needed for accident causation research (e.g. type of use).

Cohort study
Because traffic accidents have a relatively low occurrence, cohort studies are inefficient. Z group of drivers should be followed for a certain period. During this period some accidents should occur in this group. Drivers with and without accidents can be compared. The presence of accidents in this group is expected to be quite low when the group is not very large or the study duration is not very long.

Case-control study
When a case-series study is extended with the collection of some sort of control group, which can be used as reference data a case-control study design is obtained, from which associations between factors can be obtained.

Because in-depth research is already a case-series study, extension to a case-control study is therefore the most logical approach. A recent example is the European Motorcycle Accident In-Depth Study (MAIDS) [OECD, 1999], carried out in five countries, in which drivers were interviewed at gas stations. Another option sometimes used is to question drivers passing through the same scene one week after the accident. In both cases analysis on environmental factors was not possible because of the chosen method. Driver cooperation was also a problematic issue. However, both methods served as a basis for the newly developed method presented in this paper.

Method

Virtual accidents

In order to compare accidents directly with exposure information in a case-control study, the data need to be in the same format. Therefore one would like to obtain the control group from normal traffic situations, which can directly be compared with the accidents: some kind of "virtual accident" (every accident that could have occurred). Traffic intensity can be used as a measure. The number of "virtual accidents" for a given location can then be calculated by:

$$N = n_m \left((n_m - 1) + n_o + \gamma \right) \quad [1]$$

Each target vehicle (in this study a truck) has a virtual accident with every other vehicle passing through the scene. The number of virtual impacts is therewith frequency induced. The more other vehicles are present, the more virtual impacts. Each vehicle can have an impact with a vehicle in the same direction as well as with a vehicle from another direction or with an environmental object (see formula [1]). Environmental information of the location has to be coded with the virtual accident. The total of all virtual impacts in all monitored traffic situations will then serve as the control group.

A main problem with the virtual impact method is that the number of impacts increases quadratically with each extra vehicle in the main stream. Therefore the number of generated virtual impacts is limited to ten thousand per location. The maximum allowed number of vehicles for this maximum number of virtual impacts that has to be sampled for the main stream (A_m) and the other directions (A_o) is calculated according to formula [2].

$$f = \frac{n_m}{n_o} = \frac{A_m}{A_o}$$
$$M = A_m \cdot \left(\left(A_m - 1 \right) + A_o + \gamma \right) \Rightarrow$$
$$A_m = \sqrt{\frac{f}{f+1} M} \qquad [2]$$

This maximum allowed number of vehicles is sampled randomly from the video sample to acquire the distribution of vehicle types on a specific location.

Interaction model

In practice the traffic system is rather complex. Driver, vehicle and environment interact in unknown ways (see Figure 2a). All information needed for the control group should therefore be investigated at the same time.

The “virtual accidents” have to be collected in the same area in which the accidents are collected and should represent the normal traffic situations in that area. Therefore, the inspection of the locations should be completely randomized over the research area, such that it represents the conditions in the accident collection area. The samples should be taken equally over the duration of the study and at random times. The method for sampling the locations is shown later.

The traffic counts can be obtained from video, together with driven speeds, manoeuvers, distances to other road users, color, etc. Extra information can be obtained by license plate detection, coupled with vehicle registration information. From all the vehicles passing through the location the drivers should be observed and interviewed, and the vehicles should be inspected similar to the accidents investigations. This imposes a practical problem, because not all drivers can be stopped and interviewed in a monitored scene. Even if a sample could be taken, this means that needs to be on an involuntary bases, otherwise biases are introduced. In many occasions this is not possible. However, there is a way around this when some conditions and assumptions are met.

Model assumptions

Two possibilities now arise. For (semi) permanent physical conditions for which we expect that they have no relation with the environmental conditions (e.g. gender, illness, etc.) a control group could be gathered at any given location, because the conditions will be randomly distributed over the environment (not necessarily over the vehicles). Transient conditions which may have an interaction with or are induced by the environment are more complicated (attention diversion, using cruise control on motorways). Non-environment related issues can be investigated by interviewing people about the frequency of use or habits. Environmentally related habits can also be investigated on a more global level by asking about frequency of use under specific circumstances. This method suffers (less) from the same problems as global indicators (see Exposure).

The other option could be to form a cohort of random drivers whose behavior is recorded/logged under occurring circumstances. This option is more complicated and time consuming, but the result is likely better. For financial reasons the interviewing method is chosen for this study.

The driver information will be obtained at convenient locations which are sufficiently randomized. Drivers are interviewed and vehicles inspected.

The idea that is now used is the following (see Figure 2b): The driver interviews and detailed vehicle inspections are treated as missing values in the data from the video observations. These missing data are imputed from the separate vehicle

inspections and interviews on the video information. The random imputation is conditioned (matched) on variables collected in both inspections (so-called conditional random imputation). These values may not be treated as real values per accident but can be used to make appropriate inferences and to show statistical associations. The vehicle type distribution does not have to match the one found from the video observations and can deliberately be biased towards groups of interest to obtain the most useful information. The parameter for which the bias is introduced should also be measured from the video data (vehicle type, color, etc.). It would not lead to a biased sample, because the distribution of generated virtual accidents would stay the same, only with more or less details and statistical certainty for specific groups.

In future projects multiple imputation techniques can be used in order to improve the prediction of the missing values [RUBIN, 1987]. For this pilot study this method has not been used yet.

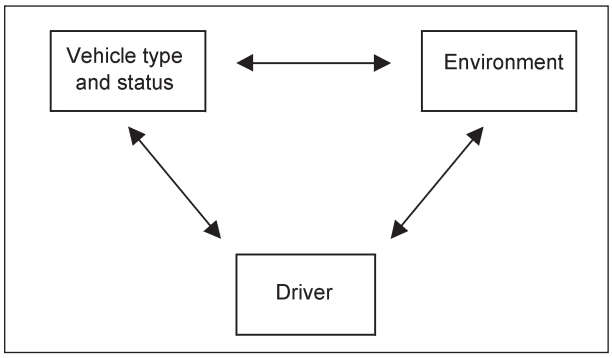

Figure 2a: Main interactions

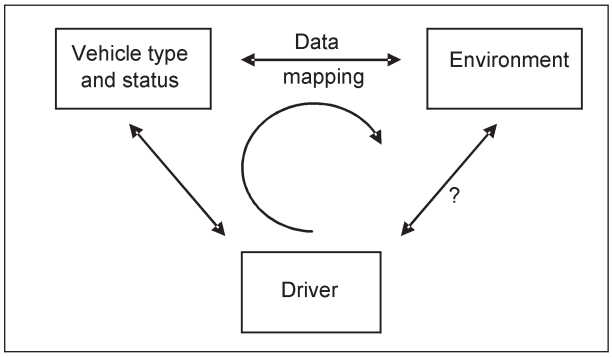

Figure 2b: Model assumptions and imputation

Practical Implementation

Selection of randomized locations

Information on all Dutch roads is available from a Geographic Information System (GIS) database (see Figure 3). Most other countries also have mapped their (main) roads into vector based files that are compatible with GIS [GIS, 2004]. The location selection has been split into crossings and segments without crossings. In order to sample all roads equally, all segments have been divided into 25 meter sections. Sampling the number of lanes instead of the number of roads is more appropriate, so that the traffic flow on one-way streets and multiple-lane streets are sampled just as much as on two-way streets.

For crossings the following approach is taken: Each lane into the crossing is counted as an intersection, because vehicles may come from all directions. Roundabouts are considered to be sets of T-crossings with two in-coming lanes and one out-going lane. This is the same as considering it to be one crossing, due to the fact that the number of manoeuvers is limited in the multiple T-crossing approach.

Figure 3: Example of a GIS-location and the low visibility video pole clamped to a lamp-post

A (disproportionate) stratified sampling plan may be used if specific road types are of more interest to the researcher. For this study stratification was used. The road types that were expected to have a higher accident occurrence, were sampled with higher frequency to obtain more detail for those roads.

Obtaining the control group information

In Figure 4 the workflow is depicted. Traffic lanes are randomly selected in the accident collection region and will serve as the main traffic streams. All lanes and the scenes are recorded on video for approximately 30 minutes from a high location (a low visibility extendable beam; see Figure 3) to reduce parallax effects in the analysis of speeds and distances. With special developed software and markings on the road with known distance to each other, the speed and distance to other vehicles can be obtained from the video. For trucks extra information is recorded that will improve the conditioning (matching) for the required conditional random imputation.

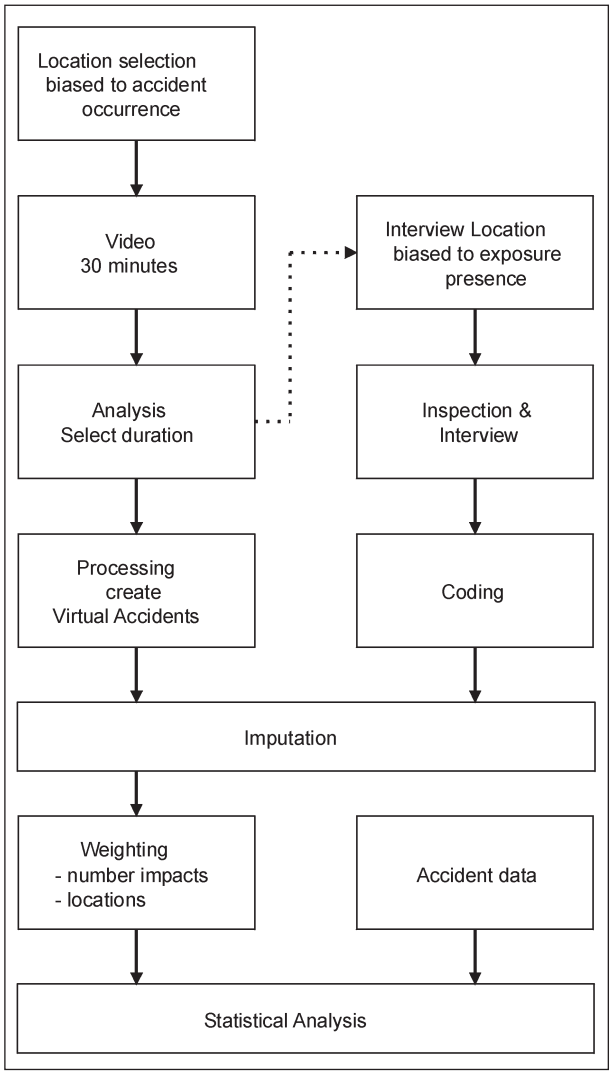

Figure 4: Work flow diagram

The interviews were done at restaurants, gas stations, distribution centers and companies using (specific) trucks. The selection of truck types was matched to the distribution observed in the accident sample, again to obtain a maximum of detail with minimum effort. In doing so, a bias was introduced in the interviews. This only results in a larger sample with a higher confidence level for the vehicles of interest, and less certain information for the other vehicles. A comparison will have to be made between accident trucks of the same type (e.g. on mirror adjustments) in the analysis, for which a sample with higher confidence level is beneficial. Not all truck types could be investigated with the limited sample.

Data weighting

A problem which manifested itself is that due to several reasons the video duration of 30 minutes was in more than one occasion not 30 minutes. Another aspect is that analyzing 30 minutes of motorway video is time consuming and not really necessary to obtain a stable distribution. Thirdly, the number of virtual impacts that are created is limited. Therefore a sample of the video is taken. This made it necessary to assign a weight factor to each vehicle according to the following formula:

$$W_i = \frac{N_i^V}{N_i^A} \cdot \frac{30}{D} \qquad [3]$$

The sum of vehicles of type i in the maximum allowed sample (N_i^A) equals the sum of the allowed number of vehicles from formula [2]:

$$\sum_i N_i^A = A_m + A_o \qquad [4]$$

The weight factor for each virtual impact dependents on the involved vehicles and was established in the following way:

$$W_{ij}^{\ I} = W_i \cdot W_j \qquad [5]$$

After establishing these weight factors per vehicle and virtual impact on a given location, the locations had to be weighted towards their real presence in the sampling region:

$$W_k^{\,L} = \frac{L_k}{N_k^{\,s}}$$
$$W_{ijk}^{\,LI} = \left(W_{ij}^{\,I}\right)_k \cdot W_k^{\,L}$$
$$W_{ik}^{\,LV} = \left(W_i^{\,V}\right)_k \cdot W_k^{\,L} \qquad [6]$$

These weight factors were used throughout the analysis.

Results

The authors want to stress that the results presented in the following sections can only serve as an indication and merely show the analysis possibilities because the sample of accidents is limited.

Accidents and control group population

For this study 30 truck accidents were investigated. The Dutch Accident Research Team (DART) was notified by the technical police departments in the province of Zuid-Holland. The technical police departments in these regions are notified of all truck accidents and measure accident (skid) marks in detail. All accidents for which TNO was notified were investigated by the investigation team. The total sample is biased towards more severe accidents. Comparison with the control group therefore gives information about severe accidents.

Of these accidents, 15 occurred on straight segments and 15 on crossings. The road types (urban, motorway, etc.) of the control group population were matched with the accident occurrence road types and weighted afterwards to obtain a maximum of accuracy and detail.

Example analysis

In Table 1 the frequencies of collision partners in the accident cases and in the virtual accidents (control) on intersections are shown. The adjusted residuals indicate whether the cases and controls differ significantly and may be interpreted as follows: an absolute value larger than two indicates for normal distributions a 95% certainty that a significant difference is present [SPSS, 1998]. The sample in this pilot-study is unfortunately too small to satisfy the condition of normality, therefore only indications can be given. From the presented table it can be read that motorized two-wheelers and bicycles are more present in the accident cases than in the control group. If the sample were large enough, this would indicate that the probability to be in an accident as a bicyclist or motorized two-wheeler is higher than for example a car driver.

A classification tree analysis [SPSS, 1998] with a forced split on vulnerable road users (motorized two-wheelers, bicycles and pedestrians) was used to identify aspects in truck-vulnerable road user impacts, which were found to have a high probability of occurrence (see Appendix). The variables in the classification tree were the manoeuvers of the truck and vulnerable road user with respect to each other.

Compared with the virtual truck accidents (frequency of occurrence of normal meetings) with vulnerable road users (VRU) two groups can be identified with differences in occurrence:

- Truck turning right, with the original driving direction identical to the VRU driving direction. This situation was never observed in the virtual accidents. Not shown is that this occurs in all investigated cases on local small roads and that the VRU is going straight or is turning right. This

			Case		
			Case	Control	Total
OV1: Object type	Truck	Count	0	14158	14158
		% within Case	.0%	6.6%	6.6%
		Adjusted Residual	-1.0	1.0	
	Car	Count	3	170786	170789
		% within Case	20.0%	79.6%	79.6%
		Adjusted Residual	-5.7	5.7	
	Motorised 2-wheeler	Count	3	354	357
		% within Case	20.0%	.2%	.2%
		Adjusted Residual	18.8	-18.8	
	Bicycle	Count	6	1008	1014
		% within Case	40.0%	.5%	.5%
		Adjusted Residual	22.3	-22.3	
	Bus	Count	1	3145	3146
		% within Case	6.7%	1.5%	1.5%
		Adjusted Residual	1.7	-1.7	
	Van	Count	1	24900	24901
		% within Case	6.7%	11.6%	11.6%
		Adjusted Residual	-.6	.6	
	Other vehicle	Count	0	8	8
		% within Case	.0%	.0%	.0%
		Adjusted Residual	.0	.0	
	Pedestrian	Count	0	12	12
		% within Case	.0%	.0%	.0%
		Adjusted Residual	.0	.0	
	Stationary object	Count	1	166	167
		% within Case	6.7%	.1%	.1%
		Adjusted Residual	9.2	-9.2	
Total		Count	15	214537	214552
		% within Case	100.0%	100.0%	100.0%
a. Junction = Yes					

Table 1: Comparison between collision partners in accident cases and in virtual accidents on intersections (The distributions differ significantly. Chi^2-test: $p<0.05$). Cases represent accidents and controls represent virtual accidents

is typically the situation in which blind angle aspects are considered relevant.

- Truck is going in the opposite direction from where it comes from (driving backwards); while in the control group this situation is never observed. This is possibly also a blind angle aspect, but on the rear of the truck.

Discussion

On the method

An aspect that is more difficult to investigate with this study setup is the environmentally related driver behavior: whether an association exists between certain driver behavior and the environment (e.g. cruise control on motorways, use of mirrors at certain locations). When it is expected that these factors play an important role, the relationship can be investigated by implementing questions regarding these relationships in the interviews. Frequency of use under various conditions can be asked. Another more expensive method already mentioned is to form a cohort of random drivers for which the behavior and actions are recorded in some way, possibly by actually monitoring the driver and recording and coding the behavior. Again it is not required that the population matches the exposure information from video.

Night time control samples caused some problems. Video information with “night shot”-mode was not of very good quality.

On the results

From the classification tree in the Appendix it can be seen that right turning trucks and VRU’s coming from the same direction have a relatively high number of cases (3) with respect to the control group (0). At the top of the classification tree this was (10 cases/ 250 controls). However, many more virtual bicycle accidents would have been present if due to chance a location near a school was sampled in the small control group or sampled at hours at which children and students bike to school. Therefore no real conclusions may be drawn from this sample. When the sample would have been large enough and the same situation would persist it might have been concluded for these cases that situations with right turning trucks and VRU’s coming from the same direction impose a greater accident probability than other cases. It then could be suggested that this relates to blind angle aspects.

With more cases the classification tree analysis could go further. If any control group cases would be present for these typically dangerous situations a comparison between environmental, driver or truck-related issues could be made to show typical problems for these locations. But at this time, no control data is available and the number of accidents is clearly limited.

Although potentially influenced by the coincidental choice of control locations, the method of analysis seems to indicate a potential risk factor that was also identified in national statistics [de VRIES, 2000]. The conclusions from the national statistics were based on assumptions about exposure. This in-depth analysis shows that this can be supported objectively with control-group information. Details concerning mirror adjustment, road layout can give more details about exact causation-related aspects. Again comparison with control group information can show discrepancies between the two data sets. This information can be further supported by objective and subjective descriptive information.

Risk adaptation and secondary safety

This method could also identify certain driver behavior and driver risk assessment. The exposure data and injury probability data can be used to calculate driver risk, the risk a driver “feels”:

Relative risk (*K* in accident type)=
P (*K*| accident type) χ *P* (accient type) [7]

The relative risk for being killed (*K*) in a certain accident type equals the probability to get killed in a certain accident type times the probability that such an accident occurs. The relative probability for occurrence can be obtained from the exposure data. If this occurrence probability is very low, but the consequences still high, a driver might still feel quite safe. When the occurrence probability is high and the probability is also high this will be perceived as dangerous.

A certain safety feature could induce more-unsafe driving (risk adaptation). Exposure data can show that this may be the case if discrepancies exist between the accidents population and the exposure population in the presence of secondary safety features. Suppose the degree of implementation in

the normal population of a secondary safety feature is found to be 50% (e.g. frontal airbag), one would expect to find a same or less relative accident probability (number of accidents divided by the number of virtual accidents in that category) for cars equipped and cars not equipped with airbag. Airbags are supposed to reduce injuries, so some accidents will not be reported to the police anymore or will not be included in the study sample, therefore a lower accident probability is expected. If one would find a higher accident probability for cars equipped with airbags, but still a lower probability to get injured one may conclude that risk adaptation has occurred, reducing the expected benefit. Measuring the degree of implementation based on for example car sell rates or kilometers driven should be done only with extreme care (see Introduction).

Conclusions

From the literature study it was learned that a case-control study is best suited for in-depth traffic accident research at this time. No good documented case-control study could be found which includes environmental, driver and vehicle information. Therefore the new method was developed.

- A case-control group study with real and “virtual accidents” was developed and tested on 30 accident cases and 30 random locations.
- Data imputation could technically be realized. A validation was not yet possible because of the small sample.
- Injury causation analysis can be done in great detail. A large amount of data analysis possibilities exist. The analysis possibilities seem to give good information and indications to find problems in accident and injury causation from which new solutions may be derived.
- Risk adaptation for primary and secondary safety features can be assessed.
- Environmental, human and vehicle factors can be investigated together, taking into account the relationship between the factors.
- The results from this study, although limited, are in line with results from other studies.

Recommendations

When defining measures for improved safety, it is recommended to include a dedicated exposure evaluation in order to determine with statistical significance whether, and up to what extent, actual safety improvements can be expected. The case control method presented here is a good approach.

The exposure method should be evaluated on a wider scale, preferably European-wide to effectively indicate risk factors. European projects like SafetyNet or TRACE might provide a good basis.

References

S.L. BANGDIWALA (2000): Methodological Considerations in the Analysis of Injury Data: A challenge for the Injury Research Community, Injury Prevention and Control, Taulor & Francis, ISBN 0-748-40959-9.

DOT (2005): http://www.minvenw.nl/dgg/dodehoek/Zichtveldprobleem/Feitenencijfers/#feiten

ELSEVIER (1997): De kans op een ongeluk, 6-9-1997

GIS (2004): http://www.gis.com/whatisgis/index.html

OECD (2001): Motorcycles: Common international methodology for in-depth accident investigations

D.B. RUBIN (1987): Multiple imputation for non-response in surveys, John Wiley & Sons, New York.

SPSS (1998): AnswerTree 2.0 user’s Guide, ISBN 1-56827-254-5

Y.W.R de VRIES (2000): Accident analysis of heavy trucks, TNO Report, 00.OR.BV.053.1/YdV

Y.W.R de VRIES (2005): A method for control group data to find accident and injury causation factors in in-depth traffic accident studies, TNO Report, 05.OR.SA.034.1/YdV

Appendix

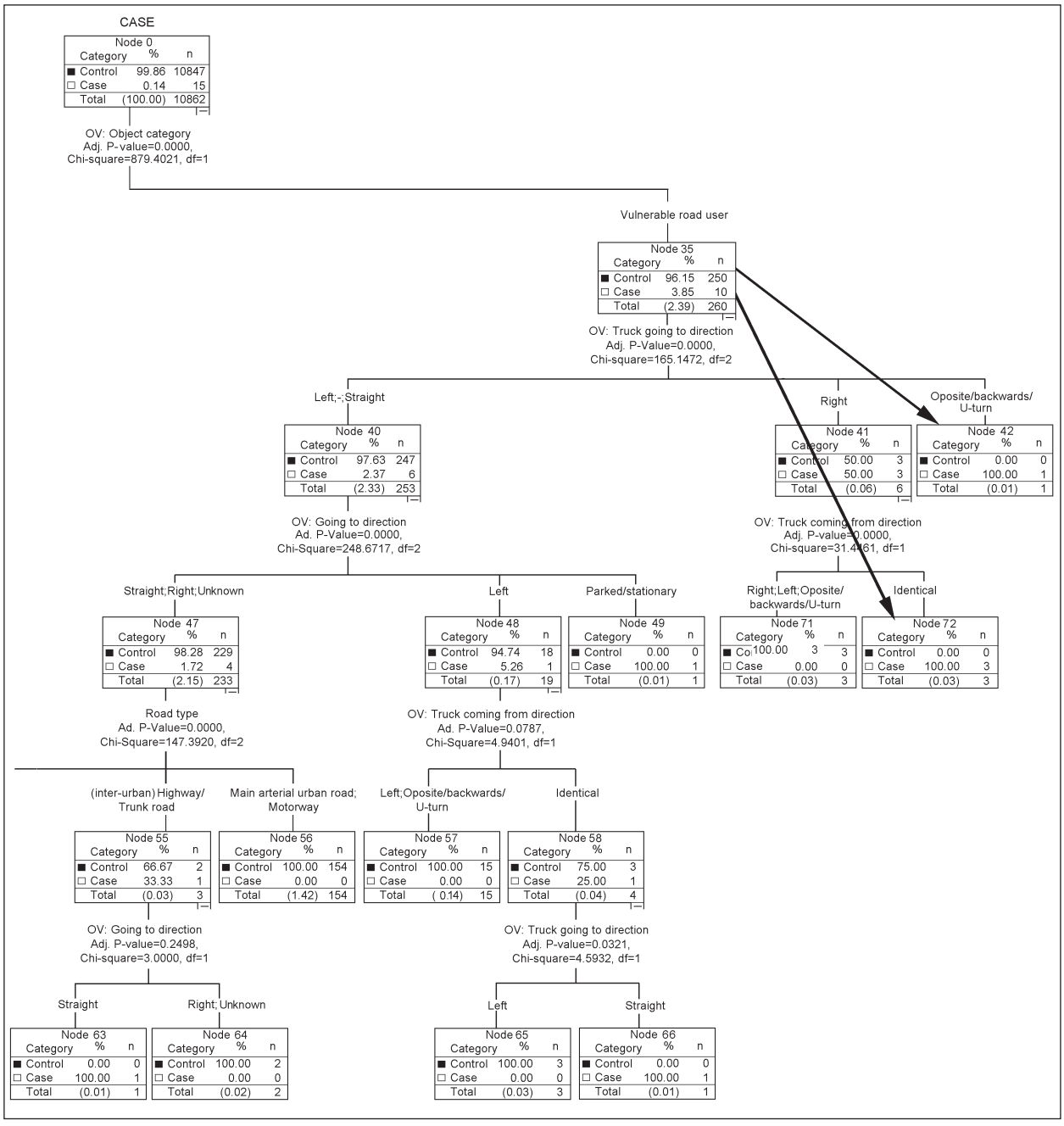

Figure 5: Classification tree. Explanation of the used terminology: underneath vulnerable road users: from the Other Vehicle (OV) perspective, the truck is going into a certain direction relative to the OV. Suppose the truck is turning right, then read: from the OV perspective the truck is coming from, e.g., the identical direction as the OV