About the covariance between the number of accidents and the number of victims

F.D. Bijleveld

R-2002-24

About the covariance between the number of accidents and the number of victims

R-2002-24
F.D. Bijleveld
Leidschendam, 2002
SWOV Institute for Road Safety Research, the Netherlands

Report documentation

Number:	R-2002-24
Title:	About the covariance between the number of accidents and the number of victims
Author(s):	F.D. Bijleveld
Research theme:	Road safety analysis
Theme leader:	Dr. J.J.F. Commandeur
Project number SWOV:	37.112
Keywords:	Accident rate, fatality, injury, statistics, covariance
Contents of the project:	Traffic safety is not only indicated by the number of accidents, but also by numerous accident-related outcomes like the number of people killed, the number of people seriously injured, the amount of material damage, etc. In this study some statistical issues involved in the simultaneous analysis of accident-related outcomes (the number of victims, fatalities and accidents) of the traffic process were studied. The main focus of this study was the covariation of the outcomes.
Number of pages:	26 + 13
Price:	€ 10,-
Published by:	SWOV, Leidschendam, 2002

SWOV Institute for Road Safety Research
P.O. Box 1090
2260 BB Leidschendam
The Netherlands
Telephone +31-703173333
Telefax +31-703201261

Summary

Traffic safety is not only indicated by the number of accidents, but also by numerous accident-related outcomes like the number of people killed, the number of people seriously injured, the amount of material damage, etc. When assessing the possible effect of, for example, a road safety measure on traffic safety, it is therefore important to be able to investigate the differentiated effect of such a measure on accidents and accident-related outcomes.

In this study some statistical issues involved in the simultaneous analysis of accident-related outcomes (such as the number of victims, fatalities or accidents) of the traffic process were studied. The main focus of this study was the covariation of the outcomes: the interdependencies of accident-related outcomes were investigated by establishing their (theoretical) covariance structure. Estimates of the covariances of normal approximations of joint distributions were derived for the following cases:

a. The total number of accidents, victims and fatalities in a certain class. Examples of classes are monthly data, car-only accidents.
b. The logarithm of the total number of accidents, victims and fatalities in a certain class.
c. The logarithm of the total number of accidents, the logarithm of the ratio of the number of victims to the number of accidents, the logarithm of the ratio of the number of fatalities to the number of victims.

The quality of these estimates was evaluated using samples of real-life data from the Netherlands.
Distributional aspects like effects on the estimates of small numbers, of small numbers of fatalities per accident, and of different types of accidents were also investigated in this study. It turns out that deviations are generally modest in most cases but may become serious when the counts are smaller.

The following results were found:
- It is possible to derive relatively simple expressions for the variances and covariances of (logarithms and ratios of) accidents and victim counts. As regards usability, some information needed to compute estimates of the covariance matrices may not be available over a longer period of time. However, in some cases this information can be estimated.
- When performing a multivariate analysis using numbers of accidents, victims, and fatalities as outcome variables, or any of the other outcome variables mentioned above under b) and c), all three variables must be used. This follows from the finding that each of the three variables carries unique information that cannot be estimated from the other two.
- The logarithm of the total number of accidents is (approximately) uncorrelated to the other two variables mentioned above under c). This means that the effect of explanatory variables on the logarithm of the total number of accidents can be (approximately) assessed independently, with no regard to the other two variables.

- The approximation of the logarithm needed when log-counts of accidents victims or other accident counts are analysed, is usually sufficiently close. In practice but depending on circumstances, problems caused by the approximation are unlikely when counts are higher than 30.

Contents

1. Introduction

This research was conducted within the framework of the SWOV research theme ‘Road safety analysis’.

One part of the research in this theme consists of related and supplementary studies intended to enhance the technical possibilities for the analysis of the development of road safety in general, and for the projects within the theme in particular. This part of the research is intended to enhance both applicability and reliability of the analysis.

Other studies in this theme involve the disaggregation into road, traffic and victim characteristics; use of economic and other explanatory variables, among other things.

The main objective of the research theme is to find explanations for the observed developments in road safety by studying the relationships between these developments and (developments of) explanatory variables such as road safety measures, changes in traffic volume, economic developments, etc.
Obviously, road safety is not a one-dimensional concept. It not only consists of the number of accidents, but also of numerous accident-related outcomes like the number of people killed, the number of people seriously injured, the amount of material damage, etc.
When assessing the possible effect of a road safety measure on traffic safety, it is therefore important to be able to investigate the differentiated effect of such a measure on accidents and accident-related outcomes. This typically calls for a multivariate multiple regression type of analysis, where accidents and accident-related outcomes are the dependent variables in the analysis. Moreover, when measures are introduced in a relatively short period of time, without an analysis of the differentiated effects, it is sometimes impossible to determine which traffic safety effect is caused by which road safety measure.

In a classical multivariate analysis the dependent variables are assumed to be unrelated to one another. This is reflected in the fact that in a classical multivariate regression analysis with p dependent variables for example, exactly the same regression weights are obtained as when p separate univariate regression analyses are performed.
However, numbers of accidents and numbers of accident-related outcomes are clearly related to one another. For instance, the number of road fatalities is related to the number of accidents: zero accidents imply zero fatalities, while an increase in the number of accidents is usually associated with an increase in the number of fatalities. See for instance Cameron & Trivedi (1998, p. 260) for a similar example. Failing to incorporate these dependencies in a multivariate analysis may result in false conclusions from the analysis.

In the present study the interdependencies between numbers of accidents and (numbers of) accident-related outcomes are investigated by establishing their (theoretical) covariance structure. The covariance structure proposed in this report can be used to correct the dependencies

between numbers of accidents and accident-related outcomes in a multivariate analysis, thus yielding more reliable results.

The proposed method is not restricted to numbers of accidents, victims and fatalities, but can be generally applied to any other type of outcome of independent accidents, like, for example, medical costs and the total length of traffic queues resulting from accidents.

2. Problem analysis and set-up of the study

2.1. Problem analysis

2.1.1. *The analysis of the (impacts on the) development of traffic safety needs more than just one dimension*

The development of traffic safety or the lack of safety (‘unsafety' for short) is very often analysed following the development of one single relevant phenomenon in relation to explanatory variables. However, in order to get a better understanding of for instance the effectiveness of road safety measures it would be better to study the potentially differentiated effect on the different dependent variables as well. For example, it is likely that the compulsory use of seat belts mainly has an effect on the consequences of an accident, whereas measures aiming to reduce the occurrence of driving under the influence of alcohol mainly have an effect on the number of accidents. Speed reducing measures are supposed to have an effect on both the number of accidents and the consequences of an accident. Particular theories, such as for example the risk homeostasis theory (Wilde, 1994) and the zero risk theory (Summala & Näätäänen, 1988) state that theoretically likely developments may be counteracted because of behavioural adaptation. An example of this is the use of seat belts which may, according to some theories, result in higher speeds and other more dangerous behaviour, so that the expected reduction in the number of injuries is undone by the fact that the number of accidents increases. Thus in general it can be argued that road safety developments cannot be measured by just one such accident consequence alone. Factors influencing safety are likely to affect more than just one kind of consequence. Hence, road safety analysis models would need to take account of these multi-outcome characteristics.

2.1.2. *The impacts on the development of traffic safety need to be differentiated in time (and ‘space')*

Firstly, for a good understanding of the road safety developments and the differential contribution of various types of road safety measures, it is important not only to look at the development of single road safety indicators, but also to look into the coherence of effects of different indicators and explanatory factors. Additionally, the development over time should be considered, as changes in traffic safety should also be related in timing to the implementation of the road safety measures that caused them. A similar argument can be used to look into location aspects (for instance road types). Confidence in the conclusions concerning the question of effectivity of road safety measures is increased when the pattern of influences reflects theoretical predictions (or the reverse).
The main goal of this study is directed at the first issue: the influence of road safety measures on different types of unsafety outcomes. The second and third issues (time and location specific measures) are not within the scope of this study. This study focuses on how to analyse related outcomes simultaneously.

The analysis of the development of multiple consequences of traffic unsafety over time is most likely best conducted using multivariate time series. The multivariate character of time series analysis is in this case not only reflected in the multiple explanation variables (e.g. traffic volume, economic factors, weather conditions, and safety measures; to name a few), but also by the fact that more than one consequence is studied. Moreover, as is obvious in empirical sciences, the explanation variables are based on estimates of some kind, sometimes a missing value. Its true value is relevant and should in some cases be the same for all the consequences under study. However, in order to be able to do so, a number of methodological problems has to be solved.

2.1.3. *The issue of covariance*

One important problem is that multivariate observations observe some covariance. The most notable example is the fact that no more fatal accidents can occur in a period of time than the total number of fatalities. A more general example is that when more traffic accidents occur in a year, it is likely that more people get injured in traffic as well. The former aspect is not developed completely in this study, rather, an approximation is made. This is done by developing an expression for the covariance (matrix) of the (possible logarithms of) counts of unsafety outcomes. As such, the usability of the covariance structure is not restricted to time series analysis.

In the context of the last (more general) example, it can be observed that when a relatively large number of accidents occurs in a certain year, quite often a relatively large number of people gets injured as well in that same year. This would mean a positive covariance between the number of accidents and the number of injured. Now suppose two hypothetical models, model ‘A' and model ‘B' are to be compared. The fact that model ‘A' predicts say 10 more accidents and 10 more injured than are actually observed in that year has to be interpreted quite differently than the fact that model ‘B' also predicts 10 more accidents in that year but 10 less injured. Roughly speaking both models are equally likely when covariance is ignored and model ‘A' is more likely when the positive covariance between the number of accidents and the number of injured is taken into account. Even worse, if model ‘B' had predicted 5% less injured, it would have been deemed more likely than model ‘A' if covariance is ignored, possibly resulting in false conclusions about the significance of the effectivity of measures studied.

2.2. **Report set-up**

Chapter 3 describes and discusses the results of an analytic derivation of a mathematical description of the covariance between the total number of accidents (not necessarily injury accidents) and the total number of victims in a time period.

The analytic results were compared to results based on simulation, sampling from real accident data from the Netherlands dating from 1980 to and including 1999. The purpose is to assess the consequences in terms of accuracy of some of the simplifications used in the theoretical derivations. Two versions of simulations are described: one based on car-only accidents

and one based on all accidents. The derivation itself and details of the analysis involved are in the *Appendix* to this report.

Chapter 4 describes the results of a trial study of the simultaneous development since 1964 of some of the main accident outcomes in relation to major explanatory variables and three road safety measures. In this trial study the covariances found in *Chapter 3* are used. No claim is made to be complete at this point, specifically with respect to the use of explanatory variables, as this development is subject to more elaborate studies in later stages in this research theme. The purpose of this trial study is to evaluate the usability of the technique.
The method is applied to identify the potentially differentiated effect of three road safety measures:

- the introduction of the legal BAC limit of 0.5 ‰ (introduced in November 1974). This measure is expected to have had an impact on the number of accidents, mainly accident frequency.
- the introduction of the compulsory use of seat belts in front seats (introduced in June 1975). This measure is expected to have had an impact on accident severity.
- the introduction in September 1990 of the administrative settlement of minor road traffic violations, the Mulder Law. This measure may have had an impact on the occurrence of various accident outcomes.

Finally, in *Chapter 5* the results are discussed and recommendations are made.

3. Covariance structure of the traffic unsafety counts

This chapter summarizes the results derived from the study into the covariance structure of the traffic unsafety counts. The main results are given in *Table 3.2*. Details can be found in the *Appendix* to this document.

3.1. Method

Using the assumption of a Poisson distributed number of accidents as a starting point, expressions for central moments of the number of victims (and as a special case: fatalities) are derived using characteristic functions. This is done using a method similar to the method used by Feller (Feller, 1968). Using these moments variance estimates are derived. Next, using multivariate characteristic functions covariances are derived. These results are used to derive (co)variances for logarithmic-transformed counts.

In addition to the Poisson assumption it is further assumed that the number of victims (or fatalities) per accident is identically and independently distributed over the accidents. The consequences of deviation from this last assumption in practice, mainly because different types of accidents may have different distributions of the number of victims, are studied using simulation techniques. The results thereof and other comparisons are presented in the *Appendix A.3*.

3.2. Results

Based on the results of the derivations as reported in *Appendix A.1* and *A.2* of this document, covariance matrices based on normal approximations can be formulated between either the count variables, the logarithms of those count variables or the logarithms of ratios of count variables. In principle, other combinations are possible too, but have been omitted in this study.

The following cases were developed:

a. The total number of accidents, victims and fatalities in a certain class. Examples of classes are monthly data, car-only accidents.
b. The logarithm of the total number of accidents, victims and fatalities in a certain class.
c. The logarithm of the total number of accidents, the logarithm of the ratio of the number of victims to the number of accidents, the logarithm of the ratio of the number of fatalities to the number of victims.

The basic assumption made in this study is that the number of accidents in a period of time is Poisson distributed. Under this assumption the variance of the number of accidents N is equal to its expected value λ: Var(N) = λ. Both quantities are estimated by the observed number of accidents n. *Table 3.2* provides an overview of all results while *Table 3.1* contains an explanation of abbreviations used in *Table 3.2.*

Number of	Realisation	Usually available	Abbreviation
Accidents (acc)	n	Yes	n
Victims in accident i	v_i	No	
Fatalities in accident i	f_i	No	
Sum over all accidents of the number of	Estimate	Usually available	Abbreviation
Victims (vic)	$\sum_{i=1}^{n} v_i$	Yes	$\sum v$
Fatalities (fat)	$\sum_{i=1}^{n} f_i$	Yes	$\sum f$
Sum over all accidents of the square of the number of	Estimate	Usually available	Abbreviation
Victims	$\sum_{i=1}^{n} v_i^2$	No	$\sum v^2$
Fatalities	$\sum_{i=1}^{n} f_i^2$	No	$\sum f^2$
Sum over all accidents of the cross product of the numbers of	Estimate	Usually available	Abbreviation
Victims and fatalities	$\sum_{i=1}^{n} v_i f_i$	No	$\sum f v$

Table 3.1. *Abbreviations used in the derived equations for variances and covariances and estimates.*

From *Table 3.1* it can be seen that not all information is available in standard publications on accidents. This is indicated by “No” in the “Usually available” column. More detailed sources on individual accidents are needed to get more precise estimates. In that case the individual fatality f_i and victim v_i counts per accident will be available. Then the variance of the total number of fatalities for instance can be computed as the sum of the squared fatality counts as indicated in *Table 3.2*. In cases where it can be assumed that no substantial changes over time should occur, values constant for an entire period might be estimated. This should amount to a solution similar to estimating a ‘scale' in generalized linear models.

As regards the structure of the covariance matrices, no useful dependencies can be found. This means that it is not possible to capture the information on the number of accidents, victims and fatalities (or any of the other two triplets) by using just two of the indicators (for instance accidents and fatalities). In order to make sure not to lose relevant information, all three indicators are necessary. For more information on this see *Appendix A.3.2.*

One interesting result (that can be derived from the last part of *Table 3.2*) however is the fact that the logarithm of the number of accidents is uncorrelated to the logarithms of ratios of the number of victims to the number of accidents, and the number of fatalities to the number of victims. This has advantages in certain statistical techniques. This case has been studied in a trial study (see *Chapter 4*). The covariance matrix of these three components is essentially taken proportional to the reciprocal of the (expected) number of accidents. However, one reservation had to be made

in this case, because the variance of the logarithm of the ratio of the number of fatalities to the number of victims is increasing even stronger than can be attributed to the decrease in the number of accidents. This seems to be caused by a decrease in the number of fatalities per accident (see *Figure A.6* in *Appendix A.3.2*).

Results based on counts	
Variance of	Estimate
the total number of accidents	n
the total number of victims [a)]	$\sum v^2$
the total number of fatalities	$\sum f^2$
Covariance of	Estimate
the total number of accidents and victims	$\sum v$
the total number of accidents and fatalities	$\sum f$
the total number of victims and fatalities	$\sum f v$
Results based on logarithms of counts	
Variance of	Estimate
the total number of accidents	$1/n$
the total number of victims	$\sum v^2/(\sum v)^2$
the total number of fatalities	$\sum f^2/(\sum f)^2$
Covariance of	Estimate
the total number of accidents and victims	$1/n$
the total number of accidents and fatalities	$1/n$
the total number of victims and fatalities	$\sum f v \,/\, (\sum v \times \sum f)$
Results based on logarithms of ratios of counts	
Variance of	Estimate
the total number of victims to accidents	$\sum v^2/(\sum v)^2 - 1/n$
the total number of fatalities to victims	$\sum f^2/(\sum f)^2 + \sum v^2/(\sum v)^2 - 2\sum f v/(\sum v \sum f)$
Covariance of the logarithm of the total number of accidents and	Estimate
the total number of victims to accidents	0
the total number of fatalities to victims	0
Covariance of the logarithm of the ratio of the total number of victims to accidents and	Estimate
the total number of fatalities to victims	$(\sum f v)/(\sum v \sum f) - (\sum v^2)/(\sum v)^2$

a) Consequences of this result and the result for fatalities will be the subject of a separate study. In case of all accidents, the variance of the total number of victims is up to about 50% higher than the total number of victims.

Table 3.2. *Derived equations for variances and covariances and estimates.*

4. Trial study of multivariate traffic unsafety counts

The example in this chapter concerns a trial study intended to demonstrate the use of differentiated effect estimates of road safety measures and uses one of the derived covariance matrices in *Chapter 3*. This example is neither designed nor intended to give a definitive estimate of any of the road safety measures included in the analysis. It is for demonstrative purposes only, aiming to assess the usability of the technique employed. The data used are selected on both possible relevance and the fact that the data are readily available disaggregated into monthly data.
In this trial study the following traffic safety indicators were analysed simultaneously (data starting in 1964 - 1999).

- the logarithm of the number of injury accidents,
- the logarithm of the number of victims per injury accident.
- the logarithm of the number of fatalities per victim.

The following explanatory variables were considered:

- (average) traffic index (index),
- (average) temperature in 'De Bilt' (T), which is a town situated in the middle of the Netherlands, where the Meteorological Office is located
- duration of precipitation in De Bilt (D).

The following intervention time points were considered:

- the introduction of the legal BAC limit of 0.5 ‰ (introduced in November 1974),
- the introduction of the compulsory use of seat belts on front seats (introduced in June 1975),
- the introduction of the administrative settlement of minor road traffic violations, the so-called Mulder Law (introduced in September 1990).

4.1. Method

A multivariate state space model was fitted to these data. State space models are described in Harvey (1989) and, in the context of traffic safety, in Harvey & Durbin (1986) and COST329 (In preparation).

The general description of the model is as follows. The three traffic safety indicators are each explained using the product of a 'risk' development (not a very well chosen name) and the three explanatory variables $index$, T and D. The interventions are assumed to affect only the 'risk' development as they should affect safety, causing, if anything at all, a shift in the level of the risk. The risk development is defined by means of a state space. A state space is a series of vectors that is defined using a set of equations, defining the next vector based on the current vector and possibly other information, such as in this example interventions. In the currently used equations the interventions may (if at all) cause a change of the level in three of the components in the state space. These components denote the level of the development of the 'risk' for each of the three traffic safety indicators, and are called the level components. This equation is called the 'state equation'. These individual vectors are also related to the three traffic safety indicators through yet another equation, called the measurement equation in which the indicators are explained using the state vectors and in this case explanatory information ($index$, T and D). In this example only linear equations are used.

It is this flexible set of equations that makes this method so useful because it can be adapted to many practical cases in traffic safety.
A multiplicative model is built by means of an additive model on the logarithms of the accident data.

In order to give a more precise description of the model (but still very rough), first the vector of logarithms of the traffic safety indicators is denoted by y_t. The development of each of the indicators is thus assumed to be related to some unique 'risk' development and exogenous influence as explained by the explanatory variables. The unique 'risk' development consists of a level and a drift component denoted together by 'trend' and, except for annual data, a seasonal component. The level component is used to model the level of the risk while the drift denotes the systematic change (drift) in the level of the risk. Other kinds of components are possible but are not used here. This means (i = 1,2,3):

$$y_{ti} = \text{trend}_{ti} + \text{seasonal}_{ti} + \log(\text{index}_{ti}) + \tau_i \times \text{T}_{ti} + \delta_i \times \text{D}_{ti} + \text{error}_{ti}.$$

τ_i and δ_i are unknown parameters. Whether or not the (average) temperature 'T' or the duration of precipitation in De Bilt 'D' should be log-transformed has not been studied, but should be in a definitive study, as can be stated for the fact that the $\log(\text{index}_{ti})$ has no coefficient.

The vectors y_t are linked to the state space as follows:

$$y_t = H_t z_t + d_t + \varepsilon_t$$

where z_t denotes an unobserved series of vectors (called state vectors) which elements consist of trend, seasonal and possibly other components. The vector ε_t is assumed to be normally distributed with covariance matrix R_t as estimated using *Table 3.2* (or equation *A.20*) and independent of all other random components. The matrix H_t is the appropriately dimensioned measurement matrix. The vector d_t is the vector of exogenous contributions, by component it is:

$$d_{ti} = \log(\text{index}_{ti}) + \tau_i \times \text{T}_{ti} + \delta_i \times \text{D}_{ti},$$

It is assumed that the state vector z_t contains all relevant information up to time t. This information is a combination of the prediction of the current state based on the previous state and the most recent observation. The equation used for prediction of the current state z_t based on the previous state z_{t-1} is

$$z_t = F_t z_{t-1} + c_t + \omega_t$$

The vector ω_t is assumed to be normally distributed with covariance matrix Q_t. The matrix Q_t usually is to be estimated. The matrix F_t is the appropriately dimensioned transition matrix. It defines how the current state is carried over to the next. The vectors ε_t and ω_t are independent of each other and all other time points. Some components in z_t are used to quantify the level of the 'risk'. It is these components that are influenced by the

interventions by means of the vector c_t, allowing a certain level change at the time points at which the respective interventions became effective. No effort was made here in this study to verify the time points at which the interventions became effective. Level changes are assumed to be permanent.

Further details on state space models can be found in Harvey (1989), and in the context of traffic safety, in Harvey & Durbin (1986), among others.

4.2. **Results**

The accident data were analysed using monthly, quarterly and annual data. The purpose of this is to determine which of the three is best suited for analysis, mainly as a result of the effect of including a seasonal pattern in the state space.

The accident data were analysed using data from 1964 - 1999. Not all data were used. Rather, the observations of the last few (10, 5 and 3) years were temporarily withdrawn from analysis and subsequently used for comparison with prognosis from the models. This means that for the monthly data the last 120, 60 and 36 months, for the quarterly data 40, 20 and 12 quarters, and for the annual data the last 10, 5 and 3 years were held back in the estimation procedure and were thus not used in the prognosis resulting on that estimation. The held back observations were compared to the prognosis results. In this study the length of prognosis was called the ‘lead'.

In *Figure 4.1* the development of the number of accidents is graphed. The development is modelled per year in uninterrupted lines, three months ‘quarterly’ observations (lines with larger dashes) and per month (short-dashed line). Results for quarterly and monthly data were aggregated into annual data. The first four vertical gridlines denote the following time points respectively: introductions of the alcohol law; the seat belt law; the beginning of the last 10 years of observation and the Mulder Law. Don't be tempted to interpret the effectivity of the measures just by the local development of the number of accidents in the vicinity of the time points. Other influences, like traffic volume, also influence the development. The last vertical gridlines denote the beginning of the last 5 and 3 years respectively. The ‘prediction' for the outcome observations in the estimation period was attained using a smoothed estimate of the state, which resulted in a rather close approximation of the outcome data. This should not be confused with a good fit. Rather, the forecasts are better suited for this. *Figure 4.2* for victims per accidents and *Figure 4.3* for fatalities per victim have a similar layout.

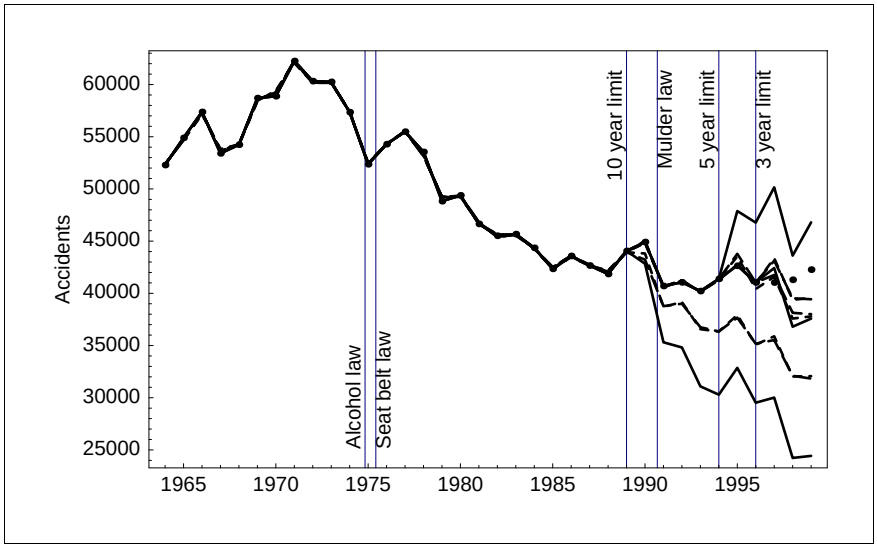

Figure 4.1. *The development of the number of accidents modelled per year (uninterrupted lines, one line for each lead), quarter (dashed lines, one line for each lead) and per month (short-dashed lines, one line for each lead), dots are observed values. Results for quarterly and monthly data were aggregated into annual data. The vertical gridlines denote the following time points respectively: introductions of the alcohol law; the seat belt law; the beginning of the last 10 years; introduction of the Mulder Law; beginning of the last 5 and 3 years.*

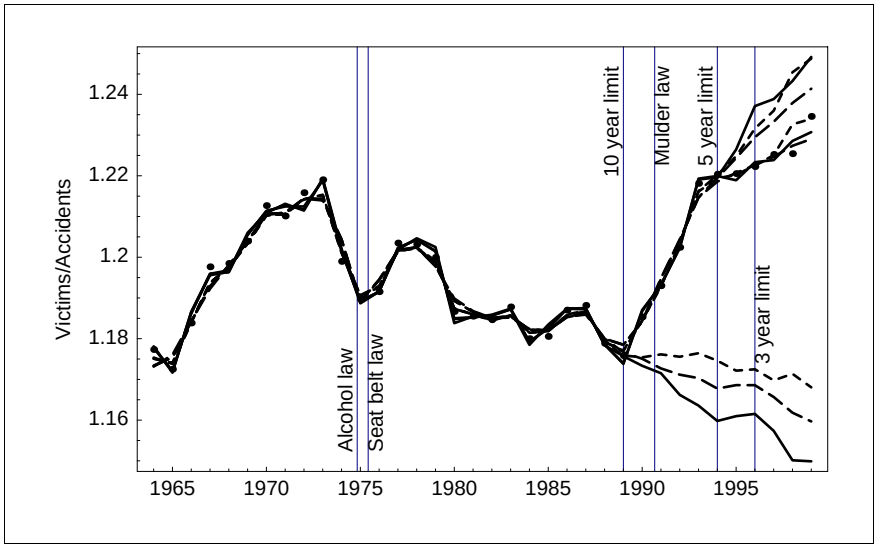

Figure 4.2. *The development of the number of victims per accident modelled per year (uninterrupted lines, one line for each lead), quarter (dashed lines, one line for each lead) and per month (short-dashed lines, one line for each lead), dots are observed values. Results for quarterly and monthly data were averaged into annual data. The vertical gridlines denote the following time points respectively: introductions of the alcohol law; the seat belt law; the beginning of the last 10 years; introduction of the Mulder law; beginning of the last 5 and 3 years.*

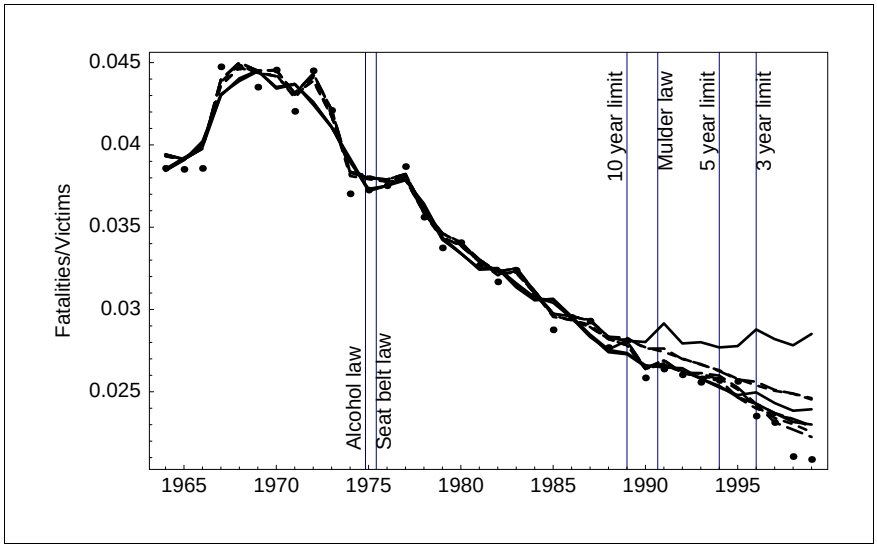

Figure 4.3. *The development of the number of fatalities per victim modelled per year (uninterrupted lines, one line for each lead), quarter (dashed lines, one line for each lead) and per month (short-dashed lines, one line for each lead), dots are observed values. Results for quarterly and monthly data were averaged into annual data. The vertical gridlines denote the following time points respectively: introductions of the alcohol law; the seat belt law; the beginning of the last 10 years; introduction of the Mulder Law; beginning of the last 5 and 3 years*.

The parameter estimates for the intervention effects are given in *Tables 4.1* (for the alcohol law) *4.2* (seat belt law) and *4.3* (Mulder Law). These intervention effects are the changes in the level components of the respective 'risks': one for accidents (third column), one for victims/accident (fourth column) and one for fatalities/victim (fifth column). The first column list the time unit. This means that the records starting with a "Y" ("Y" for year) are based on results of an analysis based on annual data (January through December), "Q" ("Q" for quarter, a three month period starting in January, April, July or October) are based on an analysis based on quarterly data and "M" ("M" for month, entire months).

The second column indicates the number of time units that have been left out of the respective analysis. In the first case this represents the last 3 years left out, in the second case the last 5 years left out and in the third case the last 10 years left out. In the fourth case this pattern is repeated as twelve quarters are left out and so on.

The actual values in the tables indicate the shift in the level of the respective risk components. Positive values indicate that the intervention resulted in an increase of the corresponding traffic safety indicator, while negative values are associated with a decrease in the corresponding traffic safety indicator. Thus, the value 0.042363 in *Table 4.1* indicates a slight permanent relative increase in the risk component of the number of accidents. If this value was significant, it would indicate an increase in the number of accidents at the time point at which the alcohol law was introduced. The fact that some observations in *Table 4.3* are missing is due to the fact that the Mulder Law was introduced within the last 10 years.

Time unit	Lead (units)	Estimates of coefficients		
		Accidents	Victims / Accident	Fatalities / Victim
Y	3	0.042363	0.005597	0.057084
Y	5	0.084330	0.005751	0.059915
Y	10	0.066302	0.004015	0.054438
Q	12	-0.097782	-0.006697	0.044883
Q	20	-0.098809	-0.006559	0.046923
Q	40	-0.092517	-0.006965	0.044060
M	36	-0.012037	-0.012914	0.002874
M	60	-0.008473	-0.012578	0.004251
M	120	0.000836	-0.013501	0.004696

Table 4.1. *Estimates of coefficients (level changes) due to the alcohol law on level components for Accidents, Victims/Accident and Fatalities/Victim in state space. For more information see explanation on page 19.*

Time unit	Lead (units)	Estimates of coefficients		
		Accidents	Victims / Accident	Fatalities / Victim
Y	3	0.007556	0.012203	0.087772
Y	5	0.044960	0.012601	0.087893
Y	10	0.031894	0.010919	0.088046
Q	12	0.077444	0.003290	-0.038024
Q	20	0.077247	0.003404	-0.037288
Q	40	0.076404	0.002857	-0.045700
M	36	0.060929	0.005483	0.001025
M	60	0.061250	0.005688	-0.001087
M	120	0.061067	0.005126	-0.004251

Table 4.2 *Estimates of coefficients (level changes) due to the seat belt law on level components for Accidents, Victims/Accident and Fatalities/Victim in state space. For more information see explanation on page 19.*

Time unit	Lead (units)	Estimates of coefficients		
		Accidents	Victims / Accident	Fatalities / Victim
Y	3	-0.011283	0.005636	0.040535
Y	5	-0.036550	0.006007	0.037103
Y	10	-	-	-
Q	12	0.023712	0.000733	0.043955
Q	20	0.023821	-0.000080	0.041790
Q	40	-	-	-
M	36	0.003289	0.007349	0.066259
M	60	0.008918	0.006800	0.065327
M	120	-	-	-

Table 4.3 *Estimated coefficients (level changes) due to the Mulder Law on level components for Accidents, Victims/Accident and Fatalities/Victim in state space. For more information see explanation on page 19.*

The forecasts for 10, 5 and 3 years ahead are compared to annual observations. To this end, forecasts based on quarterly and monthly data were aggregated for accidents (*Figure 4.1*) and averaged for the other two outcomes (*Figure 4.2* and *Figure 4.3*). A quick inspection reveals that the accuracy of the prognosis decreases as time proceeds. This is an obvious property; longer term prognoses are generally worse compared to short term prognoses.

Furthermore, it seems that the prognosis based on quarterly data is generally (no formal proof, hypothesis based on observations) better than the prognosis based on annual data and as good as the prognosis based on monthly data. In this case ‘better' means closer to the observed annual data. The parameter estimates for the intervention effects in *Tables 4.1* (alcohol law) *4.2* (seat belt law) and *4.3* (Mulder Law) suggest a similar conclusion. The ‘sign' patterns differ substantially between annual and monthly data. This last fact by itself could be caused by the fact that the resolution of the annual data is not enough for the interventions, as they are not introduced at the beginning of years. A clear example is the introduction of the alcohol law in November 1974 and the seat belt law in June 1975. It is difficult to support the assumption that all development in 1974 can be attributed to the alcohol law, as it was in force only in the last two months. If instead the intervention time point is chosen to be 1975, then its effects can hardly be distinguished from those of the seat belt law. To some extent this phenomenon is also relevant for quarterly data, but does not seem to have that much influence in that case.

In *Figure 4.2* it can be seen that the ratio of the number of victims to the number of accidents started to increase in the beginning of the 1990s. The reason for this is not understood nor has it been explained by any of the relevant models excluding the last 10 years as of 1989. This can be seen by the substantial difference between the observed ratios and the prognosed ratios of all models excluding the last 10 years as of 1989. Although the upsurge resembles a similar upsurge starting in 1965, it is much more extensive and continues over a longer period of time. A possible explanation could be a selective change in the level of accident registration, resulting in police reporting of more serious accidents, in this case accidents with more victims. No proof of this is available. This change in registration could hardly explain the sustained and quite systematic increase for the first few years after its onset. Another explanation might have been the introduction of the Mulder Law, as it was introduced approximately at the same time as the onset of the increase. Although no definitive conclusions can be drawn from this study, estimated coefficients for Mulder Law on the level component of the victims per accident in *Table 4.3* do not support such an effect either. This intervention also cannot explain the sustained increase for the first few years after its onset. Noteworthy is that even in this case, where no model seems to yield good prognoses 10 years ahead, the monthly and quarterly data seem to result in better prognoses than the annual data.

In *Figure 4.3* the development of the number of fatalities per victim is shown. It can be seen that the prognosis based on the annual development excluding the last 10 years fails to predict the last 10 years rather severely compared to the others. This result could be interpreted as resulting from a further decrease in the trend in the last decade, suggesting that the ratio

would have followed the annual development if no change had taken place. However, both the results for monthly and quarterly analysis do not reveal this effect. It is likely that the lack of a seasonal pattern in the model for annual development is involved in this difference.

If parameter significance was given it would be possible to assess the effects of the three interventions from *Tables 4.1*, *4.2*, and *4.3*.

The results given in those tables seem to suggest to disregard the analysis based on annual data as the results of the quarterly and monthly based analysis seem to be more in line with each other and, as was already noted, the quarterly and monthly based models are better suited to fit the interventions that all took place somewhere in the middle of a year. Each coefficient in the tables should be tested for significance first, but this has not been done yet. The results are also subject to changes in the time points at which interventions took place.

From the actual estimates it should be noted that all coefficients seem to be quite stable with respect to varying the length of the estimation period (as indicated by ‘lead' in the tables). The magnitude of the parameters should be compared to the level components or may be expressed as relative changes in either the level components or the real life outcomes the components are designed for. However, coefficients can be compared on a per column basis in the tables. That is, between other interventions over the same level components. If anything can be concluded from the results it would be the following:

- The analysis based on quarterly data seems to give the most stable results in the sense that parameter estimates seem to be less sensitive to the length of the period used for estimation, as indicated by the lead. This does not necessarily indicate that analysis based on quarterly data is better than analysis based on monthly or annual data.

- Since the majority of the parameters in the third and fourth column of *Table 4.1* are negative while those in the fifth column are all positive, the introduction of alcohol law decreased both the number of accidents and the number of victims per accident, but did not decrease the number of fatalities per victim.

- Similarly inspecting the signs of the parameter values in *Table 4.2*, the seat belt law only had a beneficial effect on the number of fatalities per victim. Note that both interventions took place in a relatively short time frame; therefore, results are likely to be entangled. In this case it would be very advantageous to know in advance what relative benefits should occur and test this relative effect, both in terms of timing and in terms of magnitude. Note that the effects of the seat belt law and the alcohol law on the ratio of the number of fatalities to victims almost level out in the 10 year lead analysis of both monthly and quarterly data. As the alcohol law was introduced ahead of the seat belt law, this means that the ratio went up just ahead of the seat belt law. It is not clear how this phenomenon is to be explained. In the case of the number of accidents and the ratio of the number of victims to accidents, the development was reversed: first a shift down and then a (smaller) shift up. In all cases it is unclear how to interpret these developments. One explanation might be

that the up or down shifts at the time of the seat belt interventions were not related to the introduction of the seat belt law at all. This could mean that the effects turn out to be different when another time point for the up shift is chosen. This suggests that the estimated effects of the alcohol law are sensitive to such timing issues. This is an issue that should be resolved.

The Mulder Law does not seem to have had a beneficial effect on any of the outcomes.

4.3. Conclusion

This study shows moderate success of the state space approach by Harvey & Durbin (1986) in determining the effect of various traffic safety measures, although no definitive conclusions can be drawn. Other features of the state space approach, such, as prognosis have not explicitely been addressed in this study.
On the basis of this study it seems reasonable to conclude that it is crucial, in disentangling the effects of measures, to either be able to quantify as many properties of the interventions in terms of effectiveness as possible, or alternatively be able to estimate those properties. Two very important properties are the timing and, more generally, the development over time of the effects. Ideally both approaches, estimation, and quantification should be taken. Mis-specification of the shape and timing may lead to biased estimates of effects of measures.

5. Conclusions and recommendations

5.1. Conclusions

In this study some statistical issues involved in the simultaneous analysis of accident-related outcomes (such as the number of victims, fatalities or accidents) of the traffic process were studied. The main focus of this study was on estimation and how to deal with the supposed covariation of the outcomes. Correction for covariation is needed in order to enhance the statistical reliability of techniques applied to the simultaneous analysis of accident-related outcomes. The scope of techniques is not restricted to the analysis of time dependent data.

Estimates of the covariances of normal approximations of joint distributions were derived in this study. This has been done for the following cases:

a. The total number of accidents, victims and fatalities in a certain class. Examples of classes are monthly data, car-only accidents.
b. The logarithm of the total number of accidents, victims and fatalities in a certain class.
c. The logarithm of the total number of accidents, the logarithm of the ratio of the number of victims to the number of accidents, the logarithm of the ratio of the number of fatalities to the number of victims.

The results are listed in *Table 3.2* in *Chapter 3*, details are described in the *Appendix A.1 and A.2*. The alternative to normal approximations, so-called ‘exact' results were not studied.

The following conclusions are drawn from *Chapter 3* and the *Appendix*:

- It is possible to derive relatively simple expressions for the variances and covariances of (logarithms and ratios of) accidents and victim counts. The results are in *Table 3.2*. Some information needed to compute estimates of the covariance matrices may not be available over a longer period of time.

- When it is intended to analyse traffic unsafety outcomes like accidents, victims and fatalities or any of the other combinations as mentioned above simultaneously, all three unsafety outcomes have to be used: each unsafety outcome has unique information that cannot be estimated from the other two outcomes.

- It turns out that the logarithm of the total number of accidents is (approximately) uncorrelated to the others.

- The quality of the estimates is evaluated using samples of real life data from the Netherlands. The following aspects have been studied:
  - simulation studies based on real life data support results; deviations from the assumptions do not have serious consequences. More results are in *Appendix A.3*.
  - the approximation of the logarithm needed when log-numbers are analysed is studied for accidents only. In practice problems are unlikely when counts are higher than approximately 30.

The following conclusions are drawn from *Chapter 4*:

- The trial study shows moderate success of the state space approach by Harvey & Durbin (1986) in determining the effect of various traffic safety measures, although no definitive conclusions can be drawn.

- The trial study suggests that it is better to analyse quarterly or even monthly data than annual data, even if annual data need to be predicted.

- The estimated coefficients in the trial study show differentiated effects of the alcohol law and seat belt law. Both interventions took place in a rather short time frame. It is essential to accurately determine the timepoint at which each of the interventions took place and, if possible, to determine the relative effects the intervention should have on the relevant components of the state space. This would allow for a better determination of the timing of the interventions and would allow for better disentangling of the intervention effects.

5.2. **Recommendations**

Based on the results of this study it can be recommended not to use so-called Poisson approximations of the variance of for instance the number of victims in a year by estimating its value using the observed number of victims. The variance may be substantially larger. The amount of ‘extra' variance depends on the distribution of the number of victims per accident. When more victims tend to occur in certain types of accident the variance of the number of victims tends to be higher. It seems better to approximate this variance by the sum of the square of the number of victims per accident rather than by the sum of the number of victims per accident.

This figure however may not be available for older observations, as data on the accidents level may not be available for older accidents. The few examples studied here do not indicate a substantial change in the ratio of for instance the sum of the square of the number of victims per accident to the sum of the number of victims per accident over the last 25 years. This suggests that it may be possible to estimate this ratio and use the (recorded) total number of victims instead.

In multivariate analysis of the traffic safety indicators investigated in this report it is recommended to use the covariances as described in *Table 3.2.*

References

Cameron, A.C. & Trivedi, P.K. (1998). *Regression analysis of count data*. Cambridge University Press, Cambridge.

COST329. *Draft report of cost-329, models for traffic and safety development and interventions*. Technical report, European Union, Directorate General for Transport, Brussels. [In preparation].

Feller, W. (1968). *An introduction to probability theory and its applications, volume I*. Third edition. John Wiley & Sons, Inc., New York.

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, Cambridge.

Harvey, A.C. & Durbin, J. (1986). *The effects of seat belt legislation on british road casualties: A case study in structural time series modelling.* Journal of the Royal Statistical Society A, 149 (3): pp. 187-227.

Summala, H. & Näätäänen (1988). *The zero-risk theory and overtaking decisions*. In: Rothengatter, J. A. & Bruin, R. A. de (editors). Road User Behaviour; Theory and research, pp. 82-92. Van Gorkum, Assen/Maastricht.

Wilde, G.J.S. (1994). *Target risk.* PDE Publications, Toronto.

Appendix Technical details

A.1. Derivation for counts

This section describes how the covariance (matrix) between the number of injury accidents and victims is derived. It is assumed that a basic simplification can be used: the number of victims per accident is equally distributed for all accidents, although it may be possible to relax this assumption. No further assumptions on the shape have been made. The derivation extends the result in Feller (1968).

A.1.1. *The expected value and variance of the number of accidents and victims*

Define N as the number of accidents in a certain period of time. N is assumed to be Poisson distributed with parameter λ.

The stochastic variables V_i (i = 1,...,N) denote the number of victims in accident i. The V_i are assumed to be independently identically distributed. The distribution of the V_i has characteristic function $\varphi(t)$ with expected value μ. The symbol v_i is used to denote a realisation of the number of victims in accident i. Similarly, the symbol f_i is used to denote a realisation of the number of fatalities in accident i.

Let $V = \Sigma^{N}_{i=1} V_i$, thus V is a sum over a random number (N) of accidents. Define $\Phi(t)$ the characteristic function of V then

$$\Phi(t) = \mathrm{E}\,(e^{i\,t\,V}) = \mathrm{E}\,(\mathrm{E}\,(e^{i\,t\,V}|N))$$

where i is the imaginary number ($i^2 = -1$).

$$E(e^{itV}\,|\,N = n) = E\left(e^{it\sum_{k=1}^{n} V_k}\,|\,N = n\right)$$

$$= E\left(\prod_{k=1}^{n} e^{itV_k}\right) = \prod_{k=1}^{n} \phi(t) = \phi^n(t) \qquad \text{(A.1)}$$

then (because N follows a Poisson distribution)

$$\Phi(t) = E\left(\phi^N(t)\right) = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n\phi(t)^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda\phi(t))^n}{n!}$$

Using $e^x = \Sigma^{\infty}_{n=0}\,(x^n)/n!$ the result is (Feller, 1968, page 286):

$$\Phi(t) = e^{-\lambda+(\lambda\varphi(t))} = e^{\lambda(\varphi(t)-1)} \qquad \text{(A.2)}$$

As $E(|V|^3)$ exists, $E(V) = i^{-1}\Phi'(0)$ and $E(V^2) = -\Phi''(0)$, using $\varphi(0) = 1$, $\Phi(0) = 1$, $\varphi'(0) = i\,E(V_k) = i\,\mu$ and $\varphi''(0) = -E(V_k^2)$, we get the following expected value for V:

$$E(V) = i^{-1}[\lambda\varphi'(t)\Phi(t)]_{t=0} = i^{-1}\lambda\varphi'(0) = \lambda\mu \tag{A.3}$$

This can (no surprise here) be estimated using:

$$\hat{E}(V) = \sum_{k=1}^{N} v_k \tag{A.4}$$

The variance of V is $\sigma^2(V) = E(V^2) - E^2(V)$, thus

$$E(V^2) = -[\lambda\varphi''(t)\Phi(t) + (\lambda\varphi'(t))^2\Phi(t)]_{t=0} = -\lambda\varphi''(0) - (\lambda\varphi'(0))^2$$

resulting in

$$\sigma^2(V) = \lambda E(V_k^2). \tag{A.5}$$

This can be estimated using

$$\hat{\sigma}^2(V) = \sum_{k=1}^{N} v_k^2 \tag{A.6}$$

The expected value and the variance of the number of fatalities can be derived in the same way.

A.1.2 *The covariance between the number of accidents and the number of victims*

The covariance between the number of injury accidents and the number of victims is more complicated. Its derivation is based on the same characteristic function argument as used above. The characteristic function of the random vector (N,V) is defined as

$$\Phi(s,t) = E(e^{isN + itV}) \equiv E(f(N) \times g(V))$$

using the same conditional expectation trick:

$$E(E(f(N) \times g(V)|N)) = E(f(N)\,E(g(V)|N))$$

then using (*A.1*) we get

$$\Phi(s,t) = E\left(e^{isN}\phi^N(t)\right) =$$

$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\phi(t)e^{is})^k}{k!} = \tag{A.7}$$

$$= e^{-\lambda}e^{\lambda\phi(t)e^{is}} = e^{\lambda(\phi(t)e^{is}-1)} \tag{A.8}$$

In order to derive the covariance, $\mathrm{E}(N) = \lambda$ by the Poisson law and $\mathrm{E}(V) = \lambda\mu$ is already available in (*A.3*). Needed is:

$$E(NV) = -\left[\frac{\partial^2 \Phi(s,t)}{\partial s\,\partial t}\right]_{s=t=0} \tag{A.9}$$

Continuing, it is found

$$\frac{\partial \Phi(s,t)}{\partial s} = i\lambda\phi(t)\Phi(s,t)$$

$$\frac{\partial \Phi(s,t)}{\partial t} = \lambda e^{is}\phi'(t)\Phi(s,t)$$

so

$$\frac{\partial^2 \Phi(s,t)}{\partial s\,\partial t} = i\lambda\phi(t)\lambda e^{is}\phi'(t)\Phi(s,t) + i\lambda\phi'(t)\Phi(s,t) =$$

$$= i\lambda\Phi(s,t)\phi'(t)(\phi(t)\lambda e^{is} + 1)$$

because $\Phi(0,0) = \varphi(0) = 1$, $\varphi'(0) = i\mu$ it follows that

$$\left[\frac{\partial^2 \Phi(s,t)}{\partial s\,\partial t}\right]_{s=t=0} = i^2\lambda\mu(\lambda + 1)$$

and

$$\mathrm{E}(N\,V) = \lambda\mu\,(\lambda+1) \tag{A.10}$$

and thus

$$\mathrm{Cov}(N,V) = \lambda^2\mu + \lambda\mu - \lambda\lambda\mu = \lambda\mu \tag{A.11}$$

A.1.3. *The covariance between the number of victims and the number of fatalities*

Let the random variable F_i be the number of fatalities in accident i. Let $F = \Sigma^N_{i=1} F_i$. Define $\Psi(s,t)$ the characteristic function of the random vector (V,F) and $\psi(s,t)$ the characteristic function of the random vector (V_i,F_i), thus for each i and j, (V_i,F_i) is independent of (V_j,F_j) if $i \neq j$ but the V_i and F_i are not independent, as almost surely $V_i \geq F_i$. Then

$$\Psi(s,t) = \mathrm{E}(e^{itV + isF}) = \mathrm{E}(\mathrm{E}(e^{itV + isF}|N)),$$

then similar to (*A.1*),

$$E\left(e^{itV+isF} \mid N = n\right) = E\left(e^{\sum_{k=1}^{n}(itV_k+isF_k)} \mid N = n\right)$$

$$= E\left(\prod_{k=1}^{n} e^{itV_k+isF_k}\right) = \prod_{k=1}^{n}\psi(s,t) = \psi^n(s,t) \quad \text{(A.12)}$$

also *e*

$$\Psi(s,t) = E\left(\psi^N(s,t)\right) = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\psi^n(s,t)\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda\psi(s,t))^n}{n!}$$

$$\Psi(s,t) = e^{-\lambda+(\lambda\psi(s,t))} = e^{\lambda(\psi(s,t)-1)}. \quad \text{(A.13)}$$

Similar to (*A.10*) it follows:

$$\mathrm{E}(V\,F) = \lambda^2\mathrm{E}(F_i)\,\mathrm{E}(V_i) + \lambda\mathrm{E}(F_i\,V_i) \quad \text{(A.14)}$$

and:

$$\mathrm{Cov}(V,F) = \lambda\mathrm{E}(F_i\,V_i). \quad \text{(A.15)}$$

This can be estimated with

$$\hat{Cov}(V,F) = \sum_{i=1}^{N} f_i v_i.$$

A.2. **Derivation for the logarithm of counts**

A.2.1. *The expected value and variance of the logarithms of number of accidents and victims*

Unfortunately, it is not possible to derive an explicit characteristic function as simple as the one given in equation (*A.2*) in the case of the logarithm of the number of accidents and victims. For that reason, approximations need to be made in order to get a useful expression for the covariance between the logarithm of the number of accidents and victims. This is done using a method often called the 'delta' method. The basic idea is that the logarithms of N and V are approximated by a series expansion of order k (usually order one) about their expected values. This would mean that log(N) is approximated by a polynomial in N of order k, say
$\log(\mathrm{N}) \approx a_0 + a_1(N-\lambda) + \ldots\ldots + a_k(N-\lambda)^k$.

To be precise, a first order approximation about the expected value (λ) of the number of accidents is:

$$\log(N) \approx \log(\lambda) + \frac{N-\lambda}{\lambda} = \log(\lambda) - 1 + \frac{N}{\lambda}$$

This makes the expected value of this first order approximation equal to $\log(\lambda)$: $\mathrm{E}(\log(N)) \approx \log(\lambda)$. Similarly, $\mathrm{E}(\log^2(N)) \approx 1/\lambda + \log^2(\lambda)$ which together makes

$$\hat{\sigma}^2(\log(N)) \approx \frac{1}{\lambda} \qquad \text{(A.17)}$$

where $\approx$ means variance of the first order approximation.

In the case of $\log(V)$, approximations are about the expected value $\lambda\mu$ of V: $\log(N) \approx (\log(\lambda\mu)\text{-}1) + V/\lambda\mu$. For that reason $\mathrm{E}(\log(V)) \approx \log(\lambda\mu)$ and using (*A.5*)

$$\sigma^2(\log(V)) \approx \frac{\sigma^2(V)}{(\lambda\mu)^2} = \frac{E\left(V_k^2\right)}{\lambda\mu^2}$$

Results for fatalities are derived in a similar way.

Obviously, both results for the logarithmic case are approximations. From *Figure A.1* it can be seen that the relative approximation error for the variance of the number of accidents may turn out substantial if λ is less than about 20-30. Similar results will hold for victims and fatalities.

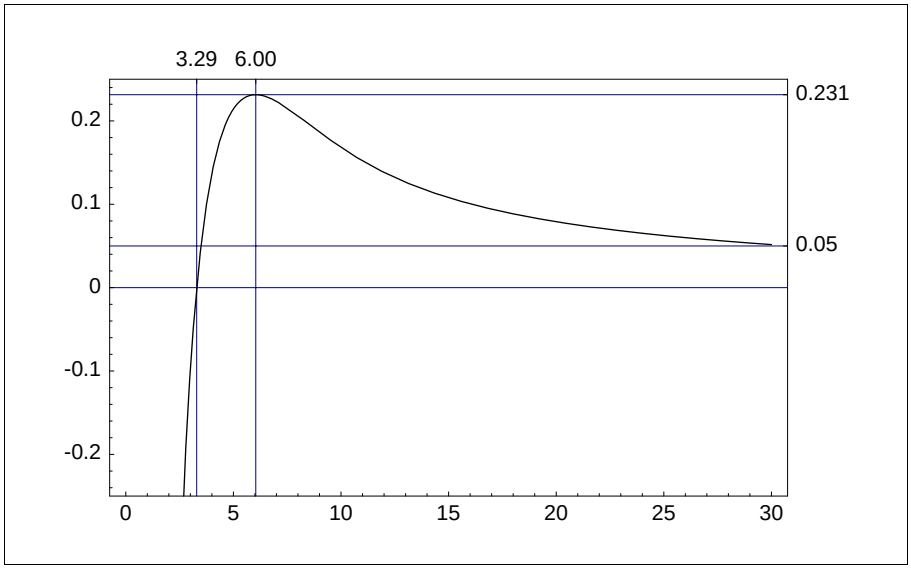

Figure A.1. *The relative error of the variance estimate of the logarithm of a Poisson distributed random variable as a function of its expected value λ (horizontal axis). The relative error is computed as (exact-estimate)/exact*.

A.2.2. *The covariance between the logarithm of the number of injury accidents and the logarithm of the number of victims and fatalities*

Extending the first order approximations of both log-accident counts and log-victims, it can be seen that

$$Cov(\log(N),\log(V)) = E(\log(N)\log(V)) - E(\log(N))E(\log(V))$$

$$\approx \sum_{i,j=0}^{k} a_i b_j E(N^i V^j) - \left(\sum_{i=0}^{k} a_i E(N^i)\right)\left(\sum_{j=0}^{k} b_j E(V^j)\right)$$

Expressions like $E(N^i V^j)$, $E(N^i)$ and $E(V^i)$ can be evaluated similar to (*A.9*) using higher order derivatives of the characteristic function in (*A.8*).
The resulting equations are derived in *Table A.2*. They are based on first order approximations. *Table A.1* gives an explanation of abbreviations used in *Table A.2*.

Number of	Stochastic symbol	Realisation	Usually available	Abbreviation
Accidents (acc)	N	n	Yes	n
Victims in accident i	V_i	v_i	No	
Fatalities in accident i	F_i	f_i	No	
Sum over all accidents of the number of	Stochastic symbol	Estimate	Usually available	Abbreviation
Victims (vic)	V	$\Sigma_{i=1}^{n} v_i$	Yes	Σv
Fatalities (fat)	F	$\Sigma_{i=1}^{n} f_i$	Yes	Σf
Sum over all accidents of the square of the number of	Expected value	Estimate	Usually available	Abbreviation
Victims	$\lambda E(V_i^2)$	$\Sigma_{i=1}^{n} v_i^2$	No	Σv^2
Fatalities	$\lambda E(F_i^2)$	$\Sigma_{i=1}^{n} f_i^2$	No	Σf^2
Sum over all accidents of the cross product of the numbers of	Expected value	Estimate	Usually available	Abbreviation
Victims and fatalities	$\lambda E(V_i F_i)$	$\Sigma_{i=1}^{n} v_i f_i$	No	$\Sigma f v$

Table A.1. *Abbreviations used in the derived equations for variances and covariances and estimates*.

Results based on counts		
Variance of	Value	Estimate
acc	λ	n
vic	$\lambda \mathrm{E}(V_k^2)$	$\sum v^2$
fat	$\lambda \mathrm{E}(F_k^2)$	$\sum f^2$
Covariance of	Value	Estimate
acc & vic	$\lambda \mathrm{E}(F_k V_k)$	$\sum v$
acc & vic	$\lambda \mathrm{E}(V_k)$	$\sum f$
vic & fat	$\lambda \mathrm{E}(F_k)$	$\sum f v$
Results based on logarithms of counts		
Variance of	Value	Estimate
acc	$1/\lambda$	$1/n$
vic	$\mathrm{E}(V_k^2)/\lambda \mathrm{E}(V_k)^2$	$\sum v^2/(\sum v)^2$
fat	$\mathrm{E}(F_k^2)/\lambda \mathrm{E}(F_k)^2$	$\sum f^2/(\sum f)^2$
Covariance of	Value	Estimate
acc & vic	$1/\lambda$	$1/n$
acc & fat	$1/\lambda$	$1/n$
vic & fat	$\mathrm{E}(F_k V_k)/(\lambda \mathrm{E}(F_k) \mathrm{E}(V_k))$	$\sum f v/(\sum v \times \sum f)$
Results based on logarithms of ratios of counts		
Variance of	Value	Estimate
acc/vic	$\sigma^2(V)/\lambda \mathrm{E}(V)^2$	$\sum v^2/(\sum v)^2 - 1/n$
fat/vic	$\sigma^2(F)/\lambda \mathrm{E}(F)^2 - 2\mathrm{Cov}(F,V)/(\lambda \mathrm{E}(F) \mathrm{E}(V)) + \sigma^2(V)/\lambda \mathrm{E}(V)^2$	$\sum f^2/(\sum f)^2 + \sum v^2/(\sum v)^2 - 2\sum f v/(\sum v \sum f)$
Covariance of	Value	Estimate
acc & acc/vic	0	0
acc & fat/vic	0	0
acc/vic & fat/vic	$\mathrm{Cov}(F,V)/(\lambda \mathrm{E}(F) \mathrm{E}(V)) - \sigma^2(V)/\lambda \mathrm{E}(V)^2$	$(\sum f v)/(\sum v \sum f) - (\sum v^2)/(\sum v)^2$

Table A.2. *Derived equations for variances and covariances and estimates.*

A.3. **Simulation studies**

To support the findings, a simulation study was conducted. From injury accidents that occurred in the Netherlands in the years 1980 - 1999 the number of victims (at least one) and the number of fatalities was recorded for each individual accident as well as the month and year it occurred. A separate simulation was done using accidents that only involved cars, one that involved only fatal accidents and one that involved fatal car-only accidents. All were performed by selecting with replacement a random number of accident records from a specific month. The number to be selected was a random number sampled from a Poisson distribution with expected value equal to the number of accidents that actually occurred that

particular month. Such a selection should look similar to what could have happened that month. For each sample created that way the total number of accidents, victims and fatalities have been computed, as well as some other statistics. Covariances have been computed using a large number of such samples (at least 40,000) that were created that way.

It should be noted that this sampling scheme implies a Poisson distribution of the number of accidents and that, therefore, the following checks on the estimation of the variance of the number of accidents are more or less, but not intended to be, a check on the performance of the random number generator.

Table A.3 compares results for car-only simulations with all vehicle type simulations. It seems that the results are similar except for log-fatalities. This is not well understood, at least not yet. It may be that this has at least something to do with the approximation, as it does not occur in the cases of var(fat) and relatively little fatalities occur in this case. See also *Figure A.1*.

Measure	Only cars	All accidents
Var(acc)	0.02246	0.02160
var(vic)	0.02216	0.02266
var(fat)	0.02210	0.02261
cov(acc,vic)	0.02394	0.02334
cov(acc,fat)	0.13625	0.09611
cov(vic,fat)	0.06753	0.06803
var(log(acc))	0.02257	0.02168
var(log(vic))	0.02215	0.02272
var(log(fat))	0.05702	0.02297
cov(log(acc),log(vic))	0.02397	0.02343
cov(log(acc),log(fat))	0.13783	0.09549
cov(log(vic),log(fat))	0.04260	0.06740

Table A.3. *Standard deviations of the relative differences* $(\hat{e}\text{-}e)/\hat{e}$ *between simulation sample estimates of the measures 'ê' and computed estimates 'e', 40,000 simulations, 1980-1999.*

A.3.1. *Graphical comparison of derived results and simulations*

Besides computing covariances between the number of accidents, victims and fatalities from the samples and comparing those to the derived estimates, it may be wise to look at how well the covariances describe the (co)variability of the data.

At least from one point of view (i.e., the fact that no less fatalities can occur than fatal accidents) it can be argued that the applicability of the covariance estimates as a measure of covariability is limited.

This section is intended to indicate this problem. To this end the data of only 1 month (December 1999) out of 240 are used to compare the derived covariance results with the sampled data in a graphical way.

It is expected that the description of the variability as derived in this document works best when numbers are larger. At the one end (with the larger numbers) the number of injury accidents and the number of victims are compared. At the other end the number of fatal car accidents and the number of fatalities are compared.

From *Figure A.2* it can be seen that the covariance between the number of victims (horizontal) and the number of injury accidents is substantial. One could be tempted to conclude that this means that one could use only one to describe both. This may not be the correct conclusion. Changes over time are the subject of study and therefore it is necessary to know how much (co)variation can be expected by nature.

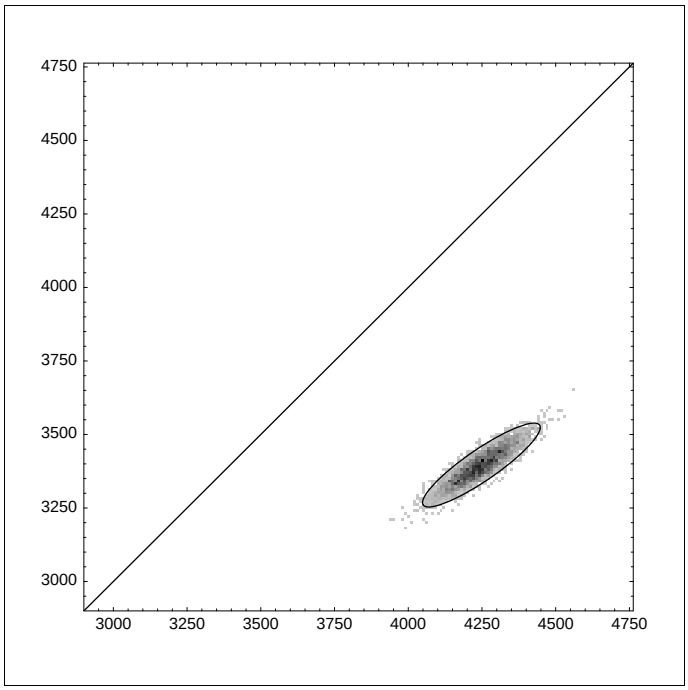

Figure A.2 *Density plot of frequencies of the number of victims (horizontal) and the number of injury accidents from the samples of injury accidents. The inside of the ellipsis is the 95% confidence region based on a multivariate normal distribution with expected values and covariance from Table A.2*.

From *Figure A.3* it can be seen that the confidence region extends into the zone in which more fatal accidents ‘occur' than fatalities. Based on the approximation there would be a chance of about 0.16 that more fatal accidents occur than the number of fatalities. Also, variance seems to be spread more than is estimated (see *Tables A.2*, *A.3*). This phenomenon is less clear in earlier months in which more accidents and fatalities occurred. How serious the consequences of this spread are has not been studied yet.

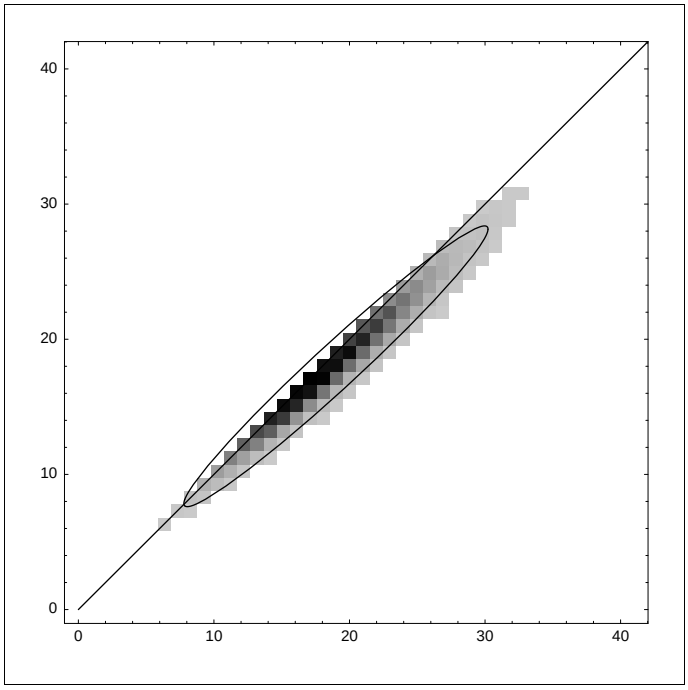

Figure A.3. *Density plot of frequencies of the number of fatalities (horizontal) and the number of fatal accidents from the samples of fatal car-only accidents. The inside of the ellipsis is the 95% confidence region based on a multivariate normal distribution with expected values and covariance from Table A.2*.

A.3.2. *Some interpretations on the derived covariance matrices*

Using *Table 3.2* it is possible to study the covariance matrices (of logarithms) of the number of accidents, victims and fatalities in more detail.

The covariance matrix for counts

The estimate of the covariance matrix Σ can be computed as follows. For each accident i denote the vector $c_i = (1, v_i, f_i)'$ of counts of accidents (= 1), victims and fatalities involved in the accident.

Then

$$\hat{\Sigma} = \sum_{i=1}^{n} c_i c_i'.$$

The components $c_i\, c_i'$ are rank-one matrices. Therefore it may in principle be possible in some cases to represent the space of accidents, victims and fatalities in less than three dimensions. If this were possible, it would mean that not all three components need to be explained if the safety development is to be analysed, simplifying models. In order for this to be true however, it would require limited variability in the vectors c_i. For instance $f_i \equiv 0$ for all accidents would be one obvious case. Usually, however, there will not be a clear simplification available.

The covariance matrix for log-counts

Compiling the results for *Table 3.2* on logarithms the following is found (approximately):

$$\frac{1}{\lambda}\begin{pmatrix} 1 & 1 & 1 \\ 1 & \dfrac{E(V^2)}{E^2(V)} & \dfrac{E(FV)}{E(F)E(V)} \\ 1 & \dfrac{E(FV)}{E(F)E(V)} & \dfrac{E(F^2)}{E^2(F)} \end{pmatrix} \tag{A.18}$$

No simplifications have been found here besides the ones that will apply in the previous (count based) matrix.

When, instead of logarithms of the number of accidents, injuries and fatalities the logarithms of the number of accidents, the number of injuries divided by the number of accidents and the number of fatalities divided by the number of injuries are modelled, the following covariance matrix is obtained (approximately):

$$\begin{pmatrix} \dfrac{1}{\lambda} & 0 & 0 \\ 0 & \dfrac{E(V^2)}{\lambda E^2(V)} - \dfrac{1}{\lambda} & \dfrac{E(FV)}{\lambda E(F)E(V)} - \dfrac{E(V^2)}{\lambda E^2(V)} \\ 0 & \dfrac{E(FV)}{\lambda E(F)E(V)} - \dfrac{E(V^2)}{\lambda E^2(V)} & \dfrac{E(F^2)}{\lambda E^2(F)} - \dfrac{2E(FV)}{\lambda E(F)E(V)} + \dfrac{E(V^2)}{\lambda E^2(V)} \end{pmatrix} \tag{A.19}$$

This matrix can be rearranged:

$$\frac{1}{\lambda}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \dfrac{\sigma^2(V^2)}{E^2(V)} & \dfrac{Cov(FV)}{E(F)E(V)} - \dfrac{\sigma^2(V^2)}{E^2(V)} \\ 0 & \dfrac{Cov(FV)}{E(F)E(V)} - \dfrac{\sigma^2(V^2)}{E^2(V)} & \dfrac{\sigma^2(F^2)}{E^2(F)} - \dfrac{2Cov(FV)}{E(F)E(V)} + \dfrac{\sigma^2(V^2)}{E^2(V)} \end{pmatrix} \tag{A.20}$$

Hence - at least in first order approximation - the logarithm of number of accidents and the logarithm of the ratio of either accidents to injuries or injuries to fatalities are uncorrelated.

Using data of 1980 through 1999 car-only accidents), the quantities $\mathrm{E}(V^2)$, $\mathrm{E}(F^2)$ and $\mathrm{Cov}(F,V)$ as well as $\mathrm{E}(V)$, $\mathrm{E}(F)$ have been estimated.

For practical purposes however, first estimates for the quantities $\sigma^2(V)/\mathrm{E}^2(V)$, $\mathrm{Cov}(F,V)/(\mathrm{E}(F)\,\mathrm{E}(V)) - \sigma^2(V)/\mathrm{E}^2(V)$ and $\sigma^2(F)/\mathrm{E}^2(F) - 2\mathrm{Cov}(F,V)/(\mathrm{E}(F)\,\mathrm{E}(V)) + \sigma^2(V)/\mathrm{E}^2(V)$ are graphed in the *Figures A.4, A.5* and *A.6* respectively. These are the relevant components of the covariance matrix (*A.15*) multiplied by λ.

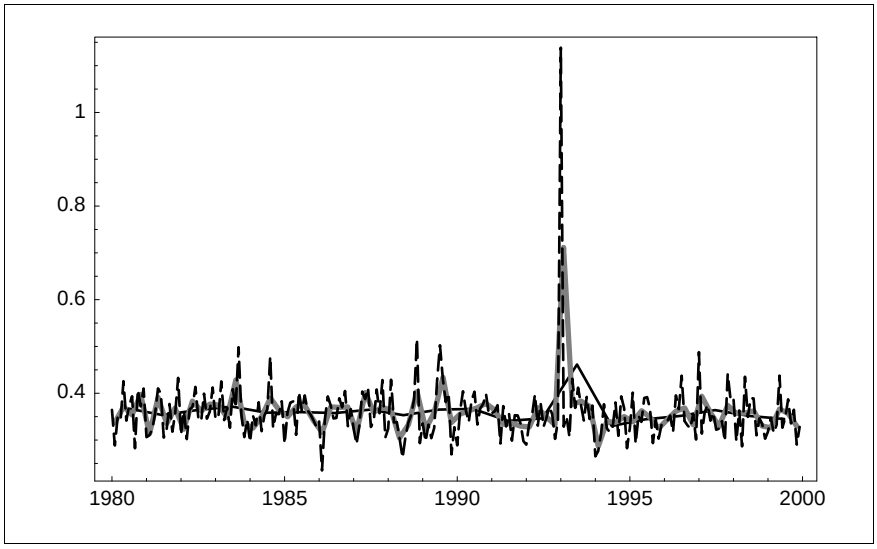

Figure A.4. *Development of $\sigma^2(V)/E^2(V)$ over time for car-only accidents. Annual data: uninterrupted line, Quarterly data: uninterrupted grey line, Monthly data: dashed line. The peak in the beginning of 1993 is probably entirely caused by a number of fog related pile-ups in the first few hours of 1993*.

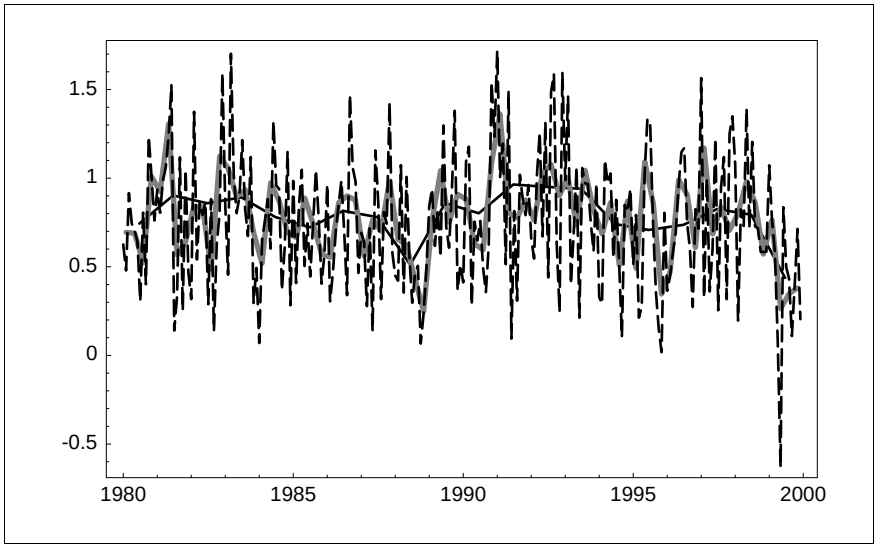

Figure A.5. *Development of $Cov(F,V)/(E(F)\,E(V))$ - $\sigma^2(V)/E^2(V)$ over time for car-only accidents. Annual data: uninterrupted line, Quarterly data: uninterrupted grey line, Monthly data: dashed line*.

The increasing tendency in *Figure A.6* seems to be caused by $\sigma^2(V)/\mathrm{E}^2(V)$ only, as *Figures A.4* and *A.5* do not observe this tendency.

One option to explain this phenomenon could have been an actual increase in the variance of the number of fatalities per accident while the mean number of fatalities per accident approximately remains the same. This however is contradicted by the results in *Figure A.7*. In this figure, the indexed annual development over time of $\mathrm{E}^2(F)$ and $\sigma^2(F)$ are graphed.

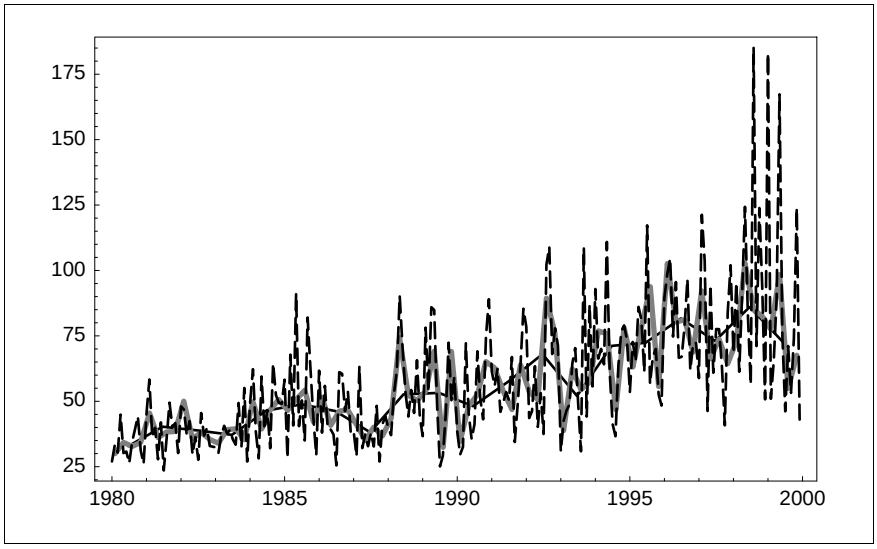

Figure A.6. *Development of $\sigma^2(F)/E^2(F) - 2Cov(F,V)/(E(F)\,E(V)) + \sigma^2(V)/E^2(V)$ over time for car-only accidents. Annual data: uninterrupted line, Quarterly data: uninterrupted grey line, Monthly data: dashed line.*

It shows that both components decline but $E^2(F)$ declines more rapidly so their ratio actually increases. Apparently, somehow the square of the average number of fatalities per accident dropped steeper than the average of the square of the number of fatalities per accident.

One consequence of this is that it should be checked whether or not the variance can be approximated using an estimate that is not dependent on not generally available components like $\sigma^2(F)$.

Figure A.7. *Annual indexed (1980) development of $E^2(F)$ (uninterrupted line) and $\sigma^2(F)$ (dashed line) over time*.