International orientation on methodologies for modelling developments in road safety

Martine Reurings & Jacques Commandeur

R-2006-34

SWOV
INSTITUTE FOR
ROAD SAFETY RESEARCH

International orientation on methodologies for modelling developments in road safety

R-2006-34

Martine Reurings & Jacques Commandeur

Leidschendam, 2007
SWOV Institute for Road Safety Research, The Netherlands

Report documentation

Number:	R-2006-34
Title:	International orientation on methodologies for modelling developments in road safety
Author(s):	Martine Reurings & Jacques Commandeur
Project leader:	Paul Wesemann
Project number SWOV:	40.103
Keywords:	Safety, mathematical model, forecast, injury, development, accident rate, method, statistics, United Kingdom, Belgium, Canada, France, Sweden.
Contents of the project:	This report gives an overview of the models developed in Belgium, Canada, France, Great Britain and Sweden to evaluate past developments in road traffic safety and to obtain estimates of these developments in the future.
Number of pages:	47
Price:	€ 11,25
Published by:	SWOV, Leidschendam, 2007

This publication contains public information.
However, reproduction is only permitted with due acknowledgement.

SWOV Institute for Road Safety Research
P.O. Box 1090
2260 BB Leidschendam
The Netherlands
Telephone +31 70 317 33 33
Telefax +31 70 320 12 61
E-mail info@swov.nl
Internet www.swov.nl

Summary

This report gives an overview of the models developed in countries other than the Netherlands to evaluate past developments in road traffic safety and to obtain estimates of these developments in the future. These models include classical linear regression and loglinear models as applied in Great Britain, and the ARIMA and DRAG models used in Belgium, Canada, France and Sweden.

The linear regression models for Great Britain were used to forecast the number of road crash casualties of different severities in a future year (2000 and 2010). In the model used to predict the number of casualties in 2000, the year 1983 played an important role. In this year compulsatory seat-belt wearing was introduced and this turned out to be of great influence on the number of casualties. To predict the number of casualties in 2010, the effect of three road safety measures was first examined, and the number of casualties over the years was estimated if these measures had not yet been introduced. Based on these results a prognosis was made for 2010. A totally different model, i.e. not a linear regression model, was used to forecast the number of crashes and casualties for drivers older than 60 over 20 years. This model consisted of three submodels, describing the predicted number of older drivers, the predicted number of crashes involving older drivers, and the number of casualties in crashes involving older drivers respectively.

An important problem with classical linear and loglinear regression applied to time series data is the assumption of independence of the observations. However, repeated observations over time are usually not independent at all, since last year's number of casualties is often quite a good predictor for current year's number of casualties. In a classical linear regression this is reflected in residuals that are serially correlated. This results in statistical tests whose standard errors are too small, and therefore in overoptimistic conclusions about the relations between variables that evolve over time. This in turn results in forecasts that are flawed. ARIMA and DRAG models, on the other hand, do take the dependencies between observations into account.

All the developed DRAG models have basically the same structure, they all consist of several layers. The first layer describes the road demand, either expressed as the total road mileage or as the total fuel sales. The next layer is dedicated to the explanation of the number of crashes and victims. Finally, in the last layer the severity of the crashes is explained, where the severity is expressed as the number of persons injured per crash with bodily injury and the number of persons killed in an crash with bodily injury. Each layer consists of one or more models, each of them containing a large amount of explanatory variables, varying from weather conditions to economic activities.

DRAG models have several disadvantages. First, being extended ARIMA models, the observations should be stationary. This means that they must have constant mean and variance over time. Since this is not the case for time series, the observations should be filtered first. Another disadvantage is the large amount of explanatory variables. For forecasting purposes the future values of all these variables need be modelled separately.

SWOV uses structural time series are used to describe, explain and forecast developments in Dutch road traffic safety. This type of time series does not have the disadvantages described above: structural time series do not require stationarity, and the explanatory and dependent variables are modelled simultaneously.

Contents

1. Introduction

Two objectives of the SWOV's Road Safety Assessment Department are:

- to build explanatory models for disaggregated developments of road traffic safety in the Netherlands;
- to obtain forecasts for the future development of road traffic safety based on the modelled disaggregated developments in the past.

Two aspects are essential in setting up explanatory models for the analysis of developments in road safety. The first is that a theoretical (conceptual) model needs to be designed, indicating which explanatory variables are assumed to influence developments in road safety; the second aspect is the choice of the analysis technique that is most suitable for the empirical validation of the relations hypothesized in the theoretical model. In the current research program we have worked on these two aspects in separate projects.

In a different project from the one presented in this report, a *theoretical model* was developed based on existing knowledge about the factors that affect the occurrence of road accidents of a certain severity. At the core of this model is the following equation: accidents = exposition x risk, as applied within a certain time and space interval. The objective is to add explanatory variables to this equation which affect road safety within that same time and space interval. Since an explanatory variable usually does not affect all types of accidents, it is considered important to disaggregate the entire traffic process.

In the present study, the main focus has been on obtaining an inventory of the *analysis techniques* used internationally to empirically validate the relations between exposure and risk on the one hand, and explanatory variables on the other, for disaggregated types of accidents. The objective was to see whether we could learn from the experiences of researchers in other countries on this subject.

In presenting the relevant studies we will not only discuss which analysis techniques were used, but also at what level of disaggregation they were applied, and which explanatory variables were included. Studies solely concerned with the analysis of accident data (and which did not even consider exposure data) were excluded from the inventory.

We were particularly interested to see whether other countries had used structural time series techniques for the analysis of their road safety data, since previous SWOV research found that this technique is best suited for the analysis of time series data (Bijleveld & Commandeur, 2006).

Several studies corresponding to the two objectives of SWOV's Road Safety Assessment Department were found. The methodological approaches used in Great Britain are the most helpful in achieving these objectives and will hence be extensively discussed. The methodologies used in Canada, Belgium, Sweden, France, and Spain will also be reviewed and presented.

In this report a model is defined as one or more mathematical equations which relate road safety to several variables. A method or methodology will entail the way in which the model is developed, for example which technique is used to estimate the parameters of the models.

2. Great Britain

For Great Britain three models will be discussed which forecast the road safety at different disaggregation levels. These models were developed by Broughton (1991), Broughton et al. (2000) and Maycock (2001).

Broughton (1991) developed four models to forecast the number of road crash casualties of different severities in 2000. For this purpose he studied British casualty trends since 1949. He paid special attention to certain statistical questions affecting the confidence intervals for the forecasts. The model and the methodology used are described in *Section 2.1*.

Section 2.2 discusses the methodology of Broughton et al. (2000). This was used to forecast the number of casualties at three severity levels in 2010 for five road user groups and two road type categories. This methodology takes into account the effect of several road safety measures.

Finally, Maycock (2001) developed a model to forecast the road safety of a specific group of the population in Great Britain, the over-60s. This model consists of three submodels, predicting the number of older drivers, the number of crashes they are involved in and the resulting number of casualties, disaggregated by gender and age group. This model is discussed in *Section 2.3*.

2.1. Forecasting the number of casualties in the year 2000

2.1.1. *The basic model*

Several authors have noted that for several decades the fatality rate per hundred million vehicle kilometres travelled has been decreasing by an almost constant percentage each year. The statistical strength of this relation in Great Britain was shown by Broughton (1988), with the model taking two specific safety measures into account: the introduction of compulsory seat belt wearing in 1983 and the very short impact of drink-driving legislation in 1968.

Broughton (1991) extended the analysis of Broughton (1988) to data from the period 1986-1989 using the basic model:

$$\log\left(\frac{C}{T}\right) = a + b \cdot y + s + \varepsilon, \tag{2.1}$$

where C is the total number of casualties in year y, T is the total traffic in year y, a, b are the model parameters to be estimated, ε is the error term and s measures the effect of the compulsory seat belt wearing, i.e., $s = 0$ prior to 1983 and $s = s'$ from 1983. The short impact of the drink-driving legislation in 1968 was allowed for by removing this year from the data.

Four separate models of the form *(2.1)* were fitted for:

- the number of fatalities;
- the number of killed and seriously injured persons (KSI);
- the number of all casualties;
- the number of all injury crashes.

The model parameters were determined by ordinary linear regression. Confidence intervals were computed for the parameter estimates and for the annual rate of decline of the casualty rate, which can be computed as:

$$B = 100 \cdot e^{-b}.$$

The fit of the models to the data was good. The proportion of variance explained was higher than 0.98 for each model. This high value can also be a consequence of misinterpreting purely random variation as systematic variation, see Fridstrøm et al. (1995).

2.1.2. *Analysis of the residuals*

Before using the models for forecasting purposes, the residuals were analysed. This analysis consisted of two parts: studying whether the rate of decline did change and checking for autocorrelation.

The first part of the analysis is important because model *(2.1)* assumes that the rate of decline is constant. So if the rate of decline did change, the model should be altered in such a manner that it takes this change into account. This was the case for the model for the number of fatalities and for the model for KSIs, the killed and seriously injured. Indeed, residuals indicated that these numbers decreased more rapidly over recent years starting in 1983. To test this the model *(2.1)* was altered into

$$\log\left(\frac{C}{T}\right) = a + b \cdot y + c \cdot (y + 3) + s + \varepsilon, \tag{2.2}$$

where $c = 0$ prior to 1983 and $c = c'$ afterwards. This model assumes that the annual change in the dependent variable was b before and $b + c'$ after 1983. A c' deviating significantly from zero hence indicates that a change in trend did occur. Such a c' was found for the models for the fatality and KSI rate.

The presence of autocorrelation can be tested by plotting the residuals against time. For the casualty and injury crash rate the residuals resulting from the original model were used, whereas for the fatality and KSI rate the residuals from model *(2.2)* were used. The model for the KSI rate seemed to overestimate the rate for certain periods and to underestimate it for others, which is evidence for autocorrelation or misspecification. Additional (formal) tests were conducted, which confirmed that the model for the KSI rate indeed was affected by autocorrelation. The tests also showed that the models for the casualty and injury crash rate were affected by autocorrelation. This means that for these three models the parameter estimates can still be quite accurate, but that the computed standard errors may be too small.

The problem of autocorrelation occurring in the models for the casualty and injury crash rate were solved by transforming the dependent variable according to the following transformation:

$$z'(y) = z(y) - \rho z(y - 1),$$

where z is a variable depending on time y, ρ is a measure for the degree of autocorrelation estimated from the data and z' is the transformed variable. The residuals of the transformed models were found to be free of autocorrelation, so the estimates resulting from the models can be accepted. The parameter estimates changed a little, whereas the standard errors were 2.5 times larger than the standard errors for the original estimates.

2.1.3. *The forecasts of casualties*

The developed models were used to forecast the casualty rates for future years, under the assumption that road safety measures were introduced at the same rate and with the same effectiveness as in the past. For the forecasts of fatality and KSI rates model *(2.2)* was used, whereas model *(2.1)* was used for the other two rates. From the predicted casualty rates the predicted number of casualties can be computed if the future traffic growth is known or can be estimated.

Estimates of the future traffic growth were based on a prediction of the Department of Transport. This prediction was that the mileage would grow between 1989 and 2000 by 40% and 23% under two economic scenarios that represent optimistic and pessimistic combinations of assumptions about growth in the Gross Domestic Product and fuel prices. The resulting casualty forecasts were expressed as percentages of the averages for 1981-1985, which is a baseline often used in Great Britain for evaluating progress in reducing casualties.

2.2. **Forecasting the number of casualties in the year 2010**

In October 1997 the Department of Transport (DOT) of Great Britain announced the intention of setting a new national casualty reduction target for 2010. Following this announcement the DOT established the ‘Safety Targets and Accident Reduction’ (STAR) Group, consisting of eight subgroups, to work on the development of the new road safety strategy and the numerical targets. The results derived by the subgroup concerned with the numerical context for the new targets were reported by Broughton et al. (2000). The numerical context was provided by forecasting the number of casualties in 2010 under appropriate assumptions about the traffic volume in that year. The used forecasting method will be discussed in the next subsections.

2.2.1. *The basic model*

Broughton et al. (2000) adopted the linear relationship between the logarithm of the casualty rate (defined as the number of casualties of a specific severity per measure of traffic volume) and time found in previous studies, including Broughton (1991). The basic model of Broughton et al. (2000) is hence of the form

$$\log\left(\frac{F_t}{M_t}\right) = a + b\,t + \varepsilon_t, \quad t = 1, \ldots, n, \tag{2.3}$$

where F_t is the number of traffic casualties of a specific severity in year t and M_t is the amount of traffic volume in year t.

Model *(2.3)* can be used to calculate the number of casualties for future years. This deduction involves estimating the parameters a and b based on data from the past by ordinary linear regression and then computing

$\log(F_t/M_t)$ for a year in the future using the estimated values of a and b. The simple equality

$$M_t \cdot e^{\log\left(\frac{F_t}{M_t}\right)} = M_t \cdot \frac{F_t}{M_t} = F_t.$$

shows that the number of casualties in year t can be computed once M_t is known.

2.2.2. *Disaggregations*

Broughton et al. (2000) developed separate models, all of the form given in *(2.3)*, for the following disaggregations:

- severity of injuries, consisting of three categories: killed, seriously injured, slightly injured;
- group of road user, consisting of five categories: car occupants, pedestrians, bicyclists, motorcyclists (including users of mopeds and scooters and other two-wheeled motor vehicles), other (a small group including people travelling by bus, coach, van or lorry);
- road type, consisting of two categories: urban and rural.

This adds up to a total of thirty models.

To deduce the number of casualties from these models, different traffic volumes M_t were used for different groups of road users. It was defined to be

- the volume of car traffic in year t for car occupants;
- the total traffic volume in year t for pedestrians, bicyclists and other;
- the volume of motorcycle traffic in year t for motorcyclists.

2.2.3. *Road safety measures and their effects on road safety*

It is not sufficient to just extrapolate the developed basic models to forecast the number of casualties in 2010. Indeed, these basic models just show the trend in the past and they do not take explicit account of road safety policies. This section will describe how future measures can be incorporated in the forecasting method.

There is a wide range of possible road safety measures which influence the number of casualties. It is impossible to take them all into account. Therefore Broughton et al. (2000) only chose three types of measures which are known to have a significant effect on the casualty rates. These types of measures are:

- improved standards of passive safety in cars;
- measures to reduce the level of drink-driving;
- road safety engineering.

Broughton et al. (2000) described what the number of casualties since 1983 might have been if these measures had not yet been introduced. These adjusted numbers indicated the effect of all other road safety activities together. These other road safety activities were referred to as the core road safety activities.

Here it will only be explained in which way the effect of measures to reduce the level of drink-driving and the effect of the core road safety activities can be quantified. The computations will be illustrated with the Dutch data in *Table 2.1*.

Year	F_t	F_t^{alc}	p_t^{alc}	$\hat{F}_t^{\text{alc}}$	% increase of casualties due to alcohol without measures	M_t
1985	1408	540	0.383523	540.00	0.00	60.6
1986	1442	586	0.406380	553.04	-5.62	63.2
1987	1355	405	0.298893	519.67	28.31	64.8
1988	1315	382	0.290494	504.33	32.02	70.6
1989	1334	437	0.327586	511.62	17.08	71.2
1990	1310	387	0.295420	502.42	29.82	73.6
1991	1271	384	0.302124	487.46	26.94	72.1
1992	1195	297	0.248536	458.31	54.31	76.4
1993	1145	317	0.276856	439.13	38.53	73.8
1994	1141	333	0.291849	437.60	31.41	73.6
1995	1296	370	0.285494	497.05	34.34	77.5
1996	1297	383	0.295297	497.43	29.88	78.7
1997	1243	370	0.297667	476.72	28.84	80.3
1998	1175	326	0.277447	450.64	38.23	82.2
1999	1275	386	0.302745	488.99	26.68	85.3
2000	1269	322	0.253743	486.69	51.15	85.9
2001	1194	281	0.235343	457.93	62.96	87.0
2002	1172	297	0.253413	449.49	51.34	88.7
2003	1184	330	0.278716	454.09	37.60	89.4

Table 2.1. *Example data for the computation of the effect of alcohol on casualties*

The basic assumption in the computations is that the proportion of casualties from crashes in which alcohol played a part should be equal for the years following 1985 if no measures against drink-driving are implemented after 1985. Because the proportion of alcohol-related casualties in 1985 is equal to

$$p_{1985}^{\text{alc}} = \frac{F_{1985}^{\text{alc}}}{F_{1985}} = \frac{540}{1408} = 0.383523,$$

the number of alcohol-related casualties for $t = 1986, \ldots, 2003$ without measures having been taken can be computed as

$$\hat{F}_t^{\text{alc}} = p_{1985}^{\text{alc}} \cdot F_t = 0.383523 \cdot F_t.$$

This implies for example that $\hat{F}_{2003}^{\text{alc}} = 0.383523 \cdot 1184 = 454.09$, and hence an increase of $454.09 - 330 = 124.09$ casualties due to alcohol if no measures are taken.

The sixth column of *Table 2.1* contains the yearly percentages with which the real number of casualties due to alcohol should be multiplied to get the expected number of casualties due to alcohol if no measures against drink-driving would have been implemented. These percentages can be computed as

$$100 \cdot \left(\frac{\hat{F}_t^{\text{alc}} - F_t^{\text{alc}}}{F_t^{\text{alc}}} \right).$$

One of the conclusions which can be drawn from *Table 2.1* is that the measures against alcohol in traffic have led to a reduction of 37.60% of casualties in the year 2003.

Analogously the effects of safety of cars and the improvement of infrastructure were estimated.

Now that the effects of measures against drink driving, of safety of cars and of the improvement of infrastructure are known, the effects of all other road safety activities together (the core road safety activities) can be estimated. For this estimation, the mobilities M_t for the example data in 1985 and 2003 are needed. The mobilities in 1985 and 2003 are M_{1985} $=$60.6 and M_{2003} $=$89.4. If in the period 1985-2003 there had been no improvement in road safety the casualty rate F_{2003}/M_{2003} should be equal to the casualty rate F_{1985}/M_{1985} which is computed to be 23.23. Under the assumption that the number of casualties is reduced by respectively 2%, 7% and 3% due to the three types of safety measures mentioned before, the casualty rate in 2003 would not be 23.23 but

$$\frac{F_{1985}}{M_{1985}} \cdot (1 - \frac{2}{100}) \cdot (1 - \frac{7}{100}) \cdot (1 - \frac{3}{100}) = 20.54.$$

This formula is based on the assumption that effects of measures are independent. So every reduction percentage is applied to the number of casualties already reduced by the other two types of measures. For the unknown reduction percentage x of the core road safety activities the following equation should hold:

$$\frac{F_{2003}}{M_{2003}} = \frac{F_{1985}}{M_{1985}} \cdot (1 - \frac{2}{100}) \cdot (1 - \frac{7}{100}) \cdot (1 - \frac{3}{100}) \cdot (1 - \frac{x}{100}),$$

from which it follows that

$$x = 100 \cdot \left(1 - \frac{\frac{F_{2003}}{M_{2003}}}{\frac{F_{1985}}{M_{1985}} \cdot (1 - \frac{2}{100}) \cdot (1 - \frac{7}{100}) \cdot (1 - \frac{3}{100})} \right) = 35.52.$$

So the reduction percentage of the core road safety activities is, in this example, equal to 35.52%.

It should be remarked here that this estimate of x is only based on the data for the years 1985 and 2003. The data for all the other years is ignored. Hence it can be expected that a better estimate can be found by using all the data.

2.2.4. *Baseline prognoses for 2010*

From the data in *Table 2.1* it is possible to compute the total number of casualties under the assumption that no measures against drink-driving were implemented. Indeed, this number, denoted by $\hat{F}_t$, equals

$$\hat{F}_t = F_t - F_t^{\text{alc}} + \hat{F}_t^{\text{alc}}.$$

Now two models, both of the form *(2.3)*, can be considered. The first one is based on the real data F_t and the other one on the estimated data $\hat{F}_t$. Linear regression gives the following results:

$$\log\left(\frac{F_t}{M_t}\right) = 58.726 - 0.028t + \varepsilon_t,$$

$$\log\left(\frac{\hat{F}_t}{M_t}\right) = 49.156 - 0.023t + \varepsilon_t.$$

Based on these two models the prognosis for the casualty rate in 2010 can be made under the assumption that the core road safety activities stay on the level of 2003 and that no more measures against drink-driving will be taken. This prognosis is made by drawing a line from the point on the first regression line corresponding to $t = 2003$ up to $t = 2010$ parallel to the second regression line, see *Figure 2.1*.

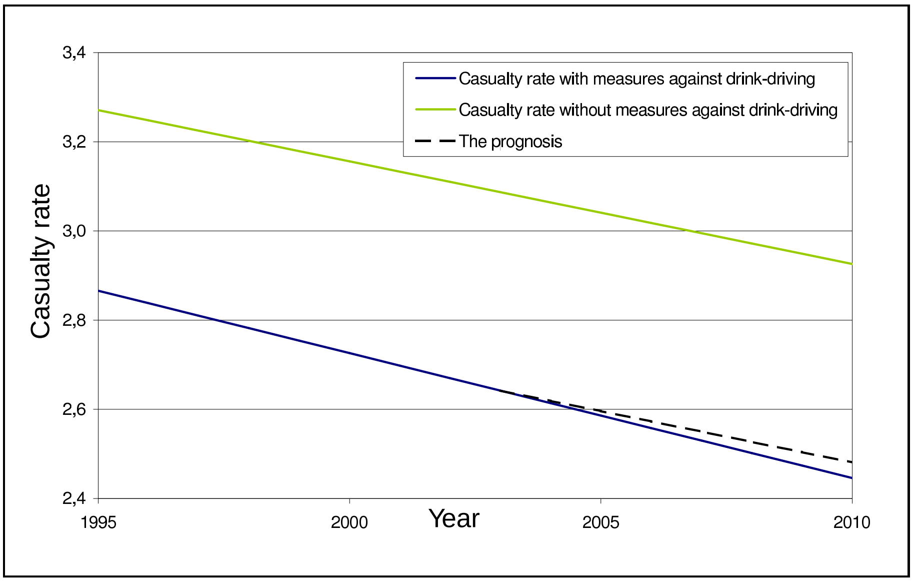

Figure 2.1. *The casualty rates with and without measures against drink-driving and the prognosis for 2003 and further years based on them.*

The prognosis for the casualty rate in 2010 can be used to deduce the expected number of casualties in 2010. The expected number of casualties depends on the chosen future scenario for the M_t. Broughton et al. (2000) identified several possibilities for the future developments of the mobility of the different groups of road user. For car traffic the following four scenarios were investigated:

- a fast increase of mobility;
- a central increase of moblity;
- a slow increase of mobility;
- no change in mobility.

These forecast come from the 1997 national road traffic forecasts from the Department of the Environment, Transport and the Regions. The central increase is the most likely outcome, the fast and slow increases correspond to some confidence bounderies.

For the other road user groups similar scenarios were identified, based on knowledge of past trends. The combined scenario for all mobilities should not contradict the linkages which exists between the different road user groups. For example, if it is assumed that the car traffic stays on the same level, then it is reasonable to assume that there will be more walking and cycling.

2.2.5. *Adding extra measures to the baseline prognoses for 2010*

After the prognoses have been determined for the numbers of casualties in 2010 for several subgroups, the effects of additional measures can be computed as:

$$\hat{F}_{2010} \cdot (1 - \mu_1) \cdot (1 - \mu_2) \cdot \ldots \cdot (1 - \mu_m).$$

Here μ_i is the estimated reduction factor of measure i. So if measure 2 is expected to reduce the number of casualties by 2% then $\mu_2 =$0.02. The number of casualties in 2010 when measure 2 is implemented is then equal to $\hat{F}_{2010} \cdot 0.98.$

2.3. **Forecasting older driver crashes and casualties**

The number of older drivers will increase considerably over the next years, due to several reasons. First of all, during the last decades the life expectancy increased, which resulted in a faster growth (in terms of percentage) of the older groups in the population than the population as a whole. Secondly, the proportion of the older members of the population who have a driving licence increased over the years and is expected to continue to do so. This is particularly true for women.

This increase in the number of older drivers will cause a road safety problem, because older drivers have a higher crash rate than other drivers. The British Department of the Environment, Transport and the Regions funded a study of older drivers. One part of this study, which is the subject of this section, aimed at forecasting the numbers of crashes and casualties for older drivers (over 60) by gender and age group over the next 20 years (Maycock, 2001).

2.3.1. *The structure of the model*

The model used by Maycock (2001) for the prediction of the number of crashes involving older drivers and the resulting casualties consists of three submodels. The first submodel is given by:

> Predicted number of older drivers = (Predicted number of people in the relevant sector of the population) $\times$ (Predicted proportion of drivers in this sector).

The predictions for the population were published by the Government Actuary's Office. The prediction of the proportion of drivers was based on the data for the years 1973 up to 1997. The methodology for this prediction will be discussed later.

Using the outcome of the previous submodel, the number of crashes involving an older driver was computed via:

> Predicted number of crashes involving older drivers = (Predicted number of older drivers) × (Predicted crash rate of older drivers).

The crash rate of drivers is defined as the number of crashes per year a driver is expected to be involved in as a driver. It was estimated from past national crash data taking into account the facts that crash rates have changed over the years and that the mileage that drivers drive has increased over the years.

Finally, the predicted number of crashes involving older drivers was used to compute the casualties caused by older drivers as follows:

> Predicted number of casualties caused by older drivers= (Predicted number of crashes involving older drivers) × (Predicted casualty rate in this type of crashes)

Here the casualty rate is defined as the number of casualties per accident.

2.3.2. *Disaggregations*

The predicted number of drivers, crash involvements and casualties was computed for both male and female drivers. For each gender the computations were carried out for seven age groups: 60-64, 65-69, 70-74, 75-79, 80-84, 85-89 and 90 and older. The number of crashes was computed for two severities separately: KSI (killed and seriously injured) crashes and slight crashes. The same disaggregation was used for the number of casualties.

2.3.3. *Predicting the proportion of licensed drivers*

From the structure of the model sketched in *Section 2.3.1* it follows that the first step in forecasting the road safety of older drivers was predicting the proportion of drivers for both genders and each older age group. The horizontal thick line in *Figure 2.2* divides the drivers in older and younger groups, whereas the vertical thick line divides time in past and future. Hence the first step was to predict the proportion of drivers in the cells in the right bottom rectangular of the figure. The development of the proportion of drivers can be described in two ways, which are illustrated by the horizontal and diagonal arrows in *Figure 2.2*.

The horizontal arrows in *Figure 2.2* illustrate the change of the proportion of drivers in a certain age group, say for example the group 60-64, over the years. The relationship between the proportion of drivers in the population and time was assumed to be a logistic curve described by:

$$P_A(t) = \frac{1}{k + be^{-at}}, \tag{2.4}$$

where P_A is the proportion of drivers in a certain age group A and $a > 0, b$ and k are model parameters. So in year t the proportion of drivers in age group 60-64 is denoted by $P_{60-64}(t)$. The upper limit of the proportion of drivers in each age group is given by:

$$P_{A,\text{lim}} = \lim_{t\to\infty} P_A(t) = \frac{1}{k}.$$

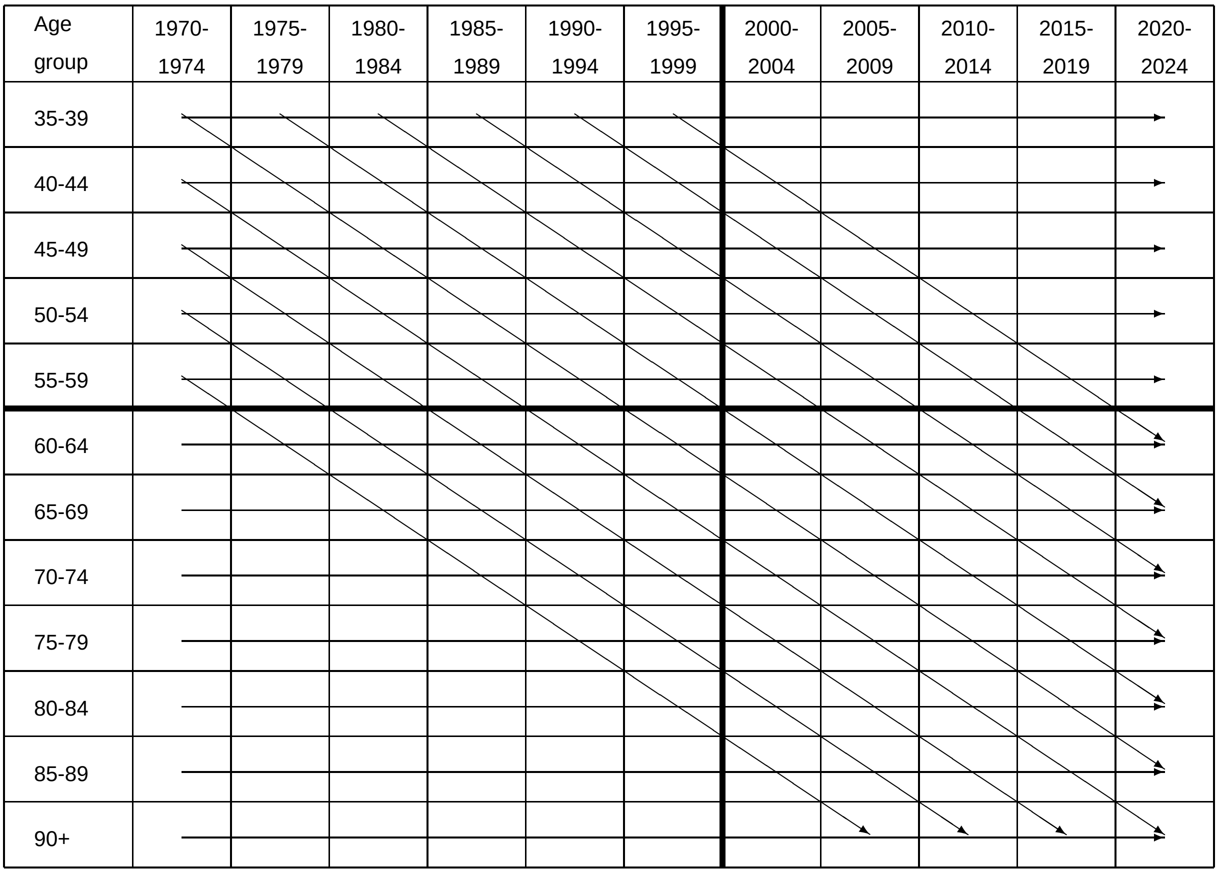

Figure 2.2. *Illustrating the principles involved in the prediction of the proportion of drivers in the population.*

The diagonal arrows in *Figure 2.2* illustrate the development of a cohort, which is defined to be a certain age group in a certain 5-year period. For example, the group 35-39 in the period 1970-1974 is the same cohort as the age group 40-44 in the period 1975-1979 and so on. The proportion of drivers in a cohort changes due to people getting their licence and people giving up driving. At first the proportion increases rapidly because most people are getting their licence just after becoming 17. Then the increase becomes smaller, until it balances the rate at which drivers are giving up driving. At this time the proportion of drivers in the cohort has reached its maximum, denoted by $P_{C,\text{max}}$. After reaching this maximum the proportion of drivers will decrease. This decrease was assumed to be represented by a curve in which $1 - P_C/P_{C,\text{max}}$ increases exponentially, where P_C is the proportion of drivers in a certain cohort C. In formula:

$$1 - \frac{P_C}{P_{C,\text{max}}} = de^{c(\text{Age} - A_0)}, \tag{2.5}$$

where A_0 is the age at which the maximum proportion is reached. It turned out that the age of 65 is appropriate.

From *Figure 2.2* it follows that the data to the left of the vertical thick line can be used to estimate the parameters in *(2.4)* and *(2.5)*, and hence to predict the proportion of drivers in the future. The logistic form is the most convenient form to predict the future trends, but it was hard to determine the value of $P_{A,\text{lim}}$ for many of the age/gender groups. This problem was solved by using the cohort approach, which makes it possible to estimate the proportion of drivers in 2020-2024. So the proportion of drivers in all age groups was known for the periods up to and including 1995-1999 and it was estimated by the cohort approach for 2020-2024. Now the logistic function passing through these two points could be calibrated, resulting in a convincing estimate of $P_{A,\text{lim}}$.

The process of estimating the proportion of drivers in 2020-2024 by the cohort method consisted of two steps: first the value of $P_{C,\text{max}}$ was estimated, then the parameters A_0, c and d were be determined. The proportion of drivers in each age group was known for the 5-year periods before 2000. A quadratic regression of the proportion of drivers in an age group on time was carried out on this data. The resulting relationship was used to calculate $P_{C,\text{max}}$. Plotting the values of $1/P_{C,\text{max}}$ against the cohorts implied that $P_{C,\text{max}}$ followed a logistic curve. The results of fitting a curve to these points are:

$$\frac{1}{P_{C,\text{max}}} = \begin{cases} 1.03 + 0.022e^{0.070\cdot C}, & \text{for male drivers,} \\ 1.03 + 0.103e^{0.086\cdot C}, & \text{for female drivers,} \end{cases}$$

where C is a number corresponding to each cohort, such that $C = 0$ for the cohort of age 35-39 in 1995-1999 and C increases by units of five for each cohort. Thus $C = 55$ for the cohort of age 90+ in 1995-1999. In order to be able to predict the proportion of drivers for each age group over 60 the values of $P_{C,\text{max}}$ should be calculated for the cohorts with age 35-39 up to 65-95 in 1995-1999, which can be done using the equations above.

As mentioned earlier the variable $P_{C,\text{max}}$ denotes the highest possible proportion of drivers in cohort C. Once $P_{C,\text{max}}$ is reached the proportion of drivers in the cohort will decrease. This decrease is described by the function in *(2.5)*. The identification of the parameters in this function was done two steps:

1. First the ratios $P_C/P_{C,\text{max}}$ were calculated for the cohorts with age 70-74 up to 90+ in 1995-1999, using the estimated values of $P_{C,\text{max}}$. The ratios exceeding 1 were omitted.
2. The parameters of the function were determined by plotting $\ln(1 - P_C/P_{C,\text{max}})$ for the same cohorts as in the first step against the mid-point age of the cohort minus a suitable value for A_0. The value for A_0 was chosen to be 65.

There turned out to be no statistical significant difference between the plots for female and male drivers, hence the data were combined. The function which describes the development of the proportion of drivers (male and female together) in a cohort is given by

$$P_C = P_{C,\text{max}}(1 - 0.02e^{0.14(\text{AGE}-65)}),$$

where AGE is the mid-point age of the age groups. This function was expected to change over time due to an increase in life expectancy. This increase was expected to be 2.2 years between 1994 and 2031 for men and women, which corresponds to 0.3 years per five year period. By including a term in the driving increment function this change was taken into account. The driving increment function becomes

$$P_C = P_{C,\text{max}}(1 - 0.02e^{0.14(\text{AGE}-65-0.3N)}),$$

where N is the number of future 5-year periods with $N = 0$ for 1995-1999.

Using this function the proportion of drivers in each age group above 60 years of age in 2020-2024 was estimated.

Next the logistic prediction curve given by *(2.4)* was computed. Because this curve had three parameters (a, b, k) and had to pass through two points, the value of one parameter had to be chosen. In this case a was supposed to be the same as for the curves based on the given initial data. Next b and k were estimated for each age group, for male and female drivers separately.

2.3.4. *Predicting the crash rate of drivers*

As mentioned before, the crash rate is defined as the number of crashes per year a driver is expected to be involved in. Because crash rate changes over time, it should be modelled. Under the assumption that the crash rate decreases with a constant multiplier per year, the model for the crash rate is given by:

$$\text{Crash rate} = A_0(\text{Age, Gender})e^{-b\text{Year}},$$

where b is the fractional reduction per year, Year is the year in which the crashes took place and A_0 is a constant which is dependent on the age group and gender being considered. The value of b was assumed to be the same for each age group but different for the two genders. The values of A_0 and b were estimated on basis of crash data by severity and crash type over the years 1986-1997.

At first a total of 90 models were fitted, namely separate models for male and female drivers, the nine age groups and for the following five categories:

- KSI crashes;
- slight crashes;
- single-vehicle crashes with a pedestrian (all severities);
- single-vehicle crashes with a cyclist (all severities);
- single-vehicle crashes with any other object (all severities).

It was tested if the age effect for the different types of crashes were different. This turned out not to be the case, so the last three categories were no longer considered as different crash types.

It is logical to expect that the crash rate also depends on the mileage travelled by a certain age/gender group in a year, so the previous crash rate models could be improved by including the average annual mileage appropriate to each driver group. To do so, the average annual mileage should be known. The models for the average annual mileage were based on several National Traffic Surveys for male and female licence holders separately and for the various age groups. The most appropriate models are:

$$\text{Annual mileage} = (32.014 - 564\text{Age} + 2.4\text{Age}^2)e^{0.000237\text{Age}\cdot\text{Year}},$$

for male licence holders, and

$$\text{Annual mileage} = (9.557 - 129\text{Age} + 0.32\text{Age}^2)e^{0.000237\text{Age}\cdot\text{Year}},$$

for female licence holders, where Year is the actual year minus 1992. The mileage was included in the models for crash rate in two ways:

$$\text{Crash rate} = A_1(\text{Age, Gender})Me^{-b\text{Year}}$$

and

$$\text{Crash rate} = A_2(\text{Age, Gender})M^{0.3}e^{-b\text{Year}},$$

where M is the average annual mileage appropriate to each gender and age driver group and A_1 and A_2 are constants depending on the age group and gender of the driver. The second model is far less sensitive to mileage and errors in mileage than the first model and was hence used for prediction purposes. This corresponds to what was done by for example Maycock, Lockwood & Lester (1991) and Forsyth, Maycock & Sexton (1995).

2.3.5. *Predicting the casualty rate*

Recall that the casualty rate is defined as the number of casualties per accident. Two questions are of interest:
– Do older driver crashes result in a different pattern of casualties from those of younger drivers?
– Has the pattern of casualties changed over the years?
By studying the data for the years 1979, 1983, 1987, 1992 and 1997 it was concluded that there were small differences in casualty patterns between the age groups and between the genders, but the differences were not large in practical terms. Furthermore, the only notable change in crash pattern over the years was the fall in motorcycle casualties, but this was very probably a consequence of the decrease in motorcycling.

Because the casualty patterns were not very different for different age groups and years, the type of crash did not have to be taken into account when estimating the casualty rates for the two severities (KSI and slight) separately. It was only necessary to determine if the casualty rate depends on age group and gender and whether or not it had changed over the years. In order to do so crash and casualty data for the same five years as before were analysed. For each year two ratios were computed:

$$\frac{\text{KSI casualties}}{\text{KSI crashes}} \quad \text{and} \quad \frac{\text{slight casualties}}{\text{all crashes}}.$$

There is a difference in the severity of the crashes by which the number of casualties is divided, because slight injuries can arise from crashes of all severities, while KSI casualties can only arise from KSI crashes. The analysis of these ratios lead to the following conclusions:
– both for KSI and slight casualty rates there was no significant difference between the genders;
– for KSI casualty rates there was no trend over time and the difference between the age groups was of no practical importance;
– for the slight casualty rates there was a trend over time and an age effect.
For the prediction of the number of KSI casualties it was hence assumed that the KSI casualty rate was constant over the years and equal for each age group. This constant value was taken to be the overall average value of the computed ratios, which was 1.22 KSI casualties per KSI accident. The slight casualty rate was described by

$$\text{Slight casualties per accident} = 1.34e^{-0.002\text{Age}+0.01(\text{Year}-1992)}.$$

Now all the necessary information was available to predict the future number of crashes and the resulting casualties for age groups 60-64 up to 90+.

3. Canada

Gaudry (1984) and Fournier & Simard (2002) used what are known as DRAG models to explain the road use demand, the crashes and their severity in the province of Quebec. The first DRAG model, now referred to as DRAG-1, was developed by Gaudry (1984). The Société de l'Assurance Automobile du Québec (SAAQ), requested further development, which resulted in the DRAG-2 model, developed by Fournier & Simard (2002). An overview of both DRAG models is given in this section.

3.1. The basic structure of the DRAG model

The underlying structure of both the DRAG-1 and the DRAG-2 model is quite similar. It can be described as follows:

$$DR \longleftarrow (X^{\text{dr}}), \tag{3.1}$$

$$VI \longleftarrow (DR, X^{\text{vi}}), \tag{3.2}$$

where the arrows means "determines in a certain way" and X^{dr} and X^{vi} stand for collections of explanatory variables for road demand (DR) and for victims (VI) respectively. If a variable belongs to both collections it has a direct effect on VI via the relation in *(3.2)* and an indirect effect via its impact on DR as described in *(3.1)*. Two types of victims were considered, namely the number of injured persons and the number of those killed. One of the differences between DRAG-1 and DRAG-2 is the road use demand. In the DRAG-1 model this is expressed in the consumption of gasoline and diesel fuel, whereas in the DRAG-2 model it is measured by the distance travelled by gasoline-powered and diesel-powered cars separately. A consequence of this difference is that the detailed structure of both models is slightly different.

Both models consist of several layers and equations. In the first two layers of the DRAG-1 model the fuel sales for highway use (gasoline and diesel separately) are extracted from the data on total fuel sales. This is done because the total fuel sales also include the fuel sales for off-highway use, for example agriculture and building sites, which is not a measure for the road demand. The extraction can be done using the following linear form:

$$DC = DNR + DR = \sum_i \beta_i X_i^{\text{dnr}} + \sum_j \beta_j X_j^{\text{dr}} + e^{\text{dr}},$$

where DC is the total fuel sales measured, DNR is the fuel sales for off-highway uses, DR is the fuel sales for highway uses, X_i^{dnr} are the explanatory variables of the sales for off-highway uses, X^{dr} are the explanatory variables for highway uses and e^{dr} is the residual error.

The first layer of the DRAG-2 model consists of two equations which explain the distance travelled by gasoline- and diesel-powered cars separately.

The next three layers in both models are dedicated to the explanation of the number of crashes and victims. The first of these layers explains the following three crash categories:

- property damage only (PDO) crashes;

– crashes with at least one person injured;
– and crashes with at least one person killed;

It does this together with two aggregations, namely:

– crashes with bodily injuries;
– the total number of crashes.

In the following layer the severity of the crashes is explained, expressed in the morbidity (the number of persons injured per crash with bodily injury) and the mortality (the number of persons killed per crash with bodily injury). Now the total number of victims can be computed. For example, the number of injuries is computed as the product of the number of crashes with bodily injury and the morbidity of these crashes. So the three layers explaining the number of victims can be written as:

$$\begin{aligned} AC &\longleftarrow (DR, X^{\text{vi}}) \\ GR &\longleftarrow (DR, X^{\text{vi}}) \\ VI &= AC \cdot GR, \end{aligned}$$

where AC is the number of crashes of a certain type and GR is a certain severity.

The equations in the last layer (i.e., $VI = AC \cdot GR$) are the only deterministic equations of both models. All the other equations are stochastic. The mathematical formulation for each of these stochastic equations is:

$$y_t^{(\lambda_y)} = \sum_{k=1}^{K} \beta_k X_{k,t}^{(\lambda_x)} + u_t, \tag{3.3}$$

where $X_{k,t}$ is the value of the k-th explanatory variable in month t and $y^{(\lambda)}$ denotes the Box-Cox transformation of a variable, which is:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \ln(y), & \lambda \to 0, \end{cases}$$

and where:

$$\begin{aligned} u_t &= \left(e^{\delta_0 + \sum_{m=1}^{M} \delta_m Z_{m,t}^{(\lambda_{zm})}}\right)^{-1/2} v_t, \\ v_t &= \sum_{l=1}^{r} \rho_l v_{t-l} + w_t. \end{aligned}$$

Here $Z_{m,t}, m = 1, \ldots, M$, are the explanatory variables for the variance of u_t. They are chosen from the set of explanatory variables $X_{k,t}$. Furthermore, w_t is white noise. The model parameters β_i, δ_i and ρ_i are estimated simultaneously by the maximisation of the log likelihood function.

3.2. Explanatory variables

DRAG models usually involve a large number of explanatory variables. Therefore it is useful to classify these variables. The classes (with examples between brackets) for the DRAG-1 and DRAG-2 model were:

1. the dependent variables of the first (two) layer(s), which are the gasoline and diesel fuel consumption in the DRAG-1 model and the distance travelled by gas-powered and diesel-powered cars in the DRAG-2 model;
2. the prices (of the fuel, public transport and car maintenance);

3. motorisation:
   - quantity (the number of motor vehicles of various categories);
   - vehicles characteristics (the size of vehicles and presence of seat belts);
4. networks:
   - laws, regulations, police (the speed limit and patrol frequency);
   - levels of services of transports modes (the transit wait time);
   - infrastructure, climate (rain and snowfall per day);
5. consumers:
   - general characteristics (the number of driver's licenses per car);
   - age (the proportion of drivers between age 16 and 24 compared to the total number of driver's license holders);
   - gender (the proportion of pregnant women in the group of women holding a driver's license);
   - ebriety or vigilance (drugs per driver's license);
6. final economical activities and intermediates (the vacation index and Expo 1967);
7. et cetera:
   - administrative decisions that affect the measurement;
   - aggregation (month composition);
   - seasonal and constant (regression constant).

The meaning of a certain explanatory variable is not the same for all layers where it can be found. For example, a variable such as snow explains the level of road use demand in the first (two) layer(s), whereas in the other layers it explains the change in crash probability and crash severity to a given road demand.

3.3. Representation of the results

In this section the results of the DRAG models will be discussed. The results of the DRAG-2 model will be discussed in more detail.

The appropriate Box-Cox transformation and the presence of heteroscedasticity were identified by means of likelihood ratio tests. This test was not used to evaluate the contribution of particular variables. Instead, this was done by comparing the log likelihood L for a reference model to the one obtained by adding one or more variables. The Student's t test was used to examine if the coefficient of a particular variable deviates from zero. It was decided that a coefficient significantly deviates from zero if the Student's t is larger than 2.

The goodness-of-fit of the models (both DRAG-1 and DRAG-2) was measured by two forms of the Pseudo-R^2, which are defined as:

$$\text{Pseudo-(L)-}R^2 = 1 - e^{\frac{2}{N}(L_0 - L)},$$
$$\text{Pseudo-(E)-}R^2 = 1 - \frac{\sum_{t=1+r}^{N}(y_t - \mathbb{E}(y_t))^2}{\sum_{t=1+r}^{N}(y_t - \bar{y})^2}.$$

Here L is the log likelihood of the considered model and L_0 is the log likelihood of the model developed under the assumption that the model is linear, that the results are homoscedastic and independent and that all the coefficients β_k, except the constant, are equal to zero. Furthermore, $\mathbb{E}(y_t)$ is the mathematical expectation of y_t and $\bar{y}$ is the sample mean.

Table 3.1 summarises the results of the DRAG-2 model. It shows the distribution of the explanatory variables by means of their $t-$values. Also the number of autocorrelation coefficients $\rho_1, \ldots, \rho_r$, the values of the Box-Cox transformation parameters λ_x and λ_y, the Pseudo-(E)-R^2 and the number of used observations are given. Observations that were not used correspond to the autocorrelation coefficient of the highest order.

	Distance		Crashes			Severity		Victims	
Elements examined	Gas	Diesel	PDO	Injury	Fatal	Morbidity	Mortality	Injury	Dead
# explanatory var.	37	32	48	47	46	47	48	47	48
# $t-$values ≥ 2	14	8	18	18	17	8	4	16	16
# $t-$values $\in (1, 2)$	3	8	11	14	10	12	13	17	16
# $t-$values ≤ 2	20	16	19	15	19	27	31	14	16
Autocorrelation parameters	5	5	3	4	2	2	2	3	1
Value of λ_y	0.213	0.174	0.241	0.279	0.360	0.539	1.567	0.201	0.367
Value of λ_x	0.213	0.174	0.241	0.279	0.360	0.539	1.567	0.201	0.367
Pseudo-(E)-R^2	0.994	0.993	0.963	0.968	0.898	0.696	0.331	0.964	0.889
# observations	445	445	445	445	445	445	445	445	445
# used observations	432	431	432	433	439	433	434	433	439

Table 3.1. *Summary of the results of the DRAG-2 model.*

The main conclusions that can be drawn from *Table 3.1* are:

1. For the equations concerning the distance travelled almost half of the variables had a low Student t. This means that the corresponding coefficients did not significantly deviate from zero. The Pseudo-(E)-R^2 was over 99% for both equations, which means that the models reproduced the data very accurately. Both λ_x and λ_y were approximately equal to 0.2 which indicates that the models were close to the logarithmic function.
2. For the crashes and victims a large proportion of Student t values was higher than 1. For PDO crashes, bodily crashes and injured persons the Pseudo-(E)-R^2 was over 96%, so these models fitted the data reasonably well. The values of λ_x and λ_y were between 0.20 and 0.28 for the non-fatal crashes and injured victims, hence the corresponding models were close to the logarithmic function. For fatal crashes and killed victims the values of λ_x and λ_y were close to 0.36. However, the distance travelled was treated differently than the other explanatory variables. It was included in the model in linear form ($\lambda = 1$) and in quadratic form ($\lambda = 2$).
3. For morbidity and mortality there was a high proportion of unimportant variables. The values of these variables did not fluctuate much. Also the goodness-of-fit was not very good, which is shown by the Pseudo-(E)-R^2 values of 33% and 69%. The values λ_x and λ_y were equal to 0.539 for the morbidity. This indicates that the model was halfway between the logarithmic and linear form.

3.4. Forecasts for the period of 1997-2004

Fournier & Simard (2002) obtained forecasts for the years 1997-2004. For this purpose, the values of the model parameters had to be known, as well as the values of the explanatory variables for the forecast period. The values of the model parameters were determined by using data from December 1956 up to December 1993. Then by using the data from explanatory variables for the future, forecasts were obtained.

For the two types of distance travelled, for PDO crashes, for bodily injury crashes, and for injured persons these forecasts were realistic and acceptable. However, the predicted numbers for fatal crashes and for killed persons were very small and hence unlikely to be the right numbers. Therefore, data for the years 1994, 1995 and 1996 were used to re-estimate all the parameters of the models equations, using the same explanatory variables, the same autocorrelation structure and the same values for the Box-Cox parameters. Only the value of λ corresponding to distance travelled was allowed to be different than its initial value of 2. Indeed, the predicted number of fatal crashes and killed persons became higher, while the values of all the other dependent variables remained almost the same.

4. Belgium

In Belgium two models were developed describing the influence of several variables on the number of crashes and their severity. The first model is based on the DRAG model, which was described in *Chapter 3*, and was developed by Van den Bossche & Wets (2003). However, there are some differences with the original DRAG model: exposure was included only as an explanatory variable, not as a dependent variable and the severity of crashes was not defined as morbidity and mortality, but as the number of killed and injured persons.

The other model was developed by Van den Bossche, Wets & Brijs (2004). Instead of a DRAG model, a multiple regression model with ARIMA errors was used. Not only was it attempted to explain the past development of traffic safety, forecasts were also made for a twelve months out-of-data-set period. Because data was not available, exposure was not included in the model as an explanatory variable.

The models will be discussed simultaneously in this chapter.

4.1. The data

To keep a balance between availability of information and variability in the variables monthly data were used. For the DRAG type model data was available from 1986 up to 2000, whereas the model with ARIMA errors was fitted on data from January 1974 up to December 2000. Here the last year was used for forecasting purposes.

Both models had four dependent variables, but they were slightly different. The dependent variables of the DRAG type model were:

- the number of crashes with injured persons;
- the number of crashes with persons killed;
- the number of persons injured;
- the number of persons killed;

And the variables of the ARIMA model were:

- the number of crashes with slightly injured persons;
- the number of crashes with persons killed or seriously injured;
- the number of persons slightly injured;
- the number of persons killed or seriously injured.

Table 4.1 gives an overview of the explanatory variables which were used in both models. The dummy variables used to indicate the month in the DRAG type model were included to cope with seasonality. The model with ARIMA errors also included dummy variables. These variables corresponded to January 1979, January 1984, January 1985 and February 1997 and were added to the model because the numbers of crashes and victims in these months were extremely low.

4.2. Methodology

It is possible to model the relations between the dependent and explanatory variables listed above with multiple linear regression. In theoretical linear regression several assumptions are made about the explanatory variables

Category	Independent variables	Model
Exposure	Fuel consumption (for gasoline and diesel)	DRAG
Prices	Price of vehicle maintenance	DRAG
	Price of adult public transit	DRAG
	Average fuel price and tax	DRAG
Laws and regulations	Mandatory seat belt use in front seat in 1975	ARIMA
	Introduction of 30 km/h zones in 1988	Both
	Improvement of position vulnerable road users in 1991	DRAG
	Several regulations introduced in 1992	Both
	The imposement of 0.05% alcohol level in 1994	Both
	Introduction of right of way for pedestrians in 1996	Both
Weather conditions	Quantity of precipitation (ml/100)	Both
	Average temperature	DRAG
	Number of days with precipitation	Both
	Number of days with sunlight	Both
	Number of days with frost	Both
	Number of days with snow	Both
	Number of days with thunderstorms	Both
	Quantity of precipitation per precipitation day	DRAG
	Number of sunlight hours	ARIMA
Economic activity	Inflation (in percentage)	Both
	Percentage of unemployed people	Both
	Net exports (Exports - Imports)	DRAG
	Number of car registrations	ARIMA
	Percentage of second hand car registrations	ARIMA
Time	Number of workdays	DRAG
	Number of Saturdays	DRAG
	Number of Sundays and public holidays	DRAG
	Dummy variable for indication of the month	DRAG
	Months since first observations divided by 100	DRAG

Table 4.1. *Summary of the independent variables in Belgium.*

and the error term. However, some of these assumptions are frequently violated when linear regression is applied to time series. These assumptions are that:

– the explanatory variables may not be perfectly correlated;
– the error terms should be uncorrelated over time;
– the error terms should be identically distributed with mean zero and constant variance.

The situation were the first assumption is violated, is called multicollinearity. In this case it is not possible to isolate the effect of a single explanatory variable. When the error terms are correlated among themselves, i.e., if the second assumption is violated, then there is autocorrelation. In this case the variance of error terms and the standard deviations of the regression parameters may be underestimated and confidence intervals, $t-$tests or

$F-$tests are no longer strictly applicable. It leads to the possibility that a coefficient is assumed to be significantly different from zero, while it is not. This is also the case if the third assumption is violated, which is called heteroscedasticity.

Van den Bossche & Wets (2003) and Van den Bossche, Wets & Brijs (2004) solved the problems mentioned above in two different ways. The DRAG type model of Van den Bossche & Wets (2003) allowed for flexible functional forms and corrected for possible autocorrelation and heteroscedasticity. Because it is not always necessary to correct for these problems, a gradual modelling approach was used. First an ordinary regression was carried out. By verifying the condition indices, high multicollinearity was identified in all equations. Hence, the variables in the collinear relation were transformed by Box-Cox parameters. If there was no appropriate transformation the variable was dropped from the model. Then the new model was tested for heteroscedasticity, using the Spearman Rank Test and the Modified Levene Test. The conclusion was that heteroscedasticity was not a serious problem. Finally, the Portmanteau Q^*-statistic indicated that the null hypothesis of white noise residuals could not be rejected. However, some autocorrelation terms were added to the model because of the (Partial) Autocorrelation Function and the corresponding confidence interval.

The model with ARIMA error terms, used by Van den Bossche, Wets & Brijs (2004), only corrected for autocorrelation. The obtained models were tested for multicollinearity and heteroscedasticity, but they turned out to be no serious problems.

4.3. Results of the DRAG type model

The quality of the different models for the four dependent variables is described by the values in *Table 4.2*. Note that the models for the number of crashes with persons injured and the number of persons injured performed better than the other two models. The number of observations is the total number of observations minus 12, which was the maximum autocorrelation level in the models.

	Crashes with injuries	Crashes with killed	Persons injured	Persons killed
Log likelihood	-1,113.82	-634.36	-1.192.02	-649.152
Pseudo-(E)-R^2	0.9003	0.6459	0.8735	0.6961
Pseudo-(E)-R^2 adjusted	0.8785	0.5835	0.8501	0.6401
Number of observations	168	168	168	168

Table 4.2. *Goodness of fit of the DRAG model for Belgium.*

4.4. Forecasting with the ARIMA model

The developed model was used to forecast the dependent variables for the year 2000. The values of the explanatory variables were available. Because the observed values of the dependent variables were known for 2000 a comparison could be made between the observed and the predicted values. They were quite close, but crashes were predicted better than victims.

5. Sweden

In January 1991 The Stockholm Traffic Agreement, also called The Dennis Traffic Agreement, was signed. This agreement contained three measures to solve urban transportation problems in Stockholm and its surroundings. These measures are:

- a ring road system with one completed inner ring and one horse-shoe-shaped outer ring;
- a public transport commuter rail upgrading and a tangential light rail line;
- a system of vehicle tolls at the inner ring and on the outer western by-pass to reduce road traffic and to finance the road program.

During the process of implementation of these measures the authorities wanted to be able to monitor the impacts of the measures.

For this purpose two consultancy bureaus were requested to carry out the MAD-project, where MAD stands for ”Measurement and Analysis of the Dennis Agreement”. The aim of this project was to give a description of and an analytic tool for assessing the past development of several aspects of traffic in the Stockholm region during the last 25 years.

Tegnér et al. (2002) developed two models which were used in the MAD-project to study the traffic safety development in the Stockholm region. These models are called DRAG-Stockholm-1 and DRAG-Stockholm-2, where the latter is an improvement of the first. Tegnér (2000) developed a third model, the DRAG-Stockholm-3 model. These DRAG-models are discussed in *Section 5.1*.

A far more simple model has been developed by Brüde (1995). This model used time series analysis covering the years 1977-1991 to forecast the number of road fatalities up to the year 2000 in the whole of Sweden. Only two explanatory variables were used, namely time and traffic. *Section 5.2* discusses this simple model.

5.1. The DRAG-type models

5.1.1. *The structure of the model*

The DRAG-Stockholm-1 model consists of three typical DRAG submodels:

- an exposure model of total road mileage (vehicle kilometres) for gasoline driven cars;
- a frequency model of the total number of injury and fatal crashes;
- three severity models for the numbers of slightly injured, severely injured and fatalities per accident.

The other two DRAG models have a similar structure. However, not only the gasoline vehicle kilometres were included, but also the diesel vehicle kilometres. Moreover, the crash frequency model of DRAG-Stockholm-2 consists of three submodels:

- the number of slight injury crashes;
- the number of severe injury crashes;
- the number of fatal crashes.

The numbers in the third submodel of the DRAG-Stockholm-1 model were changed correspondingly into:

- slightly injured persons per bodily injury accident;
- severely injured persons per severe and fatal accident;
- fatalities per fatal accident.

5.1.2. *Explanatory variables*

For the DRAG-Stockholm-1 and -2 model monthly data for a total of a hundred variables was collected for the period 1970-1995. This period was extended to 1970-1998 for the DRAG-Stockholm-3 model. Not all variables were available or defined for the Stockholm region, so national data was used for these variables instead. The used explanatory variables described the economic activities, the vehicle fleet, prices and public transport, road network and restrictions, climate and calendar, special events and health. In the DRAG-Stockholm-2 model more explanatory variables were tested than in the DRAG-Stockholm-1 model.

5.1.3. *Results*

The modelling results of the DRAG-Stockholm-1 model are described in *Table 5.1*. The overall correspondence between observed and estimated vehicle kilometres was also tested and turned out to be quite good.

	vhc-km	Road crashes	Slight injuries	Severe injuries	Fatalities
Number of expl. variables	20	29	29	29	29
Number of $t-$values ≥ 2	15	10	3	4	2
Number of $t-$values $\in (1, 2)$	1	10	13	12	12
Number of $t-$values ≤ 1	4	9	13	13	15
Autocorrelation parameters	3	3	3	3	3
λ_y	0.53	0.24	0.13	0.53	0.57
λ_{x_1}	0.53	0.24	0.13	0.53	0.57
λ_{x_2}	1.51	2.00	2.00	2.00	2.00
Log likelihood at opt. form	-2878	-1282	389	508	825
Sample size	300	288	288	288	288

Table 5.1. *Function form, stochastic specification and other summary statistics.*

5.1.4. *Prognoses*

The DRAG-Stockholm-3 model was also used to forecast the number of crashes and their severity for the year 2015. Several assumptions were made:
- the population increases from 1.78 to 2.09 million;
- the employment grows by 30%;
- shopping activities grow by 37%;
- the car park grows by 18%;
- car traffic volume increases by 26%.

Various scenarios, which describe the development of the other explanatory variables, have been analysed with DRAG-Stockholm-3. The conclusions are that:
- the growth of the Stockholm region leads to three more fatal crashes per

year;
- new city highways would reduce the number of severe injuries by 30 each year;
- powerful countermeasures are necessary for safer bicycle traffic;
- a 1% yearly reduction of the average speed up to 2015 may lead to 6 fewer fatalities and 17 fewer severe injuries;
- the number of fatalities is forecast to be reduced by 21% if the road and street illumination is improved by 1%.

5.2. The more simple model

The form of the model chosen by Brüde (1995) is

$$\text{Fatalities} = a \cdot b^{\textit{Year}} \cdot \textit{Traffic}^{c},$$

where *Year* $= 1$ for 1977, *Year* $= 2$ for 1978 et cetera and *Traffic* is equal to the traffic (mileage) index with 1977 as base year, i.e. *Traffic* $= 100$ for 1977. The given model has three advantages:
- it is simple and interpretable;
- it immediately shows the number of fatalities instead of the death rate (fatalities/traffic);
- it permits a non-proportional relationship to traffic volume for fatalities.

The model parameters a, b and c were estimated by using generalised linear models under the assumption that the number of fatalities follows a Poisson distribution and using the data for 1977-1991. The estimated model fitted the data reasonably well: it explained 95% of the variation in the number of fatalities. The obtained model predicted 743 fatalities for 1992 and 647 fatalities for 1993, while the actual values were 759 and 632 respectively. Hence, the predictions were very accurate.

The forecasts, however, were uncertain in several aspects. They were based on the assumption that future road safety work would be as extensive and successful as before. Also, the fact that the model fitted the data quite well is no guarantee that the model will be reliable in the future. Furthermore, making the forecasts required to extrapolate the regression model outside the area where the observations were made.

6. France

Several attempts were made in France to explain the past developments of road safety. Three DRAG-type models were developed, with their own characteristics. Hence the models will be discussed separately in *Sections 6.1 - 6.3*. Furthermore, two models will be discussed which tried to explain the effect of specific variables, namely climate variables (*Section 6.4*) and the presidential amnesties of 1988 and 1995 (*Section 6.5*).

6.1. The TAG-1 model

Jaeger & Lassarre (2002) developed the TAG-1 model for France. TAG stands for Traffic, Crash and Gravity, which illustrates that the TAG model is inspired on the DRAG model. However, several adaptations were made: explanatory variables concerning specific French road safety conditions were included in the model and behaviour was added to the model as a fourth dimension, next to traffic, crash and gravity.

6.1.1. *The structure of the model*

The TAG-1 model consists of the following four dimensions:

1. the exposure to risk measured as the number of kilometres travelled;
2. the risk behaviour measured in terms of the average speed driven on the inter-urban network;
3. the number of injury crashes, both for fatal and for non-fatal crashes;
4. the crash severity in terms of the rate of fatalities, minor injuries and severe injuries per injury accident.

The resulting model consists of seven equations of the form *(3.3)*. Let $y_{1,t}$ be the total mileage, $y_{2,t}$ the average inter-urban speed, $y_{31,t}, y_{32,t}$ the number of fatal and non-fatal crashes respectively, $y_{41,t}, y_{42,t}, y_{43,t}$ the number of fatal, serious and slight severity rate respectively and $x_{i,t}, i = 1, \ldots, k$ the explanatory variables, all in year t. Then the model is given by

$$\begin{cases} y_{1,t} &= f_1(x_{i,t}; u_{1,t}), \\ y_{2,t} &= f_2(y_{1,t}, x_{i,t}; u_{2,t}), \\ y_{31,t} &= f_3(y_{1,t}, y_{2,t}, x_{i,t}; u_{31,t}), \\ y_{32,t} &= f_4(y_{1,t}, y_{2,t}, x_{i,t}; u_{32,t}), \\ y_{41,t} &= f_5(y_{1,t}, y_{2,t}, x_{i,t}; u_{41,t}), \\ y_{42,t} &= f_6(y_{1,t}, y_{2,t}, x_{i,t}; u_{42,t}), \\ y_{43,t} &= f_7(y_{1,t}, y_{2,t}, x_{i,t}; u_{43,t}), \end{cases} \tag{6.1}$$

where $u_{i,t}$ denotes white noise. The advantage of this structure is that the direct and indirect effects of the explanatory variables can be identified and that the compensation effects between the numbers of fatal and non-fatal crashes and between the numbers of fatalities and injuries can be studied.

6.1.2. *The explanatory variables*

Distinction was made between internal and external variables. The internal variables were linked to the characteristics of vehicles, drivers, and the road infrastructure, whereas external variables were related to the system environment. The external variables usually have an indirect effect on the performance of the road transport system through their effect on its three components: vehicle, driver, road. The main explanatory variables are given in *Table 6.1*.

Vehicles	1. Stock	– PC/HVG/motorised two-wheeler breakdown
Drivers	2. Characteristics	– Proportion of young adults
	3. Behavioural variables	– Rate of seat belt wearing
		– Taxed wine consumption
Infrastructure	4. Networks	– Share of traffic on motorways and autoroutes and on main roads
Demography	5. Population	– Proportion of young adults
Economy	6. Price	– Price of fuel per km
		– Price of a car, public road, rail and air transport
	7. Unemployment	– Proportion of unemployed
	8. Reasons for people's journeys	– Working population
		– Household consumption
		– Vacations
		– Taxed wine consumption
	9. Reasons for good transport	– Industrial activity
Government	10. Road safety laws	– Mandatory technical inspections
		– Behavioural legislation
Climatic	11. Climatic variables	– Snow, etc

Table 6.1. *Classification of the main explanatory variables integrated in the TAG model.*

6.1.3. *Methodology*

The developed models were based on the monthly series of the average speed and the number of crashes and casualties from 1967 to 1993. The total number of kilometres travelled on the French network was not known on a monthly basis for such an extended period of time. Hence a methodology was developed for calculating the total mileage travelled on the entire French road network by all road vehicles.

The L-1.5 algorithm of the TRIO software programme was used to set or estimate the values of the Box-Cox parameters of the dependent and explanatory variables in *(6.1)*. The risk of multicollinearity was minimised by including the variables in a stepwise manner in the model.

6.2. **The RES model**

The TAG model was designed for the whole French road network and used a large amount of explanatory variables. The RES model, which is the subject of this section, is an extension of the TAG model into a vectorial, i.e., a multivariate, framework. The number of explanatory variables was limited to about ten for the sake of robustness and the model was limited to two types of roads: toll motorways and main roads.

6.2.1. *The basic model*

The dependent variables for the models to be developed were the traffic, the number of crashes and the number of deaths, all three for both toll motorways and main roads. Bergel & Girard (2002) gave two possible model forms: a univariate and a vectorial specification. The first one is of the form given in *(3.3)*. It was tested if the model with the constraints $\lambda_y = 0$ and $\lambda_x = 0$ for the primary explanatory variables was significantly different from the model with only the first constraint. The result was that the constraint $\lambda_x = 0$ could not be rejected from a statistical point of view. Hence, not the primary explanatory variables but their logarithms were included in the vectorial specification.

The vectorial formulation makes it possible to take into account the (immediate or delayed) correlation between the residuals corresponding to two equations on the same level, for example the number of crashes on main roads and on toll motorways. The general form of the autoregressive vectorial formulation is as follows:

$$\Theta(B)^{-1}\Phi(B)Y_t = \Theta(B)^{-1}\Psi(B)Z_t + W_t, \tag{6.2}$$

where B is the delay operator defined as $B^n x_t = x_{t-n}$ and

$Y_t =$ the vector of the p dependent variables;
$Z_t =$ the explanatory vector with q components;
$W_t =$ white noise vector with p components;
$\Phi, \Psi, \Theta =$ matrices of which the entries are polynomials.

The vector Y_t can be considered to consist of two parts: an exogenous part (explained by the explanatory variables) and its own dynamics. These dynamics can be described in the form of an unobservable autoregressive vector X_t. This leads to the following, representation, named state space:

$$\begin{aligned} X_{t+1} &= FX_t + KW_t, \\ Y_t &= D(B)Z_t + HX_t + W_t, \end{aligned}$$

where D is a polynomial matrix in B and F, K and H are three real matrices of appropriate dimensions. The vector X_t is called the state vector.

6.2.2. *The explanatory variables*

The available database contained monthly data for the years 1975-1993 for the following variables:

- the number of personal injury crashes (fatal and non-fatal) and victims (killed, seriously injured and slightly injured), separately for both toll motorways and main roads;
- the number of kilometres travelled on motorways and on main roads;
- transport supply (road network length, fuel prices and motorway tolls);
- transport demand (gross domestic product, industrial production and household consumption expenditures);
- climatic variables relating to temperature, the occurrence of frost and the height of rain;
- behavioural variables: speed and seat belt use.

The primary explanatory variables for the number of crashes and deaths were the numbers of kilometres travelled on toll motorways and on main roads. The primary explanatory variables for the volume of traffic were the household consumption expenditures, the fuel prices, and the length of the motorway network.

6.2.3. *Methodology*

The methodology for the univariate model is similar to the one used in the ARMAX model and will be discussed in detail in *Section 6.3*. For the vectorial model it was decided to estimate the parameters in the state-space model instead of the parameters in *(6.2)*, because then a distinction could be made between the effects of the explanatory variables and the dynamics of Y_t itself. The parameter estimation involved three steps:

1. In this step the dependent variable Y_t was corrected for the effects of the explanatory variables to isolate the dynamics of Y_t. This results in the following corrected vector YC_t :

   $$YC_t = Y_t - D(B)Z_t.$$

2. This step concerned the modelling of the process YC_t :

   $$\begin{aligned} X_{t+1} &= FX_t + KW_t, \\ YC_t &= HX_t + W_t. \end{aligned}$$

   First the system's rank was determined using the Hankel matrix of YC_t. Then the matrices F, K and H were estimated using the Desai and Pal algorithm.
3. Finally the obtained values of F, K and H were used to re-estimate the effects of the explanatory variables, that is $D(B)$, by the maximum likelihood method on the initial vector Y_t.

6.3. **ARMAX models**

The aim of Depire (1999) was to develop explanatory and predictive models for road safety. He dealt with time series analysis of risk and severity for toll motorways and main roads on a monthly basis over the years 1975-1998. For the year 1999 forecasts were made. The developed models are of the DRAG type.

6.3.1. *The basic model*

Four models were developed: two models for the number of crashes with bodily injuries and the number of fatalities on toll motorways and two models for the same crashes on main roads. Each of the four submodels has the following form:

$$\begin{aligned} \log Y_t &= \beta_1 \frac{X_{1,t}^{\lambda} - 1}{\lambda} + \sum_{i=2}^{I} \beta_i X_{i,t} + \varepsilon_t, \\ \varepsilon_t &= u_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_p \varepsilon_{t-p}, \\ u_t &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{6.3}$$

where $X_{1,t}$ denotes the number of vehicle kilometres. Note that this model is indeed of the DRAG form given in *(3.3)*. However, DRAG models generally describe the exposure to risk, the risk itself and the severity. Depire (1999) did not model the exposure to risk, but only used it as an explanatory variable for risk and severity.

The maximum order p of the autoregressive coefficients was assumed to be equal to 14. This assumption was based on the following model for the error term ε_t :

$$\begin{aligned}(1-&\theta_1 B - \theta_2 B^2)(1-\theta_{12}B^{12})\varepsilon_t = \\ &= (1-\theta_1 B - \theta_2 B^2 - \theta_{12}B^{12} - \theta_{13}B^{13} - \theta_{14}B^{14})\varepsilon_t = u_t,\end{aligned} \tag{6.4}$$

where B is the backward shift operator. This model takes into account the short memory (at the most two months) of the dependent variable, the short term effect of the independent variables and the (annual) seasonality of the time series.

6.3.2. *Explanatory variables*

Probably the most important explanatory variable was the exposure to risk, measured in vehicle kilometres for each of the two considered road types.

Furthermore, it seems quite logical that weather conditions also have an influence on the road safety. Hence, the amount of precipitation per month, the monthly average temperature and the number of days with frost per month were included as explanatory variables. The temperature was split into two variables, namely the summer temperature (TE) and winter temperature (TH). They are defined as follows:

$$TH = \begin{cases} TMAX, & \text{from October up to March,} \\ 0, & \text{for other months,} \end{cases}$$

$$TE = \begin{cases} TMAX, & \text{from April up to September,} \\ 0, & \text{for other months,} \end{cases}$$

where $TMAX$ is the maximum temperature in a month. Two variables involving temperature instead of one were included in the model because of the idea that the influence of the temperature on the number of crashes and fatalities is different in different seasons.

As already mentioned, the number of vehicle kilometres was the most important explanatory variable. However, this was not known for the ‘future’ year 1999. Hence other explanatory variables should be included in the model. It was assumed that there were two economic factors which could be used for this purpose, namely the households consumption expenditures and the prices of petrol. The first one was only available with a quarterly periodicity, but was computed for this study with a monthly periodicity on basis of the households consumption of manufactured products issued by l’Institute National de la Statistique et des Études Économiques (INSEE). The obtained series was expressed in constant prices of 1980 and was included in the model with a three month delay. The price of petrol was defined as the mean price per litre, expressed in constant prices of 1980, of regular gas, super with and without lead and diesel. The price index was then set to 1.00 for the year 1980.

6.3.3. *Methodology*

Two methodologies were used to develop the models. The idea of the first approach is to estimate the parameter λ in the same way as the other parameters in the model. Because this model is not linear, the NLIN procedure in SAS can be used. The autoregressive structure of the time series under consideration should be taken into account. Indeed, if a classical regression is used, the residuals will show an autoregressive structure which makes it impossible to use the obtained estimates of the model parameters. To confirm the autoregressive structure given in *(6.4)* two SAS procedures (AUTOREG and ARIMA) are available. The first approach can be summarised as follows:

1. The SAS procedure NLIN is used in order to obtain the residuals ε_t. NLIN computes the estimates of the parameters $\lambda, \beta_1, \ldots, \beta_I$. The residuals can then be estimated by

$$\varepsilon_t = y_t - \left(\hat{\beta}_1 \frac{X_{1,t}^{\hat{\lambda}} - 1}{\hat{\lambda}} + \sum_{i=2}^{I} \hat{\beta}_i X_{i,t} \right).$$

2. One of the procedures AUTOREG or ARIMA is applied to ε_t in order to get an idea of its autoregressive behaviour. The autoregressive parameter p in

$$\varepsilon_t = u_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_p \varepsilon_{t-p}$$

is determined.

3. The SAS procedure NLIN is applied once more, but know with the value of p obtained in the previous step. The parameters of the equation

$$Y_t = f(Y_{t-1}, \ldots, Y_{t-p}; X_{1,t}, \ldots, X_{1,t-p}; \ldots; X_{I,t}, \ldots, X_{I,t-p})$$

are reestimated by the coefficients in the following equation:

$$Y_t = \hat{\beta}_1 \frac{X_{1,t}^{\hat{\lambda}} - 1}{\hat{\lambda}} + \sum_{i=1}^{I} \hat{\beta}_i X_{i,t} - \sum_{j=1}^{p} \left(\hat{\theta}_j Y_{t-j} - \hat{\theta}_j \sum_{i=1}^{I} \hat{\beta}_i X_{i,t-j} \right) + u_t.$$

There are several problems with this approach. Firstly, the procedure is based on the assumption that the first step does not have an impact on the estimates in the second step. However, this assumption can hardly be verified. Secondly, the coefficients are not estimated simultaneously but successively. Furthermore, the estimated value of λ is very unstable: it depends strongly on its initial value λ_0. Finally, some of the coefficients can not be estimated with this method, due to divergence. Because of these problems, it was decided to use another approach.

In this alternative approach the value of λ is not estimated, but it is fixed. The optimal value is determined by the least square method. A grid is used for λ, denoted by $\lambda_0, \ldots, \lambda_n$, so that there is no instability because of the initial conditions. In short the approach involves the following steps, starting with λ_0 :

1. The Box-Cox transformation of $X_{1,t}$ corresponding to λ_i is set.
2. The SAS procedure AUTOREG is used with the transformed variable to estimate the other model parameters.

3. The model error, expressed as the sum of squared errors, is computed. The sum of squared errors is defined by

$$SSE = \sum_{i=1}^{N} u_t^2.$$

4. If $i \neq n$ the proceedings steps are repeated for λ_{i+1}.

This results in a set of sum of squared errors, one for each possible value of λ. The optimal value of λ is the value corresponding to the minimal sum of squared errors. Two methods were used to estimate the parameters: the Yule-Walker algorithm and the method of maximum likelihood. The first method diverged sometimes, due to a convergence problem in the algorithm. The second one seemed to work fine, but it was very slow. The results of both methods were very similar when the Yule-Walker algorithm converged.

6.3.4. *Results*

The parameters for the four models were estimated based on the period January 1975 up to January 1994. Taking into account the three months delay of the households consumption expenditures this came to a total of 226 months.

The results for the four models are extensively discussed in Depire (1999). First, the optimal value of λ was presented, together with the “dynamics” of the model, which is described by the indices of θ_i for which θ_i is not equal to zero. Then the stability of the estimated parameters was tested. This was done by reestimating the parameters for the extended period of January 1975 - December 1998 and checking if the new values were contained in the confidence intervals of the parameters in the original models. It was also tested if the model given in *(6.3)* with $\lambda = \hat{\lambda}$ deviated significantly from the model with $\lambda = 0$ and with $\lambda = 1$. For this purpose the Lagrange multiplier test was used. The autoregressive model was compared with the classical regression model by means of a plot of the predicted values and the observed values of the dependent variables. The residuals were subjected to a number of tests. These tests are:

– the sign test, which is used to test the hypothesis that the residuals come from the same distribution;
– the Shapiro-Wilk test, which is used to test the hypothesis that the residuals are normally distributed;
– the Fisher test and the Bartlett-Kolmogorov-Smirnov test are used to test the hypothesis that the residuals are white noise.

The results of all those tests are summarised in *Table 6.2*.

Both hypotheses $H_0 : \lambda = 0$ and $H_0 : \lambda = 1$ were accepted for crashes on main roads. However, the significance of the first one was larger than the significance of the second, so $\lambda = 0$ was preferred.

The hypotheses for the residuals were accepted or rejected on the basis of a 5% confidence level. For the residuals of each of the four models at least three tests were accepted. If the confidence level had been 1% instead, the hypothesis that the residuals corresponding to crashes on toll motorways form a sample would not have been rejected. The same hypothesis corresponding to fatalities on main roads would already have been accepted if the confidence level was set to 4%. Furthermore, the tests

	Crashes on toll motorways	Crashes on main roads	Fatalities on toll motorways	Fatalities on main roads
Value of $\hat{\lambda}$	-0.0404	-0.2828	-0.3636	-1.333
Autoregr. coeff.	$\theta_1, \theta_{12}, \theta_{14}$	$\theta_1, \theta_2, \theta_{12}, \theta_{14}$	θ_{12}	$\theta_1, \theta_2, \theta_{12}, \theta_{14}$
Stable coefficients	Yes	Yes	Except vhc-km	Yes
$H_0 : \lambda = 0$	Accept	Accept	Accept	Accept
$H_0 : \lambda = 1$	Reject	Accept	Reject	Reject
Sample test	Reject	Accept	Accept	Reject
Normality test	Accept	Reject	Accept	Accept
Fisher test	Accept	Accept	Accept	Accept
Bartlett test	Accept	Accept	Accept	Accept
SSE	2.935156	1.052959	23.54387	2.91741
R^2	68.58%	33.12%	45.54%	26.96%

Table 6.2. *The results for the four different models.*

for normality and the Fisher and Bartlett tests contradicted each other in the case of crashes on main roads. The first one rejected the hypothesis that the residuals are normally distributed, while the Fisher and Bartlett tests accepted this hypothesis.

6.3.5. *Prognoses for 1998 and 1999*

Depire (1999) not only explained the past development of road safety, but also made predictions for 1998 and 1999. The real number of crashes and fatalities in 1998 was known, hence a comparison could be made between the predicted and observed values. The values of the explanatory variables were also known for 1998.

There are two types of prediction. Namely the prediction of Y_{t+h} step-by-step, denoted by $\hat{Y}_t(h)$, and the prediction of Y_{t+h} with $h > 1$ in one step, denoted by $\tilde{Y}_t(h)$. They are defined as:

$$\hat{Y}_t(h) = \mathbb{E}(Y_{t+h} | \hat{Y}_t(h-1), \ldots) \text{ where } \hat{Y}_t(h) = Y_t \text{ if } h \leq 0,$$
$$\tilde{Y}_t(h) = \mathbb{E}(Y_{t+h} | Y_t, \ldots).$$

In the case under consideration predictions are required for Z_t where

$$Z_t = \log Y_t = \beta_0 + \beta_1 \log X_{i,t} + \sum_{i=2}^{I} \beta_i X_{i,t} + \varepsilon_t,$$
$$\varepsilon_t = u_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_p \varepsilon_{t-p}.$$

Because of the results described in the previous section, the Box-Cox parameter of Y_t and $X_{1,t}$ is now set to $\lambda = 0$.

The prediction of Z_{t+1} is given by:

$$\hat{Z}_t(1) = \hat{\beta}_0 + \hat{\beta}_1 \log X_{1,t+1} + \sum_{i=2}^{I} \hat{\beta}_i X_{i,t+1} + \hat{\theta}_1 \varepsilon_t + \ldots + \hat{\theta}_p \varepsilon_{t-p+1}.$$

By induction it can be proved that the step-by-step prediction of Z_t for h steps ahead is given by:

$$\hat{Z}_t(h) = \hat{\beta}_0 + \sum_{i=1}^{I} \hat{\beta}_i \log X_{i,t+h} + \sum_{j=1}^{J} \hat{\gamma}_j X_{j,t+h} + \\ + \hat{\theta}_1 \hat{\varepsilon}_t(h-1) + \ldots + \hat{\theta}_p \hat{\varepsilon}_{t-p+1}(h-p).$$

It is also possible to compute the prediction $\tilde{Z}_t(h)$ of Z_{t+h} :

$$\tilde{Z}_t(h) = \hat{\beta}_0 + \sum_{i=1}^{I} \hat{\beta}_i \log X_{i,t+1} + \sum_{j=1}^{J} \hat{\gamma}_j X_{j,t+1} + \\ + \hat{\theta}_1 \hat{\varepsilon}_t(h-1) + \ldots + \hat{\theta}_p \hat{\varepsilon}_{t-p+1}(h-p).$$

From this it follows that $\hat{Z}_t(h) \neq \tilde{Z}_t(h)$ except for $h = 1$. Depire (1999) preferred $\tilde{Z}_t(h)$.

The results of the predictions for 1998 were expressed in terms of δ_t and δ_t^* which are defined as:

$$\delta_t = \frac{\log \tilde{Y}_t - \log Y_t}{\log Y_t}, \quad \delta_t^* = \frac{\tilde{Y}_t - Y_t}{Y_t}.$$

The relative differences δ_t for the number of crashes on toll motorways were almost all smaller than 5%, which indicated a very accurate prediction. The prediction of the number of crashes on main roads was even better, because all the relative differences were of the order 1%. Far less satisfactory were the predictions for the number of fatalities on toll motorways. The reason for this was that the observed number of fatalities varied a lot in 1998 and hence could not be modelled very well. Finally, the predictions for the number of fatalities on main roads were reasonably good with relative differences smaller than 5%.

To be able to predict the number of crashes and fatalities in 1999 assumptions had to made about the development of the explanatory variables. These assumptions are:

– The household consumption expenditures, denoted by CFM, increase with 2.5%;
– The price index for patrol ($ICARB$) increases with 1.1%;
– The weather variables are equal to the normal monthly values, defined as the mean of the mean monthly values of the last thirty years.

The variables CFM and $ICARB$ were included in the model instead of the number of vehicle kilometres, because this last number was not known for 1999. For the development of CFM and $ICARB$ there were two scenario's. The first scenario is

$$CFM_i^{99} = (1 + 2.5\%) \cdot CFM_i^{98}, \\ ICARB_i^{99} = (1 + 1.1\%) \cdot ICARB_i^{98},$$

where i denotes the month. However, this scenario has a drawback: it does not respect the continuity of $ICARB$. Furthermore, the values of $ICARB$ were available for the first semester of 1999. Hence, it was decided to consider a second scenario which uses this knowledge. In order to have a yearly growth of 1.1% of $ICARB$ the values of $ICARB$ for the second semester should be determined using the following equation:

$$(1+1.1\%)\sum_{i=1}^{12} ICARB_i^{98} = \sum_{i=1}^{6} ICARB_i^{99} + (1+\tau)\sum_{i=7}^{12} ICARB_i^{98},$$

hence

$$\tau = \frac{(1+1.1\%)\sum_{i=1}^{12} ICARB_i^{98} - \sum_{i=1}^{6} ICARB_i^{99}}{\sum_{i=7}^{12} ICARB_i^{98}} - 1.$$

So the second scenario is:

$$ICARB_i^{99} = (1+\tau)ICARB_i^{98}, \qquad i = 7, \ldots, 12.$$

A similar derivation holds for CFM. The results for both scenarios were very similar.

6.3.6. *Extension to the entire network*

In the previous sections models were discussed for toll motorways and for main roads. In this section the generalisation of these models to the entire French road network is discussed. Because the number of vehicle kilometres was not known for the entire network on a monthly basis, the variables CFM and $ICARB$ were introduced into the model. The resulting model takes the form:

$$\begin{aligned}\log Y_t = \beta_0 + \beta_1 \log CFM_{t-3} + \beta_2 \log ICARB_t + \beta_3 TE_t + \\ + \beta_4 TH_t + \beta_5 NGEL_t + \beta_6 HPLUIE_t + \varepsilon_t,\end{aligned}$$

where Y_t denotes the number of crashes on the entire network or the number of fatalities on the entire network. The error terms ε_t again had an autoregressive structure. To find the estimates of the model parameters the software programme MODEST was used. The modelling of the number of crashes was straightforward, the modelling of the number of fatalities required more work. It turned out that the residuals were rather large for eight months, so eight dummy variables were introduced. The inclusion of these variables in the model improved the fit of the model. The obtained models were applied to the year 1999, under the same assumption as in the previous section.

6.4. **The effect of climate variables**

Bergel & Depire (2002) aimed at analysing the influence of climate on the number of injury crashes and fatalities, aggregated for the whole of France and for each of the four main network categories.

6.4.1. *The explanatory variables*

Monthly data for the period 1975-1999 were collected for the following variables:

- risk exposure;
- the highest temperature of the day;
- the occurrence of frost;
- the daily rainfall height;

– atypical meteorology variables.

The first variable was measured with the traffic volume (expressed in hundreds of billion vehicle kilometres) on the main network (consisting of main roads and motorways). On the secondary and urban networks the traffic volume was not known on a monthly basis, so it was modelled in that case with the help of several explanatory variables.

The climate variables were averaged over the whole territory and over the month. First, daily climate variables were computed by averaging the daily variables measured at several points in France. Then they were aggregated or averaged over the month to construct monthly variables. The atypical meteorology variables code the number of days in a month with extreme climate values.

6.4.2. *The basic model structure*

A total of five models was developed, one for the entire network and one for each of the four road categories: main roads, motorways, secondary roads and urban roads. The models have the following form:

$$\Phi(B)\left(\log Y_t - \sum_{i=1}^{I} \beta_i \log X_{i,t} - \sum_{j=1}^{J} \beta_j X_{j,t} - \mu\right) = \Theta(B) u_t,$$

where Y_t is the two-dimensional vector with the number of injury crashes and the number of fatalities in year t as its components, $X_{i,t}$ are the main explanatory variables measuring the risk exposure, $X_{j,t}$ are explanatory variables measuring the climate, Φ, Ψ are two polynomials in the delay operator B and u_t is a white noise not correlated with the past of $Y_t, X_{i,t}$ and $X_{j,t}$. All the variables, dependent as well as explanatory, were obtained by filtering the initial data with $I - B^{12}$ in order to get a stationary dependent variable which is corrected for the effects of the explanatory variables.

For main roads and motorways a model was also developed for the risk exposure, with the same climate variables. This made it possible to identify not only the direct effect of the explanatory variables on the number of crashes and fatalities, but also their indirect effect, through the traffic volume.

6.4.3. *Methodology*

First models were developed with the monthly averaged variables. Next the atypical meteorology variables were also added to the models. The model parameters were estimated with the AUTOREG and ARIMA procedures of SAS by means of maximising the log likelihood. It was assumed that the dependent variables were Gaussian. Although the calculation of the likelihood can be questioned if this normality assumption does not hold, the parameter estimates can still have good asymptotic convergence properties.

6.5. **The effect of presidential amnesties**

In France it is a tradition that a newly elected president pronounces a general amnesty of traffic offences. Because this is published by the media in advance, drivers are encouraged to ignore traffic regulations. Bergel (2002) attempted to detect the effect of the amnesties in 1988 and 1995 on the number of traffic fatalities in France.

6.5.1. *The structure of the model*

Intervention analysis was applied to a model of a similar form as the model in the previous section. To be precise, the developed model is of the following form:

$$\Phi(B)(I-B^{12})\left(\log Y_t - \alpha_1 \log C_t - \sum_{k=1}^{K} \beta_k X_{k,t} - \right.$$
$$\left. - \sum_{i=0}^{n} \gamma_i P_{t-i}^{[T_{0,1}]} - \sum_{i=0}^{m} \delta_i P_{t-i}^{[T_{0,2}]} \right) = \mu + \Theta(B) a_t,$$

where

$Y_t =$ the number of fatalities in year t,
$C_t =$ the fuel consumption in year t,
$X_{k,t} =$ the explanatory variables of climatic and calendar nature,
$P_t^{[T_{0,i}]} =$ dummy variable defined by$P_t^{[T_{0,i}]} = 1$ when $t = T_{0,i}$ and 0 elsewhere for $i = 1, 2$,
$T_{0,1}, T_{0,2} =$ the two months in which the publicity first appeared in the media,
$n, m =$ the number of months of intervention following the announcement in the media,
$\Phi, \Theta =$ polynomials in the delay operator B,
$\mu =$ the intercept,
$a_t =$ white noise.

The sum involving $P^{[T_{0,1}]}$ is only unequal to 0 if t is in the interval $[T_{0,1}, T_{0,1} + n]$ in which case the sum equals γ_j for $t = T_{0,1} + j$ where $j = 0, \ldots, n$. A similar statement holds for the sum involving $P^{[T_{0,2}]}$. Assuming that $\gamma_0 = \ldots = \gamma_n$ and $\delta_0 = \ldots = \delta_n$ leads to the following simplified model:

$$\Phi(B)(I-B^{12})(\log Y_t - \alpha_1 \log C_t - \sum_{k=1}^{K} \beta_k X_{k,t}$$
$$- \gamma Step_{1,t} - \delta Step_{2,t}) = \mu + \Theta(B) a_t,$$

where $Step_{1,t}$ and $Step_{2,t}$ are dummy variables equal to 1 on the interval $[T_{0,1}, T_{0,1} + n]$ and $[T_{0,2}, T_{0,2} + n]$ respectively, and 0 elsewhere.

6.5.2. *Methodology*

The model was applied using different dates for the beginning and the end of each of the intervention periods. Then the best models were applied again but then without the explanatory variables which were not significant.

7. Conclusion

This report gives an overview of studies that were performed in countries other than the Netherlands in order to obtain explanatory models for the analysis and forecasting of developments in road safety. Several types of disaggregation of the national traffic process were used in these studies. All studies used some form of disaggregation in terms of accident severity. Some studies also disaggregated the traffic process by type of road user, road type, gender, and/or age group.

Moreover, apart from exposure data, several types of explanatory variables were considered for inclusion in these studies. In the studies performed in Great Britain these variables were drink driving, passive safety in cars, road safety engineering, and what are called core road safety activities. In the studies performed in Canada, Belgium, Sweden, and France many explanatory variables were investigated. Examples are: prices, laws and regulations, weather conditions, economic variables, and calendar variables.

In terms of analysis techniques, classical linear regression and log-linear regression were applied in the studies performed in Great Britain, while the ARIMA and DRAG approaches were used in the studies of Belgium, Canada, France, and Sweden.

An important problem with classical linear and loglinear regression applied to time series data is the assumption of independence of the observations. However, repeated observations over time are usually not independent at all, since last year's number of casualties is often quite a good predictor for current year's number of casualties. In a classical linear regression this is reflected in residuals that are serially correlated (see *Section 2.1.2*). This, in turn, results in statistical tests whose standard errors are too small, and therefore in overoptimistic conclusions about the relations between variables that evolve over time, and also in forecasts that are flawed.

The ARIMA and DRAG models discussed in *Chapters 3 - 6*, on the other hand, are dedicated time series models, and therefore do take the dependencies between the observations in time series data into account. An important requirement of ARIMA models, however, is that time series must be *stationary* (meaning among other things that they must have constant mean and variance over time) before the actual analysis can be carried out. Since most observed time series are made up of possibly non-stationary components like trend and seasonal (if measured on a quarterly or monthly basis, for example), cycle, and calendar variation, in the ARIMA approach the observations usually have to be filtered first (that is, trend and seasonal variation need be removed from the series by a process called differencing) in order to obtain stationarity. The actual analysis then consists of determining which autoregressive moving average (ARMA) specification yields the best fitting model for the stationary series, at the same time checking that the residuals are identically and independently distributed.

As discussed in the present report, DRAG models are extended versions of the standard ARIMA approach where

– exposure, crashes, and crash severity are modelled in separate analyses;

- many explanatory variables are added to the models;
- both the dependent and independent variables are allowed to be transformed by Box-Cox transformations.

On top of the stationarity requirements already mentioned for ARMA models, therefore, specific disadvantages of DRAG models are that

- exposure, crashes, and crash severity are not modelled simultaneously;
- the large amount of explanatory variables is at odds with the principle of parsimony;
- the Box-Cox transformations (although these are presented as an advantage) of the variables lead to all sorts of identification and interpretation problems.

A more recent development in the field of time series analysis consists of the so-called structural time series models (Harvey, 1989; Durbin & Koopman, 2001). Applications of these models to the analysis of road safety can be found in Harvey & Durbin (1986) and Ernst & Brüning (1990) on the effect of the seat belt law in Great Britain and Germany, respectively, and in Lassarre (2001) who compares the developments of road traffic safety in ten European countries. In Ord & Young (2003) structural time series models are used to analyse developments in air and rail traffic. However, we did not encounter studies in other countries where the structural time series approach was used to empirically validate the hypothesized relations between road accidents and explanatory variables, at least not with the scope intended in the Road Safety Assessment Department of the SWOV.

In contrast with ARIMA and DRAG models, structural time series models do not require stationarity and allow for the explicit decomposition of an observed time series in trend, seasonal, cycle, and calendar variation. Moreover, the latter models are very flexible, transparently handle missing data, and are easily extended to the multivariate analysis of time series data. They also allow for a simple and direct comparison with the results obtained with classical linear regression models, since the latter are just a special case of structural time series models where the components are treated deterministically instead of stochastically.

Another advantage of structural time series concerns the forecasting of explanatory variables. If for example a DRAG model is used for forecasting purposes, the values of the explanatory variables for the forecast period must be known. These values are obtained by modelling the explanatory variables separately and the modelled future values are then used in the DRAG model. In structural time series however the explanatory and dependent variables are not modelled and forecast separately, but simultaneously.

It is for these reasons that we favour the use of structural time series models for the description, explanation and forecasting of (disaggregated) developments in Dutch road traffic safety.

References

Bergel, R. (2002). *Effects of the presidential amnesties in 1988 and 1995 on the number of road traffic fatalities in France*. Paper presented at the ICTSA meeting, 25-26 April 2002, Technical University of Gdansk, Poland.

Bergel, R. & Depire, A. (2002). *Climate, road traffic and road risk: an aggregate approach*. Paper presented at the ICTSA meeting, 25-26 April 2002, Technical University of Gdansk, Poland.

Bergel, R. & Girard, B. (2002). *The RES model by road type in France*. In: Gaudry, M. & Lassarre, S. (eds.), Structural road accident models; The international DRAG family. Pergamon, p. 237–250.

Bijleveld, F. & Commandeur, J. (2006). *The basic evaluation model*. D-2006-2, SWOV, Leidschendam, The Netherlands.

Bossche, F. van den & Wets, G. (2003). *A structural road accident model for Belgium*. RA-2003-21, Steunpunt Verkeersveiligheid bij Stijgende Mobiliteit.

Bossche, F. van den, Wets, G. & Brijs, T. (2004). *A regression model with ARIMA errors to investigate the frequency and severity of road traffic accidents*. RA-2004-35, Steunpunt Verkeersveiligheid.

Broughton, J. (1988). *Predictive models of road accident fatalities*. In: Traffic Engineering and Control, 29, p. 296–300.

Broughton, J. (1991). *Forecasting road accident casualties in Great Britain*. In: Accident Analysis & Prevention, 23(5), p. 353–363.

Broughton, J., Allsop, R., Lynam, D. & McMahon, C. (2000). *The numerical context for setting national casualty reduction targets*. TRL Report 382, Transport Research Laboratory.

Brüde, U. (1995). *What is happening to the number of fatalities in road accidents? A model for forecasts and continuous monitorin of development up to the year 2000*. In: Accident Analysis & Prevention, 27(3), p. 405–410.

Depire, A. (1999). *Modélisation économétrique d'indicateurs de risque et de gravité des accidents de la circulation routière*. Ph.D. thesis, Université Pierre et Marie Curie.

Durbin, J. & Koopman, S. (2001). *Time series analysis by state space methods*. Number 24 in Oxford statistical science series, Oxford University Press.

Ernst, G. & Brüning, E. (1990). *Fünf Jahre danach: Wirksamkeit der Gurtanlegepflicht für Pkw Insassen ab 1. 8. 1984*. In: Zeitschrift für Verkehrssicherheit, 36(1), p. 2–13.

Forsyth, E., Maycock, G. & Sexton, B. (1995). *Cohort study of learner and novice drivers, part 3: Accidents, offences and driving experience in the first three years of driving*. TRL Report PR 111, Transport Research Laboratory.

Fournier, F. & Simard, R. (2002). *The DRAG-2 model for Quebec*. In: Gaudry, M. & Lassarre, S. (eds.), Structural road accident models; The international DRAG family. Pergamon, p. 37–66.

Fridstrøm, L., Ifver, J., Ingebrigsten, S., Kulmala, R. & Thomsen, L. (1995). *Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts*. In: Accident Analysis & Prevention, 27(1), p. 1–20.

Gaudry, M. (1984). *DRAG, model of the demand for road use accidents and their severity, applied in Quebec from 1956 to 1982*. Université de Montréal.

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

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

Jaeger, L. & Lassarre, S. (2002). *The TAG-1 model for France*. In: Gaudry, M. & Lassarre, S. (eds.), Structural road accident models; The international DRAG family. Pergamon, p. 157–184.

Lassarre, S. (2001). *Analysis of progress in road safety in then European countries*. In: Accident Analysis and Prevention, 33, p. 743–751.

Maycock, G. (2001). *Forecasting older driver accidents and casualties*. Road Safety Research Report 23, Department of the Environment, Transport and the Regions.

Maycock, G., Lockwood, C. & Lester, J. (1991). *The accident liability of car drivers*. TRL Report PR 315, Transport Research Laboratory.

Ord, K. & Young, P. (2003). *Estimating the Impact of Recent Interventions on Transportation Indicators*. In: Journal of Transportation and Statistics, 7.

Tegnér, G. (2000). *An analysis of urban road traffic safety in the city of Stockholm. the use of aggregate time-series models with the TRIO programme*. In: Proceedings of the Conference Road Safety on Three Continents, Pretoria, South Africa, 20–22 September 2000, VTI Konferenz, 15A, p. 730–785.

Tegnér, G., Holmberg, J., Loncar-Lucassi, V. & Nilsson, C. (2002). *The DRAG-Stockholm-2 model*. In: Gaudry, M. & Lassarre, S. (eds.), Structural road accident models; The international DRAG family. Pergamon, p. 127–156.