Multivariate nonlinear time series modelling of exposure and risk in road safety research

Frits Bijleveld$^{(a)}$, Jacques Commandeur$^{(a)}$,
Siem Jan Koopman$^{(b)}$ and Kees van Montfort$^{(b)*}$

$^{(a)}$ *SWOV Institute for Road Safety Research, Leidschendam, Netherlands*
$^{(b)}$ *Department of Econometrics, Vrije Universiteit Amsterdam, Netherlands*

Abstract

In this paper we consider a multivariate nonlinear time series model for the analysis of traffic volumes and road casualties inside and outside urban areas. The model consists of dynamic unobserved factors for exposure and risk that are related in a nonlinear way. The multivariate dimension of the model is due to the inclusion of different time series for inside and outside urban areas. The analysis is based on the extended Kalman filter. Quasi-maximum likelihood methods are utilised for the estimation of unknown parameters. The latent factors are estimated by extended smoothing methods. We present a case study of yearly time series of numbers of fatal accidents (inside and outside urban areas) and numbers of driven kilometers by motor vehicles in the Netherlands between 1961 and 2000. The analysis accounts for missing entries in the disaggregated numbers of driven kilometres although the aggregated numbers are observed throughout. It is concluded that the salient features of the observed time series are captured by the model in a satisfactory way.

Keywords: Extended Kalman filter; Quasi-maximum likelihood; Nonlinear dynamic factor analysis; Road casualties; State space model; Unobserved components.

[*] Corresponding author: Dr. K. van Montfort, Department of Econometrics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands. Email: `kvmontfort@feweb.vu.nl`. This version: November 28, 2005.

1 Introduction

This paper considers a multivariate nonlinear time series model for the analysis of traffic volume and road accident data. The model is based on the class of multivariate unobserved components time series models and is modified to allow for nonlinear relationships between components. The analysis relies on disaggregated and aggregated data and can account for missing entries in the data set. Missing observations are quite usual in road safety analysis where disaggregated data is not available throughout the sample period but data at the aggregated level is available for a longer period. The nonlinear nature of the model arises from the fact that the expected number of fatal accidents equals risk times exposure. This multiplicative relationship can be made additive by taking logarithms in the usual way. However since the analysis is based on aggregated and disaggregated data, summing constraints need to be considered as well. This mixture of multiplicative and additive relations in the model calls for a nonlinear analysis. Furthermore, the analysis is for a vector of time series and the model consists of multiple latent variables. Therefore, we adopt multivariate nonlinear state space methods for the analysis of road accidents.

The empirical motivation is to analyse the development of road safety inside and outside urban areas in the Netherlands between 1961 and 2000. The expected annual number of fatal accidents is defined by risk times exposure. Both risk and exposure are treated simultaneously as latent or unobserved components. The expected number of vehicle kilometres driven (traffic volume) is set equal to the latent exposure component. The observed traffic volume and the observed number of fatal accidents are available for inside and outside urban areas in the Netherlands. However, for some periods only the total number of vehicle kilometres driven (the sum of numbers for inside and outside urban areas) is available. For these periods, the expected total number of vehicle kilometres is set equal to the sum of the latent exposure components for inside and outside urban areas.

Since the seminal paper of Smeed (1949), time ordered accident data is analysed in many studies in road safety. In Smeed (1949) it is argued that the annual number of fatalities per registered motor vehicle can be explained by means of the motorization, measured by the number of registered motor vehicles per capita. The availability of more detailed time series data have led to advanced and interesting statistical studies on road safety. An example is the introduction of the use of traffic volume data. Traffic volume (e.g. vehicle kilometres driven, sometimes travel kilometres) is currently assumed to be one of the most important factors available for the explanation of accident counts. Appel (1982) found an exponentially decaying risk when he decomposed the (expected) number of accidents in a risk component (accidents per kilometres driven) and exposure (kilometres driven). Similar approaches have been adopted by Broughton (1991) and Oppe (1989, 1991). These models are univariate (one dependent variable) and some consist of just one explanatory variable measuring traffic volume. Time-dependencies in the error structure are ignored and estimation is based on classical methods.

Various time series analysis techniques, on the other hand, do take time-dependencies in the error structure into account. For example, autoregressive integrated moving average (ARIMA) techniques with explanatory variables (ARIMAX) as developed by Box and Jenkins (1976) are used in the DRAG (Demand for Road use, Accidents and their Gravity) analyses of Gaudry (1984) and Gaudry and Lassarre (2000). A DRAG analysis consists of three stages: first the traffic volume is modelled, next the accidents using the estimated traffic volume, and then the number of victims per accident (severity). Such a DRAG analysis is focussed on explaining the underlying factors of road safety while earlier studies were more focussed on forecasting. The DRAG approach allows for a non-linear transformation of the data by means of Box-Cox transforms. The time series structure however is linear. The model in this paper disentangles exposure and risk by unobserved components that are estimated simultaneously rather than estimated by separate stages.

An alternative method to analysing road safety data was proposed by Harvey and Durbin (1986) and is based on a structural time series model with interventions. This approach has been applied in road safety analysis by a number of authors. Ernst and Brüning (1990), for example, used a structural time series model to assess the effect of a German seat belt law while Lassarre (2001) applied structural time series models to compare the road safety developments in a number of countries. The method of Harvey and Durbin (1986) can also be extended to the simultaneous modelling of traffic volume, road safety and severity, see Bijleveld, Commandeur, Gould, and Koopman (2005). In these approaches linear Gaussian time series techniques such as the Kalman filter are used for estimation, analysis and forecasting. In the present paper we need to adopt a nonlinear equivalent of a structural time series model. Linear estimation techniques can not be used as a result and therefore we rely on extended (nonlinear) Kalman filter techniques. Related approaches based on univariate counts and with latent factors were discussed by Johansson (1996).

In road safety analysis, the use of disaggregated data is useful when the separate series can be modelled more effectively than the original aggregated time series. For instance, the composition of transport modes inside urban areas is usually different from that outside urban areas. Therefore, traffic volume and safety are different in these two parts of the traffic system. The present paper implements a model-based simultaneous treatment of traffic volume and fatal accidents for inside and outside urban areas. An important feature of the method is that it can handle the temporal unavailability of traffic volume data at the disaggregated level, while still providing estimates of the disaggregated exposure and risk for the full sample.

The paper is organised as follows. Section 2 presents the data used in the application part of the paper. The relation between observed and unobserved factors within a multivariate

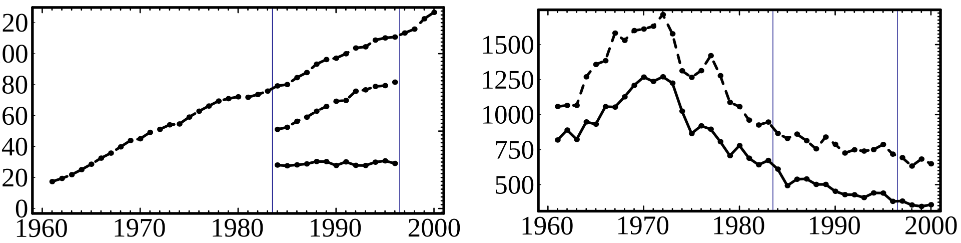

Figure 1: Traffic volume in billions of motor vehicle kilometres (left panel) and the number of fatal accidents (right panel) for inside urban areas (solid line) and outside urban areas (dashed line). The total traffic volume in the left panel is marked by a dashed line over the whole period.

nonlinear time series model is described in detail in Section 3, by first introducing the model and then providing a state space formulation of the model. A description of the estimation methods is given in Section 4. The main empirical results are presented in Section 5, and in Section 6 implications for road safety research are discussed. Section 7 concludes.

2 Data description

In the empirical study we analyse annual road traffic statistics from the Netherlands consisting of numbers of fatal accidents and traffic volume, defined as kilometres driven by motor vehicles, in the period 1961 up to and including 2000, both separated into inside and outside urban areas. This yields the following five annual time series:

y_{1t} the traffic volume inside urban areas

y_{2t} the traffic volume outside urban areas

y_{3t} the total traffic volume in the Netherlands

x_{1t} the number of fatal accidents inside urban areas

x_{2t} the number of fatal accidents outside urban areas

where time index $t = 1, \ldots, n$ represents the range of years from 1961 up to and including 2000. The total number of time points is therefore $n = 40$ in each series. All data were obtained from the Dutch Ministry of Transport and Statistics Netherlands while the accident information originated from police records.

The five time series are presented in Figure 1 with two displays. The left hand display shows the development of the motor vehicle kilometres in the Netherlands. Disaggregated figures of

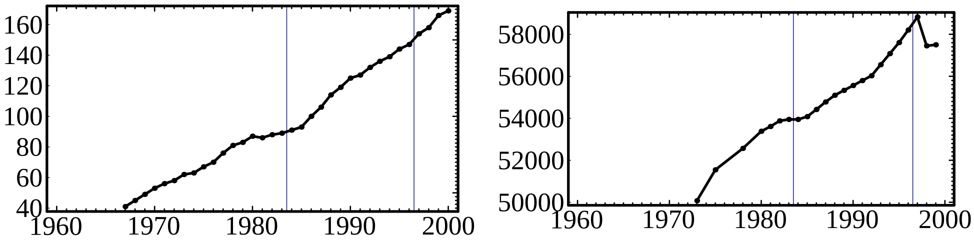

Figure 2: Traffic intensity index (left panel) and total length in kilometers of main roads outside urban areas (right panel).

traffic volume y_{1t} and y_{2t} are missing for the periods 1961 up to and including 1983 and 1997 up to and including 2000. For these years only the total traffic volume y_{3t} is available. Only modest deviations from an almost linear increase can be noticed from the traffic volume figures. These deviations are most likely caused by economic factors. The right hand display in Figure 1 shows the development of the number of fatal accidents in the Netherlands, both for inside and outside urban areas. The total number of fatal accidents has increased since the second world war. From the early 1970s the two series are decreasing but seem to level off near the end of the series.

The results of the empirical analysis in Section 5 will be validated against an alternative estimate of the traffic volume outside urban areas. This estimate is composed of indexed figures on traffic intensity on main roads multiplied by the length of the road system outside urban areas as obtained from a survey of municipalities. These two time series are presented in Figure 2. The data of the last years are considered to be inconsistent due to changes in registration. The product of the latter two series should be roughly equal to the development of motor vehicle kilometres outside urban areas when it is assumed that the development of the traffic intensity outside urban areas is approximately proportional to the intensity on the main roads.

3 The multivariate nonlinear time series model

3.1 Specification of model and assumptions

The multivariate nonlinear time series model is based on two unobserved components: a component for exposure (traffic volume) and a component for risk. Each component is bivariate to disentangle the effects for inside and outside urban areas. The statistical specification of the components is based on linear dynamic processes. It is assumed that the observed time series of fatal accidents and driven motor vehicle kilometres depend on these factors in the following ways:

1. The number of fatal accidents depends on the product of risk and exposure.

2. The number of driven motor vehicle kilometres in an area is proportional to the unobserved factor exposure. The proportionality can not be uniquely identified. As a consequence, the proportionality of the exposure is fixed at one.

3. The total number of driven motor vehicle kilometres is proportional to the sum of the unobserved factors of exposure inside and outside urban areas.

Disaggregated time series data for inside and outside urban areas is available for fatal accidents and driven kilometres although for the latter series this data is not available for the full sample. However the yearly series of total number of driven kilometres is available for the full sample. The five time series (partially missing for a number of years) are modelled simultaneously. A log-linear model can be considered to handle the multiplicative dependencies. However, it cannot at the same time handle the additive part for the missing disaggregated data. Therefore we adopt a multivariate nonlinear time series model.

The dynamic specification of the unobserved components is based on the following assumptions:

1. The unobservables are smooth functions of time, jumps and outliers are captured by interventions.

2. The exposure factors are trending (positive growth).

3. The risk factors decay exponentially over time. The log-risk factors are therefore trending (negative growth).

The latter item introduces a further nonlinear aspect of the model. The assumptions are partly motivated by the fact that both log-risk (see Appel, 1982) and exposure behave approximately linearly. This specification is well suited to fit the development of the number of fatal accidents inside and outside urban areas in Figure 1. Assume that exposure is the linear function of time $a \cdot t + b$ and risk is the exponential function of time $\exp(c \cdot t + d)$ for fixed scalars $a > 0$, b, $c < 0$ and d where t is the time-index. In a deterministic setting, the number of accidents is given by $(a \cdot t + b) \exp(c \cdot t + d)$ which implies that it is a function of $(a^* \cdot t + b^*) \exp(-t)$, where a^* and b^* are functions of a, b, c and d. The latter curve has a maximum at time $t^* = (a^* - b^*)/a^*$ if $a^* > 0$. Thus if the mobility is rising and the risk is decaying exponentially, the predicted number of fatal accidents has a maximum at some point in time. In our case the number of fatal accidents has a maximum in the early 1970s. In Figure 3 this relationship is shown for $a = 1$ and $b = 0$.

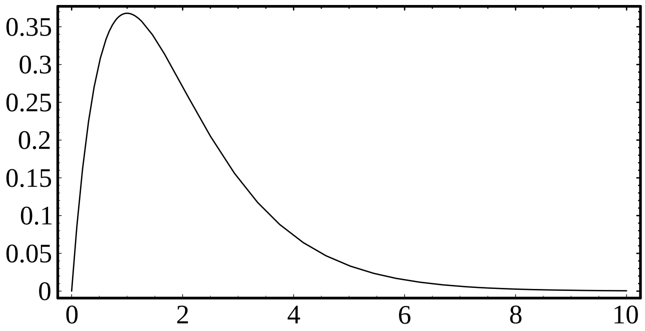

Figure 3: Graph of $t\exp(-t)$, (approximately) resembling the development of exposure times risk when exposure develops linearly over time and risk develops as an exponential transform of a linear development.

3.2 Unobserved stochastic linear trend factors

The deterministic trend specifications for exposure and log-risk are too rigid in practice because trends will not be constant over time in a long period of forty years. A time-varying trend is more flexible. A possible stochastic specification for a time-varying trend μ_t is the local linear trend model that is given by

$$\mu_{t+1} = \mu_t + \beta_t + \eta_t, \qquad \beta_{t+1} = \beta_t + \zeta_t, \qquad t = 1, \ldots, n, \tag{1}$$

where the disturbances η_t and ζ_t are normally distributed with mean zero and variances σ^2_η and σ^2_ζ, respectively. The disturbances η_t and ζ_s are mutually and serially independent of each other at all time points $t, s = 1, \ldots, n$. The initial values of μ_1 and β_1 can be regarded as realisations from a diffuse distribution or as fixed unknown coefficients, see the discussion in Durbin and Koopman (2001). The special case of $\sigma^2_\eta = \sigma^2_\zeta = 0$ is the deterministic linear trend $\mu_t = \mu_1 + \beta_1 \cdot t$ while the case of $\sigma^2_\eta > 0$ and $\sigma^2_\zeta = 0$ is the random walk plus fixed drift $\Delta\mu_t = \beta_1 + \eta_{t-1}$ for $t = 2, \ldots, n$. Further it is established that a smooth stochastic function of time μ_t is obtained by $\sigma^2_\eta = 0$ and $\sigma^2_\zeta > 0$, see Harvey (1989).

The observed time series y_t can be modelled with a time-varying trend as in

$$y_t = \mu_t + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{2}$$

where observation disturbance ε_t is normally distributed with mean zero and variance σ^2_ε. The disturbance ε_t is serially independent and mutually independent of the other disturbances η_s and ζ_s at all time points $t, s = 1, \ldots, n$. The linear trend-noise decomposition model in shorthand notation is then given by

$$y_t = \mu_t + \varepsilon_t, \qquad \mu_t \sim \text{LLT}, \qquad \varepsilon_t \sim \text{WN}, \tag{3}$$

where LLT refers to the local linear trend component (1) and WN to a Gaussian white noise sequence.

The unobserved exposure factors for inside and outside urban areas are indicated by μ_{1t} and μ_{2t}, respectively. The log-risk factors for inside and outside urban areas are indicated by δ_{1t} and δ_{2t}, respectively. Given the discussion in the previous section and to gain flexibility in modelling, we consider local linear trend specifications for the unobserved factors, that is

$$\mu_{it} \sim \text{LLT}, \qquad \delta_{it} \sim \text{LLT}, \qquad i = 1, 2, \qquad t = 1, \ldots, n.$$

All disturbance sequences driving the four unobserved factors are mutually independent of each other.

3.3 Observation equation

The dynamic mutual dependencies of the five observed time series are specified solely through the four unobserved and independent factors. This leads to a relatively simple model specification for the observed time series. Given the discussion of the model in section 3.1, the model equations for the observed traffic volume for inside and outside urban areas are given by

$$y_{it} = \mu_{it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \text{WN}, \qquad i = 1, 2, \qquad t = 1, \ldots, n, \tag{4}$$

whereas for the total traffic volume we have

$$y_{3t} = y_{1t} + y_{2t} = \mu_{1t} + \mu_{2t} + \varepsilon_{1t} + \varepsilon_{2t}.$$

These observation equations are linear and can be regarded as a special trivariate common trends model with two independent stochastic trends. It should be noted that when no observations are available for y_{1t} and y_{2t}, the disturbances ε_{1t} and ε_{2t} can not be identified separately. The sum $\varepsilon_{1t} + \varepsilon_{2t}$ can be identified when only y_{3t} is observed. Therefore, in this case we take $\varepsilon_{1t} + \varepsilon_{2t}$ as a Gaussian white noise sequence with mean zero and variance $\sigma^2_{\varepsilon,1} + \sigma^2_{\varepsilon,2}$ where $\sigma^2_{\varepsilon,i}$ is the variance of ε_{it} for $i = 1, 2$.

The statistical model specification for the number of fatal accidents in and outside urban areas is given by

$$x_{it} = \mu_{it} \cdot \exp(\delta_{it}) + \xi_{it}, \qquad \xi_{it} \sim \text{WN}, \qquad i = 1, 2, \qquad t = 1, \ldots, n. \tag{5}$$

This relationship is nonlinear in both μ_{it} and δ_{it}. There is no need to assume that all unobserved factors are independent of each other. Correlation between risk inside and outside urban areas can be estimated. This also applies to exposure.

3.4 State space model formulation

The local linear trend model (1) can be formulated in state space form by

$$\begin{pmatrix} \mu_{t+1} \\ \beta_{t+1} \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} + \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix},$$

and the observation equation (2) can be specified as

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix} + \varepsilon_t.$$

The general linear state space model is given by

$$\alpha_{t+1} = T\alpha_t + Hu_t, \qquad y_t = Z\alpha_t + Gu_t, \qquad u_t \sim \text{WN}, \tag{6}$$

where α_t is the state vector and u_t is the disturbance vector with mean zero and variance matrix V. The matrices and vectors T, H, Z and G are system matrices. The local linear trend model in the general setting is given by

$$\alpha_t = (\mu_t, \beta_t)', \qquad u_t = (\eta_t, \zeta_t, \varepsilon_t)',$$

with system matrices

$$T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad Z = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

The initial state vector is taken as a realisation from a diffuse density but can also be regarded as fixed and unknown coefficients.

The local linear trend models for exposure and log-risk, inside and outside urban areas, can be simultaneously put in state space form by placing the trends μ_{it} and δ_{it} with their associating slope terms, for $i = 1, 2$, in the state vector α_t. The disturbance terms are put in u_t. The system matrices for the state equation are given by

$$T = I_4 \otimes \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad H = I_4 \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

where I_4 is the 4×4 identity matrix.

The multivariate observation equation for traffic volume is linear. In terms of the observation vector $y_t = (y_{1t}, y_{2t}, y_{3t})'$ and the state vector α_t it follows that

$$y_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} \alpha_t + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \otimes \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} u_t. \tag{7}$$

When observations y_{1t} and y_{2t} are missing, the system does not need to be adjusted since the necessary adjustments are made within the estimation methods that are going to be employed.

The observation equation for the number of fatal accidents, inside and outside urban areas, can also be formulated in terms of the state vector α_t but it requires a nonlinear specification. Define $x_t = (x_{1t}, x_{2t})'$ and consider the nonlinear observation equation

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = Z(\alpha_t) + Gu_t, \qquad t = 1, \ldots, n, \tag{8}$$

where $Z(\cdot)$ is a continuous 5×1 vector function. The observation equation for y_t remains linear and is as given by (7). The observation equation for x_t is given by

$$x_t = \begin{pmatrix} \mu_{1t} \exp \delta_{1t} \\ \mu_{2t} \exp \delta_{2t} \end{pmatrix} + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} u_t, \tag{9}$$

where

$$\begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} \alpha_t, \qquad \begin{pmatrix} \delta_{1t} \\ \delta_{2t} \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} \alpha_t.$$

This completes the state space formulation of the multivariate nonlinear model that is the basis of the empirical study discussed in Section 5.

4 Estimation of parameters and latent factors

The linear Gaussian state space model may contain unknown parameters such as the variances of the disturbances σ_η^2, σ_ζ^2 and σ_ε^2 in the local linear trend model of section 3.2. These unknown parameters can be estimated by the method of maximum likelihood. For a linear model, the Gaussian log-likelihood function is evaluated by the Kalman filter and is maximised numerically, see Harvey (1989) and Durbin and Koopman (2001) for recent discussions on the maximum likelihood approach of estimating state space models.

The Kalman filter recursively evaluates the estimator of the state vector conditional on past observations $Y_{t-1} = \{y_1, x_1, \ldots, y_{t-1}, x_{t-1}\}$. The conditional estimator of the state vector is denoted by $a_{t|t-1} = \mathrm{E}(\alpha_t | Y_{t-1})$ and its conditional variance matrix $P_{t|t-1} = \mathrm{var}(\alpha_t | Y_{t-1})$. The Kalman filter is given by the set of vector and matrix equations

$$\begin{aligned} v_t &= y_t - Za_{t|t-1}, & F_t &= ZP_{t|t-1}Z' + GG', \\ & & K_t &= (TP_{t|t-1}Z' + HG')F_t^{-1}, \\ a_{t+1|t} &= Ta_{t|t-1} + K_t v_t, & P_{t+1|t} &= TP_{t|t-1}T' - K_t F_t^{-1} K_t' + HH', \end{aligned} \tag{10}$$

for $t = 1, \ldots, n$ and where $a_{1|0}$ and $P_{1|0}$ are the unconditional mean and variance of the initial state vector, respectively. When an initial state element is taken as a realisation from a diffuse density, we can take its mean as zero and its variance as a very large value. Exact treatments of diffuse initialisations are discussed in Durbin and Koopman (2001). The vector v_t is the one-step ahead prediction error with variance matrix F_t. The optimal weighting for filtering is determined by the Kalman gain matrix K_t. The joint density of the observations can be expressed as a product of predictive densities via the prediction error decomposition. As a result, the log-likelihood function can be constructed via the Kalman filter and is given by

$$\ell = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log|F_t| - \frac{1}{2}\sum_{t=1}^{n} v_t' F_t^{-1} v_t. \tag{11}$$

With diffuse state elements, the log-likelihood function requires some modifications. For a linear Gaussian state space model, the log-likelihood function ℓ is exact.

When a value for y_t is not available, it is taken as a missing value. The Kalman filter can handle missing values in a straightforward way. A direct consequence of a missing value y_t is that innovation v_t can not be computed and is unknown. This can be reflected by taking $v_t = 0$ and variance matrix $F_t \to \infty I$ such that $F_t^{-1} \to 0$ and $K_t \to 0$. It follows that the state update equations become

$$a_{t+1|t} = Ta_{t|t-1}, \qquad P_{t+1|t} = TP_{t|t-1}T' + HH',$$

when y_t is missing. These computations are repeated for different time indices t when a number of (consecutive) observations are missing. This solution also applies to out-of-sample forecasting or back-casting computations. A missing value does not enter the log-likelihood expression of (11). In a multivariate context, when an element of y_t is missing, the same element of v_t is taken as zero and the associating rows and columns of F_t^{-1} and K_t are taken as zero vectors.

The nonlinearities in the multivariate model are treated by the extended Kalman filter that is based on a first-order Taylor expansion of the nonlinear relation. Since the nonlinearity is limited to the observation vector, we only require the linearisation of $\mu_{it}\exp\delta_{it}$ around some known values $(\mu_{it}^*, \delta_{it}^*)$, that is

$$\begin{aligned}
\mu_{it}\exp\delta_{it} &\approx \mu_{it}^*\exp\delta_{it}^* + (\partial\mu_{it}\exp\delta_{it}/\partial\mu_{it})_{|(\mu_{it}=\mu_{it}^*,\delta_{it}=\delta_{it}^*)}(\mu_{it}-\mu_{it}^*) + \\
&\quad (\partial\mu_{it}\exp\delta_{it}/\partial\delta_{it})_{|(\mu_{it}=\mu_{it}^*,\delta_{it}=\delta_{it}^*)}(\delta_{it}-\delta_{it}^*) \\
&\approx \mu_{it}^*\exp\delta_{it}^* + \exp\delta_{it}^*(\mu_{it}-\mu_{it}^*) + \mu_{it}^*\exp\delta_{it}^*(\delta_{it}-\delta_{it}^*) \\
&\approx \exp\delta_{it}^*(-\mu_{it}^*\delta_{it}^* + \mu_{it} + \mu_{it}^*\delta_{it}),
\end{aligned}$$

for $i = 1, 2$ and $t = 1, \ldots, n$. The linearisation is more accurate when the value of $(\mu_{it}^*, \delta_{it}^*)$ is close to (μ_{it}, δ_{it}). Within the Kalman filter, the nonlinear function $Z(\cdot)$ is linearised in this way with an expansion around the filtered state vector. It implies that μ_{it}^* and δ_{it}^* are taken from the appropriate elements in $a_{t|t-1}$. The necessary amendments of the Kalman filter lead to matrix Z becoming time-varying, Z_t, and requires the replacement of the first three equations of the Kalman filter by

$$v_t = y_t - c_t - Z_t a_{t|t-1}, \qquad F_t = Z_t P_{t|t-1} Z_t' + GG', \qquad K_t = (TP_{t|t-1}Z_t' + HG')F_t^{-1},$$

where

$$c_t = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\mu^*_{1t}\delta^*_{1t}\exp\delta^*_{1t} \\ -\mu^*_{2t}\delta^*_{2t}\exp\delta^*_{2t} \end{pmatrix}, \qquad Z_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \exp\delta^*_{1t} & 0 & \mu^*_{1t}\exp\delta^*_{1t} & 0 \\ 0 & \exp\delta^*_{2t} & 0 & \mu^*_{2t}\exp\delta^*_{2t} \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix},$$

and with

$$\begin{pmatrix} \mu^*_{1t} \\ \mu^*_{2t} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} a_{t|t-1}, \qquad \begin{pmatrix} \delta^*_{1t} \\ \delta^*_{2t} \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0 \end{pmatrix} a_{t|t-1}.$$

The extended Kalman filter approximates the nonlinear features of the model. The prediction error is therefore not evaluated exactly and the log-likelihood function (11) is an approximation. Parameter estimation via the maximisation of this log-likelihood is referred to as quasi-maximum likelihood.

The smoothed estimate of a latent factor is the conditional mean given all available observations in the sample. The smoothed estimate of the state vector is denoted by $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ with its variance matrix $V_t = \mathrm{var}(\alpha_t|Y_n)$. Once the Kalman filter is carried out, the smoothed estimates can be computed via the backward recursions

$$\begin{array}{rclcrcl} r_{t-1} & = & Z_t'F_t^{-1}v_t + L_t'r_{t-1}, & \qquad & N_{t-1} & = & Z_t'F_t^{-1}Z_t + L_t'N_{t-1}L_t, \\ \hat{\alpha}_t & = & a_{t|t-1} + P_{t|t-1}r_{t-1}, & & V_t & = & P_{t|t-1} - P_{t|t-1}N_{t-1}P_{t|t-1}, \end{array} \tag{12}$$

where $L_t = T - K_tZ_t$ and with initialisations $r_n = 0$ and $N_n = 0$. The algorithm is a variation of the fixed interval smoothing method of Anderson and Moore (1979) and is developed by de Jong (1989) and Kohn and Ansley (1989). The smoothing recursions apply to the linear Gaussian state space model. However, since we have explicitly used a time-varying Z_t, the computations can also be carried out in conjunction with the extended Kalman filter. We note that smoothing requires the storage of all Kalman filter quantities, including the time-varying values of c_t and Z_t, for $t = 1, \ldots, n$.

5 Empirical results: estimation and model selection

We consider the five time series described in Section 2 for the years 1961–2000. The disaggregated time series of traffic volume is only observed for the sample 1984–1996 and therefore we need to deal with many missing values in the data set. The traffic volume series y_t are modelled by local linear trend models while the number of fatal accidents x_t are subject to a nonlinear relation, that is

$$y_{it} = \mu_{it} + \varepsilon_{it}, \qquad x_{it} = \mu_{it}\exp(\delta_{it}) + \xi_{it}, \qquad i = 1, 2,$$

for $t = 1, \ldots, n$ with $\mu_{it}, \delta_{it} \overset{\text{i.i.d.}}{\sim}$ LLT and $\varepsilon_{it}, \xi_{it} \overset{\text{i.i.d.}}{\sim}$ WN. Note that $i = 1$ refers to outside urban areas and $i = 2$ refers to inside urban areas. The total traffic volume is simply considered as $y_{3t} = y_{1t} + y_{2t}$. Each local linear trend model requires the estimation of two variances while the observation disturbances also have unknown variances.

The Poisson counts of number of yearly accidents are approximated by a normal distribution. To some extent we can account for this by taking the variance of the observation disturbances for x_{it} equal to the mean for which we take x_{it} as a proxy, with $i = 1, 2$. This leads to a time-varying sequence for matrix G in (6). To deal with the overdispersion of count data when a Gaussian approximation is used, the observation variance for x_{it} is scaled by a factor larger than 1. The variance of the observation disturbances for x_{it} is given by $x_{it}(1 + \exp\theta)$ where θ is estimated as part of the maximum likelihood procedure.

After obtaining the first estimation results, diagnostic tests based on the so-called auxiliary residuals (see Durbin and Koopman, 2001) indicated that several trend breaks can be identified. By including dummy intervention variables for trend breaks in the model and re-estimating the model with the intervention variables, a satisfactory multivariate nonlinear model for the time series was obtained. The estimation results of the latter model are discussed below.

5.1 Parameter estimation results

Table 1 presents the estimates of the parameters in the final model. Together with the estimates of variances and regression coefficients for the intervention variables, the table reports the 95% lower and upper limits of the confidence intervals. The confidence intervals are based on the approximation discussed in Harvey (1989, page 142). Since variance parameters are restricted to be non-negative, the logged variances are estimated and related confidence intervals are therefore asymmetric.

The overdispersion of Poisson counts in Gaussian approximation does not appear to be significant since the maximum likelihood estimate of parameter θ defined in the previous section is found to be a very negative number. Since the dispersion is modelled by $(1 + \exp\theta)$ and $\exp\theta \approx 0$ when θ is very negative, this parameter was removed from the model.

The estimates of the variances of the level disturbances of the exposure components μ_{it}, for $i = 1, 2$, are also found not to deviate from zero. This leads to a so-called smooth trend specification for the μ_{it}. As indicated by the maximum likelihood estimates of the variances corresponding to the slope components of the exposures in Table 1, the variation in the growth of exposure is estimated to be larger for outside urban areas (≈ 1.70) than for inside urban areas (≈ 0.03). These estimates rely on the limited sample period 1984–1996. The time series plots in Figure 1 confirm that the traffic volume inside urban areas is almost constant over these years while the growth of traffic volume outside urban areas has increased more rapidly in the

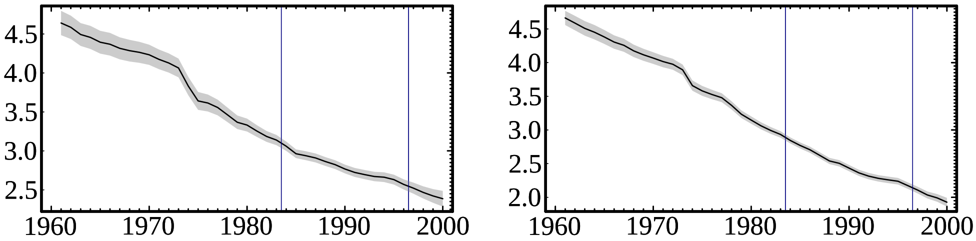

Figure 4: Time series plots of estimated trends of the log-risk component δ_{it} for inside (left panel) and outside (right panel) urban areas. Disaggregated traffic volume data is available in the period within vertical lines. The shaded areas indicate 95% confidence intervals.

period before 1990 than the period after 1990. The variances of the observation disturbances appear to be smaller for traffic volume outside urban areas (≈ 0.08) than for inside urban areas (≈ 0.75).

The slope variances of the log-risk components δ_{it} for $i = 1, 2$, are estimated as zero which reduces the log-risk trends to random walk processes with fixed growth terms. Since the level variances are estimated to be small ($\approx$ 0.001) for both inside and outside urban areas, the log-risks are close to a fixed trend. However, it should be emphasised that although these estimated variances are small they still appear to deviate from zero significantly.

The maximum likelihood estimates of the regression coefficients for the intervention variables in Table 1 imply, significant breaks in the log-risk trends for the years 1974 and 1975. They can partly be attributed to the global 'oil crisis' in 1974 and the introduction of alcohol legislation in the Netherlands. This legislation was officially introduced in November 1974. In the following year legislation on wearing moped helmets (February 1975) and seat belt legislation (June 1975) was introduced. To disentangle the effects of these measures, more detailed accident and mobility data is required. For example, the availability of time series with quarterly or monthly frequencies and of disaggregated data with a longer time horizon may be useful in this respect. Further research in this direction is however beyond the scope of this study since relevant data is not easily available for the Netherlands.

5.2 Signal extraction: trends for exposure and risk

Figure 4 presents the estimated trends for the risk of inside and outside urban areas. The apparent accelerated decrease in the trend of the risk for inside urban areas is the result of the interventions in 1974 and 1975 whereas for outside urban areas it is the result of the effect of the intervention in 1974. These trends represent the estimates of the unobserved log-risk component δ_{it}.

Figure 5 displays the trends of exposure inside and outside urban areas. The exposure

Table 1: Estimation results of variances and interventions in equations for inside and outside urban areas. Only those parameter estimates are reported that significantly deviate from zero. The lower and upper limits of the asymmetric 95% confidence interval are given below the estimated value, in brackets.

Parameter		*Estimated value*	
		inside	*outside*
Variance of			
Slope exposure μ_{it}		0.0312	1.6999
		(0.0191 0.0509)	(1.1974 2.4132)
Level log-risk δ_{it}		0.0012	0.0012
		(0.0007 0.0020)	(0.0005 0.0025)
Irregular traffic volume y_{it}		0.7492	0.0794
		(0.5177 1.0844)	(0.0183 0.3451)
Intervention in			
	1974	−0.1822	−0.1705
		(−0.2850 −0.0794)	(−0.2567 −0.0842)
	1975	−0.1387	
		(−0.2464 −0.0310)	

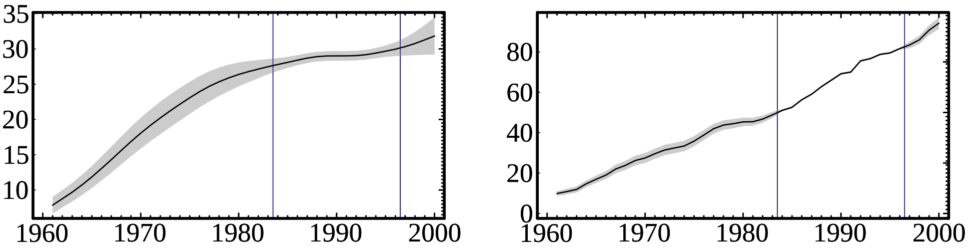

Figure 5: Time series plots of estimated trends of the exposure component μ_{it} for inside (left panel) and outside (right panel) urban areas. Disaggregated traffic volume data is available in the period within vertical lines. The shaded areas indicate 95% confidence intervals.

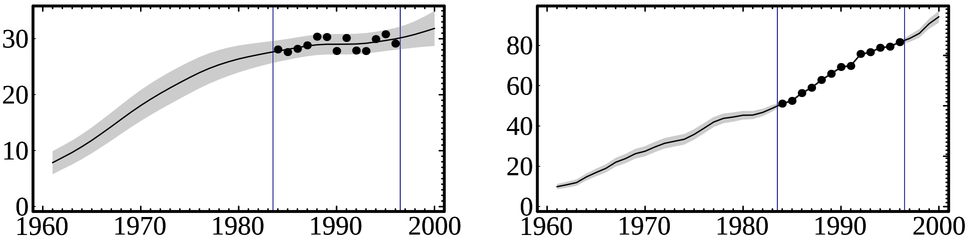

Figure 6: Estimated versus observed motor vehicle kilometres for inside (left panel) and outside (right panel) urban areas. Solid lines represent the model estimates and dots are the observed values. Disaggregated traffic volume data is available in the period within vertical lines. The shaded areas are 95% confidence intervals.

inside urban areas increases steadily from the 1960s onwards until it levels off at the end of the 1970s. It starts slowly increasing again from the 1990s onwards. It may be noted that the stabilisation of the exposure inside urban areas in the 1970s takes place before the period for which disaggregated traffic volume data is available. This shows that the methodology enables the recognition of such changes before disaggregated data is available. In comparison with the trend of exposure inside urban areas, the margin of confidence in the trend of exposure outside urban areas is small. Moreover, the outside trend is growing more consistently over the years although some minor temporary fluctuations of trend increases can be observed. Such fluctuations are detected even at time points where traffic volume data outside urban areas is not available. For example, in the period before 1980, when the outside exposure trend levels off, such fluctuations are identified partly through the aggregated traffic volume, see Figure 1. On the other hand, the fluctuations do not appear in the inside exposure trend notwithstanding its larger confidence interval. This can be explained by the fact that the estimated exposure trends also rely on the observed time series of number of fatal accidents. Since more fatal accidents occur outside urban areas, it is apparently more likely that the fluctuations in the number of accidents affect outside exposure more than inside exposure.

5.3 The fit of the model

This section concentrates on the ability of the multivariate nonlinear model to fit the time series of motor vehicle kilometres and fatal accidents, inside and outside urban areas. In Figures 6 and 7 the model predictions are represented as solid lines, with 95% confidence intervals represented by shaded areas, and the observed data is represented as enlarged dots. The confidence intervals are based on the estimated variances of the disturbances.

The estimated values for the motor vehicle kilometres in Figure 6 are equal to the trends μ_{it} for exposure discussed in the previous section. The fit of the estimated model is quite

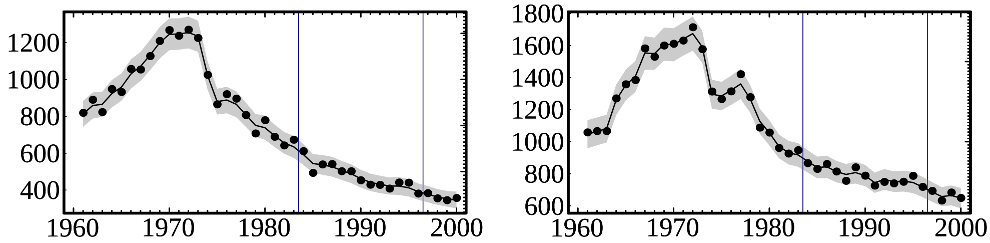

Figure 7: Estimated versus observed number of fatal accidents for inside (left panel) and outside (right panel) urban areas. Solid lines represent the model estimates and dots are the observed values. Disaggregated traffic volume data is available in the period within vertical lines. The shaded areas indicate 95% confidence intervals.

satisfactory. The estimated number of fatal accidents in Figure 7 is based on the nonlinear function $\mu_{it} \exp \delta_{it}$. The effectiveness of this simple nonlinear relationship is convincing given the good fit of the estimated number of accidents to the data. Apart from some small differences, the estimates for inside and outside urban areas show similar patterns. It is encouraging that the model has identified the sudden increase in the number of fatal accidents outside urban areas in 1975–1977 whereas the number of accidents inside urban areas continues to decrease in this period.

5.4 External validation

To further validate the estimates obtained by the model, we consider the estimated trend for the exposure outside urban areas displayed in the right panel of Figure 5. These estimates are also presented as the solid line in Figure 8. Since traffic volume data outside urban areas is only available for the years 1984 up to 1996, the fit between the observed volume data outside urban areas and the estimated trend can only be evaluated for this 13 year period. However, as mentioned in Section 2, an alternative indicator for exposure outside urban areas is available which extends beyond the 13 year period. This alternative indicator is obtained by multiplying the indexed traffic intensity on main roads in the Netherlands with the total length of roads outside urban areas. Since this alternative indicator is measured on a different scale from the motor vehicle kilometers driven outside urban areas, the values of the latter observations were regressed on the alternative indicator observations for the years 1984 up to 2000. The predicted values of this simple regression without intercept yield properly re-scaled alternative indicator observations and are plotted as dots in Figure 8. As the figure shows, the estimated trend for exposure outside urban areas is quite consistent with the alternative indicator values, even in the eleven year period from 1973 through 1983 for which no motor vehicle kilometres driven were available.

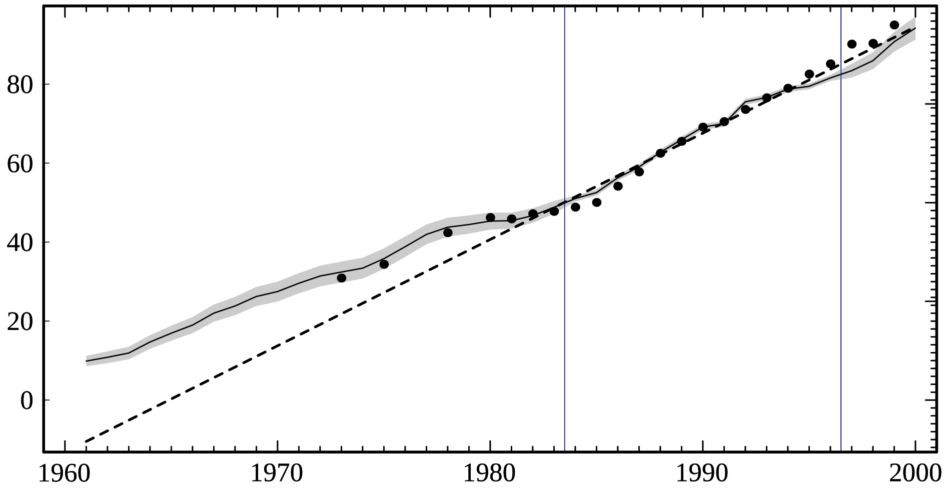

Figure 8: The fit of traffic volume outside urban areas: extrapolation and validation of the model. The fit implied by the multivariate nonlinear model is represented by a solid line. Traffic volume outside urban areas is only available in the period within vertical lines. The alternative indicator observations of volume are represented by the dots. The dashed line reflects the linear extrapolation of the traffic volume data outside urban areas.

Finally, alternative back-casts and forecasts can be produced by the linear extrapolation of traffic volume outside urban areas. These back- and forecasts are shown in Figure 8 as a dashed line. Especially the back-casts of the nonlinear state space model are clearly superior to a simple extrapolation of the traffic volume data.

6 Implications for road safety research

The current results offer the possibility to interpret the disaggregated developments of road safety over a much longer period of time than the 13 year period of 1984 up to and including 1996 for which all disaggregated data is available. Previous research by Appel (1982, for Germany) and Oppe (1989, for the Netherlands) of macroscopic developments in road safety led to the conclusion that the risk of road safety expressed as the number of persons killed per motor vehicle kilometre (which strongly resembles the development of the number of fatal accidents per motor vehicle kilometre) suggested an exponentially decreasing tendency. This conclusion is supported by our model-based approach of analysing disaggregated risk.

From Figure 6 we learn that the development of Dutch traffic volume has increased since the 1960s. Disaggregating the traffic volume for inside and outside urban areas shows that the traffic volume inside urban areas continued to increase until the end of the 1970s. It started to increase again from the 1990s. On the other hand, the traffic volume outside urban areas kept on growing more consistently and strongly with the largest acceleration between about 1983 and 1992. Although the increase of traffic mobility outside urban areas was limited in the early 1960s, it has increased more dramatically from the end of the 1960s when comparing it to mobility inside urban areas. It can therefore be concluded that this development was a dominant factor in the total traffic volume long before the beginning of the new century.

7 Conclusions

The model-based treatment of exponential and multiplicative relationships between number of accidents and factors such as exposure and risk has proven to be effective. A multivariate nonlinear time series model is estimated using a partially disaggregated data set of traffic volume and number of accidents. The estimation methods are based on extended versions of the standard multivariate Kalman filter and related algorithms. We have shown that a state space methodology in a multivariate and nonlinear setting with many missing observations is feasible and that it can lead to interesting empirical results. The empirical study consists of the analysis of road safety in the Netherlands by simultaneous consideration of two sections of the total traffic system: inside and outside urban areas. It is assumed that the development of road safety inside urban areas is different from the development of road safety outside urban areas due to differences in road infrastructure and changes in the use of road transport inside and outside urban areas over the years.

The empirical results show that developments of exposure inside and outside urban areas have roughly kept up with each other up to 1980. After this period, a decline of the growth in exposure inside urban areas occurred and lasted until approximately 1990. Then exposure inside urban areas started to increase again. In contrast, the exposure outside urban areas has steadily increased since 1980. The model has successfully reconstructed the development of traffic volume outside urban areas for a long time period. This is confirmed by considering an alternative estimate of traffic volume outside urban areas, based on the product of the index of traffic intensity and an estimate of the total road length, both outside urban areas. The similarity between these alternative data-driven estimates and the model estimates is convincing.

Although the empirical results are satisfactory, the methodology of this paper can be improved further. For example, the model may need to allow for covariances between the disaggregated values. Furthermore, introducing common components in the model may lead to statistically more significant dynamic relations between the series. Finally, the consideration of non-Gaussian features in the model may enhance the applicability of the current methodology in cases where small counts are observed.

References

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Appel, H. (1982). Strategische aspekten zur erhöhung der sicherheit im Straßenverkehr. *Automobil-Industrie 3*, 347–356.

Bijleveld, F. D., J. J. F. Commandeur, P. G. Gould, and S. J. Koopman (2005). Model-based measurement of latent risk in time series with applications. submitted.

Box, G. E. P. and G. M. Jenkins (1976). *Time series analysis*. San Francisco: Holden-Day.

Broughton, J. (1991). Forecasting road accident casualties in Great Britain. *Accident Analysis and Prevention 23*(5), 353–362.

de Jong, P. (1989). Smoothing and interpolation with the state-space model. *Journal of the American Statistical Association 84*(408), 1085–1088.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

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

Gaudry, M. (1984). Drag, un modèle de la demande routière, des accidents et leur gravité, appliqué au québec de 1956–1986. Technical Report Publication CRT-359, Centre de recherche sur les Transports, et Cahier #8432, Département de sciences économiques, Université de Montréal.

Gaudry, M. and S. Lassarre (Eds.) (2000). *Structural Road Accident Models: The International DRAG Family*. Oxford: Elsevier Science Ltd.

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

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

Johansson, P. (1996). Speed limitation motorway casualties: a time series count data regression approach. *Accident Analysis and Prevention 28*(1), 73–87.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*(1), 65–79.

Lassarre, S. (2001). Analysis of progress in road safety in ten european countries. *Accident Analysis and Prevention 33*, 743–751.

Oppe, S. (1989). Macroscopic models for traffic and traffic safety. *Accident Analysis and Prevention 21*, 225–232.

Oppe, S. (1991). Development of traffic and traffic safety: Global trends and incidental fluctuations. *Accident Analysis and Prevention 23*(5), 413–422.

Smeed, R. J. (1949). Some statistical aspects of road safety research. *Journal of the Royal Statistical Society A 112*(1), 1–34.