SWOV Fact sheet

SWOV
INSTITUTE FOR
ROAD SAFETY RESEARCH

Time series analysis

Summary
Time series analysis can be used to quantitatively monitor, describe, explain, and predict road safety developments. Time series analysis techniques offer the possibility of quantitatively modelling road safety developments in such a way that the dependencies between the observations of time series data are taken into account. This fact sheet explains time series analysis and discusses the functionality of ARMA, ARIMA, DRAG, and state space models. These models are illustrated by applying them to relatively simple descriptive analyses of time series.

Background and content
One important task of road safety research is to study, describe, explain, and predict road safety developments. If this is done quantitatively, analysis techniques are needed to do this reliably. As the term 'developments' implies, we are dealing with the analysis of a very special type of data. This data always consists of observations sequentially made over time of a particular aspect of the traffic process. An example is the annual number of road fatalities as observed over a number of years. Such a variable is known as a *time series*.

An important feature of the observations in a time series is that they are usually *not* independent of each other: after all, the observed number of road fatalities of last year is usually a fairly good indicator of the number of road fatalities this year. Given that standard techniques assume independent observations, the analysis of time series using standard techniques very often results in residuals that are also mutually correlated. At the same time, statistical tests and confidence limits are based on the crucial assumption that the residuals obtained from the analysis are random, and thus independent of each other.

This fact sheet gives an overview of the available techniques for the analysis of time series. In addition, we compare these techniques with each other in terms of their suitability for modelling road safety developments. This fact sheet is meant for researchers interested in time series analysis in various fields. For a proper understanding of the contents of this fact sheet, a basic knowledge of statistics and particularly of classical linear regression is required (see for example McCall, 1998).

What are time series analyses a solution for?
Time series analysis allows for a quantitative modelling of road safety developments in such a way that the dependencies in the observations are properly taken into account.

Who are time series analysis techniques meant for?
Time series analysis techniques are meant for researchers in all fields where repeated measurements over time are carried out. This certainly does not only apply to road safety, but also to economics, history, the social sciences, medicine, biology, et cetera.

How do time series analyses work?
Several techniques have been - and are still being - used for the analysis of time series. These techniques have in common that, in principle, they are not only capable of describing the development in an observed time series, but that they can also be used to find explanations for the development in an observed time series, and to predict future values of the development in an observed time series. However, some techniques are much better suited for these purposes than others.

This fact sheet gives an overview of the various options, and specifically discusses the advantages and disadvantages of each approach. Throughout, we will illustrate the various techniques by applying them to the analysis of the following time series: the annual number of road fatalities in the Netherlands during the period 1950-2003, as shown at the top of *Figure 1*.

Classical linear and non-linear regression models
Of old, classical linear and non-linear regression models have been used to analyse time series. In descriptive classical linear and non-linear regression analysis, time is used as an independent variable for modelling the trend, while independent dummy variables are used to model any seasonal effects if the time series consists of, for instance, quarterly or monthly observations.
Without seasonal effect, the descriptive classical linear regression model is very simply:

$$y_t = a + bt + \varepsilon_t, \qquad\qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \tag{1}$$

for time points t = 1, …, n. In (1) y_t is the dependent variable, i.e. the observed time series, t is the independent variable, i.e. time, a and b are the unknown parameters to be estimated, and ε_t is the residual. Parameter a is the intercept, and indicates the point where the regression line intersects with the y-axis; parameter b is the regression coefficient, and measures the slope of the regression line with respect to the x-axis. Behind the regression equation in (1) the *assumptions* of the model are given: the residuals are assumed to be normally and independently distributed (*NID*) with mean equal to zero, and variance equal to σ_ε^2. In addition, the residuals of the model are assumed to be homoscedastic, i.e. to have a constant variance in time. The order of importance of these three assumptions for correct statistical conclusions is as follows: first is independence, second is homoscedasticity, and last is normality.

Figure 1 shows the result of the linear regression of the logarithm of the annual number of road fatalities in the Netherlands on the variable 'time'. This figure shows that the regression line not only has large deviations from the observations in the time series, but that, in addition, the residuals to be found in the bottom graph of *Figure 1* are not independent at all: in the 1950-1960 period the observations in the series are consistently overestimated, in the 1961-1984 period they are consistently underestimated, and in the 1985-2003 period they are once more consistently overestimated. Since the calculation of the 95% confidence interval on either side of the regression line is carried out under the assumption of independent residuals, the confidence interval in the upper graph of *Figure* 1 is also completely inaccurate. This is confirmed by the fact that 40 of the 54 observations fall outside the 95% confidence interval, whereas according to chance, this should be about three observations.
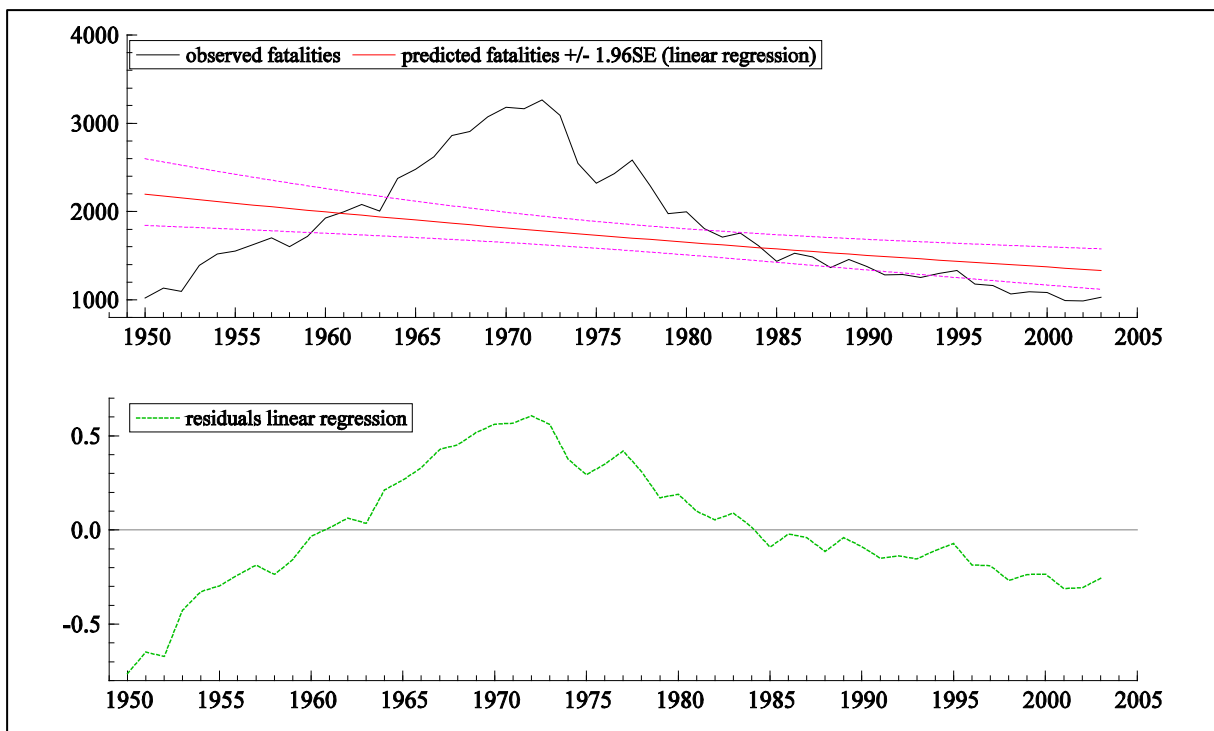

Figure 1. *Results of analysis with simple linear regression*

Time series analysis with non-linear regression models also often results in residuals that do not satisfy the assumption of independence. In this fact sheet the term 'non-linear model' refers to both a model in which the relation between the dependent and independent variables is non-linear, but in which the model itself may be 'linear in the parameters', and to a model that is non-linear in the parameters. For example, the analysis of the logarithm of the time series in *Figure 1* with what is known as a cubic trend

$$ y_t = a + b_1 t + b_2 t^2 + b_3 t^3 + \varepsilon_t, \qquad\qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \qquad\qquad (2) $$

for time points $t = 1,\ldots,n$, where a, b_1, b_2, and b_3 are the unknown parameters, gives the results shown in *Figure 2*. Although the residuals of this model are smaller than those in *Figure 1*, as appears from comparison of the scale on the *y*-axis of the plots with residuals in *Figures 1* and *2*, and the model predictions fit the data better, the residuals are still serially correlated as can be seen in the plot at the bottom of *Figure 2*. The residuals in the successive periods are once more consistently over- or underestimated. This again means that the 95% confidence limit on either side of the cubic trend in *Figure 2* is inaccurate. This is also illustrated by the fact that far too many observations still fall outside this interval.
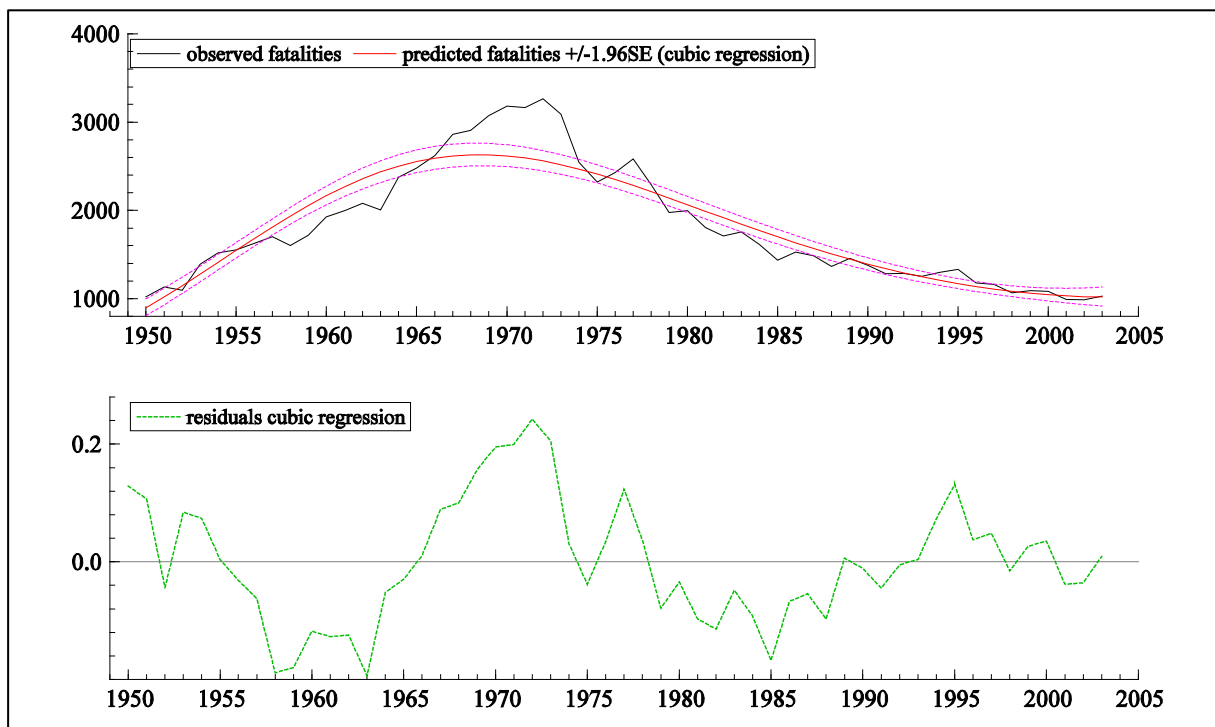

Figure 2. *Results of analysis with cubic regression*

Another example of the problems encountered when applying non-linear regression models to time series data can be found in Commandeur & Koornstra (2001). They modelled the developments in motor vehicle kilometres in the Netherlands with an S-shaped growth curve known as the Gompertz curve, and the fatality rate, defined as the number of fatalities divided by the motor vehicle kilometres, was modelled with an exponentially decaying curve. They predicted the number of fatalities as the product of the predictions obtained with these two models. These analyses showed that the residuals were also strongly serially correlated, and thus did not satisfy the assumption of independence. Broughton et al. (2000) also recognized the problem of serially correlated residuals in the application of linear regression models to developments in British road safety.

ARMA, ARIMA and DRAG models
ARMA and ARIMA models (Box and Jenkins, 1976) and DRAG models (Gaudry, 1984; Gaudry & Lassarre, 2000) have been specifically developed for the analysis of time series, and are therefore much better suited to handle the dependencies in the observations than classical linear and non-linear regression models.

Algebraically, the ARMA model can be written as follows:

$$y_t = b_1 y_{t-1} + \ldots + b_p y_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \ldots + a_q \varepsilon_{t-q}, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \qquad (3)$$

for time points t = 1, …, n, where b_1, …, b_p and a_1, …, a_q are the $p+q$ unknown parameters. In the first part on the right of the equals sign of (3), the dependent variable y_t is predicted by its own values shifted in time. This is called the *autoregressive* (AR) part of the equation. In the second part right of the equals sign the dependent variable is predicted on the basis of the shifted values in time of the residuals, and this is the *moving average* (MA) part of the equation. For a given time series it is the task of the researcher to determine an ARMA(p,q) model specification that strikes a balance between a good fit, and a parsimonious use of parameters. Furthermore it must result in residuals that satisfy the assumptions of independence, homoscedasticity, and normality.

However, an important requirement of ARMA models is also their weak point: the analysed times series must be *stationary*, i.e. it must have a constant mean and a constant variance in time, before the actual analysis can be carried out. Since time series in practice often consist of non-stationary components such as a trend or a seasonal effect (in the case of quarterly or monthly data), with ARMA models the observations usually first need to be filtered. This means that the trend and the seasonal effect first have to be removed from the series by subtracting consecutive observations from each other in order to achieve stationarity. This then results in what is known as ARIMA(p,d,q) models in which the value of d indicates how often consecutive observations must be subtracted from each other before stationarity of the time series is achieved. The actual analysis then consists of determining which ARMA(p,q) specification fits the filtered, and hopefully stationary, time series best.

The analysis of the logarithm of the annual number of road fatalities in the Netherlands during the 1950-2003 period with ARIMA models indicates that an ARIMA(0,2,2) specification provides a good description of this time series. In *Figure 1* it can be seen clearly that the time series is not stationary: first the annual number of road fatalities almost continuously increases, and then it almost continuously decreases. This time series must therefore first be differenced until stationarity is achieved. In this case, stationarity is approximately achieved by taking second order differences between the successive observations of the original series, so that in this case d equals 2. This filtered version of the time series can be seen in the upper part of *Figure 3*, together with the predictions obtained with the ARIMA (0,2,2) model. The result of the ARIMA (0,2,2) model specification, expressed in terms of the original time series, is shown at the bottom of *Figure 3* and the residuals are in the middle figure. For the first time, the residuals of the analysis do satisfy the important assumption of independence. This is also confirmed by the fact that most observations now fall within the 95% confidence limits.
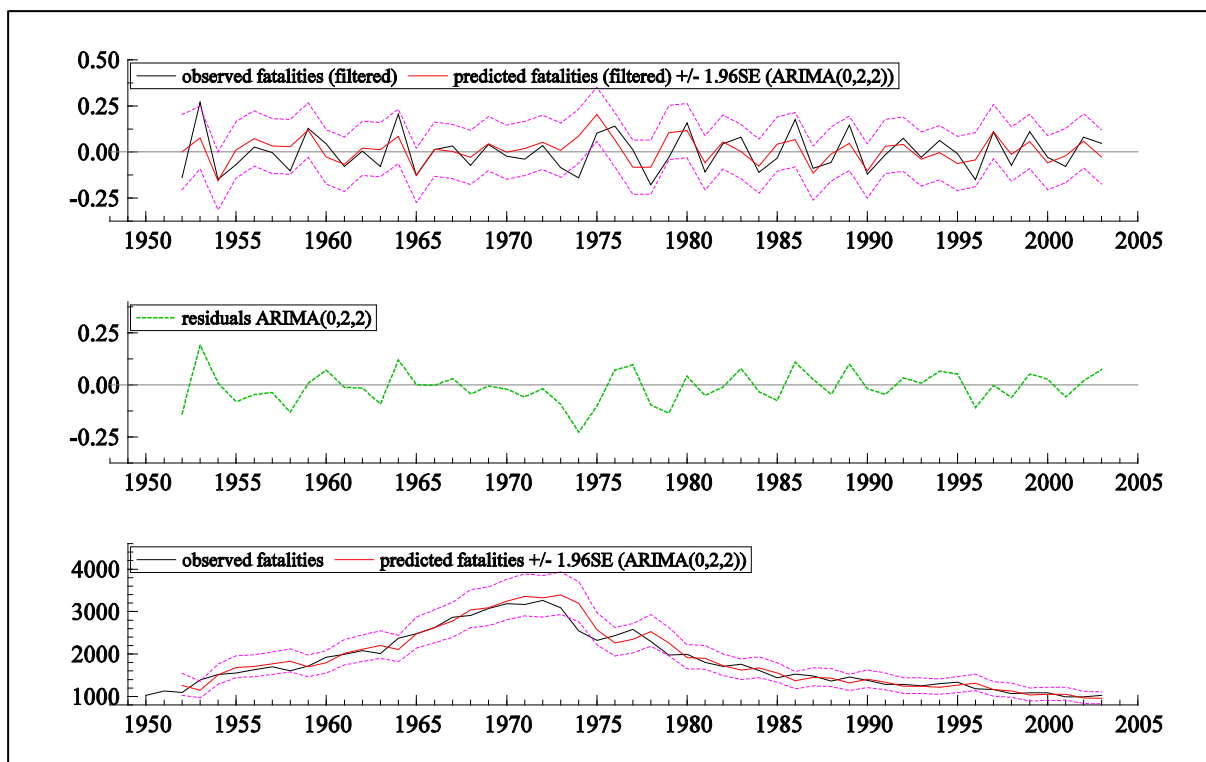
Figure 3. *Results of analysis with the ARIMA(0,2,2) model*

DRAG models (Gaudry, 1984; Gaudry & Lassarre, 2000) are extensions of ARIMA models where:
– exposure, accidents, and accident severity are modelled in three separate analysis steps;
– many explanatory variables are added to each of these three steps;
– both the dependent and independent variables may be transformed by what are called Box-Cox transformations.

In addition to the already mentioned stationarity requirements for ARIMA models, the disadvantages of DRAG models are therefore that:
– exposure, accidents, and accident severity are not modelled simultaneously;
– the large number of explanatory variables appears to be at odds with the principle of parsimony;
– the Box-Cox transformations of the variables, that are claimed to be an advantage, can lead to identification and interpretation problems.

State space models
A more recent development in the field of time series analysis are what are known as structural time series models (Harvey, 1989; Durbin & Koopman, 2012; Commandeur & Koopman, 2007). They are also called unobserved component models or state space models. In contrast to the ARIMA and DRAG models discussed above, structural time series models assume no stationarity of time series, and offer the possibility to explicitly decompose time series into components like a trend and a seasonal effect.

In general, all state space models can be expressed in a general notation. However, as this notation requires familiarity with matrix algebra, we will limit ourselves here to discussing only one state space model. We have chosen that model that provides a good description of the development in the time series of annual road fatalities in the Netherlands: the local linear trend model.

The local linear trend model is obtained by allowing the intercept *a* and the regression weight *b* in the classical linear regression model (1) to *vary in time* in the following way:

$$y_t = a_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$

$$a_{t+1} = a_t + b_t + \xi_t, \qquad \xi_t \sim NID(0, \sigma_\xi^2)$$

$$b_{t+1} = b_t + \zeta_t, \qquad \varsigma_t \sim NID(0, \sigma_\varsigma^2) \qquad (4)$$

for time points t = 1, ..., n. Here, a_t is the trend that consists of the sum of a time varying intercept and of a time varying slope b_t. The local linear trend model can thus be regarded as a generalization of the classical linear regression model, with the important difference that the parameters that are global or fixed in model (1), i.e. they may not vary in time, are made local in model (4), and thus may vary from time point to time point. The unknown parameters in the local linear trend model are the error variances σ_ε^2, σ_ξ^2 and σ_ς^2, and the initial values a_1 and b_1 at time point t = 1 of the intercept and of the slope.
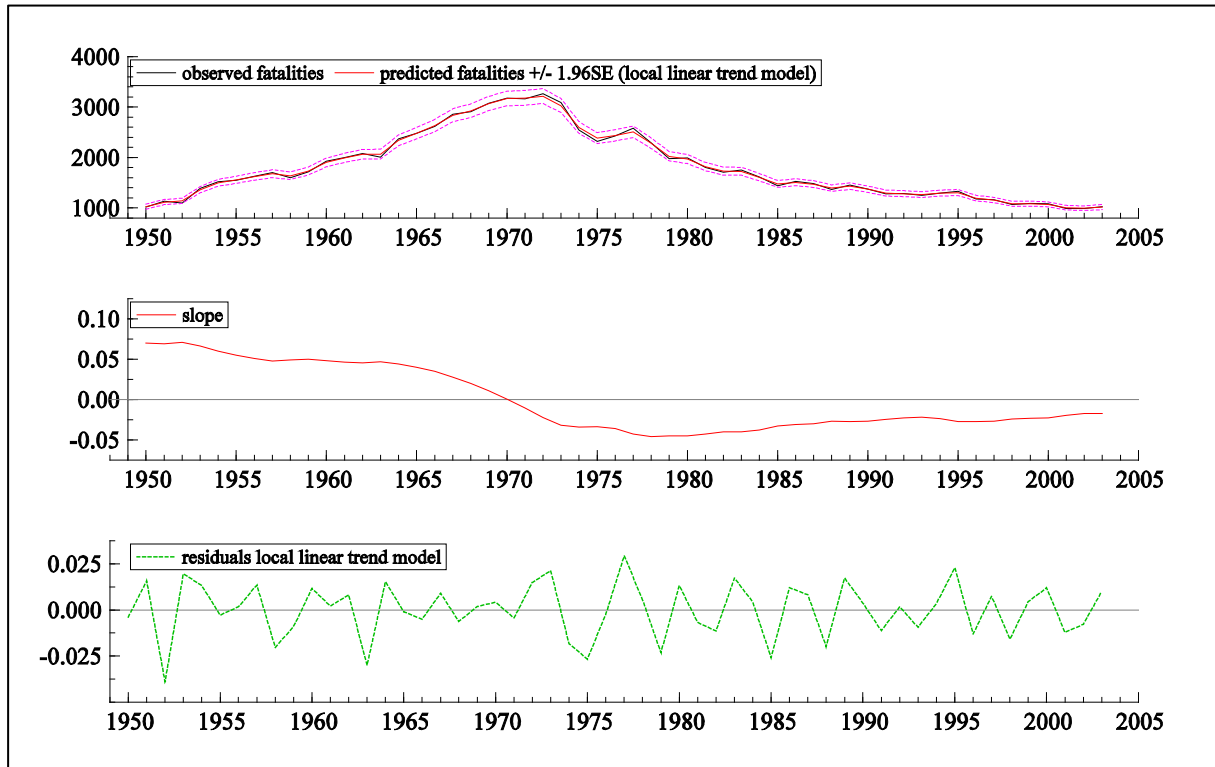

Figure 4. *Results of analysis with the local linear trend model*

The results of the analysis of the logarithm of the series of annual road fatalities in the Netherlands with model (4) are shown in *Figure 4*. The top figure shows the observations and the time varying trend, together with the 95% confidence interval; the middle figure shows the time varying slope; and the bottom figure shows the residuals of the model. The positive value of the slope in the 1950-1970 period indicates that the number of road fatalities during this period was continuously increasing, whereas the negative value of the slope shows that this trend continually decreased during the years after 1970. The fact that the slope gradually reaches zero at the end of the series means that the decreasing development is steadily levelling off. The residuals at the bottom of *Figure 4* meet the requirements of all model assumptions of independence, homoscedasticity, and normality, so that the 95% confidence limits on either side of the model predictions at the top of the figure are also correct. State space models are very flexible: they can easily cope with missing observations, and can be easily expanded to multivariate time series analysis models. Another advantage of structural time series models is that they, as mentioned earlier, have classical regression as a special case. By fixing the error variances σ_ξ^2 and σ_ς^2 in (4) at zero, this model collapses to the classical linear model (1). This makes it possible to directly compare the results of analyses with these two types of models. ARIMA models can also be put in state space form, and therefore fitted by state space methods. In fact, all the analyses presented in this fact sheet were therefore carried out with state space methods.

It is for these reasons that structural time series models are preferable to other models if the purpose is to describe, explain, and predict road safety developments. Applications of these models in road safety research can, for example, be found in Harvey & Durbin (1986) where the effect of making seat belts obligatory in Great Britain was assessed, and in Ernst & Brüning (1990) where the same was

done for Germany. Based on their analysis of monthly data, Harvey and Durbin concluded that, among other matters, the introduction in Great Britain of the seat belt law on January 31st 1984 resulted in a 23% decrease in the number drivers being killed or seriously injured, and in a decrease of 18% in the number of drivers being killed. For West-Germany Ernst & Brüning found a decrease of 28% in the number of car passengers being killed and of 21% the number of car passengers being seriously injured as a result of the August 1st 1984 seat belt law.

Lassarre (2001) has used structural time series models for comparing the road safety developments among 10 European countries. An analysis of annual data for road fatalities and motor vehicle kilometres travelled of the 10 European countries showed an average annual decrease of the fatality rate of 6% in 1994. The smallest decrease was 4.7% in France and Finland, and the largest decrease was 13.4% in Spain.

Ord & Young (2003) applied structural time series models for the analysis of the effects of the attacks in New York on '9/11' 2001 on the developments in air and rail traffic. The results showed that a combination of three effects, i.e. a short term, a temporary, and a permanent shift, can trace the changes in the United States after the attacks quickest at the monthly level. SWOV has used structural time series models, among other things, to predict the number of road fatalities in the EU member states in 2010 (Ecorys Transport & SWOV, 2005) and the number of road fatalities in the Netherlands in 2020 (Wesemann, Van Norden & Stipdonk, 2010). For this purpose, the multivariate structural time series model as published in Bijleveld et al. (2008), and as discussed in Bijleveld (2008) has been used. In the framework of the European DaCoTa project, these models have also been applied for making prognoses for the number of road fatalities for all Member States of the European Union (see Martensen & Dupont, 2010; Lassarre, Dupont & Antoniou, 2012).

What do time series analyses produce?

For didactical reasons we have limited this fact sheet to relatively simple descriptive time series analyses. However, time series analysis can also be used to:
1. assess the effects of measures and of other explanatory variables on road safety development;
2. study whether newly published data deviates from that expected, when based on the past;
3. make predictions about future road safety developments.

Conclusion

If we compare the merits of classical (non-)linear regression models, ARMA, ARIMA and DRAG models, and state space models for time series analysis, then state space models have preference. Not only can they be used to explicitly decompose a time series in, for research, interesting components such as a trend or seasonal effect, but due to their flexibility they are usually also well able to handle the dependencies in time series observations. Furthermore, they work transparently with missing data and are easily generalised to the multivariate analysis of time series.

Publications en sources

Bijleveld, F. (2008). *Time series analysis in road safety research using state space methods*. PhD Thesis, Vrije Universiteit VU, Amsterdam. SWOV Dissertation series, Leidschendam.

Bijleveld, F.D., Commandeur, J.J.F., Gould, P. & Koopman, S.J. (2008). *Model-based measurement of latent risk in time series with applications*. In: Journal of the Royal Statistical Society Series A, vol. 171, nr. 1, p. 265-277.

Box, G.E.P. & Jenkins, G.M. (1976). *Time series analysis, forecasting and control*. ISBN 0816211043. Holden-Day, San Francisco.

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

Commandeur, J.J.F. & Koopman, S.J. (2007). *An introduction to state space time series analysis*. Oxford Practical Econometrics Series, nr. 1. Oxford University Press, Oxford.

Commandeur, J.J.F. & Koornstra, M.J. (2001). *Prognoses voor de verkeersveiligheid in 2010; mobiliteit en slachtofferrisico op grond van de ontwikkelingen in de jaren 1948-1998.* R-2001-9. SWOV, Leidschendam.

Durbin, J. & Koopman, S.J. (2012). *Time series analysis by state space methods.* Second Edition. Oxford University Press, Oxford.

Ecorys Transport & SWOV (2005). *Impact Assessment Road Safety Action Programme. Assessment for mid term review.* Ecorys Transport/SWOV, Rotterdam/ Leidschendam.

Ernst, G. & Brüning, E. (1990). *Fünf Jahre danach: Wirksamkeit der Gurtanlegepflicht für Pkw Insassen ab 1.8.1984.* In: Zeitschrift für Verkehrssicherheit, vol. 36, nr. 1, p. 2-13.

Gaudry, M. (1984). *DRAG, un modèle de la Demande Routière, des Accidents et de leur Gravité, appliqué au Québec de 1956 à 1982.* Publication CRT-359. Centre de Recherche sur les Transports CRT, Université de Montréal, Montréal.

Gaudry, M. & Lassarre, S. (eds.) (2000). *Structural road accident models. The international DRAG family.* Pergamon, Amsterdam.

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

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

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

Lassarre, S., Dupont, E. & Antoniou, C. (eds.) (2012). *Forecasting road traffic fatalities in European countries: Towards an integrated European model.* Deliverable 4.7 of the EC FP7 project DaCoTA. European Commission, Brussels.

Martensen, H. & Dupont, E. (eds.) (2010). *Forecasting road traffic fatalities in European countries: Model and first results.* Deliverable 4.2 of the EC FP7 project DaCoTA. European Commission, Brussels.

McCall, R.B. (1998). *Fundamental statistics for behavioral sciences.* Seventh Edition. Brooks/Cole Publishing Company, Pacific Grove.

Ord, K. & Young, P. (2004). *Estimating the impact of recent interventions on transportation indicators.* In: Journal of Transportation and Statistics. vol. 7, nr. 1.

Wesemann, P., Norden, Y. van & Stipdonk, H. (2010). *An outlook on Dutch road safety in 2020; future developments of exposure, crashes and policy.* In: Safety Science, vol. 48, nr. 9, p. 1098-1105.