Accident Analysis and Prevention 60 (2013) 424–434

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

On statistical inference in time series analysis of the evolution of road safety

Jacques J.F. Commandeur[a,b,*], Frits D. Bijleveld[b], Ruth Bergel-Hayat[c], Constantinos Antoniou[d], George Yannis[d], Eleonora Papadimitriou[d]

[a] VU University Amsterdam, The Netherlands
[b] SWOV Institute for Road Safety Research, P.O. Box 1090, 2260 BB Leidschendam, The Netherlands
[c] UPE IFSTTAR GRETTIA French Institute of Science and Technology for Transport, Development and Networks, Le Descartes 2, 2 rue de la Butte Verte, Marne la Vallée, 77166, France
[d] NTUA National Technical University of Athens, Greece

ARTICLE INFO

Article history:
Received 22 July 2011
Received in revised form 6 October 2012
Accepted 8 November 2012

Keywords:
Road safety
Time series
Regression
ARIMA models
DRAG models
State space methods
Structural time series models
Statistical theory

ABSTRACT

Data collected for building a road safety observatory usually include observations made sequentially through time. Examples of such data, called time series data, include annual (or monthly) number of road traffic accidents, traffic fatalities or vehicle kilometers driven in a country, as well as the corresponding values of safety performance indicators (e.g., data on speeding, seat belt use, alcohol use, etc.). Some commonly used statistical techniques imply assumptions that are often violated by the special properties of time series data, namely serial dependency among disturbances associated with the observations. The first objective of this paper is to demonstrate the impact of such violations to the applicability of standard methods of statistical inference, which leads to an under or overestimation of the standard error and consequently may produce erroneous inferences. Moreover, having established the adverse consequences of ignoring serial dependency issues, the paper aims to describe rigorous statistical techniques used to overcome them. In particular, appropriate time series analysis techniques of varying complexity are employed to describe the development over time, relating the accident-occurrences to explanatory factors such as exposure measures or safety performance indicators, and forecasting the development into the near future. Traditional regression models (whether they are linear, generalized linear or nonlinear) are shown not to naturally capture the inherent dependencies in time series data. Dedicated time series analysis techniques, such as the ARMA-type and DRAG approaches are discussed next, followed by structural time series models, which are a subclass of state space methods. The paper concludes with general recommendations and practice guidelines for the use of time series models in road safety research.

1. Introduction

Time series analysis is used in road transport and road safety research for describing, explaining and forecasting trends at an aggregate level. The technique is applied to road safety indicators aggregated over an area (a country, road type or accident type, or combination thereof) and regular time intervals. Road safety indicators such as the number of injury accidents and victims – among which the number of road fatalities in particular – are chosen for measuring road safety at for instance a national level.

Since the 1980s, a systemic methodological framework for modeling the road risk process has emerged: it consists of relating risk indicators to all of their determinants and to account for road safety measures simultaneously (Hakim et al., 1991). To this end, risk indicators and risk factors have been defined at different levels of the road risk process: in the DRAG approach (Gaudry, 1984; Lassarre, 1994; Gaudry and Lassarre, 2000) these are road demand, accident risk, and accident severity.[1] The three-level approach refers to the two dimensions of road risk (the risk that an accident occurs and the risk that a person is injured in an accident), on the one hand, and to the fact that exposure to risk is the essential, primary risk factor on the other hand.

A review of time series analysis of road safety trends as performed at the national level in Europe since the 1980s highlights a progress in the time series analysis techniques: from descriptive toward explanatory models (see Bergel-Hayat, 2008, 2012), and from deterministic toward stochastic models under the form of structural models, see Harvey (1989), Durbin and Koopman (2012), Commandeur and Koopman (2007), and Bijleveld et al. (2008).

* Corresponding author at: SWOV Institute for Road Safety Research, P.O. Box 1090, 2260 BB Leidschendam, The Netherlands. Tel.:+31 070 317 33 67; fax:+31 070 320 12 61.
E-mail address: jacques.commandeur@swov.nl (J.J.F. Commandeur).

[1] The name DRAG is formed by the acronym of the French words "Demande Routière, Accidents, et leur Gravité", which translates into "Demand for Road Use, Accidents and their Severity".

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.11.006

Research streams stemming from the former COST 329 project (COST329, 2004) and the International Cooperation on Time Series Analysis (ICTSA, 2000–2006) network converged in a coherent common approach for modeling and comparing the development in road safety trends among different countries. This common approach was formalized within Work Package 7 of the SafetyNet project dealing with "Data Analysis and Synthesis".

Not all types of models are appropriate for analyzing changes in sequential measurements of casualty data sets in the road safety field. The present paper contains a review of the different types of time series analysis techniques studied and applied within the SafetyNet project, with a focus on how to overcome the problem of dependencies – or serial correlations – between model residuals. The importance of this issue for proper statistical inferences is outlined in Section 2.

For all considered techniques, whether they handle time dependencies explicitly or not, a standard approach was followed in describing the successive steps (objective of the technique, model definition and assumption, data set and research problem, model fit, estimation, diagnostic and interpretation of application results) in conducting the modeling. Detailed information for understanding each of these steps can be found in the Methodology report (Dupont and Martensen, 2007) and in the Manual (Dupont and Martensen, 2007) produced by the members of Work Package 7 of the SafetyNet project. In this paper, the main features of each technique are presented in a structured way, illustrating their use, outcomes, and interest through some applications.

The paper is structured as follows. The impact of ignoring dependencies in time series data is first highlighted in Section 2. Classical regression techniques and their extensions are then shortly discussed in Section 3. Dedicated techniques that handle time dependencies explicitly are presented in some detail in Section 4. The paper concludes with a summary, along with recommendations for analyzing road safety developments at an aggregate level.

2. The impact of time dependencies

Many road traffic data consist of time series: sets of observations that are sequentially ordered over time. Examples are the annual or monthly number of road traffic accidents in a country, its annual or monthly number of road traffic fatalities, its annual or monthly number of vehicle kilometers driven, its annual or monthly values on safety performance indicators, etc., all sequentially ordered over time.

Whenever one is interested in studying and analyzing such sequentially ordered observations, special issues arise. In this section we illustrate with a simple example what these special issues are, and how they can be dealt with by using a special family of analysis techniques collectively known as *time series models*.

The example consists of the log of the total annual number of road traffic fatalities observed in Norway for the period 1970–2009, as displayed with circles in Fig. 1. Since the period spans 40 years, there are $n = 40$ observations. In order to try and capture the dynamics of this time series, we first naively perform a classical linear regression of these 40 sequentially ordered observations on time.

Typically, in simple classical linear regression a linear relationship is assumed between a criterion or dependent or endogenous variable y, and a predictor or independent or exogenous variable x such that

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{NID}(0, \sigma_\varepsilon^2) \tag{1}$$

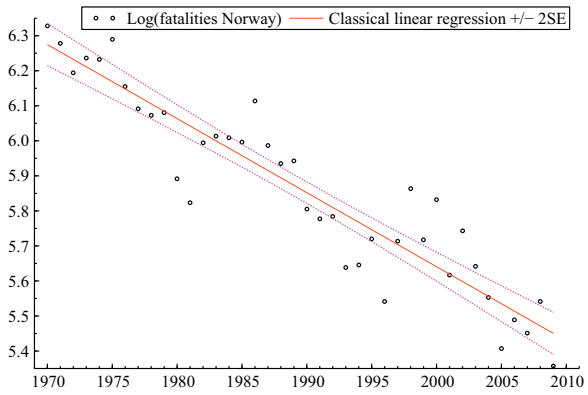

Fig. 1. Classical linear regression results for log of annual number of Norwegian fatalities, including 95% confidence limits.

where $i = 1, \ldots, n$ and n is the number of observations. The algebraic notation

$$\varepsilon_i \sim \mathrm{NID}(0, \sigma_\varepsilon^2) \tag{2}$$

in (1) expresses that the residuals ε_i are assumed to be normally and independently distributed with mean equal to zero and variance equal to σ_ε^2.

Now suppose that the dependent variable y in (1) is the just mentioned series of the log of Norwegian road traffic fatalities. Also, suppose that the independent variable x in (1) consists of the numbered consecutive time points in the series (thus, $x = 1, 2, \ldots, 40$). The usual scatter plot of these two variables – including the best fitting line according to classical linear regression model (1) and its 95% confidence limits – is shown in Fig. 1. The equation of the regression line in Fig. 1 is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = 6.2958 - 0.02114\, x_i,$$

with error variance $\widehat{\sigma}_\varepsilon^2 = 0.00936$.

The standard t-test for establishing whether the regression coefficient $\hat{\beta} = -0.02114$ deviates from zero yields

$$t = \frac{\hat{\beta}}{\sqrt{(\hat{\sigma}_\varepsilon^2)/(\sum_{i=1}^{n}(x_i - \bar{x})^2)}} = \frac{-0.02114}{\sqrt{0.00936/5330}} = \frac{-0.02114}{0.00132} = -15.95.$$

Since the value of this t-test is associated with a p-value of 2×10^{-18}, the linear relationship between the criterion variable y and the predictor variable x is extremely significant. When the assumptions for classical linear regression are valid we may conclude that time is a highly significant predictor of the log of the number of Norwegian road traffic fatalities, and that there is a negative relation between these two variables: as time proceeds the log of the number of fatalities decreases.

However, one issue has completely been overlooked in this analysis. The just mentioned t-test was based on the fundamental assumption that the 40 observations in the time series, after their correction for the intercept α and the exogenous variable x, are *independent* of each other, as implied by (2). That the observations are not independent becomes more obvious by connecting them with lines, as has been done in the top graph of Fig. 2. Inspection of the latter graph shows that the observation in a certain year has the tendency to be more similar to the observation of the previous year than to other earlier observations.

The dependencies between the observations are also reflected in the residuals of model (1) displayed at the bottom of Fig. 2. Positive values of the residuals tend to be followed by further positive