Predicting road crashes using calendar data

F. Van den Bossche, G. Wets and T. Brijs

PROMOTOR ▶ Prof. dr. Geert Wets
ONDERZOEKSLIJN ▶ Kennis verkeersonveiligheid
ONDERZOEKSGROEP ▶ LUC, PHL, VUB, Vito
RAPPORTNUMMER ▶ RA-2005-71

UNIVERSITAIRE CAMPUS
GEBOUW D
B 3590 DIEPENBEEK
T ▶ 011 26 87 05
F ▶ 011 26 87 00
E ▶ info@steunpuntverkeersveiligheid.be
I ▶ www.steunpuntverkeersveiligheid.be

Predicting road crashes using calendar data

RA-2005-71

F. Van den Bossche, G. Wets and T. Brijs

Onderzoekslijn Kennis verkeersonveiligheid

DIEPENBEEK, 2005.
STEUNPUNT VERKEERSVEILIGHEID.

Documentbeschrijving

Rapportnummer:	RA-2005-71
Titel:	Predicting road crashes using calendar data
Ondertitel:	
Auteur(s):	F. Van den Bossche, G. Wets and T. Brijs
Promotor:	Prof. dr. Geert Wets
Onderzoekslijn:	Kennis verkeersonveiligheid
Partner:	Universiteit Hasselt
Aantal pagina's:	25
Trefwoorden:	Road safety, regression, ARIMA, trading day, calendar data
Projectnummer Steunpunt:	1.3
Projectinhoud:	Analyse van de impact van mobiliteit op de verkeersveiligheid

Uitgave: Steunpunt Verkeersveiligheid, December 2005.

Steunpunt Verkeersveiligheid
Universitaire Campus
Gebouw D
B 3590 Diepenbeek

T 011 26 87 05
F 011 26 87 00
E info@steunpuntverkeersveiligheid.be
I www.steunpuntverkeersveiligheid.be

Samenvatting

In verkeersveiligheidsonderzoek worden macroscopische modellen gebruikt om kwantitatieve doelstellingen te evalueren. Deze doelstellingen zijn gebaseerd op verwachte waarden van slachtoffers en ongevallen die worden bepaald met behulp van verkeersveiligheidsmodellen. Bij het opzetten van deze modellen duiken meestal volgende problemen op: het ontbreken van relevante gegevens, een beperkte tijdshorizon en het ontbreken van toekomstige waarden voor de variabelen in het model.

Als oplossing voor deze problemen wordt in dit onderzoek het gebruik van kalendervariabelen voorgesteld. Hiertoe behoren een trend, een "trading day" patroon, dummy variabelen voor de maanden en een maat voor druk verkeer. In dit rapport testen we de relevantie van kalendervariabelen bij het voorspellen van de verkeersveiligheid. ARIMA modellen en regressiemodellen met ARMA foutentermen en kalendervariabelen worden opgezet. Met beide modellen worden een aantal voorspellingen gemaakt en de kwaliteit van de voorspellingen wordt vergeleken.

In deze studie gebruiken we Belgische gegevens (1990-2002) op maandbasis om modellen te ontwikkelen voor doden en zwaargewonden, lichtgewonden en de overeenkomstige aantallen ongevallen.

De resultaten tonen dat de regressiemodellen een betere fit hebben dan de pure ARIMA modellen. De trend en het trading day patroon hebben een significant effect op het aantal doden en zwaargewonden, en op de overeenkomstige aantallen ongevallen. De maat voor druk verkeer is significant in alle modellen. De voorspellingen van de regressiemodellen zijn beter dan die van de ARIMA modellen, in het bijzonder voor de (ongevallen met) lichtgewonden.

Summary

In road safety, macroscopic models are developed to support the quantitative targets in safety programmes. Targets are based on estimated numbers of fatalities and crashes that are derived from models. When constructing these models, typical problems are the lack of relevant data, the limited time horizon and the availability of future values for explanatory variables.

As a solution to these restrictions, we suggest the use of calendar data. These include a trend, a trading day pattern, dummy variables for the months and a heavy traffic measure. In this paper, we test the relevance of calendar data for the prediction of road safety. ARIMA models and regression models with ARMA errors and calendar variables are built. Predictions are made by both models and the quality of the predictions is compared.

We use Belgian monthly crash data (1990-2002) to develop models for the number of persons killed or seriously injured, the number of persons lightly injured and the corresponding number of crashes.

The regression models fit better than the pure ARIMA models. The trend and trading day variables are significant for the outcomes related to killed or seriously injured persons, while the heavy traffic measure is significant in all models. The predictions made by the regression models are better than those from the ARIMA models, especially for the lightly injured outcomes.

Table of contents

1. INTRODUCTION

For many years, traffic growth and the increasing importance of efficient road transportation led to a large number of road crashes. Crashes are the result of various influences at a certain location and time. In an OECD report (1), some broad categories of factors influencing road crashes are listed. The number of crashes depends on autonomous factors that cannot be influenced on a short-term (e.g. weather and technology), economic conditions (e.g. unemployment and income), the size and the structure of the transportation sector (exposure, infrastructure, vehicle park,…), the accident countermeasures, the data collection system and the random variation in crash counts. Although it is intuitively appealing to assume that these factors have an influence on the number of road crashes, it would be instructive to get a confirmation of this influence. Given the large number of possible factors, however, this is not an easy task. The main condition to develop these explanatory models is the availability of data. In Belgium, data are rarely available in a format that can be used for this purpose. Especially exposure measures are sometimes very hard to find, and the data quality is often very low. Moreover, if traffic safety is to be predicted with an explanatory model, future values of the explanatory factors are necessary. This implies another set of predictions that must be made beforehand. Therefore, descriptive models are often used to investigate the evolution in road safety and to make predictions. These models describe a time series in terms of the general trend and a possible seasonal pattern, without providing any explanatory power. However, some simple variables can be found that provide insight in the series and that are always available, for the past, the present and the future. These are variables that are related to the seasons and to the calendar. It is not unrealistic to assume that these variables can help in understanding road safety time series.

The objective of this study is threefold. First, we develop an ARIMA (Auto-Regressive Integrated Moving Average) model for four road safety outcomes, namely the number of persons killed or seriously injured (NPERKSI), the number of persons lightly injured (NPERLI) and the corresponding counts of crashes (NACCKSI and NACCLI). Note that accidents with persons killed or seriously injured and persons lightly injured are only counted in NACCKSI, and not in NACCLI. We use monthly Belgian data for a period from January 1990 to December 2000. Second, we develop multiple regression models with ARMA errors to test whether calendar variables can provide an added value to the pure ARIMA models for road crashes and victims. We include seasonal dummy variables, a trend indicator, a trading day pattern and a measure of heavy traffic. Third, we will verify whether the calendar variables included in the model improve the statistical fit and the forecasting accuracy of the regression models, compared to the classical ARIMA models. We make forecasts with both models for the years 2001 and 2002, and compare their performance. If predictions are better in a model with calendar variables, then policy makers have an easy-to-use means to improve their prediction models for traffic safety.

This text is organized as follows. First, some background information is given on the application of the kind of models used in the paper. Next, the main ideas of regression models with ARIMA errors are discussed. Then, an overview of the data is given. In the results section, the model outcomes and the forecasts are presented and discussed. Also some general conclusions and topics for model improvement are provided.

2. BACKGROUND

In economic time series, calendar adjustment methods are frequently used. These methods take into account the seasonality of a series and the presence of specific influences that are caused by the structure of the calendar. They mostly consist of a trend, seasonal variables and a trading pattern. Trading day effects reflect variations in monthly time series due to the changing composition of months with respect to the number of times each day of the week occurs (2). In each month, there are four weeks plus usually one, two or three more days. Each weekday occurs at least 4 times in a month, but some days will occur 5 times. The composition of the calendar will affect the data for the month. If, for example, a shop is only open on weekdays, then sales will be higher if some weekdays occur five times in a month. Especially monthly time series that are totals of daily activities (like the records of road crashes), are often influenced by the weekday composition of the month. Details on the construction of trading day models are given in Findley et al. (3), based on Young (4), Cleveland et al. (5) and Bell (6). The technique was subsequently used in the Census X11-ARIMA (7) and X12-ARIMA (3) seasonal adjustment methods for time series. The combination of trading day regression analysis and ARIMA time series modeling is also presented in Bell et al. (8). Trading day patterns can be included in different forms, as is shown in Soukup et al. (9). In our study, we prefer a parsimonious form of the trading day pattern, as provided by Gómez et al. (10). This form captures the trading day effect in one variable. A whole series of applications of adjustment methods can be found, mainly in the field of economics. For example, Rooijakkers et al. (11) use trading day variables to adjust monthly data on Dutch consumer spending, based on an X12-ARIMA time series decomposition. In Cano et al. (12), regression analysis with ARMA errors is used to adjust employment time series for calendar effects. The use of dummy variables to represent a deterministic seasonal pattern is widely explained in many textbooks (see for example Neter et al. (13), Makridakis et al. (14) and Pankratz (15)).

In traffic safety research, the use of calendar data to improve road safety forecasts seems less widespread. To our knowledge, the impact of trading day patterns on the evolution in traffic safety has never been investigated. However, in many studies, time series of road crashes are analyzed, using a variety of models (16, 17). More recently, Raeside et al. (18) used ARIMA models to analyze monthly time series on pedestrian casualties and fatalities in Great-Britain. Applications of regression models with ARMA errors are found in Van den Bossche et al. (19, 20). One class of explanatory time series models is known as the DRAG family (21). These are structural models, including a relatively large number of explanatory variables, whose effects on exposure, the frequency and the severity of road crashes are estimated by econometric methods. In the DRAG-2 model for Quebec (22) and the SNUS-2.5 model for Germany (23), some calendar variables like the number of working days, Saturdays, Sundays and holidays in a month are included. Apart from these examples, calendar data are rarely used to enrich the models.

However, there is a lot to be said for studying these variables. First, there is a seasonal pattern present in accident data. Some months always have a higher number of road crashes and victims than others. This is also related to the length of the month. It is to be expected that a 28-day month (February) will have a lower number of road crashes than a 31-day month, given the almost 10% difference in the length. Indicator variables for the months can capture these patterns. Also, it is known that the exposure to crashes is higher is some months than in others, like for example during holidays. These peak moments in traffic can explain the number of crashes and victims during a given month. By including variables that reflect peak exposure we can partly account for these effects. Second, given the problem of weekend crashes in Belgium, it is to be expected that crash counts are higher in months with more weekend days. Third, the planning of official holiday periods can influence the exposure in a month, and thus the number of crashes. Easter holiday can shift between March and April, and starting weekends of holiday periods always lead to higher amounts of traffic. Based on the planning of official

holiday periods, it is possible to foresee the weekends of high travel in a year. Fourth, the calendar variables (like the number of weekdays and weekend days in a month) are known for every year in the future. These properties of calendar variables make them quite appealing to practitioners, because they enrich the model and allow predictions without a heavy effort of data collection and cleaning.

3. METHODOLOGY

In this study, four traffic safety outcomes are expressed in terms of independent variables that are related to the calendar. Multiple linear regression can be used to model a relationship between a dependent variable and one or more independent variables. If the observations are measured over time, the model is called a time series regression. The resulting statistical relationship can be used to predict future values of the target. To reduce the presence of autocorrelation, the regression model will be extended with an ARIMA structure in the error term. The construction of this kind of models is discussed here. For an overview of regression models, the reader is referred to Neter et al. (13). In Makridakis et al. (14), an introduction to time series analysis is given. Regression models with ARMA errors are described in Pankratz (15).

3.1 Multiple Regression

The multiple regression model can usually be written as the equation $Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + N_t$, where Y_t is the t-th observation of the dependent variable, and $X_{1,t}, \ldots, X_{k,t}$ are the corresponding observations of the explanatory variables. The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are fixed but unknown, and N_t is the unknown error term, which is assumed to be normally distributed. Using classical estimation techniques, estimates for the unknown parameters are obtained. If the estimated values for $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are given by $b_0, b_1, b_2, \ldots, b_k$, then the dependent variable is estimated as $Y_{est,t} = b_0 + b_1 X_{1,t} + b_2 X_{2,t} + \ldots + b_k X_{k,t}$, and the estimate $N_{est,t}$ for the error term N_t is calculated as the difference between the observed and predicted value of the dependent variable: $N_{est,t} = Y_t - Y_{est,t}$.

In the theoretical model, several assumptions are made about the explanatory variables and the error term. These include absence of high multicollinearity, heteroscedasticity and autocorrelation. Heteroscedasticity is achieved by modeling log-transformed dependent variables. The autocorrelation assumption is likely to be violated in regression models with time series data. In a regression with autocorrelated errors, the errors will probably contain information that is not captured by the explanatory variables, and it is necessary to extract this information to finally end up with uncorrelated ("white noise") residuals. Typically, the Autocorrelation Function (*ACF*) and the Partial Autocorrelation Function (*PACF*) are used to detect autocorrelation among residuals (14). Autocorrelation can be taken into account by adding more complex autoregressive (AR) or moving average (MA) structures to the regression equation, as will be explained further in this text. When the assumptions are satisfied, the estimators are unbiased and have minimum variance among all linear unbiased estimators.

3.2 ARMA Modeling

The ARIMA modeling approach expresses a variable as a weighted average of its own past values. The model is in most cases a combination of an autoregressive (AR) part and a moving average (MA) part. Suppose a variable N_t is modeled as an autoregressive process, AR(p). Then, N_t can be expressed as a regression in terms of its own passed values: $N_t = C + \phi_1 N_{t-1} + \phi_2 N_{t-2} + \ldots + \phi_p N_{t-p} + a_t$, where C is a constant term, ϕ_i ($i = 1, \ldots, p$) are the weights for the autoregressive terms and a_t is a random term, which is assumed to be normally distributed "white noise", containing no further information. Using a backshift operator B^i on N_t, defined as $B^i N_t = N_{t-i}$ ($i = 1,2,\ldots$), this process can be written as $N_t = C + \phi_1 B N_t + \phi_2 B^2 N_t + \ldots + \phi_p B^p N_t + a_t$, or $(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) N_t = C + a_t$.

The series N_t can also be expressed in terms of the random errors of its past values, which is then a moving average MA(q) model: $N_t = C - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} + a_t$, where θ_j ($j=1,\ldots,q$) are the weights for the moving average terms. Using the backshift operator, this equals $N_t = C - \theta_1 B a_t - \theta_2 B^2 a_t - \ldots - \theta_q B^q a_t + a_t$, or $N_t = C + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) a_t$. In a more general setting, it is possible to include AR and MA terms in one equation, leading to an

ARMA(p, q) model: $(1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)N_t=C+(1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q)a_t$, where a_t is again assumed to be "white noise".

An ARMA model cannot, however, be applied in all circumstances. It is required that the series be stationary. For practical purposes, it is sufficient to have *weak* stationarity, which means that the data is in equilibrium around the mean and that the variance around the mean remains constant over time (14). If a series is non-stationary because the variance is not constant, it often helps to log-transform the data, as is done in this text. To have a series that is stationary in the mean, differencing is used. Instead of working with the original series, successive changes in the series are modeled. When an ARMA model is built on differenced data, it is called an ARIMA model, where "I" indicates the differencing.

3.3 Regression with ARMA errors

The ARMA modeling approach can now be applied to the multiple regression equation to model the information that remains in the error terms. Assume a regression model with one explanatory variable, denoted as $Y_t=\beta_0+\beta_1 X_{1,t}+N_t$. Suppose further that the error terms are autocorrelated, and that they can be appropriately described by an ARMA(1,1) process. This model can then be written as: $Y_t=\beta_0+\beta_1 X_{1,t}+N_t$, with $(1-\phi_1 B)N_t=(1-\theta_1 B)a_t$, and a_t is assumed to be white noise. Substituting the correction for the error term into the regression equation gives:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \frac{(1-\theta_1 B)}{(1-\phi_1 B)} a_t .$$

Because of the specific form in the error terms, the classical least squares methods are not appropriate to estimate the parameters of this equation. Instead, Maximum Likelihood estimation is done using Marquardt's method via nonlinear least squares estimation (24).

If differencing is applied to the errors in a multiple regression, all corresponding series (both of the dependent and the explanatory variables) should be differenced (15). This can be seen from our small regression example. Differencing the error terms results in the following expression, with the ARMA(1,1) model now in the differenced error terms:

$$\nabla N_t = \frac{(1-\theta_1 B)}{(1-\phi_1 B)} a_t \Leftrightarrow N_t = \frac{(1-\theta_1 B)}{\nabla(1-\phi_1 B)} a_t .$$

Substituting back this expression into the regression equation gives:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \frac{(1-\theta_1 B)}{\nabla(1-\phi_1 B)} a_t \Leftrightarrow \nabla Y_t = \beta_0^{'} + \beta_1 \nabla X_{1,t} + \frac{(1-\theta_1 B)}{(1-\phi_1 B)} a_t .$$

The intercept is now possibly different, but the (theoretical) regression coefficient β1 is not affected by the differencing operation. Its estimated value may differ slightly, since the estimation is done on different (although related) time series.

3.4 Forecasting

Regression models can easily be used for forecasting purposes. After the model has been developed, estimated values for the dependent variable can be obtained. In order to produce forecasts with a regression model with ARMA errors, the two parts of the equation need to be predicted. First, for the regression part, future values of the explanatory variables should be available. Since we use calendar data in this study, the availability of future data is guaranteed. Second, in the ARMA error part, the errors should be replaced by their estimated values. To depict uncertainty in the predicted values, 95% confidence intervals are provided.

4. DATA

4.1 Dependent variables

For this study, official monthly data on Belgian road crashes are available from January 1990 up to December 2002. The model is developed on data from 1990 to 2000, while the last two years are used for forecasting purposes. Four dependent variables will be modeled: the number of crashes with lightly injured persons (NACCLI), the number of crashes with persons killed or seriously injured (NACCKSI) and the corresponding number of victims (NPERLI and NPERKSI). Note that, in the models, log-transformations are used in order to achieve homoscedastic error variances (the variables are then named LNACCLI, LNACCKSI, LNPERLI and LNPERKSI, respectively). The evolution in time of the dependent variables, before log-transformation, is shown in FIGURE 1.

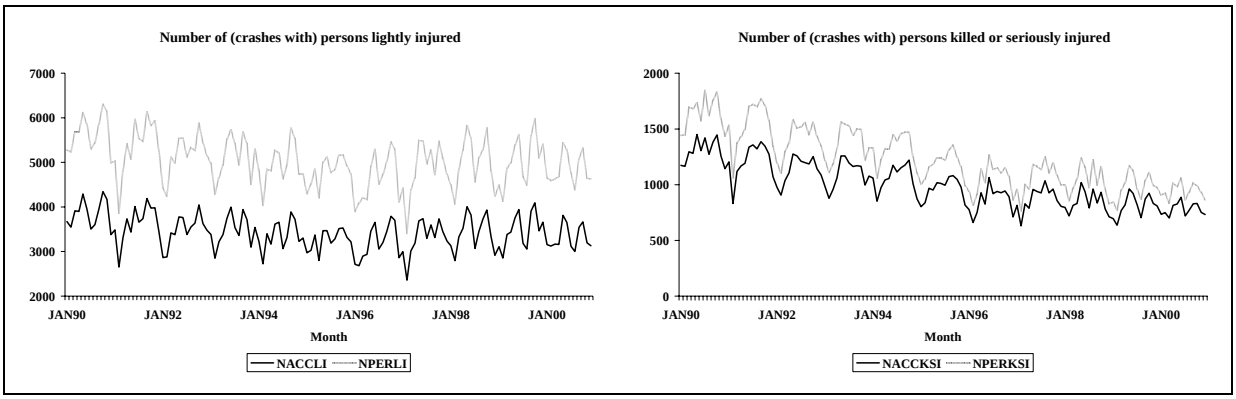

FIGURE 1: Dependent Variables NACCLI, NPERLI, NACCKSI and NPERKSI.

4.2 Independent variables

All independent variables in this study are based on the calendar. This offers the undeniable advantage of availability of the data. First, a time variable T_t is included to reflect a linear trend, where t is an index for the time period. This variable equals the time point of observation.

Second, seasonal dummy variables are used to handle seasonality. Since we have monthly data, the seasonal pattern is represented by the variables JAN_t, FEB_t, …, NOV_t, where each of these equals 1 in the given month, and zero otherwise. To avoid multicollinearity problems, no variable is included for the month December. The coefficients for the other months reflect the average difference in the dependent variable between the given month and the omitted month. Note that by the use of seasonal dummy variables, it is implicitly assumed that the seasonal component is unchanging from year to year.

A third variable is the trading day variable TD_t. The number of road crashes and the corresponding number of victims may vary according to the day of the week. To correctly forecast the number of road crashes and victims, we take into account the number of Mondays, Tuesdays, etc. in each month. Since we are primarily interested in the difference between weekdays and weekends, we propose a trading day variable TD_t that is defined as follows (9):

$$TD_t = \sum_{j=Mon}^{Fri} D_{jt} - \frac{5}{2} \sum_{j=Sat}^{Sun} D_{jt} ,$$

where D_{jt} indicates the number of times the j-th day occurs in month t. This formula forces the weights of the different days of the week to sum up to one. It also requires

that the weights for all weekdays and all weekend days are the same. If we have for example that TD_t=-0.005, then each weekday is given a negative weight of -0.005, while each weekend day is weighted as (-2.5)*(-0.005)=0.0125, which indicates that months with more weekend days may be more dangerous than months with more weekdays.

A last variable, H_t, is the a measure for heavy traffic. Some periods of the year are characterized by more traffic than other periods. The traditional holiday periods often cause days of holiday rush, especially on the highways to the tourist locations in Belgium and the neighboring countries. Also holidays related to Christmas and Easter and the public holidays cause a very specific traffic pattern. It is to be expected that periods with heavy traffic have a different road safety profile than other periods. To account for these differences, we include a variable that is based on the "density indicator" developed by the VAB (Flemish Automobile Association). This measure classifies each day of the month as "normal traffic" or "heavy traffic", and then sums up the days with heavy traffic. These days are determined based on experience of the VAB road experts and the spread of public holidays, religious feasts and school holidays over the months.

5. RESULTS

In

TABLE 1, the results of the different models are presented. All models are estimated using the SAS statistical software package (25). For both the ARIMA and the regression models, the parameter estimations are shown. The significance of the coefficients is given between brackets. If a coefficient is not significant at the 95% level, the corresponding variable is dropped from the model, except for the seasonal dummies, that are kept together. If differencing is done, the order is indicated. For all models the Box-Ljung chi-square test statistic is used to assess the level of autocorrelation in the error terms at various lags. Normality of the error terms is tested by the Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests, all available in the SAS software. If necessary, outliers were filtered by including dummy correction variables for certain months. For all models, the normality assumption could not be rejected by any of the tests. First we discuss the model fit for the ARIMA and regression models. Then the interpretation of the parameter estimates is given. Finally, the forecasting accuracy of the various models is compared.

5.1 Evaluation of model fit

Both the ARIMA models and the regression models are acceptable from a statistical point of view. That is, all necessary conditions are fulfilled to end up with valid models. Therefore, we need additional criteria to choose the best from these models. To evaluate the fit of the models for each dependent variable, we compute two information criteria, Akaike's Information Criterion or AIC (26) and Bayesian Information Criterion or BIC (27). They can be used to compare models that fit the same series (25). The AIC is computed as $-2\log(L)+2k$, where L is the likelihood function and k is the number of free parameters. The BIC is computed as $-2\log(L)+\ln(n)k$, where n is the number of residuals that can be computed from the model. Both criteria are likelihood-based and represent a trade-off between model fit and parsimony. The BIC tends to favor more parsimonious models compared to the AIC (28). For both criteria, the model is said to fit the data better when the AIC and BIC are lower.

From the AIC and BIC values in

TABLE 1, we see that the regression models fit better for all dependent variables. It seems that including the months as deterministic variables, together with trading day patterns and a heavy traffic measure, captures more variability in road safety outcomes than do ARIMA models. This is interesting since the ARIMA models allow to filter away seasonal patterns, but they cannot distinguish between months with more or less weekend days, nor can they account for periods with heavy traffic. It seems that the seasonal pattern is regular enough to be represented by seasonal dummies, and that other fluctuations can be captured by the extra calendar variables.

TABLE 1: Estimation Results for ARIMA and Regression Models

	LNPERLI	LNPERKSI	LNACCLI	LNACCKSI
ARIMA Models				
Differencing	1	12	1	12
Constant	-	-0.0559 *(0.000)*	-	-0.0521 *(0.000)*
AR(12)	0.6198 *(0.000)*	-	0.7157 *(0.000)*	-
MA(1)	0.5525 *(0.000)*	-0.1766 *(0.044)*	0.8255 *(0.000)*	-
MA(2)	0.4562 *(0.000)*	-0.3546 *(0.000)*	-	-0.3190 *(0.000)*
MA(4)	0.3589 *(0.000)*	-	0.2712 *(0.002)*	-
MA(12)	-	0.8007 *(0.000)*	-	0.9136 *(0.000)*
FEB91	-0.1799 *(0.000)*	-	-0.1631 *(0.002)*	
FEB97	-0.2362 *(0.000)*	-	-	
DEC98	-	-	-0.1628 *(0.002)*	
AIC	-319.70	-265.44	-309.23	-276.81
BIC	-302.4	-254.29	-294.85	-268.44
MAPE	7.73%	6.33%	7.99%	5.98%
Theil's U	91.62%	82.76%	79.42%	75.99%
Regression Models				
Differencing	1	0	1	0
Constant	-0.0015 *(0.141)*	7.3985 *(0.000)*	-0.0013 *(0.081)*	7.1701 *(0.000)*
JAN	-0.0940 *(0.001)*	-0.0674 *(0.050)*	-0.0860 *(0.000)*	-0.0707 *(0.029)*
FEB	-0.1838 *(0.000)*	-0.1803 *(0.000)*	-0.1538 *(0.000)*	-0.1851 *(0.000)*
MAR	-0.0622 *(0.045)*	-0.0572 *(0.130)*	-0.0481 *(0.072)*	-0.0562 *(0.115)*
APR	0.0108 *(0.674)*	0.0916 *(0.004)*	0.0253 *(0.299)*	0.0743 *(0.013)*
MAY	0.1087 *(0.000)*	0.1582 *(0.000)*	0.1244 *(0.000)*	0.1593 *(0.000)*
JUN	0.0924 *(0.001)*	0.1410 *(0.000)*	0.1040 *(0.000)*	0.1454 *(0.000)*
JUL	0.1032 *(0.030)*	0.2455 *(0.000)*	0.0847 *(0.034)*	0.2313 *(0.000)*
AUG	0.1116 *(0.014)*	0.2648 *(0.000)*	0.1102 *(0.005)*	0.2437 *(0.000)*
SEP	0.0720 *(0.014)*	0.1235 *(0.001)*	0.0872 *(0.001)*	0.1191 *(0.000)*
OCT	0.0925 *(0.010)*	0.1243 *(0.002)*	0.1001 *(0.002)*	0.1206 *(0.002)*
NOV	0.0295 *(0.286)*	0.0511 *(0.136)*	0.0312 *(0.173)*	0.0485 *(0.133)*
T_t	n.s.	-0.0047 *(0.000)*	n.s.	-0.0044 *(0.000)*
TD_t	n.s.	-0.0057 *(0.008)*	n.s.	-0.0047 *(0.023)*
H_t	-0.0101 *(0.024)*	-0.0124 *(0.017)*	-0.0103 *(0.002)*	-0.0121 *(0.015)*
FEB91			-0.1524 *(0.002)*	
FEB97			-0.1844 *(0.000)*	
APR95			-0.1336 *(0.006)*	
AR(1)			-0.6070 *(0.000)*	
MA(1)	0.8203 *(0.000)*			
MA(2)		-0.3678 *(0.000)*	0.7528 *(0.000)*	-0.3281 *(0.000)*
AIC	-337.94	-318.36	-373.31	-332.72
BIC	-297.69	-272.24	-321.56	-286.60
MAPE	5.72%	6.46%	5.86%	6.11%
Theil's U	69.50%	78.09%	61.90%	74.60%

5.2 Interpretation of parameter estimates

5.2.1 Estimates of the ARIMA structures

In the ARIMA model, each outcome is a weighted average of past observations, expressed in terms of autoregressive and/or moving average components. The weights of observations further in the past on the current observation decline, which is in line with common sense. The past observations that produce a weight for the current one is determined by the order of the ARMA terms in the model.

It is interesting to see how the ARMA part of the model changes when explanatory variables are added. The ARMA structure tends to be much more simple, since only terms of order 1 or 2 are left. This indicates that the explanatory variables capture some

information that is present in the ARIMA structure and help in explaining the variation in the dependent variable. Also the differencing that is necessary to get a stationary process changes when calendar regression variables are added (especially the months). In the ARIMA models for LNPERKSI and LNACCKSI, a 12-period difference was included, while in the corresponding regression model no differencing was done. For LNPERLI and LNACCLI, a first difference is still needed in the regression, because the seasonal dummies cannot capture the period-to-period changes.

When differencing is done, the interpretation of the mean term is especially important. A mean term is estimated for the LNPERKSI and LNACCKSI models, where a 12-period difference was necessary in order to obtain a stationary series. An intercept in the seasonally differenced data corresponds to a linear deterministic trend in the original series (15). This is an acceptable assumption, as can be seen from the graphs in FIGURE 1. For LNPERLI and LNACCLI, a 1-period difference was taken. Since no real trend can be observed for these series, the mean term turned out to be insignificant and was dropped from the differenced series. The graphs show that a linear trend would, indeed, be inappropriate here. Although the series LNPERLI and LNACCLI have a clear seasonal pattern, seasonal differences are not taken. The reason is to be found in the stability of the estimates. When we take seasonal differences, the estimated moving average parameter of order 12 approaches unity, leading to an uninvertible solution. It is as if this estimate wants to cancel the differencing.

5.2.2 Estimates of the regression model

In the regression part of the model, the explanatory variables can be tested for their influence on the road safety outcomes. First, most of the monthly dummy variables are highly significant. Each of the road safety outcomes shows a strong seasonal pattern that can be represented fairly well by globally constant (i.e. regardless of the year) seasonal factors. Recall that the December coefficient is not estimated. The effect of the last month is given by the constant. Although this constant is not significant for LNPERLI and LNACCLI, it is nevertheless included in the models to serve as the base for the seasonal dummies. For all dependent variables, the first months of the year have a significantly lower value than the December value, while November is not significantly different.

Second, the time trend variable T_t is only significant for the number of crashes with persons killed or seriously injured, and the corresponding number of victims. This is in line with the results in the ARIMA models, where the constant was only significant for the differenced data in these models, indicating a deterministic linear trend in the original data. Also the graphs in FIGURE 1 support this conclusion, as the lightly injured outcomes do not show a trend.

A similar conclusion can be drawn for the trading day variables. These also are only significant for LNPERKSI and LNACCKSI. The composition of a month in terms of weekdays and weekend days therefore influences the number of persons killed or seriously injured, which confirms our expectations for Belgium. Weekend crashes are frequently observed, mostly with fatal consequences. The trading day variable can be used to quantify the number of fatalities expected from an extra weekend day in the month. As an example, compare the months of August in the years 1997 and 2000. In the first year, we count 21 weekdays and 10 weekend days. For this month, the variable TD_{AUG97} equals -4. The same month in 2000 has 23 weekdays and only 8 weekend days. Therefore, TD_{AUG00} equals 3. Given a parameter estimate of -0.0057 for the model LNPERKSI, we have an effect on the logarithm of 0.0228 for August 1997, and -0.0171 for August 2000. Applying the exponential function results in an increase of persons killed and seriously injured of exp(0.0228)-1=2.3% for August 1997, and a decrease of 1-exp(-0.0171)=1.7% for August 2000. Note that this is only the effect attributable to the trading day pattern. Comparing two months with 9 and 10 weekend days respectively, results in a global increase in victims of 2%, all other things being equal. Given the high number of victims in Belgium (a monthly average of 1169 killed or

seriously injured over the analysis period), this percentage is not negligible. This is an interesting instrument for policy makers. The models allow measuring the number of victims that can be expected based on calendar structure. If we start a month with more weekend days, safety campaigns can be directed towards the group of people that is likely to be on the road on Saturdays or Sundays.

The last variable in the model is the measure of "heavy traffic". This variable is significant for all road safety outcomes, with a negative sign. For example, if a month counts one extra day with heavy traffic, the number of fatalities decreases by 1-exp(-0.0124)=1.23%, all other things being equal. This may look counterintuitive at first sight, but can easily be explained. In periods of heavy traffic, caused by public or national holidays and Christian holy days, large concentrations of traffic can be found in Belgium, mainly on the main roads. Although many people travel, the road seems to be quite safe because of congestion or slow traffic. Our variable therefore is not a real measure of traffic exposure, but an indication of the level of traffic concentration. Clearly the months with less concentration are also less safe.

5.3 Evaluation of forecast accuracy

To test the forecasting accuracy of our models, we kept the years 2001 and 2002 out of the training set, and these years are only used for forecasting. The forecasts of the models for the years 2001 and 2002 are shown in FIGURE 2. Although it is difficult to assess forecasting accuracy on the basis of a graphical representation, it can be seen that the confidence intervals for the predictions from the regression models are smaller than those from the ARIMA models. To quantify the forecast accuracy, we computed two measures that can be used to compare the forecasts of different models on the same data set. The first one is called the Mean Absolute Percentage Error or the MAPE (14). It is calculated as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|PE_t\right| = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - F_t}{Y_t}\right|,$$

where n is the number of observations for which the forecasts are made and PE_t is the Percentage Error for the t-th observation, that is the relative difference between the observed value Y_t and the forecasted value F_t.

The MAPE is an average of the absolute deviances between the observed and predicted values, and is expressed as a percentage. The lower the MAPE, the better the predictions of the model are. A second measure is called Theil's U (29). It is based on the comparison between the forecasts of the model and the forecasts of a very simple, "naïve" forecast method, that takes the value of the current period as a forecast for next. It is derived as follows:

$$U = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{F_{t+1} - Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1} - Y_t}{Y_t}\right)^2}},$$

with F_t, Y_t and n as defined above. Theil's U compares the MAPE of a given forecasting method with the MAPE of the naïve forecast. It gives more weight to larger errors, and provides a relative basis for comparison with naïve methods. Also, the interpretation of the statistic is quite straightforward. If U=1, then the predictions made by the model are as good as those made by a naïve forecast. The model performs better than the naïve forecast if U<1. When U>1, there is no point in using the model, since the naïve model can even make better forecasts.

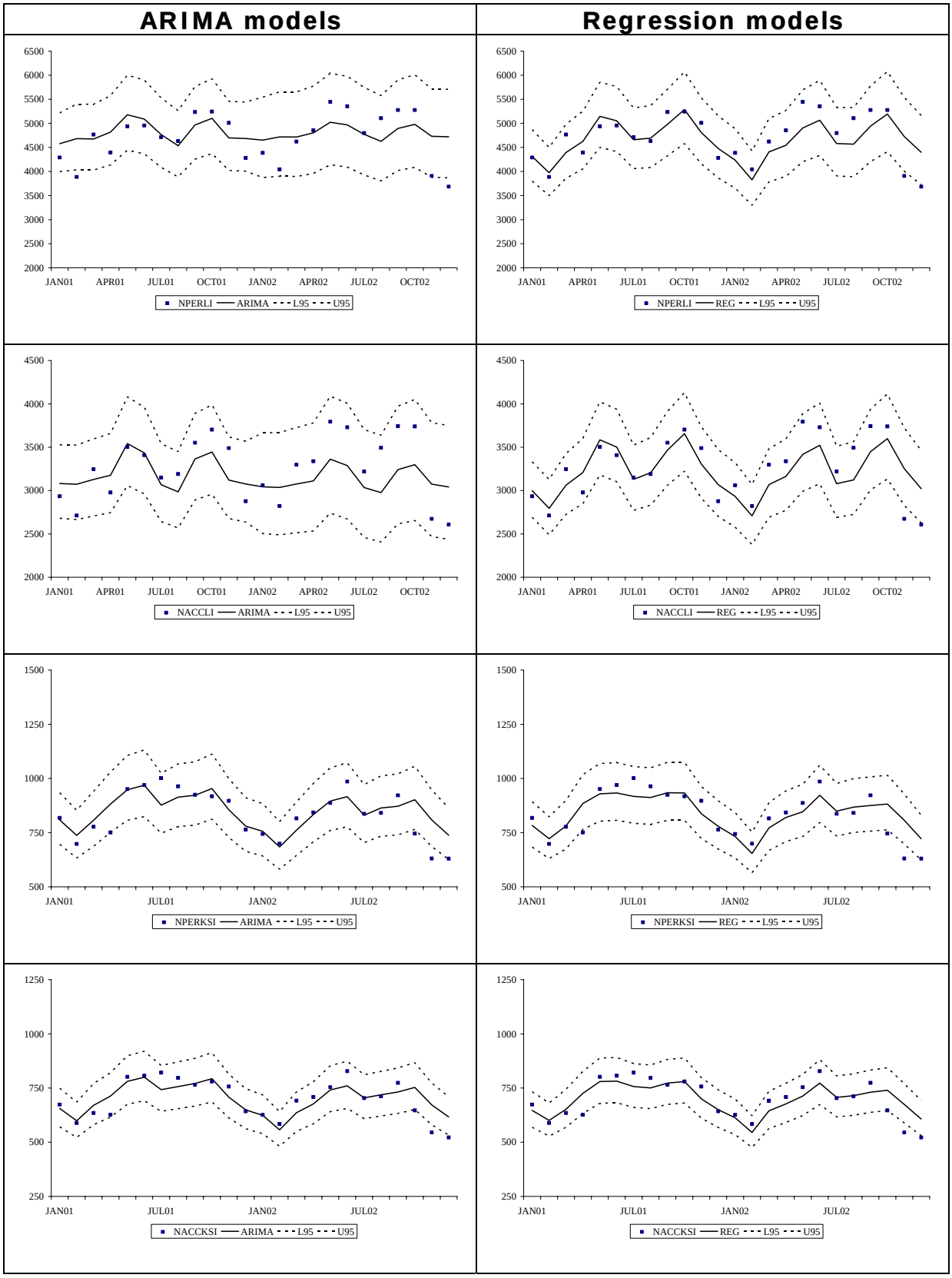

FIGURE 2: Comparison of ARIMA and Regression Forecasts.

TABLE 1, the test statistics are shown for the different models. Based on the MAPE, the regression models perform better than the corresponding ARIMA models for LNPERLI and LNACCLI. For LNPERKSI and LNACCKSI, both models perform almost equally. Theil's U shows a small improvement of the regression models over the ARIMA models. For the other models, adding explanatory calendar variables substantially improves the predictions. This is an attractive conclusion for road safety workers who have to predict monthly safety outcomes without having any data available. Since calendar data are known for the future, no extra prediction efforts are needed and future safety outcomes can be assessed. Based on the accuracy measures, prediction quality can be monitored and models can be regularly updated.

6. CONCLUSIONS AND FURTHER RESEARCH

In this study, ARIMA models and regression models with calendar variables are built for four road safety outcomes. The model fit and the accuracy of the produced forecasts are compared by means of some objective criteria. All models produced were acceptable from a statistical point of view. The comparison showed that the regression models performed better, both in terms of model fit and forecasting accuracy. The improvement was especially clear for the models LNPERLI and LNACCLI, where the ARIMA models have problems with the highly regular seasonal pattern. For the outcomes related to persons killed and seriously injured, the improvement was less pronounced, although still present. The ARIMA models were also less easy to handle in the identification phase, since they have a more complex structure. Furthermore, the regression models offers the opportunity to measure the "impact" of some calendar variables on road safety outcomes. For LNPERKSI and LNACCKSI, a clear trend could be modeled and a trading day pattern was present. The "heavy traffic" measure was significant for all dependent variables. Periods of high traffic concentration generally have lower crash counts.

Although the calendar variables indeed have the power to improve the models, they should be treated cautiously. First, including a linear trend is not always a natural assumption. For short-term predictions, the fit improves by including the trend, but long-term predictions may suffer from systematic deviances. Then a non-linear trend may be more appropriate. Second, by including deterministic dummy variables for the months, it is assumed that the effect of each month is always equally large, irrespective of the year of analysis. If there is evidence for a changing seasonal pattern, this assumption should be relaxed. The same goes for the trading day variables. The effect of the calendar composition may change over time, as is shown for example in Bell et al. (2). Third, the measure of heavy traffic is based on the scheduling of holidays and special days during the years. Before projecting this measure into the future, one has to verify whether the holiday periods will remain the same. For Belgium, we can be quite sure that this is the case, since holiday periods have been officially determined at the beginning of the nineties.

In short, the regression model with calendar variables shows potential to improve road safety forecasts, without running into data availability problems. Also the model fit improves and the calendar variables are intuitively appealing. If the models are updated regularly, road safety workers are offered a tool to estimate calendar influences and to make short-term predictions that are superior to the ARIMA results, and that can be developed without problems of data availability.

7. REFERENCES

(1) OECD Road Transport Research*. Road Safety Principles and Models: Review of Descriptive, Predictive, Risk and Accident Consequence Models*. OCDE/GD(97)153, OCDE, Paris, 1997.

(2) Bell, W. and Martin, D. Modeling time-varying trading-day effects in monthly time series. In: *Proceedings of the Joint Statistical Meetings*, August 8-12, Toronto, Canada, 2004.

(3) Findley, D., Monsell, B., Bell, W., Otto, M. and Chen, B. New Capabilities and Methods of the X-12-ARIMA Seasonal Adjustment Program (with discussion). *Journal of Business and Economic Statistics*, Vol. 16, No. 2, pp. 127-177, 1998.

(4) Young, A.H., *Estimating Trading Day Variation in Monthly Economic Time Series*, Technical Paper 12, U. S. Department of Commerce, Bureau of the Census, Washington, D.C., 1965.

(5) Cleveland, W. and Devlin, S., Calendar Effects in Monthly Time Series: Detection by Spectrum Analysis and Graphical Methods. *Journal of the American Statistical Association*, 75, pp. 487-496, 1980.

(6) Bell, W.R. (1984), *Seasonal Decomposition of Deterministic Effects*, Research Report 84/01, Bureau of the Census, Washington, DC, Statistical Research Division.

(7) Dagum, E. *The X11-ARIMA Seasonal Adjustment Method*. Statistics Canada, Ottawa, Catalogue No. 12-564E, 1980.

(8) Bell, William R. and Hillmer, Steven C., Modeling time series with calendar variation, *Journal of the American Statistical Association*, 78, 526-534, 1983.

(9) Soukup, R. and Findley, D. Detection and Modeling of Trading Day Effects, *ICES proceedings*, 2000.

(10) Gómez, V. and Maravall, A. *Programs TRAMO and SEATS, Instructions for the User (Beta Version: September 1996)*, Banco de España – Servicio de Estudios, Documento de Trabajo no. 9628 (English version), 1996.

(11) Rooijakkers, B. and Bouchehati, M. *Trading day adjustment for the consumption of Dutch households (methodological note)*, Statistics Netherlands, 2005.

(12) Cano, S., P. Getz, J. Kropf, S. Scott, and G. Stamas, 1996, Adjusting for A Calendar Effect in Employment Time Series, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1996.

(13) Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. *Applied Linear Statistical Models*. WCB/McGraw-Hill, 1996.

(14) Makridakis, S., Wheelwright, S. and Hyndman, R. *Forecasting: Methods and Applications*. Third edition, John Wiley and Sons, 1998.

(15) Pankratz, A. *Forecasting With Dynamic Regression Models*. John Wiley & Sons, 1991.

(16) *COST329: Models for Traffic and Safety Development and Interventions. Final Report of the Action*, European Commission, 1999.

(17) Van den Bossche, F. and Wets, G. *Macro Models in Traffic Safety and the DRAG Family: Literature Review*. Report RA-2003-08, Flemish Policy Research Center for Traffic Safety, Diepenbeek, 2003.

(18) Raeside, R. and White, D. Predicting Casualty Numbers in Great Britain, *Transportation Research Record: Journal of the Transportation Research Board*, No. 1897, TRB, National Research Council, Washington, DC., pp. 142-147, 2004.

(19) Van den Bossche, F., Wets,G. and Brijs,T. A Regression Model with ARMA Errors to Investigate the Frequency and Severity of Road Traffic Accidents. In: *Proceedings of the 83rd annual Meeting of the Transportation Research Board*, USA, 2004.

(20) Van den Bossche F., Wets G., and Brijs T. The role of exposure in the analysis of road accidents: a Belgian case-study. Forthcoming in: *Transportation Research Record, Journal of the Transportation Research Board*, 2005.

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

(22) Fournier, F. and Simard, R. The DRAG-2 Model for Quebec. In: *Structural Road Accident Models: The International DRAG Family* (Gaudry, M. and Lassarre, S. Eds.), Chap. 2, pp. 37-66, Elsevier Science, Oxford, 2000.

(23) Blum, U. and Gaudry, M. The SNUS-2.5 Model for Germany. In: *Structural Road Accident Models: The International DRAG Family* (Gaudry, M. and Lassarre, S. Eds.). Chap. 3, pp. 67-96, Elsevier Science, Oxford, 2000.

(24) Marquardt D.W. Generalized inverses, ridge regression, biased linear estimation and non-linear estimation. *Technometrics*, 12, 1970, pp. 591-612.

(25) SAS Institute Inc. *SAS/ETS/9.1 User's Guide*. Cary, NC: SAS Institute Inc, 2004.

(26) Akaike, H., A New Look at the Statistical Model Identification, *IEEE Transaction on Automatic Control*, AC-19, 716-723, 1974.

(27) Schwarz, G., Estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464, 1978.

(28) Verbeek, M. *A Guide to Modern Econometrics*, John Wiley and Sons, 2000.

(29) Theil, H. *Applied Economic Forecasting*, Amsterdam: North-Holland, 26-32, 1966.

Acknowledgements

This study is supported by the Flemish Government via a research grant offered to the Policy Research Center for Traffic Safety. We are also grateful to Mr. M. Van Damme from the Flemish Automobile Association (VAB), for providing the data on the "heavy traffic" measure.