INVESTIGATING THE EFFECT OF HOLIDAYS ON DAILY TRAFFIC COUNTS: A TIME SERIES APPROACH

Mario Cools, Elke Moons, Geert Wets*

Transportation Research Institute
Hasselt University
Wetenschapspark 5, bus 6
BE-3590 Diepenbeek
Belgium
Fax.:+32(0)11 26 91 99
Tel.:+32(0)11 26 91{31, 26, 58}
Email: {mario.cools, elke.moons, geert.wets}@uhasselt.be

* Corresponding author

Number of words = 5421
Number of Figures = 5
Number of Tables = 3
Words counted:5421 + 8*250 = 7421 words

Revised paper submitted: November 15, 2006

ABSTRACT

In this paper, different modeling philosophies are explored in order to explain and forecast daily traffic counts. The main objectives of this study are the analysis of the impact of holidays on daily traffic, and the forecasting of future traffic counts. Data coming from single inductive loop detectors, collected in 2003, 2004 and 2005, were used for the analysis. The different models that were investigated showed that the variation in daily traffic counts could be explained by weekly cycles. The Box-Tiao modeling approach was applied to quantify the effect of holidays on daily traffic. The results showed that traffic counts were significantly lower for holiday periods. When the different modeling techniques were compared with respect to forecasting with a large forecast horizon, Box-Tiao modeling clearly outperformed the other modeling strategies. Simultaneous modeling of both the underlying reasons of travel, and revealed traffic patterns, certainly is a challenge for further research.

1 BACKGROUND

In today’s society, mobility is one of the driving forces of human development. The motives for travel trips are not confined to work or educational purposes, but reach a spectrum of diverse goals. Mobility is more than a cornerstone for economic growth; it is a social need that offers people the opportunity for self-fulfillment and relaxation (*1*).

Governments recognize this significance of mobility. This is evidenced by the mobility plans that are formulated by government agencies at different policy levels, e.g. at European level the European Commission’s White paper “European transport policy for 2010: time to decide” (*2*), and at Belgian regional level the “Mobiliteitsplan Vlaanderen” (Mobility plan Flanders (*3*)), and evidenced by the transportation research that is directly or indirectly funded by governments.

In order to lead an efficient policy, governments require reliable predictions of travel behavior, traffic performance, and traffic safety. A better understanding in the events that influence travel behavior and traffic performance, will lead to better forecasts and consequently policy measures can be based upon more accurate data. This allows policy makers to provide more precise travel information and adapt the dynamic traffic management systems, so that an important goal, more acceptable and reliable travel times (*1*), can be achieved.

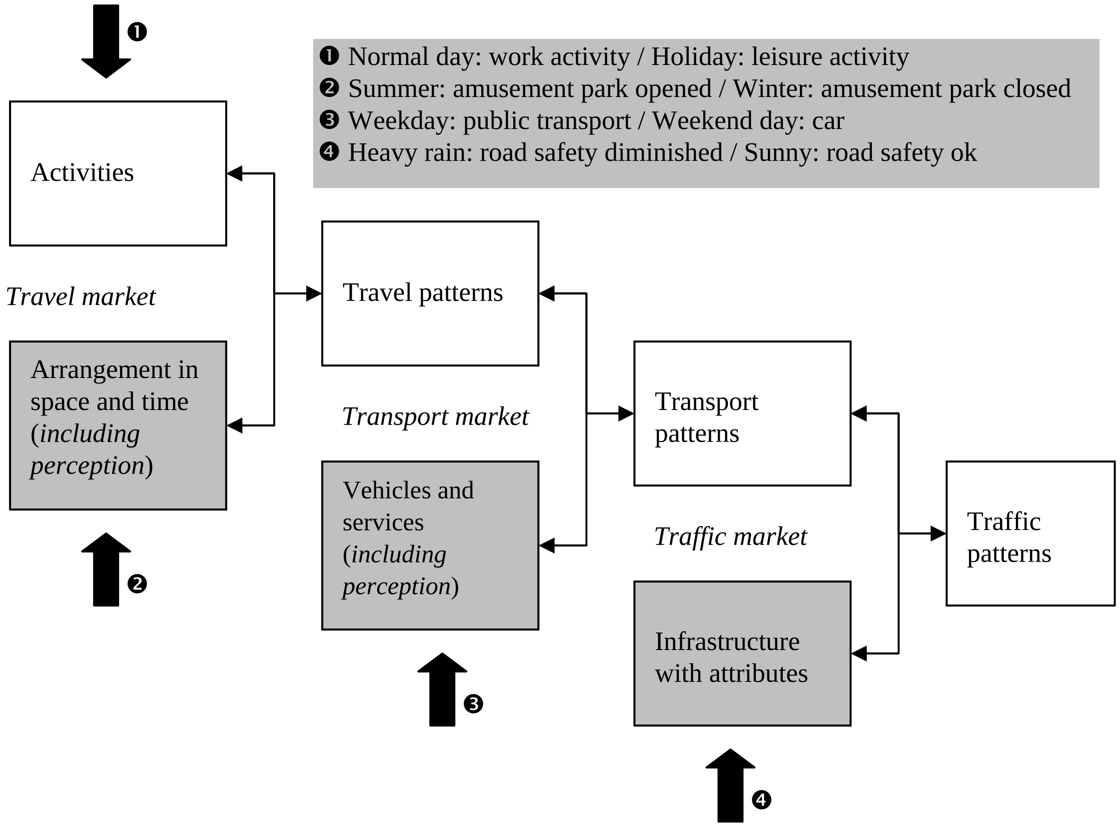
❶
❶ Normal day: work activity / Holiday: leisure activity
❷ Summer: amusement park opened / Winter: amusement park closed
❸ Weekday: public transport / Weekend day: car
❹ Heavy rain: road safety diminished / Sunny: road safety ok
Activities
Travel market
Arrangement in space and time (*including perception*)
❷
Travel patterns
Transport market
Vehicles and services (*including perception*)
❸
Transport patterns
Traffic market
Infrastructure with attributes
❹
Traffic patterns

FIGURE 1 Three market model and effects that influence mobility.

Events such as special holidays (e.g. Christmas, New Year's day), school holidays (e.g. in July and August), socio-demographic changes, and weather, can have an influence on mobility in different ways, as is illustrated by Figure 1 (*4*). First, they can influence the travel market. This is the market where the demand for activities and the supply of activity opportunities in space and time result in travel patterns. Second, these events can have an influence on the transport market. At this market, the demanded travel patterns and the supply of transport options come together in a transport pattern that assigns passenger- and good trips to vehicles and transport services. Finally, these events can have an effect on the traffic market, where the required transport patterns are confronted with the actual supply of infrastructure and their associated management systems, resulting in an actual use of the infrastructure, revealed by the traffic patterns.

When the list of examples, which is given in Figure 1, is considered, one can notice that people might perform other activities during holidays, than during normal days. During holidays, people for example go to the beach, while during normal days, people go to work. Another effect that is indicated by Figure 1, is the closing period of amusement parks during the winter. People wishing to visit the park during the winter, obviously can't, and will perform another activity, for instance ice skating. These are merely two examples of how holidays and seasonal effects influence the activities that people pursue and in turn, these activities have an impact on the travel market. Another example shows how mode choice can be influenced by the type of day, and this can have an impact of the transport market, while the fourth illustration demonstrates how the environment can have an impact on the traffic market. Note that the list of examples, given in Figure 1, is not limitative, but is meant as an exemplification of how the three markets and hence the mobility can be influenced by various events.

The main objectives of this study are the identification of the effects of holidays on daily traffic, and the prediction of future traffic volumes. A Box-Tiao model is used to quantify these effects. The combination of a regression model with ARMA (Auto-Regressive Moving Average) errors raises the opportunity to build a model with desirable statistical properties, and thus to minimize the risk of erroneous model interpretation (*5*).

The text is organized in the following way. First, an overview of the data is given, and the imputation strategy that was applied is discussed. Then, the methodology of the different models used in the analysis is explained. Next, the model outcomes and the forecast are presented. Finally, some general discussion and avenues for further research are provided.

2 DATA

The impact of holidays on daily traffic will be analyzed by studying the effect on daily highway traffic counts. In this section, first, the dependent variable (daily traffic count) is further explored. Then, the different covariates, called interventions in Box-Tiao terminology, are described.

2.1 Daily Traffic

The aggregated daily traffic counts originate from minute data of two single inductive loop detectors, located on the E314 Highway in the direction of Brussels in Gasthuisberg (Leuven, Belgium), collected in 2003, 2004 and 2005 by the Vlaams Verkeercentrum (Flemish Traffic

Control Center). Figure 2 pin-points the traffic count location under study. The highway that is analyzed is one of the entranceways of Brussels, and thus excessively used by commuters.

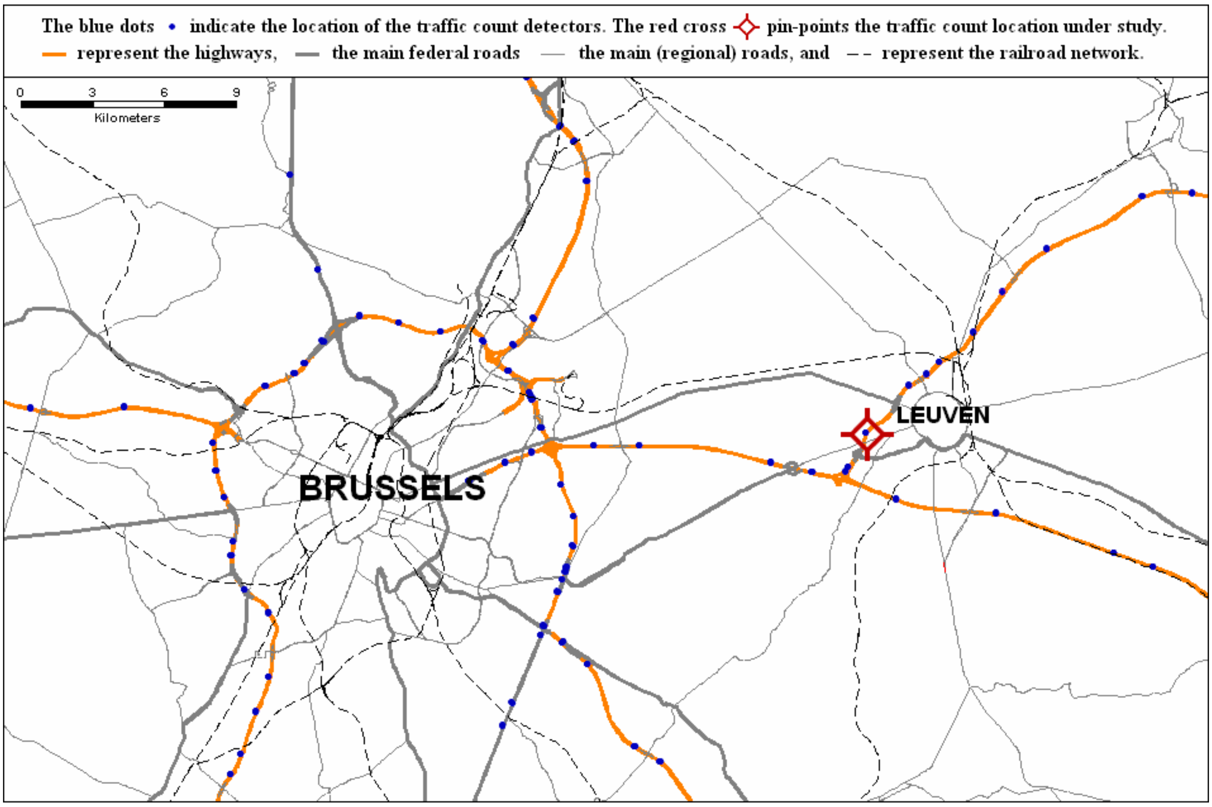
The blue dots • indicate the location of the traffic count detectors. The red cross ◇ pin-points the traffic count location under study.
— represent the highways, — the main federal roads — the main (regional) roads, and – – represent the railroad network.
0 3 6 9
Kilometers
BRUSSELS
LEUVEN

FIGURE 2 Geographical representation of the traffic count location under study.

Every minute, the loop detectors output four variables: the number of cars driven by, the number of trucks, the occupancy of the detector and the time-mean speed of all vehicles (*6*). The number of cars and trucks are added up for both detectors, yielding a total traffic count for each minute. The aggregation on daily basis of these minute data can only be done when there are no missing data that day. When some, or all of the minute data are missing, a defendable imputation strategy must be applied.

TABLE 1 Missing Data Analysis and Corresponding Imputation Strategy

Quality Assessment	Number of days	% of all days	Imputation strategy
No minutes missing	552	50,36%	no strategy
1-60 minutes missing	342	31,20%	strategy 1
61-240 minutes missing	48	4,38%	strategy 1
241-720 minutes missing	68	6,20%	strategy 1
721-1320 minutes missing	32	2,92%	strategy 1
1321-1439 minutes missing	3	0,27%	strategy 2
Entire day missing	51	4,65%	strategy 2
Total	1096	100,00%	

About half of all the days, that were analyzed, contained no missing data, as is shown in Table 1. Obviously, for these days no imputation strategy needed to be applied. This, however means that for the other half, there were some (41,78%) or a lot (7,84%) of the minute count data missing. When at least two hours of data, so at least 120 of the 1440 data points, were available, an imputation strategy was applied that is very similar to the "reference days"-method proposed by Bellemans (*7*). When there were fewer than 120 data points available in a day, a more general imputation strategy was applied.

2.1.1 Imputation strategy 1

Bellemans (*7*) assumed in his work the existence of an a priori known reference day that is representative of the day for which missing values have to be estimated. The imputed value is then calculated by scaling the reference measurement such that it corresponds to the traffic dynamics of the day under study. In his study, the scaling factor was the fraction of the measurement and the reference measurement, in the previous minute.

The imputation strategy applied in this study uses the ideas of the reference days and the use of a scaling factor. The new measurements $x_{new}\left(t\right)$ are calculated in the following way:

$$x_{new}\left(t\right)=\delta x_{ref}\left(t\right)$$

where $x_{ref}\left(t\right)$ is the reference measurement and δ the scaling factor. For determining the reference measurement, 21 reference days (7 days for each of 3 holiday statuses) were used. For each reference day, the reference measurements were defined as the average of the modus, median and mean of the available days that corresponded to the reference day. The average of these three measures of central tendency was taken, because each of them has its own unique attributes (central location, robustness, highest selection probability), and favoring one could obscure model interpretation. The scaling factor δ is calculated as follows:

$$\delta=\frac{\sum_{t=1}^{1440}d_t}{\sum_{t=1}^{1440}m_t},$$

$$\text{where } d_t=\begin{cases}\dfrac{x\left(t\right)}{x_{ref}\left(t\right)}\Leftrightarrow x\left(t\right)\text{ not missing}\\ 0\Leftrightarrow x\left(t\right)\text{ missing}\end{cases},m_t=\begin{cases}1\Leftrightarrow x\left(t\right)\text{ not missing}\\ 0\Leftrightarrow x\left(t\right)\text{ missing}\end{cases}$$

In the above equations, $x\left(t\right)$ is the measurement at minute *t* and $x_{ref}\left(t\right)$ the reference measurement at minute *t*.

2.1.2 Imputation strategy 2

For the above described imputation strategy, a scaling factor was required to match the reference measurement to the day under study. When all, or almost all, of the data points are missing, the scaling factor could not be calculated. In this case, the missing values are replaced by the reference measurements, which is equivalent with setting the scaling factor equal to 1.

2.1.3 Evaluation of implemented imputation strategies

Circumspection is essential when applying imputation strategies, as imputation processes encompass the risk of distorting the distributions of the data, and thus biasing the results. The magnitude of the risk must be indicated and potential patterns of the missing data need to be analyzed.

When the risk of distortion of the data is addressed, a thorough look at the minute data places the risk in the correct context. Of the 1578240 minutes (1096 days multiplied by 1440 minutes a day) that were aggregated on a daily basis, 140860 minutes (8,93%) were missing. Communication errors (e.g. due to system failures) account for 135654 minutes (8,60%) of missing data. The remaining 5206 minutes (0,33%) were due to other reasons such as physical errors of the loop detectors, disturbances in the electronic systems of the substations and inaccurate measurements.

When the imputation strategies are evaluated on the daily level, a first observation is that 81,56% (50,36% + 31,20%) of the days contains at least 95,83% (more then 1380 of the 1440 data points) of the data points that day. Thus, the imputation strategy has nearly no effect on these days. For the days (4,65% + 0,27%) that contained nearly no information (less than 120 of the 1440 data points available), just a measure of central tendency was used as imputed value, taking into account the day type (which day of the week and holiday or not). For 10,58% (4,38% + 6,20%) of the days between 50% and 95,83% of the data points were available, so the scaling factor used for the imputation strategy was still based upon a reliable amount of data. Only for 2,92% of the days, less than 50%, but at least 8,33% of the data points were available. It might be judged that the imputation strategy could result here in significantly distorted values. Different imputation strategies could be applied to this part of the data to assess the effect of the chosen strategy. However, since it is only a very small part of the entire data set, it was judged not to have a significant impact on the remainder of the study.

It is important to stress that the imputation strategies applied use a measure of central tendency that takes into account the day of week and the holiday status. Thus, the significance of these variables (day of week, holiday status) is not affected by the choice of the measure of central tendency. It is fair to recapitulate and infer that the implemented imputation strategies had no significant distorting effect on the results or conclusions.

2.1.4 Plot of the data

The following figure visualizes the aggregated daily traffic count data, taking into account the imputation strategies that were implemented. A similar pattern is visible over the three years. A drop in the number of passing vehicles at the beginning and end of each year is noticed, and during summer holidays, the intensity of daily traffic clearly is lower than during the other months.

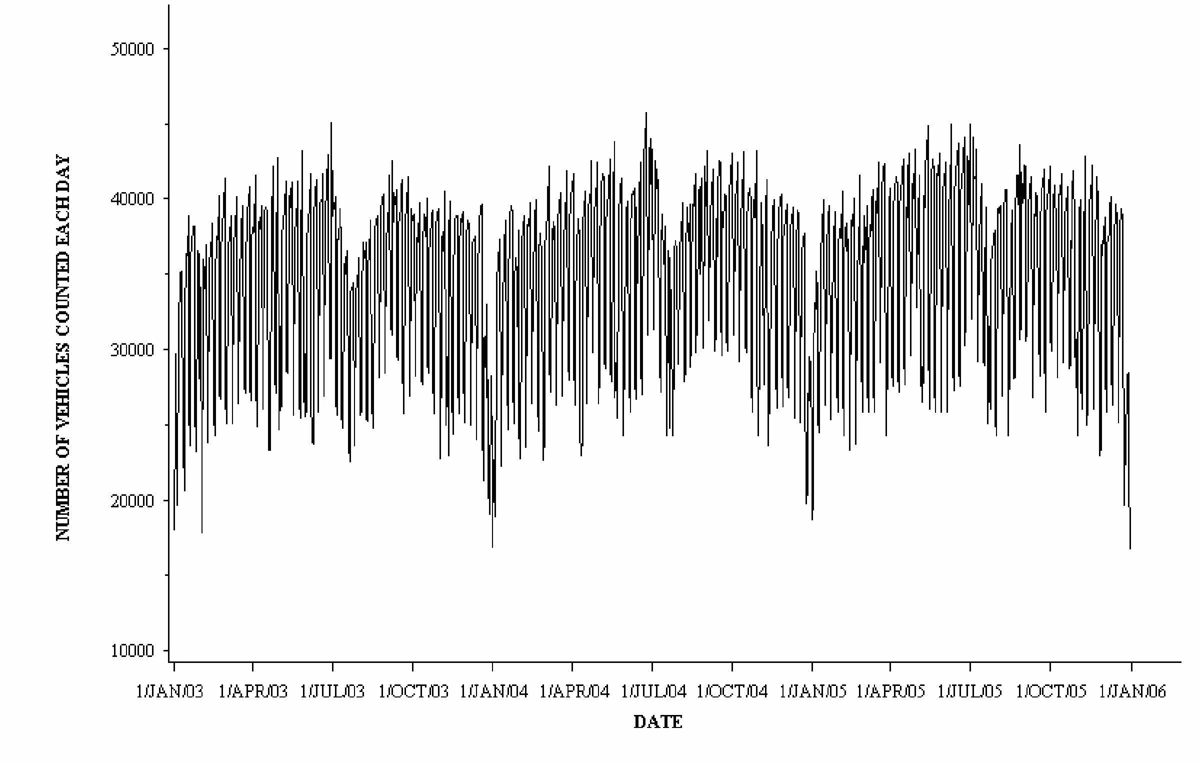

FIGURE 3 Evolution in time of daily traffic counts.

Next the different interventions will be briefly summarized.

2.2 Holiday Effect

A dummy variable was created in order to model the effect of holidays. "Normal" days were coded zero, and holidays were coded one. The following holidays were considered: Christmas vacation, spring half-term, Easter vacation, Labor Day, Ascension Day, Whit Monday, vacation of the construction industry (three weeks, starting the second Monday of July), Our Blessed Lady Ascension, fall break (including All Saints' Day and All Soul's Day), and finally Remembrance Day. Note that for all these holidays, the adjacent weekends, were considered to be a holiday too. For holidays occurring on a Tuesday or on a Thursday, respectively the Monday and weekend before, and the Friday and weekend after, were also defined as a holiday, because often people have a day-off at those days, and thus have a leave of several days, which might be used to go on a long weekend or on a short holiday.

2.3 Day Effects

Six dummy variables were created in order to model the day-of-week effect. Note that in general you have to create $k-1$ dummy variables, if you want to analyze the effect of a categorical variable with k classes (*8*). Since, there are seven days in a week, the first six days (Monday until Saturday) were each represented by one of the dummies, equal to one for the days they represent, and zero elsewhere. The reference day was Sunday, so for all traffic counts that were collected on a Sunday, the corresponding six dummies were coded zero.

3 METHODOLOGY

In this study, two main philosophies were explored in order to model the daily traffic counts. The first philosophy is based on the fact that consecutive traffic counts are correlated, and that therefore present and future values can be explained by past values. Two types of models that use this philosophy are investigated in this paper, namely exponential smoothing and ARMA modeling. The second philosophy is the regression philosophy, which postulates the idea that the dependent variable, in this study the daily traffic counts, could be explained by other variables. Since different assumptions have to be met before the linear regression model yields interpretable parameter estimates, also the Box-Tiao-model is investigated. The latter is capable of taking into account dependencies between error terms. For an introduction on time series analysis, the reader is referred to Yaffee and McGee (*9*). In Neter et al. (*8*) a comprehensive overview of regression models is given.

3.1 Exponential Smoothing

3.1.1 Simple Exponential Smoothing

Simple exponential smoothing is a way of forecasting future observations, by producing a time trend forecast, where the parameters are allowed to change gradually over time, and where recent observations are given more weight than observations further in the past (*9*). The technique assumes that the data fluctuate around a reasonably stable mean. The formula for simple exponential smoothing is: $S_t = \alpha Y_t + \left(1-\alpha\right) S_{t-1}$, where each new smoothed value S_t is computed as the weighted average of the current observation Y_t and the previous smoothed observation S_{t-1}. The magnitude of the smoothing constant α, ranges between zero and one. If the constant equals to one, then the previous observations are ignored entirely. If the constant equals to zero, then the current observation is ignored entirely, and the smoothed value consists entirely of the previous smoothed value, thus, as a consequence, all smoothed values will be equal to the initial smoothed value S_0.

3.1.2 Multiplicative Holt-Winters Exponential Smoothing

In order to accommodate the simple exponential smoothing model to account for regular seasonal fluctuations, the Holt-Winters method combines a time trend with multiplicative seasonal factors (*10*). The general formula for the multiplicative Holt-Winters model is $\hat{Y}_{t+h} = \left(\mu_t + b_t h\right) S_{t-p+h}$, where $\hat{Y}_{t+h}$ is the estimated value for the time series at time $t+h$, *h* the number of periods into the forecast horizon, μ_t the permanent component at time *t*, b_t the trend component at time *t*, S_{t-p+h} the multiplicative seasonal component at time $t-p+h$, and *p* the periodicity of the seasonality (the number of periods in one cycle of seasons).

Each of the three parameters (μ_t, b_t, S_t) is updated with its own exponential smoothing equation (*9*). The permanent component is updated by the following equation:

$$\mu_t = \alpha\left(\frac{Y_t}{S_{t-p}}\right) + (1-\alpha)(\mu_{t-1} + b_{t-1}).$$

Dividing the series Y_t by its seasonal component at its periodic lag, removes the seasonality from the data. Therefore, only the trend component and the prior value of the permanent component enter into the updating process for μ_t. The updating equation for the trend component is given by $b_t = \gamma(\mu_t - \mu_{t-1}) + (1-\gamma)b_{t-1}$. Thus, the trend component is simply the smoothed difference between two successive estimates of the deseasonalized level. The last parameter, the multiplicative seasonal component, is updated by the following smoothing equation:

$$S_t = \delta\left(\frac{Y_t}{\mu_t}\right) + (1-\delta)S_{t-p}.$$

Thus, the seasonal component is updated by a portion of the ratio of the series value over the average, plus a smoothed portion of the seasonality at its periodic lag.

3.2 ARMA Modeling

Like exponential smoothing, also the ARMA modeling approach tries to explain current and future values of a variable as a weighted average of its own past values. In most cases, the model consists of a combination of an autoregressive (AR) part and a moving average (MA) part.

When Y_t is modeled as an autoregressive process AR(*p*), then Y_t can be expressed in terms of its own passed values. Suppose Y_t is modeled as an autoregressive process of order two, AR(2), then $Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$, where ϕ_1, ϕ_2 are the weights for the autoregressive terms, c a constant and e_t a new random term. Using a backshift operator B_i on Y_t, defined as $B^i(Y_t) = Y_{t-i}$, this process can be written as $Y_t = c + \phi_1 BY_t + \phi_2 B^2 Y_t + e_t$ or $\left(1 - \phi_1 B - \phi_2 B^2\right)Y_t = .$ $c + e_t$

When the series Y_t is modeled as a moving average process MA(*q*), then Y_t can be expressed in terms of current and past errors, also called shocks. Suppose Y_t is modeled as a moving average process of order two, MA(2), then $Y_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$, where θ_1, θ_2 are the weights for the moving average terms. Using the backshift operator, previously defined, this process can be written as $Y_t = c + \left(1 - \theta_1 B - \theta_2 B^2\right)e_t$.

In the cases that a series Y_t is modeled as combination of an autoregressive process of order *p*, AR(*p*), and a moving average process of order *q*, MA(*q*), the combined process is called an ARMA(*p*,*q*) process. The model is then given by:

$$\left(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\right)Y_t = c + \left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right)e_t.$$

Note that the ARMA-model is only valid, when the series satisfies the requirement of weak stationarity. A time series is weakly stationary when the mean value function is constant and does not depend on time, and that the variance around the mean remains constant over time (*11*). If the variance of the series does not remain constant over time, a transformation, like taking the logarithm or the square root of the series, often proves itself be a good remedial measure to achieve constancy (*8*). To achieve stationarity in terms of the mean, it sometimes is required to difference the original series. Successive changes in the series are then modeled instead of the original series. When differencing is applied, the ARMA model is called an AR*I*MA model where "*I*" indicates that the series is differenced.

3.3 Regression Modeling

Instead of modeling a series Y_t as a combination of its past values, the regression approach tries to explain the series Y_t with other covariates. Formally, the multiple linear regression model can be represented by the following equation:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \varepsilon_t \,,$$

where Y_t is the *t*-th observation of the dependent variable, and $X_{1,t}, X_{2,t}, \ldots, X_{k,t}$ are the corresponding observations of the explanatory variables. $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the parameters of the regression model, which are fixed, but unknown, and ε_t is the unknown random error (*8*).

Estimates for the unknown parameters can be obtained by using classical estimation techniques. If $b_0, b_1, b_2, \ldots, b_k$ are the estimates for $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, then the estimated value for the dependent variable Y_t is given by $\hat{Y}_t = b_0 + b_1 X_{1,t} + b_2 X_{2,t} + \ldots + b_k X_{k,t}$.When the error terms are independently and identically normally distributed with mean 0 and variance σ^2, then the estimators for the parameters are BLUE (Best Linear Unbiased Estimators).

3.4 Box-Tiao Modeling

When regression modeling is applied to time series, the assumption of independence of the error terms is often violated because of autocorrelation (the error terms being correlated among themselves). This violation of one of the underlying assumptions of linear regression increases the risk for erroneous model interpretation, because the true variance of the parameter estimates may be seriously underestimated (*8*).

Box-Tiao modeling can be used to solve this problem of autocorrelation. A Box-Tiao model corrects for autocorrelation by describing the errors terms of the linear regression model by an ARMA(*p*,*q*) process. Let $Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + N_t$ be the regression model, where $\left(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\right) N_t = \left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right) \varepsilon_t$ and ε_t is assumed to be white noise, then the Box-Tiao model can then be represented by the following equation:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \frac{\left(1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q\right)}{\left(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p\right)} \varepsilon_t \,.$$

The parameters in this equation are estimated using Maximum Likelihood. Studies, comparing least squares methods with maximum likelihood methods for this kind of models, show that maximum likelihood estimation gives more accurate results (*12*). The Likelihood function is maximized via nonlinear least squares using Marquardt's method (*10*). When differencing of the error terms is required to obtain stationarity, all dependent and independent variables should be differenced (*5*,*13*).

3.5 Model Evaluation

Since different types of models are considered to estimate the daily traffic counts, it is required that an objective criterion is used to determine which model performs better (*14*). The following criteria were used to determine the appropriateness off the models: the Akaike Information Criterion (AIC), the Mean Square Error (MSE) and the Mean Absolute Percentage Error (MAPE). Note that the models were constructed on a training data set containing the first 75% of the observations. The remaining 25% of the observations make up the validation or test data set that can be used to assess the performance of the models, by calculating the MSE and MAPE for the forecasts. The choice of these percentages is arbitrary, but common practice in validation studies (see e.g. Wets et al. (*15)* or Moons (*16*)).

The Akaike Information Criterion (AIC) is defined as $AIC = \text{-}2 \times log\ likelihood + 2 \times number\ of\ free\ parameters$. Models with a lower value for this criterion are considered to be the more appropriate ones (*17*). The Mean Square Error (MSE) equals the Sum of all Squared Errors (SSE) divided by its degrees of freedom, which are calculated by subtracting the number of parameters in the model from the number of observations. The Mean Absolute Percentage Error (MAPE) is defined as the average of the absolute values of the proportion of error at a given point of time.

4 RESULTS

In this Section, the results are presented. The parameter estimates of the models are interpreted, and the different models are compared with each other. Predictions of the daily traffic counts are graphically displayed. A distinction is made between the predictions that are based on the training data (Figure 4), and the predictions that are based on the test data (Figure 5).

4.1 Holt-Winters Multiplicative Exponential Smoothing

The best Holt-Winters model, in terms of AIC, was obtained when a cycle of seven seasons (the seven seasons correspond to the seven days of the week), combined with a linear trend, was considered. In this model, nine (seven plus two) parameters had to be estimated: the parameter for the permanent component ($\hat{\mu}_1 = 35154$), the parameter for the linear component ($\hat{b}_1 = -64.56$), and the seven factors of the seasonal component. The estimated seasonal parameters are given by $\hat{S}_1 = 1.122,\ \hat{S}_2 = 1.128,\quad \hat{S}_3 = 1.137,\quad \hat{S}_4 = 0.781,\quad \hat{S}_5 = 0.731,$ $\hat{S}_6 = 1.009,\ \hat{S}_7 = 1.091,$ where $i = 1, 2, ..., 7$ represents the ordering of the seasonal parameters. The average of these seven parameters must be equal to one (*9*). Note that these seasonal factors correspond to the different days of a week. Since the first observation in the data set was a Wednesday (January 1, 2003), the first seasonal factor also represents a Wednesday. Similarly, the other seasonal factors represent the other days of the week. Recall that the Holt-Winters method uses smoothing equations for updating the parameters. The smoothing parameters for the permanent component and the linear component are given by $\alpha = \gamma = 0.106$ and the smoothing parameter for the seasonal component is given by $\delta = 0.25$.

When the estimates for the seasonal parameters are compared, the difference between the components that correspond to the weekend-days and the components that correspond to the week-days is appealing. The results indicate that during weekend-days the daily traffic count will be much lower. This tendency can also be observed in Figure 4.

4.2 ARMA Modeling

In order to obtain stationarity, the ARMA model was developed on differenced data. The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) of the residuals were investigated to determine which Autoregressive (AR) and moving average (MA) factors were required to build the model. Let $\nabla_1 Y_t$ denote the first difference of the data ($Y_t - Y_{t-1}$), then the obtained model could be written as:

$$\nabla_1 Y_t = \frac{(1-0.79134B)(1-0.94883B^4)(1-0.9998B^7)}{(1-0.308B)(1-0.87767B^4)(1-B^7)}\varepsilon_t .$$

This model contains three multiplicative autoregressive and three multiplicative moving average factors. Notice that if the model is worked out, other autoregressive and moving factors also play a role. When the parameter estimates for the ARMA factors are investigated, it can be seen that the estimates for the terms of the seventh order are very close or equal to one. This is an indication for the weekly cyclic behavior, which was also evidenced by the Holt-Winters model. The high parameters estimates for the ARMA factors of the fourth order might be evidence of some half-week recurring pattern in daily traffic counts. The dependency on the previous day was much smaller, yet significant.

4.3 Box-Tiao Modeling

The classical linear regression modeling approach did not yield valid results, because of the problem of autocorrelation of the error terms. As is indicated in Section 3.4, Box-Tiao modeling is an approach that can tackle this problem of autocorrelation.

Like for the ARMA modeling, it was also for the Box-Tiao modeling required to take the first difference of the data to obtain stationarity. Note that for both the ARMA model and the Box-Tiao model the intercept was dropped from the equations. When differencing is done, the intercept is interpreted as a deterministic trend, and that is not always realistic (*13*). The final error terms obtained were accepted to be ‘white noise’ according to the Ljung-Box Q^*-statistics (*18*). The final Box-Tiao model obtained is given by the following equation:

$$\nabla_1 Y_t = \begin{cases} -4130\nabla_1 X_{Holiday,t} \\ +9176\nabla_1 X_{Monday,t} + 10812\nabla_1 X_{Tuesday,t} + 11546\nabla_1 X_{Wednesday,t} \\ +11295\nabla_1 X_{Thursday,t} + 12053\nabla_1 X_{Friday,t} + 1450\nabla_1 X_{Saturday,t} \\ +\dfrac{(1-0.858\ B)(1-0.937\ B^4)}{(1-0.272\ B)(1-0.893\ B^4)}\varepsilon_t \end{cases}$$

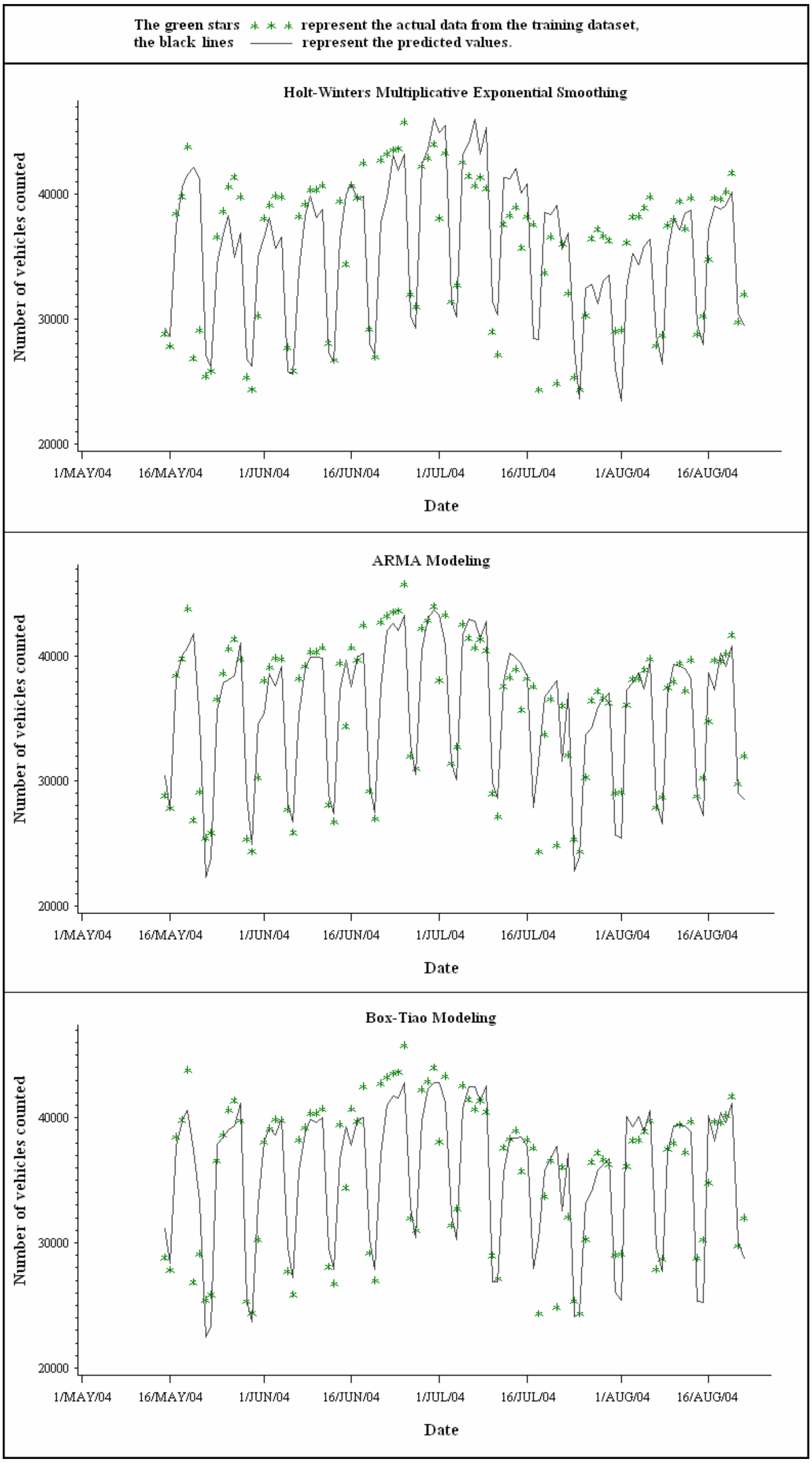

FIGURE 4 Daily traffic counts and their corresponding predicted values (a subset of the training data set).

The six dummy variables to model the day-of-week effect, and the dummy variable of the holiday effect were all very significant (p-value < 0.0001) as can be seen from Table 2. This evidences that the daily traffic counts are influenced by holidays. Interpretation of the parameter estimates is not straightforward since both the dependent and independent variables were differenced.

TABLE 2 Parameter Estimates for the Box-Tiao Model

Parameter	Estimate	Standard Error	t-value	p-value
Moving Average (Lag 1)	0.858	0.026	32,7	< 0.0001
Moving Average (Lag 4)	0.937	0.047	20,1	< 0.0001
Auto Regressive (Lag 1)	0.272	0.046	5,9	< 0.0001
Auto Regressive (Lag 4)	0.893	0.061	14,5	< 0.0001
Holiday	-4130	303.67	-13,6	< 0.0001
Monday	9176	264.38	34,7	< 0.0001
Tuesday	10812	299.24	36,1	< 0.0001
Wednesday	11546	307.85	37,5	< 0.0001
Thursday	11295	306.17	36,9	< 0.0001
Friday	12053	296.99	40,6	< 0.0001
Saturday	1450	260.18	5,6	< 0.0001

The parameter estimate for the holiday effect could be interpreted in the following way. When the holiday starts (the differenced holiday dummy equals one), the daily traffic count will be 4130 vehicles lower than the day before. The day after the holiday (the differenced holiday dummy equals minus one), the daily traffic count will increase again with 4130 vehicles. Note that for all other days the differenced holiday dummy equals zero.

For the interpretation of the parameter estimates for the day-of-week effects, the Wednesdays are taken as an example. On a Wednesday, the differenced dummy of the Wednesday-effect equals one, and the differenced dummy of the Tuesday-effect equals minus one. All other differenced day-of-week dummies equal zero for a Wednesday. Thus, on a Wednesday, the traffic count will be 734 (11546-10812) vehicles higher than the day before (obviously the Tuesday before).

4.4 Model Comparison

When the different models are compared, the weekly cyclic behavior was exposed by all three models. In the Holt-Winters Exponential Smoothing model this cyclicality was revealed by the seasonal component, in the ARMA model by the high estimates for the seventh order autoregressive and moving average factors, and in the Box-Tiao model by the clearly significant day-of-week effect. Differences between different weekdays were also discovered by Weijermars & Berkum (*19*). In their work, they used cluster analysis techniques that revealed the differences.

In order to determine whether predicting daily traffic counts with other covariates, such as the holiday effect and the day-of-week effects, adds insight, different criteria that assess the model fit are shown in Table 3.

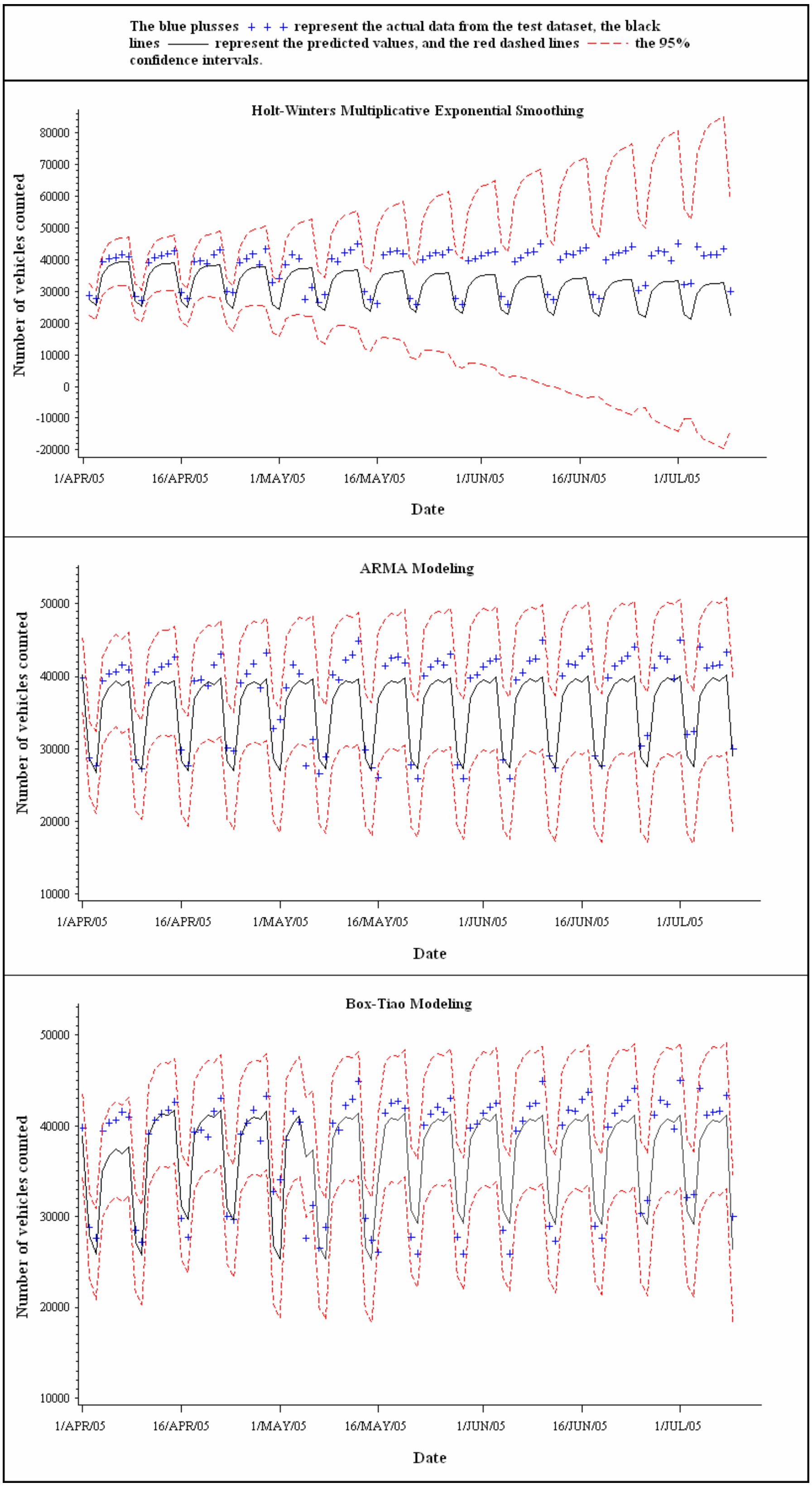

FIGURE 5 Daily traffic counts and their corresponding predicted values and confidence bounds (a subset of the test data set).

TABLE 3 Criteria for Model Comparisons

Criterion	Holt-Winters	ARMA	Box-Tiao
Comparison based on training data set			
AIC(Model)	13,295.7	15,284.8	15,095.2
MSE(Model)	10,469,329	6 708 034	5,573,994
MAPE(Model)	6.788	5.373	4.976
Comparison based on test data set			
MSE(Forecast)	125,737,853	14,638,087	9,375,331
MAPE(Forecast)	27.683	7.685	6.482

According to the AIC the best model is the Holt-Winters model, but when the other criteria are assessed, the Box-Tiao model outperforms the other models, indicating that considering a holiday effect and day-of-week effects with a Box-Tiao model really adds insight into the cyclicality of daily traffic counts. Note that Liu and Sharma (*20*) also identified a significant holiday effect.

Figure 4 shows that the predictions that are based upon the training data set are comparable for the three modeling strategies. The MSE and MAPE criteria for these predictions indicate that the ARMA and Box-Tiao models perform better, however the AIC favors the Holt-Winters model.

When the different models are validated by a test data set it can be seen from Figure 5 that the Box-Tiao model performs best. The ARMA model also performs quite well, but the Holt-Winters model performs only well for a very short forecast horizon. The MSE and MAPE criteria demonstrate that the ARMA and Box-Tiao model approaches outperform the Holt-Winters Exponential Smoothing model, favoring the Box-Tiao model.

5 CONCLUSIONS AND FURTHER RESEARCH

In this study, different modeling approaches were considered to predict daily traffic counts. The different techniques pointed out the significance of the day-of-week effects: weekly cycles seem to determine the variation of daily traffic flows. In the weekends the daily traffic flows turn out to be lower than during the week. The Box-Tiao model approach demonstrated that during holidays the daily traffic flows are significantly lower.

When forecasting of daily traffic flows is required, the Box-Tiao model appears to be an approach that performs reasonably well. Smoothing techniques, like the Holt-Winters Exponential Smoothing model, are to be avoided for predictions with a large forecast horizon.

These findings can be used by policy makers to fine-tune current policy measures. More precise travel information can be provided and the dynamic traffic management systems can be improved. In this way, the findings of this study contribute in achieving an important goal, i.e. more acceptable and reliable travel times.

The analysis of day-of-week and holiday effects in this study was done on the revealed traffic patterns. Generalization of the discussed results is possible, when traffic patterns of other parts of the road network are analyzed. In order to get more insight in how holidays affect mobility, further analysis is required. The different modeling techniques described in this paper could be applied on data from national travel surveys, to determine potential effects on travel behavior. Simultaneous modeling of both the underlying reasons of travel, and revealed traffic patterns, certainly is a challenge for further research.

6 ACKNOWLEDGEMENTS

The authors would like to thank the Vlaams Verkeerscentrum (Flemish Traffic Control Center) for providing the data used in this study.

7 REFERENCES

(1) Ministerie van Verkeer en Waterstaat. *Nota Mobiliteit*. Den Haag, 2004.

(2) European Commission. *White Paper on European Transport Policy 2010: Time to Decide*. 2004. http://ec.europa.eu/transport/white_paper/index_en.html. Accessed, July 31, 2006.

(3) Ministerie van de Vlaamse Gemeenschap. *Ontwerp Mobiliteitsplan Vlaanderen*. Departement Leefmilieu en Infrastructuur, Mobiliteitscel, 2001.

(4) Egeter, B., van de Riet, O.A.W.T. *Systeemdiagram voor het Beleidsveld Vervoer en Verkeer*. Rapportnummer Inro/VVG, 1998-02. TNO Inro/Rand Europe & TU Delft, 1998.

(5) Van den Bossche, F, Wets, G., Brijs, T. A Regression Model with ARMA Errors to Investigate the Frequency and Severity of Road Traffic Accidents. *Proceedings of the 83th annual meeting of the Transportation Research Board*. CDROM. Transportation Research Board of the National Academies, Washington, D.C., 2004.

(6) Maerivoet, S. *Modeling Traffic on Motorways: State-of-the-art, Numerical Data Analysis, and Dynamic Traffic Assignment*, Ph.D. Thesis. Katholieke Universiteit Leuven, Department of Electrical Engineering ESAT-SCD (SISTA), 2006.

(7) Bellemans, T. *Traffic Control on Motorways*, Ph.D. Thesis. Katholieke Universiteit Leuven, Department of Electrical Engineering ESAT-SCD (SISTA), 2003.

(8) Neter, J., et al. *Applied Linear Statistical Models*. WCB/McGraw-Hill, 1996.

(9) Yaffee, R.A., McGee, M. *Introduction to Time Series Analysis and Forecasting, with Applications of SAS and SPSS.* Academic Press, San Diego, 2000.

(10) SAS Institute Inc. *SAS/ETS® 9.1 User's Guide.* SAS Institute Inc., 2004, pp. 363-485.

(11) Shumway, R.H., Stoffer, D.S. *Time Series Analysis and Its Applications*. Springer, 2000.

(12) Brocklebank, J,C., Dickey, D.A. *SAS for Forecasting Time Series, 2nd Edition.* SAS Institute, 2003.

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

(14) Makridakis, S., Wheelwright, S. and Hyndman, R. *Forecasting: Methods and Applications – 3rd* edition. John Wiley and Sons, 1998.

(15) Wets, G., Vanhoof, K., Arentze, T. and Timmermans, H. Identifying Decision Structures Underlying Activity Patterns: An Exploration Of Data Mining Algorithms. *Transportation Research Record,* No. 1718, 2000, 1-9.

(16) Moons, E. *Modeling Activity-Diary Data: Complexity or Parsimony?*, Ph.D. Thesis. Limburgs Universitair Centrum, Faculteit Wetenschappen, 2005.

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

(18) Ljung, G. and Box, G. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, Vol. 65, No. 2, 1978, 297-303.

(19) Wijermars, W.A.M., van Berkum, E.C. Analyzing Highway Flow Patterns Using Cluster Analysis. *Proceedings of the 2005 IEEE Intelligent Transportation System Conference (ITSC).* CDROM. Vienna, Austria, 2005

(20) Liu, Z., Sharma, S., Prediction of Directional Design Hourly Volume Based on Statutory Holiday Traffic. *Proceedings of the 85th annual meeting of the Transportation Research Board*. CDROM. Transportation Research Board of the National Academies, Washington, D.C., 2006.