Forecasting Vehicle Kilometers Traveled: Estimating an Autoregressive Integrated Moving Average (ARIMA) Model with Exogenous Variables

Mohsen Nazemi¹, Mohammad-Sajjad Shafiei², Meead Saberi² *, Majid Sarvi²

¹Department of Civil and Environmental Engineering, Tarbiat Modares University, Tehran, Iran
²Institute of Transport Studies, Department of Civil Engineering, Monash University, Melbourne, Australia

ABSTRACT

Vehicle kilometers traveled (VKT or VMT) is a key variable in long-term transportation planning and policy making, especially for road infrastructure investment. In this paper, we estimate an Autoregressive Integrated Moving Average (ARIMA) model with exogenous variables to forecast VKT in Australia. The estimated model produces forecasts based on lagged values in the time series and the errors made by previous predictions, which typically allows the model to rapidly adjust for sudden changes in trend, and at the same time includes exogenous variables resulting in more accurate forecasts. We found that the effects of unemployment rate, fuel price over Gross Domestic Product (GDP) per capita, and Consumer Price Index (CPI) on VKT are all statistically significant and negative. More importantly, number of vehicles per capita appears to be highly statistically significant with positive impact on VKT. We postulate that policies or behavioral changes towards less car ownership are more likely to have greater impact on VKT rather than fluctuations in fuel price, unemployment rate, or individual’s purchasing power.

Keywords: vehicle kilometers traveled (VKT), vehicle miles traveled (VMT), peak car, time series modeling, ARIMAX
1. INTRODUCTION

Long-term transportation planning and policy making requires a proper understanding of future travel demand. Measuring and forecasting traffic growth, in terms of Vehicle Kilometers Traveled (VKT) or Vehicle Miles Traveled (VMT), is a key element in strategic transportation planning process for resource allocation, infrastructure investment, and estimating emissions and fuel consumption. For many years, specifically after World War II, VKT per capita has increased steadily. In 2000s, a sudden change in the trend of VKT per capita, known as “peak car”, occurred for the first time in many developed countries. Similarly in Australia, the total number of privately owned vehicles multiplied and demand for mobility increased significantly during the last few decades and reached its highest point in 2003-2004 (BITRE, 2012).

Focusing on economic and demographic trends, several studies in the literature suggest that a period of decline in vehicle travel may have happened due to various reasons. Existing models for predicting VKT for planning and policy purposes use a variety of factors including urban development patterns, transportation supply, economic activity, socio-demographics, and telecommuting. In study (Cervero and Kockelman, 1997) authors characterized development patterns by Density, Diversity, and Design (3Ds). They used travel diary data and land-use records from 1990’s to examine how these factors affect trip rates. They found that density, land-use diversity, and pedestrian-oriented designs generally reduce trip rates and encourage non-auto travel. Highway expansion and public transportation are also known to be two of the most strongly supply-related variables associated with changes in VKT over the past few decades (Hymel et al., 2010; McIntosh et al., 2014; Noland, 2001).

Other studies (Ewing et al., 2013; Jakobsson et al., 2002; Khan et al., 2014; McMullen and Eckstein, 2012; Zhang et al., 2009) reviewed the effects of economic disincentives on VKT. Fuel price, road tolls, and parking fees along with annual household income, Gross Domestic Product (GDP), and Consumer Price Index (CPI) are known as the common descriptive variables for VKT. Travelers may respond to pricing policies through short-term actions such as shifting to non-auto modes and longer-term actions such as purchasing more fuel-efficient vehicles and/or residential and job relocation. Population and young adult proportion are also commonly used as socio-demographics variables in many VKT studies (Collia et al., 2003; Polzin et al., 2014). Other researches (Choo et al., 2005; Delbosc and Currie, 2014; Polzin et al., 2014) considered more complex social trend changes such as licensing behavior of young adults, cultural trend in having children and the effect of telecommuting in transition from studying and living at home to full-time employment.

Time series data are widely used in VKT studies (Graham-Rowe et al., 2011; Salon et al., 2012). However, very few studies have applied appropriate time series statistical models to forecast VKT. Times series data often have specific statistical features and thus, requires well-developed statistical tests to ensure the reliability and validity of the estimated models and the proposed model-driven policies. The two common approaches in the past were: (a) estimating models with different relating variables to VKT that usually overlook the time dependency of the variables and (b) forecasting a future event purely on the basis of historical information of the same event such as using Autoregressive Moving Average (ARMA) or Autoregressive Integrated Moving Average (ARIMA) modeling (Li et al., 2010). In this paper, we apply a rigorous statistical approach to forecast VKT in Australia. We examine the impact of several exogenous variables on VKT per capita through a multivariate time series analysis using aggregate nationwide data spanning 1965–2012. The study demonstrates that an ARIMA model with exogenous variables,
which is also known as ARIMAX, can identify the underlying patterns of time series data and quantify the impact of various variables on VKT. ARIMAX models produce forecasts based on prior values in the time series and the errors made by previous predictions, which typically allows the model to rapidly adjust for sudden changes in trend, and at the same time includes other exogenous variables resulting in more accurate forecasts. While the inclusion of exogenous variables adds complexity to the model development process, the model can capture the impact of relevant independent factors.

The rest of the paper is organized as follows. The next section describes the data used in this study. The third section discusses the methodology and the overall model-estimation process followed by introducing the applied measures of goodness of fit for model validation. Section four summarizes the model estimation results followed by a discussion and scenario analysis. The last section discusses policy recommendations and suggests future research directions.

2. DATA DESCRIPTION

Aggregate nationwide time series data spanning 1965–2012 are used in this study to examine the possible impact of explanatory variables on VKT per capita in Australia including kilometers travelled by all different modes such as light-duty autos and heavy-duty trucks. Data are provided by the Bureau of Infrastructure, Transport, and Regional Economics (BITRE, 2012). VKT per capita is preferred to VKT as it accounts for population growth or decline providing a better understanding of the underlying behavioral dynamics of car use by individuals.

Figure 1 illustrates Australian VKT and VKT per capita over time. VKT per capita follows a general increasing pattern until 2003 with some minor fluctuations. Analyzing more closely, there has been a sharp increasing phase lasting until 1978. After a rapid growth in the 70’s, traffic growth has consistently slowed down while still gradually rising until 2003. VKT per capita peaked in 2003-2004 and reached its highest point at a level slightly over 7,178 kilometers traveled per person. VKT per capita in 2012 was 6,440, 10.3% smaller than the peak value.

![Fig. 1 Australian VKT and VKT per capita trend (1965-2012)](image-url)
Several key explanatory factors, which frequently appear in VKT forecast models in the literature, can be categorized into two main groups:

- Economic activity: Gross Domestic Product (GDP), real disposable income, employment and unemployment rates, Consumer Price Index (CPI), and real fuel price.
- Socio-demographics: population, household size, number of licensed drivers, number of personal vehicles, and proportion of the population living in suburban areas.

Twelve exogenous variables are selected to include in the model as summarized in Table 1. Figure 2 illustrates the time series trends for the selected variables. Most of the explanatory variables exhibit an increasing trend. The most notable exception is fuel price/GDP per capita. This unit-less variable represents the connection between transportation cost and economic activity and is constructed after numerous model estimations to represent the purchasing power of users that supposedly has a direct impact on private car use. We further discuss this in next section.

Table 1 Statistical summary of selected exogenous variables

<table>
<thead>
<tr>
<th>Description</th>
<th>Unit</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP per capita</td>
<td>(thousand) dollars</td>
<td>18.739</td>
<td>15.461</td>
<td>2.281</td>
<td>67.436</td>
</tr>
<tr>
<td>Annual income</td>
<td>(thousand) dollars</td>
<td>30.545</td>
<td>21.052</td>
<td>3.193</td>
<td>72.436</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>percent</td>
<td>5.85</td>
<td>2.57</td>
<td>1.50</td>
<td>10.87</td>
</tr>
<tr>
<td>Consumer Price Index (CPI)</td>
<td>percent</td>
<td>50.34</td>
<td>30.73</td>
<td>8.69</td>
<td>100.00</td>
</tr>
<tr>
<td>Median housing price</td>
<td>(thousand) dollars</td>
<td>166.920</td>
<td>162.369</td>
<td>9.400</td>
<td>562.500</td>
</tr>
<tr>
<td>Real fuel price</td>
<td>cents/liter</td>
<td>79.87</td>
<td>20.08</td>
<td>55.47</td>
<td>142.90</td>
</tr>
<tr>
<td>Annual fuel budget per person</td>
<td>dollars</td>
<td>443.48</td>
<td>304.22</td>
<td>31.96</td>
<td>993.22</td>
</tr>
<tr>
<td>Australian population</td>
<td>(million) people</td>
<td>16.922</td>
<td>3.167</td>
<td>11.784</td>
<td>22.918</td>
</tr>
<tr>
<td>Young adult proportion (18-34)</td>
<td>percent</td>
<td>25.83</td>
<td>1.76</td>
<td>23.23</td>
<td>28.52</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>(million) vehicles</td>
<td>9.690</td>
<td>3.715</td>
<td>3.739</td>
<td>16.900</td>
</tr>
<tr>
<td>Number of vehicles per capita</td>
<td>percent</td>
<td>55.10</td>
<td>11.86</td>
<td>31.70</td>
<td>73.80</td>
</tr>
<tr>
<td>Fuel price/GDP per capita</td>
<td>-</td>
<td>0.008</td>
<td>0.007</td>
<td>0.002</td>
<td>0.026</td>
</tr>
</tbody>
</table>
3. METHODOLOGY

In this study, we use the widely practiced (Box and Jenkins, 1976) approach for time series analysis to obtain the most parsimonious model that is still an adequate representation of the data to yield a comprehensive model for forecasting purposes. Box and Jenkins approach consists of three steps: identification, estimation, and diagnostic checking. Identification involves investigating a tentative formulation for the model as a starting point with the specific objective of obtaining the initial estimates of the required parameters used in the next steps. These parameters are initially suggested by patterns either in the autocorrelation function and partial autocorrelation function of the raw series itself, or in the residuals from a previously estimated model, while judgment is exercised. The output model of this step leads to application of more formal and efficient parameter estimation methods in the estimation step via least-squares, maximum likelihood, or other approaches using one among several special-purpose routines devoted to time series analysis. After estimating the parameters, the fitted model will be subject to diagnostic checks of the residuals and relevant tests of goodness of fit to examine if there are any autocorrelation or time-dependent trends left that indicate an incorrect or incomplete specification. The three-step process is repeated until such pattern is achieved. Although model-building, diagnostic checking, and model correction occurs in a sequence, generally all parameters from earlier steps are re-estimated simultaneously with parameters related to the current step. This makes the most efficient use of the data, and allows all parameters to be estimated as precisely (with the greatest confidence) as possible (Box and Jenkins, 1976).

The time series models allow for the inclusion of external predictor variables to explain some of the historical variation of the dependent variable along with subtle time series dynamics that can be handled with ARIMA technique to produce more accurate forecasts (Andrews and Dean, 2013). These models that conceptually combine regression models and ARIMA models are called regression models with ARMA errors or ARIMAX models. ARIMAX is an autoregressive integrated moving-average model with exogenous variables. “Integration” in this context means that the time series has first been transformed by differencing. It is a logical extension of pure ARIMA modeling that incorporates independent variables. When the AR and MA terms in an ARIMA model are not sufficient to provide an acceptable high goodness of fit, it is reasonable to search for relevant explanatory variables whose influence over time is not sufficiently embedded in the historical values of the dependent time series. In other words, an ARIMAX model can be viewed as a multiple regression model with autoregressive (AR) and moving average (MA) terms; where AR terms are merely statistically significant lagged values of the dependent variable, and MA terms are residuals (i.e., lagged errors) resulting from previously produced estimates. Therefore, an ARIMAX model merges both the trailing time series values ($Y_t$) and the trailing model errors ($\epsilon_t$) with exogenous variables.

Here, we consider an original (undifferenced, $d = 0$) ARIMAX model with $p$ autoregressive term and $q$ moving average term in the following form:

$$\phi(L)Y_t = \phi(L)(\delta + \sum \beta_l X_{lt}) + \theta(L)\epsilon_t$$

where
t represents time,
$\delta$ is a constant,
$X_{lt}$ is the $i$th exogenous variable at time $t$,
L is the lag (backshift) operator; that is \( LX_t = X_{t-1} \), 
\( \phi(L) \) is the autoregressive operator, represented as a polynomial in the lag operator; that is  
\[ \phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p, \]
\( \theta(L) \) is the autoregressive operator, represented as a polynomial in the lag operator; that is  
\[ \theta(L) = 1 - \theta_1 L - \cdots - \theta_q L^q, \]
and  
\( \beta_i \) is coefficient of the \( i \)th variable.

An important consideration in ARIMAX model is that the dependent variable and all the predictor variables in the model must first be stationary, before model building begins. Otherwise, the estimated coefficients are incorrect. Practically speaking, a stationary process has a mean and variance that do not change over time, and the process has no trends. In other words, non-stationarity may occur from the presence of outliers, random walk, drift, trend, or changing variance and series must be examined in order to ascertain whether any of these phenomena inhere within the series.

Following the assessment of model fit and predictive validation, the best model is chosen to forecast over the forecast horizon. To forecast a regression model with ARIMA errors, both the regression part of the model and the ARIMA part of the model should be forecasted, and then combined. This requires a prior forecast or assumption for the unknown predictors extended in the forecast horizon.

4. ESTIMATION RESULTS

This section presents the results of the VKT per capita modeling using multivariate time series analysis described in the previous section. The econometric software package SAS 9.4 was used to estimate the ARIMAX models and to conduct relative statistical tests to validate the results. There are several statistical assumptions that must be examined in the classic Box and Jenkins methodology to ensure that the resulting model is valid at each stage of its evolution. The first step in analyzing a series is to make sure it is stationary. The validity and reliability of the results lies at the root of this assumption. The data used in the current study were non-stationary in its raw form as seen earlier in Figure 2. Differencing as a common transformation method was applied to the raw series. The first-order differenced form of the series, \( d = 1 \), achieved stationarity. It is often desirable to maintain the form of the relationship between the dependent variable and the predictors to preserve interpretability. Consequently, it is common to apply the same differencing scheme to all exogenous variables. This leads to a model called “model in differences” in comparison to a “model in levels” which is observed when the original data are used without differencing.

A common test to evaluate the degree of stationarity of series is the Augmented Dickey–Fuller (ADF) unit root test. Table 2 summarizes the results of ADF test conducted on the dependent variable, VKT per capita.

<table>
<thead>
<tr>
<th>Type</th>
<th>Lags</th>
<th>Rho</th>
<th>Pr &lt; Rho</th>
<th>Tau</th>
<th>Pr &lt; Tau</th>
<th>F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero Mean</td>
<td>1</td>
<td>-11.9059</td>
<td>0.0129</td>
<td>-2.5073</td>
<td>0.0132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Mean</td>
<td>1</td>
<td>-15.8437</td>
<td>0.0209</td>
<td>-2.7521</td>
<td>0.0732</td>
<td>3.8363</td>
<td>0.1253</td>
</tr>
<tr>
<td>Trend</td>
<td>1</td>
<td>-57.4620</td>
<td>&lt;0.0001</td>
<td>-5.0409</td>
<td>0.0009</td>
<td>12.7141</td>
<td>&lt;0.0010</td>
</tr>
</tbody>
</table>
Once a time series is statistically identified stationary, time series model estimation begins. In the identification step, we check Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF), to gain insight into the nature of the serial correlation. Furthermore, $p$ and $q$ values are determined by inspecting the graphs of the appropriately differenced series ($d = 1$ in current study). If ACF and PACF graphs cut-off after a few lags, the last lag with a large value suggests the initial values of $p$ and $q$. Figure 3 depicts the ACF and PACF correlograms of the transformed series. Based on visual inspection of the correlogram, ACF shows the sudden drop-off after first lag and the values of PACF are all in the non-significant area. It is worth mentioning that several other combinations for $p$, $q$, and $d$ were adopted on a trial and error basis which led to inferior results, compared to the final model presented herein.

![Fig. 3 (a) ACF and (b) PACF correlograms (d = 1)](image)

One of the early tasks in ARIMAX model estimation is to identify and preliminarily evaluate the logical/statistical attractiveness of exogenous variables. The selected exogenous variables display a significant relationship with the dependent variable. The forward/backward stepwise procedure provides an iterative approach to ARIMAX model estimation that both adds significant variables to the model and removes variables from the model that are insignificant. The final model fitted to the data includes statistically significant variables producing superior goodness of fit measures, as summarized in Table 3. The rest of the exogenous variables are excluded in the model building process due to statistical insignificance. The estimated coefficients for the variables are significantly different than zero, as judged by their t-statistics ($p$-values) while the regression residuals are white noise.

Table 3 provides the coefficients, t-statistics, and $p$-values of the best-fitted ARIMAX model. This model consists of four explanatory variables, one AR term, and a constant value. The Akaike’s Information Criterion (AIC) equals to 551.2 and the Durbin-Watson test statistic equals to 2.01, which suggests that the model is not auto-correlated. Moreover, other statistical tests were conducted to ensure that the model satisfies all the conditions required.

| Parameter                      | Estimate  | Standard Error | t value | Approx. Pr > | | Lag |
|--------------------------------|-----------|----------------|---------|---------------|---|
| MU                             | 39.32913  | 48.6983        | 0.81    | 0.4240        | 0 |
| AR1.1 ($p = 1$)                | 0.58831   | 0.1275         | 4.61    | <0.0001       | 1 |
| Number of vehicles per capita | 95.01454  | 30.77918       | 3.09    | 0.0036        | 0 |
| Unemployment rate              | -26.74465 | 13.20607       | -2.03   | 0.0494        | 0 |
| Fuel price/GDP per capita      | -27.30581 | 15.66144       | -1.74   | 0.0887        | 0 |
| Consumer Price Index (CPI)     | -33.30709 | 14.74818       | -2.26   | 0.0293        | 0 |
The estimated coefficients all have the expected sign. A negative sign for unemployment rate complies with the fact that more jobs increases the number of daily trips which consequently increases the VKT per capita. Changes in CPI are used to assess price changes associated with the cost of living. Therefore, an increase in CPI has the converse effect on VKT per capita. Fuel price/GDP per capita is statistically significant with a logical negative sign. Number of vehicles per capita, representing car ownership rate, is highly significant and appears with a positive sign; which is in complete agreement with the idea that higher number of vehicles per person induces a greater possibility to drive. This is consistent with several studies in the literature suggesting that younger generation is less likely to own a private vehicle and thus, less likely to drive a car.

In the diagnostic checking step, the residual series must not exhibit significant serial correlation, otherwise appropriate AR and MA terms in accordance with ACF and PACF correlograms should be added to assure the residuals are white noise. In addition, it should be noted that proper interpretation of the significance levels (p-values) of regression model coefficients requires that the residuals produced by the model under scrutiny are normally distributed. The residuals should be normally distributed so that the t-statistics used to assess the significance of variable terms are valid. Figure 4 depicts the ACF and PACF plots of the residuals. The correlograms reveal no significant autocorrelation in the residuals. The points in the Q–Q plot approximately lie on the straight line, which suggests that the residuals follow a normal distribution.

4. DISCUSSION

Based on the estimated model, number of vehicles per capita plays a significant role in changes in VKT per capita in comparison with other variables. Here, we examine two different scenarios to forecast traffic growth in Australia. In the first scenario, the forecast is made based on a resumption of growth in traffic on Australian roads under the “business-as-usual” condition. It is
assumed that all of the exogenous variables follow their prevailing trend by forecasting four exogenous variables’ values from 2013 to 2020 with pure ARIMA models. The second scenario considers the same assumption except for the number of vehicles per capita to remain constant during the selected period. Figure 5 illustrates VKT per capita forecasts from 2013 to 2020 for both scenarios. The results are based on the 95% confidence interval for all the variables. The first scenario predicts an upward trend during the upcoming years while the second scenario maintains a declining tendency in the forecast period. In the first scenario, VKT per capita reaches 6,838 kilometers per person in 2020 compared to 6,166 kilometers per person in the second scenario, showing a considerable reduction which reveals the influence of car ownership.

5. CONCLUSION

In this paper, we estimated a time series model of VKT per capita in Australia incorporating multiple exogenous variables. A total of twelve explanatory variables were included in the model. Four significant exogenous variables remained in the final model including number of vehicles per capita, unemployment rate, fuel price/GDP per capita, and Consumer Price Index (CPI). The coefficient for the number of vehicles per capita has a positive sign while unemployment rate, fuel price/GDP per capita, and CPI are estimated to have coefficients with negative signs, as expected.

The estimated model demonstrates that although fuel price/GDP per capita, representing individual’s fuel budget, is statistically significant; however, its variation contributes limited alteration to VKT per capita. This suggests that fuel price is not the direct and main determinant of VKT per capita in Australia. However, car ownership was found to have larger statistically significant impact on VKT. Therefore, we postulate that policies or behavioral changes towards less car ownership are more likely to have greater impact on VKT rather than changes in fuel price, unemployment rate, or individual’s purchasing power.

For future research, once could apply more advanced methods such as Artificial Neural Networks (ANN) to model VKT per capita and compare the results with other modeling
techniques. Including transportation supply and urban form variables associated with VKT in the models could also provide further insights.

REFERENCES


Bureau of Infrastructure Transport and Regional Economics (BITRE), 2012. Traffic Growth in Australia, Report 127, Canberra ACT.

Box, G., Jenkins, G., 1976. Time series analysis, control, and forecasting. San Fr. CA Holden Day.


