Day on day dependencies in travel
First results using ARIMA modelling

Author(s): Fraschini, Elena; Axhausen, Kay W.

Publication Date: 2001-02

Permanent Link: https://doi.org/10.3929/ethz-a-004232828

Rights / License: In Copyright - Non-Commercial Use Permitted
Day on Day Dependencies in Travel: First Results using ARIMA Modelling

E. M. Fraschini
K.W. Axhausen

Arbeitsbericht
Verkehrs- und Raumplanung

February 2001
Day on Day Dependencies in Travel: First Results using ARIMA Modelling

E. M. Fraschini and K. W. Axhausen
IVT
ETH
CH-8093 Zürich
Telephone: +41-1-633 39 43
Fax: +41-1-633 10 57
E-Mail: axhausen@ivt.baug.ethz.ch

February 2001

Abstract

In this paper, we have used a technique for the analysis of a 42-day continuos record of travel behaviour data that has not been used before for travel diary and transport planning data. The travel diary data of the MobIDrive dataset has been examined from a different perspective with the use of a specific class of stochastic models called ARIMA (Box and Jenkins, 1976), which have been used to extrapolate day on day correlations and weekly patterns from the data collected. An ARIMA modelling approach has been conducted on an individual basis for each subject of the study and for a set of variables related to trips, distances and durations of daily activities. This approach has allowed a preliminary classification of the subjects of the MobIDrive study into different model categories, and can be used as a starting point or a more detailed categorisation for the application of different analysis techniques.

Keywords

Time series analysis – Travel Diary – ARIMA Models – ETH Zürich – Institut für Verkehrsplanung und Transporttechnik, Strassen- und Eisenbahnbau (IVT)
1. Introduction

In this paper, we will make use of a special technique for analysing time series data that has not been used for travel diary data before but rather for business, economics, industrial, and natural sciences data.

The MobiDrive dataset records 42 days of various items of information about the travel behaviour of a sample of respondents. It is therefore possible to consider a sequence of observations for a single individual as a time series. In fact, a *time series* is a sequence of measurements of a variable taken at equally spaced time intervals (hours, days, weeks, months...) where successive values are dependent on each other. The investigation of this dependence and the creation of a mathematical models that can adequately describe it, are therefore the main goals of this kind of approach in order to:

(a) Identify the nature of the phenomenon represented by the observations  
(b) Forecast, that is predicting future values of the same variable or fill in missing data  
(c) Determine the effect of external interventions on the system

For this purpose, it is necessary to identify first and properly describe the *pattern* of the time series, so that it can be interpreted and possibly integrated with other data. Once the pattern is established, it is then possible to extrapolate it and predict future events or identify the events that influence its behaviour, with more or less confidence depending on the validity of the pattern interpretation and of our understanding of the time series.

Within the information collected during the MobiDrive project, a number of variables have been selected to implement this analysis, and they are

- Number of person trips
- Number of leisure related trips
- Number of person journeys
- Total daily out-of-home activity duration
- Total daily trip duration
• Total daily distances

The length of the time series is of great importance since it affects the reliability of the pattern identification and the efficiency of the parameter estimates. In the case of the MobiDrive data, the length of the study corresponds to six weeks of continuous data (42 consecutive daily observations). With this length it is necessary to understand that some of the parameters and functions calculated during the analysis must be carefully interpreted, and evaluated mainly for their relative magnitude rather than their absolute value.

However, it has been possible to extract interesting information from the travel diary data of the MobiDrive Study using this alternative approach for data analysis, and the initial results suggest that for certain categories of persons this method could give significantly more insight into their behaviour, than currently used method.

2. Time Series Analysis

The idea behind time series analysis is Wold's decomposition theorem (Wold, 1938), which proves that any stationary time series can be decomposed into two different components, one self-deterministic (that is related to the memory of the process) and one stochastic (that is a random and uncorrelated component, which represents the unpredictable part of any process and usually makes the pattern difficult to identify).

This decomposition is performed by stochastic models that are used to estimate the probability that the next observation of the variable will be located between two specified limits, which characterise the confidence or the reliability of the forecast value.
One of the main concerns in the choice of the model form is the *stationarity* of the series itself. A series is considered stationary if its statistics do not vary over time, that is, if the series has a constant mean over time. Non-stationarity can happen both because of a general growing trend over time (e.g. the constantly growing number of airline passengers per year) or because of a periodic behaviour of the observed variables (e.g. a weekly travel behaviour or a yearly cycle of mean water temperatures).

It is therefore necessary to choose a special class of models that is able to deal with the problem of non-stationarity and among the class of non-stationary models, we have chosen the family of ARIMA (AutoRegressive Integrated Moving-Average) models (Box and Jenkins, 1976).

### 2.1 ARIMA Models

ARIMA Models have initially been developed for the use in the fields of business, industry, and economics and have been later applied to all kinds of time series. ARIMA is an acronym for *autoregressive integrated moving-average*. The acronym indicates that the model is composed of different terms, an autoregressive term (AR), a moving-average term (MA), and an integrating term (I) that accounts for the non-stationarity of the time series. The *order* of an ARIMA model represents the number of parameters included in each of the different components. Hence a model is usually indicated as ARIMA ($p, d, q$), where $p$ is the order of the
autoregressive term, $d$ is the order of the integrating or differencing term, and $q$ is the order of the moving-average term.

The different components of the ARIMA model are described below.

### 2.1.1 Auto regressive process

The value of the variable at the time $t$ is expressed as a linear combination of $p$ previous values of the same variable (observations that have occurred previously to time $t$), and a random value $\varepsilon_t$, also called white noise process given its assumed statistical properties (a sequence of uncorrelated random variables, each with zero mean and $\sigma^2$ variance). The autoregressive model can be written as:

$$\phi(B)z_t = \varepsilon_t + \mu$$

Where

- $z_t$ is the value of the variable at time $t$;
- $B$ is the backward shift operator, that is: $B z_t = z_{t-1}$;

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots B^p$$ is defined as the autoregressive operator of order $p$ and

$\phi_1, \phi_2, \ldots, \phi_p$ are the $p$ parameters of the autoregressive model;

- $\varepsilon_t$ is the white noise process;

$$\mu = E[z_t]$$ is the average of the process $z_t$.

The autoregressive model contains $p+2$ parameters (the average $\mu$, the $p$ autoregressive parameters, and the variance of the error term) that must be estimated from the data.
An autoregressive process can be stationary or not stationary. The condition for stationarity is that the roots of the autoregressive operator \( \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_n B^n = 0 \), considered as a polynomial in \( B \), must lie outside the unit circle.

### 2.1.2 Moving-average process

The moving-average process similarly consists of a linear combination of random values or shocks \( \varepsilon_t \), and can be written as

\[
z_t = \theta(B) \varepsilon_t + \mu.
\]

Where:

- \( z_t \) is the value of the variable at time \( t \);
- \( B \) is the backward shift operator, that is: \( B \varepsilon_t = \varepsilon_{t-1} \);
- \( \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \) is the *moving average operator* of order \( q \), and \( \theta_1, \theta_2, \ldots, \theta_q \) are the \( q \) parameters of the linear combination of \( \varepsilon_t \) values;
- \( \varepsilon_t \) is the white noise process;
- \( \mu = E[z_t] \) is the average of the process \( z_t \).

The moving average process contains \( q+2 \) parameters (the average \( \mu \), the \( q \) autoregressive parameters, and the variance of the error term) that must be estimated from the data.
2.1.3 Autoregressive moving-average

In order to better understand the time series pattern, and obtain a better fit for the model, it is sometimes effective to adopt mixed models that incorporate both an autoregressive and a moving-average component:

\[
\phi(B) z_t = \theta(B) \varepsilon_t + \mu
\]

which contains \(p+q+2\) unknown parameters (the \(p\) autoregressive parameters, the \(q\) moving average parameters, the average \(\mu\), and the variance of the error term).

2.1.4 Non Stationarity

It often happens that a time series shows a non-stationary or a seasonal behaviour, or even a combination of the two (Figure 2). Non-stationary behaviour can be taken into account by making use of a different autoregressive operator \(\phi(B)\): this operator is constructed so that a number \(d\) (\(\leq p\)) of the roots of the polynomial \(\phi(B) = 0\) lie on the unit circle. The operator can be written as:

\[
\phi(B) = \phi(B)(1 - B)^d.
\]

Where \(\phi(B) = 1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p\) is the original stationary autoregressive operator whose roots are greater than one in their absolute value, and \(B\) is the backward shift operator.

An ARIMA model that can represent non-stationary behaviour can therefore be written as:

\[
\phi(B) z_t = \phi(B)(1 - B)^d z_t = \theta(B) \varepsilon_t.
\]

Or better:

\[
\phi(B) w_t = \theta(B) \varepsilon_t.
\]
With the introduction of the *backward difference operator* $\nabla = (1 - B)$, we can then write $w_t = \nabla^d z_t = (1 - B)^d z_t$.

Thus, the non-stationarity of the series is translated into the hypothesis that the $d^{th}$ difference of the data series can be represented by a stationary process. Generally, the differencing operator is utilised as any times as it is necessary to turn the series into a stationary one. In practical cases, it is not necessary to differentiate the series more than 1 or 2 times, therefore it is often $w_t = \nabla z_t = z_t - z_{t-1}$.

Figure 1 shows an example 42-day data series from the MobiDrive dataset for the variable total out-of-home daily activity duration. The presence of a linear trend component is obvious, and emphasised by the trendline in Figure 1a, while Figure 1b shows the same series after the removal of the trend component by means of a single differencing. The series can be now dealt with as a stationary time series.

Figure 2 Total daily activity duration data (City of Halle, household number 194, and person number 2): (a) original data series, (b) data series after removing the linear trend

If the non-stationarity of the series is given by a seasonal pattern (e.g. a yearly, monthly or weekly cycle), it will then be necessary to difference the series at a lag corresponding to the length of the cycle of seasons. This is the case for many individuals of the MobiDrive dataset, where the weekly cycle is clearly discernible (Figure 3) and needs to be taken into account with a 7-day lag seasonal differencing:
\[ w_t = z_t - z_{t-\gamma}. \]

It must be said, however, that the differencing operation eliminates a number of observation corresponding to the differencing order, and therefore further reduces the length of the data series available for the estimation of the parameters. Hence, the weekly differencing has been performed only on datasets without any missing data to avoid the loss of precious information, and if the trend or seasonal effect was very regular, we have considered the introduction of other explanatory variables as an alternative to differencing.

Figure 3  Total daily activity duration data: (a) original data series with 7-day cycle, (b) data series after the removal of the weekly non-stationary component.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>500</td>
<td>-500</td>
</tr>
<tr>
<td>14</td>
<td>1000</td>
<td>-250</td>
</tr>
<tr>
<td>21</td>
<td>1500</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>35</td>
<td>2500</td>
<td>1000</td>
</tr>
<tr>
<td>42</td>
<td>3000</td>
<td>1500</td>
</tr>
</tbody>
</table>

2.1.5  Model Identification

The first step in the construction of an ARIMA model is the choice of the orders of its different components and the identification of the order of differencing.

For this latter aspect, the 361 individual data series of the MobiDrive dataset have been consequently analysed for stationarity, and a check for the presence of a seasonal pattern has been performed.

The choice of the orders of the AR and MA parts requires a detailed analysis of the Autocorrelation Function and of the Partial Autocorrelation Function (ACF and PACF, respectively), whose shape and value will determine the order of the model.
The autocorrelation function is defined as the correlation between \( z_t \) and \( z_{t+k} \):

\[
\rho(h) = \text{Corr}(z_t, z_{t+k})
\]

and can be estimated at different values of the lag \( k \) from the observations of the time series.

The partial autocorrelation function at lag \( k = 1, 2, 3, \ldots \) is the autocorrelation between \( z_t \) and \( z_{t-k} \) calculated for different lags \( k \), conditional upon the given correlations of \( z_t \) with \( z_{t-1}, \ldots, z_{t-k+1} \). In other words, if the time series \( z_t \) is approximated with different AR processes of order \( k \):

\[
z_t = \phi_{11} z_{t-1} + \epsilon_t,
\]

\[
z_t = \phi_{21} z_{t-1} + \phi_{22} z_{t-2} + \epsilon_t,
\]

\[
z_t = \phi_{31} z_{t-1} + \phi_{32} z_{t-2} + \phi_{33} z_{t-3} + \epsilon_t,
\]

\[
\vdots
\]

\[
z_t = \phi_{k1} z_{t-1} + \phi_{k2} z_{t-2} + \ldots + \phi_{kk} z_{t-k} + \epsilon_t,
\]

then the partial autocorrelations are the last coefficient in each of these autoregressive processes \( \phi_{11}, \phi_{22}, \ldots, \phi_{kk} \). There exist various algorithms, not discussed here, for computing the partial autocorrelation based on the sample autocorrelations among which the Yule-Walker equations:

\[
\begin{bmatrix}
1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\
\rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\
\rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1
\end{bmatrix}
\begin{bmatrix}
\phi_{k1} \\
\phi_{k2} \\
\phi_{k3} \\
\vdots \\
\phi_{kk}
\end{bmatrix}
=
\begin{bmatrix}
\rho_1 \\
\rho_2 \\
\rho_3 \\
\vdots \\
\rho_k
\end{bmatrix}
\]

which, solved for \( k = 1, 2, 3, \ldots \), lead to the formulation of the autocorrelation functions \( \phi_{11}, \phi_{22}, \ldots, \phi_{kk} \). See Box and Jenkins (1976) or Brockwell and Davis (1987) for the mathematical details of this and other procedures.
The autocorrelation and the partial autocorrelation functions serve as a useful indicator of the correlation or of the dependence between the values of the time series at different time steps. Their shape and value is strictly related to the nature of the process and a comparison with the behaviour of the theoretical functions described in the literature allows one to create a set of rules, which should be considered in the choice of the orders of the model (Figure 4).

So for example, the behaviour of the theoretical PACF of a pure autoregressive process of order $p$ (i.e. $q=0$) is significantly different from zero for all orders up to $p$, while the theoretical ACF tails off geometrically. For a pure moving average process of order $q$ (i.e. $p=0$), the PACF tails off geometrically, while the ACF has a cut-off after order $q$. For a mixed process ARMA $(p, q)$ both the ACF and PACF tail off respectively after lag $q+1$ and after lag $p-q$.

Figure 4  Autocorrelation and partial autocorrelation functions $\rho_k$ and $\phi_{kk}$ for various ARMA (1,1) models.

Source: Box et al. (1994), 82.

However, using a finite time series, we have only a finite data set available, and we can thus only obtain estimates of the ACF and PACF. The quality of these estimates is strongly asso-
associated with the sample size and they often do not reproduce theoretical behaviours in detail, resulting in irregular patterns that cannot be easily identified. The ideal course of action would be a trial and error method where one model is chosen, identified and tested, and if the performance is not satisfactory then some parameters are added or removed in order to “tune” the model with the data. Since the MobiDrives dataset consists of 360 data series, and an analysis of each one of them is not the purpose of this study, we have decided to use a specific automated method to tentatively identify the orders of the model. The SCAN method (Smallest CANonical correlation method) can identify the orders of a stationary or non-stationary ARIMA process based on the calculation of the eigenvalues of the correlation matrix of the ARIMA process. A SCAN table is then constructed with penalty functions assigned to the various orders of the AR and MA processes and the order of the model is chosen according to the value of these functions. For each time series, the method provides a choice of up to three different models. The choice among them has been determined by parsimony that is the model chosen had always a minimum number of parameters. This technique has been proposed by Tsay and Tiao (1985) and details about the algorithm can be found in Box et al. (1994) and Choi (1990).

2.1.6 Model Estimation

The number of parameters to be estimated corresponds to the orders of the model. For an ARIMA \((p, d, q)\), one would have to estimate \(p+q+2\) parameters corresponding to the \(p\) parameters of the AR process, the \(q\) parameters of the MA process, the constant term \(\mu\), and the variance of the error term \(\sigma^2\). These parameters are estimated using specific methods such as the Maximum Likelihood Method or the Conditional/Unconditional Least Squares Method. Without going into the details of these methods (see Box at al. (1994)), it is enough to say that the parameters of the model are estimated based on an initial guess and the minimisation of the total sum of squared error terms (that is the difference between the observed value of the variable \(z_i\) and the value \(\hat{z}_i\), calculated by the model with the estimated parameters):

\[
\min \sum \varepsilon_i^2 = \min \sum (z_i - \hat{z}_i)^2.
\]
2.1.7 Diagnostic checks

Once the parameters of the model have been estimated, it is necessary to decide whether this model is satisfactory, that is testing how good the model fits the observed data. This process includes testing if the statistics of the data series are correctly represented, if a simpler or a more complex model provide a better description of the process, and, most important, if the series of the residuals is uncorrelated.

The techniques used in model checking are not different from those used in model identification. The initial hypothesis of the random noise series being a white noise is tested by way of calculation of the ACF and PACF of the residual series $\hat{e}_t$. If this hypothesis is rejected then the model has to be modified in order to find a more appropriate one.

3. Results

The ARIMA modelling approach has been applied to the Mobi\textit{Drive} dataset. For each of the variable listed in Section 1, one ARIMA model has been estimated and tested for each subject and a tentative initial forecast has been conducted for some subjects where the estimated model showed a good performance. The identification and estimation of the ARIMA models has been performed by the SAS System for Windows (SAS, 1999).

The hypothesis that the time series exhibits non-stationary or seasonal pattern has been tested by means of the Dickey-Fuller test. This test is based on the fact that the autoregressive operator $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 0$, considered as a polynomial in $B$, will exhibit a unit root if the time series is not stationary, therefore the hypothesis of the existence of a unit root at lag $d$ has been tested against the alternative of stationarity (refer to Dickey and Fuller (1979) and Dickey et al. (1984) for details about the test algorithm). Table 1 illustrates the results of the testing for the different variables and indicates the absence of a linear trend in the time series (Axhausen et al. (2000) also report about the absence of reporting fatigue), but makes clear that the 7-day cycle must be considered in the modelling procedure.
<table>
<thead>
<tr>
<th></th>
<th>Lag 1</th>
<th></th>
<th>Lag 7</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>[%]</td>
<td>No.</td>
<td>[%]</td>
</tr>
<tr>
<td>Number of journeys</td>
<td>1</td>
<td>0.3%</td>
<td>273</td>
<td>75.8%</td>
</tr>
<tr>
<td>Number of trips</td>
<td>2</td>
<td>0.5%</td>
<td>275</td>
<td>76.3%</td>
</tr>
<tr>
<td>Number of leisure related trips</td>
<td>1</td>
<td>0.3%</td>
<td>258</td>
<td>71.6%</td>
</tr>
<tr>
<td>Total daily out-of-home act. duration</td>
<td>3</td>
<td>0.8%</td>
<td>280</td>
<td>77.7%</td>
</tr>
<tr>
<td>Total daily trip duration</td>
<td>2</td>
<td>0.5%</td>
<td>269</td>
<td>74.2%</td>
</tr>
<tr>
<td>Total daily distance</td>
<td>4</td>
<td>1.1%</td>
<td>255</td>
<td>70.8%</td>
</tr>
</tbody>
</table>

Initially, the subjects have been grouped according to the model category identified, independent of the order of the model itself:

- Simple autoregressive models (AR (p), no moving-average component),
- Simple moving-average models (MA (q), no autoregressive component),
- Mixed autoregressive moving-average models (ARMA (p, q), both components are present),
- White noise signal or random behaviour (WN, subjects where a model could not be identified with a reasonable confidence, and the observations did not reveal any pattern according to the ARIMA classification).

The next step tried to understand this subdivision and extract additional information. Figure 5 shows for instance the subdivision of subjects in different model categories. It is evident, that for both variables number of trips (Fig. 5a) and number of leisure related trips (Fig. 5b) the difference in frequency between males and females is minimal and that the sex of the subjects does not give supplementary information about the model classification.

On the other hand, there is, regardless of sex, a relatively high frequency of subjects categorised into white noise signals (Fig. 5a and 5b). This could be explained by the non stationarity of the time series, and a quick look at Fig. 6 shows that once the seasonal differencing is introduced, the relative distribution within the groups is quite different and that the removal of the weekly cycle allows one to identify an autoregressive pattern in the time series. The results for the model category classification for all the variables are reported in Table 2.
Figure 5  Frequency of distribution of female and male subjects within the different model categories: (a) number of trips, (b) number of leisure related trips.

(a) Number of trips

(b) Number of leisure related trips

Table 2  Model type distribution for all the variables and for Lag-1 and Lag-7 differencing lags, number of models estimated for each category

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Not differenced</th>
<th>Differenced Lag-1</th>
<th>Differenced Lag-7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of journeys</td>
<td>AR</td>
<td>81</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>28</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>205</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Number of trips</td>
<td>AR</td>
<td>85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>35</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>201</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Number of leisure related trips</td>
<td>AR</td>
<td>81</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>210</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total daily out-of-home activities duration</td>
<td>AR</td>
<td>79</td>
<td>74</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>57</td>
<td>156</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>17</td>
<td>61</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>205</td>
<td>9</td>
<td>148</td>
</tr>
<tr>
<td>Total daily trips duration</td>
<td>AR</td>
<td>74</td>
<td>121</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>23</td>
<td>177</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>35</td>
<td>56</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>226</td>
<td>4</td>
<td>177</td>
</tr>
<tr>
<td>Total daily distance</td>
<td>AR</td>
<td>68</td>
<td>88</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>MA</td>
<td>30</td>
<td>209</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>ARMA</td>
<td>29</td>
<td>51</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>WN</td>
<td>231</td>
<td>10</td>
<td>181</td>
</tr>
</tbody>
</table>
Once the individuals have been classified according to the model estimated, it was interesting to observe the distribution of the different kind of models for the same individual across similar variables. Figure 7 shows how the individuals are distributed in the model classes space, that is how many subjects present which combination of models for a specific set of variables. It is evident that the interrelationships are very similar for the two different variables groups, with either one of the models classified as white noise for a big share of the sample, and with both models classified as white noise for more that 40% of the sample.

This could be explained by the fact that more than the variable type, it is the person’s behaviour that is somehow uniform and that, independently from the variable used to monitor it,
(trips, journeys, distances or durations) it can be characterised by a certain model type, which then remains constant across the variables. The same pattern can be observed in Table 3, where the combination of the models across the three trip and journey related variables (number of trips, number of journeys, and number of leisure related trips) is listed. The percentage of subjects with at least two models of the same kind in the tuplet represents more than 70% of the total.

Table 3 Model type distribution across the trip and journey related variables: number of trips, number of journeys and number of leisure related trips

<table>
<thead>
<tr>
<th>Type of model</th>
<th>Share of sample [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 equal models</td>
<td>29%</td>
</tr>
<tr>
<td>2 equal models</td>
<td>57%</td>
</tr>
<tr>
<td>3 different models</td>
<td>14%</td>
</tr>
</tbody>
</table>

Given these results, a possible development of this kind of analysis would be to try to identify if one or more socio-demographic variables (such as age, income, sex, number of household members, type of employment and so on) that could somehow explain the variability and the subdivision of different classes of subjects into different model categories, and therefore characterise the subjects which belong to the same group.

An initial analysis for different age classes, showed that the different model categories are indeed associated with the different ages. Figure 8 shows the results for the variable number of leisure related trips. The category AR and ARMA are present with a higher frequency in the age classes between 35 and 65 years old. This is an indication of a more regular behaviour, and it is probably linked with a fixed employment and a more settled lifestyle. On the other hand, MA models are mainly identified for younger subjects, indicating the presence of external influences on the subject behaviour. The MA models try indeed to explain the observed pattern of the time series with a linear combination of random signals and suitable weights, and this approach is necessary every time the correlation of the series from one day to the other is not enough to define a suitable AR model. Younger subjects could be influenced by the behaviour of other household members, or for example by fixed commitments or by the weather. The distribution of the WN model, on the other hand, does not seem to be strongly influenced by the age and but it is more or less regularly distributed over the different age classes.

A similar pattern exists for the number of trips and journeys, while for the distance and duration variables (total daily out-of-home activities duration and total daily trips duration) the
major part of younger subjects fall into the AR and ARMA classes, perhaps because of presence of the regular weekly school activities. The same behaviour is observed for the lag-7 differenced variables.

Figure 8 Frequency of distribution of different model categories within the different age classes: number of leisure related trips.

For each individual of the dataset and each of the variables considered in this study, one ARIMA model has been estimated. It has not been possible to perform a proper validation of the models due to the short data series available for the estimation of the parameters, but for some of the subjects an initial tentative forecast has been performed for a 7-day and 14-day period. The estimation of the model parameters has been performed by means of the Maximum Likelihood method, and due to the short time series available, the method did not always converge to a stable solution. To give an idea of the figures, Table 4 and Table 5 summarise the model parameters and their statistics for some of the models correctly estimated, and Figure 9 shows one example of a time series model and forecasts.
<table>
<thead>
<tr>
<th>Table 4</th>
<th>Statistics of the parameters of some of the estimated models for the original (non differenced) variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>AR (1)</strong></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
</tr>
<tr>
<td>Number of journeys</td>
<td>Min 0.96, Max 3.63, Mean 1.97, Std Deviation 0.80</td>
</tr>
<tr>
<td></td>
<td>Min 0.49, Max 0.85, Mean 1.83, Std Deviation 0.74</td>
</tr>
<tr>
<td>Number of trips</td>
<td>Min 1.04, Max 6.71, Mean 3.18, Std Deviation 1.58</td>
</tr>
<tr>
<td>Number of leisure related trips</td>
<td>Min 0.14, Max 1.20, Mean 0.63, Std Deviation 0.29</td>
</tr>
<tr>
<td>Total daily out-of-home activity duration</td>
<td>Min 966.27, Max 1469.76, Mean 1303.23, Std Deviation 110.07</td>
</tr>
<tr>
<td>Total daily trips duration</td>
<td>Min 33.52, Max 138.09, Mean 85.02, Std Deviation 30.28</td>
</tr>
<tr>
<td>Total daily distance</td>
<td>Min 2.84, Max 102.82, Mean 36.71, Std Deviation 27.4</td>
</tr>
</tbody>
</table>
Table 5  Statistics of the parameters of some of the estimated models for the lag-7 differenced variables

<table>
<thead>
<tr>
<th></th>
<th>AR (1) Constant</th>
<th>AR parameter</th>
<th>MA (1) Constant</th>
<th>MA parameter</th>
<th>ARMA (1, 1) Constant</th>
<th>AR parameter</th>
<th>MA parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total daily out-of-home activity duration</td>
<td>Min</td>
<td>-177.26</td>
<td>-0.69</td>
<td>-76.13</td>
<td>-0.68</td>
<td>-171.28</td>
<td>-0.97</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>1018.39</td>
<td>0.59</td>
<td>1.29</td>
<td>0.99</td>
<td>-3.19</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>-5.49</td>
<td>-0.27</td>
<td>-24.37</td>
<td>0.49</td>
<td>-59.14</td>
<td>-0.47</td>
</tr>
<tr>
<td></td>
<td>Std Deviation</td>
<td>166.5</td>
<td>0.33</td>
<td>24.92</td>
<td>0.60</td>
<td>68.27</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>-10.42</td>
<td>-0.49</td>
<td>-6.97</td>
<td>-0.99</td>
<td>-3.45</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>16.43</td>
<td>0.52</td>
<td>27.7</td>
<td>-0.16</td>
<td>-1.51</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1.75</td>
<td>-0.01</td>
<td>2.15</td>
<td>-0.48</td>
<td>-2.48</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>Std Deviation</td>
<td>5.58</td>
<td>0.38</td>
<td>9.63</td>
<td>0.22</td>
<td>1.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Figure 9  One example of ARIMA (0,7,1) modelling and forecasting. Total daily trips duration (city of Halle, household number 50, and person number 1).
4. Conclusions

The preliminary results of an ARIMA modelling approach for travel diary data of the Mobi-Drive dataset have been presented in this paper. For each of the variables considered, the subjects have been grouped according to the model category that best represented their behaviour. Due to the short time series available, the model estimation has not been performed under ideal conditions, therefore model parameters showed high variances and their estimation not always converged. An initial classification based on sex and age classes has been also carried out, in order to identify tentatively one or more socio-demographic variable that could classify the subjects into different categories.

This kind of analysis could be performed in greater details by means of discriminant analysis. Discriminant analysis should be able to determine which, among the socio-demographic variables, is sufficient to discriminate the subjects among the different model classes, and this analysis could be a starting point for further investigations into this direction.

A different possible approach, could be making use of a different class of ARIMA models, Periodic ARMA (PARMA). These stochastic models are especially suited to model series with a periodic component and actually estimate a different ARMA model for each period of the cycle. The presence of a different pattern for the weekdays and the weekends could indeed create problems during the estimation of an autoregressive model. In the case of the MobiDrive series, where the weekly cycle is evident for the majority of the subjects, one would estimate a model for each day of the week and use the same model for the same day of the following week. The problem with this kind of approach applied to our data set is once again the length of the time series. One can use only the data belonging to a specific period to estimate each PARMA model (i.e. only Mondays to estimate the Monday model) and this further reduces the length of the time series to a mere six days for each period, which is not enough to estimate the parameters of the model. One possible solution could be that of pulling the data together across the subjects and therefore creating a longer data series, but this must be carefully done, since the behaviour of the individuals can be very different and therefore a smoothing of the series can be unsafe. The results of the ARIMA modelling and the discriminant analysis classification could be, in this regard, a good indicator for the formation of each homogeneous group.
5. References


Wold, H. O., (1938) *A Study in the Analysis of Stationary time series*, Almqvist & Wiksell, Uppsala.

Eine vollständige Liste der Berichte kann vom Institut angefordert werden:

IVT ETHZ
ETH Hönggerberg (HIL)
CH - 8093 Zürich

Telefon: +41 (0)1 633 31 05
Fax: +41 (0)1 633 10 57
E-Mail: hotz@ivt.baug.ethz.ch
WWW: www.ivt.baug.ethz.ch

Der Katalog kann auch abgerufen werden:

http://www.ivt.baug.ethz.ch/veroeffent_arbeitsbericht.html

The Working Papers Traffic and Spatial Planning are intended for the quick dissemination of the results of the members and guests of the Institut. Their content is the sole responsibility of the authors.

A complete catalogue of the papers can be obtained from:

IVT ETHZ
ETH Hönggerberg (HIL)
CH - 8093 Zürich

Telephone: +41 (0)1 633 31 05
Fax: +41 (0)1 633 10 57
E-Mail: hotz@ivt.baug.ethz.ch
WWW: www.ivt.baug.ethz.ch

The catalogue can also be obtained from:

http://www.ivt.baug.ethz.ch/veroeffent_arbeitsbericht.html