Modelling Bicycle Demand Using Autoregressive and Moving Average Models

Jakub Starczewski 1, Dariusz Grzesica 1, Vojtech Jirsa 2

1 Krakow University of Technology, Institute of Road and Railway Engineering, Department of Transportation Systems, 31155 Warszawska 24 street, Poland
2 University of Pardubice, Jan Perner Transport Faculty, Department of Transport Technology and Control, Studentská street 95, 532 10 Pardubice, Czech Republic
dgrzesica@pk.edu.pl

Abstract. Bicycle transport is gaining importance in modern cities, therefore the modeling of its demand is most needed for sustainable city development. In this paper the measurement of number of bicyclists at one of the Czech Republic street was conducted. Monthly data in hourly periods was used. The autoregressive and moving average models of bicycle demand was calculated. Depending on ARMA parameters, to estimate the relative quality of statistical models, the Akaike Information Criterion (AIC) was used for a given set of data. In order to verify the correctness of the ARMA model, the residual autocorrelation analysis was carried out. The verification of the model's fit to the real data was made using the correlation coefficient.

1. Introduction

The congestion is one of the main problems of today's cities. It has a lot of negative effects, affecting many areas of everyday life. It is complicated and difficult to solve it through one action. Nevertheless, the biking is helpful in solve a case. Where the congestion exists, it often happens that cyclists can reach their destination faster than by car. Research shows that cycling has many other, good consequences. Frequent cycling has a positive effect on human's health, it is reducing environmental pollution and saving the money [1]. However, often the creation of bicycle traffic systems as an integral part of the city, takes places without forecasting its demand or taking it as a percentage of whole trips. It comes from the fact is it difficult to study bicycle traffic because of the small volume of data. This is a problem not only for researchers, but also for infrastructure investment decisions. Only in a few countries in Europe, reliable researches are making on cycling and its demand which have influence on the transport system in the urbanized area. As a consequence of the lack of data, many models have been created to forecast bicycle traffic. For example, these are: regression models [2], logit models [3], maximization of entropy models [4] and based on the Taylor series models [5]. Most of models focus on a 24-hour period of time and all of motivation needs due to bikes are relatively low part of traffic. Well-known and proven methods for estimating average annual daily traffic only from short-term car demand do not work in the estimation of bicycle demand [6]. There are 9 determining factors in the literature that influence of using a bicycle [7, 8] the weather, the public perception of biking, the safety, the household income, the sociodemographic attributes of the travel, the size and population density of cities, the relative cost of cars and public transport and the cycling infrastructure. These parameters have been tested in many independent studies in various parts
of the world and in many cases correlations between parameters and demand for bicycles have been confirmed [9]. For example, cyclists most often choose the shortest route between origin and destination. This is such an important parameter, that the extension of the path by 10%, decreases the number of cyclists by 60%, while at the extension by 70%, decreases the number of cyclists by 97% [10]. Moreover, people more likely take a trip by bike (1.46 times) when is it sunny weather or men appear to be 2 times more likely to use bike than women [11]. In the other hand a high population density, high job density, public facilities, services are associated with higher probability of biking [12]. Nevertheless, as noted by [13], in many respects the research on the choice of a bicycle is incomplete and sometimes contradictory. This is because there are many latent variables in research (for example in the human behaviour) that are often overlooked by researchers. That is why new research is needed for this topic.

2. Methodology

2.1. Data collection

Bicycle traffic flow data was obtained from the automatic bicycle counting system in the Czech capital Prague. The system in Prague operates over 7 years and consists of 23 counting locations. Cyclists are detected by induction loops (wires) located under or within the road/bicycle paths surface. Changes in inducted electrical current are recorded and evaluated by a data unit located in the nearby protective box. The data are then transferred to the web collection and date management platform. The transfer is performed automatically and remotely by means of GSM transmission. The data may be downloaded in CSV format (comma-separated values) for the selected period. The tables in the files contain the following columns: date time, number of cyclists in one direction, number of cyclists in the second direction, total number of cyclists, average temperature.

For the purpose of this study data collected in 2014 in the location “Podolské nábřeží - stezka” was used. It is the third busiest counting location in Prague with the average of 318 000 cyclists per year in both directions (Figure 1). This location has been monitored since 2010. The facility is located at the bike paths at the right bank of Elbe river. It is an important link between the downtown and the southern part of the city. The riverside provides the comfort of limited altitude changes during the ride, which in generally hilly landscape of Prague, increases the use of this particular path.

![Figure 1. Counting facility Podolské nábřeží - stezka (unicam.camea.cz, 2018)](image)

According to the hourly variability of traffic flows showing significant morning and afternoon peak in the traffic counts, it was estimated that the prevailing purpose of trips in this location is utilitarian [14].
2.2. **ARMA model construction**

Modelling time variables, as well as forecasting, it is assumed that the time series is generated by a linear combination of random variables. They can be used in stationary time series, i.e. in which there are only random variations around the average. Non-stationary time series can be reduced to being stationary through the process of differentiation. The time series therefore has an autoregressive representation.

Modelling time series, an autoregressive process of \( p \) order is often used – AR(\( p \)) and expressed by a formula:

\[
Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t
\]  

(1)

where: \( \varphi_0, \varphi_1, \ldots, \varphi_p \) – model parameters, \( \varepsilon_t \) - random variable with a normal distribution \( N(0, \sigma^2_\varepsilon) \), \( p \) – the value of delay.

The moving average part MA(\( q \)) is explained as \( q \) initial weights of the linear process different from zero and it is presented as follows:

\[
Y_t = \mu + \varepsilon_t - \vartheta_1 \varepsilon_{t-1} - \vartheta_2 \varepsilon_{t-2} - \cdots - \vartheta_q \varepsilon_{t-q}
\]  

(2)

where: \( \varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{t-q} \) – random disturbances in periods \( t, t-1, \ldots, t-q \); \( \mu, \vartheta_1, \vartheta_2, \ldots, \vartheta_q \) – model parameters, \( q \) – the value of delay.

However, it is not always possible to correctly match the model to real data using the AR(\( p \)) or MA(\( q \)) model. In order to increase the flexibility in matching, these two models are combined into one ARMA(\( p, q \)) with the formula:

\[
Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} - \vartheta_1 \varepsilon_{t-1} - \vartheta_2 \varepsilon_{t-2} - \cdots - \vartheta_q \varepsilon_{t-q} + \varepsilon_t
\]  

(3)

The ARMA(\( p, q \)) autocorrelation function depends on \( p \) and \( q \) parameters. If \( q - p < 0 \) it consists of a combination of disappearing exponential functions or suppressed sine waves whose properties are determined by a polynomial \( \Phi_p(B) \) and initial conditions, where \( B \) is associated with the autocorrelation function as follows: \( Y_t = \mu + \Phi_q(B) \varepsilon_t \). If \( q - p \geq 0 \) then \( q - p + 1 \) initial values of \( \rho_0, \rho_1, \ldots, \rho_{q-p} \) not subject to this scheme. The ARMA(\( p, q \)) partial autocorrelation function extends indefinitely, but for large intervals it behaves as a partial autocorrelation function of the moving average process.

2.3. **Model parameters selection**

One of the most important problems is to select model parameter. Firstly, the \( p \) parameter must be determine using autocorrelation function, and the \( q \) parameter by partial autocorrelation function [15]. Both \( p \) and \( q \) are selected based on the subjective decision of the forecaster. That’s why it’s hard to say which ARMA model in the best way reflects real data. The correctness of the chosen parameters is established on the basis of the autocorrelation and partial autocorrelation function of the rests.

2.4. **Stationary condition**

In order to apply ARMA model, the time series must be stationary, i.e. the constant time variance and the expected value of the process. To identify the stationarity of time series the unit root test was used. Taking into account possible autocorrelation of the random component the Augmented Dickey Fuller (ADF) test was used. When the time series is non-stationary it is necessary to bring a time series to stationarity by calculating the following differences: \( \Delta y_t = y_t - y_{t-1} \) - the first differences \( (d = 1) \), \( \Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \) - the second differences \( (d = 2) \). If the calculation of first differences is sufficient
to achieve stationarity the series is integrated of order one, it means that ARMA(\(p, q\)) model turns into ARIMA(\(p, d, q\)) as Autoregressive integrated moving average.

2.5. ARMA model selection

In selecting ARMA model proper parameters, it is necessary to increase the order until the modeling error is minimized. One of well none methods is Akaike Information Criterion (AIC) which provides a means for model selection [16]. It is an estimator of the relative quality of statistical models for a given set of data. For the several competing models the minimum information theoretical criterion is defined by the model and the maximum likelihood estimates of the parameters which give the minimum of AIC is defined by AIC = (-2) log (maximum likelihood) + 2(number of independently adjusted parameters within the model). The ARMA model parameters with minimum magnitude of AIC will be selected.

3. Results and discussions

Data of bicycle demand is from April 2015 measured in hourly periods. Figure 2 shows the variability of collected data.

![Figure 2. Bicycle demand in April 2015 at “Podolské nábřeží – stezka” point](image)

From all 720 observations, for constructing the model 715 were used. The last 5 observation will be used for constructing the forecast. The presented time series has been tested by ADF test. The value of \(p = 8.999e^{-07}\) and it’s less than the significant level \(p = 0.05\). The studied time series is stationary.

The model parameters were obtained from ACF and PACF functions presented in figure 3.
The maximum $p$ parameter which can be used was taken as 7 and the maximum $q$ parameter as 3. The Akaike Information Criterion along with the mean absolute percentage error (MAPE) checking the accuracy of a built model is shown in table 1.

Table 1. ARMA models with Akaike Information Criterion and MAPE level

| ARMA(p,q) model parameters | AIC   | MAPE  | ARMA(p,q) model parameters | AIC   | MAPE  | ARMA(p,q) model parameters | AIC   | MAPE  |
|----------------------------|-------|-------|----------------------------|-------|-------|----------------------------|-------|-------|
| (1,1)                      | 6526,053 | 86,895 | (1,2)                      | 6482,605 | 62,915 | (1,3)                      | 6482,007 | 55,066 |
| (2,1)                      | 6475,432 | 24,671 | (2,2)                      | 6465,555 | 11,702 | (2,3)                      | 6455,115 | 13,416 |
| (3,1)                      | 6466,053 | 32,106 | (3,2)                      | 6477,996 | 30,377 | (3,3)                      | 6455,922 | 12,409 |
| (4,1)                      | 6478,037 | 31,630 | (4,2)                      | 6461,706 | 10,635 | (4,3)                      | 6454,706 | 17,111 |
| (5,1)                      | 6452,169 | 17,432 | (5,2)                      | a       | a     | (5,3)                      | 6448,601 | 13,223 |
| (6,1)                      | 6453,794 | 18,547 | (6,2)                      | 6454,753 | 19,171 | (6,3)                      | 6456,635 | 18,3   |
| (7,1)                      | 6455,524 | 17,574 | (7,2)                      | 6456,731 | 19,016 | (7,3)                      | 6450,903 | 18,333 |

* the convergence criterion has not been achieved

The lowest MAPE value is for ARMA(4,2) and ARMA(2,2) and ARMA(3,3). The next one is ARMA(5,3) and it is also the model with the lowest AIC value. The ARMA(5,3) model is formed as:

\[
Y_t = 1,9Y_{t-1} - 1,991Y_{t-2} + 1,87Y_{t-3} - 1,043Y_{t-4} + 0,151Y_{t-5} - 0,7355e_{t-1} + 0,853e_{t-2} - 0,795e_{t-3}
\]

(4)

The model with forecast and real data are shown in Figure 4.
Validation of the ARMA(5,3) model is through the study of the autocorrelation of the model residues. The model shows no autocorrelation of residuals in the case when the coefficients do not differ significantly from zero (Fig. 5). The ARMA(4,2) and ARMA(2,2) and ARMA(3,3) models are invalid, because the model residues are significantly different from zero autocorrelation coefficients for subsequent lags indicate that the model parameters were correctly selected.
The obtained model can be used to prepare predictions. Forecast for 5 next periods is shown in Figure 6.

From the figure 6 it can be seen that the forecast fairly well cover the real data and it’s located in the middle of 95 significance level.

4. Conclusions
The selection of ARMA model parameters is subjective. Using Akaike Information Criterion there is a possibility to choose the best suitable one, which have the lowest MAPE level. Depending on ACF and PACF functions in this case there is no clear choice between values of their magnitude. Second most importing thing is that authors shows the benefits of using autoregressive and moving average models in identifying bicycle demand. Lots of papers declares that there is lack of methods that describe the bicycle demand. In this paper, it can be concluded that ARIMA models provide high quality of prepared forecasts. The calculated values are close to real. Seeing the seasonal changes in time series of the bicycle demand, the next step will be to create a model including the seasonality.

References
[1] C. Portilla, F. Valencia, J. Espinosa, A. Nunez, B. De Schutter, “Model-based predictive control for bicycling in urban intersections,” Transportation Research Part C, 2016.
[2] S. Hanson, P. Hanson, “Problems in integrating bicycle travel into the urban transportation planning process,” TRB TRR 629, Washington D.C., 1976.
[3] D.Y. Lott et al, “Bicycle Transportation for Downtown Work Trips: A Case Study in Davis, California,” TRR 629, TRB TRR Washington D.C., 1976.
[4] M.M. Richards, “Voorhees Associates of Disaggregate Technique to Calibrate a Trip Distribution and Modal Split,” TRB 592 NRC Washington D.C., 1976.
[5] C.J. Ruijgrok, “A method for making short term predictions of passenger transport,” M.F.A.M. Maarseveen Traffic & Transportation Group TNO, Delft 1984.
[6] N. Fournier, E. Chistofa, M. Knodler Jr., “A sinusoidal model for seasonal bicycle demand estimation,” Transportation Research Part D, 2017.
[7] J. Pucher, R. Buehler, M. Seinen, “Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies,” Transportation Research Part A, 2011.
[8] J. Pucher, R. Buehler, Making Cycling Irresistible, “Lessons from The Netherlands, Denmark and Germany,” Transport Reviews, 2008.
[9] A. Osama, T. Sayed, A. Bigazzi, “Models for estimating zone-level bike kilometers traveled
using bike network, land use, and road facility variables,” *Transportation Research Part A*, 2017.

[10] A. Bello Morales, J.M. Fonseca, “Manual para el plenamiento proyecto y ejecucion de pistas ciclistas,” *Association Espanola Permanente de los Congresos de Carreteras*, Madrid 1985.

[11] F. Godefroy, C. Morency, “Estimating latent cycling trias in Montreal,” *TRB Annual Meeting*, 2012.

[12] C. Zhao, T. Nielsen, A. Olafsson, T. Carstensen, X. Meng, “Urban form, demographic and socio-economic correlates of walking, cycling, and e-biking: Evidence from eight neighborhoods in Beijing,” *Transport Policy*, 2018.

[13] K. Habib, J. Mann, M. Mahmoud, A. Weiss, “Synopsis of bicycle demand in the City of Toronto: Investigating the effects of perception, consciousness and comfortability on the purpose of biking and bike ownership,” *Transportation Research Part A*, 2014.

[14] V. Jirsa., Y. Susilo, “Estimating the hourly variability of bicycle trip patterns and characteristics from automatic bicycle counters: Case study in Prague, Czech Republic,” *International Conference on traffic and Transport Engineering ICTTE*. 2016. Belgrade, Serbia. ISBN 978-86-916153-3-8.

[15] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, “Time Series Analysis: Forecasting and Control,” John Wiley & Sons, pp. 198, 2015.

[16] H. Akaike, “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, vol. 19/6, 1974.