Optimal Forecast Model For Erbil Traffic Road Data

Amin S. Mohammed
Software Department, College of Engineering, Salahaddin University, Erbil, Kurdistan Region, Iraq.

ARTICLE INFO

Article History:
Received: 28/03/2017
Accepted: 05/07/2017
Published: 28/12/2017

Keywords:
Transportation system Traffic Forecast model ARIMA

*Corresponding Author:
Amin S. Mohammed
Amin.mohammed@su.edu.krd

ABSTRACT

The paper is a study of the urgent problems of Erbil city - the issue of Transportation System. The paper analyzes traffic data in one of the main entrances street of the city, offers a model based on the theory of time series, to solve problems and improve the Erbil city's transportation system. There is verifying the behavior of the traffic flows and propose the best model for the traffic. Then offer an optimal forecast model. The models will help to structure the problem, to reorganize the transportation system and manage the endless process of development based on the principles of forming continuity. It is trying to predict transport traffic, which is a significant interest in areas such as traffic monitoring, control traffic flows, and management. Carefully forecasting and selected model can identify and predict the most important characteristics of traffic flow in the city.

1. INTRODUCTION

Currently, in most developed countries of the world there is a serious imbalance of supply and demand in the transportation system sector: traffic counts, spatially in large cities, increasing the throughput close to the limits of the capacity of existing roads, and sometimes exceed them in the rush hour. For a long time, the virtual monopoly of solving problems was extensive way, aimed at expanding the existing infrastructure of the city. However, this option is not rational. In its place comes high-tech approaches, involves the construction of new types of transports and a new model for traffic roads. Evaluating the effectiveness and reliability of such model is an important task in the design phase.

In recent years, Erbil, as the capital of the Kurdistan Region and as one of the most ancient cities, has faced with significant development. As a result, there was a need for the introduction of new ideas in all directions of developments, one of the main is a road transportation system. The primary task of this paper is to analyze traffics flow for one of the main entrance roads to Erbil City (Pirmam St) through time series theory and then predict the traffic route data (Fig. 1).
2. The literature review

At the moment, there is diverse literature on the study and modeling of road traffic flows and transportation systems. Basics of mathematical modeling of transportation systems were laid down by the RVOJAK (Jacques et al.). The primary objective, which served as the development of modeling of traffic flows, was to analyze the throughput of highways and intersections.

The first task of the transport flow theory historically was to seek independent of time relationships (links) between velocity, density, and bandwidth (stream), the so-called fundamental diagrams. A description of these relationships (links) discussed in the works of T. F. Golob, W. W. Recker and F.L. Hall (Hall, 1996), (Golob et al., 2003). The solution to this problem is possible only for small time intervals. The second step in the development of traffic flows modeling is the introduction of dynamics, i.e., describe the time-dependent. This was achieved by 2013-2014 Wang and Tettamanti (Wang et al., 2013, Tettamanti et al., 2014). In parallel, Gupta, S. Sharma, and P. Redhu offered Burgers equation as a model for the description of traffic flow and provide automatic measurement of the number of transport data (Gupta et al., 2014). At the moment, there is a diverse literature on the study and modeling of road traffic flows (Bus, 2015, Wang, 2010, Zheng et al., 2006, Pavlyuk, 2017, Neath et al., 2012, Wang, 2016, Ye, 2016, Barfield et al., 2014).

As mentioned above the aim of this paper is analyze traffic flow for one of the main entrance roads to Erbil City (Pirmam St) through time series theory and then forecast the data and optimal it as a part of the transportation system of the city.

2.1. Identify the model of traffic

To identify the model for traffic data is a multi-component task that required to find solutions build various types of mathematical models for part of transportation system (Hürlimann, 2013.). To develop and model the car traffic data it is using time series theory. The most crucial steps are to identify and create a model based on the available data. To reach this goal, we are following four steps:

1. Plot the time series data;
   - Check for Stationarity, Outliers, Seasonality
   - Interpretation of the Run Sequence Plot
2. Draw autocorrelation plot to test and examine the sample autocorrelations.
   - Compute and test the sample ACF and the sample PACF of the original series to further confirm a necessary degree of differencing.
   - Compute and review the sample ACF and PACF of the appropriately transformed and differenced series.
3. Examine the sample autocorrelations of the differenced data.
   Information-based criteria, such as the Akaike Information Criterion can be used to automate the choice of an appropriate model.

3. Model for Erbil traffic road data

3.1. Procedure of building model

As discussed in the previews section the first step of model determination is analyzing the data set of the transports. Fig. 2 shows the series of data, which is the daily number of traffic passes through the Pirmam Street. The data consist of 365 daily observations of consecutive days between January 1 to December 31 in 2009. In the next step is analyzing the data set to define the model,
which will be used for forecasting. To exactly say which model is more appropriate and close to the real data set several models have been tested, and the best one will be selected. At the end, by modification and changing some parameters build the optimal model. This section presents a variety of real-world traffic data to illustrate the method of model identification.

The next step is a calculation of the autocorrelation coefficient of 1st order, autoregression equation parameters, and sample averages. Parameters of the regression equation:

\[ x = \frac{\sum x_i}{n} = 1570.3260; \]
\[ y = \frac{\sum y_i}{n} = 1569.795; \]
\[ \bar{x}y = \frac{\sum xy_i}{n} = 3670387.723 \]  

then sample variance and standard deviation will be equal to:

\[ S^2(x) = \frac{\sum x_i^2}{n} - \bar{x}^2 = 1300232.700; \]
\[ S^2(y) = \frac{\sum y_i^2}{n} - \bar{y}^2 = 1300959.337; \]
\[ S(x) = \sqrt{S^2(x)} = 1140.277; \]
\[ S(y) = \sqrt{S^2(y)} = 1140.596. \]

The correlation coefficient \( b \) can be found by the formula, without solving the system directly:

\[ r_{t,t-1} = \frac{x_t \cdot y_{t-1}}{S(x_t)S(x_{t-1})} = 0.759 \]  

As is known, the linear correlation coefficient takes values from -1 to +1. Links between the signs may be weak and strong (tight). Thus, the connection between the rows is tall and straight (table 1).

| x    | y    | x²   | y²   | x*y  |
|------|------|------|------|------|
| 346  | 479  | 119716 | 229441 | 165734 |
| 479  | 566  | 229441 | 320356 | 271114 |
| …    | …    | …    | …    | …    |

Now we have to define the significance of the autocorrelation coefficient.

\[ t_{mon} = r_t \sqrt{\frac{n-2}{1-r_t^2}} = 6.92 \]  

Where \( t_{mon} \) - monitoring time. According to the table Student with a significance level \( \alpha = 0.05 \) and degrees of freedom
where \( m = 1 \) - the number of explanatory variables; \( t_{\text{crit}} \) – Student t criterion.

The value of the \( t_{\text{crit}} \) obtained from the observation data will compare with the tabular (\( t_{\text{critic}} \)-critical) value determined from the tables of the Student's distribution. If \( t_{\text{mon}} > t_{\text{critic}} \), the autocorrelation coefficient obtained value is recognized significant (the null hypothesis, which asserts the equality to zero autocorrelation coefficient rejected). Since \( t_{\text{mon}} > t_{\text{critic}} \), then we dismiss the hypothesis of equality of the autocorrelation coefficient of 0. In other words, the autocorrelation coefficient is statistically – significant.

Interval estimation of the correlation coefficient (confidence interval).
\[
(\bar{r} - t_{\text{crit}} \frac{1-r^2}{\sqrt{n}}; \bar{r} + t_{\text{crit}} \frac{1-r^2}{\sqrt{n}})
\]
Then \( r(0.57;0.95) \), and partial correlation coefficient \( \varphi_1 = r_1; \)

\( \text{Lag}=1, r_{1,t-1}=0.7593 \)

From the Fig. 3, it is clear that sample autocorrelations are very strong and positive and decay very slowly.

And the result is that in this range there is a tendency dynamic \( (r_{t,t-1} = 0.759 \rightarrow 1) \) and the autocorrelation indicates that the process is non-stationary and suggests an ARIMA model (Brockwell et al., 2016). The next section is to difference the data.

3.2. Determine the best model for traffic data

After analyzing the sequence plot of the differenced data, it will be easy to understand that the mean of the differenced data is close to zero, and differenced data less autocorrelated than the original data(Fig. 4).
The autocorrelation and a run sequence of the differenced data suggest that the differenced data are stationary. Based on the autocorrelation plot, an MA(2) model is proposed for the differenced data. To examine other possible models, we produce the partial autocorrelation plot of the differenced data.

![Residual Autocorrelations for CarTraffic ARIMA(2,1,2)](image)

**Figure (5a): Differenced data autocorrelation**

To determine the appropriate forecast model, one should consider the analysis of examines trends, Linear trend, Brown's linear, exponential Holt's linear exponential, etc. (Sbrana et al., 2014, Önder et al., 2014.). That are sustained movements in the variable of interest in a specific direction. To have a best and clear analyze and also choose the best model once should calculate all value of RMSE, MAE, MAPE, ME, MPE, AIC, HQC, and SBIC. The result of this calculation is illustrated in Table 2.

The table compares the results of seventeen different forecasting models. Looking at the error statistics, the model with the smallest root-mean-square error (RMSE) during the estimation period is model ARIMA (2,1,2). The model with the smallest mean absolute error (MAE) is model ARIMA (1,1,1). The model with the smallest mean absolute percentage error (MAPE) is Random Walk model. For ME it is model Holt's linear, exponential smoothing and so on. To select the most appropriate model. In the next step to determine the adequate of each model tests have been run on the residuals to determine whether each model is adequate for the data. Table 3 summarizes the results of these tests. Table 3 shows the results of fitting models to the data. The model with the lowest value of the Akaike Information Criterion (AIC) is model M, which will be used to generate the forecast. In the table, an OK means that the model passes the test. One * means that it fails at the 95% confidence level. Two *'s means that it fails at the 99% level of trust. Three *'s means that it fails at the 99.9% confidence level. Note that the current model M (ARIMA (2,1,2)) passes more tests.

![Residual Partial Autocorrelations for CarTraffic ARIMA(2,1,2)](image)

**Figure (5b): Differenced data partial autocorrelation's**

Fig. 5.a shows the sample autocorrelations of the differenced data. It is evident from the Fig. 5.b the partial autocorrelations of the first and second lag are significant. This suggests an AR (1) model for the differenced data.
### Table 2: Estimation Period (Using different model)

| Model                          | RMSE  | MAE   | MAPE  | ME     | MPE    | AIC   | HQC   | SBIC |
|-------------------------------|-------|-------|-------|--------|--------|-------|-------|------|
| (A) Random walk               | 436.797 | 267.17 | 19.529 | 0.417582 | -4.74325 | 12.1589 | 12.1589 | 12.1589 |
| (B) Random walk with drift = 0.417582 | 437.398 | 267.138 | 19.5259 | 7.49584E-15 | -4.78707 | 12.1672 | 12.1714 | 12.1779 |
| (C) Constant                  | 1141.84 | 908.984 | 92.3101 | 3.13963E-13 | -64.7812 | 14.0863 | 14.0905 | 14.097 |
| (D) Linear trend              | 1122.32 | 906.564 | 94.7732 | -7.35071E-14 | -65.5171 | 14.0573 | 14.0658 | 14.0786 |
| (E) Quadratic                 | 1120.44 | 913.416 | 96.7875 | -2.92783E-14 | -66.5104 | 14.0594 | 14.0721 | 14.0915 |
| (F) Exponential               | 1186.44 | 850.901 | 69.1739 | 347.929 | -28.4503 | 14.1684 | 14.1769 | 14.1898 |
| (G) S-curve trend             | 1195.82 | 841.013 | 67.6774 | 349.255 | -28.1346 | 14.1841 | 14.1926 | 14.2055 |
| (H) Simple moving average of 2 terms | 460.545 | 291.355 | 22.3386 | 0.247934 | -6.81627 | 12.2703 | 12.2745 | 12.281 |
| (I) Simple exponential smoothing with alpha = 0.8223 | 431.379 | 267.389 | 19.883 | 0.386818 | -5.57669 | 12.1395 | 12.1437 | 12.1501 |
| (J) Brown's linear exp. smoothing with alpha = 0.403 | 477.12 | 308.703 | 24.1087 | 2.32875 | -2.93748 | 12.341 | 12.3453 | 12.3517 |
| (K) Holt's linear exp. smoothing with alpha = 0.812 and beta = 0.0031 | 433.579 | 267.941 | 20.4329 | -28.5969 | -8.40993 | 12.1551 | 12.1636 | 12.1765 |
| (L) Brown's quadratic exp. smoothing with alpha = 0.212 | 513.336 | 342.616 | 27.3453 | 3.36929 | -3.5367 | 12.4873 | 12.4916 | 12.498 |
| (M) ARIMA(2,1,2)              | 425.003 | 267.174 | 20.5633 | -0.158286 | -7.22408 | 12.1261 | 12.1431 | 12.1689 |
| (N) ARIMA(1,1,1)              | 427.702 | 264.621 | 19.9659 | 0.257257 | -6.38129 | 12.1287 | 12.1363 | 12.1492 |
| (O) ARIMA(1,0,1)              | 430.149 | 267.757 | 19.6121 | 32.8303 | -3.2764 | 12.1392 | 12.1477 | 12.1606 |
| (P) ARIMA(1,0,2)              | 429.587 | 265.808 | 19.5949 | 29.5277 | -3.88179 | 12.1421 | 12.1548 | 12.1741 |
| (Q) ARIMA(2,0,0)              | 430.772 | 268.358 | 19.6057 | 34.497 | -2.99422 | 12.1421 | 12.1506 | 12.1635 |

### Table 3: Test for fitting models to the data

| Model                          | RMSE  | RUNS | RUNM | AUTO | MEAN | VAR |
|-------------------------------|-------|------|------|------|------|-----|
| (A) Random walk               | 436.797 | OK   | OK   | ***  | OK   | **  |
| (B) Random walk with drift = 0.417582 | 437.398 | OK   | OK   | ***  | OK   | **  |
| (C) Constant                  | 1141.84 | ***  | ***  | ***  | ***  | *** |
| (D) Linear trend              | 1122.32 | ***  | ***  | ***  | OK   | *** |
| (E) Quadratic                 | 1120.44 | ***  | ***  | ***  | OK   | *** |
| (F) Exponential               | 1186.44 | ***  | ***  | ***  | **   | *** |
| (G) S-curve trend             | 1195.82 | ***  | ***  | ***  | ***  | *** |
| (H) Simple moving average of 2 terms | 460.545 | ***  | **   | ***  | OK   | OK  |
| (I) Simple exponential smoothing with alpha = 0.8223 | 431.379 | OK   | OK   | ***  | OK   | *   |
| (J) Brown's linear exp. smoothing with alpha = 0.403 | 477.12 | **   | ***  | ***  | OK   | *   |
| (K) Holt's linear exp. smoothing with alpha = 0.812 and beta = 0.0031 | 433.579 | OK   | OK   | ***  | OK   | *   |
| (L) Brown's quadratic exp. smoothing with alpha = 0.212 | 513.336 | ***  | ***  | ***  | OK   | OK  |
| (M) ARIMA(2,1,2)              | 425.003 | OK   | OK   | ***  | OK   | *   |
| (N) ARIMA(1,1,1)              | 427.702 | OK   | OK   | ***  | OK   | *   |
| (O) ARIMA(1,0,1)              | 430.149 | OK   | OK   | ***  | OK   | *   |
| (P) ARIMA(1,0,2)              | 429.587 | OK   | OK   | ***  | OK   | **  |
| (Q) ARIMA(2,0,0)              | 430.772 | OK   | OK   | ***  | OK   | *   |
4. Forecast traffic data using the best model

This procedure will forecast future values of car traffic. The data cover 365 time periods. Currently, an autoregressive integrated moving average ARIMA (2,1,2) model has been selected. This model assumes that the best forecast for future data is given by a parametric model relating the most recent data value to previous data values. The output summarizes the statistical significance of the terms in the forecasting model. Terms with P-values less than 0.05 are statistically significantly different from zero at the 95.0% confidence level. The value of $P$ is less than 0.05, so it is significantly different from 0. The P-value for the MA (2) term is less than 0.05, so it is significantly different from 0. The estimated standard deviation of the input white noise equals 425.033. The table also summarizes the performance of the currently selected forecast model in fitting the historical data.

Each of the statistics is based on the one-ahead forecast errors, which are the differences between the data value at time t and its forecast at time t-1. The first three statistics measure the magnitude of the errors. A better model will give a smaller value. The last two statistics measure bias. A better forecast model should give a value close to 0.

Table 4: Effects ARIMA Model Summary

| Parameter | Estimate | Std. Error | $t$ | P-value |
|-----------|----------|------------|-----|---------|
| AR(1)    | 0.181141 | 0.0189525  | 0.957115 | 0.339152 |
| AR(2)    | 0.593371 | 0.121772   | 4.87281  | 0.000002 |
| MA(1)    | 0.028417 | 0.189562   | 0.149913 | 0.880917 |
| MA(2)    | 0.683237 | 0.144279   | 4.73551  | 0.000003 |

Table 5 and Fig. 6 show the forecasted values for traffic data. During the period where actual data is available, and the predicted values from the fitted model and the residuals are displayed (data-forecast). For time periods beyond the end of the series, it shows 95.0% prediction limits for the forecasts. These limits indicate where the actual data value at a selected future time is likely to be with 95.0% confidence, assuming the fitted model is appropriate for the data.

Table 5: Forecast values for traffic data

| Period | Forecast | Lower 95.0% | Upper 95.0% |
|--------|----------|-------------|-------------|
| 366.0  | 1080.85  | 244.987     | 1916.71     |
| 367.0  | 1077.33  | 11.8808     | 2142.78     |
| 368.0  | 1046.43  | -184.936    | 2277.79     |
| 369.0  | 1049.94  | -287.2      | 2387.08     |
| 370.0  | 1030.97  | -401.165    | 2463.1      |
| 371.0  | 1036.49  | -466.914    | 2539.89     |
| 372.0  | 1024.23  | -548.704    | 2597.17     |
| 373.0  | 1029.73  | -600.568    | 2660.02     |
| 374.0  | 1021.46  | -666.598    | 2709.51     |
| 375.0  | 1026.22  | -712.356    | 2764.79     |
| 376.0  | 1020.45  | -769.42     | 2810.32     |
| 377.0  | 1024.32  | -812.095    | 2860.73     |

5. Optimal ARIMA model

As a result of the test, several models are adequate to the initial traffic data, to optimal and final choice, it should take into account two factors:
- Improving the accuracy (quality of fit of the model);
- Reducing the number of model parameters.

Thus, the value of one parameter in models ARIMA (p, d, q) are already known. In part 4 after receiving a stationary series explores the behavior of the sample ACF and PACF and hypothesize about the values of the parameters p (order autoregression) and q (the order of the moving average), then the final selection of the model. To configure the optimal model two mentioned above factors going to take into account.

In this work, the choice of optimal model will produce by changing some parameters of the selected model. For that, Schwarz Bayesian Information Criterion (SBIC) [11] has been chosen which strengthened demand reduction of the number of model parameters:

\[ SBIC = \ln \left( \frac{\sum_{t=1}^{n} e_t^2}{n} \right) + \frac{(p+q)\ln(n)}{n} \]  

(5)

The model ARIMA (2,1,2) can construct a point, and interval forecast for L steps forward. To assess the accuracy of the forecast uses some standard indicators. The mean absolute percentage error (MAPE):

\[ MAPE = \frac{100\%}{L} \sum_{t=1}^{L} \left| \frac{Z_t - \hat{Z}_t}{Z_t} \right| ; \quad MAPE = 18.529\% \]  

(6)

Where \( Z_t \) - is real data values, \( \hat{Z}_t \) - forecast values, and L-Forecast interval. If \( MAPE < 10\% \), Then the forecast is made with high accuracy; If \( 10\% < MAPE < 20\% \), then the forecast is good, for the \( 20\% < MAPE < 50\% \)-forecast is acceptable and for \( MAPE > 50\% \) - it is a not acceptable forecasting. In chosen model \( MAPE = 18.529\% \), that means there is a good forecasting (Fig. 7.a).

5.1. Experimental procedure

The experiment was to identify the following algorithm made the minimum required the number of parameters of forecast model for adequate prediction:
- Original discrete time series of \( X_t \), appropriate traffic stood the so-called training area;
- at this training site evaluated parameters, proactive model;
- The results of the number of parameters in each model, active training site collected in tables.
- From the resulting table is selected the most common model;
- By the model of a randomly selected training portion of the original series, evaluated the predictive model parameters, and formed forecast of \( X_{t+1} \) values of some \( X_t \), following the end of the training site;
- It was recorded result of absolute prediction error;

Using above algorithm once can identify the model with a minimal number of parameters, which allows making the optimal forecast. Fig. 7 represents transport traffic with a dedicated training area for the Piramam St before and after using optimal forecast algorithm.

As it illustrated the \( MAPE \) value is reduced , which means the forecast is made with high accuracy and the model with the smallest root-mean-square error.

6. CONCLUSION
In this paper, we consider a model for the traffic flow of main entrance streets of the Erbil city. For this purpose, traffic data was collected for the Pirmam street during the year. After survey and analysis of the traffic data, the model of ARIMA (2,1,2) has selected as the most suitable model for this traffic, stationarity and relatively high efficiency were identified. Using this model procedure of forecasting was carried out. All the obtained values and graphs have subjected accuracy of the model. At the end of the work, it was proposed an algorithm for optimizing the chosen model. By reducing the number of model parameters, this goal has been reached.

The developed model makes it possible to analyze the current loading of the street and trace the reaction of transport networks to possible changes in transport demand (the construction of new flow-forming facilities) and the transport offer.

In the feature work, it is necessary to include the number of streets of the city, as well as possibly the entire town streets. And also, take into account other parameters of this system, such as the speed, type of the transporter, and etc.

REFERENCES

JACQUES, L. V., ROBERT, V. & MEHDI, D. Multifractal description of road traffic structure. 7th IFAC/IFOR.

HALL, F. L. 1996.Traffic stream characteristics. Traffic Flow Theory US Fed. Highw. Adm.

GOLOB, T. F. & RECKER, W. W. 2003.Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng., vol. 129, no. 4, pp. 342–353.

WANG, C., QUDDUS, M. A., & ISON, S. G. 2013. The effect of traffic and road characteristics on road safety: A review and future research direction. Saf. Sci., vol. 57, pp. 264–275.

TETTAMANTI, T., LUSPAY, T., KULCSAR, B., PENI, T. & VARGA, I. 2014. Robust control for urban road traffic networks. IEEE Trans. Intell. Transp. Syst., vol. 15, no. 1, pp. 385–398.

GUPTA, A. K., SHARMA, S. & REDHU, P. 2014. Analyses of lattice traffic flow model on a gradient highway. Commun. Theor. Phys., vol. 62, no. 3, p. 393.

BUS, J. 2015. Conservative forecasting with the damped trend. Res., vol. 68, no. 8, pp. 1739–1741, 2015.

WANG, F.Y. 2010. Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. IEEE Trans. Intell. Transp. Syst., vol. 11, no. 3, pp. 630–638.

ZHENG, W., LEE, H. & SHI, Q. 2006. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. J. Transp. Eng., vol. 132, no. 2, pp. 114–121.

PAVLYUK, D. 2017. Short-term Traffic Forecasting Using Multivariate Autoregressive Models. Procedia Eng., vol. 178, pp. 57–66.

NEATH, A. A. & CAVANAUGH, J. E. 2012. The Bayesian information criterion: background, derivation, and applications. Wiley Interdiscip. Rev. Comput. Stat., vol. 4, no. 2, pp. 199–203.

WANG, C. & YE, Z. 2016. Traffic flow forecasting based on a hybrid model. J. Intell. Transp. Syst., vol. 20, no. 5, pp. 428–437.

BARFIELD, W. & DINGUS, T. A. 2014. Human factors in intelligent transportation systems. Psychology Press.

HÜRLIMANN, T. 2013. Mathematical Modeling and Optimization: An Essay for the Design of Computer-Based Modeling Tools. Springer Science & Business Media.

BROCKWELL, P. J. & DAVIS, R. A. 2016. Introduction to time series and forecasting. Springer.

SBRANA, G. & SILVESTRINI, A. 2014. Random switching exponential smoothing and inventory forecasting. Int. J. Prod. Econ., vol. 156, pp. 283–294.

ÖNDER, E. AND KUZU, S. 2014. Forecasting air traffic volumes using smoothing techniques.