https://doi.org/10.26552/com.C.2011.4.87-91

Ondrej Cyprich *

APPLICATION OF UNIVARIATE TIME SERIES THEORY TO PASSENGER DEMAND FORECASTING

The methods, which are used for the purpose of passenger demand forecasting by Slovak transportation companies at the present time, are considerably simplified, and what is more, they are not already considered to be accurate. These limitations might be caused by insufficient research in this area over last years. Purpose of this paper is to identify a statistical model of passenger demand for suburban bus transport which satisfies the statistical significance of its parameters and randomness of its residuals. Three different methodologies – exponential smoothing, multiple linear regression and autoregressive models were used in order to identify more accurate and reliable statistical model compared with nowadays used ones.

Keywords: *Passenger demand. Demand modelling. Short-term demand forecasting.*

1. Introduction

Statistical modelling and forecasting of passenger demand by using univariate time series theory is probably one of the most common forecasting methods used for work with periodic time series data. This methodology has been successfully applied in the sphere of urban transport [1, 2] and in recently published models of passenger (carried per school reduced [3, 4] and normal fare [3, 5, 6]) demand for suburban bus transport. The main goal of this paper is to introduce method of the statistical modelling of passenger (carried per school reduced fare) demand by using univariate time series theory which appears to be more accurate and reliable alternative to automated forecasting procedures published in the literature [3]. In accordance with the main goal of the paper there was designed a statistical model which is suitable for short-term (forecast horizon $h \leq 1$ year) forecasting of passenger (carried per school reduced fare) demand for suburb bus transport in Zilina region. The most of analyses, modelling and forecasting procedures of the time series mentioned in this paper were worked out by using SAS LE 4.1 [7] and SAS 9.3.1 [8] software.

2. Materials and Methods

Properties of the used data, methods of its analysis, modelling and testing are briefly described in this section.

2.1. Properties and Adjustments of Input Data

Input data of experiments presented in the paper were counts of carried pupils and students collected by the cooperating carrier. These values were aggregated by summing so that an output of the aggregation process was monthly time series of passenger demand carried per school reduced fare $\{Q_p(t);\ 1 \leq t \leq 96\}$ (for period of months 1/2000-12/2007) in the Zilina Region.

Values in such a manner designed time series $Q_P(t)$ were considered to be spatially and substantially homogeneous as the carrier had changed neither his geographic scope nor transportation technology in the range affecting substantial and spatial aspects of the analysed time series within the specified period of months. "Trading day effects" were eliminated by own [9], passenger demand properties respecting, modification by Cipra [10] described calendar adjustment procedures. The output of the calendar adjustment process was fully homogeneous time series of passenger (carried per school reduced fare) demand for suburban bus transport $\{Q(t);\ 1 \leq t \leq 96\}$.

At first there were by subjective methods identified and later by objective methods properly confirmed – constant trend, monthly additive seasonality of $Q(t)$ time series in pre-forecasting analyses [9]. The models presented in this paper respect these properties completely.

2.2. Methods

Multiple regression, exponential smoothing and autoregressive models were used in order to statistical modelling of $Q(t)$ time series. The seasonal exponential smoothing model (method A) was developed and fitted by using exponential smoothing methodology. Smoothing state at time $t = 0$ of the model was obtained by Chatfield's backcasting method [11]. Smoothing weights (level α, sea-

* **Ondrej Cyprich**
Department of Road and Urban Transport, Faculty of Operation and Economics of Transport and Communications, University of Zilina, Slovakia.
E-mail: cyprich@fpedas.uniza.sk

sonal δ) were determined so as to minimize the sum of squared one-step-ahead prediction errors:

$$\sum_{t=1}^{n} \varepsilon_t^2 \rightarrow \min . \qquad (1)$$

Multiple regression was used in combination with Box-Jenkins methodology. The multiple regression (constant term with seasonal dummies) model combined with an autoregressive process of order $p = 1$ (AR(1) - method B) was used for the first time and then in the case of the multiple regression (constant term with seasonal dummies) model combined with an autoregressive/moving average process (ARMA (1,1) – method C). There were used practices and principles of linear stochastic models designing [10, 12] in the process of developing and fitting of $Q(t)$ time series models by using Box-Jenkins methodology. Applying this methodology were designed three autoregressive integrated moving average models of seasonal time series (ARIMA$(1,0,1)(0,1,1)_{12}$ – method D, ARIMA$(1,0,1)(2,1,0)_{12}$ – method E and ARIMA$(1,0,1)(1,1,0)_{12}$ – method F) – all without intercept parameter.

The statistical models presented in the paper were tested for compliance with the requirements imposed on mutual linear independence, stationarity and the normality of probability distribution of their standardized residuals ($\varepsilon_t = 1, ..., 96$). Mutual linear independence of models ε_t was tested by Bartlett´s test for autocorrelation [13] and Ljung-Box's χ^2 statistics [14]. Stationarity of the residual components was evaluated by augmented Dickey-Fuller's tests (ADF tests) [15] and Dickey-Fuller's unit root tests of seasonal time series (SDF tests) [16]. Normality of the standardized residuals probability distribution was tested by Shapiro-Wilk's (SW) [17] and by D'Agostino [18], Prins [19] and Filiben [13] described Kolmogorov-Smirnov's (K-S), Anderson-Darling's (A-D) and Cramér von Mises's (C-M) tests. Statistical significance of estimated parameters of the models was tested by Student's t-test [20]. These tests were conducted at significance level $\alpha = 0.05$ (except for the normality tests where higher values of significance levels ($\alpha = 0.15$) were used because of tests detection abilities).

3. Empirical results

The outputs of the forecasting procedures (analyses, modelling, testing) presented in the paper are goodness-of-fit statistics (Tab. 1), outputs of the randomness tests (Tab. 2) as well as evaluation of statistical significance of model parameters (Tab. 3). According to high volume of available outputs of computations they are presented only in considerably reduced form in the paper. Full outputs for all models including estimates of model parameters and their statistical significance evaluation, goodness-of-fit statistics, point and interval forecasts are part of dissertation thesis [9] and in the case of model estimated by using method E also in Perner's contacts [5].

In order to measure how well different models (methods A-F) fit the data there was computed traditional (root mean square error – RMSE, mean absolute percent error – MAPE) and penalty (Akaike's information criterion – AIC [21], Schwarz Bayesian information criterion – SBIC [22]) as well as extrapolational (MAPE$_3$, MAPE$_{12}$) goodness-of-fit statistics. Computed values of these measures see Tab. 1.

Based on the results of the tests for mutual linear independence, stationarity, normality of probability distribution and statistical significance of estimated parameters of the models seems the method E as the only one suitable for forecasting (ex-post, ex-ante) of $Q(t)$. The model (2) estimated by the method E showed very well fitting ability for actual data by its forecasts compared with other ones. Estimated values of its parameters with standard errors and outputs of their statistical significance tests (see Tab. 3).

Computed values of the goodness-of-fit statistics Tab.1

statistic	unit	method					
		A	B	C	D	E	F
RMSE	[1000 passengers]	4.795	4.746	4.690	5.772	5.691	5.965
MAPE	[%]	0.511	0.500	0.489	0.640	0.638	0.657
AIC	[-]	304.992	325.004	324.710	300.500	300.142	306.045
SBIC	[-]	310.121	358.340	360.611	307.792	309.865	313.338
MAPE12	[%]	1.076	0.866	0.989	.	0.977	0.939
MAPE3	[%]	0.806	0.802	0.776	0.748	0.626	0.744

Note: RMSE, MAPE, AIC and SBIC were computed by using the actual and forecasted values of observations in the period of evaluation (for a period of months 1/2000 – 12/2007), parameters of the models used for forecasting were estimated by applying observations from the same period of time. MAPE$_3$ and MAPE$_{12}$ were computed by using actual and forecasted values of observations in the period of evaluation (for a period of months 1/2007 -12/2007 – MAPE$_{12}$ and 10/2007 – 12/2007 – MAPE$_3$), parameters of the models used for forecasting were estimated by applying observations for a period of months 1/2000 – 12/2006 – MAPE$_{12}$ and 1/2000 – 9/2007 – MAPE$_3$.

Evaluation of tests for randomness of ε_t and statistical significance of estimated parameters of the models Tab. 2

Method	Statistical significance	Linear independence			Stationarity		Normality			
		BT-ACF	BT-PACF	LB	ADF	SDF	S-W	K-S	C-M	A-D
A	-	+	+	-	+	+	-	-	-	-
B	+	-	-	+/-	+	+	-	-	-	-
C	+	+	+/-	-	+	+	-	-	-	-
D	+	+	+	-	+	+	+	+	+	+
E	+	+	+	+/-	+	+	+	+	+	+
F	+	+	-	+	+	+	-	+	+	+

Note: Statistical tests provided "+"- satisfactory "-"- unsatisfactory "+/-"-boundary (satisfactory) results.

$$(1-B^{12})Q(t)=\frac{\theta(B)}{\varphi(B)\Phi(B^{12})}a_t \qquad (2)$$

where:
B is the backshift operator, that is, $BQ(t) = Q(t-1)$,
$\theta(B)$ is the moving-average operator, represented as the polynomial in the backshift operator: $\theta(B) = 1 + \theta_1 B$,
$\varphi(B)$ is the autoregressive operator, represented as the polynomial in the backshift operator: $\varphi(B) = 1 - \varphi_1 B$,
$\Phi(B)$ is the seasonal autoregressive operator, represented as the polynomial in the backshift operator: $\Phi(B) = 1 - \Phi_1 B^{12} - \Phi_2 B^{24}$,
a_t is independent disturbance (random error) at time t.

Inappropriateness of other models to produce forecasts resulted from confirming autocorrelation of their ε_t by Bartlett's tests (BT ACF, BT PACF) or Ljung-Box's χ^2 statistics (LB).

Further use of the forecasted values requires consideration of the fact that model (method E) systematically underestimates reality. This accrues from the value of mean percentage error ($MPE =$ $= 0.087\%$) of this model. True values $Q(t)$; $t = n + 1, ..., n + h$ are likely higher than forecasted values.

Graphical output of modelling and forecasting by using the method E (see Fig. 1) where estimated values are expressed by smooth curve and empirical values by black points. The graphical interpretation of the actual (empirical) and forecasted values show that this model accurately describes the variability of empirical values of $Q(t)$. This fact is also supported by low levels of residuals of the model (displayed by the bar diagram in Fig. 1).

It was objectively proved that it is possible to reduce the confidence interval (3) around the estimator $\hat{Q}(t)$; $t = n + 1, ..., n + h$ (at the confidence level of 0.95), from ±200 to ±16 thous. passengers carried, compared with outputs of computations published by Konečný [3].

$$P(L_{95t} \le \hat{Q}(t) \le U_{95t}) = 1 - \alpha \qquad (3)$$

where:
L_{95t} is lower limit of the confidence interval,
U_{95t} is upper limit of the confidence interval,
$1-\alpha$ is given probability, called confidence level of the interval,
$\hat{Q}(t)$ is estimated value of passenger demand.

More detailed comparison of forecasting abilities and statistical properties of the method presented in the paper with statistical model designed by Konečný [3] in view of goodness-of-fit statistics inaccessibility was not possible. It is obvious that the increase of statistical model (method E) reliability defined by the reduced confidence interval (3) is also the attendant phenomenon of its increasing interpolation accuracy.

Estimates of SARIMA model parameters and outputs of their statistical significance tests Tab.3

Model parameter	Estimate	Standard error	t-test criterion	p-value
MA(1) - θ_1	0.33996	0.1222	2.7825	0.0067
AR(1) - ϕ_1	0.91483	0.0503	18.1999	<.0001
SAR(1) - Φ_1	-0.77398	0.1084	-7.1430	<.0001
SAR(2) - Φ_2	-0.42544	0.1068	-3.9824	0.0001

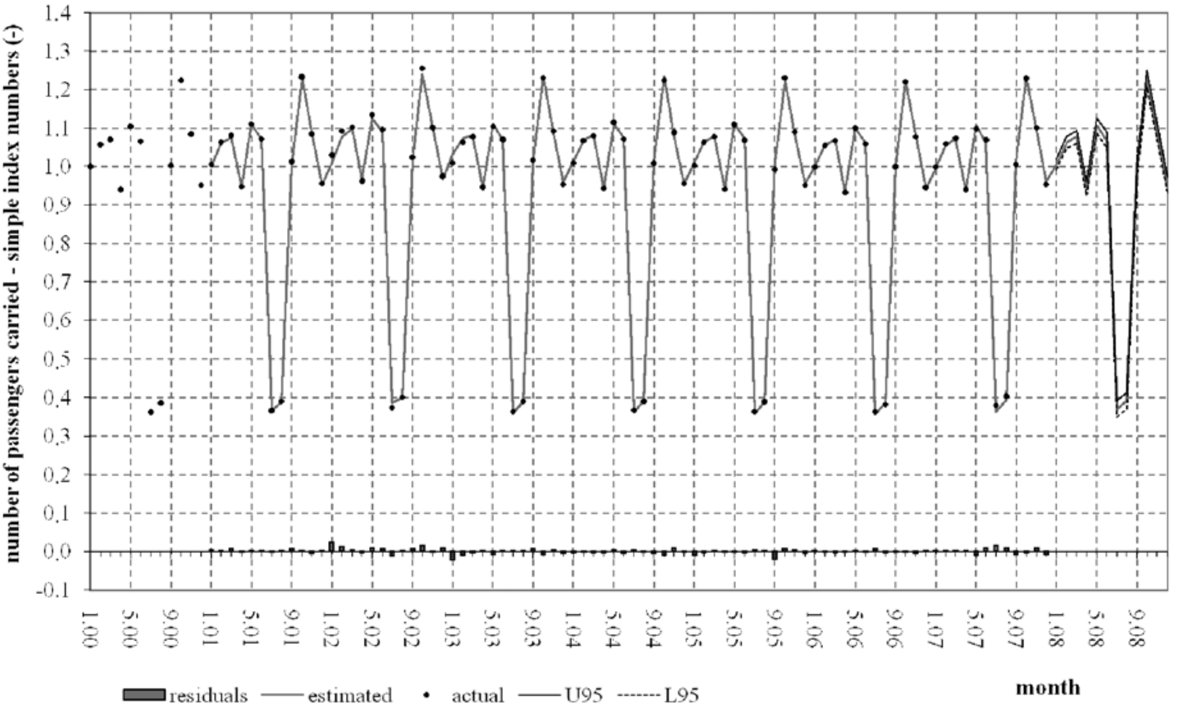

Fig. 1 Actual and estimated values of the Q(t) time series

Note: In order to protect interests of the cooperating bus transport company are values of Fig.1 presented in the form of simple index numbers ($I_{n/0}$). Reference value of the variable Q(t) is expressed in the base 1.0 in reference situation (t = 1, January 2000).

4. Conclusion

Outputs of the statistical tests of standardized residuals randomness and the values of goodness-of-fit statistics proved that the autoregressive integrated moving average model of seasonal time series ARIMA(1,0,1)(2,1,0)$_{12}$ without intercept parameter (method E) fulfils the requirements for statistical significance of its parameters, and what is more, mutual linear independence, stationarity and normality of probability distribution of its standardized residuals. The model presented in the paper is also because of these facts very good alternative to nonperiodic passengers demand time series forecasting methodologies [23, 24] and moreover provides more detailed monthly multi-step ahead forecasts. This model with respect to cross-regional differences cannot be considered as universally applicable throughout the Slovak Republic, but only in the Zilina region.

ARIMA(1,0,1)(2,1,0)$_{12}$ without intercept parameter presented in this paper despite the abovementioned restriction represents more reliable and more accurate passenger demand forecasting method in comparison with up to this time used ones. The attendant phenomenon of application in the paper described model in relevant transport company management is the reduction of manager's decisions uncertainty, and what is more, it can result in increase of company´s revenues.

References

[1] CYPRICH, O., LISCAK, S.: Modelling of Residuals and its Influence on Quality of Passenger Demand Multiple Regression Model (in Slovak). In: *Transport and Communications* [online]. 1. ed. 2010-1 [cit. 5. March 2011], p. 16–24. Available from: <http://fpedas.uniza.sk/dopravaaspoje/2010/1/cyprichliscak.pdf> ISSN 1336-7676.

[2] GNAP, J., CYPRICH, O.: Analysis of Hourly Traffic Flow of Passengers Seasonal Pattern in Mass Urban Transport of Bardejov City (in Slovak). In: *Transport and Communications* [online], 1. ed., 2010-1 [cit. 25. March 2011], p. 1–15. Available from: <http://fpedas.uniza.sk/dopravaaspoje/2010/1/gnapcyprich.pdf> ISSN 1336-7676.

[3] KONECNY, V.: Forecasting Passenger Demand for Suburb Bus Transport by using Univariate Time Series Theory (in Slovak). In: *Perner's Contacts* [online]. 1. ed.. Sept. 2009, vol. 15, No. III, [cit. 8. Jan. 2011], p. 130–136. Available from: <http://pernerscontacts.upce.cz/15_2009/Konecny.pdf > ISSN 1801-674X.

[4] CYPRICH, O.: ARIMA $(1,0,0)(1,1,0)_{12}$ Model of Passenger Demand for Suburb Bus Transport (in Slovak). In: *Perner's Contacts* [online], 1. ed. Nov. 2010, vol. 19, No. III [cit. 8. Jan. 2011], p. 26–34. Available from: <http://pernerscontacts.upce.cz/19_2010/Cyprich.pdf> ISSN 1801-674X.

[5] CYPRICH, O.: SARIMA Model of Students Demand for Suburb Bus Transport (in Slovak). In: *Perner's Contacts* [online]. Vol. 21, No. I [cit. 9. March 2011], p. 36–44. Available from: <http://pernerscontacts.upce.cz/21_2011/Cyprich.pdf> ISSN 1801-674X.

[6] GNAP, J., POLIAK, M., KONECNY, V.: *Forecast of Passenger Demand for Districts of Zilina Region Served by SAD Zilina (in Slovak)* [research study]. Zilina : Univesity of Zilina. 2008.

[7] *SAS LE 4.1* [software]. Cary, NC : SAS Institute Inc. 2006. Last actualisation 22. Dec. 2008.

[8] *SAS 9.1.3* [software]. Cary, NC : SAS Institute Inc. 2003. Last actualisation 22. Dec. 2008.

[9] CYPRICH, O.: *Modelling of Passenger Demand for Suburb Bus Transport (in Slovak).* [Dissertation thesis concept], Advisor: Liscak, S. Zilina : University of Zilina, 2010.

[10] CIPRA, T.: *Analysis of Time Series with Applications in Economy (in Czech).* 2. ed., Praha : STNL; Bratislava : ALFA, 1986. 248 p. ISBN 04-012-86.

[11] CHATFIELD, CH., YAR, M.: Holt-Winters Forecasting: Some Practical Issues. In: *The Statistician.* 1988, vol. 37, p. 129–140.

[12] ARLT, J., ARLTOVA, M.: *Economic Time Series (in Czech),*1. ed., Praha : Professional Publishing, 2009, p. 290, ISBN 978-80-86946-85-6.

[13] FILIBEN, J. J., HECKERT, A.: Exploratory Data Analysis. In: *NIST/SEMATECH e-Handbook of Statistical Methods* [online], [s.l.]: NIST/SEMATECH, published 6 July 2003. Last actualisation 23. June 2010. [cit. 8. Jan. 2011], chap. 1.3.5 Quantitative Techniques. Available at: <http://www.itl.nist.gov/div898/handbook/>.

[14] LJUNG, G. M., BOX, G. E. P.: On the Measure of Lack Fit in Time Series Models. In: *Biometrika,* 1978, vol. 65, pp. 297–303.

[15] HAMILTON, J. D.: *Time Series Analysis.* Princeton : Princeton University Press. 1. ed. 1994, p. 813, ISBN 0-691-042289-6.

[16] DICKEY, D. A., HASZA, D. P. A FULLER, W.A.: Testing for Unit Roots in Seasonal Time Series. In: *J. of the American Statistical Association,* June 1984, vol. 79, No. 386, p. 355–367.

[17] SHAPIRO, S. S., WILK, M. B.: An Analysis of Variance Test for Normality (complete samples). In: *Biometrika,* 1965, vol. 52, pp. 591–611.

[18] D'AGOSTINO, R. B., STEPHENS, M. A.: *Goodness-of-Fit Techniques.* New York : Marcel Dekker, 1. ed., 1994, p. 586, ISBN 978-0824774875.

[19] PRINS, J. et al.: *Product and Process Comparisons.* In: *NIST/SEMATECH e-Handbook of Statistical Methods* [online], [s.l.]: NIST/SEMATECH. published 6 July 2003. Last actualisation 23. June 2010 [cit. 8. Jan. 2011], chap. 1.3.5 Quantitative Techniques. Available at: http://www.itl.nist.gov/div898/handbook/.

[20] MARCEK, D., MARCEK, M.: *Analysis, Modelling and Forecasting of Time Series with Applications in Economics (in Slovak,* 1. ed., Zilina : EDIS, 2001, p. 282, ISBN 80-7100-870-2.

[21] AKAIKE, H.: Factor Analysis and AIC. In: *Psychometrika*, 1987, vol. 52, p. 317-332.

[22] SCHWARZ, G.: Estimating the Dimension of the Model. In: *The Annals of Statistics,* 1978, vol. 6, No. 2, pp. 461–464.

[23] DICOVA, J., ONDRUS, J.: Prediction of Development Public Passenger Transport. In: *J. of Information, Control and Management Systems,* 2010, vol. 8, pp. 475–482.

[24] DICOVA, J., ONDRUS, J.: Trend of Public Mass Transport Indicators – as a Tool of Transport Management and Development of Regions. In: *Communications - Scientific Letters of the University of Zilina,* vol. 12, No. 3A, 2010, pp. 475–482.