Generalised Intrinsic Characteristics as a Forecasting Tool: A dynamic perspective

Radosav B. Jovanović

***Abstract*— The trip generation model presented combines sequentially cross-sectional and time-series data analyses, utilising only basic statistical tools. It exemplifies one possible way of tackling one of the frequent forecasting-related issues: restricted sample size – either spatially (small number of observational units) or/and longitudinally (insufficiently long time-series). The model developed makes use of the observed autocorrelation of the regression residuals over time, and by including a dynamic element into the formerly known theory of intrinsic characteristics, takes account of the gross effect of extraneous factors on air passenger traffic volume. The modification applied significantly improves the forecasting accuracy of the original cross-sectional model.**

***Index Terms*—Autocorrelation, cross-sectional analysis, intrinsic characteristics, regression analysis, time-series, trip generation model.**

I. Introduction

Modelling transport, in general, is a context-dependent process, and no universal technique exists that might be successfully applied to every problem, it is rather that every single model should be tailored to its particular context. Naturally, theoretical soundness of the model proposed is a necessary condition, still, one should bear in mind that accuracy is only one aspect of forecasting, and that cost, ease of use, utility of output and ease of interpretation are almost as important [1], [2]. Therefore, balance ought to be sought between theoretical consistency and expedience in the forecasting process [3]. In many cases, relatively simple (inexpensive and readily applicable) modelling techniques may provide satisfactory forecasts with respect to most of the criteria previously mentioned.

This paper presents the results of such an application, blending two related, but so far separate concepts — to obtain an air trip generation forecast for a set of European countries. The first concept is purely statistical - autocorrelation of the error terms in a time-series regression, whereas the second, a theory of intrinsic characteristics, but with a dynamic element incorporated, has a reasonable practical interpretation. Cross-sectional and time-series approaches were combined, with only basic statistical tools utilised for parameter estimation.

The trip generation model developed is inevitably concerned with the rest of the modelling process, being an integral part, thus a brief overview of the broader modelling context is thought to be helpful in recognizing the framework in which the model has been built.

II. Context

The model described in this paper has been developed as a part of the *"Serbia and Montenegro (SCG) Air Traffic Analysis and Forecast"* study. The study ordered by the Serbia and Montenegro Air Traffic Services Agency (SMATSA) was undertaken by the Faculty of Transport and Traffic Engineering, University of Belgrade, and supported by EUROCONTROL, in the period December 2004 – June 2005. Its objective was the forecast of annual and peak day SCG air traffic volume for years 2007, 2010 and 2015 [4].

A classical sequential four-phase transportation demand model structure has been used for that purpose. Due to the absence of serious intermodal competition in the area of interest a mode-specific approach has been employed. The trip generation model was used to estimate the air trip production of a set of European zones identified to be the principal generators of the SCG transit air traffic (Figure 1). Its output has served as one of the inputs for the trip distribution model, the output of which represented itself one of the inputs for the final, route-choice phase.

Four scenarios concerning the design years' levels of explanatory variables in the route-choice phase were provided by the study customer, thus any branching in earlier stages would have led to an impracticably large number of variants at the very end. Therefore, the point forecast of the trip production level of each of the zones was needed.

A. Data Availability

It is well known that data availability and modelling approach constrain and condition each other [5]. Whereas it is sometimes difficult to decide on the predominant direction of this two-way relationship, in our case the model structure was clearly heavily influenced by the data

Manuscript received February 26, 2006.

Radosav B. Jovanović is a part-time research assistant at the Air Transport Department, Faculty of Transport and Traffic Engineering, University of Belgrade, Belgrade, Serbia and Montenegro (phone: +381 64 1256 829, +381 11 3091 352; fax: +381 11 2496 476; e-mail: rgrozni@eunet.yu).

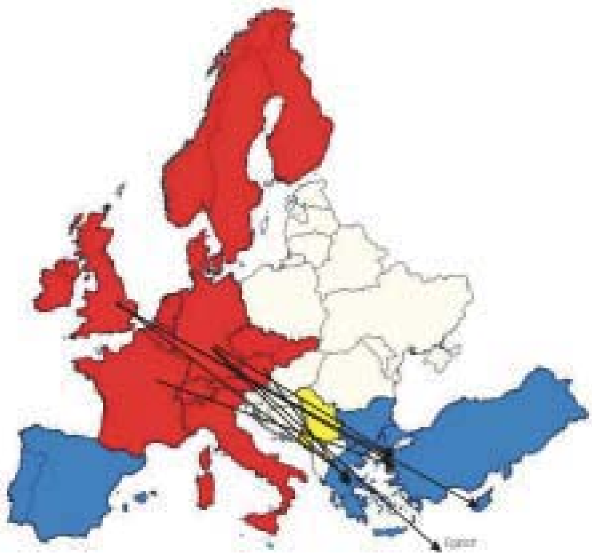


Fig. 1. Main air traffic flows crossing SCG airspace
Source: [4]

availability. Apart from the problems of obtaining a long enough historical data series for each of the zones considered, as a ten-year forecast of the zonal passenger traffic was the required model output, the equivalent length of a credible forecast of any would-be explanatory variable was subsequently needed too. This turned out to be a highly restrictive prerequisite, with only a few variables having the desired properties.

Historical data on the volume of international air passenger traffic by reporting country were available for the 1995-2003 period. Data on GDP levels, a few tourism indicators and on the volume of external trade by country were collected. The importance of the air fare level as a measure of travel impedance was recognized too, however, both lack of systematic historical data on ticket prices, and non-existence of reliable forecasts on their future level made the inclusion of this factor impracticable.

B. Variable Choice

Having conducted preliminary analyses based on available data sets, being in principle inclined towards developing a causal rather than a univariate model, and bearing in mind the required form of the model output, it was finally decided to model the *zonal volume of international air passenger traffic production* (IP) using zonal GDP as the only explanatory variable. It should be mentioned that the other exogenous variables tested have also showed statistical significance with respect to the zonal traffic production, but were also highly correlated with the GDP level of the zones considered, which posed a problem of multicollinearity.

A literature review also showed that GDP is most often considered a main driver for passenger growth [6] - [8]. Moreover, such an approach is in line with the practice followed by ICAO [9] and EUROSTAT [10], which correlate passengers to GDP only.

III. Trip Generation Model Concept

It should be stressed at the very beginning of this section, that no strict trip generation model concept had been set in advance of the data analysis process. The only point fixed was the goal of the analysis, that is, the required form of the model output, thus freedom existed in that sense to make different methodological choices at different points of the process, in order to approach the desired goal as close as possible. To keep some cohesion in the discussion below, the comments on emerging theoretical (and practical) issues at the points where alternative options were possible will be given right after a particular choice is made, rather than in a separate section.

A. Cross-sectional vs. Time-series Approach

Having a nine-year long time series of both IP and GDP values for 14 zones, and opting for the regression analysis as a forecasting tool, it was possible either to try to set the common relationship for all zones, or, alternatively, to model an IP versus GDP temporal relationship (multivariate model) for each of the zones considered. Pros and cons for each of the alternatives have been widely discussed in the referent literature.

It is generally argued that the kind of behaviour measured from cross-sectional data is typically long-run in nature, whereas time-series data tend to yield short-run responses [11], [12]. It is certain that by establishing a common pattern, cross-sectional models have advantages in terms of generality and ease of interpretation. However, the gain in generality is usually offset by a corresponding loss of predictive accuracy, since it is fairly unlikely that the static cross-sectional structure can satisfactorily reflect the changing nature of the underlying process for each of the observational units. On the other hand, the impact of sample size on the choice of the modelling approach has to be recognized too, and a time-series approach typically requires relatively large numbers of temporal observations to draw statistically valid conclusions [13].

Both approaches are, however, concerned with certain theoretical problems when using linear regression for model parameter estimation. Cross-sectional observations are frequently *heteroscedastic*, that is, the assumption of equal variance over observations is often violated, usually because of differing factors related to the size of the different cross-sectional entities or varying background conditions [14]. On the other hand, time-series models typically violate the assumption of independence of the error terms, i.e. they suffer from *autocorrelation* [15], [16]. The primary cause of this problem is considered to be the failure to include one or more important regressors in the model when there happens to be a high degree of temporal correlation in their cumulative effect [2]. The violation of the assumption of *normally distributed disturbances* can generally also pose a problem, yet in cases of point forecasts needed “ignoring normality will not hinder the ability to make predictions” [17]. The presence of these and some related problems might affect the efficiency and unbiasedness of parameter estimates and hence lead to invalid conclusions, unless they are appropriately tackled [14], [18].

B. *Original Model Concept*

Cross-analysing the EUROSTAT air passenger data and the corresponding data on the zonal GDP, it turned out that the number of international passengers flying to/from the zones (IP) is highly correlated with their GDP level. The power curve

$$\mathbf{IP = \alpha \times GDP^{\beta}} \qquad (1)$$

proved to facilitate this relationship most appropriately, with a multiple determination coefficient (R^2) ranging from 0.86 to 0.88 for every year from 1997 to 2003.

Having decided on the form of the relationship, the following step was to select the base year for the model parameter estimation. For more than one reason our choice was 1997. Firstly, it was a relatively calm (free of external disturbances) year concerning the air transport industry. The economic recession of the early 1990s was already a few years away. In addition, the corresponding statistical database turned out to be richer relative to other years. Finally, it was considered an advantage to be able to test the model on annual data both preceding and succeeding the base modelling year. This was thought to be helpful in both detecting its possible shortcomings and at the same time testing its temporal (un)stability.

After linearising, equation (1) becomes:

$$\mathbf{IP' = \alpha' + \beta \times GDP'} \qquad (2)$$

where:

$$IP' = \ln(IP),$$
$$\alpha' = \ln\alpha, \text{ and}$$
$$GDP' = \ln(GDP).$$

The summary of the model (2) parameter estimation for 1997 is shown in Table I.

TABLE I
TRIP GENERATION MODEL PARAMETER ESTIMATION, 1997

Expl. Var.	**Coeff.**	**Est. Coeff.**	**t-stat.**
Constant	α'	10.2425	14.061
GDP'	β	0.5358	9.337
R^2_{adj}	0.869		

Taking anti-logarithm of (2) the original trip generation model finally becomes:

$$\mathbf{IP = 28072 \ x\ GDP^{0.5358}} \qquad (3)$$

where:

IP stands for the volume of international air passengers flying to/from the zone considered, and

GDP stands for the gross domestic product (in millions of 2003 US$) of the corresponding zone.

Figure 2 shows the fitted regression curve (3) for the 1997 data. Although GDP itself explained 87% of the total variance, which can be considered a fairly good result, it can be seen that in a few cases a difference between estimated and observed traffic was beyond tolerable, especially having in mind the forecast to be made. Traffic was significantly underestimated for two zones (UKI and SPA, see Table III for the list of zones and labels used), while Italian trip production was noticeably overestimated.

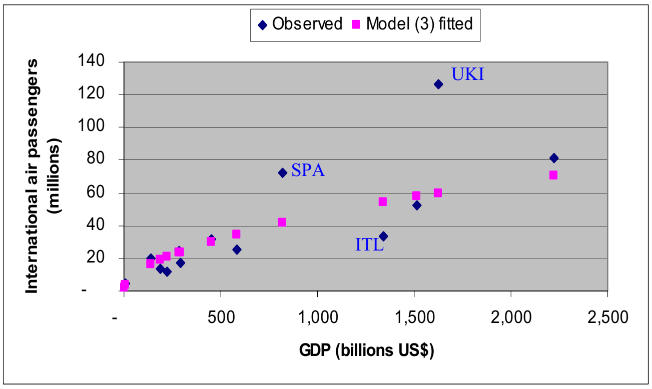


Fig. 2. Trip generation model estimation, 1997 data.

C. *Checking Regression Assumptions*

A plot of *model fitted values vs. residuals* was inspected first to assess homoscedasticity [17]. There seemed to be no severe violation of the homoscedasticity assumption, i.e. no consistent pattern of the disturbances magnitude with respect to the magnitude of the fitted dependent variable was observed.

The normality, on the other hand, appeared to be somewhat violated, which was ascertained through the analyses of disturbances' summary statistics and normal probability Q-Q plots, as well as by conducting the Kolmogorov-Smirnov test. However, it has to be stressed that the sample size (only 14 observations in this case) directly influences the degree to which disturbances exhibit "normal" behaviour, hence normally distributed small samples can exhibit "apparently non-normal behaviour as viewed through diagnostic tools" [17]. Anyway, as already mentioned, in cases when prediction of the point estimates is the only modelling purpose, the impact of non-normality on accuracy is practically negligible, but it could partly explain the presence of the regression outliers [1].

The regression outliers could be dealt with in different ways. They could simply be removed from the regression and modelled separately whereas the original model would be re-estimated, with a likely improvement of fit, but also a certain loss of coherence. Alternatively, the outliers could be left in, and some other way sought to resolve the problem and make the model fit the data [17]. However, prior to opting for any of these choices, there still remained a temporal dimension of the problem to be examined.

D.Residual Analysis

Testing (3) on a number of annual datasets, a certain regularity has been observed. For most of the zones, there turned out to be a high degree of consistency in the behaviour of the residuals (conveniently transformed to relative terms - ratios of observed/estimated IP) over time, i.e. a clear monotony has been observed in the way that the estimated number of passengers differs from the observed one for the zone considered (Figure 3). On the other hand, for BLX and SWI zones, the variations of this ratio have coincided with the crises and eventual bankruptcy of Sabena and Swissair in 2001 and 2002.

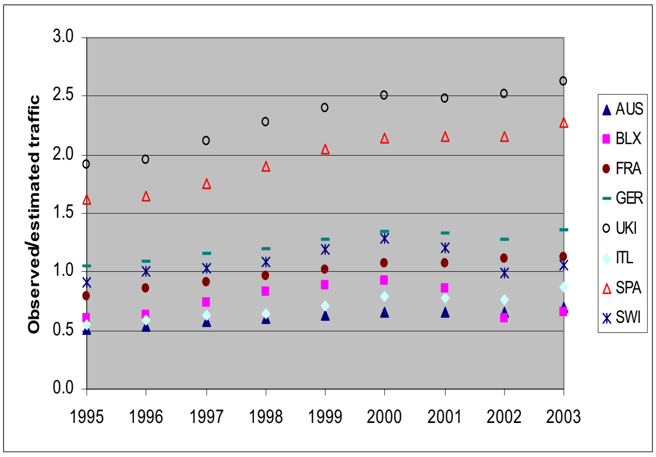


Fig. 3. Zonal residual time-series.

It has already been mentioned that using regression for time series analysis could potentially pose the problem of the linear association of successive residuals, referred to as autocorrelation. But it is not necessarily bad news. Put another way, use could be made of the strong dependence observed. On the one hand, it has been recognized that the error terms contain the information on the influence of the factors not explicitly included in the model [13]. A number of methods have been developed attempting to exploit this opportunity, e.g. "cross - sectionally heteroscedastic and timewise autoregressive model" [18], "error components model" [19], "regression models with a time-series error component" [20], etc. These generally fall under the *panel (or pooled) data analysis* category, as they deal with cross-sections of observational units observed over time. Most of them are, however, associated with relatively complex statistical procedures for parameter estimation, yet little is known on their performance with small samples [13].

On the other hand, a simple transformation of original error terms into *ratios of observed/fitted values* (see y-axis, Figure 3) implies an analogy with what is known as the theory of intrinsic characteristics.

E. The Theory of Intrinsic Characteristics

The idea has originally been employed in a model estimating the trip distribution pattern for a set of North-American cities. It points out the inability of a general transportation demand model to account for all the differences in the trip production and attraction characteristics of zones. Realizing that the error term may include the effect of all omitted variables it has been suggested that this might be used for forecasting purposes. For that reason, the error term has been decomposed into two independent components: "*intrinsic characteristic*" (reflecting the specificity of the corresponding city-pair) and the *random error*. Both Generalised Least Squares (GLS) and Ordinary Least Squares (OLS) methods were used to estimate the intrinsic characteristics (IC), with GLS performing somewhat better. Finally, the original application of this theory assumed that the intrinsic characteristics do not change substantially over time [21].

F. Critique, Generalisation and Synthesis

More than a few authors, however, oppose the constant specification models. Chatfield, for instance, argues that "a local model, which changes through time, may be preferred to a global model, that has constant parameters" [1]. Kanafani suggests that the model specifications can be altered to permit variable elasticities and "to allow for a levelling off of air traffic growth as related to the growth in the demand variables" [6].

With all of this put together, the possibility of generalisation of the original concept of intrinsic characteristics to the trip generation model seemed rather straightforward. The basic hypothesis, that the effects of the conventional demand variables on intercity travel vary with the economic and social characteristics of the city, applies equally well to the traffic generating potential of different zones. It is both intuitively clear and in accordance with empirical evidence that traffic growth is driven by a different (both qualitatively and quantitatively) blend of factors for different geographical entities. The extent to which factors other than GDP influence the trip production volume of different zones could therefore be represented by zonal intrinsic characteristics. In addition, it is to be expected that the intensity of this aggregate effect changes with time, reflecting the temporal changes of the factors constituting it (e.g. transport supply, tourism, external trade, etc.).

In summary, the following approach was adopted: to exploit the strong temporal correlation between the successive zonal relative residuals, referred to as *generalised zonal intrinsic characteristics* (GIC), and derive a separate trend for each zone's GIC series, which would finally be integrated into the original model. Hence, a dynamic element has been incorporated into the authentic IC theory, in order to reflect the changing nature of the underlying process with time.

It turned out that in most of the cases a simple linear approximation of the GIC time-series provided well-fitted sub-models (Figure 4). Although it might be more reasonable to assume that the regression coefficients in the sub-models evolve stochastically with time, giving rise to what is called a "stochastic trend" [1], the practical issues (i.e. point forecast needed) were again overwhelming.

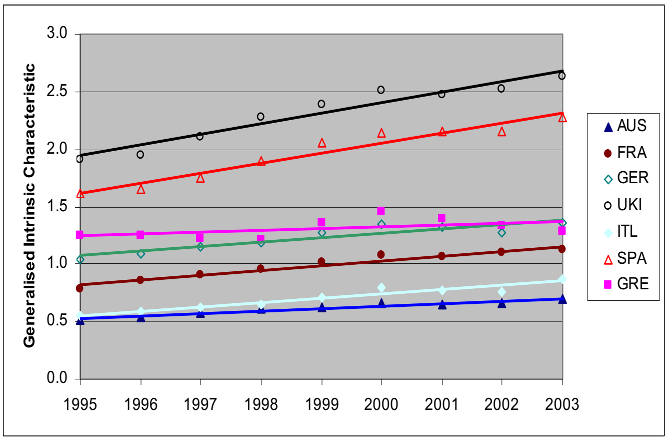


Fig. 4. Zonal generalised intrinsic characteristics, linear approximation.

Thus the deterministic GIC trends were finally incorporated into the original model, so that the modified trip generation model became:

$$\mathbf{IP_{zt} = 28072 \times GDP_{zt}^{\ 0.5358} \times GIC_{zt}} \qquad (4)$$

where:

IP$_{zt}$ is the number of international air passengers flying to/from zone **z** in a year **t**;

GDP$_{zt}$ is the gross domestic product (in millions of US$) of a zone **z** in a year **t**, while

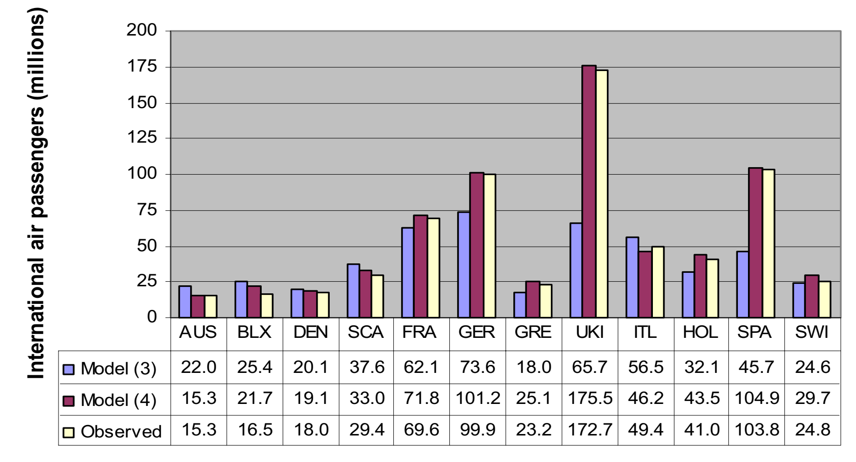
GIC$_{zt}$ stands for the generalised intrinsic characteristic of a zone **z** in a year **t**.

IV. Model Validation

Model (4) validation was originally conducted using 2003 figures on air passenger traffic and the GDP levels, whereas zonal GICs where derived based on 1995-2002 time-series. The series has been shortened one year so that the model prediction could be compared with information not used during the process of the model estimation [3].

The forecasting output for 2003 is presented in Figure 5, along with the observed values and the model (3) predictions. The gain in accuracy is more than obvious, with *Mean Average Percentage Error* (MAPE) being 8.3% for model (4) estimates, while the overall 2003 traffic volume for 14 zones considered was overestimated by 3.6%, Table II.

Table III presents the comparison of the model (4) estimates and in the interim released observed traffic volumes in 2004.



	AUS	BLX	DEN	SCA	FRA	GER	GRE	UKI	ITL	HOL	SPA	SWI
□ Model (3)	22.0	25.4	20.1	37.6	62.1	73.6	18.0	65.7	56.5	32.1	45.7	24.6
■ Model (4)	15.3	21.7	19.1	33.0	71.8	101.2	25.1	175.5	46.2	43.5	104.9	29.7
□ Observed	15.3	16.5	18.0	29.4	69.6	99.9	23.2	172.7	49.4	41.0	103.8	24.8

Fig. 5. Model validation, 2003 data.

Table II
Model validation – summary statistics

Year	2003			2004		
	MAPE (%)	Total traffic volume prediction error (%)	R^2_{adj}	MAPE (%)	Total traffic volume prediction error (%)	R^2_{adj}
Model (3)	28.2	-27.1	0.646	-	-	-
Model (4)	8.3	+3.6	0.998	4.95	-1.6	0.997

TABLE III
MODEL (4) VALIDATION – 2004 DATA
(millions of international air passengers)

Zone label	AUS	BLX	DEN	SCA	FRA	GER	GRE	UKI	ITL	HOL	SPA	SWI	Total
Constituting countries	Austria	Belgium & Luxemb.	Denmark	Sweden, Finland & Norway	France	Germany	Greece	UK & Ireland	Italy	Holland	Spain & Portugal	Switzerland	
Observed	17.7	19.0	19.4	32.3	76.0	114.5	24.0	186.7	57.0	44.4	111.7	25.8	728.6
Estimated	16.0	20.2	19.9	32.5	75.4	105.5	25.6	185.3	50.6	45.6	111.1	28.5	716.1

V. SUMMARY AND CONCLUSIONS

To obtain a forecast of the international air passenger traffic production for a set of European zones, a sequential combination of cross-sectional and time-series data analysis has been employed, resulting in a common trip generation model established which also allows for the specificity of the different geographical entities considered. This has been accomplished by incorporating a dynamic element into the theory of intrinsic characteristics, accounting that way for the time-varying gross effect of the causal variables not explicitly specified in the model.

The modification applied significantly improved the explanatory power and predictive accuracy of the original cross-sectional model, producing fairly accurate forecasts as a result. Two out-of-sample validation tests performed showed the reduction of MAPE from 28% to 5-8%, with the corresponding R^2_{adj} value rise from 0.65 to 0.99.

The model specification allows for a straightforward re-calibration once any new historical data are released, with a consequent likely improvement in its forecasting accuracy. This feature is in accordance with theoretical recommendations on the use of macro-models with constant elasticities for transport demand forecasting [6].

However, the model presented should primarily be seen in the light of its particular context, i.e. as a convenient practical tool to obtain short-to-medium run forecasts in situations with small samples and/or relatively short historical data time-series.

Further generalisations of the concept are, nevertheless, possible, and more advanced methods could be employed to examine in more detail the behaviour of zonal generalised intrinsic characteristics over time. That would generally enable a plausible interpretation of the different impact of extraneous factors on the predicted variable, with a likely improvement of the overall effectiveness of the forecasting system. Examining the effects of the integration of such methodology into the concept presented might be an interesting problem for further research.

REFERENCES

[1] C. Chatfield, *The Analysis of Time Series: An Introduction.* 6th ed. Boca Raton: Chapman & Hall/CRC Press, 2004.

[2] D. C. Montgomery, L. A. Johnson, and J. S. Gardiner, *Forecasting and Time Series Analysis.* 2nd ed. New York: McGraw-Hill, 1990.

[3] J. de D. Ortuzar and L. G. Willumsen, *Modelling Transport.* 2nd ed. Chichester: John Wiley & Sons, 1994.

[4] Faculty of Transport and Traffic Engineering, University of Belgrade, *Serbia and Montenegro (SCG) Air Traffic Analysis and Forecast.* Final report. Belgrade: FTTE, 2005.

[5] D. Hensher and K. Button (eds.), *Handbook of Transport Modelling.* Amsterdam: Pergamon, 2000.

[6] A. Kanafani, *Transportation Demand Analysis*, New York: McGraw-Hill, 1983.

[7] N. J. Ashford and P. H. Wright, *Airport Engineering.* 3rd ed. New York: John Wiley & Sons, 1992.

[8] L. Castelli, S. Schiratti, and W. Ukovich, "Analysis of passenger demand for air transportation", Work Package 1 of the EUROCONTROL *CARE Innovative Action Project: "Innovative Route Charging Schemes"*. Final Report, 2002.

[9] International Civil Aviation Organization, *Annual Review of Civil Aviation*, 2002.

[10] EUROSTAT, "Statistics in Focus, Highlights of the Panorama of Transport 1970 – 1999", in *Transport* 3/2002, 2002.

[11] E. Kuh and J. Meyer, "How extraneous are extraneous estimates?", in *The Review of Economics and Statistics.* Vol. 39, No. 4, 1957, pp. 380-393.

[12] B. Baltagi and M. Griffin, "Short and long run effects in pooled models", in *International Economic Review.* Vol. 25, No. 3, 1984, pp. 631-645.

[13] T. E. Dielman, *Pooled Cross-Sectional and Time Series Data Analysis.* New York: Marcel Dekker, Inc, 1989.

[14] D. L. Harnett and J. L. Murphy, *Statistical Analysis for Business and Economics.* 3rd ed, Reading, Massachusetts: Addison-Wesley Publishing Company, 1985.

[15] T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics for Business and Economics.* New York: John Wiley & Sons, 1972.

[16] C. Chatfield, *Time-Series Forecasting.* Boca Raton: Chapman & Hall/CRC Press, 2001.

[17] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis.* Boca Raton: Chapman & Hall/CRC, 2003.

[18] J. Kmenta, *Elements of Econometrics.* 2nd ed. New York: Macmillan Publishing Company, 1986.

[19] T. Wallace and A. Hussain, "The use of error components models in combining cross-section with time-series data", in *Econometrica.* Vol. 37, No. 1, 1969, pp. 55-72.

[20] A. H. Choudhury, R. Hubata, and R. D. St. Louis, "Understanding time-series regression estimators," in *The American Statistician.* Vol. 53, No. 4, 1999, pp. 342-348.

[21] R. E. Quandt and K. H. Young, "Cross-sectional travel demand models: estimates and tests," in *Journal of Regional Science.* Vol. 9, No. 2, 1969, pp. 201-214.

Radosav B. Jovanović (1978) is at the moment a full-time student at the MSc Transport Planning and Management programme, Transport Studies Group, University of Westminster, London. He graduated from the Air Transport Department, Faculty of Transport and Traffic Engineering, University of Belgrade, Serbia and Montenegro, in 2004. The principal fields of his interest include transport demand analysis, transport planning and transport economics, in particular air transport.