TIME SERIES MODELS TO FORECAST AIR TRANSPORT DEMAND: A STUDY ABOUT A REGIONAL AIRPORT

Alberto Andreoni, Maria Nadia Postorino*

Department of Computer Science Mathematics Electronics and Transportation
Engineering Faculty- Mediterranea University of Reggio Calabria (Italy)
*Email: alberto.andreoni@.unirc.it; * corresponding author: npostorino@unirc.it*

Abstract: In this work, both univariate and multivariate air transport demand ARIMA models are proposed to estimate the demand levels about Reggio Calabria regional airport (South of Italy), in order to make available a tool able to analyse the impact of recent modifications in the supply (new links, new destinations and lower fares) that are expected to produce an increase in the air transport demand, also due to induced trips generated by the new supply. Differences between the univariate and multivariate models are highlighted. The models calibrations have been obtained by using the Box-Jenkins procedure.

Keywords: ARMA, Air Traffic, Time series analysis; Forecasts; Trends.

1. INTRODUCTION

The increase in air transport demand in the last few decades, also helped by the deregulation policy, has had the major effect of increased transport services offered by different air carriers and has resulted in increasing congestion levels both in the airways and at airports. As an immediate effect of deregulation, the service offered to users in terms of trip organization and costs has changed rapidly and various alliances and mergers have occurred, together with the emergence of new air carriers on the market (Janic, 2000). The above considerations show the importance of correct analyses of the demand characteristics and its evolution in time, thereby allowing the service supplied by both airport managers and air carriers to be designed more effectively. Usually, the estimation of the air transport demand can be obtained by different models/methods, among which national(multi-mode) models, time series models and market surveys (RP – SP methods).
Multi-mode models proposed in the literature to forecast travel demand at a national level can be used to obtain an estimate of the air travel demand, by using explicitly behavioural models in the class of random utility models to simulate mode choice (see for example Cascetta *et al.*, 1995). However, they do not analyse the temporal evolution of demand nor the specificity of the regional airports subject to increases in specific demand segments.

Time series models have been widely used, among the others Melville (1998), Karlaftis and Papastavrou (1998), Abed *et al.* (2001), Postorino and Russo (2001), Hensher (2002), Postorino (2003), Inglada and Rey (2004), Ling Lai and Li Lu (2005). However, mainly simple autoregressive time series models have been used, even if with explanatory variables, while there are very few examples of ARIMA models and no one on the calibration of univariate and multivariate ARIMA models in the specific topic of the air transport demand simulation for a regional airport.
Other techniques that could be used are the Neural Networks (NN) models, a well known and widely treated tool inspired to human brain they try to simulate. Recently, also fuzzy-NNs approaches have been used in the transport field, mainly to forecast mode choice (Pribyl and Goulias, 2003; Sadek, *et al.*, 2003; Postorino and Versaci, 2006). However, NNs do not allow the explicit values of the parameters to be obtained, so the interpretation of the model in terms of elasticity values, parameters ratios and so on cannot be achieved.
The purpose of this paper is to calibrate and compare univariate and multivariate trend models to estimate the demand levels about Reggio Calabria regional airport (located in the South of Italy). The application is particularly interesting because due to the fare policy adopted by the main airline company operating at the airport and the variations in the number and schedule of the flights, the passenger demand at the

airport changed in the last ten years against expectations (a promising positive trend has been followed by a very strong demand reduction). Recent modifications started by the local airport authority in the supply (new links, new destinations and lower fares) are expected to produce an increase in the air transport demand, also due to induced trips generated by the new supply.

This study was partially developed within the research project on "Guidelines for the planning of the development of Italian regional airports", financed by the Italian Ministry of Higher Education and Research.

2. TIME SERIES MODELS

A time series is a stochastic process where the time index takes on a finite or countable infinite set of values. A stochastic process is an ordered and infinite sequence of random variables: if the time index t assumes only integer values, then it is a discrete stochastic process. To describe such a stochastic process, its mean and its variance are used as well as two functions, i.e. the AutoCorrelation Function (ACF) ρ_k, k being the lag, and the Partial AutoCorrelation Function (PACF) π_k, k being the lag. The ACF is a measure of the correlation between two variables composing the stochastic process, which are k temporal lag far away; the PACF measures the net correlation between two variables which are k temporal lag far away.

ARMA (AutoRegressive Moving Average) models are a class of stochastic processes expressed as (Box and Jenkins, 1970):

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = a_t - \sum_{j=1}^{q} \vartheta_j a_{t-j}$$

where ϕ and θ are the model parameters, p and q are the order of the AutoRegressive (AR) and Moving Average (MA) processes respectively. If the B operator such as $X_{t-1} = BX_t$ is introduced, the general form of an ARMA model can be written as:

$$\phi(B) \cdot X_t = \theta(B) \cdot a_t$$

To estimate these models, some conditions should verify: the series must be stationary and ACF and PACF must be time-independent. The non stationarity in variance can be removed if the series is transformed with the logarithmic function. The non stationarity in mean can be removed by using the operator $\nabla = 1\text{-}B$ applied d times in order to make the series stationary. In this way, the ARMA model becomes an ARIMA (AR Integrated MA) model:

$$\nabla^d \phi(B) \cdot X_t = \vartheta(B) \cdot a_t$$

For a given set of data, the Box-Jenkins approach (Box and Jenkins, 1970) is the most known method to find an ARIMA model that effectively can reproduce the data generating the process. The method requires three stage: identification, estimation and diagnostic checking (fig. 1).

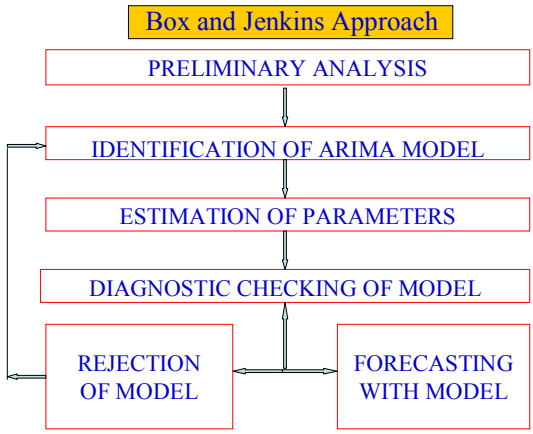

Fig. 1. The Box-Jenkins methodology.

Preliminarily a data analysis should be carried out in order to verify the presence of outliers. The identification stage provides an initial ARIMA model specified on the basis of the estimated ACF and PACF, starting from the original data:

1. if the autocorrelations decrease slowly or do not vanish, there is non stationarity and the series should be differenced until stationarity is obtained. Then, an ARIMA model can be identified for the differenced series;
2. if the process underlying the collected series is an MA(q), then the ACF ρ_k is zero for k>q and the PACF is decreasing;
3. if the process underlying the collected series is an AR(p), then the PACF π_k is zero for k>p and the ACF is decreasing;
4. if there is no evidence for an MA or an AR then a mixture ARMA model may be adequate.

Several statistical tests have been developed in the literature to verify if a series is stationary, among these, the most widely used is the Dickey-Fuller test (Makridakis *et al.*, 1998). After an initial model has been identified, the AR and MA parameters have to be estimated, generally by using least squares (LS) or maximum likelihood (ML) methods. The choice of the AR component order derives from the analysis of the PACF correlogram; for large sample size, if the order of the AR is p, the estimate of the partial autocorrelations π_k are approximately normally distributed with mean zero and variance 1/N for k>p, where N is the sample size. The significance of the residual autocorrelations is often checked by verifying if the obtained values are within two standard error bounds, $\pm 2/\sqrt{N}$, where N is the sample size (Judge *et al.*, 1988). If the residual autocorrelations at the first N/4 lags are close to the critical bounds, the reliability of the model should be verified. Another test that can be used is the Ljung and Box one (1978):

$$Q = N \cdot (N+2) \cdot \sum_{k=1}^{m} (N-k)^{-1} \cdot \left[\rho_{\hat{a}}(k)\right]^2$$

where $\rho_{\hat{a}}(k)$ are the autocorrelations of estimation residuals and k is a prefixed number of lags. For an ARMA (p, q) process this statistic is approximately χ^2 distributed with (k-p-q) degrees of freedom if the orders p and q are specified correctly.

To check the normality of the residuals, the Jarque-Bera test (JB) can be used:

$$JB = \frac{N - n_p}{6} \cdot \left(S^2 + \frac{(K-3)^2}{4} \right)$$

where S is a measure of skewness, K is a measure of Kurtosis, n_p is the number of parameters and N is the sample size. This test verifies if the skewness and kurtosis of the time series are different from those expected for a normal distribution. Under the null hypothesis of normal distribution, the JB test is approximately χ^2 distributed with two degrees of freedom.

Starting from a univariate ARIMA model, some explanatory (or independent) variables can be inserted. In this case, the dependent variable x_t depends on lagged values of the independent variables. The length of the lag may sometimes be known a priori, but usually it is unknown and in some cases it is assumed to be infinite. Generally, if one dependent variable and one explanatory variable are considered, then the model has the form:

$$x_t = \alpha + \beta_0 y_t + \beta_1 y_{t-1} + ... + \beta_P y_{t-P} + e_t \qquad (2.8)$$

To reduce the number of parameters, it can be assumed that $\beta_i=0$ for i greater than a finite number P, called lag length. The obtained models are then called finite distributed lag models, because the lagged effect of a change in the independent variable is distributed into a finite number of time periods.

If $e \sim (0, \sigma^2 I)$ and y_t are fixed, then, based on the sample information **X**, the LS estimator is the best linear unbiased estimator of β_i. If the true lag length P is unknown but an upper bound M is known, then the LS estimator of $\boldsymbol{\beta}' = (\alpha, \beta_0, \beta_1, \ldots, \beta_M)^T$ is inefficient since it ignores the restrictions $\beta_{P+1} = \ldots = \beta_M = 0$. In order to compute P, these sequential hypotheses can be set up:

$$H_0^i : P = M - i, \Rightarrow \beta_{M-i+1} = 0$$

versus

$$H_a^i : P = M - i + 1, \Rightarrow \beta_{M-i+1} \neq 0 \,\big|\, H_0^1, H_0^2, ..., H_0^{i-1}$$

The null hypotheses are tested sequentially beginning from the first one. The testing sequence ends when one of the null hypotheses of the sequence is rejected for the first time. The likelihood ratio statistic to test the i-*th* null hypothesis can be written as:

$$\lambda_i = \frac{SSE_{M-i} - SSE_{M-i+1}}{\hat{\sigma}^2_{M-i+1}} \qquad (2.10)$$

where SSE_P is the sum of the squared errors for a model with lag length P. This statistic has an F-distribution with 1 and (T-M+i-3) degrees of freedom if $H_0^1, H_0^2, ..., H_0^i$ are true.

When the lag has been computed, the explanatory variable can be inserted in the univariate model, in order to derive a so-called multivariate ARIMAX model. In the general case of more than one explanatory variables, the model has the form:

$$\nabla^d \Phi(B) \cdot X_t = \vartheta(B)_t \cdot a_t + \sum_{i=0}^{P_1} \beta_{t-i}^{(1)} y_{t-i}^{(1)} + \sum_{i=0}^{P_2} \beta_{t-i}^{(2)} y_{t-i}^{(2)} + ...$$

where: $y^{(j)}_{t-i}$ is the j-*th* independent variable at the time (t-i) and $\beta^{(j)}_{t-i}$ is the corresponding parameter.

3. THE APPLICATION

Both univariate and multivariate models have been estimated by using the same data (Table 1) that refer to planed /enplaned passengers at the Reggio Calabria airport (sources: Italian Official Statistic Institute ISTAT; Ministry of Infrastructure and Transport; Association of Italian Airports: Assaeroporti).

Table 1 Planed/enplaned passenger at the airport

Year	Pax	Year	Pax
1989	157225	1997	464161
1990	245711	1998	461091
1991	222571	1999	543041
1992	246306	2000	535264
1993	266782	2001	480287
1994	260539	2002	457862
1995	252294	2003	441099
1996	364036	2004	273758

Note that data referring to 2004 should be considered as outlier, because the airport was closed during the months of March, April and May 2004 for some adjustment work on the runway.

Following the Box-Jenkins approach, first of all the ACF and PACF have been estimated (see fig. 2 and 3). The analysis of the correlograms shows that the ACF decreases linearly and the value of the PACF at lag 1 is close to 1, i.e. there is a non stationarity in mean.

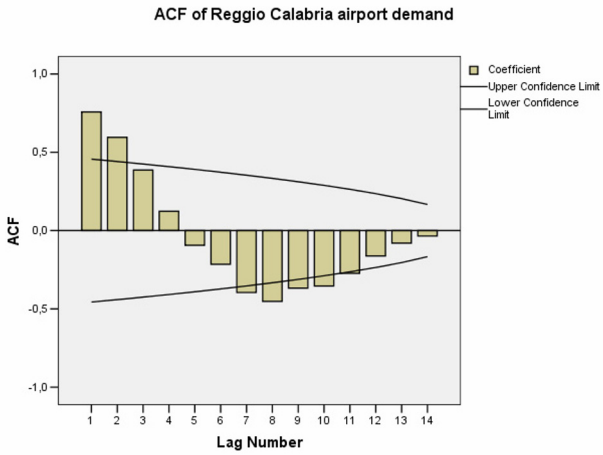

Fig. 2. Correlogram of the ACF.

In order to remove the non-stationarity, the series has been differenced once (in practice, it is almost never necessary to go beyond second order differences, because real data generally involve non-stationarity of only the first or second level) and verified by using

the Dickey-Fuller test.

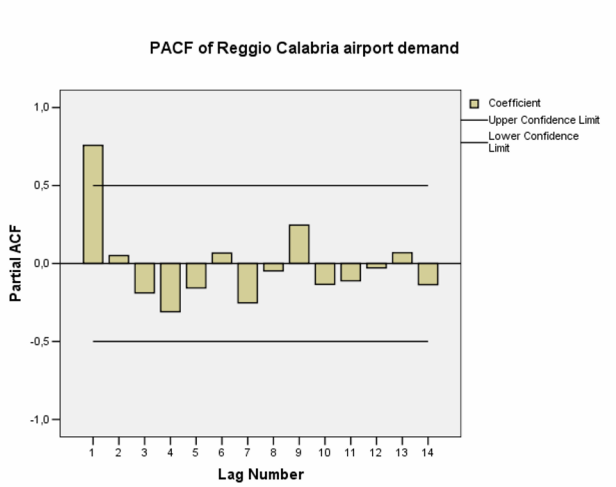

Fig. 3. Correlogram of the PACF.

To remove the non stationarity in variance the time series has been transformed by using the logarithmic function. The estimate of the partial autocorrelation coefficients shows that only π_1 does not fall within the two standard error bounds $\pm 2/\sqrt{N}$, and therefore the order 1 has been chosen for the AR component. The same procedure is applied for the choice of the MA component order, by using the correlogram of the ACF, that suggests a MA(2) component. Then, the identified general model is a ARIMA(1,1,2).

3.1 Univariate models

The presence of the outlier at the year 2004 has suggested to estimate two univariate ARIMA (1,1,2) models: in the first one the outlier has been removed ($ARIMA_{SO}$); in the second one the outlier has been suitably modified ($ARIMA_{CO}$) in order to better follow the natural trend of the series.
The parameters estimates of the $ARIMA_{SO}$ model are reported in Table 2, and the estimated series w.r.t. the true one is depicted in fig. 4 (note that the figure reports forecast till 2006, used for the testing application described in section 3.3). The residuals series obtained has been used to carry out the diagnostic checking, particularly the Ljung-Box test can be considered satisfied for k=12 and the Jarque-Bera test that provides a value of 0,31, i.e. the null hypothesis of residual normality can be accepted.

Table 2 $ARIMA_{SO}$: model parameters

	Parameter	Value
AR1	Φ	0,909
MA1	θ_1	0,858
MA2	θ_2	-0,053

The $ARIMA_{CO}$ requires a suitable estimate of the outlier. This has been performed by means of a monthly AR(1) model (table 3), using the time series monthly data of the latest years. Data referred to 2004 have been forecast by the model and then the $ARIMA_{CO}$ model has been estimated (table 4, fig. 4).

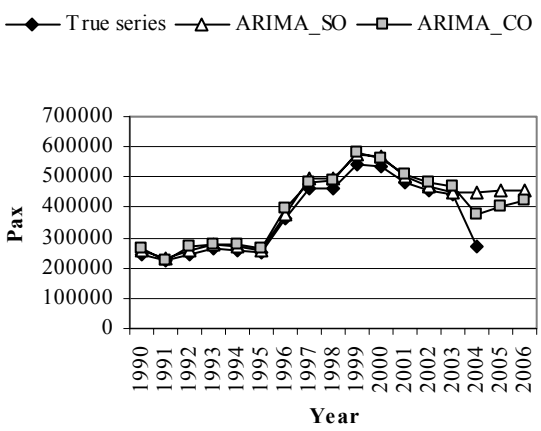

Fig. 4. True and ARIMA estimated series.

Again, the Ljung-Box test can be considered satisfied for k=12, and the Jarque-Bera test provides a value of 0,11, i.e. acceptation of the null hypothesis.

Table 3 AR(1) model parameters

	Parameter	Value
AR1	φ	0,665
Constant	c	10220

Table 4 $ARIMA_{CO}$ model parameters

	Parameter	Value
AR1	Φ	-0,410
MA1	θ_1	-0,470
MA2	θ_2	-0,846
Constant	c	0,056

The comparison between the univariate models shows that both models fit well the true series and are statistically satisfactory. It is important to notice that the estimated value 2006 is the same for both models, so they validate mutually (fig. 4). Moreover, the official passenger data for the year 2005 for the airport of Reggio Calabria, not used to estimate the models and then considered a hold-out sample data, is 398089 (source: Assaeroporti, www.assaeroporti.it), which confirms the goodness of the estimated models.

Table 5 Parameters of the ARIMAX model

	Parameter	Value		Parameter	Value
AR1	Φ	0,44	$\ln I_{t-2}$	α_3	0,21
MA1	θ_1	1,08	$\ln I_{t-3}$	α_4	-1,33
MA2	θ_2	-0,09	$\ln I_{t-4}$	α_5	1,22
$\ln m_t$	δ	1,22	$\ln I_{t-5}$	α_6	0,26
$\ln I_t$	α_1	-0,004	$\ln I_{t-6}$	α_7	0,002
$\ln I_{t-1}$	α_2	3,45	Constant	κ	-0,08

3.2 Multivariate model

Starting from the univariate models, a multivariate model with two explanatory variables (*pro capite* income, I_t, and yearly number of movements from/to the Reggio Calabria airport, m_t) has been considered. The sequential testing procedure previously described

allows the *P* values for both variables to be identified; particularly, the demand in the year *t* depends on the movements in the same year *t* and on the income from year *t* to year *t*-6. The results of the model estimation are reported in Table 5 and fig. 5.

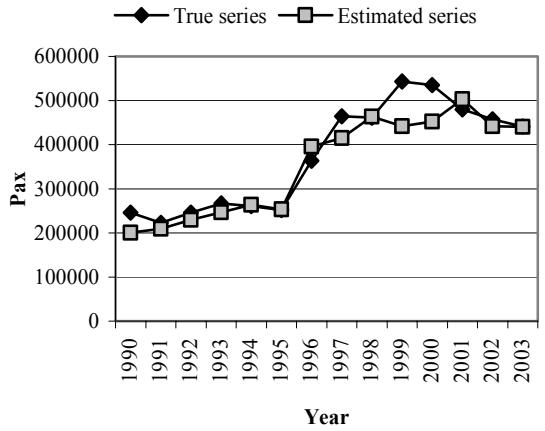

Fig. 5. True and ARIMAX estimated series.

3.3 Some application tests

Starting from the estimated models, particularly the multivariate models, some tests have been carried out in order to provide forecasting of the demand level both for the current year and the next year, and to verify the potential demand increase due to the new offered services with the capacity in terms of offered seats. At the end of 2005 and beginning of 2006, the local airport authority has promoted the introduction of new links with low-cost carriers from/to the airport of Reggio Calabria. Specifically, from the end of December 2005, the low-cost carriers Interstate Airlines and AlpiEagles have set up new links to Venice, Turin and Bologna. At the beginning of 2006, new links connect the city of Reggio Calabria to Bergamo (air carrier: MyAir.com) and to Rome, Genova and Pisa (air carrier: Interstate Airlines).

Estimation of the demand level for the current year 2006 and for the next year 2007 by using the multivariate model requires the knowledge of the income and movements explanatory variables.

The income for the years 2006 and 2007 has been obtained by means of an ARIMA model ($ARIMA_{inc}$). Specifically, the available data on income have been used to build an income time series univariate model, starting from the hypothesis that the boundary conditions are stable. As before in this paper, the model has been estimated by using the Box-Jenkins methodology. The analysis of the income series shows the presence of non stationarity, that have been removed by differencing twice. The Dickey-Fuller test applied to the differenced series confirms its stationarity.

Table 6 Parameters of $ARIMA_{inc}$

	Symbol	Parameter
AR1	Φ	-0,979
MA1	θ_1	-0,800

The identification of the $ARIMA_{inc}$ model has been performed by analyzing the ACF and PACF correlograms, that suggest an AR(1) component and a MA(1) component. The results of the estimated $ARIMA_{inc}$ are reported in Table 6.

With reference to the second explanatory variable, the number of movements for the year 2004 and 2005 has been obtained by official data (source: Assaeroporti, www.assaeroporti.it), while the number of movements for the current year 2006 has been computed by assuming the available data for the first forth months, and assuming that the remaining months have the same number of movements of April. This can be considered realistic, because at the moment there is not any developing plan about new links (more destinations) or flights (increase in the frequency). The same value of movements has been considered for the year 2007, starting from the previous considerations.

The ARIMA model applied to forecast the demand level for the current and the next years provides respectively 709468 passengers (year 2006) and 741248 passengers (year 2007).

Following the estimated values, there is a growth of 78% w.r.t. 2005, whereas for the year 2007 there is an increase of 4,5% w.r.t. 2006. These results can be considered reliable, because the first months of the year 2006 have registered a considerable increase in the number of planed/enplaned passengers at the Reggio Calabria airport, due to the significant increase of the air transport offer, particularly if compared with the offered service in the previous ten years. After this encouraging answer of the demand to the new supply system for the current year, the successive more moderate increase of 4% fully agrees with the forecast demand rate provided by Eurocontrol (www.eurocontrol.int) and IATA (www.iata.org) for the European market.

Finally, in order to verify the ratio between the forecast demand level and the capacity in terms of offered seats, an estimate of the seat number for the year 2006 has been done; such value has been obtained by assuming the same hypotheses as before, i.e. the number of flights on April 2006 remain the same until December 2006. Then, 1164522 seats have been estimated; given that the estimated demand level for the same year is 709468, the average load factor is 0,61. Then, the actual supply can be considered sufficient to satisfy the forecast demand.

Furthermore, the number of movements for the current year is completely beneath the runway capacity, and generally the landside capacity, thus suggesting a potential for growing both in terms of supply and demand.

4. CONCLUSIONS

The comparison between the univariate and the ARIMAX models shows that both the models provide satisfactory results, even if the univariate models fit better than the ARIMAX model when there are some peaks (fig. 6). However, it is not possible to assert in absolute that the estimated univariate models are

better than the multivariate models and vice versa. As obtained in this study, the better forecasting power of univariate models is offset by its limits of validity, which depends on the stability of the boundary conditions. On the other hand, even if the multivariate model solves this problem with explanatory variables, however their time series are often difficult to find.

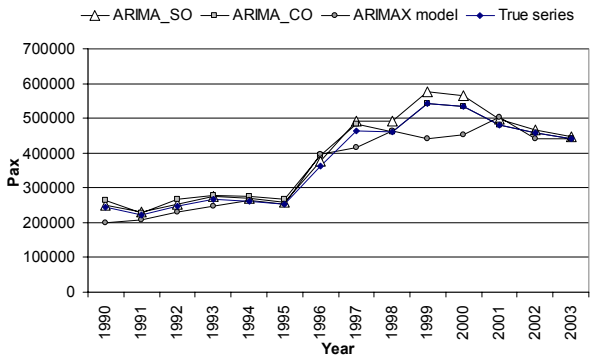

Fig. 6. Trends of the original and estimated series.

That is the reason why no more independent variables (as for example the fare) have been introduced, given that the data are not always available for the examined period. In any case, the estimated model shows a reasonable explanation power and it can be used to test policies about the development of the Reggio Calabria airport. Particularly, some tests have been carried out following the current development plan of the airport authorities, in order to verify the limits to the airport growth.

On the contrary, the univariate models can only forecast the demand level all the underlying conditions being the same and then they cannot be used to simulate the effects of different policies.

Further developments concern the use of more explanatory variables, such as the number of served destinations and the fare (this latter should be estimated by using suitable approaches), as well as the implementation of a specific procedure to test and evaluate the effects of different developing policies, particularly in terms of both landside capacity (due to the forecasted demand increase) and airside capacity (due to infrastructure characteristics and ATC systems that set a limit to the maximum number of movements at the airports).

REFERENCES

Abed S. Y., Ba-Fail A. O., Jasimuddin S. M. (2001). An econometric analysis of international air travel demand in Saudi Arabia. *Journal of Air Transport Management 7*, 143-148.

Box G. E. P., Jenkins G. M. (1970). *Time-Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Cascetta, E., Biggiero, L., Nuzzolo, A., Russo, F.(1995). Passenger and freight demand models for the Italian national transportation system. *Proceedings of the 7th WCTR*, Sydney, Australia.

Hensher D. A. (2002). Determining passenger potential for a regional airline hub at Canberra International Airport. *Journal of Air Transport Management 8*, 301–311.

Inglada V., Rey B. (2004). Spanish air travel and September 11 terrorist attacks: a note. *Journal of Air Transport Management 10*, 441–443.

Italian Ministry of Transport (1998). *National Transport Analysis*. Istituto Poligrafico di Stato, Rome.

Janic M., (2000). An assessment of risk and safety in civil aviation. *Journal of Air Transport Management 6* (**1**), 43-50.

Jarque C.M., Bera A.K., (1982). Model specification tests: a simultaneous approach. *Journal of econometrics 20*, 59-82.

Judge G. G. (1988). *Introduction to the Theory and Practice of Econometrics, 2nd Edition*. Wiley & Sons, Inc.

Karlaftis, M.G., Papastavrou, J.D.(1998). Demand characteristics for charter air-travel. *International Journal of Transport Economics, XXV* (**3**), 19-35.

Lim C., McAleer M. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management 23*, 389-396.

Ling Lai, Li Lu (2005). Impact analysis of September 11 on air travel demand in the USA. *Journal of Air Transport Management 11*, 455–458.

Ljung M. G. and Box G. E. P. (1978). On a measure of lack of fit in time series model. *Biometrika*, 65, 297-303.

Makridakis S., Wheelwright S.C., Hyndman R.J. (1998). *Forecasting. Methods and Applications. Third Edition*. Wiley & Sons, Inc.

Melville, J.A.(1998). An empirical model of the demand for international air travel for the Caribbean region, *International Journal of Transport Economics XXV* (**3**), 313-336.

Postorino M.N., Russo F. (2001). Time series uni-mode or random utility multi-mode approach in national passenger models: the impact on the Italian air demand forecast. *Proceedings of the European Transport Conference (PTRC), Cambridge.*

Postorino M.N. (2003). A comparison among different approaches for the evaluation of the air traffic demand elasticity. *Proceedings of Sustainable Planning and Development Conference, WIT press*.

Postorino M.N., Versaci M. (2006). A Neuro-Fuzzy Approach to Simulate the User Mode Choice Behaviour in a Travel Decision Framework. Forthcoming on *International Journal of Modelling and Simulation*, ACTA Press.

Pribyl O., Goulias K. G. (2003). Application of adaptive neuro-fuzzy inference system to analysis of travel behaviour. *Transportation Research Record 1854*, 180-188.

Sadek A. W., Spring G., Smith B. L. (2003). Towards more effective transportation applications of computational intelligence paradigms. *Transportation Research Record 1836*, 57-63.