SENTIMENT ANALYSIS AND ARTIFICIAL NEURAL NETWORKS-BASED ECONOMETRIC MODELS FOR TOURISM DEMAND FORECASTING

Raffaella Folgieri
Tea Baldigara
Maja Mamula

Abstract
Purpose – This is the second step of a previous paper (Folgieri et al., 2017), where we modelled and applied a backpropagation Artificial Neural Network (ANN) to forecast tourists arrivals in Croatia. Tourism is a very important sector of current Countries' economies, and forcasting assumes even more an significant issue to lead the local tourist offer. In this context, early prediction on the tourist inflow represents a challenge as it is an opportunity in developing tourist income. Applying a Machine Learning Method for Decision Support and Pattern Discovery such as ANN, represents an occasion to achieve a greater accuracy if compared to results usually obtained by other methods, such as Linear Regression.
Design – In this paper, we extended the model of the previously used backpropagation Artificial Neural Network, including data from sentiment analysis collected through social networks on the Internet.
Methodology –The accuracy of the neural network has been measured by the Mean Squared Error (MSE) and compared to results obtained applying the ANN without data coming from the sentiment analysis.
Approach – Our approach consists in combining ideas from Tourism Economics and Information Technology, in particular Artificial Intelligence methods, such as Machine Learning and sentiment analysis, throught the Artificial Neural Networks (ANN) we used in our study.
Findings – The results showed that including also data from sentiment analysis, the neural network model to predict tourists arrivals outperforms the previous obtained results.
Originality of the research –The idea to use ANN as a Decision Making tool to improve tourist services in a proactive way or in case of unexpected events is innovative. Adding data from sentiment analysis, we can add also tourists' preferences so considering collective intelligence and collective trends as factors which could influence a prediction.
Keywords Artificial Neural Networks, Econometrics, Forecasting, Artificial Intelligence, Machine Learning, Prediction

INTRODUCTION

The appealing of a country for tourists is determined not only by the advertising of touristic attraction, but also by opinion registered on the web and on the indices related to the general lifestyle of a country. Indeed, visitors' perception about situations, safety when travelling and others' travellers opinions on staying at a destination influences tourists choices. This is the main reason for what we decided to include also these latter variables in our work, aimed at studying a method to obtain an accurate forecast for the touristic demand. To realize this further step, we enlarged the Artificial Neural Network

(ANN) used in our previous work (Folgieri et al., 2017) with the results from sentiment analysis algorithms applied to two selected websites:

- The official https:www//croatia.hr
- https://www.europeanbestdestination.com

As in the previous study, we refer to other studies performed with the aim of obtaining accurate forecasts, such as in (Athanasopuoulus and Hyndman, 2008; Santos and Fernandes, 2001; Peng, Song, Crouch and Witt 2014). Despite of the effort done to find an optimal method for tourism forecast, in most of the research works the previsioning of incoming visitors is based on incomplete data, due to the difficulties in collecting omogeneous indices from official tourism statistics. In our research, we considered and classical considered indices, such as, for example, cumulative data or number of night stays and arrivals, including also other significant variables, i.e. terrorism index, safety, temperature, mood on the internet/social networks, GDP of the visitors' countries, degree of preservation of local natural resources, environment protection, pollution indices and so on. The analysis of moods generated by visitors on the social networks represents, for example, a very significant source of information for tourism forecast (Folgieri and Bait, 2014; Bait, Folgieri and Scarpello, 2015; Folgieri, Bait and Carrion, 2016;). This latter approach allows to collect information having a great impact on tourist demand forecasting.

In this work, we applied ANN to predict tourism demand. Our research hypotheses is that through ANN and including sentiment analysis data, we could improve current used methods in tourism forecast. ANN are currenty investigated by several scholars for tourism forecast referring to different countries (Law and Au, 1999; Fernandes and Teixeira, 2008; Claveria and Torra, 2014). Tourist arrivals and in some cases, the number of overnight stays, are the most used variables to detect tourist demand (Gunter and Önder, 2015, Law, 2000; Fernandes et al., 2008; Cunha and Abrantes, 2013; Teixeira and Fernandes, 2014. In our work we aimed at overcoming this limitation. Consequently, as a a second step of our previous study, we used a backpropagation ANN to model and forecast tourism demand in terms of total overnight stays in Croatia, expanding the datased with sentiment analysis indices provided by the analysis of touristic websites on Croatia.

In paragraph one the adopted ANN approach is shortly described, including used data from statistics and from sentiment analysis; in paragraph two the results are presented and discussed; final paragraph is conclusion with considerations and ideas for further possible development.

1. THE BACKPROPAGATION ANN APPROACH AND THE DATASET

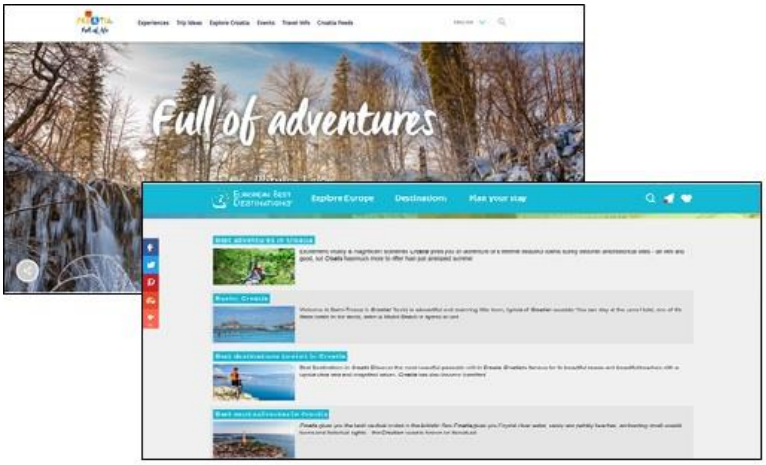
In our work we used an Artificial Neural Network (ANN) as the main algorithm to analyze collected data. ANN have been inspired by biological neural model (Palmer, Montaño and Sesé, 2006) and they present interesting features making these algorithms very attractive: high parallelism, nonlinearity, noise tolerance, learning and generalization characteristics (Basheer and Haimeer, 2000). For detailed information about the ANN used in this work, please refer to our previous one (Folgieri et al., 2017).

Here, we want to recall that we used a backpropagation ANN, consisting in a gradient descent technique that minimizes some error criteria *E*.

As told in the previous paragraph, for our considered ANN, in addition to statistical data, we also used input from sentiment analysis performed on two touristic website on Croatia:

- The official https:www//croatia.hr
- https://www.europeanbestdestination.com

Figure 1: **images from the tourism websites https:www//croatia.hr and https://www.europeanbestdestination.com**



We performed the sentiment analysis (Pang and Lee, 2008; Godbole, Srinivasaiah and Skiena, 2007) on the contents of the selected websites (in the second on the section related to Croatia). Sentiment analysis is an extensive algorithm aiming at detecting the general mood of the content of a website or social networks (i.e. comments posted by users). It allows to determine if the terms used on the web, related to a specific event of place or news, elicit positive, neutral or negative emotions. Sentiment analysis is a tool from Artificial Intelligence (A.I.) providing an objective evaluation of the mood of a text (Nasukawa and Yi, 2003; Lohr, 2012), presenting two significative advantages: on one side it provides better results than the same analysis conduced by humans, often influenced by personal opinion and culture; and on the other side it is useful to analyze large amount of data in a short time, resulting, in this way, more cost effective if compared to the same task performed by individuals.

To perform the sentiment analysis of the content of the magazines, we used two online free tools providing sentiment analysis engines. Specifically, we used free sentiment analysis tools from https:// www.danielsoper.com and https:// www.werfamous.com.

1.1. Dataset

In this work we considered the previous dataset (Folgieri et al., 2017), enlarged with the data coming from the sentiment analysis.

The original dataset is composed by statistical data from the European Union Official statistical website, the Croatian Bureau of Statistics, the ISTAT (Italian National Institute for Statistics), from Trading Economics (http://www.Tradingeconomics.com) and from the United Nations Sustainable Development Solution Network. The considered period starts from 1st January 2007 and ends up with 31st December 2012, as in this range variables are more omogeneous and complete.

The considered variables are listed below:

- TempC, from the Croatian Bureau of Statistics: average temperature, in degrees Celsius, monthly recorded in Croatia;
- HICP, from ISTAT: serves to measure inflation in the Euro area, measured by the MUICP (Monetary Union Index of Consumer Prices) index, as defined in the Council Regulation (EC) No 2494/95 of 23 October 1995 that is the official aggregate of the Euro Area.
- EORD, EUROSTAT source: percentage of contracts received by orders online, considering all the commercial activities with more than 10 employees.
- GDPEP, from EUROSTAT: Percentage of GDP (Gross Domestic Product) reserved for the environmental protection.
- ECUEUR, source EUROSTAT: ECU/EUR exchange rates versus national currency - 1 ECU/EUR = n units of national currency (annual average).
- FTA and DTA, source Croatian Bureau of Statistics: number of Foreign Tourists Arrivals and Domestic Tourists Arrivals, respectively.
- GTI, source Tradingeconomics.com: Croatia Terrorism Index 2002-2015.
- ROH, source United Nations Sustainable Development Solution Network: Ranking Of Happiness.
- Overnight stays, from the Croatian Bureau of Statistics: total overnight stays in Croatian tourist accommodation sector, monthly recorded.

For a complete description, see our previous work (Folgieri et al., 2017).

We wish to recall that we built the matrix of Pearson correlations (Table 1), with the aim to select, as the input layer, the variables resulting more correlated with the variable to predict (i.e. the number of overnight stays in accommodations structures) and less with each other were selected and tested for the input layer.

Table 1: **Matrix of Pearson correlation coefficient.**

	time	tempC	HICP	EORD	GDPEP	ECUEUR	FTA	DTA	GTI	RoH	nights
time	1										
tempC	0.01	1									
HICP	-0.03	-0.01	1								
EORD	0.62	-0.008	-0.41	1							
GDPEP	-0.11	0.03	0.29	-0.46	1						
ECUEUR	0.24	0.04	0.02	0.33	0.63	1					
FTA	0.06	0.87	-0.01	0.03	0.02	0.05	1				
DTA	-0.11	0.91	0.03	-0.16	-0.0008	-0.12	0.90	1			
GTI	-0.25	-0.01	-0.001	0.09	-0.86	-0.71	-0.03	0.07	1		
RoH	0.25	0.04	0.17	0.42	0.26	0.74	0.04	-0.09	-0.38	1	
nights	0.05	0.82	-0.01	0.01	0.01	0.03	0.99	0,87	-0.02	0,03	1

Additional data come from the sentiment analysis performed on the websites https:www//croatia.hr and https://www.europeanbestdestination.com. We considered:

- AMS: the average mood score: 28%
- CS: the confidence (in score): 53 %
- PDW: the number of positive detected words (the average between the two websites): 260

Figure 2: **an example of the results obtained with the sentiment analysis tools**

1.2. The ANN model and implementation

We implementd the backpropagation Artificial Neural Network using the free statistical software environment R (https://www.r-project.org/).

The ANN had the following characteristics:

- 12 input nodes, corresponding to the variable tempC (average monthly temperature, in degrees Celsius), HICP (Croatian Harmonised Indices of Consumer Prices), EORD (percentage of contracts received by orders online), GDPEP (percentage of Gross Domestic Product reserved for the Environmental Protection), ECUEUR (exchange rates versus national currencies), FTA (number of monthly Foreign Tourists Arrivals), DTA (number of monthly Domestic Tourists Arrivals), GTI (Croatia Terrorism Index), RoH (Croatian Ranking Of Happiness), AMS, CS, PDW (the latter three coming from the sentiment analysis);
- 1 hidden layers with 2 nodes;
- 1 output node, for the predictions of the number of monthly overnight stays of incoming tourists in Croatian tourist accommodation sector;
- a sigmoidal activation function in the hidden layer and a linear activation function at the output layer;
- initial random weights.

1.3. Methods

We performed the same steps and adopted the same approach had in our previous work (Folgieri et al., 2017), to compare the obtained results;

- aiming at a monthly prediction, we built one ANN per month. We applied a linear regression to the same data, to allow the comparison of the two approach.

The Mean Squared Error (MSE) has been used to determine the performance of the ANN. The MSE consists in the average of the squares of the errors or deviations that is, rougly speaking, the difference between the estimator and estimated value. It is always not negative and values closer to zero denote a better performance of the ANN.

The linear regression has been also developed through the software environment R.

The data preprocessing consisted in normalizing data through the min-max method and scaling data in the interval [0,1] (this latter interval has been adopted as usually provides better results).

Data were divided into a training-set (70% of the records, randomly selected) and a test-set (the remaining 30% of the records, randomly selected), and then used to run the ANN. We validate both the ANN and the linear regression using a 5-fold cross validation.

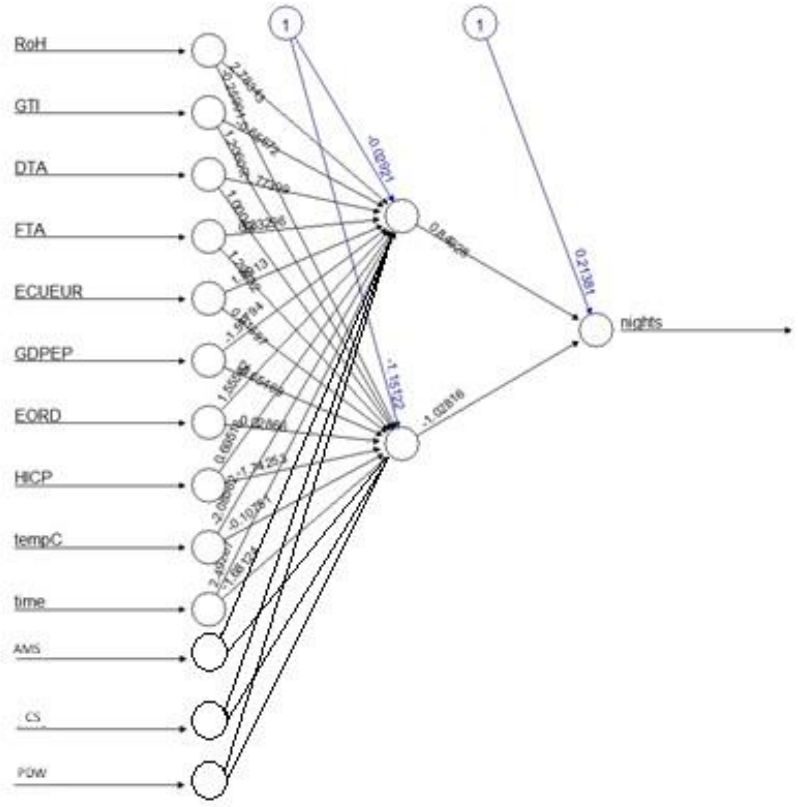
Furthermore, to evaluate if different kinds of data have impact on the results, we also applied ANN and linear regression on the dataset reduced only to number of overnight stays and weather conditions.

2. RESULTS AND DISCUSSION

In the first step, we obtained 12 ANN – one per month – considering all the variables of the data set from 2007 to 2012. In this case the error resulted, in average, 0.0000247.

In the following figure 3, we show, as an example of the outcome of the procedure, one of the ANN applied to the dataset.

Figure 3: **The ANN resulting for prediction of tourist arrivals in August (Error 0.000035) – dataset from 1st January 2007 to 31st December 2012.**



For about the linear regression, we considered the R^2, as it is comparable to the MSE of the ANNs. We obtained, in average, a value of 1 and MSE in average 0.005655. The following Table 2, shows the comparison between all the values of the MSE registered for each ANN and the values of R^2 for each Linear Regression performed on all the dataset for each considered month (forecast).

Table 2: **Comparison between the MSE of the ANNs and the MSE and R^2 of the linear regressions – dataset from 1st January 2007 to 31st December 2012.**

Month	**MSE ANN**	**MSE Linear Regression**	**R^2 Linear Regression**
January	0.000023	0.00567	1
February	0.000037	0.020133	1
March	0.000031	0.01902	1
April	0.00008	0.00899	1
May	0.000002	0.001034	1
June	0.000041	0.006719	1
July	0.000013	0.005042	1
August	0.000025	0.007214	1
September	0.000026	0.002708	1
October	0.000006	0.00150	1
November	0.000004	0.004329	1
December	0.000028	0.005634	1
average	**0.0000247**	**0.005655**	**1**

We can state that in all the considered montly forecasts, the ANN outperform the linear regression technique.

With the aim of evaluating the impact of different kinds of data on the adopted approach, we also trained the ANNs with 7 variables and for the period 2002-2014. Specifically, we considered tempC (average monthly temperature, in degrees Celsius), HICP (Croatian Harmonised Indices of Consumer Prices), GTI (Croatia Terrorism Index), nights (number of monthly nights spent in Croatian touristic accommodation by incoming tourists) and, of course, the variable deriving from the sentiment analysis.

The obtained average error (across the monthly datasets) has been about 0.003176, higher than the previous one, but still lower than the one obtained applying the linear regression technique. This result confirm our consideration in our previous work (Folgieri et al., 2017), showing that the accuracy of the prediction is strongly linked to the amount of information collected.

Furthermore, similar results have been obtained when we considered only the variables temperature and nights (in the period 2002-2014), adding the values obtained from the sentiment analysis: the average error (across the monthly dataset) has been about 0.02 that is, specifically, an average MSE of 0.067, corresponding to an average R^2 equal to 0.9.

CONCLUSION

This paper is the second step of a wider study previously started (Folgieri et al., 2017). To further validate our previous investigation on the use of Artificial Neural Networks as an effective model in predicting tourists arrivals, we enlarged the dataset (consisting in statistical data collected from official tourism websites), including indices from sentiment analysis. This latter has been performed on the websites https:www//croatia.hr and https://www.europeanbestdestination.com.

The inclusion of data from sentiment analysis determined an improvement of the method, confirming on one side that the Artificial Neural Network model in predicting tourists arrivals is a robust method thanks to its low error rate and because it outperforms the linear regression technique; on the other side that the more data are considered, the more accuracy is achieved. Considering the limitation of our study (in future we intend to expand the sentiment analysis), this is an important scientific contribution to the field, opening new possibilities in tourism forecast and, furthermore, giving inspiration to include other hetereogeneous data from different sources (not only statistical and sentiment analysis ones) in future application of this approach.

The presented study also demonstrates, through examples of application to subsets of the collected data, different in period and number of considered variables, that ANNs perform efficiently on heterogeneous data, such as classical statistical values and data coming from the analysis of the mood registered on the Internet (sentiment analysis).

In this paper, we considered the mood elicited by the content of the two website analyzed through the sentiment analysis tools, but in future investigations we could consider also data coming from a reputation analysis on the Internet, including, for example, indices on specific accommodation structures, as we could consider reaction of tourists to unespected events (for example analyzing data from Social Networks such as Facebook or Twitter), having, in this way, the possibility to realize a Decision Making support, for proactive and reactive response in improving tourist services, as we already stated in our previous papers.

REFERENCES

Athanasopoulos, G. and Hyndman, R. (2008). “Modelling and fore-casting Australian domestic tourism”. *Tourism Management*, 29, 19-31.

Bait, M., Folgieri, R. andScarpello, O. (2015). “The use of agent-based models in cognitive linguistics: an approach to Chomsky’s linguistics through the clarion model”. *Journal of Foreign Language Teaching and Applied Linguistics, 1*(3).

Basheer, I. and Hajmeer, M. (2000). “Artificial neural networks:Fundamentals, computing, design, and application”. *Journal of Microbiological Methods*, 43, 3-31.

Claveria, O. and Torra, S. (2014). “Forecasting tourism demand toCatalonia: Neural networks vs. time series models”. *EconomicModelling*, 36, 220-228.

Cunha, L. and Abrantes, A. (2013). *Introdução ao turismo*, Lisboa(5th Ed.).

Fernandes, P. and Teixeira, J. (2008). “Previsão da Série TemporalTurismo com Redes Neuronais Artificiais”. *5. Congresso Luso-Moçambicano de Engenharia - CLME’ 2008 - A Engenharia noCombate à Pobreza, pelo Desenvolvimento e Competitividade*, Maputo-Moçambique.

Fernandes, P., Teixeira, J., Ferreira, J. and Azevedo, S. (2008). "Modelling tourism demand: A comparative study between artificialneural networks and the Box-Jenkins methodology". *Romanian Journal of Economic Forecasting*, 5(3), 30-50.

Folgieri, R. and Bait, M. (2014, January). "The new profile of the virtual tourist- traveler: communicative perspectives and technological challenges". In *Faculty of Tourism and Hospitality Management in Opatija. Biennial International Congress. Tourism & Hospitality Industry* (p. 408). University of Rijeka, Faculty of Tourism & Hospitality Management.

Folgieri, R., Bait, M. and Carrion, J. P. M. (2016). "A Cognitive Linguistic and Sentiment Analysis of Blogs: Monterosso 2011 Flooding". In *Tourism and Culture in the Age of Innovation* (pp. 499-522). Springer International Publishing.

Folgieri, R., Baldigara, T. and Mamula, M., (2017). "Artificial Neural Networks-Based Econometric Models for Tourism Demand Forecasting". In *Tourism in Southern and Eastern Europe 2017*.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). "Large-Scale Sentiment Analysis for News and Blogs". ICWSM, 7, 21.

Law, R. and Au, N. (1999). "A neural network model to forecastJapanese demand for travel to Hong Kong". *Tourism Management*, 20(1), 89-97.

Lohr, S. (2012). "The age of big data". New York Times, 11.

Nasukawa, T., & Yi, J. (2003, October). "Sentiment analysis: Capturing favorability using natural language processing". In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77). ACM.

Palmer, A., Montaño, J. J. and Sesé, A. (2006). "Designing an artifi-cial neural network for forecasting tourism time series". *TourismManagement*, 27(5), 781-790.

Pang, B., & Lee, L. (2008). "Opinion mining and sentiment analysis. Foundations and trends in information retrieval", 2(1-2), 1-135.

Peng, B., Song, H., Crouch, G. and Witt, S. (2014)."A meta-analysisof International tourism demand elasticities". *Journal of TravelResearch*, 1-23.

Santos, N. and Fernandes, P. (2011)." Modelação e caracterizaçãoda procura turística: o caso da região Norte de Portugal".*TÉKHNE - Review of Applied Management Studies*, 9(16), 118-137.

Teixeira, J. and Fernandes, P. (2014). "Tourism time series forecastwith artificial neural networks". *Tékhne Review of Applied Man-agement Studies*, 12, 26-36.

Raffaella Folgieri, PhD, Assistant Professor
Università degli Studi di Milano
Department of Philosophy
Via Festa del Perdono 7, Milan, Italy
+39 02 50312739
E-mail: Raffaella.Folgieri@unimi.it

Tea Baldigara, PhD, Full Professor
University or Rijeka
Faculty of Tourism and Hospitality Management
Department for Quantitative Economy
Primorska 42, 51410 Opatija, Croatia
+385 51 294 684
E-mail: teab@fthm.hr

Maja Mamula, PhD, Assistant Professor
University or Rijeka
Faculty of Tourism and Hospitality Management
Department for Quantitative Economy
Primorska 42, 51410 Opatija, Croatia
+385 51 294 684
E-mail: majam@fthm.hr