Közgazdász Fórum
Forum on Economics and Business
21(135), 66–78.
2018/2

Forecasting the regional unemployment rate based on the Box-Jenkins methodology vs. the Artificial Neural Network approach. Case study of Brașov and Harghita counties

SZILÁRD MADARAS[1]

This paper presents different methods for the forecasting of unemployment rates in two Romanian counties. The stationarity of the monthly unemployment rate time series between January 2000 and November 2016 was examined using the ADF and KPSS tests. Based on time series, a forecast was estimated using two approaches: the Box-Jenkins methodology and the Artificial Neural Network-based NAR model. Results showed a decreasing trend by the end of the forecasted period in all cases, except for the NAR model of Harghita County. Comparing the forecasted values with the officially registered unemployment rates from the same period, we observed that, by the end of the period, the differences between the real and predicted values became higher in the NAR model than in the ARMA model-based forecasting. These results indicate that, in these particular cases, NAR neuron network model-based forecasts fit well if values are estimated for a short-term period, while for medium-term forecasts the ARMA model-based forecasting is more precise.

Keywords: regional unemployment, time-series models, forecasting and prediction methods, Box-Jenkins methodology, Artificial Neural Network.

JEL codes: C32, C53, R15.

Introduction

Time series analysis is an actual topic in regional studies. Approaches differ in the assumptions and models used for testing, i.e. the regions are studied as unique cases using the Box-Jenkins methodology or the group of regions or counties form a panel data base structure. Both of these are generally used in regional time series forecasting, while Artificial Neural Networks (ANN) currently represent a new approach in economic research.

The present study examines the unemployment rate monthly time series in Brașov and Harghita counties (NUTS3 level territorial units for statistics) from Romania. The differences between the two case studies were verified using the main regional indicators, while the employment and unemployment analysis

[1] PhD, assistant professor, Sapientia Hungarian University of Transylvania, Faculty of Economics, Socio-Human Sciences and Engineering, e-mail: madarasszilard@uni.sapientia.ro.

Publisher: Hungarian Economists' Society of Romania, the Department of Economics and Business Administration in Hungarian Language at the Babeş-Bolyai University, and the Partium Christian University
ISSN: 1582-1986 http://econ.ubbcluj.ro/feb/

proved the special situation of unemployment in those counties. In Brașov, a typical urban development-related employment was observed in the services and industry sectors, while Harghita, as a mainly rural county, was characterised by high agricultural employment.

The time series analysis and the forecasting are focusing on these two case studies. The ADF (Augmented Dickey-Fuller) and the KPSS Lagrange Multiplier tests were used to analyse the stationarity of the unemployment rate time series. The unemployment rate was forecasted using the Box-Jenkins methodology and, secondly, an Artificial Neural Network-based NAR model was set up and used for this purpose.

Literature review

Spatial differences, as evidenced by the spatial modelling of unemployment, are one of the actual topics tackled by regional unemployment research. Schanne et al. (2008) forecasted regional unemployment using a spatial GVAR model in the case of the German regions. Madaras (2009) modelled the unemployment rate in the Central Region (NUTS2 level territorial units for statistics) of Romania using the random effect panel regression model. Kryńska (2014) discussed regional employment forecasting methods and presented different forecasting case studies from the regions of Poland, while Mayor et al. (2007) set up shift-share and ARIMA models for forecasting employment in the Spanish regions.

Using the Box-Jenkins methodology as an ARIMA (1, 1, 4) process, Madaras (2014) modelled the number of the unemployed in Romania for the period January 2005–June 2013 and, based on that, performed a medium-term forecasting. The Box-Jenkins methodology was also used for the time-series forecasting of macroeconomic indicators in Romania (Morariu et al. 2009), to forecast regional tourism demand in Spain (Fernandes et al. 2008), and to forecast regional employment in Germany (Longhi et al. 2005).

The Artificial Neural Network (ANN)-based forecasting of the regional tourism demand time series was used by Fernandes et. al. (2008) compared with an ARIMA model estimation. Longhi et al. 2005 used ANN models for regional employment forecasting in Germany and proved that those were useful forecasting tools compared with the maximum likelihood random effect estimator [ML]-based panel model forecasting.

Research methodology

Stationarity analysis is one of the primary subjects of time series analysis, while time dataset-based model identification, which represents a forecasting instrument, is another important research topic.

In this paper, we used two of the most commonly known unit root tests, the ADF (Augmented Dickey-Fuller) and the KPSS Lagrange Multiplier test. Those tests present significant differences regarding the null hypothesis: the first one has the null of non-stationarity, while the second one has the null of stationarity (Kirchgässner–Wolters 2007).

The Box-Jenkins methodology has a long and prestigious past in the field of time series research. The p-order auto-regressive model (AR_p) is based on the assumption of a given time t, the endogenous y_t variable depends on its time delayed values of the previous 1, 2, ... p periods (Kirchgässner–Wolters 2007; Pecican 2006):

$$y_t = a_0 + a_1 y_{t-1} + ... + a_p y_{t-p} + u_t,$$

with u_t being the error term.

The q-order moving average process (MA_q) describes the y_t, as

$$y_t = \mu + b_1 u_{t-1} + ... + b_p u_{t-q},$$

where μ is the mean and $u_{t-1},...u_{t-q}$ are pure random processes, for the previous 1, 2, … q periods (Kirchgässner–Wolters 2007; Pecican 2006).

The autoregressive moving average process (ARMA) with AR order p and MA order q are:

$$y_t = a_0 + a_1 y_{t-1} + ... + a_p y_{t-p} + \mu + b_1 u_{t-1} + ... + b_p u_{t-q} + u_t$$

and the ARIMA (p,d,q) autoregressive integrated moving average process refers to an I (integrated) process.

This model, initially developed by Box and Jenkins in the 1970s, is constructed with the following steps: model identification, i.e. the determination of p, q values, using the autocorrelation function (ACF) and the partial autocorrelation function (PACF), and the d of the I(1) process are specified testing the stationarity of the time series. These tests are followed by the estimation of the model coefficients and model validation (Kirchgässner–Wolters 2007; Pecican 2006).

The selection of the best fitting model is usually based on the Akaike Information Criterion (AIC) values (Pecican 2006). And, in the last step, the thus constructed model is used for the short- or medium-term forecasting of the time series.

The ANN has a wide application in research and the statistical perspective of its implementation was discussed, among others, by Cheng and Titterington (1994) and by Warner and Misra (1996).

In financial and economic time series analyses, the ANN is present as a useful nonlinear, semiparametric model.

The *feed-forward* Artificial Neural Networks are those where the inputs have forward connections to the neurons in the (one or more) hidden layers, reaching the output layer at the end. The information from one layer to other is transmitted by the activation function f_j, generally a logistic function:

$f_j(Z) = \frac{e^z}{1 + e^z}$, where j represents the *j*th node in the hidden layer,

and the feed-forward network is defined as:

$h_j = f_j(\alpha_{0j} + \Sigma_{i \to j} w_{ij} x_i)$, where w_{ij} represents the weights and the $i \to j$ summation, which include all input nodes feeding into j, and α_{oj} is the bias (Tsay 2005).

The first phase of the ANN construction is network building, i.e. determining the number of hidden layers and nodes. The second phase is the training process and, as a result, we have the estimated best fitting biases and weights of the nodes, according to the selected criterion.

In time series estimation, the ANN approach of nonlinear autoregressive models (NAR) has the *d*-period delayed values of y(t) as input:

$$y(t) = f(y(t-1), ..., y(t-d))$$

The time series analysis of the unemployment rate was performed in two counties (Brașov and Harghita) from the Central Region of Romania. These two counties were selected because, according to many regional socio-economic indicators, they were rather different: in Brașov, the share of the inhabitants living in the urban area, the regional gross domestic product, and the number of enterprises per 1,000 inhabitants are all higher than in Harghita County. Major differences could be observed among them in the activity rate, the employment rate and the unemployment rate.

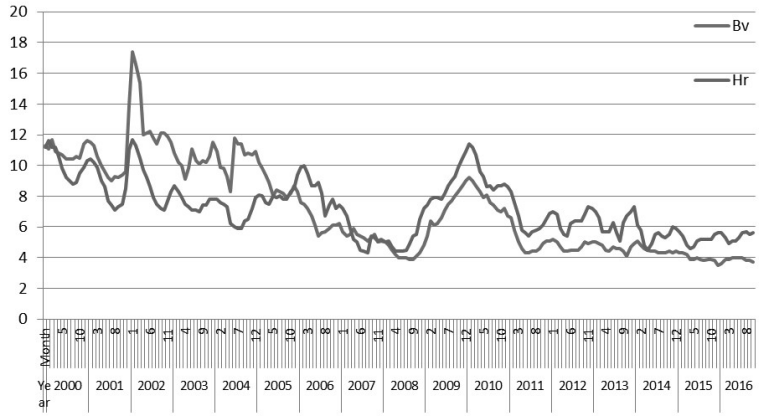
In Brașov County, the high unemployment rate (7.2%) recorded in 2010 was presumably due to the local consequences of the 2008 world financial crisis, but it decreased to 3.6% in 2016. In Harghita County, the unemployment rate was higher (8.8% in 2010 and 5.8% in 2016), due to the greater vulnerability of the local labour market to the same economic impact. Although activity rates and employment rates in Brașov and Harghita counties reached relatively similar

levels in 2016, there were major differences in the structure of the employed population: a high share of agricultural employment in Harghita (23.67%), as opposed to a high share of employment in services in Brașov (51.54%).

The time series of monthly unemployment rates in Brașov and Harghita counties for the period January 2000 – November 2016, used for the calculations below, were obtained from the Tempo Online database of the Romanian National Institute of Statistics (INSSE 2018).

Results

The time series analysis contains the stationarity tests, the Box-Jenkins method and the Artificial Neural Network analysis, as described below. The evolution of unemployment rates in the two counties followed similar trends in the studied period (Figure 1).



Source: author's own design based on INSSE (2018) data

Figure 1. Evolution of unemployment rates in Brașov and Harghita counties (January 2000 – November 2016)

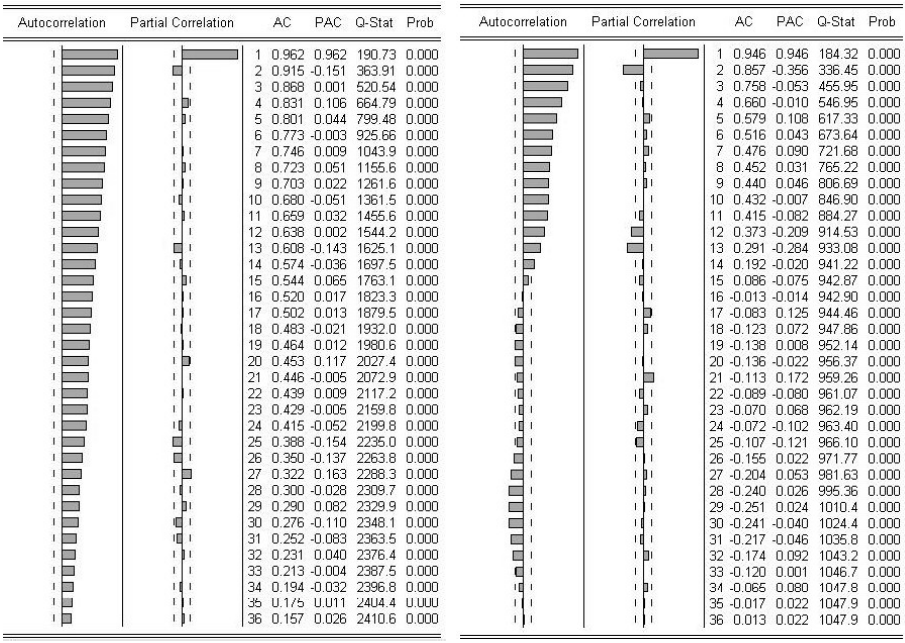
The stationarity of the series was examined using the ADF and KPSS tests (Table 1). Both of the univariate unit root tests suggest that the unemployment rate monthly time series in Brașov is a non-stationary, eventually I(1) series, while in Harghita the results of the ADF test suggest an AR(p) process. The results of the KPSS test are similar.

Table 1. Univariate unit root tests of unemployment rate time series in Brașov and Harghita counties (January 2000 – November 2016)

County	Level		First Difference	
	ADF	KPSS	ADF	KPSS
Brașov	-2.008759(1)	1.435880(11)***	-11.09876(0)***	0.045059(13)
Harghita	-3.724851(1)***	0.081944(10)	-8.765453(0)***	0.035965(1)

Source: author's own calculations based on INSSE (2018) data

In the next step, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) values were calculated for model identification. For the unemployment rate time series from Brașov County, results indicated an AR(2) process (Figure 2a), while for Harghita County the partial autocorrelation test indicated an AR(2) process (Figure 2b). ARMA or ARIMA models were also considered and more tests had to be computed for the identification of the most appropriate model.



Source: author's own calculations based on INSSE (2018) data

Figure 2. Correlogram of unemployment rate time series in Brașov (a) and Harghita (b) counties

The best fitting ARMA model was chosen based on the AIC values: ARMA(1,1) for both counties (Table 2).

Table 2. Akaike information criterion (AIC) values for the estimated models of the unemployment rate time series in Brașov and Harghita counties

Brașov County		Harghita County	
ARIMA	**AIC**	**ARMA**	**AIC**
(2,0,0)	2.891793	(2,0)	2.499080
(1,0,1)	1.899506	(1,1)	1.214645
(1,1,1)	1.920951	(1,2)	1.382784
(2,1,1)	1.966818	(1,3)	1.420065
(2,0,1)	1.961739	(2,1)	1.394078
(2,0,2)	2.900252	(2,2)	2.314901
(2,0,3)	2.862625	(2,3)	2.435688

Source: author's own calculations based on INSSE (2018) data

The ARMA(1,1) model statistics of the unemployment rate time series from Brașov and Harghita counties are presented in Table 3 and we can see that, in the first case, the R-squared is equal to 0.96, while in the second case the R-squared is 0.94, which means that both estimated models explain the time series well. In the next step, the models are used for a medium-term forecast of the time series.

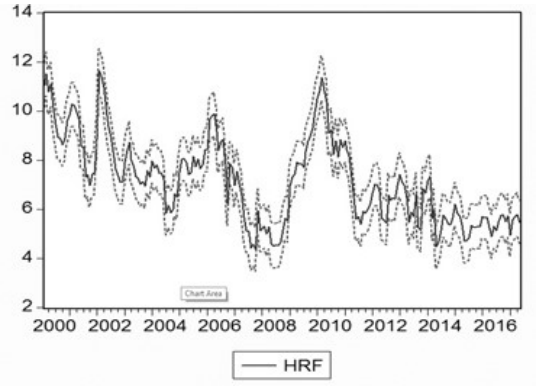
Table 3. ARMA models of monthly unemployment rate time series in Brașov and Harghita counties (January 2000 – November 2016)

Dependent variable	**Unemployment rate in Brașov County**			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.140487	1.643046	3.737259	0.0002
AR(1)	0.964376	0.018738	51.46694	0.0000
MA(1)	0.260068	0.069814	3.725161	0.0003
R-squared	0.956994			
Adj. R-squared	0.956562			
AIC	1.899506			

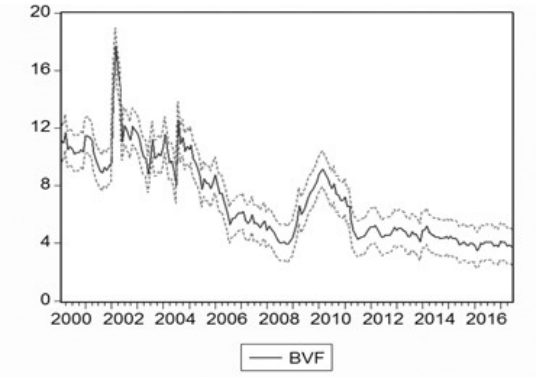
Dependent variable	**Unemployment rate in Harghita County**			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.844177	0.661996	10.33869	0.0000
AR(1)	0.931595	0.024561	37.93010	0.0000
MA(1)	0.428266	0.065383	6.550110	0.0000
R-squared	0.941462			
Adj. R-squared	0.940874			
AIC	1.214645			

Source: author's own calculations based on INSSE (2018) data

The ARIMA-based forecasting of the unemployment rates was carried out for the period December 2016 – May 2017. The same trends resulted as observed in the previous years: high values in first months of the year and decreasing values by the end of the period (Figure 3).



a.



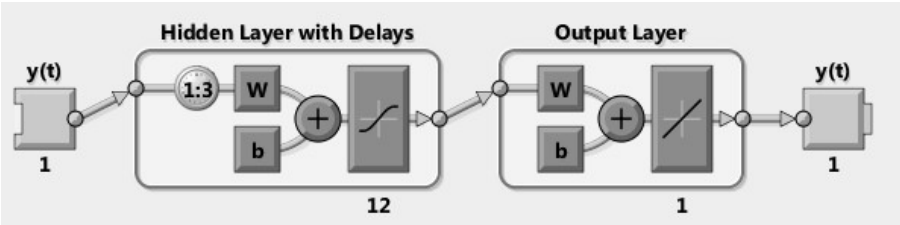
b.

Source: author's own calculations based on INSSE (2018) data

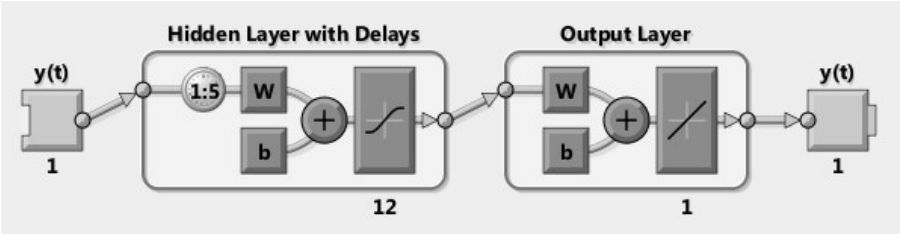
Figure 3. ARIMA-based forecasting of monthly unemployment rates in Brașov (a) and Harghita (b) counties

For the prediction of the natural logarithmed values of unemployment rate time series in Brașov and Harghita counties, I built up the ANN-based NAR model. The time series was divided into three groups: the training group with 173 observations, the validation group with 10 observations, and the testing group

with 20 observation values. In the case of the logarithmed unemployment rate in Brașov County, the network architecture was set to 1 input, 12 hidden neurons and d = 3 number of delay, while in the case of the logarithmed unemployment rate in Harghita County it was set to 1 input, 12 hidden neurons and d = 5 number of delay (Figure 4). With this neuron network architecture and lag values, the errors are autocorrelated.



a.



b.

Source: author's own design

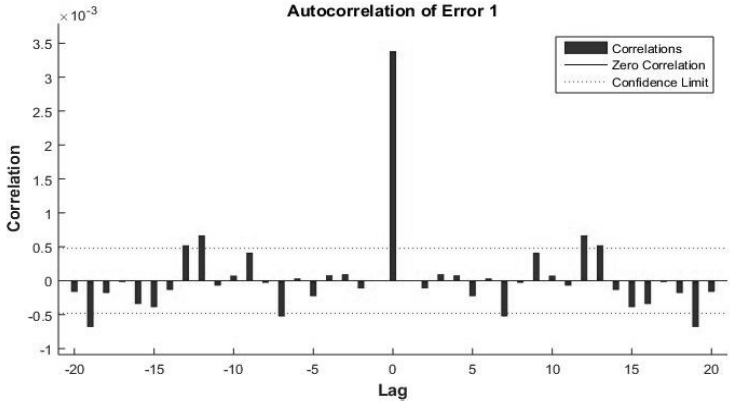
Figure 4. NAR neuron network construction of monthly unemployment rates in Brașov (a) and Harghita (b) counties

The time steps were divided into three groups: the training group (85%), the validation group (5%), and the testing group (10%). In both Artificial Neural Networks, the Levenberg-Marquardt training algorithm was used. The prediction errors became uncorrelated, after a retraining process (Figure 5), and this way the final form of the ANN models was validated for the second forecasting.

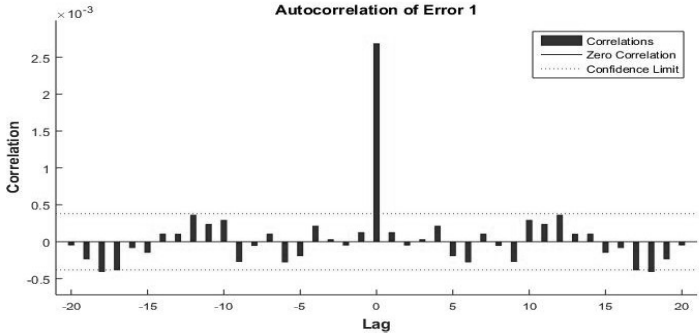
The two neuron network models presented above were used for a medium-term forecasting of unemployment rates in Brașov and Harghita counties.

Both the ARIMA model and the NAR model forecasted higher unemployment rates for Harghita than for Brașov County (Table 4), as that was the most common

characteristic between January 2000 and November 2016 (Figure 1). In all cases, with the exception of the NAR model for Harghita County, results show a decreasing trend by the end of the forecasting period, which is similar to the previous years' periodicity: higher unemployment rates in the winter and lower in the summer.



a.



b.

Source: author's own calculations based on INSSE (2018) data

Figure 5. Autocorrelation errors of logarithmed unemployment rate time series in Brașov (a) and Harghita (b) counties

Table 4. Results of the Box-Jenkins and the Artificial Neural Network forecasting of monthly unemployment rates in Brașov and Harghita counties

Month	ARMA models		NAR models	
	Brașov	Harghita	Brașov	Harghita
	%	%	%	%
2016M12	3.740316	5.736696	3.8851	5.4722
2017M01	4.143789	4.946636	3.7602	5.7280
2017M02	4.038858	5.556964	4.0287	5.3310
2017M03	4.066147	5.703539	3.8241	5.8240
2017M04	3.814162	5.776752	4.1868	5.0325
2017M05	3.879695	5.473426	3.8578	6.0922

Source: author's own calculations based on INSSE (2018) data

In the end, I compared the forecasted values to the officially registered unemployment rates from the same period (Table 5).

Table 5. Comparison of forecasted values with the officially registered unemployment rates

Month	Registered value of the unemployment rate		Difference to ARMA model forecasting		Difference to NAR model forecasting	
	Brașov	Harghita	Brașov	Harghita	Brașov	Harghita
	%	%	%	%	%	%
2016M12	3.60	5.80	-0.14	0.06	-0.29	0.33
2017M01	3.60	5.80	-0.54	0.85	-0.16	0.07
2017M02	3.60	5.90	-0.44	0.34	-0.43	0.57
2017M03	3.50	5.20	-0.57	-0.50	-0.32	-0.62
2017M04	3.20	4.90	-0.61	-0.88	-0.99	-0.13
2017M05	3.20	4.80	-0.68	-0.67	-0.66	-1.29

Source: author's own calculations based on INSSE (2018) data

We can observe that the officially registered unemployment rates follow the same trend as in the two estimated models. By the end of the forecasting period, the differences between the real and predicted values were higher for the NAR model-based forecasting than for the ARMA model-based forecasting, while at the beginning they were almost the same.

Conclusions

In this paper, two forecasting models were developed for predicting monthly unemployment rates in Brașov and Harghita counties, using the time series for the period January 2000 – November 2016. Based on the Box-Jenkins methodology,

ARMA(1,1) type models resulted for both counties. Secondly, NAR neuron network models were constructed by 1 input and 12 hidden neurons for both counties, but with different numbers of delay. The error autocorrelation test results indicated that these types of NAR models were most appropriate for the time series.

Results showed a decreasing trend by the end of the forecasted period in all cases, except for the NAR model of Harghita County.

Comparing the forecasted values with the officially registered unemployment rates from the same period, we can observe that, by the end of the forecasting period, the differences between the real and predicted values became higher for the NAR model than for the ARMA model-based forecasting. These results indicate that neuron network model-based forecasts fit well if values are estimated for a short-term period, while for medium-term forecasts the ARMA model-based forecasting is more precise.

My results confirm the findings of Fernandes et al. (2008) that the NAR or other neuron network-based models are useful alternatives to the Box-Jenkins methodology for regional economic data time series forecasting.

In future studies, the different types of neuron network models are recommended to be analysed in comparison to the commonly used Box-Jenkins methodology to identify their usefulness and limitations in regional economic research.

References

Cheng, B.–Titterington, D. M. 1994. Artificial Neural Networks: A review from a statistical perspective. *Statistical Science* 9, 2–54.

Fernandes, P.–Teixeira, J.–Ferreira, J. M.–Azevedo, S. G. 2008. Modelling tourism demand: A comparative study between Artificial Neural Networks and the Box-Jenkins methodology. *Romanian Journal of Economic Forecasting* 2008(3), 30–50.

Goschin, Z.–Constantin, D.–Roman, M.–Ileanu, B. 2008. The current state and dynamics of regional disparities in Romania. *Romanian Journal of Regional Science* 2(2), 80–105.

INSSE 2018. http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table, downloaded: 18.02.2018.

Kirchgässner, G.–Wolters, J. 2007. *Introduction to Modern Time Series Analysis*. Berlin, Heidelberg: Springer–Verlag.

Kryńska, E. 2014. *Employment Forecasting. Social Policy Thematic Issue No 2.* Warsaw: Institute of Labour and Social Studies.

Longhi, S.–Nijkamp, P.–Reggianni, A.–Maierhofer, E. 2005. Neural Network Modeling as a Tool for Forecasting Regional Employment Patterns. *International Regional Science Review* 28(3), 330–346.

Madaras, Sz. 2009. A munkanélküliségi helyzet elemzése a Központi Régió megyéiben. *Közgazdász Fórum* 12(5), 45–58.

Madaras, Sz. 2014. A gazdasági válság hatása a munkanélküliség alakulására országos és megyei szinten Romániában. *Közgazdász Fórum* 17(1–2), 136–149.

Mayor, M.–López, A. J.–Pérez, R. 2007. Forecasting Regional Employment with Shift–Share and ARIMA Modelling. *Regional Studies* 41(4), 543–551.

Morariu, N.–Iancu, E.–Vlad, S. 2009. A Neural Network Model for Time-Series Forecasting. *Romanian Journal of Economic Forecasting* 2009(4), 213–223.

Pecican, Ș. E. 2006. *Econometrie*. București: Editura Beck.

Schanne, N.–Wapler, R.–Weyh, A. 2008. Regional unemployment forecasts with spatial interdependencies. *IAB-Discussion Paper* 28/2008, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-307640, downloaded: 18.02.2018.

Tsay, R. S. 2015. *Analysis of Financial Time Series*. Hoboken, NJ: John Wiley & Sons.

Warner, B.–Misra, M. 1996. Understanding Artificial Neural Networks as Statistical Tools. *The American Statistician* 50(4), 284–293.