Economic Insights – Trends and Challenges	Vol.IV(LXVII) No. 2/2015	53 – 61

Selecting the Best Model to Forecast Romanian Employment in Industry

Cristian Marinoiu

Petroleum-Gas University of Ploieşti, Bd. Bucureşti 39, 100680, Ploieşti, Romania
e-mail: marinoiu_c@yahoo.com

Abstract

The political changes after 1989 together with Romania's EU integration are inevitably reflected in the changes which have been taking place in the structure of the Romania's employment. Studying the evolution of the proportion of employees in industry in the total employment can provide important clues regarding our expectations related to the desire of the Romania's society transformation into a society with genuine European attributes. In this respect, the econometric modelling of the time series indicator Employment in industry as percent of total employment (EI) can provide us with useful working instruments. In this paper we build three such models on the basis of annual values of the EI time series for the time period 1980-2012. The model we finally propose is the model with the slightest forecast error.

Keywords: *employment; forecast; accuracy forecast; ARIMA; exponential smoothing; neural networks*

JEL Classification: *E24, C53, C52, C51*

Introduction

According to [4], in 2012, the distribution of the Romania's employment, from the point of view of the activities of the national economy was the following: 29% of the total employment was in the agricultural sector, 29.60% was concentrated in industry and the remaining 42.4% in services. The report reveals major discrepancies compared to European averages in the agricultural and services sectors and also highlights the closeness to the European average in case of the industry sector. In this paper we forecast the evolution of the proportion of employees in industry in the total Romania's employment in the years 2013-2022, using an econometric model built on data provided by [5]. The data presented in [5] regarding the industry "corresponds to divisions 2-5 (ISIC revision 2) or tabulation categories C-F (ISIC revision 3) and includes mining and quarrying (including oil production), manufacturing, construction, and public utilities (electricity, gas, and water)"[5]. These data recorded annually over the time period 1980 -2012 form a time series that we refer to as EI (from Employment in Industry) and can be regarded as realizations of a random process y_t. EI time series graph is represented in figure 1.

Fig. 1. EI time series graph
Source: made by the author in R using data from [5]

For modelling and forecasting the EI time series values, we will use three methods: Exponential smoothing, Autoregressive Integrated Moving Average (ARIMA) model and Neural Networks. The model with the smallest forecast error will be selected as the best forecast model for this series.

Exponential Smoothing

The class of the exponential smoothing models includes many variants that can be used for modelling and forecasting among which [1]: Simple exponential smoothing, Holt's exponential smoothing and Holt-Winter exponential smoothing. The graphical representation of EI time series in figure 1 reveals a downtrend and together with the lack of sesonality suggests Holt's exponential smoothing method as the most appropriate in this case.

Figure 2 shows the EI time series (black colour) and also its in-sample forecasts values (red colour). The differences between the values of the original series and the forecasts values represent the series of in-sample forecast errors or, in more simple terms, residuals. One can easily see that in-sample forecasts values are quite close to the original values. The parameters of the obtained model have the following values: alpha= 1 and beta= 0.2745121. The maximum value of the alpha parameter shows that the estimated value of the level depends on the most recent observations, while the relatively small value of the beta parameter shows that the estimated value of the trend depends mainly on older observations.

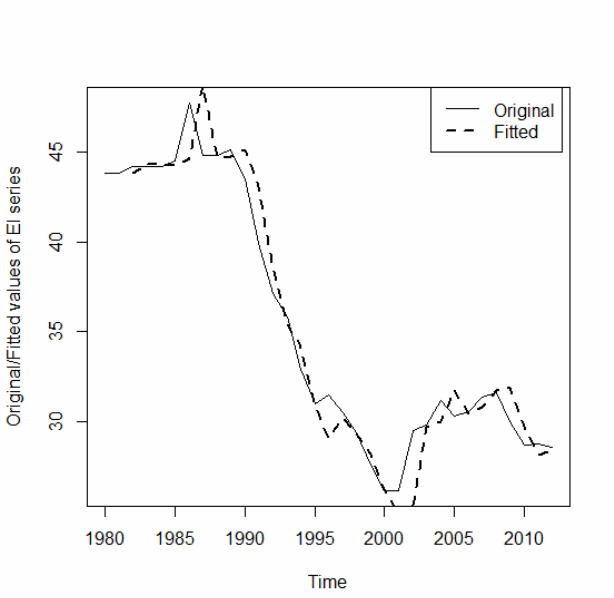

Fig. 2. EI time series: original values plot (solid lines plot) and in-sample forecasts plot (dashed lines plot)
Source: made by the author in R using data from [5]

We use the obtained model to achieve the time series forecasts values for the period 2013-2020. The model is considered valid if the obtained forecast errors (residuals) fulfil these conditions [1]: they are uncorrelated, are normally distributed with mean zero and constant variance. In order to verify that the above conditions are respected the following types of graphs (figure 3) will provide us useful information: the graph of residuals, the correlogram of residuals and the histogram of residuals.

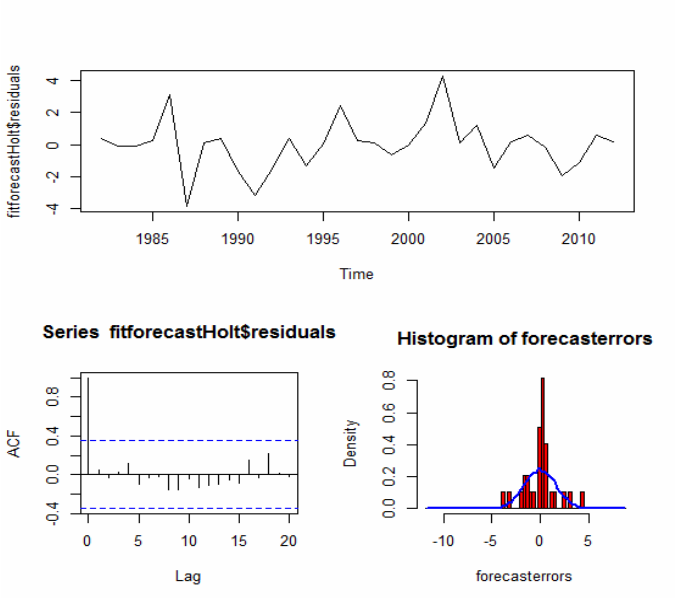

Fig. 3. Top: the graph of residuals. Bottom left: the correlogram of residuals (the 95% significance bounds are marked by dashed lines). Bottom right: the histogram of residuals with overlaid normal curve
Source: made by the author in R

By analysing the graph representations from figure 3 we can make the following observations:

- Residuals can be considered as uncorrelated as they do not signal the presence of a specific pattern. In addition, the graph of the correlogram in figure 3 shows that the autocorrelation coefficients values do not exceed the 95% significance bounds. We can confirm this by applying the Ljung-Box test [6] as the p-value (equal to 0.9281) is higher than the significance level of the test (equal to 0.05) and therefore we accept the hypothesis that residuals are uncorrelated;
- It seems plausible that the variance of residuals is constant because the fluctuations of the residuals (around zero) are roughly constant in size over time;
- The histogram of residuals shows that residuals are normally distributed with mean zero and constant variance.

Consequently, the model obtained using Holt's exponential smoothing is valid.

Autoregressive Integrated Moving Average (ARIMA)

To obtain the ARIMA model we use standard Box–Jenkins methodology [3]. Important steps of this methodology are implemented [8] in R, which enables automatic selection of the best ARIMA model based on AIC, AICC and BIC criteria. In this case the best selected ARIMA model is ARIMA (0,1,0) with drift (random-walk-with-drift), described by the equation $y_t = y_{t-1} -0.475$. Corresponding values AIC, AICC and BIC are AIC=118.51, AICC=118.92, BIC=121.44. The graphic representation (figure 4) of the original values (black plot) of the EI series and in-sample forecasts plot (red plot) reveals a good adequacy of the forecast model to the original data.

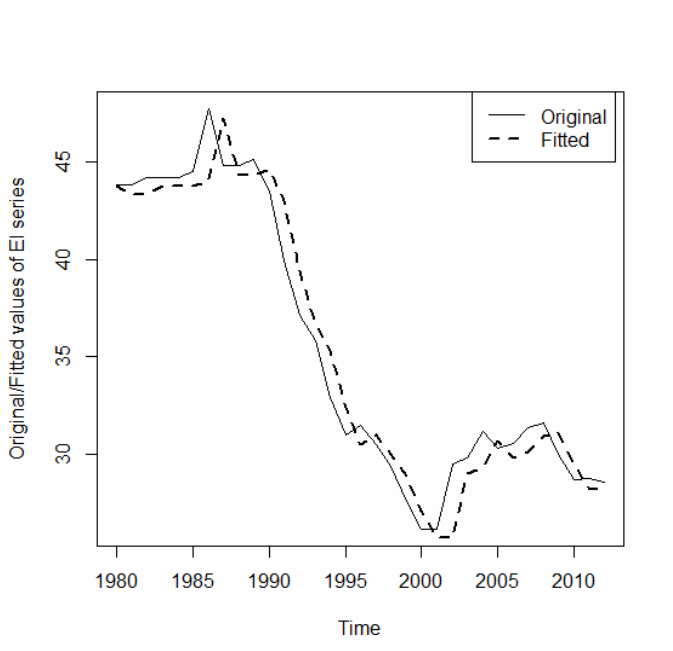

Fig. 4. EI time series: original values plot (solid lines plot) and in-sample forecasts plot (dashed lines plot) Source: made by the author in R using data from [5]

In order to validate the model we check if the residuals are uncorrelated and normally distributed with zero mean and constant variance.

The graphical representation of residuals in figure 5 (top) does not reveal a specific pattern, and therefore we consider that they are uncorrelated. Ljung-Box test [6] used to test the hypothesis of errors uncorrelation provides a p-value = 0.2064 which is higher than the significance level 0.05. This means that we accept the hypothesis that errors are uncorrelated. This decision is strengthened by the correlogram represented in figure 5 (bottom left) which reveals that all the autocorrelation coefficients values do not exceed the 95% significance bounds marked with dashed lines. We also notice that (figure 5 top) the residuals values fluctuate roughly constant over time and therefore their variance may be regarded as constant. The normality of residuals is proven by the shape of the histogram of residuals (figure 5 bottom right) which is close to normal distribution shape with zero mean.

In conclusion the ARIMA obtained model is valid.

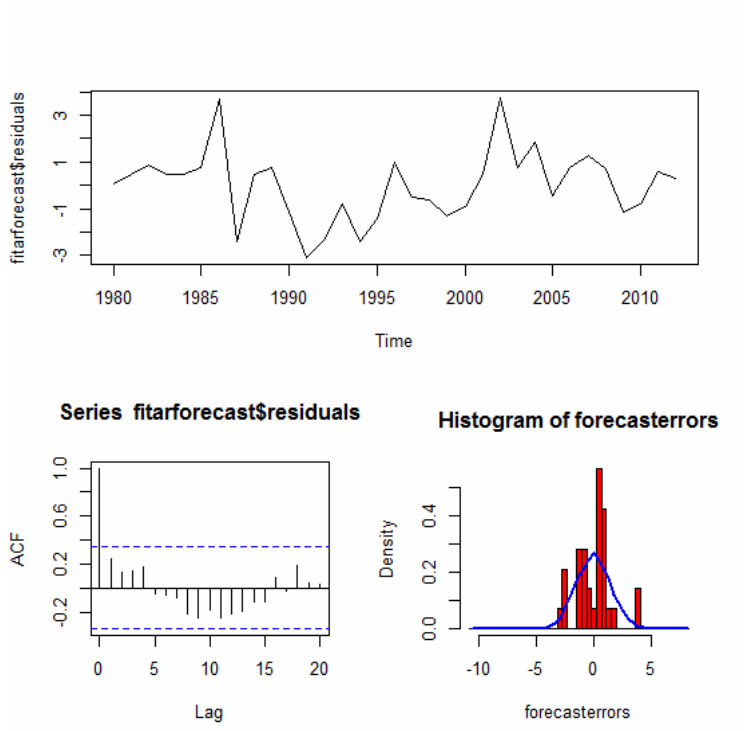

Fig. 5. Top: the graph of residuals. Bottom left: the correlogram of residuals (the 95% significance bounds are marked by dashed lines). Bottom right: the histogram of residuals with overlaid normal curve. Source: made by the author in R

Neural Network

From the class of forecast models related to time series that use neural networks we will continue to use a model based on feed-forward neural network with a single hidden layer implemented in R [7].

Figure 6 represents both the graph of EI time series and the graph of the in-sample forecasts obtained with the neural network model. As with the previous two methods, with the forecast model based on the use of neural networks, we can observe that there are no major differences between the original values of the EI series and the in-sample forecasts values.

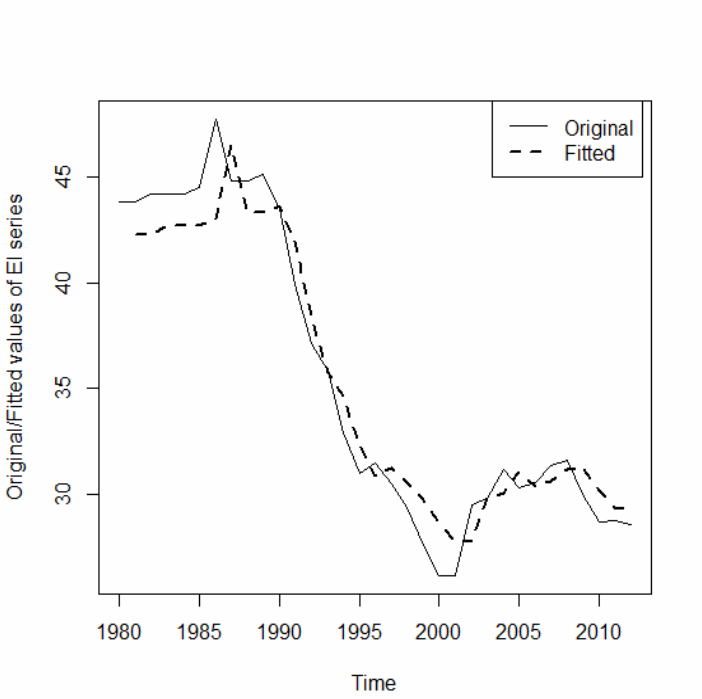

Fig. 6. EI time series: original values plot (solid lines plot) and in-sample forecasts plot (dashed lines plot)

Source: made by the author in R using data from [5]

The graph of residuals (figure 7. top) shows that residuals fluctuate roughly constantly around zero and therefore we can say that their mean is equal to zero and the variance is constant. It is also plausible that residuals are not correlated because:

- the autocorrelation coefficients values do not exceed the 95% significance bounds (figure 7. bottom left);
- by applying Ljung-Box test [6] we obtain the p-value 0.4138 higher than 0.05 significance level and therefore we accept the hypothesis that the residuals are uncorrelated.

In addition, the histogram of residuals (figure 7 bottom left) shows that the residuals follow a normal distribution and therefore the model is valid.

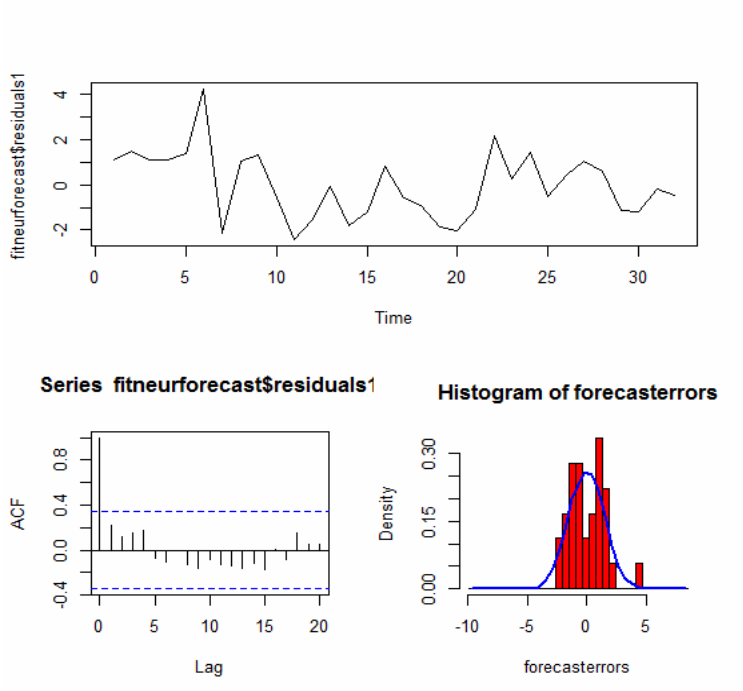

Fig. 7. Top: the graph of residuals. Bottom left: the corelogram of residuals (the 95% significance bounds are marked by dashed lines). Bottom right: the histogram of residuals with overlaid normal curve

Source: made by the author in R

Results and Discussions

The three methods presented are valid candidates for modelling and forecasting EI times series. The best of them in terms of forecast accuracy is the one which has the lowest forecast error. There are several ways to estimate the forecast error. Among the most representative we can mention [2]: RMSE (root mean squared error), MAE (Mean Absolute Error), MAPE (Mean absolute percentage error), MASE (Mean Absolute Error Scaled). Table 1 contains the estimated forecast errors values for each of the three forecast methods and for each estimation.

Table 1. Estimated values of the forecast error for the models discussed

Forecast error calculation methods / **Models**	**RMSE**	**MAE**	**MAPE**	**MASE**
Holt's model	**4.11**	3.99	13.13	4.13
ARIMA model	8.68	8.39	27.79	8.68
Neural Network model	11.54	8.13	27.54	8.41

Source*:* made by the author with results obtained in R

The estimation of forecast error was obtained by applying the following steps[2] :

- EI series values for the years 1980-2001 were used as training data to build forecast models;
- EI series values for the years 2002-2012 were used as test data in order to estimate the predictive accuracy of the models obtained in the previous step.

The values in Table 1 show that, in our case, the most suitable model is Holt model because it has the lowest values for all the estimators of the forecast error considered (RMSE, MAE, MAPE and MASE).

Table 2 presents annual values of the EI series for the period 2013-2022 and also their confidence intervals values. In figure 8 we represent the graph of these values.

Table 2. Forecast values and their 80% (Lo 80, Hi 80) and 95% (Lo 95, Hi 95) prediction intervals

Year	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2013	28.29649	26.23646	30.35652	25.145944	31.44703
2014	27.99298	24.65574	31.33021	22.889112	33.09684
2015	27.68946	23.07212	32.30681	20.627846	34.75108
2016	27.38595	21.43352	33.33838	18.282491	36.48941
2017	27.08244	19.72639	34.43849	15.832328	38.33255
2018	26.77893	17.94719	35.61066	13.271948	40.28590
2019	26.47541	16.09585	36.85498	10.601241	42.34959
2020	26.17190	14.17364	38.17016	7.822142	44.52166
2021	25.86839	12.18233	39.55444	4.937375	46.79940
2022	25.56488	10.12388	41.00587	1.949915	49.17984

Source: Results obtained by the author using functions from R

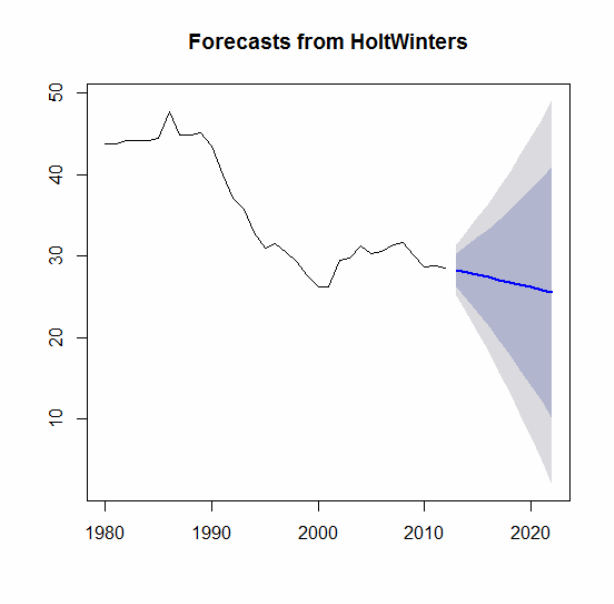

Fig. 8. The graph of the EI time series values, the forecasts for the years 2013-2020 (the line included in the shaded area) as well as the 80% (the dark shaded area) and the 95% prediction intervals (the light shaded area).

Source: made by the author using the functions from R

Conclusions

In this paper we built a forecast model for EI time series. The model we propose was selected from three models as the best candidate in terms of forecast accuracy from the point of view of all criteria used. The forecasts obtained with this model show that, for the period 2013-2020, the proportion of employment in industrial jobs compared to the total Romania's employment will continuously decrease.

References

1. C o g h l a n , A., *A litle book of R for times series*, available at https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf, [accesed at 3.06.2015]
2. H y n d m a n , R. J. and A t h a n a s o p o u l o s , G., *Forecasting: Principles and Practices*, available at https://www.otexts.org/fpp, [accessed at 2.06.2015]
3. *** *NIST/SEMATECH e-Handbook of Statistical Methods*, available at http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc445.htm, [accesed at 2.06.2015]
4. *** Strategia națională pentru ocuparea forței de muncă 2014-2020, available at http://www.mmuncii.ro/j33/images/Documente/Munca/2014-DOES/2014-01-31_Anexa1_Strategia_de_Ocupare.pdf, [accesed at 2.06.2015]
5. *** http://data.worldbank.org/country/romania, [accesed at 2.06.2015]
6. *** http://stat.ethz.ch/R-manual/R-patched/library/stats/html/box.test.html, [accesed at 4.06.2015]
7. *** http://www.inside-r.org/packages/cran/forecast/docs/nnetar, [accesed at 2.06.2015]
8. *** http://www.inside-r.org/packages/cran/forecast/docs/auto.arima, [accesed at 2.06.2015]