The Review of Regional Studies, Vol. 37, No. 1, 2007, pp. 64 – 81 *ISSN 1553-0892*

The Review of Regional Studies
Oklahoma State University

A Rank-Order Test on the Statistical Performance of Neural Network Models for Regional Labor Market Forecasts

Roberto Patuelli
Department of Spatial Economics, VU University of Amsterdam, The Netherlands, email: rpatuelli@feweb.vu.nl

Simonetta Longhi
ISER, University of Essex, UK, email: slonghi@essex.ac.uk

Aura Reggiani
Department of Economics, Faculty of Statistics, University of Bologna, Italy, email: aura.reggiani@unibo.it

Peter Nijkamp
Department of Spatial Economics, VU University of Amsterdam, The Netherlands, email: pnijkamp@feweb.vu.nl

Uwe Blien
Institut fuer Arbeitsmarkt und Berufsforschung (IAB), Nuremberg, Germany, email: uwe.blien@iab.de

Abstract

Using a panel of 439 German regions, we evaluate and compare the performance of various Neural Network (NN) models as forecasting tools for regional employment growth. Because of relevant differences in data availability between the former East and West Germany, the NN models are computed separately for the two parts of the country. The comparisons of the models and their *ex post* forecasts are carried out by means of a non-parametric test: *viz*. the Friedman statistic. The Friedman statistic tests the consistency of model results obtained in terms of their rank order. Since there is no normal distribution assumption, this methodology is an interesting substitute for a standard analysis of variance.

Keywords: Forecasts; Regional employment; Learning algorithms; Rank order test

JEL classification: C23; E27; R12

1. INTRODUCTION

Regional labor markets play a crucial role in the socio-economic development of the space-economy. Their performance (e.g., in terms of (un)employment or labor participation rates) is the outgrowth of a complex (multi-)regional force field, while their functioning is decisive for a balanced growth of a regional system. Therefore, it is of strategic importance to have sufficient insight into future developments of regional labor markets. From this perspective, it is important to predict regional economic growth as well as trends that might affect regional socio-economic development. To provide better guidelines for both national and local policies, forecasts of socio-economic variables at a disaggregated (e.g., sectoral or regional) level are becoming extremely relevant. Employment is undoubtedly a key variable in this regard: employment directly influences private expenditures, savings, and many other parts of the economy. It is therefore important to be able to produce reliable forecasts of its variation in time and space. In this paper, we aim to evaluate different statistical models used to compute forecasts of the regional employment evolution in Germany.

Given the spatial dynamic evolution of Germany in the past two decades, much interest has arisen in the functioning of its regional labor markets. For example, Blien and Tassinopoulos (2001) and Bade (2006) have proposed new methodologies to produce labor market forecasts for German regions. More specifically, Blien and Tassinopoulos (2001) suggest a combination of top-down and bottom-up techniques to compute short-term forecasts for West German regions. Their forecasts take into account regional autonomous trends that are next combined with expectations about the development of single industrial sectors, by means of an entropy-optimizing procedure. Bade (2006) uses an extension of the ARIMA approach to forecast the long-term development of regional shares in the national employment. However, both methodologies are subjected to a number of constraints and strict economic, as well as econometric, assumptions.

Similarly to Blien and Tassinopoulos (2001), Longhi et al. (2005) and Patuelli et al. (2007) proposed a set of models to compute short-term forecasts of employment at the regional level using data on West German regions. These models are based on a learning/forecasting algorithm called Neural Networks (NNs). In economic research, NNs have been applied to several fields ranging from financial forecasts to transport economics (see, e.g., Himanen, Nijkamp, and Reggiani 1998; Reggiani, Nijkamp, and Sabella 2000). Over the years, the NN models proposed in Longhi et al. (2005) and Patuelli et al. (2007) have been constantly updated with newly acquired data and evaluated on their ability to forecast regional employment variations using statistical indicators and benchmark forecasting models for comparison.

The choice of a proper forecasting model is not an easy task. For example, the stochastic nature of NN models may produce unstable forecasts over time, data sets, and statistical indicators. Various statistical model applications may be envisaged and assessed, thus requiring a methodology able to rank the statistical performance of these models on the basis of their computational outcomes. Patuelli et al. (2003), for instance,

apply Multicriteria Analysis (MCA) for choosing the best among NN models with the aim of defining a set of models that would perform better overall than the others when several relevant test features of the statistical performance are taken into account. They consider several dimensions of NN performance: indicators assessing the error levels of the forecasts, their computational complexity, the stability of their in-sample forecasts, and their generalization properties (out-of-sample forecasts).

However, further statistical aspects should be taken into account when evaluating a forecasting technique. In addition to assessing the "quality" of forecasts in terms of errors from the observed values, it is also necessary to examine their consistency by analyzing, e.g., whether the NN models show a stable rank order in their performance over time for various statistical indicators or other performance measures. In this paper we extend the analysis by Patuelli et al. (2003) by using the Friedman statistic to assess the presence of regularities in the rank orders of the NN models' aggregate and disaggregate performance.

The paper unfolds as follows. Section 2 briefly illustrates the NN models based on Longhi et al. (2005) and Patuelli et al. (2007) as well as their results. Section 3 introduces the Friedman rank-order statistic, while Section 4 shows the empirical application of the test. Section 5 summarizes the results and suggests future research directions.

2. NEURAL NETWORKS AS FORECASTING MODELS

NNs are calculation algorithms originally developed to mimic the operations of the human brain (see, e.g., Rumelhart and McClelland 1986). The attractiveness of NNs is mainly based on their ability to approximate mathematical relationships between variables, even when these relationships are not known *a priori*. Given their ability to obtain a good fit in data sets of any complexity, NNs are applied nowadays to a wide range of problems, in particular in those situations where the simulation of complex relationships between independent variables is necessary as, for example, in applied finance or environmental analysis. For a review of NN applications in economics we refer the reader to, among others, Herbrich et al. (1999). Comprehensive descriptions of the NN technique can be found, among others, in Bossomaier (2000). NN models have often been compared to more traditional forecasting methodologies with contrasting results (see, among others, Cheng and Titterington 1994; Swanson and White 1997a, b; and Fischer 1998, 2001a).

In an NN model, the statistical calculation is distributed over a high number of simple units working in parallel, the neurons. Neurons are organized in layers and are internally connected through a set of weights (w_j; w_{jn}). In the models evaluated in this paper (called feedforward NNs), the neurons are organized in only three groups of layers: one input layer, some hidden layers, and one output layer. Neurons from each layer are connected to every neuron of the next layer. Every input neuron (the explanatory variables in a regression framework, x_n) is connected only to the neurons of the first hidden layer, while only the neurons of the last hidden layer are connected with the neurons in the output

layer (the dependent variable in a regression framework, E_{rt}). Only in case of no hidden layers are input and output neurons directly connected.

The data entering each neuron is aggregated by an activation function (Ψ; φ_j) that computes the output of the unit before passing the result to the units of the next layer. In case of only one hidden layer (see Fischer 2001a, b), we have:

$$(1) \qquad E_{rt} = \Psi\left[\sum_j w_j \varphi_j \left(\sum_n w_{jn} x_n\right)\right] + \varepsilon_{rt}.$$

The "training process" of an NN in the present paper is based on the computation of the network weights via a recursive algorithm called Back-Propagation Algorithm (BPA). The BPA modifies the network weights step by step by minimizing the difference between the observed data and the output obtained using the current set of weights. To identify the best NN architecture and to generate the final forecasts, the data set is divided into three (mutually exclusive) sub-sets: the first, called the "training set," is used for parameter estimation; the second, called the "validation set," is used in the process of fine-tuning of the parameters; and the third, called the "test set," is used to assess the performance of the model.

The models developed by Longhi et al. (2005) and Patuelli et al. (2007) employed two distinct but similar data sets, one for West Germany and one for East Germany. The data set, provided by the German Institute for Employment Research (Institut für Arbeitsmarkt und Berufsforschung – IAB), is essentially a panel of 439 regions containing data on the number of individuals employed full-time each year on June 30, subdivided into nine economic sectors,[1] as well as the average of their daily wages at the regional level. The data for the 326 West German regions is available for the period 1987–2003, while the data for the 113 East German regions is available only for the period 1993–2003. Information on the year of collection of the data as well as on the level of urbanization of each region (the "type of region," see Blien and Tassinopoulos 2001) is also available. Such data is used to produce forecasts of the growth rate of East and West German employment over a two-year period.

Inputs for the models are the growth rate of sectoral employment, the growth rate of daily wages, an indicator of the year of data collection, and an indicator of the degree of urbanization of the region concerned. Information identifying the year was used in two different ways, as a set of dummy variables (one dummy for each year) or as a string variable.[2] The degree of urbanization was used a string variable.

---

[1] The nine economic sectors are: 1) primary sector, 2) industry goods, 3) consumer goods, 4) food manufacturing, 5) construction, 6) distributive services, 7) financial services, 8) household services, 9) services for society.

[2] The commercial software used for our experiments, Neuralyst, allows the use of non-numeric (string) input and output variables. The software associates numeric values between 0 and 1 with the values of each variable. The interpretation and mapping of the values are automatically taken care of by the algorithm.

TABLE 1

NN Models Compared

Model's Name	Input Variables
Model A	Growth rate of sectoral employment; year: dummies
Model B	Growth rate of sectoral employment; year: string
Model D	Growth rate of sectoral employment; year: dummies; urbanization level: string
Model AW	Growth rate of sectoral employment; growth rate of daily wages; year: dummies
Model DW	Growth rate of sectoral employment; growth rate of daily wages; year: dummies; urbanization level: string

In our computational analysis we have used different statistical specifications of NNs. The five NN models compared in this paper are summarized in Table 1. We will now concisely describe the different features of these models.

All models use the growth rate of sectoral employment as well as information about the year of collection of the data. This last variable is represented by dummy variables in all models except Model B, where it is used as a string. Model D uses the same inputs as Model A, plus the information on the urbanization level. Model DW and Model AW add the growth rate of daily wages to the inputs of Model D and Model A, respectively.

All these models have been used to compute *ex post* forecasts for the years 2001, 2002, and 2003. The results of these models are used in our empirical application in Section 4.

An assessment of the performance of NN models was carried out by Longhi et al. (2005) using only statistical indicators such as the Mean Absolute Error (MAE), the Mean Square Error (MSE), and the Mean Absolute Percentage Error (MAPE), defined respectively as:

$$MAE = 1/N * \left(\sum_i \left| y_i - y_i^f \right|\right); \tag{2}$$

$$MSE = 1/N * \left[\sum_i \left(y_i - y_i^f\right)^2\right]; \tag{3}$$

$$MAPE = 1/N * \left(\sum_i \left| y_i - y_i^f \right| * 100/y_i\right); \tag{4}$$

where y_i is the observed value in region i (target), y_i^f is the forecast of the model adopted, and N is the number of regions for which the forecast is computed. The common interpretation of these indicators is: the better estimation, the closer the value to zero.

Patuelli et al. (2003) have added new criteria to evaluate other aspects of the NN performance. The generalization properties of the NN models are evaluated using the "generalization" criterion (*Gen*). The criterion aims to compare the performance of the models – measured by the three indicators mentioned above – over different parts of the data sets: the part used to compute the network weights (*train* set) and the part used to compute the *ex post* forecasts (*test* set). The *Gen* criterion is calculated as follows.

$$Gen = \frac{MSE_{train} - MSE_{test}}{\overline{MSE}} + \frac{MAE_{train} - MAE_{test}}{\overline{MAE}} + \frac{MAPE_{train} - MAPE_{test}}{\overline{MAPE}}, \tag{5}$$

where $\overline{MSE}$, $\overline{MAE}$, and $\overline{MAPE}$ are, for each of the three statistical indicators, the average of the values computed for the train set and test set. These indicators are used to normalize for the average scale of each type of error so as to sum the components of the criterion. *Gen* can assume positive or negative values. A large negative value of *Gen* may suggest that the model overfitted the learning data, thus being unable to compute reliable *ex post* forecasts. A value of *Gen* close to 0 indicates that similar error levels are found in the two parts of the data set, which would suggest a model's good generalization power. Finally, a higher (positive) value of the indicator indicates lower errors in the *ex post* forecasts. (The test data actually fits the computed parameters better than the training data.) A combined assessment of the statistical performance of these various models – on the basis of several test criteria – was undertaken by using multicriteria analysis (MCA).

Patuelli et al. (2003) found that Model AW appeared to be the most robust model, showing the best performance for two out of three statistical indicators. However, in more recent forecasts such model appeared to show a weaker performance, being outperformed by Model B. Furthermore, none of the proposed models had been able to correctly predict the decrease in full employment in Germany in the beginning of the 2000s. (On this topic, see Statistisches Bundesamt 2002.)

The MCA approach adopted in the above test studies has neglected two important aspects of the models' performance. The first relates to the question of whether such models are able to correctly reproduce/forecast the ranking of regions in terms of growth rates of employment. The second relates to the question of whether the models' rankings are consistent or robust over time (after updating the data set used for the analysis) and for different statistical indicators. A test based on the Friedman statistic as proposed in the next section should enable a better evaluation of the models' forecasting performance.

3. RANK ORDER TEST: THE FRIEDMAN STATISTIC

The main objective of the test based on the Friedman (1937) statistic is the "*isolation of factors which account for variation in the variable studied*" (p. 675). The test based on the Friedman statistic is commonly implemented in modern data mining software and extensively used in the statistical literature. Applications can be found in biology (e.g., Edgar 2004), medicine (e.g., Efficace et al., 2004) and economics (e.g., Frees 1995).

Profit and Tschemig (1998) use an iterative Friedman test to rank results from a survey on possible solutions to German unemployment, while Scarelli and Venzi (1997) proposed the Friedman test for the hierarchization of the alternatives in MCA.

In our study, we use the Friedman test to assess the correlation among measurements – e.g., statistical indicators – of a set of alternatives (the NN models). The Friedman test is a distribution-free measure of the variability observed for one variable over two or more factors and is computed over intra-sample rankings as:

$$S^* = \frac{12S}{Kn(n+1)}, \tag{6}$$

with

$$S = \sum_{j=1}^{n} \left(S_j - S_e\right)^2, \tag{7}$$

where n is the number of observations of the variable and K is the number of factors according to which the variable is evaluated. In our case, n is equal to the number of NN models compared, while the K factors are the statistical indicators. S is the deviance of the sums of the rankings (S_j) of each j^{th} observation (in our case, the NN models' ranks are summed on the basis of the factors) from the average rank sum S_e. Therefore, if the j^{th} observation is ranked first under two hypothetical factors, the sum (S_j) of its rankings is two. For a sufficient number of cases the S^* statistic has a χ^2 distribution, with $(n - 1)$ degrees of freedom. When $K = 3$ and $n \leq 15$, or $K = 4$ and $n \leq 8$, the critical values table prepared by Friedman should be consulted.

In the computation of a Friedman test, the alternatives can be organized in a two-way table as in Table 2, where each row represents a factor according to which the alternative is measured, while each column contains the rank positions of the alternatives according to the factors. The null hypothesis H_0 is that the rankings by row are uncorrelated. If H_0 is true, then the sums (by column) of the rankings S_j will differ only because of sampling error. When the value of the Friedman statistic exceeds the critical values for the chosen significance level, H_0 is rejected, thus indicating that the rankings by row are unlikely to be uncorrelated.

A non-significant value of the test might imply two different scenarios: when the differences among the rank sums S_j are too small to be statistically significant, the rankings could be either uncorrelated or, in the opposite case, inversely correlated. In case of two factors and four alternatives, for example, rankings like (1, 2, 3, 4) and (4, 3, 2, 1) generate equal rank sums. Although the null hypothesis is not rejected, the rankings are not uncorrelated but inversely correlated.

TABLE 2

An Example of the Data Used for the Friedman Test

	x_1	x_2	x_3
Factor 1	x_{11}	x_{21}	x_{31}
Factor 2	x_{12}	x_{22}	x_{32}
…	…	…	…
Factor K	x_{1K}	x_{2K}	x_{3K}
S_j	S_1	S_2	S_3

Note: Each alternative x_j has a rank x_{jk} for each factor k. S_j is the sum of each alternative's rank scores

The Friedman test can be considered as a generalization of the Spearman rank correlation test to the case when more than two factors are compared. Since the Friedman test is exclusively based on rank-order positions, the observations can be multiplied by any arbitrary number without affecting the test results. This property is useful when dealing with a linear or logarithmic transformation of the data.

The Friedman test can be employed as a substitute for the analysis of variance when the observations are not normally distributed because it does not require normally distributed data. This is particularly advantageous in economics, where the normality condition is rarely met. Furthermore, in case of large data sets, the Friedman test has been found to be four times faster to compute than the analysis of variance, with an information loss ranging from 9 to 36 percent, which decreases with the number of factors considered (Friedman, 1937). In light of this consideration, it would be desirable to employ additional factors in our analysis in the course of future research.

4. COMPARING THE PERFORMANCE OF NEURAL NETWORK MODELS

4.1 Rank Order Tests for Regional Performance

The first aspect to be investigated regards the ability of the NN models to correctly predict the hierarchy on the changes in employment at the district level in Germany. As a preliminary test on our analysis, a Friedman test carried out on observed employment variations in 2001, 2002, and 2003 suggested that there are hierarchies among the German districts that are consistent over the years. The test computed over the years 2001, 2002, and 2003 rejected the null hypothesis H_0 with a 99 percent confidence level, suggesting that (fairly) stable patterns of growth among the districts are present. The result is not trivial because, in the presence of regional change, the rank position of the districts in terms of (employment) growth might change greatly. Similar results were also

obtained when the rankings were split in broad groups – of 30 and 10 districts, for West and East Germany, respectively.[3]

The above analysis suggests that there is in fact a ranking of the districts to be forecasted. For the purpose of testing the NN models' ability to reproduce this hierarchy, we separately compare the rankings for 2001, 2002, and 2003 computed on real employment variations to the rankings generated *ex post* by each forecasting model. A test was carried out for each NN model, for each year of forecast (2001, 2002, and 2003), separately for East and West Germany. The results are shown in Table 3.

The Friedman test computed on the observed growth rate of employment in the West German regions suggests that the rankings of regions are consistent over the three *ex post* forecasts. The test computed for the forecasts for the years 2001, 2002, and 2003 rejects the null hypothesis of random correlation at 99 percent confidence level, thus suggesting that the models are able to correctly predict the pattern of growth among the regions. Contrasting regional evolutions that might significantly modify the rank position of each region might be correctly identified by the NN models evaluated.

The Friedman test for East German regions rejects the hypothesis of uncorrelated rankings for 7 out of 15 models. This failure to correctly predict the ranking of regions in terms of employment growth rate can be attributed to the high errors in the *ex post* forecasts for the East German regions, these being primarily a consequence of the shorter time span available. In 2001, all models fail to reproduce the rankings of regions, while their performance seems to improve in 2002 and 2003, where the hypothesis of uncorrelated rankings is rejected at a 95per cent confidence level. The inferior performance of the models in the year 2001 may be caused by an exogenous shock that changed regional employment evolutions, thereby making forecasts more difficult. Such a finding for 2001 appears to be counterintuitive, as it would more appropriately be expected for the West German NN models. The year 2001 was in fact a turning point in West Germany after a few years of aggregate employment gains and before a new decreasing trend. Relatively stable losses were instead observed in the same years in East Germany. Reasons behind this finding should be investigated in future research. In particular, the above results do not say very much about the quality of the forecasts in terms of relative error; in this regard, the error probabilities in Table 3 will be used as (partial) evaluation criteria in the rank order analysis of the NN models in the next section.

A visual representation of the observed and forecasted rankings for the three years is shown in Figures 1, 2, and 3 where Model B has been selected as an example. The left map plots the observed rankings, while the right map plots the rankings generated by the *ex post* forecasts of Model B. While for West Germany the rank of each region seems to be predicted with a small error, the errors for East German regions seem bigger. In fact,

[3] The ranking of the districts belonging to each layer was structured so that every district within a certain layer would have the same position. Consequently, the 30 West German districts with the higher growth rates were ranked 1st, followed by 30 more districts ranked 31st, and so on.

while West Germany shows a rather heterogeneous performance over the districts, East German districts tend to show a more homogeneous (negative) performance, which is more difficult to reproduce in terms of rank order.

4.2 Rank Order Tests on the Model Results

The second aspect of the model's performance in which we are interested is the rank order of NN models. We test whether the rankings of the NN models are consistent over the three *ex post* forecasts and five statistical indicators. The NN models are evaluated and compared separately for East and West Germany, and the three *ex post* forecasts, using the four statistical indicators introduced in Section 2, in addition to the results of the S-Tests computed in the previous subsection and shown in Table 3. The NN model ranked first for a given year is the one that shows the lowest error or, in the case of the S-Test, the lowest error probability when the null hypothesis is rejected.

TABLE 3

Districts' Rankings by Year and the Friedman Statistic

West Germany				East Germany			
Years	Models	S^*	Prob.	Years	Models	S^*	Prob.
2001	Model A	440.026^{***}	$2.13 * 10^{-5}$	2001	Model A	117.7792	0.3358
	Model AW	446.5338^{***}	$8.48 * 10^{-6}$		Model AW	108.4741	0.5767
	Model B	444.2226^{***}	$1.18 * 10^{-5}$		Model B	116.0121	0.3785
	Model D	436.9179^{***}	$3.26 * 10^{-5}$		Model D	116.6204	0.3636
	Model DW	434.5176^{***}	$4.51 * 10^{-5}$		Model DW	88.2692	0.9524
2002	Model A	434.8807^{***}	$4.29 * 10^{-5}$	2002	Model A	139.1886^{**}	0.0417
	Model AW	437.5326^{***}	$3.00 * 10^{-5}$		Model AW	139.8547^{**}	0.0384
	Model B	430.5571^{***}	$7.62 * 10^{-5}$		Model B	138.4807^{**}	0.0455
	Model D	437.2602^{***}	$3.11 * 10^{-5}$		Model D	127.2986	0.1531
	Model DW	443.2864^{***}	$1.38 * 10^{-5}$		Model DW	129.6665	0.1215
2003	Model A	391.4998^{***}	0.0067	2003	Model A	139.9823^{**}	0.0378
	Model AW	389.2529^{***}	0.0083		Model AW	139.5668^{**}	0.0398
	Model B	382.9143^{**}	0.0148		Model B	137.9870^{**}	0.0483
	Model D	394.4091^{*}	0.0050		Model D	137.0545	0.0540
	Model DW	405.5177^{***}	0.0016		Model DW	143.2781^{**}	0.0247
*** significant at 1%; ** significant at 5%				*** significant at 1%; ** significant at 5%			

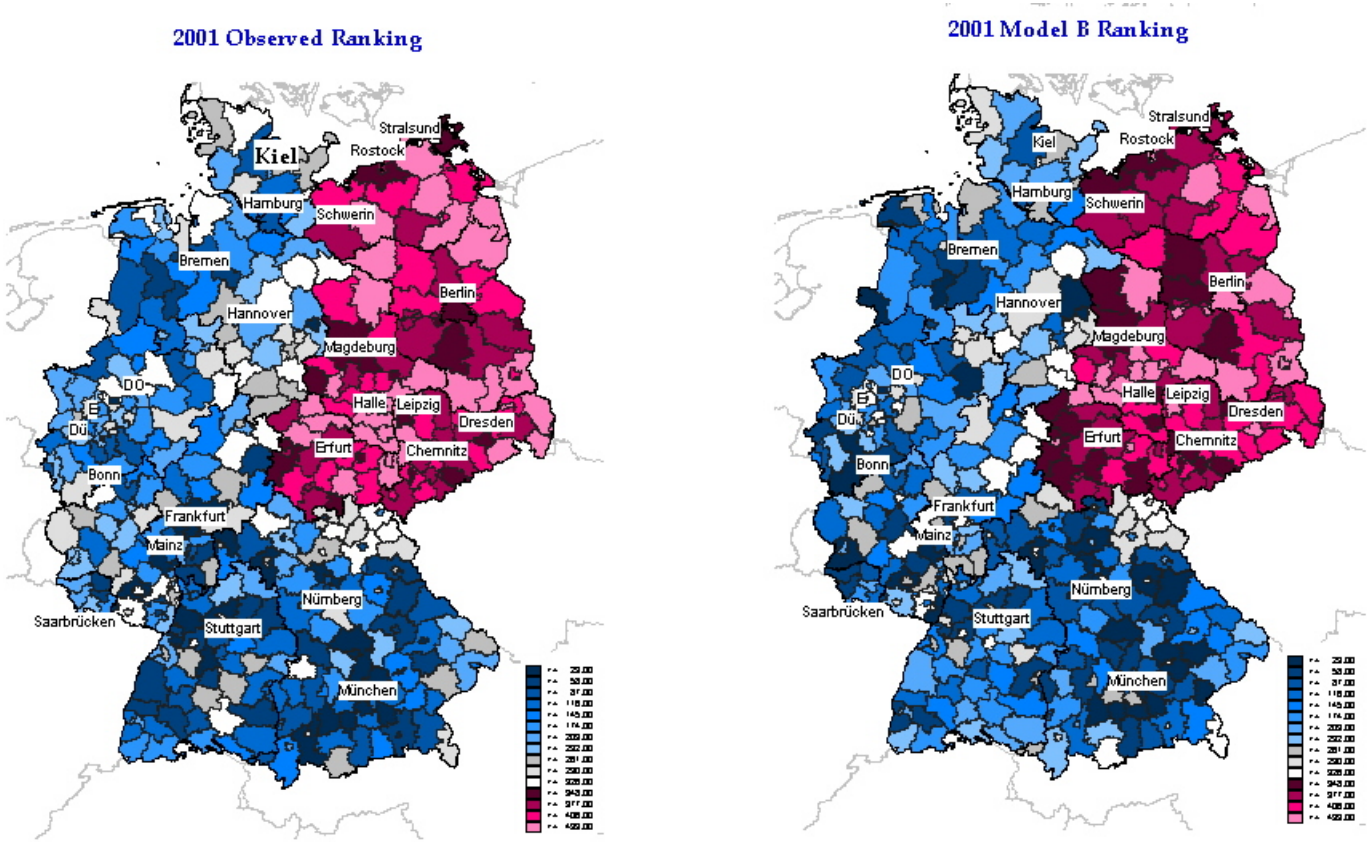

FIGURE 1. Graphical Representation of Observed and Forecasted Growth Rate District Rankings for the Year 2001 (Model B)

Note: The districts' rankings are here split in layers containing about 30 districts each. The colors for West and East Germany are blue and red, respectively; the shading of the layers goes from dark to light, from the first-ranked to last-ranked districts. The legend shows the cumulative rank positions of the districts. The layers for East Germany are listed subsequently to the West German ones.

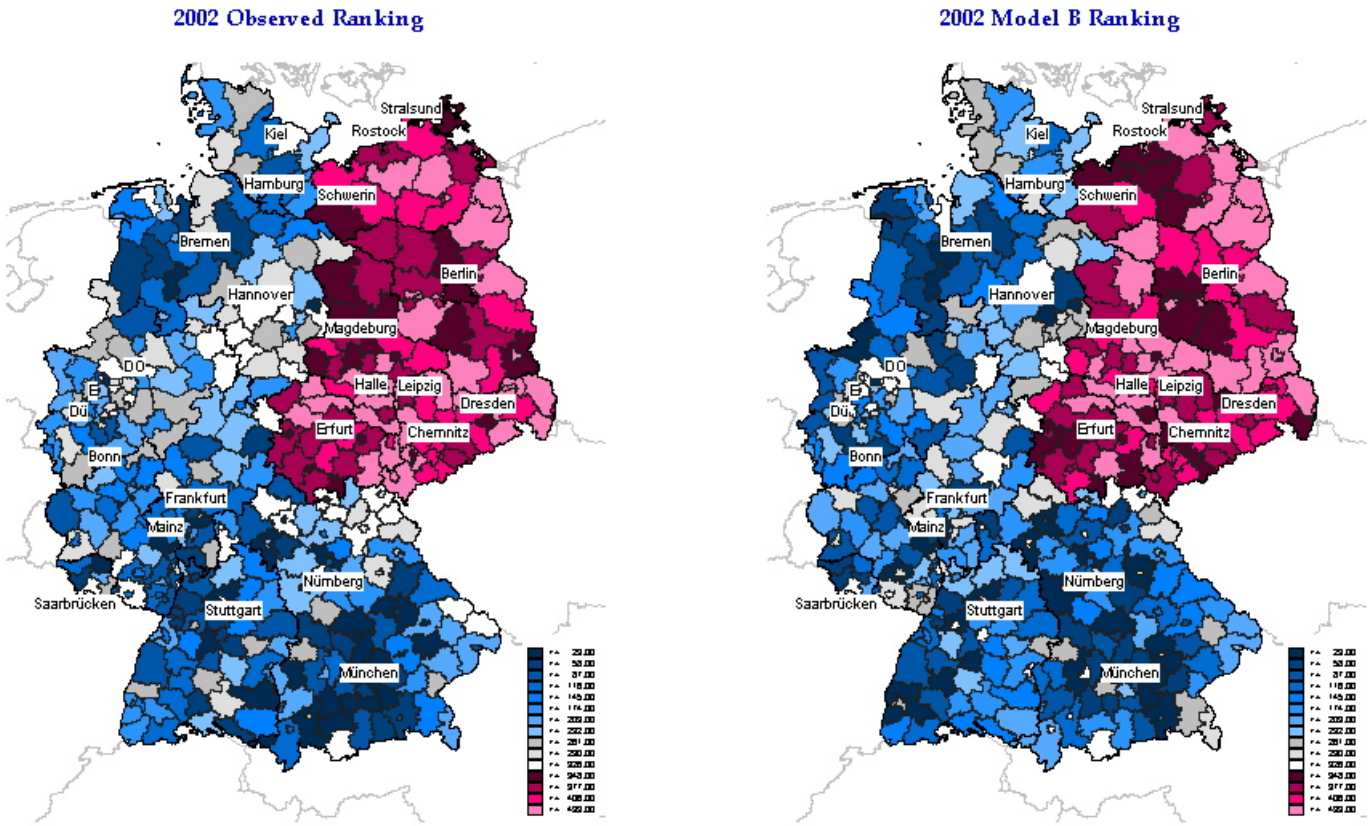

FIGURE 2. Graphical Representation of Observed and Forecasted Growth Rate District Rankings for the Year 2002 (Model B)

Note: The districts' rankings are here split in layers containing about 30 districts each. The colors for West and East Germany are blue and red, respectively; the shading of the layers goes from dark to light, from the first-ranked to last-ranked districts. The legend shows the cumulative rank positions of the districts. The layers for East Germany are listed subsequently to the West German ones.

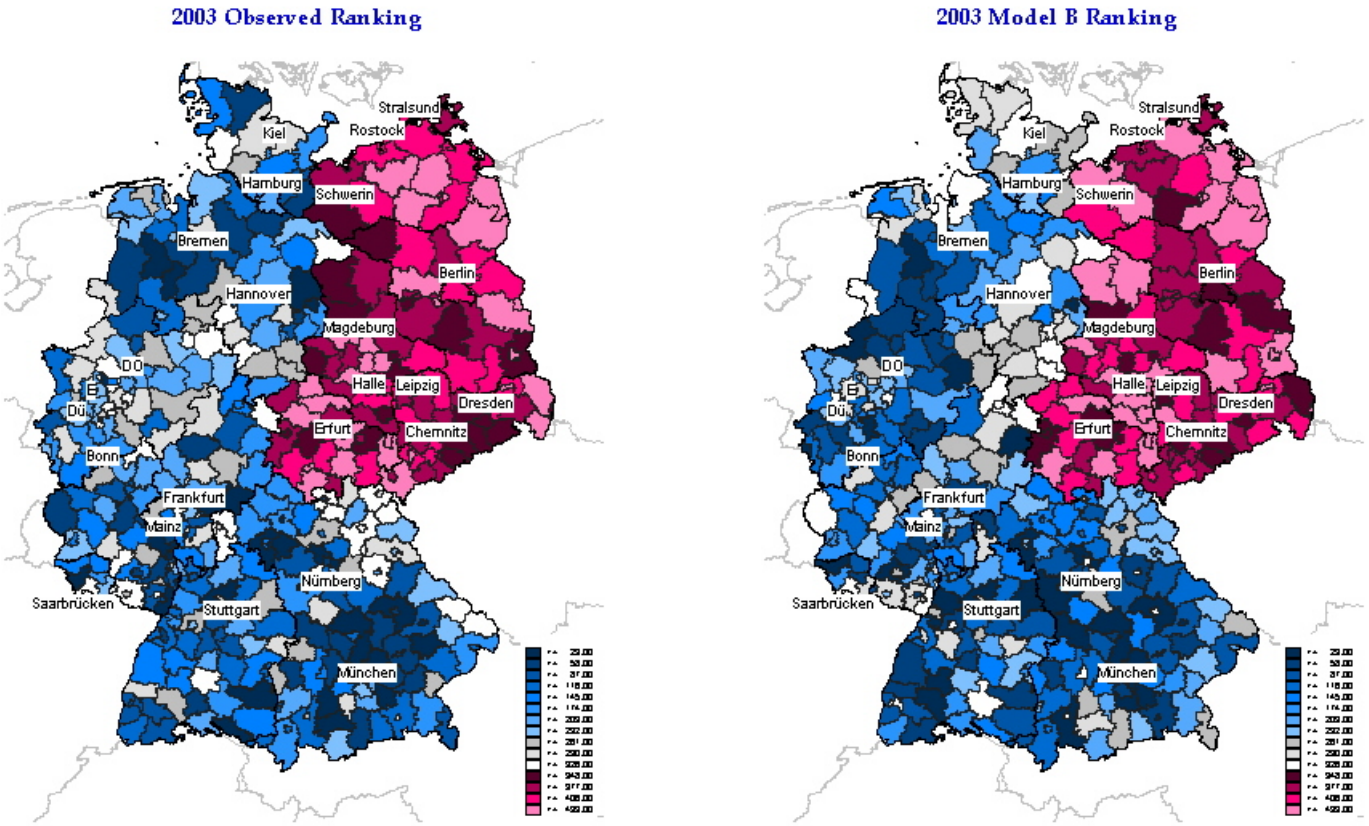

FIGURE 3. Graphical Representation of Observed and Forecasted Growth Rate District Rankings for the Year 2003 (Model B)

Note: The districts' rankings are here split in layers containing about 30 districts each. The colors for West and East Germany are blue and red, respectively; the shading of the layers goes from dark to light, from the first-ranked to last-ranked districts. The legend shows the cumulative rank positions of the districts. The layers for East Germany are listed subsequently to the West German ones.

Table 4 show the rank orders of each NN model for the five criteria. The first panel of Table 4 shows the rankings of the models computed on data for West Germany, while the second panel shows the rankings of the models computed on data for East Germany. The last row of each panel – West-Sum and East-Sum – shows the sums of the models' rankings that are used to compute the Friedman test. When the sums differ between the NN models, the test is significant and suggests that the rankings analyzed are correlated. The last row of Table 4 shows the results for the test carried out on all models: those computed for West Germany and those computed for East Germany. Columns from 1 to 5 show the ranking of each model for each criterion. Column 6 shows the critical value of the Friedman tests, while column 7 shows the associated probability of uncorrelated rankings.

The Friedman test rejects the null hypothesis of uncorrelated rankings for both West and East Germany. The test computed for West Germany has a value of 10.5067 and is statistically significant at the 5 percent level; the test for East Germany has a value of 23.0933 and is statistically significant at the 1 percent level. Similar to the results for East and West, the test computed on the whole country, shown in the last row of Table 4, suggests the existence of a correlation among the rankings observed. Note that the statistical significance of the separate tests for West and East Germany does not imply statistically significant results for the combined data set, since the NN models might rank differently for West and East Germany. This might bring non-significant results for the analysis on the whole country. The significance of the test for the combined data set further supports the previous results.

Because of how the Friedman test is computed, the results are sensitive to the number of criteria employed. In a small data set, rank order variations tend to have a bigger weight in the determination of the rank order sums, therefore making the identification of clear-cut hierarchies more difficult. Further, the Friedman test follows a χ^2 distribution only for a sufficient number of alternatives and criteria.

The implications of the above considerations were evident once we carried out further analyses that employed only a portion of the factors examined earlier. The aim of these analyses was to observe the correlation of the rank orders for smaller sets of criteria. More in detail, we carried out additional analyses to verify the existence of consistent hierarchies between the NN models' performances over time and for different statistical indicators. The selected NN models were tested separately by year – carrying out a comparison of the rank orders obtained in terms of MSE, MAE, and MAPE for each year – and by statistical indicator – comparing each indicator's rank orders over the 2001-2003 interval. Both analyses were carried out separately for the former West and East Germany, as well as for Germany as a whole. The analyses provided mixed results. These additional analyses are examples of how a limited number of factors may influence the analysis and may complicate the identification of significant results, also because of a higher loss of information (see Section 2).

TABLE 4

NN Models' Rankings by East and West Germany, By Year and Indicator

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Model A	Model AW	Model B	Model D	Model DW		
				Rankings			S*	Prob.
West Germany	MSE 2001	4	3	1	5	2		
	MSE 2002	4	2	1	3	5		
	MSE 2003	2	4	1	5	3		
	MAE 2001	4	1	3	5	2		
	MAE 2002	4	2	1	3	5		
	MAE 2003	2	4	1	5	3		
	MAPE 2001	3	1	5	2	4		
	MAPE 2002	4	2	1	3	5		
	MAPE 2003	2	4	1	5	3		
	Gen 2001	4	5	1	3	2		
	Gen 2002	4	2	1	3	5		
	Gen 2003	2	4	1	5	3		
	S-Test 2001	3	1	2	4	5		
	S-Test 2002	4	2	5	3	1		
	S-Test 2003	3	4	5	2	1		
	West-Sum	49	41	30	56	49	10.5067**	0.03271
East Germany	MSE 2001	3	5	2	1	4		
	MSE 2002	1	2	4	5	3		
	MSE 2003	2	3	1	5	4		
	MAE 2001	3	2	4	5	1		
	MAE 2002	2	3	1	5	4		
	MAE 2003	2	3	1	5	4		
	MAPE 2001	3	1	4	5	2		
	MAPE 2002	2	3	1	5	4		
	MAPE 2003	2	3	1	5	4		
	Gen 2001	2	4	3	5	1		
	Gen 2002	2	3	1	5	4		
	Gen 2003	2	3	1	5	4		
	S-Test 2001	1	4	3	2	5		
	S-Test 2002	2	1	3	5	4		
	S-Test 2003	2	3	4	5	1		
	East-Sum	31	43	34	68	49	23.0933***	0.00012
Germany	EW-Sum	80	84	64	124	98	27.0933***	0.00002

*** significant at 1%; ** significant at 5%

In summary, our analyses showed two different aspects of the reliability of NN models. The district-level estimates proved to be consistent with the hierarchies incorporated in the (observed) data. This suggests that NN models are able to correctly identify the contrasting performances of districts in terms of ranking.

The "aggregate" analyses showed that the rankings of the NN models are not uncorrelated. It follows that the NN models tested here show a pattern of results, where some models perform better than others. The order of "preference" of these models is proven to be consistent for different data sets and, to a certain degree, over time and different evaluation indices.

5. CONCLUSIONS

This paper has statistically analyzed the consistency of five Neural Network (NN) models to forecast employment in Germany at the regional level. We used the Friedman statistic as a test to verify the ability of the NN models to reproduce the observed rankings of regions in terms of employment growth. The results suggest that the models are able to effectively reproduce the rankings of German regions in terms of employment growth rates. This suggests that the NN models do not level out the variability among the districts into a bland average performance; to a certain extent they are able to correctly predict the rankings observed among the regions' employment growth rates.

The second set of analyses of our study centered around the aggregate performance of the NN models. We analyzed whether the NN models showed a consistent performance (in terms of rank order) over time and for various statistical indicators. The tests on the models computed for West Germany reject the hypothesis of uncorrelated rankings. The tests on the models computed for East Germany are insignificant in the first year (2001) and significant for the subsequent two years (2002 and 2003). The test on all NN models – those computed for West Germany and those computed for East Germany – rejects the hypothesis of uncorrelated rankings. This suggests a correlation between the rank orders found for the West and East German NN models, which were developed separately. The resulting rankings of the NN models still seem to be correlated according to the evaluation criteria used.

When a wide set of mutually complementary criteria is used, the models show fairly stable rankings. The tests carried out for East and West Germany separately and on Germany as a whole rejected the null hypothesis of uncorrelated rankings. Additional analyses should use more evaluation criteria (such as spatio-temporal autocorrelation or the computation of indicators for both East and West Germany) and different NN models. Also, a comparison of the Friedman test to other rank order estimators might provide a wider look at the implications of using distribution-based or non-parametric tests (such as Friedman's) or the possibility of employing pairwise comparisons instead of summarizing measures like the sum of the rank positions, which was employed in computing our tests.

Finally, more reflection and empirical research is needed on how to generate more reliable and consistent NN models and how to evaluate them. Future research should address these issues by expanding the set of tools used in evaluating the NN models, e.g., shift-share analysis or different classification approaches, such as the one recently proposed by Berardi, Patuwo, and Hu (2004). We may conclude that the NN methodology still opens up many new and intriguing research issues.

REFERENCES

Bade, F.J., 2006. ′Evolution of Regional Employment in Germany: Forecast 2001 to 2010," in A. Reggiani and P. Nijkamp (eds), *Spatial Dynamics, Networks and Modelling* (pp. 297–323). Cheltenham (UK): Edward Elgar.

Berardi, V.L., B.E. Patuwo, and M.Y. Hu, 2004. "A Principled Approach for Building and Evaluating Neural Network Classification Models," *Decision Support Systems* 38(2), 233–46.

Blien, U., and A. Tassinopoulos, 2001. "Forecasting Regional Employment with the ENTROP Method," *Regional Studies* 35(2), 113–124.

Bossomaier, T.R.J., 2000. "Complexity and Neural Networks," in T.R.J. Bossomaier and D.G. Green (eds), *Complex Systems* (pp. 368–406). Cambridge (UK): Cambridge University Press.

Cheng, B. and D.M. Titterington, 1994. "Neural Networks: A Review from a Statistical Perspective," *Statistical Science* 9(1), 2–30.

Edgar, R.C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput," *Nucleic Acids Research* 32(5), 1792–7.

Efficace, F., L. Biganzoli, M. Piccart, C. Coens, K. Van Steen, T. Cufer, R.E. Coleman, H.A. Calvert, T. Gamucci, C. Twelves, P. Fargeot, and A. Bottomley, 2004. "Baseline Health-Related Quality-of-life Data as Prognostic Factors in a Phase III Multicentre Study of Women with Metastatic Breast Cancer," *European Journal of Cancer* 40(7), 1021–30.

Fischer, M.M., 1998. "Computational Neural Networks: An Attractive Class of Mathematical Models for Transportation Research," in V. Himanen, P. Nijkamp, and A. Reggiani (eds), *Neural Networks in Transport Applications* (pp. 3–20). Aldershot, England: Ashgate Publishing Ltd.

_____, 2001a. "Central Issues in Neural Spatial Interaction Modeling: The Model Selection and the Parameter Estimation Problem," in M. Gastaldi and A. Reggiani (eds), *New Analytical Advances in Transportation and Spatial Dynamics* (pp. 3–19). Aldershot, England: Ashgate.

_____, 2001b. "Computational Neural Networks – Tools for Spatial Data Analysis," in M.M. Fischer and Y. Leung (eds), *GeoComputational Modelling. Techniques and Applications* (pp. 15–34). Berlin, Germany: Springer-Verlag.

Frees, E.W., 1995. "Assessing Cross-Sectional Correlation in Panel-Data," *Journal of Econometrics* 69(2), 393–414.

Friedman, M., 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association* 32, 675–701.

Herbrich, R., M. Keilbach, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer, 1999. "Neural Networks in Economics. Background, Applications and New Developments," in T. Brenner (ed.), *Computational Techniques for Modelling Learning in Economics* (pp. 169–93). Dordrecht: Kluwer Academic Publisher.

Himanen, V., P. Nijkamp, and A. Reggiani (eds), 1998. *Neural Networks in Transport Application*: Ashgate, Aldershot.

Longhi, S., P. Nijkamp, A. Reggiani, and E. Maierhofer, 2005. "Neural Network as a Tool for Forecasting Regional Employment Patterns in West Germany," *International Regional Science Review* 28(3), 330–46.

Patuelli, R., S. Longhi, A. Reggiani, and P. Nijkamp, 2003. "A Comparative Assessment of Neural Network Performance by Means of Multicriteria Analysis: An Application to German Regional Labor Markets," *Studies in Regional Science* 33(3), 205–29.

_____, 2007. "Forecasting Regional Employment in Germany by Means of Neural Networks and Genetic Algorithms," *Environment and Planning A* forthcoming.

Profit, S. and R. Tschemig, 1998. "Germany's Labor Market Problems: What to Do and What Not to Do?" *Ifo Studien* 44(3), 307–25.

Reggiani, A., P. Nijkamp, and E. Sabella, 2000. "A Comparative Analysis of the Performance of Evolutionary Algorithms," in A. Reggiani (ed.), *Spatial Economic Science. New Frontiers in Theory and Methodology* (pp. 332–54). Berlin: Springer-Verlag.

Rumelhart, D.E. and J.L. McClelland, 1986. *Parallel Distribute Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Scarelli, A. and L. Venzi, 1997. "Nonparametric Statistics in Multicriteria Analysis," *Theory and Decision*, 43, 89–105.

Statistisches Bundesamt, 2002. *6,8 Mill. Menschen in Deutschland arbeiten Teilzeit*. Retrieved November 3rd, 2004, from http://www.destatis.de/presse/deutsch/pm2002/p3770031.htm

Swanson, N.R. and H. White, 1997a. "Forecasting Economic Time Series Using Flexible versus Fixed Specification and Linear versus Nonlinear Econometric Models," *International Journal of Forecasting* 13, 439–61.

_____, 1997b. "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," *The Review of Economic and Statistics* 79, 540–50.