Multicriteria Analysis of Neural Network Forecasting Models: An Application to German Regional Labour Markets

Roberto Patuelli[a], Simonetta Longhi[b], Aura Reggiani[c], Peter Nijkamp[d]

[a] School of Public Policy, George Mason University, Fairfax, Virginia; e-mail: rpatuel1@gmu.edu
[b] Tinbergen Institute of Amsterdam, Keizersgracht 482, 1017 EG Amsterdam (The Netherlands); e-mail: longhi@tinbergen.nl
[c] Department of Economics, Faculty of Statistics, University of Bologna, Piazza Scaravilli 2, 40126 Bologna (Italy); e-mail: reggiani@economia.unibo.it
[d] Department of Spatial Economics, Free University, De Boelelaan 1105, 1085 HV Amsterdam (The Netherlands); e-mail: pnijkamp@feweb.vu.nl

ABSTRACT

This paper develops a flexible multi-dimensional assessment method for the comparison of different statistical-econometric techniques based on learning mechanisms, with a view to analysing and forecasting regional labour markets.
The aim of this paper is twofold. A first major objective is to explore the use of a standard choice tool, namely Multicriteria Analysis (MCA), in order to cope with the intrinsic methodological uncertainty on the choice of a suitable statistical-econometric learning technique for regional labour market analysis. MCA is applied here to support choices on the performance of various models – based on classes of Neural Network (NN) techniques – that serve to generate employment forecasts in West Germany at a regional/district level. A second objective of the paper is to analyse the methodological potential of a blend of approaches (NN-MCA) in order to extend the analysis framework to other economic research domains, where formal models are not available, but where a variety of statistical data is present. The paper offers a basis for a more balanced judgement of the performance of rival statistical tests.

1 Need for a New Statistical Test Framework

The modern information age has dramatically increased the scientific potential to handle large scale data sets. Simulation of 'big models' has become a popular modelling activity, as the computational capacity of modern computers has exhibited a sky-rocketing pathway. The good old days of statistics and econometrics, which were for researchers a 'serious play to estimate one model a day' using standard ordinary least squares techniques, have passed by. We are now able to estimate an enormous range of model specifications under different background conditions, with a large set of sensitivity tests, and with the help of different aggregation levels of endogenous variables. Illustrative for this new situation is the title of a recent article by Sala-i-Martin (1997) on "*I Just Ran Two Million Regressions*".

The new data situation has prompted new challenges to both researchers and policy-makers. Researchers have to be selective regarding the choice of a method that is suitable for analysis and forecasting, while policy-makers have to be alert on the results – and in particular the robustness – of predictions offered to them.

The great rise in estimating alternative models has prompted a new interest in econometric-statistical model tests. The presence of a great diversity of model runs – and related results – leads to the inevitable question: which statistical model performs best? Modern standard statistical software packages offer a variety of test statistics, starting from R^2-values or t-values to more sophisticated statistical test values. The problem is then that the values of these statistical measures often mirror only a part of the statistical performance of models, so that essentially a multicriteria problem emerges where alternative model results have to be evaluated in terms of a multidimensional assessment scheme comprising different statistical indicators.

Nowadays, a modelling experiment is normally accompanied by a range of performance tests. And hence, we are essentially facing a situation of multicriteria analysis, where a set of alternatives (i.e., alternative model specifications) has to be judged on the basis of a set of rival criteria (i.e., different statistical test indicators). This is a challenging research question, as we are increasingly facing forecasting problems with large data sets, but without a formally specified and estimated structural model.

The present paper will address the above issue of robustness of statistical performance of large data systems in regard to alternative test possibilities. Examples are housing market data, transport behaviour data, stock market data or labour market data. We will deploy here – by way of illustration – a large database on German regional labour market conditions, which has been used by means of Neural Network (NN) methods in order to estimate regional labour market forecasting

models[1]. They will use a range of adjusted statistical tools, e.g. genetic algorithms. A range of different tools will next be applied to the above data base, each of them leading to a vector of different statistical performance test values. MCA is then used to develop an overall multidimensional assessment scheme. In the next section we will briefly describe some prominent methods in MCA, based on pairwise comparison. Then we will offer a description of the database. Subsequently, the statistical-econometric tests are carried out, followed by a presentation of the MCA method on the test results. The paper will be concluded with some retrospective and prospective remarks.

2 Multicriteria Analysis Methods: An Introduction

2.1 General Remarks

The present paper centres around the choice of a proper methodology for forecasting against the background of different and competing techniques, which can be judged by means of different statistical performance criteria. This is essentially a multi-dimensional choice problem.

Multicriteria analysis (MCA) is a choice-support tool developed for systematic evaluation of complex problems (see, among others, Nijkamp and Voogd, 1985). This kind of methodology is nowadays largely applied because of its many advantages in evaluation experiments. Specifically, MCA permits to choose between – and to identify a ranking of – different alternatives (called *alternatives*) when there is not a clear dominance of one alternative over the others.

Representing an analytical and multidisciplinary support for the policy analyst, MCA provides a solid base for the analysis of complex policy and choice problems. One of the principal characteristics of MCA is, in fact, the possibility '*not to end up with a single and "forced" solution dictated by a researcher but with a spectrum of feasible solutions from which a choice can be made*' (Hinloopen and Nijkamp, 1990, p. 2)[2]. MCA provides an array of dominant alternatives, which will be subject to the judgement of the policy-maker.

In order to evaluate conflicting alternatives, it is necessary to define a set of so called *criteria*, which represent the relevant aspects influencing the choice between the alternatives. The vectors containing the values of each alternative for all the criteria form the *impact matrix*, which therefore contains the entirety of the information available. Criteria can contain – depending from the MCA

[1] Explanation for the empirical application of Neural Network Analysis and implementation of Genetic Algorithms is later presented in Annex A.

[2] See also Nijkamp and Voogd (1985, p. 63): *'… the meaning of systematic evaluation for public decision making is not primarily the identification of 'the optimal and unambiguous solution'…'.*

method that is going to be applied – different kinds of information, both quantitative and qualitative, either empirically acquired or subjective values. The flexibility of MCA, which is able to incorporate different types of decisional criteria, opens up to discussion about heterogeneous approaches to the decisional process and to evaluation. As a consequence, a broader range of agents – and knowledge – can be involved in the process, in order to come with a complete set of criteria/attributes.

Depending on the number of alternatives that methods are able to evaluate, a general classification is usually done between continuous and discrete methods (respectively for infinite and finite number of alternatives). In the remaining part of this section, we will discuss the main characteristics of different types of discrete methods.

Concordance Analysis (CA), which is one of the principal families of discrete methods based on pairwise comparisons, has been mainly developed by Bernard Roy and is based '*on the definition of the individual preferences system as a base for defining the meaning to be attributed to the decision rule*' (translation from De Montis, 2001, p. 50). In fact, the ranks of the alternatives for each criterion – at different levels of preference (De Montis, 2001) – represent the leading classifying rule. Consequently, the analyst role is fundamental in choosing which criteria are useful in the analysis and which are not, since non-significant criteria tend to bias results. The main critique to CA regards this possibility, because of the influence the analyst can have on the decision-maker by choosing the criteria.

CA is often compared to another prominent class of discrete methods, the Multi-Attribute Utility Theory (MA). This methodology was firstly developed by Keeney and Raiffa (1976) and was inspired by the seminal work of Von Neumann and Morgenstern (1947). The subjective expected utility theory, on which MA is based, involves indeed the presence of a decision-maker who expresses his preferences through utility functions.

Although commonly used, MA theory is often criticized owing to the difficulties in studying the utility functions and using formal mathematical relations. Furthermore, CA is frequently preferred to MA in the field of regional and environmental planning, since incomparability or indifference relationships between alternatives better fit to uncertain phenomena, which are frequent in economics. In CA the analyst plays an active role in tuning the instruments for the particular objective. While in the MA theory the attributes correspond to the characteristics of the alternatives, criteria (in CA) refer to the entirety of the consequences associated to each alternative.

CA is characterised by the presence of concordance/discordance indexes, which are used in order to rank the alternatives, through one-on-one comparisons. Differences about the way in which these comparisons are led gave birth to several different methods in CA.

A first distinction in CA can be made between the class of quantitative[3] and qualitative approach techniques. The first ones are usually able to deal with cardinally expressed criteria, while the techniques belonging to the second class can employ qualitative information criteria.

The next section will introduce a related method, called Regime Method, which has been used in our framework.

2.2 The Regime Method

The methodology applied in this paper, in the context of MCA, is called *Regime Method* (RM). RM (Hinloopen *et al.*, 2002) belongs to the class of discrete decision-making methods, in particular to the one of CA.

Although categorised between the qualitative methods, RM is instead able to employ both cardinal and ordinal criteria. These mixed values are homogenised through standardised scales[4] referring to the relative position of each alternative in the range of values of each criterion.

In order to assess the dominance relationships between the alternatives, RM introduces paired comparisons between the alternatives for each criterion. Different criteria are made comparable through a standardisation process. Being S_{ij} the value of alternative *i* for criterion *k*, its standardised value $V_j(S_{ij})$ is:

$$V_k(S_{ik}) = \frac{\max\{S\min_k, \min\{S\max_k, S_{ik}\}\} - S\min_k}{S\max_k - S\min_k} \tag{1}$$

where $S\min_k$ and $S\max_k$ are the minimum and maximum values observed (or accepted) for criterion *k* respectively.

The difference between the standardised values of alternatives *i* and *i'* for criterion *k* is then calculated as:

$$D_k(S_{ik}, S_{i'k}) = V_k(S_{ik}) - V_k(S_{i'k}) \tag{2}$$

Consequently, the sum $D(S_i, S_{i'})$ of the values $D_k(S_{ik}, S_{i'k})$ for all of the *k* criteria represents the aggregated dominance relationship between alternatives *i* and *i'*. When both the types of signs are present in the addends – so that there is not a certain winner, assigning a weight to each criterion is useful to determine dominance relations. The weight vector defines the importance of each criterion. The resulting equation is then:

---

[3] In this framework a well-known software is ELECTRE (see Roy, 1991).

[4] In case of mixed data (partly ordinal and partly cardinal) in the criteria, the ordinal elements are re-calculated through a standardisation process (see Hinloopen *et al.*, 2002) in a scale from –1 to +1, in order to be compatible with cardinal data.

$$D(S_i, S_{i'}) = \sum_k w_k D_k(S_{ik}, S_{i'k}) \tag{3}$$

where w_k is the value of the weight vector for criterion k. If $D(S_i, S_{i'}) > 0$, then the alternative i is preferred to the alternative i'. In the case of ordinal criteria, $D(S_i, S_{i'})$ is stochastic, so that its values are associated to a probability distribution[5] and $p_{ii'} = prob(D(S_i, S_{i'}) > 0)$ represents the probability that alternative *i* is winning a comparison from alternative *i'* (Hinloopen *et al.*, 2002).

Finally, in order to assess the preference probability of each alternative, the probability value p_i is calculated as follows:

$$p_i = \frac{\sum_{i' \neq i} p_{ii'}}{N - 1} \tag{4}$$

It is important to highlight that different weight vector ranks can be used for different choice possibilities. In particular, each choice represents the different priorities given to the many aspects of the evaluation problem.

In conclusion, MCA may be a meaningful tool in selecting a best performing alternative from a range of competing options. It may then be helpful in identifying a proper forecasting tool for complex data situations.

3 A Concise Introduction to Neural Network Analysis

3.1 Introduction

Large data sets have become rather common in social science research. They often reveal a hidden structure, which has to be identified in order to use them for spatio-temporal forecasts. In various cases, formal econometric models are not available. Traditionally, spatio-temporal time series analysis (e.g., based on ARIMA or VAR techniques) has been used. More recently, NN approaches have gained much popularity.

The present paper will address only the potential of computational NN methods.

As pointed out in Section 1, the NN method, which in a second stage we extended by means of a Genetic Algorithm (GA) approach, will be applied here in order to offer short-term employment forecasts for the regional labour market in West Germany. Since the main aim of the present paper is an MCA application in order to identify the most 'suitable' NN models for forecasting purposes,

[5] See Hinloopen *et al*. (2002) for details about the probability assumptions for ordinal data.

the introduction of NNs and GAs will be restricted to the main characteristics of these two approaches. The related illustration will be outlined in the next two sections.

3.2 Neural Network Methodology

NN methods are essentially statistical goodness-of-fit techniques based on learning principles, where, through repetitive experiments of individual data, a hidden structure is identified. NN models, initially developed to explain and imitate the functioning of the human brain (see e.g. Rumelhart and McClelland, 1986), have been applied to a large variety of problems ranging from pattern recognition to transportation (Himanen *et al.*, 1998; Reggiani *et al.*, 2000). For an historical review of the NN methodology we refer, among others, to Taylor (1997); for an overview of NN applications in the economic field we refer to Herbrich *et al.* (1999).

Like in the human brain computation, an artificial NN is based on the principle of distribution of the activity in a high number of calculation units (the neurons) strictly connected and working in parallel. More in detail[6], neurons are organized in layers: one input layer receiving the information to be processed, one output layer providing the final output of the network, and a certain number of layers of 'hidden' neurones[7]. The one-to-one connections between neurons are represented by means of weights. Each unit processes the information received from the preceding layer and transmits the results to the succeeding layer. In many NNs, learning takes place by recursively modifying the weights (the initial set of weights is randomly chosen), with the aim to find the set of weights that offers the most appropriate results. The so-called supervised NN is able to learn the pattern linking input and output on the basis of a set of previously solved empirical examples (the training set). After a successful training, the NN should be able to generalize the example proposed and to offer the right output pattern.

The most popular way to find the best set of weights is the back propagation (BP) algorithm, which is composed of two steps. In the first step the input pattern (namely the set of training examples) is analysed by the network on the basis of the current set of weights, to compute 'provisional' results. The provisional results are then compared to the expected (from the set of solved examples) ones, and the error is computed. The error is then backpropagated from the last to the first layer and then the weights are modified in such a way to minimise the average error produced. The algorithm is re-iterated up to the point where the error reaches an acceptably low

[6] For simplicity we restrict this NN introduction to the so-called feed-forward NN since this is the NN model we used in our application on regional labour market forecasts. For more details on different types of NNs we refer, among others, to Sarle (1997).

[7] The number of hidden layers may virtually vary from zero to infinite. However, Kuan *et al.* (1994) demonstrated that a three layer NN – with only one hidden layer – is able to approximate almost any function.

value, or the process reaches the pre-defined number of iterations (number of epochs). One of the main inconveniences of the BP algorithm is that it may get stuck into local minima; some suggestions on how to avoid this problem can be found in Fischer (2001a).

Other difficulties we encountered in our empirical application consist in the choice of the NN architecture and in the possibility of overfitting the data. As pointed out by Fischer (2001b) an inappropriate choice of the NN architecture – namely the number of hidden layers, hidden neurones and some other learning parameters – or an inadequate learning procedure – concerning, for example, the choice of the training set and of number of epochs – can cause the failure of the NN in generalising the pattern of examples presented. More in detail, we have overfitting when the model is only able to perfectly represent the random fluctuations present in the data and therefore fails in the process of generalising the results to make them useful for out-of-sample analyses and forecasts. In order to reduce the possibilities of overfitting the data, we used the technique of 'early stopping' (see Sarle, 1997), in which the NN is trained until the error on a further validation data set deteriorates. For this purpose the data set has been split in three sub-sets: training set, validation set and test set. The training set has been used to find the best set of weights; the validation set has been used to tune the NN parameters and to find the best architecture; the test set has been used to evaluate the performance of the models proposed. Concerning the selection of the NN architecture, since no exact rules helping the choice exist, we adopted a large number of different NN architectures, until the most suitable one emerged. As it will become clearer in the next sub-section, this procedure may be turned into a more automatic process by enhancing the NN by means of the GA algorithm.

In the next sub-section we will briefly illustrate the GA methodology we will use in our empirical application.

3.3 Genetic Algorithms

GAs belong to a class of computer-aided optimisation tools named Evolutionary Algorithms (EAs). EAs are search methods of human behaviour mimicking natural biological evolution (Reggiani *et al.*, 2000), since these methods employ – in the social sciences – computational models trying to map out the design and structure of evolutionary biological processes.

In particular, GAs represent one of the most widely used classes of EAs. GAs are stochastic global search methods, which imitate the genetic evolution processes, on the basis of the well-known Darwinian law of 'survival of the fittest' (see, e.g., Holland, 1975). In fact, the algorithms stochastically explore a population of individuals, which represent the potential solutions of the given problem – in our case the different configurations of NN parameters – by identifying and

creating individuals 'better' fitting the objective at hand. Particularly, GAs use selection, mutation and recombination operators to generate new sample points in a search space (Fischer and Leung, 1998). The strength of GAs is therefore given by their ability to update an entire population during each iteration of the algorithm (Reggiani *et al.*, 2001).

The evaluation of the individuals' performance is usually given by a fitness function. GAs are in fact able to search for the individuals that minimise this function. At every iteration the fitness function is calculated, and a new generation of individuals is created by the action of genetic operators on a set of individuals from the previous generation. For an in depth overview of GAs methods and hypotheses we refer to Fischer and Leung (1998) and Reggiani *et al.* (2000, 2001).

In this case study, GAs are used in order to optimise NN performance. The aim is to obtain better generalisation properties from the NN (for an explanation of criteria evaluating generalisation properties we refer to Section 5.1). A brief description of the results emerging from the implementation of GA within NN models is presented in Annex A.1.

The next section will describe the empirical application concerning the case study and the use of MCA in order to evaluate the NN models according to their forecasting and computational potential. Thus, the main aim of the paper is not to find out which time series method in general has the best forecasting potential, but to identify, from the class of neural network and genetic algorithm methods, the one with the best predictive potential, by using MCA techniques.

4 Empirical Applications: The Case of West German Labour Market and the Application of Multicriteria Analysis

4.1 The Data Set

The statistical experiment in this section is rather straightforward. We start with an extensive spatio-temporal data base on labour market conditions. Then we apply NN and GA methods to make a 'forecast' for the last year for which we have data available. Subsequently, we compare these forecasts with actual realisations, and calculate various statistical performance measures. Since we use various statistical methods, the resulting problem is an MCA problem, leading to the question: which statistical learning method gives the best overall predictive performance?

Before dealing with the empirical application, we offer here some details on the data set at hand. The data set available[8] is organised as a panel of 327 districts and 13 years (from 1987 to 1999),

[8] For further information about the data set we refer to Longhi *et al.* (2002).

containing information about the total number of persons employed[9] every year on June 30th. Following the BfLR/BBR-typology, the 327 districts can be clustered by means of a cross tabulation of centrality and population density in 9 economic regions[10] (see, for details, Blien and Tassinopoulos, 1999). Information about daily wages is available as well. Thus, the data base comprises data on employment, wages for German regions and sectors. The reason why we focus on German labour market data is that with rising unemployment levels, the German government wants to know the related social security expenditures, and hence needs to have reliable forecasts of (un)employment in the next year.

On the basis of this employment and wages data, Longhi *et al.* (2002) proposed several NN models to make short-term forecasts of the total number of employees at a district level for West Germany. A brief introduction on the NN applications – also embedding GAs techniques – is given in Annex A.1-2. Once more, it ought to be noticed that NN and GA methods do not require a fully specified econometric model, but only an extensive data set with different variables describing the relevant issues. Their aim is to identify unknown patterns in such data in order to use them for forecasts by means of learning principles.

In the next sub-sections the performance of both NN models and NNs employing GAs will be considered as a basis for a series of MCA experiments, aiming to evaluate the forecasting ability of these NN techniques.

4.2 Evaluation Criteria and Assessment of Models

The assessment of the statistical-econometric performance of models is fraught with many difficulties. Whereas in the past only a few simple tests (such as the t-square, the b-value and the standard deviation) were deployed, we deserve nowadays – as a consequence of the large scale computing potential of modern computers – an avalanche of statistical indicators which all serve to assess the reliability, robustness or predictive precision of models. This is essentially an MCA

---

[9] The total number of employees can be subdivided in 9 economic sectors:

1- Primary sector	**2-** Industry goods	**3-** Consumer goods
4- Food manufacture	**5-** Construction	**6-** Distributive services
7- Financial services	**8-** Household services	**9-** Services for society

[10] The BfLR/BBR (BBR is the Bundesanstalt für Bauwesen und Raumordnung, Bonn, which former name was Bundesforschungsanstalt für Raumordnung und Landeskunde (BfLR)) district typologies are:

A. Regions with urban agglomeration	**1.** Central cities	**2.** Highly urbanised districts
	3. Urbanised districts	**4.** Rural districts
B. Regions with tendencies towards agglomeration	**5.** Central cities	**6.** Highly urbanised districts
	7. Rural districts	
C. Regions with rural features	**8.** Urbanised districts	**9.** Rural districts

Source: IAB - Institute for Employment Research, Nuremberg, Germany

problem, as the performance of alternative estimations models is judged against various competing test statistics.

In our comparative study, we aim to compare the above mentioned models on the basis of eight criteria listed in Table 1. The first three criteria – Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) – are statistical indicators referring to the error incorporated in the estimated models used for ex-post (or retrospective) forecasts; in general, a forecast is considered to be better the closer the value of the indicator at hand is to zero.

Since the presence of a large number of weights and/or epochs may slow down the training process of the NN, the need for criteria able to take also into account the differences between "light" and "heavy" models emerges. Consequently, two more criteria have been introduced: the number of epochs (NE) and the number of weights (NW). Low values for these criteria indicate models that require fewer computations and can be trained in a shorter time.

The two subsequent criteria, the stability (STAB) and the generalisation indicator (GEN), intend to assess the reliability of the NN models. More in detail, STAB is an indicator of the dissimilarity in the performance between the first and the second test year[11], measured by the absolute difference of the values of the statistical indicators concerned. This indicator relies on the assumption that a small difference between the two test years may signify a more stable behaviour of the network. The GEN criterion indicates whether the models are able to efficiently generalise the information contained in the training set. This criterion, which is the only one that can have both negative and positive values, is built as a sum of differences among the indicators calculated on the training and test sets, respectively.

The last criterion, namely daily wages (WAGE), refers to the economic relationship existing between the level of employment and earnings. From an economic-theoretical perspective, a model comprising information about wages is supposed to be more easily interpretable.

Once the criteria have been defined for each alternative, the impact matrix in an MCA setting is filled in with the values of the criteria concerned. It is important to note that not all scenarios evaluated make use of all criteria.

The next section will present an MCA experiment based on the models forecasting the year 2000. In the subsequent section, an MCA will be performed for models forecasting the year 2001.

[11] The choice of the NN structures has been originally based on the performance of the models on a two-years test set (1997/98 and 1998/98). This procedure was carried out only for models making forecasts for the year 2000. Models forecasting 2001 were trained using only one test year.

Table 1 – Evaluation Criteria

Criteria	*Definitions*[12]
MSE	Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NE	Number of Training Epochs (see Annex A.3)
NW	Number of NN Weights (see Annex A.3)
STAB	Difference in the models' performance in the two test sets used (1997/98 and 1998/99)
GEN	Generalization Ability
WAGE	Inclusion of the variable 'wage'

4.3 Comparing Models: Forecasts for the Year 2000

In this section the characteristics and the results of the models developed and used forecasting the employment in the year 2000 will be analysed through the use of MCA. The impact matrix used for this purpose is shown in Table 2.

The values reported for the first three criteria show that there would be no doubt about which model to choose if the statistical indicators were the only choice parameter, since model B clearly outranks – although sometimes with minimum differences – the other models. This result is acceptable for alternatives providing similar criteria, since the MCA aims to propose a set of valid (better) alternatives, which are finally subjected to the evaluation of the responsible actor.

Since the information embodied in the first three criteria are somewhat similar it may be useful – in order to have a more complete analysis – to add further criteria relative to the features of the models compared. The introduction of more criteria has two main outcomes: on the one hand it increases the amount of information about the intrinsic characteristics of the models and their final statistical performance. On the other hand, by introducing more conflicting criteria, it increases the complexity of the choice. MCAs are then necessary in determining a rank among the alternatives.

---

[12] The models are compared using the following statistical indicators:

- Mean Absolute Error (x_1): $MAE = [\Sigma_i |y_i - y_i^f|]$
- Mean Square Error (x_2): $MSE = [\Sigma_i (y_i - y_i^f)^2] /N$
- Mean Absolute Percentage Error (x_3): $MAPE = [\Sigma_i |y_i - y_i^f| / y_i] *100/N$

where: y_i is the observed value (target); y_i^f is the forecast of the model adopted (NN); y^a is the average of the observed values; N is the number of observations/examples.

Furthermore, the STAB and GEN indicators are calculated as follows:

- Stability criterion: $STAB = 0.5 * (|ARV_t – ARV_{t-1}| + |MSE_t – MSE_{t-1}|)$

where t-1 is the growth rate of employment between 1997 and 1998; t is the growth rate of employment between 1998 and 1999 and the Average Relative Variance is $ARV = [\Sigma_i (y_i - y_i^f)^2] / [\Sigma_i (y_i - y^a)^2]$

- Generalisation criterion: $GEN = \Sigma^3_i ({}^{train}x_i – {}^{test}x_i) / (0.5 * ({}^{train}x_i – {}^{test}x_i))$

where: ${}^{train}x_i$ is the value of the i^{th} indicator calculated on the train set; ${}^{test}x_i$ is the value of the i^{th} indicator calculated on the test set.

Table 2 – Impact matrix for models making forecasts for the year 2000

Criteria	*Model A*	*Model B*	*Model C*	*Model D*	*Model E*	*Model AW*	*Model DW*
MSE	1890058	1743597	2728751	2269495	15802724	2807580	3398917
MAE	709.147	704.580	829.573	772.586	2496.202	808.317	950.148
MAPE	1.161	1.156	1.278	1.228	5.360	1.211	1.435
NE	500	800	150	350	350	200	200
NW	220	55	207	230	310	230	216
STAB	0.212	1.991	0.165	0.184	0.229	0.306	0.451
GEN	-0.019	0.579	-0.201	-0.183	-1.331	-0.177	-0.425
WAGE	1	1	1	1	1	2	2

Criteria	*Model A-GA*	*Model B-GA*	*Model C-GA*	*Model D-GA*	*Model E-GA*	*Model AW-GA*	*Model DW-GA*
MSE	2883405	1902278	2370246	3766407	2228661	1892359	2696994
MAE	874.076	719.859	775.467	1055.872	753.103	712.334	791.299
MAPE	1.341	1.169	1.215	1.601	1.168	1.161	1.188
NE	1100	700	700	200	800	800	900
NW	660	231	3456	698	558	529	720
STAB	0.821	1.449	0.501	0.430	0.214	0.350	0.395
GEN	-0.376	1.569	-0.269	0.235	-0.279	-0.115	-0.286
WAGE	1	1	1	1	1	2	2

Note: models employing wages assume value '2' for the WAGE criterion.

Before applying the MCA to our comparative case study, we calculated, as suggested by Scarelli and Venzi (1997), a non-parametric statistic, called Friedman statistic, whose aim is to confirm whether the ranks of the alternatives – made on the basis of each criterion – differ significantly. If this is statistically confirmed, then the need of a multicriteria approach is proved. If, instead the test reveals no significant differences in the rank orders of the alternatives, then the multiple criteria ranking is not necessary. The base assumption is the following: if the alternatives have dissimilar evaluations on the criteria, then the sums of the ranks of the alternatives for each criterion will be different. The null hypothesis (H_0) to be tested is that there are no systematic differences between the alternatives. If this hypothesis is valid, then the ranking of the alternatives – on the basis of the criteria used – is essentially random. The Friedman statistic which defines the test and which, for a large number of cases, has a c^2 distribution with (N – 1) degrees of freedom, is:

$$S^* = \frac{12S}{KN(N+1)} \tag{5}$$

where N is the number of alternatives evaluated and K is the number of the criteria considered. S is an indicator of the variability of the alternatives' ranking sums[13], calculated as:

$$S = \sum_j \left(S_j - S_e\right)^2 \tag{6}$$

and where S_j is the sum of the ranks that alternative j has for each criterion k and S_e is the value of the expected sum of ranks. In our case study the values of S^*, which is 27.67, is significant at (almost) 1% level, suggesting that MCA is necessary in order to define a ranking of the alternatives. This also prompts the need to specify weights for the best indicators.

In this framework four different weight vectors, and therefore four different scenarios, summarized in Table 3, have been proposed and compared.

In 'Scenario 1', every criterion has been given equal weight. Since this implies that none of the criteria is preferred to the others, this is the simplest analysis that may be carried out. By giving equal weight to each criterion, a possible bias is introduced in the analysis. It can be noted, for example, that the statistical indicators provide essentially the same information on the performance of the models, and that the NE and NW criteria evaluate in two different ways the same aspect – namely the computational complexity – of the alternatives. In this way the weight given to the models' performance is indeed multiplied by three, while the weight given to the models' computational complexity is multiplied by two. The next scenarios try to address this problem.

In order to define the weights in the remaining three scenarios, the eight criteria have been grouped in four clusters. Since they give essentially the same kind of information about the goodness of fit of the models, we clustered the statistical indicators (MSE, MAE, MAPE) in the first group. The second group comprises the criteria NE and NW, which may be seen as a measure of the above mentioned model complexity. The indicators STAB and GEN, which are supposed to measure the stability and reliability of the models compared, are grouped in the third cluster. Finally, the remaining criterion – WAGE – is the only one belonging to the fourth group. In 'Scenario 2' each group has equal weight; the second column of Table 3 shows the resulting weights attached to each criterion. One possible criticism on 'Scenario 2' is that it does not provide sufficient importance to standard statistical indicators, which are usually the most conventional way to evaluate the quality of a model. In order to assess these requirements, two more scenarios – 'Scenario 3' and 'Scenario 4' – have been proposed. These two last scenarios weight the group of the statistical indicators with 40% and 50% of the total, respectively. The remaining weights have

[13] In detail, S_j is given by the sum of the rank of the j^{th} alternative for each criterion, where $r_{jk} > r_{hk}$ if the alternative k has a better value than alternative h for the k^{th} criterion.

been subdivided in equal parts between the remaining criteria, as shown in the third and fourth columns of Table 3.

Table 3 – Weights of each criterion for the different scenarios

	Scenario 1	*Scenario 2*	*Scenario 3*	*Scenario 4*
MSE	0.125	0.083	0.133	0.167
MAE	0.125	0.083	0.133	0.167
MAPE	0.125	0.083	0.133	0.167
NE	0.125	0.125	0.12	0.1
NW	0.125	0.125	0.12	0.1
STAB	0.125	0.125	0.12	0.1
GEN	0.125	0.125	0.12	0.1
WAGE	0.125	0.25	0.12	0.1

On the basis of these scenarios, various MCA experiments analyses have been carried out. Since it was not possible to analyse all 14 models simultaneously, due to software limitations[14], some preliminary MCAs have been performed. Particularly, in order to choose the best models belonging to the NN group as well as the best models belonging to the group of the NN-GAs, the first two MCAs have been carried out on NN and NN-GA models separately. The results are reported in Table 4.

Concerning the choice of the NNs, the preferred model seems to be different for each scenario. The results show similarities between 'Scenario 1' and 'Scenario 2', in which there is a dominance of model C and model AW with respect to all others. Likewise, also 'Scenario 3' and 'Scenario 4' have similar results, since the models with the best values for the statistical indicators (model A and model B) appear to be winners. Preliminary analyses (not shown here) were also carried out on the NN models by considering only the statistical indicators or the other criteria. The first MCA (on the statistical indicators) showed the dominance of models A, B, C, D and AW, while the second analysis (using only the other criteria) showed model C as dominant, suggesting that model C will have particularly favourable results in scenarios with a high weight given to the second group of criteria.

Concerning the NN-GA models, the MCAs show that model AW-GA is dominant in any case. Separate analyses using only the statistical indicators or only the remaining criteria show model AW-GA as the best one. models B-GA, E-GA and DW-GA represent, in each analysis, the second best choices.

[14] The Samisoft software (developed by Vreeker and Nijkamp (2001)) used for our analyses allows to carry out MCAs with a maximum of 10 alternatives. The questions arising from this constraint regard the influence that the models left

On the basis of these results, the 5 best NN and the 5 best NN-GA models emerging from 'Scenario 3' have been chosen for the final MCAs. The choice of 'Scenario 3' as the preferred one rests on the consideration that this scenario offers a good compromise between the weights given to the statistical indicators versus the weights given to the other criteria. The hypothesis implicit in this choice (of the best five models among the NNs and of the best five models among the NN-GAs) is that the models classified as the least preferred ones in the separate rankings would probably also be at the bottom of the table in a MCA in which we would include all available models. Consequently, the absence of the least preferred models will possibly not influence significantly the results, because of the fewer dominance positions they have on the other models.

Table 4 – Results from MCAs on NN and NN-GA models separately

	NN Models				*NN-GA Models*			
Rank	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 1	Scenario 2	Scenario 3	Scenario 4
1	Model C (0.79)	Model AW (0.89)	Model A (0.79)	Model B (0.83)	Model AW-GA (0.89)	Model AW-GA (0.94)	Model AW-GA (1)	Model AW-GA (1)
2	Model AW (0.7)	Model C (0.77)	Model B (0.72)	Model A (0.78)	Model B-GA (0.74)	Model B-GA (0.63)	Model E-GA (0.78)	Model B-GA (0.83)
3	Model A (0.59)	Model DW (0.58)	Model C (0.62)	Model D (0.62)	Model E-GA (0.68)	Model E-GA (0.59)	Model B-GA (0.72)	Model E-GA (0.67)
4	Model D (0.57)	Model D (0.48)	Model D (0.59)	Model C (0.61)	Model DW-GA (0.53)	Model DW-GA (0.58)	Model DW-GA (0.5)	Model DW-GA (0.43)
5	Model B (0.49)	Model A (0.44)	Model AW (0.58)	Model AW (0.46)	Model D-GA (0.42)	Model D-GA (0.53)	Model C-GA (0.27)	Model C-GA (0.33)
6	Model DW (0.35)	Model B (0.31)	Model DW (0.19)	Model DW (0.19)	Model C-GA (0.17)	Model C-GA (0.13)	Model D-GA (0.12)	Model D-GA (0.14)
7	Model E (0)	Model E (0.04)	Model E (0)	Model E (0)	Model A-GA (0.08)	Model A-GA (0.09)	Model A-GA (0.11)	Model A-GA (0.10)

In summary, in the final multicriteria experiments performed in this study, 50% of the alternatives is represented by NN models, while the remaining 50% is represented by NN-GA models. Particularly, the chosen models are models A, B, C, D and AW for the NNs group, and models B-GA, C-GA, E-GA, AW-GA, DW-GA for the NN-GAs group. The results of the MCAs carried out on these models (summarized in Table 5) show that the NN models seem to be for a greater part winning over NN-GA models. Because of their good values for criteria regarding the network complexity and reliability, model C and model AW are dominant in the first two scenarios. In 'Scenario 3' and 'Scenario 4', where the statistical indicators have comparatively higher weights, model B is the highest-ranked and model A always represents the second best choice. Furthermore,

out of the analysis could have had on the results, since the regime vectors are built, for each alternative, on the dominance relations with all other alternatives.

model AW-GA is always ranked third. Finally, the rankings assigned by these analyses confirm the dominance relations previously observed inside the NN and NN-GA groups. Concluding, it appears that the introduction of the GA in the NN models did not bring to a significant improvement in the models.

One reason for the dominance of NNs on NN-GAs may be the computational burden of the NN-GA models. In fact, the NN structures proposed by the GA approach – the NN-GA models – are often more complicated than the manually-chosen structures of the NN models. This finding was also confirmed by further analyses carried out using as criteria the statistical indicators and the other parameters separately, which show that the NN-GA models are noticeably outranked by the NN models for the second group of criteria.

Table 5 – Results from MCAs on NN and NN-GA models

Rank	*Scenario 1*	*Scenario 2*	*Scenario 3*	*Scenario 4*
1	Model C (0.83)	Model AW (0.93)	Model B (0.97)	Model B (0.91)
2	Model AW (0.8)	Model C (0.84)	Model A (0.88)	Model A (0.85)
3	Model A (0.72)	Model D (0.67)	Model AW-GA (0.67)	Model AW-GA (0.62)
4	Model D (0.69)	Model A (0.65)	Model AW (0.55)	Model D (0.56)
5	Model B (0.56)	Model AW-GA (0.54)	Model B-GA (0.48)	Model C (0.53)
6	Model AW-GA (0.48)	Model B (0.45)	Model E-GA (0.45)	Model B-GA (0.5)
7	Model B-GA (39)	Model B-GA (0.32)	Model D (0.4)	Model AW (0.46)
8	Model E-GA (0.29)	Model DW-GA (0.3)	Model C (0.34)	Model E-GA (0.36)
9	Model DW-GA (0.2)	Model E-GA (0.25)	Model DW-GA (0.21)	Model DW-GA (0.14)
10	Model C-GA (0.05)	Model C-GA (0.05)	Model C-GA (0.06)	Model C-GA (0.09)

Having conducted, in this section, different analyses on the models providing forecasts for the year 2000, we will offer a similar analysis of models providing forecasts for the year 2001 in the next section.

4.4 Comparing Models: Forecasts for the year 2001

This session will present additional multiple criteria experiments on the NN and NN-GA models forecasting the employment for the year 2001. Scenarios similar to the ones used in the latter section will be introduced and evaluated.

Like in the previous analysis, the impact matrix is built on the basis of the criteria illustrated in Section 5.1, although, as anticipated, one of the criteria (STAB) can not be used. In fact, since in this case the choice of the networks' structure has been based on a one-year test set, it was not possible to evaluate this criterion. Therefore, the number of criteria that have been used for our comparative analyses is only 7.

The number of models considered has also been reduced in this new experiment. Because of this, the models evaluated in this section are 10 (while in the previous section there were 14), equally divided between NN models and NN-GA models. For simplicity, the "2001" specification is not used in this section, since the experiments are based on 2001 models only.

Table 6 – Impact matrix for models making forecasts for the year 2001

Criteria	*Model A 2001*	*Model B 2001*	*Model D 2001*	*Model AW 2001*	*Model DW 2001*
MSE	4151746	8398732	3679887	3322544	3356230
MAE	1112.172	1663.389	1054.36	1011.711	1003.7
MAPE	1.8977	2.598186	1.832554	1.782615	1.757255
NE	400	350	550	850	450
NW	210	55	220	198	230
GEN	0.603	1.186	1.034	1.108	0.987
WAGE	1	1	1	2	2

Criteria	Model A-GA-2001	Model B-GA-2001	Model D-GA-2001	Model AW-GA-2001	Model DW-GA-2001
MSE	14256551	8993670	5519886	5777180	16198822
MAE	2180.079	1740.735	1223.313	1345.53	2250.689
MAPE	3.30698	2.706966	1.939949	2.186076	3.37501
NE	1700	1000	3100	200	1700
NW	441	10	198	909	506
GEN	-2.216	0.989	-0.405	0.538	-2.370
WAGE	1	1	1	2	2

Note: models employing wages assume value '2' for the WAGE criterion.

The impact matrix, containing the values of the criteria for all models (see Table 6), shows that NN models perform, on average, better than NN-GA models. In fact, only models D-GA and AW-GA seem to be competitive with the first group of models.

The statistical indicators regarding NN models present predominantly better values, as well as the GEN criterion. Like in the previous section, the criteria regarding the computational complexity of the networks show that NN-GA models are based on much more complicated networks, as they present higher values for both the NE and NW criteria.

Once again, the Friedman statistic has been calculated and tested in order to assess significant differences between the alternatives. The obtained value for S^* was 26.59, which equals to a significance level of 1%, permitting to refuse the null hypothesis of no difference between the evaluations of the alternatives.

This result once more confirms the need to evaluate the rank order of the presented models by means of MCA, since significant differences between them have been found. In order to do this, four scenarios have again been used. Slight differences can be found for the respective scenarios that have been used in the previous section, because of the different number of criteria.

As previously stated, 'Scenario 1' presents an equal weight for each criterion. The absence of the STAB criterion results, in comparison with the previous section, in a higher weight for each criterion and, particularly, for the group of the statistical indicators, which now represent 3/7 of the total weight importance.

In 'Scenario 2' the criteria have been subdivided in 4 groups of equal weight. The first group contains the statistical indicators criteria, while the second one comprises criteria NE and NW, representing the computational complexity. The third and fourth groups are respectively represented by the criteria GEN and WAGE. Note that in the previous section the third group also contained the criterion STAB.

Again, in 'Scenario 3' and 'Scenario 4', the group of the standard statistical indicators (the first three criteria) has been weighed as 40% and 50% of the total weight, respectively.

Table 7 - Weights of each criterion for the different scenarios

	Scenario 1	*Scenario 2*	*Scenario 3*	*Scenario 4*
MSE	0.143	0.083	0.133	0.167
MAE	0.143	0.083	0.133	0.167
MAPE	0.143	0.083	0.133	0.167
NE	0.143	0.125	0.15	0.125
NW	0.143	0.125	0.15	0.125
GEN	0.143	0.25	0.15	0.125
WAGE	0.143	0.25	0.15	0.125

Like in the previous section, separate analyses have been made on NN models and NN-GA models (see Table 8 for the results). The first analyses, carried out on the NN models, show, in the first three scenarios, the dominance of model DW. This is due to generally good values for nearly all criteria. Even in 'Scenario 4', in which model AW turns out to be dominant, model DW performs well, because of its two out of three dominant positions in the statistical indicators

group[15], which is given half of the total weight in this scenario. Furthermore, a good performance can be observed for model AW, mainly due to dominant positions for three criteria.

The analyses on the NN-GA models provide similar results. In fact, model AW-GA is clearly dominant in every scenario, despite of having only two winning criteria. Separate analyses based on the statistical indicators and the other criteria respectively showed that model AW-GA is dominant in both cases. A secondary choice is represented by model B-GA, which is ranked as second on three of the four scenarios.

Table 8 – Results from MCAs on NN and NN-GA models separately

	NN Models				*NN-GA Models*			
Rank	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 1	Scenario 2	Scenario 3	Scenario 4
1	Model DW (0.91)	Model DW (0.93)	Model DW (0.9)	Model AW (0.96)	Model AW-GA (0.85)	Model AW-GA (0.96)	Model AW-GA (0.98)	Model AW-GA (0.92)
2	Model AW (0.61)	Model AW (0.6)	Model B (0.67)	Model DW (0.79)	Model B-GA (0.77)	Model B-GA (0.78)	Model B-GA (0.77)	Model D-GA (0.71)
3	Model B (0.35)	Model B (0.55)	Model AW (0.51)	Model D (0.46)	Model D-GA (0.63)	Model D-GA (0.51)	Model D-GA (0.5)	Model B-GA (0.62)
4	Model A (0.32)	Model A (0.22)	Model A (0.25)	Model A (0.24)	Model DW-GA (0.13)	Model DW-GA (0.18)	Model DW-GA (0.15)	Model A-GA (0.18)
5	Model D (0.31)	Model D (0.2)	Model D (0.17)	Model B (0.05)	Model A-GA (0.13)	Model A-GA (0.07)	Model A-GA (0.1)	Model DW-GA (0.07)

In line with the results of separated NN and NN-GA analyses, results from the MCA, comprehensive of both types of models (see Table 9), show that model DW and, next, model AW appear to be the highest-ranked alternatives. As in the previous analyses, model AW surpasses model DW in 'Scenario 4'. The rankings obtained in the previous (separate) analyses are confirmed, although with only slight differences, due to very similar probabilities – as previously explained. The NN-GA models do not provide good performance, as it was evident from the impact matrix, because of their difficulties in generalising (which are likely to be a cause of inefficient statistical indicators) and because of their heavy structure, generated by the GA. The models that were classified at the bottom of the NN-GA analysis were found to be the worst of both types of models. The best performing GA model is AW-GA, which is ranked fourth in two scenarios.

In conclusion, this analysis showed that the introduction of the GA in the definition process of the network structure did not bring about significant improvements in the quality of the estimates. Besides, also other criteria show the relatively lower adequacy of NN-GA models, as they present the lowest values for the GEN criterion and for the NE and NW criteria, representing the

[15] An analysis carried out using the statistical indicators only as criteria showed that model DW is winning.

computational complexity of the models. Consequently, models AW and DW seem to be the ones that satisfy most of the requirements incorporated in the chosen criteria.

Table 9 – Results from MCAs on NN and NN-GA models

Rank	*Scenario 1*	*Scenario 2*	*Scenario 3*	*Scenario 4*
1	Model DW (0.96)	Model DW (0.96)	Model DW (0.95)	Model AW (0.98)
2	Model AW (0.81)	Model AW (0.8)	Model AW (0.76)	Model DW (0.91)
3	Model A (0.64)	Model B (0.67)	Model B (0.72)	Model D (0.76)
4	Model D (0.63)	Model AW-GA (0.64)	Model AW-GA (0.6)	Model A (0.66)
5	Model B (0.57)	Model D (0.54)	Model A (0.57)	Model AW-GA (0.48)
6	Model AW-GA (0.56)	Model A (0.54)	Model D (0.55)	Model D-GA (0.42)
7	Model B-GA (0.45)	Model B-GA (0.51)	Model B-GA (0.52)	Model B (0.41)
8	Model D-GA (0.28)	Model D-GA (0.23)	Model D-GA (0.22)	Model B-GA (0.27)
9	Model DW-GA (0.06)	Model DW-GA (0.08)	Model DW-GA (0.08)	Model A-GA (0.1)
10	Model A-GA (0.05)	Model A-GA (0.03)	Model A-GA (0.03)	Model DW-GA (0.01)

5 Conclusions

In the present paper, MCA, as a technique for evaluating the NN models' performance – in the framework of forecast experiments for the West German regional labour markets – has been explored. An additional goal of the paper was to test the potential of extended GA models in comparison with the conventional NN models.

On the basis of NN configurations already adopted in previous experiments (see Longhi *et al.*, 2002), analogous NN models have been developed here by employing GA techniques in order to automatically control the complexity level of NNs (quantity of layers and hidden neurons). Our results showed that NN-GA architectures, due to their complex structure, demanded high computational needs.

Several experiments concerning MCA were next carried out. While the alternatives focused on the above mentioned NN and NN-GA models, the criteria considered – on both a cardinal and ordinal scale – were addressing the main characteristics of the adopted models, like statistical

relevance, economic meaning and computational efficiency. This allowed us to obtain a broader view on the models' performance than the usual – more limited – statistical information.

In our application, four scenarios have been developed attaching different importance to the statistical judgement criteria. The configuration of the scenarios ranged from equal weights for all criteria to different group types of criteria.

Results emerging from the models forecasting the year 2000 appeared to be somewhat contradictory, since the winning models changed their position according to the chosen scenario. Furthermore, a less favourable performance was shown by NN-GA models, which, although presenting satisfactory statistical results, were indeed held down by their computational requirements and generalisation properties.

On the contrary, MCA results on models forecasting the year 2001 outlined more homogeneity, since two models (particularly, DW and AW) resulted to be dominant under all given scenarios. It is noteworthy that a good performance by the NN-GA models was not found, because of substandard values found for most of the criteria.

In summary, our analyses showed that, in an uncertain situation with contradictory values in the statistical impact table, MCA can be a useful tool in evaluating the NN models' performance and providing a related ranking based on priorities attached to each statistical test indicator individually. Obviously, if one or two models are clearly dominant, MCA will indeed confirm this dominance.

Further research directions would have to address an in-depth investigation of the use of GAs, since the extended GA models adopted in our empirical application lacked in performance, mostly due to their structural complexity. Furthermore, the identification of additional appropriate criteria could also offer more insight into the models' performance. Next, a comparison between the obtained rankings and the forecasting differences in the real employment volumes could undoubtedly be able to test better the power of this novel joint NN-MCA approach. And finally, it would be useful to extend the prediction range by not only investigating the performance of these statistical models for predictions one year ahead, but for several years ahead. Although basically the same approach could be used, this approach would allow us to test the robustness of predictions over a longer period.

Acknowledgements

The authors wish to thank the Institute for Employment Research (IAB), Nuremberg (Germany), and particularly Uwe Blien and Erich Maierhofer, for kindly providing the data.

The first author wishes to thank Michel Beuthe (Group Transport & Mobility (GTM), Facultés Universitaires Catholique de Mons, Belgium) for the useful discussions and organisational support, in the framework of a fellowship assigned to him by GTM.

The authors also wish to thank two anonymous referees for the useful comments.

References

Blien U., A. Tassinopoulos (1999), Forecasting Regional Employment with the Entrop Method, Paper presented at the European Congress of the Regional Science Association, Dublin, Ireland.

De Montis A. (2000), *Analisi Multicriteri e Valutazione per la Pianificazione Territoriale*, CUEC, Cagliari.

Fischer M.M. (2001a), Neural Spatial Interaction Models, in Fischer M. M., Leung Y. (eds.) *GeoComputational Modelling, Techniques and Applications*, Springer-Verlag Berlin, pp. 195-219.

Fischer M.M. (2001b), Central Issues in Neural Spatial Interaction Modeling: the Model Selection and the Parameter Problem, in Gastaldi M., Reggiani A. (eds.) *New Analytical Advances in Transportation and Spatial Dynamics*, Ashgate, Aldershot, UK, pp. 3-19.

Fischer M.M., Y. Leung (1998), A Genetic-algorithms Based Evolutionary Computational Neural Network for Modelling Spatial Interaction Data, *Annals of Regional Science* 32, pp. 437-458.

Herbrich R., M. Keilbach, T. Graepel, P. Bollmann-Sdorra, K. Obermayer (1999), Neural Networks in Economics. Background, Applications and New Developments, in Brenner T. (ed.) *Computational Techniques for Modelling Learning in Economis*, Kluwer Academic Publisher, Dordrecht, pp.169-193.

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

Hinloopen E., P. Nijkamp, P. Rietveld (2002), Multiple Criteria Choice Models for Quantitative and Qualitative Data, in Trzaskalik T., J. Michnik (eds.) *Advances in Soft Computing*, Physica-Verlag, Heidelberg , pp. 61-85.

Hinloopen E., P. Nijkamp (1990), Qualitative Multiple Criteria Choice Analysis, *Quality and Quantity* 24, pp. 37-56.

Holland, J.H. (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbour.

Keeney R.L., H. Raiffa (1976), *Decisions with Multiple Objectives*, John Wiley, New York.

Longhi S., P. Nijkamp, A. Reggiani, U. Blien (2002), Forecasting Regional Labour Market in Germany; an Evaluation of the Performance of Neural Network Analysis, Discussion Paper, Tinbergen Institute, Amsterdam.

Nijkamp P., H. Voogd (1985), An Informal Introduction to Multicriteria Evaluation, in Fandel G. (ed.), *Multiple Criteria Decision Methods and Applications: Selected Readings of the First International Summer School, Acireale, Sicily, September 1983*, in collaboration with Benedetto Matarazzo, Springer-Verlag Berlin, pp. 61-84.

Reggiani A., P. Nijkamp, E. Sabella (2000), A Comparative Analysis of the Performance of Evolutionary Algorithms, in Reggiani A. (ed.), *Spatial Economic Science. New Frontiers in Theory and Methodology*, Springer-Verlag, Berlin, pp. 332-354.

Reggiani A., P. Nijkamp, E. Sabella (2001), New Advances in Spatial Network Modelling: Towards Evolutionary Algorithms, *European Journal of Operational Research* 128 (2), pp. 385-401.

Roy B. (1991), The Outranking Approach and the Foundation of the ELECTRE Methods, *Theory and Decision* 31, pp. 49-73.

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

Sala-i-Martin, X.X. (1997), I Just Ran Two Million Regressions, *American Economic Review* 87, pp. 178-183.

Sarle W.S., (ed.) (1997), Neural Network FAQ, part 1 of 7: Introduction, periodic posting to the Usenet newsgroup comp.ai.neural-nets, URL: ftp://ftp.sas.com/pub/neural/FAQ.html

Scarelli A., L. Venzi (1997), Nonparametric Statistics in Multicriteria Analysis, *Theory and Decision* 43, pp. 89-105.

Taylor J.C. (1997), The Historical Background, *Handbook of Neural Computation* 1, pp. A1.1:1-A1.1:7, URL: http://library.thinkquest.org/18242/data/resources/nnhistory.pdf?tqskip1=1&tqtime=0429.

Von Neumann J., O. Morgenstern (1947), *Theory of Games and Economic Behaviour*, 2nd Ed., Princetown University Press, Princetown, NJ.

Vreeker R., Nijkamp P. (2001), Samisoft Software, Department of Spatial Economics, Free University, Amsterdam.

Annex A – The Empirical Application by means of Neural Network Models and Genetic Algorithms

A.1 Application of Neural Network Methods

The models developed in Longhi *et al.* (2002) aimed at making short-term employment forecasts at a regional level. These methods were developed/tuned to forecast the total number of employees in 2000 and in 2001. For reasons related to the training phase, the statistical models forecasting 2000 have been developed separately from the models forecasting 2001. For convergence reasons the models were tuned on employment growth rates between t and t+1, for the forecast in 2000, and on employment growth rates between t and t+2, for the forecast relative to the year 2001. In both groups of models many different NN architectures have been developed and (roughly) compared in order to choose the best way of introducing information and variables in the models. A summary of all models proposed in Longhi *et al.* (2002) and compared in the subsequent sections is given in Annex 3, Table A.1 and Table A.2, and is briefly described here. The results relative to the models' performance can be found in sections 4.3 and 4.4.

All statistical methods, proposed in the upper part of both Tables A.1 and A.2, use as input variables the (one year) lagged growth rate of sectoral employment. Other input variables are 'time', 'type', 'district' and 'wages'. More in detail, one of the main problems encountered in developing the NN models was the high number of cross-sections to be estimated. As a consequence, since considering the information as a panel would require too many weights, information about time had to be introduced in the models as a new variable. This was done alternatively by means of dummy variables (Model A) or by means of a qualitative[16] variable (Model B). Model C has the same inputs of Model A, plus a qualitative variable able to distinguish among the 327 districts. This can be seen as the correspondent of cross sectional fixed effects in a panel model. Model D and Model E have the same inputs of Model A, plus the variable 'type of economic region'. The main difference between the two models is that the new variable has been introduced as qualitative variable in Model D, and as dummies in Model E. Finally, information about daily wages has been introduced as new input variable in Model A, obtaining Model A+W, and in Model D, obtaining Model D+W.

The models developed to make employment forecasts with a time span of two (instead of one) years have names similar to the previous models, plus the suffix '2001'. They share the same inputs (not necessarily the same architecture) of the models with the same name that were developed to forecast 2000. Other than the lagged (two years) growth rate of sectoral employment, Model A-2001 and Model B-2001 also introduce information about time by means of dummy variables (Model A-2001) or by means of qualitative variable (Model B-2001). Model D-2001 has the same inputs of Model A-2001, plus the variable 'type of economic region', introduced as qualitative variable. Finally, Model AW-2001 and Model DW-2001, have the same inputs of Model A and Model D respectively, plus information about daily wages.

A concise characterisation of the various NN and GA methods employed in the present study can be found in Annex A.3.

A.2 Application of Neural Networks Implemented with Genetic Algorithms

In parallel to the experiments depicted in the previous sub-section, a new set of statistical models – employing GA as an optimisation tool – has been built. GAs have been used here in order to define the optimal configuration of the NN architecture, since GAs can automatically modify and propose new network parameters (e.g. the learning rate).

Since the input data have not been changed (even though the NN structure is different), the NN-GA models' performance can be directly compared to the NN performance (for example, Model A-GA is comparable to Model A).

It can easily be seen that the NN structures proposed by the GA techniques are often much more complex than the traditional NN models in terms of their NN structure[17]. On the one hand, this complexity mirrors the broader experimental possibilities of GAs, while on the other hand it is evident that a bigger and more complex architecture of the model may certainly slow down the elaboration process because of the need for more computational power. In all cases, we would always obtain 327 predictions. Thus, at the end we have a series of predictions generated by means of various NN and GA statistical models.

[16] The adopted NN software includes the possibility of introducing in the data set – as inputs – qualitative variables, which can also be defined in a non-numeric format.

[17] GA-enhanced models often present a higher number of layers and hidden neurons.

A.3 Details on Model Experiments

Table A.1: Summary of the models proposed in order to make forecasts for 2000

Model's Name	Models Characteristics
Model A *	Basic NN with time defined as dummies
Model B *	Basic NN with time defined as qualitative
Model C *	NN time as dummies + fixed effects
Model D *	NN time as dummies + type as qualitative
Model E *	NN time as dummies + type as dummies
Model AW *	Basic NN with time defined as dummies + wage
Model DW *	NN time as dummies + type as qualitative + wage
Model A-GA +	GA-developed basic NN with time defined as dummies
Model B-GA +	GA-developed basic NN with time defined as qualitative
Model C-GA +	GA-developed NN time as dummies + fixed effects
Model D-GA +	GA-developed NN time as dummies + type as qualitative
Model E-GA +	GA-developed NN time as dummies + type as dummies
Model AW-GA +	GA-developed basic NN with time defined as dummies + wage
Model DW-GA +	GA-developed NN time as dummies + type as qualitative + wage

Note: models with (*) have been developed by Longhi *et al*. (2002); models with (+) have been developed by the authors of the present paper.

Model A is a three-layer NN with 21 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 500 epochs to avoid overfitting.
Model B is a three-layer NN with 10 inputs, 5 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 800 epochs to avoid overfitting. The learning rate is set at 0.5.
Model C is a three-layer NN with 22 inputs, 9 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 150 epochs to avoid overfitting.
Model D is a three-layer NN with 22 inputs, 10 hidden neurones and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 350 epochs to avoid overfitting.
Model E is a three-layer NN with 30 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 350 epochs to avoid overfitting. The main difference between model D and model E is the way in which the qualitative variable "type" is introduced in the models. While model D treats the variable as qualitative information, model E treats it as a number of dummies variables.
Model A+W is a three-layer NN with 22 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 200 epochs to avoid overfitting.
Model D + W is a three-layer NN with 23 inputs, 9 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 200 epochs to avoid overfitting.
Model A-GA is a three-layer GA-developed NN with 21 inputs, 30 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 1100 epochs to avoid overfitting.
Model B-GA is a three-layer GA-developed NN with 10 inputs, 21 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 700 epochs to avoid overfitting.
Model C-GA is a three-layer GA-developed NN with 22 inputs, 15 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 700 epochs to avoid overfitting.
Model D-GA is a four-layer GA-developed NN with 22 inputs, 23 and 8 hidden neurones and 1 output. The activation function is a sigmo id, and the learning process was forced to stop after 200 epochs to avoid overfitting.
Model E-GA is a three-layer GA-developed NN with 30 inputs, 18 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 800 epochs to avoid overfitting. The main difference between model D-GA and model E-GA is the way in which the qualitative variable "type" is introduced in the models. While model D-GA treats the variable as qualitative information, model E-GA treats it as a number of dummies variables.
Model AW-GA is a three-layer GA-developed NN with 22 inputs, 23 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 800 epochs to avoid overfitting.
Model DW-GA is a three-layer GA-developed NN with 23 inputs, 30 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 900 epochs to avoid overfitting.

Table A.2: Summary of the models proposed in order to make forecasts for 2001

Model's Name	Models Characteristics
Model A2001 *	Basic NN with time defined as dummies for the year 2001
Model B2001 *	Basic NN with time defined as qualitative for the year 2001
Model D2001 *	NN time as dummies + type as qualitative for the year 2001
Model AW2001 *	Basic NN with time defined as dummies + wage for the year 2001
Model DW2001 *	NN time as dummies + wage + type as qualitative for the year 2001
Model A-GA2001 +	GA-developed basic NN with time defined as dummies for the year 2001
Model B-GA2001 +	GA-developed basic NN with time defined as qualitative for the year 2001
Model D-GA2001 +	GA-developed NN time as dummies + type as qualitative for the year 2001
Model AW-GA2001 +	GA-developed basic NN with time defined as dummies + wage for the year 2001
Model DW-GA2001 *	GA-developed NN time as dummies + wage + type as qualitative for the year 2001

Note: models with (*) have been developed by Longhi *et al*. (2002); models with (+) have been developed by the authors of the present paper.

Model A2001 is a three-layer NN with 20 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 400 epochs to avoid overfitting.
Model B2001 is a three-layer NN with 10 inputs, 5 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 350 epochs to avoid overfitting.
Model D2001 is a three-layer NN with 21 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 550 epochs to avoid overfitting.
Model AW2001 is a three-layer NN with 21 inputs, 9 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 850 epochs to avoid overfitting.
Model DW2001 is a three-layer NN with 22 inputs, 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 450 epochs to avoid overfitting.
Model A-GA2001 is a three-layer GA-developed NN with 20 inputs, 21 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 1700 epochs to avoid overfitting.
Model B-GA2001 is a two-layer GA-developed NN with 10 inputs and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 1000 epochs to avoid overfitting.
Model D-GA2001 is a three-layer GA-developed NN with 21 inputs, 9 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 3100 epochs to avoid overfitting.
Model AW-GA2001 is a four-layer GA-developed NN with 21 inputs, 29 and 10 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 200 epochs to avoid overfitting.
Model DW-GA2001 is a three-layer GA-developed NN with 22 inputs, 22 hidden neurons and 1 output. The activation function is a sigmoid, and the learning process was forced to stop after 1700 epochs to avoid overfitting.