NEW STATISTICAL TECHNOLOGIES APPLIED TO THE ESTIMATION OF THE FREE HOUSING PRICES: ARTIFICIAL NEURAL NETWORKS.

José María Montero Lorenzo and Noelia García Rubio

José María Montero Lorenzo
Facultad de Ciencias Jurídicas y Sociales de Toledo
Universidad de Castilla-La Mancha
E-mail: jlorenzo@correo.jur-to.uclm.es

Noelia García Rubio
Facultad de Ciencias Jurídicas y Sociales de Toledo
Universidad de Castilla-La Mancha
E-mail: ngarcia@jur-to.uclm.es

ABSTRACT.

The aim of this research is the use of the artificial neural networks models, specifically Multilayer Perceptrons trained by the algorithm known as Backpropagation to estimate the free housing prices. This methodology allows, through the training of the backpropagated nets, to estimate the houses prices on the basis of some variables, related to the houses, which are considered relevant (location, age, surface, quality, ...), overcoming the linear restrictions characteristic of the traditional statistical models used for this objective. The authors purpose is to show the forecasting power of these techniques as well as to suggest a change in the design of databases which have to do with the real estate market in such a way that the generalized use of this procedure is possible.

INTRODUCTION.

Real Estate price and, in particular, housing price is, by any reckoning, a very important parameter not only for city dweller but for the economic agents involved in housing markets and economic authorities which have powers on the subject of housing policy. That's why unchallengeable efforts have been made in order to estimate them even though, in most cases, with doubtful results.

Like in all estimation processes, the success in the estimation of housing prices depends on the amount of information, the quality of this information and the statistical techique which has been applied. In Spain, we don't have difficulties related to the amount of information. The Fomento Ministry has a database about prices of valued free housing which, in terms of quantity, could be assessed as excellent. However, there are problems concerning to the quality of this information. We don't hesitate about the truthfulness of the prices in the database, but it only provides aditional useful information related to the age and surface. Moreover, the database includes the houses location but refered to the postal code. So this information is not useful to make inferences (specially when we deal with kriged estimations).

The aim of this paper is to show a statistical technique which we consider very useful in order to estimate housing prices. This tool could be incorporated to the catalogue of estimation methods which are used nowadays: artificial neural networks. An artificial neural network is an estimation device which uses learning algorithms. The goal is to characterize a house by its price and a number of measures about it.Once it has been done, the network should be capable of associate a new house with some pattern that has been recognized during the learning process.

Artificial neural networks, which have had a great success in many estimation and classification tasks, constitutes one of the most powerful tools for housing price estimation. We think they could help the expert valuation agencies to save a great amount of money. However, at this moment, the results are not completely satisfactory for Spain. As we have mentioned before, an artificial neural network needs a large amount of information related to a high number of features about the houses. Unfortunately, the only available information in Spain is related to the age and surface, but it would be desirable know its exact geographic location, if it possesses garage or not, its floor, orientation, quality and so on.

Therefore, we are going to show the basic fundamentals of artificial neural networks focusing on feedforward networks trained through error backpropagation. Following, we are going to present the results of an empirical analysis from a database with 475 houses in Albacete city which has been collected by us and contains information regarding to the price by square meter, location, surface, age, quality and parking facilities.

ARTIFICIAL NEURAL NETWORKS.

Artificial Neural Networks provide a very powerful technique for pattern recognition tasks. One of these tasks is the regression problem in which we want to predict the values of one or more than one variables called outputs in function of a set of independent variables called inputs. The main advantage of this kind of models is their flexibility which avoids represent non-linear mappings from several inputs to several outputs. This property is specially useful when we don't know the form of the real relation between both set of variables.

The most common architecture is formed by the feed forward network named Multilayer Perceptron. This model consists of an input layer, whose nodes collect the interesting information from the outside world, also an output layer with units that supply the output signal of the network, and a number of hidden layers between the input and the output layers. The units in these layers play a very important role since they allow the network to solve non-linear problems. In this kind of networks the information signal is sent from the units in the input layer, through the hidden layers, right to the point where it reaches the output units. The information is always sent forward without cycles of links. In order to send out the signal, each neuron is connected to the units in the next layer through synaptic junctions or synapses, whose weights show the strength of the connections between two units.

The output given by the network is

$$\hat{y}_k = g_k\left(\sum_j \omega_{jk}\, z_j\right)$$

where g_k is the activation function for the output units, ω_{jk} denotes the weight between the hidden node j and the node k in the output layer and z_j represents the activation value of the node j in the hidden layer, given by the expression

$$z_j = f_j\left(\sum_i \omega_{ij} x_i\right)$$

where f_j is the activation function for the hidden units[1] ω_{ij} denotes the weight between the input unit i and the hidden node j and x_i represents the input i.

One of the most important features of these artificial systems is its learning ability. The learning process is shown through changes in the strength of connections. The goal is to find suitable values for the synaptic weights in order to optimize some of the performance criteria such as an error function. The most widely criterion used as an error function, induced by its analytical simplicity, is the sum of squares errors

$$E(\boldsymbol{\omega}) = \frac{1}{2} \sum_{n} \sum_{k=1}^{m} (y_k^{\,n} - \hat{y}_k^{\,n})^2$$

where y_{ki} is the target output value for pattern i and node k in the output layer and $\hat{y}_{ik}$ is the output given by the network.

In this case, the learning process operates in such a way that the weights are modified with the aim to minimize the sum of the squared differences between the target and the output values given by the network on each output unit. This process is known as supervised learning and the best known mechanism for weight adaptation is the backpropagation algorithm[2], a gradient descent algorithm whose task relies on the calculation of the derivatives of the error function with respect to the adaptive network parameters. The backpropagation algorithm for SSE function and sigmoid output units gives rise to the following rule for the weights from hidden to output units

$$\Delta\omega_{jk} = -\eta \frac{\partial E(\boldsymbol{\omega})}{\partial \omega_{jk}} = -\eta \sum_{n} \delta_k^{\,n} z_j^{\,n}$$

where, in this case,

[1] The activation function for units in the hidden layer should be non-linear in order to guarantee the network can solve non-linear problems.

[2] This algorithm has been attributed to Rumelhart because this author popularized it in 1986, although other researchers such as Werbos, in his doctoral thesis (1974) or Parker (1985) had been developed it earlier.

$$\delta_k^n = g'\left(\sum_j \omega_{jk} z_j\right)\left(\hat{y}_k^{\,n} - y_k^{\,n}\right) = \hat{y}_k^{\,n}\left(1-\hat{y}_k^{\,n}\right)\left(\hat{y}_k^{\,n} - y_k^{\,n}\right)$$

and η denotes the learning rate[3].

The update rule for the weights from input to hidden layers is

$$\Delta\omega_{ij} = -\eta \frac{\partial E(\boldsymbol{\omega})}{\partial \omega_{ij}} = -\eta \sum_n \delta_j^n \, x_i^{\,n}$$

where

$$\delta_j^n = z_j\left(1-z_j\right)\sum_k \delta_k^n \, \omega_{jk}$$

and

$$\delta_k^n = \hat{y}_k^{\,n}\left(1-\hat{y}_k^{\,n}\right)\left(\hat{y}_k^{\,n} - y_k^{\,n}\right)$$

The process convergence may be accelerated by certain changes of the basic algorithm. One way may be the addition of a momentum term. It causes the backpropagation algorithm to "pick up speed" if a number of consecutive steps change the weights in the same direction. The new update rule for becomes

$$\Delta\omega_{ij}(t+1) = -\eta \frac{\partial E(\omega)}{\partial \omega_{ij}} + \alpha \, \Delta\omega_{ij}(t)$$

where α is a constant in the range (0,1).

When a neural network is trained, the goal is not so much to get the network memorizes the data but optimize its generalization performance. In this sense, some stopping criteria must be set. The most common one lies in divide the whole sample into two subsets (training and verification sets) and stop the learning process once the error mesasured in the verification set starts to increase. Sometimes a third subset is created with the purpose of assessing the real performance capabilities of the trained network.

[3] A larger learning rate may lead to faster training but it could give rise to oscillations making impossible to reach the convergence. Usually, a time dependent parameter is set from a higher value when the process starts a lower one as the number of epochs increases.

ESTIMATION OF THE FREE HOUSING PRICE IN THE CITY OF ALBACETE USING ARTIFICIAL NEURAL NETWORKS.

The application starts with the description of the sampling. Available information was obtained by a sampling procedure from data supplied by real state agencies due to the lack of official information related to variables such as quality, parking facilities or the location of the houses. The official surveys only provides information about age, surface and the postal district. The sample contains very rich information about the housing market in Albacete in 1997. It covers a wide range of houses which includes 475 records concerning the following explanatory variables:

- *Location* of each house in Albacete city plan.
- *Age*, expressed in years.
- *Surface*, measured in useful squared meters.
- *Parking facilities*, a categorical variable that indicates whether a house possesses it or not .
- *Quality*, a nominal variable with five categories:
    - · *Bad*, for old houses built with bad material and in very bad habitability and conservation conditions.
    - · *Substandard*, for new or semi-new houses that were built below the standard of quality.
    - · *Good*, an intermediate level corresponding to the standard quality.
    - · *Very good*, comparable to the new houses of first quality.
    - · And finally, *Luxury*.

The output variable is the price by square meter.

The architecture of the multilayer perceptron proposed has been selected after a large number of trials. The number of nodes in the input and the output layer is determined by the structure of our analysis, i.e., number of explanatory and output variables respectively. On the other hand, the number of hidden layers and the elements in each one of them have been chosen taking as criterion the construction of a network with the less complexity as possible. This aim

has lead us to select a 6:10-5-1:1 network, i.e., an input layer with six nodes[4], that have been preprocessed in ten nodes[5], one hidden layer with five elements and finally, an output layer with one node.

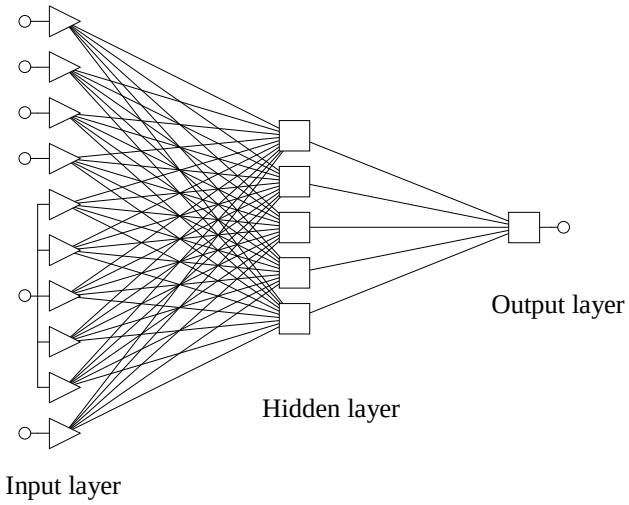

Figure 1. Network architecture.

The learning parameters have been set as follows:

Maximum number of epochs: 5000.

Adaptive learning rate: 0.05 to 0.001.

Momentum term: 0.7

The evolution of the sum of squares error for the training and verification subsets can be seen in figure 2.

[4] All explanatory variables were selected after applying a sensitivity analysis on the inputs to the neural network.This proccess allows to prune out input variables with low sensitivity.

[5] All variables have been preprocessed before their introduction into the network. The numerical variables have been scaled to produce new variables in the range 0-1. The variable *parking facilities* has been encoded by the two-state technique in a single input variable. Finally, the variable *quality*has been converted using the one-of-N method. This technique uses a set of variables, one for each possible nominal value. In our case, there are five categories of quality, so the total number of variables changes from 6 to 10.

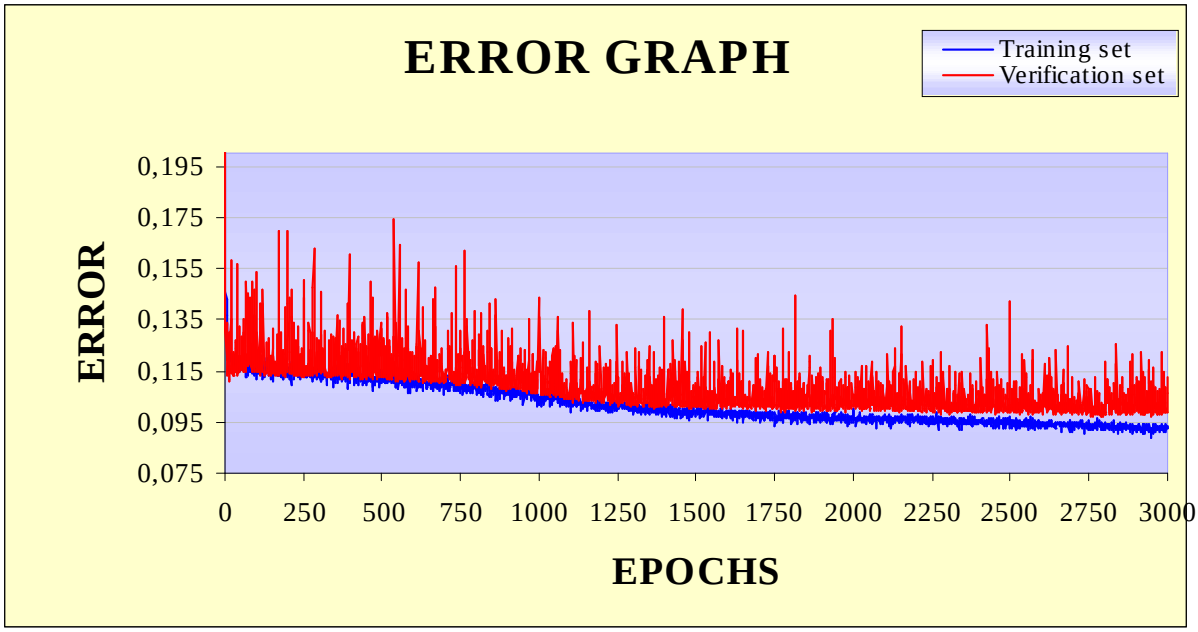

Figura 2. Evolution of the sum of squares error.

The training process has been stopped after 2758 epochs since this is the number of iterations from which the verification set error starts to increase.

We can analyse the degree of predictive accuracy through the regression statistics shown in the following table:

	Training	Verification	Test
Data Mean	125668,9	119434,3	119405,9
Data S.D.	38327,6	41189,3	37631,4
Error Mean	3705,1	1289,3	3704,2
Error S.D.	22818,3	25068,5	22509,3
Abs. E. Mean	18188,4	18565,4	18250,7
S.D. Ratio	59,5	60,9	59,8
Correlation	80,3	79,9	80,1

The most significant value is the prediction error standard deviation (Error S.D.). If this is no better than the training data standard deviation, then the network has performed no better

than a simple mean estimator. We can analyse the explained variance of the model through the ratio of the prediction error SD to the training data SD. A value significantly below 1.0 indicates good regression performance.

Another way to assess the performance of the model is the standard Pearson-R correlation coefficients between the actual and predicted outputs. Although a correlation coefficient of 1.0 does not imply a perfect fit or prediction, it is a good indicator of performance in practice.

CONCLUSION

In sight of the preceding regression statistics we could conclude that the network results are not optimum. The reason is that, as we have mentioned before, the statistical information is not enough and we should include variables like the geographical orientation, the number of the floor and so on.

However, the main objective was to show an unknown technique for the real estate market and to emphasize its power in this area if we have enough statistical information. If you consider this procedure is interesting, we should make an effort in order to get the involved authorities include the largest number of housing features. Without doubt, it would improve the results enormously.

BIBLIOGRAPHY.

BISHOP, C. M. (1995): *"Neural Networks for Pattern Recognition"*. Oxford New York.

GOLDEN, R.M. (1996): *"Mathematical Methods for Neural Network Analysis and Design"*. MIT Press.

HASSOUN, M. (1995): *"Fundamentals of Artificial Neural Networks"*. MIT Press.

HERTZ, KROGH, & PALMER (1991): *"Introduction to the theory of Neural Computation"*. Santa Fe Institute studiesin the sciences of complexity. Lecture notes: v.I

HINTON, G. E. (1989): *"Connectionist learning procedures"*. Artificial Intelligence 40, 185-234.

KAY, J.W. & TITTERINGTON, D.M. (1999): *"Statistics and Neural Network: Advances at the Interface"*. Oxford University Press.

LIPPMAN, R. (1987): *"An Introduction to Computing with Neural Nets"*. IEEE ASSP Magazine, Vol. 3, nº 4, p. 4-22.

RIPLEY, B. D. (1996): *"Pattern Recognition and Neural Networks*". Cambridge University Press.

RUMELHART, D. E., HINTON, G. H. & WILLIAMS, R. J. (1986 A): *"Learning internal representations by error propagation". Parallel Distributed Processing: explorations in the Microstructures of Cognition,* Vol. 1, D.E. Rumelhart and J.L. McClelland (Editors.), Cambridge, MA: MIT Press, pp. 318-362.

RUMELHART, D. E., HINTON, G. H., WILLIAMS, R. J. (1986 B): *"Learning representations by back-propagating errors". Nature* 323: 533-536.

THEODORIDIS, S. & KOUTROUMBAS, K. (1999): *"Pattern Recognition"*. Academic Press, San Diego.

ZAPRANIS, A. & REFENES A.P. (1999): *"Principles of Neural Model Identification, Selection and Adequacy"*. Springer-Verlag. London.