ARTIFICIAL NEURAL NETWORKS AS A TOOL FOR SITE SELECTION WITHIN GIS

T.A. Yanar [a] , Z. Akyürek [a, *]

[a] METU, Geodetic and Geographic Information Technologies, Natural and Applied Sciences, 06531 Ankara, Turkey
e122330@metu.edu.tr
zakyurek@metu.edu.tr

Commission II, WG II-7

KEY WORDS: Artificial Neural Networks, Decision-Making, Site Selection, Fuzzy Logic, FuzzyCell, GIS

ABSTRACT:

To obtain more flexibility and more effective capability of handling and processing imprecise information about the real world, fuzzy set theory is introduced into GIS. FuzzyCell is a system designed and implemented to enhance conventional GIS software (ArcMap®) with fuzzy set theory. Extending GIS with fuzzy logic (a linguistic approach as the model of human thinking) not only offers a way to represent and handle uncertainty present in the continuous real world but also assist GIS user to make decisions using experts' experiences in decision-making process. The cost of finding solutions to decision-making problems by models which enable decision-makers to express their constraints and imprecise concepts that are used with geographic data (i.e., fuzzy logic) for large volume is high. For such cases, artificial neural networks (ANNs), which can solve complex problems and can "learn" from prior applications, can be used. In this study, a fuzzy rule based system (FuzzyCell) was used to model site selection problem by capturing rules from human experts. The ANNs were trained by obtained fuzzy measures against input data to recognize patterns for reproduction of relevant sites for new locations and were tested whether ANNs can produce reasonable guesses for locations other than training sites when compared to results obtained from FuzzyCell. Two metrics, Kolmogorov-Simirnov test metric and root mean squared error values, were used for testing. It was found in this study that ANNs provide reasonable guesses for locations other than training sites.

1. INTRODUCTION

One of the main tasks in Geographic Information Systems (GIS) is making decisions using information from different layers. In conventional decision-making process, a common type of operation is the threshold model. In the threshold model, low and high threshold values limit the exact boundaries of criteria. For each criterion the study area is classified into two subregions describing whether a property value of a specific location is in the defined limits or not. Then maps produced for each criterion are overlaid using logical connectives (i.e., Boolean overlay). Such models can cause problems since they are inherently rigid. On the other hand, FuzzyCell (Yanar and Akyürek, 2006) can be used to make decisions capturing uncertain information using fuzzy set methodologies.

FuzzyCell is generic and enables decision-makers to express their constraints and imprecise concepts that are used with geographic data through the use of natural language interfaces (Yanar and Akyürek, 2006).

Artificial neural networks (ANNs) provide a mechanism for learning from data and mapping of data. They can solve complex problems and can "learn" from prior applications. ANNs can be trained to provide an alternative approach for problems that are difficult to solve such as decision-making problems in GIS.

In this study, FuzzyCell is used to model site selection problems by capturing rules from human experts using its natural language interfaces. Then, FuzzyCell converts these rules to fuzzy IF-THEN rules. Implication of these fuzzy rules to input data produces suitable sites for searching criteria. If ANNs are trained by obtained data defining suitable sites against input data, then ANNs can recognize patterns for reproduction of relevant sites for new locations faster.

The objective of this study is to test whether ANNs trained by data obtained from fuzzy models can produce reasonable guesses for locations other than training sites when compared to results obtained from fuzzy rule based system (FuzzyCell).

2. MATERIALS AND METHODS

2.1 Fuzzy Set Theory

Fuzzy logic (Zadeh, 1965) generalizes crisp logic to allow truth-values to take partial degrees. Since bivalent membership functions of crisp logic are replaced by fuzzy membership functions, the degree of truth-values in fuzzy logic becomes a matter of degree, which is a number between 0 and 1.

An important advantage of using fuzzy models is that, they are capable of incorporating knowledge from human experts naturally and conveniently, while traditional models fail to do so (Yen and Langari, 1999). Other important properties of fuzzy models are their ability to handle nonlinearity and interpretability feature of the models (Yen and Langari, 1999).

[*] Corresponding author. Zuhal Akyürek

2.2 Artificial Neural Networks

Originally motivated by the biological structures in the brains of human and animals, which are powerful for such tasks as information processing, learning and adaptation (Nelles, 2000). The most important characteristics of neural networks include (Nelles, 2000):

- Large number of simple units
- Highly parallel units
- Strongly connected units
- Robustness against the failure of single units
- Learning from data

Neural networks provide a mechanism for learning from data and mapping of data. However, models built using neural networks are difficult to interpret; they are essentially a "black box" model (Yen and Langari, 1999).

2.3 Comparison Methods and Tests

Comparison between fuzzy rule based method and ANNs needs comparison metrics. These metrics should compare predicted data set with target data set and give higher results when predicted data and target data are more similar. In this work, non-parametric tests for ratio scale data, Kolmogorov-Simirnov and root mean squared error are used for comparison metrics.

Non-parametric tests do not use predictable distribution of sample means to infer whether samples are from same population and used if the data are grossly non-normal (McKillup, 2006) as data sets in this study.

The two-sample Kolmogorov-Simirnov test compares the distribution of values in the two sample data vectors representing random samples from some underlying distributions (McKillup, 2006). The objective is to determine if these two data sets have been drawn from identical population distribution functions. Therefore, null hypothesis for this test is whether the two sample data vectors are drawn from same continuous distribution. The alternative hypothesis is that they are drawn from different continuous distributions.

Root mean squared error is computed by taking square root of average of the squared differences between each computed value and its corresponding correct value. Root mean squared error formula is given below:

$$rmse = \sqrt{\left(\frac{(x_1 - \hat{x}_1)^2 + \cdots + (x_n - \hat{x}_n)^2}{n} \right)} \qquad (1)$$

2.4 Data Sets

Birecik data set is taken from study Yanar (2003). Bartın data set is taken from study Usul et al. (2002). For both regions DEMs are used to construct slope maps. Constructed slope map, closeness to roads and town maps of these regions are used in this work.

3. APPLICATION

3.1 Data

Slope, closeness to roads and closeness to town maps are used for input. Output data defines suitability values for industrial development sites which are constructed from fuzzy rule given below by using FuzzyCell.

IF Slope is flat or
Slope is gentle and
Distance to road is close and (2)
Distance to town is close
THEN site is suitable

where membership functions for linguistic terms are depicted in Figures 1-4.

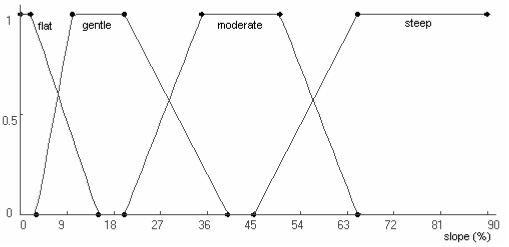

Figure 1. Membership functions flat and gentle slope

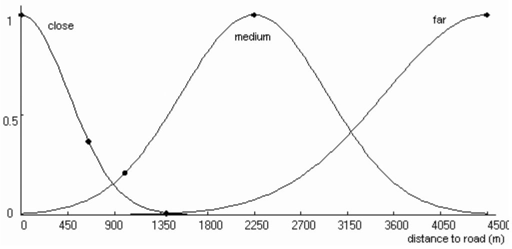

Figure 2. Membership functions for close to roads

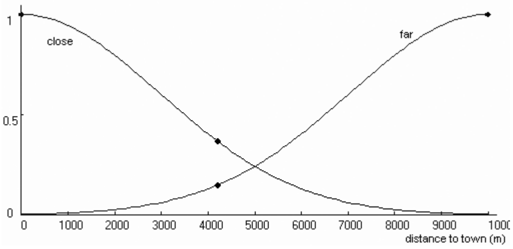

Figure 3. Membership functions for close to town

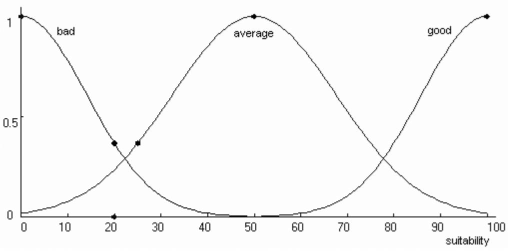

Figure 4. Membership functions for suitability

Input and target values of Birecik and Bartın zones are mapped into the interval [+1,-1] to simplify the problem of network. It also ensures that target values fall into the range that new feed forward network can reproduce. Second, the data are divided into training, validation and test sets. The training set is used for training the network. Validation set is used to validate network performance during training so that training stops early if it attempts to over fit the training data. Test set is used for an independent measure of how the network might be expected to perform on data it was not trained on. 20% data is selected for validation set and 20% for the test set, remaining 60% data is selected for training set. Individual elements are selected randomly.

3.2 Estimation

Although there are plenty of network types that can be used for geographic phenomena, feed forward back-propagation networks are used in this work. Standard back-propagation is a gradient descent algorithm in which the network weights are moved along the negative of the gradient of the performance function (Demuth et al, 2006).

A multilayer network has multiple layers of neurons. Each layer plays different roles. A layer that produces the network output is called an output layer. All other layers are called hidden layers. Since there is only one target value associated with input vectors, all networks used in this study should have one output neuron. The number of hidden layers and the number of neurons used in each layer are changed to create different network structures. The aim is to find a network structure which gives the most similar values with target (output) layer. Unfortunately, there is no defined rule to build appropriate network structure. Therefore, this step requires testing many different structures and there is no guarantee that the selected network which gives the most similar values is the best network structure among all alternatives. This is the main disadvantage of using neural networks.

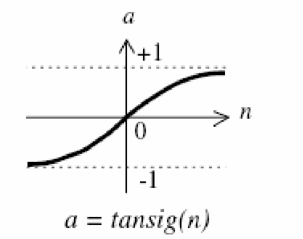

Figure 5. Tan-sigmoid transfer function

For all networks tan-sigmoid transfer function which is shown in Figure 5 is used. Once the network weights and biases are initialized the network training begins. All networks use Levenberg-Marquardt back-propagation algorithm for training and gradient descent with momentum weight and bias learning function. Mean squared error is used for network performance function which measures the network's performance according to the mean squared errors.

4. RESULTS

Prediction performances of networks trained and tested with data from Birecik and Bartın zones are shown in Tables 6 and 7. Two-sample Kolmogorov-Simirnov test metric, KSTAT, gets lower as the probability that compared sets are come from the same population increases. Therefore, lower the KSTAT values the network has better prediction performance. Similarly, lower the root mean squared error values the higher the prediction performances of networks. According to two-sample Kolmogorov-Simirnov test metric, and root mean squared error values two hidden layer one output layer network with two nodes in both hidden layers (Net221) trained and tested with Birecik zone data is the best predictor.

Model	Kolmogorov-Smirnov Test	
	KSTAT	RMSE
Net11	0.1507	2.6199
Net21	0.0846	1.0972
Net31	0.0613	0.8007
Net41	0.0614	0.9596
Net51	0.0598	0.9960
Net111	0.1507	2.6183
Net121	0.1260	2.5564
Net221	0.0346	0.5489

Table 6. Performance tests for Birecik networks

For Bartın zone data best predictor is one hidden layer one output layer network with five nodes in hidden layer (Net51).

Model	Kolmogorov-Smirnov Test	
	KSTAT	RMSE
Net11	0.3014	11.9558
Net21	0.0756	9.9959
Net31	0.0496	2.5141
Net41	0.0482	1.8934
Net51	0.0444	1.8384
Net111	0.3645	11.8817
Net121	0.3645	11.8817
Net221	0.1671	6.6730

Table 7. Performance tests for Batın networks

All metrics indicate that networks trained and tested with Birecik zone data give better prediction performance values.

5. CONCLUSION

The cost of finding solutions to decision-making problems by models which enable decision-makers to express their constraints and imprecise concepts that are used with geographic data (i.e., fuzzy logic) for large volume is high. For such cases, a small part of data, which represents the whole data much, is selected. FuzzyCell is used to construct fuzzy rules by capturing rules from human experts and to find solution to decision problem for chosen region. The ANNs were trained by obtained fuzzy measures against input data to recognize patterns for reproduction of relevant sites for whole, large volume data fast and in an effective manner. Results obtained from ANNs are compared by results obtained from fuzzy rule based system. According to Kolmogorov-Smirnov test metric and root mean squared error ANNs successfully recognize patterns of relevant sites.

Original output and outputs produced by best networks for Birecik (Net221) and Bartın (Net51) zones are given in Figures 8 - 11. Histogram graphs of original suitability values and values produced by best networks for Birecik and Bartın regions are depicted in Figures 12 – 13.

Furthermore, one disadvantage of using neural networks is that there is no defined rule to build appropriate network structure. Therefore, building appropriate network structure requires testing many different structures.

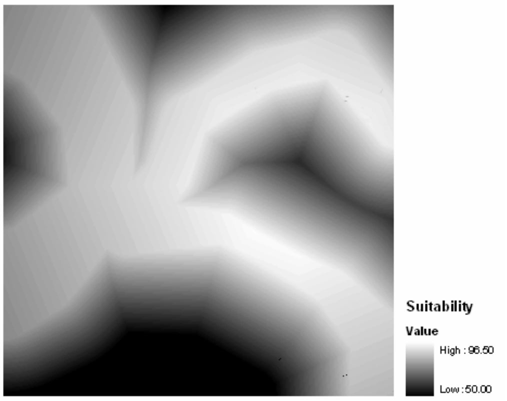

Figure 8. Birecik suitability (original output)

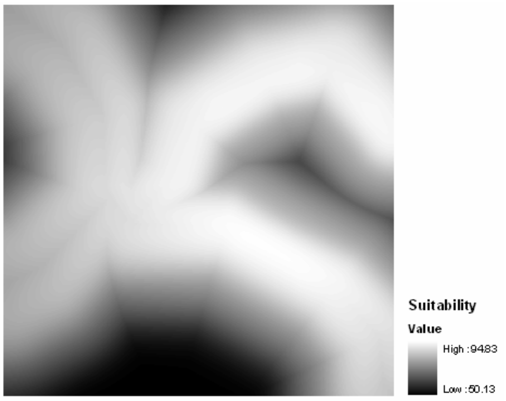

Figure 9. Output data produced by ANN Net221 for Birecik region

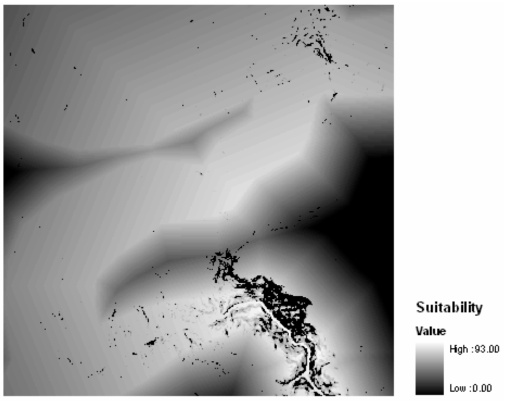

Figure 10. Bartın suitability (original output)

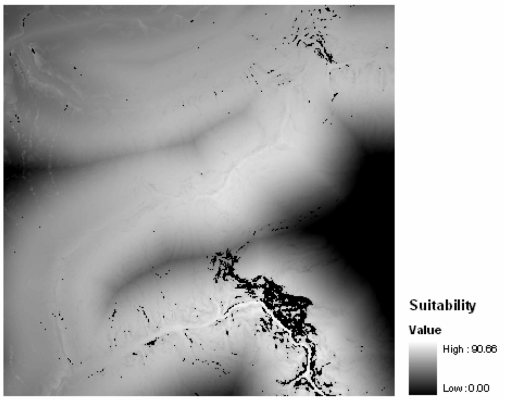

Figure 11. Output data produced by ANN Net51 for Batın region

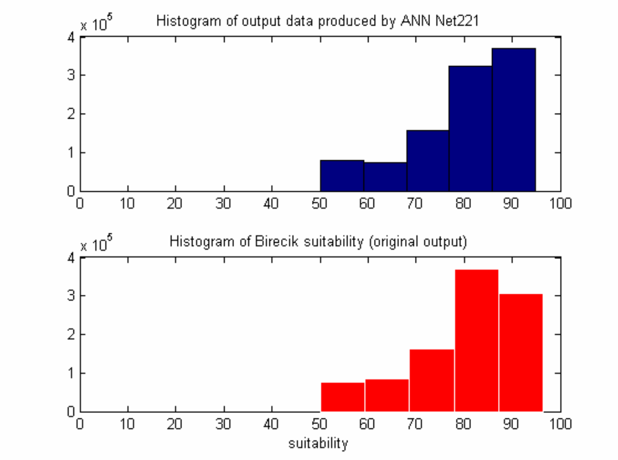

Figure 12. Histograms for Birecik suitability

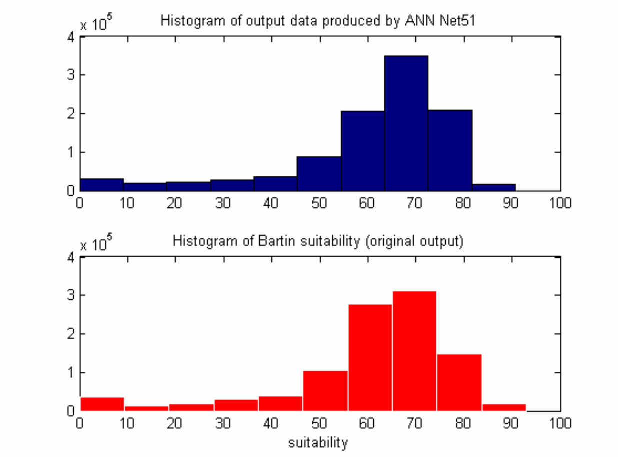

Figure 13. Histograms for Bartın suitability

References

References from Journals:

Yanar, T. A. and Akyürek, Z., 2006. The enhancement of the cell-based GIS analyses with fuzzy processing capabilities. *Information Sciences*, 176, pp. 1067-1085.

Zadeh, L. A., 1965. Fuzzy Sets, *Information and Control*, 8, pp. 338-353.

References from Books:

Demuth, H., Beale, M., and Hagan, M., 2006. *Neural Network Toolbox User's Guide Version 5*. Mathworks Inc.

McKillup, S., 2006. *Statistics Explained An Introductory Guide for Life Scientists*. Cambridge University Press.

Nelles, O., 2000. *Nonlinear System Identification*. Springer Verlag.

Yen, J. and Langari, R., 1999. *Fuzzy Logic: Intelligence, Control, and Information.* Prentice Hall.

References from Other Literature

Usul N., Akyurek Z., Sorman A. U., 2000. Project on Flood Analysis using the integration of Hydraulic modeling and GIS. AGUDOS 00070220001.

Yanar, T. A., 2003. The enhancement of the cell-based gis analyses with fuzzy processing capabilities. Master's thesis, Middle East Technical University.

Acknowledgements

The authors would like to thank Adnan Yazıcı and Şebnem Düzgün from Middle East Technical University for their constructive comments that helped to improve the ongoing work.