

# COMPARING NEURAL NETWORKS AND MULTIVARIATE DISCRIMINANT ANALYSIS IN THE SELECTION OF NEW CROP VARIETIES

**Jose L. García de Ceca and Javier Moro**

*Area de Selvicultura y Mejora Forestal, INIA-CIFOR,  
Carretera de La Coruña km 7, 28040 Madrid (Spain)  
ceca@inia.es, jmoro@inia.es*

**Abstract:** To explore new ways to help the process of selection of new crop varieties, neural networks (NN) and multivariate discriminant analysis (MDA) have been applied with the hope of making the variety selection easier and faster. It can be said that, for our case study, whilst Quadratic MDA has reported results that may be considered of interest, NN has worked better than MDA, mimicking quite acceptably the decision process of the human committee, so that NN could be considered as a support tool in the process of selection of new varieties.

**Keywords:** neural networks, multivariate discriminant analysis, crop selection

## **1 Introduction.**

The aim of this study is to explore new ways to make the selection of new crop varieties, or at least to find procedures that would help the process. The Spanish regulations state that the production and sale of seeds is restricted to varieties included in the National Commercial List. The selection of new varieties involves the making of some trials based on a statistical design and the evaluation by a human committee of the results of a multidimensional analysis of variance. While in most of the cases the choosing is clear, in others it is not easy whether a specific variety should or not be included in the commercial list. As an alternative to multidimensional analysis of variance, there are two techniques that may be applied to the process of selection of new crop varieties with the hope of making the variety selection easier and faster, even automatically. These are neural networks (NN) and multivariate discriminant analysis (MDA).

To carry out this study the winter barley crop has been chosen. 30 varieties tested from 1983 to 1988 have been used for the study. From these, an initial data set was used to train the NN and to make the classification in the MDA. This initial set had 3 varieties that were used as test, 3 that were accepted for the commercial list and 4 that were rejected. The variables used were: Year of trial, # of trial, efficiency, % moisture, ear date, plant height, damages by cold, beaten down, burning, oidium, rincosporiosis, brown mildew, a thousand seeds weight, hectoliter weight, % protein, sieve < 2.2 mm, and efficiency at 10% moisture. There are three classes to which we want to classify all the varieties: Accepted to be included in the commercial list (better than test varieties), Rejected (worse than test varieties), and Test (which means as good as test varieties but not better).

## **2 Neural networks approach.**

Traditional programming techniques require that someone create an algorithm. While for some problems this is straightforward, for many real world problems it is very difficult to develop such an algorithm. Neural networks (NN), in contrast to being programmed, are trained. This means that examples are presented to the network and the network adjusts itself by some learning rule usually based on how correct the response is to the desired response. NN are part of the field of artificial intelligence. As a difference to expert systems, when developing NN solutions to a problem, neither the knowledge nor the explicit rules for processing the knowledge are coded by the programmer. Instead, the NN learns the rules for processing the knowledge. This is done by adjusting the weight values in a highly connected network based on the example data. One can think of a NN as a very general model that is parametrized by the adjustable weights.

In statistics, one must often make many assumptions about the data, and must sometimes limit the analysis to a certain number of possible interactions. By contrast from a practical point of view, NN are basically non-parametric although in theory one can think of a NN as being parametrized by its weights. In addition, more terms can be examined for interaction by a NN since the network will place, hopefully, its emphasis on those inputs that help to predict the output. By allowing more data to be analyzed at the same time, more complex and subtle input interactions are possible. While a knowledge of statistics is excellent preparation to appreciate the capabilities of NN, NN modeling can be carried out by non-experts in statistics.

NN are built of neurons arranged in layers and connected to other neurons in other layers. Sometimes neurons are connected to other neurons in their own layer or even to themselves. Each neuron processes the input it receives via these connections and provides a continuous analog value to other neurons via its outgoing connections. As in biological systems, the strengths of these connections can change and, in fact, do change in response to the strengths of the inputs and the type of transfer function used by the neuron. Deciding how the neurons are connected in a network, how the neurons process their information and how the connection strengths are modified all go into creating a NN. Once the network is designed, it has to be fed with a set of training data containing both input and output data. The network then learns through an iterative process and shows its performance mainly by the convergence to near zero of its RMS. The network has some parameters to play with (e.g., learning coefficients, transfer function, etc.), and if the network gets to a state of good convergence then it is said that the network is trained. Once we have a trained network, it is fed on different sets of input data and it produces the output for these sets, either a prediction or a classification.

In this case, the network implemented has been a Learning Vector Quantization (LVQ), which is a classification network that assigns vectors to one of several classes. The LVQ network contains an input layer, an intermediate layer (known as a Kohonen layer) which learns and performs the classification, and an output layer. The input layer contains one neuron for each input parameter which are 19 in this case because it includes all the variables plus the information from the design of the multidimensional analysis of variance: replicate #, and block #. The output layer contains one neuron for each class (Accepted, Rejected and Test. It has been found that for the best performance the Kohonen layer needs 61 neurons per class, 183 in total.

## **2 Multivariate discriminant analysis approach.**

Among multivariate statistical analysis techniques, the discriminant analysis is concerned with separating distinct sets of objects (or observations) and with allocating new objects to previously defined groups. MDA tries to find the combination of variables that best predicts the category or group to which a case belongs. The group identification must be known for each case used in the analysis. The combination of predictor variables is called a classification function, and this function can then be used to classify new cases whose group membership is unknown. MDA presents two weak points compared

to the NN approach: It only considers complete cases (no missing nor out of range data) and the presence of outliers can severely affect the analysis. There were considered two ways of making a MDA analysis: Linear and quadratic discriminant analysis. The reason for this diversity is that discriminant methods are very much affected by the nature of the within group covariance matrices. On linear discrimination technology, the covariance matrices are assumed equal for all groups. On quadratic discrimination technology, the assumption of homogeneity of within groups is patently transgressed, although it retains the distributional assumption of multivariate normality.

Prior to carry on the analysis, some tests were made to be sure that applying MDA techniques was correct. Following the same classification that in the NN case and using the same initial data set to define the groups (Accepted, Test, Rejected), it was necessary to test if there was an effective difference among groups. Thus, a Wilks test of equality of group means was done giving a P-Value for F-Statistic of 0.0, so that the test rejected the hypothesis of equality of group means. Next verification was to decide which model of MDA was more adequate to apply. Since the key point was the homogeneity of within groups covariance, a test of homogeneity of covariance matrices was made showing a P-Value result of 0.0 again. Hence, the test rejected the null hypothesis of covariance homogeneity within groups, which means that there is no common covariance structure within groups. Quadratic classifications functions will provide better classification rates than the linear techniques. But the experience tells that it may be of interest using the linear analysis in the exploratory stage of data analysis, so that a linear stepwise analysis was used at the beginning. For the analysis, it is needed a grouping variable named "type" which gets the value 1 for Test, 2 for Accepted and 3 for Rejected.

A jackknife validation procedure was made to reduce the bias of the evaluation, consisting this procedure of classifying each case into the group with the highest posterior probability according to classification functions computed from all the data except the case being classified. In the beginning, the linear case, it has to be emphasized that only seven variables of the total set entered the classification functions. These functions led to the following results of percent number of cases correctly classified into group in the classification of the initial set of data: Test 80.9 %, Accepted 85.4 %, Rejected 82.1 %. Trying to ameliorate the classification, all variables were forced to enter in the MDA. In doing this, the variable RTO (efficiency) was eliminated due to its high correlation to RTOC (efficiency at 10% moisture), being under a tolerance level of 0.0005. The percentages of correctness for the classification were then Test 80.9 %, Accepted 87.5 %, Rejected 85.1 %. While it may be argued whether these values may be considered acceptable for an effective classification, it is clear that there is not a patent difference between these two results. Therefore, the Quadratic Discriminant Analysis was applied. In the Quadratic MDA, the percentages of cases correctly classified into group were Test 87.3 %, Accepted 95.8 %, Rejected 97.0 %. With a misclassified rate of 9.84% and an error rate (counting prior probability) of 6.63 %, the results of the analysis now are reasonably good, so it may proceed to continue the classification with all the varieties of the set.

#### **4 Conclusions.**

For the NN, the results may be considered very promising. In no case, a variety accepted by the committee has been rejected by the network. In three cases (varieties 84116, 85163 and 85181), a variety rejected by the committee has been classified as a Test by the network. Since the varieties have to be better than the Tests to be accepted, we can consider this result as a correct one. This would imply that the committee should have some discussion about whether these three varieties should be included in list or not. In one case (variety 83201), a variety rejected by the committee has been accepted by the network. The exact cause of this rejection is not known since the proceedings of the meetings were not kept until 1989. Checking the records of that variety, it has been found a higher presence of diseases than in the tests, which may be the cause of the rejection by the committee. Also and although is not the case, sometimes there are problems in the botanical characteristics of a variety that make it non-homogeneous from an identification point of view, which may be the cause of its rejection despite a good agronomical result. In any case, it could be argued that the amount of information given to the NN was not enough to

achieve a 100% efficiency, or that maybe the net could have been tuned a little bit better by essaying a more exhausting sensitivity analysis. For the MDA, in the linear case, the results may be considered weak since there is an unquestionably discrepancy between human and MDA decisions. In five cases there is a complete disagreement. Among these five cases, in four of them the discrepancy is distinct: two varieties classified by the committee as Accepted (85174 and 87236) were classified as Rejected 100% by the analysis, and two classified by the committee as Rejected (83159 and 83185) were classified as Accepted 100% by the analysis. Also, in two cases (74003 and 74005) a variety rejected by the committee was classified as Test by the analysis, which would mean -as in the NN section- that we could consider these results as correct ones and that this would imply that the committee should have some discussion about whether these varieties should be included in list or not. Something similar can be said about variety 87186 since the percentages of T-A-R are almost at a 30% each, so that the decision about including it in list should be passed to the committee. In the Quadratic case, the results are very similar to the Linear case. In respect to the committee classification, there is a gain of only one variety: Instead of five wrong classified varieties, there are four of them, which coincide them all. Only the classification of variety 87236 was significantly different, although it was changed to category Test, which means that -again- the decision should remain in the committee. Moreover, the number of varieties included in the Test category went to seven, instead of three as in the Linear case. Anew, more work to the committee. Question then is which would be the advantage -if any- of MDA over the multidimensional analysis of variance. Most probably, one cause of the disagreements in the final results of the analysis is the already mentioned presence of missing and outlier data. A more careful collecting and managing data would help in the amelioration of the results, but due to the biological nature of the process of growing the varieties this can be very difficult to get in all the occasions - impossible, so.

As a summary, it can be said that it seems that Neural Networks could be considered as a promising technique to developed support tools for the process of selection of new varieties, while Multivariate Discriminant Analysis does not offer any substantial advantage over the traditional multidimensional analysis of variance.

Acknowledgements: This work has been developed under the auspices of the European Union ADDA (Agricultural Data Dictionaries and Analysis) project AIR3-CT94-1330 (PL921330).

Table 1. Summary of results.

Variety	Comm.	N. NET		LINEAR			Decision	QUADRATIC			Decision
		Decision	%Test	%Acc.	%Rej.	%Test		%Acc.	%Rej.		
74003	R	R	100,0	0,0	0,0	T	100,0	0,0	0,0	T	
74005	R	R	100,0	0,0	0,0	T	100,0	0,0	0,0	T	
79232	R	R	0,0	42,1	57,9	R	52,6	36,8	10,5	T	
82183	R	R	0,0	45,5	54,5	R	13,6	0,0	86,4	R	
82229	A	A	0,0	85,7	14,3	A	14,3	57,1	28,6	A	
83159	R	R	0,0	100,0	0,0	A	7,7	92,3	0,0	A	
83185	R	R	0,0	100,0	0,0	A	3,4	96,6	0,0	A	
83201	R	A	0,0	71,4	28,6	A	14,3	85,7	0,0	A	
84116	R	T	8,7	0,0	91,3	R	26,1	0,0	73,9	R	
85163	R	T	28,6	0,0	71,4	R	14,3	0,0	85,7	R	
85174	A	A	0,0	0,0	100,0	R	4,3	0,0	95,7	R	
85181	R	T	0,0	0,0	100,0	R	28,6	0,0	71,4	R	
86287	A	A	0,0	100,0	0,0	A	12,5	79,2	8,3	A	
86288	A	A	4,2	70,8	25,0	A	58,3	29,2	12,5	T	
87186	A	A	31,4	31,4	37,1	R?	48,6	14,3	37,1	T	
87190	A	A	0,0	80,0	20,0	A	31,4	68,6	0,0	A	
87216	A	A	0,0	91,4	8,6	A	14,3	85,7	0,0	A	
87220	A	A	0,0	100,0	0,0	A	8,6	91,4	0,0	A	
87236	A	A	5,7	0,0	94,3	R	100,0	0,0	0,0	T	
87247	A	A	2,9	65,7	31,4	A	57,1	42,9	0,0	T	