ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Joanna Kisielińska**

APPLICATION OF DISCRIMINANT ANALYSIS AND NEURAL NETWORKS TO FORECASTING THE FINANCIAL STANDING OF FARMS

Abstract

The aim of the research was to determinate a linear discriminant function and neural network that could be applied for financial situation forecasting in polish farms sector. The construction of discriminant models was based on set of financial indicators and the classification criterion was based on the private farm's income. The investigated population was divided into two equal groups with respect to the median value of income.

The data was gathered in the period of several years that allowed examine the influence of the time on the quality of discriminant models. Also the set of indicators with large forecasting ability was determined.

The data used for the discriminant models was sourced from private farms keeping farm accountancy under auspices the Institute of Agricultural and Food Economics in the years 1992–2002. The calculations was made with help of STATISTICA and data analysis with Excel using VISUAL BASIC FOR APPLICATION.

Key words: linear discriminant function, neural networks.

1. Introduction

The aim of the research presented in the paper was to evaluate a linear discriminant function and a neutral network that could be applied for financial situation forecasting in polish farms sector. The construction of discriminant models was based on a set of financial indicators supplemented by some additional information concerning farms and the classification criterion was based on the private farms' income. The analysis was made based on data gathered in the period of several years that allowed examining the influence of

* Dr., Chair of Econometrics and Computer Science, Warsaw Agricultural University.

the time on the quality of discriminant models. Namely, a possibility of application models obtained for a given period to another year was examined.

The additional aim was to distinguish the set of indicators with large forecasting ability, i.e., such indicators that influence the values of the discriminant function in the most significant way.

The data used for the discriminant function was sourced from private farms keeping farm accountancy under auspices of the Institute of Agricultural and Food Economics in the years 1992–2001. The values for the year 2002 were taken for model verification. The calculations were made with help of STATISTICA and Excel with VISUAL BASIC FOR APPLICATION.

2. The rules of discriminant models construction

The construction of a classification model based on the discriminant analysis requires that two fundamental elements are specified. The first one is a uniqely formulated rule of assignment that results from the needs of the specific classification aim. The second is chosing in a proper way a set of features that describe classified objects and are to be the bases of classification.

The discriminant models presented here are designed for forecasting purpose to the financial situation of hauseholds. The measures of the financial standing is farmer's income. One may think that is should be the profit to be the measure but in the case of farms it is difficult to estimate the profit as it requires that some symbolic costs (like wages for farmer's work, interest of own capital or feudal rent) are assumed.

The investigated population was divided into two groups. The first group, denoted by Class I consisting of households showing weak financila condition, the second group, Class II – household regarded to be good. The division into classes was based on the median value of income. Households with income smaller then the median were classified to the Class I, the rest to Class II.

The models construction was based on a vast range of financial indicators. This set of indicators was supplemented by additional information concerning households. The set of indicators chosen following the suggestions given in papers Kulawik (1995), *Rachunek ekonomiczny i analiza finansowa* (1994) and Wyszkowska (1996) includes 29 indicators, e.g., liquidity ratios, turnover ratios, farming efficiency, financial support and those that characterize fixed capital. Additionaly, the area of farms (in hectares of cropland), age of farmers, intensivity of production, level of intensivity of production organization (as in Kopeć) were taken into account. The level of production intensivity is understood as material and financial outlays on one hectar of farmland. The level of production organization intensivity indicates how the farmers activity is organized. The way of calculating this indicator that takes into account the level stock and the structure of crops, can be found in Olko-Bagieńska and Zietara (1995).

It has got to be mentioned that in order to allow the model serve forecasting purpose the classification criterion was based of farmers income from the year ahead of that one for which the financial indicators were taken.

Because the analysis was done for data covering a period of a few years, the quantities expressed in PLN were recalculated into constant prices with respect to the year 1992.

Due to the requirements of the algorithm applied here the preliminary selection of indicators was necessary. If the correlation coefficient of two indicators was larger than 0,8 only this one was considered in further calculation that was more correlated with farmer's income. It has got to be mentioned that in various years different indicators could be eliminated. That means that the models constructed in various years were based on different sets of indicators.

From the mathematical and statistical point of view the problem of farmer households classification presented here is analogous to the forcasts made in order to alert to firm bankruptcy or to estimate credibility of individual bank clients in the loan sector. The investigated population is divided into two groups. In the case of forecasts made for warning purpose one group consists of firms that are likely to go bankrupt, the other of firms in good financial condition.

Banks are also interested in distinguishing reliable clients from those who are likely not to be able to pay the loan/credit back.

A comprehensive treatment of the above problems can be found in the literature, with the fundamental paper by Altman (1968). Some examples of construction forecasts alerting to bankruptcy can be found in Altman, Giancarlo, Varetto (1994), Hadasik (1998), harmol, Czajka, Piechocki (2004), Hołda (2001) and Mączyńska (2004), while estimation of credibility of individual bank clients can be found in Staniec (2004) and Witkowska, Staniec (2002). A comprehensive outline of systems of early aleting to bankruptcy can be found in the book by Zalewska (2002).

In case of farms there is no need of building typical systems of warning against bankruptcy. Namely, in case of farms the problem of going bankrupt does not exist. This results from a general aversion to credits/loans and low maintenance costs in case of farms (low taxes, low health insurance fees and low pension contributions). Kisielińska (2004) presents a proposal of an early warning system for households. The classification criterion was based on the farmers' income with the boundary value equal to zero.

In the above mentioned publications concerning application of discriminant analysis to bankruptcy forecasting or to evaluating the credit reliability of bank clients no influence of time was considered. The calculations were made for one year. A natural question arises. Can the models obtained in that way be used for a different period? The research presented in that paper is the trial to answer that question.

The classification models were built with application of two methods – discriminant analysis and neutral networks. The aim was to compare their effectiveness. Some examples of neutral network application to discrimination problems can be found in Altman, Giancarlo, Varetto (1994), Kisielińska (2004) and Yang, Platt, Platt (1999). The authors have not admited the advantage of network models, on the contrary, some results indicated the advantage of classical discriminant models over those built with neutral networks.

3. The results of classification done with the linear discriminant function and neutral networks

Table 1 shows the size of the data sets in the following years and the median of farmer's income in changeable and constant prices refered to 1992. The data below indicate that the median of farmer's income was lowest in 1999 and only slightly higher in the years 1994, 2000 and 2001. The largest value of income could be noticed in 1996, and only slightly lower than that in 1995, 1997 and 2002.

Table 1

The size of data sets in the following years and the median values expressed in constant and changeable prices (in PLN)

Year for which a forecast was made	Number of farms	Median of income (constant prices)	Median of income (changeable prices)
1993	663	2 822.44	38 187 570.00
1994	703	2 355.34	42 129 233.00
1995	858	3 541.91	8 096.50
1996	770	3 765.49	10 320.50
1997	997	3 060.78	9 639.00
1998	998	2 673.83	9 414.00
1999	998	1 588.62	6 001.50
2000	912	2 225.81	9 258.00
2001	851	2 468.93	10 834.00
2002	663	2 989.81	13 369.00

Source: own calculations.

In order to build forecasts of financial situation of farms for each year separately functional and network models were build. Functional models were estimated as linear discriminant models and in the sequel will be denoted by

LFD_t, on the other hand network models will be denoted by SN_t, where $t = 1993, ..., 2002$ stands for a year for which a forecast was constructed.

In the next step the data gathered from all years were joined in one data set that contained 7750 cases/samples. A functional model estimated for that set is denoted by LFD_W, and the network model by SN_W.

Linear discriminant functions were obtained using a stepwise forward analysis. In this method one introduces into the model step by step those features that influence the discrimination of classes in the most significant way.

The network models were build with help of a tool called authomatic projektor, that is capable of testing many networks and selecting both their structure and the level of complexity. The calculations for each set of data were repeated several limes and out of constructed networks the best was chosen. In eight cases the best network was a perceptron with one level hiden, in two cases networks with radial base. It has got to be mentioned that the calculations leading to building a network were long-lasting, especially those for the set of full data.

Tables 2 and 3 show the percentage of properly identified farms, for functional and network models respectively.The rows in the table represent years for which the data was collected (more precisely features describing farms). The columns describe models used for classification. The index denotes the year, for which the forecast is made (is order to assure forecasting abilities the index is greater by one then the year for which the data was sourced. The best clasification results for a given year were made bold in the tables.

The largest difference (over 15%) obtained in classification with linear discriminant models is for the year 1992. The best results were obtained with the function $LFDP_{1993}$ (the model was built for that year), the worse for $LFDP_{1997}$. The most homogeneous is the classification obtained for 1998. The difference between the best ($LFDP_{1999}$) and the worse ($LFDP_{1997}$) model is below 4%. The mean dispersion in the classification results was above 8%.

The diagonal of Table 2 shows classification results for farms done with help of a function obtained for the same year for which the data was used. It has got to be said that in five cases the result was not good. Namely, the models for subsequent years were built based on different sets of features. The features were eliminated based on the value of correlation coefficients. This is a reqiurement posed by computational algorithms that were applied. The results obtained indicate that one should pay attention to a proper feature selection as this can improve classification quality. The correlation coefficients alone do not guarantee the best set of features[1].

[1] The proper feature selection for the model can be obtained with genetic algorithms. Their application has however also drawbacks, like e.g., in the case of Statistica, random selection of parameters that control complexity of the models.

The best classification results were obtained with the function $LFDP_{1993}$. It is quite safe to apply a $LFDP_W$, obtained based on the complete data for all considered years. This function does not give the best results but those obtained with it are not the worse.

Neutral networks in majority cases gave better classification results than functional models.

The comparison of classification results obtained with different network models for the same data indicates that alike in the case of functional models one notices a considerable diversity. The largest difference occured also for 1992 (slightly more than 11%). The worse network was built for forecasting purpose in 1998 (SNP_{1998}), the best one of course was SNP_{1993}. The smallest diversity was achieved for 1993 (the difference between the best network – SNP_{1994} and the worse – SNP_{1998} was less than 5%). The mean dispersion of results was 7,5%.

One should notice that the most proper classifications appeares on the diagonal. This means that the best models are network models built on the basis of classified data. In the case of neutral networks the problem of preliminary features selection does not exist. The model alone choses the best set of features for given conditions.

Neutral network built on the basis of the complete data set, alike its functional counterpart, gave moderate results – not the best but also not the worse. It was however evidently better for all years than the fonctional model ($LFDP_W$). The smallest classification impovement was achieved for 1998 (slightly over 2%), the largest for 1994 (almost 10%). The number of proper classifications for all years was 80,80% in the case of SNP_W and hardly 75,38% for $LFDP_W$.

Summarizing one should say that application of a discriminant model obtained for a given year to data classification taken from another year can give evidently worse results.

Many authors who are aware of the advantages of discriminant analysis pay also attention to its drawbacks. M ą c z y ń s k a (2004) emphasizes „restrictions of mechanical transfering the models obtained for conditions of one specific country or sektor to another area". H a r m o l, C z a j k a and P i e c h o c k i (2004) point out that the models are sensative to changes of the sample set of firms. They also notice possibility of the model becoming outdated due to the time changes. This was confirmed by the classification results obtained in the paper.

All functional and network models were used for classifying data gathered in the year 2001 in order to obtain forecast of the farms' financial situation for the year 2002. Classification results are shown in Table 4.

Table 2

Results of farms classification done with $LFDP_t$, where t = 1993, ..., 2001 and $LFDP_W$ (in %)

Year for which the indicators were collected	Number of farms	$LFDP_{1993}$	$LFDP_{1994}$	$LFDP_{1995}$	$LFDP_{1996}$	$LFDP_{1997}$	$LFDP_{1998}$	$LFDP_{1999}$	$LFDP_{2000}$	$LFDP_{2001}$	$LFDP_W$
1992	663	**81.45**	75.72	76.92	68.48	65.91	69.38	70.74	73.30	71.49	72.10
1993	703	76.96	**76.96**	76.10	69.42	73.12	72.55	72.97	68.71	70.27	73.54
1994	858	**82.28**	76.92	80.42	74.83	69.58	71.45	72.38	72.26	71.10	72.84
1995	770	78.05	77.01	77.92	**78.31**	72.73	75.32	71.56	69.48	70.52	74.16
1996	997	75.03	75.53	76.43	**76.93**	74.82	73.02	71.01	69.01	70.21	73.92
1997	998	79.36	78.46	**81.26**	79.56	75.05	78.16	76.85	74.15	74.75	77.66
1998	998	77.15	76.35	77.76	74.85	74.45	75.75	**78.06**	76.45	75.25	77.25
1999	912	**80.92**	75.44	80.26	73.57	70.83	74.34	74.78	78.84	73.68	75.66
2000	851	**82.14**	81.32	81.20	78.97	76.26	78.38	80.96	80.26	80.49	79.67
Total	7750	**79.16**	77.08	78.80	75.33	72.79	74.48	74.57	73.77	73.23	75.38

Source: own calculations.

Table 3

Results of farms classification done with SNP_t, where t=1993,...,2001 and SNP_W (in %)

Year for which the indicators were collected	Number of farms	SNP_{1993}	SNP_{1994}	SNP_{1995}	SNP_{1996}	SNP_{1997}	SNP_{1998}	SNP_{1999}	SNP_{2000}	SNP_{2001}	SNP_W
1992	663	**83.71**	77.53	78.28	77.98	75.41	72.55	79.79	76.77	76.17	78.13
1993	703	78.24	**79.09**	75.53	74.82	74.11	74.11	78.24	75.96	77.52	77.67
1994	858	83.33	81.24	**84.62**	82.17	80.42	77.62	81.12	79.95	76.34	82.52
1995	770	78.57	77.27	80.65	**85.84**	83.38	82.34	82.34	76.75	77.66	82.34
1996	997	75.63	72.72	76.63	79.84	**81.85**	80.24	77.53	73.52	74.02	77.83
1997	998	78.76	76.55	80.66	83.27	83.37	**83.67**	81.56	79.66	78.96	82.16
1998	998	75.85	72.75	76.15	78.06	78.56	76.35	**80.66**	78.66	78.56	79.56
1999	912	80.37	76.10	79.61	78.40	80.37	78.95	79.39	**82.24**	79.17	82.24
2000	851	80.14	77.91	80.14	81.90	83.90	80.73	83.43	82.73	**84.49**	84.14
Total	7750	79.18	76.56	79.14	80.34	80.41	78.78	80.27	78.54	78.11	**80.80**

S o u r c e: own calculations.

The largest percent of correct classifications – 78.36% was obtained by two models – $LFDP_{1996}$ and SNP_{1999}. The worse result 72.13% was given by SNP_{1993} and SNP_{1994}. The dispersion of classification results is over 6%. Good results were obtained using models built based on full data set. SNP_W has classified correctly 77.16%, and $LFDP_W$ 76.68% cases.

The forecast made for 2002 is a good test of effectiveness of models, because the data from 2002 have not been used for models construction in any case.

Table 4

Results of farms classification done for forecasting their financial situation in 2002 (in %)

LFDP applied	Percent of properly classified farms with LFDP	Percent of properly classified farms with SNP	SNP applied
$LFDP_{1993}$	72.70	72.13	SNP_{1993}
$LFDP_{1994}$	77.25	72.13	SNP_{1994}
$LFDP_{1995}$	76.49	72.99	SNP_{1995}
$LFDP_{1996}$	78.39	76.49	SNP_{1996}
$LFDP_{1997}$	75.26	77.25	SNP_{1997}
$LFDP_{1998}$	75.64	77.73	SNP_{1998}
$LFDP_{1999}$	75.83	78.39	SNP_{1999}
$LFDP_{2000}$	74.03	75.36	SNP_{2000}
$LFDP_{2001}$	74.79	76.11	SNP_{2001}
$LFDP_W$	76.68	77.16	SNP_W

Source: own calculations.

4. Financial indicators with the largest forecasting power

The discrimination power of a feature is described by the standardized discriminant function coefficient. The largest the absolute value of the coefficient the largest the influence of the feature in the model. In the network models the discrimination power of an indicator is described by its rank in the so called sensativity analysis. In both cases indicators can be ranked according to their significance in individual models

In order to distinguish features with the largest discrimination power, one had to evaluate the indicators in every model. The feature with the largest rank in the network model or largest standardized coefficient in the functional model was assigned 10 points. Next indicators were assigned one point less, ect. The points were added and the total allowed to distinguish 10 indicators with the largest discrimination power (the rest obtained lower results). The features with their ranks and calculation formulas are shown in Table 5.

Area in hectare of cropland has the most influence on forecasting, profitability of own capital and profitability of sale a little lower. Next characteristics obtained distinct fewer points. They are work output indicator, level of production organization intensity, work output indicator and cash flow. Level of production intensity, quantity of fixed assets at enterprises and age of the farm manager have much lower meaning.

Out of ten features only four are financial indicators (WRKW, WRS, WCF and WWPKS), two are indicators of farming efficiency (WPZ and WWP), the rest are features that describe the farm (PUR, PIOP, PIP and WIEK). Models built based of financial indicators alone were of much lower quality and gave very inaccurate forecasts.

Table 5

Indicators with the best forecasting properties and their notations

Name of the indicator	Calculation formula	Suma rang
Area in hectare of cropland (PUR)	–	172
Profitability of own capital (WRKW)	Farmer's income/ Own capital	139
Profitability of sale (WRS)	Farmer's income / Final output brutto	135
Land productivity indicator (WPZ)	Final output netto / Farmland in hectare of cropland	74
Level of production organization intensity (PIOP)	Calculated as in B. Kopeć	66
Work output indicator (WWP)	Final output netto / Labour input (the number of hours spent on farming)	60
Cash flow (WCF)	(Farmer's income + depreciation) / Final output brutto	56
Level of production intensity (PIP)	Material and financial input on 1 ha of cropland	42
Quantity of fixed assets at enterprises (WWPKS)	Fixed capital/ Total liabilities	35
Age of the farm manager (WIEK)	–	32

Source: own calculations.

5. Conclusions and summary

The results shown in the paper allow formulating the following conclusions:

1. Discriminant models (both functional and network) built for forecasting purpose in the farms sektor on data from one year give much worse results when

applied to classification purpose in other years. The best option is to use models built on data gatherd over the period of many years. Those models may not give the best possible results but the results obtained can be regarded as satisfactory.

2. Network classification models in majority cases did better than the functional. A visable improvement of the quality of those models could be obtained thanks to numerous repeating of calculations.

3. The comparison of functional and network classification shows that the first are easier in use and more universal. Namely, functional models are equipped in formulas that do not exist in network models. Application of a network model requires specific software and a file with a prepared taught network. The time needed for building a network model is much longer than in case of the functional model. The advantage of network models is that one can improve the model (one can build many networks with different structure and chose the best one).

4. The percentage of properly classified farms to forecasting purpose in the years 1993–2001 was contained in the interval (66%, 86%). The results obtained indicate that it is necessary to improve the quality of the models by considering some additional features (e.g., farm location, education of the farmer), the devision of the population into subgroups and building separate models for group (e.g. according to the production type).

Referencens

Altman E. I. (1968), *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*, ",Journal of Finance", **23**.

Altman E. I., Giancarlo M., Varetto F. (1994), *Corporate distress diagnostic: comparison using linear discriminant analysis and neural networks (the italian experience)*, "Journal of Banking and Finance", **18**.

Hadasik D. (1998), *Upadłość przedsiębiorstw w Polsce i metody jej prognozowania*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

Harmol M., Czajka B., Piechocki A. (2004), *Analiza dyskryminacyjna. Przegląd najważniejszych modeli*, „Przegląd Organizacji", **4**.

Hołda A. (2001), *Prognozowanie bankructwa jednostki w warunkach gospodarki polskiej z wykorzystaniem funkcji dyskryminacyjnej ZH*, „Rachunkowość", **5**.

Kisielińska J. (2004), *Wykorzystanie liniowej funkcji dyskryminacyjnej oraz sieci neuronowych do prognozowania kondycji finansowej gospodarstw rolniczych*, „Przegląd Statystyczny", **2, 51**.

Kulawik J. (1995), *Wskaźniki finansowe i ich systemy w zarządzaniu gospodarstwami rolniczymi*, Studia i monografie, z. 72, IERiGŻ, Warszawa.

Mączyńska E. (2004), *Systemy wczesnego ostrzegania*, „Nowe Życie Gospodarcze", **12**.

Olko-Bagieńska T., Ziętara W. (1995), *Materiały do ćwiczeń z organizacji i ekonomiki gospodarstw rolniczych*, Wydawnictwo SGGW, Warszawa.

Rachunek ekonomiczny i analiza finansowa w przedsiębiorstwie rolniczym, Warszawa 1994.

Staniec I. (2004), *Analiza dyskryminacyjna w ocenie wiarygodności kredytobiorców*, „Taksonomia", **11**.
Witkowska D., Staniec I. (2002), *Dychotomiczna klasyfikacja kredytobiorców przy użyciu wielowymiarowej analizy dyskryminacyjnej*, „Acta Universitatis Lodzensis", Folia Oeconomica **156**.
Wyszkowska Z. (1996), *Wybrane elementy wskaźnikowej analizy finansowej na przykładzie przedsiębiorstw rolniczych*, ART, Bydgoszcz.
Yang Z. R., Platt M. B., Platt H. D. (1999), *Probabilistic neural networks in bankruptcy prediction*, „Journal of Business Research", **44**.
Zalewska M. (2002), *Ocena ekonomiczno-finansowa przedsiębiorstwa przez analityka bankowego*, Szkoła Główna Handlowa, Warszawa.

Joanna Kisielińska

Wykorzystanie analizy dyskryminacyjnej oraz sieci neuronowych do prognozowania sytuacji finansowej gospodarstw rolniczych z uwzględnieniem czasu

Celem prezentowanych badań było wyznaczenie liniowej funkcji dyskryminacyjnej oraz sieci neuronowej do tworzenia prognoz sytuacji finansowej gospodarstw rolniczych. Podstawę konstrukcji modeli dyskryminacyjnych stanowił zestaw wskaźników finansowych, natomiast kryterium klasyfikacji oparte zostało na dochodzie rolniczym. Badaną zbiorowość podzielono na dwie równoliczne klasy. Gospodarstwa osiągające dochód rolniczy mniejszy od mediany (gospodarstwa słabe) zaliczano do klasy I, natomiast o dochodzie od niej większym (gospodarstwa dobre) do II. Taki dobór kryterium klasyfikacji wynika z tego, że w przypadku gospodarstw rolniczych problem bankructwa praktycznie nie występuje, wobec czego nie można dla nich budować typowych modeli ostrzegawczych.

Analizy przeprowadzono na podstawie danych pochodzących z kilku lat, co pozwoliło na zbadanie wpływu czasu na jakość uzyskanych modeli dyskryminacyjnych. Chodziło o sprawdzenie, czy model zbudowany dla jednego roku można będzie wykorzystać w latach kolejnych.

Cel dodatkowy polegał na określeniu wskaźników finansowych o największych zdolnościach prognostycznych, czyli takich, których wpływ na wartość funkcji dyskryminacyjnej jest najistotniejszy.

Modele dyskryminacyjne utworzono w oparciu o wyniki finansowe gospodarstw rolniczych prowadzących rachunkowość rolną pod kierunkiem Instytutu Ekonomiki Rolnictwa i Gospodarki Żywnościowej w latach 1992–2001. Do obliczeń wykorzystany został pakiet STATISTICA, natomiast obróbkę danych i analizę wyników wykonano w arkuszu kalkulacyjnym EXCEL wykorzystując język VISUAL BASIC FOR APPLICATION.