The use of Machine Learning techniques to predict farm size change – An implementation in the Dutch Dairy sector

Diti Oudendag [a,b],2, Zoltán Szlávik [b], Hennie van der Veen1[a]

[a] Agricultural Economic Research Institute Alexanderveld 5, 2585 DB The Hague, The Netherlands

[b]VU University Amsterdam, Faculty of Exact Sciences, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Abstract. This paper investigates the use of several machine learning techniques in order to predict dairy farm size change in the Netherlands. The work presented is part of a larger effort to improve an agricultural model, called the Financial Economic Simulation (FES) model. The FES model simulates midterm financial economic development of farms, but until now it has not taken farm size change into account, which made it static, hence, sub-optimal when significant structural changes might occur in agriculture.

In our work, we used data from the Dutch Farm Accountancy Data Network (FADN), covering the period between 2001 and 2009. After preprocessing the data, we built models using Multiple Linear Regression (MLR) and Neural Networks (NN), and measured model performance at various prediction periods (looking ahead one to eight years in time).

Our results show that the chosen methods are able to predict farm size change effectively, and that prediction quality is best when the aim is to predict farm size four years ahead, for which we also provide a likely explanation.

Keywords: dairy farming, FADN, FES, neural networks, multiple linear regression, farm size and farm size change, forecasting

1 INTRODUCTION

Agricultural enterprises are strongly driven by economic motives, i.e., they want to maximise their production while keeping costs low. To be able to optimise agricultural processes (at field, farm and sector level), and to monitor effects of agricultural policies, a number of models have been developed (Bewley et al., 2001; Ledebur et al., 2008).

One of the economic models developed and used in the Netherlands is the Financial Economic Simulation model (FES-model) (van der Veen, (in press)). The FES-model is an economic model with which effects of policies on the primary agricultural sector can be assessed. The goal of the FES-model is to simulate midterm financial economic development of specific farm types. The model calculates financial results and performance on micro (farm) level and aggregates the results to farm types and macro level (whole sector).

FES uses data from the Farm Accountancy Data Network (FADN: European Commission (2010)). The FADN has been established in the European Union (EU) "to monitor the income and business activities of agricultural holdings and to evaluate the impacts of the Common Agricultural Policy (CAP)" (European Commission, 2011).

The FES-model was reviewed in 2010 by a scientific committee (Kleinhanss, 2010). One of the weak points was found to be the static character of the model: When simulating financial change, the FES model does not take structural changes (e.g. size change of the farm over time) into account. This is a significant limitation, thus, the modelling team decided that size change needs to be taken into account as well. However, in order to incorporate farm size

change to predict future financial situations, one also needs to be able to simulate how farm size changes over time.

The research presented in this paper investigates if farm size change can be predicted using machine learning (ML) techniques. We focus our research on one particular farm type, i.e. dairy farms, as 24% of the farms in the Netherlands are dairy farms (CBS-LEI, 2012), providing good 'coverage' for our work. Furthermore, the composition of the group of dairy farms is more homogeneous than it is for other farm types, which makes the task slightly less complex than in case of other farms. Once we can predict farm size change for dairy farms we will be able to extend our work and investigate other farm types as well.

The remainder of the paper is organized as follows. First we discuss related work concerning the prediction of farm size change, then concerning the application of machine learning techniques for similar problems. Thereafter we describe the dataset we used (Data Used) and the performed experiments (Experimental Setup). These sections are followed by the presentation of Results. We close the paper with the Conclusions and Future Work.

2 RELATED WORK

In order to be able to predict farm size change, one needs to gather information about what attributes (i.e. growth indicators) to consider, and what predictive modelling techniques to use. The next two subsections discuss related work regarding these two issues.

2.1 Indicators for farm size change

Related work tends to focus on changes in agricultural structure in general rather than on farm size change specifically. However, apart from various other aspects, agricultural structure also covers farm size change, which is our interest. Farm agricultural structure can be characterised, e.g., by the number of farms, farm type and specialisation (Goddard et al., 1993). Values of such attributes can potentially be effective in automatically building models to predict farm size change.

Goddard et al. (1993) investigated what causes structural change. They categorise factors into prices, economic growth, demographics, off-farm employment, structure of related industry and public programs. Weiss (1999, p. 103) states that there are two interrelated elements driving structural change: “entry and exit from the farm sector and the expansion and contraction of continuing farms”. Based on the work of Weiss as well as that of Goddard et al., we identified several categories of attributes that we could use in our work, including investments, efficiency, current farm size, and external financing.

We based our identification of specific attributes within each of these categories on related work and on the availability of attributes in the Dutch FADN-dataset. We used the work of Weiss (1999), Röder and Kilian (2008) and Goddard et al. (1993) for this. Weiss (1999, p. 113), for example, found that “smaller farms are growing much faster towards some minimum efficient scale of production than farms at or above this threshold size”. Röder and Kilian (2008) state that good proxies for the assessment of ending farms are the farmer’s age and recent developments such as newly rented land and investments. Furthermore, they report a negative correlation between livestock density and exit rates. Goddard et al. (1993) mention, for instance, that small farms can survive when having income from outside the sector.

2.2 Using machine learning techniques for prediction purposes

In this subsection we discuss related work on the use of machine learning techniques for prediction in agriculture.

Ahmad (2009) reports modelling poultry growth, and reports about forecasting egg production (Ahmad, 2011). In the former research, the author compared the results of the

Gompertz model and a logistic model of Nahashon et al. (2006) with results from four types of neural networks. Based on the results (Guinean fowl weight per category) he proposed neural networks for predicting poultry growth. Ahmad (2011) compared the results of three types of neural networks with linear regression and the results of an estimated Gompertz model. According to this work, the general regression neural network - GRNN, (Hannan, Manza and Ramteke, 2010) - had the best performance (correlations of 0.68 and 0.71 for the neural network models (NN-models) and 0.36 for the linear regression model).

Põldaru, Roots and Viira (2005, p. 177) concluded that "artificial neural network models (ANN models) may be used for parameter estimation of econometric models". They concluded this from a study into the use of neural networks in predicting grain yield in which they compared the use of multiple linear regression with neural networks in FADN panel data from Estonia. Their correlation score was 0.38 for linear regression and between 0.42 and 0.46 for NN-models.

Bonfiglio (2011) applied a multilayer feed forward neural network (MFNN) to be able to estimate environmental effects as a result of decoupled direct payments in an arable farm system in the Marche region of Italy for the period 2005-2007. The MFNN outperformed the multi-dimensional linear regression technique (correlations of 0.81-0.82 for MFNN, and 0.79 for multi-dimensional linear regression.)

Pao (2008) compared neural networks and multiple regression analysis in modelling capital structure. With his model he predicts debt (the total book-debt/total assets) with seven (financial) variables. He used panel data from Taiwan from 2000-2005. Pao concluded that an ANN models fit better and perform better in forecasting than regression models (Root Mean Square Error - RMSE - for regression models between 0.58 and 0.86 and for neural network models between 0.06 and 0.08).

In a research to farmers home administration and farm debt failure prediction (Douglas, Graves and Johnson, 1999) the results of a neural network (genetic-algorithm-derived) were compared with logistic regression, an OLS-model, the models of Farmers Home Administration and a model of Price Waterhouse. One of their conclusions was that "the NN-model outperforms both the OLS and logit-models based on (classification) error rates" (Douglas et al. 1999, p. 99).

Based on related work described above, we decided to use neural networks and multiple linear regression for modelling in our research, and correlation and RMSE for evaluation.

3 DATA USED

For our research, we used the Dutch FADN dataset, which is also the basis for the currently used FES model. The data used for our research covers the period from 2001 to 2009, and it contains data about dairy farms. As the FADN dataset contains thousands of attributes, a number of them had been pre-selected before they were used for model building. During pre-selection, attributes appearing in related work (see previous section) were paid particular attention to. Table 1 presents the selected independent (explanatory) attributes. They are shown grouped according to five categories, i.e. Investments, Efficiency, Farm size, External financing and Other. One attribute, "no successor", was calculated to have an indication about potential continuity perspectives which might indicate growth of farm size.

The dependent or prediction variable is farm size change. We use the European Size Units (ESU) of a farm as an economic measure for farm size and farm size change. The value of one ESU is defined as "a fixed number of EUR/ECU of Farm Gross Margin" (European Commission, 2011).

Table 1. Pre-selected attributes from the Dutch-FADN data set, with units shown in parentheses.

Investments	**Efficiency**	**Farm size**	**External financing**	**Other**
Total investments (€)	milk production per cow (litre)	European Size Units (ESU)	Subsidies (€)	Concentrates per cow (kg)
Paid interest (€)	Revenues/Cost ratio (ratio)	Number of dairy cows	External income (€)	Costs of fodder (€)
Long loans (€)	Costs per 100 kg milk (€)	Total hours of labour (h)	Fraction of external income	Income (€)
Percentage investment in machinery (%)	Revenues per 100 kg milk (€)	Number of entrepreneurs		No successor (binary)
	Yield per normalized worker (€)	Area property (ha)		Fraction of rented land
	Young animals per cow			Fraction of labour by non-farm holder(s)

After pre-selection of attributes, the data was further pre-processed. As the data was recorded in 2001-2009, attributes containing monetary information needed to be adjusted, to make values from various years comparable. Attributes having euros as unit were deflated with the Gross Domestic Product (GDP). The GDP deflator is a measure of the level of prices of all new, domestically produced, final goods and services in an economy. This transformation scaled data from various years so different years could be compared. Farm size (in ESU) is not deflated while the value of an ESU is already corrected for deflation per two years.

The raw data was organised per year, and we chose to prepare different datasets based on how many years in advance predictions were to be made. As we investigated prediction periods from one to eight years, there were eight datasets created. For instance, when the prediction period was one year, i.e. when we studied if we could predict farm growth one year in advance, every record in the corresponding new dataset contained information about a farm in a certain year, and its growth calculated based on the farm's performance the following year. When the prediction period was longer, farm size information from the appropriate years was taken into account when creating the corresponding datasets. Note that this step produced datasets of different size, containing 2048 records for one-year predictions (largest) to 168 records for eight-year predictions (smallest). The target variable (size change in ESU) was standardized so prediction performance from various prediction periods can later be compared.

Once the eight datasets were created, they were used in experiments with the aim to predict farm size change, naturally, for eight different time intervals (in years). The experiments are described in the next section.

4 EXPERIMENTAL SETUP

For each of the eight datasets, corresponding to eight prediction periods, we built three models. These three models were built using Multiple Linear Regression (MLR), and two Neural Network variants (with one and two hidden layers, respectively). The number of nodes in the hidden layer(s) of the neural networks was determined based on the recommendations by Heaton (2008), and additional fine-tuning. We refer to the neural network variants as NN1

and NN2 in the remainder of this paper. We also generated a baseline model which "predicted" no change in farm size for all farms.

For the 24 models built, we report values of two performance measures, i.e. values of the Root Mean Square Error (RMSE) and the correlation coefficient (R^2) are computed, reported and discussed (see next section). These were obtained using 10-fold cross validation, a technique that is often used to prevent overfitting (Witten et al., 2011).

As the number of attributes after pre-selection (Table 1) might still be relatively high compared to the number of records in each dataset, we made a further selection of attributes per prediction period: We applied backward selection of attributes when estimating the MLR-models. We used a p-value of 0.05 or higher as a stopping criterion during the selection process. The number of attributes remaining after selection varied over different prediction periods. It ranged from 3 for a prediction period of eight years to 17 for a period of three years. Cost of Fodder and Long loans were used for all eight prediction periods and Total Investments for seven prediction periods, indicating their importance in predicting farm size change. Fractions of rented land, income from outside and labour of non-farm holders were not used in any model, showing that they are not effective when predicting farm size change.

After selection of attributes, all three models could be built for each prediction period. The quality of the built models is discussed in the next section.

5 RESULTS

In this section we present the results of the performance of the three model types. Table 2 presents RMSE and correlation values (correlation between the observed and predicted farm growth) for the three models and eight prediction periods. The calculated RMSE for the baseline prediction is on average in between 0.997 and 1.000. The correlation between predicted and observed values is 0.

5.1 Results by Model

As the numbers marked with * show, the MLR model outperforms the other two models six out of eight cases in terms of RMSE, and five out of eight cases in terms of correlation, which indicates that a (reasonably) simple linear model tends to be better than the more complex neural network models. This might be due to the fact that relatively small datasets cause the neural network models to overfit, i.e., these models do not generalise well on unseen data. Based on the above, we believe that the linear model is able to capture enough details to make reasonably accurate predictions, and that the non-linearity of neural network models might be further exploited but more data is required to avoid overfitting.

We also compared our findings with other research that used MLR and NN (Bonfiglio, 2011; Ahmad, 2009; Ahmad, 2011; Pao, 2008 and Põldaru at al., 2005). They found that NN-models outperformed MLR-models or multi-dimensional linear regression models. We did not find this in this research and we think it might be caused by the different nature of problems investigated (e.g. egg production and farm size change might have quite different nature in terms of prediction), data structure (panel data), type of attributes, etc. Nevertheless, we believe that the methodology we used (employing cross-validation, in particular) makes our results reliable, also because, compared to related work, we produced comparable RMSE and correlation values.

Table 2 Performance indicator values for the 24 experiments; best results in terms of models are indicated in bold, while italicised numbers with * indicate best performance per prediction period.

Prediction period (years looking ahead)	Root Mean Squared Error			Correlation		
	MLR	NN1	NN2	MLR	NN1	NN2

1	0.972	1.129	*0.960**	*0.364**	0.251	0.361
2	*0.873**	0.945	0.926	*0.488**	0.476	0.471
3	*0.836**	1.080	0.912	*0.494**	0.401	0.433
4	***0.785****	0.911	0.866	**0.584**	***0.605****	**0.592**
5	*0.827**	1.114	0.858	0.488	*0.592**	0.537
6	*0.796**	1.237	0.826	0.545	0.461	*0.550**
7	0.868	**0.901**	***0.799****	*0.450**	0.425	0.443
8	*0.955**	1.021	0.992	*0.426**	0.400	0.405

5.2 Results by Prediction Period

Looking at Table 2 column-wise, it seems that the best predictions can be obtained concerning a period of four years (see bold values), i.e. farm growth can be predicted best looking ahead four years. Correlation values for all three models are highest at the period of four years, while MLR also performs best in terms of RMSE at four years.

Although it seems intuitive that one should have best predictions for shorter prediction period (e.g. looking ahead one year), better results at longer - but not too long - prediction periods can be explained. For example, an increase in the number of dairy cows will be preceded by longer term investments made in, for instance, fixed assets such as soil and stable (in our research we used the sum of all investments). After these investments, cattle herd is often extended. Cattle herd size is included in farm size; investments in fixed assets are not. In short, if someone invests in their farm, or makes changes in a year, the results tend to be not immediately visible in terms of farm size (there is an "incubation period"). According to our results, results of investments show themselves most after four years.

6 CONCLUSIONS AND FUTURE WORK

In conclusion, we have shown that farm size change can be predicted with reasonable accuracy using machine learning techniques. We have built several models and found that a prediction period of four years is when our models are most accurate.

We believe that the results presented in this paper are sufficiently good to extend this approach to include other farm types as well. Once done, this will enable the FES modelling team to improve the predictions of midterm financial economic development.

REFERENCES

Ahmad, H.A. (2009). Poultry growth modelling using neural networks and simulated data. Journal of Applied Poultry Research 18: pp.440-446.

Ahmad, H.A. (2011). Egg production forecasting: Determining efficient modelling approaches. Journal of Applied Poultry Research 20: pp.463-473.

Bewley, J., R.W. Palmer, D.B. Jackson Smith (2001). Modelling milk production and labour efficiency in modernized Wisconsin dairy herds. Journal of Dairy Science 2001 84(3): pp.705-716.

Bonfiglio, A. (2011). A neural network for evaluating environmental impact of decoupling in rural systems. Computers, Environment and Urban Systems 35: pp.65-70.

CBS-LEI (2012). Landbouwcijfers 2011. Den Haag: LEI.

Douglas, K.B., O.F. Graves and J.D. Johnson (1999). The farmers home administration and farm debt failure prediction. Journal of Accounting and Public Policy 18: pp.99-139.

European Commission (2011). FADN, Methodology, Field of Survey. Retrieved May 21, 2012 from http://ec.europa.eu/agriculture/rica/methodology_en.cfm.

European Commission (2010). Concept of FADN. Retrieved May 21, 2012 from http://ec.europa.eu/agriculture/rica/concept_en.cfm.

Goddard, E, A. Weersink, K. Chen and C.G. Turvey (1993). Economics of structural change in agriculture. Canadian Journal of Agricultural Economics 41, Issue 4: pp.475-489.

Hannan, S.A., R.R Manza and R.J. Ramteke (2010). Generalized Regression Neural Network and Radial Basis Function for Heart Disease Diagnosis. International Journal of Computer Applications, Volume 7-No. 13: pp.7-13.

Heaton, J. (2008). The Number of Neurons in the Hidden Layers. Retrieved 20 May, 2012 from

http://www.heatonresearch.com/node/707

Kleinhanss, W. and Mulder, M. (2010) Assessment of the FES-model.

Ledebur, O. von, Salamon, P., Zimmermann, A., Leeuwen, M. van, Tabeau, A. and Chantreuil, F (2008). Modelling impacts of some European biofuel measures. In: L. Bartova and R. M'Barek. Modelling agricultural and rural development policies: proceedings; 107th EAAE Seminar, Sevilla, 29th January - 1st February 2008, Luxembourg: Office for Official Publications of the European Communities, ISBN13: 978-92-79-08068-5.

Nahashon, S.N., S.E. Aggrey, N.A. Adefope and A. Amenyenu (2006). Modelling growth characteristics of meat type guinea fowl. Poultry Science 85: pp.943-946.

Pao, H.T. (2008). A comparison of neural network and multiple regression analysis in modelling capital structure. Expert Systems with Applications 35: pp.720-727.

Põldaru, R., J. Roots and A.H. Viira (2005). Artificial neural network as an alternative to multiple regression analysis for estimating the parameters of econometric models. Agronomy research 3(2), pp.177-187.

Röder, N. and S. Kilian (2008). Which parameters determine farm development in Germany. 109th Seminar EAAE-congress, Viterbo, Italy November 20-21, 2008.

Veen, H. van der, (in press.). FES: financial economic simulation. The Hague, LEI.

Weiss, C.R. (1999). Farm growth and survival: Econometric evidence for individual farms in upper Australia. American Journal of Agricultural Economics Vol. 81: pp.103-116.

Witten, I.A., Frank, E. and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Burlington: Morgan Kaufmann Publishers.