The use of Machine Learning techniques to predict farm size change

an implementation in the Dutch Dairy sector

Diti Oudendag

Supervisor: Zoltán Szlávik

Research Paper Business Analytics

The use of Machine Learning techniques to predict farm size change

an implementation in the Dutch Dairy sector

Diti Oudendag

2 June, 2012

Research Paper Business Analytics

Vrije Universiteit Amsterdam
Faculteit Exacte Wetenschappen
Studierichting Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

Research on behalf of
Agricultural Economic Research Institute
Alexanderveld 5
2585 DB The Hague
The Netherlands

Preface

Part of the Master study Business Analytics is the writing of a Research Business Analytics Paper. Goal of such a paper is that it should reflect a research combining the business, informatics and mathematic aspect of the study.

As working-student I have worked at the Agricultural Research Institute as a researcher (LEI). At LEI a lot of data is available and I wanted to apply some of the new learned techniques, within these datasets, in a meaningful goal.

Therefore I consulted one of my colleagues (Hennie van der Veen) and asked her what would be an appropriate research question. Together we defined several problems from which one finally was picked: is it possible with machine learning techniques to predict farm size change in the Dutch dairy sector. If this is possible then the results can be used in the FES-model which would improve the quality of the model.

Hereby I would like to thank:

- Hennie van der Veen for making it possible to use new techniques in LEI datasets and for helping with the Dutch FADN;
- Wietse Dol, Karl Shutes and Foppe Bouma for helping with plots in R, questions about panel data and other
- And last but not least Zoltán Szlávik who came up with new and challenging sub-questions.

Abstract

One of the models used for research purposes at Agricultural Economic Institute in the Netherlands, is FES. The goal of the model is to predict midterm financial economic development of specific farm types. The results consist of averages of financial indicators and also distributions of financial indicators. Besides it shows differences of financial indicators between firms with good and bad prospects.

Induced by a review of the FES-model it was decided to implement more dynamics in structural farm characteristics in the model. Predicting farm size change is a start. For this goal, we investigated whether machine learning techniques can be used to predict farm size changes. Hereby we focussed on Dutch dairy farms.

We used data from the Dutch FADN database for the period 2001-2009. Special attention was paid at the prediction period. We also wanted to know for which prediction period (in years) predictions would perform best.

We selected Multiple Linear Regression models and two types of Neural Networks (one with one hidden layer and one with two hidden layers). The performance is measured on the error indicators Root Mean Squared error and correlation between predicted and observed farm size changes.

Based on the results we can conclude that farm size change can be predicted with Machine Learning techniques. On average the MLR-models perform the best although in some cases the performance of NN with two hidden layers is close by. For MLR with OLS there are restrictions to be met. Least half of these restrictions are not fulfilled. Therefore we recommend to take Neural networks with 2 hidden layers into account (or other techniques). These Neural networks perform for this problem on average better than Neural networks with 1 hidden layer.

The performance of the techniques is best for a prediction period of 4 years. This is more or less in line with what happens on dairy farms. In case of growth first there will be investments in soil and stable. Thereafter cattle herd will be extended. Cattle herd size is included in farm size; investments in soil and stable are not.

Performance of Neural networks can be improved by focussing on parameter selection and optimization of model parameters. Another point for further research could be the split up of the problem in grow of farm size and decline of farm size. These structural changes might be induced by different structural and financial variables and therefore should be estimated by different models.

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

Samenvatting

Een van de onderzoeksmodellen op het LEI (Landbouw Economisch Instituut) is het FES-model. Het doel van het FES-model is het voorspellen van de continuiteit en perspectieven op middellange termijn voor verschillende bedrijfstypen. Het model levert gemiddeldelde financiele kengetallen, verdelingen van deze kengetallen en toont ook de verschillen in kengetallen tussen bedrijven met goede en minder goede perspectieven.

Naar aanleiding van een revieuw van het FES-model is besloten meer dynamiek in structuurkenmerken in het model in te brengen. Een eerste aanzet is om verandering in de bedrijfsomvang mee te nemen. Ten behoeve hiervan is in deze studie onderzocht of machine learning-technieken kunnen worden gebruikt om veranderingen van de bedrijfsgrootte te kunnen voorspellen.

Voor dit onderzoek is gebruik gemaakt van FADN-data voor de periode 2001-2009. Naast het hoofddoel van het onderzoek is ook gekeken naar het effect van de lengte van de periode waarover de bedrijfsgrootte verandering is vastgesteld. Met andere woorden: is een voorspelling over 1 jaar beter dan over bijvoorbeeld 5 jaar. De onderzoekspopulatie bestaat uit Nederlandse melkveebedrijven.

In het onderzoek zijn drie technieken gebruikt: multiple lineare regressie en twee typen neurale netwerken waarvan één met 1 tussenlaag en één met 2 tussenlagen. De werking van de technieken is beoordeeld aan de hand van de standaard deviatie en de correlatie tussen de waarnemingen en de voorspelde waarden.

Op basis van het onderzoek kunnen we vaststellen dat machine learning-technieken kunnen worden gebruikt voor het maken van voorspellingen van veranderingen in de bedrijfsgrootte. In het algemeen is de werking van de multiple regressie modellen beter dan de werking van de neurale netwerken met uitzondering van de voorspellingsperiode van vier jaar. Voor het gebruik van multiple regressie gelden echter voorwaarden waaraan niet allemaal is voldaan. Daarom zouden we moeten richten op het type neurale netwerken met 2 verborgen lagen (of andere machine learning-technieken). Deze neurale netwerken werken voor dit probleem in het algemeen beter dan de neurale netwerken met 1 laag.

De technieken werken het beste voor een voorspelperiode van vier jaar. Dit ligt in de lijn van de ontwikkelingen op een melkvee bedrijf. Bij bijvoorbeeld groei wordt eerst geinvesteerd in de grond en de stal waarna de veestapel wordt uitgebreid. De veestapel zit verwerkt in de maat voor de bedrijfsgrootte, de stal niet.

De werking van de neurale netwerken kan worden verbeterd door nog gerichter te kijken naar de voorspellende variabelen die zijn gebruikt. Verder kan door een betere keuze van de parameters waarmee de netwerken worden geschat, de werking worden verbeterd. Een ander punt voor vervolg onderzoek is het opsplitsen van het probleem in groei en afname in bedrijfsomvang. Het kan zijn dat deze veranderingen door verschillende structuur en financiele kenmerken worden verklaard. Het apart voorspellen zou dan tot betere resultaten kunnen leiden.

Kernwoorden: melkvee bedrijven, FADN, FES-model, neurale netwerken, multiple lineaire regressie, bedrijfsgrootte en bedrijfsgrootte verandering, voorspellingen.

Table of Contents

1 Introduction

1.1 Context

1.1.1 FADN

The Farm Accountancy Data Network (FADN) has been established "to monitor the income and business activities of agricultural holdings and to evaluate the impacts of the Common Agricultural Policy (CAP)" (European Commission, 2011a).

In 1965 the Council Regulation 79/65 set down the legal basis for the organisation of the FADN network. The regulation obliges member states to set up a network for the collection of accountancy data on financial indicators like income (European Commission, 2011a).

The survey does not cover all agricultural holdings in the Union but only those that could be considered commercial due to their size. The FADN per EU-member state is further restricted by the fact that at member state level the farms in the FADN should at least cover 90% of the total Standard Gross Margin (SGM) covered in the FARM Structure Survey (FSS EUROSTAT) (European Commission, 2009).

The Regulation prescribes a selection plan for the recruitment of farms and also sampling stratification according to three criteria (region, economic size and type of farming) should be taken into account. The methodology applied aims to provide representative data (European Commission, 2011a).

For 2008 the network covers about 80.000 holdings representing 5 million farms in EU25 (European Commission, 2011a).

1.1.2 FES-model

One of the models in the Netherlands using FADN data is the FES-model (Mulder, 1991; van der Veen, 2011).

The goal of the FES-model is to predict midterm financial economic development of specific farm types. The model calculates on micro (farm) level and aggregates the results to farm types and macro level (whole sector). The results at sector level are the final output of the FES-model. Results of the FES-model are averages of financial indicators and also distributions of financial indicators. Furthermore differences between firms with good and bad prospects are presented.

The scope of the first version of the FES-model was horticulture (glass houses) in the Netherlands (Mulder, 1991) but since then the model has been further developed (applicable for all farm types in the Netherlands) and became also applicable in 2010 for all farm types within EU25 (van der Veen et al, 2011).

Depending on the scope of the research goal, the model uses data from the Dutch and/or the European FADN. The Dutch FADN is deviating from the European FADN by having more data available per farm. More information about the Dutch FADN can be found in appendix 1. For more information about the FES-model see appendix 2.

1.2 Problem definition and goal

The FES-model was scientifically reviewed in 2010 (van der Veen, 2011). Results from this review were the appointments of main weak points and related recommendations.

One of the weak points is the static character. “The structural characteristics of the firms, in terms of production plan, size and location are fixed. In addition we believe that the few behavioural equations currently implemented in the model, lack sufficient theoretical and empirical underpinning. It’s our impression that they are largely based on expert judgements or incomplete statistical analysis. The reviewers advised to explore whether behavioural equations could be added which are based on economic principles……..(van der Veen, 2011,p 12)”.

Elaborated on this advice, the research goal of this paper is to find out whether it is possible to predict with machine learning techniques and Dutch FADN data change in farm size. We restrict our research to one farm type because we expect that growth of farm size differs per farm type due to difference in farm size and structure. For instance financial data (like income) might differ year by year between farm types. Also average farm size measured in ESU (European Size Units: see appendix 3) can be quite different between farm types. Dairy farming is an area consuming farm type in the Netherlands (around 50% of the utilizable cultural area is used for dairy farming and other grazing activities: CBS-LEI, 2011) and therefore we will apply our research question to dairy farms. Furthermore the composition of the group of dairy farms is more homogeneous which might lead to better results.

A sub goal of this research is to find out if the length of the prediction period (measured in years) is influencing the predicting results.

1.3 Approach

We will start by making an appropriate dataset out of the Dutch-FADN. We will use data from the years 2001 till 2009. The data before 2000 is not usable because some of the definitions in the system were changed. The data of 2000 is not usable because due to a major revision of the Dutch FADN there have been delivery problems with DG-Agri.

From the yearly datasets we will derive different datasets containing the dependent variables in year t and the change in ESU in the assumed prediction period Δt. Δt is differing from one up to eight years. Economic variables like investments, revenues and so on will be taken into account and also structural variables like the age of the head of the farm. The datasets contain discrete and continuous variables. Due to this and the type of dependent variable (change in farm growth, numerical predictor) we will use different kind of (machine learning techniques).

1.4 Structure of the paper and other issues

(Dutch) FADN data is privacy sensitive data. Farmers are willing to contribute assuming that the data is treated as private. LEI and third parties benefit (for their research) from correct and complete contributions of farmers to this FADN database. For that reason the use of FADN data is only possible under strict limitations. Therefore not all data can be presented in this report. The most important restrictions in the use of FADN data are:

- At least 10 observations per reported group of farms
- No report of minimum and maximum values per attribute

In some parts of the paper there will be a reference to this privacy issue.

We start with an overview of relevant studies in chapter two. On the whole the relevant studies can be divided into two categories: modelling farm growing based on FADN data and use of machine learning techniques in FADN data for predicting issues like wheat production. In chapter three Machine Learning Techniques will be described. Chapter four summarizes the used dataset. The experiments using MLT will be presented in chapter five. Chapter six reports the results. Conclusions can be found in chapter 7. In chapter 8 recommendations are presented.

2 Relevant studies

2.1 Structural changes of farms including farm size change

Change in farm size is one of the elements of change in agricultural structure. Agricultural structure covers various items. It can be used for the chains in agriculture, the agricultural sector and the structure of farms in the primary sector. We will focus on farm structure in the primary sector. Farm Agricultural structure can be for instance characterized by the number of farms and land, capital and labour per farm (Van Bruchem and Silvis, 2008).

Goddard at all (1993) enters the question what causes structural change. They categorize the factors into:

1. Prices: “labour saving technological change and an increase in price of labour can both lead to a fall in labour employment (Goddard at al. 1993, p. 480)”
2. Human Capital: education and the use of new technologies increase the value of human capital
3. Economic Growth: economic growth induces consumers to spend more on different products and changes the form in which consumers purchases their products. The form of the purchased products will change towards (pre-) processed food.
4. Demographics: will there be enough successors or do they prefer a (better) income outside agriculture
5. Off-farm employment: “the decision to work off the farm depends on the marginal value productivity of labour in agriculture versus then best alternative employment opportunity. (Goddard at al. 1993, p. 482)”. Off-farm earnings add to farm income to meet a given level of purchasing power. With outside farm income farms can operate at sizes not consistent with minimum costs
6. Related Industry structure: “changes in related industry sector will change relationships between producers and processors.”

7. Public programs[1]: the effect of public programs can be two sided depending on how programs are designed. Some think that the number of farms might be reduced by benefitting larger farms while most times subsidies are supplied to support smaller farms.

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". According to Weiss most empirical studies on the growth rates of surviving farms typically use Gibrat's Law as point of departure. According to Gibrat's Law farm size at time t for farm i is a linear regression with independent variable farm size at time t-1, an intercept and an error term. Other expanded this model with importance of experience, human capital and other individual characteristics of the farm. Weiss refers to a lot of other research about factors influencing structural change. From the list of Weiss we extracted for our research the factors human capital (labour), experience of the farmer, technology, national economic growth and off farm income, characteristics of the farm family (labour, education, successor and other), change in relative prices, public programs and farm debts.

Weiss (1999) performed two regressions: one for the probability of survival and one for farm growth in Austria in the period 1985-1990. In these regressions he used the characteristics of the farm family. Weiss (1999, p. 113) concluded that "smaller farms are growing much faster towards some minimum efficient scale of production than farms at or above this threshold size." Furthermore he found that multiple job holding has a significant lower probability of farm survival and growth.

Röder and Kilian (2008) investigated whether the transition from coupled to decoupled support instruments may impact the rate of structural change (in Germany). From their literature study they mention that good proxies for the assessment of farm exit rates are the farmer's age and the recent development like rented land and/or investments. Furthermore they report a negative correlation between livestock density and exit rates.

Heidhues (1966) mentions advancing technology and variations in prices for inputs and products causing continuous shifts in optimal farm organisation and therefore change in agricultural structure.

2.2 Using machine learning techniques for prediction

The previous section (2.1) reports several times about the use of multiple linear regression. So far as is known no studies have been done predicting farm size change with neural networks. However the use of machine learning techniques for prediction purposes is increasingly including comparative studies about the use of multiple linear regression and neural networks in FADN and other datasets.

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 for the period 2005-2007. The MFNN outperformed the multiple linear regression technique.

[1] Legislation is not explicitly mentioned in Goddard et all (1993) but it could be placed under Public Programs. Legislation might cause growth of farms because it could be cheaper to fulfil the legislation on larger farms.

Ahmad (2009; 2011) compared Neural Network and Multiple Linear Regression models in two research subjects. The first research subject (Ahmad, 2009) is about modelling poultry growth and the second (Ahmad, 2011) is about forecasting egg production. In the first research Ahmad compared the results of the classical Gompertz model and a logistic model (results of both models cited from Nahason et al, 2006) with results from four types of Neural networks: the neural network with one hidden layer and three hidden layers, a Ward neural network with 5 hidden slabs and a general regression neural network. He judged the performance based on the observed weights and predicted weights (for all models) and on the performance results (R^2) of the three neural network types. The performance of the last two networks was the highest and equal. Based on the results he proposed Neural networks for predicting Poultry growth.

In the research of prediction egg production Ahmad (2011) compared the results of a neural network with 1-hidden layer, a general regression neural network and a Ward-5 (5 hidden slaps) with linear regression and the results of an estimated Gompertz model. In this research the general regression neural network had the best performance (based on R^2). Based on comparing observed results with the results of the estimated Gompertz model, Ahmad concluded that the Gompertz model is not useful for predicting egg production.

Põldaru et al (2005, p. 177) concluded that "artificial neural network models (ANN models) may be used for parameter estimation of econometric models". Põldaru et al (2005) concluded this from a study to the use of neural networks in predicting grain yield in which he compared the use of multiple regressions with neural networks in FADN panel data from Estonia.

Pao (2008) compared neural networks and multiple regression analysis in modelling capital structure. In the model he predicts debt (the total book-debt/total assets) with seven (financial) variables. He used panel data from Taiwan from 2000-2005. Based on the Root Mean Squared Error values (a measure for the errors of the model) Pao concluded that an ANN models fit better and perform better in forecasting than regression models.

In an extended research to farmers home administration and farm debt failure prediction (Douglas et al, 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. Goal of the modelling was to develop better loan-making criteria for direct loans, to strengthen FmHA's lending policies. One of their conclusions was that "the NN-model outperforms both the OLS and logit models based on error rates" (Douglas et al. 1999, p. 99)

McQueen et al (1995) and Garner et al (1995) show examples of applying machine learning techniques in Agricultural data. They do not compare different techniques.

3 Data Mining and Machine Learning Techniques

3.1 Introduction

According to Chapple (2012) a definition of Data mining is: **"**Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering". Witten at all (2011, p. XXI) lies down that "machine learning (techniques) provides the technical basis of data mining. It is used to extract information from the raw data in databases". These techniques are equal to the mentioned data mining techniques in the definition of Chapple.

Machine Learning techniques can be divided into supervised and unsupervised learning. Supervised learning means that an observed outcome is available and is used to improve the model. In case of unsupervised learning there is no observed outcome which can be used in the learning process. Examples of supervised learning techniques are classifications, linear models like linear regression, logistic regression and other. Unsupervised learning techniques are for instance clustering, principal component analysis, single value decomposition and SOM (self-organizing MAP: neural network).

3.2 Machine learning techniques for numerical prediction variables

The selection of techniques in a research process is limited by the type of variable you want to predict. Is the variable numerical or non-numerical? Generally supervised learning techniques can be divided into those applicable for classification problems and those for numerical prediction. Because we want to predict the change (growth, shrinkage) in ESU which is a numerical prediction variable, we choose Multiple Linear Regression (MLR) and two types of Neural Networks (one hidden layer and two hidden layers).

3.2.1 Multiple Linear Regression

In (multiple) linear regression there is a dependent variable y and one or more independent predicting variables x. Imagine there are p predicting variables x, then the model can be written as:

$$y = X\beta + \in$$

With X a matrix of size n x p (n = number of observations and p number of predicting variables) and β a vector of p+1 by 1. Goal of the linear regression is to find the values for the matrix β (or the weight factors). The most common used way is to minimize the quadratic difference between the predicted and observed values; the so-called OLS (ordinary least squares).

The use of (multiple) linear regression techniques is in principle limited by several restrictions which also depend on the way the weights are calculated. A summary of the most important restrictions are (Osborne and Waters, 2002):

1) The dependent variable y is normally distributed
2) The variables are normally distributed

3) There's a linear relationship between the dependent and independent variables
4) The errors are normally distributed ~N(0,σ^2)
5) Homoscedasticity of the errors (constant variance)

3.2.2 Artificial Neural Network ANN

A Neural network consists at least of an input layer and an output layer. The layers consist of one or more nodes. The number of nodes in the input layer is equal to the number of inputs (the number of attributes you take into account). The number of nodes in the output-layer is equal to the number of outputs or dependent variables.

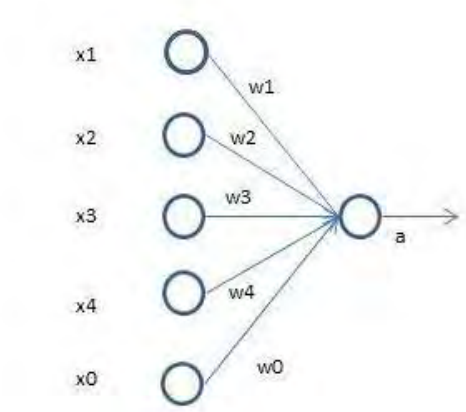

Figure 3.1 A single-layer neural network with a node for a bias

Figure 3.1 shows a neural network with four inputs and one output. An extra node has been added representing a bias (or intercept comparing with linear regression). Using neural networks means that weights of nodes are calculated in a way that the error in prediction is minimized.

We can formulate the output of the network (with an input and an output layer) as (taken into account that the bias $x_0 = 1$).

$$a = \sum_{i=1}^{4} w_i * x_i + w_0 \qquad (3.1)$$

Each node contains an activation function. This activation function transforms the input of the node to the output of the node. An example of such a function is $\sum_{i=1}^{4} w_i * x_i + w_0$ from (3.1).

The weights can be found by minimizing the errors between the predicted and observed values. There are several algorithms which can be used for this question like Adeline and Gradient search. An extended description of algorithms finding the weights can be found in Mitchel (1995) and Witten et al (2011).

Single-layer networks are appropriate for linear relations. With a multi-layer network also non-linear relations can be described (Bishop, 1995). A multi-layer network exists of an input-layer, an output layer and one or more hidden layers. You can choose the number of hidden layers in the network and the

number of nodes per layer. There are several rules of thumb for these decisions. Anon (2012) presents a broad overview including literature references.

One hidden layer appeared to be appropriate for most problems (Anon 2012).

The (main) advantage of a neural network is that the use is not limited by assumptions of distributions of attributes in the underlying dataset. It's a kind of parameter free estimation technique. A disadvantage is the "black box" image. Another disadvantage is the risk on over-fitting. This issue might be solved by using techniques such as cross validation (Lisboa et al, cited in Cerney, 2001, p4).

3.3 K-fold cross validation

An estimated model can be over-fitted when it performs well on the data-set used for the estimation but in case of new observations not in the training set it performs poor. Cross validation is a technique to prevent over-fitting in the estimation procedure. Another goal to use cross validation is to compare the performance of two or more different algorithms (Refaeilzadeh et al, 2009). 10-fold cross validation is appropriate for most models (Bishop, 1995 and Refaeilzadeh et al, 2009). Figure 3.2 demonstrates the k-cross fold validation algorithm.

Action
1. split up the data set randomly into k-partitions.
2. Do k-times a. Train on k-1 partitions b. Test with the partition not used for the training. c. Remember the test error
3. Enddo
4. Calculate the average test-error

Figure 3.2 Algorithm for the k-fold cross validation

A special form of K-fold cross validation is LOOV or leave one out validation (k=1). This method gives the most accurate and almost unbiased estimate but has high variance (Refaeilzadeh et al, 2009) and high computation time (Witten et al, 2011 and Refaeilzadeh et al, 2009). LOOV has not been used in this paper due to its computation time.

4 Description dataset

The Dutch-FADN dataset is part of the main database system at LEI. The data in this database system can be approached by different predefined views. In our research we use a view for the (Dutch)-FADN dataset called COBRA and a view for the predefined dataset for the FES-model called microwave. The data from both views can be linked by the use of a unique farm number. A description of the Dutch FADN and FES can be found in Appendix 1 and 2.

4.1 Variables in the final data set

The selection of the final variables was mainly based on literature (see chapter two). A list of these final variables is presented in table 4.1. Based on the literature from chapter two we determined five main categories of attributes causing structural change (assuming they also count for changes in farm size):

1. Investments: making investments indicates a willing to improve the farm; having more technology; expanding the farm and other
2. Efficiency: when producing efficiently, costs will be lower and probably there will be more use of (new) technology
3. Farm size: this has been discussed in chapter 2
4. Financing from outside: this has been discussed in chapter 2
5. Other

We selected attributes to be representative for what the categories are standing for. Some extra attributes were added which couldn't be classified into the first four categories. One of them was the absence of a successor: the so-called no-successor in this research. The presence of a successor on the farm indicates potential growth (Weiss, 1999). Successors are not directly available in the dataset (they are available from the Annual Census but per 4 year period) therefore a new variable was created by determining the presence of no-successor. This has been done by indicating farms as having no successor when the age of the youngest entrepreneur is 60 years or older.

There might be correlation between many variables, like long term loans, investments and paid interest. An extended correlation matrix is presented in appendix 4. From this correlation matrix it can be concluded that at least for the group of attributes CostsFodder, LongLoans and InterestPaid there are high correlations. Also the farm size in ESU and number of dairy cows is highly correlated (97%). This is called multicollinearity. Because we only use regression for predicting and estimation multicollinearity is not a problem here.

4.2 Handling outliers, missing values and scaling

In this section we discuss all kind of data operations in order to obtain the final datasets.

When combining the FADN dataset with the FES dataset, we lost about 600 observations. These observations are from 2006 till 2009 and do not have a FADN-weight factor. This means that they were already not taken into account within the FES-model.

Farms with negative feeding costs and farms with negative use of concentrates per cow were removed. Other outliers are still outliers but can occur in reality. Due to privacy reasons we cannot plot them.

Missing values were replaced by 0 except for the variable YieldPerNormalizedWorker and income. Here missing values were replaced by the average value (because there is always YieldPerNormalizedWorker and income).

Table 4.1 Summary statistics and scaling factors for the final attributes

Category	attribute	mean	standard deviation	number of missing values	unit	adaption
Investments	TotalInvestments	11.816	23.091	0	euro	/10000
	InterestPaid	3.013	2.937	26	euro	/10000
	LongLoans	65.662	66.28	0	euro	/1000
	PercInvestmMachinery	0.638	0.357	100	%	
Efficiency	milkprodpercow	75.536	12.237	*	liter per cow	
	RevenewsCostratio	82.828	16.92	0	…	
	Costsper100kgmilk	52.035	19.785	0	euro	
	RevenewsPer100kgmilk	39.374	12.251	0	euro	
	YieldPerNormalizedWorker	2.558	2.884	*	euro	/10000
	YoungAnimalsPerCow	0.303	0.122	26	young animals per cow	
Size	ESU	135.82	77.692	0	esu	
	Number of dairy cows	77.497	45.616	0	number	
	TotalHours	38.054	15.011	0	hours	/100
	Number of Entrepreneurs	1.807	0.775	0	number	
	AreaProperty	32.362	23.858	0	are	/100
Financing fom outside	Subsidies	1.48	2.248	*	euro	/10000
	FarmIncomeOutside	1.046	1.205	31	euro	/10000
	FracOutsideIncome	0.109	7.183	189	fraction	
Other	ConcentratesPerCow	20.46	5.727	0	kg	/100
	CostsFodder	1.099	1.75	32	euro	/10000
	Income	5.479	6.365	*	euro	/10000
	NoSuccessor	0.513	0.5	0	binary (0 1)	
	FractionRent (land)	0.301	0.268	*	fraction	
	FractionLaborOthers	0.072	0.128	*	fraction	

*) * 10 observations or less; … no unit

We applied some scaling of the data in order to have the values in more or less the same range. The used scaling factors are presented in table 4.1.

The attributes having euro as unity (not the ESU) were deflated with the GDP (Gross Domestic Product). The GDP deflator is a measure of the level of prices of all new, domestically produced, final goods and services in an economy. Another deflator is the CPI. The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. Both deflators are not covering the situation in agriculture but the GDP is through the relation with the production process most related with the agricultural production process. On the other

hand they are close related (figure 4.1): the correlation between both for the period 2000-2009 is 99.6%.

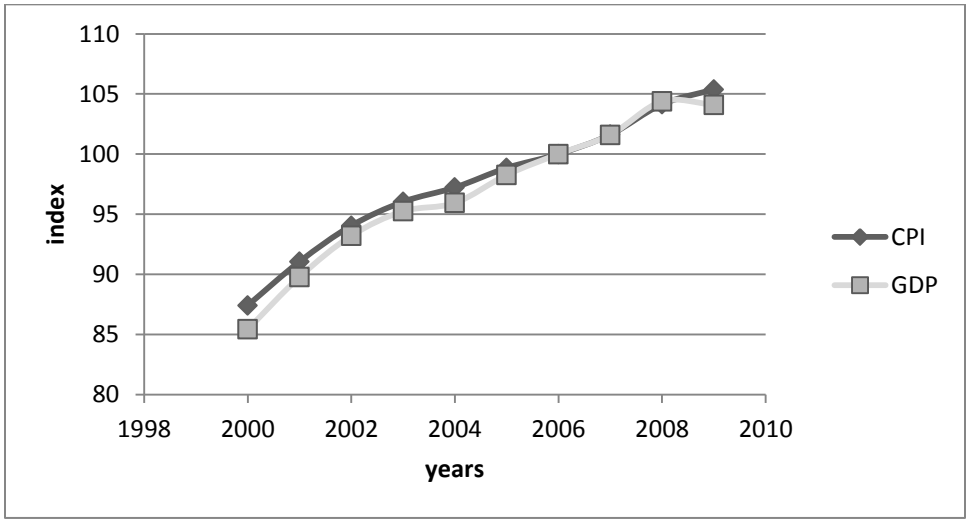

Figure 4.1 Relation between the CPI and BDP deflator for the Netherlands (2006=100)

4.3 Summary of the data

Table 4.1 presents a global overview of the values of the attributes in the dataset. Due to privacy reasons we cannot present distributions. But we can mention that some of the variables are more centralized and less skewed distribute like income then other like for instance number of cows, ESU and long loans. Especially for these skewed distributed attributes we found outliers which can exist in reality.

A correlation matrix of the variables is presented in appendix 4. As already mentioned some of the attributes correlate much with each other like LongLoans and costs for fodder.

4.4 Composition of datasets

One of the sub goals of this research is to find out if ANN-models predict and perform as well as MLR-models and for which prediction period (in number of years) we obtain the best results. Therefore we created eight datasets[2]: each dataset contained the farm data for the starting year and the net change in ESU within the prediction period. As mentioned before farms can be multiple times in one dataset. Table 4.2 presents the number of observations and the number of farms in the datasets. Furthermore table 4.2 shows that on average farms are more than one time present in datasets.

[2] Finally we didn't use the dataset with a prediction period of eight years. This dataset contained only observations for one starting year (2001).

Table 4.2 Number of observations and unique farms for different datasets

Prediction period (years)	Number of observations	Number of unique farms	Average time a farm is in the dataset
1	2048	381	5.4
2	1674	348	4.8
3	1335	284	4.7
4	1058	269	3.9
5	798	247	3.2
6	560	209	2.7
7	363	197	1.8
8	168	168	1

5 Experiments

5.1 Method selection

Machine learning techniques are a gathering of different methods. Selection of the method depends beside other factors on the type of variable to be predicted. In this research we want to predict change in ESU which is a numerical variable. Therefore two techniques applicable for numerical variables were chosen: multiple linear regressions (MLR) and an artificial neural network (ANN or NN). This choice was supported by the fact these techniques have been used several times in research (Bonfiglio, 2011; Ahmad, 2009; Ahmad, 2011; Pao, 2008, Põlderu et al, 2005 and Douglas et al, 1999).

The classical method for analyzing panel data is panel data models or time series cross section regression (TSCS). A description of panel data and the estimated models can be found in appendix 5. In this research we didn't use panel data analyses. The most important reason is that it's not applicable for the goal of this research. One of the outcomes of this research should be an algorithm which can be implemented in the FES-model and with which prediction on farmsize change can be made. Results of panel data analysis cannot be used in case of prediction for new years and new farms being not in the estimated panel data model.

In general one hidden layer in an ANN model is sufficient (Anon, 2012). We used such an ANN and for all prediction periods they performed worse than MLR. While this is not in line with other research (Bonfiglio, 2011; Ahmad, 2009; Ahmad, 2011; Pao, 2008), we also estimated models with two hidden layers.

The numbers of nodes in the hidden layers are presented in table 5.1. The number of nodes in the two hidden layer model were determined by trial and error.

Table 5.1 Number of nodes in the hidden layer for different prediction years

Prediction period	1 hidden layer	2 hidden layers*)
1	7	7 4
2	7	7 4
3	10	10 4
4	9	10 5
5	9	9 5
6	7	8 5
7	5	4 2
8	3	Na

*) the first figure for the first layer, the second figure for the second layer

MLR and ANN models were estimated with Rapid Miner with 10k-Cross validation. The ANN-models were Forwardfeeded NN with back propagation as learning algorithm. The activation function of the nodes is a sigmoid function. Therefore the input variables have been normalized (automatic by the program) to get values between 0 and 1.

The absolute change in ESU is depending on the number of prediction years (figure 5.1). This means that we cannot compare directly the performance measures between the different numbers of prediction years. Therefore we normalized (z-transform) the label attribute ESU-change.

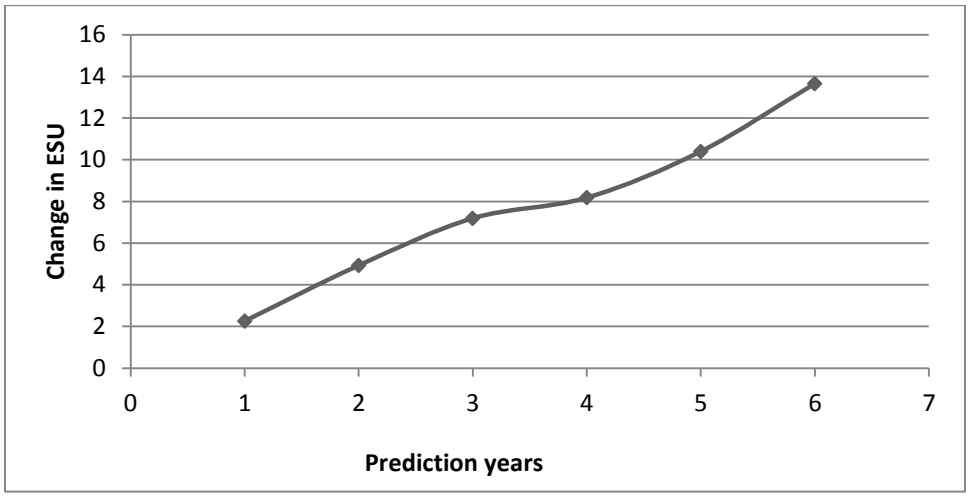

Figure 5.1 Average change in ESU per prediction period

In this research the final selected methods are indicated as MLR, NN and NNH with NN a neural network with 1 hidden layer and NNH a neural network with two hidden layers.

5.2 Selecting final variables in the model

We selected the final variables in models in by the next steps:

- Start modelling MLR, NN and NNH with all variables
- Remove based on the results for all MLR models the variables with less significance (P-level > 0.05) (this is a common technique in statistics)
- Feed the three models MLR, NN and NNH with the selected variables

The final variables used in the MLR, NN and NNH are presented in appendix 6.

We chose to do it in this way to have the same input sets for all models (per prediction period). Differences in performance can therefore not be caused by different inputs.

5.3 Performance indicators

To be able to select the best estimated model one can use performance indicators. In this research we use three performance indicators in the analysis: the Root Mean Squared Error (RMSE), the Absolute Error (AbsE) and the correlation. The definition of the error measures can be found in Witten et al (2011). The RMSE was chosen because one of the methods is linear regression with least squared errors. Therefore the method is based on minimization of the squared errors. The AbsE was chosen to have an extra indicator to compare with. According to Witten et al (2011, p. 182) “in most practical situations the best numerical prediction method is still the best no matter which error measure is used”. The measure correlation is an intuitive one. In contradiction with the other two it doesn’t measure the error but the rate of relationship between predicted and observed values.

6 Results

This chapter is about the performance of the three model types. We want to know which of the three estimated models perform the best. We use pictures of the results (section 6.1 and 6.2) and statistical tests (section 6.3) to analyse the performance results. We also checked the proper use of MLR (appendix 7).

6.1 Performance errors

When looking at the performance errors AbsE and RMSE we should take into account that the model parameters have been estimated based on minimization of the squared error. In figure 6.1 the AbsE and RMSE are presented for different prediction years. The RMSE-errors are higher than the Abs-errors.

Based on figure 6.1 one can conclude that on average the NN with 1 hidden layer performs the worst. The performance of the NN with two hidden layers is performs worse than the multiple linear regression except for a prediction period of 7 years.

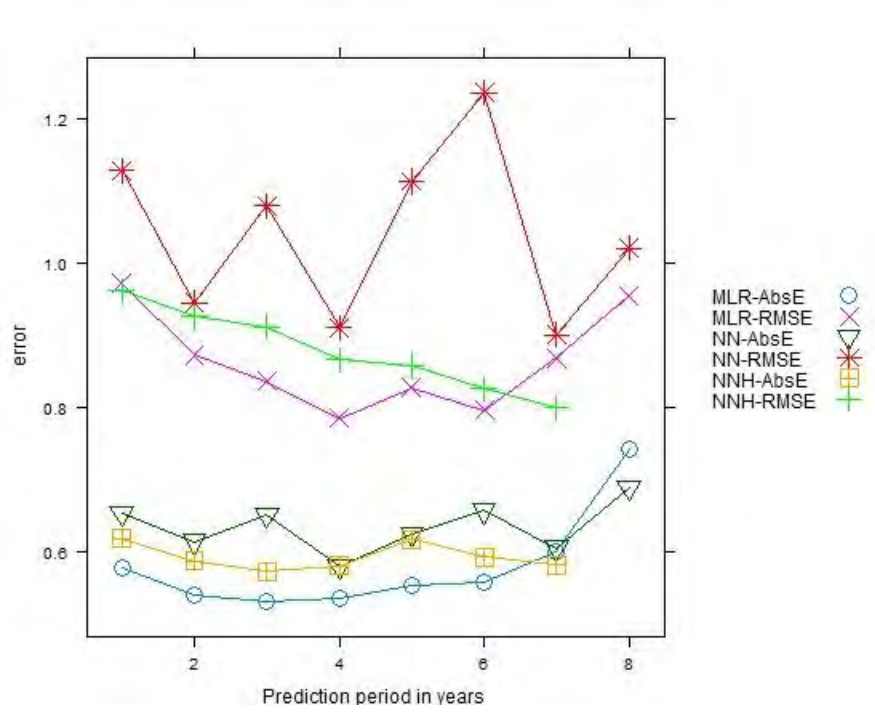

Figure 6.1 Absolute error and Root Mean Squared Error for different models for different prediction years

When we compare the performance between the years it's not directly visible for which number of prediction years the prediction performs the best. It could be four or seven years. Therefore we summed up the RMSE for the three different models en plotted them (figure 6.2).

Figure 6.2 doesn't give an exclusive answer. A prediction period of four or seven years might look equally good.

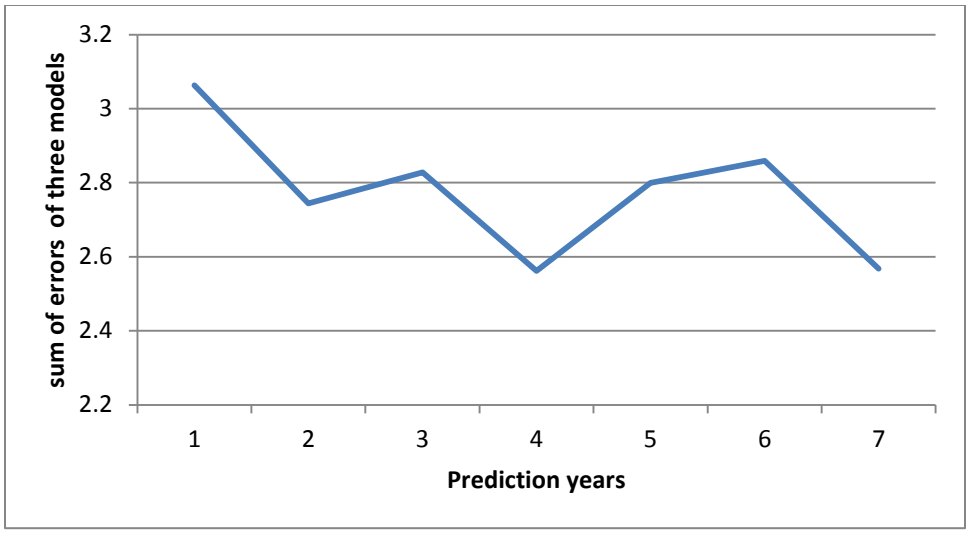

Figure 6.2 Sum of the RMSE of the three models per prediction period.

The correlation between predicted and observed values (figure 6.3) is the highest for prediction period of 4 years for all three models. The correlation of the NN-model with 1 layer is on average lower than the correlation with the MLR and NNH-models.

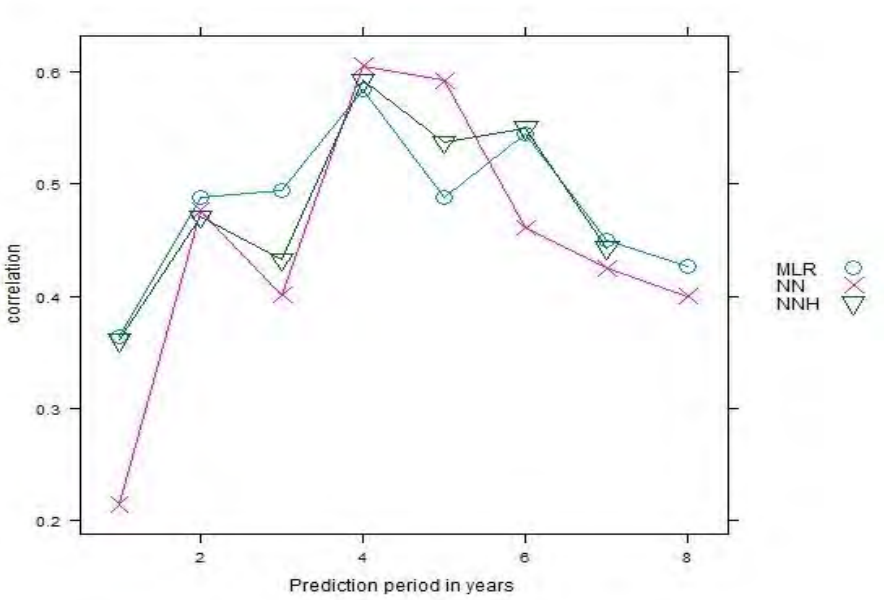

Figure 6.3 Correlation between observed and predicted values for different models for different prediction periods

6.2 95%-confidence intervals of the predictions

Figure 6.4 shows the confidence intervals size of the predictions with the three methods MLR, NN and NNH and the confidence interval of the observed values (Obs) for the different prediction years.

On average the 95%-confidence intervals of the neural network predictions with one hidden layer are the smallest. There is an exception for the prediction period of 4 years. In this case the confidence

interval of the prediction with a neural network with two hidden layers is the smallest. The size of confidence intervals increases with the prediction period, especially for the periods of 5 year and longer.

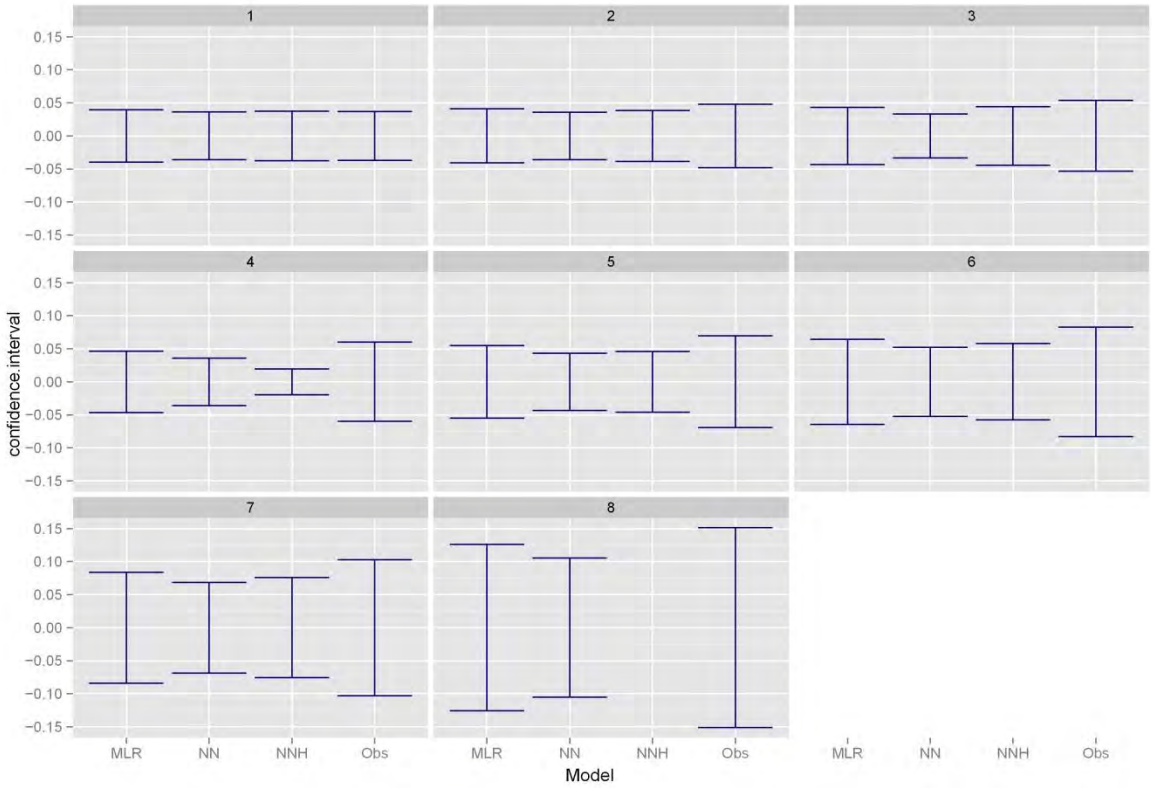

Figure 6.4 Confidence intervals per used prediction period for the used prediction methods and the observed values

This fact of a small confidence interval for the NNH-model for the prediction period of four years can also be seen figure 6.5. The patterns in confidence interval sizes per prediction year, is equal for Obs, MLR and NN. For Neural networks with 2 hidden layers (NNH) we see a “dip” for prediction year 4 and 5. The weight factors for the NNH network are presented in appendix 8.

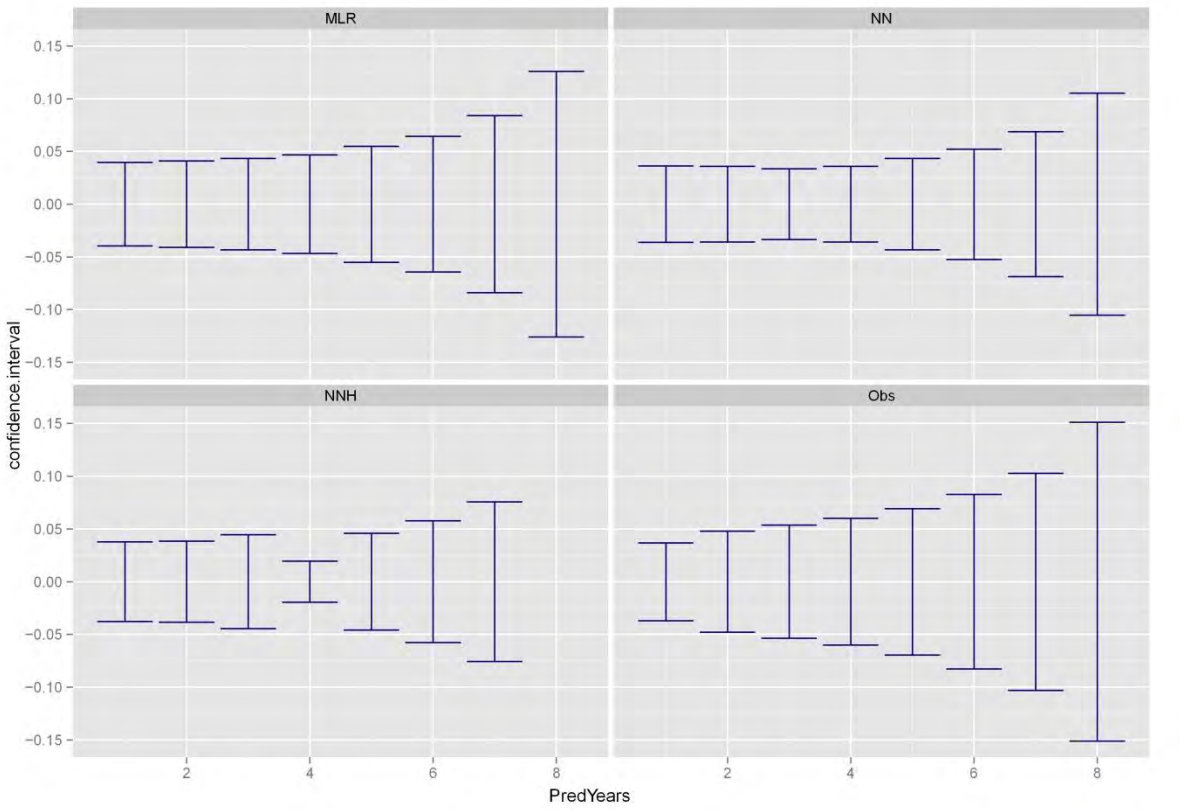

Figure 6.4 Confidence intervals per used prediction method and for the observed values for different prediction years

6.3 Performance testing

Besides the use of pictures also tests can be performed on the outcomes. In order to compare which type of model performs best, we tested with pairwise t-test per prediction period whether the distribution of the results of the models are equal to each other. Table 6.1 shows the T-values for the pairwise t-test[3].

We tested the null hypothesis H_0: the prediction results of the models are equal against the alternative hypothesis H_1: the prediction results of the models are not equal. Because we already know which model is performing better in a combination we can test one-sided with $\alpha = 0.05$.

For α=0.05 and degrees of freedom >120 the limit T statistic for one sided testing is 1.658 and for two-sided testing T=1.980. Taken these T-values into account, table 6.1 shows for two situations that the H_0-hypothesis is not rejected: prediction period 4 NNH versus MLR and prediction period 5 years NN versus MLR. In all other cases H_0 is rejected. The prediction results of the models are not equal to each other.

[3] T-test assumes normality in the data. We didn't test for all situations but on average it won't be true. So formally we should have tested with Kolmogorov-Smirnoff but taken the size of the datasets and the Central limit theory the t-test is appropriate for this situation.

Table 6.1 T-values for testing on different results for different models

Prediction years	NN <-> MLR	NN <-> NNH	NNH <-> MLR
1	53.750	-49.920	24.418
2	-11.660	16.541	28.701
3	-7.772	48.672	45.409
4	18.681	31.134	1.284
5	1.569	-23.812	-19.572
6	5.295	26.791	23.359
7	20.154	41.329	2.484
8	-7.706	Na	Na

Furthermore we can test for which prediction period the models perform best. For instance from figure 6.3 it looks like that prediction periods of four and seven years have the better performing results. Is one of these periods the best? We can test this with the t-test for non-paired observations.
We tested per model the null hypothesis H_0 the performance of the prediction of the four year period is equal to the performance of the prediction of the seven year prediction period against the alternative hypothesis the performance of the prediction of the four year period is not equal to the performance of the prediction of the seven year prediction period (α=0.05). The results are presented in table 6.2.

Taken the same limited T-values as mentioned above we only reject the H_0-hypothesis for the NN-model with one hidden layer: the performance prediction results for a four years prediction period are not equal to the performance of a prediction results of a seven years prediction period.

Table 6.2 T-values for testing on difference in results between a prediction period of four and seven years

Model	T-value	Significant diff with (α=0.05)
MLR	0	no
NN	6.434	yes
NNH	-0.739	no

7 Conclusions

We conclude from figure 6.1 and 6.2 that it looks like that the Neural network with 2 hidden layers performs better than the neural network with one hidden layer. The errors MAE and RMSE are lower for the neural network with two hidden layers compared with the neural network with one hidden layer. According to the size of the errors the Multiple Linear Regression performs on average better than both Neural Networks. The test results with the pairwise t-test (table 6.1) confirm this except for the prediction period of four years. Here the performance of the neural network with two hidden layers performs as well as the MLR-model.

One of the research questions is: for which prediction period do the models perform the best? Looking at the RMSE (figure 6.2) the lowest error is reached for a prediction period of 4 and 7 years for almost all models (MLR, NN and NNH). The highest correlation (60%, figure 6.3) is found for a prediction period of

four years for all models. It's not unlikely that changes in farmsize are seen the best in a longer period. In case of farm size grow on dairy farms there will be first investments to be made in soil and stable. There after cattle herd will be extended. Cattle herd size is included in farm size; investments in soil and stable are not. A period of four years seems appropriate although it could also have been three or five years.

We tested with the non-paired t-test per model type whether the prediction results are equal or not. The test results from table 6.2 show that for the linear model (MLR) and the neural network with two hidden layers (NNH) that the performance of the model with prediction period of four years is rather equal to the performance of the model with a prediction period of seven years.

The confidence intervals of the predictions are higher for a prediction period of seven years compared to a prediction period of four years (figure 6.4 and 6.5). The correlation is lower for a prediction period of seven years. We have doubts about the results for the prediction period of seven years. There are two issues:

1) The dataset contains two years with predicting variables 2001-2008 and 2002-2009 while with four years there five years with predicting variables. Therefore the results can be a coincidence.
2) The deviation in the results for the prediction period of seven years is larger than for four years.

Taken this into account we conclude that the prediction period of four years performs the best.

The use of the MLR-models is restricted by the assumptions to be made. Not all assumptions are full filled (appendix 7). Neural networks with two hidden layers are potentially an alternative to predict farmsize changes.

The main goal of this research is to determine if we can use machine learning techniques to predict farm size changes. Based on the results presented in chapter six we conclude that this is possible although some additional research is advisable. Hereby we think about selection of attributes by optimizing the selection, allowance of different attributes per machine learning techniques, optimizing parameters of the techniques and other.

8 Recommendations

Due to time issues not all possibilities of model selection and model definition and composition have been taken into account. The model performance (of all models) can possible be improved when attribute selection and parameter optimization are used. This will have to be tested out.

The findings have been compared with other research that used MLR and NN (Bonfiglio,2011 ; Ahmad, 2009; Ahmad, 2011; Pao, 2008). They all found that NN-models outperformed MLR-models. We did not find this in this research. This difference in expected outcomes and real outcomes might happen because the mentioned research did not address the use of cross validation so the estimated models

might be over-fitting. On the other hand we think improvement of the performance might be reached by using the options mentioned in the first section.

Another point for further research could be the split up of the problem in grow of farm size and decline of farm size. These structural changes might be induced by different structural and financial variables and therefore should be estimated by different models. This could also increase performance as well.

More theoretical reflection can be made to find out why in our investigation of predicting change in ESU a two hidden layer performs better than a one hidden layer: a NN of one hidden layer is on average appropriate for non-linear problems (Bishop, 1995).

Furthermore in case the time series of data increases, models for the prediction period of seven years can be re-researched and also for other periods.

Literature

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

Ahmad, H.A. (2009). Poultry growth modelling using neural networks ans simulated data. . J. Applied Poultry Research 18: 440-446.

Anon (2012) comp.ai.neural-nets FAQ, Part 3 of 7: Generalization. Section-How many hidden units should I use? Available at: http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-10.html. [Accessed at: 7 April 2012]

Baltussen, W.H.M., R. Hoste, H.B. van der Veen, S. Bokma, P. Bens en H. Zeewuster (2010). Economische gevolgen van bestaande regelgeving voor de Nederlandse varkenshouderij. The Hague, LEI, report 2010-010.

Baltagi, B.H. (2008). Forecasting with Panel Data. J. Forecast, 27: 153-173.

Baltagi, B.H. (2005). Econometric Analysis of Panel Data. John Wiley & Sons, Ltd. West Sussex, United Kingdom

Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, New York, United States.

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

Bruchem, C. van, H. Silvis(red) (2008). Agrarische structuur, trends en beleid: Ontwikkelingen in Nederland vanaf 1950. Den Haag, LEI 2008-060.

CBS (2012). Agriculture and fishing. Available at: http://www.cbs.nl/en-GB/menu/themas/landbouw/beschrijving/beschrijving.htm. [Accessed at: 19 May 2012]

Cerney, P.A. (2012). Data mining and Neural Networks from a Commercial Perspective. Available at: http://www.orsnz.org.nz/conf36/papers/Cerny.pdf [Accessed at: 19 May 2012]

CBS-LEI (2012). Landbouwcijfers, 2011. Den Haag, Netherlands.

Chapple, M. Data Mining. Available at: http://databases.about.com/cs/datamining/g/dmining.htm. [Accessed at: 17 May 2012]

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

European Commission (2011a). Concept of FADN. Available at: http://ec.europa.eu/agriculture/rica/concept_en.cfm. [Accessed at: 20 May 2012]

European Commission (2011b). FADN, Methodology, Field of Survey. Available at: http://ec.europa.eu/agriculture/rica/methodology_en.cfm. [Accessed at: 7 April 2012]

European Commission (2009). EU Dairy Farms Economics. Brussels, European Commission, Unit L3 D(2009).

Garner, S.R. G. Holmes, R.J. McQueen and I.H. Witten (1995). Machine Learning from Agricultural databases: practice and experience. Proceedings of the 14th NZCS Conference: pp 75-82. New Zealand Computer Society, Wellington, New Zealand.

Goddard, E, A. Weersink, K. Chen and C.G. Turvey (1993). Economics of Structural Change in Agriculture. In: Can. J. Agr. Econ. 41, Issue 4: 475-489.

HeidHues, T. (1966). A Recursive Programming Model of Farm Growth in Northern Germany. Am. J. Agr. Economics 48 (3 Part I): 668-684.

LEI (2011). Binternet: LEI's Farm Accountancy Data Network. Available at: http://www.lei.wur.nl/UK/statistics/Binternet/. [Accessed at: 7 April 2012]

Lisboa, P.J.G., B. Edibury and A. Vellido. Business Applications of Neural Networks. World Scientific, Singapore, USA, UK.

McQueen, R.J., S.R. Garner, C.G. Nevill-Manning and I.H. Witten. Applying Machine Learning to Agricultural Data. J. Computing and Electronics in Agriculture 12 (4): 275-293.

Mitchel, T.M. (1995). Machine learning. McGraw-Hill, Singapore.

Mulder, M.(1991). Financiele Analyse en Continuiteitsvoorspelling. The Hage, LEI, report 4.127.

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

Offerman, F (editor), (2011). Implementation, validation and results of the costs of production model using national FADN data bases. Braunschweig, vTI, FACEPA Deliverable No. 3.1 – January 2011.

Osborne, J.W. and E. Waters (2002). Four Assumptions of Multiple Regression That researchers Should Always Test. Available at: http://pareonline.net/getvn.asp?v=8&n=2. [Accessed at: 6 April 2012]

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

Põlderu, 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), 177-187.

Poppe, K.J. (2004). Het Bedrijven-Informatienet van A tot Z. The Hague, LEI, report 1.03.06

Refaeizadeh, P., L. Tang and H. Liu (2009) Cross-Validation. Available at:

http://www.public.asu.edu/~ltang9/papers/ency-cross-validation.pdf. [Accessed at: 5 April 2012]

Rijnhard, S., L. Van Staalduinen and M. Spijkerman (2001). Handleiding voor de mogenlijkheden en het gebruik van panel data op het LEI: het informatienet en de landbouwtelling. LEI, Den Haag, Netherlands.

Röder, N. and S. Kilian (2008). Which parameters determine farm development in Germany? Presented at the 109th EAAE Seminar, Viterbo, Italy

Shami, A.A., A. Lofi, E.Lai and S. Coleman (2011). Forecasting Macro-Knowledge Competitiveness: Integrating Panel Data and Computational Intelligence. Nottingham Trent University, Nottingham, United Kingdom.

Torres-Reyna, O. (2009). Panel Data Analysis: Fixed and Random Effects. Princeton University, Princeton, USA.

Veen, H. van der (2011). FES: Financial Economic Simulation. The Hague, LEI <under construction>

Veen, H. van der, K. Oltmer and K. Boone (2006). Het BIN-nenstebuiten: beschikbare gegevens in het Bedrijven–Informatienet Land- en Tuinbouw. LEI, Den Haag, Netherlands.

Vrolijk, H.C.J., H.B. van der Veen, J.P.M. van Dijk (2010). Sample of Dutch FADN 2008; Design principles and quality of the sample of agricultural and horticultural holdings. The Hague, LEI, report 2010-096.

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

Witten, I.H., E. Frank, M.A. Hall (2011). Data mining: Practical Machine Learning Tools and Techniques. Elsevier, Burlington, USA.

Abbreviations

AbsE	Absolute Error
ANN	Artificial Neural Network
CPI	Consumer Price Index
GDP	Gross Domestic product
ESU	European Size Units
EU10	Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovenia, Slovakia
EU15	Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Malta, Netherlands, Portugal, Spain, Sweden, United Kingdom
EU25	EU15 + EU 10
EU27	EU25 + Bulgaria + Romania
FADN	Farm Accountancy Data Network
FES	Financial Economic Simulation model
FSS	Farm Structure Survey
LEI	Landbouw Economisch Instituut
LOOV	Leaf One Out Cross validation
MAE	Mean Absolute Error
MLR	Multiple Linear Regression
NN	Neural network
OLS	Ordinary least squares (regression)
RMSE	Root Mean Squared Error
SGM	Standard Gross Margin
SO	Standard Output
TSCS	Time Series Cross Section regression

Appendices

Appendix 1 Dutch FADN

A1.1 Context

Source: Vrolijk et al (2010)

Dutch FADN is a data network consisting of accountancy data of about 1500 agricultural and horticultural farms in the Netherlands. For all farms in Dutch FADN accountancy data has been collected and for more than half of them also relevant technical, social and environmental data have been gathered yearly as well. The accountancy (or financial economic information) data is part of the legal obligation of the Netherlands to report yearly on the financial economic situation of farms to Brussels. This report contains the profit- and loss account, the balance sheet, the possible gathering of (EU) subsidies for each farm and some technical data like type and size of growing crops and/or the type of animals kept. The extra financial, technical, social and environmental data is gathered for about 67% of the 1500 available farms. This information is used for many national policy evaluations and research project such as Baltussen et al (2010).

A1.2 Farm selection

One of the main goals of the (Dutch) FADN is to be able to perform predictions/simulations for the primary agricultural sector or parts of the primary agricultural sector. Therefore the population in the dataset must be representative for the total number of agricultural holdings. This has been achieved by disproportional stratified random sampling out of the Agricultural Census (Vrolijk et al, 2010). The farms in the dataset are assigned with a weight due to the stratification process.

Although the target is to cover the agricultural population for 100%, this cannot be achieved. The selection of farms is restricted by their size in ESU and the share of income from outside the holding. The selected sample population is bounded by a lower threshold of the farm size (16 ESU) and an upper threshold of 2000 ESU (Vrolijk et al, 2010). These limitations are set to avoid small and extreme large farms and farms in the dataset. The Dutch FADN covers always at least 90% of the total production.

Farmers used to join the Dutch FADN for 5 years. After 5 years they were replaced by newly selected farmers. So each year 20% of the farmers were replaced. For this replacement new farmers have to be sampled and recruited. Through this rotating system the data in the Dutch FADN can be considered as panel data (Poppe, 2004). Since 2001/2002 this is no longer the case. Farmers can stay for more than five years in the FADN system. If they want to quit, they will be replaced by new selected farms.

The sampling is based on another dataset: the Annual Agricultural Census (CBS, 2012). This dataset contains structural variables of all farms in that year, like the grown crops, age of the farmer, composition of the livestock, and number of parcels and so on. The dataset does not contain economic variables. The sampling is executed according to a selection plan submitted to the European

Commission. The final recruitment is done taken strict privacy rules into account and is further described in Vrolijk et al (2010).

A1.3 Data in the dBase

Behind the Dutch FADN is a large database system (ARTIS) containing all (financial) facts happening on a farm. So per farm thousands of facts are recorded. The database can be accessed by different views. There are predefined views and researchers can make their own view. Two predefined views are mentioned here COBRA and microwave. COBRA generates standard data which can be used for all kind of projects. Microwave contains data especially composed for FES. Most data in microwave are processed COBRA data.

A1.4 Type of data

The data available in the Dutch FADN can be classified by subject Economic and financial issues, technical issues and issues concerning nature and environment. A detailed description can be found in Van der Veen et al (2006). All data presented in the next sections are available on a yearly base per farm.

An impression of the available data can also be found on LEI (2011).

Appendix 2 FES-model

A2.1 Introduction

The goal of the FES-model is to predict midterm financial economic development of specific farm types. The model calculates on micro (farm) level and aggregates the results to farm types and macro level (whole sector). The results at sector level are the final output of the FES-model. Results of the FES-model are averages of financial indicators and also distributions of financial indicators. Furthermore differences between firms with good and bad prospects are presented (van der Veen, 2011).

The scope of the first version of the FES-model was horticulture (glass houses) in the Netherlands (Mulder, 1991) but since then the model has been further developed (applicable for all farm types in the Netherlands) and became also applicable in 2010 for all farm types within EU25.

Depending on the scope of the research done with the model, FES uses data from the Dutch and/or the European FADN (van der Veen, 2011).

A2.2 Method

The outcome of the model is a set of financial ratios. With these financial ratios (per farm) it is possible to perform some predictions on short- and midterm continuity prospective of farms.

In the FES-model 7 types of midterm continuation prospects have been distinguished varying from poor future prospects till excellent prospects. The criteria for classifying farm continuity perspectives are:

- Liquidities
- Financial means
- Financial means after the necessary replacement investments
- Age of the farmer

The FES-model can make short term continuity prospects (table A2.1).

Table A2.1 Short term prospective

	Available liquidities	
	Sufficient	insufficient
Net cash flow +	Good	Not relevant
Net cash flow -	Sufficient	*delay of pay off can help -> sufficient * delay of pay off won't help -> insufficient

Source: Van der Veen (2011)

The model works with a reference scenario in which current developments influencing the financial situation of farms are defined. Besides the reference scenario other scenarios will have to be defined in which new policies are translated into changes in the parameters. In the next step the results of these scenarios are compared with the results of the reference scenario to determine the influence of policies on short and midterm financial situation.

Appendix 3 Economic size of farms

Source: European commission (2011).

Standard Gross Margins

The definition of the Standard Gross Margin is:

> "*The standard Gross Margin (SGM) of a crop or livestock item*
> *=*
> *the value of output from one hectare or from one animal*
> *-*
> *the cost of variable inputs required to produce that output*"

"*For each region all crop and livestock items are accorded an SGM. The Liaison Agencies calculate the SGMs themselves on the basis of empirical data collected from farms. To avoid bias caused by fluctuations, e.g. in production (due to bad weather) or in input/output prices, three year averages are taken. SGMs are expressed in Commission publications in European Currency (EUR/ECU).*

SGMs are updated every two years and are calculated on a regional basis for more than 90 separate crop and livestock items. This large number of items not only reflects the diversities of agriculture within the European Union but also indicates the level of detail that is required to ensure that the results of FADN and other surveys are both comprehensive and reliable.
In future, Standard Output (SO) will replace SGM in calculation of farm sizes. SO is monetary of the gross agricultural output at the farm-gate price."

European Size Units

"*Economic size of farms is expressed in terms of European Size Units (ESU). The value of one ESU is defined as a fixed number of EUR/ECU of Farm Gross Margin*."

"*There are five steps in the determining of farm size in ESU.*

1. *Identify the enterprises present on the farm*
2. *Determine the scale of each enterprise (hectares or number of animals)*
3. *Multiply the scale of each enterprise by the appropriate SGM to give the enterprise standard gross margin*
4. *Sum up the different enterprise standard gross margins for the farm. This gives the farm standard gross margin (i.e. the total of the enterprise standard gross margins for the farm)*
5. *Define the economic size of the farm by dividing the farm total gross margin by the value of the ESU*"

Appendix 4 Correlation between the attributes in the dataset

Table A4.1 Correlation coefficients between the attributes in the dataset(s)

	c1	*c2*	*c3*	*c4*	*c5*	*c6*	*c7*	*c8*	*c9*	*c10*	*c11*	*c12*	*c13*	*c14*	*c15*	*c16*	*c17*	*c18*	*c19*	*c20*	*c21*	*c22*	*c23*	*c24*
c1	1.00																							
c2	0.16	1.00																						
c3	-0.03	-0.19	1.00																					
c4	0.36	0.31	-0.08	1.00																				
c5	0.04	0.00	0.05	-0.37	1.00																			
c6	0.29	0.28	-0.07	0.37	0.05	1.00																		
c7	0.09	0.00	0.01	-0.19	0.11	-0.03	1.00																	
c8	0.11	0.14	0.00	0.30	0.02	0.17	-0.50	1.00																
c9	0.49	0.39	-0.03	0.75	0.05	0.46	-0.13	0.38	1.00															
c10	0.21	0.23	0.00	0.47	-0.01	0.24	-0.27	0.57	0.57	1.00														
c11	-0.06	0.01	0.00	0.01	-0.13	-0.03	0.15	-0.40	-0.03	0.04	1.00													
c12	0.87	0.22	-0.02	0.43	0.09	0.33	0.12	0.09	0.59	0.14	-0.04	1.00												
c13	0.84	0.20	-0.03	0.56	-0.03	0.31	-0.06	0.30	0.64	0.46	-0.03	0.75	1.00											
c14	-0.10	0.01	-0.04	-0.13	0.04	0.01	-0.06	0.10	-0.13	-0.21	-0.04	-0.09	-0.12	1.00										
c15	0.00	-0.01	-0.01	-0.03	0.00	0.01	-0.01	0.01	-0.04	-0.01	-0.01	-0.02	0.00	0.02	1.00									
c16	-0.37	-0.14	0.00	-0.20	-0.07	-0.13	-0.07	-0.11	-0.32	-0.14	0.09	-0.48	-0.34	0.05	0.01	1.00								
c17	0.12	-0.03	-0.01	0.04	-0.01	0.01	0.00	-0.07	0.00	-0.06	0.05	0.10	0.05	-0.02	-0.01	0.77	1.00							
c18	0.20	0.12	0.07	0.09	0.01	0.10	0.10	0.11	0.14	0.11	-0.03	0.29	0.21	-0.03	0.01	-0.24	-0.10	1.00						
c19	-0.07	-0.09	0.01	-0.06	0.05	-0.09	-0.01	-0.02	-0.08	-0.07	0.03	-0.08	-0.07	-0.02	-0.02	0.07	0.02	0.18	1.00					
c20	0.48	0.40	-0.04	0.74	0.02	0.44	-0.11	0.36	0.97	0.56	-0.03	0.58	0.64	-0.12	-0.03	-0.31	0.01	0.16	-0.15	1.00				
c21	0.38	0.39	-0.06	0.63	0.00	0.47	-0.01	0.24	0.77	0.37	0.02	0.48	0.46	-0.07	-0.02	-0.24	0.01	0.22	-0.03	0.76	1.00			
c22	0.28	0.39	-0.01	0.40	0.00	0.45	-0.03	0.18	0.66	0.30	0.00	0.33	0.33	-0.02	-0.03	-0.16	0.01	0.23	-0.11	0.68	0.63	1.00		
c23	0.40	0.49	-0.08	0.66	-0.02	0.45	0.01	0.24	0.78	0.38	0.01	0.49	0.49	-0.06	-0.01	-0.25	0.00	0.22	-0.05	0.79	0.95	0.64	1.00	
c24	-0.02	0.04	0.06	0.00	-0.03	0.02	-0.01	0.05	0.04	0.08	0.02	0.00	0.00	-0.04	0.00	-0.07	-0.09	0.56	0.23	0.04	0.08	0.03	0.06	1.00

c1	YieldPerNormalizedWorker
c2	TotalInvestments
c3	PercInvestmMachinery
c4	AreaProperty
c5	FractionRent(land)
c6	subsidies
c7	NoSuccessor
c8	number of entrepreneurs
c9	ESU
c10	TotalHours
c11	FractionLaborOthers
c12	RevenewsCostFraction
c13	Income
c14	FarmIncomeOutside
c15	FracOutsideIncome
c16	Costsper100kgmilk
c17	Revenewsper100kgmilk
c18	MilkprodperCow
c19	YoungAnimalsPerCow
c20	Number of dairycows
c21	InterestPaid
c22	CostsFodder
c23	LongLoans
c24	ConcentratesPerCow

Appendix 5 Panel data models

A5.1 Theoretical background

Source: Reinhard et al (2001)

The datasets we use are organized as panel data. Panel data is a combination of cross-section and time series datasets. Panel data contains data for a number of observation units for several time units. Our datasets consists of observations for farms for more than one year.
In balanced datasets for all firms for all years there are observations. In case of unbalanced datasets this is not holding. Our datasets are unbalanced panel datasets. The model for panel data can be formulated as:

$$Y_{it} = \alpha + X'_{it}\beta + v_{it}$$

With

Y_{it} = observation of farm i at year t
X_{it} = the value of the independent variables for farm i at year t
v_{it} = noise term for farm i and year t

The noise term should be normally distributed N(0, $\sigma^2 v$).

Summarizing: the panel data model consists of a regression $\alpha + X'_{it}\beta$ and effects for year and farm v_{it}.

The noise term v_{it} exists of a farm effect (u_i) and a year effect (w_t), therefore we can talk about a Two-way Error Component model. Because most variables correlate with year we have also a fixed effects model. We can write

$$v_{it} = u_i + w_t + \epsilon_{it}$$

With

u_i = farm effect
w_t = year effect
ϵ_{it} = error term

The error term should be normally distributed N(0, $\sigma^2 v$).

MLR with OLS does not take heterogeneity across groups or time into account (Torres-Reyna, 2009).

A5.2 Panel data models applied to predicting change in farmsize (ESU)

In this section we present an estimated panel data model for a prediction period of four years. Individual farm effects are not shown due to privacy reasons.

Results

```
summary(fixedplm1b)
Twoways effects Within Model

Call:
plm(formula = ESUChange ~ PercInvestmMachinery + subsidies +
    ESU + LongLoans + opbrengstKostenverh + costsPer100kgmilk +
    revenieuwsper100kgmilk + dcow, data = Pdata, effect = "twoways",
    index = c("BinNummer", "year"), type = "within")
```

```
Unbalanced Panel: n=267, T=1-5, N=1012

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -91.40   -4.53    0.00    4.63   97.60

Coefficients :
                        Estimate Std. Error  t-value  Pr(>|t|)
PercInvestmMachinery    3.698306   1.456093   2.5399 0.0112943 *
subsidies              -1.832724   0.930656  -1.9693 0.0492971 *
ESU                    -1.351417   0.093570 -14.4428 < 2.2e-16 ***
LongLoans               0.261015   0.035726   7.3060 7.203e-13 ***
opbrengstKostenverh    -0.468021   0.134991  -3.4671 0.0005569 ***
costsPer100kgmilk      -0.521307   0.176012  -2.9618 0.0031576 **
revenieuwsper100kgmilk  0.841834   0.246670   3.4128 0.0006783 ***
dcow                    1.266787   0.156344   8.1025 2.250e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    200450
Residual Sum of Squares: 138680
R-Squared      :  0.30816
      Adj. R-Squared :  0.2232
F-statistic: 40.8116 on 8 and 733 DF, p-value: < 2.22e-16
> summary(fixef(fixedplm1b,effect="time"))
     Estimate Std. Error t-value  Pr(>|t|)
2001  105.262     12.675  8.3045 < 2.2e-16 ***
2002  106.924     12.323  8.6770 < 2.2e-16 ***
2003  109.868     12.284  8.9438 < 2.2e-16 ***
2004  115.018     12.476  9.2194 < 2.2e-16 ***
2005  116.956     12.748  9.1744 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments

Compared with the MLR-model for a prediction period of four years (no normalizing of the dependent variable), fewer variables are significantly contributing to the regression part of the panel data model. Also some different variables are now selected like subsidies and costs en revenues per 100 kg milk.

The R^2 of the estimated regression line is with panel data analysis lower. Compared with MLR in panel data analyses part of the dependent variable estimation is covered by year and farm effects.

Baltagi (2005) and Torres-Reyna (2009) describe a lot of tests on panel data analysis. It is for instance possible to calculate the noise term and test for normality and equal variance. It is also possible to test if it is really a fixed effect or random effect model and so on.

A5.4 Predicting with panel data models

Baltagi (2008) reports that there is little literature available on forecasting with panel data.

Shami et al (2011) compared results of panel data models with MLR and two types of neural networks. One NN was fed with structured and balanced panel data, the other not. Both NN performed well but fed with balanced panel data performed slightly better. They also made a prediction for one year ahead. They used a dataset for the years 2007-2009 and predicted for 2010 based on 2009.

A5.5 Relation with main text

We decided not to use panel data models in this research because:

We want to use estimated models in the FES-model. So therefore the model should also cover for new years and/or farms in the dataset. If we would hold on to the fixed effect model we wouldn't have effects for new years and farms. If the effects of year and farm would be random then we could use this random factor. If we want to know if these effects are random or not we have to apply several test. After some literature study we observed that estimating the panel data models is a sophisticated way, is a time consuming task asking using expert knowledge. This action would be out of the scope of the BMI-projects.

Appendix 6 Coefficients of selected variables in the MLR models for different prediction years

Table A6.1 Coefficients of variables in the MLR-models for different number of prediction years

	Prediction period (years)							
Attributes	1	2	3	4	5	6	7	8
YieldPerNormalizedWorker			-0.0791	-0.0914	-0.1586	-0.1761		
TotalInvestments	0.0069	0.0064	0.0072	0.0087	0.0079	0.0098	0.0101	
PercInvestmMachinery			0.1413	0.2440	0.1782	0.2497		
AreaProperty	-0.0047	-0.0059	-0.0070	-0.0043	-0.0065			
FractionRent(land)								
subsidies		0.0534	0.0562					
NoSuccessor	0.1155							
number of entrepreneurs	0.0914		0.0724	0.1342	0.1163			
ESU	-0.0076	-0.0098	-0.0135	-0.0161	-0.0119		-0.0107	
TotalHours	-0.0042							
FractionLaborOthers								
RevenewsCostFraction			0.0169	0.0101	0.0186	0.0201		
Income	0.0360	0.0205	0.0311	0.0245	0.0392	0.0588		
FarmIncomeOutside				-0.0534	-0.0871	-0.1083	-0.1105	
FracOutsideIncome								
Costsper100kgmilk	0.0028		0.0137					
Revenewsper100kgmilk			-0.0196					
MilkprodperCow		0.0037	0.0083	0.0073				
YoungAnimalsPerCow						-0.6362		
Number of dairycows	0.0116	0.0138	0.0198	0.0219	0.0137		0.0134	
InterestPaid		-0.0697	-0.1190	-0.2210	-0.2074	-0.1256		-0.2179
CostsFodder	0.0404	0.1031	0.0809	0.1370	0.2343	0.3381	0.3446	0.2931
LongLoans	0.0031	0.0079	0.0105	0.0139	0.0136	0.0096	0.0039	0.0136
ConcentratesPerCow			-0.0202	-0.0219	-0.0163		-0.0134	
Intercept	-0.4426	-0.4865	-1.7351	-0.9951	-1.0076	-1.0027		

Appendix 7 Check on assumptions MLR

The use of (multiple) linear regression techniques is in principle limited by several restrictions which also depend on the way the weights are calculated. Most important restrictions are (Osborne and Waters, 2002):

1) The dependent variable y is normally distributed
2) The variables are normally distributed
3) There's a linear relationship between the dependent and independent variables
4) The errors are normally distributed ~N(0,σ^2)
5) Homoscedasticity of the errors (constant variance)

We checked whether the assumptions for using MLR do hold for a prediction of one, four and seven years.

The general assumption of independent observation cannot be held because in general we have more than one observation per farm.

The independent variables are on average not normally distributed (see section 4.3).

We tested with the Shapiro-Wilk test if the observed dependent values are normally distributed. The null-hypothesis is that the dependent variable is normally distributed and the alternative hypothesis is that the dependent variable is not normally distributed. We did the same for the errors (table A7.1).

Table A7.1 Tests on normality for the distribution of the dependent variable and the errors with Shapiro-Wilk test

Prediction years	Attribute	W-statistic	p-value
1	Y	0.8359	< 2.2e-16
1	ϵ	0.3890	< 2.2e-16
4	Y	0.9056	< 2.2e-16
4	ϵ	0.7818	< 2.2e-16
7	Y	0.7452	< 2.2e-16
7	ϵ	0.9191	< 4.3e-13

Appendix 8 Weight factors Neural net with two hidden layers

Hidden 1
========

Node 1 (Sigmoid)
----
tMainDataSet2_ESU: -2.400
arboprperaje: 0.934
totInvestments: -0.230
PercInvestmMachinery: 0.214
haEigendom: -1.598
nr_entrepreneurs: -0.464
opbrengstKostenverh: -0.085
income: 1.762
farmincomeOutside: 0.197
milkprodpercow: 0.786
dcow: 1.600
rentpaid: -0.955
CostsFodder: 1.497
LongLoans: 2.098
kgkrvperkoe: -0.010
Threshold: -1.207

Node 2 (Sigmoid)
----
tMainDataSet2_ESU: -1.457
arboprperaje: -2.853
totInvestments: 1.371
PercInvestmMachinery: 0.111
haEigendom: 0.688
nr_entrepreneurs: -0.207
opbrengstKostenverh: 0.337
income: -2.272
farmincomeOutside: -0.188
milkprodpercow: 1.907
dcow: 2.879
rentpaid: -0.607
CostsFodder: 0.207
LongLoans: 1.015
kgkrvperkoe: -0.645
Threshold: -1.413

Node 3 (Sigmoid)
----
tMainDataSet2_ESU: -1.243
arboprperaje: -0.097
totInvestments: -0.064

PercInvestmMachinery: 0.210
haEigendom: 0.798
nr_entrepreneurs: 0.326
opbrengstKostenverh: 0.019
income: 1.154
farmincomeOutside: -0.186
milkprodpercow: -0.108
dcow: 0.289
rentpaid: -0.665
CostsFodder: -0.560
LongLoans: 1.577
kgkrvperkoe: 0.475
Threshold: -0.896

Node 4 (Sigmoid)
---
tMainDataSet2_ESU: -1.852
arboprperaje: 0.861
totInvestments: -1.880
PercInvestmMachinery: -0.577
haEigendom: 0.419
nr_entrepreneurs: -1.293
opbrengstKostenverh: 1.724
income: -0.669
farmincomeOutside: 0.824
milkprodpercow: -2.516
dcow: -0.564
rentpaid: -3.691
CostsFodder: -0.523
LongLoans: 0.899
kgkrvperkoe: -0.302
Threshold: 0.850

Node 5 (Sigmoid)
---
tMainDataSet2_ESU: -1.191
arboprperaje: -0.992
totInvestments: 1.610
PercInvestmMachinery: 0.274
haEigendom: -0.413
nr_entrepreneurs: 0.725
opbrengstKostenverh: -0.856
income: -0.198
farmincomeOutside: -0.302
milkprodpercow: 0.308
dcow: 2.198
rentpaid: -0.144
CostsFodder: 0.923

LongLoans: 2.224
kgkrvperkoe: 0.057
Threshold: -1.100

Node 6 (Sigmoid)
---
tMainDataSet2_ESU: -1.353
arboprperaje: -0.610
totInvestments: 0.807
PercInvestmMachinery: 0.127
haEigendom: -0.149
nr_entrepreneurs: 0.451
opbrengstKostenverh: -0.332
income: -0.004
farmincomeOutside: -0.102
milkprodpercow: 0.666
dcow: 1.830
rentpaid: -0.418
CostsFodder: 0.767
LongLoans: 1.929
kgkrvperkoe: 0.129
Threshold: -1.028

Node 7 (Sigmoid)
---
tMainDataSet2_ESU: -0.976
arboprperaje: -0.796
totInvestments: 0.939
PercInvestmMachinery: 0.127
haEigendom: -0.467
nr_entrepreneurs: 0.496
opbrengstKostenverh: -0.557
income: 0.377
farmincomeOutside: 0.054
milkprodpercow: -0.243
dcow: 2.070
rentpaid: -0.115
CostsFodder: 0.636
LongLoans: 2.246
kgkrvperkoe: -0.533
Threshold: -1.418

Node 8 (Sigmoid)
---
tMainDataSet2_ESU: -1.829
arboprperaje: -1.027
totInvestments: 1.373
PercInvestmMachinery: 0.818

haEigendom: -4.088
nr_entrepreneurs: 0.992
opbrengstKostenverh: -0.887
income: 1.204
farmincomeOutside: -1.006
milkprodpercow: -1.443
dcow: 4.478
rentpaid: 1.244
CostsFodder: 0.871
LongLoans: 2.452
kgkrvperkoe: -3.267
Threshold: -0.930

Node 9 (Sigmoid)
---
tMainDataSet2_ESU: -1.161
arboprperaje: -0.123
totInvestments: -0.264
PercInvestmMachinery: -0.409
haEigendom: 0.980
nr_entrepreneurs: 0.174
opbrengstKostenverh: 0.004
income: 1.550
farmincomeOutside: 0.188
milkprodpercow: -0.475
dcow: 0.711
rentpaid: -1.159
CostsFodder: -1.262
LongLoans: 1.550
kgkrvperkoe: 0.052
Threshold: -0.901

Node 10 (Sigmoid)
---
tMainDataSet2_ESU: -1.054
arboprperaje: -1.050
totInvestments: 3.314
PercInvestmMachinery: 2.012
haEigendom: -0.629
nr_entrepreneurs: 2.179
opbrengstKostenverh: -2.294
income: 0.008
farmincomeOutside: -1.307
milkprodpercow: 1.061
dcow: 3.291
rentpaid: 0.876
CostsFodder: 3.098
LongLoans: 1.731

kgkrvperkoe: 2.442
Threshold: -1.289

Hidden 2
========

Node 1 (Sigmoid)

Node 1: -0.754
Node 2: -0.807
Node 3: -0.335
Node 4: -1.442
Node 5: -1.029
Node 6: -0.688
Node 7: -0.788
Node 8: -1.244
Node 9: -0.342
Node 10: -2.128
Threshold: -0.691

Node 2 (Sigmoid)

Node 1: -0.830
Node 2: -0.853
Node 3: -0.423
Node 4: -1.439
Node 5: -0.960
Node 6: -0.663
Node 7: -0.756
Node 8: -1.246
Node 9: -0.388
Node 10: -2.148
Threshold: -0.659

Node 3 (Sigmoid)

Node 1: -0.789
Node 2: -0.862
Node 3: -0.312
Node 4: -1.496
Node 5: -1.040
Node 6: -0.692
Node 7: -0.716
Node 8: -1.236
Node 9: -0.280
Node 10: -2.149
Threshold: -0.649

Node 4 (Sigmoid)

Node 1: -0.828
Node 2: -0.845
Node 3: -0.485
Node 4: -1.363
Node 5: -0.976
Node 6: -0.652
Node 7: -0.806
Node 8: -1.253
Node 9: -0.472
Node 10: -2.087
Threshold: -0.697

Node 5 (Sigmoid)

Node 1: -0.794
Node 2: -0.793
Node 3: -0.433
Node 4: -1.347
Node 5: -1.040
Node 6: -0.665
Node 7: -0.754
Node 8: -1.251
Node 9: -0.432
Node 10: -2.091
Threshold: -0.739

Output
======

Regression (Linear)

Node 1: -2.192
Node 2: -2.194
Node 3: -2.203
Node 4: -2.174
Node 5: -2.173
Threshold: 0.516