Estimation of Commodity Specific Production Costs Using German Farm Accountancy Data

AUTHORS
Sirak Bahta, Anja Berner and Frank Offermann

Paper prepared for presentation at the EAAE 2011 Congress
Change and Uncertainty
Challenges for Agriculture,
Food and Natural Resources

August 30 to September 2, 2011
ETH Zurich, Zurich, Switzerland

ESTIMATION OF COMMODITY SPECIFIC PRODUCTION COSTS USING GERMAN FARM ACCOUNTANCY DATA

Sirak Bahta[1]., Anja Berner[2] and Frank Offermann[3]

Abstract

A central problem in estimating per unit costs of production originates from the fact that most farms produce multiple outputs and standard farm-accounting data are only available at the whole-farm level. The seemingly unrelated regression (SUR) approach is used to estimate per unit production costs based on German farm accountancy data. Special emphasis is put on outlier detection prior to the estimation of production costs to increase the robustness of the results. Outlier observations are identified based on the Mahalanobis distance for each observation on the data set. It was observed that less negative cost coefficients are estimated after the exclusion of the outliers.

The time series analysis of cost estimation based on SUR regression shows the costs of arable crops after 2004, affected by rising prices of fertilizer, seeds and energy, while the increase of livestock production costs after 2006 is attributed to feed costs.

Keywords

Multi-output, outlier detection, production costs, Seemingly Unrelated Regression

[1] Institute of Farm Economics, Johann Heinrich von Thünen-Federal Research Institute for Rural Areas, Forestry and Fisheries, Bundesallee 50, D-38116 Braunschweig, Germany. Email: sirak.bahta@vti.bund.de.

[2] Institute of Farm Economics, Johann Heinrich von Thünen-Federal Research Institute for Rural Areas, Forestry and Fisheries, Braunschweig, Germany.

[3] Institute of Farm Economics, Johann Heinrich von Thünen-Federal Research Institute for Rural Areas, Forestry and Fisheries, Braunschweig, Germany

1. Introduction

Surveys often measure variable inputs only in value terms as total expenditure on each item, but allocations of individual fixed and variable inputs to particular enterprises are not recorded. Nevertheless, this commodity or enterprise specific information concerning the allocation of inputs, called cost allocation approach, is very important and can serve for farm planning and agricultural policy making (PEETERS and SURRY, 2003).

Various[4] researchers have been addressing the issue of input-output allocation or cost allocation using farm accountancy data. ERRINGTON (1989) proposed an ordinary Least Square (OLS) technique applied to a system of derived demand equations, where the total demand for a given input is treated as a function of output value of each enterprise. As a comment on ERRINGTON's work, MIDMORE (1990) raised several analytical concerns. Among other things, he noticed the occurrence of heteroscedasticity and remarked that due to the accounting constraint the disturbance terms are not independent, which leads to the invalidity of the OLS technique.

Later in 1992, ERRINGTON again provides a detailed study of the nature of the estimation problems and the various remedies that have been proposed (and often failed) to deal with those problems. HALLAM et al. (1999) revisit these arguments again and find that with the use of a number of panel data estimation techniques, some of the OLS's shortcomings can be overcome, even though practical and methodological problems remain. The use of panel data provides a means of monitoring for the individual farm effects (there are substantial variations among farms in relationships between inputs and outputs as a result of farm specific factors such as land quality and managerial ability) and year effects (substantial year to year variations depending on weather).

BUTAULT and CYNCYNATUS (1990) use the French FADN to estimate enterprise costs of production using the INRA/INSEE/SCEES model. They compare the performance of several single equations and systems of equation techniques for estimating the coefficients.

A number of constrained estimation methods to allocate total input expenditure between the individual enterprises were then proposed by MOXEY and TIFFIN (1994). They move into a Bayesian framework and apply the Inequality Restricted Least Squares method to ensure the non-negativity of the estimated production coefficients.

LEON et al (1999), propose the use of Generalized Maximum Entropy (GME) to estimate input-output or production coefficients. More specifically they lay out how GME can be used to deal with problems encountered earlier, such as singularity, constrained estimation and zero-observations. The study also compares the GME results with those through OLS, Bayesian and LP. However, in their conclusion, the authors admit sensitivity of parameter estimates to the support values and thus suggest further examination of the connection of GME with the Bayesian approach. Also, the problem of heteroscedasticity is still an issue.

PEETERS and SURRY (2003) explore the methodology of the LEON et al. (1999) using a data set from a sample of Saskatchewan crop farms. Even though their findings fall within the expected range, they conclude that more homogenous crop categories could improve the results and discrepancies observed.

In line with the various efforts to measure enterprise specific production costs, this study uses a set of input equations that may or may not have similar independent variables or farm outputs. This method, which is known as the SUR (seemingly unrelated regression) method, allows for correlated errors between the set of equations which improves the efficiency of the estimation. Prior to estimation of input-output equations, a multivariate outlier detection method is applied to avoid possible impact of outlier observations on the estimated cost coefficients.

This study forms part of the largest Farm Accountancy Cost Estimation and Policy Analysis of European Agriculture (FACEPA) to estimate enterprise specific farm production costs. And it aims at estimating a time series product or enterprise specific farm production costs and exploring the effect of the outlier observations on the estimated costs.

The paper is structured as follows: first, the underlying data set and methodologies for outlier detection are discussed. Then, the SUR method will be introduced followed by a comparison of time series production costs for wheat and barley, before and after the removal of outlier observations. Later, the time series production cost results based on the sample that excludes outlier observations are discussed. The paper closes with a summary and the main conclusions.

---

[4] BOONE and WISEMAN (1998) compared pig cost prices across EU using data obtained from the Farm Accountancy Data Network (FADN).
.PINGAULT and DESBOIS (2003) estimated costs of production of key agricultural products from FADN data.
NGUYEN, MCLAREN and ZHAO (2008) estimated a cost function using quasi-micro farm level data for Australia.
MACK and MANN (2008) estimated marginal cost functions for Swiss dairy production based on FADN Data.

2 Data and Methodology

Data from the German Farm Accountancy Data Network is used for this study. The sample consists about 11,000 to12,000 farms per year starting from 1995[5] to 2008, and only full time farms and farms greater than or equal to 16 ESU[6] (Economic size unit) are considered for this study. The production cost analysis includes 16 aggregated input categories, including subsidies (defined as negative input) and net value added, as well as 31 output categories. Prior to estimation of the production cost model, outlier observations are detected using a multivariate outlier detection method and excluded from the time series production cost analysis. The multivariate outlier detection method followed by the production cost model is subsequently discussed.

2.1 Outlier elimination

Most real-world data sets often contain outlier observations which appear to be inconsistent with the remainder of the data set or which have an unusually large or small value when compared with others in the data set. This could be due to some observations having different characteristics regarding a specific variable, or due to measurement errors (ESCALANTE, 2005).

In many data analysis tasks a large number of variables are recorded or sampled. One of the first steps towards obtaining a coherent analysis is the detection of outlaying observations. Detected outliers are candidates for aberrant data that may otherwise adversely lead to model misspecification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modelling and analysis (WILLIAMS et al., 2002; LIU et al., 2004).

Statistical methods for multivariate outlier detection often indicate those observations that are located relatively far from the centre of the data distribution. Several distance measures can be implemented for such a task. One of these distance measures is the Mahalanobis distance. The Mahalanobis distance depends on estimated parameters of the multivariate distribution (BEN-GAL, 2005). In this case, a multivariate outlier can be defined as a case with a large Mahalanobis distance. On the chi-square probability plot, these would appear as points in the upper right that are substantially above the line for the expected chi-square quantiles (FRIENDLY, 2008). The Mahalanobis distance is calculated as:

$$d_i^2 = (x_i - \bar{x})S^{-1}(x_i - \bar{x}) \quad \textbf{(1)}$$

Where

x_i is the i^{th} observation of the vector $x_{i1}, x_{i2}.....x_{ip}$ and the mean vector c for the total sample

S is the $p \times p$ sample variance-covariance matrix and

d^2 is the multivariate analogue of the square of the standard score of a single variable, $z_i = (xi - \bar{x})/S$ which measures the distance from the mean in standard deviation units and z_i^2 is distributed as x^2.

In this case, a few discrepant observations not only affect the mean vector, but also inflate the variance covariance matrix. To address this issue, CAUSSINUS and ROIZ (1990) propose a robust estimate for the covariance matrix, which is based on weighted observations according to their distance from the center. They also propose a method for low dimensional projections of the dataset and used Generalized Principle Component Analysis (GPCA) to reveal those dimensions which display outliers. Furthermore, HADI (1992) also addresses the problem by replacing the mean vector with a vector of variable medians to compute the covariance matrix for the subset of those observations with the smallest Mahalanobis distance. One reasonably general solution, however, is the use of multivariate trimming to calculate squared distances (FRIENDLY 1991). This is an iterative process where, on each iteration, some proportion of the observations with the largest d^2 values are temporary set aside, and the trimmed mean, x_i and trimmed variance-covariance matrix, S_i, are computed from the remaining observations. The new D^2 values are then computed using the robust mean and covariance matrix as

$$d_i^2 = (x_i - \bar{x}_{(i)})' S_{(i)}^{-1}(x_i - \bar{x}_{(i)}) \quad \textbf{(2)}$$

---

[5] 1995 refers to the economic year 1995/96, etc.

[6] ESU is defined as a fixed number of EUR/ECU of Farm Gross Margin. Over time the number of EUR/ECU per ESU has changed to reflect inflation. Refer: http://ec.europa.eu/agriculture/rica/methodology1_en.cfm

Therefore, the effect of a few extreme observations will spread through all the d^2 values and observations with large distances do not contribute to the calculations for the remaining observations. One way to carry out this process in the SAS system has been suggested by YOUNG and SARLE (1989). This process also avoids the necessity of calculating the inverse of the variance-covariance matrix by transforming the data to standardized principal component scores.

Let z_i be the vector of standardized principal component scores corresponding to x_i, then the squared distance is just the sum of squares of the elements in z_i as is described in Equation 3.

$$D_i^2 = z_j' z_i = \sum_{j=1}^{p} z_{ij}^2 \quad \textbf{(3)}$$

Based on the idea of YOUNG and SARLE (1989), a SAS outlier macro[7] has been written by Friendly (2008) which identifies points with the extreme squared distance of Mahalanobis and a chi square value of less than 5 % as an outlier observations. This SAS macro program for outlier detection is modified for this study and applied to the German FADN data set.

2.2 Production cost model

To estimate the cost-allocation coefficients from farm accounting data, a set of linear equations is considered where the derived demand from farm f for each input i is represented as a function of several outputs k (PEETERS and SURRY 2003):

The relevant microeconomic unit is assumed to be the professional farm holding, therefore the model derives the empirical estimates from the FADN statistical database. The output of the various products is denoted as $y_k(k = 1,.....,k)$ and the, $x_i(i = 1,....., I)$ representing the non-allocated costs of the production factors.

Assuming *i* inputs used by *f* farms to produce *k* outputs, the set of equations can be written as (PEETERS and SURRY 2003):

$$x_{if} = \sum_{k=1}^{K} \beta_{ik} y_{kf} + u_{if} \,, \quad \textbf{(4)}$$

where x_{if} is the total cost of input *i* paid by farm *f* (including subsidies and net value added),

y_{kf} is the total value of output *k* produced by farm *f*,

β_{ik} is the unknown technical production coefficient, which is defined as the average (for all farms) expenditure on input i required to produce one unit of output value *k*,

u_{if} is the error term specific to each input and farm.

On each farm *f*, the observed costs in input *i* differ from the theoretical costs by a random factor *ujf* of zero expectation and independent from one farm to the next. This means that the use of input *i* by a given farm is not affected by another farm use of the same input.

In order to achieve the accounting consistency of the model, we have to introduce the constraint that the sum of output values equals the sum of input costs plus net value added the model is estimated subject to:

$$\sum_{k=1}^{K} \beta_{ik} = 1 \quad \textbf{(5)}$$

This equation ensures that the production coefficients add up to one.

To estimate the model, various techniques can be applied. For example, LEON et al. (1999) use OLS regression, Bayesian, Generalised Maximum Entropy, and Linear Programming approaches. In line with the work of POLLET et al. (2001), here, the model is estimated based on the so-called seemingly unrelated regression (SUR).[8]

The method described is applied to a set of accounting data for Germany and it considers 16 aggregated input categories, including subsidies and net value added, as well as 31 output categories. The subsidies enter the

[7] The SAS outlier macro is available at http://www.math.yorku.ca/SCS/sssg/outlier.html

[8] This method is implemented employing the PROC SYSLIN Procedure in SAS (see SAS 9.2 Users Guide under http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/syslin_toc.htm).

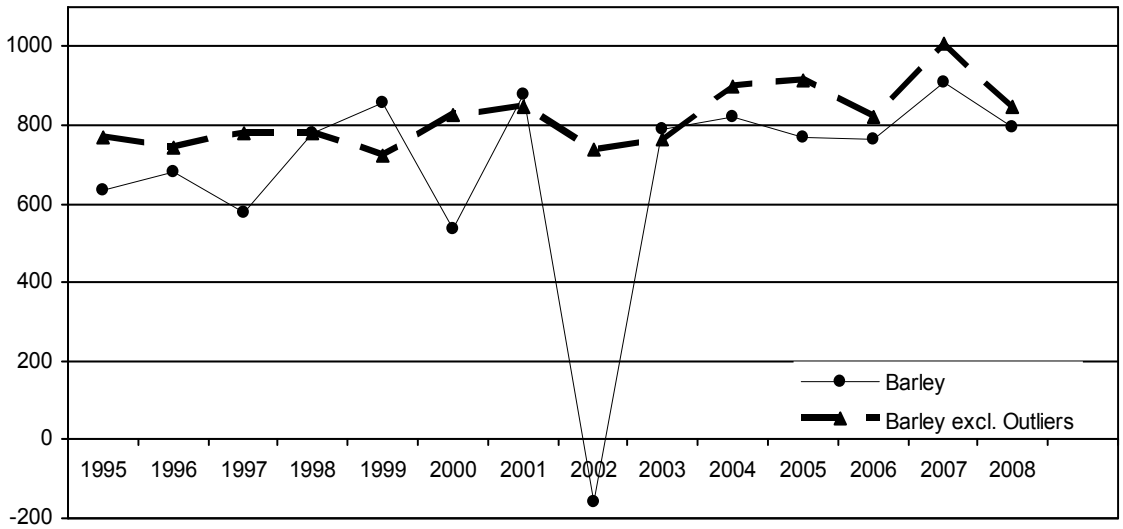
model as an independent variable with negative values. Thus, it is possible to derive the average amount of subsidies associated with the production of one unit of output value k. The net value added is composed of the sum of output value plus subsidies minus input costs. Using the aforementioned nomenclature, this relation can be written as:

$$\text{Net value added}_{\,f} = \sum_{k=1}^{K} y_{kf} - \sum_{i=1}^{I-1} x_{if} \qquad \textbf{(6)}$$

3 Results

Before the OUTLIER macro is applied for German FADN data, the output variables are converted to monetary value per hectare and LU values for crop and livestock outputs; respectively. Similarly, seed, fertilizer, crop protection costs are changed to a value per total crop output; feed and veterinary cost to a value per total livestock output and the rest (except subsidy and net value added which are not included in the outlier detection) to a value per total farm output. To demonstrate the effect of outlier observations on the production cost estimation, a comparison of cost estimates before and after the exclusion of outliers, for wheat and barley using the 1995-2008 data, are presented in Figure 1 and Figure 2, respectively.

Figure 1: Production cost of wheat with and without outliers (1995-2008)

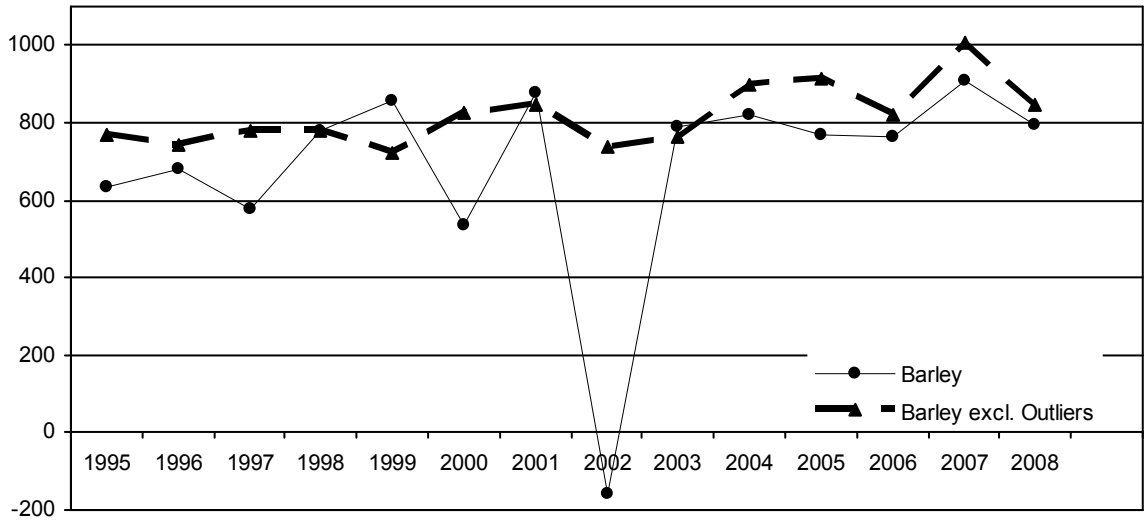


Source: BMELV-Testbetriebe and own calculations.

It can be seen that before elimination of the outliers, the production cost of wheat is underestimated for most of the years under consideration. After elimination of the outlier observations; however, there exist plausible values of production costs (Figure 1).

Furthermore, the production cost of barley before the elimination of outlier observations shows high fluctuations, and in some cases (for the year 2002) the total cost is negative. But after the removal of outlier observations, the costs become more or less stable in the range of 700 Euro/ha to 1100 Euro/ha. These two examples of wheat and barley show the extent of the impact of outliers on their estimated value of total production cost.

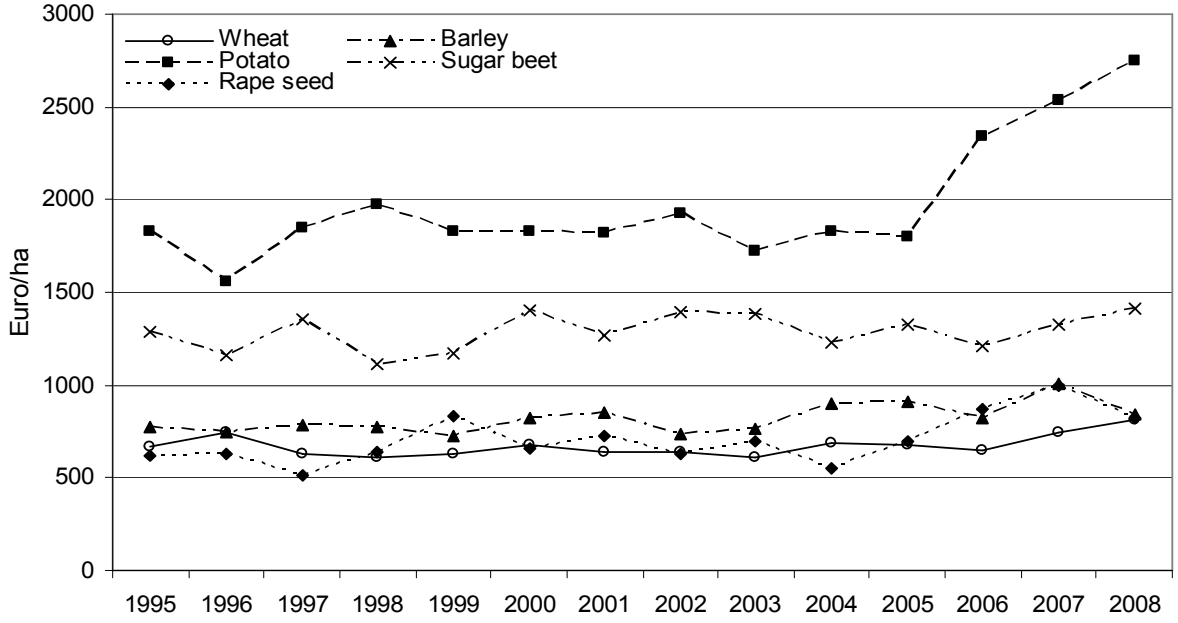
Figure 2: Production cost of barley with and without outliers (1995-2008)



Source: BMELV-Testbetriebe and own calculations.

The following figures show the development of the production costs of the different products over time after the outliers have been eliminated. In Figure 3 the production costs per hectare for wheat, barley, potato, rape seed and sugar beet are illustrated. It can be seen that wheat and barley show almost the same development, with slightly higher costs for barley.

Figure 3: Production costs arable crops per hectare



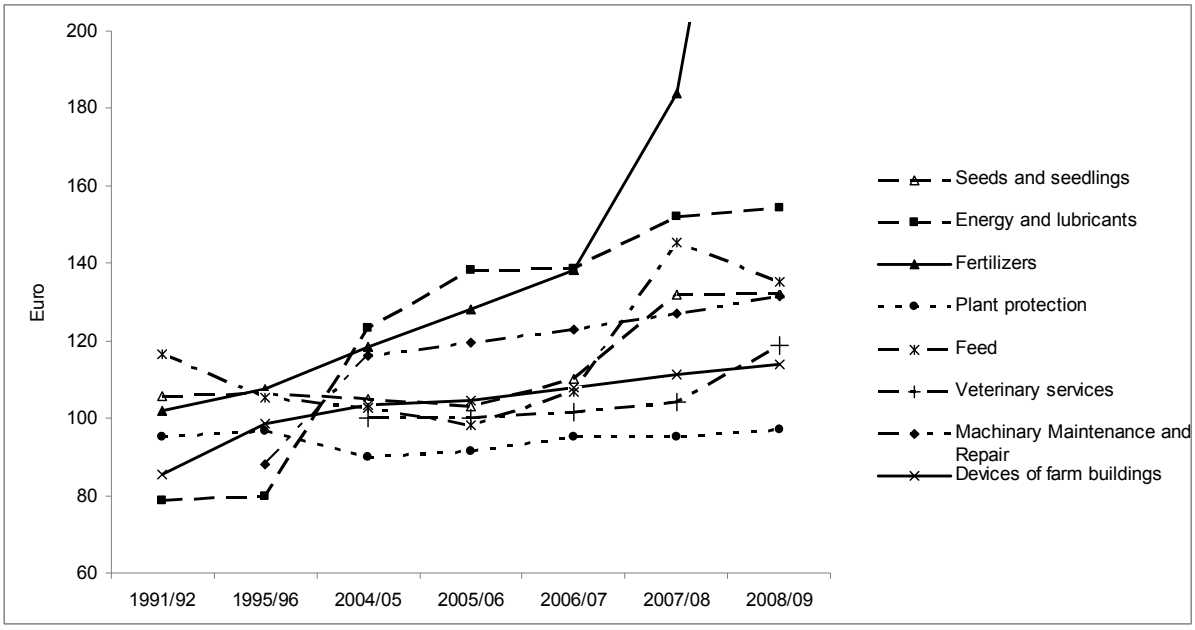
Source: BMELV-Testbetriebe and own calculations.

The costs were rather stable within the range of 500-1000 €/hectare. The production costs of rape seeds were more or less stable until 2004, and then increased until 2007, followed by a slight decrease in 2008.

The costs of sugar beet remained, however, constant within a range of 1000 to 1500 €/hectare. The production costs of potato were within the range of 1500-2000 €/hectare only until 2005, and then constantly increased and reach to production cost of more than 2500 Euro/hectare in 2008. This rise of production costs, specifically after

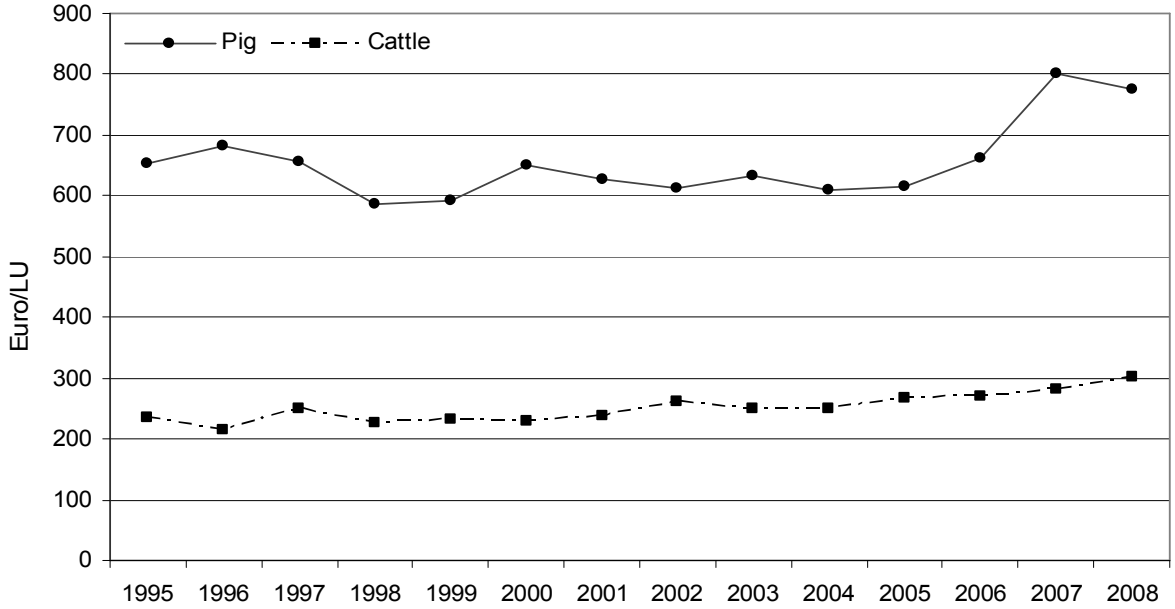
2004, is caused by high input costs. This finding is well supported by the development of input prices throughout these years (see Figure 4) where prices of fertilizers, seeds and oil and electricity increased constantly overtime.

Figure 4: Input prices trend



Source: BMEL Statistics and own calculations.

Figure 5: Production costs for beef and pigs per Livestock Unit



Source: BMELV-Testbetriebe and own calculations.

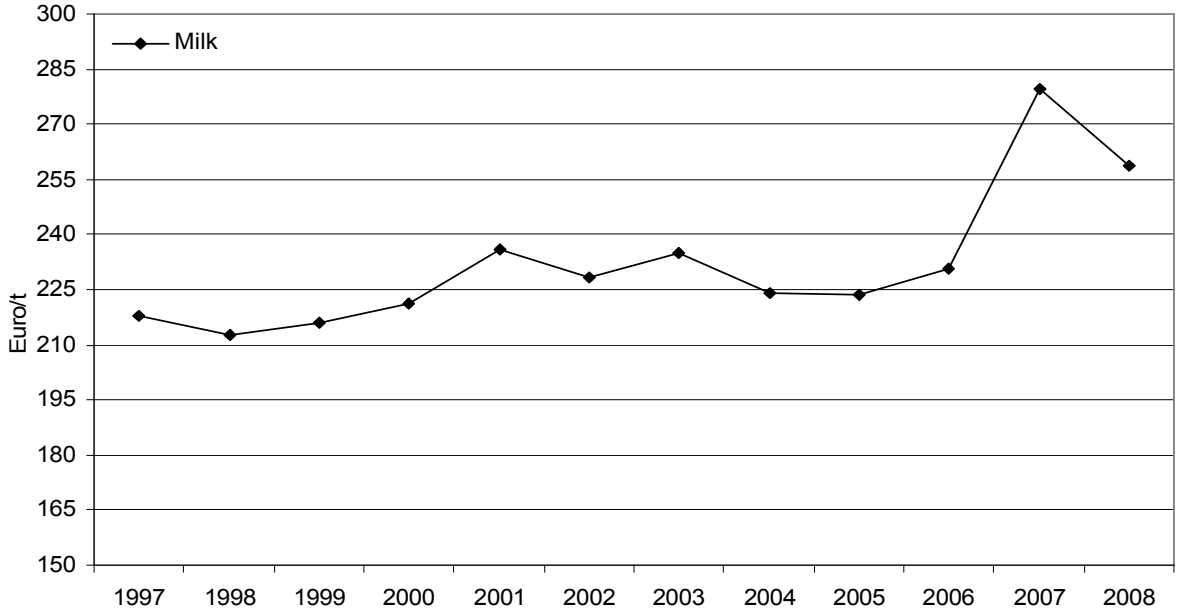
The production costs for cattle were slightly increasing after 1998 while they remained within the range of 200-300 €/LU (Livestock Unit). In 2008 the costs reached about 300 €/LU. The production costs of pigs were a little volatile, and remained within a range of 600-700 €/LU. In 2007, however, they increased to about 800 €/LU (figure 5).

Yet again, the rise of farm input prices is primarily responsible for the increase of cattle and pig production costs after 2006. More specifically, this is attributed to the increase in feed and energy prices (Figure 4).

The unit cost of production per tonne of milk shows (Figure 6) minor fluctuations. It shows a slight increase until 2001 followed by slight decrease in 2004. From 2006 onwards the production cost per tonne of milk continues to

increase, while an increase of only about 40 €/tonne is noticed in 2008. High feed as well as energy prices are the major contributors to the higher milk production costs after 2006.

Figure 6: Production costs per tonne of milk



Source: BMELV-Testbetriebe and own calculations.

4 Summary and conclusions

The study attempted to solve the problem of deriving input cost allocation coefficients from whole farm data using an econometric approach. To increase the robustness of the model, a multivariate outlier detection method is applied to the data set prior to estimation. The robust Mahalanobis distances are calculated for each observation in a data set. The points with the extreme squared distance of Mahalanobis and a chi square value of less than 5 % are identified as outlier observations. Production cost estimates before and after the removal of outlier are compared and discussed for the year 1998 and the results clearly show that the presence of outlier observations can lead to biased parameter estimation, over-or underestimation of results and violation of model specifications.

The time series production cost analysis for major products shows that production cost of wheat, barley and potato is highly affected by high fertilizer and energy prices. However, the cost of production for rape seed shows a slight increase over time. The cost of production of cattle was slightly decreasing until 2004 and then increased afterwards. The production cost of pig remained in the range of 600-700 €/LU until 2005/2006 and jumped to the range of 700-800 €/LU from 2006/2007 onwards. Similarly, milk production was stable with minor fluctuations until 2004, which then rise constantly afterwards until 2007. High feed and energy prices after 2004 are primarily responsible for the increase in production cost cattle, pig and milk.

References

BEN-GAL, I. (2005): Outlier detection, In: Maimon, O. and Rockach, L (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005, ISBN 0-387-24435-2.

BMELV (2010): Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz, to be found at <http://www.bmelv-statistik.de> [quoted March 2010]

BOONE, J.A. and J.H. WISMAN (1998): Cost prices in pig production: Experiences with an EU-wide comparison, in Pig News and Information 19 (1998) 1, Landbouw Economisch Institute (Agricultural Economics Research Institute), Wageningen.

BUTAULT, J.P. and M. CYNCYNATUS (1990): Les Cmits de Production des Pnncipaux Produits Ap.coles dam la Communauti Europke en 1984-85-86. Paris: INRA. In Hallam, D. et al. (1999): Estimating Input Use and Production Costs From Farm Survey Panel Data. Journal of Agricultural Economics, 1999, vol. 50, issue 3, 440-449.

CAUSSINUS, H. and A. ROIZ (1990): Interesting projections of multidimensional data by means of generalized component analysis, In Compstat 90, 121-126, Heidelberg.

ERRINGTON, A. (1989): Estimating Enterprise Input-Output Coefficients from Regional Farm Data. Journal of Agricultural Economics, 40, 52-56.

ESCALANTE, H. J. (2005). A comparison of outlier detection algorithms for machine learning. In *Proceedings of the International Conference on Communications in Computing*.

FILZMOSER, P. (2004): A multivariate outlier detection method. In: S. Aivazian, P. Filzmoser and Yu. Kharin, editors, Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling, volume 1, 18-22, Belarusian State University, Minsk.

FRIENDLY, M. (1991): SAS system for graphics. Cary, NC, USA: SAS Institute Inc.

FRIENDLY, M. (2008): Robust multivariate outlier detection, to be found at <http://www.math.yorku.ca/SCS/sssg/outlier.html> [quoted January 2010]

.HADI, A.S. (1992): Identifying Multiple Outliers in Multivariate Data. Journal of the Royal Statistical Society, Series (B), 54, 761-771.

HALLAM, D., BAILEY, A.S., JONES, J. and A. ERRINGTON (1999): Estimating Input Use and Production Costs From Farm Survey Panel Data. Journal of Agricultural Economics, 1999, vol. 50, issue 3, 440-449.

LEÓN, Y., PEETERS, L., QUINQU M. and Y. SURRY (1999): The Use of Maximum Entropy to Estimate Input-Output Coefficients from Regional Farm Accounting Data. In: Journal of Agricultural Economics 50 (3): 425-439.

LIU, H., SHAH, S. and W. JIANG (2004): On-line outlier detection and data cleaning, Computers and Chemical Engineering, 28, 1635–1647.

MACK and MANN (2008): Defining elasticities for PMP models by estimating marginal cost functions based on FADN Data - the case of Swiss dairy production, 107th EAAE Seminar, Seville, Spain.

MIDMORE, P. (1990): Estimating Input-Output Coefficients from Regional Farm Data: A Comment. Journal of Agricultural Economics, 41, 108-111.

MOXEY, A. and R. TIFFIN (1994): "Estimating linear production coefficients from farm business survey data: a note" Journal of Agricultural Economics, of Agricultural Economics, Blackwell publishing, volume 45(3), pages 381-385.

NGUYEN, D.T.M., MCLAREN, K. and X. ZHAO (2008): Multi-output broadacre agricultural production: estimating a cost function using quasi-micro farm level data from Australia, AARES 52nd annual conference.

PEETERS, L. and Y. SURRY (2003): Farm Cost Allocation Based on the Maximum Entropy Methodology: The Case of Saskatchewan Crop Farms, Agriculture and Agri-Food Cananda, Technical Report 2121/E, Ottawa, Ontario.

PINGAULT and DESBOIS (2003): Estimation des coûts de production des principaux produits agricoles à partir du RICA, NEE n.19, 9-51.

POLLET, P., BUTAULT, J.P. and E. CHANTRY (2001): The Agricultural Production Costs Model. Agricultural Division, INSEE. ESTAT-2001-03195-00-00-EN-TRA-00 (FR).

WILLIAMS, G. J. and Z. HUANG (1997): Mining the knowledge mine: The hot spots methodology for mining large real world databases. In: Abdul Sattar, editor, Advanced Topics in Artificial Intelligence, volume 1342 of Lecture Notes in Artificial Intelligence, 340–348, Springer.

YOUNG, F.W. and W.S. SARLE (1989): Multivariate statistical methods: practical applications course notes. Cary, NC: SAS institute INC.