Drivers of Farm Performance: Empirical and econometric framework

by Pr Johannes Sauer
(from OECD report TAD/CA/APM/WP(2018)16)

1. This note outlines the methodological steps applied to empirically identify and econometrically approximate the different technology classes for each national farm type. Furthermore, it describes the statistical procedure that has been used to represent a variety of farm classes within the number of classes determined empirically based on a combination of differences in multiple farm specific characteristics as well as multiple netput (i.e. output and input) variables (see in more detail Sauer and Morrison-Paul, 2013).

Technology model

2. The first part of the econometric modelling exercise consists in choosing a technology function to approximate the production process of a farm. Depending on theoretical considerations and data availability, different netput functions (e.g. production, cost, profit, distance or transformation function) and functional forms can be chosen for this purpose. From a purely theoretical perspective, (dual) functional representations that allow for the inclusion of price-related information are desirable in order to map the technical and allocative behaviour of farm managers. However, the availability of multi-output related information seems problematic for many national farm accounting systems. Furthermore, in order to avoid the disadvantages of normalising by one input or output as required for a distance function representation and therefore implying econometric endogeneity problems (as the right-hand side variables are expressed as ratios with respect to the left-hand side variables, see for example Paul and Nehring, 2005), a single-output based production function representation applying a second order approximation in the form of a flexible translog functional form is preferred.

3. The analysis considers a production function model representing the most output producible from a given input base and existing production conditions (representing the feasible production set). In general form, this function can be written as $0 = F(Y, \mathbf{X}, \mathbf{T})$, where Y is the farm's output, $\mathbf{X}$ is a vector of production related inputs and $\mathbf{T}$ is a vector of shift variables reflecting external production conditions. Applying the implicit function theorem F can be explicitly specified with one of the arguments on the left-hand side of the equation. Hence, the production function $Y = G(\mathbf{X}, \mathbf{T})$ can be estimated with Y as the output of the farm. This specification of the farm's production technology does not reflect endogeneity of output and input choices, but simply represents the most farm output that can technologically be produced given the levels of the other arguments of the $F(\cdot)$ function. The production function is approximated by a flexible functional form (second-

order approximation), to accommodate various interactions among the arguments of the function including non-constant returns to scale and technical change biases.

4. This second order flexible production function model can be formulated as:

$$Y_{it} = F(\boldsymbol{X}_{it}, \boldsymbol{T}) = \alpha_0 + \sum_{k=1}^{n} \beta_k lnX_k + \frac{1}{2}\sum_{k=1}^{n} \beta_{kk} lnX_k lnX_k + \sum_{k=1}^{n-1}\sum_{l=i+1}^{n} \gamma_{kl} lnX_k lnX_l + \delta_T T + \delta_{TT} TT + \sum_{k=1}^{n} \delta_{kT} lnX_k T \quad [1]$$

for farm i in period t with Y = total milk (crop) output, **X** is a vector of X_k inputs depending on the type of production, and a time trend T as the only component of the vector **T**. By using such a flexible functional form, observable technology differences among production units are accommodated to a certain extent as derived measures (such as output elasticities) allow for different netput mixes, hence, will differ per observation.

5. Unobservable technology heterogeneity is further partly accommodated by the error term in the estimation model. However, the factors leading to technology heterogeneity between farms are not directly represented by estimating [1] alone and therefore parameter estimates might be biased (Griliches, 1957). Consequently, derived policy conclusions remain at a very general level. Recognising and evaluating heterogeneity among production systems and exploring differences in technical change developments requires a more explicit approach, consisting in estimating the technology separately for different groups or 'classes' of farms. Hence, the estimation of production technology as outlined by [1] will be combined with a probabilistic approach that allows to simultaneously consider multiple characteristics of farms operating in a specific production system. This approach will result in an adequate approximation of the individual farm's production technology by considering a multitude of characteristics and therefore robustly identifying various farm groups or classes along these characteristics, for which technologies are then estimated. Hence, we combine the estimation of the production structure as outlined in [1] with the estimation of a latent class structure (see for example Greene, 2002 and 2005; Orea and Kumbhakar, 2004; Sauer and Morrison-Paul, 2013).

Class identification model

6. Different methods can be applied to explicitly consider technological heterogeneity in farm level production (Bravo-Ureta, 1986; Tauer, 1998; Newman and Matthews, 2006; Gillespie et al., 2009; Kumbhakar et al., 2009). The data sample can simply be chosen based on some homogenous production criteria (e.g. a norm technology defined by the average technology in the sample) or can be divided in sub-samples to estimate different technologies based on a particular characteristic (e.g. conventional versus organic or small scale versus large scale). At a higher methodological level, multiple criteria based cluster analysis can be applied to divide the sample according to similar farm or production related characteristics (using between versus within variances to group observations). Furthermore, random coefficient estimators have been used to model each farm as a unique technology based on continuous parameter variation (Alvarez et al., 2008; Greene, 2005).

7. The application of latent class structures (LCM) to empirically identify and estimate heterogeneous classes of observations (farms or firms) results in a separation of the data into multiple technological classes (groups or categories). This separation is based on estimated probabilities of class memberships considering multiple pre-specified criteria. Each farm is then assigned to a specific class based on these probabilities while both the estimated technological (flexible TL function) as well as the estimated probability relationships are considered (Sauer and Morrison-Paul, 2013; Balcombe et al., 2006). Hence, a latent class modelling approach overcomes possible estimation bias due to omitted variables with respect to the class identification vector and also effectively addresses endogeneity suspicions by a simultaneous estimation approach (i.e. a technology model and class identification model). In more detail, the LCM estimates a multi-nomial logit model together with the technological structure (whereas the number of parameters to be estimated might be limited by available degrees of freedom). Statistical tests can be performed to choose the most adequate number of classes/technologies to be considered. Furthermore, in addition to multiple technologies, a flexible functional form with a random effects panel estimation routine can be applied (Greene, 2005; Alvarez and delCorral, 2010) to capture farm heterogeneity over time. In this project the focus is explicitly on measuring productivity instead of unobserved inefficiency (based on a frontier specification) to reflect the specific interest in relative productivity levels between farms considering country level contextual specificities.

8. The latent class model in a more general form can be formally denoted as the technology model (outlined in equation [1]) for class c:

$$Y_{it} = F(\boldsymbol{X}_{it}, \boldsymbol{T}) = \left(\alpha_0 + \sum_{k=1}^{n} \beta_k lnX_k + \frac{1}{2} \sum_{k=1}^{n} \beta_{kk} lnX_k lnX_k + \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} \gamma_{kl} lnX_k lnX_l + \delta_T T + \delta_{TT} TT + \sum_{k=1}^{n} \delta_{kT} lnX_k T \right) | \, c \qquad [2]$$

where c denotes the class including farm i implying a different technology function for each class c. Assuming a normal distribution for the error term, the likelihood function for farm i at time t for class c, LF_{ict}, has the standard Ordinary Least Squares (OLS) form. The unconditional likelihood function for farm i in class c, LF_{ic}, is the product of the likelihood functions in each period t, and the likelihood function for each farm, LF_i, which is the weighted sum of the likelihood functions for each class c (with the prior probabilities of class c membership as the weights), i.e. $LF_i = \sum_c P_{ic} LF_{ic}$. These prior probabilities P_{ic} are parameterised using a multinomial logit model (MNL) consisting of indicators to describe the different dimensions of farm performances and characteristics and which are used to determine the probabilities of class memberships or separate technologies (separating or q-variables q_i).

9. Hence, the MNL parameters θ_c are estimated for each technology class (relative to one class serving as numeraire)

$$P_{ic} = exp(\theta_c q_i) \Big/ \left[\sum_c exp(\theta_c q_i) \right] = exp\left(\theta_{0c} + \sum_n \theta_{nc} q_{nit} \right) \Big/ \left[\sum_i exp\left(\theta_{0c} + \sum_n \theta_{nc} q_{nit} \right) \right]$$

[3]

where the q_{nit} denote the N q-variables/indicators for farm i in time period t.

Multi-dimensional indices

10. As outlined earlier farms are production units which differ along multiple characteristics: production structure, environmental impact and sustainability, innovation behaviour, commercialisation focus, openness towards cooperation, input intensity and capital endowment, diversity of production, individual characteristics such as age or education, as well as locational conditions. A multitude of continuous or binary variables in level form can be used to directly approximate these farm characteristics as elements of the class identification vector. However, including all those variables would lead to scaling and weighting problems and also, depending on sample size, most probably to limitations regarding the number of variables that can be considered due to missing degrees-of-freedom. Hence, multi-dimensional indexes are defined and statistically estimated, to then be incorporated as elements of the class identification vector **q**.

11. The various indices are chosen for their potential to contribute to robustly identify and distinguish individual farms. These multi-dimensional indexes consist of different variables that measure underlying farm characteristics relevant for the dimension of the specific index to approximate. These individual index components can be equally weighted with regard to their importance for the overall index score. Further, the weights for these components could be chosen following specific expert guidance or based on trial-and-error procedures applying statistical significance criteria with respect to the parameters estimated for the g-vector elements. However, the principal components analysis (PCA) is applied as a statistically well-defined and empirically tested multivariate method to estimate significant and robust weights for the indices' components. The PCA is a method to conduct a conceptual factor analysis that will then create statistically robust indexes based on different variables (for an overview of PCA, see Jackson, 2003 or Afifi et al., 2012).

12. PCA is a multivariate statistical technique used for data reduction. The leading eigenvectors (i.e. principal components) from the eigen decomposition of the correlation or covariance matrix of the variables (here index components) describe a series of uncorrelated linear combinations of the variables that contain most of the variance. In addition to data reduction, the eigenvectors from a PCA can be further inspected to learn more about the underlying structure of the data. Hence, in a first step such a PCA is run for each farm related dimension (e.g. production structure) resulting in the eigenvalues for the individual components (e.g. share of family labour and area or herd size). The eigenvalue for each component represents how much of variance the component explains (i.e. factor loading). Subsequently, the factor loadings are used to calculate the index score for each observation via an optimally-weighted linear combination of the factor scores for the individual components- characteristics.

13. Accordingly, up to seven different farm indices are defined and estimated for each observation of the respective sample using the deviations of each index component from the sample mean to adequately consider differences between member countries' farm structures and conditions (For example, an average family farm in Italy in terms of family labour share may be very different from an average family farm in terms of family labour share located in Estonia). Scaling issues between different components (e.g. share of family

labour versus herd size or acreage) are further addressed by calculating the z-score based deviations for these components, which are then used for the PCA based index creation following the statistical procedure outlined above. For subsequent analyses up to seven multi-dimensional indexes are chosen to identify and measure class membership per farm and year. These indexes are estimated as outlined above subject to type of production and data availability. Table 1 provides an overview of the choice of indexes' components. The significance and the posterior probabilities of resulting q-variable coefficient's estimates are evaluated for the individual farm classes. Furthermore, statistical tests are applied to robustly determine the number of classes (for example, the Akaike Information Criterion/Schwarz and Bayesian information criterion [AIC/SBIC] tests, described in Greene, 2005) by testing down (i.e. to verify if fewer classes would be statistically supported).

Table 1. Indexes for farm classification

Components for multi-dimensional indices as elements of class identification vector q, see equation [3].

Indexes / Components	Index 1 Structure	Index 2 Sustainability	Index 3 Innovation/ Commerce/Coop	Index 4 Intensity	Index 5 Diversity	Index 6 Individual	Index 7 Location
Acreage	x						
Age						x	
Agritourism income			x				
Altitude							x
Biofuel income			x				
Capital per cow				x			
Capital per labour				x			
Chemicals usage		x					
Education						x	
Environmental subsidies		x					
Experience						x	
Family labour share	x						
Forestry production					x		
Gender						x	
Herd size	x						
Herfindahl index					x		
Investment subsidies			x				
Labour per cow				x			
Land irrigated share			x				
Land rented share			x				
Less-favoured-area							x
Material per land				x			
Natura2000							x
Net investment ratio			x				
Organic production		x					
Ownership	x						
Production diversity					x		
Stocking density		x					
Total assets				x			

Note: Final choice of indices' components depends on production type and data availability.

Full model specifications

14. The combined (technology and class identification) model can be estimated in a cross-sectional or a panel form whereas for the full-model specification a random effects based estimator can be applied (Sauer and Morrison-Paul, 2013; Greene, 2005). The panel data related specification of the model is then:

$$Y_{it|c} = \alpha_0 + \sum_{k=1}^{n} \beta_{k,c} lnx_{kit} + \frac{1}{2}\sum_{k=1}^{n} \beta_{kk,c} lnX_{kit} lnX_{kit} + \sum_{k=1}^{n-1}\sum_{l=k+1}^{n} \gamma_{kl,c} lnx_{kit} lnx_{lit} + \delta_{T,c} t_{it} + \delta_{TT,c} t_{it} t_{it} + \sum_{k=1}^{n} \delta_{kT,c} lnx_{kit} t_{it} + \varepsilon_{it|c} \quad [4]$$

with farm i in time period t and class c and ε denoting an independent and identically distributed (iid) stochastic term. For an alternative specification each observation is considered as a separate entity and the model is then estimated as a cross-sectional specification. This model allows farms to switch between technology classes and hence, changes in production systems over the time period can be approximated.

$$Y_{i|c} = \alpha_0 + \sum_{k=1}^{n} \beta_{k,c} lnx_{ki} + \frac{1}{2}\sum_{k=1}^{n} \beta_{kk,c} lnX_{ki} lnX_{ki} + \sum_{k=1}^{n-1}\sum_{l=k+1}^{n} \gamma_{kl,c} lnx_{ki} lnx_{li} + \delta_{T,c} t_i + \delta_{TT,c} t_i t_i + \sum_{k=1}^{n} \delta_{kT,c} lnx_{ki} t_i + \varepsilon_{i|c} \quad [5]$$

with farm i, class c and ε denoting again the independent and identically distributed (iid) stochastic term.

15. As both model components (technology related and class identification related) are simultaneously estimated the probabilities P_{ic} (see equation [3]) are functions of the parameters of the MNL model and the log-likelihoods LF_{ic} are functions of the technology parameters for class c farms. Accordingly, the overall likelihood function for farm i in class c consists of both sets of parameters whereas the overall log-likelihood function for the complete model is maximised based on the sum of the individual log-likelihood functions. Finally, due to degrees-of-freedom problems related to the parameter intensive LCM specification, as done in Sauer and Morrison-Paul (2013), the models in [4] and [5] are estimated as a reduced (or constrained) form approximation to the underlying translog functional form. Thus, the resulting (first-order and own second-order) elasticities represent the average contributions of each input to production, as well as overall technical change and returns to scale for each class. To accommodate and measure the second-order effects involving input technical change biases and substitution, the full TransLog (TL) form for the full sample and the separate classes will also be estimated. If the distinctions among classes capture key differences in technology, as found for all country cases investigated, the elasticities for the constrained and fully flexible functional forms

will be comparable, but incorporating the interaction terms will allow assessment of cross effects between inputs.

Performance measures

16. Several performance measures derived from the technology related component of the combined estimation model outlined in equation [3] and [4] or [5] are explored. In a first step, the relative *levels of productivity* are estimated among the different identified farm classes based on the predicted output levels for a given amount of inputs at the sample means (Alvarez and Corral, 2010). The hypothetical productivity levels are then estimated for each class assuming an alternative technology and the differences between real and hypothetical technologies are compared. In a second step, productivity dynamics, more commonly noted as *technical change*, is considered per class and technology. Such technical change is measured by shifts in the overall production frontier over time using the output elasticity with respect to T

$$\epsilon_{y,T\,|c} = \frac{\partial lnY}{\partial T}\,|c = \delta_{T,c} + 2 * \delta_{TT,c} t_i + \sum_{k=1}^{n} \delta_{kT,c} lnx_{ki} \qquad [6]$$

Technical change is estimated for each class at the sample means using the estimated parameters and the elasticity formula given by equation [6]. The hypothetical technical change rate is also estimated for each class assuming an alternative technology and the differences between real and hypothetical rates of technical change compared. These two core measures deliver evidence on the distribution of productivity and technical change over different farm classes and also allow inferences with regard to potential productivity increases as well as technical change rate accelerations by facilitating farms' switch to more productive technologies over time.

17. The next analytical performance measure that is derived from the constrained flexible and fully flexible TL production functions are *first-order elasticities* with respect to the primary output (e.g. dairy or crop related output) for each class c. These first-order elasticities in terms of primary output Y represent the (proportional) shape of the production function (given other inputs) for input X_k - or input contributions to primary output respectively. The estimated output elasticity with respect to input k

$$\epsilon_{y,k\,|c} = \left(\frac{\partial Y}{\partial X_k} * \left[\frac{X_k}{Y}\right]\right)|c = \beta_{k,c} + \frac{1}{2}\beta_{kk,c} lnx_{ki} + \sum_{l=k+1}^{n} \gamma_{kl,c} lnx_{li} + \delta_{kT,c} t_i \qquad [7]$$

would be expected to be positive, with its magnitude representing the (proportional) marginal productivity of X_k. Second-order own-elasticities may be computed to confirm that the curvature of these functions satisfies regularity conditions; the marginal productivity is expected to increase at a decreasing rate, so second derivatives with respect to X_k are expected to be negative to fulfil the concept of a well-defined functional representation of the production problem under consideration. Input elasticities give insight

into the relative productivity of different inputs given the production context. The policy maker is, hence, able to evaluate the marginal contribution of each input to overall production at farm and sectoral levels and therefore its relative importance for the type of production. Linked to the technology class related analysis performed here, input elasticities enable policy makers to evaluate different technologies with respect to their relative input intensity and dependence.

18. Based on the derived first-order elasticities *returns to scale* are estimated as a linear combination of the input elasticities with respect to the primary output. These are simply defined as the sum of the input elasticities as follows

$$s\epsilon_{y,X\,|c} = \sum_{k=1}^{n}\left(\frac{\partial Y}{\partial X_k} * \left[\frac{X_k}{Y}\right]\right)|c = \sum_{k=1}^{n}\left(\beta_{k,c} + \frac{1}{2}\beta_{kk,c} lnx_{ki} + \sum_{l=k+1}^{n}\gamma_{kl,c} lnx_{li} + \delta_{kT,c} t_i\right)$$

[8]

Returns to scale allow for empirically informed inferences about the "cost of scale" with respect to a type of production at farm and sectoral level. Increasing returns to scale suggest extending the production of the specific output to increase the profitability of production via lower average costs. Decreasing returns suggest the opposite, i.e. reducing the scale of production to increase profitability via lower average costs, and finally constant returns suggest that the actual scale of production is approximately near the optimal — most efficient — point of scale for the specific firm or sector (*ceteris paribus*). Policy makers are therefore able to design more efficient programmes and measures to more effectively enable economies of scale where relevant based on these measures. As a result of the simultaneously estimated farm classes, policy makers are able to design such programmes more efficiently as the latter are also farm class specific depending on the class identification vector (see above).

19. Finally, *second-order or cross-elasticities* with respect to input substitution as well as input-using or input-saving technical change (biases) can be estimated based on the flexible TL production function. These performance measures involve second-order derivatives such as, for input substitution,

$$\epsilon_{k,l\,|c} = \left(\frac{\partial^2 Y}{\partial X_k \partial X_l}\right) * \left[\frac{X_l}{\left(\frac{\partial Y}{\partial X_k}\right)}\right]|c = \left(\frac{\partial MP_{Y,k}}{\partial X_l}\right) * \left[\frac{X_l}{MP_{Y,k}}\right]|c = \gamma_{kl,c}$$

[9]

where $MP_{Y,k}$ refers to the marginal product of Y with respect to X_k. The elasticity in [9] represents the extent to which the marginal product of X_k changes due to changes in X_l. The corresponding technical change measure

$$\epsilon_{k,T\,|c} = \left(\frac{\partial^2 Y}{\partial X_k \partial T}\right) * \left[\frac{1}{\left(\frac{\partial Y}{\partial X_k}\right)}\right]|c = \left(\frac{\partial MP_{Y,k}}{\partial T}\right) * \left[\frac{1}{MP_{Y,k}}\right]|c = \delta_{kT,c}$$

[10]

represents the bias in technical change, i.e. whether such technical change is input k-using or input k-saving. Accordingly, the input k intensity for farms in class c is increasing or decreasing over the time period investigated. Finally, returns to scale (see equation [8]) can be analysed whether they are increasing or decreasing over time (depending on technical change) for each identified class of farms following:

$$s\epsilon_{y,X,T\,|c} = \frac{\partial \sum_{k=1}^{n} \left(\frac{\partial Y}{\partial X_k} * \left[\frac{X_k}{Y} \right] \right)}{\partial T} \,|c\, = \sum_{k=1}^{n} \left(\delta_{kT,c} \right)$$

[11]

These second-order performance measures rely on the unconstrained flexible functional form and deliver empirical evidence on the input substitution patterns and technical change biases per class. Policy makers might want to know which type of farm is most effective in substituting less sustainable inputs by more sustainable inputs as a reaction to specific incentives or regulatory measures.

References

Afifi, A. A., S. May, and V. A. Clark (2012), Practical Multivariate Analysis. 5th ed. Boca Raton, FL: CRC Press.

Alvarez, A. and J. del Corral (2010), "Identifying different technologies within a sample using a latent class model: extensive versus intensive dairy farms", *European Review of Agricultural Economics*, Vol. 37, pp. 231–50.

Alvarez, A., J. del Corral, D. Solis, and J. A. Perez (2008), "Does intensification improve the economic efficiency of dairy farms?", *Journal of Dairy Science*, Vol. 91, pp. 3693–8.

Balcombe, K., I. Fraser, and J. H. Kim (2006), "Estimating technical efficiency of Australian dairy farms using alternative frontier methodologies", *Applied Economics*, Vol. 38, pp. 2221–36.

Bravo-Ureta, B. E. (1986), "Technical efficiency measures for dairy farms based on a probabilistic frontier function model", *Canadian Journal of Agricultural Economics*, Vol. 34, pp. 400–15.

Gillespie, J., R. Nehring, C. Hallahan, C. J. Morrison Paul, and C. Sandretto (2009), Economics and productivity of organic versus non-organic dairy farms in the United States, *manuscript*, ERS/USDA.

Greene, W. (2002), "Alternative panel data estimators for stochastic frontier models", *Working Paper*, Department of Economics, Stern School of Business, NYU.

Greene, W. (2005), "Reconsidering heterogeneity in panel data estimators of the stochastic frontier model", *Journal of Econometrics*, Vol. 126, pp. 269–303.

Griliches, Z. (1957), "Specification bias in estimates of production functions", *Journal of Farm Economics*, Vol. 49, pp. 8–20.

Jackson, J. E. (2003), A User's Guide to Principal Components. New York: Wiley.

Newman, C. and A. Matthews (2006), "The productivity performance of Irish dairy farms 1984–2000: A multiple output distance function approach", *Journal of Productivity Analysis*, Vol. 26, pp. 191–205.

Nicholson, W. and C. Snyder (2008), *Microeconomic Theory*, 10th edition, Thomson South-Western Publishers, Ohio.

Orea, L. and S. C. Kumbhakar (2004), "Efficiency measure-ment using a latent class stochastic frontier model", *Empirical Economics*, Vol. 29, pp. 169–83.

Paul, C. J. M. and R. Nehring (2005), "Product diversification, production systems, and economic performance in US agricultural production", *Journal of Econometrics*, Vol. 126, pp. 525–48.

Sauer, J. and C.J. Morrison Paul (2013), "The empirical identification of heterogeneous technologies and technical change", *Applied Economics*, Vol. 45, pp. 11.

Tauer, L. W. (1998), "Cost of production for stanchion versus parlor milking in New York", *Journal of Dairy Science*, Vol. 81, pp. 567–9.