Exploring the distribution of conditional quantiles estimation ranges: an application to specific costs of pig production in the European Union

Dominique Desbois
UMR Economie publique, INRAE-AgroParisTech

This paper uses symbolic data analysis tools to visualize conditional quantile estimation intervals, applying it to the problem of cost allocation in agriculture. After recalling the conceptual framework of the estimation of agricultural production costs, the first part presents the empirical model, the quantile regression approach and the interval data processing techniques used as symbolic data analysis tools. The second part presents the comparative analysis of the econometric results between twelve European Member States, using the principal components analysis and the hierarchical grouping of the estimation intervals, by discussing the relevance of the exploratory graphs obtained for the international comparisons.

Keywords: *input-output model, pig production cost, micro-economics, quantile regression, factor analysis and hierarchic clustering of interval estimates.*

"Applied economists increasingly want to know what is happening to an entire distribution, to the relative winners and losers, as well as to averages." (Angrist and Pischke, 2009)

1. Introduction

The successive reforms of the Common Agricultural Policy (CAP), the integration of the agricultural systems of the Member States resulting from the enlargement process of the European Union (EU), both in the context of competitive markets and markets subject to regulation generates recurrent needs to estimate the production costs of major agricultural products. The analysis of agricultural production costs, whether retrospective or prospective, is also a tool for analyzing farmers' margins. It makes it possible to evaluate the price competitiveness of farmers, one of the major elements of the development or maintenance of agro-food chains in certain European regions. Thus, the estimation of production costs provides some partial but essential insights into the questions posed by the adaptation of European agriculture to the context of agricultural markets, whether national, European or international, both from the point of view of the regulation of international trade in agricultural products (see the proposals for measures to combat market im-

[1] http://agriculture.gouv.fr/etude-sur-les-mesures-contre-les-desequilibres-de-marche-quelles-perspectives-pour-lapres-quotas

balances in the post-quota dairy sector[1]), and the successive reforms of the CAP (see the debate on future CAP in 2020[2]) or new challenges for European agriculture caused by environmental factors (climate change, environmental and biodiversity management[3]).

Confronted more directly with price risks since the abolition of production quotas in 2015[4], European producers with few opportunities for differentiation opt for cost reduction strategies, seeking either to reduce structural costs by playing on the volume of production, either to reduce specific costs by optimizing the management of inputs or opting for low-input technical routes. However, structural adjustment is not always possible due to constraints (herd management, rights to produce, availability) that can restrict access to the three main production factors of land (e.g. mountain areas), working capital (financing conditions) or work, whether salaried or self-employed. On the other hand, the adjustment on specific inputs offers more flexibility as shown by the adoption of reasoned practices leading to savings on the main items of expenditure such as animal feed and veterinary fees. The evolution of specific costs, not only globally but also by product, thus constitutes an important indicator for pig farmers in terms of technical management of the herd and adjustment of their product mix to the demands of the agricultural markets, taking into account the resources and competitiveness factors available to them.

Given these different issues, in contexts either ex ante scenario development or ex post evaluation of measures concerning possible agricultural public policy options, we must be able to provide information as suggested (Angrist and Pischke, 2009) across the entire distribution of production costs, thus making it possible to meet the needs of simulations or impact analysis within the various common organizations of the market. In this perspective, from the observation of asymmetry and heterogeneity of their empirical distribution, we propose a methodology adapted to the problem of the estimation of the specific costs of production relative to the main agricultural reference products in a European context where farm holdings remain predominantly multi-commodity, despite a preponderance of specialized farms in some more integrated sectors of agricultural production. In this multi-product context, it is strategic to generate for each of the main agricultural products the central estimates of the cost distribution, but also the lower or higher quantiles with a view to selectivity of the instruments for regulating agricultural markets for production, and evaluation of public policies.

Given the heterogeneity of agricultural production structures and productive choices in Europe, how can the maximum amount of information be used to estimate agricultural production costs? In response to this concern, we propose an estimation methodology that can provide information on the overall distribution of specific production costs for the main agricultural reference products in a European context. In order to overcome the constraint of average estimators, sensitive to the asymmetry or the extreme values of the distributions of interest and likely to mask the inter-structural differences, it is necessary to generate for each of the main agricultural products not only the median estimates of cost distribution but also lower or higher quantiles. To this end, we propose using a methodology to obtain estimates of these quantiles of specific costs that are conditioned by the product mix of farmers (Desbois et al., 2017a). In order to demonstrate the relevance of this approach, we will then apply this methodology to estimate the specific costs of pig, given its place in the world

[2] https://www.sfer.asso.fr/source/Coll-trajectoire-2018/Programme-Future-of-CAP-30-05.pdf

[3] http://agriculture.gouv.fr/lagriculture-et-les-forets-au-coeur-de-la-cop23

[4] Cf. EU Milk Margin Estimate up to 2016, n°16: "Gross margins: a lot of instability and a record low level in third quarter of 2016", https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/farming/documents/agri-farm-economics-brief-16_en.pdf

[5] In 2017, pig production produced by the 28 European countries accounted for 20% in weight of pigmeat produced at the world level (according to https://ec.europa.eu/agriculture).

production by the EU28[5], on a set of twelve European states (EU12) where these productions are significant in 2006, the base year chosen for the period.

We first present the empirical model for estimating the specific costs of production, derived from an econometric cost allocation approach, initially developed by Aufrant (1983) using microeconomic data to build an input-output matrix (Divay and Meunier, 1980). Then, we introduce the estimation methodology according to the conditional quantiles proposed by Koenker and Bassett Jr (1978), extended by Koenker (2005). Next, we present the symbolic data analysis procedures used to explore the empirical estimates of conditional quantile distribution intervals based on the concepts and methods provided by the symbolic approach (Bock and Diday, 2000; Billard and Diday, 2006). Then, we present the graphs from the analysis tools for symbolic data applied to the estimation intervals of the conditional quantiles. Finally, we conclude on the relevance of this approach applied to the pig production, proposing an extension of this type of analysis at the regional level.

2. Conceptual framework and methodological aspects of cost allocation

Surveys specific to large agricultural commodities are conducted according to the production workshop to provide detailed data on operational production costs, such as that used by the French pig Institute (IFIP) on specialized pig producers for France[6]. However, these technical and economic surveys are relatively expensive, making their generalization to all European pig farms financially unbearable. Also, this work is situated in the cost allocation framework of the factors to multiple productions, initiated on a European scale by INRAE[7] works (Butault et al., 1988) financed by the European Commission (EC), allowing to estimate production costs on the basis of the Accounting Information Network (FADN), accounting survey harmonized on definitions about the professional holdings and the accounting, technical and financial aggregates.

2.1. The empirical model for estimating the specific costs of production

In EU agricultural accounting systems, the recording of charges is done at the farm level and does not provide a direct estimate of the production costs incurred by that farm for each of the agricultural crops undertaken. From the accounting records, the farm holding data sheet[8] of the FADN survey provides individually by farm the amount of the gross products generated by the various speculations and the one of specific costs, the sum of the recorded input purchases. So, by regression of specific costs on gross products, it becomes possible to estimate the allocation coefficients of expenditure to the main agricultural products, called 'specific coefficients of production'. The gross margin M_i of the farm holding i is defined as a difference between the sum of the gross products x_i and the sum Y_i of the specific costs[9]: $M_i = x_i - Y_i$.

The sum of specific costs is linearly decomposed according to each production j, as follows :

$$Y_i = \sum_{j=1}^{p} \gamma_j x_i^j + \varepsilon_i \quad \text{with} \quad \varepsilon_i \quad \text{i.i.d.} \qquad (1)$$

[6] https://www.ifip.asso.fr/fr/resultats-economiques-elevages-de-porc.html

[7] Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, formerly Institut national de la Recherche agronomique (INRA).

[8] The questionnaire used to establish this farm holding data sheet and the methodology of the FADN survey are available at: http://wwww.agreste.agriculture.gouv.fr/enquetes/reseau-d-information-comptable-610/reseau-d-information-comptable.

[9] Throughout this text, we use the classical convention in mathematical statistics to denote the endogeneous variable by the Y symbol and the exogeneous variables by x. Conversely in a previous paper published to present the empirical model (Desbois et al., 2013), the econometrical convention using the x symbol for inputs and the Y symbol for outputs have been used.

implying: $M_i = \sum_{j=1}^{p} x_i^j - \sum_{j=1}^{p} \gamma_j x_i^j + \varepsilon_i = \sum_{j=1}^{p} (1-\gamma_j) x_i^j + \varepsilon_i$.

Thus, the allocation of the specific costs of the farm holding i to the set J of the productions carried out by this conceptual model makes it possible, because of the complementation to the unit, to deduct the unit rates of gross margin $\hat{\alpha}$ from the estimate of the specific production coefficients $\hat{\gamma}_j$ for each of the J considered productions: $\hat{\alpha}_j = (1 - \hat{\gamma}_j) \quad j = 1, ..., p$

The linear decomposition of the gross margin leads us to estimate the specific production coefficients of the stochastic equation 1 for comparison:

- on the one hand, according to the Gaussian regression methodology, the ordinary least squares estimate $\hat{\mu}_{OLS}(Y_i) = \sum_{j=1}^{p} \hat{\gamma}_j^{OLS} x_i^j$ coincides with conditional expectation;
- on the other hand, according to the quantile regression theory, the estimate $\hat{\mu}_q(Y_i) = \sum_{j=1}^{p} \hat{\beta}_j^{(q)} Y_i^j$ is obtained by solving an optimisation problem, cf. infra eq. 6, and expressed as the conditional quantiles of order q, in order to take into account the intrinsic heterogeneity of the distribution of specific costs as shown in the following section on the estimation methodology.

2.2. *The interest of conditional quantiles in the estimation of agricultural production costs*

The standard specification of the classical regression based on conditional expectation raises certain problems which would be risky to neglect in the perspective of the establishment of benchmarks on production costs, into account the challenges of competitiveness for the various sectors. Firstly, in a context of using the European FADN as an empirical basis for estimating the specific production costs, the stochastic assumptions of the Gaussian linear model may not be satisfied: indeed, the asymmetry of the distributions of specific costs (concentration for lower values and dispersion of values higher than average, or vice versa) lead us to rejecting the assumption of normality of errors. In addition, given the selection method specific to each national FADN (for example, the French FADN is a survey administered according to the quota method), the accounting data are not always collected according to a stratified random sampling design allowing to deliver inferences such as interval estimation based on a parametric distribution, even in the asymptotic case.

The conditional estimation of quantiles was developed in Koenker and Bassett Jr (1978) under the name of 'quantile regression' in order to take into account the heterogeneity of the set of values of an endogenous variable x in the context of a linear model. When looking at farms, this econometric method yields an estimated distribution of specific costs for major agricultural products and thus complements the estimates obtained by classical mean regression, which only provides an average value (expectation) of these same costs. Instead of an interval estimate built on a normality assumption, the quantile process provides an empirical distribution of the estimates without having to make assumptions about the nature of this distribution or to follow a stratified random sampling design. For a continuous random variable x, under the assumption that F_x, the cumulative distribution function (CDF), is strictly monotonous, the q^{th} quantile of the population is the value μ_q such as x is less than or equal to μ_q with probability q :

$$q = Pr[Y \leq \mu_q] = F_Y(\mu_q) \tag{2}$$

where F_Y is the CDF of Y. The q^{th} quantile is then defined as the image of the value q by the CDF reciprocal function:

$$\mu_q(Y) = F_Y^{-1}(q) \tag{3}$$

In quantile regression, the q^{th} conditional quantile of the production cost Y conditioned by all the exogenous variables x determining input consumption is the indexed function $\mu_q(Y|x)$ ordered by q. Thus, we can formally define the q^{th} conditional quantile by the following

expression:

$$\mu_q(Y|x) = F^{-1}_{Y|x}(q) \tag{4}$$

where $F_{Y|x}$ is the CDF of Y conditioned by x.
Following Cameron and Trivedi (2005), suppose that the data generating process is a linear model with multiplicative heteroscedasticity:

$$Y = x'\beta + u \quad \text{with} \quad u = x'\alpha \times \varepsilon \quad \text{and} \quad x'\alpha > 0 \tag{5}$$

where $\varepsilon \sim$ i.i.d.$(0, \sigma^2)$ is an identically and independently distributed random vector of zero mean and constant variance σ^2, and where α and β are the parameters of interest.
Under this hypothesis , $\mu_q(Y|x, \beta, \alpha)$, the q^{th} conditional quantile of the production cost Y conditioned by x, α and β, is analytically deduced as follows:

$$\mu_q(x, \beta, \alpha) = x'[\beta + \alpha \times F^{-1}_{\varepsilon}(q)]$$

where F_ε is the CDF of the random error ε.
Thus, for a data generating process following a linear model with multiplicative heteroscedasticity (i.e. $u = X'\alpha \times \varepsilon$, the q^{th} conditional quantile of the production cost Y conditioned by the x exogenous factors is linear in x. Based on the parameters of interest, the q^{th} quantile estimate of the production cost converges to $\beta + \alpha \times F^{-1}_{\varepsilon}(q)$ and therefore behaves monotonically with respect to the quantile order q, depending on the quantile function of the error term, $F^{-1}_{\varepsilon}(q)$.
Following a typology proposed by d'Haultfoeuille and Givord (2014), several models can be distinguished:

i. $Y = x'\beta + u$ with $u = K\epsilon$ with homoscedastic residues $(V(\epsilon|x) = \sigma^2)$ designated as the linear model of homogeneous slope conditional quantile ('location shift model'). The case where $x'\alpha = K$ is constant, corresponds to conditional quantiles differing only by a constant $(\mu_q(Y|x, \beta, \alpha) = x'\beta + KF^{-1}_{\varepsilon}(q)$, all showing the same slope and growing uniformly as the q order of the quantile increases;
ii. $Y = x'\beta + (x'\alpha)\varepsilon$ with $x'\alpha > 0$ with heteroscedastic residues, referred to as the heterogeneous-slope conditional quantile linear model ('location-scale shift model'). The case where $x'\alpha > 0$ corresponds to heterogeneous and increasing slopes (as functions of q): $\mu_q(Y|x, \beta, alpha) = x'(\beta + \alpha\mu_q(\varepsilon))$, involving fixed linear effects $\gamma_q = \beta + \alpha\mu_q(\varepsilon)$;
iii. $Y = x'\gamma_U$ with U random variable independent of Y following a uniform distribution over the interval [0,1] and such that the function $u \longrightarrow x'\gamma_U$ is strictly increasing whatever x , is designated as the random coefficient model. U corresponds to an unobserved random component determining the rank of the individual within the Y distribution. Under the distribution invariance assumption of ranks, which is considered as a strong hypothesis in the scientific literature, the random coefficient γ_q would represent the effect of a marginal change in x for farms at the q^{th} quantile of the ε distribution, based on unobserved characteristics. For example, this distributional assumption of rank invariance is equivalent to assuming that median farms (q = 0,5) in terms of input productivity would maintain this rank, regardless of the different levels of production.

2.3. *Estimation and test procedures*

The Ordinary Least Squares (OLS) estimator can be written as a solution to an optimization problem that minimizes the sum of the squared deviations (L_2 norm):

$$\begin{aligned}\hat{\beta}_{OLS} &= \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_i (y_i - x_i'\beta)^2 \right\} \\ &= \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ e'\delta^2(Y - x'\beta) \right\}\end{aligned} \tag{6}$$

where $e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ is the unit vector in $\mathbb{R}^n$, space of observations, and $\delta^2(Y - x'\beta)$, the vector of quadratic differences.
Similarly, the quantile regression is defined for each quantile order q as the solution of the problem of minimizing the sum of weighted absolute deviations (norm L_1):

$$\hat{\beta}(q) = \arg\min_{\beta \in \mathbb{R}^p} \{ \sum_{i \in \{i | y_i \geq x_i'\beta\}} q|y_i - x_i'\beta| + \sum_{i \in \{i | y_i < x_i'\beta\}} (1-q)|y_i - x_i'\beta| \} \quad (7)$$

which can be written in matrix form as follows:

$$\hat{\beta}(q) = \arg\min_{\beta \in \mathbb{R}^p} \{ qe'(Y - x'\beta \geq 0)\delta^1[Y - x'\beta] + (1-q)e'(x'\beta - Y \geq 0)\delta^1[x'\beta - Y] \},$$

with $e'(Y - x'\beta \geq 0)$, indicator vector of observations i such as $y_i - x'i\beta \geq 0$, and δ^1, the vector of absolute deviations.
Let $x_i = y_i'\beta_q + e_i$ with $e_i = u_i - v_i$, $u_i = e_i \mathbb{1}(e_i > 0)$, $v_i = |e_i| \mathbb{1}(e_i < 0)$. Then, like the L_1 regression (Barrodale and Roberts, 1973), quantile regression can be formulated as a primal problem of linear optimization, which is expressed in matrix form as follows:

$$\hat{\beta}(q) = \arg\min_{\beta \in \mathbb{R}^p, (u,v) \in \mathbb{R}^{n \times n}} \{ qe'u + (1-q)e'v \} \quad (8)$$

under the constraint $Y = x'\beta + u - v$.
This program can be reformulated as an equivalent dual optimization problem:

$$Max_z \{ y'z \} \quad (9)$$

under the constraint $xz = (1-q)xe$ for $z \in [0,1]^n$
Thus, the methods for solving the linear optimization problem developed for the L_1 regression easily extend to quantile regression (Koenker and d'Orey, 1994). The simplex method (Dantzig, 1948) has an algorithmic complexity in $O(n^6)$, the 'interior point' method (Karmarkar, 1984) of algorithmic complexity $O(n^{3.5})$ is preferable in practice as soon as the sample size is important. For large samples, Portnoy et al. (1997) showed that a combination of the 'interior point' algorithm[10] and the Madsen and Nielsen (1993) smoothing algorithm for objective function makes quantile regression computationally competitive with least squares regression.
The weighted conditional quantiles have been proposed as L-estimates in linear heteroscedastic models by Koenker and Zhao (1994). They are defined by a set of weights $\omega_i; i = 1, \ldots, n$ and the following minimization problem:

$$\hat{\beta}_\omega(q) = \arg\min_{\beta \in \mathbb{R}^p} \{ \sum_{i \in \{i | y_i \geq x_i'\beta\}} \omega_i q|y_i - x_i'\beta| + \sum_{i \in \{i | y_i < x_i'\beta\}} \omega_i (1-q)|y_i - x_i'\beta| \} \quad (10)$$

The weighted estimation procedure uses the 'predictor-corrector' implementation of the primal-dual algorithm proposed by Lustig et al. (1992). Let us assume the following regularity conditions:

i. The cumulative distribution functions $F_i(Y)$ of input expenditures for a given product mix are absolutely continuous with densities $f_i(Y)$ continuous and uniformly bounded on $]0, +\infty[$ at $\xi = \mu_q(Y|x_i)$;
ii. $\sum_0 = lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i x'i$ exists and is positive definite;
iii. $\sum_1 = lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_i(\xi_i) x_i x'i$ exists and is positive definite;
iv. $Sup_{i=1,\ldots,n} \|x_i\| \sim 0(\sqrt{n})$, as a normalization factor;
Pollard (1991) shows that under conditions i and ii, $\hat{\beta}_q \longrightarrow \beta_q$, the estimator converges in probability. In addition, under the set of conditions i, ii, iii et iv, Pollard (1991) obtains the asymptotic normality:

$$\sqrt{n}(\hat{\beta}_q - \beta_q) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}(0, q(1-q)\Sigma_1^{-1}\Sigma_0\Sigma_1) \quad (11)$$

[10] The weighting ω is introduced by the standard instruction weight into the QUANTREG procedure of the SAS 9.2 software.

Finally, under the equality assumption $f_i(\xi_i) = f_\varepsilon(0)$ and the independence of ξ_i , this result is simplified as follows:

$$\sqrt{n}(\hat{\beta}_q - \beta_q) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}(0, \sigma^2(q)\Sigma_0^{-1}) \qquad (12)$$

with $\sigma(q) = \frac{\sqrt{q(1-q)}}{f_\varepsilon(0)}$.
In addition, under conditions iii and iv, if the errors attached to the i^{th} observation $\varepsilon_i = y_i - x_i'\beta$ are identically and independently distributed, with distributions F_i admitting a density $f = F$ such as $f(F^{-1}(q)) > 0$ in the neighbourhood of q, then Koenker and Bassett Jr (1982) show that:

$$\sqrt{n}(\hat{\beta}_q - \beta_q) \underset{\mathcal{L}}{\rightarrow} \mathcal{N}(0, \omega^2(q, F)\Omega^{-1}) \qquad (13)$$

with $\omega(q, F) = \frac{\sqrt{q(1-q)}}{f(F^{-1}(q))}$ and $\Omega = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i'$.
These results can be used to construct confidence intervals for estimates using three procedures: inverse density function, rank method, or resampling algorithm. The inverse density function estimation is the most direct and the fastest method, but it is sensitive to the hypothesis of identically and independently distributed data (iid). For data that are not iid, the rank method, which calculates confidence intervals by reversing the rank score test, is preferred. However, based on the simplex method, the rank method generates significant computation times for large datasets. The resampling method, based on the bootstrap technique, makes it possible to overcome all assumptions but is unstable for small samples. Given the size of the FADN sample, its non-random selection and the existence of three distinct a priori sub-populations (specialized or non-pig Types of Farming), we opted for the resampling method, based on the use of a Markov chain marginal bootstrap (MCMB) because, without distributional hypothesis on the error term, this method gives robust empirical confidence intervals in a reasonable computing time (He and Hu, 2002).

[11] Vertex Principal Component Analysis

2.4. Symbolic analysis of empirical distributions of specific costs

2.4.1 Principal component analysis of distributions

The PCA of the interval endpoints

In the extension of the principal components analysis (PCA) to interval data proposed by Cazes et al. (1997), called V-PCA[11] , a normalised PCA of the interval endpoint array is carried out. In this way, the vertices of the hyper-rectangles are vectors of $\mathbb{R}^p$, while the estimates of the conditional quantiles are elements of $\mathbb{R}^{\mathcal{N}}$. Thus, the V-PCA provides a dual representation of the specific empirical cost distributions represented by their estimation intervals, which are the symbolic objects, and conditional quantiles which are the descriptors. As in classical PCA, the proper subspace (optimal for the dual representation) is structured by orthonormal axes $v_m (1 \leq m \leq p)$, maximizing the sum of squares of vertex coordinates $\psi_m = Zv_m$ and satisfying in $\mathbb{R}^{\mathcal{N}}$ the following equations involving the characteristic eigenvector equation v_m and eigenvalues λ_m of the matrix $\frac{1}{N}Z'Z$:

$$\frac{1}{N}Z'Zv_m = \lambda_m v_m \qquad (14)$$

The dual analysis $\mathbb{R}^p$ leads to a similar equation

$$\frac{1}{N}ZZ'w_m = \lambda_m w_m \qquad (15)$$

having the same non-zero eigenvalues but eigenvectors w m such that: $v_m = \lambda_m^{-1/2} Z'w_m$.
The interpretation of the axes of the V-PCA is based on the conditional quantiles (variables of the V- PCA) presenting the strongest contributions. In normalized PCA, the contribution to the inertia of the variable j to the axis m is calculated as the square of the correlation between the factorial axis and the variable (factorial coordinates). The coordinates of the projections of the estimation interval endpoints of the conditional quantiles (vertices s (i)) of the empirical distribution ω_i specific costs (symbolic

object) on the main factorial axes are provided by the relation:

$$\psi_{i,m} = Z_i v_m \tag{16}$$

The representation of the empirical distribution ω_i on the factorial axis m is provided by the projections of the estimation interval endpoints (hyper-rectangle of maximum inaccuracy, HRIM). The projection of the HRIM on a factorial plane provides a maximum imprecision rectangle (RIM) for the empirical distribution represented by the symbolic object ω_i.

In order to avoid the over-dimensioning projection bias of projected rectangles, Chouakria et al. (1998) propose to retain among the representations those whose vertices are best represented by using the relative contribution (CTR) as a criterion defined in cosine terms, that is, for an estimated interval endpoint $s(i)$ of the quantile distribution i:

$$CRT_{s(i),m} = \sum_{j=1}^{p} (z_{s(i),j} v_m)^2 / \sum_{j=1}^{p} z_{s(i),j}^2 \tag{17}$$

The problems of symbolic object representation are studied by Verde and De Angelis (1997) who proposed a better adjustment of convex envelopes.

The PCA's estimate interval centers

Let be $z_i = [\underline{y_i^j}; \overline{y_i^j}]$, the interval estimate of the $j_t h$ conditional quantile for the empirical distribution i of specific costs. This interval estimate can be represented by the data of the couple $(m_{ij}; r_{ij})$ where $m_{ij} = \frac{\underline{y_i^j} + \overline{y_i^j}}{2}$ is the interval midpoint and $r_{ij} = \frac{\underline{y_i^j} - \overline{y_i^j}}{2}$ its radius. The T matrix of the interval data is then constituted by the concatenation of the matrix of the interval estimate midpoints with the matrix of the interval estimate radii.

The PCA of the interval centers (PCA-IC) in $\mathbb{R}^N$, the space of the specific cost distributions, corresponds to the following eigenvector equation:

$$\frac{1}{n} \tilde{M}' \tilde{M} v_m = \lambda_m v_m \tag{18}$$

where $\tilde{M}$ is the matrix $M = [m_{ij}]$ standardized by the standard deviation of the interval centers, v_m and λ_m are respectively the eigenvectors and the eigenvalues associated with the inertia operator $\frac{1}{n}\tilde{M}'\tilde{M}$. It is therefore the diagonalization of the matrix $\frac{1}{n}\tilde{M}'\tilde{M}$ of quantile interval estimates across all specific cost distributions. The centers of quantile intervals estimates[12] are projected on the factorial planes, with possibly the endpoints of the intervals estimates (vertices) as additional projections.

In the vertice-based analysis (PCA-V), these are considered as independent statistical units. In order not to loose information on the size and shape of the hyper-rectangles, Lauro and Palumbo (2000) introduce a constraint of cohesion between the vertices. The method is based on maximizing the variance between symbolic inter-objects. Let A be the Boolean matrix indicating the membership of the N estimation interval ends to the n empirical distributions. The expression of the variance between symbolic objects is given by:

$$\frac{1}{N} Z' A (A'A)^{-1} A' Z \tag{19}$$

If all the empirical distributions have the same number of intervals estimates, they have the same number of vertices $\frac{N}{n} = 2p$ (in our case study, the same number of conditional quantiles were estimated for each empirical distribution), then $A'A = 2^p I_n$.

Let P_A be the orthogonal projector associated with the A matrix on the reference sub-space:

$$P_A = A(A'A)^{-1}A' \tag{20}$$

[12] It should be noted that Markov chain marginal estimation intervals (MBMC) are not symmetrical in contrast to asymptotic estimation intervals, whereas interval-based PCA-IC assumes in its representation the symmetry with respect to the midpoint. Nevertheless, we propose to use the point estimate as a center and to introduce the concept of lower radius and upper radius to locate the endpoints of the MBMC interval.

In the space $\mathbb{R}^N$, the factorial axes of inertia are obtained as a solution of the following eigenvalue equation:

$$\frac{1}{N} Z' P_A Z \tilde{v}_m = \tilde{\lambda}_m \tilde{v}_m \tag{21}$$

where $\tilde{\lambda}_m$ and $\tilde{v}_m$ are the eigenvalues and the eigenvectors associated with the inertia operator $\frac{1}{N} Z' P_A Z$.
The coordinates of the hyper-rectangle associated with the empirical distribution are then computed as follows:

$$\tilde{\psi}_{i,m} = Z_i \tilde{v}_m \tag{22}$$

The analysis in $\mathbb{R}^p$ is equivalent to solving the following equation for the eigenvalues:

$$(A'A)^{-1/2}(A'ZZ'A)(A'A)^{-1/2}\tilde{w}_m = \tilde{\lambda}_m \tilde{w}_m \tag{23}$$

The relative contributions of the variables (CRT) are defined in the same way as for the V-PCA. These CRTs are also used to select the empirical distributions to be represented on the factorial graphs. As in V-PCA, representations of empirical distributions as a symbolic object are constructed by the Maximum Covering Area Rectangular (MCAR). If one compares the V-PCA with the IC-PCA proposed by Cazes et al. (1997), where the PCA is performed only on the interval data centers standardized with the correlation matrix of the centers, several improvements were introduced: first, the data are standardized by the standard deviation of the vertices considered as active units in the analysis, whereas they are considered in the IC-PCA as additional units; more generally, this can be considered as an improvement of the IC-PCA because it can be applied to data constrained by logical or hierarchical relationships.

PCA ranges of interval estimates

Partial PCA can also be used to better emphasize the differences between symbolic objects. The following section shows a partial PCA where the vertices of the hyper-rectangles are centered relative to the Inf value.

In order to only take into account only the sizes and shapes of the hyper-rectangles associated with the descriptions, Lauro and Palumbo (2000) proposed a PCA based on scaling interval data that summarizes useful information to describe the size and shape of symbolic objects, with the following transformation $\mu(z_i^j) = \overline{x_i^j} - \underline{x_i^j}$. The *Description Potential – DP* (De Carvalho, 1992, 1997), is the hyper-volume associated to the description of the i_{th} empirical distribution, domain defined by the cartesian product of the ranges $Z_i = z_i^1 \times \cdots \times z_i^j \ldots z_i^p$ of p parameters associated with the assertion[13] a_i of the symbolic object ω_i. Its measurement is defined by: $\pi(a_i) = \Pi_{j=1}^p \mu(z_i^j)$ where $\mu(z_i^j) = z_i^j / s_j$ is the normalized range relative to the domain $D^j = z_i^j; i \in I$ of the descriptor Z^j. However, if the measure of one of the descriptors tends to zero, then the description potential tends to zero. To overcome this drawback, we use the linear measure of the potential of an assertion a_i of the symbolic object $\omega_i : \sigma(a_i) = \sum_{j=1}^p \mu(z_i^j)$.

Let all the assertions of the empirical distributions of specific costs be $a_1, \ldots, a_i, \ldots, a_n$ and X be the n times p matrix of general term $x_i = \sqrt{z_i^j}$, then the PCA of ranges estimates (range transformation PCA) is defined by the factor decomposition of the total linear description potential $LDP = \sum_{i=1}^n \sigma(a_i)$, allowing a different geometrical representation of the vertices than in the V-PCA.

The transformation of the data into a range corresponds to an affine translation where the minima $\underline{x_i^1}, \ldots, \underline{x_i^j}, \ldots, \underline{x_i^p}$ are all located at the origin. Thus, the search for an optimal representation subspace for the size and shape of each symbolic object is made from a non-centered PCA of maxima $\overline{x_i^1}, \ldots, \overline{x_i^j}, \ldots, \overline{x_i^p}$

This Range Transformation PCA (RT-PCA)

[13] In the formalism introduced by symbolic analysis, an 'assertion' is the formal description of a symbolic object based on a conjunction of properties expressed by variables whose associated functions apply to individuals.

breaks down the criterion

$$LDP = tr(X'X) = tr(XX') = \sum_{i=1}^{n} \sigma(a_i) \quad (24)$$

according to the following characteristic eigenvector equations:

$$X'Xt_m = \mu_m t_m \quad (25)$$

and

$$XX'u_m = \mu_m u_m \quad (26)$$

Thus, the sum of the eigenvalues μ_m associated with the eigenvectors t_m in $\mathbb{R}^n$ and u_m in $\mathbb{R}^p$ corresponds to the factorial decomposition of the linear description potential:

$$\sum_{m=1}^{p} \lambda_m = \sum_{i=1}^{n} \sigma(a_i). \quad (27)$$

The factorial coordinates of the representation of the specific cost distributions in the optimal subspace are given by:

$$\phi_m = Xt_m \quad (28)$$

The absolute contribution (CTA), as the ratio between the factorial coordinate and the eigenvalue, measures the contribution of the empirical distribution of specific costs to the potential of description of the m th factorial axis; it is defined by:

$$CTA_{i,m} = \frac{\phi_{i,m}^2}{\mu_m} \quad (29)$$

The relative contribution (CTR) measures the representation quality of the empirical distribution in the chosen representation factorial subspace:

$$CTR_{i,m} = \frac{\sum_{m=1}^{M*} \phi_{i,m}^2}{\sum_{j=1}^{p} x_{i,j}^2} \quad (30)$$

The interpretation of the factorial axes is performed according to the contributions (factorial coordinates) of the estimated quantiles estimated for the empirical distributions of specific costs, as descriptors of the symbolic objects:

$$CTA_{j,m} = t_{j,m}^2 \quad (31)$$

The RT-PCA can be represented by the projection of the factorial coordinates of the maxima. The distributions described by conditional quantile estimates, share representations in hyper-rectangles similar in size and shape if they are projected in the same neighbourhood. If all the terms of matrix X are positive then the first eigenvector u_1 and the associated factor

$$\phi_1 = Xt_1 \quad (32)$$

are positive (Lauro and Palumbo, 2000). The first major component can therefore be interpreted as a size factor, while the higher order factors order the empirical distributions according to their shape characteristics.

Mixed Strategy PCA of Estimation Intervals

The mixed strategy in principal component analysis of symbolic objects (*SO-PCA*) combines the vertex PCA (V-PCA) and the range PCA (RT-PCA) in a three-step approach to account for differences in scale and shape between empirical distributions of specific costs:

i. PCA ranges to extract the main axes that best represent the scales and forms of empirical distributions of conditional quantiles;
ii. Projection from Z to $\hat{Z} = P_A Z$ in order to take into account the relations between the different extrema, given the order relationships between the different conditional quantiles of the distribution of specific costs ;
iii. PCA of the line projections $\hat{Z}_i$ on the sub-space of optimal representation $\phi = \phi_1, \ldots, \phi_m, \ldots, \phi_{M*}$ by the projector $P_\phi = \phi(\phi'\phi)^{-1}\phi'$.

The mixed analysis strategy is therefore based on the solution of the following eigenvector equation:

$$\begin{aligned} \hat{Z}' P_\phi \hat{Z} &= Z'A(A'A)^{-1/2} P_\phi (A'A)^{-1/2} A'Z s_m \\ &= \rho_m s_m \end{aligned} \quad (33)$$

where the diagonal matrix $(A'A)^{-1}$ is broken down into the product $(A'A)^{-1/2}(A'A)^{-1/2}$

for symmetry reasons, with respectively ρ_m and s_m, eigenvalues and eigenvectors associated with the usual ortho-normality conditions. The interpretation of the results of the analysis depends on the choice of the projection operator P_ϕ, whose diagonal term, interpretable as a normalized weight, is equal to:

$$\phi_i(\phi_i'\phi_i)^{-1}\phi_i' = \sum_{m=1}^{M*} \phi_{i,m}^2/\mu_m \qquad (34)$$

2.4.2 Automatic clustering of empirical distributions of specific costs

All empirical distributions of specific costs $\Omega = \omega_1,\ldots,\omega_i,\ldots\omega_n$ are described as symbolic objects by a set of $p = 6$ descriptors[14] which are the conditional quantiles $X = \tilde{Q}_{0.10}, \tilde{Q}_{0.25}, \tilde{Q}_{0.50}, \tilde{Q}_{0.75}, \tilde{Q}_{0.90} = x_1,\ldots,x_j,\ldots,x_p$. On the basis of interval estimates of conditional quantiles $z_i^j = [Inf = \underline{x_i^j}; Sup = \overline{x_i^j}]$, the dissimilarities between interval estimates of the j_th conditional quantile $z_i = [x_i; \overline{x_i}]$ and $z_k = [x_k; \overline{x_k}]$ respectively associated with the distributions characterizing country i and country k, can be computed according to the following three standards:

- L_1 Metric (sum of absolute differences)[15]:
  $\delta_1(z_i^j, z_k^j) = |\underline{x_i^j} - \underline{x_k^j}| + |\overline{x_i^j} - \overline{x_k^j}|$
- L_2 Metric (sum of quadratic differences)[16]:
  $\delta_2(z_i^j, z_k^j) = \sqrt{(\underline{x_i^j} - \underline{x_k^j})^2 + (\overline{x_i^j} - \overline{x_k^j})^2}$
- L_∞ Metric ((Chebyshev's distance)[17]:
  $\delta_2(z_i^j, z_k^j) = Sup(\underline{x_i^j} - \underline{x_k^j}); (\overline{x_i^j} - \overline{x_k^j})$

For each of these metrics M on $\mathbb{R}$, a dissimilarity between empirical distributions based on the differences between intervals estimates of the conditional quantiles can be calculated according to a quadratic criterion: $d(\omega_i, \omega_k) = (\sum_{j=1}^p \delta_M^2(z_i^j, z_k^j))^{1/2}$.

The matrix of dissimilarities between national empirical distributions of specific costs can be used to directly apply the classical methods of automatic clustering based on dissimilarities such as the minimum ultrametric (*single linkage*), the maximum ultrametric (*complete linkage*), and the centroid, in a way similar to the Ward's method. Among the automatic clustering procedures developed for interval data, Chavent et al. (2007) proposes a divisive hierarchical clustering algorithm on symbolic data (DIVCLUS-T) as an extension of the DIV procedure (Chavent, 1998), valid for both interval data and categorical data. Subsequently, we detail the principles on which the this automatic clustering procedure is based for interval data.

The divisive hierarchical clustering algorithm recursively splits each cluster into two subclusters, starting from the whole set of symbolic objects $\Omega = \omega_1,\ldots,\omega_i,\ldots\omega_n$. At each partition in k symbolic clusters $P_k = C_1,\ldots,C_k$, a cluster has to be divided in order to get a partition P_{k+1}, with $k+1$ clusters, optimizing the selected adequacy criterion based on the inertia.

The inertia of a cluster is defined by $I(C_l) = \sum_{i\in C_l} \mu_i d_M^2(z_i, g(C_l))$ where w_l is the weight of the symbolic object i and $g(C_l)$ is the cluster centroid defined by:

$$g(C_l) = \frac{1}{\sum_{i\in C_l}\mu_i} \sum_{i\in C_l} \mu_i z_i$$

The within-cluster inertia is defined by the sum of the inertias over all clusters:

$$W(P_k) = \sum_{l=1}^{k} I(C_l).$$

The between-cluster inertia is defined by the inertia of the centroids with regards to the g

[14] This choice of a small number of descriptors was made for comparative convenience with some more classical graphic approaches Desbois et al. (2013) and Desbois et al. (2017b); however, like these earlier works, it could be extended without disadvantage to sets of cardinality descriptors $p = 9$ (deciles), or even $p = 99$ (percentiles) if the analysis objectives required it.

[15] Labelled 'Type L1' in SCLUST.

[16] Labelled 'Euclidean' in SCLUST.

[17] Labelled 'Hausdorff' in SCLUST.

overall centroid of Ω, such as follows:

$$B(P_k) = \sum_{l=1}^{k} \mu_k d_M^2(g(C_l), g) \qquad (35)$$

where $\mu_k = \sum_{l=1}^{k} \mu_l$.
For a partition P_k, the total inertia is the sum of the within-inertia with the between-inertia: $I_(\Omega) = W(P_k) + B(P_k)$. Hence, minimizing the heterogeneity (measured by W) is equivalent to maximizing the homogeneity (measured by B).
Generated by the logical binary choice (*yes/no*) to a numerical binary question $Q = [Is \quad X^j \leq c?]$, let us denote by A_l, A_l the induced bipartition of a cluster C_l formed of n_l objects. In order to choose among the $n_l - 1$ possible bipartitions of the C_l cluster, a discriminating criterion can be defined by the following ratio:

$$D(Q) = \frac{B^j(A_l, \overline{A_l})}{I^j(C_l)} = 1 - \frac{W^j(A_l, \overline{A_l})}{I^j(C_l)}, \quad (36)$$

where the between-cluster inertia $B_j(A_l, A_l)$ and the inertia $I^j(C_l)$ are computed with regards to the j^{th} variable of the X matrix. Hence, minimizing the within-cluster inertia $W(A_l, A_l)$ is equivalent to maximizing the between-cluster inertia $B(A_l, A_l)$, and therefore the $D(Q)$ discriminating criterion.
As in Ward method, the *'upper hierarchy'* (Mirkin, 2005) of a partition P_k is indexed by the height h of a cluster C_l , defined par its between-cluster inertia as follows:

$$\begin{aligned} h_(C_l) &= B(A_l, \overline{A_l}) \\ &= \frac{\mu(A_l)\mu(\overline{A_l})}{\mu(A_l) + \mu(\overline{A_l})} d^2(g(A_l), g(\overline{A_l})) \end{aligned} \quad (37)$$

The DIVCLUS algorithm splits the cluster C_l^* that maximises $h(C_l)$, ensuring that the next partition $P_{k+1} = P_k \cup \{A_l, \overline{A_l}\} - C_l$ has the minimum within-cluster inertia value, with respect to the rule

$$W(P_{k+1}) = W(P_k) - h(C_l) \qquad (38)$$

3. Data collection and distributional analysis of specific agricultural costs in the EU

3.1. European RICA, the model, the aggregates and the countries studied

Since its establishment in 1965[18], the European RICA has been defined by European regulations specifying the implementation modalities and their revisions, the most recent being the EC Regulation n ° 1217/2009 published in JOE L328 of 15/12/2009 for an entrance into force on 01/04/2010. Together with the Census of Agriculture and the Structural Surveys, it completes the tripod of the Community achievement (*'Acquis communautaire'*) on agricultural statistics, which makes it possible to define the population of agricultural holdings, to follow the evolution of their productive structures, and finally to evaluate variations in their income. Focused from the outset on monitoring the income of so-called 'professional' farmers and analyzing the economic functioning of their farms, it has gradually established itself as a vital database for ex ante and ex-post analysis of the impact of agricultural policy measures, in particular those related to the reforms of the Common Agricultural Policy (CAP). As underlined (Chantry, 1998 and 2003), the European FADN is the result of a process of adaptation and harmonization of pre-existing national arrangements within the Member States. European FADN as a database is fed by national FADN which despite the harmonization of accounting and technical- economic concepts carried out[19] under the auspices of the Directorate-General for Agriculture (DG Agri), presents a certain number of specificities relating mainly to the sample selection (sampling plan, selection method, economic size thresholds) and the management of the survey. For each Member State, the data for each holding ('record') collected at European level ('Community record') are derived from data collected at national level ('national record').

[18] European regulation n°79/65/CEE, of 15 June 1965.
[19] by the unit L3 in charge of the relations with the national operators of the FADN.

For some Member States such as Belgium and the Netherlands, the national FADN survey questionnaire collects more information than the European FADN questionnaire. Conversely, for other Member States, the 'European file' incorporates as missing data the possibilities of exemption provided for by the Regulation due to limitations or constraints related to national FADN. However, as regards the accounting aggregates used in our work (gross products and specific expenses), the definitions are harmonized in both plant and animal production; the elements of differentiation that can influence estimates via weighting are mainly in the reference population of farms defined as professional (economic size thresholds) and in the sampling methodology (random selection versus quota selection).

Equation (5) is the basic model for estimating conditional quantiles of the direct costs specific to the studied products, including the pig, our product of interest. Dependent variable of the empirical model, denoted by Y, the specific costs[20] are defined as the sum of the following terms:

- Crop-specific inputs, i.e. items of expenditure on seeds and seedlings, fertilizers and amendments, plant protection products, and other crop-specific costs ;
- Livestock-specific inputs, which include herbivore, granivore, and other animal-specific expense items ;
- inputs specific to forestry activities.

Independent variables of the empirical model, for each of the productions implemented by the multi- product farm, the raw products[21] (denoted by X) relate to all plant, animal and animal products, or even forest products, where appropriate, with the following breakdown into fifteen aggregates: wheat, other cereals, industrial crops, protein crops, oilseeds, horticultural productions, fruit, wine, other vegetable or forest products, cattle, pig, poultry, dairy milk, other animal products, other raw products. The sub-populations of farms selected as the basis of estimation are those corresponding to the following European FADN samples: for 2006, the following twelve Member States were selected (Austria, Belgium, Denmark, France, Germany, Hungary , Italy, Netherlands, Poland, United Kingdom and Sweden) together noted EU 12. The weighted conditional quantile estimation is carried out using the SAS software, by the QUANTREG procedure associated with the WEIGHT instruction, for each of the countries but also for each of the dimension classes.

3.2. Distributional characteristics of specific agricultural costs

According to Angrist and Pischke (2009), 'For better or worse, 95% of estimates in econometrics are provided by averages' however 'applied economists want more and more to know what's going on, not just on average, but for the whole distribution, the losers as the winners'. Thus, in many evaluative and prospective studies, it is often useful to be able to compare results across a large number of sub- populations, to reflect the heterogeneity of the populations studied, and to be able to propose more realistic adjustments.

The non-parametric estimate of the density of specific costs by the kernel method (Figure 1) highlights the asymmetry (2,377), indicated by the difference between the median (€ 33,930) and the average (€ 47,446) pushed towards the distribution right tail by the weight of extrema. This asymmetry, in addition to the dispersion of cultivated areas, reveals the underlying heterogeneity in the choice of specific production factors.

For such skewed distributions, it is well known that the median is a better estimator of central tendency than the arithmetic mean, with regards to the mean absolute error. However, very often specific cost distributions are not

[20] The specific costs are recorded by the European FADN under the variable label SE281.

[21] The gross product is defined, with the variations of stock, as the total gross production from which the total of the intra- consumptions is subtracted.

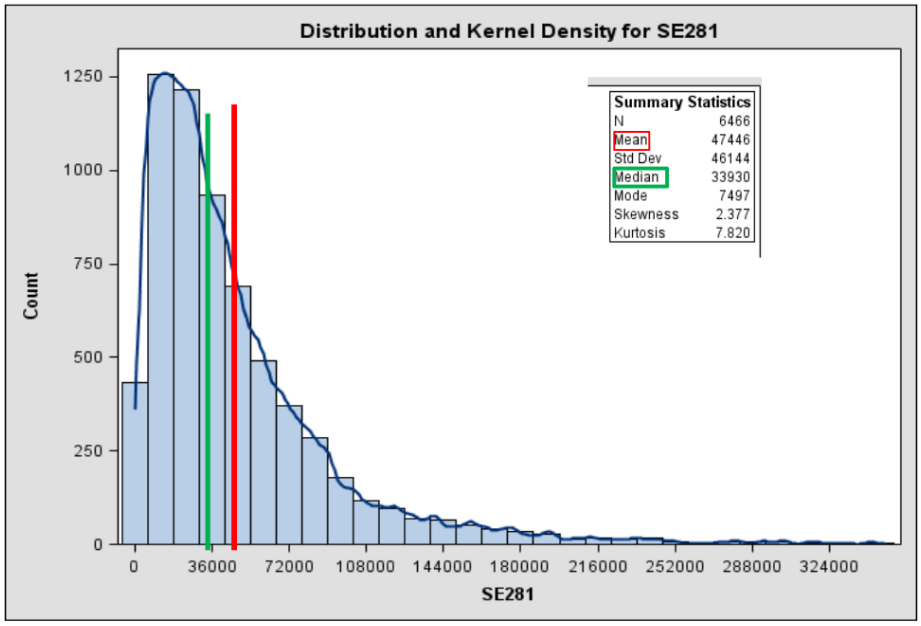

Figure 1: Specific costs, empirical distribution of French FADN, 2006.
Source: author's processing, from French FADN 2006.

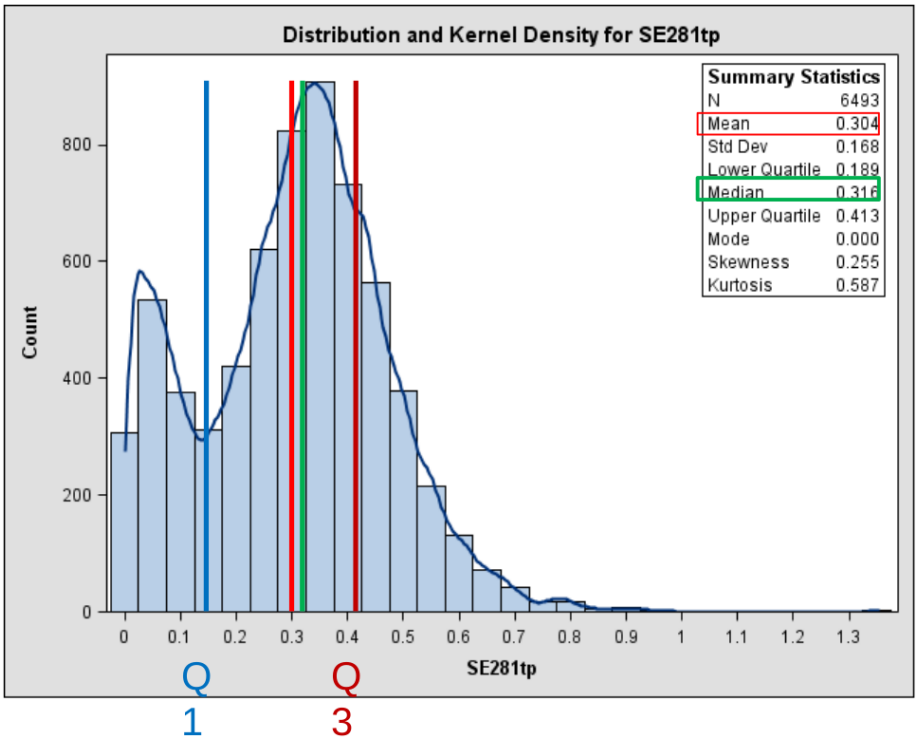

Figure 2: Specific costs per unit of gross product, empirical distribution of French FADN, 2006.
Source: author's processing, from French FADN 2006.

unimodal, as is the case for example for the distribution of specific costs per unit of gross product (Figure 2) that we define as specific costs. Other values may then be needed to better characterize the form of the empirical distribution of specific costs, such as those of the lower quartile (Q1 = 0.189) and higher quartile (Q3 = 0.413) providing more precise information on the empirical distribution of specific costs than that given by the single estimate provided by the average.

First of all, if we compare the countries of the EU12 set in terms of central tendency of specific costs per farm (Table 1), the median varies from € 2,700 (Italy) to € 10,500 (Austria) for a first group also comprising Poland. Then, we distinguish an intermediate group with France and Sweden between € 25,000 and € 29,000, respectively. Finally, the third group comprising the other countries with median costs per farm varying from € 34,500 for Denmark to € 56,000 for the Netherlands, and also including Germany, Belgium and the United Kingdom.

Secondly, if we compare these EU12 countries in terms of dispersion, the coefficient of variation (CoV)[22] is between 134% for Belgium, a country with the most homogeneous production structures (followed by Austria, France and Sweden) and 1023% for Hungary which appears with Italy as the most heterogeneous country.

Although these distributions (Figure 3) are all right-skewed with a large number of extreme values , they differ however in their form: more skewed with large right tails (greater dispersion of the values above the median) as in Italy, Spain, Poland or Hungary, with a skewness between 5 and 6; or less skewed as in Denmark, Austria, the Netherlands, Belgium or France (with a skewness of 1.5 to 2.5), with the United Kingdom having an intermediate skewness of a factor of 4. In addition, the kurtosis varies from a minimum of 17 for the Netherlands, or 23 for Denmark or 37 for Belgium, to a maximum of 210 to 315 for Poland, Spain or Italy.
For very asymmetric distributions with extreme values, the interquartile ratio of dispersion (IRD)[23] is preferable to the CoV for measuring relative dispersion: Austria, France and Germany (between 130 and 150%) have the lowest relative dispersion, while Denmark has the highest relative dispersion of specific costs (420%). By ignoring the extreme values of inherited farms from increasingly marginal collective structures in its production, the IRD reduces the relative dispersion of Hungary to that of Italy.
Thus, the lowest levels and dispersions per farm are found in southern and eastern European countries, with more right skewed distributions, while the highest levels and dispersions are observed in countries in the North and West of Europe, with distributions less skewed than the previous ones. Since specific cost distributions have many extreme values, it is more appropriate to use quartiles to locate the scale of specific costs, as well as the interquartile range or interquartile ratio of dispersion to measure dispersion rather than the mean, standard deviation and coefficient of variation are weight-sensitive to these extreme values.
The ratio of the specific costs to the raw products allows to analyze the productivity of the inputs and to compare it with that of the other production factors. Therefore, it is interesting to be able to describe by structural type the structural differences from the point of view of the specific costs between countries where the production is located: this angle of analysis is therefore developed in the rest of the presentation.

4. Econometric results: national estimates of specific costs

As we have shown in the methodological section, the estimation according to the condi-

[22] Expressed as a percentage, the coefficient of variation reports the value of the standard deviation to the mean : $CoV = \sigma/\mu$

[23] As a ratio of interquartile dispersion at the median level, the quartile dispersion coefficient $IRD = (Q_3 - Q_1)/Q_2$ provides a non-parametric measure of relative dispersion.

Table 1: National distributions of specific costs per farm, EU 12.
Source: author's processing, from EU-FADN 2006. Nota bene: (*) As a ratio of interquartile dispersion at the median level, the interquartile dispersion coefficient $IRD = ((Q_3 - Q_1))/Q_2$ *provides a non-parametric measure of relative dispersion.*

Country	Sample	Mean	CoV	Skewness	Kurtosis	D1	Q1	Median	Q3	D9	IRD*
Austria	1 790	16 870	139 %	6.5	86.7	3 500	5 840	10 430	19 700	37 350	133 %
Belgium	1 040	74 150	134 %	4.5	37.1	11 270	21 980	43 660	90 370	166 150	157 %
Denmark	1 690	112 200	241 %	3.4	23.1	4 670	10 810	34 620	155 180	314 640	417 %
France	6 510	39 310	160 %	5.9	63.1	5 620	12 290	24 910	47 500	83 000	141 %
Germany	6 750	63 420	261 %	6.6	67.1	11 080	19 590	38 170	75 730	137 550	147 %
Hungary	1 690	14 850	1023 %	7.5	85.7	1 070	1 880	4 350	10 240	25 460	192 %
Italy	13 200	12 180	939 %	14.4	314.5	700	1 320	2 670	7 200	20 860	220 %
Netherlands	1 340	124 330	218 %	3.4	17.1	9 870	25 350	56 100	138 300	294 040	201 %
Poland	11 000	7 010	383 %	10.6	209.5	1 470	2 220	3 660	7 180	14 300	136 %
Sweden	850	53 970	187 %	8.5	111.6	7 300	15 840	28 850	67 030	122 760	177 %
United-Kingdom	2 590	82 620	210 %	7.9	97.8	14 300	23 090	44 220	93 150	177 050	158 %
Total	56 180	22 250	570 %	9.0	135.9	1 010	2 010	5 050	18 070	51 490	318 %

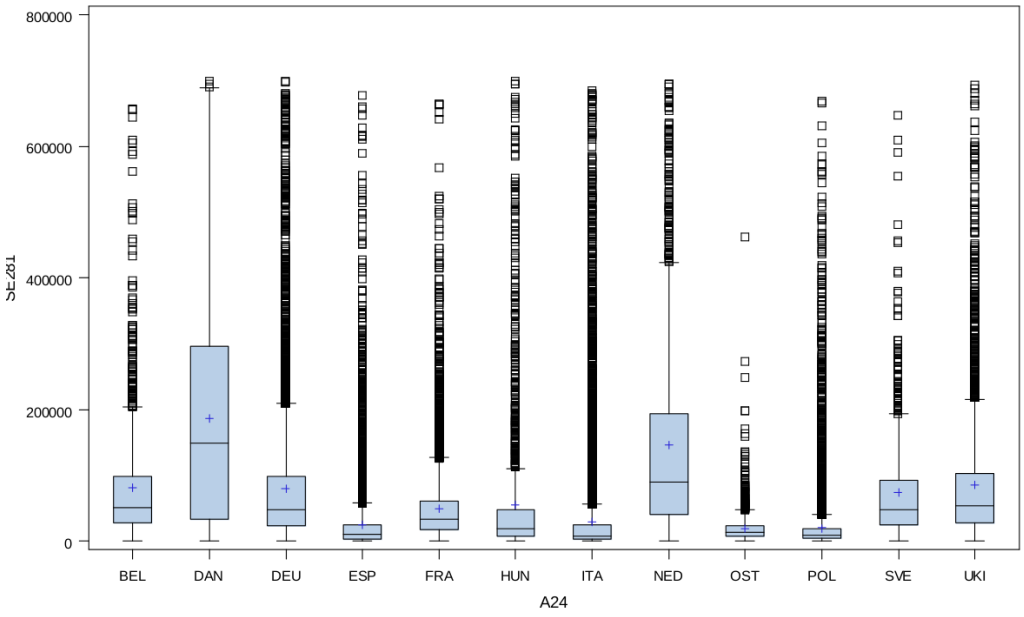

Figure 3: Distribution of the specific costs per farm (SE 281 < € 750,000) by country, EU12.
Source: author's processing, from EU-FADN 2006.

tional quantiles makes it possible to carry out a conditional allocation of the specific costs by products, allowing the comparison of the different workshops within the framework of a multi-product exploitation based on gross margin, its complement to the gross product. We use this conditional allocation to provide specific cost estimates to answer farm competitiveness measurement questions, which are asked by ex-ante or ex-post design and evaluation of different agricultural policy options. In the framework of the FACEPA research project, the choice made by the managers in charge of the Knowledge Based Bio- Economy program (KBBE) of the 7th PCRD was made for feasibility reasons on the three main agricultural commodities that are wheat, milk and pig, produced at a sufficient broad level in the European area to allow cross-country comparisons. Quantile estimates are therefore computed for each of the EU12 Member States in order to test the national differentiation of the productive framework at European level. We have chosen to analyze the estimates obtained for the year 2006[24], in order to compare the results of the conditional quantile approach later with those of the Seemingly Unrelated Regressions Equations (SURE) approach. Initially proposed by Zellner (1962), the latter approach is the standard procedure for estimating the GECOM model of the FACEPA project.

Thus, we analyze the results obtained in particular for the pig, one of the conveniences selected in the framework of the FACEPA project[25].

4.1. Comparative description of gross products for pig between twelve EU countries

In 2010, according to EuroStat estimates[26], the EU-27 accounts for 24.4% of world pig production. The European Union is behind China the second largest producer in the world with 26 million tonnes in 2010. The number of pigs slaughtered in 2010 was 302.6 million heads, or 23% of the herd. The studied countries are among the main producers in terms of tons of carcasses produced, in descending order: Germany (21.58%), Spain (13.33%), France (8.29%), Poland (7.44%), Denmark (6.61%), Italy (6.15%), the Netherlands (5.18%), Belgium (4.26%), the United Kingdom United Kingdom (2.91%), Austria (2.09%), Hungary (1.67%) and Sweden (0.98%), or 80.5% of European production.

Even if the correlation with national statistics is less good[27], the hierarchy of raw products observed within the European FADN (table 2) remains in line with the hierarchy of national statistics on pig production[28], ranking differences exist for the Netherlands (overestimation of 6%), France (overestimation of 4%) and Spain (underestimation of 3%).

4.2. Factor analysis of estimated range distributions

Table 3 presents the main estimates of conditional quantiles (lower decile D1, lower quartile Q1, median Q2, upper quartile Q3, upper decile D9) for pig, derived from quantile regression and ordinary least squares regression (OLS) for the specific costs of agricultural production (accounting aggregate SE281 of the European FADN) from a breakdown of the gross product into fifteen aggregates (cf. III.1), for the subset of 12 European countries selected in 2006. Among the results that may be encountered in estimates of conditional specific cost quantiles for pig (table 3), the estimated gross product shares for pig from the standard FACEPA model, cf. table A3.3 in Kleinhanß et al. (2011), show consistent ranking: in fact,

[24] The analysis over the entire period is the subject of work in progress to adapt the quantile estimates approach to a panel data structure.

[25] The results obtained simultaneously on the other productions are the subject of analyses in progress, conducted in parallel.

[26] According to Focus on the Common Agricultural Policy, Eurostat 2012.

[27] Coefficient of correlation: r = 0.92.

[28] Coefficient of correlation r = 0.98.

[29] Austria was not included in the FACEPA report.

Table 2: Pig, distribution of gross product by country, EU12.
Source: author's processing, from EU-FADN 2006

Country	Population	D1	Q1	Median	Q3	D9	Mean
Austria	29 040	64	126	506	33 031	81 340	24 698
Belgium	5 140	19 243	85 255	184 581	319 312	512 674	228 803
Denmark	6 950	5 544	65 231	213 944	506 288	902 508	354 530
France	13 370	1 620	41 845	123 884	271 555	450 627	194 908
Germany	52 980	1 341	11 946	51 488	135 958	230 833	91 407
Hungary	19 330	761	1 318	2 483	6 843	16 751	16 008
Italia	17 310	400	650	1 530	6 897	132 600	95 450
Netherlands	6 530	45 457	101 343	251 186	460 617	751 500	359 503
Poland	422 190	402	980	2 100	4 624	10 928	5 275
Spain	27 700	632	5 290	14 265	58 749	217 633	79 458
Sweden	3 100	3 209	13 388	54 656	173 638	319 639	124 786
United-Kingdom	2 980	2 739	23 329	133 288	310 707	658 173	228 966
Total	56 180	606 610	368	1 088	10 765	77 316	35 625

in 11 EU Member States[29], the first (Q1) and second (Q2) conditional quartile estimators are significantly correlated with the linearly constrained estimator of FACEPA (the levels are close to the level reached by the OLS estimator[30]).
The visualization of the specific cost estimates is done on the graph in Figure 4, showing the conditional quantile estimates in ascending order for each country. For 2006, this graph of the conditional quantile estimates of specific pig costs by country identifies four types of distributional scales. In the first type, we find Italy (ITA) and Spain (ESP) having a similar behaviour (high inter-quantile growth with an inter- decile difference D1-D9 larger than € 400) despite distinct locations (the minimum of the differentials between respective quantiles is larger than € 200); in the second type, Austria (OST) opposes the previous model with an inter-decile gap D1-D9 of 100 €; in the third type, the United Kingdom presenting a distributional scale with significant inter-quantile growth (inter-decile gap D1-D9 larger than € 300); and in the fourth type, a subset of countries with moderate inter-quantile growth (inter- decile range of between € 200 and € 300). The conditional median (Q2) estimated levels are also a second criterion for distinguishing between these different distributional scales with two subsets: on the one hand, Italy (ITA) and Austria (OST) with median estimates less than € 450; on the other hand, all the other countries whose conditional median estimates are between € 500 and € 600.

Among the differences that can be identified in 2006, let us first note the significant difference between two similar distributional scales with heterogeneous slopes, Spain (ESP)[31] and Italy (ITA)[32], Figure 4 confirming this separation of distributional scales for all conditional quantiles. This is an illustration of the linear model of conditional quantile with heterogeneous slope (cf. above § II.2.ii). Less easy to identify, we secondly note the absence of overlapping distributional scales of Belgium (BEL), Denmark (DAN) and Austria (OST) whose separation of confidence intervals can be seen on Figure 4. Apart from certain differences in precision for the estimation of the upper conditional decile (D9) between Belgium and Austria on the one hand, and Denmark on the other hand, this illustrates the linear model of con-

[30]The rank correlation levels of Spearman are increasing from $corr(SURE, D1) = 0,62$ to $corr(SURE, D9) = 0,69$, comparable to $corr(SURE, MCO) = 0,72$.

[31]For which, the differences between extreme conditional quantiles overlap those between Comunitat Valenciana (Product of Designated Origin - PDO Jamon de Teruel) at the highest costs and Extremadura at the lowest costs.

[32]Whose highest quantile estimates correspond to those recorded in Emilia-Romagna (PDO Prosciutto di Parma) or Venetto (PDO Veneto Berico-Eugeano), which oppose the lowest quartile and decile estimates in Lombardia.

Table 3: Pig, specific costs for 1,000 € of gross product, EU12.
Source: author's processing, from EU-FADN 2006.

Country	D1	Q1	Q2	Q3	D9	OLS
Austria	[347.2 ; 369.2]	[397.3 ; 409.1]	[425.6 ; 447.4]	[463.5 ; 485.3]	[523.1 ; 562.5]	[433.7 ; 442.1]
Belgium	[539.9 ; 566.1]	[561.2 ; 579.8]	[591.5 ; 608.7]	[642.7 ; 674.7]	[684.6 ; 707.4]	[630.9 ; 641.8]
Denmark	[445.2 ; 458.6]	[503.0 ; 515.4]	[558.9 ; 570.9]	[617.7 ;632.1]	[654.7 ; 671.3]	[535.2 ; 542.6]
France	[470.9 ; 493.5]	[509.9 ; 527.7]	[547.9 ; 561.7]	[577.6 ; 594.2]	[610.8 ; 644.6]	[541.6 ; 547.4]
Germany	[444.3 ; 454.7]	[475.6 ; 485.4]	[514.8 ; 526.4]	[567.7 ; 582.7]	[593.4 ; 618.8]	[493.6 ; 502.7]
Hungary	[369.0 ; 451.8]	[459.3 ; 568.1]	[589.2 ; 662.2]	[633.3 ; 681.5]	[647.8 ; 737.0]	[605.1 ; 620.7]
Italy	[116.5 ; 170.1]	[162.2 ; 245.2]	[325.1 ; 386.5]	[559.6 ; 633.0]	[627.9 ; 718.3]	[300.7 ; 307.8]
Netherlands	[487.4 ; 506.2]	[528.2 ; 550.4]	[584.6 ;602.4]	[639.9 ; 661.5]	[676.0 ; 721.6]	[573.2 ; 595.1]
Poland	[471.3 ; 483.3]	[541.8 ; 552.2]	[603.0 ; 618.0]	[655.1 ; 674.7]	[704.5 ; 727.3]	[641.7 ; 648.1]
Spain	[191.5 ; 285.3]	[369.8 ; 441.6]	[552.3 ; 638.5]	[743.8 ; 802.2]	[824.6 ; 893.4]	[449.9 ; 456.7]
Sweden	[396.3 ; 443.9]	[507.2 ; 533.1]	[533.1 ; 578.1]	[547.5 ; 619.5]	[641.7 ; 722.9]	[528.1 ; 543.2]
United-Kingdom	[376.8 ; 559.2]	[548.4 ; 596.0]	[599.2 ; 629.6]	[641.3 ; 712.9]	[723.2 ; 805.4]	[565.7 ; 588.4]

Numbers are presented in [Min ; Max] interval.

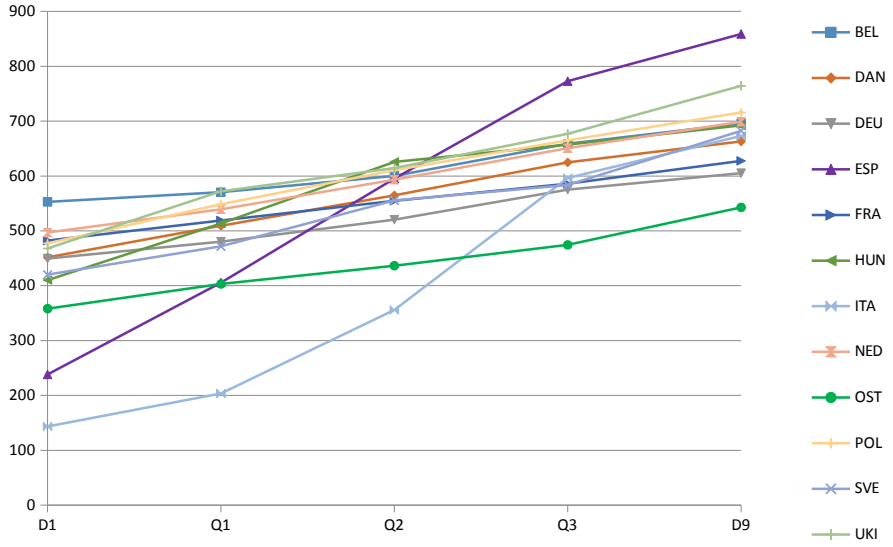

Figure 4: Pig, estimation of conditional quantiles for 12 EU member states (2006).
Source: author's processing, from EU-FADN 2006.

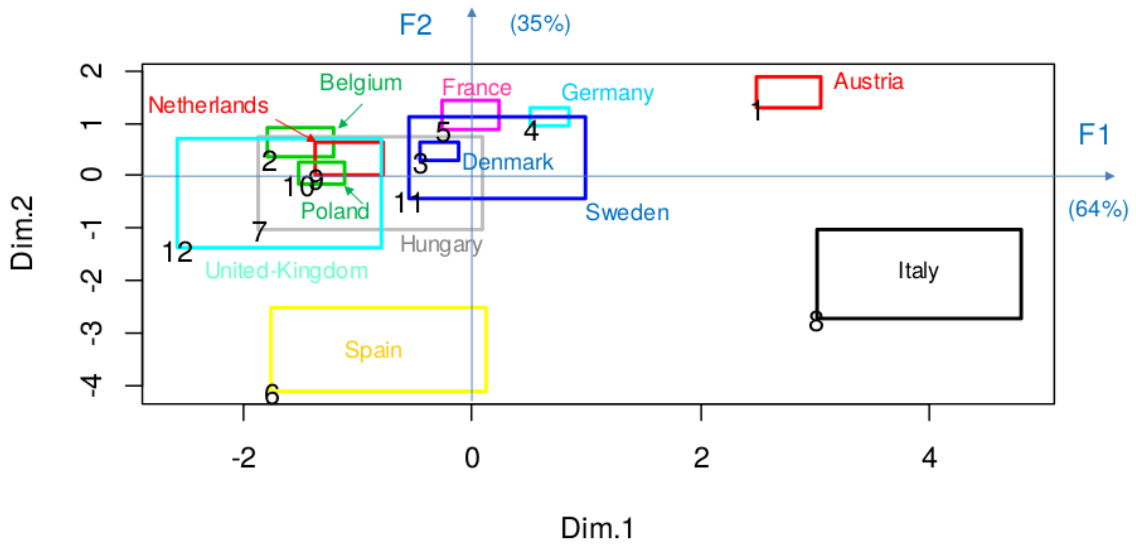

Figure 5: Pig, quantile estimation interval SO-PCA, factorial plane F1xF2 of EU12 countries (2006). *Source: author's processing, from EU-FADN 2006.*

ditional quantile with homogeneous slope (cf. above § II.2.i).

The SO-PCA interval estimates of conditional quantiles allow to specify this distributional structure. The first principal component F1 (Axis Dim1) representing 64% of the inertia, is negatively correlated with the first conditional quantiles (decile D1 and quartile Q1 highly correlated). The second principal component F2 (Axis Dim2), representing 35% of the inertia, is positively correlated with the upper decile (D9) and the third quartile (Q3). The median Q2 is also correlated with the first two major components. The SO-PCA F1xF2 first factorial factor plane, representing 99% of the country variability, makes it possible to identify two distinct groups of countries (Figure 5) differentiated according to the level of the conditional estimate of the first quantiles (D1 and Q1): on the other hand, in F1> 0, Italy with the first quantiles lower than 205 € and, on the other hand in F1 <0, all the other countries for which the first quantiles are larger than 235 €. The second main component makes it possible to distinguish three groups: on the one hand, Austria in F2> 0 with the most homogeneous quantile estimates situated between € 350 and € 430; on the other hand, Spain with the highest estimates (from € 770 for Q3 to € 860 for D9); and all other countries in the quadrant (F1 <0, F2 <0) or close to it.

Thus, the SO-PCA identifies Austria's cost homogeneity model and distinguishes two models of cost heterogeneity, one by lower cost quantiles (D1 and Q1) for Italy, the second by the higher costs (D9 and Q3) for Spain. Finally, taking into account additional parameters could make it possible to better separate two putative subgroups: on the one hand, Denmark, France, Hungary and Sweden; on the other hand, Germany, Belgium, Poland and the United Kingdom.

Hierarchical descending clustering (DIV)[33] allows the cost structure to be specified by country class (Figure 6). First, there is a major distinction in the location of distributional scales: Austria (OST) is separated from other countries by an upper quartile estimate Q3 < € 516.50; Italy stands out with a median estimate

[33]Unsupervised clustering algorithm on the MBMC confidence interval table at 95% quantile estimates (SODAS 2.5 software).

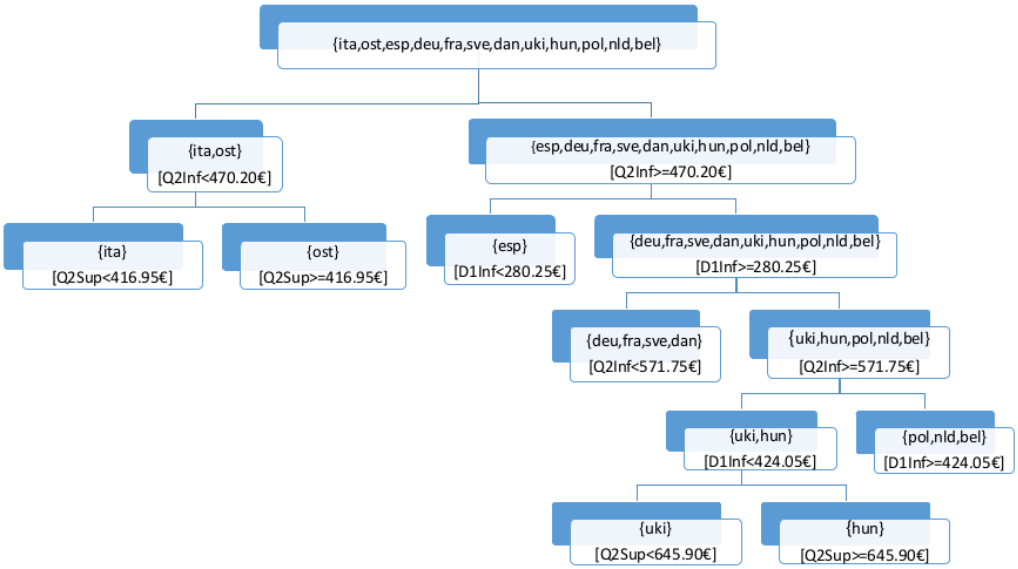

Figure 6: Pig, specific costs for € 1,000 gross product, country clustering, EU12.
Source: author's processing, from EU-FADN 2006.

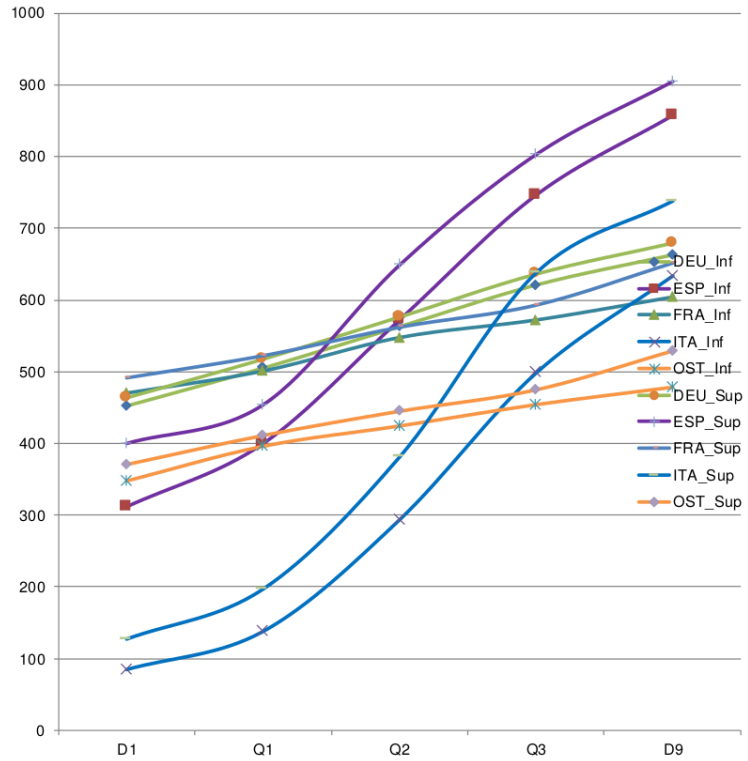

Figure 7: Pig, location shift model (FRA-DEU/OST) versus location-scale shift model (ITA/ESP).
Source: author's processing, from EU-FADN 2006.

Q2 < € 429.75; Spain, Denmark and Sweden are distinguished by an estimate of the first quartile Q1 < € 491; the United Kingdom is characterized by a higher decile estimate is less than this value; on the other hand, among these other countries, a supplementary distinction must be made between those whose estimate of the last decile (D9) exceeds € 738; Hungary is distinguished by an estimate of the first decile (D1) less than € 442.25. The other countries are divided into two sub-groups: on the one hand, France and Germany are characterized by an estimate of the upper quartile (Q3) less than € 636; on the other hand, Belgium, the Netherlands, and Poland, which are distinguished by a quartile estimate (Q3) of over € 636.

This descending hierarchy shows that the set of quantile estimates is ordered by all of these discriminant values, which implies keeping all the parameters describing the distribution, and possibly extending it with a finer quantile scale allowing some of the national distributions to be better distinguished.

5. Results discussion

The heterogeneity of national distributions of specific costs covers the combined effect of different dispersal factors, including the economic dimension of farms, that should be analyzed. In fact, the studied European countries have neither the same composition in terms of the economic dimension of the farms, nor the same thresholds to define a professional farm holding. Thus, the heterogeneity of the quantile estimates of specific costs within national distributions, either in Italy or in Spain, probably covers those of very different production structures both in their economic dimension and in the production technology used.

Regionalised estimates make it possible to specify national situations that are not homogeneous: the Spanish region Communitat Valenciana is distinguished by the maximum estimate of the median quantile of specific costs; in contrast, the Spanish region of Extremadura and the Italian region of Venetto are distinguished by the lowest overall levels of quantile estimates, especially for the median quantile; the Spanish region Andalucia and the Italian region Emilia-Romagna are characterized by higher quantile estimates (Q3 and D9); the central Swedish region Skogs- och mellanbygdslan, the central Hungarian region Kozép-Magyarorszàg, the French region Basse-Normandie, and the German region Sachsen-Anhalt are associated with lower quantile estimates (D1 and Q1) among the highest.

It cannot be ruled out that the high values of some estimates may in some cases come from artifacts related to the estimation methodology for countries where pig production correlates with other production facilities on the farm. Indeed, the size of the pig workshop compared to the other workshops, according to the more or less pronounced productive specialization of the farms, can produce artifacts resulting from the productive correlations at the level of mixed technical-economic orientations to the extent that the weight of the costs specifically related to the other workshops would lead, depending on the hierarchy of specific costs, either to an underestimation bias for minority pig workshops compared to other products with a smaller production detour or conversely to an overestimation bias for productions presenting with the detour of more important production such as pig production.

However, the existence of very high specific costs may also signal the maintenance of technically less efficient producers in less favourable areas because of the existence of comprehensive income support measures (Barkaoui, Daniel and Butault, 2009), or even agri-environmental measures specific to certain productive contexts, in particular those aimed at maintaining agricultural production in certain territories. On the other hand, the lower estimates can point to either the presence

of intensive farms that perform better technically, such as for pig producers in western France, or the presence of productive systems based on less demanding input and output techniques as in piedmont and mountain areas.

6. Conclusions

On the basis of European FADN, we have tested the feasibility of the micro-econometric estimation methodology of the specific production costs using the conditional quantiles, and we have illustrated its relevance to take into account the intrinsic heteroscedasticity of these distributions for one of the major commodities of the European market, the pig. The lessons learned from these analyzes are relatively consistent for the pig: the lower quantiles (D1 and Q1) and, respectively, the higher quantiles (Q3 and D9) are the specific cost parameters that can differentiate national productions according to their cost distributions based on observed regional differences.

The analysis of these estimates makes it possible to identify types of national distributions of specific costs. The main producing countries are located in a two-dimensional graph based on a principal component analysis of the conditional quantile interval estimates that provides an exploratory test for the differences found between national distribution scales according to their respective conditional quantiles. Differences and similarities between countries are exploited using hierarchical top-down clustering to produce country classes with comparable costs. The differences between these groups of countries are delimited by thresholds expressed according to the conditional quantiles in terms of the gross product. These thresholds can be used to segment farm populations to analyze the differential effects of agricultural policy measurement. These analyses therefore allow to identify different models of distributional scale, notably that of the location shift one opposite that of the location-scale shift one (Figure 6).

We hypothesize that the differentiation of these national distributions takes place on the one hand between specialized and input-intensive farms and, on the other hand, input-extensive and/or multiproduct farms. We also consider the perspective of pursuing and valuing the conditional quantile estimation methodology in the context of an input-output analysis of European agriculture. The unit estimates given in terms of the share of the specific costs in the gross product that we have privileged in this paper can be used in the context of the computation of standard gross margins, either at the normative level to provide a statistical basis for the estimation to feed the input-output matrices of the particular agri-food sector to a set of EU countries, or even certain groups of European regions, to implement sensitivity analyses for possible options of agricultural policy through social and environmental accounts matrices (Léon and Surry, 2009). In the current context of the 'greening' of the Pac, the proposed national typology for the pig could be applied to carry out simulations aimed at exploring the relocation of pig production in mountain areas or in intermediate regions.

Acknowledgements

This paper is an adaptation of some of the author's work done during the preparation of his PhD thesis (Desbois, 2015), co-directed by Y. Surry and JC. Bureau, supported by the Farm Accountancy Cost Estimation and Policy Analysis project. of European 7th Framework Program of the European Community (FP7 / 2007 2013, Approval No. 212292). This mention does not imply any approval of the persons and organizations mentioned, the author assuming full responsibility for the text. This work is dedicated to the memory of Yves Surry, Professor Emeritus at the Sverigeslantbruksuniversitet in Uppsala, who passed away recently.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Barrodale, I. and Roberts, F. D. (1973). An improved algorithm for discrete *l*_1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848.

Billard, L. and Diday, E. (2006). Symbolic data analysis : Conceptual statistics and data mining. In *Symbolic Data Analysis*, Wiley Series in Computational Statistics, pages 1–321. Wiley Interscience.

Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin, Springer-Verlag.

Butault, J., Hassan, C., and Reignier, E. (1988). Les coûts de production des principaux produits agricoles dans la cee. *Luxembourg: Office of Official Publications of the European Communities*.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge University Press.

Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique appliquée*, 45(3):5–24.

Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, 19(11):989–996.

Chavent, M., Lechevallier, Y., and Briant, O. (2007). Divclus-t: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52(2):687–701.

Chouakria, A., Diday, E., and Cazes, P. (1998). Vertices principal components analysis with an improved factorial representation. In *Advances in data science and classification*, pages 397–402. Springer.

Dantzig, G. B. (1948). Programming in a linear structure. *Econometrica*, 17:73–74.

De Carvalho, F. (1992). *Méthodes descriptives en analyse de données symboliques*. PhD thesis, Paris Dauphine University, Paris.

De Carvalho, F. (1997). Clustering of constrained symbolic objects based on dissimilarity functions. In *Indo–French Workshop on Symbolic Data Analysis and its Applications*. University of Paris IX, Paris.

Desbois, D. (2015). *Estimation des coûts de production agricoles: approches économétriques*. PhD thesis.

Desbois, D., Butault, J.-P., and Surry, Y. (2013). Estimation des coûts de production en phytosanitaires pour les grandes cultures. une approche par la régression quantile. *Économie rurale. Agricultures, alimentations, territoires*, (333):27–49.

Desbois, D., Butault, J.-P., and Surry, Y. (2017a). Distribution des coûts spécifiques de production dans l'agriculture de l'union européenne: une approche reposant sur la régression quantile. *Économie rurale. Agricultures, alimentations, territoires*, (361):3–22.

Desbois, D., Butault, J.-P., and Surry, Y. (2017b). Distribution des coûts spécifiques de production dans l'agriculture de l'union européenne: une approche reposant sur la régression quantile. *Journées de Recherches en Sciences Sociales*. 34 pages.

d'Haultfoeuille, X. and Givord, P. (2014). La régression quantile en pratique. *Économie et Statistique*, 471(1):85–111.

Divay, J.-F. and Meunier, F. (1980). Deux méthodes de confection du tableau "entrées-sorties". *Annales de l'INSEE*, 37:59–109.

He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97(459):783–795.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311.

Kleinhanß, W., Offermann, F., Butault, J.-P., and Surry, Y. (2011). Cost of production estimates for wheat, milk and pigs in selected eu member states. *Arbeitsberichte aus der vTI-Agrarökonomie*, 7/2011. 31 pages.

Koenker, R. (2005). Quantile regression. *Cambridge University Press*.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1):33–50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, 50(1):43–61.

Koenker, R. and d'Orey, V. (1994). Remark as r92: A remark on algorithm as 229: Computing dual regression quantiles and regression rank scores. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(2):410–414.

Koenker, R. and Zhao, Q. (1994). L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, 3(3-4):223–235.

Lauro, C. N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational statistics*, 15(1):73–87.

Léon, Y. and Surry, Y. (2009). Les effets d'entraînement du complexe agroalimentaire au niveau local. *Politiques agricoles et territoires, Versailles, Éditions Quae*, pages 21–48.

Lustig, I. J., Marsten, R. E., and Shanno, D. F. (1992). On implementing mehrotra's predictor–corrector interior-point method for linear programming. *SIAM Journal on Optimization*, 2(3):435–449.

Madsen, K. and Nielsen, H. B. (1993). A finite smoothing algorithm for linear l_1 estimation. *SIAM Journal on Optimization*, 3(2):223–235.

Mirkin, B. (2005). *Clustering for data mining: A data recovery approach*. Chapman & Hall, CRC Press, London, Boca Raton, FL.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, pages 186–199.

Portnoy, S., Koenker, R., et al. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300.

Verde, R. and De Angelis, P. (1997). Symbolic objects recognition on a factorial plan. *NGUS'97*.

Correspondence: dominique.desbois@inrae.fr.