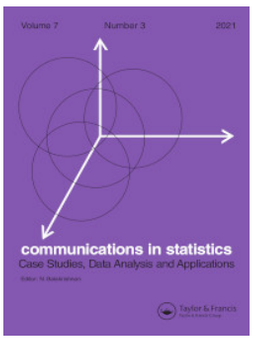
Applying interval PCA and clustering to quantile estimates: empirical distributions of fertilizer cost estimates for yearly crops in European Countries

Dominique Desbois



Taylor & Francis
Taylor & Francis Group

Communications in Statistics: Case Studies, Data Analysis and Applications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucas20

Applying interval PCA and clustering to quantile estimates: empirical distributions of fertilizer cost estimates for yearly crops in European Countries

Dominique Desbois

To cite this article: Dominique Desbois (2021): Applying interval PCA and clustering to quantile estimates: empirical distributions of fertilizer cost estimates for yearly crops in European Countries, Communications in Statistics: Case Studies, Data Analysis and Applications, DOI: 10.1080/23737484.2021.1972875

To link to this article: https://doi.org/10.1080/23737484.2021.1972875

Published online: 14 Sep 2021.

COMMUNICATIONS IN STATISTICS: CASE STUDIES, DATA ANALYSIS AND APPLICATIONS
https://doi.org/10.1080/23737484.2021.1972875

Applying interval PCA and clustering to quantile estimates: empirical distributions of fertilizer cost estimates for yearly crops in European Countries*

Dominique Desbois

UMR Economie publique, INRAE-AgroParisTech, Université Paris-Saclay, Paris Cedex 05, France

ABSTRACT

The decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation but also based on the performances of the productive systems used by farm holdings, assessing their environmental impacts through the cost of fertilizer resources used. This communication uses the symbolic clustering tools in order to analyze the conditional quantile estimates of the fertilizer costs of yearly crop productions in agriculture, as a replacement proxy for internal soil erosion costs. After recalling the conceptual framework of the estimation of agricultural production costs, we present the empirical data model, the quantile regression approach and the interval principal component analysis clustering tools used to obtain typologies of European countries on the basis of the conditional quantile distributions of fertilizer cost empirical estimates. The comparative analysis of econometric results for yearly crops between European countries illustrates the relevance of the typologies obtained for international comparisons to assess land management alternatives based on their impact on agricultural carbon sequestration in soils.

ARTICLE HISTORY
Received November 2020
Accepted August 2021

KEYWORDS
Principal component analysis; hierarchic clustering; interval estimates; quantile regression; input–output model; symbolic data analysis; agricultural production cost; fertilizer; yearly crops; micro-economics

> 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. Economics of agricultural carbon sequestration in soils

Signatory States to the 2015 Paris Agreement have set a common goal of achieving carbon neutrality. According to a logic of net emissions flow adopted by several European countries, France has adopted a Climate Plan in July 2017

CONTACT Dominique Desbois dominique.desbois@inrae.fr UMR Economie publique, INRAE-AgroParisTech, Université Paris-Saclay, 16 rue Claude Bernard, F-75231, Paris Cedex 05, France

*This text is the continuation of some of author's works done preparing his PhD dissertation (Desbois 2015), co-directed by Y. Surry and J.C. Bureau. It is, supported by the "*impActs and feedbackS between climate and Soil affected by EroSion: cost in terms of carbon Storage in Mediterranean regions*" (ASSESS) project (*ANR-16-NME1-0008)* of the *OTE-Med* Eranet.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

with a target of zero net emissions (ZEN) of greenhouse gases, at the 2050 horizon (Quinet 2019).

Carbon sequestration in soils is one of the means proposed to achieve common goals of reducing greenhouse gas emissions: the "4 per 1,000 Initiative: Soils for food security and climate"[1] was launched in 2015 to increase soil organic carbon sequestration. In addition to their soil carbon storage capacity, some sustainable land management technologies can benefit farmers by increasing yields and reducing production costs.

For the European Union, a group of experts from the European Commission on agricultural markets also proposes to encourage farmers to store carbon on the basis of adapted agricultural practices. However, on one hand, the evolution of the CAP's regulatory frameworks by 2020 shows that the proposed instruments alone cannot support large-scale projects on the agricultural soil carbon storage in Europe: in fact, there is very little likely that the future CAP budget is sufficient (Jevnaker and Wettestad 2017). Hence, the decision to adopt one or another of the sustainable land management alternatives should not be based solely on their respective benefits in terms of climate change mitigation but also based on the consideration of the farmers, assessing comprehensively productivity, resource utilization and environmental impact of the productive system.

In the framework of the OTE-Med Eranet, the ASSES project proposes to better assess the economic cost of erosion for farmers by estimating the costs of restoring soil fertility, conceived as an ecosystem service for the benefit of agriculture. The economic evaluation of erosion distinguishes between two types of costs: on-site and off-site costs: in this paper, we focus on the on-site costs and in particular the costs induced by the resulting loss of nutrients. A review of the literature shows that estimates of the soil erosion cost due to nutrient loss are significant and vary greatly depending on the type of crops grown and the production regions. In order to evaluate erosion costs due to nutrient losses, we estimate the production costs of fertilizers using an input/output methodology.

The integration of agriculture in the 27 Member States of the European Union (EU) have raised both in the context of competitive markets as markets subject to regulation, recurring needs in estimating costs of production for major agricultural products, all along the successive reforms of the Common Agricultural Policy (CAP). The analysis of agricultural production costs is a tool for analyzing economic results of farmers: it allows to assess the price competitiveness of farmers, one of the major elements for development and sustainability of food chains in the European regions. To meet the needs of simulations and impact assessment in the various common market organizations, we must be able to provide information on the entire distribution of production costs for the

[1]The "4 per 1,000" initiative aims to unite all public and private stakeholders in achieving a 4‰ annual growth rate for the carbon stored in the first 30 cm of soil, in order to help limit the rise in temperature to +2°C.

assessment of public agricultural policy options. Based on the observation of asymmetry and heterogeneity within the empirical distribution of agricultural inputs, we propose a methodology adapted to the problem of estimating the empirical distributions of fertilizer costs for the main agricultural products in a European context where agricultural holdings remain mainly oriented toward multiple productions (Desbois, Butault, and Surry 2017).

We first present the empirical model for estimating the fertilizer costs of production, derived from an econometric cost allocation approach inspired by Divay and Meunier (1980) using microeconomic data to build an input-output matrix. Then, we introduce the estimation methodology according to the conditional quantiles proposed by Koenker and Bassett (1978). 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 Billard and Diday symbolic approach (Billard and Diday 2006). Then, we interpret the graphs of results from the analysis tools for symbolic data applied to the estimation intervals of the conditional quantiles. Eventually, we conclude on the relevance of this approach applied to yearly crops, suggesting an extension of this type of analysis at the regional level.

2. Conceptual framework and methodological aspects

First, we present the methodology for estimating input costs (Desbois, Butault, and Surry 2017), among which the fertilizer costs. Secondly, we introduce the factorial analysis and the clustering procedure of the estimation intervals in the formalism of the symbolic data.

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

Inspired by Divay and Meunier (1980), the allocation of the sum x_i of the input costs[2] for farm holding i is made by linear decomposition along the gross products Y_i^j of farm holding i for each production j, where u_i is a random vector with a zero mathematical expectation:

$$x_i = \sum_{j=1}^{p} \beta_j Y_i^j + u_i \tag{1}$$

As Cameron and Trivedi (2005), we assume that the data generator process is a linear model with multiplicative heteroscedasticity characterized in matrix form by:

$$x = Y^{'}\beta + u \quad \text{with } u = Y^{'}\alpha \times \varepsilon \quad \text{and} \quad Y'\alpha > 0 \tag{2}$$

where $\varepsilon \sim iid[0, \sigma]$ is an identically and independently distributed random-vector of zero mean and constant variance σ^2. Under this assumption,

[2]Throughout this text, we follow the classical convention in economics using the *x* symbol for inputs and the *Y* symbol for outputs.

$\mu_q(x|Y,\beta,\alpha)$, the q^{th} conditional quantile of the production cost x, conditioned by Y and the α and β parameters, is derived analytically as follows:

$$\mu_q(x|Y,\beta,\alpha) = Y^{'}\left[\beta + \alpha \times F_{\varepsilon}^{-1}(q)\right] = Y^{'}\gamma \tag{3}$$

where F_ε is the cumulative distribution function (CDF) of the errors.

The technical coefficient for the j^{th} product of the q^{th} quantile of the fertilizer cost is defined by the j^{th} component of the multivariate slope vector:

$$\gamma^j(q) = \left[\beta + \alpha \times F_{\varepsilon}^{-1}(q)\right]^j \tag{4}$$

Following D'Haultfoeuille and Givord (2014), three models can be *derived*:

i) $x = Y^{'}\beta + u$ with $u = K\epsilon$, homoscedastic errors $V(\epsilon|Y) = \sigma^2$, denoted as the *location-shift* model, *i.e.* the linear model of conditional quantile with homogeneous slopes; while $Y^{'}\alpha = K$ is constant, the conditional quantiles $\mu_q(x|Y,\beta,\alpha) = Y^{'}\beta + KF_e^{-1}(q)$ get all the same β slope, but differ only by a constant gap, growing as q, the quantile order, increases;

ii) $x = Y^{'}\beta + (Y'\alpha)\varepsilon$ and $Y^{'}\alpha > 0$ with heteroscedastic residuals, referred as the *location-scale shift* model, *i.e.* the linear model of heterogeneous conditional quantile slopes. The case where $x^{'}\alpha > 0$ corresponds to heterogeneous slopes as growing functions of q;

iii) $X = Y^{'}\gamma_\xi$ with ξ random variable independent of Y following a uniform distribution over the interval [0,1] such as $\xi \longrightarrow Y'\gamma_\xi$ be strictly increasing whatever Y, designated as the random coefficient model. ξ corresponds to a random component determining the rank of the individual within the distribution of X. Under the strong distributional hypothesis of rank invariance, the random coefficient γ_q represents the effect of a marginal change in Y for agricultural holdings located at the q^{th} quantile of the ξ distribution. This distributional assumption of rank invariance means that median farms in terms of input productivity would maintain the $q = 0.5$ rank, regardless of the different levels of production Y_i registered for the i^{th} farm holding.

2.2. The procedures for estimating and testing conditional quantiles

The quantile regression is defined for each quantile of order q as the solution of a problem minimizing the sum of weighted absolute deviations (L_1 norm):

$$\hat{\beta}(q) = \underset{\beta\in\Re^p}{\arg\min}\left\{\sum_{i\in\{i/x_i\geq y_i'\beta\}} q\left|x_i - y_i'\beta\right| + \sum_{i\in\{i/x_i<y_i'\beta\}}(1-q)\left|x_i - y_i'\beta\right|\right\} \tag{5}$$

can be written in matrix form:

$$\hat{\beta}(q) = \arg\min_{\beta \in \Re^p} \left\{ qe'(X - Y'\beta \geq 0)\delta^1[X - Y'\beta] + (1-q)e'(Y'\beta - X \geq 0)\delta^1[Y'\beta - X] \right\} \quad (6)$$

with $e(\mathbf{X} - \mathbf{Y}'\boldsymbol{\beta} \geq 0)$, indicator vector of farms i such as $x_i - y'_i\beta \geq 0$, and δ^1, vector of absolute deviations.

Then, the linear optimization problem solving methods developed for the L_1 regression easily extend to quantile regression (Koenker and d'Orey 1994). Although the simplex method (Dantzig 1949) has an algorithmic complexity in $O(n^6)$, the Karmarkar's, method of the "interior-point" (Karmarkar 1984) is in practice preferable as soon as the sample size becoming large, because of its reduced algorithmic complexity to $O(n^{3.5})$. For large samples, Portnoy and Koenker (1977) have shown that a combination of the "interior-point" algorithm and a smoothing algorithm for the objective function by Madsen and Nielsen (1993) makes quantile regression calculations competitive with those of least squares regression.

The weighted conditional quantiles have been proposed by Koenker and Zhao (1994) as L-estimates[3] in linear heteroscedastic models. The $W = \{w_i, i = 1, \ldots, n\}$ weighting of the observations leads to a quantile regression scheme solving the following minimization problem:

$$\hat{\beta}_\omega(q) = \arg\min_{\beta \in \Re^p} \left\{ \sum_{i \in \{i/x_i \geq y'_i\beta\}} w_i q \left| x_i - y'_i\beta \right| + \sum_{i \in \{i/x_i < y'_i\beta\}} w_i(1-q) \left| x_i - y'_i\beta \right| \right\} \quad (7)$$

The weighted estimation procedure uses the "predictor–corrector" implementation of the primal–dual algorithm proposed by Lustig, Marsden and Shanno (1992).

Given the size of the Farm Accounting Data Network (FADN) sample, its nonrandom selection and the existence a priori of distinct subpopulations (e.g. specialized types of farming), we opted for the resampling method, based on the Markov Chain Marginal Bootstrap (MCMB) technique. Without distributional assumption, this method yields robust empirical confidence intervals in a reasonable computation time (He and Hu 2002).

For a given product j_0 such as yield crops and the l^{th} European country, the estimation interval of technical coefficients for the q^{th} conditional quantile of the fertilizer costs

$$z_l^q = \left[Inf\_\hat{\gamma}_l^{j_0}(q); Sup\_\hat{\gamma}_l^{j_0}(q) \right] = \left[\underline{z_l^q}; \overline{z_l^q} \right] \quad (8)$$

is obtained by MCMB.

[3] An L-estimate is an estimate defined by a linear combination of ordinal statistics.

2.3. Symbolic PCA of the fertilizer cost distributions

The symbolic approach has been introduced by Diday (2006) in order to take in account several values rather a single one attached to a variable into the framework of exploratory methods of data analysis. Within this conceptual framework of symbolic data analysis, the extension of principal component analysis (PCA) to interval data was initially proposed by Cazes et al. (1997) and later improved by Chouakria, Diday, and Cazes (1998) with the Vertex and the Center methods using either the vertices or the center of the hyper-rectangle defined by interval values as a multidimensional support for the initial PCA. In this paper, we propose to assess different PCA variants around the Vertex or the Center Methods, proposed by Garro and Rodriguez (2019) in order to maximize the variance of the projections or to minimize the distance between the vertices and the projections of the hyper-rectangle, on the basis of distributional data.

As symbolic objects, the L national distributions $\Omega = \{\omega_1, \cdots, \omega_l, \cdots \omega_L\}$ are described by a set of $Q = 5$ descriptors[4], which are the estimation intervals of $\left\{z^{0.10}, z^{0.25}, z^{0.50}, z^{0.75}, z^{0.90}\right\}$, coding for the D1 and D9 deciles combined with the three quartiles Q1, Q2 and Q3.

Let define the set of $L \times Q$ "within interval"-value matrices,

$$\mathcal{M} = \left\{ Z \in M_{L\times Q} \,|\, z_l^q \in \left[\underline{z_l^q}; \overline{z_l^q}\right] \right\}.$$

2.3.1. The center-PCA of the interval distribution for quantile estimates

Let us define $U \in M$, the center-interval matrix of Z, by:

$$U = \left[U^1, \cdots, U^q, \cdots, U^Q\right] = \begin{bmatrix} u_1^1 & \cdots & u_1^Q \\ \vdots & u_l^q & \vdots \\ u_L^1 & \cdots & u_L^Q \end{bmatrix} \text{ with } u_l^q = \frac{\overline{z_l^q} + \underline{z_l^q}}{2};$$

$$V = \begin{bmatrix} v_1^1 & \cdots & v_1^Q \\ \vdots & v_l^q & \vdots \\ v_L^1 & \cdots & v_L^Q \end{bmatrix} \text{ with } v_l^q = \left[\frac{\underline{z_l^q} - \hat{\mu}^q}{\sqrt{L}\hat{\sigma}^q}; \frac{\overline{z_l^q} - \hat{\mu}^q}{\sqrt{L}\hat{\sigma}^q}\right]$$

where $\hat{\mu}^q$ and $\hat{\sigma}^q$ are, respectively, the mean and the standard deviation of the q^{th} column vector U^q of the matrix U.

According to Cazes et al. (1997), the interval principal components are defined by the following equations:

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

$$\underline{\varphi}_l^q = \sum_{k=1,K;\zeta_k^q<0} \left(\overline{u}_l^k - \hat{\mu}^k\right)\zeta_k^q + \sum_{k=1,K;\zeta_k^q\geq 0} \left(\underline{u}_l^k - \hat{\mu}^k\right)\zeta_k^q \tag{9}$$

$$\underline{\varphi}_l^q = \sum_{k=1,K;\zeta_k^q<0} \left(\underline{u}_l^k - \hat{\mu}^k\right)\zeta_k^q + \sum_{k=1,K;\zeta_k^q\geq 0} \left(\overline{u}_l^k - \hat{\mu}^k\right)\zeta_k^q \tag{10}$$

where ζ_k^q is the q^{th} coordinate of the k^{th} eigenvector of $U'U$, the variance-covariance matrix of U.

According to Rodriguez, Diday, and Winsberg (2000), the pattern of duality in the center-PCA implies the following relationships:

$$\underline{\varphi}_h^q = max\left[\sum_{k=1,\ldots,Q;\zeta_k^q<0} \overline{v}_h^k \acute{\zeta}_k^q + \sum_{k=1,K;\zeta_k^q\geq 0} \underline{v}_h^k \acute{\zeta}_k^q; -1\right] \tag{11}$$

$$\overline{\varphi}_h^q = min\left[\sum_{k=1,\ldots,Q;\zeta_k^q<0} \underline{v}_h^k \acute{\zeta}_k^q + \sum_{k=1,K;\zeta_k^q\geq 0} \overline{v}_h^k \acute{\zeta}_k^q; 1\right] \tag{12}$$

where $\acute{\zeta}_k^q$ is the q^{th} coordinate of the h^{th} eigenvector of VV' the inertia matrix of V, and $\overline{v}_h^k = \underset{l_h \in L}{Sup}\left\{v_{l_h}^k\right\}$ respectively $\underline{v}_h^k = \underset{l_h \in L}{Inf}\left\{v_{l_h}^k\right\}$. This duality pattern determines the infimum and the supremum of the hyper-rectangle defined by the projection of the qth vector of V in the direction of the h^{th} principal component of VV'.

2.3.2. The "best point" PCA of the interval distribution for quantile estimates

In the bivariate case ($q = 2$) with the Q1 ($Z^{0.25}$) and Q3 ($Z^{0.75}$) quartiles, the vertex submatrix $\overline{\underline{Z}}_l$ associated with the l^{th} country, is defining the $n = 2^q = 4$ vertices of a Q1 by Q3 rectangle H_l (cf. Figure 1):

$$\overline{\underline{Z}}_l = \begin{Bmatrix} \overline{\underline{z}_l^{0.25}} & \overline{\underline{z}_l^{0.75}} \\ \overline{z_l^{0.25}} & \overline{z_l^{0.75}} \\ z_l^{0.25} & z_l^{0.75} \\ \overline{\overline{z_l^{0.25}}} & \overline{\overline{z_l^{0.75}}} \\ \underline{z_l^{0.25}} & \overline{\overline{z_l^{0.75}}} \end{Bmatrix} \tag{13}$$

Via a similar process for $l = 1, \ldots, L$, let us define $\overline{\underline{Z}} = \left(\overline{\underline{Z}}_1, \ldots, \overline{\underline{Z}}_l, \ldots, \overline{\underline{Z}}_L\right)'$, the vertex-interval matrix, by its submatrices $\overline{\underline{Z}}_l$ of the l^{th} country ω_l, represented by

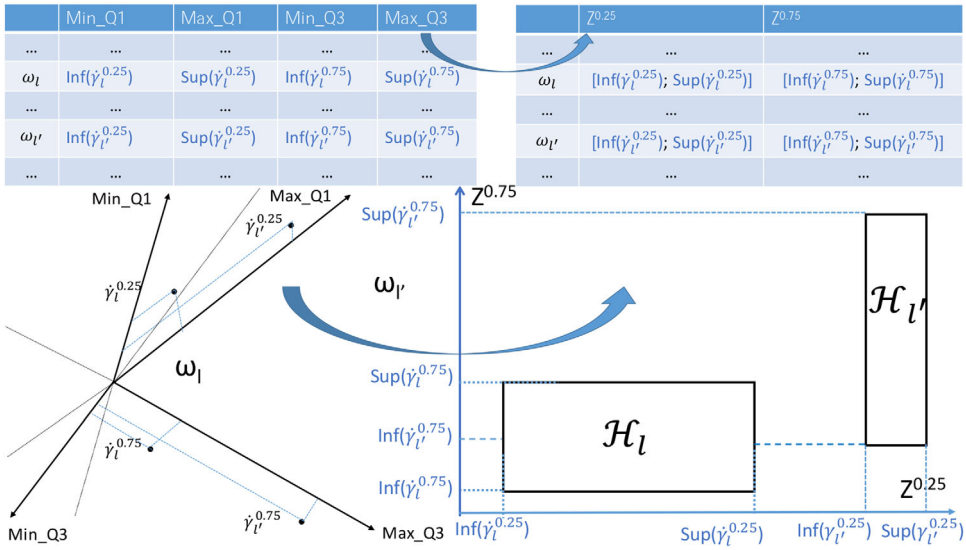


Figure 1. The symbolic coding of the estimation intervals for the technical coefficients of the lower (Q1) and higher (Q3) quartiles of fertilizer costs.

H_l the hyper-rectangle build with $n_l = 2^{q_l}$ vertices of the q_l non-trivial intervals.

$$\overline{Z}_l = \begin{bmatrix} \overline{z^1_{s_1}} & \cdots & \overline{z^q_{s_1}} & \cdots & \overline{z^{q'}_{s_1}} & \cdots & \overline{z^Q_{s_1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{z^1_{s_h}} & \cdots & \underline{z^q_{s_h}} & \cdots & \overline{z^{q'}_{s_h}} & \cdots & \underline{z^Q_{s_h}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{z^1_{s_{n_l}}} & \cdots & \underline{z^q_{s_{n_l}}} & \cdots & \underline{z^{q'}_{s_{n_l}}} & \cdots & \underline{z^Q_{s_{n_l}}} \end{bmatrix}$$

In this way, the vertices of hyper-rectangles $\mathcal{H}_l$ are vectors of $\mathbb{R}^Q$, while the Q estimates of the conditional quantiles are elements of $\mathbb{R}^N$, with $N = \sum_{l=1}^{L} n_l$.

Let us apply PCA to $Z \in \mathcal{M}$, a within-interval value matrix. The k^{th} principal component of the l^{th} country is given by:

$$\psi_l^k = \sum_{q=1}^{Q} \left(z_l^q - \mu_q \right) w_q^k \tag{14}$$

where $\mu_q = \frac{1}{L} \sum_{l=1}^{L} z_l^q$ is the average of the qth conditional quantile of cost estimates and w_q^k, the q^{th} coordinate of the k^{th} eigenvector of the variance-covariance matrix of Z.

Defining the supplementary normalized vertex $\tilde{Z} = \left(\tilde{Z}_1, \ldots, \tilde{Z}_l, \ldots, \tilde{Z}_L\right)'$ by its l^{th} submatrix, where σ_q is the standard deviation of Z^q

$$\tilde{Z}_l = \begin{bmatrix} \frac{\overline{z_l^1} - \mu_1}{\sqrt{L}\sigma_1} & \cdots & \frac{\overline{z_l^q} - \mu_q}{\sqrt{L}\sigma_q} & \cdots & \frac{\overline{z_l^Q} - \mu_Q}{\sqrt{L}\sigma_Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\underline{z_l^1} - \mu_1}{\sqrt{L}\sigma_1} & \cdots & \frac{\underline{z_l^q} - \mu_q}{\sqrt{L}\sigma_q} & \cdots & \frac{\underline{z_l^Q} - \mu_Q}{\sqrt{L}\sigma_Q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\underline{z_l^1} - \mu_1}{\sqrt{L}\sigma_1} & \cdots & \frac{\overline{z_l^q} - \mu_q}{\sqrt{L}\sigma_q} & \cdots & \frac{\overline{z_l^Q} - \mu_Q}{\sqrt{L}\sigma_Q} \end{bmatrix}$$

Each s_h vertex of hyper-rectangle of the lth national distribution of fertilizer cost estimate $\tilde{Z}_l$ can be projected on the principal components of the Z-PCA, with the following k^{th} coordinates:

$$c_{s_h}^k = \sum_{q=1}^{Q} \tilde{z}_{s_h}^q w_q^k \tag{15}$$

According to Rodriguez (2000), the minimum and maximum of the k^{th} coordinate for each estimation interval for the l^{th} country can be computed as follows:

$$\underline{\psi}_l^k = \underset{s_h = 1, \ldots, n_l}{Inf}\left\{c_{s_h}^k\right\} = \sum_{\left\{q | w_q^k < 0\right\}} \left(\overline{z}_l^q - \mu_q\right) w_q^k + \sum_{\left\{q | w_q^k \geq 0\right\}} \left(\underline{z}_l^q - \mu_q\right) w_q^k \tag{16}$$

$$\overline{\psi}_l^k = \underset{s_h = 1, \ldots, n_l}{Sup}\left\{c_{s_h}^k\right\} = \sum_{\left\{q | w_q^k < 0\right\}} \left(\underline{z}_l^q - \mu_q\right) w_q^k + \sum_{\left\{q | w_q^k \geq 0\right\}} \left(\overline{z}_l^q - \mu_q\right) w_q^k \tag{17}$$

Let us denote t_h the eigenvectors of $\tilde{Z}\tilde{Z}'$ for $h = 1, \ldots, H$, the coordinate of the qth quantile estimates on the hth principal component is given by

$$r_h^q = \sum_{s=1}^{N} \tilde{Z}_q^{'s} t_s^h \tag{18}$$

According to Garro and Rodriguez (2019), by projection of the q^{th} quantile estimate on the h^{th} principal component in the direction of t_h, the infimum and supremum values of the hyper-rectangle H_l are computed as follows:

$$\underline{\chi}_l^q = \underset{s_l = 1, \ldots, n_l}{Inf}\left\{r_{s_l}^q\right\} = \sum_{\left\{s | t_s^h < 0\right\}} \overline{\tilde{z}'}_l^s t_s^h + \sum_{\left\{s | t_s^h \geq 0\right\}} \underline{\tilde{z}'}_l^s t_s^h \tag{19}$$

$$\overline{\chi}_l^q = \underset{s_l = 1, \ldots, n_l}{Sup} \left\{ r_{s_l}^q \right\} = \sum_{\left\{ s | t_s^h < 0 \right\}} \underline{\tilde{z}'_l{}^s} t_s^h + \sum_{\left\{ s | t_s^h \geq 0 \right\}} \overline{\tilde{z}'_l{}^s} t_s^h \tag{20}$$

Thus, the Z-PCA provides a dual representation of the fertilizer empirical cost distributions represented by their estimation intervals, which are the symbolic objects, and conditional quantiles which are the descriptors of these symbolic objects.

Let us define $(Z) = \left\{ w_1^Z, \ldots w_s^Z, \ldots, w_S^Z \right\}$, the orthonormal basis of eigenvectors issued from the variance-covariance matrix of Z, and the function

$$\Psi(Z) : \mathcal{M} \to \mathbb{R}^+ \cup \{0\} \text{ based on the Euclidean norm } \|.\|,$$

$$\text{such as } \Psi(Z) = \sum_{l=1}^{L} \left\| \tilde{Z}_l - Pr_{(Z)} \left(\tilde{Z}_l \right) \right\|^2$$

and where $Pr_{(Z)} \left(\tilde{Z}_l \right)$ is the projection of the sub-matrix $\tilde{Z}_l$, coding the vertices of the hyper-rectangle H_l, on (Z), as an appropriate orthonormal basis.

The interval-valued matrix Z^* that solves the optimization problem

$$Min \underset{Z \in M}{\Psi(Z)} \tag{21}$$

is estimated through Procedure (below), using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970) in order to find the minimal distance to $\tilde{Z}$, the vertex matrix.

Procedure 1. Minimizing the squared distance

Input:

$Z \in \mathcal{M}$, a $L \times Q$ matrix with s principal components;

TOL, a numerical threshold of tolerance;

ITER, a maximum number of iterations.

i) $Z \leftarrow U$, the center matrix as the initial value;

ii) $Z^* \leftarrow$ *lbfgs* (Z, objective $= \Psi(Z)$, *TOL*, *ITER*) ;

iii) Compute the $\left[\underline{\psi_L^*}; \overline{\psi_L^*} \right]$ coordinates, applying (16) and (17) duality relationships;

iv) **Return** $\left[\underline{\psi_L^*}; \overline{\psi_L^*} \right]$

Source: adapted from Garro and Rodriguez (2019).

Nota bene: *lbfgs* is a function implementing BFGS algorithm, from the *nloptr* package by Ypma (2020).

Let us define the function $\Lambda(Z, s) : \mathcal{M} \times N \to \mathbb{R}^+$ such as $\Lambda(Z, s) = \sum_{h=1}^{s} \lambda_h$, the variance of the first s components issued from the PCA of Z, where λ_h is the h^{th} eigenvalue associated to the h^{th} eigenvector of (Z).

The interval-valued matrix Z^s that solves the optimization problem

$$Max \begin{array}{l} \Lambda(Z,s) \\ Z \in \mathcal{M} \end{array} \tag{22}$$

is estimated through Procedure 2 (below) using the BFGS algorithm, in order to maximize the variance of the first s components.

Procedure 2. Maximizing the variance of the first components
Input: $Z \in M$, a $L \times Q$ matrix, with s principal components;
TOL, a numerical threshold of tolerance;
$ITER$, a maximum number of iterations.

i) $Z \leftarrow \quad U$, the center matrix as an initial value;
ii) $Z^s \leftarrow lbfgs(\mathrm{Z}, \text{objective} = \Lambda(Z,s), TOL, ITER)$;
iii) Compute the $\left[\underline{\chi_L^*}; \overline{\chi_L^*}\right]$ coordinates, applying (19) and (20) duality relationships;
iv) **Return** $\left[\underline{\chi_L^*}; \overline{\chi_L^*}\right]$.

Source: adapted from Garro and Rodriguez (2019).
Nota bene: lbfgs is a function implementing BFGS algorithm, from the *nloptr* package by Ypma (2020).

2.4. Symbolic clustering analysis of the fertilizer cost distributions

The local dissimilarities between country l and country l',associated with these estimation intervals of technical coefficients for the q^{th} conditional quantile, are computed according to the Euclidean distance metric:

$$\delta_M\left(z_l^q, z_{l'}^q\right) = \sqrt{\left(Inf\_\hat{\gamma}_l^{j_0}(q) - Inf\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2 + \left(Sup\_\hat{\gamma}_l^{j_0}(q) - Sup\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2} \tag{23}$$

For this metric M, a global dissimilarity between country l and country l' based on the differences over the national distributions of estimation intervals for the technical coefficients is computed according to the following quadratic criterion:

$$d\left(\omega_l, \omega_{l'}\right) = \left(\sum_{q=1}^{Q} \delta_M^2\left(z_l^q, z_{l'}^q\right)\right)^{1/2}. \tag{24}$$

Given a matrix of dissimilarities between national empirical distributions of fertilizer costs issued from the previous computations, we can use the methods of unsupervised clustering. In a way similar to the Ward's method, Chavent, Lechevallier, and Briant (2007) proposed a divisive hierarchical clustering algorithm on symbolic data (DIVCLUS-T), valid for both interval data and categorical data. Subsequently, we detail for interval data the principles on which the operations of this unsupervised clustering procedure are based.

The divisive hierarchical clustering algorithm recursively splits each cluster into two sub-clusters, starting from the whole set of countries as symbolic objects

$$\Omega = \{\omega_1, \cdots, \omega_l, \cdots \omega_L\}.$$

At each partition in k symbolic clusters $\mathrm{P}_K = \{C_1, \cdots, C_k \cdots, C_K\}$, 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 the kth cluster is defined by $I(C_k) = \sum_{l \in C_k} \mu_l d_M^2 \left(z_l, g(C_k)\right)$ where μ_l is the weight of the lth country and $g(C_l)$ is the cluster centroid defined as:

$$g(C_k) = \frac{1}{\sum_{l \in C_k} \mu_l} \sum_{l \in C_k} \mu_l z_l \tag{25}$$

The intra inertia is defined by the sum of the inertias of the clusters to their centroids:

$$W(P_K) = \sum_{k=1,\ldots,K} I(C_k) \tag{26}$$

The inter inertia is defined by the inertia of the centroids with regards to the g overall centroid of Ω, as follows:

$$B(P_K) = \sum_{k=1,\ldots,K} \mu_k d_M^2 \left(g(C_k), g\right) \text{ where } \mu_k = \sum_{l=1,\ldots,k} \mu_l \tag{27}$$

For a partition P_K, the total inertia sums the intra inertia with the inter inertia:

$$I(\Omega) = W(P_K) + B(P_K) \tag{28}$$

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 $\Psi = [Is\ z^q \leq c?]$, let us denote $\left\{A_k, \overline{A_k}\right\}$ the induced bipartition of a cluster C_k formed of n_k objects. In order to choose among the $n_k - 1$ possible bipartitions of the C_k cluster, a discriminating criterion can be defined by the following ratio:

$$D(\Psi) = \frac{B^q\left(A_k, \overline{A}_k\right)}{I^j(C_k)} = 1 - \frac{W^j\left(A_k, \overline{A}_k\right)}{I^q(C_k)} \tag{29}$$

where the inter inertia $B^q\left(A_k, \overline{A}_k\right)$ and the inertia $I^q(C_k)$ are computed with regard to the q^{th} conditional quantile. Hence, minimizing the intra inertia $\mathrm{W}\left\{A_k, \overline{A_k}\right\}$ is equivalent to maximizing the inter inertia $\mathrm{B}\left\{A_k, \overline{A_k}\right\}$ and, as a result, to the $D(\Psi)$ discriminating criterion.

As in Ward method, the "upper hierarchy" (Mirkin 2005) of partition P_K is indexed by the height h of a cluster C_K, defined by its inter inertia as follows:

$$h(C_k) = B\left(A_k, \overline{A_k}\right) = \frac{\mu(A_k)\,\mu\left(\overline{A_k}\right)}{\mu(A_k) + \mu\left(\overline{A_k}\right)} d^2\left(g(A_k), g\left(\overline{A_k}\right)\right) \tag{30}$$

The DIVCLUS-T algorithm splits the cluster C_K^* that maximizes $h(C_K)$, ensuring that the next partition $P_{K+1} = P_K \bigcup \left\{A_K, \overline{A_K}\right\} - C_K^*$ has the minimum intra inertia value, with respect to the rule

$$W(P_{K+1}) = W(P_K) - h\left(C_K^*\right) \tag{31}$$

In order to determine an optimal clustering, we use as the internal quality index for each partition P_K, the log of the determinant ratio computed as follows:

$$\varkappa_K = N \log\left(\frac{det(T)}{det\left(WG^{(K)}\right)}\right) \tag{32}$$

where $T = Z'Z$ is the total scatter matrix (N times the total variance-covariance matrix) and $WG^{(K)} = \sum_{k=1}^{K} W^{(k)}$ the sum of the within-group scatter matrices, $W^{(k)}$ for each group C_k of the partition P_K in K groups.

The optimal score for the quality index is given by the *min_diff* decision rule:

$$K^* = \arg\min_K \{\partial_K - \partial_{K-1}\}$$

with $\partial_K = \varkappa_{K+1} - \varkappa_K$, using procedure *ClusterCrit* proposed by Desgraupes (2017) for needed computations.

3. Results

Based on the gross product, the estimation according to the quantiles provides a conditional allocation of the fertilizer costs by main products, within the framework of a multi-product exploitation. In the framework of the Farm Accountancy Cost Estimation and Policy Analysis project (FACEPA) research project, the managers in charge of the Knowledge Based Bio-Economy project of the 7th EU Framework Program of Research has chosen to focus on the main agricultural commodities produced at a level sufficiently broad at the European level to allow meaningful cross-country comparisons for the twelve European Member States which are the main producers (EU12), choosing 2006 as a baseline for comparison convenience.

We analyze the results obtained in particular for the yield crops about fertilizer inputs. The figures are estimated from a quantile regression of the fertilizer inputs on a decomposition of the gross product into five product aggregates (yearly crops, permanent crops, pasture livestock, off-ground livestock, others) for the set of twelve European countries (UE12) selected on 2006.

Table 1 presents for yield crops the estimation intervals of conditional quantiles (lower decile D1, lower quartile Q1, median Q2, upper quartile Q3, upper decile D9) of the fertilizer inputs of agricultural production.

The pre-visualization of the fertilizer cost estimates is done according to the graph in Figure 2, showing the conditional quantile point estimates in ascending order for each country. This graph of point estimates of conditional quantiles of fertilizer costs for yield crop by country highlights some distributional facts.

Table 1. Yield Crops, estimation intervals for technical coefficients of quantile fertilizer costs for €1 of gross product, EU12.

	C1			C2			C3			C4			C5		
Contr. (%)	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar
Bel	7	11	9	3	1	3	0	0	0	2	3	0	2	2	0
Dan	1	1	1	5	4	1	16	8	5	0	9	58	0	0	1
Deu	0	0	0	10	7	4	16	1	6	1	15	10	1	1	0
Esp	1	1	1	0	0	0	12	3	25	5	17	28	45	31	0
Fra	18	17	20	1	1	3	5	2	6	21	13	0	23	31	46
Hun	11	13	8	11	11	1	0	0	16	0	3	0	15	17	48
Ita	8	7	7	0	0	0	1	0	3	2	3	0	1	0	0
Ned	41	35	40	0	1	0	2	1	11	6	9	0	2	4	4
Ost	0	0	1	1	0	15	26	77	3	48	7	0	0	2	0
Pol	9	10	8	5	2	1	13	2	12	3	17	0	3	3	0
Sve	4	3	3	64	69	70	7	2	2	0	0	3	7	6	0
Uki	1	2	1	0	3	0	3	5	11	11	4	0	2	3	0

Source: Author's processing, from EU-FADN 2006.

Below 3%, the overall level of the Netherlands distribution curbs (*iNED* and *sNED* on Figure 1) is the lowest of the twelve European countries studied, with the exception of the lower bound (*i*) of the first decile (*D1*) in Sweden (*SVE*) which is negative. The Netherlands distribution is also the flattest of the twelve European distributions analyzed, followed by the distributions for Italy and Belgium, which have fairly moderate slopes and overall estimation levels below 13%. The Netherlands distribution illustrates the *location shift* linear model of conditional quantile with homogeneous slopes.

Conversely, the maximum and minimum curves of the Swedish distribution (*iSVE* and *sSVE*) are the steepest (from 1.6% to near 30%), immediately followed by those of France (*iFRA* and *sFRA*) and Poland (*iPOL* and *sPOL*). These three countries illustrate the *location-scale shift* linear model of conditional quantile with heterogeneous slopes.

Next, Hungary (*iHUN* and *sHUN*), Germany (*iDEU* and *sDEU*), Austria (*iOST* and *sOST*), the United Kingdom (*iUKI* and *sUKI*) and Spain (*iESP* and *sESP*) form an intermediate group where, on the basis of this first graph, it becomes difficult to distinguish clear differences between these national distributions.

3.1. The interval PCAs of fertilizer cost estimates

Applying Equations (9) and (10), the "centers" option of the interval PCA shows a correlation circle displaying the estimate quantile coordinates on the first two principal components with the highest negatives correlation for D1, Q1 and Q2 quantiles (Figure 3). The larger fans which indicate the greater infimum-supremum intervals of estimation are found for D1 and Q1 quantiles meanwhile Q2, Q3 and D9 quantiles displays the smallest which indicate the lower interval ranges of estimates.

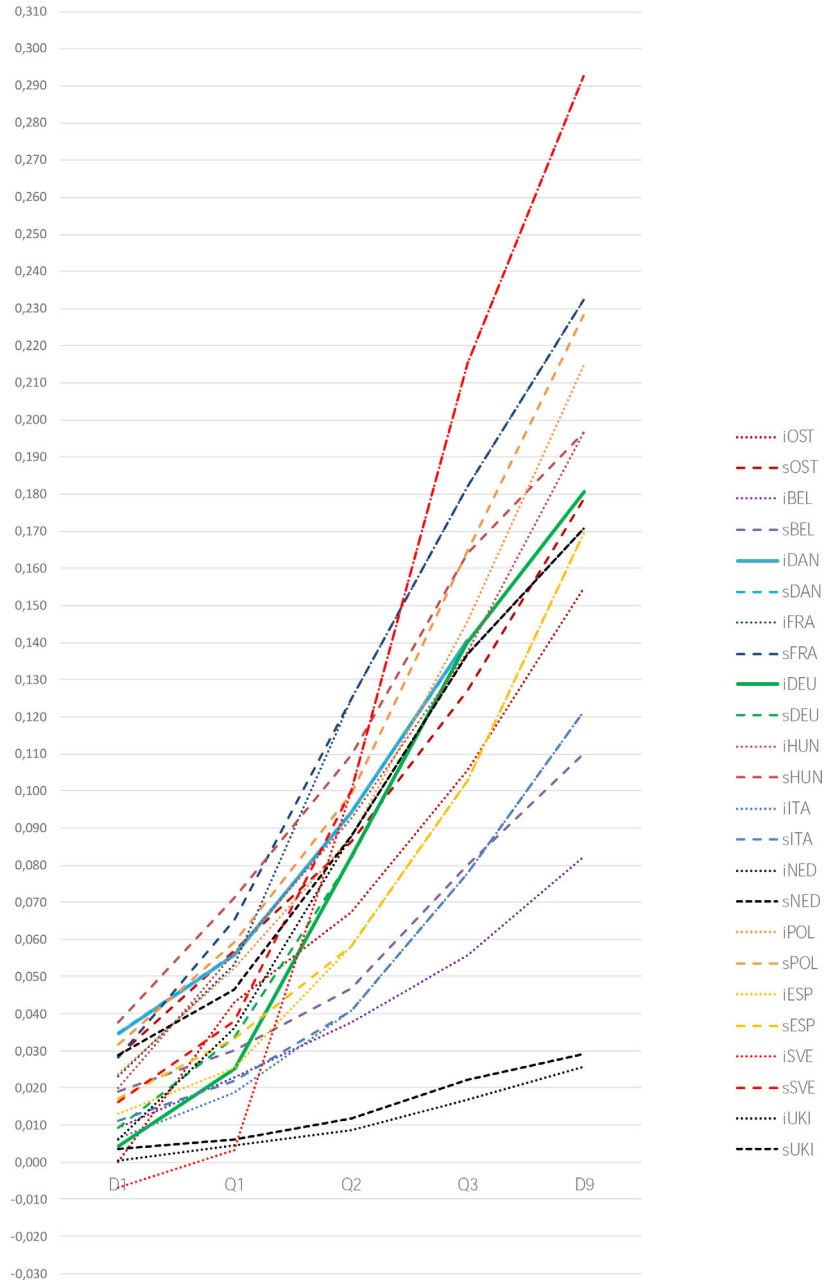


Figure 2. Yield Crops, interval estimation for fertilizer coefficients of conditional quantiles for 12 EU member States; *iOST* stands for Austria infimum, respectively, *sOST* for Austria supremum. Source: author's processing, from EU-FADN 2006.

In the factorial plane of the first two components C1 × C2 (Figure 4), the Netherlands are plotted with Belgium and Italy in the first quadrant (C1C2 > 0), indicating a lower general level of fertilizer estimates. In the opposed half-plan (C1 < 0), France, Hungary, and Poland are plotted with Sweden, indicating

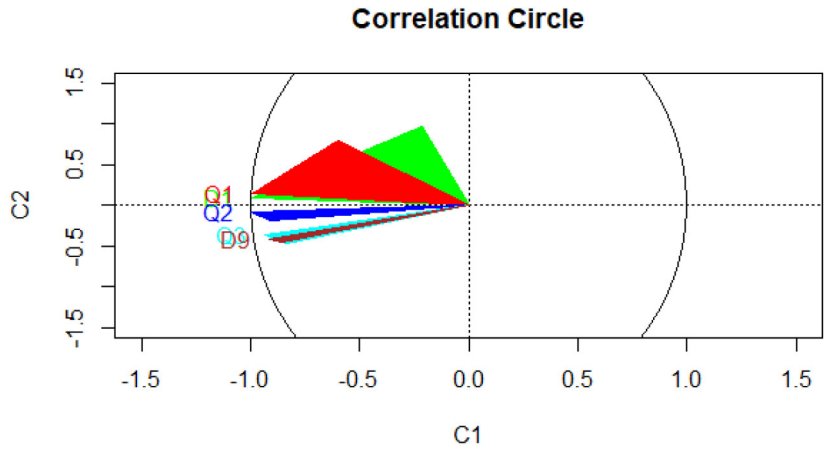


Figure 3. Symbolic PCA ("centers" option) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

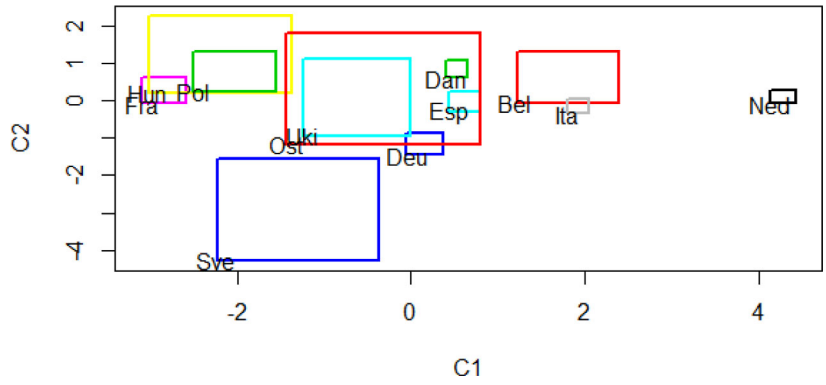


Figure 4. Symbolic PCA ("centers" option) for Quantile Estimates, factorial plane C1xC2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

much higher general levels of fertilizer estimates. Along the C1 component negative side, Austria and the United Kingdom are nearer from Sweden while Germany (Deu), Denmark (Dan) and Spain (Esp) along the positive side of C1 component are plotted nearer Belgium.

Along the C2 component, Sweden is clearly opposed to the other countries, taking in account its extreme D9 estimates.

Countries symbolized by a larger rectangle are Austria, Sweden, Hungary, the United Kingdom, Belgium, and Poland, which correspond to those with greater interval range. Conversely, countries symbolized by a smaller rectangle are Denmark, Italy, the Netherlands, Spain, Germany, and France, which are characterized by a narrower range of estimate intervals.

For individuals, alternate projections are provided by the "best point" PCA options, the optimized distance option on one hand, and on the other hand the optimized variance option.

Table 2. Comparison of the percentage of cumulative variance between the principal components of the three following PCA options: classical PCA (Classic), optimized distance (Optdist), and optimized variance (Optvar).

l%_cum_var	Classic	Optdist	Optvar
C1	75.2	69.5	65.4
C2	97.4	94.9	98.7
C3	99.3	99.2	100.0
C4	99.9	99.9	100.0
C5	100.0	100.0	100.0

Source: author's processing, from EU-FADN 2006.

Table 3. Comparison of the mean absolute deviation (MAD) between the principal components of three PCA options: centers PCA (Centers), optimized distance (Optdist), and optimized variance (Optvar). Source: author's processing, from EU-FADN 2006.

Method	**MAD_C1**	**MAD_C2**	**MAD_C3**	**MAD_C4**	**MAD_C5**
Centers	0.93	1.27	1.21	0.89	0.33
Optdist	0.61	0.83	0.88	0.32	0.27
Optvar	0.67	1.17	0.73	0.39	1.06

As shown in Table 2, the optimized variance option of the PCA maximizes the variance of the first components since the cumulative percentage of variance of the first factorial plan is the highest (98.7%) compared to the optimal distance option (94.9%) and to the classical PCA (97.4%). So, the optimized variance option provides a more complete summary in two dimensions.

Except for the third principal component (C3), the optimized distance option of the interval PCA displays the minimum absolute deviation (MAD) between supremum and infimum vertices over the principal components, compared to the centers option and the optimized variance options (Table 3). So the optimized distance option provides a narrower display of interval estimates for quantile.

In the first factorial plane, the optimized distance and the optimized variance options display a pattern of correlations between quantile estimates and principal components very similar to those of the classic PCA on the two first principal components. As shown by their contributions to inertia (Table 4), the first two principal components have roughly the same definition in terms of quantile. The correlations between quantile estimates and the other principal components (C3, C4 and C5) are different from the classical PCA for the optimized variance option, however without few practical implications due to the very small level of inertia (below 5%) expressed by this these components.

The contributions to inertia for the national distributions of fertilizer estimates (Table 5) show similar patterns on the two first components between the optimized options and the classic PCA, with the exception of Poland opposed to Sweden in the optimized variance option, instead of Hungary in classic and optimized distance options for the C2 component.

Table 4. Comparison of the relative contribution to inertia (Contr.) between the principal components of the three PCA options: classic PCA (Classic), optimized distance (Optdist), and optimized variance (Optvar). Source: author's processing, from EU-FADN 2006.

	C1			C2			C3			C4			C5		
Contr. (%)	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar
D1	13	11	5	43	41	49	35	48	16	9	0	3	0	0	27
Q1	19	15	16	22	29	28	26	49	5	30	5	0	3	2	50
Q2	25	28	30	2	1	1	20	1	29	31	45	23	22	25	18
Q3	22	24	25	16	12	11	0	1	1	3	2	63	58	60	0
D9	21	22	24	17	17	11	19	1	49	26	47	12	17	13	4

Table 5. Comparison of the relative contribution to inertia (Contr.) between the principal components of the three PCA options: classic PCA (Classic), optimized distance (Optdist), and optimized variance (Optvar). Source: author's processing, from EU-FADN 2006.

	C1			C2			C3			C4			C5		
Contr. (%)	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar	Classic	Optdist	Optvar
Bel	7	11	9	3	1	3	0	0	0	2	3	0	2	2	0
Dan	1	1	1	5	4	1	16	8	5	0	9	58	0	0	1
Deu	0	0	0	10	7	4	16	1	6	1	15	10	1	1	0
Esp	1	1	1	0	0	0	12	3	25	5	17	28	45	31	0
Fra	18	17	20	1	1	3	5	2	6	21	13	0	23	31	46
Hun	11	13	8	11	11	1	0	0	16	0	3	0	15	17	48
Ita	8	7	7	0	0	0	1	0	3	2	3	0	1	0	0
Ned	41	35	40	0	1	0	2	1	11	6	9	0	2	4	4
Ost	0	0	1	1	0	15	26	77	3	48	7	0	0	2	0
Pol	9	10	8	5	2	1	13	2	12	3	17	0	3	3	0
Sve	4	3	3	64	69	70	7	2	2	0	0	3	7	6	0
Uki	1	2	1	0	3	0	3	5	11	11	4	0	2	3	0

As summarized by the mean absolute deviation in Table 3, the display of all country rectangle projections is the largest into the centers option (Figure 4) and the smallest into the optimized distance option (Figure 6) while the display of the optimized variance option (Figure 5) is of medium range between the two previous options, both in the lengths (dimension 1 of the first principal component) and the widths (dimension 2 of the second principal component).

By the relative sizes and locations of their hyper-rectangle projections, these three factorial representations (Figures 4–6) distinguish clearly Netherlands on the first principal component, as the archetype of the *location shift* model, and Sweden, on the second principal component, as the archetype of the *location scale shift* model.

3.2. The divisive hierarchy of fertilizer cost estimates

Showned by Figure 7, the divisive hierarchy obtained with Euclidean distance option shows that the set of D1, Q1, Q2 and Q3 quantile estimates is used by the discriminant values, which implies keeping these parameters to describe the distribution, and possibly extending it by a finer quantile scale allowing some of the national distributions to be better distinguished.

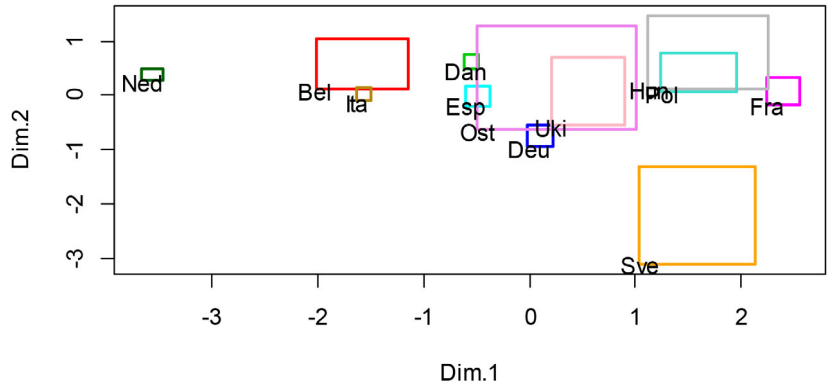


Figure 5. Symbolic PCA ("optimized.distance" option) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

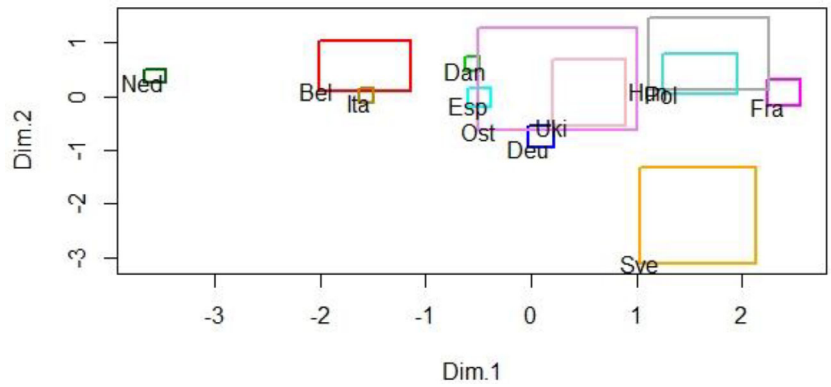


Figure 6. Symbolic PCA ("optimized.variance" option) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

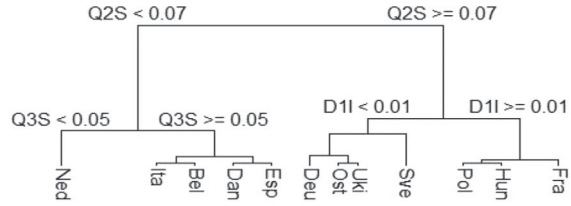


Figure 7. Symbolic Divisive Clustering ("Euclidean distance" option) for Quantile Estimates, EU12 countries. Source: author's processing, from EU-FADN 2006.

The first partition in two clusters corresponds to the supremum level of the median estimate (Q2S).

At the top of the divisive hierarchy, the clustering procedure allows to identify two contrasted models for empirical distributions of the fertilizers technical coefficients for yearly crops production costs used to €1,000 of gross product.

As the first cluster, the Netherlands (*Ned*) and the group of Italy (*Ita*), Belgium, (*Bel*), Denmark (*Dan*) and Spain (*Esp*) grouped by their supremum

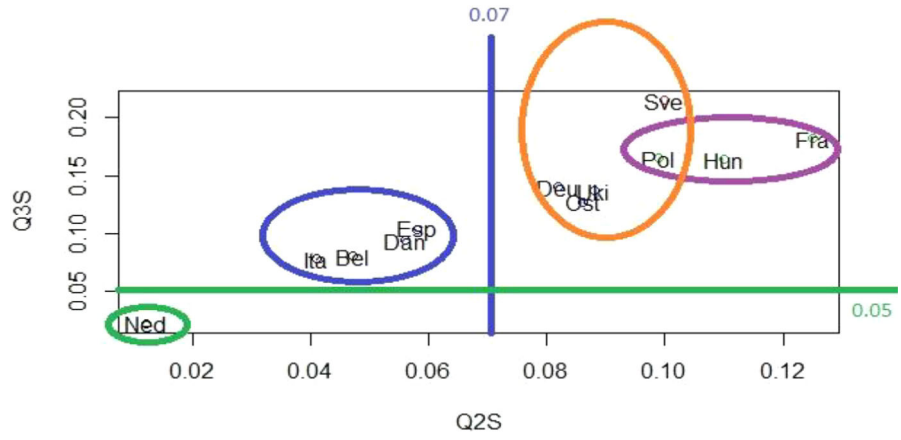


Figure 8. Symbolic Divisive Clustering (C4 optimal partition for Determinant Ratio Criterion) for Quantile Estimates, factorial plane F1xF2 of EU12 countries. Source: author's processing, from EU-FADN 2006.

median (*Q2S*) levels which are lower than €7, are split in the following divisive step by the supremum higher quartile (*Q3S*) level of €5 which identifies the Netherlands as the less intensive in fertilizer input. The Netherlands is the archetype of the *location-shift* model formalizing the assumption of homogeneous producers in their fertilizer costs.

As the second cluster, for which their supremum median (*Q2S*) of fertilizer cost is greater than 7€, is split into two groups: first, the group for which the fertilizer first decile input is greater than €1, i.e. the subgroup formed by Poland (*Pol*), and Hungary (*Hun*) aggregated with France (*Fra*); second, the group formed by Sweden (*Sve*) aggregated with the subgroup formed by Germany (*Deu*), Austria (*Ost*) and the United Kingdom (*Uki*), on the basis of their fertilizer first decile lesser than *€1 input* level. This latter group illustrates the *location-scale shift* model, formalizing the assumption of heterogeneous producers in their fertilizer costs.

The partition into four groups displays by Figure 8 is the optimal partition for the minimum difference in the logarithm of the ratio of determinants (package *ClusterCrit*), which is a consistent rule with the criterion of the DIVCLUS-T algorithm.

4. Conclusions

Based on quantile regression and symbolic data analysis, this paper presents a global methodology which aims to keep as much as possible relevant information for the policy design, all along the econometric process of estimating and analyzing agricultural fertilizer costs for yearly crops production. The different properties of three options of interval PCA (centers, optimized distance and optimized variance) are described allowing to identify different models of

distributional scale, notably that of the *location shift* model opposite that of the *location-scale shift* one. Differences and similarities between interval estimates are exploited using divisive hierarchical clustering to produce two country clusters identifying through quantile cost thresholds the archetypes of the *location shift* model and the *location-scale shift* one. The differences between four groups of countries are delimited by optimal thresholds expressed according to the conditional quantiles in unitary terms of the gross product. These thresholds can be used for segmenting farm populations to later analyze the differential impacts of agricultural policy measures. We will apply this methodology at the second level of the European Nomenclature of Territorial Units for Statistics (NUTS 2, 281 regions).

Funding

This work was supported by Agence Nationale de la Recherche on behalf of the "impActs and feedbackS between climate and Soil affected by EroSion: cost in terms of carbon Storage in Mediterranean regions" (ASSESS) project (ANR-16-NME1-0008) of the OTE-Med Eranet.

ORCID

Dominique Desbois http://orcid.org/0000-0001-6198-454X

References

Angrist, J., and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton Univ Press.

Billard, L., and E. Diday. 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Chichester, UK: Wiley Interscience.

Broyden, C. G. 1970. "The Convergence of a Class of Double-Rank Minimization Algorithms." *IMA Journal of Applied Mathematics* 6 (1):76–90.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics. Methods and Applications*. Cambridge University Press.

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

Chavent, Marie, Yves Lechevallier, and Olivier Briant. 2007. "DIVCLUS-T: A Monothetic Divisive Hierarchical Clustering Method." *Computational Statistics & Data Analysis* 52 (2): 687–701.

Chouakria, A., E. Diday, and P. Cazes. 1998. "An Improved Factorial Representation of Symbolic Objects." In *Studies and Research, Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA'98)*. Luxembourg: Office for Official Publications of the European Communities, 276–289.

D'Haultfoeuille, X., and P. Givord. 2014. "La Régression Quantile en Pratique." *Economie et Statistique* 471 (1):85–111.

Dantzig, G. B. 1949. "Programming in a Linear Structure. *Econometrica* 17:73–74.

Desbois, D. 2015. "Estimation Des Coûts de Production Agricoles: approches Économétriques." PhD dissertation directed by J. C. Bureau and Y. Surry, Paris: ABIES-AgroParisTech.

Desbois, D., J.-P. Butault, and Y. Surry. 2017. "Distribution Des Coûts Spécifiques de Production Dans L'agriculture de L'Union Européenne: une Approche Reposant Sur la Méthode de Régression Quantile." *Économie Rurale* 361 (361):3–22.

Desgraupes, B. 2017. *Clustering Indices*. Lab Modal'X, Paris-Ouest University, 22pp. November, Nanterre.
Diday, E. 2006. "Thinking by Classes in Data Science: The Symbolic Data Analysis Paradigm." *WIREs Computational Statistics* 8:171–205.
Divay, J. F., and F. Meunier. 1980. "Deux Méthodes de Confection du Tableau Entrées-Sorties." *Annales de l'INSEE* 37:59–109.
Fletcher, R. 1970. "A New Approach to Variable Metric Algorithms." *The Computer Journal* 13 (3):317–322.
Garro, J., and O. Rodriguez. 2019. "Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognition." *Entropy* 21 (9):1016. doi:10.3390/e21101016
Goldfarb, D. 1970. "A Family of Variable Metric Updates Derived by Variational Means." *Mathematics of Computation* 24 (109):23–26.
He, X., and F. Hu. 2002. "Markov Chain Marginal Bootstrap." *Journal of the American Statistical Association* 97 (459):783–795.
Jevnaker, T., and J. Wettestad. 2017. "Ratcheting up Carbon Trade: The Politics of Reforming EU Emissions Trading." *Global Environmental Politics* 17 (2):105–124.
Karmarkar, R. 1984. "A New Polynomial-Time Algorithm for Linear Programming." *Combinatorica* 4 (4):373–395.
Koenker, R., and G. Bassett. 1978. "Regression Quantiles." *Econometrica* 46 (1):33–50.
Koenker, R., and Q. Zhao. 1994. "L-Estimation for Linear Heteroscedastic Models." *Journal of Nonparametric Statistics* 3 (3/4):223–235.
Koenker, R., and V. d'Orey. 1994. "Remark as R92: A Remark on Algorithm as 229: Computing Dual Regression Quantiles and Regression Rank Scores." *Applied Statistics* 43 (2):410–414.
Lustig, I. J., R. E. Marsten, and D. F. Shanno. 1992. "On Implementing Mehrotra's Predictor-Corrector Interior-Point Method for Linear Programming." *SIAM Journal on Optimization* 2 (3):435–449.
Madsen, K., and H. B. Nielsen. 1993. "A Finite Smoothing Algorithm for Linear Estimation." *SIAM Journal on Optimization* 3 (2):223–235.
Mirkin, B. 2005. *Clustering for Data Mining. A Data Recovery Approach*. London: Chapman & Hall.
Portnoy, S., and R. Koenker. 1977. "The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-Errors vs. absolute-Errors Estimators." *Statistical Science* 1:279–300.
Quinet, A. 2019. "La Valeur de L'action Pour le Climat. Une Valeur Tutélaire du Carbone Pour Évaluer Les Investissements et Les Politiques Publiques." France Stratégie.
Rodriguez, O. 2000. "Classification et Modèles Linéaires en Analyse Des Données Symboliques." PhD Thesis. Paris IX-Dauphine University, France.
Rodriguez, O., E. Diday, and S. Winsberg. 2000. *Generalizations of Principal Component Analysis*. Paris: Paris-Dauphine University.
Shanno, D. F. 1970. "Conditioning of Quasi-Newton Methods for Function Minimization." *Mathematics of Computation* 24 (111):647–656.
Ypma, J. 2020. Package 'nloptr', CRAN repository.