Application de la classification symbolique au problème d'estimation des coûts de production agricole

Dominique DESBOIS
UMR Economie publique, INRA-AgroParisTech, Université Paris Saclay

"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 et Pischke, 2009)

XXVI^e Rencontres SFC

Nancy,04/09/2019

INRA SCIENCE & IMPACT

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Problem:

How to keep the maximum amount of information for clustering estimated parameter distributions?

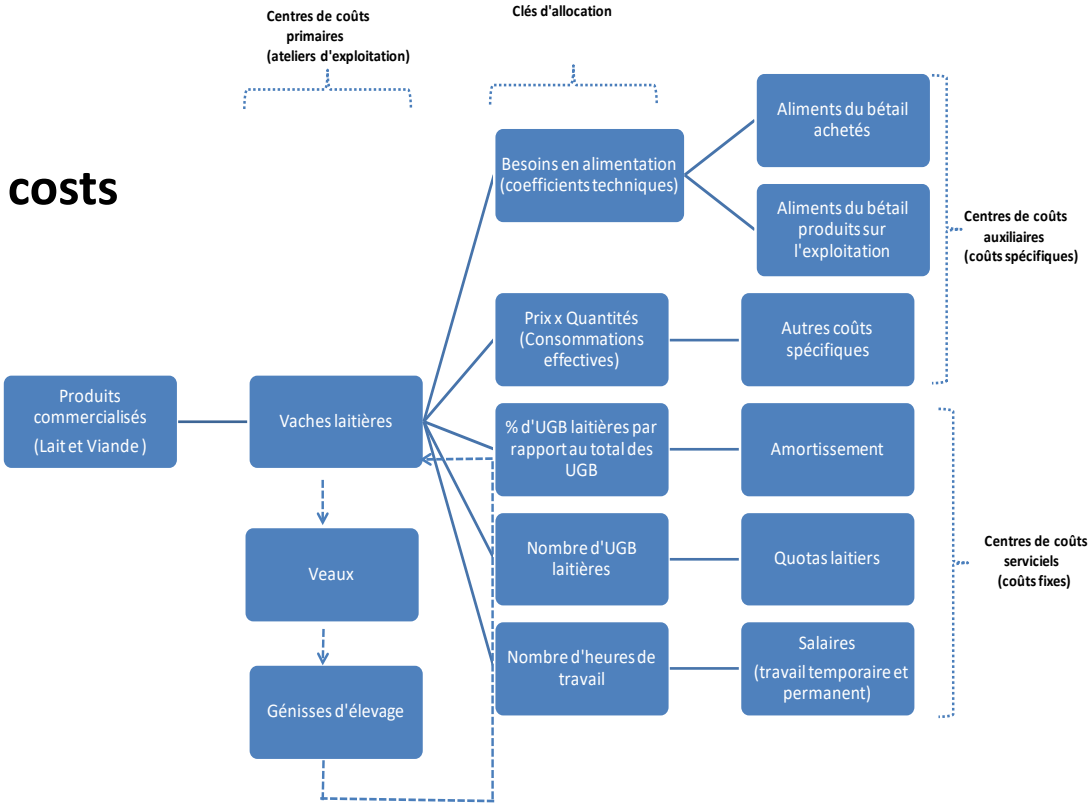
Outline :

1. **The estimation by intervals of conditional quantiles;**
2. **Divisive clustering analysis on estimation intervals;**
3. **Viewing estimated conditional quantiles distributions;**
4. **Discussion of the results.**

I.1 Evaluation of Agricultural production costs

1. Technical accounting of Production costs

Choice of a method of Technical-Economic Accounting



2. Econometric Modeling of production costs

Choosing a Model with coefficients

PRODUITS / CHARGES	Y_{1h}	…	Y_{kh}	…	Y_{Kh}	TOTAL CHARGE
X_{1h}	a_{1h}^{1}	…	a_{1h}^{k}	…	a_{1h}^{K}	$\sum X_{1h}$
⋮	⋮		⋮		⋮	⋮
X_{ih}	a_{ih}^{1}	…	a_{ih}^{k}	…	a_{ih}^{K}	$\sum X_{ih}$
⋮	⋮		⋮		⋮	⋮
X_{Ih}	a_{Ih}^{1}	…	a_{Ih}^{k}	…	a_{Ih}^{K}	$\sum X_{Ih}$
TOTAL PRODUIT	$\sum Y_{1h}$	…	$\sum Y_{kh}$	…	$\sum Y_{Kh}$	$\sum_k Y_{kh} = \sum_i X_{ih}$

$$X_{ih} = \sum_{k=1}^{K} \alpha_{ih}^{k} Y_{kh} + \varepsilon_{ih} \quad with \quad \varepsilon_{ih} \quad i.i.d.$$

I.2 Quantitative estimation of specific cost of agricultural production

Problem : In the face of the heterogeneity of agricultural production structures and production behaviours in Europe, how to retain the maximum useful information when you are estimating production costs then clustering distributions parameter distributions?

Conceptual model : The Input-Output table

Methods : **i)** The L1 standard regression allows, by estimating conditional quantiles, to generate a distribution of specific production costs and gross margins ;

ii) The clustering of interval estimation data gives an optimal tree with regards of the scale and shape of the distributions of estimated parameters.

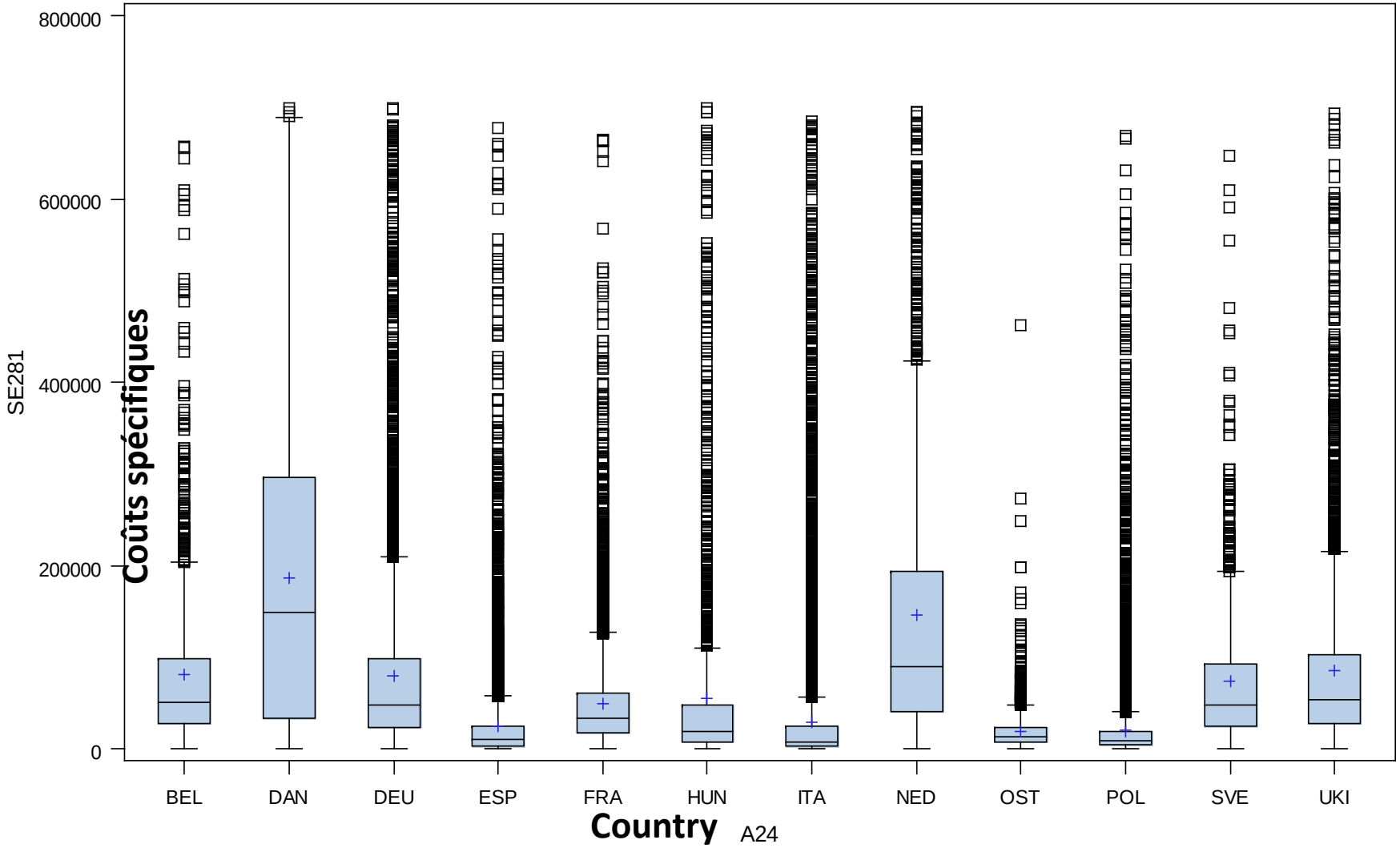
Data : The Farm Accounting Data Network (FADN) in France and Europe

Application :

Estimation of specific costs to deduct gross margin (needed for insurance and solidarity fundings)

I.3 Distribution of specific inputs of agricultural production in the EU

Distribution of specific inputs (< 750 €0) by country, RICA 2006, Eu12



I.4 Estimation of conditional quantiles.

The estimation of conditional quantiles is obtained by solving a problem of minimizing the following loss function for each quantile Q, according to the mean absolute deviation (MAD): criterion

$$\sum_{i;y_i \geq x_i'\beta} q\left|x_i - y_i'\beta\right| + \sum_{i;y_i < x_i'\beta} (1-q)\left|x_i - y_i'\beta\right|$$

For a data generator process

(Cameron & Trivedi, 2005)

$X=Y'\beta+u$ as $u=Y'\alpha*\varepsilon$

Following a linear model with multiplicative heteroscedasticity (i.e. u = Y * e), under the assumption $Y'\alpha>0$

The conditional quantile **q^th^** of cost X conditionally to Y is equal to:

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

So linear in *Y*.

The **technical coefficient** of the q^{th} quantile of specific costs for the j^{th} product is estimated by bootstrap (He & Hue, 2002):

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

II.1 Interval clustering of specific cost distributions

For the product j_0 (pig) and the l^{th} european country, the estimation intervals of the technical coefficients for the specific costs are :

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

The L country distributions $\Omega = \{\omega_1, \cdots, \omega_l, \cdots \omega_L\}$ are described by a set of $Q = 5$ quantiles, namely $Z = \{z^1, \ldots, z^q, \ldots, z^Q\} = \{z^{0.10}, z^{0.25}, z^{0.50}, z^{0.75}, z^{0.90}\}$.

The δ_M *distance* between country *l* and country *l'*, with regards to the q^{th} quantile:

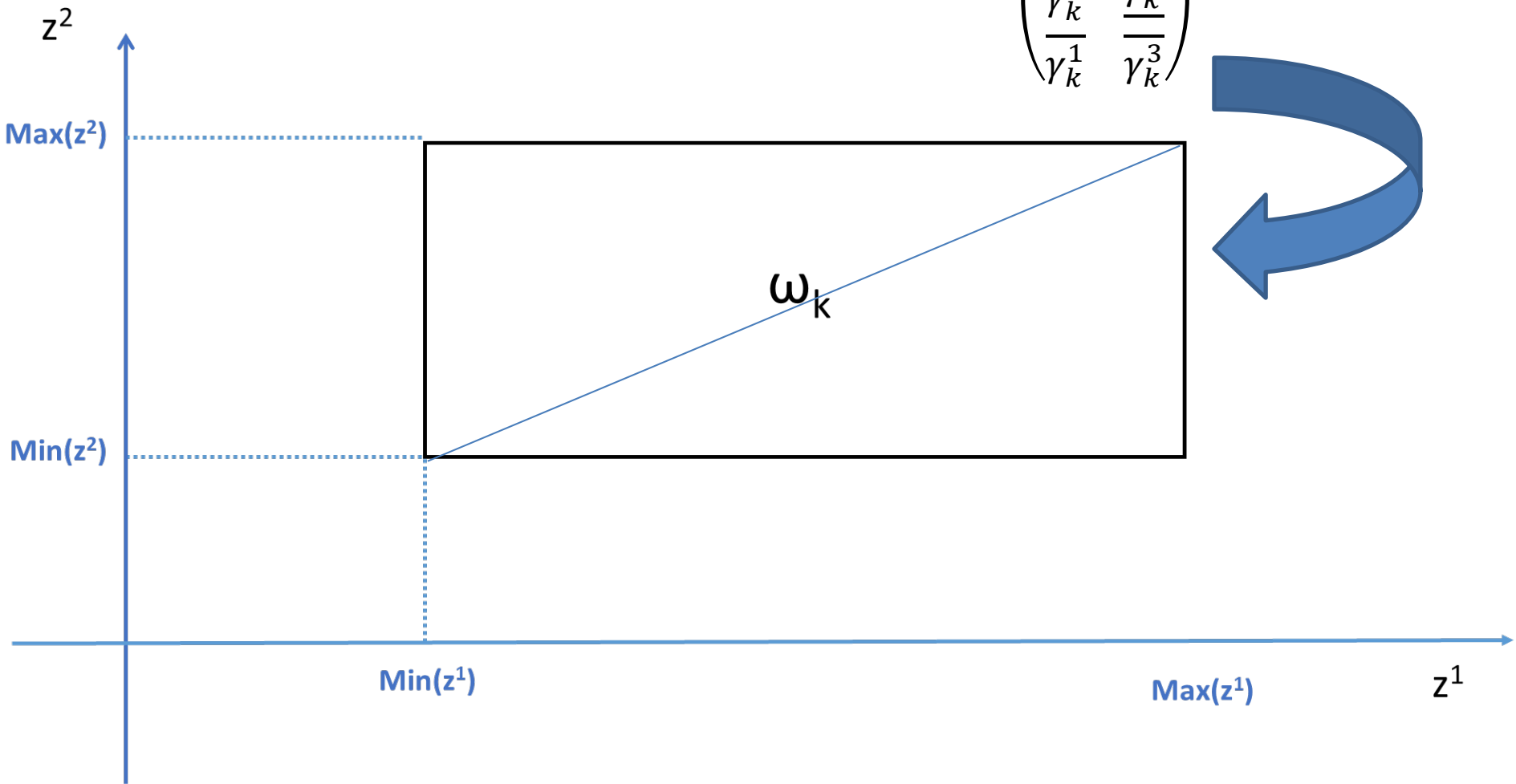
$$\delta_M\left(z_l^q, z_{l'}^q\right) = \sqrt{\left(Inf\_\hat{\gamma}_l^{j_0}(q) - Inf\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2 + \left(Sup\_\hat{\gamma}_l^{j_0}(q) - Sup\_\hat{\gamma}_{l'}^{j_0}(q)\right)^2}$$

The global dissimilarity between country *l* and country *l'* , based on the δ_M metric:

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

II.2 Estimation Intervals of technical coefficients γ: Mapping in 2 dimensions

$$\gamma_k = \left(\left[\underline{\gamma_k^1} \;;\; \overline{\gamma_k^1} \right] \left[\underline{\gamma_k^3} \;;\; \overline{\gamma_k^3} \right] \right) \longleftrightarrow Z_k = \begin{pmatrix} z^1 & z^3 \\ \underline{\gamma_k^1} & \underline{\gamma_k^3} \\ \overline{\gamma_k^1} & \overline{\gamma_k^3} \\ \underline{\gamma_k^1} & \underline{\gamma_k^3} \\ \overline{\gamma_k^1} & \overline{\gamma_k^3} \end{pmatrix}$$



II.3 divisive clustering algorithm

Generated by the binary answer (*yes*/*n*) to the question $\Psi = [z^q \leq c\,?\,]$,
Let us note $\{A_k, \overline{A_k}\}$ the issued bipartition of the class C_k grouping n_k objects.
To choose among the $n_k - 1$ possible bipartitions of C_k,
the discriminant criterion is defined by:

:

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

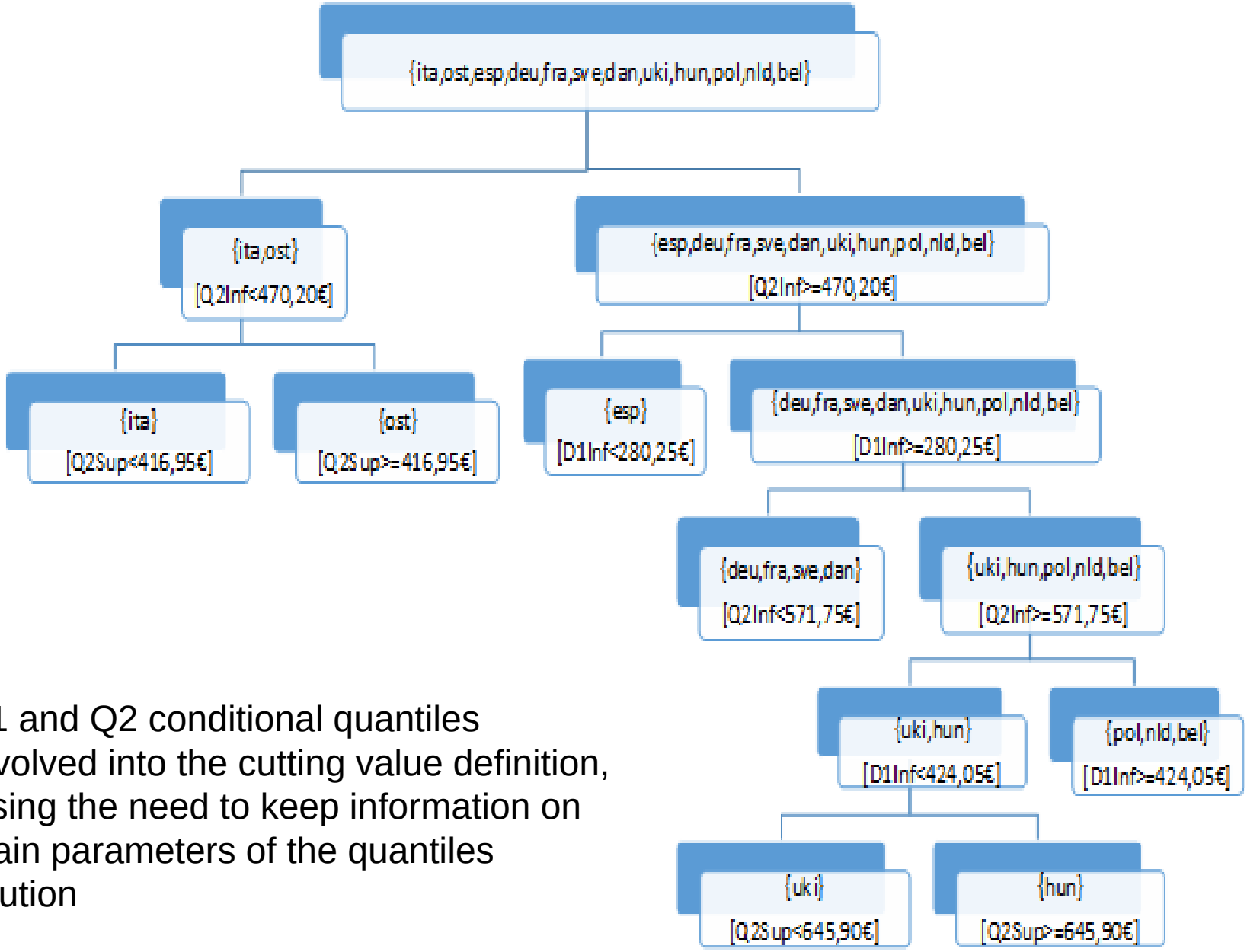
The DIVCLUS-T algorithm cuts the class C_K^* which maximizes its « height » $h(C_K)$:

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

Such as the next partition $P_{K+1} = P_K \cup \{A_K, \overline{A_K}\} - C_K^*$ shows the minimal value of the intra-Inertia, according to:

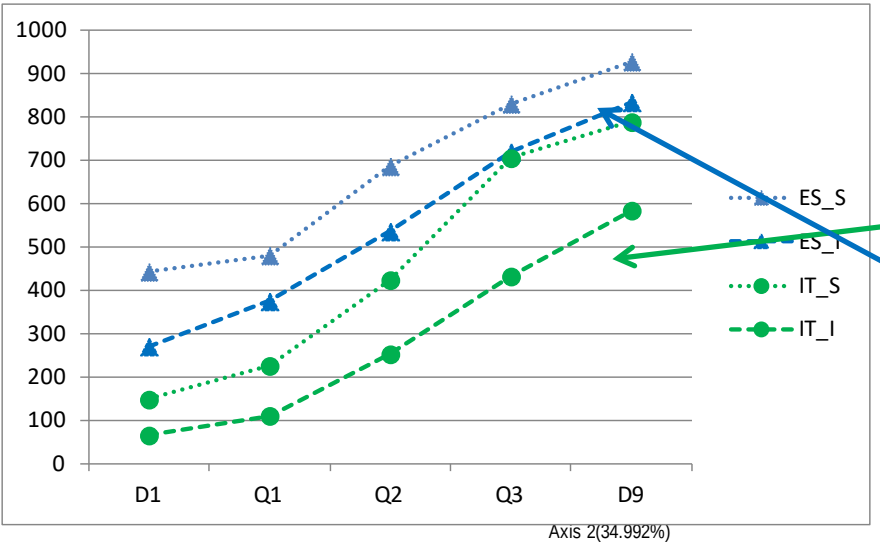
$$W(P_{K+1}) = W(P_K) - h(C_K^*)$$

III.1 Quantile estimates of production costs: pig hierarchical divisive clustering on the estimation intervals



the D1 and Q2 conditional quantiles are involved into the cutting value definition, stressing the need to keep information on the main parameters of the quantiles distribution

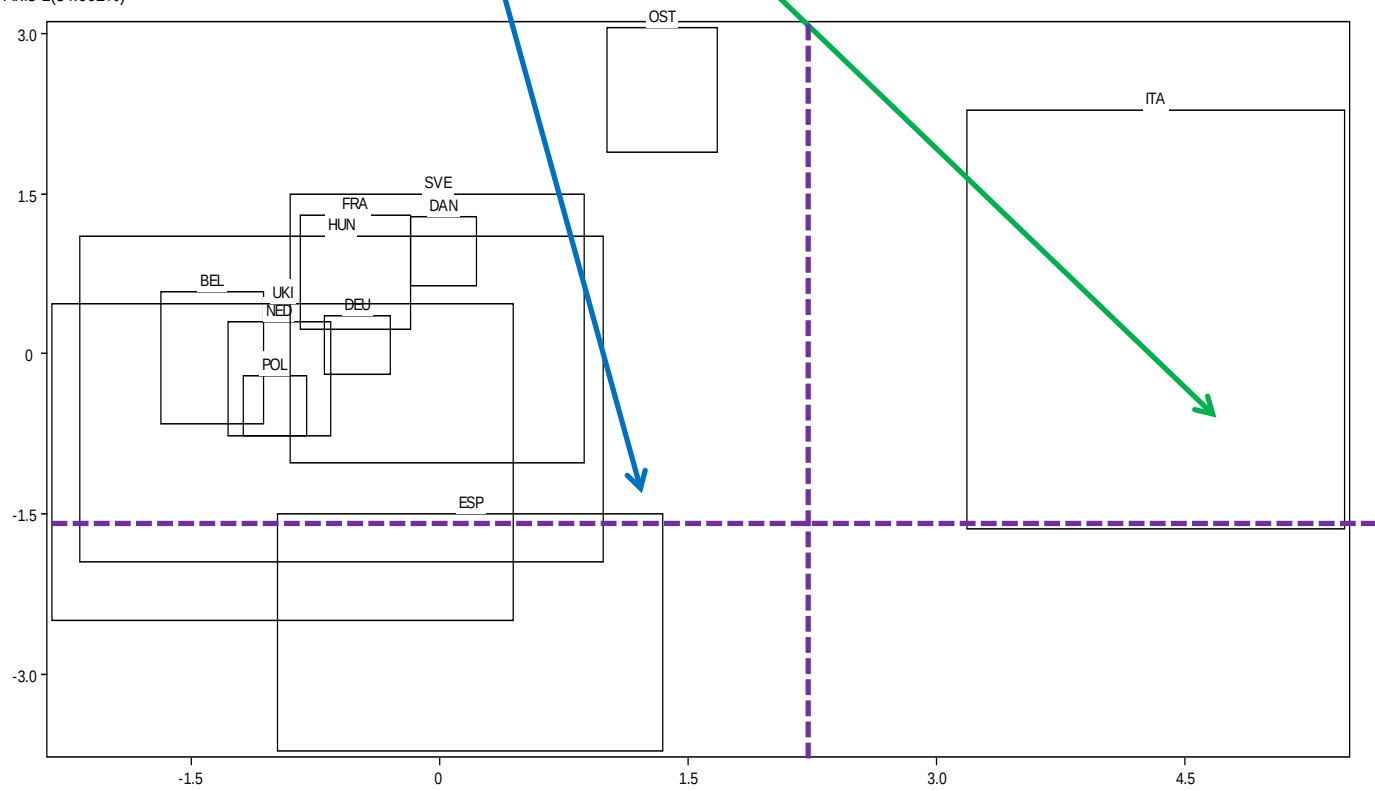
III.2 Pig-2006: Conditional Quantile Estimates of 12 EU-Members, SPCA



Graphical evidences of significant differences between :

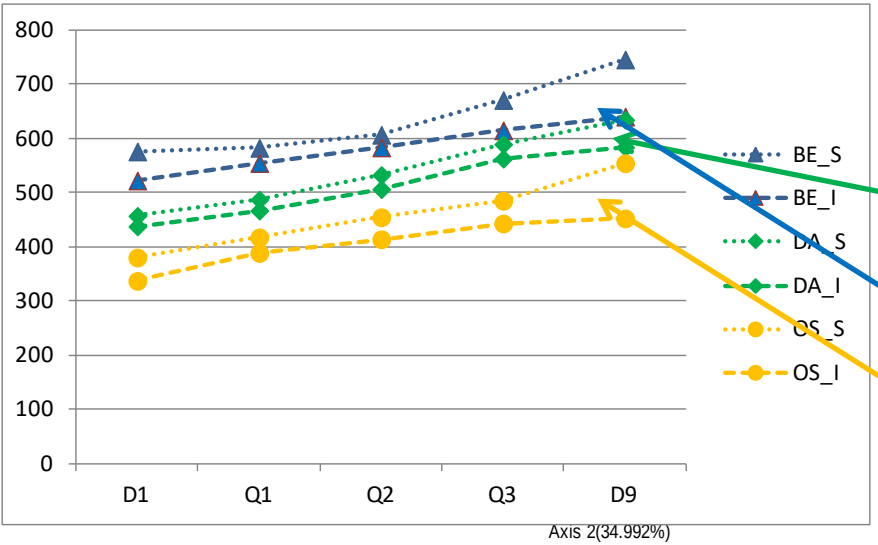
ITA Italia (heterogeneous)

ESP Espana (heterogeneous)



Data source: EU-FADN – DG AGRI

III.3 Pig-2006: Conditional Quantile Estimates of 12 EU-Members, SPCA

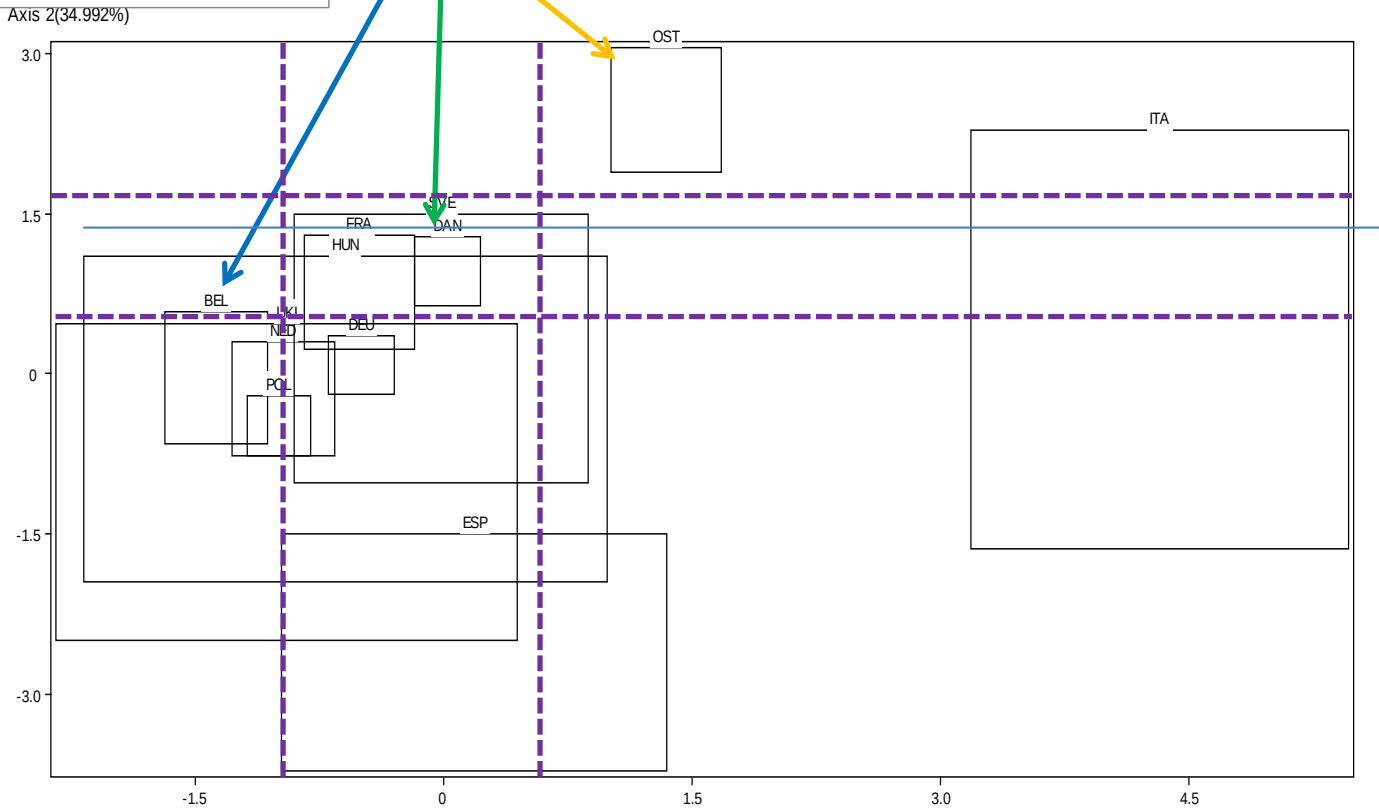


Graphical evidence of significant differences between :

DAN Danemark (homogeneous)

BEL Belgium (homogeneous)

OST Österreich (homogeneous)



Data source: EU-FADN – DG AGRI

IV. Conditionnal Quantiles of Specific Costs for Pig Production Conclusions

Specific costs: in UE12, for the pig

- The national context is a significant factor of heterogeneity distinguishing northern Europe (homogeneous production systems) from Southern Europe (heterogeneous production systems) ;

- The specific costs of the pig workshop appear to be lower overall for small specialized granivoreous farms, while they are generally higher when this pig workshop is located in mixed farms of medium or large economic dimension ;

- There are significant differences between the southern regions of the EU12 European countries (Italy, Spain) due to regional specializations in certain types of pig production.

References

- Desbois D. (2017) Displaying empirical distributions of conditional quantile estimates: an application of symbolic data analysis to the cost allocation problem in agriculture », *Proceedings of the 17th Conference of the Applied Stochastic Models and Data Analysis International Society*, Londres, pp. 189-202.

- Desbois D., Butault JP., Surry Y. (2017) Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la régression quantile, *Économie rurale*, vol. 361, n°5, pp. 3-22.

- Desbois D. (2015) Estimation des coûts de production agricoles : approches économétriques. *PhD Thesis*, ABIES-AgroParisTech, dirigé par J.C. Bureau et Y.Surry, 249 p.

- Desbois D., Butault J.-P., et Surry Y. (2013) Estimation des coûts de production en phytosanitaires pour les grandes cultures. Une approche par la régression quantile. *Economie Rurale*, n° 333, 27-49.

- Surry Y., Desbois D., et Butault J.-P. (2012) Quantile Estimation of Specific Costs of Production. *FACEPA*, D8.2, 49 p.

Collaborative research platform

Sign up for a FREE account

Search

Search

authors titles abstracts

Numéro 45

MODULAD 2018. *Modulad - Le Monde des Utilisateurs de L'Analyse de Données* PDF Bib
vol.Modulad 45
Editors : Dominique Desbois, Edwin Diday

https://editions-rnti.fr/

[SFdS] Revue de Modulad : appel à contributions pour un numéro spécial « Apprentissage et Sémantique »

Suite à l'organisation d'une Action nationale de formation (http://devlog.cnrs.fr/apsem2018) sous l'égide du réseau de Développement logiciel (Devlog) du CNRS avec l'appui du dispositif d'ingénierie numérique de l'Inra (https://www.ingenum.inra.fr) consacrée à l'apprentissage et à la sémantique des données (*ApSem*), la rédaction de la *Revue de Modulad* appelle aux contributions dans le champ de l'apprentissage et de la sémantique pour l'étude des systèmes complexes.

Si la science des données se définit comme celle de l'extraction d'informations intelligibles et de connaissances à partir d'ensembles de données, cette discipline s'appuie à la fois sur des concepts mathématiques, des méthodologies statistiques et des outils informatiques(1). Héritiers de sciences basées essentiellement sur le recueil de données expérimentales, nous assistons désormais à l'émergence de sciences basées sur les possibilités d'exploration et de simulation offertes par les données massivement observées(2). Aussi, sommes-nous confrontés à un changement de paradigme suscité par l'afflux de données issues de réseaux de capteurs ou d'objets connectés et par la disponibilité de moyens de calcul toujours plus puissants avec un essor notable du parallélisme et des architectures vectorielles. L'émergence d'e-infrastructures numériques offre l'intermédiation de services rendant opérationnel l'accès à ces nouvelles ressources. D'ores et déjà, les nouvelles possibilités de « fouille des données » permettent d'envisager des changements d'échelle dans l'étude des systèmes complexes.

L'objectif de ce numéro spécial est de faire le point sur l'apport croisé des méthodologies de l'apprentissage et des technologies du web sémantique :
1> l'apport des concepts, méthodes et outils de l'apprentissage pour tirer parti des représentations sémantiques des données
2> l'apport des concepts, méthodes et outils des mécanismes inférentiels pour générer de nouvelles représentations des connaissances.

Dans le prolongement du programme du séminaire *ApSem*, les contributions se répartiront selon les quatre thématiques suivantes :
i) Apprentissage par les données et science des systèmes complexes ;
ii) Ingénierie des connaissances et infrastructure web des données ;
iii) Modalités de convergence entre apprentissage automatique et sémantique des données ;
iv) Ateliers logiciels, études de cas et retours d'expérience.

Dans un contexte de science ouverte, nous encourageons les contributions originales et inédites en langue anglaise ou française privilégiant la reproductibilité des résultats présentés, si possible par la publication conjointe d'un extrait des données et d'une implantation des algorithmes, dans la continuité de la ligne éditoriale de la *Revue de Modulad*. Les textes déjà publiés en version courte (e.g., actes de conférence) peuvent être soumis à la *Revue de Modulad* pour publication.

Ces contributions anonymisables (noms, affiliations et adresses des auteurs figurant sur la page de garde) devront être soumises conjointement aux éditeurs de ce numéro spécial (christophe.biernacki@inria.fr ; pascal.dayre@enseeiht.fr ; dominique.desbois@inra.fr) en indiquant un auteur correspondant unique, pour être transmises à trois relecteurs indépendants aux fins d'évaluation par les pairs.