Functional regression to model the link between livestock farms' performance and their size

A. de la Foye*,[1] and P. Veysset[2]

[1,2]UMR1213 Herbivores, INRA, 63122 Saint-Genès-Champanelle, France

November 24, 2019

Abstract

Over the last decades, livestock farms' sizes have increased in Western Europe although the benefits obtained by the farms and agricultural workers are far from being obvious nor demonstrated.

The purpose of this paper is to propose the use of functional regression to study the relationship between farm size and farm performance trends in long-term panel data. While a linear Partial Least Squares regression is performed as a reference model, a regression for functional data enrich the modelling with the chronological information contained in panel data, in particular evolution shapes of performance and size.

The proposed methods are applied on a long-term structural, technical and economic database of cattle beef farms in the Charolais region (France). The achieved regression models, whether classical or functional, suggest that, in our case study, the farm size is not sufficient to explain its performance.

We now expect our work to be a starting point for further analysis. In particular, other structural, economical and technical characteristics could contribute to better model livestock farms' evolution.

1 Introduction

While the sizes of beef cattle farms have continually increased in France for at least twenty years, the technical and economical performances have stagnated over the same period ([15], [14]). These observations challenge two commonly held opinions. The first is that the expansion of farms should benefit their productivity through economies of scale. The second is that technological progress and agricultural policy should benefit agricultural workers.

In fact, there is no consensus about the relationship between farm size and farm performance. Indeed, in his literature review on family farm size and efficiency, [2] wrote that in business literature, the corresponding correlation was often estimated as positive. However, he added that in agriculture and development economics papers, a negative correlation between farm size and productivity was often shown. The results are anyway often mixed. [4] found that though the negative correlation observed in 1982-2008 in Nicaragua gradually declined, it still remained. [8] also challenged the opinion expressed in previous studies that small-to-medium-sized businesses would deteriorate economically. Yet, [8] noted the good economic health of larger farms over the decade 2002-2011 in Western Australia. [2] added that the complexity of the mechanisms governing farm behaviour makes it difficult to draw theoretical conclusions. [7] suggested that the observed negative correlation between farm size and productivity in developping countries is probably due to the labour market imperfections. In western Europe, to our knowledge, the issue remains unexplored though it seems crucial to ensure the adequacy of agricultural policy.

In order to model the link between farm performance as a response and size as a predictor, long-term observations on a sample of independant farms are the most appropriate. On this type of data, a first idea is to carry out a panel regression analysis as [3], [7] et [4] did. However, [1] rejected panel regression analysis in favour of quantile regression analysis for 3 reasons : 1) the marginal effects of changes in some of the variables in estimated model can differ according to the technical efficiency scores, 2) the restrictive assumption of the distribution of error terms in Ordinary Least Squares (OLS) and 3) the better robustness of quantile regression to heteroskedasticity of the residuals.

None of these models, though, take into account the chronological link between years in performance and size evolutions when this information may be essential. In contrast, functional models, as studied by [11], make it possible to include evolution curves as predictors (scalar-on-function regression), as a response (function-on-scalar regression) or both (function-on-function regression). Function-on-function regressions can thus predict evolution curves with other evolution curves. Numerous functional regression methods have been developped as reviewed by [9]. Yet, the latter noted that there had been comparatively little work

*Corresponding author

on function-on-function regression which interests us in the first place. The objective of this paper is to show how functional regression can be used to analyse the links between the evolution of farm size and the evolution of its performance comparing Partial Least Squares (PLS) regression ([12]) and functional regression. One concern in Functional Data Analysis (FDA) is the transformation of usually discrete observed data to functional data, which we address at first. A second concern is the estimation of the functional regression coefficients which are infinite-dimensional by nature as [11] noted. To deal with the undetermination issue arising from the induced infinite number of solutions, we perform a functional PLS regression, as developped by [10].

The paper is structured as follows. The second section presents our methodological approach. The third section describes the long-term structural, technical and economic database of beef cattle farms in the French Massif Central to which our methodology is applied. The farms' performance modelling results of our application study are reported in Section 4. The fifth section is devoted to the discussion and conclusion.

2 Methodology

In the following, we assume that relevant performance and size indicators are observed on a constant sample of N farms over T years.

Two methods are proposed to model farms' performances with their size. On the one hand, a PLS regression models the links between T performance variables and T size variables over the same T years. On the other hand, a more complex model based on evolution curves such as function-on-function regression could be needed to exploit the chronology of developments.

2.1 PLS linear regression

PLS regression ([12]) was developped to analyse complex linear relationships between a matrix of output correlated variables Y and a matrix of input correlated variables X. It consists in extracting orthogonal linear combinations U from X in the one hand and orthogonal linear combinations V from Y in the other hand so that the covariances between U_t and V_t columns are iteratively maximized for $t \in [\![1, T]\!]$. Matrix U contains the PLS components of the model.

In our case, we aim at modelling $N * T$ matrix Y as a response where Y_{it} is the performance of farm i in year t. The predictor is $N * T$ matrix X where X_{it} is the size of farm i in year t :

$$Y = A + XB + E$$

B is a $T * T$ coefficient matrix and E is the $N * T$ error matrix.

U and V are also of dimension $N * T$, but an aim of PLS regression is to reduce their dimension to $N * h$, $h < T$, so that only useful information is retained in a parsimonious model. The selection of the optimal number of PLS components $\tilde{h}$ and the estimate of the predictive quality of the model $R^2_{\hat{Y}(\tilde{h})}$ are processed via a double Leave-One-Out (LOO) procedure as advised by [6] and [13].

Thus,

$$R^2_{\hat{Y}(\tilde{h})} = 1 - \frac{PRESS_{\hat{Y}(\tilde{h})}}{SST_Y}$$

where

- $PRESS_{\hat{Y}(h)} = \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \hat{Y}_{(-i)t}(h))^2$ is the Predicted Sum of Squared Residuals after the outer LOO loop for a model with h PLS components
- $\hat{Y}_{(-i)t}(h)$ is the predicted value of Y_{it} after the outer LOO loop, that is on a model with h PLS components estimated with all observations but i.
- $\tilde{h} = \arg\min_h PRESS_{\hat{Y}(h)}$ the optimal number of PLS components as estimated after the inner LOO loop executed on all farms but i
- $SST_Y = \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \bar{Y}_t)^2$ is the Sum of Squares Total, expressing the dispersion of variables Y_t to their means $\bar{Y}_t$.

By construction, when the prediction error due to the regression model is greater than the dispersion of Y from the mean, $R^2_{\hat{Y}}$ can be negative. In this case, the model is useless.

2.2 Functional linear regression

In the PLS regression model described above, nor chronological information between observed years, neither any of the specific information contained in the evolution curves for performance and size was taken into account. Assuming these evolution curves were at our disposal, the performance of farm i at any time $s \in \mathcal{T}$ could in particular be modelled according to farm i's size at any time $t \in \mathcal{T}$, where $\mathcal{T}$ is the continuous interval of time $[1, T]$:

$$y_i(s) = \alpha(s) + \int_{\mathcal{T}} \beta(s,t)x_i(t)dt + \epsilon_{it} \forall i \in [\![1, N]\!] \quad (1)$$

This is a functional regression model where both the predictor and response are functional.

2.2.1 Preliminary step: computing curves from discrete data

The number T of observations on a farm i is necessarily finite by construction and thus do not consist of curves. Therefore, as a first step before performing functional regression, following [11] and [10], our observed size data X_{it} are first smoothed into curves x_i which are L_2-continuous functions. These functions are further assumed to be the real information in the whole $\mathcal{T}$ interval.

The smoothing is achieved using B-splines basis functions $\phi_k(t)$ which are known for being well adapted to non periodic data ([11]). The underlying function x_i can then be assumed as a finite sum of parametric functions :

$$x_i(t) = \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$$

$\xi_{ik}, k = 1, ..., K$ are the coefficients of the basis B-spline functions ϕ_k associated with x_i. The dimension of functions x_i is given by the basis size K.

How do we choose the B-splines basis functions $\phi_k(t)$? Indeed, a B-spline is a piecewise polynomial function which depends in particular on the degree of the polynoms and on the number of polynomials (and sub-intervals) defined. Under the usual conditions in which we operate, the basis size K is one more than the highest power of the polynoms added to the number of polynomials. Furthermore, in order to avoid overfitting, it is recommended to use roughness penalties for the smoothing. This leads to 2 additional parameters to choose from : the penalty as well as the differential operator to be penalised.

As there is no a priori on the degree of smoothness of functions x_i which would be the most relevant for the modelling, several functional variables x are to be tested. On the one hand, higher K favor x_i estimations approaching smooth interpolations of X_i, which can be relevant when the data are not noisy. On the other hand, lower K could allow for more parsimonious and thus robust models. We fit the x_i according to each possible basis size in $[\![2; T]\!]$. For each fixed basis size K, we compute 1 to 200 x curves, varying, when relevant, the degree of the polynoms d from 1 to 5, the number of polynoms b from 1 to $T + 1 - d$, the differential operator D to penalize (no penalized differential operator, a penalized first derivative and a penalized second derivative) and 28 different roughness penalties λ (from 0.01 to 1). For each K, we select the best parameters combination (d, b, D, λ) according to the integrated Mean Squared Error of x considering that :

- the Mean Squared Error (MSE) is based on the differences between $x_i(t)$ and X_{it} at each observed year t : $MSE_x = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_i(t) - X_{it})^2$,
- the integrated Mean Squared Error (MSE^{int}) also takes into account the noise which could be added between observed points due to overfitting : $MSE_x^{int} = \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}}(x_i(t) - \tilde{x}_i(t))^2dt$ where $\tilde{x}$ is the linear interpolation of X, that is a curve obtained by joining the observed points by linear segments. Thus, the lower the distance to the adjusted curve to the linear interpolation, the lower the MSE^{int}.

The same smoothing process is achieved for Y, leading to a set of computed y such that :

$$y(t) = \sum_{k=1}^{K'} \upsilon_{ik}\psi_k(t)$$

with υ_{ik} the coefficients of the basis functions ψ_k associated with y_i.

Thus, $T-1$ smoothed functions are selected for X and for Y. The smoothing procedures are processed via the fda R package.

2.2.2 PLS functional linear regression

Each of the $(T-1)^2$ combinations of functional variables couples (x, y) gives rise to a functional linear model as in equation (1).

For each model, coefficient function β in equation (1) is estimated via functional PLS regression as proposed by [10]. The latter showed the equivalence between the functional PLS regression of y on x and the PLS regression of basis coefficients matrices $(\upsilon)_{ik}$ on basis coefficients $(\xi)_{ik}$. Consequently, $(T-1)^2$ combinations of estimated (x, y) result in $(T-1)^2$ PLS regression models on corresponding basis coefficients matrices $((\upsilon)_{ik}, (\xi)_{ik})$.

Furthermore, for each model induced by (x, y), the number h of PLS components can vary between 0 and T, leading to $T + 1$ potential sub-models. Thus, a total of $(T + 1)(T - 1)^2$ models are to be evaluated. The selection of the optimal $\widetilde{(x, y, h)}$

combination and the estimate of the predictive quality of the model $R^2_{\hat{Y}((\widetilde{x,y,h}))}$ are processed via a double Leave-One-Out (LOO) procedure as in previous subsection ([6], [13]).

Thus,

$$R^2_{\hat{Y}((\widetilde{x,y,h}))} = 1 - \frac{PRESS_{\hat{Y}((\widetilde{x,y,h}))}}{SST_Y}$$

where

- $PRESS_{\hat{Y}(x,y,h)} = \sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - \hat{y}_{(-i)}(t,x,y,h))^2$ is the Predicted Sum of Squared Residuals after outer LOO loop for a model with h components and smoothings x and y,
- $\hat{y}_{(-i)}(t,x,y,h)$ is the predicted value of $y_i(t)$ after the outer LOO loop, that is on a model estimated with all observations but i with h components and x and y smoothings,
- $(\widetilde{x,y,h}) = \arg\min_{(x,y,h)} PRESS_{\hat{Y}_{(x,y,h)}}$ is the optimal combination of (x,y,h) as estimated after the inner LOO loop executed on all farms but i.

By construction, when the prediction error due to the regression model is greater than the dispersion of Y from the mean, $R^2_{\hat{Y}((\widetilde{x,y,h}))}$ can be negative. In this case, the model is useless.

3 Data: INRA network of Charolais beef cattle farms

In order to understand the determinants of macrotrends in beef cattle system farms and capture and analyze the technical and economic potentials of their production systems, an Economics team from INRA UMRH has set and sample-surveyed a Charolais-region farm network since the 1970s. Data have been collected on labour, structure, hectarage and land allocation scheme, herd, intermediate inputs, sales, aids and subsidies, investments and borrowing and allow long-term observational statistics (see [15]). All nominal-value prices internalize currency depreciation using the national consumer price index as deflator.

In the following study, we focus on the farms' income per worker (Y) and on the farm's UAA per worker (X). The farms' income per worker is a particularly complex performance indicator as it includes all the products, the subsidies, the operational and fixed costs (including depreciation costs). However, this indicator represents what is left to pay for labour and own capital and it is thus crucial to understand the determinants of its evolution.

From more than 50 farms performance-monitored per year over the T = 25 years from 1992 to 2016, we formed a constant subsample population of N = 38 farms, all located in the Charolais-area pool. Over this period, the UAA per worker (annual work unit, AWU) has been gradually rising (+ 62% on average), as can be seen in Figure 1a. By way of contrast, the farm income per worker evolution has been chaotic and even slightly decreasing in the last years (Figure 1b).

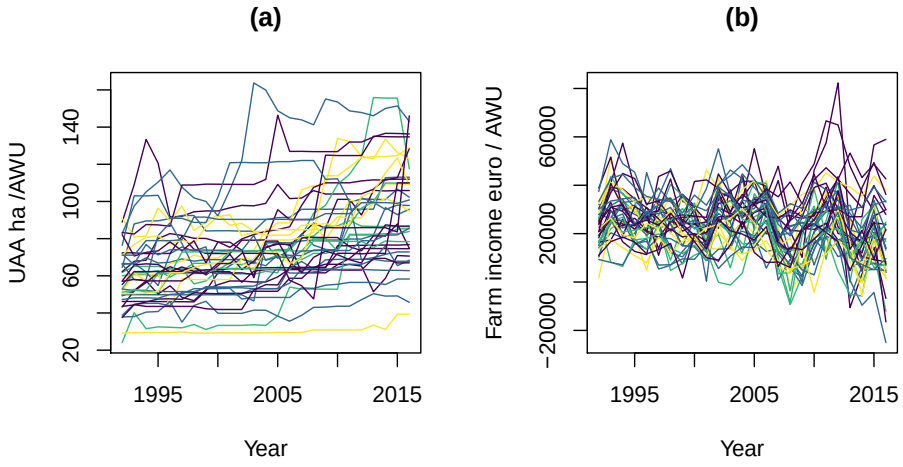

Figure 1: Evolution of observed (a) UAA per worker and (b) income according to the year

The observed yearly points are linearly interpolated

4 Application study: livestock farms' income per worker as a function of its UAA per worker

4.1 Descriptive analysis

As can be seen in Fig. 2, the relationship between income per worker and UAA per worker is positively linear in the first years but no so much after year 2004. We note that the correlation coefficient even becomes slightly negative and the relationship not obviously linear after year 2012 . These results show the need for a more complex model than one only taking into account both characteristics at the same time points.

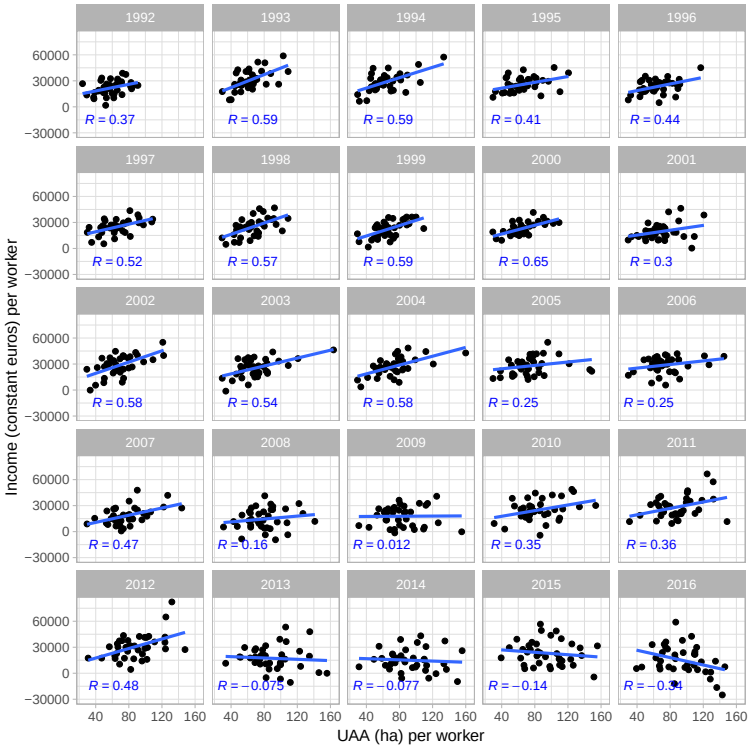

Figure 2: Income per worker as a function of UAA per worker and the year

R : Pearson correlation coefficient

4.2 PLS linear regression

A linear PLS model is fitted on the 25 income per worker variables (1 per year) as responses and 25 UAA per worker variables as predictors.

PLS components 1 and 2 gather 91% of the information generated by the UAA per worker year variables as mentioned in the correlation plot (Fig. 3). The latter also shows that the UAA per worker year-variables are globally positivally intercorrelated and more particularly years 2009 to 2016 on the one hand, and 1992 to 2008 on the other hand. By construction, PLS components 1 and 2 allow the best representation of the income per worker year variables in the UAA per worker space. The most positivally correlated income per worker year variables to PLS component 1 (and thus globally to UAA per worker year variables) are 1999, 2004, 2000, 2012, 1993, 1997. However, the correlations between both sets of variables are quite low. On the opposite side of PLS component 1, 2015 and 2016 income per worker are shown slightly and negatively correlated to the UAA per worker variables. This is consistent with what was seen in Fig. 2.

Thus, as could be expected, the estimated PLS model predictive performance is very low with a predicted total R^2 of -0.07.

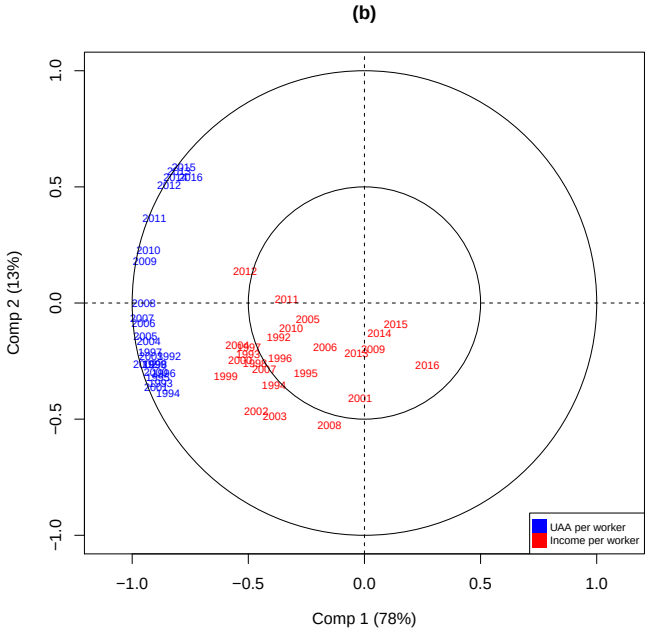

Figure 3: PLS regression model plot

Correlation plot of income per worker and UAA per worker on PLS components 1 and 2

4.3 Computing curves from discrete data

4259 curves were computed for the UAA per worker (vs income per worker) evolution, varying according to the number of b-splines basis functions used (2 to 25) as well as according to 4 other parameters as detailed in section 2. For each b-spline basis size, the optimal curve according to MSE^{int} criterion was selected for further analysis. Thus, 24 curves were selected for the UAA per worker as well as for income per worker evolution.

Figures 4 and 5 show that, as expected, the adjustment quality according to MSE_y and MSE_y^{int} is generally better when the function is adjusted with a larger number of B-splines functions. The maximal (vs minimal) basis size, which allows the finest (vs poorest) fitting to the observed data is shown in Figure 6. An intermediate basis size of 12 (vs 15) appears to be a priori a good compromise between parsimony and fitting quality for the UAA per worker (vs income per worker) according to figure 4 (vs 5) and is also represented on Figure 6.

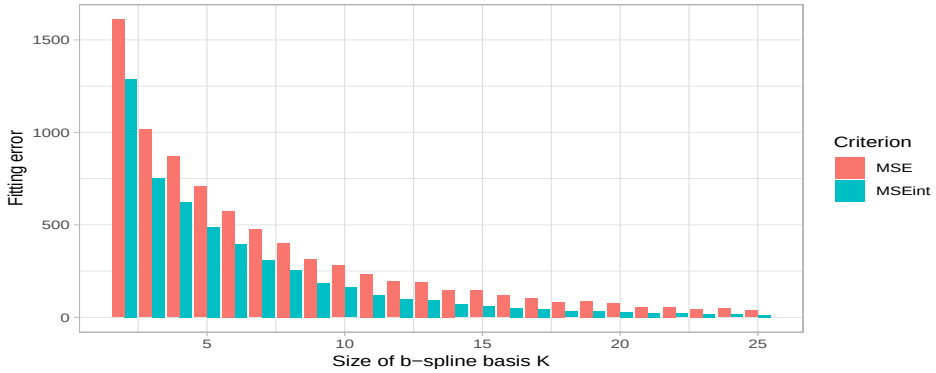

Figure 4: Fitting error of the UAA per worker estimated evolution according to the number of B-spline basis functions K

MSE stands for means squared error - MSEint stands for means squared integrated error

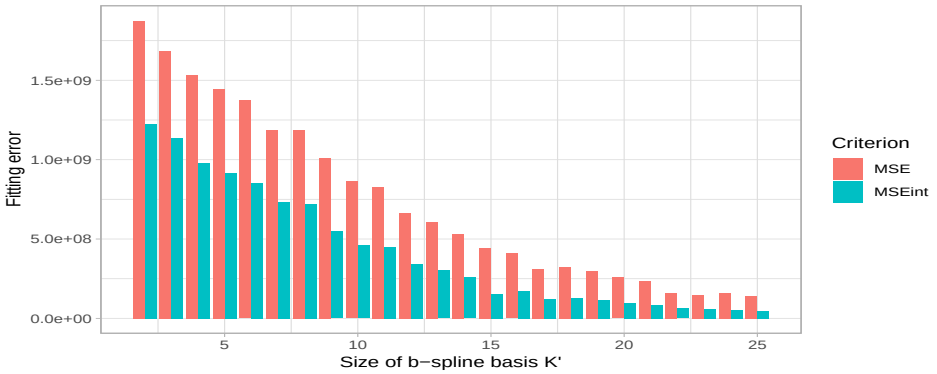

Figure 5: Fitting error of the income per worker estimated evolution according to the number of B-spline basis functions K'

MSE stands for means squared error - MSEint stands for means squared integrated error

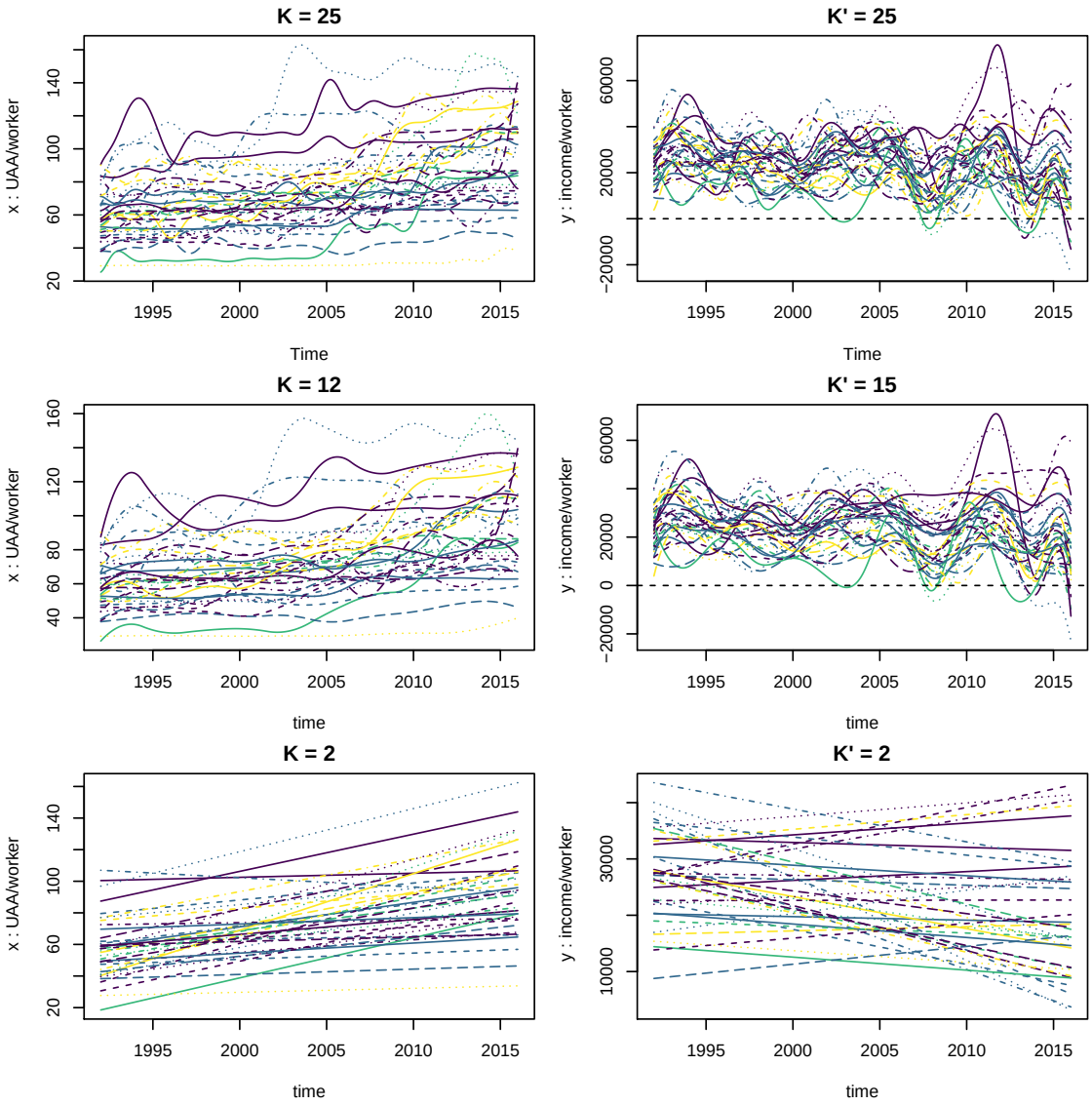

Figure 6: Fitted x (column 1) and y (column 2) according to the time for 3 parameter combinations (rows 1 to 3)

(Row 1) Basis sizes K and K' of 25 - (Row 2) Basis sizes K and K' of 12 and 15 - (Row 3) Basis sizes K and K' of 2

4.4 PLS functional regression

A functional regression model was fitted on the income per worker evolutions as a functional response and the UAA per worker evolutions as a functional predictor. 14400 parameters combinations for the modelling were computed, varying according to the UAA curves fitted ($K \in [\![1, 24]\!]$), the income per worker curves fitted ($K' \in [\![1, 24]\!]$) and the number of PLS components kept

($h \in [\![0, 25]\!]$).

The inner LOO loop allowed the selection of 38 optimal parameters combinations $(\widetilde{x, y, h})$, that is one for each excluded farm and are shown in Table 1. In this table, parameters x and y are represented by their degrees of smoothness K and K'. The optimal combinations are quite stable as 24 out of 38 are identical, as shown in Table 1. Indeed, the most frequent optimal model includes income per worker curves estimated with $K' = 22$ B-splines functions, UAA per worker curves estimated with $K = 12$ B-splines functions and 3 PLS components. Furthermore, the other optimal models have quite close parameters values. Thus, on the one hand, 36 of the optimal models use response curves close to linear interpolations of the observed incomes per worker, with $K' > 21$. On the other hand, in these models, all the observed UAA/worker are fitted with K varying from 9 to 17, which stand for a moderate smoothing. Besides, 3 or 4 PLS components are systematically chosen.

$\tilde{K'}$	$\tilde{K}$	$\tilde{h}$	Frequency
15	14	4	1
17	9	3	1
22	12	3	1
22	12	4	1
22	14	3	24
25	12	4	9
25	17	4	1

Table 1: Optimal parameters estimated from the inner LOO loop for functional PLS regression

However, the satisfying consistency of these results are to be put into perspective. Indeed, the predictive $R^2_{\hat{y}}$ for optimal h, x, y is estimated after the outer LOO loop at about 0.01. In other words, our functional model (1) is useless as it stands and needs further improvement.

5 Discussion and conclusion

In this paper, we propose a method that could be relevant in the modelling of farm performance evolution. In our case study, though, none of the models tested, including functional data regression, make it possible to predict correctly the farms' income per worker.

Concerning the two-steps functional regression method used, one limit raised by [5] is that since the fitted continuous curves are further considered as if they were observed curves, the measurement and fitting errors are not taken into account in the functional regression model. In contrast, [5] proposes a framework which allows direct modelling of the observed data and expands all model terms in suitable basis expansions. In our case, though our models did not, our final prediction quality criterion has been estimated against the real observed values.

This paper introduces the application of functional regression in agricultural economics and in particular to the income per worker modelling. As mentioned in the literature, the link between farms' performances and their size is complex. Moreover, the income per worker is also a complex performance indicator including subsidies. Indeed, the evolution of farms' UAA per worker is found to be insufficient to explain the evolution of the farms' income per worker, as was expected. It will therefore be necessary to improve the model. A first obvious option would be to introduce other explanatory functions corresponding to the evolution of other structural, technical or economic characteristics - like beef price. A second option would be to make the model more parsimonious by restricting the time interval on which performance is linked by predictor functions. At last, evolutions' derivatives could be relevant to predict the performances evolution.

More research is thus needed to understand the unsatisfactory evolution of farms' income per worker evolution and more widely farms' performances evolution. Functional regression is promising in the new perspectives it can brings in this field and will hopefully contribute to the political debate on the relevance of increasing farms' size.

References

[1] S. Bojnec and I. Ferto. Farm income sources , farm size and farm technical efficiency in Slovenia. *Post-Communist Economies*, 25(3):343–356, 2013.

[2] S. Bojnec and L. Latruffe. Farm Size and Efficiency during Transition: Insights from Slovenian Farms. *Transformations in Business & Economics*, 10(3), 2011.

[3] Š. Bojnec and L. Latruffe. Farm size, agricultural subsidies and farm performance in Slovenia. *Land Use Policy*, 32:207–217, 2013.

[4] K. Deininger, S. Jin, Y. Liu, and S. K. Singh. Can Labor-Market Imperfections Explain Changes in the Inverse Farm Size - Productivity Relationship ? Longitudinal Evidence from Rural India. *Land Economics*, pages 239–259, 2018.

[5] S. Greven and F. Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2):1–35, 2017.

[6] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. 2009.

[7] H. Henderson. Considering Technical and Allocative Efficiency in the Inverse Farm Size - Productivity Relationship. *Journal of Agricultural Economics*, 66(2):442–469, 2015.

[8] T. Lefroy, J. Key, and R. Kingwell. An Examination of Broadacre Farm Size and Performance in Western Australia. *The Australian Economic Review*, 50(1):52–65, 2017.

[9] J. S. Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2:321 – 359, 2015.

[10] C. Preda and J. Schiltz. Functional PLS regression with functional response : the basis expansion approach . In *Proceedings of the 14th Applied Stochastic Models and Data Analysis Conference*, pages 1126–1133. Università di Roma La Spienza, 2011.

[11] J. O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and Matlab*. 2009.

[12] M. Tenenhaus. *La régression PLS : théorie et pratique*. Ed. Technip, 1998.

[13] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 2006.

[14] P. Veysset, M. Lherm, P. Natier, and J. P. Boussemart. Formation et répartition des gains de productivité en élevage bovin viande . Qui sont les gagnants et les perdants sur les 35 dernières années ? (1):11–14, 2017.

[15] P. Veysset, M. Lherm, M. Roulenc, C. Troquier, and D. Bébin. Productivity and technical efficiency of suckler beef production systems: Trends for the period 1990 to 2012. *Animal*, 9(12):2050–2059, 2015.