MODELLING INDIVIDUAL DECISIONS TO SUPPORT THE EUROPEAN POLICIES RELATED TO AGRICULTURE

Deliverable D3.4: Report on modelling crop management practices and interfaces to the MIND STEP model toolbox

AUTHORS	Fabienne Femenia (INRAE), Alain Carpentier (INRAE), Obafemi Philippe Koutchade (INRAE), Elodie Letort (INRAE), Scarlett Wang (WU), Frederic Ang (WU), Paolo Sckokai (UNICATT), Alessandro Varacca (UNICATT).
APPROVED BY WP MANAGER:	John Helming (WR)
DATE OF APPROVAL:	27-02-27 (v1) 21-04-2023 (v2)
APPROVED BY PROJECT COORDINATOR:	Hans van Meijl (WR)
DATE OF APPROVAL:	28-02-2023 (v1) 21-04-2023 (v2)

CALL H2020-RUR-2018-2	Rural Renaissance
WORK PROGRAMME Topic RUR-04-2018	Analytical tools and models to support policies related to agriculture and food - RIA Research and Innovation action
PROJECT WEB SITE:	https://mind-step.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773901.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

TABLE OF CONTENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

LIST OF FIGURES

LIST OF TABLES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

1. INTRODUCTION

One of the main objectives of the MIND STEP project is to include individual decision making (IDM) unit in policy models. Among these IDM units, are innovative microeconomic models of farmers' production choices that have been developed in task 3.4. Part of these models are micro-econometric models and aim at empirically analyzing crop farmers' choices in terms of yields, chemical input uses, acreages and crop management practices (CMP) and dairy farmers' choices in terms of feeding strategy and land allocation. Those models are primarily specified for exploiting the information contained in available cost accounting datasets and are estimated for two main purposes: being used directly as simulation models, or providing behavioral parameters to – more complex – policy simulation models (e.g., IFM-CAP, GLOBIOM, CAPRI, MAGNET). Another type of model, based on a Data Envelopment Analysis (DEA) framework has been developed in task 3.4 in order to investigate the potential of optimizing land allocation between crops and grassland to reduce greenhouse gas (GHG) emissions.

A significant part of task 3.4 builds on the micro-econometric multi-crop (MEMC) models developed in recent years by partner INRAE. These models are random – farm specific – parameter models which, once estimated, allow the calibration of technical and behavioral parameters at the farm level based on a well-defined statistical background. Within task 3.4, these models have been extended to account for the decision of farmers to choose to produce specific subsets of crops among all the crops they could produce. Significant work has also been undertaken during the MIND STEP project to allow the estimation of MEMC models by other partners of the project. A first set of work has been carried out to enable the estimation of these models on data available for European member states, i.e. FADN data. Indeed, the estimation of MEMC models requires information on input uses per crop, while the FADN data only contain input expenditures at the farm level. An original procedure has consequently been proposed to allocate input uses at the farm level to input uses per crop based. A second set of work has been done to alleviate the significant estimation burden of MEMC models and enabling their use on a routine basis, as well as for incorporating CMP choice in these models.

In addition to these works on MEMC models, research has been conducted in Task 3.4 to identify the heterogeneity in the flexibility of dairy farms based on their observed short run responses, in terms of feed concentrate uses and acreage adjustments, to input and output prices. This work aims at providing an economic explanation to the relative rigidity of dairy farms, which may in fact be due to the existence of unobserved adjustment costs, and at identifying, among a sample of farmers, those who appear to be most flexible in the short run. For this purpose, an analytical framework, based on random parameter panel smooth transition regression (PSTR) model implicitly accounting for the impacts of input adjustment costs on dairy farmers' production decisions, has been proposed and estimated, as an illustrative purpose, on a sample of French dairy farms.

Finally, partner WU has developed an integrated multi-production technology framework to investigate the potential of land optimization in dairy farms to reduce GHG emissions. Indeed, circular farming has been proposed as a cost-effective way to reduce GHG emissions by the Dutch Ministry of Agriculture in 2019. Dutch dairy farms have already incorporated circular

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

principles in their farming activities, e.g. upcycle manure as crop fertilizers, use crop residuals for animal feed. However, no study until now has quantified the technical and environmental efficiency of dairy farms which incorporate these circularity principles, and explores land reallocation. The framework proposed by WU combines a by-production approach with a network DEA model to answer these questions.

The key messages of D3.4 are the following:

- Methodological developments

First, micro-econometric models of farmers' production choices can be substantially improved by accounting for unobserved heterogeneity, adjustment costs or production practice choices. Second, mathematical programming models and micro-econometric models are complementary for improving simulation models based on individual farms. For instance, parameter estimates obtained from random parameter micro-econometric models can be used for calibrating mathematical programming models at the farm level. Micro-econometric models of input allocations to crops can be used for estimating input uses at the crop level that can be in turn used to feed mathematical programming models.

- Policy implications of obtained empirical results

The empirical results presented in D3.4 show that farmers tend to be efficient from a technical viewpoint given their current production technologies. This result is welcome as famers' efficiency underlies the economic models considered for policy analysis. It also suggests that solving environmental issues implies to solve economic trade-offs that involve changes in production technologies. The results obtained in D3.4 also show that farmers respond to economic incentives, even if their responsiveness display significant heterogeneity, implying that economic policy instruments could be useful for achieving the objectives of the EU Green Deal. Finally, some empirical results demonstrate the importance and heterogeneity of farmers' adaptability in response price variations on agricultural markets. This heterogeneity can mostly be attributed to factors that are specific to each farmers and not observable in economic data. Further investigation, through targeted surveys for instance, might reveal some of these factors and allow designing policies, targeted on higher education or training for instance, which could improve farmers' adaptability.

- Data needs

The work presented in D3.4 show that public authorities should invest in more accurate data collection. Of course, research regularly complain about data lacking. Yet, the results we obtain document the adverse consequences of these missing data issues on the ability to assess the effects of agri-environmental policies. For instance, the work on input cost allocation tends to show that micro-econometric models allow to obtain reliable estimates of input uses for major crops but much more questionable estimates of input uses for minor crops. Indeed, the fact that FADN data report input uses at the farm level (standard accountancy data) instead of at the crop and farm level (cost accounting data) can only be imperfectly overcome. This provides interesting research topics for micro-econometricians but lead to imprecise inputs to simulation models. Of course, collecting cost accounting data would be costly. A less costly option would be to collect mean input use levels at the NUTS1 or NUTS2 level. This information could be useful for improving the cost allocation modules proposed in D3.4. This information, however, is unavailable for many Member States. The same observation holds for innovative production practices, whether agronomic practices

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(e.g., low input practices, biocontrol techniques) or precision agriculture techniques. This is unfortunate for economists involved in the quantitative assessment of agri-environmental policies but also for farmers interested in changing their production practices.

Research conducted in Task 3.4 for accounting for crop choices in MEMC models is presented in section 2. The proposed approach for allocating input costs to crops and hence enabling the estimation of these models on FADN data is presented in section 3, while section 4 is devoted to the work undertaken to alleviate their estimation burden. The PSTR model of dairy farms' production decisions is presented in section 5. Finally, section 6 is devoted to the presentation of the efficiency model developed by WU in the task.

2. MICRO-ECONOMETRIC MULTI-CROP MODEL WITH ENDOGENOUS REGIME CHOICE[1]

2.1. Introduction

Market prices and agricultural policies impact crop supplies through their effects on input uses and yield levels, and acreage choices. Starting in the eighties with the pioneering work of Just et al (1983), Chambers and Just (1989) and Chavas and Holt (1990), agricultural production economists developed micro-econometric multi-crop (MEMC) models for analyzing and quantifying these effects with farm accountancy data. These models have then been widely applied during the last decades.

In MEMC models, farmers are assumed to allocate their cropland to the crops of a given crop set in order to maximize their expected profit or the expected utility of their profit. This ensures the economic consistency of the resulting models. However, currently available MEMC models ignore or poorly describe an important decision of crop producers: their choice to produce a subset of crops among the set of crops they can produce and sell. Indeed, applications of MEMC models frequently ignore null acreages by relying either on very specific farm samples (e.g., Just et al, 1983, 1990; Bayramoglu and Chakir, 2016) or on crop aggregation that eliminate null crop acreages (e.g., Oude Lansing and Peerlings, 1996; Serra et al, 2005; Oude Lansink, 2008; Carpentier and Letort, 2012, 2014). Yet, sample selection prevents extrapolation of the estimation results to farmers not producing all considered crops while crop aggregation induces information losses regarding production choices at the crop level.

A few recent MEMC models explicitly account for null crop acreages (e.g., Sckokai and Moro, 2006, 2009; Lacroix and Thomas, 2011; Bateman and Fezzi, 2011; Platoni et al, 2012). These models are designed as censored regression (CR) systems and are estimated following two-step approaches inspired by that initially proposed by Shonkwiler and Yen (1999). These MCEM based on censored

[1] The work presented in this is section has led to a publication in the American Journal of Agricultural Economics (AJAE) (Koutchadé et al, 2021). A significant part of the results presented in the paper have been obtained within the MIND STEP project. This is reflected in the dates of submission and resubmission of the paper: the initial version, submitted prior to the MIND STEP project (in February 2019), did not contain the empirical application focusing on the evaluation of the EU support to protein peas while the revised version submitted in May 2020 did. As acknowledgement to MIND STEP is missing in the publication, INRAE is in contact with the AJAE editor to correct this.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

regressions (CR-MCEM) suitably account for null acreages from a statistical viewpoint but display severe micro-economic inconsistencies.

Indeed, these models are conceived as statistical versions – featuring error terms and accounting for mass points at 0 – of theoretical micro-economic models ignoring null acreages. Their main shortcoming is due to their relying on a single crop acreage choice model, whatever the subset of crops actually produced. These models thus fail to recognize that the crop acreage choices of a farmer structurally depend on the composition of the set of crops actually produced by this farmer. For instance, farmers are unlikely to consider the prices of the crops they don't produce when choosing the acreages of the crops they produce. Of course, the lack of micro-economic coherency of CR-MCEM models substantially undermines their ability to yield consistent estimates of crop acreage responses to economic incentives. The composition of the produced crop sets, which displays substantial variability when null acreages are frequent, deeply impacts the structure of farmers' crop acreage choices. These effects of farmers' crop set choices are ignored in CR-MCEM models.

The recent articles addressing the issue raised by null crop acreages from a statistical viewpoint by considering CR-MEMC models don't focus either on corner solutions in acreage choices or on farmers' crop choices. By contrast, the main objective of this article is to develop a consistent modelling framework for analyzing farmers' crop set choices and, as a result, for handling null crop acreages in MEMC models. More precisely, the main aims of this article are (a) to revisit the null acreage issue in multi-crop models from a theoretical viewpoint, (b) to propose an original MEMC model that accounts for farmers' crop choices in a way that is consistent from an economic viewpoint, together with a suitable estimation approach, and (c) to show, by means of an empirical application focused on crop diversification choices, that considering crop set choices significantly enriches micro-econometric analyses of farmers' crop supply.

Our multi-crop micro-economic modelling framework is based on an expected profit maximization problem considering land as an allocable quasi-fixed input. This problem includes the usual crop acreage non-negativity constraints but also production regime fixed costs. The production regime chosen by a farmer is defined by the subset of crops that this farmer decides to plant. The regime fixed costs consist of unobserved costs – such as unmeasured marketing costs or implicit labor and machinery management costs – that depend on the set of crops that are grown but that don't depend on the acreages of these crops. Accordingly, our modelling framework assumes that farmers choose the production regime maximizing their expected profit, regime fixed costs included, as well as the related optimal crop acreage, yield and input use levels. Importantly, our considering regime fixed costs implies that null acreages are not necessarily due to binding non-negativity constraints.

Based on this micro-economic background, we design our MEMC model as an endogenous regime switching (ERS) multivariate model with multiple regimes. This endogenous regime switching micro-econometric multi-crop (ERS-MEMC) model consists of a probabilistic regime choice model coupled with a set of regime specific MEMC models. As estimating multivariate ERS models with multiple regimes is challenging and considering regime specific fixed costs increases the estimation burden, choosing relevant functional forms for the per regime MEMC models appears crucial. Thanks to their specific properties, the Multinomial Logit (MNL) acreage choice models proposed by Carpentier and Letort (2014) are particularly well suited in that respect. They yield simple and well-behaved functional forms for important components of our ERS-MEMC model, thereby significantly reducing its estimation cost. Relying on the MNL acreage choice

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

models also enables us, following Koutchadé et al (2018), to go one step further and to account for farmers' unobserved heterogeneity by considering a random parameter version of our ERS-MEMC model. Estimating ERS models with multiple regimes is challenging mostly because their likelihood function involves integration of expectations over the probability distribution of multivariate latent error terms (e.g., Pudney, 1989). Also, the likelihood function of our ERS-MEMC model needs to be integrated over the probability distribution of its random parameters. Our estimation approach combines tools from the micro-econometrics and computational statistics literatures.

We illustrate the empirical tractability of our approach by estimating our model for a panel data sample of French arable crop producers. Our results tend to demonstrate that our random parameter ERS-MEMC model performs well according to standard fit criteria. They also tend to show that regime specific fixed costs significantly matter in farmers' crop choice, along with crop expected returns. Importantly, these results also demonstrate that acreage choices' responses to economic incentives strongly depend on the production regime choices. The elasticity of crop acreages in crop prices increases in the number of produced crops, a pattern that cannot be reproduced by CR-MEMC models. Finally, our simulation results show that the acreage of minor crops respond non-linearly to increases in their prices due to production regime changes.

Our contributions are twofold. First, ERS-MEMC model presented in this article accounts for null crop acreages while relying on a well-defined micro-economic background. As a result, it is the first theoretically coherent response to an issue that is pervasive when analyzing crop production with farm level data. Other ERS-MEMC models could be considered, but the one presented here allows to consider production regime fixed costs as well as farm specific parameters while remaining empirically tractable. Second, this model allows to disentangle the effects of the main economic drivers of farmers' crop supply choices. It accounts for intensive and extensive margin choices, including the effects on crop set choices at the extensive margin. This unique feature is of special interest for investigating future agri-environmental policies. In particular, owing to its positive agronomic effects, crop diversification is a key feature of environmentally friendly crop production systems (e.g., Matson et al, 1997; Tilman et al, 2002; Lin, 2011; Kremen et al, 2012; Bowman and Zilberman, 2013). Our modelling framework is especially well-suited for analyzing samples containing both specialized and diversified farms as well as for simulating the effects of policy instruments aimed to foster crop diversification.

The rest of this article is organized as follows. The approach proposed to account for crop choices in micro-economic models of acreage decisions is presented in the first section. The structure of the corresponding ERS-MEMC model is described in the second section. The main features of our estimation strategy are presented in the third section, with a specific focus on the main issues arising with random parameter ERS-MEMC models. Illustrative estimation and simulation results are provided in the fourth section. Finally, we conclude.

2.2. Regime switching in multi-crop models: corners, kinks and jumps

This section presents the theoretical modelling framework we propose for dealing with null crop acreages in micro-econometric acreage choice models. We proceed in three steps. First, we present the micro-economic crop acreage choice model underlying our ERS-MEMC model. Second, we compare this model to the models that have been proposed for modelling multiple binding non-negativity constraints or regime switching. We focus on the ability of these models

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

to cope with corners, kinks and jumps in farmers acreage choices. Third, we present the functional form of the crop acreage choice models used in our ERS-MEMC model.

2.2.1. Crop choices and crop acreages

We assume that farmers can allocate their fixed cropland area to K crops. Accordingly, set $\mathcal{K} = \{1,...,K\}$ denotes the set of crops that any considered farmer can produce and sell and farmers' problem consists of optimally choosing a crop acreage share vector $\mathbf{s} = (s_k : k \in \mathcal{K})$ satisfying $\mathbf{s} \geq \mathbf{0}$ and $\mathbf{s}'\boldsymbol{\iota} = 1$, term $\boldsymbol{\iota}$ being the dimension K unitary column vector.

We now introduce notions and notations aimed at describing farmers' decisions to produce a subset of crops among crop set $\mathcal{K}$. Set $\mathcal{R} = \{1,...,R\}$ denotes the set of feasible production regimes. A production regime is defined by the subset of crops with strictly positive acreages. Set $\mathcal{K}^{+}(r)$ denotes the subset of crops planted in regime r while $\mathcal{K}^{0}(r)$ denotes its complement to $\mathcal{K}$, that is to say the subset of crops that are not planted in regime r. Finally, function $r(\mathbf{s})$ defines the regime of the acreage share vector $\mathbf{s}$.

We assume that farmers are risk neutral. In year t farmer i is assumed to choose her/his crop acreages by solving the following expected profit maximization problem:

$$(1) \qquad \max_{\mathbf{s}} \left\{ \mathbf{s}'\boldsymbol{\pi}_{it} - C_{it}(\mathbf{s}) - D_{it}(r(\mathbf{s})) \text{ s.t. } \mathbf{s} \geq \mathbf{0} \text{ and } \mathbf{s}'\boldsymbol{\iota} = 1 \right\}$$

Term $\boldsymbol{\pi}_{it} = (p_{k,it} : k \in \mathcal{K})$ is the vector of crop returns expected by farmer i when choosing $\mathbf{s}$ in year t. Function $C_{it}(\mathbf{s})$ is the implicit management cost of acreage $\mathbf{s}$ and $D_{it}(r)$ is the fixed cost of production regime r incurred by farmer i in t. This cost is fixed in the sense that it doesn't depend on $\mathbf{s}$.

Acreage management costs $C_{it}(\mathbf{s})$ are costs not included in the crop gross margins that vary in $\mathbf{s}$. They include unobserved variable input costs. They also account for the implicit costs related to constraints on acreage choices due to limiting quantities of machinery or labor, or to agronomic factors. These constraints providing motives for diversifying crop acreages, function $C_{it}(\mathbf{s})$ is assumed to be convex in $\mathbf{s}$. In order to ensure that the solution in $\mathbf{s}$ to problem (1) is unique, we strengthen this assumption by assuming that function $C_{it}(\mathbf{s})$ is strictly convex in $\mathbf{s}$.[2] These crop acreage management costs prevent farmers to solely produce the most profitable crop.

Regime fixed cost terms $D_{it}(r)$ introduce discrete elements, and thus severe discontinuities, in farmers' acreage choices. These costs do not depend on the chosen acreage in a given regime, they only depend on the crop set defining this regime. They account for the hidden fixed costs incurred by the farmer for any acreage choice in the considered regime, such as fixed costs related to the marketing process of the crop products or those incurred when purchasing specific variable inputs, when renting specific machines, when seeking crop specific advises, *etc*. These regime fixed costs may also depend on characteristics of crop

[2] Analogous cost functions are used in the Positive Mathematical Programming literature (*e.g.*, Mérel and Howitt, 2014; Heckelei *et al*, 2012) and in the multi-crop econometric literature (*e.g.*, Heckeleï and Wolff, 2003; Carpentier and Letort, 2012, 2014).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

biological cycles. For instance, part-time farmers may decide not to produce a given crop because the management of this crop is not compatible with their other non-farming activities.

The smooth acreage management cost function $C_{it}(\mathbf{s})$ and the discontinuous regime fixed cost function $D_{it}(r(\mathbf{s}))$ are expected to impact farmers' crop diversification in opposite directions. While limiting quantities of quasi-fixed factors impose constraints fostering crop diversification, regime fixed costs are expected to foster crop specialization. In particular, the regime fixed costs are expected to be non-decreasing in the number of produced crops.[3]

We solve farmers' expected profit maximization problem following a standard backward induction approach according to which farmers choose their production regime after examining their expected profit in each possible production regime.

First, the acreage choice problem is solved for each potential regime. This yields the regime specific optimal acreage shares:

(2a) $\mathbf{s}_{it}(r) = \arg\max_{\mathbf{s}} \{\mathbf{s}'\boldsymbol{\pi}_{it} - C_{it}(\mathbf{s}) \text{ s.t. } \mathbf{s} \geq \mathbf{0},\ \mathbf{s}'\boldsymbol{\iota} = 1 \text{ and } s_k = 0 \text{ if } k \in \mathcal{K}^0(r)\}$

and the regime specific optimal expected profit levels (regime specific fixed costs excluded):

(2b) $\Pi_{it}(r) = \max_{\mathbf{s}} \{\mathbf{s}'\boldsymbol{\pi}_{it} - C_{it}(\mathbf{s}) \text{ s.t. } \mathbf{s} \geq \mathbf{0}, \mathbf{s}'\boldsymbol{\iota} = 1 \text{ and } s_k = 0 \text{ if } k \in \mathcal{K}^0(r)\}$.

for $r \in \mathcal{R}$.

Second, the optimal production regime r_{it} is determined by comparing the regime specific expected profit levels while accounting for the production regime fixed costs. Accordingly, the expected profit maximizing production regime r_{it} is defined as the solution in r to a simple discrete maximization problem with:

(3) $r_{it} = \arg\max_{r \in \mathcal{R}} \{\Pi_{it}(r) - D_{it}(r(\mathbf{s}_{it}(r)))\}$.

Assuming that optimal regime r_{it} is unique, optimal acreage choice $\mathbf{s}_{it}$ is obtained by combining equations (3) and (2a), with:

(4a) $\mathbf{s}_{it} = \mathbf{s}_{it}(r_{it})$.

Similarly, equations (3) and (2b) yield the expected profit level Π_{it}, with:

(4b) $\Pi_{it} = \Pi_{it}(r_{it})$.

Regime specific acreage choices $\mathbf{s}_{it}(r)$ are derived from optimization problems that differ from one regime to the other due to nullity constraints on crop acreages. These constraints significantly impact how the acreage choices of the produced crops respond to market

[3] Note however that in specific empirical settings the $D_{it}(r)$ terms may also capture the effects of exogenous factors preventing farmer *i* to produce specific crops, *e.g.* due to unsuitable soils or to lacking outlets. In the empirical application presented in section 4, such features are unlikely to occur. Our sample covers a limited geographical area and we only consider crops which can be profitably produced in this area.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

conditions. For instance, the regime *r* acreage choice, $\mathbf{s}_{it}(r)$, doesn't respond to changes in the expected returns of the crops not produced in regime *r*. Similarly, acreages of produced crops are expected to be more responsive to economic incentives in regimes containing numerous crops than in regimes containing only a few crops, crop acreage substitution opportunities being more limited with small crop sets.

2.2.2. Corners, kinks and jumps in acreage choice models

Our micro-economic crop acreage choice model is an example of ERS multivariate model with multiple regimes. To our knowledge, ERS models for multiple choices have been mostly used for demand systems, either for consumption goods (*e.g.*, Wales and Woodland, 1983; Lee and Pitt, 1986; Kao *et al*, 2001; Millimet and Tchernis, 2008) or for production factors (*e.g.*, Lee and Pitt, 1987; Arndt, 1999: Chakir and Thomas, 2003). Most of these studies rely on the dual modelling framework proposed by Lee and Pitt (1986).

The main differences between the approaches that can be considered for handling null acreages in MEMC models are illustrated schematically in Figure 1. Panels (a)-(c) depict how the crop acreage of a given crop depends on its expected return according to three multi-crop acreage models. These models differ on how they handle null acreage choices – based either on ERS models or on CR systems – and on whether they account for crop or regime production fixed costs or not. Indeed, Figure 1 shows that this comparison is all about "corners", "kinks" and "jumps".

Models that account for null acreages and don't account for crop production fixed costs are defined as systems of standard Tobit models (*e.g.*, Moore and Negri, 1992; Moore *et al*, 1994). They define null acreages as corner solutions at zero. Their crop acreage models display one kink at the crop return level at which the non-negativity constraint of the considered crop just bind, as illustrated in panel (a).

Panel (b) depicts patterns allowed by models that account for null acreages based on CR systems as well as for crop production fixed costs. These crop acreage choice models display one kink and, potentially, a jump at the crop return level where farmers are indifferent between planting the considered crop or not. Being based on extensions of generalized Tobit models, recent CR-MEMC models (*e.g.*, Sckokai and Moro, 2006, 2009; Lacroix and Thomas, 2011; Bateman and Fezzi, 2011; Platoni *et al*, 2012) implicitly account for production regime costs.

Crop acreage choices patterns allowed in our ERS-MEMC model are depicted in panel (c). Due to the effects of the regime choices on acreage choices, crop acreages may display several kinks. A kink occurs wherever changes in the expected return of the considered crop induce a regime switch. The first kink occurs at the crop return level above which farmers decide to plant the considered crop while others occur at regime switch points concerning the decision to produce or not to produce other crops. Our ERS-MEMC may also induce jumps at regime switch points, these jumps being due to threshold effects induced by regime fixed costs. According to our knowledge, this is the first MEMC model allowing such crop choice patterns.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

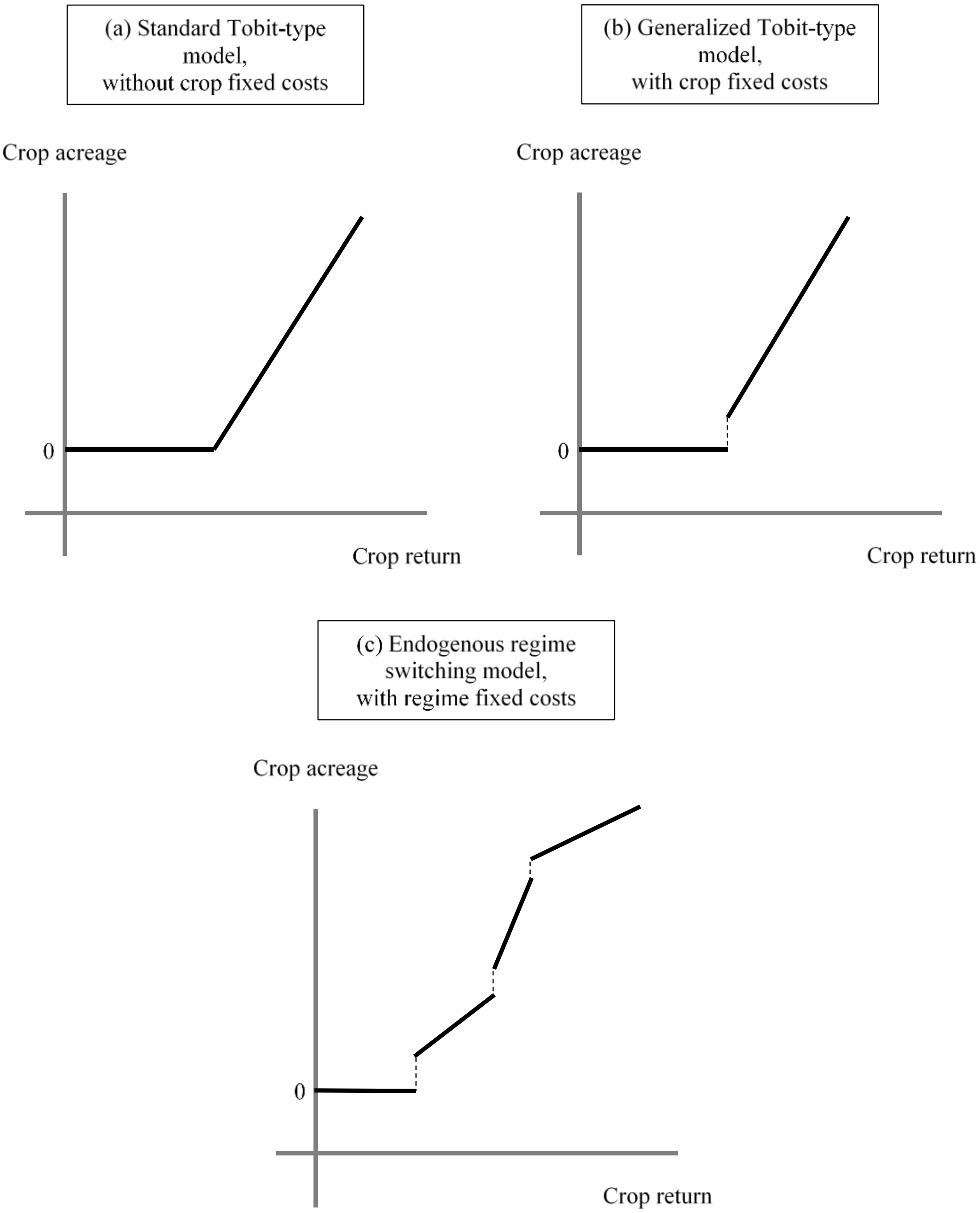

Figure 1. Typical multi-crop acreage models handling null crop acreages

2.2.3. Crop choices and MNL acreage choice models

The regime fixed cost considered in the maximization problem (3) determining the optimal regime r_{it} is $D_{it}\left(r\left(\mathbf{s}_{it}(r)\right)\right)$ rather than simply $D_{it}(r)$. In effect, the production regime of $\mathbf{s}_{it}(r)$ may not be regime r, depending on the functional form chosen for the cost function $C_{it}(\mathbf{s})$. The regime of $\mathbf{s}_{it}(r)$ is only guaranteed to be a regime "included" in regime r as elements of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$\mathbf{s}_{it}(r)$ may be null due to binding non-negativity constraints. The production regime of $\mathbf{s}_{it}(r)$ is regime *r* if and only if $s_{k,it}(r)$ is an interior solution to problem (2a) for any $k \in \mathcal{K}^{+}(r)$.

For instance, if $C_{it}(\mathbf{s})$ is quadratic in **s** then $s_{k,it}(r)$ is null if $\pi_{k,it}$ is sufficiently low. Moreover, neither crop acreage $\mathbf{s}_{it}(r)$ nor expected profit $\Pi_{it}(r)$ are obtained in analytical closed form in the quadratic case, precisely because elements of $\mathbf{s}_{it}(r)$ may be corner solutions at 0.

By contrast, the Multinomial Logit (MNL) crop acreage share models proposed by Carpentier and Letort (2014) appear especially convenient in this context.[4] This modelling framework relies on a family of acreage management cost functions ensuring that optimal crop acreage shares $\mathbf{s}_{it}(r)$ and expected profit levels $\Pi_{it}(r)$ satisfy two important conditions for any regime *r*.

First, these terms are obtained in analytical closed forms. For instance, if the acreage management cost function is assumed to have the linear-entropic functional form $C_{it}(\mathbf{s}) = \sum_{k \in \mathcal{K}^{+}(r)} s_k \beta^{s}_{k,it} + (\alpha^{s}_{i})^{-1} \sum_{k \in \mathcal{K}^{+}(r)} s_k \ln s_k$ with $\alpha_i > 0$ then the regime specific acreage share vectors $\mathbf{s}_{it}(r)$ are given by Standard MNL acreage share models:

$$(5) \qquad s_{k,it}(r) = \frac{\varphi_k(r)\exp\left(\alpha^{s}_{i}(\pi_{k,it} - \beta^{s}_{k,it})\right)}{\sum_{l \in \mathcal{K}} \varphi_l(r)\exp\left(\alpha^{s}_{i}(\pi_{l,it} - \beta^{s}_{l,it})\right)} \text{ for } k \in \mathcal{K}.$$

where function $\varphi_k(r)$ indicates whether crop *k* belongs to regime *r* or not; with $\varphi_k(r) = 1$ if $k \in \mathcal{K}^{+}(r)$ and $\varphi_k(r) = 0$ otherwise. Second, it is easily seen from equation (5) that, for Standard MNL acreage share models, if crop *k* belongs to regime *r* then the optimal acreage share of crop *k* in regime *r* is ensured to be strictly positive. More generally, considering Standard or Nested MNL crop acreage models ensures that the production regime of $\mathbf{s}_{it}(r)$ is regime *r*.

The fact that $s_{k,it}(r)$ cannot be null means that null crop acreages are handled in a specific way in the MNL modelling framework. Crop acreage non-negativity constraints never bind when deriving MNL acreage share models.[5] These constraints just imply that the optimal acreage shares of the least profitable crops (acreage management cost included) are very small when they are much less profitable than other crops of the considered crop set.[6] The acreage shares

[4] Of course, choosing functional forms for their being convenient is unwarranted. Yet, their estimation being particularly challenging, all specifications of ERS models with multiple regimes that were used in empirical studies exploit, to some extent, properties of specific functional forms (e.g., Wales and Woodland, 1983; Lee and Pitt, 1986, 1987; Arndt et al, 1999). Also, other properties of MNL acreage share models make them empirically relevant for modelling production choices of arable crop producers (Carpentier and Letort, 2014).

[5] This property comes from properties of the entropy terms that appear in the acreage cost management functions leading to MNL acreage share models (Carpentier and Letort, 2014). Term $-s_k \ln s_k$ tends to 0 as s_k decreases to 0 (we have $s_k \ln s_k = 0$ if $s_k = 0$ according to a standard extension by continuity result) while its derivative in s_k tends to infinity as s_k decreases to 0.

[6] It is easily seen, from equation (5), that $s_{k,it}(r)$ decreases to 0 as $\pi_{k,it}$ decreases to $-\infty$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

of the least profitable crops may only become null when farmers choose their production regime. Farmers exclude these crops from their production plans when they can get higher expected profit level without planting them. Incidentally, this feature of MNL acreage choice models prevents their use in CR-MEMC models.

2.3. ERS-MEMC model with regime specific fixed costs: micro-economic structure

This section presents the structure of the ERS-MEMC model considered in the empirical application presented in the next section. This model is composed, on the one hand, of yield supply functions, variable input demand functions and acreage share choice models for each produced crop, and on the other hand, of a probabilistic production regime choice model. This MEMC model can be interpreted as an extension to an ERS framework with regime fixed costs of the model proposed by Carpentier and Letort (2014).

As in Koutchadé *et al* (2018) we adopt a random parameter approach for accounting for farmers' and farms' unobserved heterogeneity. We assume that the parameters of farmers' production choices, including those driving farmers' responses to economic incentives, are farm specific. Accordingly, the main aim of the estimation procedure is to recover their distribution across the farmers' population represented by the considered sample.

The considered ERS-MEMC model is presented in three steps. First, we present the production choice models defined at the crop level, *i.e.* the yield supply and variable input demand models. Second, we present the per regime acreage share choice models. Finally, we describe the production regime choice model. This presentation is organized following the structure of the model: yield supply and variable input demand models are used for defining expected crop return models. These models are then used for defining crop acreage share models, which are themselves used for defining the production regime choice model.

2.3.1. Yield supply and variable input demand models

We assume that farmers produce crop k from a variable input aggregate under a quadratic technological constraint. *I.e.*, we assume that the yield of crop k obtained by farmer i in year t is given by:

(6) $$y_{k,it} = b^{y}_{k,it} - 1/2 \times (a^{x}_{k,i})^{-1}(b^{x}_{k,it} - x_{k,it})^{2}$$

where $x_{k,it}$ denotes the variable input use level. Parameter $a^{x}_{k,i}$ is required to be (strictly) positive for the production function to be (strictly) concave in $x_{k,it}$. It determines the extent to which the yield supply and the input demand of crop k respond to the input and crop prices.

Terms $b^{y}_{k,it}$ and $b^{x}_{k,it}$ have direct interpretations in the considered yield function. Term $b^{y}_{k,it}$ is the yield level that can be potentially achieved by farmer i in year t while $b^{x}_{k,it}$ is the input quantity required to achieve this potential yield level. These parameters are decomposed as $b^{y}_{k,it} = b^{y}_{k,i} + (\boldsymbol{\delta}^{y}_{k,0})'\mathbf{c}^{y}_{k,it} + e^{y}_{k,it}$ and $b^{x}_{k,it} = b^{x}_{k,i} + (\boldsymbol{\delta}^{x}_{k,0})'\mathbf{c}^{x}_{k,it} + e^{x}_{k,it}$ where terms $\mathbf{c}^{y}_{k,it}$ and $\mathbf{c}^{x}_{k,it}$ are observed variable vectors used to control for observed farm heterogeneity (*i.e.*, farm size and capital endowment per unit of land) and climatic conditions (*i.e.*, temperature and rainfall). The $b^{y}_{k,it}$ and $b^{x}_{k,it}$ terms are farmer specific parameters aimed at capturing unobserved

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

heterogeneity across farms and farmers. These terms, as well as the $a^x_{k,i}$ random parameter, mainly capture three kinds of effects: those of the natural and material factor endowment of farms (*e.g.*, soil quality, machinery quality), of farmers' practice choices (*e.g.*, crop management practices, cropping systems) and of the skills of farmers. Terms $e^y_{k,it}$ and $e^x_{k,it}$ are standard error terms aimed to capture the effects on production of stochastic events (*e.g.*, climatic conditions, and pest and weed problems). We assume that farmer *i* is aware of the content of $e^x_{k,it}$ when deciding his variable input uses.

Assuming that farmer *i* maximizes the expected return to variable input uses of each crop, we can easily derive the demand of the variable input for crop *k*:

(7a) $$y_{k,it} = b^y_{k,i} + (\boldsymbol{\delta}^y_{k,0})'\mathbf{c}^y_{k,it} - 1/2 \times a^x_{k,i} w^2_{k,it} p^{-2}_{k,it} + e^y_{k,it}$$

and the corresponding yield supply:

(7b) $$x_{k,it} = b^x_{k,i} + (\boldsymbol{\delta}^x_{k,0})'\mathbf{c}^x_{k,it} - a^x_{k,i} w_{k,it} p^{-1}_{k,it} + e^x_{k,it} \ .$$

Terms $p_{k,it}$ and $w_{k,it}$ respectively denote the expected output and input prices of crop *k*. Assuming that the expectations of $e^y_{k,it}$ and $e^x_{k,it}$ of farmer *i* are null at the beginning of the cropping season,[7] this farmer expects the following return to the variable input:

(8) $$\pi_{k,it} = p_{k,it}\left(b^y_{k,i} + (\boldsymbol{\delta}^y_{k,0})'\mathbf{c}^{ys}_{k,it}\right) - w_{k,it}\left(b^x_{k,i} + (\boldsymbol{\delta}^x_{k,0})'\mathbf{c}^{xs}_{k,it}\right) + 1/2 \times a^x_{k,i} w^2_{k,it} p^{-1}_{k,it}$$

for crop *k* when she/he chooses her/his acreage shares. Vector $(\mathbf{c}^{ys}_{k,it}, \mathbf{c}^{xs}_{k,it})$ is defined by replacing in vector $(\mathbf{c}^{y}_{k,it}, \mathbf{c}^{x}_{k,it})$ the climatic variables by their expectations.

2.3.2. Acreage share choice models

As discussed in Carpentier and Letort (2014), the Standard MNL crop acreage model given in equation (5) appears to be rather rigid because it treats the different crops symmetrically. Indeed, arable crops can often be grouped according to their competing for the use of quasi-fixed factors or according to their agronomic characteristics. The ERS-MEMC model considered in our application presented in the next section is based on a 3 level Nested Multinomial Logit (NMNL) acreage share model.

For sake of simplification, we consider a 2 level NMNL acreage share model in this section.[8] Crop set K is partitioned into *G* mutually exclusive groups of crops. Term $\mathcal{G} = \{1,\ldots,G\}$ defines the considered group set. Group $g \in \mathcal{G}$ defines the crop subset $\mathcal{K}(g)$. Crops belonging to a same group are assumed to share similar agronomic characteristics and to compete more for farmers' limiting quantities of quasi-fixed factors than they compete with crops of other groups. The corresponding acreage management cost function is given by:

---

[7] As discussed below, this assumption can be relaxed, *e.g*. for accounting for potential correlations between the $e^y_{k,it}$ and $e^x_{k,it}$ error terms on the one hand, and the $e^s_{k,it}$ error terms on the other hand.

[8] The model used in our application is presented in the Online Appendix.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(9) $$C_{it}(\mathbf{s}) = \sum_{k\in\mathcal{K}} s_k \beta^s_{k,it} + \sum_{g=1}^{G} (\alpha^s_i)^{-1} s_{(g)} \ln s_{(g)} + \sum_{g\in\mathcal{G}} s_{(g)} (\alpha^s_{(g),i})^{-1} \sum_{m\in\mathcal{K}(g)} s_{m|(g)} \ln s_{m|(g)}$$

where $s_{(g)}$ denotes the acreage share of group g and $s_{m,(g)}$ that of crop m in group g. Terms α^s_i and $\alpha^s_{(g),i}$ are farm specific parameters determining the flexibility of farmers' acreage choices.[9] The larger they are, the more the acreage share choice respond to economic incentives (because the less management costs matter). Condition $\alpha^s_{(g),i} \geq \alpha^s_i > 0$ is sufficient for cost function $C_{it}(\mathbf{s})$ to be strictly convex in $\mathbf{s}$.

The linear terms of the cost function $C_{it}(\mathbf{s})$ are decomposed as $\beta^s_{k,it} = \beta^s_{k,i} + (\boldsymbol{\delta}^s_{k,0})'\mathbf{z}^s_{k,it} + \varepsilon^s_{k,it}$ where $\mathbf{z}^s_{k,it}$ are explanatory variable vectors used to control for observed heterogeneous factors and climatic events. Farm specific parameters $\beta^s_{k,i}$ account for unobserved heterogeneity effects. Error terms $\varepsilon^s_{k,it}$ capture the effects of stochastic variations of the cost due to random events such as unobserved interactions of climatic events and soil characteristics impacting the soil state at planting. Farmers are assumed to know these terms when choosing their acreages. Error terms $\varepsilon^s_{k,it}$ are assumed to be independent from the error terms of the yield supply and input demand equations, $\varepsilon^y_{k,it}$ and $\varepsilon^x_{k,it}$.

Farmers' optimal crop acreage choices as given by equation (2a) can be derived for any production regime. It suffices to solve the maximization problem given in equations (3). For instance, eight acreage share subsystems are considered in our empirical application, one for each production regime present in the data. Of course, the functional form of the derived acreage choice function depends on the subset of crops produced in the considered regime. Assuming that crop k belongs to group g, we obtain:

(10) $$s_{k,it}(r) = \frac{\varphi_k(r)\exp\left(\alpha^s_{(g),i}(p_{k,it} - \beta^s_{k,it})\right)\left(\sum_{\ell\in\mathcal{K}(g)} \varphi_\ell(r)\exp\left(\alpha^s_{(g),i}(p_{\ell,it} - \beta^s_{\ell,it})\right)\right)^{\alpha^s_i(\alpha^s_{(g),i})^{-1} - 1}}{\sum_{h\in\mathcal{G}}\left(\sum_{\ell\in\mathcal{K}(h)} \varphi_\ell(r)\exp\left(\alpha^s_{(h),i}(p_{\ell,it} - \beta^s_{\ell,it})\right)\right)^{\alpha^s_i(\alpha^s_{(h),i})^{-1}}}$$

and:

(11) $$\Pi_{it}(r) = (\alpha^s_i)^{-1} \ln \sum_{h\in\mathcal{G}}\left(\sum_{\ell\in\mathcal{K}(h)} \varphi_\ell(r)\exp\left(\alpha^s_{(h),i}(p_{\ell,it} - \beta^s_{\ell,it})\right)\right)^{\alpha^s_i(\alpha^s_{(h),i})^{-1}}.$$

Parameter α^s_i drives the land allocation to crop group acreages while parameters $\alpha^s_{(g),i}$ drive the allocation of the crop group acreages to crop acreages.

Production regime choice model

Observing that the regime specific optimal acreage choice $\mathbf{s}_{it}(r)$ necessarily belongs to regime r in the MNL case considered here, the regime specific expected profit levels $\Pi_{it}(r)$ can be used for defining a regime choice model based to the choice problem described in equation (4). Let define the regime fixed costs as $D_{it}(r) = d_i(r) - \sigma_i^{-1}\varepsilon_{r,it}$. The farm specific parameters

[9] We have we have $\alpha^s_{(g),i} = \alpha^s_i$ if group g contains a single crop.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$d_i(r)$ aim to capture the effects of unobserved factors affecting the regime fixed costs. The error terms $e_{r,it}$ aim to capture the effects of stochastic factors and define the regime choice model as a probabilistic discrete choice model, with:

(12) $$r_{it} = \arg\max_{r \in \mathcal{R}} \{\Pi_{it}(r) - d_i(r) + \sigma_i^{-1} e_{r,it}\}.$$

Scale parameter σ_i determines the extent to which the regime expected profit levels (*i.e*. the $\Pi_{it}(r) - d_i(r)$ terms) explain the production regime choice as regards to the effects of the $e_{r,it}$ idiosyncratic terms. The higher σ_i, the more the expected profit levels impact the observed regime choices.

Regime fixed costs $d_i(r)$ can be specified in different ways. These costs are expected to increase with the number of crops. Transaction costs and labor requirements related to a production regime increase with the number of crops produced in that regime. Indeed, one way to specify $d_i(r)$ is to consider a sum of fixed costs associated to each crop produced in the considered production regime, with $d_i(r) = \sum_{k \in \mathcal{K}^+(r)} b_{k,i}^c$ where $b_{k,i}^c$ is the fixed costs related to crop *k*. Interestingly, this specification allows computing the fixed costs of regimes which are not observed in the data. This is of particular interest for simulation purposes. For example, changes in market conditions can lead farms to adopt new production regimes. This regime fixed cost specification is used in our empirical application.

This specification of the regime fixed costs can be usefully compared with more general ones. Farmers may purchase inputs specific to different crops from the same supplier, implying savings in the related transaction costs. Moreover, different crops may generate work peak loads during the same periods, implying that can concentrate their workload (or that of their employees) during these periods if they wish so. In these cases, the regime fixed costs are sub-additive in the crop fixed costs. One way to deal with this pattern consists of directly specifying these fixed costs as famers specific constant terms on a regime per regime basis, with $d_i(r) = d_{r,i}$ (given that the fixed cost of a "benchmark regime" needs to be normalized). Of course, the costs corresponding to regimes that are not observed in the data can't be recovered, thereby constraining the regime set that can be simulated to be equal to the one that is observed in the data.

2.3.3. Overall structure of the ERS-MEMC model

The ERS-MEMC model is composed of three main parts: a subsystem of yield supply and input demand equations (7), a set of per regime subsystems of acreage share equations (10) and a probabilistic regime choice model (12).

The set of dependent variables of this model contains the crop level production choices. These consist of the yield levels, input use levels and acreage shares of each crop that are produced by for farmer *i* in year *t*. These are collected in vector $(\mathbf{y}_{it}^+, \mathbf{x}_{it}^+, \mathbf{s}_{it}^+)$. Production regime r_{it} is the last dependent variable of the model.

The set of explanatory variables contains crop prices, variable input prices and the control variable vectors used in the crop yield supply, input demand and acreage share equations for

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

all crops. These variable are collected in vector $\mathbf{z}_{it}$, which defines the information set of the ERS-MEMC model.

The sole fixed parameters appearing in the model equations are the coefficients of the control variable coefficient vectors for all crops, which are collected in vector $\boldsymbol{\delta}_0$.

The considered ERS-MEMC model contains two main subsets of random components: a vector of random parameters and a vector of error terms.

Vector $\boldsymbol{\gamma}_i$ collects the farm specific parameters of the model, with $\boldsymbol{\gamma}_i = (\boldsymbol{\beta}_i, \boldsymbol{\alpha}_i, s_i)$. This vector contains the potential yield parameters, the input requirement parameters, the cost function linear parameters and the crop fixed costs parameters for all crops. These random parameters are collected in vector $\boldsymbol{\beta}_i$. It also contains the input use flexibility parameters for all crops and the acreage choice flexibility parameters, which are collected in vector $\boldsymbol{\alpha}_i$. Finally, $\boldsymbol{\gamma}_i$ contains the scale parameter, s_i, of the regime choice model.

In the error term vector $\boldsymbol{\varepsilon}_{it} = (\boldsymbol{\varepsilon}_{it}^{yx}, \boldsymbol{\varepsilon}_{it}^{s})$, sub-vector $\boldsymbol{\varepsilon}_{it}^{yx}$ collects the error terms error terms of crop yield supply and input demand equations for all crops while sub-vector $\boldsymbol{\varepsilon}_{it}^{s}$ collects those of the acreage share equations. Finally, vector $\mathbf{e}_{it}$ collects the error terms of the regime choice model (*i.e.*, $e_{r,it}$ for $r \in \mathcal{R}$).

2.4. ERS-MEMC model with regime specific fixed costs: estimation strategy

This section presents the main features of the estimation strategy adopted for estimating the ERS-MEMC model described above. As this model involve multiple endogenous regimes, considers numerous interrelated production choices and features random parameters, we impose parametric distributional assumptions on its random components (*i.e.* error terms and random parameters) that ensure its empirical tractability. We also impose simplifying assumptions regarding the dynamics of farmers' choices and the multi-crop production technology. These assumptions are presented and discussed first. Then, we present how the main parameters of interest of our ERS-MEMC model are recovered from the data. Finally, we briefly describe our estimation strategy. More specifically, we present the main estimation issues that we face when estimating our random parameter ERS-MEMC model and the approaches chosen for overcoming these issues. A detailed description of our estimation procedure is provided in Appendix 2A. This procedure combines techniques found in the micro-econometrics and computational statistics literatures.

2.4.1. Main probabilistic assumptions

We assume that terms $(\boldsymbol{\varepsilon}_{is}, \mathbf{e}_{is})$, $\boldsymbol{\gamma}_i$ and $\mathbf{z}_{it}$ are independently distributed for any pair (t,s). This implies that the explanatory variables vector, $\mathbf{z}_{it}$, is assumed to be (*i*) strictly exogenous with respect to the error term vectors and (*ii*) independent of the random parameters $\boldsymbol{\gamma}_i$. This latter assumption, which is standard in random parameter models, defines $\boldsymbol{\gamma}_i$ as a term that captures heterogeneity effects not captured by control variables $\mathbf{z}_{it}$.

We further assume that error term $(\boldsymbol{\varepsilon}_{it}, \mathbf{e}_{it})$ vectors are independently distributed across

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

time. Combined with the fact that vector $\mathbf{z}_{it}$ doesn't contain any lagged endogenous variable, this serial independence assumption implies that our MEMC model can be interpreted as a reduced form model as regards the dynamic features of the modelled choices. Indeed, we hypothesize that random parameters $\boldsymbol{\gamma}_i$ capture the effects on farmers' production choices and performances of the stable crop rotation schemes that these farmers rely on.[10] Koutchadé *et al* (2018) provide empirical results confirming this hypothesis with a sample of arable crop producers located in an area contiguous to the one considered in our application.

Finally, we assume that the error term vectors $\boldsymbol{\varepsilon}_{it}^{yx}$, $\boldsymbol{\varepsilon}_{it}^{s}$ and $\mathbf{e}_{it}$ are independent. Relaxing this independence assumption for $\boldsymbol{\varepsilon}_{it}^{s}$ and $\boldsymbol{\varepsilon}_{it}^{yx}$ is possible but significantly increases the estimation burden and Koutchadé *et al* (2018), in a similar context, found that error terms $\boldsymbol{\varepsilon}_{it}^{s}$ and $\boldsymbol{\varepsilon}_{it}^{yx}$ were not significantly correlated.

2.4.2. Distributional assumptions

Random parameter vectors $\boldsymbol{\gamma}_i$ are assumed independent across farms. For sake of simplification, we assume here that these random parameter vectors are normally distributed, with $\boldsymbol{\gamma}_i : \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Omega}_0)$. Various transformations of elements of $\boldsymbol{\gamma}_i$ actually allow for other distribution choices for these elements while keeping the multivariate structure of the probability distribution of $\boldsymbol{\gamma}_i$ (*e.g.*, Stanfield *et al*, 1996). For example, considering log-transformations of $\boldsymbol{\alpha}_i$ and s_i in $\boldsymbol{\gamma}_i$ implies that these random parameters, which are required to be positive, are jointly log-normality distributed. We used this log-transformation in the ESR-MEMC model used for our empirical application. Robustness checks demonstrated that other probability distribution choices have a limited impact on the main results.[11]

We make the usual assumptions stating that error term vectors $\boldsymbol{\varepsilon}_{it}$ are independent across farms (and years) and normally distributed, with $\boldsymbol{\varepsilon}_{it} : \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_0)$.[12]

Finally, we assume that the regime choice model error terms $e_{r,it}$ are independent across regimes and distributed according to a type I extreme value distribution. This assumption implies that the considered regime choice model is a standard Multinomial Logit discrete choice model conditionally on the scale parameter and on the regime specific expected profit levels and fixed costs. The corresponding conditional probability of the observed the regime choices is given by:

[10] In that, we rely on well-known features of heterogeneous dynamic processes: those implying that empirically disentangling the effects of unobserved heterogeneity from those of unobserved persistent dynamic features is notably difficult. Accounting for dynamic features of multi-crop production technologies and of farmers' choices is challenging, and largely beyond the scope of this article.

[11] We tested specifications assuming that $\boldsymbol{\beta}_i$ is log-normally distributed and/or that $\boldsymbol{\alpha}_i$ follows a bounded Johnson distribution (*e.g.*, Stanfield *et al*, 1996).

[12] Matrix $\boldsymbol{\Psi}_0$ is block-diagonal under the assumption stating that $\boldsymbol{\varepsilon}_{it}^{s}$ and $\boldsymbol{\varepsilon}_{it}^{yx}$ are independent.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(13) $$P(r_{it} \mid \boldsymbol{\varepsilon}_{it}^{s}, \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\delta}_0) = \frac{\exp(\sigma_i(\Pi_{it}(r_{it}) - d_i(r_{it})))}{\sum_{r \in \mathcal{R}} \exp(\sigma_i(\Pi_{it}(r) - d_i(r)))}.$$

This probability is defined as a function of $(\boldsymbol{\varepsilon}_{it}^{s}, \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\delta}_0)$ because the vector of regime specific expected profit levels, $\Pi_{it}(r) - d_i(r)$ for $r \in \mathcal{R}$, is a function of all the terms contained in $(\boldsymbol{\varepsilon}_{it}^{s}, \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\delta}_0)$, scale parameter σ_i excepted.

2.4.3. Identification

We consider here identification of the probability distribution of main random parameters of interest: the production choice flexibility parameters and the parameters of the regime choice model.

Under the considered assumptions the probability distribution of farmers' responses to economic incentives, $\boldsymbol{\alpha}_i$, are identified through two main channels. Identification of the probability distribution of the variable input use flexibility parameters, $\alpha_{k,i}^{x}$ for $k \in \mathcal{K}$, mostly relies on the variations of the corresponding input to crop price ratios. Identification of the probability distribution of the acreage choice flexibility parameters, α_i^{s} and $\alpha_{(g),i}^{s}$ for $g \in \mathcal{G}$, mainly relies on the variations of the expected crop return terms, $\pi_{k,it}$ for $k \in \mathcal{K}$. Importantly, the expected crop returns are defined as functions of random parameters (*i.e.*, $\beta_{k,i}^{y}$, $\beta_{k,i}^{x}$ and $\alpha_{k,i}^{x}$ for $k \in \mathcal{K}$) that may be correlated with the acreage choice flexibility parameters. The "full" variance-covariance matrix of the joint probability distribution of the random parameters $\boldsymbol{\gamma}_i$ takes into account these potential correlations.

Scale parameter σ_i, which is the random coefficient associated to the regime specific expected profit levels $\Pi_{it}(r)$ in the regime choice model, is mainly identified by the variations in these variables. Crop fixed costs $\beta_{k,i}^{c}$ are entailed in the regime fixed costs $d_i(r) = \sum_{k \in \mathcal{K}^{+}(r)} \beta_{k,i}^{c}$. Importantly, the fixed costs of the crops that are always produced cannot be identified because these crops are part of any regime present in the data. Therefore, the fixed costs of these crops are normalized at zero. The joint probability distribution of the identifiable crop fixed cost vector is mainly identified by the variations in the differences in the regime specific expected profit levels $\Pi_{it}(r)$ across the production regimes. The potential correlations between, on the one hand, the random parameters that are part of the expected profit levels and, on the other hand, the crop fixed costs and the scale parameter are taken into account in the distribution of $\boldsymbol{\gamma}_i$.

2.4.4. Estimation issues and sketch of the estimation procedure

The considered ERS-MCEM model being fully parametric, we consider a Maximum Likelihood (ML) estimator for efficiently estimating its parameters. These parameters are collected in $\boldsymbol{\theta}_0 = (\boldsymbol{\delta}_0, \boldsymbol{\Psi}_0, \boldsymbol{\mu}_0, \boldsymbol{\Omega}_0)$. Contribution of farmer *i* to the likelihood function of the model corresponds to the probability density function (pdf) of her/his sequence of production choices conditional on the sequence of exogenous variables characterizing this choice

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

sequence. Assuming that the considered pdf is parameterized by $\boldsymbol{\eta}$, let function $f(\mathbf{u}|\mathbf{v};\boldsymbol{\eta})$ generically denotes the pdf of $\mathbf{u}_{it}$ conditional on $\mathbf{v}_{it}=\mathbf{v}$ at $\mathbf{u}_{it}=\mathbf{u}$. And let function $\varphi(\mathbf{u};\boldsymbol{\Omega})$ denote the pdf of $\mathcal{N}(\mathbf{0},\boldsymbol{\Omega})$ at $\mathbf{u}$. Given the probabilistic assumptions defining the parametric version of the random parameter ERS-MEMC model, contribution of farmer *i* to the likelihood function at $\boldsymbol{\theta}$ is given by:

(14) $$\ell_i(\boldsymbol{\theta})=\int\left(\prod_{t=1}^{T} f(\mathbf{y}_{it}^{+},\mathbf{x}_{it}^{+},\mathbf{s}_{it}^{+},r_{it}\,|\,\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})\right)\varphi(\boldsymbol{\gamma}-\boldsymbol{\mu};\boldsymbol{\Omega})\,d\boldsymbol{\gamma}.$$

Likelihood function $\ell_i(\boldsymbol{\theta})$ can be obtained neither analytically nor numerically due to its integration over the probability distribution of the random parameters $\boldsymbol{\gamma}_i$.

Micro-econometricians generally solve this problem by integrating $\ell_i(\boldsymbol{\theta})$ *via* direct simulation methods for computing Simulated ML (SML) estimators of $\boldsymbol{\theta}_0$. Yet, implementing this approach is particularly challenging with ERS-MEMC models due to the dimension of parameter $\boldsymbol{\theta}_0$ and the complexity of the simulated version of the likelihood functions $\ell_i(\boldsymbol{\theta})$. For instance, the ERS-MEMC model of our empirical application considers 22 production choices. It features 80 control variables, 37 random parameters and 20 error terms. Vector $\boldsymbol{\theta}_0$ contains 786 parameters while our dataset describes 40,192 observed production choices (16.5 per observation on average).

Integration of $\ell_i(\boldsymbol{\theta})$ over the random parameter distribution is thus the first estimation issue that we have to deal with. We compute the ML estimator of $\boldsymbol{\theta}_0$ by devising a Stochastic Approximate Expectation-Maximization (SAEM) algorithm. SAEM algorithms were proposed by Delyon et *al* (1999) for computing ML estimators of models featuring continuous random parameters. These algorithms rely on simulation methods for integrating proxies of the sample log-likelihood of the considered model. They appear to use simulations more efficiently than competing alternatives (*e.g.*, McLachlan and Krishnan, 2007; Lavielle, 2014), which is a particularly relevant property when considering large samples, large multivariate models and/or large random parameter vectors. The structure of the SAEM algorithm that we propose for estimating random parameter ERS-MEMC models is described in ppendix 2A. Here, we consider its main step, the Maximization (M) step. This allows us to demonstrate the main advantages of our approach and, in the sequel, to illustrate the other two main estimation issues that we face.

At each of its iteration, the considered SAEM algorithm solves two maximization problems for updating estimates of $\boldsymbol{\theta}_0$. These problems have the form of weighted ML problems that are much simpler to solve than the corresponding SML problem. The first problem to be solved in the M step of our SAEM algorithm aims at updating the estimated value of $(\boldsymbol{\mu}_0,\boldsymbol{\Omega}_0)$, the parameter of the pdf of the model random parameters. It is of the form:

(15) $$\max_{(\boldsymbol{\mu},\boldsymbol{\Omega})}\sum_{i=1}^{N}\sum_{j=1}^{J}\tilde{\pi}_i^j\ln\varphi(\tilde{\boldsymbol{\gamma}}_i^j-\boldsymbol{\mu};\boldsymbol{\Omega})$$

where terms $\tilde{\boldsymbol{\gamma}}_i^j$ are random draws of $\boldsymbol{\gamma}_i$ from a pdf defined by the preceding iteration results and $\tilde{\pi}_i^j$ are weighting terms attached to these draws. The solution in $\boldsymbol{\mu}$ is the empirical weighted mean of the random draws $\tilde{\boldsymbol{\gamma}}_i^j$ while the solution in $\boldsymbol{\Omega}$ is their empirical weighted

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

variance-covariance matrix.

The second part of the M step of our SAEM algorithm updates the estimate of $(\boldsymbol{\delta}_0,\boldsymbol{\Psi}_0)$. It considers functions of the form $\tilde{W}(\boldsymbol{\delta},\boldsymbol{\Psi}) = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{J}\tilde{\pi}_i^j \ln f(\mathbf{w}_{it}^+, r_{it} \mid \mathbf{z}_{it}, \tilde{\boldsymbol{\gamma}}_i^j;\boldsymbol{\delta},\boldsymbol{\Psi})$ where $\mathbf{w}_{it}^+ = (\mathbf{y}_{it}^+,\mathbf{x}_{it}^+,\mathbf{s}_{it}^+)$. These functions have the functional forms of a weighted log-likelihood function of the production choices of a sample of "simulated farmers". Assuming that $(\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\Psi}})$ is the preceding iteration estimate of $(\boldsymbol{\delta}_0,\boldsymbol{\Psi}_0)$, it consists of solving either of the two following problems (*a*) $\max_{(\boldsymbol{\delta},\boldsymbol{\Psi})} \tilde{W}(\boldsymbol{\delta},\boldsymbol{\Psi})$ or (*b*) find $(\boldsymbol{\delta},\boldsymbol{\Psi})$ such that $\tilde{W}(\boldsymbol{\delta},\boldsymbol{\Psi}) \geq \tilde{W}(\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\Psi}})$. Unfortunately, solving problem (*a*) or even simpler search problem (*b*) is difficult due to the complexity of the conditional likelihood function $f(\mathbf{w}_{it}^+, r_{it} \mid \mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$.

Decomposing this function demonstrate that the problem is indeed twofold. Using Bayes's law and the structure and distributional assumptions of the ERS-MCEM model, we obtain:

(16) $f(\mathbf{w}_{it}^+, r_{it} \mid \mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi}) = P(r_{it} \mid \mathbf{s}_{it}^+,\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi}) f(\mathbf{w}_{it}^+ \mid \mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$,

Vector $\mathbf{s}_{it}^+$ collects the acreage shares of the crops produced in regime r_{it}.[13] Function $f(\mathbf{w}_{it}^+ \mid \mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$ is the likelihood of crop level choice vector $\mathbf{w}_{it}^+$ conditional on $(\mathbf{z}_{it},\boldsymbol{\gamma}_i = \boldsymbol{\gamma})$ and $P(r_{it} \mid \mathbf{s}_{it}^+,\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$ is the probability function of regime r_{it} conditional on $(\mathbf{s}_{it}^+,\mathbf{z}_{it},\boldsymbol{\gamma}_i = \boldsymbol{\gamma})$. Yet, both functions raise estimation issues.

Given the structure of our MEMC model, function $\ln f(\mathbf{w}_{it}^+ \mid \mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$ is the likelihood function of a Gaussian Seemingly Unrelated Regression (SUR) system with observations missing at random (up an additive term that doesn't depend on $(\boldsymbol{\delta},\boldsymbol{\Psi})$). The missing observations are the yield level, input use and acreage share of the crops that are not produced in regime r_{it}. Ruud (1991) discussed the use of Expectation-Maximization (EM) algorithms for alleviating the computation burden of ML estimators of models based on latent Gaussian SUR systems with missing observations. Based on Ruud's insights we devised an EM type approach for updating the estimates of $(\boldsymbol{\delta}_0,\boldsymbol{\Psi}_0)$ in the M step of our SAEM algorithm.

Our last main estimation issue is due to the computation of the regime choice probability function $P(r_{it} \mid \mathbf{s}_{it}^+,\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$. Given the structure of our MEMC model, this probability function can be defined as a function of the error terms of the acreage share equations. Let vector $\boldsymbol{\varepsilon}_{it}^{s,+}$ collect the error terms of the acreage share models of the crops produced in regime r_{it} and vector $\boldsymbol{\varepsilon}_{it}^{s,0}$ collect those of the crops that are not produced. Vector $\boldsymbol{\varepsilon}_{it}^{s,+}$ can be recovered from the acreage share model and the data, the observed crop acreages of the produced crops in particular. Let function $\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}(\boldsymbol{\gamma},\boldsymbol{\delta})$ denote the residual function corresponding to error term

[13] Yield supply and input demand levels, $(\mathbf{y}_{it}^+,\mathbf{x}_{it}^+)$, and regime choices, r_{it}, are independent conditionally on acreage choices, exogenous variables and random parameters, $(\mathbf{s}_{it}^+,\mathbf{z}_{it},\boldsymbol{\gamma}_i = \boldsymbol{\gamma})$, since error terms $(\boldsymbol{\varepsilon}_{it}^y,\boldsymbol{\varepsilon}_{it}^x)$, $\boldsymbol{\varepsilon}_{it}^s$ and $\mathbf{e}_{it}$ are assumed to be mutually independent.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$\boldsymbol{\varepsilon}_{it}^{s,+}$. The structure of our MEMC model and equation (13) yield:

$$(17)\qquad P(r_{it}\,|\,\mathbf{s}_{it}^{+},\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi}) = \int P(r_{it}\,|\,\boldsymbol{\gamma},\mathbf{z}_{it},\boldsymbol{\varepsilon}_{it}^{s};\boldsymbol{\delta})f\left(\boldsymbol{\varepsilon}_{it}^{s,0}\,|\,\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}(\boldsymbol{\gamma},\boldsymbol{\delta});\boldsymbol{\delta},\boldsymbol{\Psi}\right)d\boldsymbol{\varepsilon}_{it}^{s,0}$$

where $f(\boldsymbol{\varepsilon}_{it}^{s,0}\,|\,\hat{\boldsymbol{\varepsilon}}^{s,+};\boldsymbol{\delta}_0,\boldsymbol{\Psi}_0)$ denotes the pdf of $\boldsymbol{\varepsilon}_{it}^{s,0}$ conditional on $\boldsymbol{\varepsilon}_{it}^{s,+} = \hat{\boldsymbol{\varepsilon}}^{s,+}$, which is normal. Vector $\boldsymbol{\varepsilon}_{it}^{s,0}$ must be considered as missing variables in the estimation process because it cannot be recovered by combining the model and the data. The Multinomial Logit functional form of function $P(r_{it}\,|\,\boldsymbol{\gamma},\mathbf{z}_{it},\boldsymbol{\varepsilon}_{it}^{s};\boldsymbol{\delta})$ prevents its integration over the probability distribution of $\boldsymbol{\varepsilon}_{it}^{s,0}$, either analytically or numerically. Building on the work of Harding and Hausman (2007), we use Laplace approximates of the regime choice probability functions $P(r_{it}\,|\,\mathbf{s}_{it}^{+},\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$ for computing the likelihood function of our model.[14]

The fact that production regime choices and acreage choices depend on $\boldsymbol{\varepsilon}_{it}^{s}$ constitutes the first source of endogeneity of the regime choices in our ERS-MEMC model.[15] Random parameter $\boldsymbol{\gamma}_i$ constitutes a supplementary source of regime choice endogeneity in our ERS-MEMC model.

2.5. Empirical application: crop diversification of French arable crop producers

This section presents an application aimed to illustrate the empirical tractability of our modelling approach as well as to demonstrate the role of crop set choices in analyzes of farmers' production choices.

2.5.1. Data and model specification details

The model is estimated on an unbalanced panel data set containing 2276 observations of 415 French grain crop producers in the North and North-East of France, over the years 2006 to 2011. This sample has been extracted from data provided by an accounting agency located in the French territorial division *La Marne*. It contains detailed information about crop production for each farm (acreages, yields, input uses and crop prices at the farm gate). We consider seven crops: sugar beet, alfalfa, protein pea, rapeseed, winter wheat, corn and spring barley, which represent more than 80% of the total acreage in the considered area.[16]

The variable input aggregate accounts for the use of fertilizers, pesticides and seeds. The corresponding price index is computed as a standard Tornqvist index. When a farmer doesn't

[14] This approach relies on a second order Taylor expansion in $\boldsymbol{\varepsilon}_{it}^{s,0}$ of function $P(r_{it}\,|\,\boldsymbol{\gamma},\mathbf{z}_{it},\boldsymbol{\varepsilon}_{it}^{s};\boldsymbol{\delta})f\left(\boldsymbol{\varepsilon}_{it}^{s,0}\,|\,\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}(\boldsymbol{\gamma},\boldsymbol{\delta});\boldsymbol{\delta},\boldsymbol{\Psi}\right)$ around an optimally chosen value of $\boldsymbol{\varepsilon}_{it}^{s,0}$. Using simulation methods for integrating function $P(r_{it}\,|\,\mathbf{s}_{it}^{+},\mathbf{z}_{it},\boldsymbol{\gamma};\boldsymbol{\delta},\boldsymbol{\Psi})$ would be inconvenient in our case due to our using such methods for dealing with random parameters.

[15] Indeed, the endogeneity issues raised by $\boldsymbol{\varepsilon}_{it}^{s}$ in our ERS-MEMC model are analogous, from an econometric viewpoint, to those raised by the demand function error terms in demand systems with binding non-negativity constraints (*e.g.*, Wales and Woodland, 1983; Lee and Pitt; 1986).

[16] The EU sugar beet subsidy scheme requires limited adjustments in our application because the actual sugar beet production largely exceeds the subsidized quota for all sugar beet producers of our sample.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

produce a crop the corresponding output and input prices are unobserved. These missing prices were approximated by the yearly average of the corresponding observed prices. All prices are deflated by the hired production services price index (base 1 in 2006) obtained from the French department of Agriculture. This aggregated price index mainly depends on the price indices of machinery, fuel and hired labor, the main inputs involved in the implicit acreage management cost function. Climatic variables are provided at the municipality level by Météo France, the French national meteorological service.

Farmers' crop price expectations are defined by the corresponding lagged prices, according to a naïve anticipation scheme. Robustness checks demonstrate that anticipation scheme choices mostly impact estimates of the probability distribution of input use flexibility parameters $a_{k,i}^{x}$, with very limited effects on our main results.

Figure 2 depicts the three levels nesting structure that we adopt for the seven crops. In a first level we distinguish a cereal group composed of wheat, corn and barley, and a group of rotation entry crops: sugar beet, alfalfa, peas and rapeseed. This structure is intended to reflect the basic rotation scheme of grain and industrial crop producers in France. In a second level, the cereal group is split into two subgroups: winter cereals on the one hand and spring cereals on the other hand, in order to account for the differences in planting seasons. The 'rotation entry crop' group is split into an 'oilseeds and protein crops' subgroup and a subgroup including only sugar beet (the only root crop considered here). Wheat, which is the only winter cereal, is used as the benchmark crop. Based on these seven crops, 127 regimes could theoretically be chosen by farmers. The 8 most frequently observed regimes, out of 78 regimes present in the original dataset, were considered for selecting our estimation sample.[17]

[17] Considering a small regime set allowed us to estimate our ERS-MEMC model with regime specific fixed costs that are not defined as sums of crop fixed costs. This specification of the regime fixed costs is more flexible but only yields a modest improvement in the fit performance of the regime choice model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

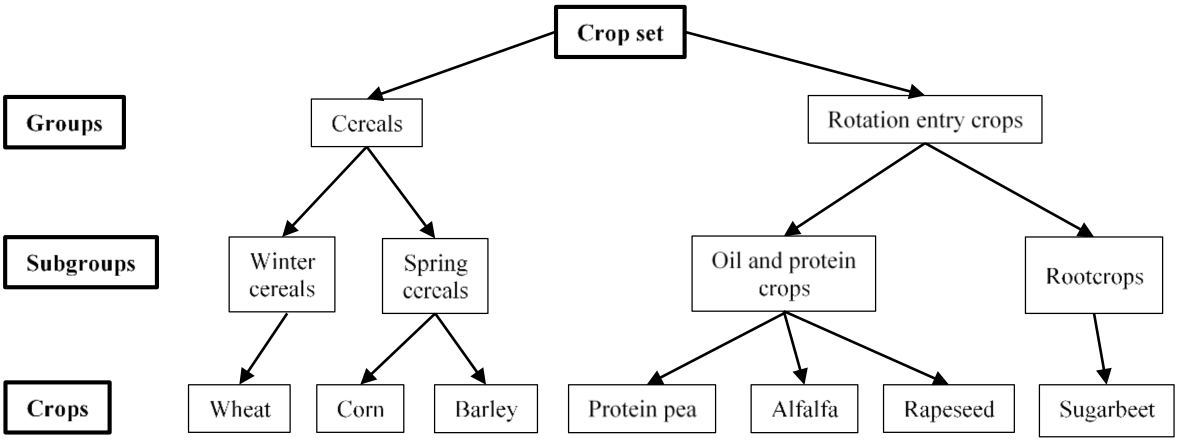

Figure 2. Nesting structure of the acreage choice model

All farmers grow winter cereals, (spring) barley, (winter) rapeseed and most of them (91.7%) grow at least two additional crops. The most frequent regimes in the sample (regimes 2, 3 and 4) include five or six crops. Table 1 provides descriptive statistics concerning the production regimes observed in the data. Most farmers adopt different production regimes over the 6 years of our sample: only

8 out of 415 farmers have not changed their production regime. The average gross margins associated to each regime are reported in the last column of Table 1. An interesting feature appears here: the most frequently chosen regimes are not the ones that lead to the highest average gross margin per hectare. For instance, regime 2 – which excludes corn – is characterized by the highest observed gross margin on average, but has been adopted in only 21.5% of the observations. This comes to illustrate the fact that farmers' choices of production regime are driven by factors other than gross returns, such as the acreage management and regime fixed costs represented in our model.

Because we assume that regime costs are equal to a sum of fixed costs associated to each crop produced in the considered regime, the fixed costs associated to winter cereals, spring barley and rapeseed, which are always produced in our sample, are set to zero for normalization purpose. Interestingly, our data configuration illustrates an important advantage of this regime fixed cost specification. According to Table 1, the less frequently produced crop (*i.e.*, corn) is produced in at least 24% of our observations while 3 production regimes (*i.e.*, regimes 5, 6 and 8) out of 8 are adopted in less than 3% of our observations. The probability distribution of fixed costs cannot be estimated accurately with our dataset on a pure per regime basis. But, that of crop fixed costs can be.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Table 1. Descriptibe statistics

Regime Number	Average crops acreage shares per regime							Regime frequency	Average gross margin (€/ha)[b]
	Winter wheat	*Corn*	*Spring Barley*	*Sugarbeet*	*Alfalfa*	*Protein pea*	*Rapeseed*		
1	0.38	0.07	0.15	0.12	0.09	0.06	0.13	6.6%	953
2	0.37		0.16	0.15	0.11	0.06	0.15	21.5%	1014
3	0.38	0.07	0.17	0.14	0.10		0.14	11.8%	930
4	0.37		0.20	0.16	0.11		0.16	48.6%	1007
5	0.41	0.14	0.19	0.10			0.15	2.8%	989
6	0.50	0.14	0.14				0.22	2.5%	825
7	0.44		0.23	0.14			0.19	4.9%	970
8	0.58		0.15				0.27	1.3%	738
Production frequency	100%	24%	100%	96%	88%	28%	100%		
Average acreage share[a]	0.38 (0.09)	0.02 (0.05)	0.18 (0.07)	0.15 (0.06)	0.10 (0.05)	0.02 (0.03)	0.16 (0.06)		
Average acreage share[a] if produced[a]	0.38 (0.09)	0.08 (0.07)	0.18 (0.07)	0.15 (0.06)	0.11 (0.04)	0.06 (0.03)	0.16 (0.06)		
Average gross margin (€/ha)[a,b]	843 (327)	872 (449)	756 (287)	1789[d] (379)	562 (286)	663 (269)	843 (311)		
Average yield (t/ha)[a]	8.58[b]	9.23	6.82	95.19	12.62	4.72	3.89		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

	(0.88)	(1.73)	(1.21)	(13.01)	(1.92)	(1.28)	(0.64)
Average price (€/t)[a]	149[b] (31)	131 (34)	155 (35)	25[c] (3)	72 (15)	198 (25)	323 (64)
Average fertilization and crop protection costs[a]	431 (91)	308 (74)	294 (70)	547 (126)	350 (125)	246 (66)	415 (83)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

2.5.2. Estimation results

The parameter estimates of the yield, input demand, acreage shares and regime choice equations are reported in Tables 2 to 4. As shown in Table 2, the expectations of random parameters representing potential yields, $b^{y}_{k,i}$, are precisely estimated for all crops and their values lie in reasonable ranges regarding the average yields observed in the sample (Table 1). More importantly, the variances of their distributions are also statistically different from zero for all crops. These parameters thus significantly vary across farms, despite the fact that we control for observed factors characterizing farm heterogeneity (land and capital endowments and climatic conditions). This comes to illustrate the importance of unobserved farm heterogeneity in our sample.

The parameter estimates of the input demand equations, also reported in Table 2, confirm this result: the probability distribution of their farm specific parameters is precisely estimated and displays significant heterogeneity. This is true for the random intercepts $b^{x}_{k,i}$ (the input use requirement) but also for the random slope parameters, $a^{x}_{k,i}$, which represents the response of farmers to change in netput prices.

Turning to the parameter estimates of the acreage share equations in Table 3, again, the expectations and variance of random parameters are precisely estimated. Ranges of expectations of the acreage flexibility parameters are theoretically consistent. Conditions $a^{s}_{m|(g),i}{}^{3}\ a^{s}_{(g),i}{}^{3}\ a^{s}_{i} > 0$ hold on average. These are sufficient conditions for the acreage model to be well-behaved.

Finally, as shown in Table 4, the regime costs associated to crops, $d^{c}_{k,i}$, and the scale parameter, s_{i}, of the regime choice equation are significantly estimated and heterogeneous across the sample. The mean value of the scale parameter, 1.80, is large, reflecting the importance of regime profit and fixed cost levels in production regime choices. Simulation results provided in the next sub-section illustrate this point. Estimated mean fixed cost of alfalfa is negative on average. Two main reasons might explain this result. First, alfalfa is planted for at least two years. This crop requires farmers' intervention mostly at planting and harvesting. In the *Marne* region, the alfalfa downstream (dehydration) industry generally takes on harvest operations, which comes to decrease farmers' workload significantly. Second, being a legume alfalfa exhibits good agronomic properties, especially when used as a previous crop for cereals. Crop fixed cost estimates should, however, be considered cautiously given their high variability across farms.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Table 2. Selected Parameter Estimates of Yield Supply and Input demand Models[a]

	Winter wheat	*Corn*	*Spring Barley*	*Sugar beet*	*Alfalfa*	*Protein pea*	*Rape-seed*
Yield supply model							
Error term $e^{y}_{k,it}$							
Std dev	0.66*	1.83*	0.95*	9.70*	2.96*	1.72*	0.49*
	(0.02)	(0.07)	(0.02)	(0.02)	(0.02)	(0.03)	(0.016)
Potential yield $b^{y}_{k,i}$							
Mean	8.71*	9.06*	6.81*	95.60*	12.23*	4.15*	4.04*
	(0.02)	(0.04)	(0.02)	(0.32)	(0.04)	(0.03)	(0.01)
Std dev	0.26*	0.65*	0.33*	5.7*	0.69*	0.51*	0.24*
	(0.01)	(0.03)	(0.01)	(0.17)	(0.02)	(0.02)	(0.01)
Input demand model							
Error term $e^{x}_{k,it}$							
Standard deviation	0.52*	0.59*	0.41*	0.84*	0.88*	0.60*	0.58*
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)
Input requirement $b^{x}_{k,i}$							
Mean	4.36*	2.57*	2.92*	5.44*	3.15*	2.29*	4.44*
	(0.02)	(0.02)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)
Standard deviation	0.37*	0.33*	0.24*	0.54*	0.44*	0.37*	0.41*
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
Input use flexibility $a^{x}_{k,i}$							
Mean	0.43*	0.08*	0.30*	0.49*	0.25*	0.33*	0.79*
	(0.01)	(0.00)	(0.00)	(0.03)	(0.00)	(0.01)	(0.02)
Std dev	0.13*	0.09*	0.05*	0.58*	0.02	0.18*	0.31*
	(0.01)	(0.04)	(0.00)	(0.06)	(0.03)	(0.01)	(0.01)

a. Estimated standard errors of the ML estimator are in parentheses. Note: Asterisk (*) denotes a statistically significant parameter at the 5% level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Table 3. Selected Parameter Estimates of the Acreage Share Models[a]

Crop level random terms	*Winter wheat*	*Corn*	*Spring barley*	*Sugar beet*	*Alfalfa*	*Protein pea*	*Rape-seed*
Error term $e^{s}_{k,it}$							
Standard deviation	0	11.12* (0.38)	9.91* (0.19)	6.25* (0.13)	6.77* (0.15)	8.56* (0.28)	7.09* (0.16)
Acreage share shifters $b^{s}_{k,i}$							
Mean	0	17.41* (0.73)	13.88* (0.37)	24.51* (0.23)	11.15* (0.24)	18.78* (0.37)	11.07* (0.24)
Standard deviation	0	3.92* (0.02)	4.19* (0.02)	3.96* (0.03)	2.70* (0.06)	2.62* (0.01)	2.20* (0.01)

Acreage choice flexibility parameters	*Level 1* a^{s}_{i}	*Level 2 (groups)* $a^{s}_{(g),i}$		*Level 3 (subgroups)* $a^{s}_{n\|(g),i}$	
		Cereals	**Rotation heads**	**Spring cereals**	**Oil and protein crops**
Mean	0.046* (0.001)	0.053* (0.001)	0.073* (0.001)	0.530* (0.029)	0.11* (0.002)
Standard deviation	0.015* (0.001)	0.013* (0.001)	0.025* (0.001)	0.640* (0.029)	0.020* (0.002)

a. Estimated standard errors of the ML estimator are in parentheses. Note: Asterisk (*) denotes a statistically significant parameter at the 5% level.

Table 4. Parameter Estimates of Regime Choice Models

	Crop fixed costs $b^{c}_{k,i}$							**Scale parameter** s_{i}
	Winter wheat	*Corn*	*Spring barley*	*Sugar beet*	*Alfalfa*	*Peas*	*Rape-seed*	
Mean[a]	0	3.80* (0.24)	0	0.30* (0.12)	-4.70* (0.28)	1.30* (0.04)	0	1.80* (0.07)
Std dev[a]	0	4.16* (0.10)	0	2.22* (0.05)	4.40* (0.11)	0.67* (0.01)	0	1.40* (0.07)

a. Estimated standard deviation of the estimator in parentheses. Note: Asterisk (*) denotes a statistically non null parameter at the 5% level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Once we have estimated the parameters characterizing the distribution of the random parameters $\boldsymbol{\gamma}_i$, we can "statistically calibrate" those parameters for each farmer in our sample and thus obtain a set of farmer specific "calibrated" models that can then be used for simulation purposes (Koutchadé *et al*, 2018). In this study, the specific parameter $\boldsymbol{\gamma}_i$ of farm *i* is calibrated as the mode of its (simulated) probability distribution conditional on $(\mathbf{y}_{it}^{+},\mathbf{x}_{it}^{+},\mathbf{s}_{it}^{+},r_{it},\mathbf{z}_{it})$ for $t=1,...,T$ (*i.e*. according to a ML 'calibration' criterion conditionally on what is known about farm *i* in the data). One interesting feature is that this procedure also allows us to calibrate the parameters of the yield, input demand and acreage equations corresponding to crops that have not been grown by the considered farmer as well as farmer specific regime fixed costs for regimes that have never been chosen by the considered farmer.

The estimated farmer specific models allow us to compute fitting criteria, Sim-R^2, which are reported in Table 5. The Sim-R^2 criterion measures the quality of the prediction of the observed choices of farmers by the estimated models. Its construction is analogous to that of the R^2 criterion of the standard linear regression model: for a given choice variable and a given model, the Sim-R^2 criterion is defined as the ratio of the empirical variance of the prediction of this variable to the empirical variance of the observed variable.

Table 5. Fitting Criteria (Sim-R^2)

	Winter wheat	*Corn*	*Spring barley*	*Sugar beet*	*Alfalfa*	*Peas*	*Rape-seed*
Yield supply models	0.37	0.24	0.35	0.42	0.28	0.39	0.45
Input demand models	0.44	0.30	0.40	0.34	0.30	0.43	0.40
Acreage share models		0.57	0.34	0.83	0.70	0.53	0.41

These estimated criteria tend to show that the proposed model offers a satisfactory fit to our data.[18] Using the estimated farmer specific models to predict the regime choices observed in our data, we find our model to exhibit a relatively good predictive power with 72.4% of regime choices correctly predicted. Importantly, our investigations on this issue tend to demonstrate that our results are robust to various distributional assumptions related to the model random parameters.

2.5.2. Simulation results

The structure of the proposed ERS multi-crop micro-econometric model allows to investigate the relative importance of the main drivers of production regime choices. For that purpose, we consider the simulation model obtained from the estimated one by calibrating the farm specific parameters for each farm of our sample. Then we use this simulation model for investigating the prediction power of three elements of the regime choice models: the weighted sum of the expected crop gross returns $\boldsymbol{\pi}_{it}'\mathbf{s}_{it}(r)$, the acreage management costs

[18] Much better fit levels are obtained for crop supply, acreage and input demand model defined at the farm level, mostly due to the explanatory power of the cropland area variable.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$C_{it}(\mathbf{s}_{it}(r))$ and the regime fixed costs d_i^r for $r \in \mathcal{R}$. We simulate the regime choices according to each of these elements as well as combinations of these elements, and then confront them, on average, with the observed regime choices. Taken together these simulation results confirm that regime fixed costs matter, but mainly in combination with the other drivers of the regime choice model. The maximization of gross margins, or the minimization of acreage management costs or regime fixed cost alone leads to predictions of regime choices that are strongly biased on average. Considering pairs of these choice criteria only slightly improve the predictions, while considering together these three criteria unsurprisingly provides predicted choices very close, on average, to the observed ones.

To illustrate the relevance of the approach we propose to deal with corner solutions in acreage choices, we simulate the impacts of changes in expected crop prices on acreage choices. As acreage price elasticities play a crucial role in this type of exercise, we present them first. In our ERS-MEMC model these elasticities account for the impact of crop prices on both acreages within any given regime and switch in production regimes. These two effects can be distinguished by generalizing, to a multiple regime case, the decomposition proposed by McDonald and Moffit (1980) for standard Tobit models. The average acreage own price elasticities of our farm sample are reported in Table 6. They have expected signs and, because of the crop disaggregation level of our data, are larger than those commonly found in the literature. The decomposition of these elasticities shows that a large part of the price effects on acreages can be due to the inclusion or not of these crops in the production regimes chosen by farmers. For crops like corn or pea, which are minor crops in the considered area, changes in the production regimes account for about one third of the estimated price elasticities. However, changes in the production regimes can also have significant effects for frequently produced crops. For instance, they account for 11% of the sugar beet acreage own price elasticities.

Table 6. Average Own Price Elasticities of Crop Acreages

	Winter wheat	*Corn*	*Spring barley*	*Sugar beet*	*Alfalfa*	*Protein pea*	*Rape--seed*
Average crop acreage own price elasticities	0.33	4.26	0.44	1.39	0.74	1.22	0.76
Due to changes in acreages within production regimes	0.33 (100%)	2.33 (55%)	0.43 (98%)	1.24 (89%)	0.60 (81%)	0.71 (58%)	0.75 (99%)
Due to changes in production regimes	0.00 (0%)	1.93 (45%)	0.01 (2%)	0.15 (11%)	0.14 (19%)	0.51 (42%)	0.01 (1%)

Observing how crop acreage elasticities within production regimes vary across regimes allows to illustrate the main features distinguishing ERS-MEMC models from their CR-MEMC counterparts. Table 7 reports the estimated means of own price crop acreage elasticities per regime.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Table 7. Per Regime Average Own Price Crop Acreage Elasticities

Regime			Crops produced in the regime						
Number	*Frequency*	*Crop number*	*Winter wheat*	*Corn*	*Spring barley*	*Sugar beet*	*Alfalfa*	*Peas*	*Rape-seed*
1	6.6%	7	0.33	0.95	0.92	1.19	0.62	0.68	0.84
2	21.5%	6	0.31		0.32	1.17	0.61	0.67	0.82
3	11.8%		0.32	0.95	0.92	1.16	0.57		0.75
4	48.6%	5	0.30		0.32	1.14	0.56		0.74
5	2.8%		0.31	0.95	0.90	1.10			0.44
6	2.5%	4	0.29	0.95	0.90				0.35
7	4.9%		0.29		0.31	1.10			0.43
8	1.3%	3	0.27		0.30				0.30

These estimates display significant differences across production regimes. In particular, crop acreage own price elasticities grow with the number of crops produced in the considered production regime. The higher the crop number, the more farmers can make use of crop acreage substitution opportunities. For instance, the more the considered regime contains rotation starting crops, the more rapeseed acreage choices are responsive to rapeseed price. This elasticity range, on average, from 0.30, when rapeseed is the only rotation starting crop in the regime, to 0.84, in regimes with 4 rotation entry crops. Similarly, barley acreages are much more responsive to changes in barley price in regimes including corn than in regimes without corn. Corn and barley are the only spring cereals in farmers' crop set. Crop acreage models of CR-MEMC models cannot represent the substitution patterns uncovered by our estimation results. These models account for crop regimes but consider the same crop acreage model for all production regimes.

The impact of the production regime choice is further highlighted by simulating the effects of increases in the price of protein pea on its acreages. Owing to its fixing atmospheric nitrogen for themselves as well as for following crops, French agricultural scientists consider pea as a "diversification crop" of particular interest by. Yet, protein pea acreages have declined over the last decade in the considered area mostly because of lacking profitability, especially as regards to that of other rotation starting crops. The simulated impacts of increases in the price of peas on crop acreages are depicted in Figure 3.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

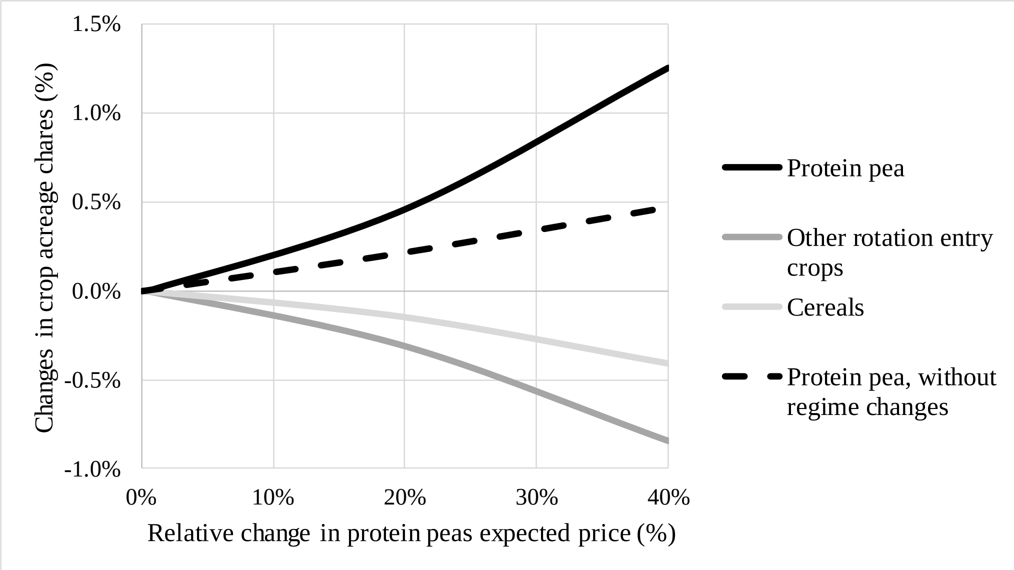

Figure 3. Estimated impacts of protein pea expected price on crop acreage shares

According to our results, a 40% increase in the expected price of pea would increase the average pea acreage share by 1.3%, from 2.0% to 3.3%. These additional pea acreages would mainly replace those of other rotation starting crops. The combined average acreage share of rapeseed, alfalfa and sugar beet would decrease by 0.9% while that of cereals would only decrease by 0.4%. This illustrates the interest in considering crop – agronomic and management – characteristics when specifying the acreage management cost function. Interestingly, about two thirds of the increase in the pea acreage would be due to new producers. This also explains another feature of our simulation results. The simulated increases in the pea acreage is not linear in the price of pea. In particular, the increase in the pea acreages is more pronounced above the 20% price increase level than below. Threshold effects due to production regime fixed costs and changes in crop acreage elasticities due to regime changes can explain this pattern. These induce kinks in farmers' pea acreage choices that are smoothed by the averaging process.

2.6. Conclusion

The main aims of this article are threefold. First, we present an original modelling framework for dealing with null acreages in MEMC models. This framework is fully consistent from an economic viewpoint and explicitly considers regime fixed costs. These features make the ERS-MEMC model proposed in this article suitable for analyzing and, to some extent, disentangling, the effects of the main drivers of farmers' acreage choices at disaggregation levels at which issues raised by null acreages are pervasive. Our estimation and simulation results notably tend to demonstrate that expected crop returns are not the sole significant drivers of farmers' crop acreage choices, at least in the short run. In particular, crop production fixed costs also matter. These results also show that crop acreages display patterns that cannot be accounted for by the CR-MCEM models currently used for handling null acreage choices. Effects of economic incentives on the crop acreage choices of a farmer strongly depend on

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

the crop set chosen by the considered farmer.

Second, the application presented in this article illustrates the empirical tractability of random parameter ERS-MEMC models for investigating farmers' production choices. Of course, estimating such models raises challenging issues. But, this is also necessary for estimating structured micro-econometric models suitably accounting for important features characterizing micro-economic agricultural production data, among which significant unobserved heterogeneity. In particular, to estimate such models enables analysts to calibrate simulation models consisting of samples of farm specific models.

Third, according to our experience, ML estimators computed with stochastic versions of EM algorithms appear to be interesting alternatives to Simulated ML estimators for relatively large systems of interrelated equations such as the random parameter ERS-MEMC models considered in our empirical application. SAEM algorithms appear to be particularly relevant.

Of course, significant specification and estimation issues remain to be addressed. First, the empirical tractability of the ERS-MEMC model proposed in this article strongly relies on properties that are specific to the MNL crop acreage share models proposed by Carpentier and Letort (2014). Adapting our modelling approach to other crop acreage choice models would widen the scope of specification search for ERS-MEMC models. Also, the ERS-MEMC model considered in our empirical application relies on restrictive assumptions regarding the dynamic features of multi-crop technologies and farmers' choice process. Finally, the estimation cost of the models proposed in this article is relatively high, due to long computing and coding times.

This article proposes solutions to methodological issues that could be used for improving micro-econometric analyzes of policies impacting crop acreage choices. For instance, Babcock (2015) noted that policies related to biofuels led to a dramatic increase in interest in the econometric analyses of crop supply response to crop prices. The ERS-MEMC models considered in this article not only allow to disentangle intensive and extensive margin effects, they also allow to investigate crop choice effects. Analyzing crop choices also appear crucial for investigating agri-environmental policies and issues. For instance, changes in the location of crop production induced by climate change are due to crop set choices made the farm level. Also, as fostering crop diversification tend to become an important agri-environmental objective in many countries, including those of the European Union, coherent model of farmers' crop set choices appear to be especially relevant. Finally, random parameter ERS-MEMC models can contribute to close the gap existing between MEMC models and mathematical programming models (*e.g.*, Heckeleï et al, 2012; Mérel and Howitt, 2014). The overall structure of our ERS-MEMC models is similar to that of mathematical programming models and their random parameter versions can be used for calibrating heterogeneous farm models.

Estimation costs appears to be among the limitations of our modelling framework that need to be addressed. Significant computing and coding costs make applied research work, such as specification search, tedious and time consuming. Relatively slight modifications of the model specification could, however, significantly reduce the estimation cost of the ERS-MEMC model presented in this article. For instance, in the model considered in the empirical application the covariance parameters of the random parameter vector $\boldsymbol{\gamma}_i$ represent more than 64% of the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(786) estimated parameters. Yet, our estimates demonstrate that the random parameters of the crop yield supply and input demand models are strongly correlated, for a given crop but also across crops. This suggests that these parameters are linked by a few farmer specific “productivity factors”. Such latent productivity factors could be used for imposing some structure on the variance-covariance matrix of the considered random parameters and, thereby, for significantly reducing the number of covariance parameters to be estimated. Also, relying on Laplace approximates for the regime choice probability function of the ESR-MEMC model involves tedious and time consuming computations. Less accurate but significantly simpler approximation approaches could dramatically reduce the estimation burden. But, these cruder approximation approaches could also impact the consistency of the estimated model. The extent of these impacts and the related – estimation burden *versus* specification approximation – trade-off is worth investigating in future work.

3. COST ALLOCATION

3.1. Variable input allocation among crops: a time-varying random parameters approach

3.1.1. Introduction

Getting information about production costs for each crop at the farm level is very important when analyzing multi-crop farms’ behaviors. It is indeed very useful to investigate variable input uses decisions of farmers for policy purpose. Production costs per crop can also be used as explanatory variables in more complex models of production choice (Letort and Carpentier, 2010). However, information these cost per crop is generally not provided in accountancy dataset, such as Farm Accountancy Data Network (FADN) data, available to agricultural economists. The information on variable input uses in these data only concerns aggregate expenditure at the farm level, and adequate statistical and\or economic modeling are necessary to allocate this aggregate information among the different crops produced by the farms.

Different approaches have been proposed in the agricultural economics literature to overcome this issue. Carpentier and Letort (2012) distinguish two groups of approaches. The first group includes approaches that consider only variable input allocation equations, in which the input allocation coefficients are treated as unknown parameters to be estimated, these parameters being either fixed, parametric functions of exogenous variables, or random (Dixon, Batte and Sonka 1984; Hornbaker et al., 1989; Just et al., 1990; Dixon and Hornbaker 1992). The second group of approaches considers input allocation equations as a part of a system of equations that includes crop yield equations, acreage functions or production equations (Just et al., 1990; Chambers and Just, 1989; Letort and Carpentier, 2012). Even if the second approach introduces a lot more economic information compared to the first approach, the first approach is the most widely used, owing to its ease of implementation using regression approaches (OLS, GLS, SUR), and to the satisfactory results it generally provide in terms of production cost predictions compared to the second approach (Just et al.,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

1990). Estimating single variable input allocation equations however raises different issues that must be addressed to ensure the consistency of this approach. First, the use of standard regression approaches does not guarantee that estimated input costs lie in reasonable ranges. These approaches can, for instance, lead to negative estimates of input costs per crop. Secondly, because input costs vary across farms, the observed, but also unobserved, heterogeneity among farms and farmers has to be taken into account. Finally, input uses per crop depend on acreage choice decisions, which are also determined by unobserved farm characteristics. This can lead to estimation issues when one seeks to account for unobserved farm heterogeneity in input allocation equations.

Given the limited information generally available in observed data samples, a general way to overcome the issues concerning the magnitude of estimated input costs is to impose constraints on parameters or to introduce additional out-of-sample information. Approaches based on inequality-restricted regression estimation (Ray, 1985; Dixon and Hornbaker, 1992), on Bayesian estimation (Moxey et Tiffin, 1994; Heckelei et al., 2008) and on Generalized Maximum Entropy estimation (Léon et al., 1999) have been proposed to this end. Issues related to the presence of unobserved farm heterogeneity in input allocation equations have also been addressed in the literature (Dixon, Batte and Sonka 1984; Hornbaker et al., 1989; Dixon and Hornbaker 1992; Hallam et al., 1999). However, as pointed out by Lence and Miller (1998), Dixon and Hornbaker (1992) and Carpentier and Letort (2012), the random parameter (RP) approaches, generally used in that case, have to deal with issues related the dependence between variable input use and acreage choice decisions. Dixon and Hornbaker (1992) propose correlation tests without, however proposing a method allowing to take this correlation into account, while Carpentier and Letort (2012) propose an approach based on control functions, which requires a simultaneous estimation of input use and acreage choices equations. To our knowledge, the different approaches proposed in the literature to estimate uniquely input allocation equations thus do not allow to simultaneously (i) control for unobserved farms and farmers heterogeneity, (ii) deal with the dependence of input uses per crop to acreage choices and (iii) guarantee consistent values of input use estimates.

Our main objective in this paper is to propose an approach allowing to address these three issues. To do so, we consider a panel data model of input allocation derived from accounting identities. We use a random parameter specification to account for farm unobserved heterogeneity. The unobserved crop input uses are viewed as time-varying random parameters, and we control for the potential correlation between crop input uses and acreage decisions by expressing these random parameters as functions of (tine-varying) exogenous variables containing acreage shares. To ensure that the estimated input uses per crop lie in reasonable ranges, we introduce additional information in the model through the distribution of random parameters. For instance, using lognormal distribution for random parameters, we enforce non-negativity constraints.

This model is estimated, using an extension of Stochastic Approximation Expectation Maximization (SAEM) algorithm (Delyon et al., 1999), on a sample French farms' accounting data. Our estimation results show that our RP estimation approach performs better in terms of input use predictions than its OLS counterpart. The rest of the paper is structured as follows. In section 3.1.2., we present our model of input use allocation. Our SAEM estimation approach is presented in section 3.1.3, and the empirical results in section 3.1.4. . Finally, we conclude.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

3.1.2. Random Parameter model of input use allocation

We consider a set of crops $C \equiv \{1,2,\ldots,C\}$ produced by a farmer i $(i=1,\ldots,N)$ in period t. We denote by $s_{c,it}$ the acreage allocated to crop $c \in C$ by farmer i in period t. In the following, we focus on one variable input used by farmer i to simplify the presentation of the model, given that the generalization to J inputs is straightforward. Let $\bar{x}_{it}$ denotes the quantity of variable input used at the farm level by farmer i at time t and $x_{c,it}$ denote the quantity of variable input used per unit of land of crop $c \in C$. The input allocation problem considered consists in recovering the input quantity $x_{c,it}$ for each crop $c \in C$, for each farmer i and each period t such that:

(1) $\bar{x}_{it} = \sum_{c \in C} s_{c,it} x_{c,it} = \mathbf{s}_{it} \mathbf{x}_{it}$,

with $\mathbf{x}_{it} = (x_{c,it} : c \in C)$ and $\mathbf{s}_{it} = (s_{c,it} : c \in C)$.

Including the (centered) measurement error term u_{it}, equation (1) becomes:

(2) $\bar{x}_{it} = \sum_{c \in C} s_{c,it} x_{c,it} + u_{it} = \mathbf{s}_{it}' \mathbf{x}_{it} + u_{it}$ with $E[u_{it}] = 0$.

In addition, the input use equation at farm level is completed by the model of crop input uses $x_{c,it}$ for $c \in C$.

3.1.2.1. Specification of $x_{c,it}$

One of the main advantages of panel data is that it allows the estimation of models accounting for the variability of unobserved determinants - called unobserved heterogeneity - of the modeled phenomena (see, e.g., Woodridge, 2002; Arellano and Bonhomme, 2011). In our case, these determinants can be unobserved characteristics of the farmers (e.g., aptitudes, motivations) and farms (e.g., soil quality, spatial distribution of the plot, available material) which do not vary or vary little over the time period considered. Here, it is assumed that crop input uses $x_{c,it}$ is a transformation of normal distributed terms $m_{c,it}$, where this transformation induces bounds. By doing so, it is easy to guarantee that the estimated crop input uses $x_{c,it}$ lies in reasonable ranges. For instance, to force the positivity of $x_{c,it}$, we can assumed that $x_{j,k,it} = \exp(m_{j,k,it})$. This allows introducing constraints on crop input uses using unconstrained parameterization. More specifically, we assume that:

(3) $x_{c,it} = h(m_{c,it})$

and

(4) $m_{c,it} = b_{c,i} + a_{c,t,0} + e_{c,it}$ with $E[e_{c,it}] = 0$,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

where h is a non-linear transformation. For example h^{-1} may be log-normal distribution, censored-normal distribution or Johnson's (1949) SB distribution (Train, 2005). The last may allows incorporating additional information provided by experts in the model.

The normal distributed term $\mu_{c,it}$ is decomposed in three components. First, the farm-specific effects $\beta_{c,i}$ denote the mean values of $\mu_{c,it}$. It is assumed that $\beta_{c,i}$ is specific to farmer/farm for $c \in C$ and vary randomly across farmers and farms. It will be normal distributed. The farm-specific effects $\beta_{c,i}$ allows accounting for, among others, the effect of farms' factor endowment (e.g., machinery, soil quality, climatic conditions) and farmers' motivations and skills. Secondly, the year-specific effect $\alpha_{c,t,0}$ denote the deviations of $\mu_{c,it}$ with respect to the farm-specific effects $\beta_{c,i}$ at time t. For identification, $\alpha_{c,t,0}$ is normalized to 0 at *t=1*. Lastly, the error terms of the model of crop input uses $\varepsilon_{c,it}$ allow accounting for the effects of stochastic events not taken into account in $\beta_{c,i}$ or $\alpha_{c,t,0}$ (*e.g.*, weather, pest infestation).

In compact form, equation (4) becomes:

(5) $\boldsymbol{\mu}_{it} = \boldsymbol{\beta}_i + \boldsymbol{\alpha}_{t,0} + \boldsymbol{\varepsilon}_{it}$ or $\boldsymbol{\mu}_{(i)} = \boldsymbol{\iota}_{T_{(i)}} \otimes \boldsymbol{\beta}_i + \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}_{(i)}$

where $\boldsymbol{\mu}_{it} = (\mu_{c,it} : c \in C)$ and $\boldsymbol{\mu}_{(i)} = (\boldsymbol{\mu}_{it} : t \in \mathcal{H}_{(i)})$ are column vector. Similarly, we define $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{\varepsilon}_{(i)}$, and $\boldsymbol{\alpha}_{t,0}$ and $\boldsymbol{\alpha}_0$.

It is assumed that:

(6) $u_{it} \sim_{iid} \mathcal{N}(0,\sigma_0^2)$, $\boldsymbol{\varepsilon}_{it} \sim_{iid} \mathcal{N}(\mathbf{0},\boldsymbol{\Omega}_0)$ and $\boldsymbol{\beta}_i = (\beta_{c,i} : c \in C) \sim_{iid} \mathcal{N}(\boldsymbol{\omega}_0,\boldsymbol{\psi}_0)$.

It is also assumed that the covariance matrices $\boldsymbol{\Omega}_0$ are diagonal matrices. All correlation in crop input uses decisions will be captured by $\boldsymbol{\beta}_i$ through the covariance $\boldsymbol{\psi}_0$, which it is assumed unrestricted. Finally, it is assumed that u_{it}, $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{\beta}_i$ are mutually independent, and u_{it}, $\boldsymbol{\varepsilon}_{it}$, and $\boldsymbol{\beta}_i$ are independent to $s_{c,it}$ for $c \in C$.

The assumption of independence between $\boldsymbol{\beta}_i$ are independent to $s_{c,it}$ may be unverified. Let $\mathbf{z}_{it}$ denotes the vectors of observed farmer's/farm's characteristics. As previously mentioned, $\boldsymbol{\varepsilon}_{it}$ capture the effects of stochastics events influencing farmers' input use decisions during the cropping season. These events are thus unknown to farmers at the time of acreage choices, in the planting season, implying that they are independent of crop acreages:

(7) $E(\boldsymbol{\varepsilon}_{it} \mid \mathbf{s}_{it}, \mathbf{z}_{it}) = E(\boldsymbol{\varepsilon}_{it} \mid \mathbf{s}_{it}) = \mathbf{0}$.

On the other hand, $\boldsymbol{\beta}_i$ captures the impacts of unobserved farmers' characteristics that may also affect their acreage choice decisions. To account for this potential link between acreage choice and crop input use decisions and avoid bias in the estimation of the model, we follow Mundlak (1978) and specify:

(8) $\beta_{c,i} = \omega_{c,0} + (s_{c,i.} - E(s_{c,i.}))\pi_{c,0} + \eta_{c,i}$

where $s_{c,i.}$ denote the means of crop acreage shares at farm level, $E(s_{c,i.})$ denotes the sample average of $s_{c,i.}$ and $\eta_{c,i}$ denote the random effects. We also consider

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(9) $$m_{c,it} = b_{c,i} + a_{c,t,0} + (s_{c,it} - E(s_{c,it}))d_{c,0} + e_{c,it},$$

where $E(s_{c,it})$ is the sample average of $s_{c,it}$. Given that, $m_{c,it}$ depends on $s_{c,it}$ and

(10) $$E(h_{c,i} \mid s_{c,i.}) = 0 \text{ and } E(e_{c,it} \mid s_{c,it}) = 0$$

More generally, it is possible to incorporate in flexible way observed control variables – including crop acreage shares and crop yield levels – in the crop input uses model:

(11) $$b_{c,i} = w_{c,0} + (\mathbf{z}_{c,i.} - E(\mathbf{z}_{c,i.}))'\boldsymbol{\pi}_{c,0} + h_{c,i}$$

(12) $$m_{c,it} = b_{c,i} + a_{c,t,0} + (\mathbf{z}_{c,it} - E(\mathbf{z}_{c,it}))'\boldsymbol{\delta}_{c,0} + e_{c,it},$$

where $\mathbf{z}_{c,it}$ include farmers' crop acreage share and farmers observed characteristics. For instance, the average yield at region levels, can be used as a proxy for the production and sanitary conditions of each region, in order to account for the specificity of production in each region. This will improve the identification of crop input uses model.

To summarize, the considered input allocation model is a random parameters model where the random parameters are time-varying and depend on both (centered) time-invariant and time-varying control variables:

(13) $$\bar{x}_{it} = \sum_{c \in C} s_{c,it} x_{c,it} + u_{it} \text{ with } u_{it} \sim_{iid} \mathcal{N}(0, s_0^2),$$

(14) $$x_{c,it} = h(m_{c,it}),$$

(15) $$m_{c,it} = w_{c,0} + (\mathbf{z}_{c,i.} - E(\mathbf{z}_{c,i.}))'\boldsymbol{\pi}_{c,0} + (\mathbf{z}_{c,it} - E(\mathbf{z}_{c,it}))'\boldsymbol{\delta}_{c,0} + h_{c,i} + a_{c,t,0} + e_{c,it}$$

where:

(16) $$\boldsymbol{\eta}_i = (h_{c,i} : c \in C) \sim_{iid} \mathcal{N}(\mathbf{0}, \boldsymbol{\psi}_0) \text{ and } \boldsymbol{\varepsilon}_{it} = (e_{c,it} : c \in C) \sim_{iid} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0).$$

The covariance matrix $\boldsymbol{\Omega}_0$ is diagonal matrix while the covariance $\boldsymbol{\psi}_0$ is unrestricted covariance matrix. It is also assumed that u_{it}, $\boldsymbol{\varepsilon}_{it}$, and $\boldsymbol{\eta}_i$ are mutually independent, and u_{it}, $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{\eta}_i$ are independent to $s_{c,it}$ for $c \in C$.

This model is estimated based on an extension of the SAEM algorithm proposed by Delyon et al. (1999). The approach used for its estimation is presented in detail in Appendix 3A.

3.1.3. Empirical application

3.1.3.1. Data

This section presents an application aimed to illustrate the empirical tractability of our modelling approach as well as to demonstrate his ability to predict variable input cost per crop for each farmer at each point in time t. The model presented above was applied to a sample of 1081 French (5028 observations) grain crop producers located in the North and North-East of France and observed from 2007 to 2014. Farmers are observed at least three consecutive years in the sample. We consider 11 crops produced in this area (wheat, winter barley, spring barley, corn, sugar beets, alfalfa, peas, rapeseed, poppy seed, potatoes, starchy potatoes). The available information includes acreage and yield levels for each crop and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

variable input use expenditures, at farms level. This sample has been extracted from data provided by an accounting agency located in the French territorial division La Marne. Here, we present fertilizer and pesticide use allocation. The advantage of the considered data is that the input costs per crop are also available, *e.g.*: Table 8 shows the average fertilizer and pesticide use expenditures (euro/ha, at base 2005 price level) per crop and per year, observed in sample (for produced crops). They have been used to validate the results of our estimations. Input prices are computed for each category of crops using he hired production services price index (base 100 in 2005) obtained from the French department of Agriculture. For pesticides, these prices vary from crop to another. However, for fertilizers, we assume the same price for all considered crops. Since these prices may differ from crop to another as in the case of pesticides, they are directly introduced in the in the estimation process. This allows accounting for price fluctuations other time.

Table 8 also shows the average crop acreage shares. These vary from one crop to another, and crops with large acreage share have generally high production frequency.

Table 8. Descriptive statistics of the sample

	Freq. of production (%)	Acreage share		Pesticide	Fertilizer
		Sample (%)	Produced (%)	euro/ha, at base 2005 price levels	
Winter wheat	100	35	35	180	187
Spring barley	87	15	18	102	139
Winter barley	65	06	10	150	160
Corn	34	05	14	103	151
Sugar beet	81	12	15	251	224
Alfalfa	62	07	11	60	207
Peas	26	02	07	141	66
Winter rapeseed	92	16	17	191	181
Blue opium poppy	08	01	07	63	102
Potato	11	01	08	704	248
Starch potato	08	01	10	480	266

3.1.3.1. Estimation results

We estimated input allocation equations for pesticides and fertilizers for the considered 11 crops that cover more that 90% of the considered farms. We considered one type of constraints on estimated crop input uses. Non-negativity constraint are imposed on crop input uses using log-normal parameterization (unconstrained parameterization). We also incorporating crop acreage shares observed in sample and other farms/farmers characteristics (e.g., average crop yields by department obtained from the French department of Agriculture) as control variables in the crop input uses models.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Our estimations are conducted by using the R package WInputAll developed for this purpose (see section 3.2). More details this package is given below. The recursive step of simulation of the SAEM algorithm is implemented using 100 draws (MCMC) at each iteration. We consider 300 iterations for the first stage of estimation where the algorithm explores parameters space without memory, tries to escape local maxima and reach quickly the neighborhood of the maximum likelihood estimator. The algorithms converges without difficulties and convergences of parameters are checked using plot of the sequences of estimated parameters at each iteration. The global convergence is also checking regarding the plot of the sequence of the estimated complete data log-likelihood functions, as it resumes all information in parameters.

Selected estimation results are reported in Table 9 and Table 10, the complete results being available from the authors upon request. These results show that the model fits relatively well to the data. Most parameters are well estimated especially the expectations and the variance parameters of random parameters. Table 9 shows the expectations and the variances of the parameters, which are statistically significant and demonstrate that unobserved heterogeneity matters in farmers' crop input uses.

Table 9. Parameters estimates: estimated distribution of random parameters

	Pesticide (%)			Fertilizer (%)		
$\ln(x_{k,it}) = w_k + \bar{\mathbf{z}}'_{k,i}\boldsymbol{\pi} + \mathbf{z}'_{k,i}\boldsymbol{\delta} + h_{k,i} + e_{k,it}$	Expecta - tion: w_k (SE)	Variance of $h_{k,i}$ (SE)	Variance of $e_{k,it}$ (SE)	Expecta - tion: w_k (SE)	Variance of $h_{k,i}$ (SE)	Variance of $e_{k,it}$ (SE)
Winter wheat	40.8 (0.8)	5.4 (0.3)	0.2 (0.0)	**18.6** (0.5)	3.0 (0.2)	0.2 (0.0)
Spring barley	-12.2 (0.3)	0.5 (0.0)	0.2 (0.0)	**11.2** (0.5)	2.2 (0.1)	0.3 (0.0)
Winter barley	54.9 (0.3)	0.6 (0.1)	0.3 (0.0)	**62.4** (0.5)	1.6 (0.1)	0.2 (0.0)
Corn	-06.8 (0.3)	0.8 (0.1)	0.3 (0.0)	**38.3** (0.3)	0.9 (0.1)	0.2 (0.0)
Sugar beet	89.3 (0.4)	1.2 (0.1)	0.2 (0.0)	**85.4** (0.4)	1.5 (0.1)	0.2 (0.0)
Alfalfa	-218.7 (2.4)	0.1 (0.0)	0.2 (0.0)	**97.1** (0.5)	1.9 (0.1)	0.2 (0.0)
Peas	23.0 (0.2)	0.3 (0.0)	0.1 (0.0)	**30.6** (0.3)	0.4 (0.0)	0.3 (0.0)
Winter rapeseed	73.1 (0.6)	2.8 (0.2)	0.3 (0.0)	**58.1** (0.5)	2.6 (0.1)	0.2 (0.0)
Blue opium poppy	-26.5 (0.4)	1.0 (0.1)	0.2 (0.0)	**33.5** (0.4)	0.7 (0.1)	0.2 (0.0)
Potato	194.4 (0.4)	0.5 (0.0)	0.2 (0.0)	**99.1** (0.5)	2.1 (0.1)	0.1 (0.0)
Starch potato	149.5 (0.2)	0.1 (0.0)	0.2 (0.0)	**82.2** (0.4)	0.8 (0.1)	0.2 (0.0)

Once we have estimated the parameters characterizing the distribution of the random parameters $\boldsymbol{\mu}_{(i)}$, we can "statistically calibrate" those parameters for each farmer in our sample and thus obtain a set of farmer specific "calibrated" models that can then be used to predict $\mathbf{x}_{it}$, $\mathbf{x}_{it} = \exp(\boldsymbol{\mu}_{it})$ for $t \in \mathcal{H}_i$. In this study, the specific parameter $\boldsymbol{\mu}_{(i)}$ of farm *i* is calibrated as the mode of its (simulated) probability distribution conditional on observed data, when it is used. One interesting feature is that this procedure also allows us to calibrate the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

potential input cost $\mathbf{x}_{it}$ corresponding to crops that have not been grown by the considered farmer. The estimated farmer input cost $\hat{\mathbf{x}}_{it}$ compared to the real values, allow us to compute fitting criteria, Sim-R², which are reported in Table 10. The Sim-R² criterion measures the quality of the prediction of the observed choices of farmers by the estimated models. It is obtained by regress the observed value on predicted value. These estimated criteria tend to show that the proposed model offers a satisfactory fit to our data. Also, the estimates crop input uses means lies in reasonable range regarding the sample average of crop input uses.

Table 10. Fitting criteria

	Pesticide				Fertilizer			
	Sim -R2	AAD	Estimate d mean (	Sample Average	Sim-R2 %	AAD	Estimate d mean (S.d.t)	Sample Average (S.d.t)
Winter wheat	54	0.31	173 (41)	180 (40)	75	0.24	148 (40)	187 (47)
Spring barley	16	0.24	97 (09)	102 (28)	61	0.19	128 (25)	139 (38)
Winter barley	05	0.41	178 (16)	150 (41)	48	0.29	207 (39)	160 (42)
Corn	03	0.37	94 (08)	103 (35)	34	0.30	146 (16)	151 (42)
Sugar beet	20	0.63	246 (29)	251 (65)	60	0.34	274 (53)	224 (71)
Alfalfa	03	0.58	11 (0.7)	60 (27)	39	0.77	299 (50)	207 (80)
Peas	07	0.47	129 (17)	141 (42)	01	0.31	131 (08)	66 (33)
Winter rapeseed	33	0.36	235 (38)	191 (49)	66	0.28	196 (35)	181 (47)
Blue opium poppy	00	0.19	74 (6)	63 (28)	00	0.26	140 (18)	102 (36)
Potato	08	1.53	708 (43)	704 (144)	13	0.66	268 (51)	248 (67)
Starch potato	39	1.01	448 (23)	480 (113)	54	0.52	245 (26)	266 (78)
	82				80			

Figures 4-6 display our results for three selected crops wheat, rapeseed and potato. Input uses are measured per ha in 100€ at the 2005 prices. Figure 4 demonstrate that we obtain reasonably good results when estimating fertilizer and pesticide input uses for winter wheat, which is produced by all sampled farmers and represents 35% of the arable crop acreage on average in our dataset. These Figures plot the estimated per hectare input use levels against their observed "true" counterparts. Of course, our estimated input use levels significantly differ from their true counterparts. But, most estimates lie within reasonable ranges around their true counterparts. For instance the average difference between the true and estimated (in absolute value, AAD) fertilizer use equals 0.37 while the average fertilizer use equals about 2 (i.e., about 200€/ha at the 2005 fertilizer prices). Yet, we underestimate fertilizer uses. Rapeseed is produced by 96% of the sampled farms but its average acreage share doesn't exceed 15%. Figure 5 shows that the estimated fertilizer and pesticide use for rapeseed are of lower quality than those for wheat, and that we overestimate pesticide uses for rapeseed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Figure 6 shows that our estimation approach fits relatively poorly the chemical input uses for potato production, which only concerns 11% of the sampled farms (for an average crop acreage of 2%).

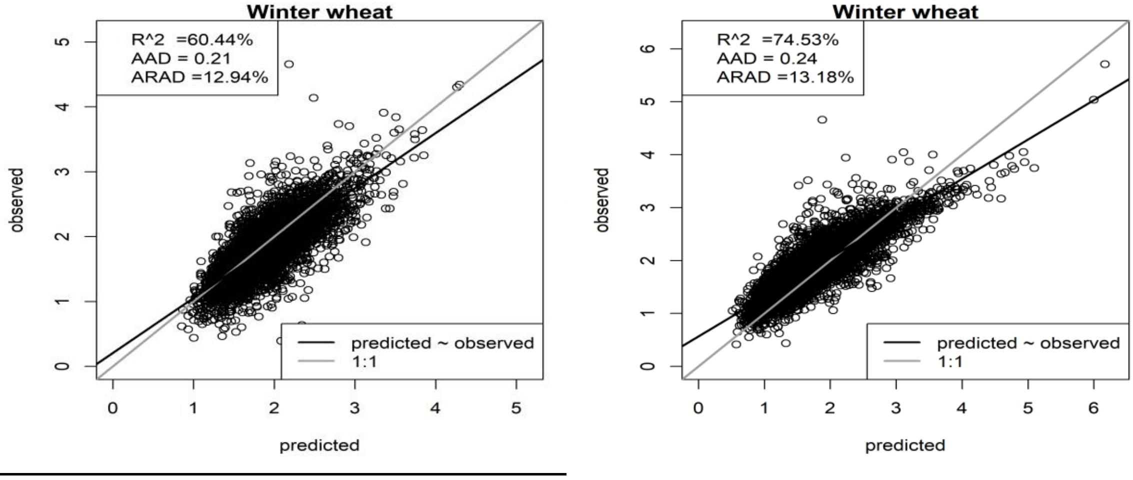

Figure 4. Observed versus estimated pesticide (left) and fertilizer (right) uses for wheat

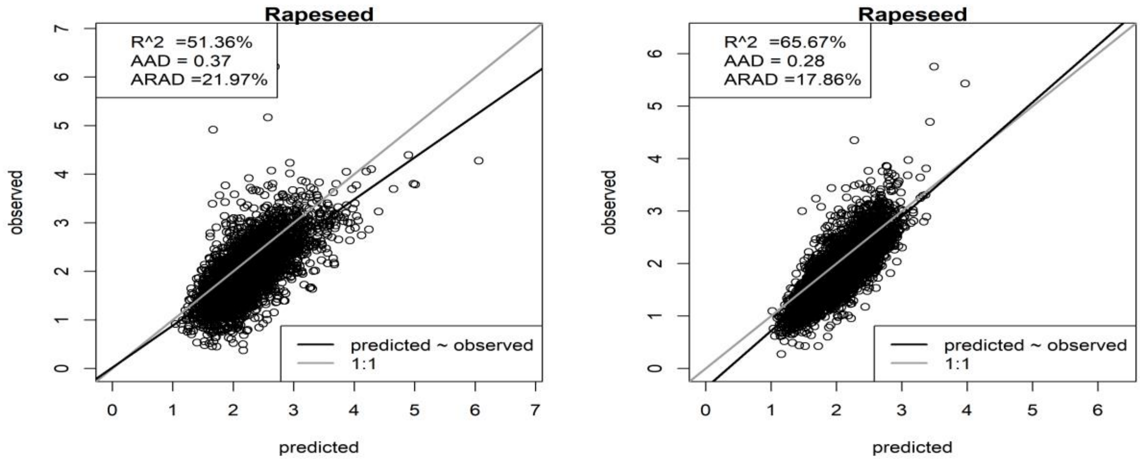

Figure 5. Observed versus estimated pesticide (left) and fertilizer (right) uses for rapeseed

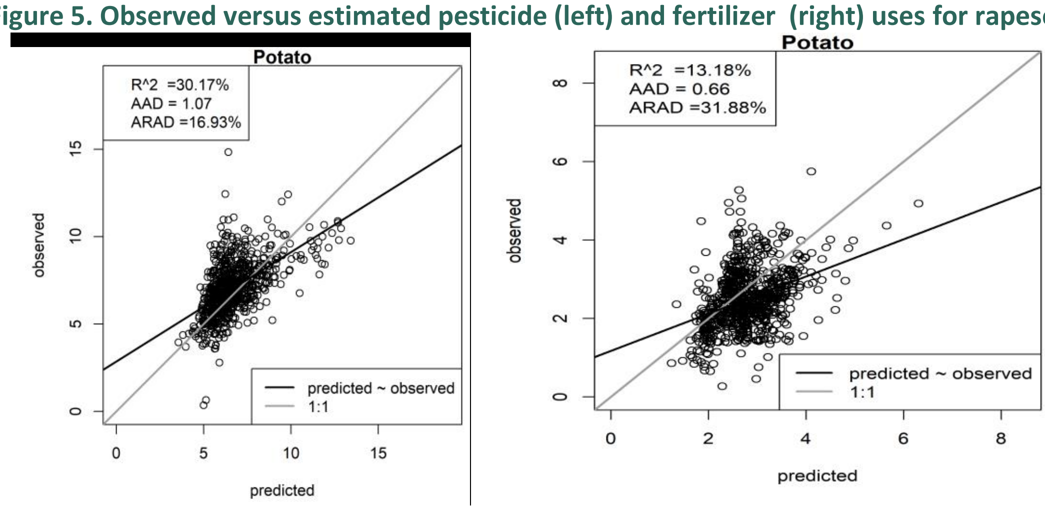

Figure 6. Observed versus estimated pesticide (left) and fertilizer (right) uses for potato

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

3.1.4. Conclusion

In this study, we consider random parameter input allocation model. It allows characterizing unobserved heterogeneity across farms and incorporating non-negativity constraints on crop input uses in flexible way. Our results show that (i) recovering pesticide uses is generally more difficult than recovering fertilizer uses, (ii) estimation accuracy decreases with the average acreage share of the considered crop and (iii) average estimated input uses are close to their true counterparts, in general. These results are promising.

We are currently investigating the effects of various constraints as means for improving crop input use estimates. Our final objective is (i) to characterize the models and constraint sets yielding the most accurate results and (ii) to devise an algorithm for estimating the considered models that is relatively easy to code, and to provide suitable ranges for its tuning parameters.

3.2. WInputAll: An R package for input cost allocation

The results presented in section 3.1 have been obtained by using an R Package developped by INRAE team within the MIND STEP project: the WinputAll package. This package is in the final testing phase and will be made available to R users on CRAN in the future.

The package thus includes the main function rpinpallEst for the single-input case. The documention of this function is presented in section 3.2.1 bellow.

3.2.1. Usage

```
## S3 method for class 'rpinpall'
print(x, ...)

## S3 method for class 'rpinpall'
summary(object, ...)

## S3 method for class 'rpinpall'
plot(x, ...)

rpinpallEst(
  data,
  id_time,
  total_input,
  crop_acreage_sh,
  crop_input_price = NULL,
  crop_indvar = NULL,
  crop_indvar_i = NULL,
  weight = NULL,
  distrib = c("lognormal", "normal", "logit-normal", "censored-normal", "probit-normal"),
  lower = -Inf,
```


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

```
 upper = Inf,
 sim.method = c("MHI", "MHRW", "MH"),
 calib.method = c("estim-sim", "cmode", "cmean"),
 saem.control = list()
)
```

3.2.2. Arguments

x	An object produced by the function rpinpallEst, to be displayed
...	Other arguments
object	An object produced by the function rpinpallEst, to be displayed
data	name of the data frame or matrix containing all the variables included in the model
id_time	first (individual) and second (time) level variables allowing characterizing panel data.
total_input	variable containing the total input used at farm level to be allocated to the different crops
crop_acreage_sh	list of variables containing the acreage shares of the different crops
crop_input_price	optional list of variables containing the input prices for considered crops. Default=NULL
crop_indvar	optional list of vector of (time-varying) variables specific to each crop used to control for observed (individual and/or temporal) characteristics in the estimation process. Default=NULL
crop_indvar_i	optional list of vector of (time-constant) variables specific to each crop used to control for observed time-constant characteristics in the estimation process. Default=NULL
weight	optional variable containing weights of individual sample farms. Default=NULL (equal weight is given to each farm). Default=NULL
distrib	assumption on the distribution of input use per crop (x_kit): "normal", "lognormal" "logit-normal", "probit-normal" or "censored-normal". Default="lognormal"
lower	lower bound when using a bounded distribution for x_kit
upper	upper bound when using a bounded distribution for x_kit
sim.method	method used to draw the random parameters in the simulation step of the estimation process: "MHI" (independant Metropolis Hasting), "MHRW" (Metropolis Hasting Random Walk) or "MH" (combined "MHI" and "MHRW"). Default= "MH"
calib.method	method used: "cmode" (conditional mode), "cmean" (conditional mean) and "estim-sim" (last simulation in the estimation process). Default="estim-sim"
saem.control	list of options for the SAEM algorithm. See 'Details

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

3.2.3. Details

An SAEM algorithm is used to perform the estimation of input uses per crop. Different options can be specified by the user for this algorithm in the saem.control argument The saem.control argument is list that can supply any of the following component :

maxit	maximum total number of iterations. Default=1000
K.burn	other arguments
K.SA	an object produced by the function rpinpallEst, to be displayed
K.RS	name of the data frame or matrix containing all the variables included in the model
tol	first (individual) and second (time) level variables allowing characterizing panel data.
rdraw.mhrw	number of random draws in the estimation process when sim.method="MHRW". Default=50
rdraw.mhi	number of random draws in the estimation process when sim.method="MHI". Default=50
calib.mult	allow providing the number of random draws in the calibration process. Default=10
stde.mult	allow providing the of random draws for computation of estimation standard errors. Default=10
p.SA	parameter determining step sizes in the Stochastic Approxiation (SA) step. Must be comprise between 0 and 1. Default=1
dec.omega.u	option forcing the final estimated variance of error terms to be quasi null (5.10-3) in order to ensure the equality between the (estimated) sum of inputs per crop and the (observed) total input. Default=FALSE
doParallels	logical.If TRUE a parallel processing is used when more than 2 cores are available. Default=FALSE
doTempering	logical. If TRUE the tempering approach proposed by (Allassonnière and Chevallier, 2021) is used to avoid convergence to local maxima. Default=TRUE
showProgress	logical. If TRUEthe evolution of the estimation process is displayed graphically at the bottom of the screen. Default=TRUE
showIterConvLL	logical. If TRUE iteration number and convergence value are displayed during the estimation process. Default=FALSE

3.2.4. Returned results

rpinpallEst returns a list with the following components:

xit_predict	matrix of predicted crop input used per ha
yit_predic	list of results of estimation: estimated parameters
est_pop	list of parameters standard errors

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

call	copy of the function call
opt	list of saem algorithm options
conv.ind.cll vecto	vector of convergence indicator
data.list	list of individual data used for estimation

3.2.5. Functions

The following methods allow display the results of estimation:

print(rpinpall)	displays the distribution of estimated crop input uses
summary(rpinpall)	displays a summary of estimated parameters
plot(rpinpall)	plot the "global" convergence indicator

4. MICRO-ECONOMETRIC MULTI-CROP MODEL: APPLICATION AND ALLEVIATING THE ESTIMATION BURDEN

4.1. (Slight) model changes for (significant) alleviation of the estimation burden

Estimating the Micro-Econometric Multi-Crop (MEMC) model presented in Section 2 is challenging for two main reasons. First, it features numerous random parameters the joint probability distribution of which needs to be estimated. If Stochastic Approximate Expectation-Maximization (SAEM) algorithms (Delyon, Lavielle and Moulines, 1999, Lavielle, 2014) enable applied statisticians (and econometricians) to efficiently estimate parametric random parameter models, the estimation burden of such models quickly increases (at a quadratic rate) in the number of random parameters featured in the considered model. We address this dimension issue by employing a common factor approach. Second, the MEMC model we consider features endogenous production regimes, which a major originality of this model. Estimating this model requires to compute regime choice probability functions that are particularly complicated. In particular, these probability functions are defined as expectations over the distribution of two sets of terms, the entire set of random parameters of the model and a sub-set of error terms of the crop acreage choice model. The solution approach we initially adopted, which is proposed by Harding and Hausman (2007) and based Laplace approximations, yields accurate approximations of the regime choice probability functions of interested. Yet, this approach is computationally intensive and cumbersome to code. We tested two competing approaches against the Laplace approximation approach: an approach based on simulation methods, which is also computationally intensive but easier to code, and an approach based on a relatively rough approximation of the considered probability function, which is easy to code and computationally (very) light. We obtained satisfactory solutions to both issues.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

We also tried to address another issue that showed up during the course of Mind Step. The MEMC model we consider assumes that farmers choose their production regime, that is to say the set of crops they actually grown, among a given set of regime choices. The empirical versions of this model we have estimated so far feature 15 regimes (that cover more than 85% of the observations of the datasets we use). Increasing the regime number is possible in theory, and could be very useful for empirical purposes, but appears to be challenging from a practical viewpoint. We have not been able to obtain satisfactory solutions to this issue so far. It is difficult to specify a reliable discrete choice model when the choice set exceeds hundreds.[19]

4.1.1. Factor structure and parameter reduction

The MEMC model presented in Section 2 implies features random parameter vector $\boldsymbol{\gamma}_i = (g_{m,i} : m = 1,...,M)$ that aims to account for farms' and farmers' unobserved heterogeneity. This random parameter vector is assumed to be multivariate normal, with

(1) $\boldsymbol{\gamma}_i : \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Omega}_0)$.[20]

Since the model we consider is a relatively large equation system the dimension of parameter $\boldsymbol{\gamma}_i$ is quite large in our applications (*e.g.*, M ranges around 70). Moreover, as $\boldsymbol{\gamma}_i$ is used for accounting for unobserved heterogeneity effects, restricting its probability distribution is unwarranted *a priori*. Therefore, we considered estimating our MEMC model while leaving variance matrix $\boldsymbol{\Omega}_0$ unrestricting in the first place.

Two observations led us to reconsider this option. First, the number of unrestricted elements of matrix $\boldsymbol{\Omega}_0$ grows in M at a quadratic rate. This number is given by $M(M+1)/2$ – that is to say M variance parameters and $M(M-1)/2$ covariance parameters – meaning hundreds of parameters to be estimated in our applications. This dimensionality issue does not rise practical computational issues (which is a major virtue of the EM type algorithms we use) but rather identification problems, especially regarding the covariance parameters among the elements of $\boldsymbol{\gamma}_i$ (*e.g.*, Cherchi and Guevara, 2012). These are numerous and often poorly estimated (which also slows down the estimation process). Second, our empirical results demonstrated specific patterns in the probability distribution of the elements of $\boldsymbol{\gamma}_i$ and their relationships. In particular, the parameters representing crop potential yield levels and the crop input levels required to achieve these potential yield levels appear to be strongly

[19] The regime number issue considered here appears to be less challenging with the (quadratic) acreage choice models considered by Heckeleï and Wolff (2003) or Carpentier and Letort (2012), which are very close to those featured in common PMP models (*e.g.*, CAPRI, IFM-CAP). In these models (which ignore production regime fixed costs), acreage choices are characterized by the first order conditions to a (constrained) quadratic programming problem. These consist of a system of in/equality conditions. Importantly, production regime choices need not be explictly modelled as they are fully characterized by the by the first order conditions related to crop acreage choices. We plan investigate the estimation of such "random parameter PMP type models with endogenous regime switching" in future projects.

[20] Note, however, that this random parameter can be partly transformed (with, *e.g.*, exponential or Logit transformations, truncations) when incorporated in the model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(positively) correlated for a given crop but also across crops. This suggests that the joint probability distribution of these farm specific "technical" parameters is partly governed by a few productivity or practice intensity common factors.

Taken together these observations led us to consider reducing the number of parameters characterising the probability distribution of random parameter vector $\boldsymbol{\gamma}_i$ on the one hand, and to adopt a simple common factor approach for doing so in the other hand. The assumptions underlying our "parameter reduction" approach are quite simple. (*i*) The correlations among the elements of $\boldsymbol{\gamma}_i$ are due a limited set of "common factors" and (*ii*) the effects of these common factors can be incorporated in the probability distribution of $\boldsymbol{\gamma}_i$. Indeed, we assume that $\boldsymbol{\gamma}_i$ can be decomposed as:

(2) $\boldsymbol{\gamma}_i = \boldsymbol{\mu}_0 + \boldsymbol{\Lambda}_0\mathbf{b}_i + \boldsymbol{\eta}_i$

where term $\mathbf{b}_i$ designates the set of latent common factors we consider while term $\boldsymbol{\Lambda}_0$ designates the factor loading parameter $M \times L$ matrix and term $\boldsymbol{\eta}_i$ is a residual term.[21] We further assume that:

(3a) $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}_L)$,

(3b) $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Psi}_0)$ where variance matrix $\boldsymbol{\Psi}_0$ is diagonal

and:

$\mathbf{b}_i$ and $\boldsymbol{\eta}_i$ are independent.

The assumptions stating that matrix $\boldsymbol{\Psi}_0$ is diagonal and the independence of $\mathbf{b}_i$ and $\boldsymbol{\eta}_i$ indicate that all correlations among the elements of $\boldsymbol{\gamma}_i$ are due to common factors $\mathbf{b}_i$ in the considered model of $\boldsymbol{\gamma}_i$. In this model, common factors $\mathbf{b}_i = (b_{l,i} : l = 1,...,L)$ basically are modelling tools that are mostly aimed to capture the links among the elements of $\boldsymbol{\gamma}_i$. Indeed, the elements of $\mathbf{b}_i$ may not be easily interpreted. They may not even capture well-defined productivity or practice intensity factors.[22] This explains why factor loading matrix $\boldsymbol{\Lambda}_0$ is left unrestricted.

Based on the specification of $\boldsymbol{\gamma}_i$ given above, the probability distribution of $\boldsymbol{\gamma}_i$ is given by:

(4) $\boldsymbol{\gamma}_i \sim \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Omega}_0^r)$

---

[21] Note that the specification of $\boldsymbol{\gamma}_i$ can easily be extended to include the effects of observed common factors in addition to latent factors $\mathbf{b}_i$.

[22] In other words, our specification of $\boldsymbol{\gamma}_i$ is essentially instrumental or exploratory in the terminology of James (2018). Indeed, estimates of factor loading matrix $\boldsymbol{\Lambda}_0$ may yield insights on the empirical content of common factors $\mathbf{b}_i$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

where:

(5) $\boldsymbol{\Omega}_0^r = \boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0' + \boldsymbol{\Psi}_0$.

The considered "common factor" structure of $\boldsymbol{\gamma}_i$ significantly reduces the number of parameters to be estimated if the dimension of $\mathbf{b}_i$, which equals L, is sufficiently smaller than that of $\boldsymbol{\gamma}_i$, which equals M.

Matrix $\boldsymbol{\Omega}_0^r$ contains $L(M+1)$ unrestricted parameter, with LM parameters in factor loading matrix $\boldsymbol{\Lambda}_0$ and M parameters in diagonal variance matrix $\boldsymbol{\Psi}_0$. Indeed, the considered "common factor" structure of $\boldsymbol{\gamma}_i$ saves covariance parameters to be estimated if and only if $L \leq M/2$. Of course, this structure is satisfactory if it does not unduly constraint the correlations of the elements of $\boldsymbol{\gamma}_i$, that is to say if $\boldsymbol{\Omega}_0^r = \boldsymbol{\Omega}_0$, a condition that can be checked empirically.

Panhard and Samson (2009) proposed a convenience factorization of the complete likelihood function of random parameter models that make the implementation of EM-type algorithms easy despite their accounting for "common factor" structures similar to the one considered here. This option is included in the R package presented below.

4.1.2. Approximating the regime choice probability

Computing the likelihood function of the MEMC model presented in Section 2 implies integrating regime choice probability functions given (using simplified notations) by:

(6) $P(r_{it};\boldsymbol{\Psi}) = \int P(r_{it} \mid \boldsymbol{\varepsilon}_{it}^{s}) f(\boldsymbol{\varepsilon}_{it}^{s,0} \mid \boldsymbol{\varepsilon}_{it}^{s,+} = \hat{\boldsymbol{\varepsilon}}_{it}^{s,+};\boldsymbol{\Psi}) d\boldsymbol{\varepsilon}_{it}^{s,0}$

where function $f(\boldsymbol{\varepsilon}_{it}^{s,0} \mid \boldsymbol{\varepsilon}_{it}^{s,+} = \hat{\boldsymbol{\varepsilon}}_{it}^{s,+};\boldsymbol{\Psi}_0)$ denotes the pdf of $\boldsymbol{\varepsilon}_{it}^{s,0}$ conditional on $\boldsymbol{\varepsilon}_{it}^{s,+} = \hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$. Term $\boldsymbol{\varepsilon}_{it}^{s} = (\boldsymbol{\varepsilon}_{it}^{s,+}, \boldsymbol{\varepsilon}_{it}^{s,0})$ is the error term vector of the considered acreage choice model, which is assumed normal with $\boldsymbol{\varepsilon}_{it}^{s} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Psi}_0)$.

Let define matrices $\boldsymbol{\Psi}_0^{++}(r_{it}) = V[\boldsymbol{\varepsilon}_{it}^{s,+}]$, $\boldsymbol{\Psi}_0^{00}(r_{it}) = V[\boldsymbol{\varepsilon}_{it}^{s,0}]$ and $\boldsymbol{\Psi}_0^{0+}(r_{it}) = Cov[\boldsymbol{\varepsilon}_{it}^{s,0}, \boldsymbol{\varepsilon}_{it}^{s,0}]$. It is easily shown that:

(7) $\boldsymbol{\varepsilon}_{it}^{s,0} \mid (\boldsymbol{\varepsilon}_{it}^{s,+} = \hat{\boldsymbol{\varepsilon}}_{it}^{s,+}) \sim \mathcal{N}(\boldsymbol{\mu}_{it}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}, r_{it}), \boldsymbol{\Psi}_0^{0|+}(r_{it}))$

where:

(8a) $\boldsymbol{\mu}_0^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}, r_{it}) = \boldsymbol{\Psi}_0^{0+}(r_{it})\boldsymbol{\Psi}_0^{++}(r_{it})^{-1}\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$

and:

(8b) $\boldsymbol{\Psi}_0^{0|+}(r_{it}) = \boldsymbol{\Psi}_0^{00}(r_{it}) - \boldsymbol{\Psi}_0^{0+}(r_{it})\boldsymbol{\Psi}_0^{++}(r_{it})^{-1}\boldsymbol{\Psi}_0^{0+}(r_{it})'$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Term $\boldsymbol{\varepsilon}_{it}^{s,+}$ relates to the acreages of crops that are grown in regime r_{it} while $\boldsymbol{\varepsilon}_{it}^{s,0}$ relates to crops that are not grown by famer *i* in year *t*. Error term $\boldsymbol{\varepsilon}_{it}^{s,+}$ can be is consistently estimated by $\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$. But, term $\boldsymbol{\varepsilon}_{it}^{s,0}$ must be considered as missing in the estimation process because it cannot be recovered by combining the model and the data. The Multinomial Logit functional form of function $P(r_{it}\,|\,\boldsymbol{\varepsilon}_{it}^{s})$ prevents its integration over the probability distribution of $\boldsymbol{\varepsilon}_{it}^{s,0}$, either analytically or numerically.

Building on the work of Harding and Hausman (2007), we first consider using Laplace approximates of the regime choice probability functions $P(r_{it}\,|\,\boldsymbol{\varepsilon}_{it}^{s})$ for computing the likelihood function of our model. This approach relies on a second order Taylor expansion in $\boldsymbol{\varepsilon}_{it}^{s,0}$ of function $P(r_{it}\,|\,\boldsymbol{\varepsilon}_{it}^{s})f(\boldsymbol{\varepsilon}_{it}^{s,0}\,|\,\boldsymbol{\varepsilon}_{it}^{s,+}=\hat{\boldsymbol{\varepsilon}}_{it}^{s,+};\boldsymbol{\Psi})$ around an optimally chosen value of $\boldsymbol{\varepsilon}_{it}^{s,0}$. It is, however, time consuming and computationally cumbersome.

Use of standard simulation methods for integrating function $P(r_{it};\boldsymbol{\Psi})$ is simple. It suffices to draw simulations of terms $\boldsymbol{\varepsilon}_{it}^{s,0}$ from $\mathcal{N}(\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\boldsymbol{\Psi}^{0|+}(r_{it}))$ and to approximate $P(r_{it};\boldsymbol{\Psi})$ by its simulated counterpart. This approach is easy to code and works relatively well. But, it does not reduce computing time and, maybe more importantly, it requires simulation draws of which we already make use massively for handling the random parameters of our MEMC model.

Indeed, the issue we face here arises because term $\boldsymbol{\varepsilon}_{it}^{s,0}$ is known to farmers – implying that it can contribute to drive their production choice – while it is not observed by the analyst. Computing probability functions $P(r_{it}\,|\,\boldsymbol{\varepsilon}_{it}^{s})$ consists in solving the multivariate integration problem given in equation (17). Any accurate solution approach to this problem is necessarily computationnally intensive and, as a result, time consuming. Accordingly, any simple solution approach supposes to accept either accuracy loss or some change in the model specification.

The best estimator of $\boldsymbol{\varepsilon}_{it}^{s,0}$ that can be computed easily by the analyst is its mean conditional on $\boldsymbol{\varepsilon}_{it}^{s,+}=\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$, that is to say $\boldsymbol{\mu}_{0}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it})=E[\boldsymbol{\varepsilon}_{it}^{s,0}\,|\,\boldsymbol{\varepsilon}_{it}^{s,+}=\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}]$. It can be shown that $P(r_{it}\,|\,\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\hat{\boldsymbol{\varepsilon}}_{it}^{s,+})$ is the first order Taylor expansion of $P(r_{it};\boldsymbol{\Psi})$ in $\boldsymbol{\varepsilon}_{it}^{s,+}$ around $\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$. Of course, term $P(r_{it}\,|\,\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\hat{\boldsymbol{\varepsilon}}_{it}^{s,+})$ is a relatively crude approximation of $P(r_{it};\boldsymbol{\Psi})$,[23] especially when variance matrix $\boldsymbol{\Psi}^{0|+}(r_{it})$ is relatively large.[24] Yet, approximating probability function $P(r_{it};\boldsymbol{\Psi})$ by term $P(r_{it}\,|\,\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\hat{\boldsymbol{\varepsilon}}_{it}^{s,+})$ appears to yield very satisfactory results according to our empirical investigations. It yields estimates of the MEMC model that very close to those obtained when integrating by the Laplace approximation approach or by standard simulation methods (at least in our applications). This may indicate that term $\boldsymbol{\varepsilon}_{it}^{s,0}$ is not a major driver of production regime choice r_{it}, function $P(r_{it}\,|\,\boldsymbol{\varepsilon}_{it}^{s})$ is almost linear in $\boldsymbol{\varepsilon}_{it}^{s,0}$

---

[23] Since $E[\boldsymbol{\varepsilon}_{it}^{s,0}-\boldsymbol{\mu}_{0}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it})\,|\,\boldsymbol{\varepsilon}_{it}^{s,+}=\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}]=\mathbf{0}$.

[24] It suffices to observe that $P(r_{it}\,|\,\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\hat{\boldsymbol{\varepsilon}}_{it}^{s,+})+tr\left(\boldsymbol{\Psi}^{0|+}(r_{it})\frac{\partial^{2}}{\partial\boldsymbol{\varepsilon}^{s,0}\partial(\boldsymbol{\varepsilon}^{s,0})'}P(r_{it}\,|\,\boldsymbol{\mu}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+},r_{it}),\hat{\boldsymbol{\varepsilon}}_{it}^{s,+})\right)$ is the second order Taylor expansion of $P(r_{it};\boldsymbol{\Psi})$ in $\boldsymbol{\varepsilon}_{it}^{s,+}$ around $\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

on its range or/and variance matrix $\mathbf{\Psi}^{0|+}(r_{it})$ is relatively small. This later case can occur if terms $\boldsymbol{\varepsilon}_{it}^{s,0}$ and $\boldsymbol{\varepsilon}_{it}^{s,+}$ are sufficiently correlated (thereby implying that $\boldsymbol{\mu}_{0}^{0|+}(\hat{\boldsymbol{\varepsilon}}_{it}^{s,+}, r_{it})$ is a relatively accurate estimator of $\boldsymbol{\varepsilon}_{it}^{s,0}$). We did not further investigate the underpinnings of these empirical observations as we were mostly interested in solving practical computational issues.

The approximate solution approach presented above is computationally much faster (as well as much easier to code) than its more accurate alternatives. This option is included in the R package presented below.

4.2. RPMultiCrop: An R package for the estimation of MEMC-ERS models

The RPMulticrop R package has been developped to allow the estimation of the ERS-MEMC model presented in section 2 on any farm cost acocunting dataset containing information on crop yeilds, acreage shares, input and output prices and, importantly, input costs per crop.

4.2.1. Package documentation

4.2.1.1. Usage

```
rpmulticropEst(
  data,
  idtime,
  crop_nest,
  crop_ref = 1,
  crop_yield,
  crop_input,
  crop_acreage_share,
  crop_price,
  crop_input_price,
  crop_subsidy = NULL,
  crop_acreage = NULL,
  uaa = NULL,
  indvar_yield_input = NULL,
  indvar_acreage = NULL,
  indvar_rp = NULL,
  distrib.method = c("modA", "modB"),
  sim.method = c("MH", "MHI", "MHRW"),
  calib.method = c("CMODE", "estim-sim", "CMEAN"),
  saem.control = list()
)
```

4.2.1.2. Arguments

x	An object produced by the function rpmulticropEst, to be displayed
...	Other arguments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

object	An object produced by the function rpmulticropEst, to be displayed
data	name of the data frame or matrix containing all the variables included in the model
idtime	first (individual) and second (time) dimensions of the panel data.
crop_nest	list that describes crop nesting structure
crop_ref	name of reference crop
crop_yield	vector of variables containing crop yields (t/ha)
crop_input	list of vectors of variables containing input uses per crop (unit/ha)
crop_acreage_share	vector of variables containing crop acreage shares
crop_price	vector of variables containing crop prices(euro/t)
crop_input_price	list of vectors of variables containing input prices per crop (euro/unit)
crop_subsidy	vector of variables containing crop subsidies (euro/ha)
crop_acreage	vector of variables containing crop acreage (ha)
uaa	variable containing utilized agricultural area (ha)
indvar_yield_input	vector of variables used to control for observed (individual and/or temporal) characteristics in yield equations
indvar_acreage	vector of variables used to control for observed (individual and/or temporal) characteristics in acreage equations
indvar_rp	vector of variables used to control for observed (individual) characteristics in random parameters model
distrib.method	assumption on the distribution of random parameters
sim.method	method used to draw the random parameters in the simulation step of the estimation process: "MH" (Metropolis Hasting) , "MHI" (independant Metropolis Hasting), "MHRW" (Metropolis Hasting Random Walk)
calib.method	method used to calibrate the random parameters for each individual once the model converged: "CMODE" (conditional mode), "CMEAN" (conditional mean) or "sim-saem" (last simulation of the saem algorithm)
saem.control	list of options for the SAEM algorithm. See 'Details

4.2.1.3. Details

An SAEM algorithm is used to perform the estimation of input uses per crop. Different options can be specified by the user for this algorithm in the saem.control argument The saem.control argument is list that can supply any of the following component

nb.iter	Maximum total number of iterations. Default=1000
nb.burn	Number of iterations of the burn-in phase where individual parameters are sampled from their conditional distribution using sim.method and the initial values for model parameters without update these parameters. Default=10

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

nb.SA	Number of iteration in the First stage of estimation where algorithm explore parameters space without memory. The parameter that controls the convergence of the algorithm is set to 1. Default=200.
nb.temp	Number of iteration of iterations where tempering approach is used. Algorithm tries to escape local maxima if doTempering=TRUE. Default=300. Note that nb.burn+nb.SA must be inferior to K.RS and nb.temp must inferior to nb.iter
toler	Tolerance value for the convergence. Default 1.10-3
rdraw.m1	number of random draws in the estimation process when sim.method="MHRW". Default=10
rdraw.m2	number of random draws in the estimation process when sim.method="MHI". Default=10
calib.mult	allow providing the number of random draws in the calibration process. Default=10
stde.mult	allow providing the of random draws for computation of estimation standard errors. Default=10
p.SA	parameter determining step sizes in the Stochastic Approxiation (SA) step. Must be comprise between 0 and 1. Default=1
doParallels	logical.If TRUE a parallel processing is used when more than 2 cores are available. Default=FALSE
doTempering	logical. If TRUE the tempering approach proposed by (Allassonnière and Chevallier, 2021) is used to avoid convergence to local maxima. Default=TRUE
showProgress	logical. If TRUEthe evolution of the estimation process is displayed graphically at the bottom of the screen. Default=TRUE
showIterConvLL	logical. If TRUE iteration number and convergence value are displayed during the estimation process. Default=FALSE

4.2.1.4. Returned results

rpmulticropEst returns a list with the following components:

est_pop	list of parameter estimates and standard errors of these estimates
call	copy of the function call
opt	list of saem algorithm options
conv.ind.cll vecto	vector of convergence indicator
data.list	list of individual data used for estimation
data.simul	data frame, to be used for simulation purpose: data frame used for the estimation supplemented by additional columns containing the values of estimated fixed parameters and calibrated random parameters for each observation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

4.2.1.5. Functions

The following functions can be used for summarizing and graphically describing the estimation results:

summary(rpmulticrop)	displays a summary of estimated parameters
plot(rpmulticrop)	plot the "global" convergence indicator

4.2.2. Notice for using the estimation results to run simulations

The data frame contained in the data.simul element of the results returned by the rpMulticropEst function, can be used to build a multi-crop model production choices for simulation purpose. This data frame actually contains observed input and output prices ($p_{k,it}$ and $w_{k,it}$), values of control variables ($\mathbf{c}^y_{k,it}, \mathbf{c}^x_{k,it}$), estimated values of fixed parameters ($\hat{\boldsymbol{\delta}}^y_{k,0}, \hat{\boldsymbol{\delta}}^x_{k,0}$) and calibrated values of the random parameters ($\hat{b}^y_{k,i}, \hat{b}^x_{k,i}, \hat{a}^x_{k,i}, \hat{a}^s_i, \hat{a}^s_{(g),i}, \hat{s}_i, \hat{b}^c_{k,i}$) for each observation of the sample.

Based on this information, the impacts any changes in input and/or output prices on yields, input uses and acreage choices can be simulated for each observation of the considered sample by proceeding as follows[25]:

For a given observation and considering new input and output prices $p^*_{k,it}$ and $w^*_{k,it}$

1. Compute predicted yield ($\tilde{y}_{k,it}$), input uses ($\tilde{x}_{k,it}$) and gross margin ($\tilde{\pi}_{k,it}$) for each crop:

$$\tilde{y}_{k,it} = \hat{b}^y_{k,i} + (\hat{\boldsymbol{\delta}}^y_{k,0})'\mathbf{c}^y_{k,it} - 1/2 \times \hat{a}^x_{k,i} w^{*2}_{k,it} p^{*-2}_{k,it}$$

$$\tilde{x}_{k,it} = \hat{b}^x_{k,i} + (\hat{\boldsymbol{\delta}}^x_{k,0})'\mathbf{c}^x_{k,it} - \hat{a}^x_{k,i} w^*_{k,it} p^{*-1}_{k,it}$$

$$\tilde{\pi}_{k,it} = p^*_{k,it}\tilde{y}_{k,it} - w^*_{k,it}\tilde{x}_{k,it}$$

2. Computed predicted indirect profit ($\tilde{\Pi}_{it}$) and probability of choice ($\tilde{P}$) for each potential crop production regime:

$$\tilde{\Pi}_{it}(r) = (\hat{a}^s_i)^{-1} \ln \sum_{h \in \mathcal{G}} \left(\sum_{\ell \in \mathcal{K}(h)} j_\ell(r) \exp\left(\hat{a}^s_{(h),i}(\tilde{\pi}_{\ell,it} - \hat{b}^s_{\ell,it})\right)\right)^{\hat{a}^s_i (\hat{a}^s_{(h),i})^{-1}}$$

$$\tilde{P}(r_{it}) = \frac{\exp\left(\hat{s}_i(\tilde{\Pi}_{it}(r_{it}) - \hat{d}_i(r_{it}))\right)}{\sum_{r \in \mathcal{R}} \exp\left(\hat{s}_i(\tilde{\Pi}_{it}(r) - \hat{d}_i(r))\right)}, \text{ with } \hat{d}_i(r) = \sum_{k \in \mathcal{K}^+(r)} \hat{b}^c_{k,i}$$

[25] Note that the procedure described here correspond to the case of a two-level nesting structure of crop acreages, as presented in section 2, but the RPMultiCrop package can also accomodate 3-level nesting strutures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

3. Keep the regime (r_{it}^{*}) with highest probability of choice and corresponding yields $\tilde{y}_{k,it}$, input uses $\tilde{x}_{k,it}$ for the crops produced in this regime and compute predicted acreage shares ($\tilde{s}_{k,it}$) within this regime:

$$\tilde{s}_{k,it}(r^{*}) = \frac{\varphi_k(r)\exp\left(\hat{\alpha}^{s}_{(g),i}(\tilde{\pi}_{k,it} - \hat{\beta}^{s}_{k,it})\right)\left(\sum_{\ell \in \mathcal{K}(g)} \varphi_\ell(r)\exp\left(\hat{\alpha}^{s}_{(g),i}(\tilde{\pi}_{\ell,it} - \hat{\beta}^{s}_{\ell,it})\right)\right)^{\hat{\alpha}^{s}_{i}(\hat{\alpha}^{s}_{(g),i})^{-1}-1}}{\sum_{h \in \mathcal{G}}\left(\sum_{\ell \in \mathcal{K}(h)} \varphi_\ell(r)\exp\left(\hat{\alpha}^{s}_{(h),i}(\tilde{\pi}_{\ell,it} - \hat{\beta}^{s}_{\ell,it})\right)\right)^{\hat{\alpha}^{s}_{i}(\hat{\alpha}^{s}_{(h),i})^{-1}}}$$

4.3. Calibration of MP models with estimated "behavioural" parameters obtained from the estimated MEMC model

The Micro-Econometric Multi-Crop (MEMC) model presented in Section 2 significantly differs from standard Positive Mathematical Programming (PMP) models. Yet, both types of models share important features that make estimated MEMC models useful for calibrating the "slope" parameters of PMP models at the farm level, provided that these models consider the same farm sample. The considered "slope" parameters largely determine crop acreage responses to changes in (expected) crop returns.

Estimating the random (farm-specific) MEMC model of Section allows obtaining farm-specific crop acreage choice elasticities. Analysts conceiving PMP models have developed efficient procedures for calibrating the "slope" parameters of their models against available activity choice elasticities. These procedures are not discussed here. The purpose of this section is to propose an alternative calibration procedure. This procedure exploits the common feature of the MEMC model and PMP models. Indeed, the conception of considered MEMC model makes use of an acreage cost management function that is analogous to the so-called PMP term of PMP models. The proposed calibration procedure is trivial as it consists in applying fairly simple "calibration formulae".

4.3.1. Problem setting

We assume that the MEMC model presented in Section 2 is estimated on a panel dataset covering a large farm sample ($i = 1,...,N$) over a T year period ($t = 1,...,T$). Let define the following notations:

$\mathcal{K} = \{1,...,K\}$: crop set

S_{it} : total crop acreage of farmer i in year t

$S_{k,it}$: acreage of crop k chosen by farmer i in year t, with $\mathbf{S}_{it} = (S_{k,it} : k \in \mathcal{K})$

$s_{k,it} = S_{k,it}S_{it}^{-1}$: acreage share of crop k chosen by farmer i in year t, with $\mathbf{s}_{it} = (s_{k,it} : k \in \mathcal{K})$

$p_{k,it}$: return of crop k expected by farmer i in year t, with $\boldsymbol{\pi}_{it} = (p_{k,it} : k \in \mathcal{K})$

We assume that the considered MEMC model is based on a profit function involving a two-level MNL/entropic acreage management cost function. Crop set $\mathcal{K}$ is partitioned into G mutually exclusive crop groups. The crop group set is denoted by $\mathcal{G} = \{1,...,G\}$. The crop subset defining group g is denoted by $\mathcal{K}_{(g)}$ .at Ignoring regime fixed costs, the considered profit function is given by

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\Pi_{it}^{MNL}(\mathbf{s};\boldsymbol{\pi},S_{it}) = S_{it}\mathbf{s}'\boldsymbol{\pi} - S_{it}C_{it}^{E}(\mathbf{s})$$

where function $C_{it}^{S}(\mathbf{s})$ defines the considered two-level MNL/entropic acreage management cost function. The functional form of this cost function is given by:

$$C_{it}^{E}(\mathbf{s}) = cst + \sum_{k\in\mathcal{K}} s_k b_{k,it}^{s} + a_i^{-1}\sum_{g\in\mathcal{G}} s_{(g)}\ln s_{(g)} + \sum_{g\in\mathcal{G}} a_{(g),i}^{-1} s_{(g)} \sum_{k\in\mathcal{K}(g)} s_{k|(g)}\ln s_{k|(g)}$$

where $s_{(g)} = \sum_{k\in\mathcal{K}(g)} s_k$ denotes the acreage share of group g and $s_{k|(g)} = s_k s_{(g)}^{-1}$ denotes the acreage share of crop k in the acreage of group g.[26] It is easily shown that:

$$S_{it}C_{it}^{E}(\mathbf{s}_{it}) = cst_{it} + C_{it}^{E}(\mathbf{S}_{it})$$

where:

$$C_{it}^{E}(\mathbf{S}_{it}) = \sum_{k\in\mathcal{K}} S_{k,it} b_{k,it}^{s} + \sum_{g\in\mathcal{G}} (a_i^{-1} - a_{(g),i}^{-1}) S_{(g),it}\ln S_{(g),it} + \sum_{g\in\mathcal{G}} a_{(g),i}^{-1} \sum_{k\in\mathcal{K}(g)} S_{k,it}\ln S_{k,it} .$$

Farm specific parameters a_i and $a_{(g),i}$ for $g\in\mathcal{G}$, which are collected in vector $\boldsymbol{\alpha}_i$, are positive. They determine to a large extent the crop acreage responses to changes in (expected) crop return levels. The larger these parameters are, the more the crop acreages of farmer *i* responds to changes in crop returns $\boldsymbol{\pi}$. Importantly, estimating the considered MEMC model and using the "statistical calibration" procedure proposed by Koutchadé *et al* (2018) yields estimates of parameter $\boldsymbol{\alpha}_i$, $\hat{\boldsymbol{\alpha}}_{i,NT}$, for each farm of the considered sample.

Let now assume that the analyst wants to calibrate a PMP model, problably describing production choices and handling policy instruments more complicated than those considered in the estimated model, on the considered farm sample (for a given year). Let further assume that the core of the PMP model used for farm *i* is given by profit function

$$\Pi_{it}^{PMP}(\mathbf{s};\boldsymbol{\pi},S_{it}) = \mathbf{S}'\boldsymbol{\pi} - C_{it}^{Q}(\mathbf{S})$$

where function $C_{it}^{Q}(\mathbf{S})$ defines the usual quadratic PMP term (possibly with a profit risk premium as in IFM-CAP). The functional form of this cost function is given by:

$$C_{it}^{Q}(\mathbf{S}) = cst + \mathbf{b}_{it}'\mathbf{S} + \mathbf{S}'\mathbf{A}_{it}\mathbf{S}/2$$

where matrix $\mathbf{A}$ is positive definite. Of course, we have

$$\frac{\partial}{\partial S_m} C_{it}^{Q}(\mathbf{S}) = b_{m,it} + \sum_{l\in\mathcal{K}} a_{ml,it} S_l$$

and

$$\frac{\partial^2}{\partial S_m \partial S_k} C_{it}^{Q}(\mathbf{S}) = a_{mk,it} .$$

[26] Following usual extension by continuity arguments we consider that $s_k\ln s_k = 0$ if $s_k = 0$ while $s_{(g)}\ln s_{(g)} = 0$ if $s_{(g)} = 0$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Parameters $a_{mk,it}$ determine to a large extent crop acreage choice responses to changes in crop return in the considered PMP model. The question we address is the following: how to use estimated parameters $\hat{\boldsymbol{\alpha}}_{i,NT}$ (and the information content of the considered dataset) for calibrating parameters $a_{mk,it}$.

4.3.2. Calibration procedure

The calibration procedure we propose here is fairly simple as it relies on the conditions stating that:

$$\frac{\partial^2}{\partial s_m \partial s_k} C^E_{it}(\mathbf{s}_{it}) = \frac{\partial^2}{\partial s_m \partial s_k} C^Q_{it}(\mathbf{s}_{it}) = a_{mk,it} \quad \text{for } (m,l) \in \mathcal{K} \times \mathcal{K}.$$

These conditions simply state the PMP farm models display crop acreage responses to economic incentives equal to those described by the corresponding farm models obtained from the estimated MEMC model in the neighbourhood of the observed choices.

It then suffices to compute the second order derivatives of entropic cost function $C^E_{it}(\mathbf{s})$ in crop acreages $\mathbf{s}$ and to use the equations given above. Let function $J:\mathcal{K} \rightarrow \mathcal{G}$ give the group $g \in \mathcal{G}$ to which belong crop $k \in \mathcal{K}$. It is easily shown that:

$$\frac{\partial^2}{\partial s_m^2} C^E_{it}(\mathbf{s}_{it}) = \alpha^{-1}_{(J(m)),i}(s^{-1}_{m,it} s_{(J(m)),it} - 1)s^{-1}_{(J(m)),it} + \alpha_i^{-1} s^{-1}_{(J(m)),it},$$

$$\frac{\partial^2}{\partial s_m \partial s_n} C^E_{it}(\mathbf{s}_{it}) = (\alpha_i^{-1} - \alpha^{-1}_{(J(m)),i}) s^{-1}_{(J(m)),it} \quad \text{if } J(m) = J(n)$$ [27]

and

$$\frac{\partial^2}{\partial s_m \partial s_n} C^E_{it}(\mathbf{s}_{it}) = 0 \quad \text{if } J(m) \neq J(n).$$ [28]

The last equality set yields the following "calibration formulae":

$$\hat{a}_{mk,it} = s^{-1}_{(J(m)),it} \times \begin{cases} 0 & \text{if } J(m) \neq J(k) \\ \hat{\alpha}_i^{-1} - \hat{\alpha}^{-1}_{(J(m)),i} & \text{if } m \neq k \text{ and } J(m) = J(k). \\ \hat{\alpha}_i^{-1} - \hat{\alpha}^{-1}_{(J(m)),i} + \hat{\alpha}^{-1}_{(J(m))} s_{(J(m)),it} s^{-1}_{m,it} & \text{if } m = k \end{cases}$$

The results presented here are easily extended to three-level MNL/entropic acreage management cost functions.

Note, in passing, that the elements of the Hessian matrix of the entropic cost function, that is to say terms $\frac{\partial^2}{\partial s_m \partial s_n} C^E_{it}(\mathbf{s}_{it})$, are non-negative if $\alpha_{(J(m)),i} \geq \alpha_i > 0$ (given that $s_{(J(m)),it} \geq 0$ and $s_{(J(m)),it} s^{-1}_{m,it} \geq 1$). This inequality conditions are satisfied for almost all sampled farms in our

[27] *I.e.*, if m and n belong to the same group.

[28] *I.e.*, if m and n belong to different groups.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

applications. They can be enforced in the estimation procedure of the MEMC model we consider here.

4.3.3. Estimating simplified versions of the MEMC models for calibrating micro-economic MP models

As discussed in Section 4.1, estimating the MEMC model considered in Section 2 is relatively complicated due to its considering that farmers' production regime choices are endogenous (*i.e.*, farmers' decisions to produce some crops and not to produce others are explicitly modelled). Of course, considering that farmers' production regime choices are exogenous greatly simplifies estimation of this model. It basically suffices to discard the regime choice sub-model from the MEMC model. This greatly alleviates the estimation burden because this eliminates the challenging computational issues raised by the integration of the regime choice probability functions (see Section 4.1.2). Yet, this also modifies the interpretation of the estimated model. The parameters of the MEMC model holding fixed farmer' production regime choices depict farmers' production choices – *i.e.*, crop acreages, input uses and yield levels – conditional on their regime choices. Consequently, the simulation model that can be obtained from the estimated MEMC model can only be used for investigating the effects of policies that do not impact farmers' production regime choices, that is to say their decisions to production certain crops or not. For other policies, simulation results can be affected by so-called selectivity biases.

The differences in the estimated models due to considering or not production regime choices may be limited, especially for datasets covering relatively homogenous regions. Our empirical investigations tend to confirm this hypothesis. In particular, considering regime choices or not when estimating the MEMC model mostly impact the estimates of parameter values related to minor crops. The decisions to produce these crops change depending on the economic conditions. Yet, since the acreages of minor crops are limited, the estimated models do differ much for the parameters related to the production choices related to major crops. Indeed, even if this point requires further investigation, estimating simplified versions of the MEMC model (*i.e.*, random parameter models considering that production regime choices are fixed) is expected to yield estimates of the "slope" parameters of the acreage management cost functions (*i.e.*, estimates of terms $\boldsymbol{\alpha}_i$) sufficiently reliable for calibrating MP models at the farm level.

Furthermore, considering versions of the MEMC model of section 2 ignoring variable input choices enables the analyst to obtain reliable estimates of crop acreage choice elasticities while overcoming a major issue occurring when using EU FADN data. These data do not report cost accounting data (*i.e.*, with input uses provided at the crop level) but rather standard accountancy data (*i.e.*, with input uses aggregated at the farm level). Two options are possible for accounting for variable input uses in the MEMC model with EU FADN data. First, one can estimate input uses at the crop level from input uses observed at the farm and use the obtained estimates for estimating the MEMC model. This is the option we explore in the MIND STEP project. Second, one can directly incorporate input allocation equations in the MEMC model (in place of systems of crop input use equations as in Section 2). Both options have merits and drawbacks. Nevertheless, the MEMC model can also be adapted for accommodating absence of any information on variable input uses. It basically suffices to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

discard input use (allocation) equations and to incorporate (expected) input use expenditures in the crop acreage choice equations (*e.g.*, Koutchadé *et al*, 2015). Such MEMC models enable to obtain reliable estimates of crop acreage choice elasticities (or the underlying slope variables) since farmers' crop acreage choice patterns are mostly identified by variations in crop prices and their effects on (expected) crop returns. Indeed, crop return levels, which largely determine farmers' crop acreage choices, are much more driven by crop yield and crop price levels than by crop input expenditure levels (at least after the mid-2000s and before 2022).

4.4. Incorporating new technologies in micro-simulation models

Accounting for adjustments in yield and variable input use levels is based on a menu of more or less intensive crop production technologies in most MP farm models. With the notable exception of FarmDyn, which considers discrete choices, most models, which are PMP models, consider pairs of crop and management intensities as separate activities. For instance, in these models wheat grown under (conventional) high-yielding practice and wheat organically grown are two activities to which farmers can devote cropland areas. The advantages of this modelling approach are twofold. First, PMP models accounting for menus of crop management practices can be developed as "extended" PMP models. Second, this approach avoids considering discrete choices in otherwise quadratic programming models. Yet, this approach rises specific calibration issues as activities considering the same crop grown under different practices are more substitutable than activities concerning different crops. Röhm and Dabbert (2023) proposed now well-known calibration devices for solving the calibration issues raised by these "extended" PMP models.

We propose here an alternative modelling approach for incorporating menus of crop production technologies in micro-simulation models. This approach considers crop production technology choices as discrete choices made at the crop level and simple "smoothing" devices are used for overcoming the issues raised by considering (crop,technologie) activities in MP models.

4.4.1. Problem setting

We assume that crop k can be grown based on set of technologies that is discrete and finite. Let define the following notations:

$\mathcal{K} = \{1,...,K\}$: crop set

$S_{k,i}$: acreage of crop k chosen by farmer i, with $\mathbf{S}_i = (S_{k,i} : k \in \mathcal{K})$

S_i : total cropland area of farmer i

$C_i(\mathbf{S})$: entropic acreage management cost function (MEMC models) or a quadratic PMP term (PMP models).

$\mathcal{M}_k = \{0,1,...,M_k\}$: set of available production technologies for crop k

$m_{k,i} \in \mathcal{M}_k$: production technology chosen for crop k by farmer i

$y_{k,i}^m$: yield level of crop k expected by farmer i if using technology m_k

$\mathbf{x}_{k,i}^m$: input use level for crop k expected by farmer i if using technology m_k

$p_{k,i}^m$: price paid for crop k to farmer i if using technology m_k

$\mathbf{w}_{k,i}$: crop k input prices paid by farmer i

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$r_{k,i}^m = p_{k,i}^m y_{k,i}^m - \mathbf{w}_{k,i}'\mathbf{x}_{k,i}^m$: return of crop k expected by farmer i if using technology m_k, with $\mathbf{r}_{k,i} = (r_{k,i}^m : m \in \mathcal{M}_k)$

$c_{k,i}^m$: implementation cost of technology m_k by farm i, with $\mathbf{c}_{k,i} = (c_{k,i}^m : m \in \mathcal{M}_k)$

Assuming that farmer i is economically rational implies that her/his crop technologies are given by:

$$m_{k,i}^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i}) = \arg\max_{m \in \mathcal{M}_k}\{r_{k,i}^m - c_{k,i}^m\} \quad \text{for } k \in \mathcal{K}$$

while her/his acreage choices are obtained as the solution to problem:

$$\mathbf{s}_i = \arg\max_{\mathbf{s} \geq \mathbf{0}}\left\{\sum_{k \in K} s_k p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i}) - C_i(\mathbf{s}) \text{ s.t. } \sum_{k \in K} s_k \leq s_i\right\}$$

where:

$$p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i}) = \max_{m \in \mathcal{M}_k}\{r_{k,i}^m - c_{k,i}^m\} = r_{k,i}^{m_k^o(\boldsymbol{\pi}_{k,i},\mathbf{c}_{k,i})} - c_{k,i}^{m_k^o(\boldsymbol{\pi}_{k,i},\mathbf{c}_{k,i})}$$

Of course, terms $m_{k,i}^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ and $p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ being defined by discrete choices does not raise severe computational issues for solving the crop acreage choice problem described above. Yet, using term $p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ allows considering crop k as a single activity in PMP models.

4.4.2. Smoothing devices

The issues raised by the severe discontinuities in $(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ of functions $m_{k,i}^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ and $p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ can be overcome by approximating these terms by functions that are smooth in $(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ (*e.g*, Bertsekas, 1996). If such approximations are of little interest in MP models assuming that farmers use a single technology for each crop, these can be useful in econometric models (*e.g.*, Devilliers *et al*, 2021) or, as discussed below, in MP models assuming that farmers can use several technologies per crop.

Term $p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i})$ can be suitably approximated by the so-called log-sum-exp term:

$$p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i}) \approx \tilde{p}_k(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k) = (\rho_k)^{-1}\ln\left(\sum_{m \in \mathcal{M}_k}\exp\{\rho_k(r_{k,i}^m - c_{k,i}^m)\}\right)$$

if tuning parameter $\rho_k > 0$ is sufficiently large.[29] An alternative solution is provided by

$$p_k^o(\mathbf{r}_{k,i},\mathbf{c}_{k,i}) \approx \sum_{m \in \mathcal{M}_k} w_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)(r_{k,i}^m - c_{k,i}^m)$$

[29] This follows from the well known property stating that $\lim_{\rho_k \to +\infty}\tilde{p}_k(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k) = \max_{m \in \mathcal{M}_k}\{r_{k,i}^m - c_{k,i}^m\}$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

where:

$$w_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)=\frac{\exp\{\rho_k(r_{k,i}^m-c_{k,i}^m)\}}{\sum_{n\in\mathcal{M}_k}\exp\{\rho_k(r_{k,i}^n-c_{k,i}^n)\}}\quad\text{for } m\in\mathcal{M}_k$$

also assuming that tuning parameter ρ_k is sufficiently large.[30]

This second solution approach can be especially useful in cases where term $\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)$ can be interpreted as the share of the acreage of crop *k* grown under technology *m*. In such cases, values of terms ρ_k and $\mathbf{c}_{k,i}$ can be adjusted for calibration purpose. The elements of $\mathbf{c}_{k,i}$ can be used as share shifters, while parameter ρ_k determines technology choice responsiveness to financial incentives given that $\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};0)=(1+M_k)^{-1}$. In such cases, however, use of terms $\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)$ requires an alternative background to that considered here.

5. PANEL SMOOTH TRANSITION REGRESSION MODEL OF DAIRY FARM PRODUCTION CHOICES[31]

5.1. Introduction

Farmers' capability to adjust their production choices in response to external events (e.g. changes in market conditions, climatic events, policy reforms), allows them to benefit from the events or to limit the profit loss induced by them. The closely related concepts of resilience, adaptive capacity and flexibility are indeed put forward as key elements of farms' economic sustainability in the face of increasing climate variability and volatility in agricultural markets (Reidsma et al., 2010; Robert et al., 2016). Although several economic studies have focused on farmers' long term adaptation to global changes, few studies have considered the adaptive capacity of farms in the short run. Yet, as pointed out by Darnhofer (2014), the ability

[30] This follows from another well known property stating that $\lim_{\rho_k\to+\infty}\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)=0$ if $r_{k,i}^m-c_{k,i}^m<r_{k,i}^n-c_{k,i}^n$ for $n\in\mathcal{M}_k/\{m\}$ while $\lim_{\rho_k\to+\infty}\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)=card(\arg\max_{m\in\mathcal{M}_k}\{r_{k,i}^m-c_{k,i}^m\})^{-1}$ if $m\in\arg\max_{m\in\mathcal{M}_k}\{r_{k,i}^m-c_{k,i}^m\}$. Of course, $\lim_{\rho_k\to+\infty}\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)=1$ if $r_{k,i}^m-c_{k,i}^m>r_{k,i}^n-c_{k,i}^n$ for $n\in\mathcal{M}_k/\{m\}$. Note also that $\sum_{n\in\mathcal{M}_k}\tilde{w}_k^n(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)=1$ and $\tilde{w}_k^n(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)\in(0,1)$ for $n\in\mathcal{M}_k$. Indeed, the functional form of term $\tilde{w}_k^m(\mathbf{r}_{k,i},\mathbf{c}_{k,i};\rho_k)$ is that of the probability functions the well-known Multinomial Logit discrete choice model. This probabilistic model was proposed D. McFadden in the 70s. It now is the workhorse of discrete choice econometric analysis.

[31] This entire section has been submitted, and is currently under review in the European Review of Agricultural Economics, under the name Heterogeneous farmers' responses to price variations: Identifying dairy farms flexibility using a panel smooth transition regression approach". The authors are Elodie Letort and Fabienne Femenia.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

of farmers to temporarily reallocate their resources when a disruption occurs, such as a sudden price decrease or a short drought, is also an important issue.

The short-run adaptive capacity of farms, which we refer to here as "farm flexibility", is of particular interest in the case of dairy farming for several reasons. First, milk and feed concentrate prices tend to vary a lot from one year to the other, implying that dairy farms have to cope with significant variations in both output and input prices. Second, although dairy production can be impacted by various external events, the impact of these events depends on the management strategy of the farms, in particular their feeding strategy. Maize silage-based production systems, which are highly represented in our sample, are both dependent on imported (concentrate) feed used to supplement maize and on climatic hazards which have significant impacts on yields. Yet, as shown in the technical livestock production literature (Peyraud et al., 2010), dairy farms do in fact have some degree of flexibility to adjust their feeding strategy in the short run in order to be more resilient to price and climate shocks, by adopting mixed feeding systems, diversifying their pasture, and using concentrates when necessary.

While flexibility appears as a key aspect of the economic performance of farms, economic studies of farm production decisions generally find low elasticities of input uses and acreage choices in the short run. Most of these studies actually focus on crop production decisions. Böcker and Finger (2017), for instance, conduct a meta-analysis of the price elasticities of pesticides demand and conclude to the inelasticity of this demand with elasticity values significantly lower than 1 in absolute terms. Crop acreages are also found to be inelastic to crop price changes by Carpentier and Letort (2012) with elasticities ranging from 0.1 to 0.3. Only few economic studies have been conducted on the price responses of input and acreage choices in livestock farms. One exception is Suh and Moss (2017) who also find input use to be price inelastic in dairy farms with an elasticity of demand for silage maize equal to -0.23.

This apparent rigidity of farm production decisions in the short run should however be considered in the light of the heterogeneity of farm behaviors. Price responses revealed in the above-mentioned studies indeed correspond to average behaviors, common to all farms for a given specialization in a given region. Yet these behaviors can differ from one farm to the other due to several (often unobserved) factors.

Our objective in this paper is to identify the heterogeneity in the flexibility of dairy farms based on their observed short run responses to input and output prices. Recent empirical works dealing with dairy farm heterogeneity have essentially focused on the heterogeneity in production technologies among farms. The most standard approach used in these studies is to make an *ex ante* classification of farms based on *a priori* knowledge and assumptions about differences in technologies, and, in a second step, to estimate parameters characterizing the production technology for each of group of farms (see e.g., Kumbhakar 2009). More elaborated approaches have also been proposed to simultaneously estimate production technologies and the probabilities for each farm to belong to a technological group (Alvarez and del Corral 2010, Sauer and Paul 2013, Renner et al. 2021). What all these methods have is common is that the identification of different technological groups is based on the use of observed criteria aimed at characterizing farm production practices. The approach we propose here goes further in the analysis of farms' and farmers' production heterogeneity. It allows revealing the heterogeneity of farms based on econometric estimations, without having to a priori define *ad hoc* criteria characterizing this heterogeneity. Our approach allows

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

for a differentiation of farms based on their observed responses to price changes, and enables an *ex-post* analysis for identifying potential specific practices or farm structures characterizing different groups of farms emerging from estimation results.

Very few works have in fact focused on the heterogeneity of farmers' responses to market prices. Koutchadé et al. (2018) is one exception. These authors use a random parameter approach to account for the unobserved heterogeneity of farmers' behaviors and show that price elasticities of farmers' yield, input use and acreage choices do in fact display a significant heterogeneity among farms, even in a sample of relatively homogeneous farm specialized in crop production in a small geographical region. Our approach differs from that of Koutchadé et al. (2018), who address the heterogeneity in farms' behavior from a statistical view point, in that we aim at providing an economic explanation to the relative rigidity or flexibility of farms. A lack of flexibility in the responses of farm production decisions to market prices can in fact be due to the existence of adjustment costs related to the rigidity of quasi-fixed inputs or to additional administrative or transaction costs incurred by the farm when adjusting its variable input levels.

Different methodological frameworks, which are reviewed in the first section of this article, have actually been developed in the economic literature to account for the impacts of adjustment costs on farm production decisions. However, these frameworks essentially aim at analyzing farm behaviors in the medium or long term by focusing on the adjustment costs of capital (e.g. Gardebroek and Oude Lansink, 2004; Pietola and Myers, 2000) but do not consider the adjustment costs incurred by farms when adjusting their variable inputs in the short run. The approach we propose here relies on the works of Önel (2018a and 2018b), who proposes a model of factor demand that implicitly accounts for the existence of adjustment costs in the U.S. industrial sector. This approach recognizes that, following a change in input prices, firms have to adjust their quantity of variable inputs, which may be costly due to the existence of adjustment costs. Önel argues that, in the presence of such adjustment costs, the adjustment of input quantities will be different depending on the size of the price change observed in the market. His empirical model, based on the threshold regression model of Hansen (2000), allows for the representation of firm behaviors in two regimes of inputs adjustment, one regime corresponding to their behavior in the face of small variations in input prices and the other to their behavior in the face of large variations. There is thus a switch between the two regimes of input adjustment, depending on whether the observed input price variations are above or below a certain threshold. This estimated threshold corresponds to the level of price change leading to different firm behaviors in terms of inputs adjustment. An industrial sector in which firms are able to rapidly adjust their production decisions in the face of a price shock is assumed to have lower adjustment costs.

Our main contribution compared to the works of Önel is to consider individual threshold levels of price variation in order to account for the heterogeneity of adjustment costs among farm. We consider that, in the face of a price shock, farmers' ability to adapt their production decisions reflect the level of adjustment costs they face and their degree of flexibility, and that this flexibility can be characterized by different ways of switching between regimes of input uses that are specific to each farmer. From an empirical viewpoint, we rely on the panel smooth transition regression (PSTR) model proposed and developed by Gonzales et al. (2005) and Fok et al. (2005) as an extension of Hansen (2000)'s threshold regression model allowing for a smooth transition between two extreme regimes. The unobserved heterogeneity of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

farmers in their adaptation to price variations is captured in our model by considering farm specific random parameters in the transition function representing the way farmers switch from one regime to the other. Relying on farm specific transition function was also proposed by Fok et al. (2005). However, these authors rely on a simulated maximum likelihood approach to estimate their model. This approach is appropriate in the empirical case they consider, but is not suitable and may generate convergence issues, for the type of panel data generally used to estimate agricultural production models, such as those available to us, which have a large individual dimension but a rather limited time dimension. We thus rely on a different approach, a stochastic version of the expectation maximization (EM) algorithms originally proposed by Dempster et al. (1977), to estimate our model, which is another original contribution of our work.

Our approach is applied to a panel of dairy farms located in the west of France over the period 2007-2018. Our empirical results show differences in the feeding strategy adopted by these farms depending on the magnitude of price variations they face. They actually tend to substitute three sources of animal feed (feed concentrate, fodder maize and grassland) in response relatively small variations in market prices but become less flexible in their acreage adjustment, and thus essentially adjust their quantities of feed concentrates, in response to larger price variations. Our estimation results also reveal a significant heterogeneity among farms regarding the level of price variation up to which they keep adjusting their feeding strategy in a flexible way. An *ex post* analysis of our results allows highlighting some specific features of the most flexible farms, which appear to be more autonomous financially and more self-sufficient in terms of animal feed.

The rest of the paper is organized as follows. A first section is devoted to a literature review on the structure of adjustment costs in agricultural sectors in presented. Our empirical model of dairy farm input use decisions allowing to identify their flexibility at individual levels is presented in a second section. The estimation strategy used to estimate this model is described in a third section. Our empirical results are presented and discussed in a fourth section. Finally we conclude.

5.2. Literature review on farm adjustment costs

Different methodological frameworks have been developed in the economic literature to account for the impacts of adjustment costs on firm production decisions.

The first category of approaches focuses on firm investment decisions using dynamic models. In the agricultural economics literature, the existence of adjustment costs prevent farms from immediately adjusting their capital or labor stock, hence their output level and variable input quantities (e.g. Gardebroek and Oude Lansink, 2004; Pietola and Myers, 2000). The rigidity of capital is confirmed in livestock sectors, especially for pig farms, which have particularly been studied given the important investments they require to expend (Pietola and Myers, 2000; Gardebroek and Oude Lansink, 2004; Boetel et al. 2007). The growing literature on investment has resulted in the improvement of theoretical and empirical models of farm investment decisions. As a result, the structure of adjustment costs is increasingly discussed. Lansink and Gardebroek (2004) notably propose to relax the assumption usually made that all farms exhibit the same structure of capital adjustment costs. Their empirical application confirms the heterogeneity of adjustment costs, which allows them to identify groups of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

farms with similar structure of adjustment costs. These models of farm investment are however designed to analyze long-term behaviors of farms, which is not our purpose here. Our main concern is to obtain information on short-term farm production decisions that do not require investment or technological change.

Some micro-econometric models of short-term crop production decisions (Carpentier and Letort, 2012 and 2014; Koutchadé et al., 2018) rely on a concept borrowed from the Positive Mathematical Programming (PMP) literature (Howitt, 1995) and introduce implicit adjustment costs of acreage in their objective function in order to account for crop diversification motives in acreage choices. This type of approach implicitly considers that farmers are restricted in their acreage choices by agronomic and technical constraints. For example, an expected work peak, a lack of machines and/or a greater risk of pests prevent farmers to specialize in a single-crop farming. Koutchade et al. (2018) make a significant improvement to this modelling framework by considering farm-specific parameters, especially in the implicit acreage adjustment cost function of the model. A random parameters specification actually allow them to account for heterogeneous responses of crop producers to economic drivers. Their empirical application confirms the heterogeneity of acreage adjustments costs in cereal farming, and shows that ignoring the variability in the considered farmers' responses to economic incentives may lead to poor estimation of production decisions. However, this approach does not directly take into account the existence of costs induced by the adjustment of variable inputs, such as fertilizers or pesticides, which may be more or less important depending on the level of adjustment required, as are the costs of adjusting capital. Moreover, although all the parameters of their model are farm-specific, they implicitly consider that differences in input decision behavior among farmers can be attributed to the heterogeneity of their production technology, not by the heterogeneity of their adjustment cost structure.

A third modelling framework, recently proposed by Önel (2018a and 2018b), is of particular interest for our purpose. This approach recognizes that, following a variation in input price, firms have to adjust their quantity of inputs, which may be costly, and that, in the presence of adjustment costs, the adjustment of input quantities will differ depending on the magnitude of the price variation. Two cases are in fact possible, depending on the structure of adjustment costs faced by the firm. On the one hand, if adjustment costs are convex, these costs increase with the adjustment of input quantities, implying larger price elasticities of input for smaller price variations. On the other hand, if adjustment costs are non-convex, they are non-increasing with the adjustment of input quantities, implying larger price elasticities of input for higher price variations: in that case, a small adjustment of input quantities, following a small variation in market price, is costly compared to the profit gained from this small adjustment. Firms will thus adjust more their inputs for large price variations. Önel (2018a) uses a threshold regression model (Hansen, 2000) of input demand to implicitly account for these adjustment costs. In this model, the price elasticities of input are allowed to vary depending on observed input price variations. More precisely, the price parameters of the model take one value below a certain threshold level of price variation and another value above this threshold. For our purpose, the advantage of this approach is that adjustment costs concern quasi-fixed inputs, but also variable input uses, making this approach adapted to the analysis of short-term behaviors. Önel proposes this framework with the primary objective of highlighting the non-linearity, and potential non-convexity, of input adjustment costs. At the same time, his empirical application allows him to compare

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

adjustment costs structure between industrial sectors of the United States. Note that he assumes that the structure of adjustment costs are identical for all firms within each industrial sector.

Here we propose an approach that builds on Önel's framework and introduce some specific features in our model in order to better represent the heterogeneity of farms flexibility in terms of input uses. In particular, we do not consider only two possible regimes of input adjustment in response to price variations, common to all farms, but a continuum of regimes specific to each farm. As explained in the next section, this is achieved by relying on a PSTR model in which we introduce random parameters to represent farm-specific transition functions.

5.3. Modelling framework

Our theoretical framework, presented in the first sub-section, builds upon Chambers and Just (1989)'s farm profit maximization problem in the presence of fixed allocable inputs. It provides a reference point for the development of the empirical threshold model described in the second sub-section.

5.3.1. Model of livestock farms' production decisions

We focus here on the short-term production decisions of dairy farmers, who allocate one fixed input, land, among three feeding sources produced on the farm (fodder maize, grassland and cereals) that are complemented by feed concentrate to produce milk. Since we are dealing with short-run production decisions, we assume their herd size to be fixed. As most works focusing on heterogeneity in farm production behaviours (Alvarez and del Corral 2010, Sauer and Paul 2013, Koutchadé et al. 2018, Renner et al. 2021), we assume that farmers are risk neutral.

Our modelling framework relies on the farm profit maximization problem in the presence of fixed allocable inputs proposed by Chambers and Just (1989), and generalized by Fezzi and Bateman (2011) to fit the case of dairy farms by allowing the number of possible land allocations to be different from the number of possible farm outputs. As shown in Fezzi and Bateman (2011), by specifying the farm profit per area as a normalized quadratic function, optimal input use and acreage shares equations can be expressed as a system of reduced form equations as:

$$y_{it}^{j} = a_{0i}^{j} + \mathbf{z}_{it}'\boldsymbol{\alpha}_{1}^{j} + \mathbf{x}_{it}'\boldsymbol{\alpha}_{2}^{j} + e_{it}^{j} \tag{1}$$

Subscripts i and t respectively denote the cross-sectional and time dimensions of our panel data and superscript j belongs to $\mathcal{J}$, the set of livestock farmers' acreages and input uses choices we consider, namely cereals, grassland and fodder maize acreages and feed concentrate purchases. The vector of dependent variable $\mathbf{y}_{it} \equiv (y_{it}^{j}, j \in \mathcal{J})$ contains acreage

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

shares[32] and feed concentrate quantities. $\mathbf{x}_{it} \equiv (\mathbf{p}_{it} / w_{nit}, \mathbf{w}_{it} / w_{n,it})$ contains a set of output and input prices normalized by the price of one input *n* (pesticides in our case). The prices considered in the acreage shares equation are those observed at the time farmers take their land allocation decisions, while those entering the feed concentrate equation are those observed at the time of feed concentrate purchases. Farmers may actually adjust their use of concentrates after observing the yields cereals, fodder maize and grass produced at the farm. Other observable factors that can potentially have an impact on farmers' production decisions, such as market prices not included in $\mathbf{x}_{it}$ or weather conditions, are included in $\mathbf{z}_{it}$. The parameters included in vector $\boldsymbol{\alpha}_1^j$, respectively vector $\boldsymbol{\alpha}_2^j$, capture the effects of $\mathbf{z}_{it}$, respectively $\mathbf{x}_{it}$, on y_{it}^j. In model (1), these effects are assumed to be common to all farms and farmers. The a_{0i}^j additive term is a random farm-specific parameter aimed at capturing the effects of unobserved factors, such as farmers' skills or farms natural endowments, on y_{it}^j. Finally, e_{it}^j is a stochastic error term.

5.3.2. Threshold model of livestock farms' production decisions

The linear model of agricultural production decisions presented above does not account for potential costs incurred by farmers when adjusting their inputs in response to price variations. In fact, it assumes homogenous reactions of farmers to any level of price change. Yet, a change in input quantities following a change in input or output price might be costly for farmers. There may exist direct adjustment costs such as transportation costs, training costs, costs arising from changes in contracts, etc. Adjustment costs may also be caused by the rigidity of quasi-fixed factors, since variable input use is tied to the amounts of capital equipment, the structure of labour and the allocation of land. For instance, important decrease in the use of concentrates to feed animals requires producing more fodder crops or encouraging grazing pasture. It involves a reorganization of the farm and a new land allocation in order to convert cropland to pasture and/or fodder crops. Farmers may also need new machinery and more workers to produce on-farm animal feed. Farmers might thus significantly adjust their animal feeding strategy for large price shocks only (non-convex adjustment costs), or, on the contrary, be more responsive to smaller price shocks that imply smaller adjustments of animal feeding (convex adjustment costs).

As in Önel (2018), we modify our empirical model to implicitly account for the existence of adjustment costs that cause farmers' responses to variations in input and output prices to differ depending on the magnitude of those variations. We thus propose a modelling approach based on a threshold regression model in which parameters associated to price variables can vary according to a regime-switching mechanism that depends on a transition variable. We use the absolute variation in an input price to output price ratio compared to the previous year as transition variable. This transition variable actually allows us to represent the price signal perceived by farmers, who take their production decisions based on the evolution of both input and output prices. This variable indeed takes on small values if input

[32] Only fodder maize and grassland acreage shares are included in the model, the cereal acreage share being redundant since it can be computed by difference.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

and output prices do not vary much from year to year or if they vary in the same direction, in which case the price ratio is stable. Finally, while Önel's model relies on Hansen's threshold regression approach, we use a PSTR model based on the works of Gonzales et al. (2005) and Fok et al. (2005) to allow for individual threshold parameters.

The threshold version of our model can be written as:

$$y_{it}^{j} = a_{0i}^{j} + \mathbf{z}'_{it}\boldsymbol{\alpha}_{1}^{j} + \mathbf{x}'_{it}\boldsymbol{\beta}_{1}^{j} + G_{it}\mathbf{x}'_{it}\boldsymbol{\beta}_{2}^{j} + e_{it}^{j} \qquad (2a)$$

where $G_{it}\mathbf{x}_{it}$ is a vector containing the product of each component of $\mathbf{x}_{it}$ with G_{it}, the value taken by function G for farmer i in year t. G is a transition function, normalized to be bounded between 0 and 1. As proposed by Gonzales et al. (2005), introducing this transition function in the model allows representing a smooth transition between two regimes of responses to price variations, contrary to Hansen (2000)'s threshold regression approach that only allows for an abrupt transition between the two regimes. G is a continuous function of an observable transition variable q_{it} and depends on farm-specific random parameters, g_i and c_i, that respectively reflect the speed and the threshold of transition.

This transition function has a logistic form:

$$G(q_{it};g_i,c_i) = \left(1 + \exp[-\ g_i(q_{it} -\ c_i)]\right)^{-1} \qquad (2b)$$

With $g_i > 0$. This smooth transition regression approach may be interpreted in two distinct ways. First, we can consider that there are two extremes regimes associated with the two extremes values of the transition function: $G_{it} = 0$ and $G_{it} = 1$, and that farmers progressively move from one regime into another. The response of y^j to changes in prices contained in $\mathbf{x}$ is given by $\boldsymbol{\beta}_1^j$ in the first extreme regime and by $\boldsymbol{\beta}_1^j + \boldsymbol{\beta}_2^j$ in the second extreme regime. Second, this might be regarded as an infinity of regimes and possible values for the price response parameter are $\boldsymbol{\beta}_1^j + \boldsymbol{\beta}_2^j G_{it}$, depending on the value of q_{it}.

Our model differs from the model originally proposed by Gonzales et al (2005) in that we consider individual-specific parameters in the transition function. To our knowledge, Fok et al. (2005) is the only paper that considers individual threshold parameters. These authors examine the existence of common nonlinear business cycle features in 19 US manufacturing sectors based on a multi-level smooth transition model. However, their approach is best suited to time series panels of data (i.e., data with a large temporal dimension and a small cross-sectional dimension), whereas our data contain observations for a large number of farms over a relatively short period of time, as is typically the case for samples of farm accounting data used to estimate farmers' behavior.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

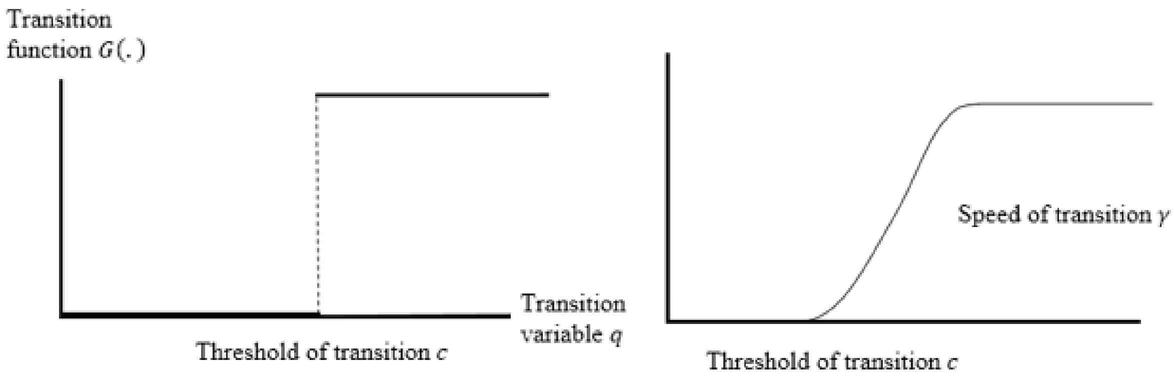

Figure 7. Approaches of Hansen and Gonzales: standard and smooth threshold regression model

Figures 7 and 8 graphically illustrate the differences between our approach and those of Hansen (2000) and Gonzales et al (2005). These graphs report the value of the transition function G according to the value of the transition variable q. For each approach, farmers are in the first regime when the transition function is equal to 0 and in the second regime when the transition function is equal to 1. The difference between the two approaches lies in the way farmers switch from the first to the second regime.

In Hansen (2000)'s model illustrated by the left-hand graph on Figure 7, the transition between the two regimes consists in a jump at a threshold level ,c, which is the same for all farmers. Gonzales et al's model, illustrated by the right-hand side graph in Figure 7, allows for a smooth transition between the two regimes, the speed of this transition being characterized by a parameter g also common to all farmers. Our approach, illustrated by the graphs in Figure 8, simultaneously allows for individual threshold and speed of transition levels.

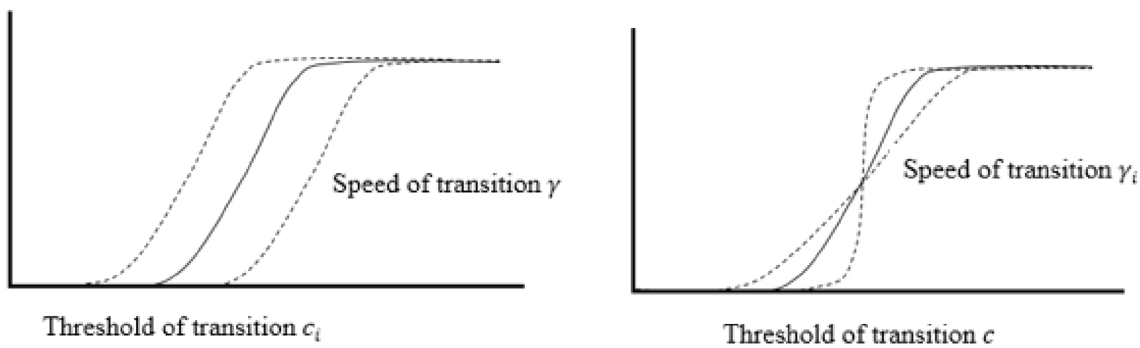

Figure 8. Our approach: individual smooth threshold regression model

Our model has the advantage to contain all others models. If the threshold of transition is not significantly different between farmers and the speed of transition tends toward infinite, our model reduces to the threshold model of Hansen (2000). If the threshold and the speed of transition are not significantly different between farmers, we obtain the smooth threshold model of Gonzales et al. Finally, if the speed of transition tends towards zero, our model reduces to a linear random parameters model.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

5.3.3. Heterogeneity of farm adjustment costs and flexibility

We explain here how our proposed modelling framework can be used to characterize adjustments costs and determine the level of flexibility of each farm.

The starting point of our analysis is the characterization of the two extreme regimes in our model. In fact, the analysis of farmers' behavior depends on the values of parameters that characterize each regime. As a reminder, the two extreme regimes are similar for all farmers. The first regime corresponds to a context of very small price variations (or, at least, to a context where both input and output prices vary in the way) and the transition function is equal to 0. The second extreme regime corresponds to a context of large variations in either output or input price (or a context where input and output prices vary, even moderately, in opposite directions) and the transition function is equal to 1. Of course, the parameters characterizing farmers' behaviors in these regimes are obtained from the estimation of the model. Two cases are possible. In the first case, farmers are more responsive to small than to large price variations. That case can arise when adjustment costs are convex and increase with the adjustment of input quantities that is required in response to a price variation. In a second case, farmers are more responsive to large than to small price variations. That case can arise if adjustment costs are non-convex and a small adjustment of input quantities, following a small price variation, is costly compared to the benefits generated by the input adjustment. Input price elasticities are higher in the first than in the second extreme regime, and conversely, in the second case input elasticities are higher in the second regime.

We can then analyze how each farmer switches from one regime to the other by comparing the two farm-specific parameters of the transition function in our model, namely the threshold c_i and speed of transition level g_i. Our interpretation is that a farmer tends to switch smoothly and slowly from one regime to the other (small g_i) when this switch requires an investment in capital or labour, or a change in production technology. In fact, as in dynamic models of investment decisions, this lack of flexibility may be explained by the rigidity of quasi-fixed inputs. Our interpretation of the threshold level c_i. is slightly different. We use this parameter to characterize the capacity of farmers to adapt their short-term production decisions. Consider for instance a case where the transition from one regime to the other is quite abrupt and quick (high g_i) for all farmers. In a case of convex structure of cost (described by a higher input use elasticity in the first regime), a farmer characterized by a high threshold level probably faces less adjustments costs than the others. In fact, she/he is able to adjust her/his short-term production decision for a wider range of price variations. She/he will switch to the second regime only when the price variations will be too large and will induce too important adjustment costs. In this case, several factors can explain the flexibility/rigidity of farmers, depending on the farming system (share of grassland, farm-produced feed, etc…), the structural features of the farm (total area, fragmented plots, etc…) or the managerial ability of the farmer.

5.4. Estimation strategy

This section presents the distributional assumptions and the approach we use to estimate the PSTR model defined by equations (2a) and (2b).

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement N° 817566.

5.4.1. Distributional assumptions

Each equation of our model of dairy farmers' production choices comprises fixed parameters $\boldsymbol{\beta}_1^j$, $\boldsymbol{\beta}_2^j$ and $\boldsymbol{\alpha}_2^j$, and two types of random components : (*i*) random parameters that includes the additive farmer-specific effects a_{0i}^j and the parameters of the transition function, g_i and c_i and (*ii*) the error term of the model, e_{it}^j. Let vector $\boldsymbol{\delta}_i \equiv (\boldsymbol{\alpha}_{0i}, g_i, c_i)$, with $\boldsymbol{\alpha}_{0i} \equiv (a_{0i}^j, j \in \mathcal{J})$, collect the model random parameter and vector $\boldsymbol{\varepsilon}_{it} \equiv (e_{it}^j, j \in \mathcal{J})$ collect the error terms. We assume that the random parameters follow a normal distribution with $\boldsymbol{\delta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$. This probability distribution describes the distribution of the random parameters across the farmers' population represented in our considered sample. We do not impose any restriction on the structure of $\boldsymbol{\Omega}$ and hence allow all farmer-specific effects (including the parameters of the transition function) to be correlated between them and across equation. This notably allows capturing the potential correlation between the production decisions of each farmer, which could for instance be attributed specific skills of farmers or to natural endowments of farms. The error term vector is assumed to be normally distributed with $\boldsymbol{\varepsilon}_{it} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. We assume the covariance matrix $\boldsymbol{\Psi}$ to be diagonal, which implies that the error terms of the model are independently distributed across time. This assumption is not too restrictive here since price and climatic shocks, which are the main elements that could potentially simultaneously impact all farmers production decisions in the short run, are captured by the effects of exogenous variables in $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ Finally, we assume that random parameter vector $\boldsymbol{\delta}_i$, error term vector $\boldsymbol{\varepsilon}_{it}$, price variables included in $\mathbf{x}_{it}$ and control variables included in $\mathbf{z}_{it}$ are mutually independent and that $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are strictly exogenous with respect to these error terms. These are standard assumptions in short panel data econometric models of farmers' production choices (see e.g., Koutchadé et al, 2018).

5.4.2. Estimation approach

The parameters we seek to estimate comprise the price effects, $\boldsymbol{\beta}_1^j$ and $\boldsymbol{\beta}_2^j$, and control variables effects, $\boldsymbol{\alpha}_2^j$, in each equation. These fixed parameters are collected in vector $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}_1^j, \boldsymbol{\beta}_2^j, \boldsymbol{\alpha}_2^j, j \in \mathcal{J})$. We also seek to estimate parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, characterizing the distribution of the random parameters, and the covariance matrix of random terms, $\boldsymbol{\Psi}$.

Our model being fully parametric, we rely on a maximum likelihood (SML) approach for its estimation. The sample log-likelihood is equal to a sum of log-likelihoods associated to each individual farm: $\ln L = \sum_i \ln \ell_i$. The individual likelihoods, can be expressed as:

$$\ell_i(\boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Omega}) = \int_{\boldsymbol{\delta}} f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\Psi}) g(\boldsymbol{\delta}; \boldsymbol{\mu}, \boldsymbol{\Omega}) d\boldsymbol{\delta} \tag{3}$$

$f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\Psi})$ denotes the probability density function of the observed sequence of production choices of farmer *i*, $\mathbf{y}_i$, conditional on exogenous variables, $\mathbf{x}_i$ and $\mathbf{z}_i$, and on individual random parameters, $\boldsymbol{\delta}_i$. $g(\boldsymbol{\delta}; \boldsymbol{\mu}, \boldsymbol{\Omega})$ denotes the probability density function of the random parameter vector, $\boldsymbol{\delta}_i$.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Maximizing the sample likelihood would involve the computation of as many two-dimensional integrals as the number of farms in our sample. Econometricians generally rely on Simulated Maximum Likelihood (SML) approaches to solve such optimization problems. This is the choice made by Fok et al. (2005) for the estimation of their model, which is comparable to ours from that point of view. The maximization of the simulated likelihood is however further complicated by the nonlinear form of the transition function G_{it} which depends on random parameters $\boldsymbol{\delta}_i$ and enters the explanatory variables of the model. To overcome this issue, Fok et al. (2005) use a two-step iterative procedure involving, in a first step, the maximization of the sample simulated likelihood for given values of the transition function parameters, and, in a second step, the solve of a numerical optimization program to find transition function parameters. This two-step procedure is however quite involving and its convergence not guaranteed, especially in our case where the individual dimension of our panel dataset is much larger than its time dimension (i.e. our sample contains a large number of individual farms for which we few observations over time). We do not use this two-step procedure here but instead rely on a Stochastic Approximate expectation-maximization (SAEM) algorithm which is a specific type of Monte Carlo expectation-maximization (MCEM) algorithms (Lavielle, 2014). MCEM algorithms are frequently used by statisticians when faced to complex likelihood maximization and allow the computation of estimators that are asymptotically equivalent to ML estimators. Technical details on these algorithms, on the SAEM algorithm in particular, and on their use for the estimation of micro-econometric random parameters of agricultural production choices can be found in Koutchadé et al. (2018 and 2021). Here the R software was used to implement the SAEM algorithm and estimate our model. The codes are available from the authors upon request.

5.5. Empirical illustration: the case of French dairy farms

5.5.1. Data

Our model is estimated on a data sample, provided by a French farm accounting agency. This unbalanced panel dataset contains 5,412 observations of 714 dairy farms located in the West of France and observed between 2007 and 2018. The three dependent variables of our model, i.e. the quantities of concentrates, the share of fodder maize acreage and the share of grassland acreage, as well as the milk and feed concentrate prices observed by farm and by year are observed in this database. The other input and output prices used as explanatory variables in the model (fertilizer, pesticide and cereal prices) are price indices provided by the French Department of Agriculture. We also use make use of data provided by the French national meteorological service (*Météo France*) to build climate indicators used as control variables in the model. Although our sample covers a relatively small area, climate conditions are in fact likely to have an impact on dairy farmers' production decisions, especially for maize, which is very sensitive to water and heat stress during the spring and summer periods[33]. We construct two cumulative rainfall indicators: one for the months of June and

[33] Climatic conditions (precipitation and temperature) may have important impacts on the yield and energy content of grass and maize fodder. In face of these climatic events, farmers will adjust the amount of concentrates in order to maintain a balanced feed ration.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

July and one for the months of August and September. A third indicator is constructed by summing the number of days during which the temperature exceeded 29 degrees Celsius, which corresponds to the maximum temperature beyond which maize development is slowed and its growth reduced (Girardin 1998).

Table 11 bellow reports some descriptive statistics of these variables. Our sample appears to be relatively homogenous in terms of production system, since all farms produce fodder maize, grassland and cereals and make use of concentrates to feed their animals. The farms in our sample are located in the French territorial division *Ille-et-Vilaine* in the Brittany region, which is the first French dairy region, producing 20% of the national production. Dairy farms in this area mostly rely intensive forage systems, characterized by fairly high levels of milk per cow, moderate use of pasture, and rather extensive use of supplementary feeding based on concentrates.

Table 11. Descriptive statistics of the sample

	Sample Average	Standard Deviation
Quantity of feed concentrate used (ton/cow)	1.20	0.56
Fodder maize acreage share	0.28	0.08
Grassland acreage share	0.45	0.12
Milk price (€/liter)	0.32	0.22
Feed concentrate price (€/kg)	0.27	0.06
Cereal price index (1 in 2015)	1.03	0.21
Fertilizer price index (1 in 2015)	1.00	0.19
Total available land (ha)	67.18	23.68
Quantity of milk produced (1000 l/cow)	7.09	1.25
Animal density (cow/ha)	1.12	0.24
Rainfall at flowering (in mm)	111.77	41.03
Rainfall at maturation (in mm)	106.49	37.48

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Heat (number of days)	1.69	0.99

As previously mentioned, regarding the threshold variable used in the transition function of the model, we do exactly follow Önel (2018a and 2018b) who uses the variation, in absolute, of input prices compared to the previous period. We instead use the variation in the input (feed concentrate) to output (milk) price ratio. The evolution of this threshold transition variable is represented on Figure 9. The main advantage of this transition variable is to better characterize the economic context faced by dairy farms. Farmers can actually face four main types of economic context depending on the evolution of animal feed and milk prices, two of these contexts lead to a stable price ratio: if the milk and feed concentrates prices do not vary a lot or if these prices evolve in the same direction (an increase in both prices, or a decrease in both prices). In these contexts, farmers' behaviors are close to the those represented in the first extreme regime. On the other and, the price ratio will be particularly high in absolute terms if only one the two prices increases or decreases significantly while the other price remains stable or if the two prices evolve in opposite directions. A sharp increase in milk price and/or decrease in concentrate price creates an economic situation particularly favorable to dairy farmers (this was the case in 2014). A sharp decrease in milk price and or increase in concentrate price leads to a bad economic situation (this was the case in 2009, 2013 and 2016). In those type of economic contexts, farmers' behaviors are close to those represented in the second extreme regime.

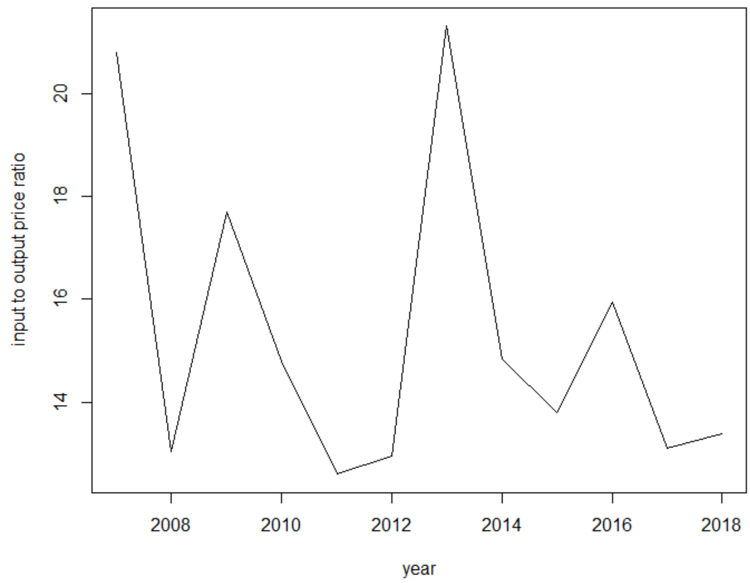

Figure 9. Evolution of the threshold transition variable

5.5.2. Estimation results

Parameter estimates of the dairy farm production decision model defined by equation system (2a) and transition function (2b) are reported in Table 12.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

The parameters of the input uses model are generally significantly estimated and lie in expected ranges. Our results show a negative impact of the price of cereals on the use of feed concentrate. Two mechanisms can indeed explain this effect depending on farmers use of their cereal production. On the one hand, if cereals are cash crops intended to be sold on the market, an increase in cereal price will incent farmers to increase their production of cereals and thus reduce their fodder maize and grassland acreages. In that case, an increase in feed concentrates will allow farmers to compensate the loss of forage area. The same mechanism besides explain the positive impact of fertilizer price on forage area (through its negative impact on cereal acreage) and negative impact on feed concentrate. On the other hand, if the cereal production is used by farmers to feed animals on their farms, an increase in the price of cereals will change the comparative advantage of feed concentrate compared to cereals in favor of concentrates.

Table 12. Parameters estimates

Input uses equation system			
	Quantity of feed concentrates	Share of fodder maize acreage	Share of grassland acreage
Distribution of individual farm effects α_{0i}			
Mean	-0.54**	0.53**	0.37**
Standard deviation	0.40**	0.03**	0.10**
Effects of price and climate variables α_1			
Fertilizer price	-0.43*	0.15**	0.04
Cereal price	4.14**	-0.14**	-0.21**
Rainfall at flowering	0.37**	-0.15**	0.09**
Rainfall at maturation	0.28*	-0.04	0.08**
Heat	0.09*	-0.04**	0.01
Effects of milk and concentrate prices in the first extreme regime β_1			
Feed price	-1.41**	0.08**	0.03
Milk price	5.14**	-0.72**	0.22**
Changes in the effects of milk and concentrate price in the second extreme regime β_2			
Feed price	-0.52**	-0.10**	-0.02
Milk price	0.37**	0.06**	0.03

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Transition function G	
Distribution of threshold parameters c_i	
Mean	0.11**
Standard deviation	0.02**
Distribution of the speed of transition parameter γ_i	
Mean	0.42**
Standard deviation	0.02**

Note: **, resp.*, denotes estimated parameters significantly different from 0 at the 5%, resp. 10%, level

Climate conditions primarily affect the composition of the feed ration. Our results suggests that an increase in precipitation and high temperatures lead to an increase in the use of concentrates. Under unfavorable conditions characterized by excess water or high temperature, maize growth can be disrupted, encouraging farmers to increase the amount of concentrates to supplement the feed ration. The impact of weather variables on crop acreage decisions is a bit less straightforward. Rainy conditions in the spring, favorable to grass growth and quality, seem to encourage farmers to produce more grass at the expense of maize fodder.

We now turn to our main parameters of interest, characterizing the flexibility of dairy farmers' in their responses to milk and feed concentrate price variations. Our estimated parameters in the first extreme regime, $\boldsymbol{\beta}_1$, representing farmers behaviors in the case of very small variations in the input to output price ratio, show a negative impact of the price of feed concentrate on the quantities of concentrates purchased by farmers, and a positive impact on the acreage share of fodder maize, which is a substitute for concentrates. The effects of milk price on both feed concentrates and grassland acreage are in that case positive. This suggests that an increase in milk price can encourage farmers to stimulate their milk production by increasing their ration of feed concentrate, and/or by exploiting the benefits of grass. In fact, early grass silage, which is rich in energy and protein, stimulates cows appetite and their production of milk. Both of these choices (increase of feed concentrate or grassland acreage) lead to a decrease in the use of fodder maize to feed animals, hence the negative of milk price on fodder maize acreage.

Estimated $\boldsymbol{\beta}_2$ parameters represent the changes in dairy farmers' behaviors when the market conditions they face change drastically. The significance of these estimates highlights the non-linearity of dairy farmers inputs adjustment in responses to input and output price variations. Our results show that when farmers face important change in input to output price ratio, the impacts of prices on quantities of purchased feed concentrates are strengthened. In fact, the second extreme regimes is characterized by more pronounced responses of feed concentrates to milk and concentrate prices, indicating a non-convex structure of adjustment costs. On the contrary, the impacts of prices on land allocation appear to be smaller, indicating a convex structure of adjustment costs in the case of acreage adjustment. This result is comparable to that found by Antle and Capalbo (2000). In an economic context characterized by small variations in the price ratio, farmers respond to price changes by adjusting the amount of concentrates feed and the allocation of their land between grassland

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

and maize. Larger changes in prices would require larger adjustments in land allocations. However, the adjustment costs of making larger changes in land allocation are too high, since farmers are constrained by the availability of equipment and labor. The only possible response to a price change is then to intensify practices and increase the use of feed concentrates.

Last, but not least, our estimation results reveal a significant heterogeneity in the behaviors of our sample farmers. The estimated standard deviations of the farm specific term additive term a_{0i} in each input use equation are actually significantly estimated, reflecting an heterogeneity, in the level or intensity of input uses by dairy farmers. This heterogeneity might be attributed to several unobserved farms and farmers characteristics such as personal skills, time availability or environmental awareness for instance. More importantly, our results also show a significant heterogeneity in the parameters characterizing the transition function G. In fact, as explained above, dairy farmers tend to substitute all their feeding sources by adjusting both feed concentrates and forage acreages for small changes in market conditions but become less flexible on acreages and essentially adjust their feed concentrate when facing large changes on the market. The way they switch from a rather flexible feeding strategy to a more constrained one however varies from one farmer to the other, as reflected by the significant estimated standard deviations of the farm specific threshold level, c_i (equal to 11% variation in price ratio on average) and speed of transition, g_i (equal to 0.42 on average).

The estimation procedure used to estimate our model, based on a SAEM algorithm, has the advantage of allowing a statistical calibration of the individual parameters of the model for each farmers in our sample. These parameters are computed as the mode of their simulated probability distribution conditional on the observed data available for each farmer (more detail on this calibration procedure can be found in Koutchadé et al., 2018). Once calibrated, the c_i and g_i parameters can notably be used to define transition functions specific to each farms.

Figure 10 represents the estimated individual transition functions. More precisely, it represents, for each farmer, the value of the estimated transition function according the level of absolute variation in price ratio. The transition from one regime to another is quite fast and the slope of the transition function rather homogeneous among farmers. In fact, although significant, the estimated standard deviation of the speed of transition parameter, l_i is relatively small (0.02) compared to its estimated average (0.42). This might be due to the fact that, the second regime being mainly characterized by an increased use of concentrates, the decision to increase the proportion of concentrates in the feed can be made immediately and does not require special additional equipment. As illustrated by the graph, the estimated threshold of transition levels, the c_i parameters, do in fact exhibit more heterogeneity, which is coherent with our estimation results since the average estimated standard deviation of c_i is equal to 2% for an estimated average value of 11%. The individual estimated threshold levels actually range from 6% to 16% in our sample farmers population. This suggests the existence of heterogeneity in farmers' responses to price variations, this heterogeneity being essentially characterized by the differences in the input to output price ratio inducing a switch between the two extreme regimes of input adjustments.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

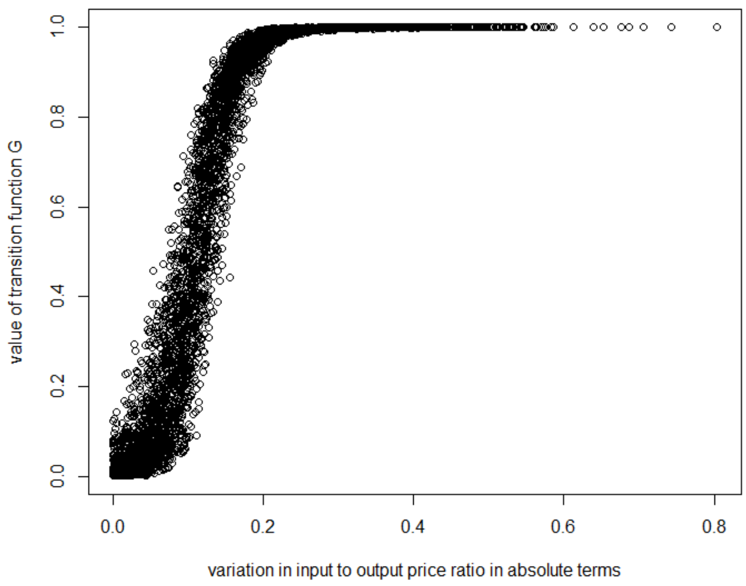

Figure 10. Estimated transition functions

To better characterize the degree of flexibility of each farm, we illustrate the differences in farm behavior between the two extreme regimes in our model by computing price elasticities of feed concentrates and acreage shares for different values of the transition function, G. These elasticities, reported in Table 13, are computed at the sample average for values of G close to 0 (first extreme regime), close to 1 (second extreme regime), and in an intermediate situation when G equals 0.5. In the first extreme regime, faced with an increase in the price of concentrates, farmers decrease the quantity of concentrates (-0.32) and slightly but significantly increase the share of land allocated to fodder maize (0.08). In the second extreme, an increase in the price of concentrates will encourage farmers to decrease the quantity of concentrates even more (-0.42), without modifying their land allocation. When the price of milk varies, the mechanisms remain the same. The price elasticity of the concentrate increases between the two extreme regimes (1.38 to 1.48), while the elasticity of the maize fodder decreases in absolute value (-0.82 to -0.75).

Table 13. Elasticities according the level of the transition function G

	Elasticity of concentrates		Elasticity of maize fodder		Elasticity of grassland	
	Concentrates price	Milk price	Concentrates price	Milk price	Concentrates price	Milk price
$G \rightarrow 0$	-0.32	1.38	0.08	-0.82	0.02	0.16
$G = 0.5$	-0.37	1.43	0.04	-0.79	0.01	0.17
$G \rightarrow 1$	-0.42	1.48	-0.01	-0.75	0.007	0.18

These results come to illustrate that the first extreme regime is characterized by a substitution between different feed sources (concentrates, grass, forage) in response to price changes,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

while in the second regime, farmers adapt to price changes by mainly modifying the quantity of concentrates used to feed to their animals. The switch in farm feeding strategy between the first and second regimes can be attributed to the existence of increasing costs associated with the adjustment in land allocation, which at point limit potential land adjustments. This suggests that farmers who move more quickly from the first extreme regime to the second extreme regime face more adjustment costs associated to acreage decisions and have therefore less degree of flexibility to adjust their feeding strategy. Since there are no additional costs of adjustment associated with the decision to adjust the amount of concentrates, the least flexible farmers seem to use them to compensate for the lack of flexibility in their land allocation choices. This implies that the least flexible farmers are also the most impacted by the highly variable price of feed concentrates.

To support this hypothesis, we build three groups of farmers based on the level of price ratio variation they need to reach the second extreme regime. Farmers in the first group reach the second extreme regime for variation in the price ratio comprised between 26 and 31%. Although these farms do not regularly observe such high price variations, they start moving to the second extreme regime earlier than other farms, their behaviors in the face of price changes are therefore more rapidly similar to those characterizing the second extreme regime. Farmers in the second group reach the second extreme regime for variations in the price ratio comprised between 31 and 35%, and the third group, for variations comprised between 35% and 51%. Farmers in the third group are considered the most flexible, as they exploit the possibilities of substitution between the different components of the feed ration in most economic contexts[34].

Table 14 presents the characteristics of these three groups of farms. The first three columns of the table present the mean and standard deviation of different variables describing farming practices and farm structure by group. The next three columns present the difference between the mean of each group for each variables. T tests of mean equality between groups were performed in order to identify differences in means that are significantly different from 0. The means between the group 1 (less flexible) and group 3 (more flexible) are statistically different for all variables. This shows some interesting features for the most flexible farms: they have better financial autonomy and food self-sufficiency. In addition, these farmers seem to have less intensive farming practices: the animal density per hectare and the quantities of concentrates purchased are significantly lower for that group of farms. Similarly, the proportion of grass in the animal feed ration is higher at the expense of concentrates. Finally, group 3 is characterized by a higher proportion of organic farms, suggesting that organic farms belong to the most flexible group of farms. The adjustment costs of the least flexible farms (group 1) seem to be related to their availability of quasi-fixed inputs (especially capital), but also to their production system: systems based on a high use of fodder maize and feed concentrates offer farmers fewer possibilities for substitutions between the different sources of animal feed.

[34] Note that the average variation of the price ratio is 15% in our and that the average of the maximum variation of the price ratio observed for each individual is 35% regardless of the group. The farmers in each group therefore face on average the same price variations.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Several studies have sought to identify the heterogeneity of production technologies in dairy farms in different European countries. Those studies generally use input use intensity, production specialization or organic farming as direct criteria to distinguish different groups of farms according to their production technology (Kumbhakar et al. 2009, Alvarez and del Corral 2010, Sauer and Paul 2013, Renner et al. 2021). These works generally conclude that the largest, most capital-intensive dairy farms with a higher animal density are the most productive ones. This type of farms appear to be relatively similar to those entering our first group, corresponding to the least flexible farms, suggesting that analysis of farm productivity may not be sufficient to evaluate the sustainability of farms and in particular their capacity to adapt to a highly variable economic and climatic context.

Table 14. Characteristics of farms according their flexibility

	Group 1 26%-31%		Group 2 31%-35%		Group 3 35%-51%		Difference in means	Difference in means	Difference in means
	mean	s.d.	mean	s.d.	mean	s.d.	G1 /G2	G2 /G3	G1 /G3
Number of farmers	181		353		178				
Total area	78	25	65	21	60	22	13**	5**	18**
Livestock density	1.16	0.24	1.11	0.23	1.10	0.21	0.05	0.01	0.06**
Yield (liter/cow)	8 034	936	7 185	1 067	6 175	1 120	849**	1010**	1859**
Share of grassland	0.35	0.08	0.44	0.10	0.56	0.10	-0.09**	-0.12**	-0.21**
Concentrates (€/cow)	478	157	370	126	254	100	108**	116**	224**
Share of farm-produced food	0.25	0.13	0.31	0.16	0.41	0.19	-0.06**	-0.1**	-0.16**
Share of organic farm	0.05	0.23	0.03	0.18	0.12	0.32	0.02	-0.09**	-0.07**
Gross margin (€/litre)	222	38	234	44	258	52	-12**	-24**	-36**
Unit of agricultural workers	1.87	0.65	1.71	0.61	1.69	0.78	0.16**	0.02	0.18**
Capital (€/1 000 litre)	907	244	903	290	973	311	4	-70**	-66**
Debts (€/1 000 litre)	390	200	347	212	348	233	43**	-1	42*

Note: **, resp.*, denotes difference in means between groups significantly different from 0 at the 5%, resp. 10%, level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

5.6. Conclusion

By relying on a panel smooth transition model, we have been able to identify heterogeneous flexibility of farmers in their short run responses to input and output prices variations. Our proposed approach to identify non-linear and heterogeneous farm behavior contributes to the literature in three main aspects. First, we propose a way to identify farm heterogeneity based on their observed short run responses to input and output prices. Our approach allows revealing farm heterogeneity, without the need to specify *ad hoc* criteria differentiating farms. Second, we propose an original framework in order to implicitly account for adjustment costs in production behavior of farmers. Our simple model allows distinguishing farmers according to their speed of reaction in response to price changes, and identifying the most flexible farmers in the short run. Third, we propose a new estimation procedure for panel smooth transition regression models with individual parameters characterizing the transition function. Our approach allows dealing with the specificity of farm accountancy panel data that generally have large individual and short time dimensions.

The results we obtain on a sample of French dairy farms confirm the interest of our approach since we identify significantly heterogeneous production behaviors. Farms face different levels of adjustment costs that constrain their ability to adapt to price variations observed on the markets. Some farmers are considered more flexible in our approach, in the sense that they adjust their feeding strategy more easily to price variations. An *ex-post* analysis confirms their specific characteristics: they use less intensive, more grass-based practices, allowing them to be more food self-sufficiency. Organic farms are also overrepresented among the most flexible farms.

Despite its originality in characterizing the heterogeneity of dairy producers in their adjustment to short-term price variations, we recognize certain limitations of our framework. First, it essentially takes into account the adjustment of the feeding strategy of farms but does not take into account potential adjustment of their herd size. This assumption actually allows improving the empirical tractability of our model and is, at least in our empirical application, not so strong because herd size varies very little in the short run in the sample we consider. Moreover, this allows our analytical framework to be directly transposable to other types of farmers' decisions, such as the choice of fertilizer or pesticide use for crop producers. Second, we focus here essentially on farmers' adjustment to price shocks, although it would also be interesting to analyze their adjustment to climate shocks. This could be done by considering a climate indicator as transition variable in the model, but would however require an index sufficiently synthetic to represent the impact of climate variations on all farmers' input choices.

As the capability to adjust their production choices appears as a key aspect of farms' economic sustainability, our approach can help identifying public actions levers in order to encourage farmers to be more reactive to price fluctuations. This is all the more important since currently input prices are high: the price of cattle feed increased by almost 30% between June 2021 and June 2022. Given the findings of our paper, the impact of these price shocks on the input market will affect farmers differently depending on their ability to react, and in particular on their ability to substitute different feed sources. In this very volatile economic context, our approach can be useful to predict the impact of external shocks to input market, and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

therefore, the effects of potential policy measures that could be taken to support farmers (Hamermesh and Pfann 1996).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

6. LAND OPTIMIZATION AND GREENHOUSE GAS EMISSION MITIGATION OF DUTCH DAIRY FARMS

6.1. Introduction

We face several major but intertwined global challenges: from climate change, to environmental degradation, global food insecurity, increasing population growth, and poverty:. In light of these challenges, the dairy sector needs to reduce its environmental impact, while continuing to produce high-quality animal products (Food and Agriculture Organization [FAO], 2019). The Dutch dairy sector is highly productive, but the environmental costs from its farming activities are substantial (Hou et al., 2016; H. J. M. van Grinsven, M. M. van Eerdt, H. Westhoek, and S. Kruitwagen, 2019; Zhu and Lansink, 2022), which is why environmental externalities need to be considered in production analyses.

To comply with the Paris Agreement on Climate Change, the Dutch government has developed its national Climate Agreement (Klimaatakkoord) (Rijksoverheid, 2022). Dairy farmers have already taken measures to reduce emissions of greenhouse gases (GHG), but there is an urgent need to accelerate the sector's response to meet the emission reduction target (Food and Agriculture Organization, 2019; H. J. Van Grinsven, M. M. van Eerdt, H. Westhoek, and S. Kruitwagen, 2019). Current policies focus on transitioning towards a more circular agriculture, which is regarded as a cost-effective means to reduce GHG emissions (Food and Agriculture Organization, 2019; Ministerie van Landbouw Natuur en Voedselkwaliteit, 2019; Wageningen University and Research, 2022).

Circular agriculture closes resource cycles by optimizing efficiency, recycling waste (e.g. manure), reducing external inputs (e.g. animal feed, artificial fertilizers, pesticides and fossil fuels), continuous systemic improvements, cross value chain collaboration, and decreasing possible emissions and negative externalities (Boer and Ittersum, 2018). Efficient production and resource optimization are cucial in the transition towards circular agriculture. In terms of land use, feeding animal left-over crops is estimated to save 25% of global crop land compared to not keeping any livestock (Van Zanten et al., 2018). Dutch dairy farmers already upcycle manure for crop fertilizers and produce their own feed on the farm. However, it is not clear to what extent land reallocation between crops and grassland can contribute to reducing GHG emissions.

This study aims to explore the potential for land optimization on dairy farms to simultaneously increase production and reduce GHG emissions. Incorporating circularity principles in an efficiency framework requires explicit modeling regarding the recycling of intermediate outputs, reallocating inputs, and reducing pollution. Focusing on U.S. dairy farms, Färe and Whittaker (1995) showed how recycled crop output can be modeled as a feed input in a livestock enterprise in an efficiency framework. Färe, Grabowski, Grosskopf, and Kraft (1997) quantified potential efficiency gains from reallocating land use inputs for a sample of Illinois grain farms. Focusing on English and Welsh farms, Ang and Kerstens (2016) combined these two aspects, and characterized the inputs as joint or output-specific following Cherchye, Rock, Dierynck, Roodhooft, and Sabbe (2013). However, none of these studies considered the recycling of manure output in the crop production process. Pollution is generated as a by-product in the agricultural production process. The current consensus for modeling pollution is the by-production approach developed by Førsund (2009) and Murty, Russell, and Levkoff (2012). Recent applications to the agricultural sector include K. H. Dakpo, Jeanneaux, and Latruffe (2017), Serra, Chambers, and Oude Lansink (2014) and Ang, Kerstens, and Sadeghi (2022). To the best of our knowledge, no study has structurally addressed these circularity aspects within one integrated multi-production technology framework. The current article addresses this research gap by developing such an efficiency framework that allows to assess the potential reduction in GHG emissions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

This study contributes to the literature in three ways. First, it extends previous work from Ang and Kerstens (2016) by explicitly considering the manure cycle, that is, distinguishing the upcycled manure as crop fertilizers and the remaining manure that is removed from the farm. Second, this is the first study to combine the by-production approach with the network DEA model to analyse land allocation decisions on dairy farms. Third, this study provides scientific evidence for policy advice on whether promoting mixed dairy farming is effective in reducing GHG emissions, under a given livestock size. Our results show that specialized dairy farms in the Netherlands can enhance technical and environmental efficiency by 5.1% simultaneously (to which land optmization only contributes 0.6%). GHG emissions can be reduced by 11.79% if production and inputs are held constant without reallocating land.

.

6.2. Method

In this section, we describe the network DEA model that is used to assess the performance of dairy farms. This model is also used to investigate the potential for land reallocation to increase production efficiency and decrease GHG emissions. We distinguish three interdependent subprocesses with their corresponding technologies. This is followed by an explanation of the axiomatic properties, model formulation and coordination inefficiency.

6.2.1. Technology

This study operationalizes two sub-technologies with intended outputs: crop production and livestock production. Crop and livestock outputs are modelled separately to optimize the land allocation between both production processes. In addition, a third residual-production technology is operationalized for GHG emissions. In the by-production approach to model pollution-generating technology, the production of intended-output sets the residual-production technology in motion, which leads to the generation of by-product (Murty et al., 2012).

The network DEA model structure is shown in Figure 11. Each dairy farm is denoted by subscript k. Crop production and livestock production processes are linked through (i) the use of upcycled manure from livestock production as fertilizer in crop production ($m_k^{L,U}$), and (ii) the use of unsold crop residuals (z_k^C as feed in addition to the purchased feed) in livestock production. The total on-farm GHG emissions (e_k) are generated by the polluting inputs ($x_k^{C,p}$, $x_k^{L,p}$ & $q_k^{J,p}$). The detailed inputs and outputs of each production technology are described in Table 15.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

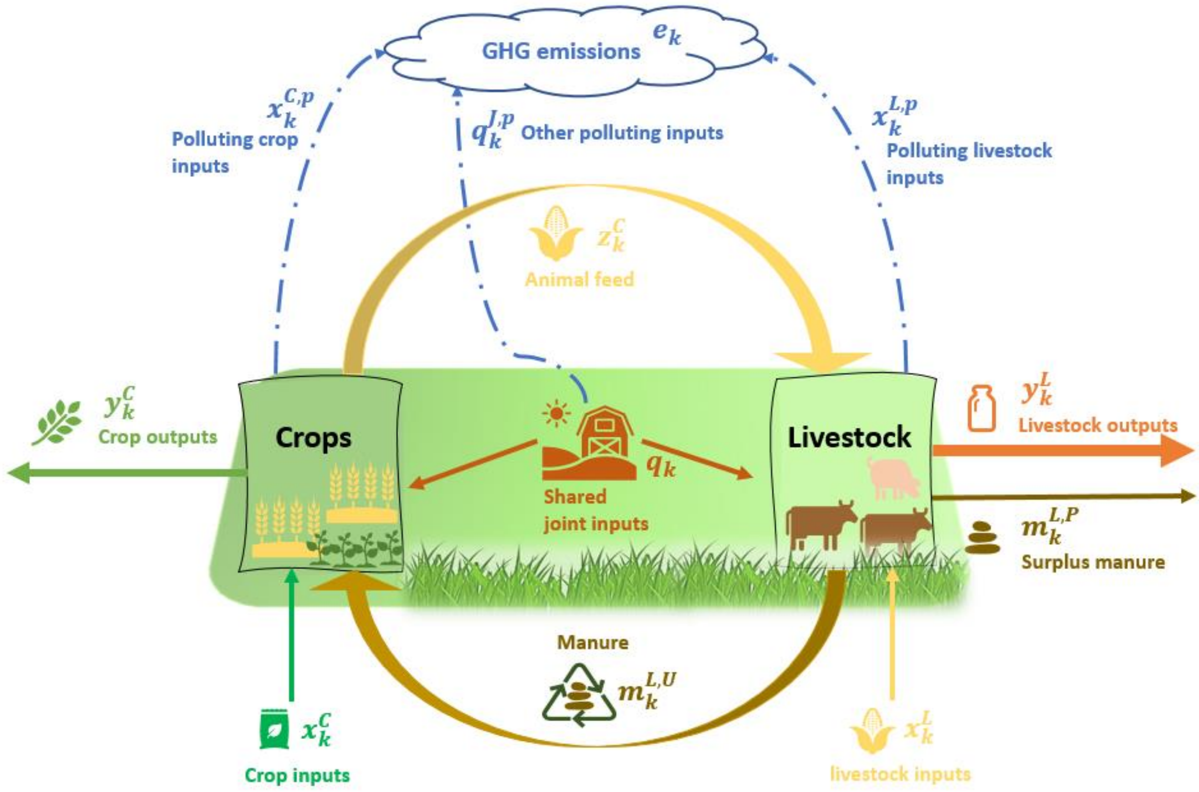

Figure 11. Network structure of specialized dairy farms

Table 15 Inputs, outputs variables for each technology

The intended crop production technology has the following inputs and outputs	
$x_k^C \in \mathbb{R}_+^{Nc}$	aggregated crop-specific inputs, including crop protection products, purchased fertilizers, and seeds.
$m_k^{L,U} \in \mathbb{R}_+$	upcycled manure used as fertilizer for crops in the same year.
$q_k \in \mathbb{R}_+^M$	shared joint inputs by crop and livestock processes, including aggregated input set (which consists of buildings, machinery & equipment, and energy consumption); as well as water use, and labor.
$y_k^C \in \mathbb{R}_+^{OC}$	aggregated crop output revenues from wheat, barley, potatoes, sugar beet, vegetables, grass seeds, folder crops, and other arable crops.
$z_k^C \in \mathbb{R}_+^{Oc}$	unsold crop residuals used as animal feed: maize & grass.
The intended livestock production technology has the following inputs and outputs	
$x_k^L \in \mathbb{R}_+^{Nl}$	aggregated livestock-specific inputs, including animal units, purchased animal feed, animal health costs and animal water use.
$z_k^C \in \mathbb{R}_+^{Oc}$	unsold crop residuals used as animal feed: maize & grass.
$q_k \in \mathbb{R}_+^M$	shared joint inputs by crop and livestock processes, including aggregated input set (which consists of buildings, machinery & equipment, and energy consumption); as well as water use, and labor.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

The intended crop production technology has the following inputs and outputs	
$y_k^L \in \mathbb{R}_+^{OL}$	aggregated livestock output revenues from milk & milk products, cattle, eggs, poultry, pigs, sheep and wool.
$m_k^{L,P} \in \mathbb{R}_+$	surplus manure removed from the farm.
$m_k^{L,U} \in \mathbb{R}_+$	upcycled manure used as fertilizer for crops in the same year.
The residual GHG emission technology has the following inputs and outputs:	
$x_k^{C,p} \in \mathbb{R}_+^{Npc}$	polluting aggregated crop-specific inputs, including crop protection products, purchased fertilizers, and seeds.
$x_k^{L,p} \in \mathbb{R}_+^{Npl}$	Polluting livestock specific inputs, including animal units, purchased animal feeds, unsold crops residuals used as animal feed.
$q_k^{J,p} \in \mathbb{R}_+^{pj}$	other polluting inputs including energy use and total manure.
$e_k \in \mathbb{R}_+$	total GHG emissions in carbon dioxide equivalent from crop and livestock production processes.

We now define the three sub-technologies with their production set as below.

The intended crop production technology is:

$$T_1 = \{(x_k^C, m_k^{L,U}, q_k) \; produces \; (y_k^C, z_k^C)\} \tag{1}$$

The intended livestock production technology is:

$$T_2 = \{(x_k^L, z_k^C, q_k) \; produces \; (y_k^L, m_k^{L,P}, m_k^{L,U})\} \tag{2}$$

The residual GHG emission production technology is:

$$T_3 = \{(x_k^{C,p}, x_k^{L,p}, q_k^{J,p}) \; produces \; (e_k)\} \tag{3}$$

The overall technology is $T = T_1 \cap T_2 \cap T_3$.

6.2.2. Axiomatic properties

The free disposability axioms apply to T_1 and T_2. T_3 satisfies the costly disposability axiom (Murty et al., 2012). Costly disposability allows inefficiencies in the generation of pollution (Murty et al., 2012). For a given level of inputs and intended outputs, there is a minimum level of pollution. Pollution above this minimum level is inefficient (K. Hervé Dakpo, Jeanneaux, and Latruffe, 2016).

T_1 is defined as:

$(x_1, y_1) \in T_1 \wedge x_1{}' \geq x_1 \rightarrow (x_1{}', y_1) \in T_1$ (Free disposability of all inputs);

$(x_1, y_1) \in T_1 \wedge y_1{}' \leq y_1 \rightarrow (x_1, y_1{}') \in T_1$ (Free disposability of all outputs).

T_2 is defined as:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$(x_2, y_2) \in T_2 \wedge x_2{'} \geq x_2 \rightarrow \left(x_2{'}, y_2\right) \in T_2$ (Free disposability of all inputs);

$(x_2, y_2) \in T_2 \wedge y_2{'} \leq y_2 \rightarrow \left(x_2, y_2{'}\right) \in T_2$ (Free disposability of all outputs, except manure);

$(x_2, y_2, m) \in T_2 \wedge 0 < \theta < 1 \rightarrow (x_2, \theta y_2, \theta m) \in T_2$ (weak disposability of manure);

$(x_2, y_2, m) \in T_2 \wedge m = 0 \Rightarrow y_2 = 0$ (null-jointness of manure and livestock production).

The combination of weakly disposable manure and null-jointness for manure is that excess manure disposal generates costs for the farmer as manure can only be upcycled and used as crop fertilizer up to a certain amount (Ronald W Shephard, 1977; Ronald William Shephard, 2012).

T_3 is defined as:

$(x^p, e) \in T_3 \wedge x^{p'} \leq x^p \rightarrow (x^{p'}, e) \in T_3$ (costly disposability of pollution-generating inputs);

$(x^p, e) \in T_3 \wedge e' \geq e \rightarrow \left(x^P, e'\right) \in T_3$ (costly disposability of GHG emissions).

6.2.3. Model formulation

For each individual farm (DMU) k = 1,..., K, the DMU under evaluation is k = i. The directional output distance function is given by:

$$D_k(x_k, y_k^C, z_k^C, y_k^L, e_k; g_k) = \sup \left\{ \beta \geq 0 : \left(x_k, y_k^C + \beta g_{y,k}^C, z_k^C + \beta g_{z,k}^C, y_k^L + \beta g_{y,k}^L, e_k - \beta g_{e,k}\right) \in T_1 \cap T_2 \cap T_3 \right\} \quad (3)$$

β is the overall technical inefficiency score. g_k is the directional vector that expands the intended outputs, y_k^C, z_k^C and y_k^L, and contracts GHG emissions, e_k. An output-oriented model is chosen as this research aims to quantify the potential of the circularity principle and of land optimization in simultaneously producing intended products and reducing residual GHG, given the level of all inputs. We have selected $g_{y,k}^C = y_k^C, g_{z,k}^C = z_k^C, g_{y,k}^L = y_k^L, g_{e,k} = e_k$ as the directional vectors, following for instance Ang and Kerstens (2016) and Chambers, Fāure, and Grosskopf (1996). β indicates the maximum proportional expansion of desirable outputs and maximum proportional contraction of undesirable outputs. x_k represents all the inputs in the directional distance function.

Land use is a shared, yet non-joint input by both crop and livestock production. Farmers have to decide how much land to use for livestock production and crop production. In line with Ang and Kerstens (2016) and Cherchye, De Rock, and Hennebel (2017), one can simultaneously further expand production and reduce GHG emissions by reallocation. Let $x_k \in \mathbb{R}_+^C$ with $C \subseteq \{1, \dots, N_C\} \cap \{1, \dots, N_L\}$ be the process-specific inputs that have to be reallocated between the crop and livestock subprocesses, such that $x_k^{C,l} + x_k^{L,l} = x_k^l \quad \forall\, l \in C$. Here, C refers to land use, common to crop and livestock, that can be reallocated among both subprocesses. Land use is treated as a re-allocatable and fixed input in line with Färe et al. (1997). The total land use on the dairy farm equals the sum of crop land and livestock land.

The DEA model that allows reallocation of land is given by equations (4), (4a) – (4z). β_i is the re-allocative technical ineffiency score for each farm i under evaluation. This model also nests

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

the model with fixed land allocation, i.e. constraints (4a) – (4y) and removing the crop and livestock land ($X_i^{C,l}$, $X_i^{L,l}$) from the optimization operand in (4). The detailed model formulation for non-reallocation can be found in appendix 6A. The resulting β from that model is the non-re-allocative technical ineffiency score for each farm i under evaluation. Note that our model makes few implicit assumptions about land use following Ang and Kerstens (2016). We assume that land use is immediately re-allocatable, (i.e. costless and substitutable) between crops and livestock on the same dairy farm.

$$\max_{\substack{\beta_i,\lambda_k,\gamma_k,\mu_k \\ X_i^{C,l}\geq 0,\, X_i^{L,l}\geq 0}} \beta_i \quad (4)$$

s.t.

$$\sum_{k=1}^{K} \lambda_k x_k^C \leq x_i^C \quad (4a)$$

$$\sum_{k=1}^{K} \lambda_k m_k^{L,u} \leq m_i^{L,u} \quad (4b)$$

$$\sum_{k=1}^{K} \lambda_k x_k^{C,l} - x_i^{C,l} \leq 0 \quad (4c)$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J1} \leq q_i^{J1} \quad (4d)$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J2} \leq q_i^{J2} \quad (4e)$$

$$\sum_{k=1}^{K} -\lambda_k y_k^C + \beta_i g_{y,k}^C \leq -y_i^C \quad (4f)$$

$$\sum_{k=1}^{K} -\lambda_k z_k^C + \beta_i g_{z,k}^C \leq -z_i^C \quad (4g)$$

$$\sum_{K=1}^{K} \lambda_k = 1 \quad (4h)$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,fh} \leq x_i^{L,fh} \quad (4i)$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,a} \leq x_i^{L,a} \quad (4j)$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,l} - x_i^{L,l} \leq 0 \quad (41)$$

$$\sum_{k=1}^{K} \gamma_k z_k^C \leq z_i^C \quad (21)$$

$$\sum_{k=1}^{K} \gamma_k q_k^{J1} \leq q_i^{J1} \quad (4m)$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\sum_{k=1}^{K} \gamma_k q_k^{J2} \ \leq \ q_i^{J2} \tag{4n}$$

$$\sum_{k=1}^{K} - \gamma_k y_k^L + \ \beta_i g_{y,k}^L \leq \ -y_i^L \tag{4o}$$

$$\sum_{K=1}^{K} \gamma_k \ = 1 \tag{3}$$

$$\sum_{k=1}^{K} \gamma_k (m_k^{L,u} \ + \ m_k^{L,p}) \ = \ m_i^{L,u} \ + \ m_i^{L,p} \tag{4q}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{C,p} \leq - x_i^{C,p} \tag{4r}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pa} \leq - x_i^{L,Pa} \tag{4}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pf} \leq - x_i^{L,Pf} \tag{4t}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pfc} \leq - x_i^{L,Pfc} \tag{4u}$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pe} \leq - q_i^{J,pe} \tag{4v}$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pm} \leq - q_i^{J,pm} \tag{4w}$$

$$\sum_{k=1}^{K} \mu_k e_k \ + \ \beta_i g_{e,k} \leq \ e_i \tag{4x}$$

$$\sum_{K=1}^{K} \mu_k \ = 1 \tag{5}$$

$$x_i^{C,l} \ \ + \ x_i^{L,l} \ = \ x_i^l \tag{4z}$$

The coordination efficiency (CI) is measured by

$$CI = \ RTIE - NRTIE \tag{5}$$

where $RTIE$ and $NRTIE$ denote re-allocative technical inefficiency and non-re-allocative technical inefficiency, respectively. CI is non-negative, as non-reallocation is always possible when reallocation is allowed. Any positive value for the CI indicates a possibility to further increase intended outputs and reduce GHG emissions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

6.3. Data

Our empirical application focuses on a sample of Dutch dairy farms over the period of 2010 to 2019. We obtained data from the Dutch Farm Accountancy Data Network (FADN) supplemented with GHG emissions data on dairy farm from Wageningen Economic Research (WEcR). The FADN is a European instrument that evaluates farm income and the impacts of the Common Agricultural Policy (van der Meer, 2019). Farmers participate in the FADN voluntarily. In the FADN, dairy farms are defined as those whose revenues from sales of milk, milk products, turnover and growth of cattle represent at least two thirds of their total revenue. The sample is unbalanced as farms stay in the sample for a period of 4–7 years, and it is statistically representative for the Dutch dairy sector. On average, there are 190 farms for each year. In this study, mixed dairy farms are not included because data on GHG emissions is not available.

We distinguish technology-specific inputs and outputs. For the crop production technology, we aggregate crop-specific costs (seeds, crop protection products and fertilizers), upcycled manure, crop land use (feed crops and cash crops), aggregated crop yields that are sold to the market, and the crop residuals used for animal feed. For the livestock-specific technology, we have livestock units, aggregated livestock specific costs (animal health costs and purchased animal feed, tap water cost), feed from own crop residuals, livestock land use (grassland), aggregated livestock production, and total manure from farm. There are joint shared inputs for the crop-production technology and the livestock-production technology: aggregated joint inputs set 1 includes energy, value of building, machinery and equipment; and joint inputs set 2 includes labor and water use irrigation. For the greenhouse technology, we have included only the pollution-generating inputs and the total on-farm GHG emissions. We aggregate the monetary inputs and outputs as implicit quantities by computing the ratio of their aggregated value to their corresponding aggregated Törnqvist price index. Price indices vary over years but not over farms. This implies that the differences in the quality of inputs and outputs are reflected by implicit quantities (Cox and Wohlgenant, 1986). The separate price indices are obtained mostly from EUROSTAT (2022) and the tap water price index from the Dutch Centraal Bureau voor de Statistiek (2022). The final dataset contains 1,896 observations for the period of 2010 to 2019. The descriptive statistics of the variables are summarized in Table 16.

Table 16 Descriptive statistics of model variables

Variables	Dimensions	Average	Std dev.
Crop-specific variable inputs $\boldsymbol{x_k^C}$; $\boldsymbol{x_k^{C,p}}$	Euros	15,147.36	14,449.90
Upcycled manure $\boldsymbol{m_k^{L,u}}$	Ton	3,598.76	2,494.82
Joint inputs set 1 $\boldsymbol{q_k^{J1}}$	Euros	598,716.17	443,149.08
Joint inputs set 2 $\boldsymbol{q_k^{J2}}$:			
Labor	Full hours	5,177.16	3,120.61
Water use irrigation	M^3	3,923.24	13,562.46
Total crop outputs as sold $\boldsymbol{y_k^C}$	Euros	6,570.74	33,309.97
Unsold crop for animal feed (maize & grass) $\boldsymbol{z_k^C}$; $\boldsymbol{x_k^{L,Pfc}}$	kVEM	728,645.37	476,513.96

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Livestock units $x_k^{L,a}$; $x_k^{L,Pa}$	Cow equivalent	171.42	105.28
Livestock-specific variable inputs $x_k^{L,fh}$	Euros	136,331.53	98,160.10
Total livestock production y_k^{L}	Euros	434,236.05	308,233.90
Animal feed expenditure $x_k^{L,Pf}$	Euros	129,869.95	95,765.49
Energy expenditure $q_k^{J,pe}$	Euros	16,469.35	12,638.89
Total manure $(m_k^{L,u} + m_k^{L,p})$; $q_k^{J,pm}$	Ton	4,333.93	2,937.53
Total crop land $x_k^{C,l}$	Hectares	12.40	14.19
Total livestock land $x_k^{L,l}$	Hectares	54.39	33.46
Total GHG emissions e_k	Ton	1,818.44	1,238.42

Note: kVEM is the energy content of the dry matter.

6.4. Results

In this section, we first present the overall technical inefficiency scores, followed by land optimization results. Scenario results and robustness check are discussed as well.

In this section, we first present the overall technical inefficiency scores, followed by land optimization results. Scenario results and robustness check are discussed as well.

6.4.1. Overall technical inefficiency scores

Table 17 depicts the yearly average results of the coordination inefficiency *CI*, technical inefficiency when land is optimally chosen ($RTIE$), and the technical inefficiency when land reallocation is not allowed between crops and livestock use ($NRTIE$). For the period 2010 to 2019, the yearly average technical inefficiency ranges from 3.0% to 7.2% when land is optimally chosen. This means on average farms could simultaneously expand production and reduce GHG emissions by 3.0% in 2010 and by 7.2% in 2016, *ceteris paribus*. When land is not allowed to be optimized, the yearly average technical inefficiency ranges from 2.3% to 6.6% for the period 2010 to 2019. This means that on average farms could gain economic and environmental efficiency by 2.3% in 2010 and by 6.6% in 2016. The difference between $RTIE$ and $NRTIE$, which is the coordination inefficiency *CI*, is on average small and ranges from 0.3% to 0.8% between 2010 and 2019.

Table 17. Average coordination inefficiency (CI) scores and average technical inefficiency scores with and without land reallocation for the full model with directional vector $(g^C_{y,k}, g^C_{z,k}, g^L_{y,k}, g_{e,k})$ per year.

Inefficiency	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
CI	0.007	0.003	0.005	0.008	0.008	0.008	0.006	0.006	0.007	0.004
$NRTIE$	0.023	0.034	0.039	0.041	0.044	0.050	0.066	0.058	0.048	0.046
$RTIE$	0.030	0.037	0.044	0.049	0.052	0.058	0.072	0.064	0.055	0.050

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

6.4.2. Land optimization

We compare actual and optimal land allocation in Figure 12. Except for the year 2010, the results suggest that more land should be allocated to crop production to reduce GHG emissions and increase production simultaneously. Our results suggest that by reallocating on average 4.5 hectares from livestock use to crop production on a Dutch dairy farm (total size of 66.8 hectares on average), farms can simultaneously increase production and reduce GHG emissions by 5.1%, with only 0.6% due to land optimization. To be specific, 0.6% efficiency gain will be achieved if crop land use takes up 25.3% of the total farm size instead of 18.6% as of the current situation. 4.5% efficiency gains can be achieved if farms try to catch up with their best performing peers.

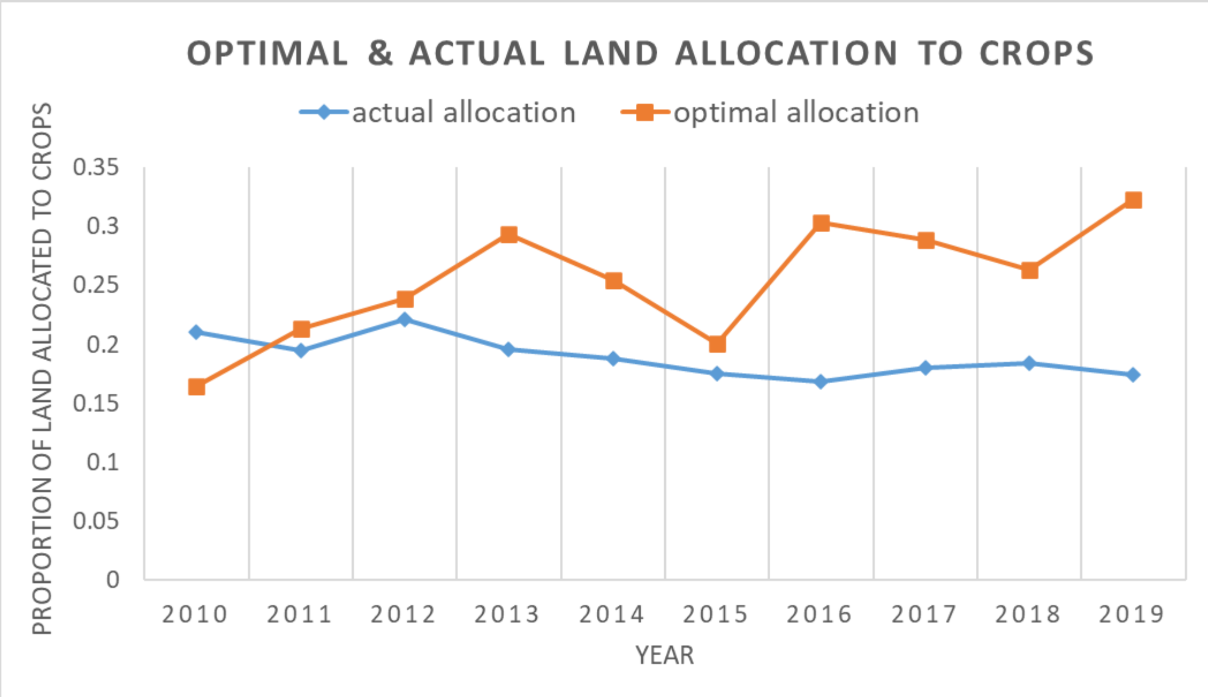

Figure 12. Distribution of optimal and actual proportion of land allocated for crop production per year

6.4.3. Alternative pathways to reduce GHG emissions

Besides the maximum proportional expansion of desirable outputs and contraction of undesirable outputs (which we denote as pathway 1), we also explored three other orientations under different directional distance vectors. The purpose is to explore the potential for further reduction of GHG emissions on dairy farms versus the potential for increased production. Table 18 illustrates the results for these four pathways. Pathway 1 shows the simultaneous results for increasing production and reducing GHG emissions, pathway 2 shows the results when only reducing GHG emissions, pathway 3 shows the results when expanding crop and livestock production, and pathway 4 shows the results when only expanding livestock production.

Under pathway 1 with the directional vector of ($g^C_{y,k} = y^C_k, g^C_{z,k} = z^C_k, g^L_{y,k} = y^L_k, g_{e,k} = e_k$), the average technical inefficiency without and with land reallocation is 4.5% and 5.1%,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

respectively. These results show that by optimizing land allocation, dairy farms can expand production and reduce GHG emissions by 5.1% on average while keeping everything else constant. The coordination efficiency gain is on average 0.6%. The efficiency gain under pathway 1 with or without land optimization is the lowest among all pathways. This implies that most specialized Dutch dairy farms are already quite efficient when it comes to proportional production expansion and GHG emissions contraction. The maximum GHG reduction potential does not come from this pathway.

Under pathway 2 with the directional vector of ($g^C_{y,k} = 0, g^C_{z,k} = 0, g^L_{y,k} = 0, g_{e,k} = e_k$), the average technical inefficiency with/without land reallocation is 11.8%, and the coordination efficiency gain is on average 0.001%. These results point out that GHG emissions can be reduced by 11.79% on average among the sample dairy farms, while keeping conventional production and all inputs constant without land reallocation. With land reallocation, the efficiency gain is only 0.001% which is very small. Land optimization does not contribute to reducing GHG emissions when inputs and conventional outputs are held constant. Nevertheless, the highest GHG reduction potential can be reached via this pathway among all pathways.

Under pathway 3 with the directional vector of ($g^C_{y,k} = y^C_k, g^C_{z,k} = z^C_k, g^L_{y,k} = y^L_k, g_{e,k} = 0$), the average technical inefficiency without land reallocation is 5.9%, and the coordination efficiency gain is on average 2.2%. Among all pathways, pathway 3 offers the highest potential to enhance both crop and livestock production, when emissions and inputs are held constant. If GHG emission and all inputs are held constant, an additional coordination efficiency gain of 2.2% can be obtained on average by optimizing land allocation across outputs. This is the highest coordination efficiency gain among all cases.

Under pathway 4 with the directional vector of ($g^C_{y,k} = 0, g^C_{z,k} = 0, g^L_{y,k} = y^L_k, g_{e,k} = 0$), the average technical inefficiency without land reallocation is 8.6%, and the coordination efficiency gain is on average 0.8% for each farm. These results show that livestock production can be increased by 8.6% on average among sample dairy farms, while crop outputs & GHG emissions, and all inputs are held constant without land allocation. If land optimization is allowed, there is an 0.8% additional efficiency gain for livestock outputs per farm on average. However, this efficiency gain is lower than for pathway 3, which indicates land reallocation does not contribute much to improve the efficiency in this case.

Given the importance of tackling climate change, it is more realistic to consider the implications of the results from the first two pathways. Overall, land optimization does not bring substantial efficiency gains as can be observed from the small value of *CI*. Interestingly, GHG emissions can be reduced with 11.8% on average with or without land reallocation if all inputs and conventional outputs are held constant. This reduction potential of GHGs decreases to 4.5% if producers are allowed to expand crop and livestock outputs at the same time, holding inputs and land use constant.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Table 18 Average technical inefficiency scores and the coordination inefficiency (CI) scores for models with different directional vectors.

Variables	Pathway 1	Pathway 2	Pathway 3	Pathway 4
Average inefficiency scores	$(y_k^C, z_k^C, y_k^L, e_k)$	$(0, 0, 0, e_k)$	$(y_k^C, z_k^C, y_k^L, 0)$	$(0, 0, y_k^L, 0)$
CI	0.006	0.000	0.022	0.008
$NRTIE$	0.045	**0.118**	**0.059**	0.086
$RTIE$	0.051	**0.118**	**0.08**1	0.094

6.4.4. Robustness check

Our DEA model used one output-specific inefficiency score for both conventional production and residual GHGs. This provides us results for simultaneous expansion and contraction in the direction of corresponding directional vectors. We investigated the robustness of the results by modeling the conventional technology and residual technology using two different output-specific inefficiency scores, i.e. a technical inefficiency score β for crop- and livestock-production technologies, and a technical ineffciency score α for the residual GHG emission technology. The detailed model formulation is added in appendix 6B.

Table 19 shows the separate inefficiency scores for conventional technology and residual GHG emission technology per year, with and without land reallocation. The last column of Table 19 shows the average score over the entire period. It is very similar to the results listed in Table 18.

Table 19 Desirable output and GHG emission specific technical inefficiency with and without land reallocation per year and the mean over the entire period.

Inefficiency	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	mean
$NRTIE$											
Desirable outputs β_i	0.041	0.051	0.053	0.057	0.065	0.076	0.107	0.082	0.072	0.083	**0.069**
GHG emissions α_i	0.085	0.098	0.120	0.130	0.098	0.117	0.124	0.127	0.122	0.116	**0.114**
$RTIE$											
Desirable outputs β_i	0.062	0.060	0.066	0.070	0.075	0.088	0.126	0.096	0.082	0.090	**0.082**
GHG emissions α_i	0.085	0.098	0.120	0.130	0.140	0.117	0.124	0.127	0.122	0.116	**0.118**

The land optimization results from the model in appendix 6A are plotted in Figure 10. In general, the distribution under separate efficiency scores follows the distribution under the identical inefficiency score, with slightly lower values. In 2014 and from 2016 to 2019, more land should have been allocated to crop production than the actual land allocation. For the years 2011 to 2013, land re-allocation would not have brought any efficiency gains. For the year 2010 and 2015, the data suggests that more land should have been allocated to livestock land use to increase efficiency. Overall, a smaller proportion of land needs to be allocated to crop production with separate inefficiency scores (on average 2.86 hectares) than considering the optimal allocation with identical inefficiency scores (on average 4.5 hectares).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

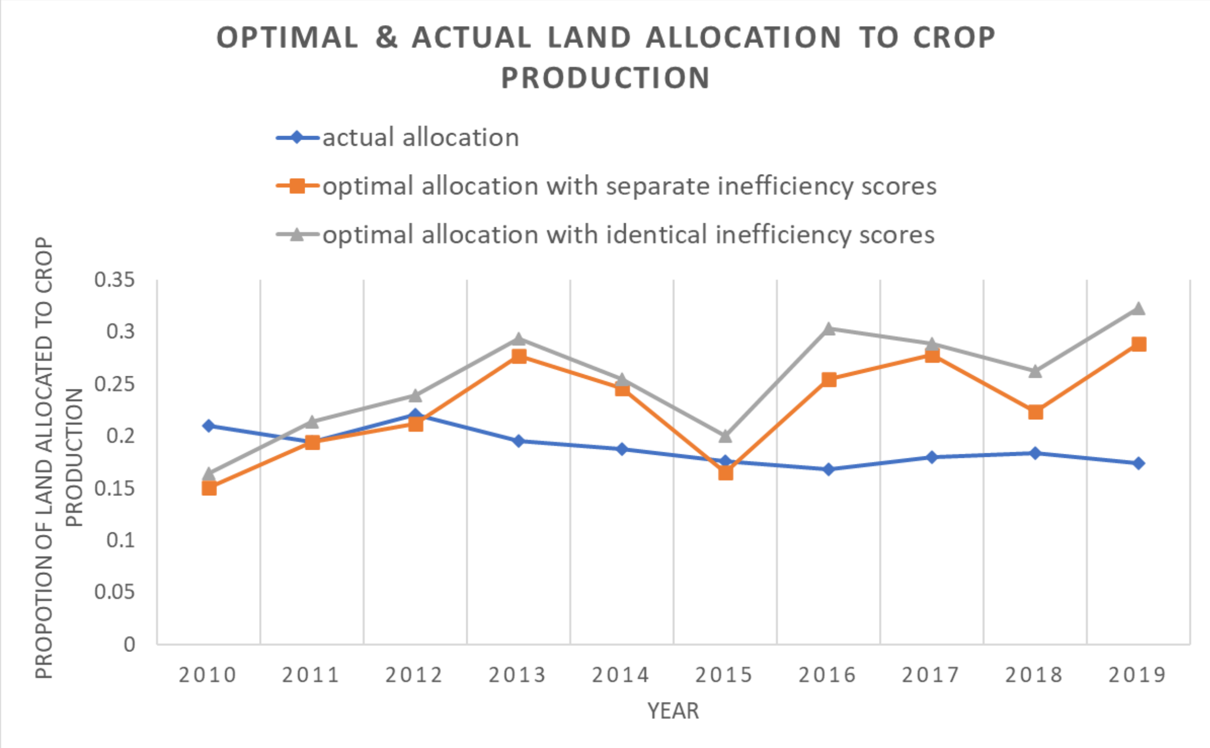

Figure 13. Distribution of optimal (under separate inefficiency scores and identical inefficiency scores) and actual land allocation for crop and livestock production per year

6.5. Discussion

This study used a network DEA model with the by-production approach to quantify the technical & environmental inefficiency of dairy farms, taking GHG emissions into account. The model also enables quantification of the efficiency gains from land reallocation between crop and livestock use. The current study found that the technical inefficiency is on average 4.5% at the farm level. Land reallocation to crop production could bring a small additional efficiency gain of 0.6% on average.

This finding is consistent with the results of Ang and Kerstens (2016), who conclude that coordination inefficient farms should in general allocate more land to crop production. However, the coordination inefficiency score obtained in this study is lower than that estimated by Ang and Kerstens (2016), which means land optimization on specialized Dutch dairy farms provides only minimal effecency gains. This difference could be explained by the fact that this study focuses exclusively on specialized dairy farms, whereas Ang and Kerstens (2016) also included mixed farms and specialized crop farms.

Only a few other studies have looked into environmental efficiency gains on dairy farms. For French suckler cow farms, K. H. Dakpo and Oude Lansink (2019) found an average technical inefficiency (*TIE*) for desirable output of 0.2%, while the average *TIE* for GHG emission was 28.4%; i.e. much lower and higher than for our study. This suggests that specialized Dutch dairy farms are more efficient in mitigating GHG emissions than French suckler cow farms. For nitrogen use, previous studies found much higher *TIE* values for Dutch dairy farms.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Reinhard, Lovell, and Thijssen (1999) found a mean *TIE* of 55.9% for nitrogen whereas Lamkowsky, Oenema, Meuwissen, and Ang (2021) found a 50% productivity gap for nitrogen.

Our findings on the efficiency gains from land reallocation cannot be directly generalized to other livestock or crop farming types or to mixed farms. Hence, additional research will be needed in order to gain a better picture of the potential contribution of land optimization to a reduction of GHG emissions and an increase in production. Our study calls investigations that include GHG emission data for mixed farms in the Netherlands.

In terms of GHG reduction potential from dairy farms operating circular dairy principles, GHG emissions per farm could be reduced by 11.8% on average if the farm production is kept constant with current input and land use. However, the GHG emissions per farm could be reduced by only 4.5% on average if crop and livestock production would be expanded by 4.5% with constant input and land use. This implies that there is a trade-off between reducing GHG emissions while keeping production constant and reducing GHG emissions while at the same time expanding production. This trade-off between environmental and economic objectives has also been found for the dairy sector of other countries (Kirilova et al., 2022; Le, Jeffrey, and An, 2020).

Beyond the scope of circular dairy, circular farming principles advocate for plant-based products to be consumed by humans before feeding it to livestock animals. This calls for a dietary shift of consumers towards more plant-based products and non-ruminants meats, and less dairy beef and other dairy products. Such dietary changes could reduce the food-related GHG emission of dairy farming as well (Kesse-Guyot et al., 2021) through mechanisms like a Pigouvian meat tax (which is set to the social cost of externality effects) or green-label education for consumers (Katare, Lawing, Park, and Wetzstein, 2020).

6.6. Conclusions

This study modelled the intended production and residual GHG emissions on Dutch dairy farms with circular principles, by combining a network DEA model with the by-production approach. The results from the directional output distance function indicate that mean inefficiency levels for Dutch dairy farms are quite small, i.e. 4.5% on average. This shows that many Dutch dairy farms are already operating quite close to the frontier. Thus, there is only limited potential for GHG emission reduction through efficiency improvement.

Although specialized dairy farms in the Netherlands should allocate more land to crop production according to the reallocation model, the potential efficiency gain would on average only be 0.6%. Hence, there is limited potential for reducing GHG emissions and increasing production by optimizing land use. While we need to remain cautious with our interpretation as our sample did not include mixed farms, the results suggest that incentivizing specialized Dutch dairy farms to become more mixed is an ineffective policy instrument to mitigate GHG emissions.

Our modeling results suggest that the largest reduction potential for GHG emissions (11.8%) can be obtained when doing so while keeping the production of crop and livestock products, and all inputs use constant. The reduction potential for GHG emissions may be even higher if production (or herd size) is to be sacrificed as shown by Le et al. (2020). However, this will come at a higher private cost for farmers if they are required by regulations to reduce the on-farm GHG emissions. In that case, policy instruments that pertain cost-sharing between the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

government and dairy producers may be needed as suggested by Le et al. (2020) and Lötjönen, Temmes, and Ollikainen (2020).

This article is a first step to structurally incorporate circularity principles in efficiency analyses. We have several recommendations for future research. In the current study, there are no interactions between individual farms, nor are waste streams from non-farm entities considered, such as urban and industry waste. Future research should consider the potential of circularity in decoupling GHG emissions from farm production at a local and/or regional level. Moreover, the behavioral and managerial determinants of high economic performance and low levels of GHG emissions will need to be investigated. Finally, additional data on GHG emissions from mixed farms should be collected to validate the findings obtained here.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

7. ACKNOWLEDGEMENTS

The Report on modelling crop management practices and interfaces to the MIND STEP model toolbox is developed as part of the H2020 MIND STEP project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

8. REFERENCES

Ang, F., Kerstens, K., and Sadeghi, J. 2022. Energy productivity and greenhouse gas emission intensity in Dutch dairy farms: A Hicks–Moorsteen by-production approach under non-convexity and convexity with equivalence results. *Journal of Agricultural Economics*. doi:10.1111/1477-9552.12511.

Ang, F., and Kerstens, P. J. 2016. To mix or specialise? A coordination productivity indicator for English and Welsh farms. *Journal of Agricultural Economics,* 67(3):779-798.

Antle, J. M., and Capalbo, S.M. 2001. Econometric-process models for integrated assessment of agricultural production systems. *American Journal of Agricultural Economics*, 83(2):389-401.

Alvarez, A., and del Corral, J. 2010. Identifying different technologies using a latent class model: extensive versus intensive dairy farms. *European Review of Agricultural Economics*, 37(2):231-250.

Arellano, M., and Bonhomme, S. 2011. Nonlinear panel data analysis. *Annuan Review of Economics*, 3(1):395-424.

Arndt, C. 1999. Demand for herbicide in corn: an entropy approach using micro-level data. *Journal of Agricultural and Resource Economics* 24(1):204–221.

Arndt, C., Liu, S., and Preckel, P.V. 1999. On dual approaches to demand systems estimation in the presence of binding quantity constraints. *Applied Economics* 31(8):999–1008.

Babcock, B.A. 2015. Extensive and intensive agricultural supply response. *Annual Review of Resources Economics* 7:333–348.

Bayramoglu, B., and Chakir, R. 2016. The impact of high crop prices on the use of agro-chemical inputs in France: A structural econometric analysis. *Land Use Policy* 55:204–211.

Bertsekas, D.P. 1996. Constrained optimization and Lagrange Multiplier Methods. 2nd Ed. Athena Scientific, Belmont, Mass., 395p.

Böcker, T. and Finger, R. 2017. A Meta-Analysis on the Elasticity of Demand for Pesticides. *Journal of Agricultural Economics*, 68 (2):518-533.

Boetel, B. L., Hoffmann, R., and Liu, D.J. 2007. Estimating investment rigidity within a threshold regression framework: the case of US hog production sector. *American Journal of Agricultural Economics*, 89(1):36-51.

Boer, I.J.M.d., and Ittersum, M.K.v. 2018. *Circularity in Agricultural production*. Retrieved from https://edepot.wur.nl/470625.

Bowman, M.S., and Zilbermann, D. 2013. Economic factors affecting diversified farming. *Ecology and Society* 18(1):33.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Caffo, B.S., Jank, W., and Jones, G.L. 2005. Ascent-based Monte Carlo expectation–maximization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):235–251.

Carpentier, A., and Letort, E. 2012. Accounting for heterogeneity in multicrop micro-econometric models: implications for variable input demand modeling. *American Journal of Agricultural Economics*, *94*(1):209-224.

Carpentier, A., and Letort, E. 2014. Multicrop production models with Multinomial Logit acreage shares. *Environmental and Resource Economics* 59(4):537–559.

Centraal Bureau voor de Statistiek. 2022. Consumer prices; price index Retrieved from https://opendata.cbs.nl/#/CBS/en/dataset/83131ENG/table?ts=1660915009992.

Chakir, R., and Thomas, A. 2003. Simulated maximum likelihood estimation of demand systems with corner solutions and panel data application to industrial energy demand. *Revue d'Economie Politique* 113(6):773–799.

Chambers, R., and Just, R. 1989. Estimating Multiouput Technologies. *American Journal of Agricultural Economics* 71(4):980-95.

Chambers, R.G., Fāure, R., and Grosskopf, S. 1996. Productivity growth in APEC countries. *Pacific Economic Review,* 1(3):181-190.

Chavas, J.-P., and M.T. Holt. 1990. Acreage decisions under risk: the case of corn and soybeans. *American Journal of Agricultural Economics* 72(3):529–538.

Cherchi, E., and C.A. Guevara, 2012. A Monte Carlo experiment to analyze the curse of dimensionality in estimating random coefficients models with a full variance–covariance matrix. T*ransportation Research Part B: Methodological*, 46(2): 321-332.

Cherchye, L., De Rock, B., and Hennebel, V. (017. Coordination efficiency in multi-output settings: a DEA approach. *Annals of Operations Research, 250*(1):205-233. doi:10.1007/s10479-015-1892-7.

Cherchye, L., Rock, B.D., Dierynck, B., Roodhooft, F., and Sabbe, J. 2013. Opening the "Black Box" of Efficiency Measurement: Input Allocation in Multioutput Settings. *Operations Research,* 61(5):1148-1165.

Cox, T. L., and Wohlgenant, M.K. 1986. Prices and quality effects in cross-sectional demand analysis. *American Journal of Agricultural Economics,* 68(4):908-919.

Coyle, B.T. 1992. Risk Aversion and Price Risk in Duality Models of Production: A Linear Mean-Variance Approach. *American Journal of Agricultural Economics* 74(4):849-859.

Coyle, B.T. 1999. Risk Aversion and Yield Uncertainty in Duality Models of Production: A Linear Mean-Variance Approach. *American Journal of Agricultural Economics* 81(3):553 -567.

Dakpo, K.H., Jeanneaux, P., and Latruffe, L. 2016. Modelling pollution-generating technologies in performance benchmarking: Recent developments, limits and future prospects in the nonparametric framework. *European Journal of Operational Research,* 250(2):347-359. doi:https://doi.org/10.1016/j.ejor.2015.07.024.

Dakpo, K.H., Jeanneaux, P., and Latruffe, L. 2017. Greenhouse gas emissions and efficiency in French sheep meat farming: A non-parametric framework of pollution-adjusted technologies. *European Review of Agricultural Economics,* 44(1):33-65.

Dakpo, K.H., and Oude Lansink, A. 2019. Dynamic pollution-adjusted inefficiency under the by-production of bad outputs. *European Journal of Operational Research,* 276(1):202-211.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Darnhofer, I. 2014. Resilience and why it matters for farm management. *European Review of Agricultural Economics*, 41(3):461-484.

Delyon, B., Lavielle, M., and Moulines, E. 1999. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 27(1):94–128.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*:1–38

Devilliers, E., Carpentie, A., and Koutchadé, O.P. 2021. Uncovering adoption and characteristics of "low-input" versus "high-yielding" crop management practices for wheat production in France: a heterogeneous hidden Markov approach. Working paper, UMR SMART, INRAE, Institut Agro Rennes-Angers.

Dixon, B. L., Batte, M. T., and Sonka, S.T. 1984. Random coefficients estimation of average total product costs for multiproduct firms. *Journal of Business & Economic Statistics*, 2(4), 360-366.

Dixon, B.L., and Hornbaker, R.H. 1992. Estimating the technology coefficients in linear programming models. *American journal of agricultural economics*, 74(4), 1029-1039.

EUROSTAT. 2022. Agricultural Prices and Price Indices Retrieved from https://ec.europa.eu/eurostat/web/main/data/database.

Färe, R., Grabowski, R., Grosskopf, S., and Kraft, S. 1997. Efficiency of a fixed but allocatable input: A non-parametric approach. *Economics Letters,* 56(2):187-193.

Färe, R., and Whittaker, G. 1995. An intermediate input model of dairy production using complex survey data. *Journal of Agricultural Economics,* 46(2):201-213. doi:10.1111/j.1477-9552.1995.tb00766.x.

Fezzi, C., and Bateman, I.J. 2011. Structural agricultural land use modeling for spatial agro-environmental policy analysis. *American Journal of Agricultural Economics* 93(4):1168–1188

Fok, D., Van Dijk, D., and Franses, P. H. 2005. A multi-level panel STAR model for US manufacturing sectors. *Journal of Applied Econometrics*, 20(6):811-827.

Food and Agriculture Organization. 2019. *Climate change and the global dairy cattle sector: The role of the dairy sector in a low-carbon future*. Retrieved from https://www.fao.org/3/CA2929EN/ca2929en.pdf.

Førsund, F.R. 2009. Good Modelling of Bad Outputs: Pollution and Multiple-Output Production. *International Review of Environmental and Resource Economics, 3*(1), 1-38. doi:10.1561/101.00000021.

Gardebroek, C., and Oude Lansink, A.G. 2004. Farm-specific adjustment costs in Dutch pig farming. *Journal of Agricultural Economics*, 55(1):3-24.

Girardin, P. 1998. Ecophysiologie du maïs. *AGPPM-Association Générale des Producteurs de Maïs* ed.

González, A., Teräsvirta, T., and Van Dijk, V. 2005. Panel smooth transition regression models. SSE/EFI *Working Paper Series in Economics and Finance*, No. 604.

Guyomard, H., Baudry, M., and Carpentier, A. 1996. Estimating crop supply response in the presence of farm programmes: application to the CAP. *European Review of Agricultural Economics* 23(4):401–420.

Hallam, D., Bailey, A., Jones, P., and Errington, A. (1999). Estimating input use and production costs from farm survey panel data. *Journal of Agricultural Economics*, *50*(3), 440-449.

Hamermesh, D. S., and Pfann, G. A. (1996). Adjustment costs in factor demand. *Journal of Economic literature*, 34(3):1264-1292.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Hansen, Bruce E. (2000) Sample splitting and threshold estimation. *Econometrica*, 68(3):575-603.

Harding, M.C., and Hausman, J. 2007. Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. *International Economic Review* 48(4):1311-1328.

Heckelei, T., Britz, W., and Zhang, Y. 2012. Positive mathematical programming approaches–recent developments in literature and applied modelling. *Bio-Based and Applied Economics* 1(1):109–124.

Heckelei, T., Mittelhammer, R.C., and Jansson, T. 2008. *A Bayesian alternative to generalized cross entropy solutions for underdetermined econometric models* (No. 1548-2016-132440).

Heckeleï, T., and Wolff, H. 2003. Estimation of constrained optimisation models for agricultural supply analysis based on generalised maximum entropy. *European Review of Agricultural Economics* 30(1):27-50.

Holt, M.T. 1999. A Linear Approximate Acreage Allocation Model. *Journal of Agricultural and Resource Economics* 24:383 -397.

Hornbaker, R.H., Dixon, B.L., and Sonka, S.T. 1989. Estimating production activity costs for multioutput firms with a random coefficient regression model. *American Journal of Agricultural Economics*, *71*(1), 167-177.

Hou, Y., Bai, Z., Lesschen, J. P., Staritsky, I. G., Sikirica, N., Ma, L., . . . Oenema, O. 2016. Feed use and nitrogen excretion of livestock in EU-27. *Agriculture, Ecosystems and Environment, 218*, 232-244. doi:https://doi.org/10.1016/j.agee.2015.11.025

Howitt, R.E. 1995. Positive mathematical programming. *American journal of agricultural economics*, 77(2):329-342.

James, J., 2018. Estimation of factor structured covariance mixed logit models. *Journal of Choice Modelling*, 28: 41-55.

Just, R.E., Zilberman, D., and Hochman, E. 1983. Estimation of Multicrop Production Functions. *American Journal of Agricultural Economics* 65(4):770-780.

Just, R.E., Zilberman, D., Hochman, E., and Bar-Shira, Z. 1990. Input Allocation in Multicrop Systems. *American Journal of Agricultural Economics* 72(1):200-209.

Katare, B., Lawing, J., Park, T., and Wetzstein, M. 2020. Toward Optimal Meat Consumption. *American Journal of Agricultural Economics,* 102(2), 662-680. doi:10.1002/ajae.12016

Kao, C., Lee, L., and Pitt, M.M. 2001. Simulated maximum likelihood estimation of the linear expenditure system with binding non-negativity constraints. *Annals of Economics and Finance* 2(1):203–223.

Keane, M.P. 2009. Simulated maximum likelihood estimation based on first-order conditions. *International Economic Review* 50:627–675.

Kesse-Guyot, E., Fouillet, H., Baudry, J., Dussot, A., Langevin, B., Allès, B., . . . Pointereau, P. 2021. Halving food-related greenhouse gas emissions can be achieved by redistributing meat consumption: Progressive optimization results of the NutriNet-Santé cohort. *Science of the Total Environment,* 789. doi:10.1016/j.scitotenv.2021.147901

Kirilova, E., Vaklieva-Bancheva, N., Vladova, R., Petrova, T., Ivanov, B., Nikolova, D., and Dzhelil, Y. 2022. An approach for a sustainable decision-making in product portfolio design of dairy supply chain in terms of environmental, economic and social criteria. *Clean Technologies and Environmental Policy,* 24(1), 213-227. doi:10.1007/s10098-021-02110-2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Koutchadé, O.P., Carpentier, A., and Femenia, F. 2018. Modeling heterogeneous farm responses to EU biofuel support with a random parameter multi-crop model. *American Journal of Agricultural Economics* 100(2):434-455.

Koutchadé, O. P., Carpentier, A., and Femenia, F. 2021. Modeling Corners, Kinks, and Jumps in Crop Acreage Choices: Impacts of the EU Support to Protein Crops. American Journal of Agricultural Economics, 103(4):1502-1524.

Koutchadé, O.P., Carpentier, A. and Femenia F., 2015a. Accounting for unobserved heterogeneity in agricultural production choice models: a random parameter approach. Conference of the International Association of Applied Econometrics. Thessalonik, Greece, June 2015.

Koutchadé, O. P., Carpentier, A. and Femenia F., 2015b. Empirical modelling of production decisions of heterogeneous farmers with mixed models. Annual Meeting of the Agricultural and Applied Economics Association. Accepted oral contribution. San Francisco, CA, July 2015.

Kremen, C., Iles A., and Bacon, C. 2012. Diversified farming systems: an agroecological, systems-based alternative to modern industrial agriculture. Ecology and Society 17(4):44.

Kuhn, E., and Lavielle, M. 2005. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* 49(4):1020–1038.

Kumbhakar, S.C., Tsionas, E.G., and Sipiläinen, T. 2009. Joint estimation of technology choice and technical efficiency: an application to organic and conventional dairy farming. Journal of *Productivity Analysis*, 31(3):151-161.

Lacroix, A., and Thomas, A. 2011. Estimating the environmental impact of land and production decisions with multivariate selection rules and panel data. *American Journal of Agricultural Economics* 93(3):784–802.

Lamkowsky, M., Oenema, O., Meuwissen, M. P. M., and Ang, F. 2021. Closing productivity gaps among Dutch dairy farms can boost profit and reduce nitrogen pollution. *Environmental Research Letters 16 (2021) 12; ISSN: 1748-9318*. Retrieved from https://edepot.wur.nl/558591

Lavielle, M. 2014. *Mixed effects models for the population approach: models, tasks, methods and tools*. New York: Chapman and Hall/CRC.

Le, S., Jeffrey, S., and An, H. 2020. Greenhouse Gas Emissions and Technical Efficiency in Alberta Dairy Production: What Are the Trade-Offs? *Journal of Agricultural and Applied Economics,* 52(2): 177-193. doi:10.1017/aae.2019.41

Lee, L.F., and Pitt, M.M. 1987. Microeconometric models of rationing, imperfect markets, and non-negativity constraints. *Journal of Econometrics* 36(1-2):89–110.

Lee, L.F., and Pitt, M.M. 1986. Microeconometric demand system with binding nonnegativity constraints: the dual approach. *Econometrica* 54(5):1237–1242.

Lin, B.B. 2011. Crop Diversification: Adaptative Management for Environmental Change. *BioScience* 61(3):183–193.

Lötjönen, S., Temmes, E., and Ollikainen, M. 2020. Dairy Farm Management when Nutrient Runoff and Climate Emissions Count. *American Journal of Agricultural Economics,* 102(3):960-981. doi:10.1002/ajae.12003

Matson, P.A., Parton, W.J., Power, A.G., and Swift, M.J. 1997. Agricultural Intensification and Ecosystem Properties. *Science* 277:504-508.

McDonald, J.F., and Moffitt, R.A. 1980. The uses of Tobit analysis. *The Review of Economics and Statistics* 62(2):318–321.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

McLachlan, G., and Krishnan, T. 2007. *The EM algorithm and extensions*, 2nd. ed. New York: John Wiley & Sons.

Mérel, P., and Howitt, R. 2014. Theory and application of positive mathematical programming in agriculture and the environment. *Annual Review of Resources Economics* 6(1):451-470.

Millimet, D.L., and Tchernis, R. 2008. Estimating high-dimensional demand systems in the presence of many binding non-negativity constraints. *Journal of Econometrics* 147(2): 384–385.

Ministerie van Landbouw Natuur en Voedselkwaliteit. 2019. Realisatieplan Visie LNV Op weg met nieuw perspectief. Retrieved from https://www.rijksoverheid.nl/documenten/publicaties/2019/06/17/realisatieplan-visie-lnv-op-weg-met-nieuw-perspectief

Moore, M., Gollehon, N. R., and Carey, M.B. 1994. Multicrop Production Decisions in Western Irrigated Agriculture: The Role of Water Price. *American Journal of Agricultural Economics* 76(4):385-395.

Moore, M.R., and Negri, D.H. 1992. A multicrop production model of irrigated agriculture, applied to water allocation policy of the Bureau of Reclamation. *Journal of Agricultural and Resource Economics* 17(1):29–43.

Moxey, A., and Tiffin, R. 1994. Estimating linear production coefficients from farm business survey data: A note. *Journal of Agricultural Economics*, 45(3), 381-385.

Murty, S., Russell, R. R., and Levkoff, S.B. 2012. On modeling pollution-generating technologies. *Journal of Environmental Economics and Management,* 64(1):117.

Önel, G. (2018a). Adjustment costs and threshold effects in factor demand relationships. Applied Economics, 50(18):2070-2086.

Önel, G. (2018b). An implicit model of adjustment costs in differential input demand systems. Theoretical and Applied Economics, 25(2).

Oude Lansink, A. 2008. Area allocation under price uncertainty on Dutch arable farms. *Journal of Agricultural Economics* 50(1): 93-105.

Oude Lansink, A., and Peerlings, J. 1996. Modelling the new EU cereals and oilseeds regime in the Netherlands. *European Review of Agricultural Economics* 23(2):161-178.

Ozarem, P.F., and Miranowski, J.A. 1994. A Dynamic Model of Acreage Allocation with General and Crop-Specific Soil Capital. *American Journal of Agricultural Economics* 76(3):385-395.

Panhard, X. and Samson, A. 2009. Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, 10 (1): 121-35.

Peyraud, J. L., van den Pol, A., Dillon, P., and Delaby, L. 2010. Producing milk from grazing to reconcile economic and environmental performances. In 23th General Meeting of the European Grassland Federation, Kiel, Germany, 29 august-02 September, 2010 (pp. 163-164).

Pietola, K. S., and Myers, R. J. 2000. Investment under uncertainty and dynamic adjustment in the Finnish pork industry. *American Journal of Agricultural Economics*, 82(4):956-967.

Platoni, S., Sckokai, P., and Moro, D. 2012. Panel data estimation techniques and farm-level data models. *American Journal of Agricultural Economics,* 94(5):1202-1217.

Pudney, S. 1989. *Modelling Individual Choice: the Econometrics of Corners, Kinks and Holes.* Oxford: Basil Blackwell.

Ray, S. C. 1985. Methods of estimating the input coefficients for linear programming models. *American journal of agricultural economics*, 67(3): 660-665.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Reidsma, P., Ewert, F., Lansink, A. O., and Leemans, R. (2010). Adaptation to climate change and climate variability in European agriculture: the importance of farm level responses. *European journal of agronomy*, 32(1):91-102.

Reinhard, S., Lovell, C. A. K., and Thijssen, G. 1999. Econometric Estimation of Technical and Environmental Efficiency: An Application to Dutch Dairy Farms. *American Journal of Agricultural Economics,* 81(1):44-60.

Renner, S., Sauer, J., and El Benni, N. 2021. Why considering technological heterogeneity is important for evaluating farm performance? *European Review of Agricultural Economics*, 48(2):415-445.

Rijksoverheid. 2022. Wat is het Klimaatakkoord? Retrieved from https://www.rijksoverheid.nl/onderwerpen/klimaatverandering/klimaatakkoord/wat-is-het-klimaatakkoord

Robert, M., Thomas, A., and Bergez, J.E. 2016. Processes of adaptation in farm decision-making models. A review. *Agronomy for sustainable development*, 36(4):64.

Röhm, O., and Dabbert, S. 2003. An extension of Positive Mathematical Programming. *American Journal of Agricultural Economics* 85(1):254–265.

Ruud, P.A. 1991. Extensions of estimation methods using the EM algorithm. *Journal of Econometrics* 49(3):305-341.

Sauer, J., and Paul, C.J.M. 2013. The empirical identification of heterogeneous technologies and technical change. *Applied Economics*, 45(11):1461-1479.

Serra, T., Zilberman, D., Goodwin, B.K., and Hyvonen, K. 2005. "Replacement of agricultural price supports by area payments in the European Union and the effects on pesticide use." *American Journal of Agricultural Economics* 87(4):870–884.

Serra, T., Chambers, R. G., and Oude Lansink, A.G.J.M. 2014. Measuring technical and environmental efficiency in a state-contingent technology. *European Journal of Operational Research, 236*(2):706-717.

Sckokai, P., and Moro, D. 2006. "Modeling the reforms of the common agricultural policy for arable crops under uncertainty." *American Journal of Agricultural Economics* 88(1):43–56.

Sckokai, P., and Moro, D. 2009. "Modelling the impact of the CAP Single Farm Payment on farm investment and output." *European Review of Agricultural Economics* 36(2): 395–423.

Shephard, R.W. 1977. Dynamic Indirect Production Functions. In *Mathematical Economics and Game Theory* (pp. 418-434): Springer.

Shephard, R.W. 2012. *Cost and production functions* (Vol. 194): Springer Science and Business Media.

Shonkwiler, J.S., and Yen, S.T. 1999. "Two-Step Estimation of a Censored System of Equations." *American Journal of Agricultural Economics* 81(3):972-982.

Stanfield, P.M., Wilson, J.R., Mirka, G.A., Glasscock, N.F., Psihogios, J.P., and Davis, J.R. 1996. "Multivariate input modeling with Johnson distributions." *Proceedings of the 28th conference on Winter simulation, IEEE Computer Society*, pp. 1457–1464.

Suh, D. H., and Moss, C. B. 2017. Decompositions of corn price effects: implications for feed grain demand and livestock supply. *Agricultural Economics*, 48(4):491-500.

Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R. and S. Polasky. 2002. "Agricultural sustainability and intensive production practices." *Nature* 418:671–677.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

van der Meer, R.W., Ge, L., and van der Veen, H.B. (2019). Sample for the Dutch FADN 2016. (Wageningen Economic Research rapport; No. 2019-020). Wageningen Economic Research. doi:https://doi.org/10.18174/471865

Van Grinsven, H.J., van Eerdt, M.M., Westhoek, H., and Kruitwagen, S. 2019. Benchmarking eco-efficiency and footprints of Dutch agriculture in European context and implications for policies for climate and environment. *Frontiers in Sustainable Food Systems, 3*, 13.

Van Zanten, H. H., Herrero, M., Van Hal, O., Röös, E., Muller, A., Garnett, T., . . . De Boer, I.J. 2018. Defining a land boundary for sustainable livestock consumption. *Global Change Biology, 24*(9), 4185-4194.

Wageningen University and Research. 2022. *Circular agriculture - A new perspective for Dutch agriculture*. Retrieved from https://www.wur.nl/nl/show/circular-agriculture-a-new-perspective-for-dutch-agriculture-2.htm

Wales, T.J., and, Woodland, A.D. 1983. “Estimation of consumer demand systems with binding non-negativity constraints.” *Journal of Econometrics* 21(3):263-285.

Wei, G.C., and Tanner, M.A. 1990. “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms.” *Journal of the American Statistical Association* 85: 699-704.

Wooldridge, J.M. 2002. Econometric analysis of cross section and panel data MIT press. *Cambridge, ma*, *108*(2): 245-254.

Wu, C.J. 1983. “On the convergence properties of the EM algorithm.” *The Annals of Statistics* 11(1):95–103.

Zhu, L., and Lansink, A.O. 2022. Dynamic sustainable productivity growth of Dutch dairy farming. *PLoS ONE 17 (2022) 2; ISSN: 1932-6203*. Retrieved from https://edepot.wur.nl/566548

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

APPENDIX 2A: SAEM ALGORITHM STRUCTURE

The aim of our estimation procedure is to compute the ML estimator of $\boldsymbol{\theta}_0$ or, at least, an estimator that is asymptotically equivalent to this estimator which is practically "infeasible". The ML estimator of $\boldsymbol{\theta}_0$ is the solution in $\boldsymbol{\theta}$ to the ML problem $\max_{\boldsymbol{\theta}} L_N(\boldsymbol{\theta})$ where $L_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \int \left(\prod_{t=1}^{T} f(\mathbf{w}_{it}^{+}, r_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}; \boldsymbol{\delta}, \boldsymbol{\Psi}) \right) \varphi(\boldsymbol{\gamma} - \boldsymbol{\mu}; \boldsymbol{\Omega}) d\boldsymbol{\gamma}$. EM algorithms were proposed by (Dempster *et al*, 1977) for maximizing the likelihood function of models involving missing information, of which random parameter models are typical examples. Extensions of the original EM algorithm were then proposed for overcoming limitations of this algorithm (e.g., McLachlan and Krishnan, 2007; Lavielle, 2014), including issues such as those raised by the integration of the likelihood function of our model.

EM type algorithms are constructed based on the expectation conditional on the "observed data" of the "complete data" sample log-likelihood function of the considered model. Contribution of farmer *i* to the likelihood function of the model corresponds to the pdf of her/his sequence of production choices conditional on the sequence of exogenous variables characterizing this choice sequence. This choice sequence is given by $(\mathbf{w}_i^{+}, \mathbf{r}_i)$ with $\mathbf{w}_i^{+} = (\mathbf{w}_{it}^{+} : t = 1,...,T)$ and $\mathbf{r}_i = (r_{it} : t = 1,...,T)$ and the corresponding conditioning set by $\mathbf{z}_i \equiv (\mathbf{z}_{it} : t = 1,...,T)$. The complete – observed and unobserved modelled variables – data related to farmer *i* consist of her/his observed production choice sequence, $(\mathbf{w}_i^{+}, \mathbf{r}_i)$, and her/his specific parameter vector, $\boldsymbol{\gamma}_i$. The complete data log-likelihood function of our model is thus given by:

$$\text{(2A.1)} \qquad \sum_{i=1}^{N} \ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i) = \sum_{i=1}^{N} \sum_{t=1}^{T} \ln f(\mathbf{w}_{it}^{+}, r_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\delta}, \boldsymbol{\Psi}) + \sum_{i=1}^{N} \ln \varphi(\boldsymbol{\gamma}_i; \boldsymbol{\mu}, \boldsymbol{\Omega}).$$

The "observed data" related to farmer *i*, thereafter denoted by $\boldsymbol{\kappa}_i$, consist of her/his observed production choice sequence, $(\mathbf{w}_i^{+}, \mathbf{r}_i)$, and of the exogenous variables conditioning these choice sequence, $\mathbf{z}_i$. That is to say, $\boldsymbol{\kappa}_i = (\mathbf{w}_i^{+}, \mathbf{r}_i, \mathbf{z}_i)$. According to our notations function $f(\boldsymbol{\gamma}_i \mid \boldsymbol{\kappa}_i; \boldsymbol{\theta}_0)$ denotes the density of $\boldsymbol{\gamma}_i$ conditional on $\boldsymbol{\kappa}_i$. Let function

$$\text{(2A.2)} \qquad E_{(\bar{\boldsymbol{\theta}})}[\ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i) \mid \boldsymbol{\kappa}_i] = \int \ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i) f(\boldsymbol{\gamma} \mid \boldsymbol{\kappa}_i; \bar{\boldsymbol{\theta}}) d\boldsymbol{\gamma}$$

denote the expectation of the $\ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i)$ conditional on $\boldsymbol{\kappa}_i$ based on the pdf $f(\boldsymbol{\gamma}_i \mid \boldsymbol{\kappa}_i; \bar{\boldsymbol{\theta}})$ where $\bar{\boldsymbol{\theta}}$ is a candidate estimate of $\boldsymbol{\theta}_0$. Function

$$\text{(2A.3)} \qquad Q(\boldsymbol{\theta} \mid \bar{\boldsymbol{\theta}}) = \sum_{i=1}^{N} E_{(\bar{\boldsymbol{\theta}})}[\ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i) \mid \boldsymbol{\kappa}_i]$$

defines the expectation of the complete data sample log-likelihood function $\sum_{i=1}^{N} \ln \ell_i^{C}(\boldsymbol{\theta}, \boldsymbol{\gamma}_i)$ conditional on $(\boldsymbol{\kappa}_i : i = 1,...,N)$ based on the pdfs $f(\boldsymbol{\gamma}_i \mid \boldsymbol{\kappa}_i; \bar{\boldsymbol{\theta}})$ for $i = 1,...,N$. This function, which can be interpreted as a well-behaved proxy of $L_N(\boldsymbol{\theta})$ when $\bar{\boldsymbol{\theta}}$ is suitably chosen, is the "engine" of EM type algorithms that can be used for estimating $\boldsymbol{\theta}_0$.

SAEM algorithms iterate three steps until numerical convergence: a Simulation (S) step, an

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

Approximation (A) step and a Maximization (M) step. They generate sequences of estimates of $\boldsymbol{\theta}_0$ that converge to maxima of $L_N(\boldsymbol{\theta})$ under mild assumptions, thereby allowing to compute ML estimators of $\boldsymbol{\theta}_0$ (Delyon *et al*, 1999; Kuhn et Lavielle, 2005; Lavielle, 2014). Assuming that the estimate of $\boldsymbol{\theta}_0$ obtained at the end of iteration n is given by $\boldsymbol{\theta}_{(n)}$, our SAEM algorithm proceeds as follows at iteration $n+1$.

The S step consists of integrating terms $E_{(n)}[\ln \ell_i^C(\boldsymbol{\theta},\boldsymbol{\gamma}_i)\,|\,\boldsymbol{\kappa}_i] = \int \ln \ell_i^C(\boldsymbol{\theta},\boldsymbol{\gamma}_i) f(\boldsymbol{\gamma}\,|\,\boldsymbol{\kappa}_i;\boldsymbol{\theta}_{(n)})d\boldsymbol{\gamma}$ with simulation methods for $i=1,\dots,N$. Building on the work of Caffo *et al* (2005), we use an Importance Sampling approach. Terms $E_{(n)}[\ln \ell_i^C(\boldsymbol{\theta},\boldsymbol{\gamma}_i)\,|\,\boldsymbol{\kappa}_i]$ are approximated with simulated weighted sums $\sum_{j=1}^{J_{(n)}} \tilde{w}_{i,(n)}^j \ln \ell_i^C(\boldsymbol{\theta},\tilde{\boldsymbol{\gamma}}_{i,(n)}^j)$ where terms $\tilde{\boldsymbol{\gamma}}_{i,(n)}^j$ are independent random draws from $\mathcal{N}(\boldsymbol{\mu}_{(n)},\boldsymbol{\Omega}_{(n)})$ while terms

$$\text{(2A.4)} \qquad \tilde{w}_{i,(n)}^j = \frac{\prod_{t=1}^{T} f(\mathbf{w}_{it}^+, r_{it}\,|\,\mathbf{z}_{it},\tilde{\boldsymbol{\gamma}}_{i,(n)}^j;\hat{\boldsymbol{\delta}}_{(n)},\hat{\boldsymbol{\Psi}}_{(n)})}{\sum_{j=1}^{J_{(n)}} \prod_{t=1}^{T} f(\mathbf{w}_{it}^+, r_{it}\,|\,\mathbf{z}_{it},\tilde{\boldsymbol{\gamma}}_{i,(n)}^j;\hat{\boldsymbol{\delta}}_{(n)},\hat{\boldsymbol{\Psi}}_{(n)})}$$

are their corresponding normalized importance weights, for $j=1,\dots,J_{(n)}$. Other proposed densities are more efficient than that of $\mathcal{N}(\boldsymbol{\mu}_{(n)},\boldsymbol{\Omega}_{(n)})$ but are more difficult to draw from.

The A step consists of constructing function $\tilde{Q}_{(n)}(\boldsymbol{\theta})$, the stochastic approximation of $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_{(n)})$, by using the following recursive formula:

$$\text{(2A.5)} \qquad \tilde{Q}_{(n)}(\boldsymbol{\theta}) = (1-\lambda_{(n)})\tilde{Q}_{(n-1)}(\boldsymbol{\theta}) + \lambda_{(n)} \sum_{i=1}^{N} \sum_{j=1}^{J_{(n)}} \tilde{w}_{i,(n)}^j \ln \ell_i^C(\boldsymbol{\theta},\tilde{\boldsymbol{\gamma}}_{i,(n)}^j).$$

Delyon *et al* (1999) and Kuhn and Lavielle (2005) provide guidelines for suitably choosing the sequence of weight terms $\lambda_{(n)}$, which must lie in $(0,1]$. Large values of $\lambda_{(n)}$ allow to explore the parameter space and yield a quick convergence to the neighborhood of a solution to the ML problem. But they also imply large simulation noise. Reducing the value of $\lambda_{(n)}$ reduces the simulation noise and allow the algorithm to converge in the neighborhood of a solution to the ML problem.

Kuhn and Lavielle (2005) also provide guidelines for suitably choosing the number of draws $J_{(n)}$. Importantly, large numbers of random draws are not needed at each iteration since function $\tilde{Q}_{(n)}(\boldsymbol{\theta})$, by construction, reuses the random draws obtained in previous iterations. Indeed, “recycling” previous iteration draws is a major advantage of SAEM algorithms over their competing alternatives such as MCEM algorithms (Delyon *et al*, 1999; Lavielle, 2014). As a matter of fact, the SAEM algorithm presented here performs significantly better that its MCEM counterpart, which corresponds to the SAEM algorithm with $\lambda_{(n)}=1$, in our empirical application.

The M step consists of updating the estimate of $\boldsymbol{\theta}_0$ by computing $\boldsymbol{\theta}_{(n+1)}$. This updated estimate is defined either as:

$$\text{(2A.6)} \qquad \boldsymbol{\theta}_{(n+1)} = \arg\max_{\boldsymbol{\theta}} \tilde{Q}_{(n)}(\boldsymbol{\theta})$$

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 817566.

or as any $\boldsymbol{\theta}_{(n+1)}$ such that condition $\tilde{L}_{(n)}(\boldsymbol{\theta}_{(n+1)}) \geq \tilde{L}_{(n)}(\boldsymbol{\theta}_{(n)})$ holds.

The main advantage of SAEM algorithms, and of other EM type algorithms, for maximizing the log-likelihood function of random parameters models is due to the following decomposition of $\tilde{L}_{(n)}(\boldsymbol{\theta})$:

(2A.7) $$\tilde{L}_{(n)}(\boldsymbol{\theta}) = \tilde{W}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) + \tilde{M}_{(n)}(\boldsymbol{\mu},\boldsymbol{\Omega})$$

where:

(2A.8) $$\tilde{W}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) = (1-\lambda_{(n)})\tilde{W}_{(n-1)}(\boldsymbol{\delta},\boldsymbol{\Psi}) + \lambda_{(n)}\sum_{i=1}^{N}\sum_{j=1}^{J_{(n)}}\tilde{\omega}^{j}_{i,(n)}\sum_{t=1}^{T}\ln f(\mathbf{w}^{+}_{it}, r_{it} \mid \mathbf{z}_{it}, \tilde{\boldsymbol{\gamma}}^{j}_{i,(n)}; \boldsymbol{\delta},\boldsymbol{\Psi})$$

and:

(2A.9) $$\tilde{M}_{(n)}(\boldsymbol{\mu},\boldsymbol{\Omega}) = (1-\lambda_{(n)})\tilde{M}_{(n-1)}(\boldsymbol{\mu},\boldsymbol{\Omega}) + \lambda_{(n)}\sum_{i=1}^{N}\sum_{j}^{J_{(n)}}\tilde{\omega}^{j}_{i,(n)}\ln\varphi(\tilde{\boldsymbol{\gamma}}^{j}_{i,(n)} - \boldsymbol{\mu}; \boldsymbol{\Omega}).$$

This decomposition enables us to separately update parameters $(\boldsymbol{\mu},\boldsymbol{\Omega})$ and $(\boldsymbol{\delta},\boldsymbol{\Psi})$. Moreover, term $(\boldsymbol{\mu}_{(n+1)},\boldsymbol{\Omega}_{(n+1)}) = \arg\max_{\boldsymbol{\theta}} \tilde{M}_{(n)}(\boldsymbol{\mu},\boldsymbol{\Omega})$ can be obtained in analytical closed form based on the "sufficient statistic approach" proposed by Delyon *et al* (1999).

Maximizing $\tilde{W}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi})$ in $(\boldsymbol{\delta},\boldsymbol{\Psi})$ appears to be much more difficult due to the functional form of $f(\mathbf{w}^{+}_{it}, r_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\eta})$. Yet, function $\tilde{W}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi})$ can be rewritten as:

(2A.10) $$\tilde{W}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) = \tilde{W}^{yxs}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) + \tilde{W}^{r}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi})$$

where:

(2A.11) $$\tilde{W}^{yxs}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) = (1-\lambda_{(n)})\tilde{W}^{yxs}_{(n-1)}(\boldsymbol{\delta},\boldsymbol{\Psi}) + \lambda_{(n)}\sum_{i=1}^{N}\sum_{j=1}^{J_{(n+1)}}\tilde{\omega}^{j}_{i,(n)}\sum_{t=1}^{T}\ln f(\mathbf{w}^{+}_{it} \mid \mathbf{z}_{it}, \tilde{\boldsymbol{\gamma}}^{j}_{i,(n)}; \boldsymbol{\delta},\boldsymbol{\Psi})$$

and:

(2A.12) $$\tilde{W}^{r}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) = (1-\lambda_{(n)})\tilde{W}^{r}_{(n-1)}(\boldsymbol{\delta},\boldsymbol{\Psi}) + \lambda_{(n)}\sum_{i=1}^{N}\sum_{j=1}^{J_{(n+1)}}\tilde{\omega}^{j}_{i,(n)}\sum_{t=1}^{T}P(r_{it} \mid \mathbf{z}_{it}, \tilde{\boldsymbol{\gamma}}^{j}_{i,(n)}, \mathbf{s}^{+}_{it}; \boldsymbol{\delta},\boldsymbol{\Psi}).$$

Terms $\ln f(\mathbf{w}^{+}_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}_i; \boldsymbol{\delta},\boldsymbol{\Psi})$ are defined – up to an additive term that doesn't depend on $(\boldsymbol{\delta},\boldsymbol{\Psi})$ – as log-likelihood functions at $(\boldsymbol{\delta},\boldsymbol{\Psi})$ of a Gaussian Seemingly Unrelated (linear) Regression (SUR) system with dependent variables missing at random (conditionally on $(\mathbf{z}_{it}, \boldsymbol{\gamma}_i)$). Ruud (1991) discussed the use of EM algorithms for computing ML estimators of models with latent Gaussian SUR systems. Based on Ruud's insights, we devised a simple EM type procedure aimed at obtaining values of $(\boldsymbol{\delta},\boldsymbol{\Psi})$ ensuring that condition $\tilde{W}^{yxs}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi}) \geq \tilde{W}^{yxs}_{(n)}(\boldsymbol{\delta}_{(n)},\boldsymbol{\Psi}_{(n)})$ holds. This procedure defines $(\boldsymbol{\delta}_{(n+1)},\boldsymbol{\Psi}_{(n+1)})$, which, together with $(\boldsymbol{\mu}_{(n+1)},\boldsymbol{\Omega}_{(n+1)})$, completes $\boldsymbol{\theta}_{(n+1)}$.

We don't consider function $\tilde{W}^{r}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi})$ when updating the estimate of $(\boldsymbol{\delta}_0,\boldsymbol{\Psi}_0)$ because this function is an awkward function of $(\boldsymbol{\delta},\boldsymbol{\Psi})$. It involves the regime choice probability functions $P(r_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}_i, \mathbf{s}^{+}_{it}; \boldsymbol{\delta},\boldsymbol{\Psi})$. Yet, our ignoring $\tilde{W}^{r}_{(n)}(\boldsymbol{\delta},\boldsymbol{\Psi})$ in the construction of $(\boldsymbol{\delta}_{(n+1)},\boldsymbol{\Psi}_{(n+1)})$ doesn't ensure that condition $\tilde{L}_{(n)}(\boldsymbol{\theta}_{(n+1)}) \geq \tilde{L}_{(n)}(\boldsymbol{\theta}_{(n)})$ holds whereas it is necessary for the convergence

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

of the SAEM algorithm. We devised a simple heuristic for coping with cases where $\boldsymbol{\theta}_{(n+1)}$ doesn't succeed in increasing $\tilde{Q}_{(n)}(\boldsymbol{\theta})$ from $\tilde{Q}_{(n)}(\boldsymbol{\theta}_{(n)})$. Yet, this heuristic was rarely activated when running this SAEM algorithm for estimating the ESR-MEMC model considered in our application. Two explanations can be put forward. If regime choice probability functions $P(r_{it} \mid \mathbf{z}_{it}, \boldsymbol{\gamma}_i, \mathbf{s}_{it}^{+}; \boldsymbol{\delta}, \boldsymbol{\Psi})$ don't have any "active" role when computing $(\boldsymbol{\delta}_{(n+1)}, \boldsymbol{\Psi}_{(n+1)})$, they have important "passive" roles though their effects as elements of IS weights $\tilde{w}_{i,(n)}^{j}$. Also, the recursive structure of the considered ERS-MEMC model implies that most statistical information needed to estimate $(\boldsymbol{\delta}_0, \boldsymbol{\Psi}_0)$ is contained in farmers' crop level choices that are considered in $\tilde{W}_{(n)}^{yxs}(\boldsymbol{\delta}, \boldsymbol{\Psi})$. Parameter $(\boldsymbol{\delta}_0, \boldsymbol{\Psi}_0)$ only impacts regime choices through its effects on the expected crop profitability levels.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

APPENDIX 3A: APPROACHES USED TO ESTIMATE THE INPUT COSTS ALLOCATION MODEL

In this Appendix, for estimation purposes, we rewrite the model in the following compact form:

(3A.1) $\bar{x}_{it} = \mathbf{s}'_{it}\mathbf{x}_{it} + u_{it}$

(3A.2) $\mathbf{x}_{it} = \mathbf{h}(\boldsymbol{\mu}_{it})$

(3A.3) $\boldsymbol{\mu}_{it} = \boldsymbol{\beta}_i + \mathbf{Z}_{(i)}\boldsymbol{\delta} + \boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{\beta}_i = \boldsymbol{\omega} + \bar{\mathbf{Z}}_i\boldsymbol{\pi} + \boldsymbol{\eta}_i$

where $u_{it} \sim_{iid} \mathcal{N}(0,\sigma_0^2)$, $\boldsymbol{\eta}_i \sim_{iid} \mathcal{N}(\mathbf{0},\boldsymbol{\psi}_0)$ and $\boldsymbol{\varepsilon}_{it} \sim_{iid} \mathcal{N}(\mathbf{0},\boldsymbol{\Omega}_0)$, and $u_{it} \perp \boldsymbol{\eta}_i \perp \boldsymbol{\varepsilon}_{it} \perp \mathbf{s}_{it}$.

The considered model is full parametric, and the parameters to be estimated are: $\boldsymbol{\theta} = (\boldsymbol{\omega},\boldsymbol{\pi},\boldsymbol{\delta},\sigma^2,\boldsymbol{\psi},\boldsymbol{\Omega})$. Under some regularity conditions and using identity specification for the transformation h, it is possible to obtain consistent estimates of $\boldsymbol{\theta}$ using a Generalized Least Square (GLS) approach applied to data that contain more observations (i.e. more time periods) for each farmer than the number of crops to which her/his variable inputs are to be allocated. Yet, this condition ($T \geq C$) may not be verified in empirical applications using farmer panel data. Other problems, such that the multi-collinearity of $\mathbf{s}_{it}$ due to acreage choices complementarity and the heteroscedastic form of the error term of the model may make tedious the identification of the model parameters with standard approaches. As explained below, our estimation approach allows tackling these issues. We propose here to use Maximum Likelihood estimation approach via an extension of EM algorithm.

Intermediary results

Here, we define some intermediate results that we will need in the estimation section. Let define, as in previous section, $\boldsymbol{\mu}_{(i)} = (\mu_{c,it} : t \in \mathcal{H}_{(i)}, c \in \mathcal{C})$ such that:

(3A.4) $\boldsymbol{\mu}_{(i)} = \boldsymbol{\iota}_{T_{(i)}} \otimes \boldsymbol{\beta}_i + \mathbf{Z}_{(i)}\boldsymbol{\delta}_0 + \boldsymbol{\varepsilon}_{(i)}$.

$\boldsymbol{\mu}_{(i)}$ follow normal distributions:

(3A.5) $\boldsymbol{\mu}_{(i)} \sim_{iid} \mathcal{N}(\boldsymbol{\iota}_{T_{(i)}} \otimes (\boldsymbol{\omega} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}) + \mathbf{Z}_{(i)}\boldsymbol{\delta}, \mathbf{G}_i)$ with $\mathbf{G}_i = \boldsymbol{\iota}_{T_{(i)}}\boldsymbol{\iota}'_{T_{(i)}} \otimes \boldsymbol{\psi} + \mathbf{I}_{T_{(i)}} \otimes \boldsymbol{\Omega}$.

The conditional distribution of $\boldsymbol{\beta} \mid \boldsymbol{\mu};\boldsymbol{\theta}_0$ is given by:

(3A.6) $\boldsymbol{\beta} \mid \boldsymbol{\mu};\boldsymbol{\theta}_0 \sim_{iid} \mathcal{N}(\mathbf{m}_{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}_0), \mathbf{V}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0))$

$$\text{with} \quad \begin{cases} \mathbf{m}_{\boldsymbol{\beta},i}(\boldsymbol{\mu};\boldsymbol{\theta}) = \mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta})\left(\boldsymbol{\Omega}^{-1}\sum_{t=1}^{T_{(i)}}(\boldsymbol{\mu}_{it} - \mathbf{Z}_{it}\boldsymbol{\delta}) + \boldsymbol{\psi}^{-1}(\boldsymbol{\omega} + \bar{\mathbf{Z}}_i\boldsymbol{\pi})\right) \\ \mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}) = (\boldsymbol{\psi}^{-1} + T_{(i)}\boldsymbol{\Omega}^{-1})^{-1} \end{cases}.$$

The distribution of $\boldsymbol{\mu} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)};\boldsymbol{\theta}_0$ has not a standard form since $\bar{\mathbf{x}}_{(i)}$ is not linear in $\boldsymbol{\mu}$. Using Bayes' formula, we have:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(3A.7) $f(\boldsymbol{\mu} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}; \boldsymbol{\theta}) \propto f(\bar{\mathbf{x}}_{(i)} \mid \boldsymbol{\mu}, \mathbf{s}_{(i)}; \sigma^2) f(\boldsymbol{\mu}; \boldsymbol{\theta}_m)$ with $\boldsymbol{\theta}_m = (\boldsymbol{\omega}, \boldsymbol{\pi}, \boldsymbol{\delta}, \boldsymbol{\Omega}, \boldsymbol{\psi})$,

where f defines the density probability function.

Maximum Likelihood estimation via TVF-SAEM algorithm

Now, let define, as in previous section, vectors $\bar{\mathbf{x}}_{(i)} = (\bar{\mathbf{x}}_{it} : t \in \mathcal{H}_i)$, $\mathbf{s}_{(i)} = (\mathbf{s}_{it} : t \in \mathcal{H}_i)$, $\mathbf{z}_{(i)} = (\mathbf{z}_{it} : t \in \mathcal{H}_i)$, $\boldsymbol{\beta}_i = (b_{c,i} : c \in \mathcal{C})$ and $\boldsymbol{\mu}_{(i)} = (\boldsymbol{\mu}_{it} : t \in \mathcal{H}_i)$. In the contest of the considered model, $\boldsymbol{\beta}_i$ and $\boldsymbol{\mu}_{(i)}$ are viewed as missing data. Then, the complete data of our model consists of the vector of observed variable $\boldsymbol{\zeta}_{(i)} = (\bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$, of the vector of unobserved variables $(\boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)})$, for $i = 1,\ldots,N$. The complete data log-likelihood function is the sample log-likelihood function of the joint model of the dependent and missing variables, $(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)})$, given the exogenous variables of the model, $\mathbf{s}_{(i)}$ and $\mathbf{z}_{(i)}$, for $i = 1,\ldots,N$. The contribution of individual i in the complete data log-likelihood function at $\boldsymbol{\theta}$ of our model is given by:

(3A.8) $\ln \ell^c(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}; \mathbf{z}_{(i)}) = \ln f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta})$

where:-

(3A.9)
$$\begin{aligned} &\ln f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}) \\ &= \sum_{t=1}^{T_{(i)}} \ln \varphi(\bar{x}_{it} - \mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}); \sigma^2) + \sum_{t=1}^{T_{(i)}} \ln \varphi(\boldsymbol{\mu}_{it} - \boldsymbol{\beta}_i - \mathbf{Z}_{it}\boldsymbol{\delta}; \boldsymbol{\Omega}) + \ln \varphi(\boldsymbol{\beta}_i - \boldsymbol{\omega} - \bar{\mathbf{Z}}_i \boldsymbol{\pi}; \boldsymbol{\psi}). \end{aligned}$$

The term $\varphi(\mathbf{A}, \mathbf{B})$ denotes the probability density function at point $\mathbf{A}$ of the standard multivariate normal distribution with variance-covariance matrix **B.** Note that the complete data log-likelihood belongs to exponential family. The corresponding observed data log-likelihood can be obtained by integrated the complete data likelihood with respect the missing data:

(3A.10) $\ln \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}) = \ln \int f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}) d(\boldsymbol{\mu}, \boldsymbol{\beta})$.

Then, the maximum likelihood estimator is obtained by maximizing the observed data log-likelihood:

(3A.11) $\boldsymbol{\theta}_{N_{tot}}^{MLE} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$.

Note that $\ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$ has not a closed form and direct maximization of the observed data log-likelihood is problematic. Indeed, it is well known that the direct maximization of this kind of observed log-likelihood using algorithms, as Newton-Raphson, is problematic. Model is non-linear in random terms. Iterative algorithms as EM algorithm and its variants are suitable to maximize $\sum_{i=1}^{N} \ln \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)})$ in our case and the complete data log-likelihood must be considered instead.

The standard EM algorithm involves two steps until convergence: the Expectation and the Maximization steps. The Expectation step of EM of standard EM algorithm of our model involves computing the following conditional expectation:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(3A.12) $q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)}) = \int (\ln \lambda^c(\boldsymbol{\theta};\bar{\mathbf{x}}_i,\boldsymbol{\mu},\boldsymbol{\beta}\,|\,\mathbf{s}_{(i)},\mathbf{z}_{(i)}))f(\boldsymbol{\mu},\boldsymbol{\beta}\,|\,\bar{\mathbf{x}}_i,\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)})d(\boldsymbol{\mu},\boldsymbol{\beta})$

where:

(3A.13) $$f(\boldsymbol{\mu},\boldsymbol{\beta}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)}) = f(\boldsymbol{\beta}\,|\,\boldsymbol{\mu};\boldsymbol{\theta}^{(n-1)})f(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)})$$

At the Maximization step, $\boldsymbol{\theta}^{(n)}$ is updated following:

(3A.14) $\boldsymbol{\theta}^{(n)} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$.

In our case, the E-step has not a simple form. However, it is possible to simplify it using some factorizations detailed below. Let consider the following factorization:

(3A.15) $q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)}) = \int q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})f(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)};\boldsymbol{\theta}^{(n-1)})d\boldsymbol{\mu}$

Where:

(3A.16) $q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)}) = \int \ln \lambda^c(\boldsymbol{\theta};\bar{\mathbf{x}}_i,\boldsymbol{\mu},\boldsymbol{\beta}\,|\,\mathbf{s}_{(i)},\mathbf{z}_{(i)})f(\boldsymbol{\beta}\,|\,\boldsymbol{\mu};\boldsymbol{\theta}^{(n-1)})d\boldsymbol{\beta}$.

It is easy to show that $q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$ has an explicit form:

(3A.18)

$$\begin{aligned} q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)}) = {} & \sum_{t=1}^{T_{(i)}} \ln \varphi\left(\bar{\mathbf{x}}_{it} - \mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it});\sigma^2\right) \\ & + \sum_{t=1}^{T_{(i)}} \ln \varphi\left(\boldsymbol{\mu}_{it} - \mathbf{m}_{\boldsymbol{\beta},i}(\boldsymbol{\mu};\boldsymbol{\theta}^{(n-1)}) - \mathbf{Z}_{it}\boldsymbol{\delta};\boldsymbol{\Omega}\right) - \tfrac{1}{2}T_{(i)}\operatorname{tr}\left(\boldsymbol{\Omega}^{-1}\mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}^{(n-1)})\right) \\ & + \ln \varphi\left(\mathbf{m}_{\boldsymbol{\beta},i}(\boldsymbol{\mu};\boldsymbol{\theta}^{(n-1)}) - \boldsymbol{\omega} - \bar{\mathbf{Z}}_i\boldsymbol{\pi};\boldsymbol{\theta}^{(n-1)});\boldsymbol{\psi}\right) - \tfrac{1}{2}\operatorname{tr}(\boldsymbol{\psi}^{-1}\mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}^{(n-1)})) \end{aligned}$$

Finally, to compute $q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$, we need to integrate $q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$ with respect to the conditional distribution $f(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)})$. As said above, $f(\boldsymbol{\mu}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)})$ has not a simple form and we need to use simulation approaches. Thus, we use approximated-SAEM algorithm (Allassonnière and Chevallier, 2021) to compute $q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$ once $q_i^{\boldsymbol{\beta}}(\boldsymbol{\mu};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)})$ is known.

The approximated-SAEM algorithm proposed by Allassonière and Chevallier (2021) is an extension of the SAEM algorithm (Delyon et al, 1999, Kuhn and Lavielle, 2004) where the simulation step is improved. It consists in three steps until convergence: the simulation (S) step, the stochastic approximation (SA) step and the maximization (M) step. At iteration (n) and given ($\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)},\boldsymbol{\theta}^{(n-1)}$), we have:

- S-step: simulate the missing data $\tilde{\boldsymbol{\mu}}_i^{(n+1)}$ under the approximated probability density function $\hat{f}_n(\boldsymbol{\mu}_{(i)}\,|\,\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)};\boldsymbol{\theta}^{(n-1)})$ for $i = 1,\dots,N$,
- SA-step: update $Q_n(\boldsymbol{\theta})$ as:

$$Q_n(\boldsymbol{\theta}) = Q_{n-1}(\boldsymbol{\theta}) + \lambda_{(n)}\left(\sum_{i=1}^{N} q_i^{\boldsymbol{\beta}}(\tilde{\boldsymbol{\mu}}_i^{(n)};\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(n-1)}) - Q_{n-1}(\boldsymbol{\theta})\right)$$

- M-step: update $\boldsymbol{\theta}^{(n)}$ according to:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\boldsymbol{\theta}^{(n)} = \arg\max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}).$$

As in standard SAEM algorithm (Delyon et al., 1999), the sequence $(\lambda_{(n)})_{n \in \mathbb{N}}$ must be a decreasing positive sequence such that $\lambda_{(1)} = 1$, $\sum_{n=1}^{+\infty} \lambda_{(n)} = +\infty$ and $\sum_{n=1}^{+\infty} \lambda_{(n)}^2 < +\infty$. This sequence defines the step of the stochastic approximation and affects the speed of the convergence of the algorithm to the Maximum Likelihood. Kuhn and Lavielle (2005) proposes to set $\lambda_{(n)} = 1$ for the first n_1 iterations of the algorithm and then gradually reduce $\lambda_{(n)}$. The approximated density $\hat{f}_n(\boldsymbol{\mu}_{(i)}, \overline{\mathbf{x}}_{(i)}, \boldsymbol{\theta}^{(n-1)})$ is defined by:

(3A.19) $\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \overline{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}; \boldsymbol{\theta}^{(n-1)}) = \dfrac{1}{C_{\boldsymbol{\theta}}(T_n)} f(\boldsymbol{\mu}_{(i)} \mid \overline{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}; \boldsymbol{\theta}^{(n-1)})^{1/T_n}$

where $(T_n)_{n \in \mathbb{N}}$ is a sequence of positive numbers such that $\lim_{n \to +\infty} T_n = 1$ and $C_{\boldsymbol{\theta}}(T_n)$ is a scaling constant. Allassonnière and Chevallier (2021) proposes to use an oscillatory temperature pattern: T_n must be oscillated around one with decreasing amplitude. They prove the convergence of their algorithm toward a (global) maximum of the (observed) likelihood for the complete data belonging to the exponential family. Their algorithm allows escaping local maxima and can be efficient in the case of high–dimensional random parameters.

Now, we describe the algorithm used in this study, which combine standard EM algorithm and approximated-SAEM algorithm. In our case, as said above, $q_i^{\beta}(\tilde{\boldsymbol{\beta}}_i; \boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n)})$ belongs to exponential family, one of the convergence conditions of the approximated-SAEM algorithm.

The considered algorithm in this section consists in three steps. At iteration n and given observed data $(\overline{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$ and $\boldsymbol{\theta}^{(n-1)}$:

- **S-step:** Simulate $\{\tilde{\boldsymbol{\beta}}_{i,r}^{(n)} : r = 1,\ldots,R\}$ according to the approximated probability density function $\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \overline{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})$ for $i = 1,\ldots,N$,
- **SA-step:** update the following sufficient statistics according to[35]:

$$\mathbf{s}_{1,i}^{(n)} = \mathbf{s}_{1,i}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\mathbf{m}_{\beta}(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)}) - \mathbf{s}_{1,i}^{(n-1)}\right) \text{ for } i,$$

$$\mathbf{s}_{2}^{(n)} = \mathbf{s}_{2}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbf{m}_{\beta}(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)})\mathbf{m}_{\beta}'(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)}) - \mathbf{s}_{2}^{(n-1)}\right) \text{ for } i,$$

$$\mathbf{s}_{3,it}^{(n)} = \mathbf{s}_{3,it}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}(\tilde{\boldsymbol{\beta}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)})) - \mathbf{s}_{3,it}^{(n-1)}\right), \text{ for } it,$$

$$\mathbf{s}_{4}^{(n)} = \mathbf{s}_{4}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}(\tilde{\boldsymbol{\beta}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)}))(\tilde{\boldsymbol{\beta}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\beta}}_{(i),r}^{(n)}; \boldsymbol{\theta}^{(n-1)}))' - \mathbf{s}_{4}^{(n-1)}\right),$$

$$s_{5,i}^{(n)} = s_{5,i}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\sum_{t=1}^{T_{(i)}} w_{it}\left(\overline{\mathbf{x}}_{it} - \mathbf{s}_{it}'\mathbf{h}(\tilde{\boldsymbol{\beta}}_{it,r}^{(n)})\right)\left(\overline{\mathbf{x}}_{it} - \mathbf{s}_{it}'\mathbf{h}(\tilde{\boldsymbol{\beta}}_{it,r}^{(n)})\right)' - s_{5,i}^{(n-1)}\right),$$

35 Differentiating $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)})$ with respect to the parameters $\boldsymbol{\theta}$ allows us to choose these minimal sufficient statistics.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

- **M-step:** update parameters $\boldsymbol{\theta}$ according to:

$$\boldsymbol{\omega}^{(n)} = N^{-1}\sum_{i=1}^{N}\left(\mathbf{s}_{1,i}^{(n)} - \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)}\right),$$

$$\boldsymbol{\pi}^{(n)} = \left(\sum_{i=1}^{N}\bar{\mathbf{Z}}_i'\boldsymbol{\psi}^{-1}\bar{\mathbf{Z}}_i\right)^{-1}\sum_{i=1}^{N}\bar{\mathbf{Z}}_i'\boldsymbol{\psi}^{-1}\left(\mathbf{s}_{1,i}^{(n)} - \boldsymbol{\omega}^{(n)}\right),$$

$$\boldsymbol{\delta}^{(n)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}\mathbf{Z}_{it}'\mathbf{Z}_{it}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}\mathbf{Z}_{it}'\mathbf{s}_{3,it}^{(n)},$$

$$\boldsymbol{\psi}^{(n)} = N^{-1}\mathbf{s}_2^{(n)} + N^{-1}\sum_{i=1}^{N}\begin{pmatrix}(\boldsymbol{\omega}^{(n)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})(\boldsymbol{\omega}^{(n)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})' + \mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}^{(n-1)}) \\ -\,\mathbf{s}_{1,i}^{(n)}(\boldsymbol{\omega}^{(n)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})' - (\boldsymbol{\omega}^{(n)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})\mathbf{s}_{1,i}^{(n)\prime}\end{pmatrix},$$

$$\boldsymbol{\Omega}^{(n)} = \operatorname{diag}\left(N_{tot}^{-1}\mathbf{s}_4^{(n)} + N_{tot}^{-1}\sum_{i=1}^{N}\begin{pmatrix}\sum_{t=1}^{T_{(i)}}\left(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)}(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})' - \mathbf{s}_{3,it}^{(n)}(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})' - (\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})\mathbf{s}_{3,it}^{(n)\prime}\right) \\ +\,T_{(i)}\mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}^{(n-1)})\end{pmatrix}\right)$$

$$\sigma^{2(n)} = N_{tot}^{-1}\sum_{i=1}^{N}s_{5,i}^{(n)}.$$

These three steps are iteratively used until convergence. The simulation step consists in sampling $\boldsymbol{\mu}$ from the approximated density $\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})$. Using Bayes' formula, this density is proportional to

(3A.20)

$$\begin{aligned}\ln\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)}) \propto &\sum_{t=1}^{T_{(i)}}\ln\varphi\left(\bar{x}_{it} - \mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}); T_{(n)}\sigma^{2(n-1)}\right) \\ &+ \ln\varphi\left(\boldsymbol{\mu}_{(i)} - \boldsymbol{\iota}_{T_{(i)}} \otimes (\boldsymbol{\omega}^{(n-1)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)}) - \mathbf{Z}_{(i)}\boldsymbol{\delta}^{(n-1)}; T_{(n)}\mathbf{G}_i^{(n-1)}\right)\end{aligned},$$

and the sequence $(T_n)_{n\in\mathbb{N}}$ can be viewed as a sequence of precision parameters. We cannot use direct simulation for $\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})$. Thus, to perform simulation step, at iteration (n) of the algorithm, we use a few (R) MCMC iterations with $\hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})$ as the stationary distribution. More precisely, for each respond (farmer or farm) *i,* (*i=1,…,N*), we use Metropolis-Hasting algorithm with two kernels (R1 iterations for the first kernel and R2 iterations for the second kernel). To generate a candidate $\boldsymbol{\mu}_{(i)}^{(k)}$, the first kernel we consider is the marginal distribution of $\boldsymbol{\mu}_{(i)}$, $\boldsymbol{\mu}_{(i)}^{(k)} \sim \varphi\left(\boldsymbol{\mu} - \boldsymbol{\alpha}^{(n-1)} - \boldsymbol{\iota}_{T_{(i)}} \otimes \boldsymbol{\omega}^{(n-1)}; T_{(n)}\mathbf{G}^{(n-1)}\right)$ and the second kernel is the random walk: $\boldsymbol{\mu}_{(i)}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_{(i)}^{(k-1)}, k\boldsymbol{\Sigma}^{(n-1)})$. $\boldsymbol{\Sigma}^{(n-1)}$ is the diagonal matrix of $T_{(n)}\mathbf{G}^{(n-1)}$ and k is adaptively chosen so that the acceptance rate is within a given range.

Let $g(\boldsymbol{\mu}^c \mid \boldsymbol{\mu}^{(r)})$ denotes the kernel of the Metropolis-Hasting algorithm where $\boldsymbol{\mu}^{(r)}$ denotes the current value of $\boldsymbol{\mu}_{(i)}$. The probability of acceptance for $\boldsymbol{\mu}^c$ given the current value $\boldsymbol{\mu}^{(r)}$ is given by:

(3A.21)
$$r(\boldsymbol{\mu}^{(r)}, \boldsymbol{\mu}^c) = \min\left(1, \frac{\hat{f}_n(\boldsymbol{\mu}^c \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})}{\hat{f}_n(\boldsymbol{\mu}^{(r)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}^{(n-1)})}\frac{g(\boldsymbol{\mu}^{(r)} \mid \boldsymbol{\mu}^c)}{g(\boldsymbol{\mu}^c \mid \boldsymbol{\mu}^{(r)})}\right).$$

For the marginal distribution kernel, the probability of acceptance is reduced to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(3A.22) $$r(\boldsymbol{\mu}^{(r)},\boldsymbol{\mu}^{c})=\min\left(1,\frac{\prod_{t=1}^{T_{(i)}}\varphi\left(\bar{\mathbf{x}}_{it}-\mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}^{c});T_n\sigma^{2(n-1)}\right)}{\prod_{t=1}^{T_{(i)}}\varphi\left(\bar{\mathbf{x}}_{it}-\mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}^{(r)});T_n\sigma^{2(n-1)}\right)}\right),$$

and for the random walk kernel, it is given by:

(3A.23)

$$r(\boldsymbol{\mu}^{(r)},\boldsymbol{\mu}^{c})$$
$$=\min\left(1,\frac{\varphi\left(\boldsymbol{\mu}^{c}-\boldsymbol{\iota}_{T_{(i)}}\otimes(\boldsymbol{\omega}^{(n-1)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)})-\mathbf{Z}_{(i)}\boldsymbol{\delta}^{(n-1)};T_n\mathbf{G}_i^{(n-1)}\right)\prod_{t=1}^{T_{(i)}}\varphi\left(\bar{\mathbf{x}}_{it}-\mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}^{c});T_n\sigma^{2(n-1)}\right)}{\varphi\left(\boldsymbol{\mu}^{(r)}-\boldsymbol{\iota}_{T_{(i)}}\otimes(\boldsymbol{\omega}^{(n-1)}+\bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)})-\mathbf{Z}_{(i)}\boldsymbol{\delta}^{(n-1)};T_n\mathbf{G}_i^{(n-1)}\right)\prod_{t=1}^{T_{(i)}}\varphi\left(\bar{\mathbf{x}}_{it}-\mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it}^{(r)});T_n\sigma^{2(n-1)}\right)}\right)$$

Accounting for weighted data

Let define, as in previous section, vectors $\bar{\mathbf{x}}_{(i)}=(\bar{\mathbf{x}}_{it}:t\in\mathcal{H}_i)$, $\mathbf{s}_{(i)}=(\mathbf{s}_{it}:t\in\mathcal{H}_i)$, $\mathbf{z}_{(i)}=(\mathbf{z}_{it}:t\in\mathcal{H}_i)$, $\boldsymbol{\beta}_i=(b_{c,i}:c\in\mathcal{C})$ and $\boldsymbol{\mu}_{(i)}=(\boldsymbol{\mu}_{it}:t\in\mathcal{H}_i)$. Now, we assume that observations may depend on weights $\mathbf{w}_{(i)}=(w_{it}:t\in\mathcal{H}_i)$ as in FADN data. Observations $\bar{\mathbf{x}}_{it}$ conditional on $\mathbf{s}_{it}$ and $\mathbf{z}_{it}$ are independently distributed with respect to the weight w_{it} .The complete data likelihood depends now on the weights through the following decomposition:

(3A.24) $$\ell^{c}(\boldsymbol{\theta};\bar{\mathbf{x}}_{(i)},\boldsymbol{\mu}_{(i)},\boldsymbol{\beta}_i\,|\,\mathbf{s}_{(i)},\mathbf{z}_{(i)},\mathbf{w}_{(i)})=f(\boldsymbol{\mu}_{(i)},\boldsymbol{\beta}_i;\boldsymbol{\theta})\prod_{t=1}^{T_{(i)}}f(\bar{\mathbf{x}}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{(i)},w_{it};\boldsymbol{\theta}).$$

In this decomposition, only the conditional distribution of observed data $\bar{\mathbf{x}}_{it}$ depends on weights w_{it} . Indeed, this is justified by the fact that the distribution of the random parameters of interest is the distribution at the population level. If we are interested in the distribution of the random parameters at the sample level, it is not necessary to introduce the weights but the hypothesis of independence of the observations remains strong.

It is also assumed that:

(3A.25) $$f(\bar{x}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{it},w_{it};\boldsymbol{\theta})=\tfrac{1}{C(w_{it})}f(\bar{x}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{it};\boldsymbol{\theta})^{w_{it}}$$

where $f(\bar{\mathbf{x}}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{it};\boldsymbol{\theta})$ is a probability density function of $\bar{\mathbf{x}}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{it}$, and $C(w_{it})$ is a normalized constant. Indeed, raise $f(\bar{\mathbf{x}}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it};\boldsymbol{\theta})$ to the power w_{it} in maximum likelihood setting is equivalent to "*observing* $\bar{\mathbf{x}}_{it}$ w_{it} *times given* $\boldsymbol{\mu}_{it}$ $\mathbf{s}_{it}$ *and* $\mathbf{z}_{it}$" in standard approaches. However, $f(\bar{\mathbf{x}}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it};\boldsymbol{\theta})^{w_{it}}$ is not a probability density function and we need to normalize it. Gebru et al., (2016) use the same approach to account for weighted data in other context. In our case:

(3A.26) $$f(\bar{x}_{it}\,|\,\boldsymbol{\mu}_{it},\mathbf{s}_{it},\mathbf{z}_{it},w_{it};\boldsymbol{\theta})\propto\varphi\left(\bar{x}_{it}-\mathbf{s}_{it}'\mathbf{h}(\boldsymbol{\mu}_{it});(1/w_{it})\sigma^{2(n-1)}\right).$$

Given equations (39) and (41), the estimation procedure will change little from that described above for unweighted observations. Only the simulation step changes. At iteration n and given observed data $(\bar{\mathbf{x}}_{(i)},\mathbf{s}_{(i)},\mathbf{z}_{(i)},\mathbf{w}_{(i)})$ and $\boldsymbol{\theta}^{(n-1)}$, the approximated probability density function depends now on $\mathbf{w}_{(i)}$ and we have:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

(3A.27)
$$\ln \hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}, \mathbf{w}_{(i)}; \boldsymbol{\theta}^{(n-1)}) \propto \sum_{t=1}^{T_{(i)}} \ln \varphi\left(\bar{x}_{it} - \mathbf{s}'_{it}\mathbf{h}(\boldsymbol{\mu}_{it}); T_{(n)}(1/w_{it})\sigma^{2(n-1)}\right) + \ln \varphi\left(\boldsymbol{\mu}_{(i)} - \boldsymbol{\iota}_{T_{(i)}} \otimes (\boldsymbol{\omega}^{(n-1)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)}) - \mathbf{Z}_{(i)}\boldsymbol{\delta}^{(n-1)}; T_{(n)}\mathbf{\Gamma}_i^{(n-1)}\right)$$

At each iteration the algorithm consists in three step until convergence :

- **S-step:** Simulate $\{\tilde{\boldsymbol{\mu}}_{i,r}^{(n)} : r = 1,\ldots,R\}$ according to the approximated probability density function $\ln \hat{f}_n(\boldsymbol{\mu}_{(i)} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)}, \mathbf{w}_{(i)}; \boldsymbol{\theta}^{(n-1)})$ for $i = 1,\ldots,N$,
- **SA-step:** update the following sufficient statistics according to[36]:

$$\mathbf{s}_{1,i}^{(n)} = \mathbf{s}_{1,i}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\mathbf{m}_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)}) - \mathbf{s}_{1,i}^{(n-1)}\right) \text{ for } i,$$

$$\mathbf{s}_{2}^{(n)} = \mathbf{s}_{2}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{i=1}^{N}\sum_{r=1}^{R}\mathbf{m}_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)})\mathbf{m}'_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)}) - \mathbf{s}_{2}^{(n-1)}\right) \text{ for } i,$$

$$\mathbf{s}_{3,it}^{(n)} = \mathbf{s}_{3,it}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}(\tilde{\boldsymbol{\mu}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)})) - \mathbf{s}_{3,it}^{(n-1)}\right), \text{ for } it,$$

$$\mathbf{s}_{4}^{(n)} = \mathbf{s}_{4}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}(\tilde{\boldsymbol{\mu}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)}))(\tilde{\boldsymbol{\mu}}_{it,r}^{(n)} - \mathbf{m}_{\beta}(\tilde{\boldsymbol{\mu}}_{(i),r}^{(n)};\boldsymbol{\theta}^{(n-1)}))' - \mathbf{s}_{4}^{(n-1)}\right),$$

$$s_{5,i}^{(n)} = s_{5,i}^{(n-1)} + \lambda_{(n)}\left(R^{-1}\sum_{r=1}^{R}\sum_{t=1}^{T_{(i)}}\left(\bar{\mathbf{x}}_{it} - \mathbf{s}'_{it}\mathbf{h}(\tilde{\boldsymbol{\mu}}_{it,r}^{(n)})\right)\left(\bar{\mathbf{x}}_{it} - \mathbf{s}'_{it}\mathbf{h}(\tilde{\boldsymbol{\mu}}_{it,r}^{(n)})\right)' - s_{5,i}^{(n-1)}\right),$$

- **M-step:** update parameters $\boldsymbol{\theta}$ according to:

$$\boldsymbol{\omega}^{(n)} = N^{-1}\sum_{i=1}^{N}(\mathbf{s}_{1,i}^{(n)} - \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n-1)}),$$

$$\boldsymbol{\pi}^{(n)} = \left(\sum_{i=1}^{N}\bar{\mathbf{Z}}'_i\boldsymbol{\Psi}^{-1}\bar{\mathbf{Z}}_i\right)^{-1}\sum_{i=1}^{N}\bar{\mathbf{Z}}'_i\boldsymbol{\Psi}^{-1}(\mathbf{s}_{1,i}^{(n)} - \boldsymbol{\omega}^{(n)}),$$

$$\boldsymbol{\delta}^{(n)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}\mathbf{Z}'_{it}\mathbf{Z}_{it}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T_{(i)}}\mathbf{Z}'_{it}\mathbf{s}_{3,it}^{(n)},$$

$$\boldsymbol{\psi}^{(n)} = N^{-1}\mathbf{s}_2^{(n)} + N^{-1}\sum_{i=1}^{N}\begin{pmatrix}(\boldsymbol{\omega}^{(n)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})(\boldsymbol{\omega}^{(n)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})' + \mathbf{V}_{\beta,i}(\boldsymbol{\theta}^{(n-1)}) \\ - \mathbf{s}_{1,i}^{(n)}(\boldsymbol{\omega}^{(n)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})' - (\boldsymbol{\omega}^{(n)} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}^{(n)})\mathbf{s}_{1,i}^{(n)\prime}\end{pmatrix},$$

$$\boldsymbol{\Omega}^{(n)} = \operatorname{diag}\left(N_{tot}^{-1}\mathbf{s}_4^{(n)} + N_{tot}^{-1}\sum_{i=1}^{N}\begin{pmatrix}\sum_{t=1}^{T_{(i)}}\left(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)}(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})' - \mathbf{s}_{3,it}^{(n)}(\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})' - (\mathbf{Z}_{it}\boldsymbol{\delta}^{(n)})\mathbf{s}_{3,it}^{(n)\prime}\right) \\ + T_{(i)}\mathbf{V}_{\beta}(\boldsymbol{\theta}^{(n-1)})\end{pmatrix}\right)$$

$$\sigma^{2(n)} = N_{tot}^{-1}\sum_{i=1}^{N}s_{5,i}^{(n)}.$$

[36] Differentiating $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n-1)})$ with respect to the parameters $\boldsymbol{\theta}$ allows us to choose these minimal sufficient statistics.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

APPENDIX 6A: MODEL FORMULATION FOR NON-REALLOCATION WITH SIMULTANEOUS INEFFICIENCY

$$\max_{\beta_i,\alpha_i,\lambda_k,\gamma_k,\mu_k} \beta_i \quad \text{(A)}$$

s.t.

$$\sum_{k=1}^{K} \lambda_k x_k^C \le x_i^C \quad \text{(6Aa)}$$

$$\sum_{k=1}^{K} \lambda_k m_k^{L,u} \le m_i^{L,u} \quad \text{(6Ab)}$$

$$\sum_{k=1}^{K} \lambda_k x_k^{C,l} - x_i^{C,l} \le 0 \quad \text{(6Ac)}$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J1} \le q_i^{J1} \quad \text{(6Ad)}$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J2} \le q_i^{J2} \quad \text{(6Ae)}$$

$$\sum_{k=1}^{K} -\lambda_k y_k^C + \beta_i g_{y,k}^C \le -y_i^C \quad \text{(6Af)}$$

$$\sum_{k=1}^{K} -\lambda_k z_k^C + \beta_i g_{z,k}^C \le -z_i^C \quad \text{(6Ag)}$$

$$\sum_{K=1}^{K} \lambda_k = 1 \quad \text{(6Ah)}$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,fh} \le x_i^{L,fh} \quad \text{(6Ai)}$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,a} \le x_i^{L,a} \quad \text{(6Aj)}$$

$$\sum_{k=1}^{K} \gamma_k x_k^{L,l} - x_i^{L,l} \le 0 \quad \text{(6A6)}$$

$$\sum_{k=1}^{K} \gamma_k z_k^C \le z_i^C \quad \text{(6Al)}$$

$$\sum_{k=1}^{K} \gamma_k q_k^{J1} \le q_i^{J1} \quad \text{(6Am)}$$

$$\sum_{k=1}^{K} \gamma_k q_k^{J2} \le q_i^{J2} \quad \text{(6An)}$$

$$\sum_{k=1}^{K} -\gamma_k y_k^L + \beta_i g_{y,k}^L \le -y_i^L \quad \text{(6Ao)}$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\sum_{K=1}^{K} \gamma_k = 1 \tag{6A7}$$

$$\sum_{k=1}^{K} \gamma_k (m_k^{L,u} + m_k^{L,p}) = m_i^{L,u} + m_i^{L,p} \tag{6Aq}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{C,p} \leq - x_i^{C,p} \tag{6Ar}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pa} \leq - x_i^{L,Pa} \tag{6A8}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pf} \leq - x_i^{L,Pf} \tag{6At}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pfc} \leq - x_i^{L,Pfc} \tag{6Au}$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pe} \leq - q_i^{J,pe} \tag{6Av}$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pm} \leq - q_i^{J,pm} \tag{6Aw}$$

$$\sum_{k=1}^{K} \mu_k e_k + \beta_i g_{e,k} \leq e_i \tag{6Ax}$$

$$\sum_{K=1}^{K} \mu_k = 1 \tag{6A9}$$

APPENDIX 6B: SEPARATE INEFFICIENCIES FOR GHG EMISSIONS AND OUTPUTS

$$\max_{\substack{\beta_i,\alpha_i,\lambda_k,\gamma_k,\mu_k \\ x_i^{C,l} \geq 0, x_i^{L,l} \geq 0}} (\beta_i + \alpha_i)/2 \tag{6B1}$$

s.t.

$$\sum_{k=1}^{K} \lambda_k x_k^C \leq x_i^C \tag{6B1a}$$

$$\sum_{k=1}^{K} \lambda_k m_k^{L,u} \leq m_i^{L,u} \tag{6B1b}$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\sum_{k=1}^{K} \lambda_k \, x_k^{C,l} - x_i^{C,l} \leq 0 \tag{6B1c}$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J1} \leq q_i^{J1} \tag{6B1d}$$

$$\sum_{k=1}^{K} \lambda_k q_k^{J2} \leq q_i^{J2} \tag{6B1e}$$

$$\sum_{k=1}^{K} -\lambda_k y_k^{C} + \beta_i g_{y,k}^{C} \leq -y_i^{C} \tag{6B1f}$$

$$\sum_{k=1}^{K} -\lambda_k z_k^{C} + \beta_i g_{z,k}^{C} \leq -z_i^{C} \tag{6B1g}$$

$$\sum_{K=1}^{K} \lambda_k = 1 \tag{6B1h}$$

$$\sum_{k=1}^{K} \gamma_k \, x_k^{L,fh} \leq x_i^{L,fh} \tag{6B1i}$$

$$\sum_{k=1}^{K} \gamma_k \, x_k^{L,a} \leq x_i^{L,a} \tag{6B1j}$$

$$\sum_{k=1}^{K} \gamma_k \, x_k^{L,l} - x_i^{L,l} \leq 0 \tag{6B1k}$$

$$\sum_{k=1}^{K} \gamma_k z_k^{C} \leq z_i^{C} \tag{6B1l}$$

$$\sum_{k=1}^{K} \gamma_k q_k^{J1} \leq q_i^{J1} \tag{6B1m}$$

$$\sum_{k=1}^{K} \gamma_k q_k^{J2} \leq q_i^{J2} \tag{6B1n}$$

$$\sum_{k=1}^{K} - \gamma_k y_k^{L} + \beta_i g_{y,k}^{L} \leq -y_i^{L} \tag{6B1o}$$

$$\sum_{K=1}^{K} \gamma_k = 1 \tag{6B1p}$$

$$\sum_{k=1}^{K} \gamma_k (m_k^{L,u} + m_k^{L,p}) = m_i^{L,u} + m_i^{L,p} \tag{6B1q}$$

$$\sum_{k=1}^{K} - \mu_k x_k^{C,p} \leq - x_i^{C,p} \tag{6B1r}$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pa} \leq - x_i^{L,Pa} \quad (6B1s)$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pf} \leq - x_i^{L,Pf} \quad (6B1t)$$

$$\sum_{k=1}^{K} - \mu_k x_k^{L,Pfc} \leq - x_i^{L,Pfc} \quad (6B1u)$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pe} \leq - q_i^{J,pe} \quad (6B1v)$$

$$\sum_{k=1}^{K} - \mu_k q_k^{J,pm} \leq - q_i^{J,pm} \quad (6B1w)$$

$$\sum_{k=1}^{K} \mu_k e_k + \alpha_i \, g_{e,k} \leq \; e_i \quad (6B1x)$$

$$\sum_{K=1}^{K} \mu_k = 1 \quad (6B1y)$$

$$x_i^{C,l} \; + \; x_i^{L,l} = \; x_i^{l} \quad (6B1z)$$

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 817566.