17èmes Journées de Recherches en Sciences Sociale, Paris-Saclay, 14 et 15 décembre 2023

Variable Input Allocation Among Crops: A Time-Varying Random Parameters Approach

Obafèmi Philippe Koutchadé[1], Alain Carpentier[1] and Fabienne Femenia[1]

(1) UMR SMART, INRAE – L'Institut Agro Rennes-Angers, 35000, Rennes, France.

Auteur de correspondance : obafemi-philippe.koutchade@inrae.fr

Abstract: Our main objective in this paper is to propose an approach to allocate variable input costs among crops produced by farmers based on panel data including information on input expenditure aggregated at the farm level. Our proposed approach simultaneously allows to control for unobserved farms and farmers heterogeneity, to address the potential dependence between variable input uses and acreage choice decisions, and to ensure consistent values of input use estimates. These are indeed major issues generally encountered in the estimation input allocation equations. This approach relies on a model of input allocation derived from accounting identities, in which unobserved input uses per crop are modeled as time-varying random parameters. Our model is estimated, using an extension of the Stochastic Approximation Expectation Maximization algorithm, on a sample of French farms' accounting data. Our estimation results show good performance of our approach compared to standard regression approaches generally used by agricultural economists.

Keywords: input cost allocation, crop production decisions, random parameter models, SAEM algorithm.

Classification JEL: C13, C18, C51, Q12

1. Introduction

Getting information about production costs per crop at the farm level is very important when analyzing multi-crop farms' behaviors. It can indeed be 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 on these costs per crop is generally not provided in farm accounting dataset, such as Farm Accountancy Data Network (FADN) data, which only include aggregate input expenditures at the farm level. Adequate statistical and\or economic modeling are thus necessary to allocate this aggregate information among the different crops produced by the farms.

Different approaches have been proposed to overcome this issue in the agricultural economics literature. Carpentier and Letort (2012) distinguish two groups of approaches. The first group includes approaches that consider only variable input allocation equations, in which 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, and to the satisfactory results it generally provide in terms of production cost predictions compared to the second approach (Just et al., 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. Second, 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 these correlations 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 do thus 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 model of input allocation derived from accounting identities. We use a random parameter specification to account for farm unobserved heterogeneity. The unobserved input uses per crop are modeled 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. We nobly enforce their non-negativity constraints by using a lognormal distribution for the random parameters.

This model is estimated, using an extension of the Stochastic Approximation Expectation Maximization (SAEM) algorithm (Delyon et al., 1999), on a sample of 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

2, we present our model of input use allocation. Our SAEM estimation approach is presented in section 3, and the empirical results in section 4. Finally, we conclude.

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 $c \in \mathsf{C}$ 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 input used at the farm level by farmer i at time t and $x_{c,it}$ denote the quantity of input used per unit of land of crop c. The input allocation problem consists in recovering the input quantity $x_{c,it}$ for each crop such that:

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

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

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

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

This input use equation at the farm level is completed by a model of crop input uses.

2.1. Model of crop input uses

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 time. Here, it is assumed that crop input uses $x_{c,it}$ are a transformation of normally distributed terms $\mu_{c,it}$, this transformation inducing bounds on the values of $x_{c,it}$. By doing so, it is easy to guarantee that the estimated crop input uses $x_{c,it}$ lie in reasonable ranges. For instance, to

force the positivity of $x_{c,it}$, we can assumed that $x_{j,k,it} = \exp(\mu_{j,k,it})$. More generally, we assume that:

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

and

(4) $\mu_{c,it} = \beta_{c,i} + \alpha_{c,t,0} + \varepsilon_{c,it}$ with $E[\varepsilon_{c,it}] = 0$,

where *h* is a non-linear transformation. h^{-1} may for instance be a log-normal distribution, a censored-normal distribution or a Johnson's (1949) SB distribution allowing to incorporate additional information in the model (Train, 2005).

As shown in (4), $\mu_{c,it}$ is decomposed into three components. First, farm-specific effects $\beta_{c,i}$ correspond to mean values of $\mu_{c,it}$. $\beta_{c,i}$ is assumed to be farm-specific for $c \in C$ and to vary randomly across farms. These farm-specific parameters allows accounting for unobserved farms' and farmers' characteristics. Second, year-specific effect $\alpha_{c,t,0}$ represent the deviations of $\mu_{c,it}$ from farm-specific effects $\beta_{c,i}$ at time *t* . For identification purpose, $\alpha_{c,t,0}$ is normalized to 0 at *t=1*. Third, 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 events, 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. $\boldsymbol{\varepsilon}_{it}$ and $\boldsymbol{\varepsilon}_{(i)}$, and $\boldsymbol{\alpha}_{t,0}$ and $\boldsymbol{\alpha}_0$ are defined similarly.

We assume 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 $\boldsymbol{\beta}_i = (\beta_{c,i} : c \in C) \sim_{iid} \mathcal{N}(\boldsymbol{\omega}_0, \boldsymbol{\psi}_0)$.

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

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:

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

with $\mathbf{z}_{it}$ a vector of observed farmer's/farm's characteristics. On the other hand, the assumption of independence between $\boldsymbol{\beta}_i$ and $s_{c,it}$ may not be verified since $\boldsymbol{\beta}_i$ 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 $\beta_{c,i}$ as a function of the average acreage share of crop c in farm i. We also introduce acreage shares as control variables in crop input uses model. It is actually possible to incorporate in a flexible way observed control variables – including crop acreage shares and crop yield levels – in the crop input uses model, such that:

$$(8) \qquad \beta_{c,i} = \omega_{c,0} + (\mathbf{z}_{c,i.} - E(\mathbf{z}_{c,i.}))' \boldsymbol{\pi}_{c,0} + \eta_{c,i}$$

$$(9) \qquad \mu_{c,it} = \beta_{c,i} + \alpha_{c,t,0} + (\mathbf{z}_{c,it} - E(\mathbf{z}_{c,it}))' \boldsymbol{\delta}_{c,0} + \varepsilon_{c,it},$$

With $\mathbf{z}_{c,i.}$ the average value of $\mathbf{z}_{c,it}$ for farm i and $E(\mathbf{z}_{c,i.})$ and $E(\mathbf{z}_{c,it})$ the sample averages of $\mathbf{z}_{c,i.}$ and $\mathbf{z}_{c,it}$. $\mathbf{z}_{c,it}$ can include farmers' crop acreage share and other farmers observed characteristics. Average yields at regional levels, can for instance be used as proxies for the production and sanitary conditions of each region, in order to account for the specificity of production in each region in order to improve the identification of the crop input uses model.

To summarize, our 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:

(10) $\bar{x}_{it} = \sum_{c \in C} s_{c,it} x_{c,it} + u_{it}$ with $u_{it} \sim_{iid} \mathcal{N}(0, \sigma_0^2)$,

(11) $x_{c,it} = h(\mu_{c,it})$,

(12) $\mu_{c,it} = \omega_{c,0} + (\mathbf{z}_{c,i.} - E(\mathbf{z}_{c,i.}))' \boldsymbol{\pi}_{c,0} + (\mathbf{z}_{c,it} - E(\mathbf{z}_{c,it}))' \boldsymbol{\delta}_{c,0} + \eta_{c,i} + \alpha_{c,t,0} + \varepsilon_{c,it}$

where:

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

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 $c \in C$.

3. Estimation approach

In this section, to simplify the presentation of the estimation procedure, we rewrite the model in the following compact form:

(14) $\bar{x}_{it} = \mathbf{s}_{it}' \mathbf{x}_{it} + u_{it}$

(15) $\mathbf{x}_{it} = \mathbf{h}(\boldsymbol{\mu}_{it})$

(16) $\boldsymbol{\mu}_{it} = \boldsymbol{\beta}_i + \mathbf{Z}_{(i)} \boldsymbol{\delta} + \boldsymbol{\varepsilon}_{it}$ 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 $u_{it} \perp \boldsymbol{\eta}_i \perp \boldsymbol{\varepsilon}_{it} \perp \mathbf{s}_{it}$.

Our 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 farm 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.

3.1. Intermediate results

Here, we present some intermediate results that are needed in the estimation section. Let define, as in previous section, $\boldsymbol{\mu}_{(i)} = (\mu_{c,it} : t \in \mathcal{H}_{(i)}, c \in C)$ such that:

(17) $$\boldsymbol{\mu}_{(i)} = \boldsymbol{\iota}_{T_{(i)}} \otimes \boldsymbol{\beta}_i + \mathbf{Z}_{(i)}\boldsymbol{\delta}_0 + \boldsymbol{\varepsilon}_{(i)}.$$

$\boldsymbol{\mu}_{(i)}$ follow normal distributions:

(18) $$\boldsymbol{\mu}_{(i)} \sim_{iid} \mathcal{N}(\boldsymbol{\iota}_{T_{(i)}} \otimes (\boldsymbol{\omega} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}) + \mathbf{Z}_{(i)}\boldsymbol{\delta}, \mathbf{G}_i) \text{ with } \mathbf{G}_i = \boldsymbol{\iota}_{T_{(i)}}\boldsymbol{\iota}'_{T_{(i)}} \otimes \boldsymbol{\psi} + \mathbf{I}_{T_{(i)}} \otimes \boldsymbol{\Omega}.$$

The conditional distribution of $\boldsymbol{\beta} \mid \boldsymbol{\mu}; \boldsymbol{\theta}_0$ is given by:

(19) $$\boldsymbol{\beta} \mid \boldsymbol{\mu}; \boldsymbol{\theta}_0 \sim_{iid} \mathcal{N}(\mathbf{m}_{\boldsymbol{\beta}}(\boldsymbol{\mu}; \boldsymbol{\theta}_0), \mathbf{V}_{\boldsymbol{\beta}}(\boldsymbol{\theta}_0))$$

with $$\begin{cases} \mathbf{m}_{\boldsymbol{\beta},i}(\boldsymbol{\mu}; \boldsymbol{\theta}) = \mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta})\ \boldsymbol{\Omega}^{-1}\sum_{t=1}^{T_{(i)}}(\boldsymbol{\mu}_{it} - \mathbf{Z}_{it}\boldsymbol{\delta}) + \boldsymbol{\psi}^{-1}(\boldsymbol{\omega} + \bar{\mathbf{Z}}_i\boldsymbol{\pi}) \\ \mathbf{V}_{\boldsymbol{\beta},i}(\boldsymbol{\theta}) = (\boldsymbol{\psi}^{-1} + T_{(i)}\boldsymbol{\Omega}^{-1})^{-1} \end{cases}.$$

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:

(20) $$f(\boldsymbol{\mu} \mid \bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}; \boldsymbol{\theta}) \propto f(\bar{\mathbf{x}}_{(i)} \mid \boldsymbol{\mu}, \mathbf{s}_{(i)}; \sigma^2)\, f(\boldsymbol{\mu}; \boldsymbol{\theta}_{\mu}) \text{ with } \boldsymbol{\theta}_{\mu} = (\boldsymbol{\omega}, \boldsymbol{\pi}, \boldsymbol{\delta}, \boldsymbol{\Omega}, \boldsymbol{\psi}),$$

where *f* defines the density probability function.

3.2. Maximum Likelihood estimation via TVF-SAEM algorithm

Now, let define, as in previous section, vectors $\bar{\mathbf{x}}_{(i)} = (\bar{\mathbf{x}}_{it} : t \in \mathcal{H}_i)$, $\mathbf{s}_{(i)} = (\mathbf{s}_{it} : t \in \mathcal{H}_i)$, $\mathbf{z}_{(i)} = (\mathbf{z}_{it} : t \in \mathcal{H}_i)$, $\boldsymbol{\beta}_i = (\beta_{c,i} : c \in C)$ and $\boldsymbol{\mu}_{(i)} = (\boldsymbol{\mu}_{it} : t \in \mathcal{H}_i)$. In our model, $\boldsymbol{\beta}_i$ and $\boldsymbol{\mu}_{(i)}$ are considered as missing data. Then, the complete data of our model consists of the vector of observed variable $\boldsymbol{\zeta}_{(i)} = (\bar{\mathbf{x}}_{(i)}, \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$ and of the vector of unobserved variables $(\boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)})$, for $i = 1,...,N$. 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 $i = 1,...,N$. The contribution of individual *i* to the complete data log-likelihood function at $\boldsymbol{\theta}$ is given by:

$$(21) \quad \ln \ell^c(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}; \mathbf{z}_{(i)}) = \ln f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta})$$

where:

$$(22) \quad \begin{aligned} &\ln f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}_i, \boldsymbol{\mu}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}) \\ &\quad = \sum_{t=1}^{T_{(i)}} \ln \varphi(\bar{x}_{it} - \mathbf{s}_{it}' \mathbf{h}(\boldsymbol{\mu}_{it}); \sigma^2) + \sum_{t=1}^{T_{(i)}} \ln \varphi(\boldsymbol{\mu}_{it} - \boldsymbol{\beta}_i - \mathbf{Z}_{it}\boldsymbol{\delta}; \boldsymbol{\Omega}) + \ln \varphi(\boldsymbol{\beta}_i - \boldsymbol{\omega} - \bar{\mathbf{Z}}_i \boldsymbol{\pi}; \boldsymbol{\psi}). \end{aligned}$$

The term $\varphi(\mathbf{A}, \mathbf{B})$ denotes the probability density function at point **A** of the standard multivariate normal distribution with variance-covariance matrix **B.** Note that the complete data log-likelihood belongs to the exponential family. The corresponding observed data log-likelihood can be obtained by integrating the complete data likelihood with respect to missing data:

$$(23) \quad \ln \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}) = \ln \int f(\bar{\mathbf{x}}_{(i)}, \boldsymbol{\beta}, \boldsymbol{\mu} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}; \boldsymbol{\theta}) d(\boldsymbol{\mu}, \boldsymbol{\beta}).$$

Then, the maximum likelihood estimator is obtained by maximizing the observed data log-likelihood:

$$(24) \quad \boldsymbol{\theta}_{N_{tot}}^{MLE} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln \ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}).$$

Note that $\ell(\boldsymbol{\theta}; \bar{\mathbf{x}}_{(i)} \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)})$ has no closed form expression and that direct maximization of the observed data log-likelihood is problematic since the model is non-linear in its random terms.

Iterative algorithms as the EM algorithm and its variants are suitable in such cases. Here we use the approximated-SAEM algorithm proposed by Allassonnière and Chevallier (2021) which is an extension of the SAEM algorithm proposed Delyon et al, (1999) and Kuhn and Lavielle (2004) in which the simulation step is improved. This algorithm consists in iterating three steps until convergence: a simulation (S) step, a stochastic approximation (SA) step and a maximization (M) step. Due to space limitation, we won't go into more detail in the description of this algorithm here.

3.3. Accounting for weighted data

Let define, as in previous section, vectors $\overline{\mathbf{x}}_{(i)} = (\overline{\mathbf{x}}_{it} : t \in \mathcal{H}_i)$, $\mathbf{s}_{(i)} = (\mathbf{s}_{it} : t \in \mathcal{H}_i)$, $\mathbf{z}_{(i)} = (\mathbf{z}_{it} : t \in \mathcal{H}_i)$, $\boldsymbol{\beta}_i = (\beta_{c,i} : c \in C)$ 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 $\overline{\mathbf{x}}_{it}$ 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:

$$(25) \qquad \ell^c(\boldsymbol{\theta}; \overline{\mathbf{x}}_{(i)}, \boldsymbol{\mu}_{(i)}, \boldsymbol{\beta}_i \mid \mathbf{s}_{(i)}, \mathbf{z}_{(i)}, \mathbf{w}_{(i)}) = f(\boldsymbol{\mu}_{(i)}, \boldsymbol{\beta}_i; \boldsymbol{\theta}) \prod\nolimits_{t=1}^{T_{(i)}} f(\overline{\mathbf{x}}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{(i)}, w_{it}; \boldsymbol{\theta}) .$$

In this decomposition, only the conditional distribution of observed data $\overline{\mathbf{x}}_{it}$ 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:

$$(26) \qquad f(\overline{x}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{it}, w_{it}; \boldsymbol{\theta}) = \tfrac{1}{C(w_{it})} f(\overline{x}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{it}; \boldsymbol{\theta})^{w_{it}}$$

where $f(\overline{\mathbf{x}}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{it}; \boldsymbol{\theta})$ is a probability density function of $\overline{\mathbf{x}}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{it}$, and $C(w_{it})$ is a normalized constant. Indeed, raise $f(\overline{\mathbf{x}}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}; \boldsymbol{\theta})$ to the power w_{it} in maximum likelihood setting is equivalent to "*observing* $\overline{\mathbf{x}}_{it}$ w_{it} *times given* $\boldsymbol{\mu}_{it}$ $\mathbf{s}_{it}$ *and* $\mathbf{z}_{it}$" in standard approaches. However,

$f(\bar{\mathbf{x}}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}; \boldsymbol{\theta})^{w_{it}}$ 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:

(27) $$f(\bar{x}_{it} \mid \boldsymbol{\mu}_{it}, \mathbf{s}_{it}, \mathbf{z}_{it}, w_{it}; \boldsymbol{\theta}) \propto \varphi\ \bar{x}_{it} - \mathbf{s}'_{it}\mathbf{h}(\boldsymbol{\mu}_{it}); (1/w_{it})\sigma^{2(n-1)}\ .$$

This specification including weights slightly change our estimation procedure in its simulation step. Here again, due to space limitation, we won't go into more detail in the description of the algorithm, the detailed procedure, together with the R package WInputAll developed to implement it on various are available from the authors upon request.

4. Empirical application

4.1 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 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 1. 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 1 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 1. 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

4.2. 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.

Our estimations are conducted by using the R package WInputAll developed for this purpose. 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 2 and Table 3, 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 2 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 2. Parameters estimates: estimated distribution of random parameters

	Pesticide (%)			**Fertilizer (%)**		
$\ln(x_{k,it}) = \omega_k + \bar{\mathbf{z}}'_{k,i}\boldsymbol{\pi} + \mathbf{z}'_{k,i}\boldsymbol{\delta} + \eta_{k,i} + \varepsilon_{k,it}$	**Expecta - tion:** ω_k **(SE)**	**Variance of** $\eta_{k,i}$ **(SE)**	**Variance of** $\varepsilon_{k,it}$ **(SE)**	**Expecta - tion:** ω_k **(SE)**	**Variance of** $\eta_{k,i}$ **(SE)**	**Variance of** $\varepsilon_{k,it}$ **(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 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-R^2, which are reported in Table 3. The Sim-R^2 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 3. Fitting criteria

	Pesticide				**Fertilizer**			
	Sim-R2 %	**AAD**	**Estimated mean (S.d.t)**	**Sample Average**	**Sim-R2 %**	**AAD**	**Estimated mean (S.d.t)**	**Sample Average (S.d.t)**
Winter wheat	54.24	0.31	173 (41)	180 (40)	74.53	0.24	148 (40)	187 (47)
Spring barley	16.27	0.24	97 (09)	102 (28)	60.96	0.19	128 (25)	139 (38)
Winter barley	5.12	0.41	178 (16)	150 (41)	47.50	0.29	207 (39)	160 (42)
Corn	2.78	0.37	94 (08)	103 (35)	33.60	0.30	146 (16)	151 (42)
Sugar beet	19.86	0.63	246 (29)	251 (65)	59.79	0.34	274 (53)	224 (71)
Alfalfa	2.51	0.58	11 (0.7)	60 (27)	38.87	0.77	299 (50)	207 (80)
Peas	6.59	0.47	129 (17)	141 (42)	1.30	0.31	131 (08)	66 (33)
Winter rapeseed	33.45	0.36	235 (38)	191 (49)	65.67	0.28	196 (35)	181 (47)
Blue opium poppy	0.28	0.19	74 (6)	63 (28)	0.17	0.26	140 (18)	102 (36)
Potato	7.83	1.53	708 (43)	704 (144)	13.18	0.66	268 (51)	248 (67)
Starch potato	39.2	1.01	448 (23)	480 (113)	53.74	0.52	245 (26)	266 (78)
Total input	82				80			

Figures 1-3 display our results for three selected crops wheat, rapeseed and potato. Input uses are measured per ha in 100€ at the 2005 prices. Figure 1 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 .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 92% of the sampled farms but its average acreage share doesn't exceed 16%. Figure 2 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. Figure 3 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%).

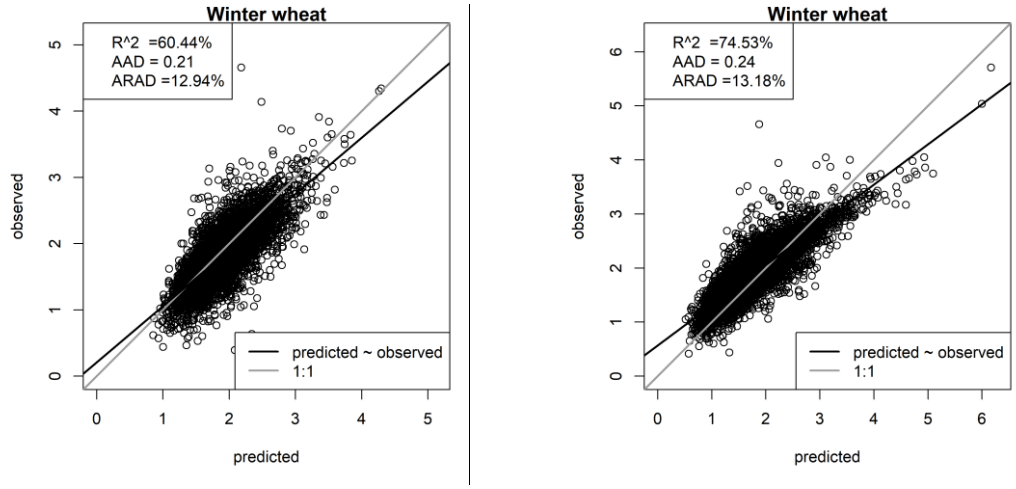

Figure 1. Observed versus Estimated pesticide (left) and fertilizer (right) uses for wheat (Marne dataset, 100€/ha, at 2005 price levels)

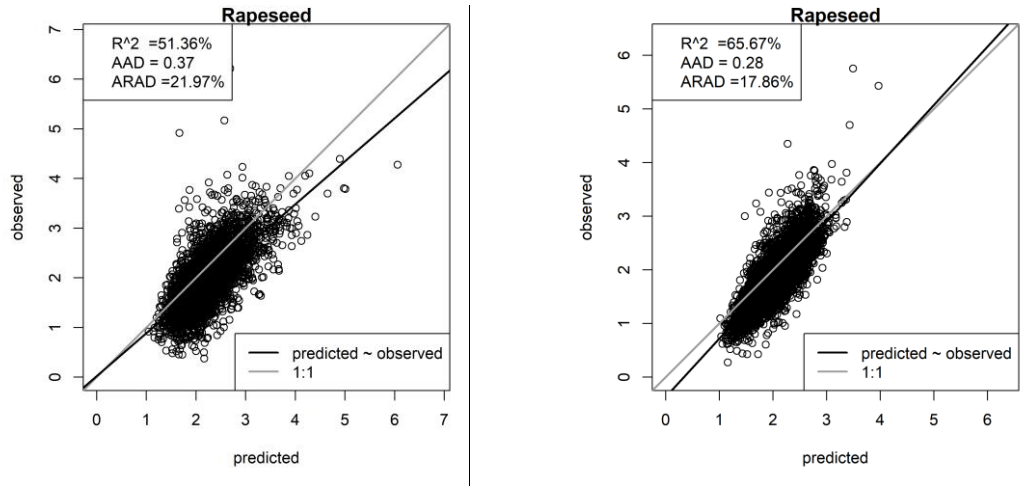

Figure 2. Observed versus Estimated pesticide (left) and fertilizer (right) uses for rapeseed (Marne dataset, 100€/ha, at 2005 price levels)

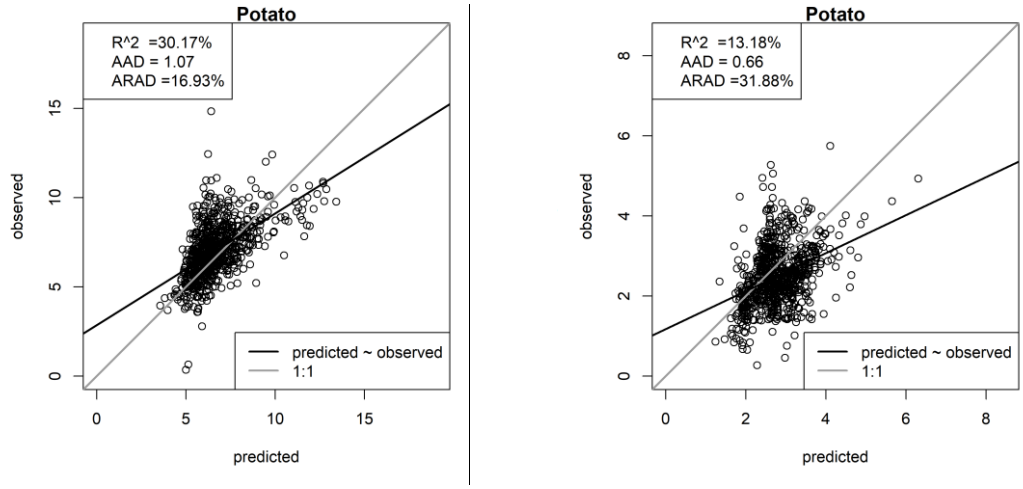

Figure 3. Observed versus Estimated pesticide (left) and fertilizer (right) uses for potato (Marne dataset, 100€/ha, at 2005 price levels)

5. 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.

6. References

Arellano, M., and Bonhomme, S. (2011). Nonlinear panel data analysis. *Annu. Rev. Econ.*, *3*(1), 395-424.

Allassonnière, S., and Chevallier, J. (2021). A new class of stochastic EM algorithms. Escaping local maxima and handling intractable sampling. *Computational Statistics and Data Analysis*, *159*, 107159.

Chambers, R. G., and Just, R. E. (1989). Estimating multioutput technologies. *American Journal of Agricultural Economics*, *71*(4), 980-995.

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.

Comets, E., Lavenu, A., & Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, 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)*, *39*(1), 1-22.

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.

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.

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).

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.

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.

Kuhn, E., and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, *49*(4), 1020-1038.

Lence, S. H., and Miller, D. J. (1998). Recovering Output-Specific Inputs from Aggregate Input Data: A Generalized Cross-Entropy Approach. *American Journal of Agricultural Economics*, *80*(4), 852-867.

Léony, Y., Peeters, L., Quinqu, M., and Surry, Y. (1999). The Use of Maximum Entropy to Estimate Input-Output Coefficients From Regional Farm Accounting Data. *Journal of Agricultural Economics*, *50*(3), 425-439.

Letort, E., and Carpentier, A. (2010). *Variable Input Allocation: Why Heterogeneity Matters?* (No. 704-2016-48235).

Meng, X. L., and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267-278.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69-85.

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.

Panhard, X., and Samson, A. (2009). Extension of the SAEM algorithm for nonlinear mixed models with 2 levels of random effects. *Biostatistics*, *10*(1), 121-135.

Ray, S. C. (1985). Methods of estimating the input coefficients for linear programming models. *American journal of agricultural economics*, *67*(3), 660-665.

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*(411), 699-704.

Wooldridge, J. M. (ed.) (2010). Econometric analysis of cross section and panel data. MIT press.