AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Corner solutions in empirical acreage choice models: an endogenous switching regime approach with regime fixed costs

Philippe Koutchade, PhD candidate, UMR SMART INRA-Agrocampus Ouest, Rennes
ARVALIS Institut du Végétal, Boigneville, Philippe.Koutchade@rennes.inra.fr

Alain Carpentier, Senior Researcher, UMR SMART INRA-Agrocampus Ouest, Rennes ;
Professor, Department of Economics, Agrocampus Ouest, Rennes,
Alain.Carpentier@rennes.inra.fr

Fabienne Féménia, Researcher, UMR SMART INRA-Agrocampus Ouest, Rennes,
Fabienne.Femenia@rennes.inra.fr

Selected Paper prepared for presentation at the 2015 Agricultural & Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28

Copyright 2015 by Philippe Koutchade, Alain Carpentier et Fabienne Féménia. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.

Corner solutions in empirical acreage choice models: an endogenous switching regime approach with regime fixed costs

Philippe Koutchade[a], Alain Carpentier[b], Fabienne Féménia[c]

a PhD candidate, UMR SMART INRA-Agrocampus Ouest, Rennes
ARVALIS Institut du Végétal, Boigneville
Philippe.Koutchade@rennes.inra.fr

b Senior Researcher, UMR SMART INRA-Agrocampus Ouest, Rennes ; Professor, Department of Economics, Agrocampus Ouest, Rennes
Alain.Carpentier@rennes.inra.fr

c Researcher, UMR SMART INRA-Agrocampus Ouest, Rennes
Fabienne.Femenia@rennes.inra.fr

This version: June 2, 2015

Abstract. Corner solution problems are pervasive in micro-econometric acreage choice models because farmers rarely produce the same crop set in a considered sample. Acreage choice models suitably accounting for corner solution need to be specified as Endogenous Regime Switching (ERS) models. Micro-econometric ERS models are however rarely used in practice because their estimation difficulty quickly grows with the dimension of the considered system. Their functional form is generally quite involved and their congruent likelihood functions need to be integrated using simulation methods in most case of interest.

We present here an ERS model specifically designed for empirically modeling acreage choices with corner solutions. This model is theoretically consistent with acreage choices based on the maximization of a profit function with non-negativity constraints and a total land use constraint. It can be combined with yield supply and variable input demand functions. Furthermore, the model accounts for regime fixed costs which represent crop specific marketing and management costs. To our knowledge, this is a unique feature for an ERS model accounting for non-negativity constraints.

The proposed ERS model defines a Nested MultiNomial Logit (NMNL) acreage choice model for each potential production regime. The regime choice is based on a standard discrete choice model according to which farmers choose the crop subset they produce by comparing the different regime profit levels. The structure of the model and the functional form of its likelihood function makes the Simulated Expectation-Maximisation algorithm especially suitable for maximizing the sample likelihood function. The empirical tractability of the model is illustrated by the estimation of a five crop production choice model for a sample of French grain crop producers.

Keywords. Corner solutions, endogenous regime switching models, agricultural production choices

JEL codes: Q12, C13, C15

Corner solutions in empirical acreage choice models: an endogenous switching regime approach with regime fixed costs

Introduction

Corner solution problems are pervasive in micro-econometric acreage choice models because farmers rarely produce the same crop set in a considered sample, even in samples considering specialized farms. Farmers' choice to produce or not a crop might be seen as purely exogenous to their production (yield, acreage) decisions, by considering, for instance, that a crop is not produced because of an absence of market opportunities or because of a lack of human and/or physical capital. However these production impossibilities only prevail in the short run. On the other hand, the choices of farmers not to produce a crop can be endogenous to their production decisions: first, a crop is grown only if it is profitable enough, *i.e.* if its associated gross margin is positive; second, fixed costs associated to the set up of a given crop can be higher than its potential profit, in that case the crop is not grown even if its gross margin is positive. Both intensive and extensive margin aspects thus intervene in the choice of farmers to produce a crop subset.

Agricultural economists usually use two approaches to cope with null crop acreages. First, crops can be aggregated to eliminate or, at least, attenuate the occurrence of null crop acreages. Of course, this approach can lead to substantial information loss. Second, corner solutions can be dealt with by specifying acreage choices as a system of censored regressions (see, *e.g.*, Platoni *et al*., 2012). However, if censored regression systems explicitly account for null crop acreages from a statistical viewpoint, they cannot consistently represent acreage

choices with corner solutions. This point was made, for consumer demand systems, by Arndt *et al*. (1999). This is easily seen by considering a simple example. Let consider a farm producing wheat but not producing barley. In a censored regression framework, the wheat acreage depends on the price of barley. This cannot occur if the considered farmer's acreage choice is the solution to a profit maximization problem (with non-negativity constraints on the acreage choices). In such a case, the price of a non produced crop cannot impact the acreages of the produced crops.

More generally, acreage choice models suitably accounting for corner solutions need to be specified as endogenous regime switching models. In such models, regimes are defined by the subsets of crops with non null acreages – *i.e.* by the subsets of actually produced crops – and the acreage choice model of a produced crop depends on the regime where this crop is produced. *E.g*., in a regime where wheat and barley are both produced, the wheat acreage depends on the price of barley whereas it doesn't depend on the price of barley in regimes in which the barley acreage is null. Micro-econometric endogenous regime switching models where mostly defined to model consumer demand systems (see, *e.g*., Kao *et al*. 2001) or firm input demand systems (see, *e.g*., Chakir et al. 2004), following the pioneering work of Wales and Woodland (1983) and of Lee and Pitt (1986). However, endogenous regime switching models are rarely used in practice because their estimation difficulty quickly grows with the dimension of the considered system. The functional form of these models is generally quite involved and their congruent likelihood functions need to be integrated using simulation methods in most case of interest – *i.e.* for models considering at least four alternatives with reasonable assumptions related to the random parts of the model (see, *e.g.*, Kao *et al*. 2001).

Our main objective in this paper is to present an endogenous regime switching model specifically designed to empirically model acreage choices with corner solutions. To our knowledge, this is the first model proposed for this purpose. This model has three main features. First, it is theoretically consistent – in its deterministic part as well as in its random parts – with acreage choices based on the maximization of a profit function with non-negativity constraints and a total land use constraint on the acreage choices. Second, this model can be combined with yield supply and variable input demand functions. Third, this model accounts for regime fixed costs. This cost accounts for marketing and management costs. To our knowledge, the ability to account for regime fixed costs is a unique feature for an endogenous regime switching model accounting for non-negativity constraints.

The proposed endogenous regime switching model is an extension of the Nested MultiNomial Logit (NMNL) acreage choice model proposed by Carpentier and Letort (2014) to model acreage choices with corner solutions. It heavily relies on the unique features of the NMNL acreage choice models: their parameter parsimony, their providing well-behaved and simple acreage choice models and their congruent profit functions. The proposed multicrop model defines an NMNL acreage choice model for each potential regime. The production regime choice is based on a standard discrete choice model according to which farmers choose the crop subset they produce by comparing the different regime profit levels, including the regime specific costs.

The considered multicrop model is fully parametric and, as a result, can be efficiently estimated within the Maximum Likelihood estimation framework. The structure of the model and the functional form of its likelihood function make the version of the Simulated Expectation-Maximization algorithm (see, *e.g.*, McLachlan and Krishnan 2008) developed by Delyon *et al.* (1999) especially suitable for maximizing the sample (simulated) likelihood function.

In order to illustrate the empirical tractability of the proposed model and assess its adequacy to observed data, a five crop production choice model is estimated for a sample of French grain crop producers covering the 1993-2007 period. This model involves six production regimes. It is composed of five yield supply functions, of four acreage (share) choices – the functional forms of which depend on the production regime – and of a probabilistic regime choice model function. Estimation results demonstrate that the model fits relatively well to the data, at least for the crops with large acreages. These first results also show that considering regime fixed costs is important to correctly model the regime choices. They are encouraging but also clearly show that the considered empirical model can be improved.

The proposed multicrop model including regime fixed costs is presented in the first section. Identification and estimation issues are discussed in the second section. The illustrative estimation results are provided in the third section. Finally we conclude.

Acreage choice modeling, corner solutions and production regime fixed costs

The main aim of this section is to present a fairly general framework for modeling acreage choices and accounting for possible corner solutions and for fixed costs related to the set of crops actually produced. It combines three elements: a given formulation of the farmers' acreage choice problem, a relevant decomposition of this problem and specific properties of the MNL framework developed by Carpentier and Letort (2014) for modeling farmers' acreage choices.

General modeling framework

Let consider a risk neutral farmer who can allocate his cropland to *K* crops. Crop *k*, with $k \in \mathcal{K} \equiv \{1,...,K\}$, provides an expected gross margin denoted by π_k. The crop gross margins are collected in the vector $\boldsymbol{\pi} \equiv (\pi_k : k \in \mathcal{K})$. His problem is to maximize the expected profit of his cropland by solving the following optimization problem:

(1) $\max_{\mathbf{s} \in \mathcal{U}} \{\mathbf{s}'\boldsymbol{\pi} - C(\mathbf{s})\}$ where $\mathcal{U} \equiv \{\mathbf{s} \geq \mathbf{0} \text{ et } \mathbf{s}'\boldsymbol{\iota} = 1\}$.

The term $\mathbf{s} \equiv (s_k : k \in \mathcal{K})$ denotes the vector of crop acreage shares and the term $\boldsymbol{\iota}$ denotes a vector of ones.[1] The $\mathcal{U}$ term defines the set of admissible acreages, *i.e.* those satisfying the non-negativity constraints $\mathbf{s} \geq \mathbf{0}$ and the total land use constraint $\mathbf{s}'\boldsymbol{\iota} = 1$. The objective function considered in problem (1) defines a trade-off between the sum of the expected gross margins weighted by their acreage shares, $\mathbf{s}'\boldsymbol{\pi} = \sum_{k \in \mathcal{K}} s_k \pi_k$, and the implicit management cost of the acreage **s**, $C(\mathbf{s})$. The function $C : \mathcal{U} \to \mathbb{R}_+$ is further described below. Such cost functions are used in the Positive Mathematical Programming (PMP) literature (see, *e.g.*, Howitt, 1995 ; Heckeleï *et al*, 2012) and in the Multicrop Econometric (ME) literature (see, *e.g.*, Capentier and Letort, 2012, 2014).

Let $\mathbf{s}^o$ denote the solution in **s** to problem (1), *i.e.*:

(2) $\mathbf{s}^o \equiv \arg\max_{\mathbf{s} \in \mathcal{U}} \{\mathbf{s}'\boldsymbol{\pi} - C(\mathbf{s})\}$.

The solution in s_k to problem (1) is a corner solution if $s_k^o = 0$, it is interior if $s_k^o > 0$. The type of solution in to problem (1) can be characterized by the production "regime" of $\mathbf{s}^o$. The regime of an acreage **s** is defined by the crop subset with strictly positive acreages, *i.e.* by the subset of crops actually produced in the acreage described by **s**. Let define by $\mathcal{R} \equiv \{0,1,...,R\}$ the set of indicators of the different subsets of $\mathcal{K}$. The term $\mathcal{K}_{(r)}$ defines the subset of crops

produced in regime *r*. *I.e.*, $\mathcal{K}_{(r)}$ is a subset of $\mathcal{K}$ and *r* is the regime of **s** if and only if $\{k \in \mathcal{K} / s_k > 0\} = \mathcal{K}_{(r)}$. The function $\rho : \mathcal{U} \to \mathcal{R}$ defines the regime of **s**, with $\rho(\mathbf{s}) = r$ if *r* is the regime of **s**.[2] Of course, we have $\mathcal{K}_{(r)} \subseteq \mathcal{K}$ for $r \in \mathcal{R}$, and we impose $\mathcal{K}_{(0)} \equiv \mathcal{K}$. The term $K_{(r)}$ defines the cardinality of $\mathcal{K}_{(r)}$. We will say that regime *j* is (strictly) included in regime *r* if and only if $\mathcal{K}_{(j)} \subseteq (\subset) \mathcal{K}_{(r)}$.

The regime of the optimal acreage choice $\mathbf{s}^o$ defines the subset of crops actually produced by the considered farmer, *i.e.* the optimal production regime. Farmers simultaneously decide the crop subset to be produced and the corresponding optimal acreage. Agricultural production economists using MP models routinely solve agricultural production choice models similar to problem (1). But they rarely consider production regimes, at least explicitly. They simply account for corner solutions in the acreage choices when they occur.

When seeking to define tractable ME models, it is tempting to decompose farmers' decision into two steps: the regime choice on the one hand, and the acreage choice conditional on the regime choice on the other hand. Most ME models explicitly accounting for corner solutions in the acreage choices rely on such a decomposition (Skockaï and Moro, 2006, 2009; Lacroix and Thomas, 2011; Fezzi and Bateman, 2011). These models define the acreage choice models as systems of censored regressions. Shonkwiler and Yen (1999) proposed a two-step estimator for censored regression systems which can be interpreted as an extension in the multivariate case of the two-step estimator proposed by Heckman (1976, 1979) for the estimation of sample selection models. The pioneering work of Shonkwiler and Yen (1999) has been the basis of numerous two-step estimators, developed in particular by Yen and his co-authors (Yen *et al.*, 2002, 2003; Yen, 2005), which were applied for numerous empirical analyses of consumption choices or of acreage choices (Skockaï and Moro, 2006;

Fezzi and Bateman, 2011) involving corner solutions. These estimators relate to econometric models in which the regime choice model is based on a set of censoring conditions on a system of virtual acreage/consumption choice. This regime choice model is estimated in a first step. The second step consists in estimating the system of virtual acreage choices conditionally on the observed regime choices.

However, this focus on the regime choice may be misleading. In particular, acreage choices cannot be consistently modeled as systems of censored regressions if these choices are the solution in **s** to profit maximization problems similar to problem (1). This point was made by Arndt *et al* (1999) for the econometric modeling of consumption demand systems. It will be discussed below for the econometric modeling of acreage choices. As a matter of fact, it will be shown that acreage choices with non-negativity constraints must be defined as endogeneous switching regime models similar to those proposed by Wales and Woodland (1983) or by Lee and Pitt (1986) for modeling consumer choices.

Nevertheless, this doesn't mean that modeling the regime choice is useless. To consider the regime choice appears to be extremely useful when the considered choices involve regime fixed costs. These regime fixed costs are specific features of the model we propose for modeling acreage choices. But modeling the regime choice is difficult unless the optimization problem defining the considered choices has specific features. This is at this point where the MNL framework proposed by Carpentier and Letort (2014) comes to play. The regime fixed costs are presented first. The properties of interest of the MNL framework are presented in a second step.

Production regime fixed costs and acreage choice modeling

The cost function $C(\mathbf{s})$ involved in problem (1) is further specialized in order to highlight the role of regime fixed costs in the acreage choice optimization problem. It is assumed that the acreage management cost function $C(\mathbf{s})$ can be decomposed into two distinct parts with

(3) $\quad C(\mathbf{s}) = D(\mathbf{s}) + b_{\rho(\mathbf{s})}$.

The function $D : \mathcal{U} \to \mathbb{R}_+$ is continuously differentiable and strictly convex in **s** on $\mathcal{U}$. The term $b_{\rho(\mathbf{s})}$ is the element of the production regime fixed cost vector $\mathbf{b} \equiv (b_r : r \in \mathcal{R})$ corresponding to the regime of **s**.

The "smooth" part of the implicit acreage management cost function $C(\mathbf{s})$, *i.e.* $D(\mathbf{s})$, accounts for the crop costs not included in the crop gross margins $\boldsymbol{\pi}$ and the implicit costs related to the constraints on the acreage choices due the limiting quantities of quasi-fixed factors. Quasi-fixed factor constraints and the associated peak load costs provide motives for diversifying crop acreages. These implicit costs imply that the function $D(\mathbf{s})$ can be assumed to be convex in **s** on $\mathcal{U}$. This function is assumed to be strictly convex in **s** for simplicity, this assumption implying that the solution in **s** to problem (1) is unique.

The fixed cost b_r is incurred by the considered farmer for any acreage choice in regime *r*. Such regime fixed costs do not depend on the acreages of the crops in the considered regime, they just depend on the crop set defining this regime. The terms collected in **b** may account for different features of the acreage management process. (*i*) They may account for transaction costs such as the fixed costs related to the marketing process of the crop products, those related to the monitoring of the pest populations of the crops, or those incurred when purchasing specific variable inputs or when renting specific machines. (*ii*) The regime fixed costs **b** may account for the fact that specific crops require farmer's availability at specific

dates during the production campaign. (*iii*) The functional of the $D(\mathbf{s})$ term has to be "smooth" in **s** and sufficiently simple for the model to be empirically tractable. The **b** terms may also partly correct for the specification errors of the $D(\mathbf{s})$ term.

The regime fixed cost $b_{\rho(\mathbf{s})}$ and the "smooth" acreage management costs $D(\mathbf{s})$ are expected to have opposite effects on crop diversification. Whereas the "smooth" acreage management costs tend to favor diversified crop acreages, the regime fixed cost $b_{\rho(\mathbf{s})}$ are expected to deter crop diversification. The transaction costs and the labor requirement related to a production regime increase in the number of crops produced in this regime. But the fixed cost of a regime may be inferior to the sum of the fixed costs of the individual crops of the considered regime. The regime fixed cost is likely to be sub-additive in the individual crop fixed costs. *E.g.*, farmers' may purchase the input specific to different crops from the same supplier, implying savings in the related transaction costs.[3]

A few remarks are in order with respect to the $D(\mathbf{s})$ and **b** terms. (*i*) They could be more explicitly defined. *E.g.* they could be defined as the virtual costs associated to the explicit "hard" constraints on work time of specific optimization problems. Our view is that such modeling exercise is of limited interest for two reasons. First, the properties of the implicit costs captured in the $D(\mathbf{s})$ and **b** terms are relatively simply interpreted. Second, these terms and their determinants are likely to significantly differ across farms.[4] These terms are assumed to be farmer specific from a theoretical viewpoint. Their heterogeneity across farms mainly is an empirical issue.[5] (*ii*) The $D(\mathbf{s})$ and $b_{\rho(\mathbf{s})}$ terms are related to the short run implementation of acreage **s** by the considered farmer. These terms do not account for investment costs. In other words, it is assumed here that the considered farmer is able to produce any crop *k* in the set $\mathcal{K}$ and to implement any acreage **s** in $\mathcal{U}$. In particular, the regime fixed costs b_r is incurred each year in which any acreage in regime *r* is implemented. The

ability of the considered farmer to implement any acreage **s** in $\mathcal{U}$ is necessary for the existence of the "smooth" acreage management cost function $D(\mathbf{s})$ whose domain is $\mathcal{U}$.[6] This limits the scope of this modeling framework for empirical purposes. The sampled farms must consider "common" crop sets and must be located in regions with suitable market opportunities. (*iii*) Finally, the regime fixed cost $b_{\rho(\mathbf{s})}$ only depends on the considered production regimes and imply that $C(\mathbf{s})$ is discontinuous in s_k at $s_k = 0$. This implies that problem (1) contains discrete choice features due to the regime fixed costs. *I.e.*, the characterization of the optimal acreage choice needs to partly rely on the characterization of the production regime choice induced by the regime fixed costs. A production regime could be optimal without these costs and sub-optimal with these costs. The discrete optimization features introduced in problem (1) by the regime fixed costs imply that a solution approach to this problem needs to rely on some mechanism aimed at comparing the outcomes related to the possible production regime choices.

As a consequence of the last remark, it appears necessary to investigate the profit optimization problem on a per regime basis, *i.e.* by restricting the acreage choices within the crop subsets $\mathcal{K}_{(r)}$ for $r \in \mathcal{R}$. Let define $\mathbf{s}_{(r)}$ the subvector of **s** containing the acreages of the crop set defining $\mathcal{K}_{(r)}$, and let define the vector $\boldsymbol{\pi}_{(r)}$ and the set $\mathcal{U}_{(r)}$ accordingly. Let also define the $K_{(r)} \times K$ selection matrix $\mathbf{Q}_{(r,+)}$ such that $\mathbf{Q}_{(r,+)}\mathbf{s} = \mathbf{s}_{(r)}$.[7] It is also easily shown that the product $\mathbf{Q}'_{(r)}\mathbf{s}_{(r)}$ defines a dimension *K* vector which can be obtained from **s** by setting at 0 its elements corresponding to the crops not contained in $\mathcal{K}_{(r)}$.[8]

Ignoring the regime fixed costs for the moment, we investigate the solutions to the profit maximization problems

(4) $\max_{\mathbf{s}_{(r)} \in \mathcal{U}_{(r)}} \{\mathbf{s}'_{(r)}\boldsymbol{\pi}_{(r)} - D(\mathbf{Q}'_{(r,+)}\mathbf{s}_{(r)})\}$

for $r \in \mathcal{R}$. The solution in $\mathbf{s}_{(r)}$ to problem (4):

(5) $\mathbf{s}^{o}_{(r)} \equiv \arg\max_{\mathbf{s}_{(r)} \in \mathcal{U}_{(r)}} \{\mathbf{s}'_{(r)}\boldsymbol{\pi}_{(r)} - D(\mathbf{Q}'_{(r,+)}\mathbf{s}_{(r)})\}$,

necessarily belongs to a production regime included in regime *r*. But it doesn't necessarily belong to regime *r*, depending on the properties of the *D* function. *E.g.*, nothing prevents some elements of $\mathbf{s}^{o}_{(r)}$ to be null in the case where $D(\mathbf{s})$ is quadratic in **s**. This frequent feature of many acreage choice models largely undermines the interest of a decomposition of the acreage problem into a sequence of two optimization problems: the production regime choice in a first step followed by the acreage choice problem conditional on the optimal production regime in a second step. The problem described above would have been worsened by considering the regime fixed costs in problem (5).

Acreage choices, production regime choices and the MNL modeling framework

As a conclusion, a difficult arises when considering regime fixed costs. Accounting for regime fixed costs requires a simple characterization of the production regime based on the profit levels obtained in the different possible regimes. But, at the same time, the properties of the solution in **s** to profit maximization problems such as problem (4) generally prevent the existence of such a simple characterization of the optimal production regime. Basically, this difficulty would not arise if the solution in $\mathbf{s}_{(r)}$ to problem (4) was guaranteed to belong to regime *r*. In that case, problem (1) could be decomposed as

(6) $\max_{r \in \mathcal{R}} \left\{ \max_{\mathbf{s}_{(r)} \in \mathcal{U}_{(r)}} \{\mathbf{s}'_{(r)}\boldsymbol{\pi}_{(r)} - D(\mathbf{Q}'_{(r,+)}\mathbf{s}_{(r)})\} - b_r \right\}$,

implying that the optimal regime choice would be given by:

(7) $r^{o} = \arg\max_{r \in \mathcal{R}} \{\Pi^{o}_{(r)} - b_r\}$

where:

(8) $\Pi^{o}_{(r)} \equiv \max_{\mathbf{s}_{(r)} \in \mathcal{U}_{(r)}} \{\mathbf{s}'_{(r)}\boldsymbol{\pi}_{(r)} - D(\mathbf{Q}'_{(r,+)}\mathbf{s}_{(r)})\}$

Of course the optimal acreage choice would be given by $\mathbf{s}^{o} = \mathbf{Q}'_{(r,+)}\mathbf{s}^{o}_{(r)}$ with $r = r^{0}$.

The MNL cost functions proposed by Carpentier and Letort (2014) have the relevant properties. These properties are presented in the case of the Standard MNL cost function, for simplification purposes. The Nested MNL cost functions basically have the same properties but are more flexible. The empirical model presented in the next section relies on a three level Nested MNL cost function.

Choosing the Standard MNL functional form for the $D(\mathbf{s})$ function implies that:

(9) $D(\mathbf{s}) \equiv A - \mathbf{s}'\mathbf{c} - \alpha^{-1} \times \mathbf{s}' \ln \mathbf{s}$ with $\alpha > 0$

where A is an unidentifiable fixed cost term and $\mathbf{c}$ is a parameter vector. This function is continuous and strictly convex in $\mathbf{s}$ on $\mathbb{R}^{+}$ and it is continuously differentiable "at will" in $\mathbf{s}$ on $\mathbb{R}^{*}_{+}$.[9] It is easily shown that

(10) $D(\mathbf{Q}'_{(r,+)}\mathbf{s}_{(r)}) \equiv A - \mathbf{s}'_{(r)}\mathbf{c}_{(r)} - \alpha^{-1} \times \mathbf{s}'_{(r)} \ln \mathbf{s}_{(r)}$ for $r \in \mathcal{R}$.

The solutions to problem (4) are given by acreage share choice with Standard MNL functional forms

(11) $s^{o}_{(r),k} = \dfrac{\exp\left(\alpha(\pi_k - c_k)\right)}{\sum_{\ell \in \mathcal{K}_{(r)}} \exp\left(\alpha(\pi_\ell - c_\ell)\right)}$ for $k \in \mathcal{K}_{(r)}$

and by indirect profit function with log-sum function forms

(12) $\Pi^{o}_{(r)} = \alpha^{-1} \ln \sum_{\ell \in \mathcal{K}_{(r)}} \exp\left(\alpha(\pi_\ell - c_\ell)\right)$.

The solution in $\mathbf{s}_{(r)}$ to problem (4), $\mathbf{s}^{o}_{(r)}$, is interior. This property is due to the entropy term of the Standard MNL cost function. This function is continuous in $\mathbf{s}$ on $\mathcal{U}$ but, due to the entrpopy term $-\alpha^{-1} \times \mathbf{s}' \ln \mathbf{s}$, its first derivative in s_k diverges to $+\infty$ as s_k goes to 0. Regime r

indirect profit function $\Pi^{o}_{(r)}$ has the so-called log-sum functional form. This finally implies that the optimal regime choice is given by

(13) $r^{o} \equiv \arg\max_{r\in\mathcal{R}} \{\alpha^{-1} \ln \sum_{\ell\in\mathcal{K}_{(r)}} \exp\left(\alpha(\pi_{\ell} - c_{\ell})\right) - b_{r}\}$,

whereas the optimal acreage choice is provided by equation (11) at $r = r^{o}$.

A few remarks are in order with respect to these results. (*i*) To rely on the MNL framework allows for decomposing problem (1) into two simple steps: the production regime choice step and the step providing the acreage choice given the optimal regime choice. (*ii*) The MNL framework has another advantage in this context. The solution to the regime choice problem requires to compute the regime indirect profit functions $\Pi^{o}_{(r)}$ for $r \in \mathcal{R}$, *i.e.* to solve problem (4) for $r \in \mathcal{R}$. The MNL framework provides analytical closed form solutions to the regime profit maximization problem (4) for $r \in \mathcal{R}$, *i.e.* it directly provides simple functional forms for the $\Pi^{o}_{(r)}$ and $\mathbf{s}^{o}_{(r)}$ terms as show by equations (11) and (12). (*iii*) Moreover these functional forms are "smooth" in the parameters $\boldsymbol{\pi}$, α and **c**. This simplifies the theoretical analysis of the properties of the statistical inference tools to be used for estimating these parameters. (*iv*) The last remark deserves a specific discussion as it relates to a specific drawback of the MNL framework. As a matter of fact, this problem can be interpreted as the price to pay for buying the desirable properties of this modeling framework in a multiple regime context, those discussed in remarks (*i*)–(*iii*).

The functional form of the Standard MNL cost function given in equation (9) implies that $\Pi^{o}_{(r)} > \Pi^{o}_{(j)}$ if regime *j* is strictly included in regime *r*. *I.e.* the Standard MNL indirect profit function given in equation (9) increases as long as crops are added to the crop set. Basically, the Standard MNL cost function tends to overstate the interest in diversifying crop acreages.

This undesirable property is due to the entropy term of the Standard MNL cost function, the one ensuring that the solution in $\mathbf{s}_{(r)}$ to problem (5) is strictly positive.

In order to illustrate these points, let consider a simple example where regime *r* is obtained from regime *j* by adding crop *k*. In this case we have $\Pi^{o}_{(r)} = \Pi^{o}_{(j)} - \alpha^{-1}\ln(1-s^{o}_{(r),k})$. Because we necessarily have $s^{o}_{(r),k} \in (0,1)$, the inequality $\Pi^{o}_{(r)} > \Pi^{o}_{(j)}$ necessarily holds. It is also easily shown that the acreage of crop *k* in production regime *r*, $s^{o}_{(r),k}$, tends to 0 as the profitability of this crop decreases relatively to the other crops, *i.e.* as $\min_{\ell \in \mathcal{K}_{(r)}}\{\pi_k - \pi_\ell\}$ tends to $-\infty$. This implies that $\Pi^{o}_{(r)} - \Pi^{o}_{(j)}$ is close to (but strictly positive) be null if crop *k* much less profitable than the other crops of regime *j*. *I.e.*, the interest in adding crop *k* in the acreage decreases as the profitability of crop *k* decreases in comparison to the other crops of regime *j*. As a result, the Standard MNL cost function tends to bias acreage choices toward diversified acreages. But, the implied biases toward crop diversification tend to decrease with respect to the relative profitability of the considered crops according to intuitive mechanisms.

This property of the $\Pi^{o}_{(r)}$ and $\mathbf{s}^{o}_{(r)}$ terms as functions of $K_{(r)}$ was discussed by Ackerberg and Rysman (2005) for cases where the Standard MNL discrete choice model is used for investigating consumer choices among sets of differentiated goods. In this context the counterparts of the $\mathbf{s}^{o}_{(r)}$ terms are the vectors of the choice probability functions. The counterparts of the $\Pi^{o}_{(r)}$ terms are used as consumer welfare measures. These mechanically increase in $K_{(r)}$, the number of goods available on the considered markets, as a consequence of the taste for the diversity of the available choice sets implied by the Standard MNL model. This feature of the Standard MNL discrete choice model implies that markets cannot be crowded out by the supply of many very similar albeit different goods. Ackerberg and Rysman (2005) propose simple devices for alleviating this drawback of the Standard MNL

discrete choice model. When adapted to the acreage choice problem, the most flexible suggestion of Ackerberg and Rysman (2005) consists in adding regime specific terms to the regime indirect profit functions $\Pi^{o}_{(r)}$. This implies that the regime fixed costs b_r simultaneously serve two purposes in our modeling framework. As agued above, the regime fixed costs **b** are mainly intended to account for regime fixed costs in the acreage choice problem. But they also contribute to alleviate a drawback of the MNL cost and indirect profit functions, their biases toward crop diversification in a multiple production regime context. Note also that in our empirical illustration, this problem is attenuated, but not completely solved, by using a (three level) Nested MNL cost function instead of a Standard MNL cost function. This allows both for more flexibility in the acreage choice model and for attenuating the biais toward crop diversification, at least when corner solutions do not occur near the root of the tree representing the nesting structure of the crop set.

A tractable multicrop micro-econometric model with corner solution

For each sampled farmer, the variable vector to be modeled is composed of the observed production regime r_i , the observed yield levels of the produced crops $\mathbf{y}_i^{+} \equiv (y_{k,i} : k \in \mathcal{K}_{(r_i)})$ and the observed acreage shares of the produced crops $\mathbf{s}_i^{+} \equiv (s_{k,i} : k \in \mathcal{K}_{(r_i)})$. The vector $\mathbf{z}_i \equiv (\mathbf{p}_i, w_i, \mathbf{v}_i)$ contains the main determinants of farmer *i* observed production choices: the expected price vector of the considered crop set $\mathbf{p}_i \equiv (p_{k,i} : k \in \mathcal{K})$, the price index of the aggregated variable input uses w_i and a vector containing aggregated climatic variables and variables used for defining time trends $\mathbf{v}_i \equiv (\mathbf{v}_{k,i} : k \in \mathcal{K})$. Farmers are assumed to have naïve price expectations, *i.e.* $\mathbf{p}_i$ is defined as the price vector obtained by farmer *i* the preceding year.[10]

The considered multicrop micro-econometric model is designed as a statistical model of $(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i)$ conditional on $\mathbf{z}_i$. It is thus composed of three parts: a system of yield supply models, a system of acreage share choice models and a probabilistic production choice model.

The yield supply, acreage share choice and production regime choice models

The model of the yield supply of crop *k*, whether crop *k* is produced by farmer *i*, is given by:

(14) $y_{k,i} = \beta_{k,i}^{y} + \mathbf{v}'_{k,i}\boldsymbol{\eta}_k - 1/2 \times \gamma_k \times w_i^2 \, p_{k,i}^{-2} + \varepsilon_{k,i}^{y}$ where $E[\varepsilon_{k,i}^{y} \mid i] = 0$ and $E[\mathbf{v}_{k,i} \mid i] = \mathbf{0}$.

The curvature parameter γ_k is required to be strictly positive for this yield supply model to be well behaved. The considered yield supply model contains two random terms. The random parameter $\beta_{k,i}^{y}$ basically captures the yield level heterogeneity across farms. It is assumed to be known to farmer *i* at the time of his acreage choice. The term $\varepsilon_{k,i}^{y}$ is unknown to farmer *i* at the time of his acreage choice. It accounts for the effects of the stochastic events not included in $\mathbf{v}_{k,i}$ and foregone by farmer *i*.

The yield supply model given in equation (14) can be obtained as the optimal expected yield level obtained by maximizing the expected gross margin of crop *k* (see, *e.g.*, Carpentier and Letort, 2012, 2014). In this modeling framework, and assuming that $E[\mathbf{v}_{k,i} \mid i] = \mathbf{0}$, the random parameter $\beta_{k,i}^{y}$ is interpreted as a measure of the maximum yield of crop *k* for farmer *i* (as it can be expected by farmer *i* at the time he chooses his acreage). Also, the congruent expected (at the time of the acreage choice) gross margin level is given by:

(15) $\pi_{k,i} = p_{k,i}\beta_{k,i}^{y} + 1/2 \times \gamma_k \times w_i^2 \, p_{k,i}^{-1} - w_i \beta_{k,i}^{x}$

where the term $\beta_{k,i}^{x}$ characterizes the variable input demand for crop *k* of farmer *i*.

Farmer *i* acreage choice is assumed to be the solution in **s** to the expected profit maximization problem with regime fixed costs, *i.e.*:

(16) $\mathbf{s}_i \equiv \arg\max_{\mathbf{s}\in\mathcal{V}}\{\mathbf{s}'\boldsymbol{\pi}_i - D_i(\mathbf{s}) - b_{\rho(\mathbf{s}),i}\}$.

where $\boldsymbol{\pi}_i \equiv (\pi_{k,i} : k \in \mathcal{K})$. The "smooth" part of the acreage management implicit cost function, $D_i(\mathbf{s})$, is a three level Nested MNL cost function. The considered three level nesting structure of the crop set based on the implicit costs of the acreage management. The crops belonging to a sub-group compete more for the farmers' quasi-fixed factors limiting quantities that they compete with crops of the other sub-groups. They also have similar agronomical roles in the crop rotations. The crop groups are also defined along these lines. The considered three level nesting structure of the crop set is formally defined as follows. The crop set $\mathcal{K}$ is assumed to be split into *G* groups. Each group $g \in \mathcal{G} \equiv \{1,...,G\}$ is itself split into M_g sub-groups of crops, with $\sum_{g\in\mathcal{G}} M_g = M$. These sub-groups are denoted by $\mathcal{K}_m$ for $m \in \mathcal{M} \equiv \{1,...,M\}$. This three level nesting structure of the crop set is depicted in Figure 1 for the illustrative application considered in the next section. The $D_i(\mathbf{s})$ term is defined as

(17)
$$\begin{aligned} D_i(\mathbf{s}) = A_i &+ \sum_{k\in\mathcal{K}} s_k c_{k,i} + \alpha^{-1}\sum_{g\in\mathcal{G}}(1-\alpha\delta_g^{-1})s_g^{\mathcal{G}}\ln s_g^{\mathcal{G}} \\ &+ \sum_{g\in\mathcal{G}}\delta_g^{-1}\sum_{m\in\mathcal{M}_g}(1-\delta_g\tau_m^{-1})s_m^{\mathcal{M}}\ln s_m^{\mathcal{M}} + \sum_{m\in\mathcal{M}}\tau_m^{-1}\sum_{k\in\mathcal{K}_m} s_k \ln s_k \end{aligned}$$

where $s_m^{\mathcal{M}}$ denotes the acreage share of sub-group, *i.e.* $s_m^{\mathcal{M}} \equiv \sum_{k\in\mathcal{K}_m} s_k$ for $m \in \mathcal{M}$ and $s_g^{\mathcal{G}}$ denotes that of group, *i.e.* $s_g^{\mathcal{G}} \equiv \sum_{m\in\mathcal{M}_g}\sum_{k\in\mathcal{K}_m} s_k$ for $g \in \mathcal{G}$. The term $s_{k|m}^{\mathcal{K}|\mathcal{M}}$ denotes the acreage share of crop *k* within the acreage share of sub-group *m*, *i.e.* $s_{k|m}^{\mathcal{K}|\mathcal{M}} \equiv s_k / s_m^{\mathcal{M}}$. Similarly, $s_{m|g}^{\mathcal{M}|\mathcal{G}}$ denotes the acreage share of sub-group *m* within the acreage acreage of group *g*, *i.e.* $s_{m|g}^{\mathcal{M}|\mathcal{G}} \equiv s_m^{\mathcal{M}} / s_g^{\mathcal{G}}$.

This cost function is strictly convex in **s** on $\mathcal{U}$ if $\tau_m \geq \delta_g \geq \alpha > 0$ for $(m,g) \in \mathcal{M}_g \times \mathcal{G}$. If sub-group m only contains a single crop then $\tau_m = \delta_g$ if $\mathcal{K}_m$ belongs to group *g*. Similarly, if group *g* only contains a single sub-group then $\delta_g = \alpha$.

Due to the total land use constraint, the $c_{k,i}$ terms are only defined up to an additive farmer specific term. The normalization constraint $c_{1,i} \equiv 0$ implicitly defines the $c_{k,i}$ terms for $k \in \mathcal{K}^-$ where $\mathcal{K}^- \equiv \{2,...,K\}$, as differences with respect to their counterparts at $k = 1$. Note that this normalization can only be used in empirical work if crop 1 is always produced.

The cost function curvature parameters, *i.e.* α, $\boldsymbol{\delta} \equiv (\delta_g : g \in \mathcal{G})$, $\boldsymbol{\tau} \equiv (\tau_m : m \in \mathcal{M})$ are assumed to be constant across farmers. Of course, this assumption is restrictive. It is maintained here for simplicity. Note however that the random parameter vector $\mathbf{c}_i \equiv (c_{k,i} : k \in \mathcal{K}^-)$ introduces some heterogeneity, admittedly in a limited amount, in the considered cost function model. These terms are random from the econometrician viewpoint but they are known to farmer *i*.

The fixed cost term A_i cannot be identified. Indeed, it cannot formally be distinguished from the production regime fixed costs. These regime fixed costs $b_{r,i}$ can only identified up to an additive constant (for farmer *i*). A convenient normalization constraint is given by $b_{0,i} \equiv 0$. Under this constraint the $b_{r,i}$ terms for $r \in \mathcal{R}^-$ where $r \in \mathcal{R}^- \equiv \{1,...,R\}$ are to be interpreted as differences in the regime costs with the cost of regime 0 as the reference. As above, the crop subset defining regime *r* is denoted by $\mathcal{K}_{(r)}$ for $r \in \mathcal{R}$ and with $\mathcal{K} \equiv \mathcal{K}_{(0)}$. The group subset $\mathcal{G}_{(r)}$, the sub-group subset $\mathcal{M}_{(r)}$, the subsets of sub-groups $\mathcal{M}_{(r),g}$ for $g \in \mathcal{G}_{(r)}$ and the crop sub-groups $\mathcal{K}_{(r),m}$ for $m \in \mathcal{M}_{(r)}$ are defined accordingly. These sets allow defining the acreage share choice of farmer *i* in regime *r* as:

(18) $s_{(r),k,i} = s^{\mathcal{K}|\mathcal{M}}_{(r),k|m,i}\, s^{\mathcal{M}|\mathcal{G}}_{(r),m|g,i}\, s^{\mathcal{G}}_{(r),g,i}$

where

(19) $s^{\mathcal{K}|\mathcal{M}}_{(r),k|m,i} \equiv \dfrac{\exp\left(\tau_m(\pi_{k,i} - c_{k,i})\right)}{\sum_{\ell \in \mathcal{K}_{(r),m}} \exp\left(\tau_m(\pi_{\ell,i} - c_{\ell,i})\right)}$,

(20) $s^{\mathcal{M}|\mathcal{G}}_{(r),m|g,i} \equiv \dfrac{\exp(\delta_g \Pi^{\mathcal{M}}_{m,i})}{\sum_{n \in \mathcal{M}_{(r),g}} \exp(\delta_g \Pi^{\mathcal{M}}_{n,i})}$ with $\Pi^{\mathcal{M}}_{(r),m,i} \equiv \tau_m^{-1} \ln \sum_{\ell \in \mathcal{K}_{(r),m}} \exp\left(\tau_m(\pi_{\ell,i} - c_{\ell,i})\right)$

and

(21) $s^{\mathcal{G}}_{(r),g,i} \equiv \dfrac{\exp(\alpha \Pi^{\mathcal{G}}_{g,i})}{\sum_{h \in \mathcal{G}_{(r),m}} \exp(\alpha \Pi^{\mathcal{G}}_{h,i})}$ with $\Pi^{\mathcal{G}}_{(r),g,i} \equiv \delta_g^{-1} \ln \sum_{n \in \mathcal{M}_{(r),g}} \exp(\delta_g \Pi^{\mathcal{M}}_{n,i})$.

for $k \in \mathcal{K}_{(r),m}$, $m \in \mathcal{M}_{(r),g}$ and $g \in \mathcal{G}_{(r)}$. This also allows determining the optimal profit level, regime fixed costs excluded, of farmer *i* in production regime *r* with:

(22) $\Pi_{(r),i} \equiv \max_{\mathbf{s}_{(r)} \in \mathcal{U}_{(r)}} \{\mathbf{s}'_{(r)} \boldsymbol{\pi}_{(r),i} - D_i(\mathbf{Q}'_{(r,+)} \mathbf{s}_{(r)})\} = \alpha^{-1} \ln \sum_{h \in \mathcal{G}_{(r),m}} \exp(\alpha \Pi^{\mathcal{G}}_{h,i})$.

All these terms but one are counterfactual for farmer *i*. They are counterfactual for $r \in \mathcal{R} \setminus \{r_i\}$ and $r \neq r_i$ where r_i is the production regime choice of farmer *i*. But these terms allow defining r_i with

(23) $r_i \equiv \arg\max_{r \in \mathcal{R}} \{\Pi_{(r),i} - b_{r,i}\}$

as well as the acreage choice of farmer *i* with

(24) $\mathbf{s}^+_i = \mathbf{s}_{(r),i} \equiv (s_{(r),k,i} : k \in \mathcal{K}_{(r)})$ and $\mathbf{s}_i = \mathbf{Q}_{(r)} \mathbf{s}_{(r),i}$ for $r = r_i$.

The multicrop micro-econometric model which is estimated in the next section differs from the one presented here due to data constraints. The joint distribution of the random parameter vector $\boldsymbol{\beta}^x_i \equiv (\beta^x_{k,i} : k \in \mathcal{K})$ of the expected crop gross margins cannot be identified with our data set, mainly because variable input expenditures are only observed at the farm level.[11] To

identify the joint distribution of $\boldsymbol{\beta}_i^x$ would require the specification of a variable input allocation equation (see, *e.g.*, Carpentier and Letort, 2012) as well as sufficient variations in the input price index w_i in the considered sample. This later condition is not met in our data set. Because the $\beta_{k,i}^x$ terms only appear in the acreage choice models together with w_i and $c_{k,i}$ terms in the considered multicrop crop models, the $c_{k,i}$ terms are assumed to stand for the $c_{k,i} + w_i \beta_{k,i}^x$ terms.

Distributional assumptions

Two tasks remain for investigating the statistical features of the considered multicrop econometric model. We need to set up tractable notations for describing the model to be estimated. And we need to define the probabilistic features of the model. The multicrop micro-econometric model described above is consistent in its deterministic parts. We need to define the assumptions related to the model random terms for this model to be also consistent in its random parts.

The random parameters $\boldsymbol{\beta}_i^y \equiv (\beta_{k,i}^y : k \in \mathcal{K})$ and $\mathbf{c}_i$ have specific roles in the considered multicrop model. The $\boldsymbol{\beta}_i^y$ term accounts for the fact that farmers have more information on their crop production process than the econometrician. The term $\beta_{k,i}^y$ is known to farmer *i* when he chooses his acreage. As a result, this term captures two kinds of effects: those of the natural factor endowment of farm *i* (*e.g.* soil quality and of standard climatic conditions) and those of skills of the farmer *i*. The $\mathbf{c}_i$ random term in the acreage management implicit cost function accounts for two kinds of effects: the heterogeneity in the quasi-fixed factor endowment, of the natural factor endowment of farm *i*, and of the human capital endowment

of farmer *i* on the one hand, and the effects of stochastic events affecting the acreage choice of farmer *i* (*e.g.* the effects of the climatic conditions which have occurred before the crop planting dates).

According to these interpretations, The random terms $\boldsymbol{\varepsilon}_i^y \equiv (\varepsilon_{k,i}^y : k \in \mathcal{K})$ and $\mathbf{c}_i$ are also closely related because they both capture the effect of year specific effects. These interpretations have three main implications. First, the $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\varepsilon}_i^y$ random terms can be assumed to be mutually independent. Second, the $\mathbf{c}_i$ and $\boldsymbol{\varepsilon}_i^y$ random terms can also be assumed to be mutually independent. The $\boldsymbol{\varepsilon}_i^y$ terms capture the effects of random events which occur after the realization of $\mathbf{c}_i$. These effects are difficult to forecast. Third, the $\boldsymbol{\beta}_i^y$ and $\mathbf{c}_i$ random parameters are closely related: they both capture the effects of the heterogeneity of the capital endowments across farms and farmers. As a result, the joint probability distribution of the $(\boldsymbol{\beta}_i^y, \boldsymbol{\varepsilon}_i^y, \mathbf{c}_i)$ term cannot be identified without further restrictions because the $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\varepsilon}_i^y$ appear as a sum in the considered model and because $\boldsymbol{\beta}_i^y$ and $\mathbf{c}_i$ are likely to be correlated.

In order to solve this identification problem the elements of the $\boldsymbol{\beta}_i^y$ term are modeled as functions of a latent productivity index denoted by e_i with

(25) $\beta_{k,i}^y = \beta_k^y + \mu_k^y e_i$ for $k \in \mathcal{K}$.

The elements of $\mathbf{c}_i$ terms are also modeled as functions of the latent productivity index e_i with additive error terms

(26) $c_{k,i} = c_k + \mu_k^s e_i - \varepsilon_{k,i}^s$ for $k \in \mathcal{K}^-$.

Of course, more flexible models could be used for $\boldsymbol{\beta}_i^y$ and $\mathbf{c}_i$ terms. The models presented in equations (25) and (26) are used for simplicity.

The regime fixed costs, *i.e.* the elements of $\mathbf{b}_i \equiv (b_{r,i} : r \in \mathcal{R}^-)$, are assumed to be simply defined as:

(27) $b_{r,i} \equiv \theta_r^\rho - \varepsilon_{r,i}^\rho$ for $r \in \mathcal{R}$.

The random term vector $\boldsymbol{\varepsilon}_i^\rho \equiv (\varepsilon_{r,i}^\rho : r \in \mathcal{R}^-)$ mainly accounts for the heterogeneity of regime fixed costs across farms, *i.e.* the $\boldsymbol{\varepsilon}_i^\rho$ terms are likely to vary much more across farms than they vary across years (at least in short time period).[12] Since regime 0 involves the entire crop set $\mathcal{K}$, the θ_k^ρ terms are expected to be negative for $r \in \mathcal{R}^-$.

The random terms $\mathbf{z}_i$, e_i, $\boldsymbol{\varepsilon}_i^y$, $\boldsymbol{\varepsilon}_i^s \equiv (\varepsilon_{k,i}^s : k \in \mathcal{K}^-)$ and $\boldsymbol{\varepsilon}_i^\rho$ are assumed to be mutually independent. In particular, $\mathbf{z}_i$ is assumed to be exogenous in the considered model. These independence assumptions allow identifying the probability distribution of the unobserved random terms e_i, $\boldsymbol{\varepsilon}_i^y$, $\boldsymbol{\varepsilon}_i^s$ and $\boldsymbol{\varepsilon}_i^\rho$. This assumption is admittedly restrictive as it constrains the functional form of the correlation between $\mathbf{c}_i$ and $\boldsymbol{\beta}_i^y$.

The crop subset defining the regime chosen by farmer *i*, *i.e.* r_i, is denoted by $\mathcal{K}_i^+$. The group subset $\mathcal{G}_i^+$, the sub-group subset $\mathcal{M}_i^+$, the subsets of sub-groups $\mathcal{M}_{g,i}^+$ for $g \in \mathcal{G}_i^+$ and the crop sub-groups $\mathcal{K}_{m,i}^+$ for $m \in \mathcal{M}_i^+$ are similarly defined. The term $\mathcal{K}_i^0$ denotes the subset of crops not produced by farmer *i*, *i.e.* $\mathcal{K}_i^0$ is the complement of $\mathcal{K}_i^+$ to $\mathcal{K}$. The following notations highlight the respective roles of the parameters and of the error terms.

The yield supply system corresponding to farmer *i* is given by:

(28) $y_{k,i} = g_k^y(\boldsymbol{\theta}_k^y, \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_{k,i}^y) = \beta_k^y + \mu_k^y e_i + \mathbf{v}'_{k,i}\boldsymbol{\eta}_k - 1/2 \times \gamma_k \times w_i^2 p_{k,i}^{-2} + \varepsilon_{k,i}^y$ for $k \in \mathcal{K}_i^+$

where $\boldsymbol{\theta}_k^y \equiv (\beta_k^y, \boldsymbol{\eta}_k, \mu_k^y, \gamma_k)$. This system depends on the parameter vector $\boldsymbol{\theta}^y \equiv (\boldsymbol{\theta}_k^y : k \in \mathcal{K})$. The functional form of the models of the yield levels $y_{k,i}$ doesn't depend on the regime r_i. Note that the yield levels of the non produced crops cannot be considered as corner solutions *per se*. A possible technical interpretation is that their observation is censored by the regime choice mechanism which determines the corner solutions, if any, of the acreage choice. Note also that expectation of $\mathbf{y}_i^0 \equiv (y_{k,i} : k \in \mathcal{K}_i^0)$ conditional on $g_k^s(\mathbf{z}_i, e_i)$ is a major determinant of the regime choice of farmer *i*.

The acreage share system corresponding to farmer *i* is given by:

(29) $s_{k,i} = g_k^s(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s; r_i)$ for $k \in \mathcal{K}_i^+$ and $s_{k,i} = 0$ for $k \in \mathcal{K}_i^0$

where $\boldsymbol{\theta}^s \equiv \left(\left((c_k, \mu_k^s) : k \in \mathcal{K}\right), \alpha, \boldsymbol{\delta}, \boldsymbol{\tau}\right)$ and $s_{k,i} = g_k^s(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s; r_i)$. This system depends on the parameter vectors $\boldsymbol{\theta}^y$ and $\boldsymbol{\theta}^s$, and on the production regime r_i. The acreage share of crop *k* chosen by farmer *i*, *i.e.* $s_{k,i}$, can now be written as

(30) $$s_{k,i} = s_{k|m,i}^{\mathcal{K}|\mathcal{M}} \, s_{m|g,i}^{\mathcal{M}|\mathcal{G}} \, s_{g,i}^{\mathcal{G}} = \frac{\exp\left(\tau_m(\pi_{k,i} - c_{k,i})\right)}{\sum_{\ell \in \mathcal{K}_{m,i}^+} \exp\left(\tau_m(\pi_{\ell,i} - c_{\ell,i})\right)} \frac{\exp(\delta_g \Pi_{m,i}^{\mathcal{M}})}{\sum_{n \in \mathcal{M}_{g,i}^+} \exp(\delta_g \Pi_{n,i}^{\mathcal{M}})} \frac{\exp(\alpha \Pi_{g,i}^{\mathcal{G}})}{\sum_{h \in \mathcal{G}_i^+} \exp(\alpha \Pi_{h,i}^{\mathcal{G}})}$$

where

(31) $$\pi_{k,i} - c_{k,i} = p_{k,i}(\beta_k^y + \mu_k^y e_i) + 1/2 \times \gamma_k w_i^2 p_{k,i}^{-1} - c_k - \mu_k^s e_i + \varepsilon_{k,i}^s$$

and

(32) $$\Pi_{m,i}^{\mathcal{M}} \equiv \tau_m^{-1} \ln \exp\left(\tau_m(\pi_{\ell,i} - c_{\ell,i})\right) \text{ and } \Pi_{g,i}^{\mathcal{G}} \equiv \delta_g^{-1} \ln \sum_{n \in \mathcal{M}_{g,i}^+} \exp(\delta_g \Pi_{n,i}^{\mathcal{M}})$$

for $k \in \mathcal{K}_{m,i}^+$, $m \in \mathcal{M}_{g,i}^+$ and $g \in \mathcal{G}_i^+$. Hopefully, the acreage choice model need not be explicitly written such as in equations (30)–(32) in the estimation criterion.

The production regime choice of farmer *i* is defined as

(33) $r_i = g_\rho(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\theta}^\rho, \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s, \boldsymbol{\varepsilon}_i^\rho) \equiv \arg\max_{r \in \mathcal{R}} \{\Pi_{(r),i} - \theta_r^\rho + \varepsilon_{r,i}^\rho\}$

where $\Pi_{(r),i} = \alpha^{-1} \ln \sum_{h \in \mathcal{G}_{(r),m}} \exp(\alpha \omega_{h,i}^{\mathcal{G}})$ and for $\boldsymbol{\theta}^\rho \equiv (\theta_k^\rho : r \in \mathcal{R}^-)$.

It is easily shown that the considered multicrop micro-econometric model is an endogenous regime switching model. First, the functional form of the acreage share choice model depends on the considered production regime (see equations (30)–(32)). Second, the regime choice is endogenous with respect to the acreage share choice model because the optimal regime depends on the random terms of the acreage share choice model, *i.e.* r_i depends on $(e_i, \boldsymbol{\varepsilon}_i^s)$. The yield levels are only censored according to the production regime. The term $y_{k,i}$ is observed if and only if crop k belongs to regime r_i but its functional form, $y_{k,i} = g_k^y(\boldsymbol{\theta}_k^y, \mathbf{z}_i, e_i, \varepsilon_{k,i}^y)$, doesn't depend on the production regime r_i.

In is also interesting to note that the yield vector $\mathbf{y}_i$ and the regime choice r_i are independent conditionally on $(\mathbf{z}_i, e_i^y)$. This implies that the unobserved yield levels, *i.e.* $\mathbf{y}_i^0 \equiv \mathbf{Q}_{(r_i,0)} \mathbf{y}_i$ where the matrix $\mathbf{Q}_{(r,0)}$ is obtained from the dimension K identity matrix by deleting its rows corresponding to the crops contained in $\mathcal{K}_{(r)}$, at missing at random conditionally on $(\mathbf{z}_i, e_i^y)$. The observed acreage choice $\mathbf{s}_i^+$ and the production regime r_i are not independent conditionally on $(\mathbf{z}_i, \mathbf{e}_i)$ because both choices depend on $\boldsymbol{\varepsilon}_i^{s,+} \equiv \mathbf{Q}_{(r_i,+)}^- \boldsymbol{\varepsilon}_i^s$, the error term of $\mathbf{s}_i^+$. The selection matrix $\mathbf{Q}_{(r_i,+)}^-$ is obtained from $\mathbf{Q}_{(r_i,+)}$ by deleting its first row and column because $\boldsymbol{\varepsilon}_i^s \equiv (\varepsilon_{k,i}^s : k \in \mathcal{K} \setminus \{1\})$ and $\boldsymbol{\varepsilon}_i^{s,+} \equiv (\varepsilon_{k,i}^s : k \in \mathcal{K}_i^+ \setminus \{1\})$. For the same reasons and because $\boldsymbol{\varepsilon}_i^{s,0} \equiv (\varepsilon_{k,i}^s : k \in \mathcal{K}_i^0)$, $\mathbf{Q}_{(r_i,0)}^-$ is obtained from $\mathbf{Q}_{(r_i,0)}$ by deleting its first row column.

Additional assumptions related to the probability distribution of the random terms of the considered multicrop micro-econometric model are required for completing its empirical specification. Because the estimation of endogenous regime switching models is particularly tedious, we choose to define a fully parametric model and to rely on rather restrictive independence assumptions as well as on convenient parametric probability distribution functions.

The e_i^y, e_i^s, $\boldsymbol{\varepsilon}_i^y$ and $\boldsymbol{\varepsilon}_i^s$ are assumed to be normally distributed with null means, *i.e.*

(34) $e_i \sim \mathcal{N}(0,1)$, $\boldsymbol{\varepsilon}_i^y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}^y)$ and $\boldsymbol{\varepsilon}_i^s \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}^s)$.

The elements of $\boldsymbol{\varepsilon}_i^\rho$ are assumed to be mutually independent and to be distributed according to a centered Gumbel distribution with $(\psi^\rho)^{-1}$ as the scale parameter. This assumption implies that the regime choice of farmer *i* is modeled, conditionally on $(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s)$, as a Standard MNL discrete choice.

As a summary we provide the main features of the functional forms of the dependent variables of the multicrop model – *i.e.* $\mathbf{y}_i^+$, $\mathbf{s}_i^+$ and r_i – and their corresponding conditional log-likelihood functions to be used for statistical inference purposes. In what follows the term $f(\mathbf{x}_i \mid \mathbf{q}_i; \mathbf{h})$ generically denotes the probability density function of $\mathbf{x}_i$ conditional on $\mathbf{q}_i$, this function being parameterized by **h**, and the term $\varphi(\mathbf{x}; \boldsymbol{\Xi})$ denotes the probability distribution function of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Xi})$ at **x**.

It easily shown that the model of $\mathbf{y}_i^+$ can be written as:

(35) $\mathbf{y}_i^+ = \mathbf{Z}_i^+(\mathbf{z}_i, e_i)'\boldsymbol{\theta}^y + \boldsymbol{\varepsilon}_i^{y,+}$ where $\boldsymbol{\varepsilon}_i^{y,+} \mid (\mathbf{z}_i, e_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_{++,i}^y)$ and $\boldsymbol{\Psi}_{++,i}^y \equiv \mathbf{Q}_{(r_i,+)} \boldsymbol{\Psi}^y \mathbf{Q}'_{(r_i,+)}$

The matrix $\mathbf{Z}_i^+(\mathbf{z}_i, e_i)$ is easily designed as a block diagonal matrix so as equation (28) to hold. As result, the probability density function of $\mathbf{y}_i^+$ conditional on $(\mathbf{z}_i, e_i)$ is given by

(36) $f(\mathbf{y}_i^+ \mid \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\Psi}^y_{++,i}) = \varphi\left(\boldsymbol{\varepsilon}_i^{y,+}(\boldsymbol{\theta}^y); \boldsymbol{\Psi}^y_{++,i}\right)$ where $\boldsymbol{\varepsilon}_i^{y,+}(\boldsymbol{\theta}^y) \equiv \mathbf{y}_i^{y,+} - \mathbf{Z}_i^+(\mathbf{z}_i, e_i)'\boldsymbol{\theta}^y$

It is also easily shown that

(37) $f(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0} \mid \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\Psi}^y) = \varphi\left(\mathbf{Q}'_{(r_i,+)}\boldsymbol{\varepsilon}_i^{y,+}(\boldsymbol{\theta}^y) + \mathbf{Q}'_{(r_i,+)}\boldsymbol{\varepsilon}_i^{y,0}; \boldsymbol{\Psi}^y\right)$.

The joint probability density function of $f(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0})$ conditional on $(\mathbf{z}_i, e_i)$ proves to be useful for implementing our estimation procedure.

According to equations (30)–(32) the model of $\mathbf{s}_i^+$ can formally be defined as a function of $(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+})$ parameterized by $(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s)$ and whose functional form depends on r_i. Let denote this functional form by . We have:

(38) $\mathbf{s}_i^+ = \mathbf{g}^+(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s; r_i)$ with $\boldsymbol{\varepsilon}_i^{s,+} \mid (\mathbf{z}_i, e_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}^s_{++,i})$ and $\boldsymbol{\Psi}^s_{++,i} \equiv \mathbf{Q}^-_{(r_i,+)}\boldsymbol{\Psi}^s(\mathbf{Q}^-_{(r_i,+)})'$.

The residual term corresponding to observation *i* at $(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s)$ is the solution in $\boldsymbol{\varepsilon}^{s,+}$ to the equation $\mathbf{g}^+(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}^{s,+}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s; r_i) = \mathbf{s}_i^+$. This residual term, denoted as $\boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s)$, can be obtained by using Berry's (1994) which, in our case, leads to:

(39)
$$\begin{aligned}\varepsilon^s_{k,i} = &-\left(p_{k,i}(\beta^y_k + \mu^y_k e_i) + 1/2 \times \gamma_k w_i^2 p_{k,i}^{-1} - c_k - \mu^s_k e_i\right) + \left(p_{1,i}(\beta^y_1 + \mu^y_1 e_i) + 1/2 \times \gamma_1 w_i^2 p_{1,i}^{-1}\right) \\ &+ \left(\tau_m^{-1} \ln s^{\mathcal{K}|\mathcal{M}}_{k|m,i} + \delta_g^{-1} \ln s^{\mathcal{M}|\mathcal{G}}_{m|g,i} + \alpha^{-1} \ln s^{\mathcal{G}}_{g,i}\right) - \left(\tau_1^{-1} \ln s^{\mathcal{K}|\mathcal{M}}_{1|1,i} - \delta_1^{-1} \ln s^{\mathcal{M}|\mathcal{G}}_{1|1,i} - \alpha^{-1} \ln s^{\mathcal{G}}_{1,i}\right)\end{aligned}$$

for $k \in \mathcal{K}_i^+$ such that $k \in \mathcal{K}^+_{m,i}$, $m \in \mathcal{M}^+_{g,i}$ and $g \in \mathcal{G}_i^+$. It is assumed that crop 1 is produced in every observation with $1 \in \mathcal{K}^+_{1,i}$ and $1 \in \mathcal{M}^+_{1,i}$. Let the terms G_i^+, $M^+_{g,i}$ and $K^+_{m,i}$ denote the cardinality of the sets $\mathcal{G}_i^+$, $\mathcal{M}^+_{g,i}$ and $\mathcal{K}^+_{m,i}$.The probability distribution function of $\mathbf{s}_i^+$ conditional on $(\mathbf{z}_i, e_i)$ is given by:

(40) $f(\mathbf{s}_i^+ \mid \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s_{++,i}) = J_i^+(\boldsymbol{\theta}^s) \times \varphi\left(\boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s); \boldsymbol{\Psi}^s_{++,i}\right)$

where

(41) $J_i^+(\boldsymbol{\theta}^s) \equiv \alpha^{1-G_i^+} \prod_{m \in \mathcal{M}^+_{g,i}} \delta_g^{1-M^+_{g,i}} \prod_{k \in \mathcal{K}^+_{m,i}} \tau_m^{1-K^+_{m,i}} \prod_{k \in \mathcal{K}_i^+} s_{k,i}^{-1}$.

The joint probability density function of $f(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$ conditional on $(\mathbf{z}_i, e_i)$

(42) $$f(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0} \mid \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) = J_i^+(\boldsymbol{\theta}^s) \times \varphi\left(\mathbf{Q}'_{(r_i,+)} \boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s) + \mathbf{Q}'_{(r_i,+)} \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}^s\right).$$

also proves to be useful for implementing our estimation procedure.[13]

Finally, it is easily shown that the regime profit levels $\Pi_{(r),i}$ for $r \in \mathcal{R}$ are functions of $(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s)$ parameterized by $(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s)$ (See equations (22), (31) and (32)). Let denote these functions by $\Pi_{(r),i}(\boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s)$ for highlighting their dependence on $\boldsymbol{\varepsilon}_i^{s,+}$ on the one hand and on $(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s)$ on the other hand. The probability of farmer *i* choosing regime *r* is given by:

(43) $$P[r_i = r \mid \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s] = \frac{\exp\left(\psi^\rho \times (\Pi_{(r),i}(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s) - \theta_r^\rho)\right)}{\sum_{j \in \mathcal{R}} \exp\left(\psi^\rho \times (\Pi_{(j),i}(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s) - \theta_j^\rho)\right)}$$

implying that the probability function of r_i conditional on $(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^s)$ is given by:

(44) $$f(r_i \mid \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}, \psi^\rho) = \frac{\exp\left(\psi^\rho \times (\Pi_{(r_i),i}(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s) - \theta_{r_i}^\rho)\right)}{\sum_{j \in \mathcal{R}} \exp\left(\psi^\rho \times (\Pi_{(j),i}(\mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,+}, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s) - \theta_j^\rho)\right)}.$$

The scaling parameter $\psi^\rho > 0$ allows accounting for the relative weights of the regime profit levels on the hand and of the $\boldsymbol{\varepsilon}_i^\rho$ random terms on the other hand.

Estimation issues

We assume that the $(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i, \mathbf{z}_i)$ terms are independently and identically distributed for $i = 1, ..., N$ where *N* is the considered sample size. The fully parametric structure of the statistical model of $(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i)$ conditional on $\mathbf{z}_i$ suggests the Maximum Likelihood (ML) inference framework for estimating its parameter vector denoted here as $\mathbf{a} \equiv \left(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\theta}^\rho, \mathbf{vech}(\boldsymbol{\Psi}^y), \mathbf{vech}(\boldsymbol{\Psi}^s), \psi^\rho\right)$. However, this model raises serious estimation issues.

The statistical model of $(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i)$ conditional on $\mathbf{z}_i$ significantly depends on unobserved variables, *i.e.* on the latent variable vector e_i (through the models of $\mathbf{y}_i^+$, of $\mathbf{s}_i^+$ and of r_i) and on the error terms of the acreage equations of the non-produced crops $\boldsymbol{\varepsilon}_i^{s,0} \equiv \mathbf{Q}_{(r_i,0)}^{-} \boldsymbol{\varepsilon}_i^s$ (through the model of r_i). This implies that the observations log-likelihood functions involve multiple integrals. Let the term $f(\mathbf{x}_i \mid \mathbf{q}_i; \mathbf{h})$ generically denotes the probability density function of $\mathbf{x}_i$ conditional on $\mathbf{q}_i$, this function being parameterized by **h**. The log-likelihood at **a** of an observation *i* is given by:

$$(45) \qquad \ln \ell_i(\mathbf{a}) \equiv \ln \int f(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i \mid \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,0}; \mathbf{a}) \, f(e_i, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^s) d(e_i, \boldsymbol{\varepsilon}_i^{s,0}) .$$

where the term $\boldsymbol{\Psi}_{00,i}^s \equiv \mathbf{Q}_{(r_i,0)}^{-} \boldsymbol{\Psi}^s (\mathbf{Q}_{(r_i,0)}^{-})$ denotes the variance matrix of $\boldsymbol{\varepsilon}_i^{s,0}$. *I.e.* $\boldsymbol{\Psi}_{00,i}^s$ is the part of variance matrix of $\boldsymbol{\varepsilon}_i^s$, $\boldsymbol{\Psi}^s$, corresponding to the element of $\boldsymbol{\varepsilon}_i^s$ missing in observation *i* (because the corresponding crops are not produced by farmer *i*). This term depends on *i* because it depend on the production regime of *i*. The integral involved in equation (45) cannot be computed, neither analytically, nor numerically. In the econometrics literature, such a problem is often dealt with by relying on simulation methods. *I.e.*, the log-likelihood functions $\ln \ell_i(\mathbf{a})$ are integrated by simulation methods for obtaining simulated approximates of $\ln \ell_i(\mathbf{a})$, these simulated log-likelihood functions being used for defining Simulated ML (SML) estimators of **a**. A SML estimator of **a** basically have the properties of the (infeasible) ML estimator of a:

$$(46) \qquad \hat{\mathbf{a}}_N^{ML} \equiv \arg\max_{\mathbf{a}} \sum_{i=1}^{N} \ln \int f(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i \mid \mathbf{z}_i, e_i, \boldsymbol{\varepsilon}_i^{s,0}; \mathbf{a}) \, f(e_i, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^s) d(e_i, \boldsymbol{\varepsilon}_i^{s,0})$$

provided that a sufficiently large number of random draws of $(e_i, \boldsymbol{\varepsilon}_i^{s,0})$ are used for approximating the log-likelihood functions $\ln \ell_i(\mathbf{a})$. We also use simulation methods for solving integration problems similar to that involved in equation (46). But we do not use the SML inference framework due to a second issue, or set of issues, to be addressed.

The log-likelihood functions $\ln \ell_i(\mathbf{a})$ are involved functions of **a** for three main reasons. (*i*) The functional forms of the acreage choice and production regime probability choice are not conventional. They unusually nonlinear in **a**. (*ii*) The functional form of the log-likelihood functions $\ln \ell_i(\mathbf{a})$ depends on the production regime of *i*. This makes the sample simulated log-likelihood function particularly awkward. (*iii*) To maximize in **a** the sample simulated log-likelihood function is a difficult task due to the problems described above and due to the dimension of **a**. In particular, this maximization problem cannot be split into a sequence of simpler optimization problems.

In such a context statisticians usually prefer to rely on Stochastic Expectation-Maximization (SEM) algorithms to compute estimators of **a** which basically have the same asymptotic properties of corresponding SML estimators (Jank and Booth, 2003; McLachlan and Krishnan, 2008). The Expectation-Maximization (EM) algorithms were proposed by Dempster *et al* (1977) for computing ML estimators in specific cases. Such algorithms basically replace an involved ML problem by an iterative scheme involving a sequence composed of an Expectation (E) step and of a Maximization step. Deterministic EM algorithms ensure to find a maximum of the considered log-likelihood function as they monotonically increase this function at each iteration. They rapidly converge to the neighborhood of a solution to the ML problem. But they are known to slowly converge to this solution within its neighborhood. EM algorithms proved to be particularly useful when the statistical model of interest involves hidden, *e.g*. missing or latent, variables (McLachlan and Krishnan, 2008). They take advantage of the specific structure of the log-likelihood function of such models.

SEM algorithms extended EM algorithms to cases where the E step cannot be performed neither analytically, nor numerically. The so-called Stochastic E steps integrate the

expectations of the EM algorithms with simulation methods. Numerous SEM algorithms were developed in the statistics literature (see, *e.g.*, McLachlan and Krishnan, 2008). They may not monotonically increase the considered (simulated) log-likelihood function. But they are expected to do so when large numbers of random draws are used for performing their E steps.

Our estimator of **a** is obtained by designing a SEM algorithm specifically adapted to our multicrop model. Other SEM algorithms could be used and might be more efficient from a numerical viewpoint. But the one we use is designed so as to be relatively easy to code and so as to only involve simple arithmetic operations. It is designed with the general framework proposed by Delyon *et al* (1999), *i.e.* it is a Stochastic Approximate EM algorithm (SAEM), and it uses the conditional maximization approach proposed by Meng and Rubin (1993) for designing the so-called Expectation Conditional Maximization (ECM) algorithms. SAEM algorithms are numerically stable when compared to other SEM. The ECM algorithms allow replacing involved M steps by a sequence of simpler Conditional Maximization steps.

The rest of this section briefly presents the SEM algorithm used for computing the estimates presented in the next section. This algorithm is presented in further details in the Technical Appendix.

We consider $(\boldsymbol{\kappa}_i, \mathbf{u}_i)$ as the complete variable vector of our SEM algorithm provided that $\boldsymbol{\kappa}_i \equiv (\mathbf{y}_i^+, \mathbf{s}_i^+, r_i)$ is our observed endogenous variable vector, that $\mathbf{u}_i \equiv (\boldsymbol{\varepsilon}_i^{y,0}, \boldsymbol{\varepsilon}_i^{s,0}, e_i)$ is our unobserved variable vector and that $\mathbf{z}_i$ is the observed exogenous variable upon which our statistical inference is conditioned. The $\boldsymbol{\varepsilon}_i^{y,0} \equiv \mathbf{Q}_{(r_i,0)} \boldsymbol{\varepsilon}_i^{y}$ term contains the yield error terms of the crops not produced by farmer *i*.

The (S)E step of our algorithm consists in computing the expectation of the complete variable log-likelihood function, $\ln \ell_i^C(\mathbf{a}) \equiv \ln f(\boldsymbol{\kappa}_i, \mathbf{u}_i \mid \mathbf{z}_i; \mathbf{a})$, of the complete observation $(\boldsymbol{\kappa}_i, \mathbf{u}_i)$ for $i = 1, ..., N$ conditional on the observed variable vector $(\boldsymbol{\kappa}_i, \mathbf{z}_i)$. This expectation is integrated over the distribution of $\mathbf{u}_i$ conditionally on $\boldsymbol{\kappa}_i$ as it is described by the last parameter update, *i.e.* $\mathbf{a}_n$ at iteration $n+1$ of the algorithm. This E step thus consists in computing:

$$(47) \qquad E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i] = \int \ln f(\boldsymbol{\kappa}_i, \mathbf{u}_i \mid \mathbf{z}_i; \mathbf{a}) f(\mathbf{u}_i \mid \mathbf{z}_i, \boldsymbol{\kappa}_i; \mathbf{a}_n) d\mathbf{u}_i$$

for $i = 1, ..., N$. The distributional assumptions underlying the considered model imply that $f(\mathbf{u}_i \mid \mathbf{z}_i, \boldsymbol{\kappa}_i; \mathbf{a}_n) = f(\boldsymbol{\varepsilon}_i^{y,0} \mid \mathbf{z}_i, \mathbf{y}_i^+, e_i; \mathbf{a}_n^y) f(\boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i, \boldsymbol{\kappa}_i; \mathbf{a}_n)$ where $\mathbf{a}_n^y \equiv \left(\boldsymbol{\theta}_n^y, \mathbf{vech}(\boldsymbol{\Psi}_n^y)\right)$ collects the parameters of the statistical model of yield supply system. *I.e.* the random terms $\boldsymbol{\varepsilon}_i^{y,0}$ and $(\boldsymbol{\varepsilon}_i^{s,0}, e_i)$ are independent conditionally on $(\mathbf{z}_i, \mathbf{y}_i^+, e_i)$. This comes from the fact that the error term vector of yield supply equation system $\boldsymbol{\varepsilon}_i^y$ is independent from the other elements of the model. This E step relies on simulation methods for the integration over the probability distribution of $(\boldsymbol{\varepsilon}_i^{s,0}, e_i)$. The integration over the probability distribution of $\boldsymbol{\varepsilon}_i^{y,0}$ is performed analytically along the lines of Ruud (1991). In fact, we have:

$$(48) \qquad E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i] = \int \ln f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i; \mathbf{a}) f(\boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i, \boldsymbol{\kappa}_i; \mathbf{a}_n) d(\boldsymbol{\varepsilon}_i^{s,0}, e_i)$$

where:

$$(49) \qquad \ln f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i; \mathbf{a}) \equiv \int \ln f(\boldsymbol{\kappa}_i, \mathbf{u}_i \mid \mathbf{z}_i; \mathbf{a}) f(\boldsymbol{\varepsilon}_i^{y,0} \mid \mathbf{z}_i, \mathbf{y}_i^+, e_i; \mathbf{a}_n^y) d\boldsymbol{\varepsilon}_i^{y,0} .$$

Our simulation method for approximating $E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i]$ relies on the following equality:

$$(50) \qquad f(\boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i, \boldsymbol{\kappa}_i; \mathbf{a}) = \varpi_i(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a}) f(\boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^s) f(e_i)$$

where:

$$(51) \qquad \varpi_i(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a}) \equiv \frac{f(\boldsymbol{\kappa}_i \mid \mathbf{z}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a})}{\int f(\boldsymbol{\kappa}_i \mid \mathbf{z}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a}) f(\boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^+) f(e_i) d(\boldsymbol{\varepsilon}_i^{s,0}, e_i)} .^{14}$$

This equality implies that:

(52) $E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i] = \int \ln f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i \mid \mathbf{z}_i; \mathbf{a}) \varpi_i(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a}_n) f(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \boldsymbol{\Psi}_{00,i,n}^s) d(\boldsymbol{\varepsilon}_i^{s,0}, e_i)$.

Provided that $\boldsymbol{\varepsilon}_i^{s,0}$ and e_i are independent and normally distributed, equations () and () suggest a simple simulator for $E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i]$. It suffices to use draws from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_{00,i,n}^s)$, the $\tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0}$ terms, and random draws from $\mathcal{N}(0,1)$, the $\tilde{e}_{i,nq}$ terms for $q = 1, ..., Q_n$. The simulator

(53) $\tilde{E}_{\mathbf{a}_n, Q_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i] \equiv Q_n^{-1} \sum_{q=1}^{Q_n} \tilde{\varpi}_{i,nq} \ln f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i, \tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0}, \tilde{e}_{i,nq} \mid \mathbf{z}_i; \mathbf{a})$

where

(54) $\tilde{\varpi}_{i,nq} \equiv \dfrac{f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i \mid \mathbf{z}_i, \tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0}, \tilde{e}_{i,nq}; \mathbf{a})}{\sum_{q=1}^{Q_n} f_{\mathbf{a}_n^y}(\boldsymbol{\kappa}_i \mid \mathbf{z}_i, \tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0}, \tilde{e}_{i,nq}; \mathbf{a})}$

converges in Q_n to $E_{\mathbf{a}_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i]$. This simulator is used by, *e.g.* Caffo *et al* (2005) and Train (2007, 2008). It uses an Importance Sampling scheme with $f(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \boldsymbol{\Psi}_{00,i,n}^s)$ as the proposal probability density function.

The corresponding M step consists in maximizing or in increasing in **a** the resulting conditional expectation of the sample (simulated) log-likelihood function:

(55) $\tilde{L}_{n,Q_n}(\mathbf{a}) \equiv \sum_{i=1}^N \tilde{E}_{\mathbf{a}_n, Q_n}[\ln \ell_i^C(\mathbf{a}) \mid \boldsymbol{\kappa}_i]$.

The update of **a** computed at iteration $n+1$ is preferably defined as $\mathbf{a}_{n+1} \equiv \arg\max_{\mathbf{a}} \tilde{L}_{n,Q_n}(\mathbf{a})$. But when the maximization problem is too demanding from a practical view point, the M step can be simplified into a Generalized M step (see, *e.g.*, Dempster *et al.*, 1977 ; Wu, 1983) consisting in finding $\mathbf{a}_{n+1}$ such that $\tilde{L}_{n,Q_n}(\mathbf{a}_{n+1}) > \tilde{L}_{n,Q_n}(\mathbf{a}_n)$, if possible. The SEM algorithms numerically converges when $\mathbf{a}_{n+1}$ is judged to be sufficiently close to $\mathbf{a}_n$ and/or when $\tilde{L}_{n,Q_n}(\mathbf{a}_{n+1})$ is judged to be sufficiently close to $\tilde{L}_{n,Q_n}(\mathbf{a}_n)$.[15]

The Conditional M steps proposed by Meng and Rubin (1993) are examples of Generalized M steps. Conditional M steps exploit the structure of the log-likelihood functions $\tilde{E}_{\mathbf{a}_n,Q_n}[\ln \ell_i^C(\mathbf{a})\,|\,\boldsymbol{\kappa}_i]$ for gradually updating the value of **a**. In our case, the complete variable vector log-likelihood function can be decomposed as follows:

$$
\begin{aligned}
\ln \ell_i^C(\mathbf{a}) &\equiv \ln f(\mathbf{y}_i^+, \mathbf{s}_i^+, r_i, e_i, \boldsymbol{\varepsilon}_i^{y,0}, \boldsymbol{\varepsilon}_i^{s,0} \,|\, \mathbf{z}_i; \mathbf{a}) \\
&= \ln f(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0} \,|\, \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\Psi}^y) + \ln f(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0} \,|\, \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s), \\
&\quad + \ln f(r_i \,|\, \mathbf{z}_i, e_i, \mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\theta}^\rho, \psi^\rho) + \ln f(e_i)
\end{aligned} \tag{56}
$$

This decomposition of $\ln \ell_i^C(\mathbf{a})$ uses Bayes' rule, exploits the independence assumptions related to the error terms and the latent variable of our model and the functional forms of the multicrop model. The probability density function of r_i is computed conditionally on $(\mathbf{z}_i, e_i, \mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0}, \mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$, that of $(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0})$ is computed conditionally on $(\mathbf{z}_i, e_i, \mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$ and finally that of $(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$ is computed conditionally on $(\mathbf{z}_i, e_i)$. The distributional assumptions of our model then allow for simplifications in the conditioning sets. These assumptions imply that r_i and $(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0})$ are independent conditionally on $(\mathbf{z}_i, e_i, \mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$, that $(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0})$ and that $(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0})$ are independent conditionally on $(\mathbf{z}_i, e_i)$ and the probability distribution of e_i is fixed.

Equation (56) highlights two main facts. First, this decomposition allows rewriting $\tilde{L}_{n,Q_n}(\mathbf{a}_{n+1})$ as:

$$
\begin{aligned}
\tilde{L}_{n,Q_n}(\mathbf{a}_{n+1}) &= \tilde{L}^y_{n,Q_n}(\boldsymbol{\theta}^y, \boldsymbol{\Psi}^y) + \tilde{L}^s_{n,Q_n}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) + \tilde{L}^r_{n,Q_n}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\theta}^\rho, \psi^\rho) \\
&\quad + \sum_{q=1}^{Q_n} \tilde{\varpi}_{i,nq} \ln f(\tilde{e}_{i,nq})
\end{aligned} \tag{57}
$$

where:

$$
\tilde{L}^y_{n,Q_n}(\boldsymbol{\theta}^y, \boldsymbol{\Psi}^y) \equiv \sum_{i=1}^{N} Q_n^{-1} \sum_{q=1}^{Q_n} \tilde{\varpi}_{i,nq} \left(\int \ln f(\mathbf{y}_i^+, \boldsymbol{\varepsilon}_i^{y,0} \,|\, \mathbf{z}_i, \tilde{e}_{i,nq}; \boldsymbol{\theta}_n^y, \boldsymbol{\Psi}^y) f(\boldsymbol{\varepsilon}_i^{y,0}; \boldsymbol{\Psi}^y_{00,i,n}) d\boldsymbol{\varepsilon}_i^{y,0} \right), \tag{58}
$$

$$
\tilde{L}^s_{n,Q_n}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) \equiv \sum_{i=1}^{N} Q_n^{-1} \sum_{q=1}^{Q_n} \tilde{\varpi}_{i,nq} \ln f(\mathbf{s}_i^+, \tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0} \,|\, \mathbf{z}_i, \tilde{e}_{i,nq}; \boldsymbol{\theta}_n^y, \boldsymbol{\theta}_n^s, \boldsymbol{\Psi}^s) \tag{59}
$$

and

(60) $$\tilde{L}^{r}_{n,Q_n}(\boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s},\boldsymbol{\theta}^{\rho},\psi^{\rho}) \equiv \sum_{i=1}^{N} Q_n^{-1} \sum_{q=1}^{Q_n} \varpi_{i,nq} \ln f(r_i \mid \mathbf{z}_i, \tilde{e}_{i,nq}, \boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s}), \tilde{\boldsymbol{\varepsilon}}_{i,nq}^{s,0}; \boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s},\boldsymbol{\theta}^{\rho},\psi^{\rho}).$$

This equation takes for granted that the observed acreage model is a function of $\boldsymbol{\varepsilon}_i^{s,+}$, *i.e.* $\mathbf{s}_i^{+} = \mathbf{g}^{s,+}(\boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s},\mathbf{z}_i,e_i,\boldsymbol{\varepsilon}_i^{s,+};r_i)$. This implies that the conditioning sets $(\mathbf{z}_i,e_i,\mathbf{s}_i^{+},\boldsymbol{\varepsilon}_i^{s,0})$ and $(\mathbf{z}_i,e_i,\boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s}),\boldsymbol{\varepsilon}_i^{s,0})$ are equivalent for the log-likelihood function of r_i at **a**. Note also that the last term of the sum of the right hand side term of equation () doesn't depend on **a**. It need not be computed because it is not involved in the M step of the SEM algorithm. Equation () illustrates the main interest of EM algorithms. They allow considering a sum of simple log-likelihood functions instead of a single involved log-likelihood function. Second, the probability distribution function of the "yield variables" $(\mathbf{y}_i^{+},\boldsymbol{\varepsilon}_i^{y,0})$ only depends on the parameter sub-vector $(\boldsymbol{\theta}^{y},\boldsymbol{\Psi}^{y})$ and the probability distribution function of the "acreage choice variables" $(\mathbf{s}_i^{+},\boldsymbol{\varepsilon}_i^{s,0})$ only depends on the parameter sub-vector $(\boldsymbol{\theta}^{y},\boldsymbol{\theta}^{s},\boldsymbol{\Psi}^{s})$. This allows for updating the value of the estimate of **a** according to following sequence of CM steps:

(61) $$\boldsymbol{\Psi}^{y}_{n+1} \equiv \arg\max_{\boldsymbol{\Psi}^{y}} \tilde{L}^{y}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\Psi}^{y})$$

(62) $$\boldsymbol{\Psi}^{s}_{n+1} \equiv \arg\max_{\boldsymbol{\Psi}^{s}} \tilde{L}^{s}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s}_{n},\boldsymbol{\Psi}^{s}),$$

(63) $$(\boldsymbol{\theta}^{\rho}_{n+1},\psi^{\rho}_{n+1}) \in \left\{(\boldsymbol{\theta}^{\rho},\psi^{\rho}) \mid \tilde{L}^{r}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s}_{n},\boldsymbol{\theta}^{\rho},\psi^{\rho}) > \tilde{L}^{r}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s}_{n},\boldsymbol{\theta}^{\rho}_{n},\psi^{\rho}_{n})\right\} \text{ if possible,}$$

(64) $$\boldsymbol{\theta}^{s}_{n+1} \in \left\{ \boldsymbol{\theta}^{s} \left| \begin{array}{c} \tilde{L}^{s}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s},\boldsymbol{\Psi}^{s}_{n+1}) + \tilde{L}^{r}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s},\boldsymbol{\theta}^{\rho}_{n+1},\psi^{\rho}_{n+1}) \\ > \\ \tilde{L}^{s}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s}_{n},\boldsymbol{\Psi}^{s}_{n+1}) + \tilde{L}^{r}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\theta}^{s}_{n},\boldsymbol{\theta}^{\rho}_{n+1},\psi^{\rho}_{n+1}) \end{array} \right. \right\} \text{ if possible}$$

and:

(65) $$\boldsymbol{\theta}^{y}_{n+1} \in \left\{ \boldsymbol{\theta}^{y} \left| \begin{array}{c} \tilde{L}_{n,Q_n}(\boldsymbol{\theta}^{y},\boldsymbol{\Psi}^{y}_{n+1}\boldsymbol{\theta}^{s}_{n+1},\boldsymbol{\Psi}^{s}_{n+1},\boldsymbol{\theta}^{\rho}_{n+1},\psi^{\rho}_{n+1}) \\ > \\ \tilde{L}_{n,Q_n}(\boldsymbol{\theta}^{y}_{n},\boldsymbol{\Psi}^{y}_{n+1}\boldsymbol{\theta}^{s}_{n+1},\boldsymbol{\Psi}^{s}_{n+1},\boldsymbol{\theta}^{\rho}_{n+1},\psi^{\rho}_{n+1}) \end{array} \right. \right\} \text{ if possible.}$$

This sequence ensures that $\tilde{L}_{n,Q_n}(\mathbf{a}_n) > \tilde{L}_{n,Q_n}(\mathbf{a}_{n+1})$ if possible, *i.e.* that this M step is an updating sequence of the elements of **a** meeting the requirements of a sequence of CM steps (Meng and Rubin, 1993 ; Delyon *et al*, 1999).

Estimation results

This last section is devoted to the presentation of the results of the first estimations that have been conducted using the endogenous regime switching model proposed in this paper. The purpose of these estimations is purely illustrative and intended to assess the empirical tractability of the model and its adequacy to observed data.

Data

We consider a dataset containing 1502 observations of French grain crop growers in the large Paris Basin over the years 1993 to 2007.[16] This dataset is obtained from the Farm Accountancy Data Network (FADN) and contains detailed information about crop production (acreages and yields). The crop prices are computed as regional crop price indices from the FADN. These indices equal 1 in 2005 in the "Ile de France" administrative *région*. The observed yield levels are computed accordingly. The variable input uses are observed at the farm level. These are not modeled for simplicity. The input price index w_i is computed as the Laspeyres price index of the aggregated variable input uses (pesticides, fertilizers and seeds). These computations are based on the variable input price indices provided by the French Department of Agriculture at the regional level. The climatic variable aggregates contained in $\mathbf{v}_i$ are defined as the first five factors obtained from a Partial Least Squares Regression of a

set of climatic variables obtained from *Météo France*, the French national meteorological service and defined as monthly averages (temperature, rainfall and sunshine) on yields.

We consider here five crops (or crop aggregates) – *i.e.*, soft wheat grain corn, other cereals (mainly barley), oilseeds (rapeseed and sunflower), protein crops (mainly peas) – which account for 80% of the arable land in the region. Figure 1 represents the three levels nesting structure that we adopt for these five crops. Corn and other cereals are nested in the first level; all cereals on the one hand, and oilseeds and protein crops on the other hand are nested in the second nest. This structure is intended to reflect the basic rotation scheme of grain producers in France, namely: oilseeds or protein crops – wheat – secondary cereal (*e.g.* corn, barley or wheat).

Based on these five crops, thirty-one different regimes could theoretically be chosen by farmers, only six of them are actually adopted in our sample. The reference regime ($r = 0$) in which all crops are grown is adopted by 25% of the farmers. Each of the other five regimes involves at least three crops and always includes wheat and oilseeds.

Illustrative application

The estimations were conducted by using the SAS software (IML procedure). The recursive step of simulation of SEM algorithm was implemented using 1000 draws. The algorithm converged after 481 iterations.

Selected parameter estimates are reported in Tables 1 to 3. The parameters representing yield levels as expected by the farmers at the time he takes his acreage decisions ($E[\beta^y_{k,i}] = \beta^y_k + \mu^y_k E[e_i]$) are reported in the first column of Table 1. They appear to be precisely estimate and lie in reasonable ranges. Notably, the μ^y_k parameters, which intend to

capture the impacts on yields of soil conditions heterogeneous among farmers, are precisely estimated and have expected signs. This is particularly true for the main crops in our sample: wheat, oilseeds and other cereals. The variance of the latent productivity index, e_i, is equal to 0.24 and is significant, which tends to reflect the presence of heterogeneity in cropping conditions between farms, even when controlling for climatic effects. On the other hand, most of the price parameters γ_k are significant but negative and the trend parameters, not reported here, are not significant. These unexpected results may reflect at least two identification issues. First, output prices exhibit time trends in the data which makes the distinction between price and time effects difficult. Second, the latent variable e_i probably captures part of the time trend in addition to the heterogeneity among farmers. Moving to panel data estimation should help solving, at least partly, these issues.

Table 2a reports the estimates of the curvature parameters of the acreage management cost function. All of them are significantly estimated and, most importantly, their estimated values satisfy a sufficient condition for the cost function to be convex: $\hat{\alpha} < \hat{\delta}_1 < \hat{\tau}_{11}$ and $\hat{\alpha} < \hat{\delta}_2$. The parameters associated to fixed costs in acreage equations are reported in Table 2b. Here some issues appear in the oilseeds and protein crops acreage estimates: their associated estimated fixed cost ($c_{k,i}$) are very large compared to the other crops and the variance of error terms in the oilseed acreage equation is ten times higher than those of the other acreage equations. These disturbing results might be related to the low estimated value of the parameter representing acreage adjustments costs for the oilseeds/protein crop nest (δ_2) compared to the cereals nest parameter (δ_1). This point requires a deeper analysis of the results.

Finally, the regime fixed costs (θ_r^ρ), reported in Table 3 are precisely estimated and have expected signs for the regimes the most represented in the database: they are negative which

implies that their associated fixed costs are lower than those associated to the reference regime where all crops are grown. These estimated fixed costs also range in reasonable values one compared to each other: regime 4, for instance contains one additional crop compared to regime 3 (protein crops) and its associated fixed cost is higher.

Conclusion

An endogenous regime switching approach in proposed in this paper to account for corner solutions in the modelling of farmers' acreage choices. One of the unique features of the proposed multicrop model is that it allows accounting for the fixed costs associated to each production regime available to the farmer at the time he takes his acreage and production decisions. We also show that this model can be estimated using a SEM algorithm specifically adapted to its structure and relatively simple to implement.

As illustrative purpose, a first set of estimations is run on a sample of French data. These first estimation results are encouraging in the sense that most of the key parameters of the model (*i.e.* parameters related to the flexibility of acreage adjustments between crops, the heterogeneity of cropping conditions and the regimes fixed costs) are significantly estimated and lie in ranges. The model thus seems to be in a relatively good adequacy with observed production choices. However, these are just preliminary results and more work still needs to be done. In a next step some fitting criteria will be computed to better assess of the adequacy of the model. Then, and more importantly, the identification issues faced in these first estimations, notably those involving time related effects, will have to be dealt with. A panel data approach seems a good alternative in that respect.

References

Ackerberg, D. A., and M. Rysman, 2005. Unobserved product differentiation in discrete choice models: estimating price elasticities and welfare effects, *RAND J. of Econ.* 36:771-788.

Arndt, C. S. Liu and P.V. Preckel, 1999. On Dual Approaches to Demand Systems Estimation in the Presence of Binding Quantity Constraints. *App. Econ.* 31:999-1008.

Berry, S.T. 1994. Estimating Discrete-Choice Models of Product Differentiation. *RAND J. of Econ.* 25:242-262.

Caffo, B. S., W. Jank and G.L. Jones, 2005. Ascent-based Monte Carlo Expectation–Maximization. *J. of Roy. Stat. Soc., Series B.* 67(2), 235–251.

Carpentier, A. and E. Letort, 2014. Multicrop models with MultiNomial Logit acreage shares. *Env. and Res. Econ.* 59:537-559.

Carpentier, A. and E. Letort, 2012. Accounting for heterogeneity in multicrop micro-econometric models: Implications for variable input demand modelling. *Am. J. of Ag. Econ.* 94(1): 209–224.

Chakir, R., A. Bousquet and N. Ladoux, 2004. Modeling Corner Solutions with Panel Data: Application to the Industrial Energy Demand in France. *Emp. Econ.* 29:193-208.

Delyon. B., M. Lavielle and E. Moulines, 1999. Convergence of a stochastic approximation version of the EM algorithm. *Ann. of Stat*. 94-128.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. of Roy. Stat. Soc., Series B.* 39(1):1-38.

Fezzi, C. and I.J. Bateman, 2011. Structural agricultural land use modelling for spatial agro-environmental policy analysis. *Am. J. of Ag. Econ.* 93(4):1168-1188.

Heckeleï, T., W. Britz and Y. Zhang, 2012. Positive mathematical programming approaches. Recent developments in literature and applied modelling. *Bio-based and App. Econ.* 1(1):109-124.

Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153-61.

Heckman, J. J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* 5(4): 475-49.

Howitt, E., 1995. Positive Mathematical Programming. *Am. J. of Ag. Econ.* 77:329-342.

Jank, W. 2006. Implementing and diagnosing the stochastic approximation EM algorithm. *J. of Comp. and Graph. Stat.* 15(4), 803–829.

Jank, W. and J. Booth, 2003. Efficiency of Monte Carlo EM and Simulated Maximum Likelihood in Two-Stage Hierarchical Models. *J. of Comp. and Graph. Stat.* 12(1), 214–229.

Kao, C., L.-F. Lee and M.M. Pitt, 2001. SML Estimation of the LES with Binding Non-negativity Constraints. *Ann. of Econ. and Fin.* 2:215-55.

Lacroix, A. and A. Thomas, 2011. Estimating the Environmental Impact of Land and Production Decisions with Multivariate Selection Rules and Panel Data. *Am. J. of Ag. Econ.* 93(3):780-798.

Lee, L.-F. and M.M. Pitt, 1986. Microeconometric demand systems with binding nonnegativity constraints: the dual approach. *Econometrica* 54:1237-1242.

McLachlan G. and T. Krishnan, 2008. *The EM algorithm and extensions. 2nd Ed*. Wiley Ed.

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

Perali, F. and J.-P. Chavas, 2000. Estimation of Censored Demand Equations from Large Cross-Section Data. *Amer. J. of Ag. Econ*. 82(4):1022-1037.

Platoni, S. P. Sckokai and D. Moro, 2012. Panel Data Estimation Techniques and Farm-level Data Models. *Am. J. of Ag. Econ.* 94(4):1202-1217.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *J. of Econometrics* 49(3):305-341.

Sckokai, P. and D. Moro, 2009. Modelling the impact of the CAP Single Farm Payment on farm investment and output *Eur. Rev. of Agric. Econ.* 36(3):395-423.

Sckokaï, P. and D. Moro, 2006. Modeling the Reforms of the Common Agricultural Policy for Arable Crops under Uncertainty. *Am. J. of Ag. Econ.* 88(1):43-56.

Shonkwiler, J.S. and S.T. Yen, 1999. Two-Step Estimation of a Censored System of Equations. *Am. J. of Ag. Econ.* 81:972-982.

Train K., 2008. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling.* 1(1), 40–69.

Train K., 2007. *A recursive estimator for random coefficient models.* University of California. Berkeley.

Wales,T.J., A.D. Woodland, 1983. Estimation of Consumer Demand Systems with Binding Non-Negativity Constraints. *J. of Econometrics* 21:263-285.

Wu, C. J., 1983. On the convergence properties of the EM algorithm. *The Ann. Of Stat.*, 11(1):95-103.

Yen, S.T., 2005. "A Multivariate Sample-Selection Model: Estimating Cigarette and Alcohol Demands with Zero Observations. *Am. J. of Ag. Econ.* 87(2):453–66.

Yen, S.T., K. Kan and S. Su, 2002. Household Demand for Fats and Oils: Two-Step Estimation of a Censored Demand System. *App. Econ.* 34(14):1799–806.

Yen, S.T., B. Lin and D.M. Smallwood, 2003. Quasi and Simulated Likelihood Approaches to Censored Demand Systems: Food Consumption by Food Stamp Recipients in the United States. *Am. J. of Ag. Econ.* 85(2):458–78.

Tables and Figures

Table 1. Selected parameter estimates, yield equations

	$\beta_k^y + \mu_k^y E[e_i]$	μ_k^y	γ_k	$Var[\varepsilon_{k,i}^y]$	Share of farmers growing the crop (%)
Wheat ($k = 1$ **)**	7.84	1.00	-0.65	1.18	100.00
	(0.11)	-	(0.30)	(0.07)	
Corn ($k = 2$ **)**	8.52	0.33	-0.89	2.40	42.68
	(0.23)	(0.15)	(0.39)	(0.17)	
Other Cereals ($k = 3$ **)**	7.26	0.74	0.45	1.42	88.55
	(0.15)	(0.08)	(0.25)	(0.09)	
Oilseeds ($k = 4$ **)**	6.62	0.59	0.58	0.53	100.00
	(0.14)	(0.07)	(0.20)	(0.08)	
Protein Crops ($k = 5$ **)**	5.91	0.48	-0.41	0.54	64.71
	(0.17)	(0.10)	(0.28)	(0.14)	

Note: standard errors are in parentheses

Figure 1: Nesting structure of the crop set

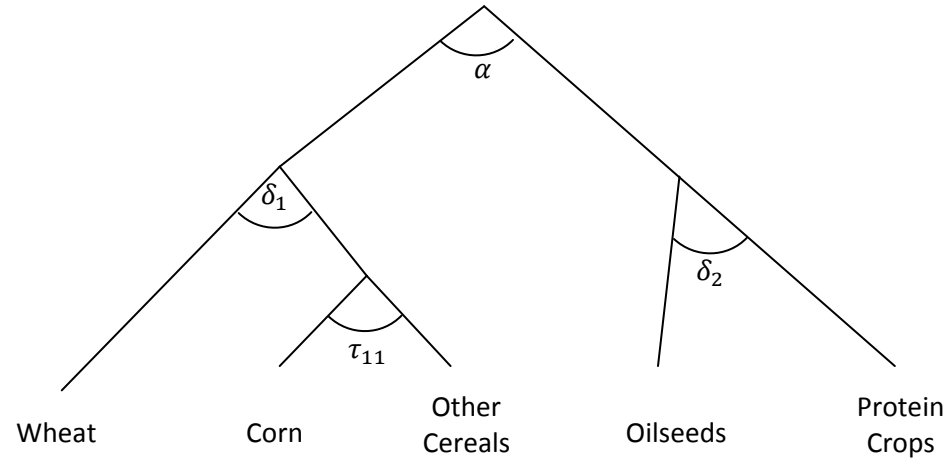

Table 2a. Selected parameter estimates, acreage share equations, curvature parameters

α	δ_1	δ_2	τ_{11}
0.03	0.15	0.06	0.26
(0.002)	(0.005)	(0.003)	(0.011)

Note: standard errors are in parentheses

Table 2b. Selected parameter estimates, acreage share equations, fixed costs terms

	c_k	μ_k^s	$Var[\varepsilon_{k,i}^s]$	Average acreage of farmers growing the crop (ha)	Average acreage in the sample (ha)	Share of farmers growing the crop (%)
Wheat	0	1	-			100.00
	(-)	(-)	-			
Corn ($k = 2$ **)**	6.68	-2.38	9.84	16.56	7.07	42.68
	(0.27)	(0.37)	(0.66)			
Other Cereals ($k = 3$ **)**	4.16	-1.02	20.27	27.26	24.14	88.55
	(0.20)	(0.53)	(1.37)			
Oilseeds ($k = 4$ **)**	37.75	28.39	178.22	26.09	26.09	100.00
	(2.83)	(2.55)	(19.14)			
Protein Crops ($k = 5$ **)**	47.44	34.89	10.54	14.77	9.56	64.71
	(3.24)	(2.50)	(1.28)			

Note: standard errors are in parentheses

Table 3. Selected parameter estimates, regime choice equations

	θ_r^ρ	Frequency of the regime (%)
Wheat – Corn – Oth. Cer. – Oilseeds – Prot. Crop. ($r = 0$ **)**	0	25.23
	(-)	
Wheat – Oilseeds – Prot. Crop. ($r = 1$ **)**	-0.90	5.13
	(0.19)	
Wheat – Oth. Cer. – Oilseeds ($r = 2$ **)**	-2.88	24.17
	(0.19)	
Wheat – Oth. Cer. – Oilseeds – Prot. Crop. ($r = 3$ **)**	-0.67	28.03
	(0.08)	
Wheat – Corn – Oilseeds – Prot. Crop. ($r = 4$ **)**	0.42	6.32
	(0.12)	
Wheat – Corn – Oth. Cer. – Oilseeds ($r = 5$ **)**	-1.44	11.12
	(0.16)	
ψ^ρ	1.26	
	(0.09)	

Note: standard errors are in parentheses

[1] With dimension *K* in equation (1).

[2] *I.e.* $\rho(\mathbf{s}) = r$ and only if $\{k \in \mathcal{K} / s_k > 0\} = \mathcal{K}_{(r)}$.

[3] Similarly, different crops may generate work peak loads at some dates. These peak loads lead to increases in the "smooth" acreage management costs represented by $D(\mathbf{s})$. But they only generate a single fixed cost. The farmer must be on his farm at these dates, whether this is due to a single crop or to several crops doesn't matter.

[4] A part-time farmer may have high regime fixed costs for regimes with numerous crops. The "smooth" acreage management costs and the regime fixed costs decrease in the quantities of quasi-fixed factor quantities available on the farmer. The transaction costs included in regime fixed costs also depend on the market opportunities available to the farmer.

[5] Depending on whether this heterogeneity can be controlled by suitable variables or not

[6] If a farmer is unable to produce crop *k* – due to his lacking necessary inputs or market opportunities – then this crop must be excluded from the crop set considered by this farmer. The equivalent mathematical convention stating that $b_r = +\infty$ if the production regime *r* contains crop *k* is of limited interest in this context.

[7] This matrix is obtained from the dimension *K* identity matrix by deleting its rows corresponding to the crops not contained in $\mathcal{K}_{(r)}$.

[8] Note that $\mathbf{Q}'_{(r,+)}\mathbf{Q}_{(r,+)}\mathbf{s} = \mathbf{s}$ if and only if $\rho(\mathbf{s}) = j$ with $\mathcal{K}_{(j)} \subseteq \mathcal{K}_{(r)}$.

[9] It achieves its minimum in $\mathbf{s}$ on $\mathcal{U}$, $A - \alpha^{-1} \ln\left(\boldsymbol{\iota}' \exp(-\alpha\mathbf{c})\right)$, at $\mathbf{s} = \exp(-\alpha\mathbf{c})\left(\boldsymbol{\iota}' \exp(-\alpha\mathbf{c})\right)^{-1}$.

[10] The variable input uses are not modeled for simplicity.

[11] With observations at the crop level variable input demand functions would have been specified for completing the considered multicrop model.

[12] This specification ignores possible correlations between the regime fixed cost terms $\mathbf{b}_i$ and the latent productivity index e_i. This simplifying assumption is admittedly restrictive.

[13] Note also that we have:

$$f(\mathbf{s}_i^+ \mid \mathbf{z}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) = J_i^+(\boldsymbol{\theta}^s) \times \varphi\left(\boldsymbol{\varepsilon}_i^{s,+}(\boldsymbol{\theta}^y, \boldsymbol{\theta}^s) - \boldsymbol{\Psi}_{+0,i}^s (\boldsymbol{\Psi}_{00,i}^s)^{-1} \boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{++,i}^s - \boldsymbol{\Psi}_{+0,i}^s (\boldsymbol{\Psi}_{00,i}^s)^{-1} \boldsymbol{\Psi}_{0+,i}^s\right)$$

where $\boldsymbol{\Psi}_{00,i}^s \equiv \mathbf{Q}_{(r_i,0)} \boldsymbol{\Psi}^s \mathbf{Q}'_{(r_i,0)}$, $\boldsymbol{\Psi}_{+0,i}^s \equiv \mathbf{Q}_{(r_i,+)} \boldsymbol{\Psi}^s \mathbf{Q}'_{(r_i,0)}$ and $\boldsymbol{\Psi}_{0+,i}^s = (\boldsymbol{\Psi}_{+0,i}^s)'$. We use here the conditioning properties of joint normal variable vectors, *i.e.* $E[\boldsymbol{\varepsilon}_i^{s,+} \mid \boldsymbol{\varepsilon}_i^{s,0}] = \boldsymbol{\Psi}_{+0,i}^s (\boldsymbol{\Psi}_{00,i}^s)^{-1} \boldsymbol{\varepsilon}_i^{s,0}$ and $V[\boldsymbol{\varepsilon}_i^{s,+} \mid \boldsymbol{\varepsilon}_i^{s,0}] = \boldsymbol{\Psi}_{++,i}^s - \boldsymbol{\Psi}_{+0,i}^s (\boldsymbol{\Psi}_{00,i}^s)^{-1} \boldsymbol{\Psi}_{0+,i}^s$. Of course we have:

$$f(\mathbf{s}_i^+, \boldsymbol{\varepsilon}_i^{s,0} \mid \mathbf{z}_i, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) = f(\mathbf{s}_i^+ \mid \mathbf{z}_i, \boldsymbol{\varepsilon}_i^{s,0}, e_i; \boldsymbol{\theta}^y, \boldsymbol{\theta}^s, \boldsymbol{\Psi}^s) f(\boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^s).$$

[14] Note that we also have:

$$\varpi_i(\boldsymbol{\varepsilon}_i^{s,0}, e_i; \mathbf{a}) = \frac{f(\boldsymbol{\kappa}_i, \boldsymbol{\varepsilon}_i^{s,0} \mid \mathbf{z}_i, e_i; \mathbf{a})}{\int f(\boldsymbol{\kappa}_i, \boldsymbol{\varepsilon}_i^{s,0} \mid \mathbf{z}_i, e_i; \mathbf{a}) f(e_i) d(\boldsymbol{\varepsilon}_i^{s,0}, e_i)} f(\boldsymbol{\varepsilon}_i^{s,0}; \boldsymbol{\Psi}_{00,i}^+)^{-1}.$$

[15] Other stopping rules can also be used (see, *e.g.*, McLachlan and Krishnan, 2008 ; Jank, 2006).

[16] Sugar beet producers are excluded from our sample because this market was still highly regulated by production quotas during most of the sample period.