Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural change

Legrand Dunold Fils Saint-Cyr, Laurent Piet

Working Papers SMART - LERECO

Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural change

Legrand D.F. SAINT-CYR, Laurent PIET

Working Paper SMART – LERECO N°15-06

September 2015

UMR INRA-Agrocampus Ouest **SMART** (Structures et Marchés Agricoles, Ressources et Territoires)
UR INRA **LERECO** (Laboratoires d'Etudes et de Recherches en Economie)

Les Working Papers SMART-LERECO ont pour vocation de diffuser les recherches conduites au sein des unités SMART et LERECO dans une forme préliminaire permettant la discussion et avant publication définitive. Selon les cas, il s'agit de travaux qui ont été acceptés ou ont déjà fait l'objet d'une présentation lors d'une conférence scientifique nationale ou internationale, qui ont été soumis pour publication dans une revue académique à comité de lecture, ou encore qui constituent un chapitre d'ouvrage académique. Bien que non revus par les pairs, chaque working paper a fait l'objet d'une relecture interne par un des scientifiques de SMART ou du LERECO et par l'un des deux éditeurs de la série. Les Working Papers SMART-LERECO n'engagent cependant que leurs auteurs.

The SMART-LERECO Working Papers are meant to promote discussion by disseminating the research of the SMART and LERECO members in a preliminary form and before their final publication. They may be papers which have been accepted or already presented in a national or international scientific conference, articles which have been submitted to a peer-reviewed academic journal, or chapters of an academic book. While not peer-reviewed, each of them has been read over by one of the scientists of SMART or LERECO and by one of the two editors of the series. However, the views expressed in the SMART-LERECO Working Papers are solely those of their authors.

Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural change

Legrand D.F. SAINT-CYR

Agrocampus Ouest, UMR1302 SMART, F-35000 Rennes, France
INRA, UMR1302 SMART, F-35000 Rennes, France

Laurent PIET

INRA, UMR1302 SMART, F-35000 Rennes, France

Corresponding author

Legrand D.F. Saint-Cyr
INRA, UMR SMART
4 allée Adolphe Bobierre, CS 61103
35011 Rennes cedex, France
Email: ldfsaint@rennes.inra.fr
Téléphone / Phone: +33 (0)2 23 48 54 18
Fax: +33 (0)2 23 48 53 80

Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural change

Abstract

This article compares the respective performance of the mover-stayer model (MSM) and the Markov chain model (MCM) to investigate whether accounting for unobserved heterogeneity in the rate of movements of farms across size categories improves the representation of the transition process. The MCM has become a popular tool in agricultural economics research to describe how farms experience structural change and to study the impact of the various drivers of this process, including public support. Even though some studies have accounted for heterogeneity across farms by letting transition probabilities depend on covariates depicting characteristics of farms and/or farmers, only observed heterogeneity has been considered so far. Assuming that structural change may also relate to unobserved characteristics of farms and/or farmers, we present an implementation of the MSM which considers a mixture of two types of farms: the `stayers' who always remain in their initial size category and the `movers' who follow a first-order Markovian process. This modeling framework relaxes the assumption of homogeneity in the transition process which the basis of the usual MCM. Then, we explain how to estimate the model using likelihood maximization and the expectation-maximization (EM) algorithm. An empirical application to a panel of French farms over 2000-2013 shows that the MSM outperforms the MCM in recovering the underlying year-on-year transition process as well as in deriving the long-run transition matrix and predicting the future distribution of farm sizes.

Keywords: structural change, unobserved heterogeneity, Markov chain, mover-stayer model, em algorithm

JEL classifications: Q12, C15, D92

"Movers" et "stayers" en agriculture : une prise en compte de l'hétérogénéité inobservée dans le changement structurel

Résumé

Nous comparons les performances respectives du modèle « mover-stayer » (MSM) et du modèle de chaîne de Markov (MCM) afin d'étudier si la prise en compte de l'hétérogénéité inobservée dans les taux de mouvement entre classes de taille améliore la représentation du processus de transition. Le MCM est devenu un outil largement utilisé en économie agricole pour représenter le changement structurel des exploitations et étudier l'impact de différents facteurs, notamment les aides publiques, sur ce processus. Même si certains travaux intègrent une certaine hétérogénéité entre exploitations en faisant dépendre les probabilités de transition de variables caractéristiques des exploitations et/ou des exploitants, seule l'hétérogénéité observée a été prise en compte jusqu'à maintenant. Partant de l'hypothèse que le changement structurel peut également dépendre de caractéristiques inobservées des exploitations et/ou des exploitants, nous présentons une application du MSM qui consiste à considérer un mélange de deux types d'exploitations : les « stayers » qui restent indéfiniment dans leur catégorie de taille initiale, et les « movers » qui suivent un processus Markovien de degré 1. Cette approche permet ainsi de relâcher l'hypothèse d'homogénéité dans le processus de transition qui sous-tend le MCM habituel. Nous présentons ensuite comment estimer le modèle grâce à la méthode du maximum de vraisemblance et à l'algorithme EM (expectation-maximization). Une application empirique à un panel d'exploitations françaises observées sur la période 2000-2013 montre que le MSM permet de mieux représenter le processus annuel de transition que le MCM et ainsi de dériver une meilleure matrice de transition à long-terme, conduisant au final à mieux prédire la distribution future des tailles des exploitations.

Mots-clés : changement structurel, hétérogénéité inobservée, chaîne de Markov, modèle mover-stayer, algorithme EM

Classifications JEL : Q12, C15, D92

Movers and stayers in the farming sector: accounting for unobserved heterogeneity in structural change

1 Introduction

The agricultural sector has faced important structural changes over the past decades. In most developed countries, particularly in Western Europe and the United States, the total number of farms has decreased significantly and the average size of farms has increased continually, implying changes in the distribution of farm sizes. According to Weiss (1999), such changes may have important consequences for equity within the agricultural sector (regarding income distribution and competitiveness among farms), for the productivity and efficiency of farming, and on the demand for government services and infrastructure and the well-being of local communities. Thus, structural change has been the subject of considerable interest among agricultural economists and policy makers. Such studies aim in particular at understanding the mechanisms underlying these changes in order to identify the key drivers that influence the observed trends, and to generate prospective scenarios.

As Zimmermann, Heckelei, and Dominguez (2009) showed, it has become quite common in the agricultural economics literature to study the way farms experience structural change using to the so-called Markov chain model (MCM). Basically, this model states that the size of a farm at a given date is the result of a probabilistic process which only depends on its size in previous periods. Methodologically, most of these studies have used 'aggregate' data, that is, cross-sectional observations of the distribution of a farm population into a finite number of size categories. Such data are often easier to obtain than individual-level data, and Lee, Judge, and Takayama (1965) and Lee, Judge, and Zellner (1977) demonstrated that robustly estimating an MCM from such aggregate data is possible. More recently, because estimating an MCM may well be an ill-posed problem since the number of parameters to be estimated is often larger than the number of observations (Karantininis, 2002), much effort has been dedicated to developing efficient ways to parameterize and estimate these models, ranging from a discrete multinomial logit formulation (MacRae, 1977; Zepeda, 1995), the maximization of a generalized cross-entropy model with instrumental variables (Karantininis, 2002; Huettel and Jongeneel, 2011; Zimmermann and Heckelei, 2012), a continuous re-parameterization (Piet, 2011), to the use of Bayesian inference (Storm, Heckelei, and Mittelhammer, 2011).

Empirically, MCMs were first used within a stationary and homogeneous approach, assuming that transition probabilities were invariant over time and that all agents in the population changed categories according to the same unique stochastic process. Despite improvements in the specification and estimation of this basic model, the resulting estimated parameters generally lead to erroneous forecasts of farm size distributions (Hallberg, 1969; Stavins and Stanton,

1980) because of this homogeneity assumption. Several studies have therefore been devoted to improving the Markov chain modeling framework. Two directions in particular have been investigated. First, assuming that transition probabilities of farms may vary over time, non-stationary MCMs have been developed in order to investigate the effects of time-varying variables on farm structural change, including agricultural policies (see Zimmermann, Heckelei, and Dominguez (2009) for a review). Second, assuming that the transition process may differ depending on certain characteristic of farms and/or farmers (such as regional location, type of farming, legal status, age group, etc.), some studies have accounted for farm heterogeneity in modeling structural change.

Our article adds to this second strand of the literature. Using MCMs, farm heterogeneity has usually been incorporated either by letting the transition probabilities depend on a set of dummy variables (see Zimmermann and Heckelei (2012) for a recent example) or by fitting the usual MCM to sub-populations, partitioned ex-ante according to certain exogenous variables (Huettel and Jongeneel, 2011). To our knowledge, only observed heterogeneity has thus been considered in these studies, implying that all farms that share the same observed characteristics follow the same stochastic process. In this article, as can be found in other strands of the economic literature (Langeheine and Van de Pol, 2002), we argue that the factors driving the evolution of the structure of the farms at the individual level may also relate to unobserved characteristics of farms and/or farmers. However, accounting for such an unobserved heterogeneity among farms requires working at the individual level rather than the 'aggregate' level. Therefore, as farm-level data has become more widely available, we propose to using a more general modeling framework than the simple MCM, namely the mixed-MCM (M-MCM), which exactly enables unobserved heterogeneity in the transition process to be accounted for. As mentioned, this extended modeling framework has so far been applied to the study of economic issues, such as labor mobility (Blumen, Kogan, and McCarthy, 1955; Fougère and Kamionka, 2003), credit rating (Frydman and Kadam, 2004; Frydman and Schuermann, 2008), income or firm size dynamics (Dutta, Sefton, and Weale, 2001; Cipollini, Ferretti, and Ganugi, 2012), but not yet to the agricultural sector.

Since structural change in agriculture refers to a long-run process (it may take time for farms to make just one transition from one size category to another), we assume that accounting for unobserved heterogeneity in the rate of movement of farms may allow the underlying data generating process to be recovered in a more efficient way than in the homogeneous MCM. This extended modeling framework should therefore also lead to better forecasts of farm size distribution and to a more efficient investigation of the effects on farm structural change of time-varying variables, including agricultural policies as well as observed individual farms and/or farmers' characteristics. As an illustration, we apply the simplest version of the M-MCM, the mover-stayer model (MSM), to estimate transition probability matrices and to perform short- to long-run out-of sample projections of the distribution of farm sizes. We do this using an unbalanced panel of 17,285 commercial French farms observed during the period 2000-2013.

The objective is to compare the performance of the MSM extended modeling framework with the simple MCM, first, in predicting the size transition probabilities of farms and, second, in performing farm size distribution forecasts over time.
With respect to the agricultural economics literature, the originality of this article is therefore threefold. First, we implement the MSM to account for unobserved heterogeneity across farms in their size change behavior. Second, we compare this extended modeling framework to the standard MCM in order to assess which model performs better in estimating the underlying transition process. Third, we use a bootstrap sampling method to compute robust standard errors for the transition probabilities we estimate and the farm size distributions we predict.
The article is structured as follows. In the first section we introduce how the traditional MCM can be generalized into the M-MCM and how the specific MSM is derived. In the next two sections we describe the method used to estimate the MSM parameters and the two measures, namely the likelihood ratio (LR) and the average marginal error (AME), used to compare the respective performances of the MCM and the MSM. Then we report our application to France, starting with a description of the data used and following with a presentation of the results. Finally, we conclude with some considerations on how to further extend the approach described here.

2 Modeling a transition process using the Markov chain framework

Consider a population of N agents which is partitioned into a finite number J of categories or 'states of nature'. Assuming that agents move from one category to another during a certain period of time r according to a stochastic process, we define the number $n_{j,t+r}$ of individuals in category j at time $t+r$ as:

$$n_{j,t+r} = \sum_{i=1}^{J} n_{i,t} p_{ij,t}^{(r)} \tag{1}$$

where $n_{i,t}$ is the number of individuals in category i at time t, and $p_{ij,t}^{(r)}$ is the probability of moving from category i to category j between t and $t+r$. As such, $p_{ij,t}^{(r)}$ is subject to the standard non-negativity and summing-up to unity constraints for probabilities:

$$\begin{array}{rl} p_{ij,t}^{(r)} & \geq 0, \quad \forall i,j,t \\ \sum_{j=1}^{J} p_{ij,t}^{(r)} & = 1, \quad \forall i,t. \end{array} \tag{2}$$

In the following, without loss of generality, we restrict our analysis to the stationary case where the r-step transition probability matrix (TPM), $\mathbf{P}_t^{(r)} = \{p_{ij,t}^{(r)}\}$, is time-invariant, *i.e.*, $\mathbf{P}_t^{(r)} = \mathbf{P}^{(r)}$ for all t. In matrix notation, equation (1) can then be rewritten as:

$$\mathbf{n}_{t+r} = \mathbf{n}_t \times \mathbf{P}^{(r)} \tag{3}$$

where $\mathbf{n}_{t+r} = \{n_{j,t+r}\}$ and $\mathbf{n}_t = \{n_{j,t}\}$ are row vectors.
Using individual level data, the observed r-step transition probabilities can then be computed from a contingency table as:

$$p_{ij}^{(r)} = \frac{\nu_{ij}^{(r)}}{\sum_j \nu_{ij}^{(r)}} \tag{4}$$

where $\nu_{ij}^{(r)}$ is the total number of r-step transitions from category i to category j during the period of observation and $\sum_j \nu_{ij}^{(r)}$ the total number of r-step transitions out of category i.

2.1 The simple Markov chain model (MCM)

The first-order Markov chain consists in assuming that the category of an agent in any period depends only on its situation in the very preceding period. Then, Anderson and Goodman (1957) showed that the maximum likelihood estimator of $\mathbf{\Pi}$, the 1-year TPM under the MCM, corresponds to the observed transition matrix, that is, $\mathbf{\Pi} = \mathbf{P}^{(1)}$.
Under the MCM and stationarity assumption, the r-step TPM ($\mathbf{\Pi}^{(r)}$) is then obtained by raising the 1-step transition matrix to the power r:

$$\mathbf{\Pi}^{(r)} = (\mathbf{\Pi})^r. \tag{5}$$

In doing so, the MCM approach assumes that agents in the population are homogeneous, *i.e.*, they all move according to the same stochastic process described by $\mathbf{\Pi}$. However, in general, while $\mathbf{\Pi}$ is an unbiased estimator of $\mathbf{P}^{(1)}$, $\mathbf{\Pi}^{(r)}$ proves to be a poor estimate of $\mathbf{P}^{(r)}$ (Blumen, Kogan, and McCarthy, 1955; Spilerman, 1972). In particular, the main diagonal elements of $\mathbf{\Pi}^{(r)}$ largely underestimate those of $\mathbf{P}^{(r)}$. This means that, in general, $\pi_{ii}^{(r)} \ll p_{ii}^{(r)}$. In other words, the simple MCM tends to overestimate the mobility of agents because of the homogeneity assumption.

2.2 Accounting for unobserved heterogeneity: the mixed Markov chain model (M-MCM)

One way to obtain a 1-step TPM which leads to a more consistent r-step estimate consists in relaxing the assumption of homogeneity in the transition process which underlies the MCM approach.
Frydman (2005) proposed grounding the source of population heterogeneity on the rate of movement of agents; agents may move across categories at various speeds, each according to one of several types of transition process. This usually constitutes unobserved heterogeneity because observing the set of transitions an agent actually experienced does not unambiguously reveal, in general, which stochastic process generated this specific sequence, hence the agent's type.

Implementing this idea leads to considering a mixture of Markov chains in order to capture the population heterogeneity. More precisely, consider that the population is partitioned (in an unobservable way) into a discrete number G of homogeneous types of agents instead of just one, each agent belonging to one and only one of these types. Assuming that each agent type is characterized by its own elementary Markov process, the general form of the M-MCM then consists in decomposing the 1-step transition matrix as:

$$\mathbf{\Phi} = \{\phi_{ij}\} = \sum_{g=1}^{G} \mathbf{S}_g \mathbf{M}_g \tag{6}$$

where $\mathbf{M}_g = \{m_{ij,g}\}$ is the TPM defining the 1-step Markov process followed by type-g agents, and $\mathbf{S}_g = \text{diag}(s_{i,g})$ is a diagonal matrix which gathers the shares of type-g agents in each category. Since every agent in the population has to belong to one and only one type g, the constraint that $\sum_{g=1}^{G} \mathbf{S}_g = \mathbf{I}$ must hold, where $\mathbf{I}$ is the $J \times J$ identity matrix.
Since we consider here the stationary case only, it is assumed that neither $\mathbf{M}_g$ nor $\mathbf{S}_g$ varies over time. Then, the r-step TPM for any future time period r can be defined as the linear combination of the r-step G processes:

$$\mathbf{\Phi}^{(r)} = \sum_{g=1}^{G} \mathbf{S}_g (\mathbf{M}_g)^r. \tag{7}$$

With the MCM and M-MCM modeling frameworks defined as above, it should be noted that: (i) the M-MCM reduces to the MCM if $G = 1$, that is, only one type of agents is considered or, equivalently, the homogeneity assumption holds; and (ii) the aggregate overall M-MCM process described by $\mathbf{\Phi}^{(r)}$ as defined by equation (7) may no longer be Markovian even if each agent type follows a specific Markov process.

2.3 A simple implementation of the M-MCM: the Mover-Stayer model (MSM)

Since the number of parameters to estimate increases rapidly with the number of homogeneous agent types (G), the estimation of matrix $\mathbf{P}$ as defined in the general case by equation (6) can quickly become an ill-posed problem.[1] Here, we stick to the simplest version of the M-MCM, namely the MSM first proposed by Blumen, Kogan, and McCarthy (1955). In this restricted approach, only two types of homogeneous agents are considered, those who always remain in their initial category (the 'stayers') and those who follow a first-order Markovian process (the 'movers'). Formally, this leads to rewriting equation (6) in a simpler form as:

$$\mathbf{\Phi} = \mathbf{S} + (\mathbf{I} - \mathbf{S})\,\mathbf{M}. \tag{8}$$

[1] To solve this issue in the general case, Frydman (2005) proposed a parameterization of the M-MCM to decrease the number of parameters to be estimated (see appendix).

With respect to the general formulation (6), this corresponds to setting $G = 2$ and defining $\mathbf{S}_1 = \mathbf{S}$ and $\mathbf{M}_1 = \mathbf{I}$ for stayers, and $\mathbf{S}_2 = (\mathbf{I} - \mathbf{S})$ and $\mathbf{M}_2 = \mathbf{M}$ for movers.[2] Thus, following equation (7), the MSM overall population r-step TPM can be expressed as:

$$\mathbf{\Phi}^{(r)} = \mathbf{S} + (\mathbf{I} - \mathbf{S})\,(\mathbf{M})^r. \tag{9}$$

3 Estimation method

In their early attempt to empirically implement the MSM, Blumen, Kogan, and McCarthy (1955) used a simple calibration method to estimate the parameters of the model. Then, since Goodman (1961) showed that the Blumen, Kogan, and McCarthy (1955) estimators are actually biased, alternative methods were developed to obtain consistent estimates using, for example, minimum chi-square (Morgan, Aneshensel, and Clark, 1983), maximum likelihood (Frydman, 1984) or Bayesian inference (Fougère and Kamionka, 2003). Frydman (2005) was the first to develop a maximum likelihood estimation method for the general M-MCM (see appendix). In the following, we present the corresponding strategy in the simplified case of the MSM, which consists of two steps: first, under complete information, that is, as if the population heterogeneity were perfectly observable; second, under incomplete information, that is, accounting for the fact that the population heterogeneity is not actually observed.

3.1 Likelihood maximization under complete information

Under complete information the status of each agent k, either stayer (denoted 'S') or mover (denoted 'M'), is perfectly known *ex-ante* and can be recorded through a dummy variable $Y_{k,S}$ where $Y_{k,S} = 1$ if agent k is a stayer and $Y_{k,S} = 0$ if agent k is a mover.
The log-likelihood of the MSM for the whole population is then:

$$\log L = \sum_{k=1}^{N} Y_{k,S} \log l_{k,S} + \sum_{k=1}^{N} (1 - Y_{k,S}) \log l_{k,M} \tag{10}$$

where the first sum on the right hand side is the overall log-likelihood associated with stayers and the second sum is the overall log-likelihood associated with movers.
At the individual level, conditional on knowing that k was initially in size category i:

- the likelihood that agent k is a stayer, $l_{k,S}$, is given by s_i the share of agents who never move out of category i during the whole period of observation (see appendix);

[2] With respect to Frydman (2005)'s specification of the M-MCM presented in the appendix, the mover-stayer model is equivalent to imposing the rate of movement for stayers as zero.

- the likelihood that agent k is a mover is (Frydman and Kadam, 2004):

$$l_{k,M} = (1 - s_i) \prod_{i \neq j} (m_{ij})^{\nu_{ij,k}} \prod_{i} (m_{ii})^{\nu_{ii,k}} \tag{11}$$

  where $\nu_{ij,k}$ is the number of transitions from category i to category j made by agent k and $\nu_{ii,k}$ is the total number of times agent k stayed in category i. On the right hand side of equation (11), the first product is thus the probability of agent k moving out of category i, while the second product is the probability of agent k staying in category i from one period to the next, even though k is a mover.

Substituting $l_{k,S}$ and $l_{k,M}$ in equation (10), the log-likelihood of the MSM for the whole population can be expressed as:

$$\log L = \sum_i n_i \log(1 - s_i) + \sum_i n_{i,S} \log(s_i/(1 - s_i)) + \sum_{i \neq j} \nu_{ij} \log(m_{ij}) + \sum_i \nu_{ii,M} \log(m_{ii}) \tag{12}$$

where n_i and $n_{i,S}$ are, respectively, the numbers of agents and stayers who were initially in category i, $\nu_{ij} = \sum_k \nu_{ij,k}$ is the total number of transitions from category i to category j, $\nu_{ii,M} = \sum_k (1 - Y_{k,s}) \nu_{ii,k}$ is the total number of times movers stayed in category i, and m_{ij} and m_{ii} are the elements of the generator matrix (M) of movers.
Then, maximizing equation (12) with respect to the unknown parameters s_i and m_{ij} leads to the optimal values of the MSM parameters:

- solving for $\partial \log L / \partial s_i = 0$ yields the optimal share of stayers in each category i:

$$\hat{s}_i = \frac{n_{i,S}}{n_i} \tag{13}$$

- solving for $\partial \log L / \partial m_{ij} = 0$ for $i \neq j$ and noting that, by definition of M as a stochastic matrix, $m_{ii} = 1 - \sum_{i \neq j} m_{ij}$ yields:

$$\hat{m}_{ii} = \frac{\nu_{ii,M}}{\nu_i + \nu_{ii,M}} \tag{14}$$

  and

$$\hat{m}_{ij} = \frac{\nu_{ij}}{\nu_i} (1 - \hat{m}_{ii}) \quad \forall i \neq j \tag{15}$$

  where $\nu_i = \sum_{j \neq i} \nu_{ij}$ is the total number of transitions out of category i.

3.2 The expectation-maximization (EM) algorithm under incomplete information

Since, as already mentioned, it is unlikely in practice that one knows beforehand which agents are stayers and which are movers, equation (12) cannot be used directly to estimate the MSM

parameters. Indeed, because the transition process is assumed to be a stochastic process, even movers may remain for a long time in their initial category before moving, so that they may not appear as movers but as stayers over the observed period. Therefore, Fuchs and Greenhouse (1988) suggested that the MSM parameters could be estimated using the EM algorithm developed by Dempster, Laird, and Rubin (1977): rather than observing the dummy variable $Y_{k,S}$, the EM algorithm allows its expected value $E(Y_{k,S})$ to be estimated, *i.e.*, the probability for each agent k to be a stayer, given agent k's initial category and observed transition sequence.
Following Frydman and Kadam (2004), the four steps of the EM algorithm are defined in the case of the MSM as follows.

(i) Initialization: Arbitrarily choose initial values s_i^0 for the shares of stayers and m_{ii}^0 for the main diagonal entries of the generator matrix ($\mathbf{M}$) of movers.

(ii) Expectation: At iteration p of the algorithm, compute the probability of observing agent k as generated by a stayer, $E^p(Y_{k,S})$. If at least one transition is observed for agent k then set $E^p(Y_{k,S}) = 0$, otherwise set it to:

$$E^p(Y_{k,S}) = \frac{s_i^p}{s_i^p + (1 - s_i^p)(m_{ii}^p)^{\nu_{ii,k}}} \tag{16.i}$$

Then compute:

- the expected value of the number of stayers in category i, $E^p(n_{i,S})$, as:

$$E^p(n_{i,S}) = \sum_k E^p(Y_{k,S}) \tag{16.ii}$$

- and the expected value of the total number of times movers remain in category i, $E^p(\nu_{ii,M})$, as:

$$E^p(\nu_{ii,M}) = \sum_k (1 - E^p(Y_{k,S}))\nu_{ii,k} \tag{16.iii}$$

(iii) Maximization: Update s_i^p and m_{ii}^p as follows:

$$s_i^{p+1} = \frac{E^p(n_{i,S})}{n_i} \quad \text{and} \quad m_{ii}^{p+1} = \frac{E^p(\nu_{ii,M})}{\nu_i + E^p(\nu_{ii,M})} \tag{16.iv}$$

(iv) Iteration: Return to expectation step (ii) using s_i^{p+1} and m_{ii}^{p+1} and iterate until convergence.

When convergence is reached, the optimal values s_i^* and m_{ii}^* are considered to be the estimators $\hat{s}_i$ and $\hat{m}_{ii}$. Then, $\hat{m}_{ij}$ derives from $\hat{m}_{ii}$ as in equation (15).
Following Frydman (2005), the standard errors attached to the MSM parameters can be computed directly from the EM equations using the method proposed by Louis (1982) (see appendix). The standard errors attached to the overall 1-year TPM $\mathbf{\Phi}$ can then be derived applying

the standard Delta method to equation (8). Finally, because it is more complicated to apply the Delta method to equation (9) as it involves powers of matrices, we used a bootstrap sampling method to compute standard errors attached to the r-step TPMs $\mathbf{M}^{(r)}$ and $\mathbf{\Phi}^{(r)}$ (Efron, 1979; Efron and Tibshirani, 1986).

4 Model comparison

Two types of analysis were performed to assess whether or not the MSM outperforms the MCM.

4.1 Likelihood ratio (LR) test

The likelihood ratio test allows the in-sample performance of the two models in recovering the data generating process to be compared. As stated by Frydman and Kadam (2004), the likelihood ratio statistic for the MSM is given by:

$$LR = \frac{L_{MCM}(\hat{\mathbf{\Pi}})}{L_{MSM}(\hat{\mathbf{S}}, \hat{\mathbf{M}})} \tag{17}$$

where L_{MCM} and L_{MSM} are the estimated maximum likelihoods for the MCM and the MSM, respectively. Theoretically, the asymptotic distribution of $-2\log(LR)$, under H_0, is chi-square with $(G-1) \times J$ degrees of freedom. In the case of the MSM, the likelihood ratio tests the hypothesis that the process follows a MCM ($H_0 : \hat{\mathbf{S}} = 0$) against the hypothesis that it is a mixture of movers and stayers ($H_1 : \hat{\mathbf{S}} \neq 0$). The observed log-likelihood for both models can be derived from equation (10), by imposing $Y_{k,S} = 0$ for all agents k for the MCM and by replacing $Y_{k,S}$ by its optimal expected value ($E^*(Y_{k,S})$) for the MSM.

4.2 Average Marginal Error (AME)

The estimated parameters were used to compute the corresponding r-step TPMs, *i.e.*, $\hat{\mathbf{\Pi}}^{(r)} = (\hat{\mathbf{\Pi}})^r$ for the MCM and $\hat{\mathbf{\Phi}}^{(r)} = \hat{\mathbf{S}} + (\mathbf{I} - \hat{\mathbf{S}})(\hat{\mathbf{M}})^r$ for the MSM. These r-step TPMs were then used to perform out-of-sample short- to long-run projections of farm distributions across size categories according to equation (1) (see below).

On the one hand, TPMs from both models were compared to the observed one, providing a second in-sample assessment complementary to the likelihood ratio test. This comparison was based on the average marginal error (AME) defined by Cipollini, Ferretti, and Ganugi (2012) as:

$$AME = \frac{1}{J \times J} \sum_{i,j} \sqrt{\left(\frac{\hat{p}_{ij}^{(r)} - p_{ij}^{(r)}}{p_{ij}^{(r)}} \right)^2} \tag{18}$$

where $\hat{p}_{ij}^{(r)}$ and $p_{ij}^{(r)}$ are the predicted and observed TPM entries, respectively:

- $\hat{p}_{ij}^{(r)} \equiv \hat{\pi}_{ij}^{(r)}$ under the MCM while $\hat{p}_{ij}^{(r)} \equiv \hat{\phi}_{ij}^{(r)}$ under the MSM
- $p_{ij}^{(r)}$ derives from equation (4).

On the other hand, AMEs were similarly computed for both the MCM and MSM projections of farm size distributions with respect to the actually observed ones, providing an out-of-sample comparison of the models.
In contrast to some dissimilarity indexes (Jafry and Schuermann, 2004) or the matrix of residuals (Frydman, Kallberg, and Kao, 1985), the AME provides a global view of the distance between the predicted TPM or population distribution across size categories and the observed ones. It can be interpreted as the average percentage of deviations on predicting the observed TPM or population distribution across size categories. Thus, the higher the AME, the more different the computed TPM or distribution with respect to the observed one. The better model is therefore the one which yields the lowest AME.

5 Data

In our empirical application, we used the 2000-2013 data of the "Réseau d'Information Comptable Agricole" (RICA), the French implementation of the Farm Accountancy Data Network (FADN). FADN is an annual survey which is defined at the European Union (EU) level and is carried out in each member state. The information collected at the individual level relates to both the physical and structural characteristics of farms and their economic and financial characteristics. Note that, to comply with accounting standards which may differ from one country to the other (*e.g.*, the recording of asset depreciation), the questionnaire defined at the EU level may be adapted at the national level, which is the case for France, but this had no consequences for our study.
In France, RICA is produced and disseminated by the statistical and foresight office of the French ministry for agriculture. It focuses on 'middle and large' farms (see below) and constitutes a stratified and rotating panel of approximately 7,000 farms surveyed each year. Some 10% of the sample is renewed every year so that, on average, farms are observed during 5 consecutive years. However, some farms may be observed only once, and others several, yet not consecutive, times. Some farms remained in the database over the whole of the studied period, *i.e.*, fourteen consecutive years. Each farm in the dataset is assigned a weighting factor which reflects its stratified sampling probability. These factors allow for extrapolation at the population level.[3]
As we considered all farms in the sample whatever their type of farming, we chose to concentrate on size as defined in economic terms. In accordance with the EU regulation (CE)

[3]To learn more about RICA France, see http://www.agreste.agriculture.gouv.fr/. The dataset used in this article directly derives from the version of the RICA publicly available at this address. To learn more about FADN in general, see http://ec.europa.eu/agriculture/rica/index.cfm.

N°1242/2008, European farms are classified into fourteen economic size (ES) categories, evaluated in terms of total standard output (SO) expressed in Euros.[4] As mentioned above, in France, RICA focuses on 'medium and large' farms, those whose SO is greater than or equal to 25,000 Euros; this corresponds to ES category 6 and above. Since size categories are not equally represented in the sample, we aggregated the nine ES categories available in RICA into five: strictly less than 50,000 Euros of SO (ES6); from 50,000 to less than 100,000 Euros of SO (ES7); from 100,000 to less than 150,000 Euros of SO (lower part of ES8); from 150,000 to less than 250,000 Euros of SO (upper part of ES8); 250,000 Euros of SO and more (ES9 to ES14).

RICA being a rotating panel, farms which either enter or leave the sample in a given year cannot be considered as actual entries into or exits from the agricultural sector. Thus, we could not work directly on the evolution of farm numbers but rather on the evolution of population shares by size categories, *i.e.*, the size distribution in the population. Table 1 presents the year-on-year evolution by size categories of farm numbers for the extrapolated population, as well as the number of farms in the sample. It also reports the average ES in thousand of Euros of SO both at the sample and extrapolated population levels.

Figure 1 shows that the share of smaller farms (below 100,000 Euros of SO) decreased from 56% to 46% between 2000-2013 while the share of larger farms (above 150,000 Euros of SO) increased from 28% to 38%, and the share of intermediate farms (100,000 to less than 150,000 Euros of SO) remained stable at 16%. As a consequence, as can be seen from table 1 and figure 1, the average economic size of French farms was multiplied by more than 1.25 over this period. Note that table 1 also reveals that the size distribution became more heterogeneous since the standard deviation of the economic size was multiplied by almost 1.5, a feature which has already been observed for the population of French farms as a whole and in other periods (Butault and Delame, 2005; Desriers, 2011). Finally, table 1 reveals that these observations also apply at the sample level, even though the latter is skewed towards larger sizes with respect to the population as a whole.

In order to assess which model performed better, we compared the MCM and the MSM on the basis of both in-sample estimation and out-of-sample size distribution forecasts. To do so, we split the RICA database into two parts: (i) observations from 2000 to 2010 were used to estimate the parameters of both models; (ii) observations from 2011 to 2013 were used to compare the actual farm size distributions with their predicted counterparts for both models. Note that, in doing so we assumed that, in the case of the MSM, eleven years is a long enough time interval to robustly estimate the transition process of movers.

[4] SO has been used as the measure of economic size since 2010. Before this date, economic size was measured in terms of standard gross margin (SGM). However, SO calculations have been retropolated for 2000 to 2009, allowing for consistent time series analysis (European Commission, 2010).

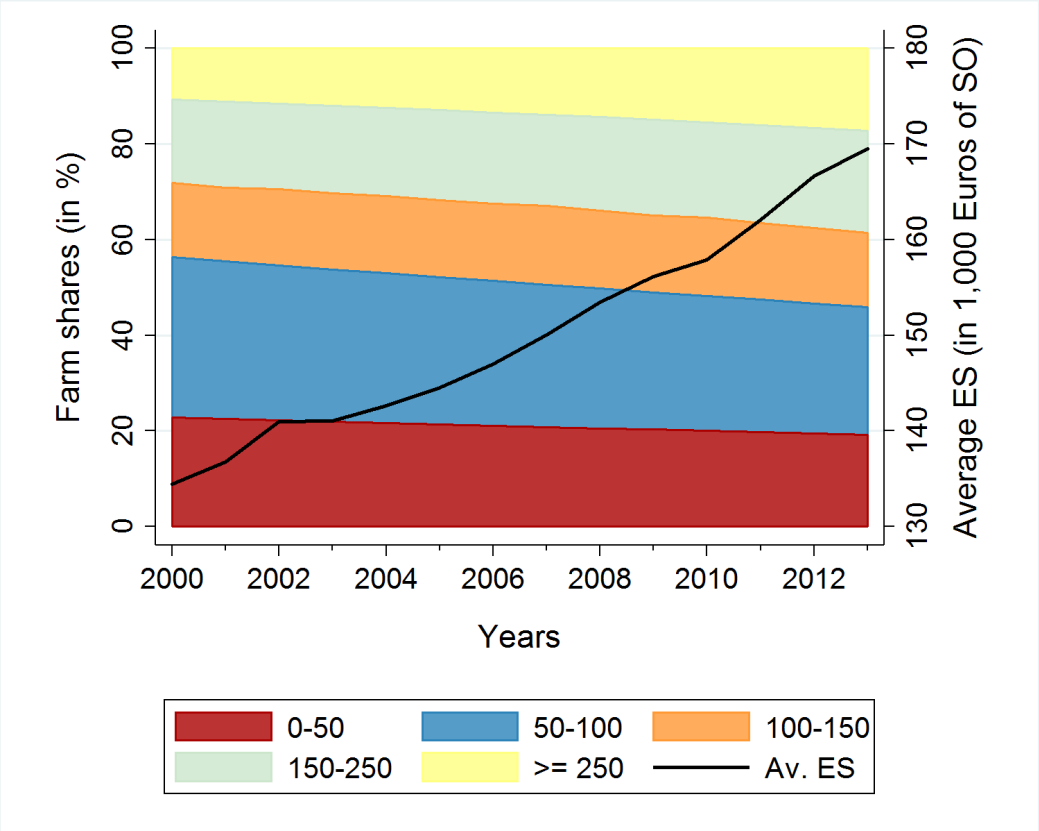
Table 1: Distribution by Economic Size (ES) and Average ES for the Studied Sample, 2000-2013

Year	Number of farms by ES category					Total	Average ES	
	0-50	50-100	100-150	150-250	≥ 250		(std. dev.)	
2000	790	2,234	1,629	1,762	1,342	7,757	168.88	(179.01)
	87,924	129,691	59,857	67,367	41,457	386,296	134.46	(151.72)
2001	746	2,231	1,625	1,817	1,382	7,801	170.98	(180.88)
	84,442	123,900	57,583	67,741	41,890	375,556	136.75	(155.04)
2002	713	2,128	1,663	1,818	1,443	7,765	177.57	(198.12)
	81,228	118,571	58,104	65,448	42,344	365,695	140.99	(184.50)
2003	690	1,975	1,562	1,693	1,393	7,313	176.27	(193.55)
	78,249	113,662	56,961	64,946	42,859	356,677	141.08	(176.08)
2004	707	1,940	1,538	1,707	1,437	7,329	177.67	(188.47)
	75,481	109,118	56,118	64,252	43,419	348,388	142.63	(169.30)
2005	741	1,927	1,516	1,711	1,467	7,362	178.03	(181.95)
	72,896	104,906	54,811	64,112	44,007	340,732	144.55	(161.46)
2006	756	1,922	1,488	1,688	1,491	7,345	181.21	(209.21)
	70,516	101,035	54,202	63,443	44,740	333,936	146.99	(171.49)
2007	774	1,845	1,552	1,694	1,511	7,376	182.27	(191.10)
	68,286	97,435	54,032	62,390	45,491	327,634	150.08	(172.33)
2008	780	1,866	1,511	1,721	1,587	7,465	185.49	(200.25)
	66,201	94,098	52,412	62,889	46,338	321,938	153.47	(185.00)
2009	778	1,816	1,517	1,734	1,624	7,469	188.43	(205.95)
	64,243	90,970	51,137	63,151	47,278	316,779	156.14	(186.03)
2010	652	1,885	1,537	1,770	1,608	7,452	190.53	(199.03)
	62,429	88,104	51,320	62,062	48,267	312,182	157.88	(174.96)
2011	638	1,856	1,468	1,791	1,658	7,411	194.89	(207.58)
	60,743	85,444	49,285	63,292	49,381	308,145	162.11	(189.23)
2012	651	1,797	1,396	1,794	1,679	7,317	200.28	(249.45)
	59,152	82,943	47,911	63,953	50,626	304,585	166.69	(227.41)
2013	658	1,769	1,361	1,804	1,701	7,293	202.41	(240.15)
	57,668	80,638	46,821	64,414	51,939	301,480	169.49	(225.53)

Note: ES in 1,000 Euros of standard output (SO). For each year, the first row reports figures at the sample level and the second row reports figures for the extrapolated population.

Source: Agreste, RICA France 2000-2013 – authors' calculations

Figure 1: Extrapolated population shares by farm size categories and average economic size (ES)



Source: Agreste, RICA France 2000-2013 – authors' calculations

Table 2: Number of Individual Farms, Observations and Transitions by Subsamples

	Subsamples									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Farms	13,123	11,291	9,322	7,680	6,257	5,107	4,112	3,448	2,801	2,170
Observations	78,434	74,770	68,863	62,295	55,180	48,280	41,315	36,003	30,180	23,870
1-year transitions										
$\forall(i,j)$	65,311	63,479	59,541	54,615	48,923	43,173	37,203	32,555	27,379	21,700
$\forall j \neq i$	7,374	7,138	6,672	6,110	5,426	4,700	4,045	3,478	2,889	2,212

Note: subsample #1 corresponds to the subset of farms which remained present in the database for at least two consecutive years; subsample #2 to those which remained for at least three consecutive years; and so forth; the third row gives the number of observed 1-year transitions from category i to category j, including remaining in the initial category; and the fourth row gives the number of observed 1-year transitions from category i to category j excluding remainning in the initial category.

Source: Agreste, RICA France 2000-2010 – authors' calculations

For in-sample estimations, we also had to restrict the subsample to farms which were present in the database for at least two consecutive years in order to observe potential transitions. The corresponding unbalanced panel then comprised 13,123 farms out of the 17,285 farms in the original database (76%), leading to 78,434 farm×year observations and 65,311 individual 1-year transitions (including staying in the same category) from 2000 to 2010. Furthermore, ten subsamples could be constructed from this unbalanced panel, according to the minimum number of consecutive years a farm remained present in the database, from two to eleven: subsample #1 thus corresponds to the subset of farms which remained for at least two consecutive years (full 2000-2010 unbalanced panel); subsample #2 to those which remained for at least three consecutive years; and so forth, up to subsample #10 which corresponds to the 2000-2010 balanced panel. Table 2 reports the corresponding numbers of individual farms and observed transitions for each subsample.

Before proceeding with the results, it should finally be noted that because we had to work with subsets of the full sample, the transition probabilities reported in the next section can only be viewed as *conditional* on having been observed over a specific number of consecutive years during the whole period under study, and therefore should not be considered to be representative for the whole population of 'medium and large' French farms.

6 Results

In this section, we first report the results of in-sample estimations, *i.e.*, the estimated 1-year TPMs for the MCM and the MSM. Then we compare the models based on in-sample results in order to assess which one performs better in recovering the underlying transition process both from a short-run and a long-run perspective. Finally, we compare the models in their ability to forecast future farm size distributions based on out-of-sample observations.

6.1 In-sample estimation results

The MCM and MSM parameters were estimated for each of the ten subsamples described in the previous section. But because it would take too much space to report the results for all of them, only those for subsample #10 are reported here, that is, using the 2000-2010 balanced panel which consists of 2,170 individual farms and 21,700 observed transitions over the 11 years (see last column of table 2). However, when any of the other subsamples are considered, the results, and hence conclusions, remain very similar to those reported here.[5]

Table 3 reports the 1-year TPM estimated from subsample #10 under the MCM assumption. As

[5]Actually, it turned out that considering subsamples which include farms remaining for a shorter period in the database added noise to the estimation of both models: the less time farms remain in the database, the more incomplete the information about them, *i.e.*, the more difficult the estimation of their true behavior. Thus, the AMEs were higher with low rank subsamples than with high rank subsamples, subsample #10 eventually yielding the lowest AMEs.

is usually found in the literature, the matrix is strongly diagonal, meaning that its main diagonal elements exhibit by far the largest values and that probabilities rapidly decrease as we move away from the main diagonal. This means that, overall, farms are more likely to remain in their initial size category from one year to the next. Note that this does not mean no size change at all but, at least, no change sufficient to move to another category as we defined them.

Table 3: Estimated 1-Year TPM under MCM ($\hat{\mathbf{\Pi}}$) (subsample #10)

		ES class				
		0-50	50-100	100-150	150-250	≥ 250
ES class	0-50	**0.917**	**0.079**	**0.002**	**0.002**	**0.001**
		(0.024)	(0.007)	(0.001)	(0.001)	(0.001)
	50-100	**0.030**	**0.898**	**0.065**	**0.005**	**0.002**
		(0.002)	(0.013)	(0.004)	(0.001)	(0.001)
	100-150	**0.002**	**0.062**	**0.854**	**0.080**	**0.002**
		(0.001)	(0.004)	(0.014)	(0.004)	(0.001)
	150-250	**0.001**	**0.004**	**0.054**	**0.886**	**0.055**
		(0.000)	(0.001)	(0.003)	(0.012)	(0.003)
	≥ 250	**0.000**	**0.001**	**0.003**	**0.048**	**0.948**
		(0.000)	(0.001)	(0.001)	(0.003)	(0.014)

Log-likelihood: $\log L_{MCM} = -8,689.36$
Note: estimated parameters in bold font, standard errors in parentheses.

Source: Agreste, RICA France 2000-2010 – authors' calculations

Table 4 reports the estimated shares of stayers and generator matrix of movers along with the corresponding 1-year TPM for the whole population under the MSM assumption, and also for subsample #10. Two main results can be drawn from this table. First, the estimated stayer shares (panel a of table 4) show that the probability of being a stayer is close to or above 30% whatever the category considered; it reaches almost 50% for farms below 50,000 Euros of SO and is even higher than 60% for farms over 250,000 Euros of SO. This means that, according to the MSM and depending on the size category, at least 30% of the farms are likely to remain in their initial category for at least 10 more years. Second, the generator matrix (panel b of table 4) reveals that even though movers are by definition expected to transit from one category to another in the next ten years, yet the highest probability for them is to remain in the same category from one year to the next. Since the average time spent by movers in a particular category is given by $1/(1 - m_{ii})$ (see appendix), it can be seen from table 4 that movers in the intermadiate ES class ($1/(1 - 0.793) = 4.8$) were likely to remain in the same category for almost five years, while those above 250,000 Euros of SO ($1/(1 - 0.875) = 8$) were likely to remain for eight years before moving. In other words, farms which remained in a particular category for quite a long time (theoretically even over the whole observation period) were not necessarily stayers but may well be movers who had not yet moved. Altogether, these two results yield a 1-year

Table 4: Estimated Stayer Shares ($\hat{s}_i$), Mover Generator Matrix ($\hat{\mathbf{M}}$) and Overall Population 1-Year TPM ($\hat{\mathbf{\Phi}}$) (subsample #10)

ES class	$\hat{s}_i$
0-50	**0.494**
	(0.036)
50-100	**0.422**
	(0.021)
100-150	**0.291**
	(0.016)
150-250	**0.371**
	(0.017)
≥ 250	**0.650**
	(0.021)

a) Stayer shares ($\hat{s}_i$)

ES class	0-50	50-100	100-150	150-250	≥ 250
0-50	**0.837**	**0.154**	**0.004**	**0.004**	**0.001**
	(0.041)	(0.012)	(0.002)	(0.002)	(0.001)
50-100	**0.055**	**0.815**	**0.118**	**0.009**	**0.003**
	(0.004)	(0.022)	(0.006)	(0.002)	(0.001)
100-150	**0.002**	**0.089**	**0.793**	**0.113**	**0.003**
	(0.001)	(0.005)	(0.020)	(0.005)	(0.001)
150-250	**0.002**	**0.007**	**0.087**	**0.816**	**0.088**
	(0.001)	(0.001)	(0.004)	(0.020)	(0.004)
≥ 250	**0.001**	**0.003**	**0.007**	**0.114**	**0.875**
	(0.001)	(0.001)	(0.002)	(0.007)	(0.027)

b) Mover generator matrix ($\hat{\mathbf{M}}$)

ES class	0-50	50-100	100-150	150-250	≥ 250
0-50	**0.917**	**0.078**	**0.002**	**0.002**	**0.001**
	(0.019)	(0.011)	(0.001)	(0.001)	(0.001)
50-100	**0.032**	**0.893**	**0.068**	**0.005**	**0.002**
	(0.003)	(0.012)	(0.005)	(0.001)	(0.001)
100-150	**0.002**	**0.062**	**0.854**	**0.080**	**0.002**
	(0.001)	(0.004)	(0.014)	(0.005)	(0.001)
150-250	**0.001**	**0.005**	**0.055**	**0.884**	**0.055**
	(0.000)	(0.001)	(0.004)	(0.012)	(0.004)
≥ 250	**0.000**	**0.001**	**0.003**	**0.040**	**0.956**
	(0.000)	(0.001)	(0.001)	(0.005)	(0.009)

c) Overall population TPM ($\hat{\mathbf{\Phi}}$)

Log-likelihood: $\log L_{MSM} = -8,384.08$
Note: estimated parameters in bold font, standard errors in parentheses.

Source: Agreste, RICA France 2000-2010 – authors' calculations

TPM for the whole population which is also highly diagonal (panel c of table 4).

6.2 In-sample model comparison

Both models are in agreement in estimating that, at the overall population level, farms were more likely to remain in their initial size category from year to year, confirming that structural change in the agricultural sector is a slow process which needs to be investigated in the long-run. In this respect, while the 1-year TPMs look very similar across both models, the resulting long-run transition model differs between the MCM, given by equation (5), and the MSM, given by equation (9). It is therefore important to assess which model performs better in recovering the true underlying transition process.

The first assessment method used, namely the likelihood ratio test, reveals that the MSM yields a better fit than the MCM: the value of the test statistic as defined by equation (17) is $-2\log(LR) = -2 \times (-8,689.36 + 8,384.08) = 610.56$, which is highly significant since the critical value of a chi-square distribution with $(G-1) \times J = 5$ degrees of freedom and the 1% significance level is $\chi^2_{0.99}(5) = 15.09$. This leads to the rejection of the H_0 assumption that the stayer shares are all zero, and thus to the conclusion that the MSM allows the data generating process to be recovered in a more efficient way than the MCM. The MSM should therefore also lead to a better approximation of transition probabilities in the long-run.

The second assessment method used allows this very point to be assessed. Estimated parameters for both models were used to derive the corresponding 10-year TPMs, namely $\hat{\mathbf{\Pi}}^{(10)} = (\hat{\mathbf{\Pi}})^{10}$ for the MCM and $\hat{\mathbf{\Phi}}^{(10)} = \hat{\mathbf{S}} + (\mathbf{I} - \hat{\mathbf{S}})(\hat{\mathbf{M}})^{10}$ for the MSM. These estimated long-run matrices were then compared to the observed one, $\mathbf{P}^{(10)}$, which is derived from equation (4) and subsample #10. It appears from visual inspection of the three corresponding panels of table 5 that the MSM 10-year matrix more closely resembles the actually observed one than the MCM 10-year matrix. In particular, we find as expected that the diagonal elements of $\hat{\mathbf{\Pi}}^{(10)}$ largely underestimate those of $\mathbf{P}^{(10)}$ while those of $\hat{\mathbf{\Phi}}^{(10)}$ are much closer. This means that the MCM largely tends overestimate the mobility of farms in the long-run, with respect to the MSM. The AMEs reported in table 6 confirm the superiority of the MSM over the MCM in modeling the long-run transition process. The AME for the overall predicted 10-year TPM is around 0.95 for the MCM while it is about 0.78 for the MSM. This means that the MSM is about 17 percentage points closer to the observed TPM than the MCM in the long-run. However, the AMEs also confirm that the improvement comes mainly from the main diagonal elements: when only these are considered, the MSM performs about five times better than the MCM ($0.292/0.057 = 5.1$), while both models are almost comparable for off-diagonal elements, with the MCM this time performing slightly better ($0.657/0.724 = 0.9$).

Finally, table 5 also shows that the standard errors associated with the elements of the estimated matrices, computed using a standard bootstrapping method with 1,000 replications, are systematically higher with the MCM than with the MSM; the MSM estimator is thus also more

Table 5: Observed 10-Year TPM and Predicted 10-Year TPMs for both Models (subsample #10).

		ES class				
		0-50	50-100	100-150	150-250	≥ 250
ES class	0-50	0.715	0.235	0.029	0.014	0.007
	50-100	0.107	0.641	0.199	0.038	0.015
	100-150	0.020	0.146	0.536	0.268	0.030
	150-250	0.010	0.032	0.096	0.630	0.232
	≥ 250	0.005	0.021	0.020	0.124	0.830

a) Observed 10-year TPM ($\mathbf{P}^{(10)}$)

		ES class				
		0-50	50-100	100-150	150-250	≥ 250
ES class	0-50	**0.476**	**0.361**	**0.106**	**0.043**	**0.014**
		(0.028)	(0.020)	(0.008)	(0.006)	(0.004)
	50-100	**0.141**	**0.467**	**0.240**	**0.116**	**0.036**
		(0.011)	(0.015)	(0.011)	(0.007)	(0.004)
	100-150	**0.044**	**0.234**	**0.338**	**0.281**	**0.103**
		(0.004)	(0.011)	(0.013)	(0.011)	(0.007)
	150-250	**0.015**	**0.082**	**0.193**	**0.428**	**0.282**
		(0.002)	(0.005)	(0.010)	(0.013)	(0.013)
	≥ 250	**0.005**	**0.026**	**0.068**	**0.245**	**0.656**
		(0.001)	(0.003)	(0.005)	(0.013)	(0.018)

b) MCM predicted 10-year TPM ($\hat{\mathbf{\Pi}}^{(10)}$)

		ES class				
		0-50	50-100	100-150	150-250	≥ 250
ES class	0-50	**0.690**	**0.140**	**0.097**	**0.053**	**0.020**
		(0.017)	(0.010)	(0.007)	(0.005)	(0.003)
	50-100	**0.060**	**0.684**	**0.126**	**0.090**	**0.040**
		(0.007)	(0.010)	(0.007)	(0.005)	(0.003)
	100-150	**0.041**	**0.119**	**0.586**	**0.164**	**0.090**
		(0.004)	(0.007)	(0.012)	(0.009)	(0.006)
	150-250	**0.018**	**0.062**	**0.117**	**0.676**	**0.127**
		(0.002)	(0.004)	(0.006)	(0.010)	(0.009)
	≥ 250	**0.005**	**0.021**	**0.048**	**0.093**	**0.833**
		(0.001)	(0.002)	(0.003)	(0.005)	(0.008)

c) MSM predicted 10-year TPM ($\hat{\mathbf{\Phi}}^{(10)}$)

Note: estimated parameters in bold font, bootstrap standard errors in parentheses (1,000 replications).

Source: Agreste, RICA France 2000-2010 – authors' calculations

Table 6: Average Marginal Error (AME) between the Predicted 10-year TPMs ($\hat{\mathbf{\Pi}}^{(10)}$ and $\hat{\mathbf{\Phi}}^{(10)}$) and the Observed TPM ($\mathbf{P}^{(10)}$) (subsample #10)

Model	Overall matrix	Main diagonal elements	Off-diagonal elements
MCM	0.949	0.292	0.657
	(0.044)	(0.010)	(0.036)
MSM	0.781	0.057	0.724
	(0.034)	(0.007)	(0.035)

Note: bootstrap standard errors in parenthesss (1,000 replications).

Source: Agreste, RICA France 2000-2010 – authors' calculations

efficient (in the econometric sense) than the MCM one in recovering the underlying transition process.

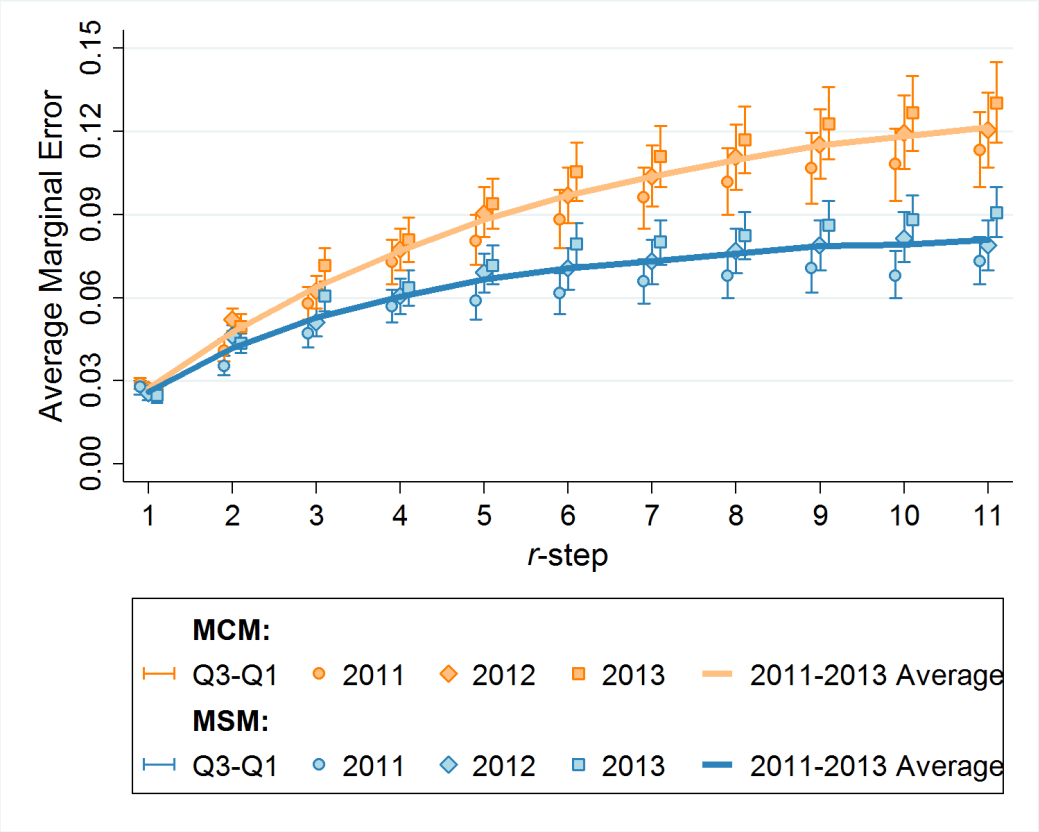
6.3 Out-of-sample projections

In-sample estimation results lead us to conclude that accounting for unobserved heterogeneity in the rate of movement of farms, as the MSM does, avoids overestimating their mobility across size categories. The MSM should therefore also lead to a more accurate prediction of the size distribution of farms in the long run, without hampering that in the short run.

To validate this point, we performed out-of-sample short- to long-run projections of the distribution of farm sizes using the parameters for both models as estimated with subsample #10. To do so, farm size distributions in 2011, 2012 and 2013 were predicted from a short- to long-run perspective, by applying the estimated r-step TPMs (for $1 \leq r \leq 11$) to the corresponding observed distributions from 2000 to 2012. In other words, distributions in 2011, 2012 and 2013 were predicted: by applying the estimated 1-year TPMs ($\hat{\mathbf{\Pi}}$ for the MCM and $\hat{\mathbf{\Phi}}$ for the MSM) to the observed distributions in 2010, 2011 and 2012, respectively; by applying the estimated 2-year TPMs ($\hat{\mathbf{\Pi}}^{(2)} = (\hat{\mathbf{\Pi}})^2$ for the MCM and $\hat{\mathbf{\Phi}}^{(2)} = \hat{\mathbf{S}} + (\mathbf{I} - \hat{\mathbf{S}})(\hat{\mathbf{M}})^2$ for the MSM) to the observed distributions in 2009, 2010 and 2011, respectively; and so forth. This process was continued by applying the estimated r-step TPMs ($\hat{\mathbf{\Pi}}^{(r)} = (\hat{\mathbf{\Pi}})^r$ for the MCM and $\hat{\mathbf{\Phi}}^{(r)} = \hat{\mathbf{S}} + (\mathbf{I} - \hat{\mathbf{S}})(\hat{\mathbf{M}})^r$ for the MSM) to the observed distributions in $(2011 - r)$, $(2012 - r)$ and $(2013 - r)$ and varying r up to eleven. Then, the resulting distributions for both models were compared to the actually observed distributions in 2011, 2012 and 2013 (see table 1). The corresponding AMEs reported in figure 2 summarize the results obtained for the 1,000 bootstrap replications.

Four conclusions can be drawn from figure 2. First, as expected, the accuracy of both models decreases when increasing the time horizon of projection: the computed AMEs are significantly smaller in the short run than they are in the medium and long run for both models. Second,

Figure 2: Average marginal error (AME) between the out-of-sample projections of farm size distributions and the actually observed ones for both models.



Note: see text for an explanation on how short- to long-run projections were obtained; interquartile ranges (Q3-Q1) obtained from the 1,000 bootstrap replications.

Source: Agreste, RICA France 2000-2013 – authors' calculations

both models almost are almost comparable in the short run, confirming that adopting the MSM modeling framework does not degrade year-on-year forecasts with respect to the MCM. Third, the MSM performs significantly better than the MCM in both the medium and long run. For example, the average AME for the MSM (0.068) is 1.3 times lower than that of the MCM (0.088) for 5-year interval projections, and 1.5 times lower for 11-year interval projections (0.082 for the MSM compared with 0.121 for the MCM). Fourth, figure 2 also shows that the accuracy and robustness of farm size distribution predictions decrease more rapidly for the MCM than for the MSM when increasing the time horizon of projections, since the AMEs as well as the interquartile ranges of the 1,000 bootstrap replications increase more rapidly for the MCM than for the MSM.

7 Concluding remarks

The modeling framework implemented here, namely the mover-stayer model (MSM), is more general than the simple Markov chain model (MCM) since it accounts for unobserved heterogeneity in the rate of movement of farms. Within this extended framework, the 1-year transition probability matrix is decomposed into a fraction of 'stayers' who remain in their initial size category and a fraction of 'movers' who follow a standard Markovian process. To estimate the model, we improved Blumen, Kogan, and McCarthy (1955)'s calibration method by using the elaborate expectation-maximization (EM algorithm) estimation method proposed by Frydman (2005) in order to account for incomplete information. Finally, we computed standard errors for long-run matrices and farm size distributions using a bootstrap method, with 1,000 replications of each estimation and projection. This allowed the models to be compared in a statistically robust manner, which adds to the existing literature. Our results show that, with respect to the MCM, accounting for unobserved farm heterogeneity enables closer estimates of both the observed transition matrix and the distribution of farms across size categories in the long-run to be derived, without degrading any short-run analysis. This result is consistent with the findings in other strands of the economic literature, namely that, by relaxing the assumption of homogeneity in the transition process which is the basis of the MCM, the MSM leads to a better representation of the underlying structural change process.

However, this modeling framework remains quite a restricted and simplified version of the more general model, the mixed-Markov chain model (M-MCM), which we presented as an introduction to the MSM. Extending the MSM further could therefore lead to even more economically sound, as well as statistically more accurate, models for the farming sector. We briefly mention some of such extensions which we think are promising. Firstly, more heterogeneity across farms could be incorporated by allowing for more than two unobserved types. For example, accounting for different types of movers could yield a better representation of the structural change process in the farming sector by allowing farms which tend mainly to enlarge to be disentangled from farms which tend mainly to shrink. Secondly, the quite strong assumption of a 'pure stayer' type could be relaxed because it may look unlikely that some farms 'never move at all', *i.e.*, will not change size category over their entire lifespan. In this respect, the robustness of the MSM to the number of years during which farms are observed could be investigated. Indeed, considering eleven years to perform in-sample estimations, our results show that movers may stay five to eight years in a given category before experiencing a large enough (positive or negative) size change to reach another category. We can then infer that the shorter the observation period, the higher the number of farms which would be inappropriately considered as 'pure stayers'. From a methodological point of view, this leads to the conclusion that long enough panel data needs to be available if the MSM is to be empirically implemented. This could be seen as a shortcoming of the MSM with respect to the MCM but it is in fact consistent with farm structural change being a long-run and slow process.

Our final recommendation is that the proposed modeling framework should be extended to account for entries and exits and that a non-stationary version of the M-MCM model should be developed. Since this should allow the transition process to be recovered in an even more efficient way, it would surely prove to be very insightful for analyzing the factors which drive structural change in the farming sector, including agricultural policies, not only from a size distribution perspective, but also as regards the evolution of farm numbers.

Acknowledgments

Legrand D.F. Saint-Cyr benefits from a research grant from Crédit Agricole en Bretagne in the framework of the chair "'Enterprises and Agricultural Economics"' created in partnership with Agrocampus Ouest.

References

Anderson, T.W., and L.A. Goodman. 1957. "Statistical Inference about Markov Chains." *Annals of Mathematical Statistics* 28:89–110.

Blumen, I., M. Kogan, and P.J. McCarthy. 1955. *The industrial mobility of labor as a probability process*, vol. VI. Cornell Studies in Industrial and Labor Relations.

Butault, J.P., and N. Delame. 2005. "Concentration de la production agricole et croissance des exploitations." *Economie et statistique* 390:47–64.

Cipollini, F., C. Ferretti, and P. Ganugi. 2012. "Firm size dynamics in an industrial district: The mover-stayer model in action." In A. Di Ciaccio, M. Coli, and J. M. Angulo Ibanez, eds. *Advanced Statistical Methods for the Analysis of Large Data-Sets*. Springer Berlin Heidelberg, pp. 443–452.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society* 39:1–38.

Desriers, M. 2011. "Farm structure. Agricultural census 2010. Production is concentrated in specialised farms." Agreste: la statistique agricole, Agreste Primeur 272.

Dutta, J., J.A. Sefton, and M.R. Weale. 2001. "Income distribution and income dynamics in the United Kingdom." *Journal of Applied Econometrics* 16:599–617.

Efron, B. 1979. "Bootstrap methods: Another look at the jackknife." *The Annals of Statistics* 7:1–26.

Efron, B., and R. Tibshirani. 1986. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." *Statistical Science* 1:54–75.

European Commission. 2010. *Farm Accounting Data Network. An A to Z of methodology*. Brussels (Belgium): DG Agri.

Fougère, D., and T. Kamionka. 2003. "Bayesian inference for the mover-stayer model in continuous time with an application to labour market transition data." *Journal of Applied Econometrics* 18:697–723.

Frydman, H. 2005. "Estimation in the mixture of Markov chains moving with different speeds." *Journal of the American Statistical Association* 100:1046–1053.

—. 1984. "Maximum likelihood estimation in the mover-stayer model." *Journal of the American Statistical Association* 79:632–638.

Frydman, H., and A. Kadam. 2004. "Estimation in the continuous time mover-stayer model with an application to bond ratings migration." *Applied Stochastic Models in Business and Industry* 20:155–170.

Frydman, H., J.G. Kallberg, and D.L. Kao. 1985. "Testing the adequacy of Markov chain and mover-stayer models as representations of credit behavior." *Operations Research* 33:1203–1214.

Frydman, H., and T. Schuermann. 2008. "Credit rating dynamics and Markov mixture models." *Journal of Banking and Finance* 32:1062–1075.

Fuchs, C., and J.B. Greenhouse. 1988. "The EM algorithm for maximum likelihood estimation in the mover-stayer model." *Biometrics* 44:605–613.

Goodman, L.A. 1961. "Statistical methods for the mover-stayer model." *Journal of the American Statistical Association* 56:841–868.

Hallberg, M.C. 1969. "Projecting the size distribution of agricultural firms. An application of a Markov process with non-stationary transition probabilities." *American Journal of Agricultural Economics* 51:289–302.

Huettel, S., and R. Jongeneel. 2011. "How has the EU milk quota affected patterns of herd-size change?" *European Review of Agricultural Economics* 38:497–527.

Jafry, Y., and T. Schuermann. 2004. "Measurement, estimation and comparison of credit migration matrices." *Journal of Banking & Finance* 28:2603–2639.

Karantininis, K. 2002. "Information-based estimators for the non-stationary transition probability matrix: An application to the Danish pork industry." *Journal of Econometrics* 107:275–290.

Langeheine, R., and F. Van de Pol. 2002. *Applied latent class analysis*, Cambridge University Press, chap. Latent markov chains. pp. 304–341.

Lee, T., G. Judge, and A. Zellner. 1977. *Estimating the parameters of the Markov probability model from aggregate time series data*. Amsterdam: North Holland.

Lee, T.C., G.G. Judge, and T. Takayama. 1965. "On estimating the transition probabilities of a Markov process." *Journal of Farm Economics* 47:742–762.

Louis, T.A. 1982. "Finding the observed information matrix when using the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 44:226–233.

MacRae, E.C. 1977. "Estimation of time-varying Markov processes with aggregate data." *Econometrica* 45:183–198.

McLachlan, G., and T. Krishnan. 2007. *The EM algorithm and extensions*, vol. 382. John Wiley & Sons.

Morgan, T.M., C.S. Aneshensel, and V.A. Clark. 1983. "Parameter estimation for mover-stayer models analyzing depression over time." *Sociological Methods & Research* 11:345–366.

Piet, L. 2011. "Assessing structural change in agriculture with a parametric Markov chain model. Illustrative applications to EU-15 and the USA." Paper presented at the XIIIth Congress of the European Association of Agricultural Economist, Zurich (Switzerland).

Spilerman, S. 1972. "The analysis of mobility processes by the introduction of independent variables into a Markov chain." *American Sociological Review* 37:277–294.

Stavins, R.N., and B.F. Stanton. 1980. *Alternative procedures for estimating the size distribution of farms*. Department of Agricultural Economics, New York State College of Agriculture and Life Sciences.

Storm, H., T. Heckelei, and R.C. Mittelhammer. 2011. "Bayesian estimation of non-stationary Markov models combining micro and macro data." Discussion Paper No. 2011:2, University of Bonn, Institute for Food and Resource Economics, Bonn (Germany).

Weiss, C.R. 1999. "Farm growth and survival: econometric evidence for individual farms in upper Austria." *American Journal of Agricultural Economics* 81:103–116.

Zepeda, L. 1995. "Asymmetry and nonstationarity in the farm size distribution of Wisconsin milk producers: An aggregate analysis." *American Journal of Agricultural Economics* 77:837–852.

Zimmermann, A., and T. Heckelei. 2012. "Structural change of European dairy farms: A cross-regional analysis." *Journal of Agricultural Economics* 63:576–603.

Zimmermann, A., T. Heckelei, and I.P. Dominguez. 2009. "Modelling farm structural change for integrated ex-ante assessment: Review of methods and determinants." *Environmental Science and Policy* 12:601–618.

A Appendix

A.1 Frydman (2005)'s specification of the mixed Markov chain model (M-MCM)

Recall that the general form of the M-MCM is given by equation (6):

$$\mathbf{\Phi} = \{\phi_{ij}\} = \sum_{g=1}^{G} \mathbf{S}_g \mathbf{M}_g \tag{19}$$

where $\mathbf{M}_g = \{m_{ij,g}\}$ is the TPM defining the 1-step Markov process followed by type-g agents, and $\mathbf{S}_g = \text{diag}(s_{i,g})$ is a diagonal matrix which gathers the shares of type-g agents in each category.

Assuming that all type g TPMs are related to a specific one, arbitrarily chosen as that of the last type G, Frydman (2005) gives the TPM of any type g as:

$$\mathbf{M}_g = \mathbf{I} - \mathbf{\Lambda}_g + \mathbf{\Lambda}_g \mathbf{M} \quad \text{for} \quad 1 \leq g \leq G-1 \tag{20}$$

where $\mathbf{I}$ is the $J \times J$ identity matrix, $\mathbf{\Lambda}_g = \text{diag}(\lambda_{i,g})$ and $\mathbf{M} = \mathbf{M}_G$ (*i.e.*, $\mathbf{\Lambda}_G = \mathbf{I}$), subject to $0 \leq \lambda_{i,g} \leq \frac{1}{1-m_{ii}}$ $(\forall i \in J)$ and $0 \leq m_{ii} \leq 1$, where m_{ii} are the main diagonal elements of matrix $\mathbf{M}$.

Within this specification, the $\lambda_{i,g}$ parameters give information about heterogeneity in the rates of movement across homogeneous agent types:

- $\lambda_{i,g} = 0$ if type-g agents originally in category i never move out of i;
- $0 < \lambda_{i,g} < 1$ if type-g agents originally in category i move at a lower rate than the generator matrix $\mathbf{M}$;
- $\lambda_{i,g} > 1$ if type-g agents originally in category i move at a higher rate than the generator matrix $\mathbf{M}$.

Then the expected time spent in category i by type-g agents is given by $\frac{1}{(\lambda_{i,g}(1-m_{ii}))}$ $(\forall \lambda_{i,g} > 0)$.

A.2 Maximum likelihood estimation of the mixed Markov chain model (M-MCM)

Consider that each agent k is observed at some discrete time points in the time interval $[0, T_k]$ with $T_k \leq T$, where T is the time horizon of all observations. According to Anderson and Goodman (1957), the likelihood that the transition sequence of agent k (X_k) was generated by the specific type-g Markov chain (*i.e.*, that k belongs to type g), conditional on knowing that k was initially in state i, is given by:

$$l_{k,g} = s_{i,g} \prod_{i \neq j} (m_{ij,g})^{\nu_{ij,k}} \prod_{i} (m_{ii,g})^{\nu_{ii,k}} \tag{21}$$

where $s_{i,g}$ is the share of type-g agents initially in category i, $\nu_{ij,k}$ is the number of transitions from i to j made by agent k (with $j \neq i$), $\nu_{ii,k}$ is the total time spent by k in category i, and $m_{ii,g}$ and $m_{ij,g}$ are the elements of $\mathbf{M}_g$.
Under Frydman (2005)'s specification of the M-MCM as defined by equation (20) the above likelihood can be rewritten as:

$$l_{k,g} = s_{i,g} \prod_{i \neq j} (\lambda_{i,g} m_{ij})^{\nu_{ij,k}} \prod_{i} (1 - \lambda_{i,g} + \lambda_{i,g} m_{ii})^{\nu_{ii,k}} \tag{22}$$

where $\lambda_{i,g}$ is the relative rate of movement of type-g agents.
Then, the log-likelihood function for the whole population is:

$$\log L = \sum_{k=1}^{N} \sum_{g=1}^{G} (Y_{k,g} \log l_{k,g}) \tag{23}$$

where $Y_{k,g}$ is an indicator variable which equals 1 if agent k belongs to type g and 0 otherwise. Note that the log-likelihoods of the MCM and MSM can easily be derived from equation (23) by stating, respectively, G=1 and $\lambda_{i,1} = 1$ for the MCM and G=2, $\lambda_{i,1} = 0$ and $\lambda_{i,2} = 1$ for the MSM.
The maximum likelihood estimators, $\hat{s}_{i,g}$, $\hat{\lambda}_{i,g}$, $\hat{m}_{ii}$ and $\hat{m}_{ij}$, are obtained using the EM algorithm in a similar way to that described in the main text.

A.3 Computing standard errors from EM algorithm equations

Two components are required to compute standard errors from the EM algorithm equations (Louis, 1982): (i) the observed information matrix given by the negative of the Hessian matrix of the log-likelihood function and; (ii) the missing information matrix obtained from the gradient vector, that is, the vector of score statistics based on complete information. Since the log-likelihood function given by equation (12) is twice differentiable with respect to the model parameters, the standard errors can be computed as follows.

Let $\mathbf{\Omega}_c(\mathbf{z}; \hat{s}_i, \hat{m}_{ij})$ and $\mathbf{\Omega}_m(\mathbf{z}; \hat{s}_i, \hat{m}_{ij})$ $(i, j = 1, ..., J)$ be the observed $(d \times d)$ information matrices in terms of complete and missing information, respectively, where $Z_k = (X_k, Y_k)$ gathers the transition sequence X_k of agent k and the unobserved agent's type dummy variable Y_k, and d is the number of estimated parameters. The observed information matrix in terms of incomplete information can then be derived as:

$$\mathbf{\Omega}(\mathbf{x}; \hat{s}_i, \hat{m}_{ij}) = \mathbf{\Omega}_c(\mathbf{z}; \hat{s}_i, \hat{m}_{ij}) - \mathbf{\Omega}_m(\mathbf{z}; \hat{s}_i, \hat{m}_{ij}) \tag{24}$$

where $\mathbf{\Omega}_m(\mathbf{z}; \hat{s}_i, \hat{m}_{ij})$ is given by:

$$\mathbf{\Omega}_m(\mathbf{z}; \hat{s}_i, \hat{m}_{ij}) = E[\mathbf{s_c}(\mathbf{z}; \hat{s}_i, \hat{m}_{ij}) \times \mathbf{s_c}(\mathbf{z}; \hat{s}_i, \hat{m}_{ij})'] \tag{25}$$

where $\mathbf{s_c}(\mathbf{z}; \hat{s}_i, \hat{m}_{ij})$ is the vector of score statistics in terms of complete information.
Therefore, if the observed information matrix in terms of incomplete information just described, $\mathbf{\Omega}(\mathbf{x}; \hat{s}_i, \hat{m}_{ij})$, is invertible, the standard errors are given by:

$$\mathbf{se} = \{\sqrt{\psi_{ll}}\} \tag{26}$$

where $\mathbf{se}$ is the $1 \times d$ vector of standard errors, $\mathbf{\Psi} = \{\psi_{ll'}\} = \mathbf{\Omega}^{-1}(\mathbf{x}; \hat{s}_i, \hat{m}_{ij})$ is defined as the asymptotic covariance matrix of the maximum likelihood estimators $\hat{s}_i$ and $\hat{m}_{ij}$ $(\forall i, j = 1, ..., J)$ under incomplete information, and $l, l' = 1, ..., d$ (McLachlan and Krishnan, 2007).

Les **Working Papers SMART – LERECO** sont produits par l'UMR SMART et l'UR LERECO

- **UMR SMART**

L'Unité Mixte de Recherche (UMR 1302) *Structures et Marchés Agricoles, Ressources et Territoires* comprend l'unité de recherche d'Economie et Sociologie Rurales de l'INRA de Rennes et les membres de l'UP Rennes du département d'Economie Gestion Société d'Agrocampus Ouest.

Adresse :
UMR SMART - INRA, 4 allée Bobierre, CS 61103, 35011 Rennes cedex
UMR SMART - Agrocampus, 65 rue de Saint Brieuc, CS 84215, 35042 Rennes cedex

- **LERECO**

Unité de Recherche *Laboratoire d'Etudes et de Recherches en Economie*
Adresse :
LERECO, INRA, Rue de la Géraudière, BP 71627 44316 Nantes Cedex 03

Site internet commun : http://www.rennes.inra.fr/smart

Liste complète des Working Papers SMART – LERECO :
http://www.rennes.inra.fr/smart/Working-Papers-Smart-Lereco
http://ideas.repec.org/s/rae/wpaper.html

The **Working Papers SMART – LERECO** are produced by UMR SMART and UR LERECO

- **UMR SMART**

The « Mixed Unit of Research » (UMR1302) *Structures and Markets in Agriculture, Resources and Territories*, is composed of the research unit of Rural Economics and Sociology of INRA Rennes and of the members of the Agrocampus Ouest's Department of Economics Management Society who are located in Rennes.

Address:
UMR SMART - INRA, 4 allée Bobierre, CS 61103, 35011 Rennes cedex, France
UMR SMART - Agrocampus, 65 rue de Saint Brieuc, CS 84215, 35042 Rennes cedex, France

- **LERECO**

Research Unit *Economic Studies and Research Lab*
Address:
LERECO, INRA, Rue de la Géraudière, BP 71627 44316 Nantes Cedex 03, France

Common website: http://www.rennes.inra.fr/smart_eng/

Full list of the Working Papers SMART – LERECO:
http://www.rennes.inra.fr/smart_eng/Working-Papers-Smart-Lereco
http://ideas.repec.org/s/rae/wpaper.html

Contact

Working Papers SMART – LERECO
INRA, UMR SMART
4 allée Adolphe Bobierre, CS 61103
35011 Rennes cedex, France
Email : smart_lereco_wp@rennes.inra.fr

2015

Working Papers SMART – LERECO

UMR INRA-Agrocampus Ouest **SMART** (Structures et Marchés Agricoles, Ressources et Territoires)

UR INRA **LERECO** (Laboratoire d'Etudes et de Recherches en Economie)

Rennes, France