Movers and Stayers in the Farming Sector: Another Look at Heterogeneity in Structural Change

Legrand D. F. SAINT-CYR[1], Laurent PIET[2]

[1] Agrocampus Ouest, UMR1302 SMART, F-35000 Rennes

[2] INRA, UMR1302 SMART, F-35000 Rennes

Paper prepared for presentation at the EAAE 2014 Congress
'Agri-Food and Rural Innovations for Healthier Societies'

August 26 to 29, 2014
Ljubljana, Slovenia

Abstract
The Markov chain model (MCM) has become a popular tool in the agricultural economics literature to study the impact of various drivers on the structural change of farms, including public support. In order to relax the process-homogeneity assumption underlying the MCM, we consider a mixture of two types of agents, the 'stayers' who always remain in their initial size category, and the 'movers' who follow a first-order Markovian process. An empirical application to a panel of commercial French farms over 2000-2012 shows that the mover-stayer model (MSM) is a better modeling framework to recover the underlying transition probability matrix.

Keywords
Structural change, Markov chain model, Mover Stayer model, Mixture model, EM algorithm

1. Introduction

As Zimmermann *et al.* (2009) show, it has become quite common in the agricultural economic literature to study the way farms experience structural change thanks to the so-called Markov chain model (MCM). Basically, this model states that, as the size of farms changes according to some stochastic process, farms move from one size category to another over time. This modeling framework has been particularly used in its non-stationary version, where transition probabilities across categories may vary with time, in order to study the impact of a variety of factors, including agricultural policies.

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 most often easier to obtain than individual-level data, and Lee *et al.* (1965) and Lee *et al.* (1977) have shown that robustly estimating a MCM from aggregate data is possible. Since then, because estimating a MCM may well be an ill-posed problem as 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 such 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 *et al.*, 2011).

However, even though some of these studies have accounted for heterogeneity across farms by considering transition probabilities covariates depicting farmer and/or farm characteristics (see Zimmermann and Heckelei (2012) for a recent example), to our knowledge, none of these studies has questioned so far the assumption of *process*-homogeneity which underlies the traditional Markov modeling framework: all of these studies define only one transition probability matrix for the whole population under study, implying that all agents follow the same and unique stochastic process. As farm-level data become more widely available, allowing for the observation of individual transitions across time, we argue that this homogeneity assumption should be relaxed. To this end, we propose to use a more general modeling framework than the MCM, namely the mixed MCM (M-MCM). As an illustration, we apply the simplest version of this extended model, the mover-stayer model (MSM), to compute the short- and long-run transition probability matrices for an unbalanced panel of 14,298 commercial French farms observed over 2000-2012.

The paper is structured as follows. Section 2 introduces how the traditional MCM can be generalized into the M-MCM. Section 3 develops the specific MSM specification along with the method used to estimate the model. Section 4 reports our application to France, first describing the data used and then presenting the results. Finally, section 5 concludes with some considerations on how to extend further the approach described here.

2. Generalizing the Markov chain model

2.1. Transition probability matrices

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

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

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

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

In the following, we restrict our analysis to the stationary case where the r-step transition probability matrix (TPM), $\mathbb{P}_t^{(r)} = \{\phi_{ij,t}^{(r)}\}$, is independent from t, *i.e.*, $\mathbb{P}_t^{(r)} = \mathbb{P}^{(r)}$ for all t. In matrix notation, equation (1) then rewrites:

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

where $\mathbf{N}_{t+r} = \{n_{j,t+r}\}$ and $\mathbf{N}_t = \{n_{j,t}\}$ are row vectors.

2.2. Markov versus mixed-Markov models

The traditional MCM approach consists in approximating $\mathbb{P}^{(r)}$ by the 1-step transition matrix $\mathbb{P}^{(1)} \equiv \mathbf{\Pi} = \{\pi_{ij}\}$ raised to the power r. The econometric model which has to be estimated thus writes:

$$\mathbf{N}_{t+r} = \mathbf{N}_t \times \mathbf{\Pi}^r + \mathbf{V}_{t+r}, \tag{4}$$

where $\mathbf{V}_{t+r} = \{v_{j,t+r}\}$ is a row vector of error terms assumed independently and identically distributed (*iid*).

In doing so, the MCM approach assumes that the individuals in the population are homogeneous, *i.e.*, they all move according to the same stochastic process described by $\mathbf{\Pi}$. However, in general, $\mathbf{\Pi}^r$ proves to be a poor estimate of $\mathbb{P}^{(r)}$ (Blumen *et al.*, 1955; Spilerman, 1972). In particular, the diagonal elements of $\mathbf{\Pi}^r$ largely underestimate those of $\mathbb{P}^{(r)}$. With the notation that $\mathbf{\Pi}^{(r)} = \mathbf{\Pi}^r$, this means that, in general, $\pi_{ii}^{(r)} \ll \phi_{ii}^{(r)}$. One way to obtain a 1-step TPM which leads to a more consistent r-step estimate, consists in relaxing the process-homogeneity assumption underlying the MCM approach. This leads to considering a mixture of time homogeneous Markov chains which captures population heterogeneity in the rate of movement among state (Frydman, 2005).

Considering that agents may follow a discrete number G of elementary Markov processes instead of just one, the general form of the mixed Markov chain model (M-MCM) consists in decomposing $\mathbb{P}^{(1)} \equiv \mathbf{P} = \{p_{ij}\}$ as:

$$\mathbf{P} = \sum_{g=1}^{G} \mathbf{S}_g \mathbf{M}_g, \tag{5}$$

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 state of nature. Since every agent in the population has to belong to one and only one type g, the constraint that $\sum_{g=1}^{G} \mathbf{S}_g = \mathbf{I}_J$ must hold, where $\mathbf{I}_J$ is the $J \times J$ identity matrix.

Under this model, equation (4) rewrites:

$$\mathbf{N}_{t+r} = \mathbf{N}_t \times \mathbf{P}^{(r)} + \mathbf{U}_{t+r} = \mathbf{N}_t \times \sum_{g=1}^{G} \mathbf{S}_g \mathbf{M}_g^r + \mathbf{U}_{t+r}, \tag{6}$$

where $\mathbf{U}_{t+r} = \{u_{j,t+r}\}$ is a row vector of *iid* error terms.

With the so-defined MCM and M-MCM modeling frameworks, it should be noted that $\mathbb{P}^{(1)} = \mathbf{\Pi} = \mathbf{P}$ but $\mathbf{\Pi}^{(r)} \neq \mathbf{P}^{(r)}$ in general, and that the M-MCM reduces to the MCM if $G = 1$.

2.3. Continuous time models

According to the structural change under study, the transition process characterizing each homogeneous type of agents can be regarded as discrete or continuous with respect to time. While several authors have used a discrete-time approach (Blumen *et al.*, 1955; Spilerman, 1972; Frydman *et al.*, 1985), a continuous-time approach is preferable if transitions may occur at any time (Lando and Skodeberg, 2002; Frydman and Kadam, 2004; Frydman and Schuermann, 2008). In this case, following Singer and Spilerman (1975), the type-g TPM is given at any time t by:

$$\mathbf{M}_g(t) = \exp(t\mathbf{Q}_g), \tag{7}$$

where $\mathbf{Q}_g = \{q_{ij,g}\}$ represents the generator matrix of the Markovian process followed by type-g agents, defined as:

- $\exp(t\mathbf{Q}_g) = \sum_{k=0}^{\infty} \frac{t^k \mathbf{Q}_g^k}{k!}$,
- $q_{ij,g} \geq 0$ for $i \neq j \quad \forall g$,
- and, by convention, $q_{ii,g} \equiv -\sum_{j \neq i} q_{ij,g} = -q_{i,g} \leq 0 \quad \forall g$.

With the generator matrix $\mathbf{Q}_g$ so defined, it is worth noting that q_{ij}/q_i is the probability that an agent in state i moves to state j, given the occurrence of a transition, and that $1/q_i$ is the expected total time an agent spends in state i.

In the agricultural economics literature, farm structural change has been so far studied mostly using the discrete-time approach. So doing, an arbitrary time interval, generally one year, is chosen to estimate the unitary, *i.e.*, annual (or 1-year) transition probability matrix which governs the process. However, even if farm sizes are observed only once a year in the best case, farms may change their size at any time during the year. Furthermore, as pointed out by Singer and Spilerman (1976), using different unitary time intervals may lead to different results. Thus, the continuous-time approach has been preferred here.

2.4. Frydman (2005)'s specification of the M-MCM

As the number of parameters to estimate increases with the number of homogeneous agent types, the estimation of equation (6) may become difficult because of an identification issue. Thus, Frydman (2005) proposed a parameterization of the M-MCM under a continuous-time approach, assuming that all type-g TPMs are related to a specific one:

$$\mathbf{Q}_g \equiv \mathbf{\Lambda}_g \mathbf{Q} \quad \forall g, \tag{8}$$

where $\mathbf{\Lambda}_g = \text{diag}(\lambda_{i,g})$ with $\lambda_{i,g} \geq 0$.

The $\lambda_{i,g}$ parameters inform about differences in the rates of movement across homogeneous agent types: $\lambda_{i,g} = 0$ if type-g agents starting in state i never move out of i; $0 < \lambda_{i,g} \leq 1$ if they move at a lower rate than the generator matrix $\mathbf{Q}$ and; $\lambda_{i,g} > 1$ if they move at a higher rate than the generator matrix $\mathbf{Q}$. The generator matrix $\mathbf{Q}$ is chosen arbitrarily as the intensity matrix for the last homogeneous agent type ($\mathbf{Q} \equiv \mathbf{Q}_G$), *i.e.*, $\mathbf{\Lambda}_G = \mathbf{I}_J$.

3. The model used

3.1. The Mover-Stayer model

In this paper, we stick to the simplest version of the M-MCM, namely the mover-stayer model (MSM) first proposed by Blumen *et al.* (1955). In this restricted approach, only two types of homogeneous agents are considered, those who always remain in the same category (the 'stayers') and those who follow a first-order Markovian process (the 'movers'). Formally, this leads to rewriting equation (5) in a simpler form as:

$$\mathbf{P}(t) = \mathbf{S} + (\mathbf{I}_J - \mathbf{S})\,\mathbf{M}(t). \tag{9}$$

With respect to the general formulation (5), this corresponds to setting $G = 2$ and defining $\mathbf{S}_1 \equiv \mathbf{S}$ and $\mathbf{M}_1 = \mathbf{I}_J$ for the stayers, and $\mathbf{S}_2 = (\mathbf{I}_J - \mathbf{S})$ and $\mathbf{M}_2 \equiv \mathbf{M}$ for the movers. With respect to Frydman (2005)'s specification of equation (8), this is equivalent to imposing $\mathbf{\Lambda}_1 = \mathbf{0}_J$ for stayers (where $\mathbf{0}_J$ is the $J \times J$ matrix with all elements set to zero), and $\mathbf{\Lambda}_2 = \mathbf{I}_J$ and $\mathbf{Q}_2 \equiv \mathbf{Q}$ for movers.

3.2. Estimation under complete information

Since Goodman (1961) has shown that Blumen *et al.* (1955) estimators for the MSM are biased, alternative methods have been developed to obtain consistent ones using maximum likelihood (Frydman, 1984, 2005) or Bayesian inference (Fougère and Kamionka, 2003). Based on the findings of Frydman (1984) and using the general formulation of equation (5) and the relation established in (8), Frydman (2005) has developed a maximum likelihood method to estimate the parameters of the M-MCM. We report this strategy, using our own notations introduced above.

Consider a population of n agents, each k of which being observed continuously on some time interval $[0, T_k]$ with $T_k \leq T$, where T the time horizon of all observations. According to Frydman and Kadam (2004) and under Frydman (2005)'s specification of the M-MCM as defined by equation (8), the likelihood that the transition history of agent k was generated by a specific Markov chain with the generator matrix $\mathbf{Q}_g$ (*i.e.*, that k belongs to type g), conditional on knowing that k was initially in state i_k, is given by:

$$l_{k,g} = s_{i_k,g} \prod_{i \neq j} (\lambda_{i,g} q_{ij})^{n_{ij,k}} \prod_{i} \exp(-\lambda_{i,g} q_i \tau_{i,k}), \tag{10}$$

where $s_{i_k,g}$ is the share of type-g agents initially in state i_k, $n_{ij,k}$ is the number of times k made a transition from i to j with $j \neq i$, and $\tau_{i,k}$ is the total time spent by k in state i (with $\tau_{i,k} \leq T_k$).

Under the MSM framework where only two type of agents are considered ('S' standing for stayers and 'M' for movers), the log-likelihood function for the whole population then writes:

$$\log L = \sum_{k=1}^{n} (Y_{k,S} \log l_{k,S} + Y_{k,M} \log l_{k,M}), \tag{11}$$

where $Y_{k,g}$ is an indicator variable which equals 1 if agent k belongs to g and 0 otherwise (with $g = \{S, M\}$).

Under complete information, all $Y_{k,g}$ are perfectly known so equation (11) rewrites:

$$\log L = \sum_{i} b_i \log(1 - s_i) + \sum_{i} b_{i,S} \log[s_i/(1 - s_i)] + \sum_{i \neq j} n_{ij} \log(q_{ij}) - \sum_{i} q_i \tau_i + \sum_{i} q_i \tau_{i,S}, \tag{12}$$

where s_i is state-i share of stayers, q_{ij} and q_i are the elements of the generator matrix $\mathbf{Q}$ of movers as defined in section 2.3, b_i is the total number of agents who were initially in state i, $b_{i,S}$ is the total number of stayers who were initially in state i, n_{ij} is the total number of transitions from state i to state j, τ_i is the total time spent in state i by all agents and $\tau_{i,S}$ is the total time spent in state i by stayers.

Then, maximizing equation (12) with respect to its unknown parameters s_i, q_{ij} and q_i leads to the following estimators:

$$\hat{s}_i = \frac{b_{i,S}}{b_i}, \quad \hat{q}_i = \frac{n_i}{\tau_{i,M}} \quad \text{and} \quad \hat{q}_{ij} = \frac{n_{ij}}{n_i} \hat{q}_i, \tag{13}$$

where n_i is the total number of transitions out of state i and $\tau_{i,M}$ is the total time spent in state i by movers (with $\tau_i = \tau_{i,S} + \tau_{i,M}$).

3.3. Estimation under incomplete information

Swensen (1996) has shown that equation (11) is actually difficult to use directly because it is unlikely that we know beforehand which agents are stayers and which are movers. This would require that we observed each agent k during a sufficiently long period T_k to reach complete information on their status.

Alternatively, Fuchs and Greenhouse (1988) and van de Pol and Langeheine (1989) suggested that the MSM parameters can be estimated using the Expectation-Maximization

(EM) algorithm developed by Dempster *et al.* (1977). Following Frydman and Kadam (2004), the EM algorithm in our case consists of the four following steps:

(i) Initialization: Arbitrarily choose initial values s_i^0 for the share of stayers and q_i^0 for the diagonal entries of the generator matrix $\mathbf{Q}$ of movers.

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

$$E^p(Y_{k,S}) = \frac{s_i^p}{s_i^p + (1 - s_i^p)\exp(-q_i^p \tau_{i,k})}.$$

Then compute:

$$E^p(b_{i,S}) = \sum_{k=1}^{n} E^p(Y_{k,S}), \quad E^p(\tau_{i,S}) = \sum_{k=1}^{n} E^p(Y_{k,S})\tau_{k,i} \quad \text{and} \quad E^p(\tau_{i,M}) = \tau_i - E^p(\tau_{i,S}).$$

(iii) Maximization: Update s_i^p and q_i^p as follows:

$$s_i^{p+1} = \frac{E^p(b_{i,S})}{b_i} \quad \text{and} \quad q_i^{p+1} = \frac{n_i}{E^p(\tau_{i,M})}.$$

(iv) Iteration: Return to step (ii) using s_i^{p+1} and q_i^{p+1} and iterate until convergence. When convergence is reached, $\hat{s}_i^*$ and $\hat{q}_i^*$ so obtained are considered as the optimal estimators, and $\hat{q}_{ij}^*$ derives from $\hat{q}_i^*$ as in equation (13).

4. Empirical application

4.1. Data used

When empirically applying the modeling frameworks presented above to the agricultural sector, 'agents' are usually farms and 'states of nature' are defined with respect to some size variable.

We applied the standard Markov (MCM) and the Mover-Stayer models (MSM) to the French strand of the EU-wide Farm Accounting Data Network (FADN) database.[1] Individual farm level data were available from 2000 to 2012 for the full sample surveyed, *i.e.*, around 7,000 farms each year. Since the FADN database is a rotating panel, farms which enter (respectively, leave) the sample a given year cannot be considered as representing actual entries into (exits from) the agricultural sector.[2] Therefore, we chose to work on size change of on-going farms, *i.e.*, without considering entries nor exits. In order to observe at least one transition for each agent, we kept only farms present during at least two consecutive years in the database. Our unbalanced panel thus counted 14,298 farms, that is 87.64% of the full sample.

[1]The French FADN is called 'Réseau d'Information Comptable Agricole' (RICA) and is produced and disseminated by the statistical and foresight service of the French ministry for agriculture. To learn more about RICA, see `http://www.agreste.agriculture.gouv.fr/`. To learn more about FADN in general, see `http://ec.europa.eu/agriculture/rica/index.cfm`.

[2]In this respect, around 10% of the French FADN sample is renewed each year.

Table 1. Distribution by economic size (ES) class and average ES for the studied sample.[a]

Years	Number of farms by ES class					Total	Average ES
	(0-50)	(50-100)	(100-150)	(150-250)	(+250)		(std. dev.)
2000	682	1,909	1,383	1,543	1,170	6,687	169.88 (183.56)
2001	730	2,147	1,571	1,757	1,320	7,525	170.51 (181.48)
2002	692	2,056	1,600	1,762	1,366	7,476	175.72 (194.42)
2003	663	1,922	1,503	1,647	1,335	7,070	175.32 (192.47)
2004	689	1,877	1,488	1,652	1,371	7,077	176.66 (187.81)
2005	707	1,869	1,467	1,653	1,388	7,084	177.00 (181.07)
2006	736	1,874	1,444	1,636	1,420	7,110	179.81 (208.74)
2007	747	1,789	1,507	1,646	1,437	7,126	180.73 (188.34)
2008	761	1,819	1,474	1,684	1,528	7,266	184.47 (199.12)
2009	752	1,774	1,493	1,694	1,570	7,283	187.45 (202.93)
2010	637	1,848	1,512	1,733	1,563	7,293	189.78 (198.67)
2011	627	1,828	1,438	1,755	1,612	7,260	194.08 (207.43)
2012	579	1,637	1,274	1,653	1,498	6,641	197.69 (248.16)

[a] ES in 1000 Euros of standard output

Source: Agreste, FADN France 2000-2012 – authors' calculations

As we considered all farms in the sample whatever their type of farming, we chose to concentrate on size as defined from an economic perspective. In accordance with the European regulation (CE) Nº1242/2008, FADN farms are classified into 14 economic size (ES) categories, evaluated in terms of total standard output (SO) expressed in Euros.[3] In France, the FADN focuses on 'commercial' farms, that is, farms whose SO is greater than or equal to 25,000 Euros; this corresponds to ES category 6 and above. We aggregated the 9 size categories available in the French FADN into 5: ES6 and below (less than 50,000 Euros of SO);[4] ES7 (between 50,000 and 100,000 Euros of SO); ES8 divided in two categories (between 100,000 and 150,000 of SO and between 150,000 and 250,000 of SO); ES9 and above (more than 250,000 Euros of SO). This led to observe 78,600 individual 1-year transitions from 2000 to 2012. Table 1 presents the evolution over the whole studied period of farm numbers by ES categories and average ES in thousand of Euros of SO for the studied panel.

Before proceeding with the results of our analysis, it should be noted that because we chose to work with a subset of the full sample, the transition probabilities reported in the next section should be viewed as size change probabilities *conditional* on having been observed at least two consecutive years during the whole period under study, and should not be considered as representative for the whole population of commercial French farms. Furthermore, because we cannot identify entries into and exits from the sector in

[3]SO is being 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 2012, allowing for consistent time series analysis (European Commission, 2010).

[4]Even if the French FADN focuses on commercial holdings, farms with less than 25,000 Euros of SO may be present at some point in the database because they are kept in the sample from year to year even if they fall below the threshold once in a while.

Table 2. Observed TPMs and size distribution in 2009.

		ES class				
		(0-50)	(50-100)	(100-150)	(150-250)	(+250)
ES class	(0-50)	0.920	0.075	0.003	0.002	0.000
	(50-100)	0.035	0.894	0.064	0.005	0.002
	(100-150)	0.003	0.064	0.851	0.080	0.002
	(150-250)	0.002	0.004	0.054	0.882	0.058
	(+250)	0.000	0.002	0.005	0.052	0.941

a) 1-year TPM ($\mathbb{P}^{(1)}$)

		ES class				
		(0-50)	(50-100)	(100-150)	(150-250)	(+250)
ES class	(0-50)	0.772	0.158	0.042	0.023	0.005
	(50-100)	0.132	0.609	0.206	0.034	0.016
	(100-150)	0.022	0.135	0.569	0.248	0.026
	(150-250)	0.010	0.032	0.094	0.639	0.226
	(+250)	0.009	0.013	0.020	0.123	0.836
D_{2009}	All	0.103	0.244	0.205	0.233	0.216

b) 9-years TPM ($\mathbb{P}^{(9)}$) and size distribution in 2009 (D_{2009})

Source: Agreste, FADN France 2000-2009 – authors' calculations

the FADN database, the distribution of farms will be analyzed in terms of shares of farms by size categories and not in terms of absolute numbers.

4.2. Results

In order to test the usefulness of the MSM and to compare its merits with respect to the MCM, we divided the database into two periods. First, we used observations from 2000 to 2009 to estimate the parameters of both models. Then, observations from 2010 to 2012 were used to compare out-sample predictions.

For the estimation phase, nine subsamples could be constructed according to the minimum number of consecutive years a farm remains present in the database, from two to ten. It appeared that the optimal subsample was the one where farms remained at least nine years in the database (not reported): with this subsample, the estimated 8-years TPM and the predicted distribution were closest to the observed ones for both model, as measured by the sum of marginal errors (SME), defined as (Frydman, 1984; Cipollini *et al.*, 2012):

$$SME = \sum \sqrt{\left(\frac{Obs - Pred}{Obs}\right)^2}.$$

The corresponding observed 1-year and 9-years TPMs were then computed (Table 2). As has been usually found in the literature, we observe that these TPMs are strongly

Table 3. Stayer shares and mover generator matrix.

	Stayers	Movers Q				
	s_{ii}	(0-50)	(50-100)	(100-150)	(150-250)	(+250)
(0-50)	0.534	-0.140	0.131	0.005	0.003	0.001
(50-100)	0.425	0.058	-0.176	0.106	0.009	0.003
(100-150)	0.169	0.004	0.079	-0.184	0.098	0.003
(150-250)	0.344	0.002	0.006	0.075	-0.163	0.080
(+250)	0.631	0.001	0.005	0.010	0.110	-0.126

Source: Agreste, FADN France 2000-2009 – authors' calculations

Table 4. Predicted 9-years TPMs and size distributions in 2009.

		ES class				
		(0-50)	(50-100)	(100-150)	(150-250)	(+250)
ES class	(0-50)	0.523	0.336	0.094	0.035	0.011
	(50-100)	0.160	0.477	0.229	0.103	0.031
	(100-150)	0.051	0.232	0.352	0.273	0.093
	(150-250)	0.018	0.075	0.188	0.446	0.274
	(+250)	0.007	0.030	0.071	0.251	0.641
D_{2009}^{MCM}	All	0.115	0.241	0.203	0.236	0.204

a) MCM TPM ($\mathbf{\Pi}^{(9)}$) and size distribution in 2009 (D_{2009}^{MCM})

		ES class				
		(0-50)	(50-100)	(100-150)	(150-250)	(+250)
ES class	(0-50)	0.708	0.167	0.079	0.035	0.012
	(50-100)	0.093	0.639	0.157	0.083	0.029
	(100-150)	0.046	0.152	0.525	0.200	0.078
	(150-250)	0.018	0.060	0.141	0.620	0.162
	(+250)	0.006	0.021	0.050	0.129	0.794
D_{2009}^{MSM}	All	0.113	0.248	0.203	0.236	0.204

b) MSM TPM ($\mathbf{P}^{(9)}$) and size distribution in 2009 (D_{2009}^{MSM})

Source: Agreste, FADN France 2000-2009 – authors' calculations

diagonal, meaning that their diagonal elements exhibit by far the largest values and that probabilities rapidly decrease as we move away from the diagonal. This means that, overall, farms are more likely to remain in their initial size category.[5]

In order to estimate the stayers proportions, **S**, and the generator matrix of movers,

[5] Which does not mean no size change at all but, at least, no sufficient change to move to another category as we defined them.

$\mathbf{Q}$, defining the MSM, we implemented the continuous-time specification and the EM algorithm estimation method developed in section 3. Table 3 reports the corresponding shares of stayers by size category and generator matrix of movers. The estimated stayer shares confirm that, for 4 categories out of the 5 ones considered, more than one third of farms do not move away from their initial category; for the intermediate category, *i.e.*, farms whose SO lies between 150,000 and 250,000 Euros, this share is less than 20%. Movers from the latter category remain about 5 years in this state while movers with less than 50,000 or more than 250,000 leave these categories after 7 to 8 years on average.[6] Such a result also shows that farms remaining in a particular state during a long time period are not necessarily stayers.

Then, Table 4 reports both the MCM 9-years TPM, $\mathbf{\Pi}^{(9)} \equiv \left(\mathbb{P}^{(1)}\right)^9$, and the MSM 9-years TPM, $\mathbf{P}^{(9)}$, obtained from $\mathbf{S}$, $\mathbf{Q}$ and equations (7) and (9), and both corresponding estimated size distributions in 2009. While both models quite compare in predicting the distribution of sizes, TPMs are obviously different, especially with respect to their diagonal elements. In particular, when compared to the actually observed 9-years TPM, $\mathbb{P}^{(9)}$ (see Table 2b), we find as expected that $\pi_{ii}^{(9)} \ll \phi_{ii}^{(9)}$ while $p_{ii}^{(9)}$ is much closer to $\phi_{ii}^{(9)}$. Overall, the MSM matrix thus appears as a better approximation of the observed matrix than the MCM matrix, which is confirmed by the respective sum of marginal errors (SME) computed with respect to the observed 9-years TPM.

Finally, out-of-sample predictions for 2010-2012 confirm the superiority of the MSM, which becomes even relatively more accurate with respect to the MCM as the projection horizon increases (not reported).

5. Concluding remarks

The empirical analysis provided in the previous section reveals that relaxing the homogeneity assumption which grounds the traditional Markov chain model (MCM) leads to a better modeling of the underlying economic process. Using a more general framework, the decomposition of the 1-year transition probability matrix into, on the one hand, a fraction of ‘stayers’ who remain in their initial size category and, on the other hand, a fraction of ‘movers’ who follow a standard Markovian process, allows to derive a closer estimate of the observed short- and long-run transition matrix as well as farm distribution across size categories.

Still, such a mover-stayer model (MSM) is quite a restricted and simplified version of the more general model which was presented in section 2. Even though we improved Blumen *et al.* (1955)’s calibration process by using the continuous-time approach and the elaborate expectation-maximization estimation method of Frydman (2005), extending Blumen *et al.* (1955)’s framework could 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 types of agents, and the quite strong assumption of a ‘pure stayer’ type could be relaxed. Secondly, with either of these two extensions put in place, Frydman (2005)’s assumption regarding the structural relation across generator matrices could be also revisited, especially in such a way that the process of structural change in the farming sector would be better represented.

Finally, the last direction towards which we would like to extend our modeling framework consists in accounting for entries and exits and developing a non-stationary version

[6] Recall that the time spent by movers in a particular category is given by $-1/q_{ii}$ (see section 2.3).

of the model. Indeed, we think that such a generalized version of the MSM approach could certainly prove very insightful for analyzing structural change in the farming sector, in particular to get a better understanding of the impact of agricultural policies on the development of farm numbers and sizes.

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

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

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

Dempster, A. P., Laird, N. M., Rubin, D. B. *et al.* (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society* 39: 1–38.

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

Fougère, D. and Kamionka, T. (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. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association* 79: 632–638.

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

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

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

Fuchs, C. and Greenhouse, J. B. (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.

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

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.

Lando, D. and Skodeberg, T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance* 26: 423–444.

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

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

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

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

Pol, F. van de and Langeheine, R. (1989). *Multiway data analysis*. Amsterdam, The Netherlands: North-Holland Publishing Co.

Singer, B. and Spilerman, S. (1975). *Identifying structural parameters of social processes using fragmentary data*. University of Wisconsin–Madison.

Singer, B. and Spilerman, S. (1976). Some methodological issues in the analysis of longitudinal surveys. *Annals of Economic and Social Measurement* 5: 447–474.

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

Storm, H., Heckelei, T. and Mittelhammer, R. C. (2011). Bayesian estimation of non-stationary Markov models combining micro and macro data. Discussion Paper 2011:2, University of Bonn, Institute for Food and Resource Economics, Bonn (Germany).

Swensen, A. R. (1996). On maximum likelihood estimation in the mover-stayer model. *Communications in statistics. Theory and methods A* 25: 1717–1728.

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

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