Assessing structural change in agriculture with a parametric Markov chain model. Illustrative applications to EU-15 and the USA

Laurent PIET
INRA, UMR1302 SMART, F-35000 Rennes
4 allée Adolphe Bobierre, CS 61103, 35011 Rennes cedex, France

Paper prepared for presentation at the EAAE 2011 Congress
Change and Uncertainty
Challenges for Agriculture,
Food and Natural Resources

August 30 to September 2, 2011
ETH Zurich, Zurich, Switzerland

ABSTRACT

The Markov chain model (MCM) has become a popular tool in the agricultural economics literature to explain the past evolution of and simulate the future developments in the number and size distribution of farms. In this paper, I show that the way MCMs have been implemented by agricultural economists so far suffers from the fact that transition probabilities are estimated as almost independent variables (up to adding-up to unity constraints). The alternative parametric MCM I propose addresses the deriving issues since (i) it is parsimonious in terms of parameters; (ii) it can be estimated with simple econometric techniques; (iii) it reveals detailed information on the structural change processes at hand. Applying it to experimentally controlled data with noise shows that the proposed model behaves well and competes with the traditional approach without any significant shortcoming. Two illustrative empirical applications, one using data from the EU-15 Farm Accounting Data Network (FADN) and the other using data from the USA Agricultural Resource Management Survey (ARMS), reveal the rich information that can be derived regarding the economic size changes experienced annually by farms in both regions.

1. INTRODUCTION

The so-called Markov chain model (MCM) has become a popular tool to study the evolution of structural change in agriculture, that is the evolution of the number and structures distribution of populations of agricultural firms (Zimmermann et al., 2009).

Basically, a Markov chain model allows to recover the number of farms in a particular category at a particular date as the sum of the transitions towards this category experienced by farms which where previously in any other categories;[1] at each time step, transitions occur only with a certain probability (*i.e.*, only a fraction of individuals move from one category to another, including staying in the same category). Formally, this can be written as $n_{j,t+1} = \sum_{k=1}^{J} p_{kj} n_{k,t}$, where J is the (finite) number of categories indexed by $j,k = \{1...J\}$; $n_{j,t}$ is the number of individuals in the j-th category at time t; and p_{kj} is the probability for the individuals in category k to move to category j between t and $t+1$; further, p_{kj} is subject to the standard probability constraints of positivity ($p_{kj} \geq 0$) and adding-up to unity ($\sum_{j=1}^{J} p_{kj} = 1$). When p_{kj} depends on t the model is said non-stationary; otherwise it is said stationary.

The task of the modeller then consists in estimating these transition probabilities. This is quite simple when individual (panel) data are available since individual transitions are directly observable and countable; it is a more complicated task when only aggregate (cross-sectional) data are available, which is the most common empirical situation –in agricultural economics at least.[2] However, Lee *et al.* (1965) and Lee *et al.* (1977) showed that econometric techniques make it possible to estimate a robust MCM from aggregate data only; since then, most of the MCM literature in agricultural economics has used such aggregate data (Zimmermann *et al.*, 2009).

[1] To my knowledge, most empirical works consider the previous date only, leading to a Markov chain process of degree 1. More general (higher degree) MCMs consider several previous dates (Berchtold, 1998).

[2] Panel data are costly and are therefore usually limited both in terms of observation dates and sample size.

The drawback of this aggregate MCM implementation –which I shall refer to as the "standard" MCM implementation in the following– is that the number of transition probabilities to estimate is usually large even when only a few categories are considered. Moreover, this number grows exponentially as the number of categories increases, since all the J^2 possible transitions have to be taken into account and the corresponding probabilities have to be estimated; this number is actually limited to $J(J-1)$ thanks to the adding-up to unity constraints but the exponential growth rate remains. Then, the number of observations needed to identify all the parameters of the model rapidly becomes prohibitive, leading to an ill-posed problem (Karantininis, 2002). In sum, the analyst is faced with a trade-off between the richness of the data he has at his disposal to estimate the model and the richness of the information he can recover from it. Two directions have been explored so far in the literature to overcome this drawback. First, arbitrary zero-constraints can be imposed on some specific probabilities, assuming that the corresponding transitions are impossible and thus reducing the number of parameters to estimate (among others, see Krenz (1964), Zepeda (1995) or Gillespie and Fulton (2001)); then, simple econometric techniques like linear seemingly unrelated regressions (SUR) or ordinary least-squares (OLS) can still be applied. Second, more elaborate econometric methods can be used such as the generalized cross-entropy (GCE) and instrumental variables GCE (IV-GCE) which take advantage of *a priori* beliefs on the magnitude of transition probabilities rather than making rigid assumptions as above (Karantininis, 2002; Stokes, 2006; Tonini and Jongeneel, 2008); however one can suspect that, even if more flexible, these exogenous priors closely drive the results in the case of such strongly under-identified models.[3] Finally, a consequence of this standard approach is the quite limited information it produces: of course, it fulfils its initial objective in the sense that it eventually permits to project the population to any arbitrary horizon, that is to simulate the number of farms in each category and as a whole (*i.e.* a relevant information for the planners) but… this is it.[4] In particular, it does not exploit the fact that in general, at least in all the works listed by Zimmermann *et al.* (2009), the dependant variable in the model, that is the criteria defining the J categories, is actually a continuous (size) variable.[5]

The structural MCM I have developed tackles all of the previous shortcomings: (i) it is parsimonious in terms of parameters; (ii) it does not require to form *a priori* assumptions or beliefs on the individual probabilities themselves; (iii) in its simpler version, it can be estimated with standard SUR techniques; and (iv) the information it brings leads to richer insights into the structural change process at hand and the distribution of the projected population.

The rest of the paper is organised as follows. Section 2 presents the proposed parametric modelling framework, emphasizing on how it departs from and enriches the standard MCM approach that was briefly outlined in this introduction. Section 3 first describes the data which were used to demonstrate the satisfactory behavior of the proposed model in a theoretic experimental design; it then describes two empirical datasets, one for EU-15 (15 Member states of the European Union) and the other for the United States of America (USA). Section 4 reports the results for both the experimental and the empirical illustrative applications before concluding remarks and directions for future work are discussed in the last section.

[3] As an illustration, Karantininis (2002) works on 19 categories and 15 census years and is so faced with the estimation of 324 probabilities from 14 transitions corresponding to 252 data points.

[4] Of course, non-stationary MCMs bring extra information regarding the impact of some explanatory variables (such as policy or market variables) on the transition probabilities but here I only refer to the "intrinsic" information regarding the structure of the population that can be extracted from a MCM.

[5] Butault and Delame (2005) are a worth noticing exception: using a large scale panel, they worked with a large number of categories not only defined upon the size of farms but also on qualitative variables such as the region, the type of farming, the legal status of the farm or the age of the operator.

2. THE PROPOSED MODEL

As in the standard approach, the population under study is partitioned into a finite number of categories J. But here, these categories are explicitly defined over the continuous quantitative size variable X as intervals $(\underline{x}_j, \overline{x}_j]$, with $j = \{1...J\}$. Dealing with aggregate data, the numbers of firms in each category at several time periods $n_{(\underline{x}_j, \overline{x}_j],t}$ are the only observations we have.

With these definitions, the standard Markov chain equation can be expressed as:

$$n_{(\underline{x}_j, \overline{x}_j],t+1} = \sum_{k=1}^{J} p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]} n_{(\underline{x}_k, \overline{x}_k],t} \tag{1}$$

where $p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]}$ is the probability to move from an initial size lying in $(\underline{x}_k, \overline{x}_k]$ to a final size lying in $(\underline{x}_j, \overline{x}_j]$ in one time-period. The standard probability constraints of positivity, $p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]} \geq 0$, and adding-up to unity, $\sum_{j=1}^{J} p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]} = 1$, apply.

Whereas in the standard approach it is an aggregate unknown, here the transition probability $p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]}$ can be explicitly derived from the individual probability $\pi(1+\delta x \mid x)$ of experiencing a relative change in size $1+\delta x$ $(-1 < \delta x < +\infty)$[6] between two consecutive dates, conditional on an initial size x:

$$\pi(1+\delta x \mid x) \equiv P\left(X_{t+1} / X_t = 1+\delta x \middle| X_t = x \right) \tag{2}$$

For an individual exhibiting the initial size $X_t = x$ lying in $(\underline{x}_k, \overline{x}_k]$, the probability to exhibit a final size $X_{t+1} = x'$ lying in $[\underline{x}_j, \overline{x}_j)$ is then given by:

$$p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]} = \int_{\underline{x}_k}^{\overline{x}_k} P(X_t = x) \left(\int_{\underline{x}_j / x}^{\overline{x}_j / x} \pi(1+\delta x \mid x)\, d(\delta x) \right) dx \tag{3}$$

In the absence of more precise information regarding the distribution of the probabilities $P(X_t = x)$, as is the case when dealing with aggregate data, the simplest assumption to be made is that of a uniform distribution over each interval $(\underline{x}_k, \overline{x}_k]$, that is, $P(X_t = x) = 1 / (\overline{x}_k - \underline{x}_k)$. Then, equation (3) simplifies to:

$$p_{(\underline{x}_k, \overline{x}_k](\underline{x}_j, \overline{x}_j]} = \frac{1}{\overline{x}_k - \underline{x}_k} \int_{\underline{x}_k}^{\overline{x}_k} \left(\int_{\underline{x}_j / x}^{\overline{x}_j / x} \pi(1+\delta x \mid x)\, d(\delta x) \right) dx \tag{4}$$

It is worth noticing that assuming a uniform distribution of sizes on $(\underline{x}_k, \overline{x}_k]$ does not imply that this distribution be uniform on the whole range of X, but only that it is piece-wise uniform. Any size distribution could thus be approximated be the number of categories sufficient. This assumption could be relaxed in empirical applications if some knowledge regarding the true size distributions at each date were available.

[6] $\delta x = -1$ would correspond to exiting the sector, a feature not taken into account so far (see Section 5).

A parametric form with a known cumulative distribution function $F(\delta x;\boldsymbol{\theta}_l(x))$, where $\boldsymbol{\theta}_l(x)$ is a set of l parameters defining F, can be chosen for $\pi(1+\delta x \mid x)$. This allows rewriting (4) in a simpler way:

$$p_{(\underline{x}_k,\bar{x}_k](\underline{x}_j,\bar{x}_j]} = \frac{1}{\bar{x}_k - \underline{x}_k}\int_{\underline{x}_k}^{\bar{x}_k}\left(F(\bar{x}_j/x;\boldsymbol{\theta}_l(x)) - F(\underline{x}_j/x;\boldsymbol{\theta}_l(x))\right)dx \tag{5}$$

Then, the parametric Markov chain model to estimate is the system of pooled equations (1) with the addition of error terms $u_{j,t+1}$ and $p_{(\underline{x}_k,\bar{x}_k](\underline{x}_j,\bar{x}_j]}$ being given by (5):

$$\begin{aligned} n_{(\underline{x}_j,\bar{x}_j],t+1} &= \sum_{k=1}^{J} p_{(\underline{x}_k,\bar{x}_k](\underline{x}_j,\bar{x}_j]} n_{(\underline{x}_k,\bar{x}_k],t} + u_{j,t+1} \\ \text{with } p_{(\underline{x}_k,\bar{x}_k](\underline{x}_j,\bar{x}_j]} &= \frac{1}{\bar{x}_k - \underline{x}_k}\int_{\underline{x}_k}^{\bar{x}_k}\left(F(\bar{x}_j/x;\boldsymbol{\theta}_l(x)) - F(\underline{x}_j/x;\boldsymbol{\theta}_l(x))\right)dx \end{aligned} \tag{6}$$

It is easy to see that the standard probability constraints of positivity and adding-up to unity are implied by (5) and the definition of $F(\delta x;\boldsymbol{\theta}_l(x))$ as a cumulative distribution function.

3. EXPERIMENTAL AND EMPIRICAL DATA

In order to demonstrate the characteristics and usefulness of the parametric MCM defined by equation (6), I first designed a hypothetical set of experimental data; the choices that I have made to do so are arbitrary but it was intended to reproduce the kind of data that are usually available in empirical studies in the agricultural sector. As an illustration of such data, I then applied the parametric MCM to two datasets, one for EU-15 and the other for the USA.

This section presents both types of data successively.

3.1. The experimental design

A hypothetical set of experimental data was used to demonstrate the behavior of the parametric MCM: a set of 100 hypothetical individuals was partitioned into ten classes defined over a hypothetical (size) variable taking its values over the domain (0;+∞). The category intervals chosen were (0;2], (2;5], (5;10], (10;20], (20;50], (50;75], (75;100], (100;200], (200;400] and (400;+∞) and the initial distribution of sizes was chosen to follow a log-normal distribution $F(X) = \Phi(\ln(X);\ln(20),\ln(2))$ where $\Phi(u;m,s)$ is the normal distribution with mean m and standard deviation s. As can be seen from Table 1, I chose (i) unequally spaced lower and upper bounds to define the size intervals and (ii) an initial size distribution which is not uniform but rather highly skewed toward small sizes; as previously mentioned, this is meant to mimic what is usually found in the empirical data.

[insert Table 1 around here]

In this experiment, a log-normal functional form $F(\delta x) = \Phi(\ln(1+\delta x);\mu,\sigma)$ for the distribution of $\pi(1+\delta x \mid x)$ was also chosen. For the sake of simplicity and without loss of generality, the μ and σ parameters were chosen not to depend on the initial size; they were both set to 0.01. Equation (5)

was then used to compute the transition probabilities $p_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)}$ and equation (1) to generate a set of 20 more observation periods; the dataset then comprised 21 time periods, or 20 transitions.[7] A second set of data was created by aggregating the initial ten categories into five ones on the following intervals: $(0;10]$, $(10;20]$, $(20;50]$, $(50;100]$ and $(100;+\infty]$; this aggregation preserves the overall log-normal shape of the initial size distribution.

Table 1 reports the 10-category population aggregation dataset and Table 2 presents the transition probability matrix for the 5-category aggregation case.

[insert Table 2 around here]

A simple white noise could not be directly added since it could have moved the data too far away from the underlying Markov process. I therefore adopted the following procedure which was applied a hundred times to both the 10- and 5-category population aggregations to obtain 100 replications of both datasets:

1. for each category and each time period, compute $n^a_{j,t} = 0.5\times\left(n_{(\underline{x}_j,\bar{x}_j],t-1} + n_{(\underline{x}_j,\bar{x}_j],t}\right)$ and $n^b_{j,t} = 0.5\times\left(n_{(\underline{x}_j,\bar{x}_j],t} + n_{(\underline{x}_j,\bar{x}_j],t+1}\right)$; for the first $(t=0)$ and last $(t=20)$ periods, take $n^a_{j,0} = 2n_{(\underline{x}_j,\bar{x}_j],0} - n^b_{j,0}$ and $n^b_{j,20} = 2n_{(\underline{x}_j,\bar{x}_j],20} - n^a_{j,20}$, respectively;
2. define $\underline{n}_{j,t} = \min\left(n^a_{j,t}, n^b_{j,t}\right)$ and $\bar{n}_{j,t} = \max\left(n^a_{j,t}, n^b_{j,t}\right)$
3. replace $n_{(\underline{x}_j,\bar{x}_j],t}$ by $\tilde{n}_{(\underline{x}_j,\bar{x}_j],t} = n_{(\underline{x}_j,\bar{x}_j],t} + \alpha.\varepsilon_{j,t}.\left(\bar{n}_{j,t} - \underline{n}_{j,t}\right)/2$ where $\varepsilon_{j,t}$ is a random number uniformly drawn in $[-1;1]$ and α is a parameter ranging from 0.1 (minimum noise) to 1.0 (maximum noise);
4. for each time period, scale back the resulting data to 100; this preserves a constant population at each date so that entries and exits are not considered, only the evolution of population shares are.

Finally, a third set of data was prepared by aggregating the disturbed 10-category data into five categories defined on the same intervals as the initial 5-category data for each replication.

Fig. 1 compares the resulting disturbance of the original data for two levels of noise intensity in the case of the $(0;2]$ category of the 10-category population and for the first replication.

[insert Fig. 1 around here]

[7] Actually, the integral in equation (5) has no closed form analytical solution when the log-normal distribution is used; a numeric approximation of the integral was performed using a simple trapeze formula and choosing a size step of 1 and an upper bound of the last interval set to 1000 as an approximation for infinity.

3.2. Empirical datasets

A first dataset was created for the fifteen "historical" Member states of the EU[8] from the Farm Accounting Data Network (FADN) data available on the website of the European Commission.[9] In the database available on-line, the total population of professional farms can be split up according to their economic size measured in Economic Size Unit (ESU); for the sake of homogeneity across countries, I only kept categories defined over a size greater than or equal to 8 ESU, leading to 4 classes: 8 to less than 16 ESU; 16 to less than 40 ESU, 40 to less than 100 ESU and 100 or more ESU. Though data are available from 1989 to 2008 on the Commission website, the time period covered here is 1995-2007 only because (i) data are not available before 1995 for Austria, Finland and Sweden who joined the EU in 1994 and (ii) data are lacking for Italy in 2008. This results in 12 observed annual transitions.

A second dataset was created for the USA from the Agricultural Resource Management Survey (ARMS) data available on the website of the US Department of Agriculture (USDA).[10] In the database available on-line, the total population of farms can be split up according to their economic size measured in US dollars of gross sales; in the ARMS terminology, I kept only categories defined over gross sales greater than or equal to US$ 100,000, leading also to 4 classes: US$ 100,000 to less than US$ 250,000; US$ 250,000 to less than US$ 500,000; US$ 500,000 to less than US 1,000,000; and US$ 1,000,000 or more. The time period covered is 1996-2009, resulting in 13 observed annual transitions.

Table 3 presents the resulting two datasets. As with the experimental dataset, the parametric MCM defined by equation (6) was run on the shares of population and not on the population numbers themselves in order to focus on structural change and not on entry/exit issues.

[insert Table 3 around here]

4. RESULTS

With only two parameters to estimate whatever the number of categories considered, the parametric model was always a well-posed problem. It could be therefore estimated with a simple non linear system of equations seemingly unrelated regression (SUR) estimation procedure.

4.1. Assessing the parametric Markov Chain Model

The parametric model was first estimated with the experimental datasets described in the previous section. The left panel of Fig. 2 shows that the estimation of the parameters μ and σ is unbiased whatever the intensity of the noise added to the original data. Still, as expected, the dispersion of the estimated coefficients, which could be assessed from the 100 replications, increases with the noise intensity. The centre panel of Fig. 2 shows that the estimation remains unbiased with more aggregated data if the aggregation occurs *before* the addition of noise; however, the rightmost panel of Fig. 2 shows that the estimation is no longer unbiased when the data are aggregated *afterwards*.

[insert Fig. 2 around here]

[8] Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden and the United Kingdom.
[9] http://ec.europa.eu/agriculture/rica/index_en.cfm
[10] http://www.ers.usda.gov/Briefing/ARMS/

In order to assess how the parametric MCM compares with the standard approach, I estimated a stationary multinomial logit model following Zepeda's (1995) specification with the 5-category data. The unconstrained logit would have required estimating $5\times(5-1)=20$ transition probabilities (thanks to the adding-up to unity constraint); since the experimental database contains 20 transitions only, this model would have been an ill-posed problem. Then, as is usually done, I imposed constraints on some transition probabilities in order to reduce the number of parameters by only setting possible: i) to stay in the same category; ii) to move to the very previous category; and iii) to move to the very next category. So constrained, the logit model encompassed 8 parameters only.

The sum of the root mean square errors over the 4 equations forming the system to estimate was used as a measure of how the models fit the data.[11] Fig. 3 reports this overall fit for both the parametric and the constrained multinomial logit MCMs. It shows that, if both models perform quite similarly whatever the level of noise, the standard model doing better on average though. However, the standard model fails to recover the true underlying transition probability matrix while the parametric model always does (not reported here).

[insert Fig. 3 around here]

4.2. Empirical results

The parametric MCM was then estimated with the two empirical datasets using four alternative functional forms for $F(\delta x)$:

- a log-normal distribution: $F(\delta x)\equiv LN(1+\delta x;\mu,\sigma)=\Phi(\ln(1+\delta x);\mu,\sigma)$;
- a Weibull distribution: $F(\delta x)\equiv WB(1+\delta x;\nu,\lambda)$ where $WB(u;\nu,\lambda)=1-e^{-(u/\lambda)^{\nu}}$ with shape $\nu>0$ and scale $\lambda>0$;
- a gamma distribution: $F(\delta x)\equiv GM(1+\delta x;\kappa,\theta)$ where $GM(u;\theta,\kappa)=\Gamma(\kappa,u/\theta)/\Gamma(\kappa)$ with shape $\kappa>0$ and scale $\theta>0$ and $\Gamma(.)$ the gamma function;
- a Gumbel distribution: $F(\delta x)\equiv GB(\delta x;\mu,\beta)$ where $GB(u;\mu,\beta)=e^{-e^{-(u-\mu)/\beta}}$ with location μ and scale $\beta>0$.

Here again, the parameters defining $F(\delta x)$ were chosen independent from the initial size whatever the specification for the functional form.[12] The number of parameters to be estimated in each case (two) was thus always much smaller than the number of observations. Table 4 reports the corresponding results; note that, since the economic size is not measured with the same indicators, the results for EU-15 and the USA cannot be directly compared to each other.

In both cases, the Weibull specification outperforms the other distributions on the ground of the total root mean square error; differences are small though from one specification to the other, especially in the USA case.

[insert Table 4 around here]

[11] Since there is neither entry nor exit in the constructed experimental dataset, we are in fact dealing with a constant population so that the number of individuals in the, say, last category is known once the number of individuals in the other categories have been determined. One equation has therefore to be dropped in the estimation process (see Zepeda, 1995).

[12] For those which must be strictly positive, the log of the parameter was actually estimated.

In the parametric MCM, the "standard" transition probabilities are not estimated directly. Rather, they derive from the underlying probability distribution of relative size change as was shown in Section 2. The standard errors associated to these probabilities are therefore not readily available from the estimation process and are not easy to compute analytically from equation (5). To overcome this issue, I implemented a Monte Carlo simulation of the transition probability matrices, replicating equation (5) 200 times by drawing in the normal distributions defined by the coefficients and standard deviations appearing in Table 4 for each parameter of the Weibull distribution. The resulting average estimated transition probabilities and their associated standard deviations are reported in the transition probability matrices of Table 5.

[insert Table 5 around here]

As is usually found in the literature, both matrices are highly diagonal: every diagonal term is close or above 0.90 and the only off-diagonal terms which are significantly different from zero are small (below 0.05). This means that most probably EU-15 and USA farms tend to stay in the same size category from one year to the other: as can be seen from the cumulative distributions reported in Fig. 4 resulting from the estimated coefficients, both the average and the median relative changes in economic size are close to zero for either datasets.

This does not mean no structural change at all though: as these curves also show, 10% of the farms experience of relative decrease of -6.0% and -10.0% or more respectively in EU-15 and in the USA, while 10% experience an increase in their economic size of respectively +5.1% and +7.7% or more. Note that this kind of information could not have been derived from the standard MCM approach.

[insert Fig. 4 around here]

Finally, some transitions appear implausible, the associated probabilities being estimated very close to zero: the estimated cumulative distributions of Fig. 4 show that this is the case for a relative change in economic size below -13.7% or above +7.7% for EU-15 and below -20.0% or above +12.1% for the USA. Note again that this could not have been inferred in the standard approach setting.

5. CONCLUDING REMARKS

In this paper, I present an original way of implementing the Markov chain model (MCM) which has been widely used in the recent academic literature to study the evolution and structural change of agricultural populations in several countries. Unlike the "standard" MCM approach which regards the transitions probabilities as almost unrelated parameters (up to adding-up to unity constraints), the method I propose takes advantage of the quantitative and continuous nature of the dependent variable used to define the categories into which the studied population is broken down. It allows to express the transition probabilities as deriving from an underlying probability distribution of the relative change in size, leading to a much richer information on the structural change process at hand. Further, no assumption has to be formed regarding the impossibility or implausibility of specific transitions: improbable transitions derive "endogenously" from the estimated probability distribution.

Yet an assumption on the shape of this distribution has to be made. Choosing a parametric functional form reduces sharply the number of parameters to be estimated. The model is thus made parsimonious and it is less likely to be an ill-posed problem; rather "simple" econometric techniques can then be employed. Moreover, several functional forms could be tested and four of them have been illustrated here. More flexible forms could be chosen, at the expense of an increasing set of parameters; this number should however stay small compared to that in the standard approach. Another important

assumption that was used here regarding the probability distribution is that its parameters were set independent of the initial size. This could be relaxed in two ways: (i) these parameters could be made dependent on the initial size category to which they apply; this would increase the number of parameters of the model and relate it directly to the number of categories; still the relation would be linear whereas it is exponential in the standard approach; (ii) a statistical relationship between the parameters and the initial size itself could be specified, adding more structure into the model; choosing a simple linear relationship would only double the number of parameters and would preserve the independence *vis-à-vis* the number of categories, maintaining the parsimonious nature of the model; but this would intuitively require to have a more detailed information regarding the distribution of sizes among the population at one's disposal or to form further assumptions regarding it. Either ways would however allow to test rigorously the relationship between the relative size change and the initial size, *i.e.* to test whether the so-called Gibrat's law holds or not.

In the illustrative empirical analyses presented here, the model was applied to the evolution of population shares. The recovered information therefore deals with structural change only and not with the developments in the absolute numbers of farms. It should be easy to extend the framework to account for entries and exits, either in the direction of Zepeda (1995) who complements the Markov chain with a net exit process, or in the direction of Karantininis (2002) and Stokes (2006) who explicitly consider entries and exits separately.

Finally, the model presented here is stationary. A non-stationary version could be built by making the probability distribution parameters depend on time-varying covariates. A simple trend would preserve parsimony but would not be much interesting from an economic and political point of view. More appealing would be to use market and policy explanatory variables as is done in the recent literature using the standard MCM approach (Zepeda, 1995; Karantininis, 2002; Stokes, 2006; Tonini and Jongeneel, 2008). The number of parameters to estimate would increase so that even the structural approach proposed here would become an ill-posed problem and generalized cross-entropy (GCE) or instrumental variables GCE (IV-GCE) techniques would be required. Yet the structural approach would still be the more parsimonious of the two methods so that more degrees of freedom would be preserved for an undoubtedly more robust covariate effects estimation.

6. REFERENCES

Berchtold, A. (1998). *Chaînes de Markov et modèles de transition : application aux sciences sociales.* Editions Hermès, Paris (France).

Butault, J.-P. and N. Delame (2005). Concentration de la production agricole et croissance des exploitations. *Economie et Statistique* 390: 47-64.

Gillespie, J. M. and J. R. Fulton (2001). A Markov chain analysis of the size of hog production firms in the United States. *Agribusiness* 17(4), 557–570.

Karantininis, K. (2002). Information-based estimators for the non-stationary transition probability matrix: an application to the Danish pork industry. *Journal of Econometrics* 107(1-2): 275-290.

Krenz, R. D. (1964). Projection of farm numbers for North Dakota with Markov chains. *Agricultural. Economics Research* 16: 77-83.

Lee, T. C., G. G. Judge and T. Takayama (1965). On estimating the transition probabilities of a Markov process. *Journal of Farm Economics* 47(3): 742-762.

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

Stokes, J. R. (2006). Entry, exit, and structural change in Pennsylvania's dairy sector. *Agricultural and Resource Economics Review* 35(2): 357-373.

Tonini, A. and R. Jongeneel (2008). The distribution of dairy farm size in Poland: a Markov approach based on information theory. *Applied Economics* 40, 1-15.

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., T. Heckeleï and I. Perez Dominguez (2009). Modelling farm structural change for integrated ex-ante assessment: review of methods and determinants. *Environmental Science and Policy* 12: 601-618.

Table 1. The experimental population

t	Category intervals									
	(0;2)	[2;5)	[5;10)	[10;20)	[20;50)	[50;75)	[75;100)	[100;200)	[200;400)	[400;+∞)
0	0.0447	2.2303	13.5905	34.1345	40.6904	6.4829	1.8149	0.9671	0.0439	0.0008
1	0.0942	2.0836	12.8150	33.6788	41.2592	7.0332	1.9458	1.0247	0.0638	0.0017
2	0.1295	1.9594	12.0929	33.1821	41.7982	7.5749	2.0888	1.0867	0.0845	0.0031
3	0.1540	1.8519	11.4216	32.6509	42.3062	8.1076	2.2432	1.1536	0.1060	0.0049
4	0.1705	1.7569	10.7981	32.0911	42.7826	8.6312	2.4080	1.2258	0.1286	0.0072
5	0.1811	1.6712	10.2193	31.5082	43.2267	9.1454	2.5826	1.3036	0.1521	0.0099
6	0.1872	1.5929	9.6816	30.9068	43.6382	9.6499	2.7661	1.3872	0.1769	0.0132
7	0.1899	1.5204	9.1820	30.2914	44.0171	10.1446	2.9577	1.4770	0.2028	0.0170
8	0.1901	1.4526	8.7173	29.6657	44.3634	10.6292	3.1569	1.5732	0.2302	0.0214
9	0.1885	1.3888	8.2846	29.0332	44.6774	11.1034	3.3629	1.6760	0.2590	0.0263
10	0.1856	1.3284	7.8812	28.3966	44.9595	11.5671	3.5750	1.7855	0.2894	0.0318
11	0.1816	1.2710	7.5046	27.7587	45.2102	12.0200	3.7926	1.9018	0.3214	0.0380
12	0.1770	1.2165	7.1526	27.1217	45.4300	12.4620	4.0151	2.0250	0.3553	0.0449
13	0.1719	1.1644	6.8231	26.4875	45.6196	12.8928	4.2419	2.1552	0.3910	0.0525
14	0.1665	1.1147	6.5142	25.8578	45.7798	13.3124	4.4724	2.2924	0.4288	0.0609
15	0.1610	1.0673	6.2242	25.2340	45.9113	13.7206	4.7062	2.4366	0.4687	0.0701
16	0.1553	1.0220	5.9516	24.6174	46.0151	14.1172	4.9426	2.5879	0.5108	0.0802
17	0.1496	0.9788	5.6950	24.0089	46.0920	14.5021	5.1812	2.7461	0.5553	0.0911
18	0.1440	0.9375	5.4532	23.4094	46.1428	14.8754	5.4214	2.9112	0.6022	0.1030
19	0.1385	0.8980	5.2251	22.8198	46.1685	15.2368	5.6629	3.0831	0.6515	0.1159
20	0.1331	0.8604	5.0095	22.2404	46.1700	15.5864	5.9052	3.2617	0.7035	0.1298

Source: author's calculations

Table 2. Transition probability matrix for the 5-category aggregation case [a]

	(0;10)	[10;20)	[20;50)	[50;100)	[100;+∞)
(0;10)	0.9579	0.0421	0.0000	0.0000	0.0000
[10;20)	0.0079	0.9500	0.0421	0.0000	0.0000
[20;50)	0.0000	0.0026	0.9782	0.0192	0.0000
[50;100)	0.0000	0.0000	0.0016	0.9765	0.0219
[100;+∞)	0.0000	0.0000	0.0000	0.0001	0.9999

[a] *The functional form of the underlying relative size change probability distribution is log-normal with parameter $\mu = 0.01$ and $\sigma = 0.01$.*

Source: author's calculations

Table 3. Farm population data for EU-15 and the USA (1,000 individuals) [a]

Year	EU-15					USA				
	8-16	16-40	40-100	> 100	Total	100-250	250-500	500-1000	>1000	Total
1995	781.8	829.8	483.4	153.9	2248.8					
1996	757.5	820.7	515.7	181.9	2275.7	206.5	100.7	40.3	22.4	369.8
1997	753.7	822.0	519.5	184.4	2279.6	206.1	82.6	34.6	18.7	342.0
1998	747.9	819.7	517.2	184.2	2269.0	197.6	96.0	43.0	24.9	361.5
1999	657.4	768.6	529.4	232.5	2187.9	199.2	81.4	38.3	26.2	345.1
2000	656.7	769.7	527.3	233.9	2187.5	202.2	82.7	41.0	21.3	347.1
2001	649.0	771.0	529.9	236.6	2186.4	191.0	87.9	39.4	27.9	346.3
2002	578.8	710.1	518.1	258.1	2065.1	187.5	88.7	42.1	27.2	345.6
2003	578.1	711.1	517.9	260.4	2067.5	170.0	87.4	45.0	28.0	330.4
2004	602.7	728.5	514.6	262.6	2108.5	167.9	88.9	44.7	34.5	336.0
2005	598.5	735.2	515.1	264.5	2113.3	165.9	89.8	43.9	35.1	334.7
2006	562.9	735.5	505.9	279.6	2083.9	165.4	90.3	45.7	35.3	336.7
2007	571.7	731.5	505.8	281.9	2090.8	147.8	96.9	72.1	47.6	364.4
2008						145.1	97.8	74.4	51.3	368.5
2009						147.2	99.0	74.4	50.1	370.7

[a] *The categories for the EU-15 are based on the economic size of the farms measured in European Size Unit (ESU); the categories for the USA are based on the economic size of farms measured in 1000 US$ of gross sales.*

Source: European Commission, FADN for the EU-15 and USDA, ARMS for the USA

Table 4. Estimation results for the four tested functional forms [a]

	EU-15				USA			
	LN	WB	GM	GB	LN	WB	GM	GB
θ_1	**-0.0020**	**0.0194*** **	**6.2158*** **	**-0.0217***	**-0.0085**	**0.0255***	**5.2791*** **	**-0.0393**
	(0.0069)	(0.0061)	(0.7379)	(0.0118)	(0.0240)	(0.0138)	(1.4936)	(0.0421)
θ_2	**-3.1077*** **	**3.3190*** **	**-6.2168*** **	**-3.3132*** **	**-2.6389*** **	**2.8413*** **	**-5.2850*** **	**-2.8435*** **
	(0.3686)	(0.3662)	(0.7343)	(0.3530)	(0.7461)	(0.7709)	(1.4775)	(0.7210)
TRMSE	*0.02805*	*0.02801*	*0.02805*	*0.02805*	*0.07636*	*0.07636*	*0.07636*	*0.07636*

[a] *"LN" stands for log-normal; the corresponding parameters are $\theta_1 = \mu$ and $\theta_2 = \ln(\sigma)$. "WB" stands for Weibull; the corresponding parameters are $\theta_1 = \ln(\lambda)$ and $\theta_2 = \ln(\nu)$."GM" stands for Gamma; the corresponding parameters are $\theta_1 = \ln(\theta)$ and $\theta_2 = \ln(\kappa)$."GB" stands for Gumbel; the corresponding parameters are $\theta_1 = \mu$ and $\theta_2 = \ln(\beta)$. "TRMSE" stands for the total root mean square error. Estimated coefficients are in bold font with the corresponding standard deviations in bracketed regular font. *** stands for significantly different from zero at the 1% level, ** for significantly different from zero at the 5% level and * for significantly different from zero at the 10% level.*

Source: author's estimates

Table 5. Estimated transition probability matrices for the EU-15 and the USA [a]

EU	8-16	16-40	40-100	100+
8-16	**0.9613**	**0.0387**	**0.0000**	**0.0000**
	(0.0065)	(0.0065)	(0.0000)	(0.0000)
16-40	**0.0147**	**0.9568**	**0.0285**	**0.0000**
	(0.0035)	(0.0053)	(0.0067)	(0.0000)
40-100	**0.0000**	**0.0131**	**0.9591**	**0.0278**
	(0.0000)	(0.0035)	(0.0057)	(0.0068)
100+	**0.0000**	**0.0000**	**0.0028**	**0.9972**
	(0.0000)	(0.0000)	(0.0006)	(0.0006)

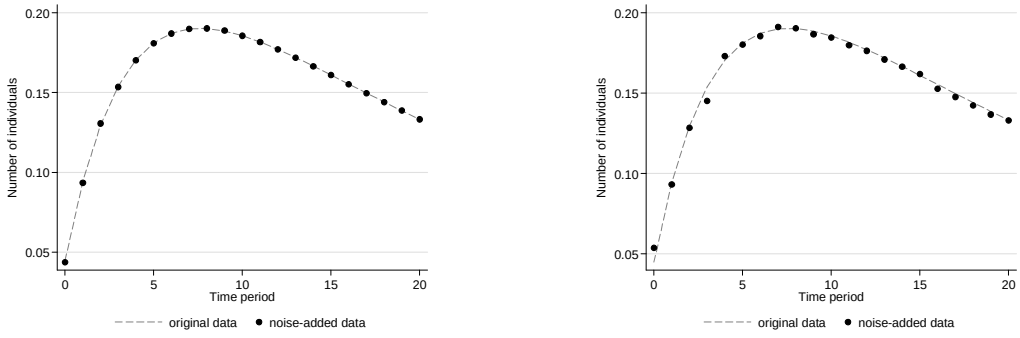
USA	100-250	250-500	500-1000	1000+
100-250	**0.9597**	**0.0403**	**0.0000**	**0.0000**
	(0.0173)	(0.0173)	(0.0000)	(0.0000)
250-500	**0.0353**	**0.9174**	**0.0472**	**0.0000**
	(0.0443)	(0.0593)	(0.0213)	(0.0000)
500-1000	**0.0000**	**0.0347**	**0.9183**	**0.0469**
	(0.0028)	(0.0427)	(0.0597)	(0.0214)
1000+	**0.0000**	**0.0000**	**0.0043**	**0.9957**
	(0.0000)	(0.0003)	(0.0052)	(0.0055)

[a] *The categories for the EU-15 are based on the economic size of the farms measured in European Size Unit (ESU); the categories for the USA are based on the economic size of farms measured in 1000 US$ of gross sales; the functional forms of the underlying relative size change probability distributions are both Weilbull with parameters $\ln(\lambda) = 0.0194$ and $\ln(\nu) = 3.3190$ for EU-15 and $\ln(\lambda) = 0.0255$ and $\ln(\nu) = 2.8413$ for the USA (see Table 4); in each cell, the bold figure is the estimated transition probability and the figure in brackets its associated standard deviation, both resulting from a Monte Carlo simulation with 200 draws (see text for further explanation); shaded cells indicate coefficients which are significantly different from zero at the 5% level at least.*

Source: author's estimates

Fig. 1. The impact of the added-noise intensity [a]

a) Noise intensity $\alpha = 0.1$ b) Noise intensity $\alpha = 1.0$



[a] *Category (0;2] of the 10-category aggregation and replication #1*

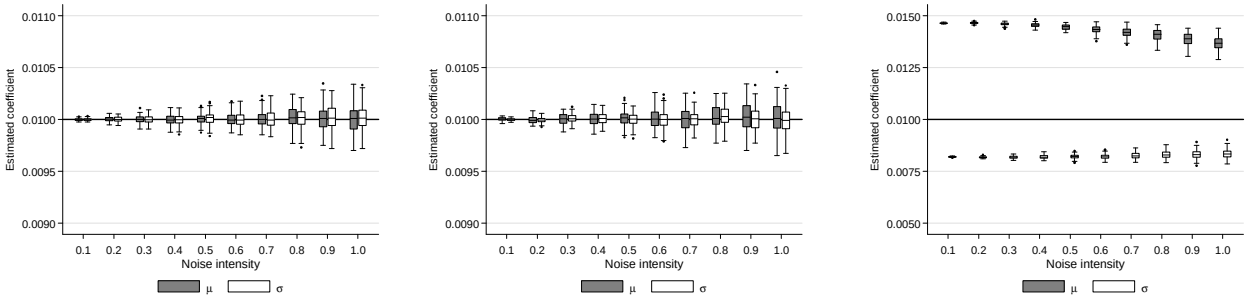
Source: author's calculations

Fig. 2. Impacts of the added-noise intensity and of the aggregation level on the estimation of the parameters μ and σ [a]

a) 10-category data with noise

b) 5-category data with noise

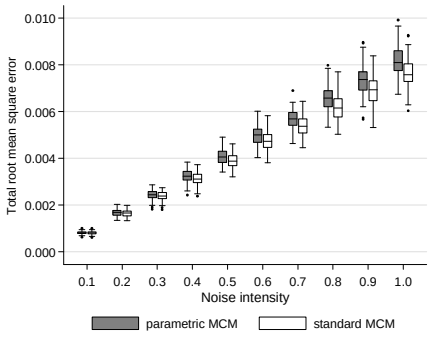
c) 5-category aggregation of 10-category data with noise



[a] *The box, whiskers and outliers have the standard definitions and summarize the results of 100 replications; the bold horizontal lines are set to the true values of μ and σ.*

Source: author's estimates

Fig. 3. Quality of the adjustment for the parametric *vs.* the standard MCM

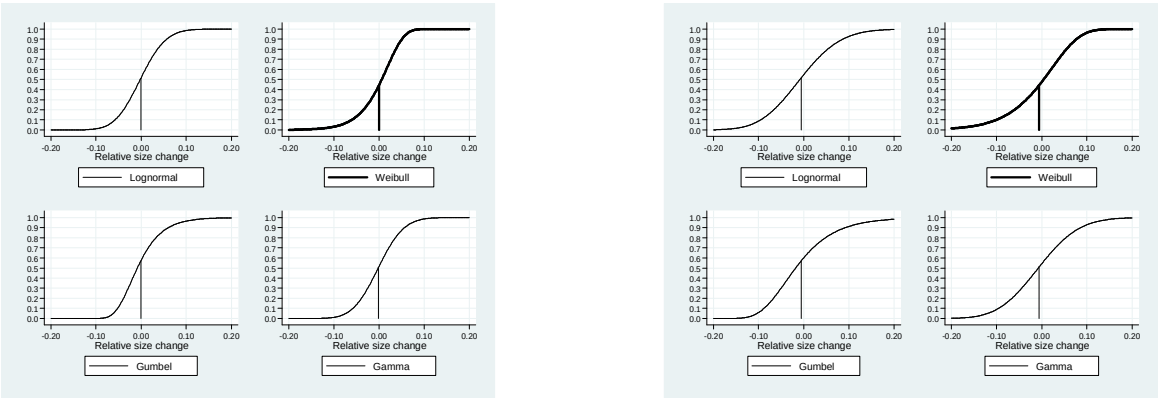


Source: author's estimates

Fig. 4. Estimated cumulative distributions of the relative size change probability for EU-15 and the USA for various functional forms [a]

a) EU-15

b) USA



[a] *The parameters defining the distributions are given in Table 4; vertical bars mark the resulting average relative size change; in either panel, the bold distribution is the one with the lowest associated total root mean square error (TRMSE).*

Source: author's estimates