# A structural approach to the Markov chain model with an application to the commercial French farms

Laurent Piet

# A structural approach to the Markov chain model with an application to the commercial French farms

**Laurent PIET**[1,2]

[1] **INRA, UMR1302 SMART, 35000 Rennes, France**

[2] **Agrocampus, UMR1302 SMART, 35000 Rennes, France**

**Adresse de correspondance : INRA, UMR1302 SMART, 4 allée Adolphe Bobierre, CS 61103 – F-35011 Rennes cedex – laurent.piet@rennes.inra.fr**


**4èmes journées de recherches en sciences sociales**

9 & 10 décembre 2010 – RENNES, France

# A structural approach to the Markov chain model with an application to the commercial French farms

Laurent PIET[1,2,*]

August 3, 2010

**Abstract**

The number and size distribution of farms are, among others, strategic control variables for public policy makers who wish to assess *ex-ante* the impact of the agriculture related policies they design. Among the various methods used in the academic literature, the Markov chain model (MCM) has become one of the most popular tool 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 the summing constraints). The alternative structural MCM I have developed 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 a richer information on the demographic processes at hand (size transitions, entries and exits). The empirical application of the model to the French strand of the Farm Accounting Data Network (FADN) shows that the structural MCM is well supported by the data and competes with the traditional approach without any significant shortcoming; moreover, it leads to the same kind of stylized facts but further permits to derive statistical indicators on the distribution of entries and exits which may interest the practitioner. A projection of the population of commercial French farms to the 2020 horizon is also presented.

*Keywords:* *Agriculture, Number and size distribution of firms, Markov Chain Model, FADN, France*

*JEL Classification:* *Q12, C15, C53*

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

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

* Corresponding address : INRA, UMR1302 Unité Mixte de Recherche SMART, 4 allée Adolphe Bobierre, CS 61103 – F-35011 Rennes cedex – tél. : 02.23.48.53.83 – fax. : 02.23.48.53.80 – laurent.piet@rennes.inra.fr

## 1. INTRODUCTION

Having a good knowledge of the population to which policy measures they design will apply is a key information for public policy makers who whish to assess *ex-ante* the potential impacts of these policies. In that perspective, the number and size distribution of farms are, among others, strategic control variables in the particular field of policies related to the agricultural production.

Zimmermann *et al.* (2009) review the various methods used in the academic literature to forecast the number of farms; as noted by these authors, farms are usually grouped into a finite number of categories which are defined on the basis of one or several criteria such as the size of farms (be it structural or economic), their location, production orientation or intensity, legal status, etc. One of the most popular tool implemented in those works is the so-called Markov chain model (MCM) applied to a population of farms grouped relative to their size only. In this paper, I show that the way MCMs have been implemented by agricultural economists so far exhibits a feature which complicates its estimation and limits the scope of its use. The alternative Markov chain modelling approach I propose simplifies the estimation of the model and leads to a richer information and a wider application range of the model.

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 toward that category experienced by farms which where previously in any other category.[1] At each time step, these transitions occur only with a certain probability (only a fraction of individuals move from one category to another) and the task of the modeller is to estimate these transition probabilities somehow. 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 situation.[2] However, Lee *et al.* (1965) and Lee *et al.* (1977) showed that

[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.

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 (Piet, 2008; Zimmermann *et al.*, 2009).

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 quite large even when only a few categories are considered. Moreover, this number grows exponentially as the number of categories increases, since all the $n^2$ possible transitions, where $n$ is the number of categories, have to be taken into account and the corresponding probabilities to be estimated; actually, this number is limited to $n(n-1)$ since summation constraints apply, but the exponential growth rate remains.[3] 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 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 the kinf of quite 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

[3] As will be made more explicit in the next section, the transition probabilities for a particular category must sum to 1, meaning that all individuals in the category experience a transition, be it moving to another category or staying in the same one.

in the case of such strongly under-identified models.[4] 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.[5] 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 categorization criteria, is actually a continuous (size) variable. [6]

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

The rest of the paper is organised as follows. The next section presents the modelling framework, emphasizing on how it departs from and enriches the standard MCM approach. Section 3 describes the empirical application of the model to the French strand of the farm accounting data network (FADN) for the period 1981-2007: a stationary annual transition probability matrix is estimated and a projection up to 2020 is simulated. Finally, the last section discusses the results and draws several directions for future work.

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

[5] Of course, the so-called 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. More on non-stationarity will be said in the last section of the paper.

[6] 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 MODEL

### *2.1 The principles and the originality of the approach*

As in the standard MCM approach described in the previous section, the population under study is broken down into a finite number of categories $J$ on the basis of a quantitative and continuous variable $X$ so that the obtained partition is complete. Said differently, categories represent intervals which are defined by a lower and an upper bound that insure continuity and a complete coverage of the definition domain of the partitioning variable.

Denoting the number of individuals in the $j$-th category at time $t$ by $n_{j,t}$, the population follows the Markov chain process of degree 1 between two observation dates $t$ and $t+1$ given by:

$$n_{j,t+1} = \sum_{k=1}^{J} p_{kj} n_{k,t} + u_{j,t} \qquad (1)$$

where $p_{kj}$ is the probability for a individual in category $k$ to move to category $j$ in one time-period $\tau$ and $u_{j,t}$ is an iid error term; further, transition probabilities, which are the parameters to be estimated, are subject to the following constraints:

$$p_{kj} \geq 0 \qquad (2)$$

$$\sum_{j=1}^{J} p_{kj} = 1 \qquad (3)$$

Here, I assumed that these probabilities do not change over time; the MCM is thus said to be stationary.[7] All together, the set of probabilities $p_{kj}$ define the (square) transition probability matrix (TPM) $\mathbf{P} = \left(p_{kj}\right)$; in matrix notation, equation (1) can thus be written as $\mathbf{N}_{t+1} = \mathbf{N}_t \mathbf{P} + \mathbf{u}_t$, where $\mathbf{N}_t = \left(n_{1,t},...n_{j,t},...,n_{J,t}\right)$, $\mathbf{N}_{t+1} = \left(n_{1,t+1},...n_{j,t+1},...,n_{J,t+1}\right)$ and $\mathbf{u}_t = \left(u_{1,t},...u_{j,t},...,u_{J,t}\right)$ are row-vectors.

In practice, in order to ensure that equation (3) holds, an "exit" category is added, stating that some individuals may "disappear" between to dates (*i.e.*, exit the agricultural sector); similarly, an

[7] As already mentioned in footnote 5, stationarity issues will be discussed in section 4.

"entry" category usually allows to account for new comers. Here, I explicitly accounted for entries and exits by rewriting equation (1) as follows:

$$n_{j,t+1} = \sum_{k=1}^{J} \left(1 - \varphi.p_k^{ex}\right) p_{kj}^{tr} n_{k,t} + \phi.p_j^{in} \sum_{k=1}^{J} \varphi.p_k^{ex} n_{k,t} + u_{j,t} \qquad (4)$$

where $p_{kj}^{tr} \geq 0$ is again the transition probability from category $k$ to category $j$, $p_k^{ex} \geq 0$ is the probability for an individual in category $k$ to exit the sector between $t$ and $t+1$, $p_j^{in} \geq 0$ is the probability for an individual to enter the sector into category $j$ between $t$ and $t+1$, and $\varphi$ and $\phi$ are scale parameters. This first part of the right hand side of equation (4) states that only the farms which did not exit may experience a change in $X$; the second part states that the number of individuals who enter the sector in each category represent a fraction of the total number of individual who exited, a formulation close to the "pool approach" adopted by Stokes (2006). The scale parameter $\phi$ determines whether the population is globally stationary ($\phi = 1$), expanding ($\phi > 1$) or shrinking ($\phi < 1$); the meaning of $\varphi$ will be explained in sub-section 2.2.

Note however that since we do not use micro-economic data and hence have no information regarding individual movements that could be used in the estimation of equation (4), $p_k^{ex}$ and $p_j^{in}$ actually are "absolute" and net exit and entry probabilities. Absolute because the overtaking of a previously existing farm by a new (*i.e.* previously un-existing) farmer –with or without a concomitant increase or decrease in the size of the farm– is treated as a single size transition and not as one exit plus one entry; then i) "absolute" exits correspond to situations where a farmer stops his activity and is not replaced by a new one, even if the land of his farm is taken over by one or more already active farmers –the number of farms actually decreases– and ii) "absolute" entries correspond to situations where a new farm settle either on previously un-operated land or on land which was previously by one or several other farmers who remain active anyway –the number of farms actually increases.[8] Net because, for a given size category, exit and entry cannot be separately

[8] For instance, this could correspond to a situation where a son would create a new farm by settling on part of his father's land.

identified. For both these reasons, entry and exit concepts used here do not directly relate to their common sense definition, a feature which has to be kept in mind when interpreting the estimation results.

So far, the assumptions made here are the same as in the standard MCM implementation. In particular, it is rarely stressed in the literature that this setting assumes the probabilities to be identical from one individual to the other and the individual transitions to be independent from each other. The strength of formulation (4) is that the "conservation" constraint of the Markov process imposes that equation (3) be now replaced by the following set of summing conditions:

$$\sum_{j=1}^{J} p_{kj}^{tr} = 1 \tag{5}$$

$$\sum_{j=1}^{J} p_{j}^{ex} = 1 \tag{6}$$

$$\sum_{j=1}^{J} p_{j}^{in} = 1 \tag{7}$$

Then, together with the fact that the variable defining the partition $J$ is a continuous variable, $p_{kj}^{tr}$, $p_{k}^{ex}$ and $p_{j}^{in}$ can all be regarded as generated from any suitable probability density function; this is where the originality of my MCM approach lies with respect to the standard one: introducing some structural information into the model thanks to the use of probability functional forms, rather than estimating each and every probability as almost independent parameters (up to the summing constraints). Though this strategy can appear less flexible at first glance, I see it as outclassing the standard implementation (provided the fit to empirical data is satisfactory) for the following reasons:

- it is parsimonious in terms of parameters: instead of being $J(J-1)$ as in the standard approach, the number of parameters to estimate depends on the functional forms chosen for the three probability distributions; in its simplest expression, assuming that each of these distributions are fully determined by two parameters, the total number of parameters is eight (three distributions times two parameters plus $\varphi$ and $\phi$) and is in this

case independent from the number of categories; on the contrary, on can expect that the more the categories, the more robust the estimation;

- then, with a limited number of parameters, the chance that enough empirical observations are available to build a well-posed problem is greater, so that simple econometric methods can be used to estimate the model;
- except for the choice of the three functional forms, no *a priori* constraint or knowledge is needed regarding impossible or implausible transitions as is the case in the standard approach; improbable transitions will "endogenously" derive from the empirically estimated forms of the distributions;
- probability *distributions* are estimated, not only *discrete* probabilities, leading to a richer information on the transition, exit and entry processes themselves; in particular, one can project the initial population into category intervals defined with upper and lower bounds which can differ from the initial ones; the number of ending categories not even needs to be the same so that the corresponding TPM will no longer be square; actually, any ex-post TPM can be constructed once the model is estimated.

In sum, all of the four issues listed in the introduction as shortcomings to the standard approach are efficiently addressed.

### *2.2 The structural model*

Expressing $p_{kj}^{tr}$, $p_k^{ex}$ and $p_j^{in}$ as deriving from probability distributions leads to reset the model expressed by equation (4) in the following way.

Instead of identifying categories thanks to their indices, we shall rather consider intervals defined over specific ranges of the dependent variable $X$ : with our previous notations, the "initial" category $k$ will be denoted by the interval $[\underline{x}_k, \bar{x}_k)$ and the "final" category $j$ by the interval $[\underline{x}_j, \bar{x}_j)$. Then, the model is expressed as:

$$n_{[\underline{x}_j,\bar{x}_j),t+1} = \sum_{[\underline{x}_1,\bar{x}_1)}^{[\underline{x}_J,\bar{x}_J]} \left(1-\varphi.p^{ex}_{[\underline{x}_k,\bar{x}_k)}\right) p^{tr}_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)} n_{[\underline{x}_k,\bar{x}_k),t} + \phi.p^{in}_{[\underline{x}_j,\bar{x}_j)} \sum_{[\underline{x}_1,\bar{x}_1)}^{[\underline{x}_J,\bar{x}_J]} \varphi.p^{ex}_{[\underline{x}_k,\bar{x}_k)} n_{[\underline{x}_k,\bar{x}_k),t} + u_{[\underline{x}_j,\bar{x}_j),t} \quad (8)$$

The probabilities $p^{tr}_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)}$, $p^{ex}_{[\underline{x}_k,\bar{x}_k)}$ and $p^{in}_{[\underline{x}_j,\bar{x}_j)}$ are given by:

$$p^{tr}_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)} = \frac{1}{\bar{x}_k - \underline{x}_k} \int_{\underline{x}_k}^{\bar{x}_k} \left[ F^{tr}\left(\frac{\bar{x}_j}{x};\boldsymbol{\theta}^{tr}\right) - F^{tr}\left(\frac{\underline{x}_j}{x};\boldsymbol{\theta}^{tr}\right)\right] dx \quad (9)$$

$$p^{ex}_{[\underline{x}_k,\bar{x}_k)} = F^{ex}\left(\bar{x}_k;\boldsymbol{\theta}^{ex}\right) - F^{ex}\left(\underline{x}_k;\boldsymbol{\theta}^{ex}\right) \quad (10)$$

$$p^{in}_{[\underline{x}_j,\bar{x}_j)} = F^{in}\left(\bar{x}_j;\boldsymbol{\theta}^{in}\right) - F^{in}\left(\underline{x}_j;\boldsymbol{\theta}^{in}\right) \quad (11)$$

where $F^{tr}()$, $F^{ex}()$ and $F^{in}()$ are the distribution functions characterizing the corresponding transitions, and $\boldsymbol{\theta}^{tr}$, $\boldsymbol{\theta}^{ex}$ and $\boldsymbol{\theta}^{in}$ the vectors of parameters unambiguously defining these distributions. As can be seen from equation (9), a transition is now expressed as a relative change in the dependent variable $X$ rather than a "simple" move from an initial to a final category. Two remarks must be made regarding the specification of the transition probability $p^{tr}_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)}$ given by equation (9).

Fist, the vector of parameters $\boldsymbol{\theta}^{tr}$ is assumed to be independent from the interval $[\underline{x}_k,\bar{x}_k)$: the probability for an individual initially exhibiting a level $X = x_k$ to experience a relative change of, say, $x_j/x_k$, is independent of the initial value $x_k$. This can sound like a strong assumption; whether it is supported by the data or not is an empirical question that will be addressed in the next section. So far, this assumption allows to express the model in its most parsimonious form as discussed earlier and it can be relaxed in two ways as is discussed in section 4.

Second, as expressed by equation (9), $p^{tr}_{[\underline{x}_k,\bar{x}_k)[\underline{x}_j,\bar{x}_j)}$ is actually the *average* probability for an individual initially lying in $[\underline{x}_k,\bar{x}_k)$ to move to $[\underline{x}_j,\bar{x}_j)$; the true transition probability for an individual initially exhibiting a level $X = x_k \in [\underline{x}_k,\bar{x}_k)$ would reduce to $p^{tr}_{x_k[\underline{x}_j,\bar{x}_j)} = F^{tr}\left(\frac{\bar{x}_j}{x};\boldsymbol{\theta}^{tr}\right) - F^{tr}\left(\frac{\underline{x}_j}{x};\boldsymbol{\theta}^{tr}\right)$. Doing so implicitly assumes that individuals are uniformly

distributed inside each interval $[\underline{x}_k, \bar{x}_k)$; it implies that a scale parameter, $\varphi$, has to be added to the model. Working with macro, aggregated, data and in the absence of more precise knowledge upon the true underlying population distribution, this is the simplest assumption to do.[9]

Eventually, assumptions must be formed regarding the functional forms of $F^{tr}()$, $F^{ex}()$ and $F^{in}()$. There is no *a priori* constraint on this choice but, in practice, the modeler will retain general enough forms so as to best fit the data. Note however that, in the general case, there is no guarantee that a closed form analytical solution exists for the integral appearing in (9). When relevant in empirical applications a numeric approximation of this integral can be performed, through a simple trapeze formula for instance, where the implied bias can be rendered as small as desired (at the expense of computation time).

The next section illustrates the implementation of the structural MCM defined by equation (8) and constraints (9) to (11) with an application to the population of commercial French farms.

## 3. AN APPLICATION TO THE COMMERCIAL FRENCH FARMS

### *3.1 The data used*

The data used in this application come from the French strand of the Farm Accountancy Data Network (FADN), an accountancy and technico-economic survey carried across whole European Union over a sample of agricultural holdings which are considered as commercial.[10]

The FADN sample is stratified using three criteria: region, type of farming and economic size. Within each stratum, a set of individuals is drawn pseudo-randomly from the population and each of the sample's farms is assigned an extrapolation coefficient based on its representativeness within

[9] Note that, in the standard MCM approach, each transition probability $p_{kj}$ can also be regarded as an average probability; but in this case, this has no direct and formal implication on the underlying population distribution.

[10] For a detailed presentation of the FADN survey: http://ec.europa.eu/agriculture/rica/index_en.cfm

the stratum knowing the total number of commercial farms present every year (Rouquette and Baschet, 2010). Yet this total is only available for certain years (Fall *et al.*, 2010): in France, it is updated each time a Farm Census or Farm Structure Survey (FSS) is issued, that is, every 2 or 3 years only. The weighting allocated to each individuals in the FADN sample is then calculated on the same frequency so that the total number of farms extrapolated from the entire sample is consistent with the known total for the corresponding years; for the years in between, these weights are determined such that the total extrapolated number of farms does not departs too much from the number found by the most recent census or FSS, all the while checking that certain aggregate economic variables such as total output value are consistent with the data in the National Agricultural Accounts for the corresponding year. As a consequence, the numbers of individuals that can be recovered from the FADN data, be them for the whole population or for sub-classes of this population, do not evolve smoothly across years but rather step-wise.

Therefore, "intermediate" years are not included in the following analysis so that the data used correspond to eleven years only: 1981, 1983, 1985, 1988, 1990, 1993, 1995, 1997, 2000, 2005 and 2007.[11] In order to estimate annual transitions the numbers of farms in each category for "intermediates" years (incl. 1980) were linearly interpolated from the list of years actually used.

Furthermore, note that the FADN observations are not panel data since farms can join or leave the sample every year for non purely demographic reasons; no explicit information is therefore available in the dataset regarding entries and exits.

For the purpose of illustration, the whole population of FADN farms were grouped into five categories, depending on their size as measured in operated hectares, defined by the following intervals: 0 to 19.99 hectares, 20 to 49.99 hectares, 50 to 99.99 hectares, 100 to 199.99 hectares and 200 or more hectares. The corresponding number of farms in each category is presented in Table 1 (the years which have been actually used appear in bold characters). As a whole, the data consist of a set of 26 transitions corresponding to 130 observation points.

[11] Due to technical reasons, the FADN coefficients were not updated in 2003 to reflect the 2003 FSS. Hence 2003 is also excluded from our list of "observed" years.

*[insert Table 1 around here]*

### *3.2 Choosing a functional form for the transition, entry and exit probability distributions*

Three functional forms for $F^{tr}(\,)$, $F^{ex}(\,)$ and $F^{in}(\,)$ were tested:

- the lognormal distribution: $L(u;\mu,\sigma)=\Phi(\ln(u);\mu,\sigma)$ where $\Phi(\mu,\sigma)$ is the normal distribution with mean $\mu$ and standard deviation $\sigma>0$, and $u>0$;
- the gamma distribution: $\Gamma(u;\theta,\kappa)$ with scale $\theta>0$ and shape $\kappa>0$, and $u\geq 0$;
- and the Weibull distribution: $\Omega(u;\nu,\lambda)$ with scale $\nu>0$ and shape $\lambda>0$, and $u\geq 0$.

All of these distributions share two interesting features: (i) they are fully defined by two parameters only so that the model encompasses eight parameters as a whole as previously explained and is therefore a well-posed problem; (ii) they can generate a wide range of distributions shapes, from symmetric ones to (potentially highly) skewed ones. The model was further designed so that the functional forms chosen for $F^{tr}(\,)$, $F^{ex}(\,)$ and $F^{in}(\,)$ could differ from one distribution to the other.

Finally, as the model depicted in equation (8) consists in a system of $J$ simultaneous equations, it was solved using the non-linear SUR (*nlsur*) procedure of the Stata 11.0 software.

### *3.3 Results*

Results showed that the most satisfactory combination of functional forms was to retain a lognormal distribution for $F^{tr}(\,)$ and Weibull distributions for both $F^{ex}(\,)$ and $F^{in}(\,)$; the gamma distribution was out-performed in any case (results not reported here).[12] The estimation results corresponding to the lognormal-Weibull-Weibull assumption are reported in Table 2. The resulting point-estimate TPM and exit and entry vectors can be derived directly from the estimated

[12] Actually, there was no significant difference on the results when choosing a lognormal distribution for $F^{in}(\,)$; for the sake of homogeneity between entries and exits, it was chosen to present the results with the Weibull distribution.

coefficients. Unfortunately, the standard deviations associated with these estimated transition, exit and entry probabilities are not easily analytically computable from equations (9) to (11); to overcome this problem, I implemented a 1000-draws Monte Carlo simulation using these equations and the coefficients and standard deviations of Table 2. The resulting average probabilities and associated standard deviations are reported in Table 3 (the point-estimate probabilities are not reported but are very close to the simulated averages).[13]

*[insert Table 2 and Table 3 around here]*

The first result worth noticing is that the $R^2$ associated with each of the simultaneous equations are all above 0.99: the model fits the data very well. In order to compare the structural MCM approach with the standard one, I estimated in parallel a constrained multinomial logit (MNL) with the same data in line with the method found in (Zepeda, 1995).[14] Both models quite compare in terms of adjustment to the data even if the standard approach slightly outperforms the structural model when examining partial and total root mean square errors (RMSE): the total RMSE is 10,929.10 with the structural MCM when it is only 9,744.90 with the constrained MNL. The corresponding TPMs are quite different though[15] and it is then a question for the analyst to decide which one seems the more plausible (or to use both in a sensitivity analysis when simulating projections).

The second result is that both $\mu^{tr}$ and $\sigma^{tr}$ are very close to zero, meaning that most probably French commercial farms experience no significant relative size change from one year to the other. This does not mean no change at all though, as can be seen in the resulting TPM: even if the matrix

[13] For convenience, the matrix presented in Table 3 is a "concatenation" of the transition matrix, the exit vector and the entry row-vector; even if, strictly speaking, the term TPM was defined in the previous sections as a subset of it only, I call this extended matrix "the TPM" in the following.

[14] As is often assumed, the constraints I imposed in the MNL imply that farms cannot move from more than one category at each step; the usual summing-up constraints also apply; as a results, 16 parameters were to be estimated.

[15] The constrained MNL TPM is not reported here; it is available from the author upon request.

is highly diagonal as is usually the case for annual transitions in the literature, some off-diagonal elements are significantly different from zero. Moreover, as was mentioned at the end of section 2.1, some transitions appear implausible but this results from the estimated parameters only, not from some *a priori* or arbitrary input to the model.

Third, the coefficient $\phi$ is strictly less than 1 ($\hat{\phi} = 0.055$), confirming the general trend of a declining population. Absolute net exit rates are at most around 5% of the population in each category (4.9% for the "0-19.99 ha" category; 5.1% for the "20-49.99 ha" category and 1.8% for the "50-99.99 ha" category); in the mean time, the sole significant absolute net entry probability is associated with the "100-199.99 ha" category (4.7%). All of the previous figures may look underestimated at first glance. They are no longer surprising when one recalls the explanation given in the section 2.1 regarding the true meaning of these "absolute" probabilities. Moreover, note that the non-overlapping of the entry and exit distributions confirms that these are net probabilities. Altogether, these results however reflect the usual feature that exits occur at a size lower than the average and entries at a size higher than the average.

But the strength of this structural model is also, as already mentioned, to bring more information than the standard MCM implementation: various distributional indicators can be easily derived from the chosen functional forms and the corresponding estimated coefficients. For example, the median of a Weibull distribution is given by $m = \nu(\ln(2))^{1/\lambda}$, its mean by $\mu = \nu.\Gamma(1+1/\lambda)$ and its standard deviation by $\sigma^2 = \nu^2.\Gamma(1+2/\lambda) - \mu^2$ (Kleiber and Kotz, 2003), where $\Gamma$ is the gamma function; with the estimated coefficients listed in Table 2, this means that:

- 50% of the absolute net exits were operating less than 33 ha before leaving the sector; the average absolute net exit size is estimated at 39 ha with a standard deviation of 28 hectares (precise figures are 33.36 ha, 39.04 ha and 27.50 ha respectively);
- 50% of the absolute net entries settled on more than 196 ha with an average of 195 ha and a standard deviation of 6 ha (precise figures are 196.28 ha, 195.41 ha and 5.68 ha respectively).

To sum up, the results show that the structural MCM I propose (i) fits the data well and to an extent that compares with the standard approach, (ii) drives to the same kind of conclusions regarding stylized facts which agricultural experts are familiar with and (iii) allows to derive a richer information useful to the practitioner. It can now be used for projection studies; this is the subject of the next sub-section.

### *3.4 Projection at 2020*

Once its parameters have been estimated, the structural MCM can be used in a standard way to forecast the state of the population at some given horizon: one simply applies iteratively equation (8) to the desired initial year using the TPM derived from the estimated coefficients.

I have simulated the distribution of the population of commercial French farms at the 2020 horizon in this manner, using each of the available observation year as a potential initial date. The result is reported in Table 4. As the model is iterated on an annual basis, intermediate figures are also available and the resulting path followed by the total number of farms until 2020 is presented in Figure 1.

*[insert Table 4 and Figure 1 around here]*

Provided that the average demographic and economic conditions that pertained over the estimation period (1981-2007) hold constant until then, the total number of commercial farms in France is estimated to decrease from 326,008 in 2007 to a little more than 233,300 individuals in 2020, or a 28% decline in 13 years; the annual rate of decrease (-2.5% per year) would thus be quite the same as observed on the most recent years (-2.3% between 2000 and 2007) and lower than observed on the previous decades (-3.9% between 1981 and 1990; -3.0% between 1990 and 2000). Looking at the influence of the starting year of the simulation, it appears that the total number of commercial farms could be lying between 217,100 and 249,200, a “confidence interval” whose range amounts for around 14% of the average value.

The model forecasts that both the number and share of smaller farms ("0-19.99" ha) will fall at an increased rate (-10.0% per year from 2007 to 2020 as compared to -4.3% per year for the period 2000-2007) while bigger farms population and share will continue to increase, though at a slower rate (+2.8% per year from 2007 to 2020 as compared to +4.5% per year for the period 2000-2007). However, the sensitivity analysis shows that, globally, the more recent the starting year, the higher the number of smaller and bigger farms; still, considering the "middle point estimates" as resulting from the simulations with the 1990-1995 initial years implies that the actual number (and share) of smaller farms could be even lower than given by the average. Finally, the "middle" sized farms ("50-99.99 ha") would still be the more numerous ones in 2020 but should be ousted by the above category ("100-199.99 hectares") soon after.

## 4. 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 summing 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. Along with a re-writing of the Markov process which explicitly accounts for absolute entry and exit, it allows to express the transition, entry and exit probabilities as deriving from underlying probability distributions, for which several functional forms can be chosen and tested. In this sense, the approach I proposed can be said "structural". Moreover, the transition probabilities which are estimated do not simply represent the likelihood to move from one category to the other (eventually

the same), but a richer and more interesting concept: the likelihood to experience a given relative change in the dependent variable.

From the practitioner's point of view this structural MCM outperforms the standard approach on four grounds: (i) it is more parsimonious in terms of parameters; (ii) as a consequence, rather standard and efficient econometric techniques can be employed; (iii) no assumption has to me formed on specific probabilities but rather on their overall shape; and (iv) it reveals a richer information on the underlying demographic processes at hand. From the empirical analyst's point of view, this structural MCM is well supported by the French FADN data used in the proposed empirical application; at least, it competes with the traditional approach without any significant shortcoming and reproduces the same and usual stylized facts. Two direct extensions of this work can be envisaged: on the one hand, the European Union-wide homogeneity of the FADN database should make it straightforward to test the method in other national contexts; on the other hand, the availability of Farm Structures Surveys should allow to easily include non-commercial farms into the analysis. On a broader perspective, the required data are fairly common and usually issued on a regular basis in most developed, transition and even some developing countries.

Yet, as any model, the proposed method relies on a set of assumptions. Actually, the main difference with the standard approach in this respect is the replacement of *several* assumptions on the *magnitude* of transition probabilities into *only three* assumptions regarding the *shape* of their distribution. As previously noted in section 2, other important assumptions made here deal with two issues and should be relatively easy to relax as will be now explained.

First, it was assumed that the parameters of the transition probability distribution function were independent from the initial value of the dependent variable; this could be relaxed by two means: (i) these parameters could be made dependent on the initial *category* to which they apply; the drawback of this simple solution is to increase sharply the number of parameters of the model

and to directly relate this number to the number of categories;[16] still this correlation is linear and not exponential as in the standard approach; (ii) a statistical relationship between the parameters and the initial *value* of the variable could be specified, adding more structure into the model; on the one hand, choosing the 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; on the other hand, this would intuitively require to have at one's disposal a more detailed information regarding the distribution of the dependent variable among the population or would imply further assumptions regarding it. Anyway, either ways of relaxing this assumption would be interesting since in both cases the statistical dependence could be rigorously tested.

Second, the model presented here is stationary in the sense that transition, exit and entry probabilities do not change over time. There again, two directions are possible to evolve toward a non-stationary model: (i) as proposed by Jongeneel and Tonini (2008), several successive TPMs could be estimated; but this simple solution is not truly dynamic and the amount of needed data is largely increased for a preserved robustness of the estimations; (ii) as is the case in the most recent published papers which use the standard MCM approach (Zepeda, 1995; Karantininis, 2002; Stokes, 2006; Tonini and Jongeneel, 2008), a really non-stationary version of the structural MCM could be built by making the distribution parameters depend on time-varying covariates; a simple trend would then 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 cited references. But then, the structural approach proposed here would face an important increase in the number of parameters, as its standard counterparts do. Simple SUR estimation procedures would be certainly no longer applicable. Yet the structural approach would still be the more parsimonious of the two methods so that more degrees of freedom would be

[16] In the empirical application presented in section 3, this would have led to estimate ten (2x5) transition parameters plus two exit and two entry parameter plus $\varphi$ and $\phi$, or a total of 16.

preserved for an undoubtedly more robust covariate effects estimation. Moreover, not only the full set of transitions could be studied (in most of the non-stationary literature, the impact of covariates is studied for only a subset of arbitrarily said “interesting” individual or aggregated transitions) but there again the derived information on the effect of market and policy would be richer for the same reasons as previously explained.

Finally, when confronted to the estimation of such an increasing number of parameters, it seems to me that, from an econometric perspective, a Bayesian inference approach would be more appropriate than the GCE or IV-GCE techniques used so far, both in terms of parameters estimation and of structural model (functional forms) selection. This sounds also like a promising direction for future research.

## 5. 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.

**Fall, M., L. Piet and M. Roger (2010).** Trends in the French commercial farm population. *Review of Agricultural and Environmental Studies / Revue d'Etudes en Agriculture et Environnement* forthcoming.

**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.

**Jongeneel, R. and A. Tonini (2008).** Dairy Quota and Farm Structural Change: A Case Study on the Netherlands. 107th Seminar of the European Association of Agricultural Economists, Sevilla(Spain), January 29-February 1, 2008, 18 p.

**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.

**Kleiber, C. and S. Kotz (2003).** *Statistical size distributions in economics and actuarial sciences*. D. J. Balding *et al.* eds., John Wiley and Sons, Hoboken (New Jersey).

**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).

**Piet, L. (2008).** The evolution of farm size distribution: revisiting the Markov chain model. XIIth Congress of the European Association of Agricultural Economists, Gent (Belgium), August 26-29, 2008, 10 p.

**Rouquette, C. and J.-F. Baschet (2010).** Le réseau d'information comptable agricole (RICA). *Analyse Centre d'Etudes et de Prospective* 23.

**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. Number of commercial farms for each of the five categories and for the whole population in the French strand of the FADN. [a]

| Years | Number of farms grouped by their size in terms of Used Agricultural Area (hectares) | | | | | Total |
|---|---|---|---|---|---|---|
| | 0-19.99 | 20-49.99 | 50-99.99 | 100-199.99 | >200 | |
| 1980 | 258,992 | 392,403 | 113,480 | 24,074 | 202 | 789,151 |
| **1981** | **233,707** | **375,147** | **112,722** | **23,386** | **813** | **745,775** |
| 1982 | 208,422 | 357,891 | 111,964 | 22,699 | 1,424 | 702,400 |
| **1983** | **183,137** | **340,635** | **111,206** | **22,011** | **2,035** | **659,024** |
| 1984 | 178,488 | 329,173 | 112,820 | 23,926 | 2,086 | 646,492 |
| **1985** | **173,839** | **317,710** | **114,433** | **25,841** | **2,136** | **633,959** |
| 1986 | 161,405 | 299,409 | 117,367 | 28,394 | 2,538 | 609,113 |
| 1987 | 148,972 | 281,108 | 120,301 | 30,946 | 2,939 | 584,266 |
| **1988** | **136,538** | **262,807** | **123,235** | **33,499** | **3,341** | **559,420** |
| 1989 | 126,187 | 248,286 | 124,979 | 36,453 | 4,629 | 540,532 |
| **1990** | **115,835** | **233,764** | **126,723** | **39,406** | **5,916** | **521,644** |
| 1991 | 108,175 | 215,480 | 127,346 | 43,796 | 6,715 | 501,513 |
| 1992 | 100,515 | 197,197 | 127,969 | 48,187 | 7,514 | 481,381 |
| **1993** | **92,855** | **178,913** | **128,592** | **52,577** | **8,313** | **461,250** |
| 1994 | 87,007 | 164,787 | 129,292 | 54,584 | 9,378 | 445,047 |
| **1995** | **81,158** | **150,661** | **129,991** | **56,591** | **10,443** | **428,844** |
| 1996 | 76,536 | 140,979 | 128,814 | 59,658 | 11,252 | 417,238 |
| **1997** | **71,914** | **131,296** | **127,637** | **62,725** | **12,060** | **405,632** |
| 1998 | 70,925 | 126,575 | 125,164 | 63,285 | 12,715 | 398,664 |
| 1999 | 69,936 | 121,853 | 122,691 | 63,846 | 13,370 | 391,696 |
| **2000** | **68,947** | **117,132** | **120,218** | **64,406** | **14,025** | **384,728** |
| 2001 | 66,423 | 110,945 | 119,237 | 65,940 | 14,481 | 377,026 |
| 2002 | 63,900 | 104,759 | 118,256 | 67,474 | 14,937 | 369,325 |
| 2003 | 61,376 | 98,572 | 117,274 | 69,007 | 15,393 | 361,623 |
| 2004 | 58,852 | 92,385 | 116,293 | 70,541 | 15,849 | 353,921 |
| **2005** | **56,329** | **86,199** | **115,312** | **72,075** | **16,305** | **346,219** |
| 2006 | 53,500 | 81,075 | 112,075 | 71,766 | 17,698 | 336,114 |
| **2007** | **50,671** | **75,951** | **108,837** | **71,457** | **19,091** | **326,008** |

[a] *FADN original data appear in bold font; normal font denotes interpolated data (see text for further explanation)*

Source: author's calculations based on FADN data

Table 2. Estimation results of the model defined by equations (8) to (11) with a lognormal distribution for $F^{tr}()$ and Weibull distributions both for $F^{ex}()$ and $F^{in}()$. [a]

| | Coefficient | Std. error | z | P>\|z\| |
|---|---|---|---|---|
| $\mu^{tr}$ | 0.0050 | 0.0002 | 21.28 | 0.000 |
| $\sigma^{tr}$ | 0.0082 | . | . | . |
| $\varphi$ | 3.6137 | 0.3252 | 11.11 | 0.000 |
| $\nu^{ex}$ | 43.0244 | 3.1625 | 13.60 | 0.000 |
| $\lambda^{ex}$ | 1.4414 | 0.1675 | 8.60 | 0.000 |
| $\phi$ | 0.0555 | 0.0188 | 2.95 | 0.003 |
| $\nu^{in}$ | 197.9466 | 1.3601 | 145.54 | 0.000 |
| $\lambda^{in}$ | 43.3801 | . | . | . |

| **SUR estimation results** | **$R^2$** | **RMSE** |
|---|---|---|
| equation 1 (« 0-19.99 ha ») | 0.9989 | 4045.03 |
| equation 2 (« 20-49.99 ha ») | 0.9997 | 3427.83 |
| equation 3 (« 50-99.99 ha ») | 0.9998 | 1548.31 |
| equation 4 (« 100-199.99 ha ») | 0.9991 | 1566.07 |
| equation 5 (« >200 ha ») | 0.9990 | 341.86 |

[a] *RMSE stands for root mean square error.*

Source: author's estimates

Table 3. Transition probability matrix resulting from the coefficients reported in Table 2. [a,b]

| $t$ \ $t+1$ | 0-19.99 | 20-49.99 | 50-99.99 | 100-199.99 | >200 | exit |
|---|---|---|---|---|---|---|
| 0-19.99 | **0.9271** (0.0090) | **0.0235** (0.0002) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0494** (0.0091) |
| 20-49.99 | **0.0000** (0.0000) | **0.9331** (0.0082) | **0.0160** (0.0001) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0509** (0.0082) |
| 50-99.99 | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.9721** (0.0024) | **0.0099** (0.000) | **0.0000** (0.0000) | **0.0179** (0.0024) |
| 100-199.99 | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.9891** (0.0019) | **0.0096** (0.0017) | **0.0013** (0.0007) |
| >200 | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0000** (0.0000) | **1.0000** (0.0000) | **0.0000** (0.0000) |
| entry | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0000** (0.0000) | **0.0468** (0.0183) | **0.0119** (0.0064) | |

[a] *In each cell, the bold figure is the average transition probability and the figure in brackets its associated standard deviation, both resulting from a Monte Carlo simulation with 1,000 draws (see text for further explanation)*

[b] *Mean values appearing in shaded cells are significantly different from zero.*

Source: author's estimates

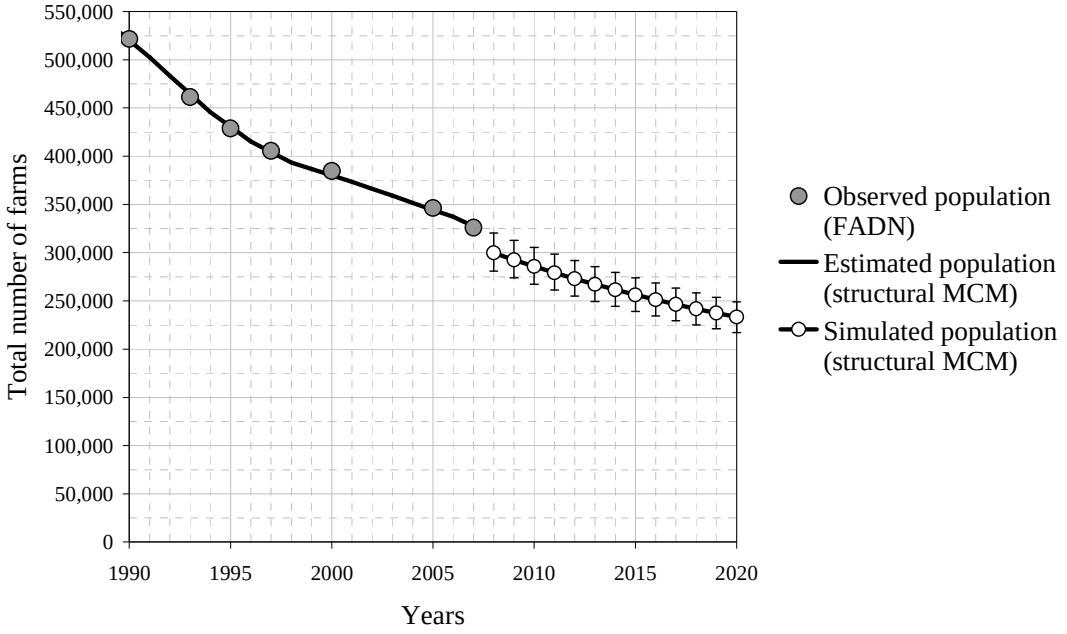
Table 4. Projected population distribution in 2020 using the point-estimate TPM derived from the coefficients reported in Table 2. [a]

| **Starting year** | Number of farms grouped by their size in terms of Used Agricultural Area (hectares) | | | | | **Total** |
|---|---|---|---|---|---|---|
| | **0-19.99** | **20-49.99** | **50-99.99** | **100-199.99** | **>200** | |
| 1981 | 11,446 | 37,379 | 84,724 | 68,818 | 26,891 | 229,258 |
| 1983 | 10,470 | 36,661 | 82,280 | 63,161 | 24,567 | 217,140 |
| 1985 | 11,601 | 38,973 | 84,047 | 64,072 | 23,974 | 222,668 |
| 1988 | 11,491 | 38,323 | 85,066 | 65,736 | 23,968 | 224,584 |
| 1990 | 11,380 | 38,154 | 85,924 | 67,712 | 25,978 | 229,149 |
| 1993 | 11,505 | 35,606 | 84,676 | 73,340 | 28,193 | 233,320 |
| 1995 | 11,738 | 34,227 | 85,092 | 74,427 | 29,164 | 234,647 |
| 1997 | 12,141 | 33,846 | 84,892 | 77,189 | 30,053 | 238,120 |
| 2000 | 14,680 | 36,794 | 84,512 | 76,086 | 29,601 | 241,672 |
| 2005 | 17,655 | 37,175 | 86,517 | 79,283 | 28,566 | 249,197 |
| 2007 | 18,539 | 36,907 | 84,541 | 77,155 | 29,464 | 246,606 |
| **Average** | **12,968** | **36,731** | **84,752** | **71,544** | **27,311** | **233,305** |
| (st. dev.) | (2,746) | (1,616) | (1,070) | (5,808) | (2,345) | (10,043) |
| *2007 pop.* | *50,672* | *75,951* | *108,837* | *71,457* | *19,092* | *326,008* |
| **2020 av. share** | **5.6%** | **15.7%** | **36.3%** | **30.7%** | **11.7%** | **100.0%** |
| *2007 share* | *15.5%* | *23.3%* | *33.4%* | *21.9%* | *5.9%* | *100.0%* |

[a] *Italic figures for 2007 are recalled or derived from Table 1 for comparison.*

Source: author's simulations

Figure 1. Projected total number of farms up to 2020 using the point-estimate TPM derived from the coefficients reported in Table 2. [a]




[a] *All the eleven "observed" years were used as starting dates for projection; thus, in the rightmost part of the graph, the hollow circles correspond to the average of the simulated populations and the vertical error bars are defined by the minimum and maximum simulated populations.*

Source: author's simulations