# Using Mixtures-of-Distributions models to inform farm size selection decisions in representative farm modelling.

by

Philip Kostov and Seamus McErlean

Working Paper, Agricultural and Food Economics, Queen's University Belfast. August 2003.

[a] Agricultural and Food Economics, School of Agriculture and Food Science, Queen's University Belfast, Newforge Lane, Belfast, BT9 5PX. Tel: +44 (0)28 9025 5237, Fax: +44 (0)28 9025 5327 e-mail: P.Kostov@qub.ac.uk

[b] Agricultural and Food Economics, School of Agriculture and Food Science, Queen's University Belfast, Newforge Lane, Belfast, BT9 5PX. Tel: +44 (0)28 9025 5621, Fax: +44 (0)28 9025 5327 e-mail: Seamus.McErlean@dardni.gov.uk

**Using Mixtures-of-Distributions models to inform farm size selection decisions in representative farm modelling.**

Abstract

The selection of 'representative' farms in farm level modelling where results are aggregated to the sector level is critically important if the effects of aggregation bias are to be reduced. The process of selecting representative farms normally involves the use of cluster analysis where the decision regarding the appropriate number of clusters (or representative farm types) is largely subjective. However, when the technique of fitting mixtures of distributions is employed as a clustering technique there is an objective test of the appropriate number of clusters. This paper demonstrates the MDM approach to cluster analysis by classifying dairy farms in Northern Ireland, based on the number of cows in each farm. The results indicate that four representative farms are needed, with a view to minimising aggregation bias, to describe the dairy sector in Northern Ireland.

JEL Classification: Q12

**Using Mixtures-of-Distributions models to inform farm size selection decisions in representative farm modelling.**

## 1. Introduction

Aggregation bias is a common problem in farm level modelling where results are aggregated to the sector level (often for the purposes of policy analysis). An important source of aggregation bias is the differences in resource endowments among farms. While this type of aggregation bias is difficult to eliminate completely it can be minimised by proper selection of 'representative' farms (Kuyvenhoven, 1998). The process of selecting representative farms normally involves the use of cluster analysis (usually hierarchical techniques). A difficulty for many cluster analysis techniques is deciding the number of clusters (or representative farm types) present in the data. For almost all of these techniques this decision is subjective. However, when the technique of fitting mixtures of distributions is employed as a clustering technique there is an objective test of the appropriate number of clusters that is available in the form of a likelihood ratio (LR) test (Everitt, 1993). This paper advocates the use of mixtures of distributions modelling as an approach to data-led farm size (and/or any other characteristic) selection in representative farm modelling, in order to help minimise problems of aggregation bias. In demonstrating the technique of fitting a mixture of distribution as a clustering technique for identifying representative farm size clusters, this paper also transforms the stylised fact that the farm size population is made up of distinct farm size groups into a testable hypothesis.

Typically, the choice of 'representative' farms in policy oriented farm models simplifies to two alternatives. The first is to choose some average farm in terms, of say, size and then to assume that the other farms are linearly related to the

characteristics of the representative farm. Theory suggests that in order to obtain consistent results for the agricultural sector it is sufficient to model only the representative farm and to aggregate accordingly. The second alternative is appropriate where the assumption that farm aggregation is linear does not hold. In this case more than one representative farm is required to represent the farm population adequately (and to minimise aggregation bias). The selection of the number of representative farms and the characteristics upon which the selection should be made is an important consideration. This paper proposes a methodology to test and determine the number of farm types.

In the next section the problem of representative farm selection is further defined. The methodology used is described in section three. An application of the methodology is presented in section four along with presentation and discussion of the results obtained. Some conclusions are drawn in section five.

## 2. Redefining the Problem of Representative Farm Selection.

An alternative way to view the question of whether a particular farm sector is adequately approximated by either a single 'representative' farm or by multiple representative farms is as follows. Is the empirical farm distribution adequately approximated by a single uni-modal statistical distribution or by a mixture of several such distributions? If the empirical farm distribution is approximated by a single uni-modal statistical distribution then the use of multiple representative farms is likely to be unnecessary. The parameters of this single approximating statistical distribution can be used to derive the rules for aggregating the representative farm results to the sector level. Where a mixture of distributions is required to represent the empirical

farm distribution then the farm structure can be viewed as consisting of several types of farms. The results for the representative farms in each of these groups can be aggregated to obtain consistent estimates for the different farm groups. Those can then be aggregated using the relative weights of the sub-samples into the total farm population. The approximation of the empirical farm distribution by a single or a combination of several uni-modal statistical distributions can be represented in terms of a statistical model to be estimated. Subsequently the choice of a number of distributions (i.e. number of farm types) becomes a model selection problem and can be resolved by standard statistical means. Once the number of the approximating distributions is estimated, the classification of the farms into the corresponding groups can be done.

In order to resolve the farm group classification problem the mixture of distributions model (MDM) is employed. The MDM relaxes the conventional assumption that an observed dataset is drawn from a single homogeneous population. Instead, it is assumed that the sample is drawn from an unknown mixture of distributions. Thus, a different set of model parameters is valid for each of the different subpopulations of the dataset. In this sense, MDMs are more flexible than conventional statistical modelling, which assumes that a single set of parameters describe all individuals in the dataset.

The existence of latent subpopulations is a real possibility in many datasets (farm size datasets are a good example). These subpopulations are solely defined by their property of being homogeneous in the sense that a particular set of parameters holds for each latent class. The latent nature of these subpopulations, where the number of classes and the observations belonging in each class are typically unknown, means

that it is not possible to directly estimate the parameter set for each subpopulation. Hence the aim of MDM is twofold: to 'unmix' the data into homogeneous subpopulations and then to estimate the parameters for each subpopulation.

## 3. Methodology

To illustrate the general structure of a MDM, let us denote the set of $n$ $d$-dimensional vectors comprising the available data by $x = \{x_1, \ldots, x_n\}$ (i.e. the sample contains $n$ observations and $d$ variables). It is assumed that each $x_i$ arises from a $d$-dimensional probability distribution with the following density:

$$f(x_i|\theta) = \sum_{k=1}^{K} p_k g(x_i|\lambda_k) \qquad (1)$$

where $p_k$ are the mixing proportions ($0<p_k<1$ for all k and $\sum_{k=1}^{K} p_k = 1$), and $g(x_i|\lambda_k)$ is some $d$-dimensional probability distribution, parameterised by $\lambda_k$. A sample of indicator variables $z = \{z_1, \ldots, z_n\}$, sometimes referred to as *labels*, can be assigned to the observed data. These are defined as: $z_i = \{z_{i1}, \ldots, z_{iK}\}$, where each $z_{ik}$ assumes the value of 1 if $x_i$ arises from the $k$-th mixture component and the value 0, otherwise. When the sample of indicator variables is known the problem is one of density estimation, where the vector of parameters to be estimated is $\theta = (p_1, \ldots p_K, \lambda_1, \ldots \lambda_K)$. When the primary interest is in estimating the indicator variables the problem is one of (classification) cluster analysis.

In this study the mixture approach to classification is used. This consists of obtaining the maximum likelihood estimate for the parameters, $\theta$, by using the Expectation

Maximisation (EM) algorithm[1] of Dempster *et al*. (1977) and then applying the 'maximum a-priori' (MAP) principle to assign a value to the indicator variables, $z_i$. The MAP involves assigning each observation $x_i$ to the mixing component based on conditional probabilities. This approach produces more consistent results than the alternative methods[2].

The EM algorithm used in the analysis consists of the following two steps, namely, the E(xpectation) step and the M(aximisation) step. In the E step the conditional probability of $z_{ik}$ being equal to one, estimated during the m-th iteration for all *i* and *k* is given by:

$$t_{ik}^{(m)} = t_k^{(m)}\left(x_i \middle| \theta^{(m-1)}\right) = \frac{p_k^{(m-1)} g\left(x_i \middle| \lambda_k^{(m-1)}\right)}{\sum_{l=1}^{K} p_l^{(m-1)} g\left(x_i \middle| \lambda_l^{(m-1)}\right)} \tag{2}$$

where the (bracketed) superscripts denote estimates for the parameters during the corresponding iteration.

In the M step the ML estimate, $\theta^{(m)}$ of $\theta$, is updated using the conditional probabilities, $t_{ik}^{(m)}$, as conditional mixing weights. This leads to maximizing:

$$F\left(\theta \middle| x, t^{(m)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}^{(m)} \ln\left(p_k g\left(x_i \middle| \lambda_k\right)\right) \tag{3}$$

[1] This is the standard algorithm for estimating MDM. Modifications of this algorithm, include the stochastic EM (SEM) (MacLachlan and Peel, 2000) and the classification EM (CEM) (Celeux and Govaert, 1992).

[2] The main alternative approach is to jointly estimate $\theta$ and *z*. In this case the indicator variables are used to weight the contributions of the individual observations to the log-likelihood function. However, the main algorithm used in this approach is the CEM algorithm, which is not expected to provide ML estimates for $\theta$ and may yield inconsistent estimates of the parameters (MacLachlan and Peel, 2000).

The updated expressions for the mixing proportions are given by:

$$p_k^{(m)} = \frac{\sum_{i=1}^{n} t_{ik}^{(m)}}{n} \tag{4}$$

The updating of $\lambda_k$ depends on the specific parametric specification and therefore, no general formula can be given.

So far we have considered estimating a mixture model for the purposes of classifying the observations into a pre-defined number of distributions (sub-samples or clusters). However, the number of clusters is typically unknown. Choosing the appropriate number of mixing distributions (clusters) is essentially a model selection problem. A popular criterion in model selection problems is the Bayesian Information Criterion (BIC) (Schwartz, 1978).

$$\mathrm{BIC}_{mK} = -2\, L_{mk} + \nu_{mK} \ln n \tag{5}$$

where $m$ is any model (thus $m$ denotes the choice of the parametric distributions $g(.)$[3]) with $K$ components, $L$ is the (maximised) log-likelihood and $\nu$ is the number of free parameters in the model. If the choice of $g(.)$ is taken for granted, then (5) suggests a strategy of consecutive estimation of $(m, K)$ models for $K$=1,2, until BIC increases. It is clear that if $(m, K)$ and $(m, K+1)$ provide essentially the same fit then the BIC for $(m, K)$ will be smaller, since it has less free parameters. In this way the BIC allows the homogeneity of the subpopulations of farms to be directly tested. The consecutive estimation strategy also ensures against the danger of over-fitting the statistical model (1).

[3] Or any combination thereof. In other words one may consider cases in which different groups of farms follow different parametric distributions.

The BIC is based on an asymptotic approximation of the integrated log-likelihood, valid under some regularity conditions. In spite of the fact that these regularity conditions are usually violated in mixture models, it has been proven that the BIC is consistent and efficient (e.g. Frealey and Raftery, 1998). The BIC is, however, essentially a criterion to choose a model specification and does not take into account the ability of a mixture model to provide evidence about the clustering nature of the data. In order to do this, the likelihood of the complete data (i.e. in a BIC-like context this means the Integrated Completed Likelihood (ICL)) must be considered. Using the MAP principle to approximate the unknown indicator variable, the ICL can be expressed (Biernacki et al. 2000) as BIC with an additional entropy penalty term as follows:

$$\mathrm{ICL}_{\mathrm{mK}} = -2\mathrm{BIC}_{\mathrm{mK}} - 2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln t_{ik} \qquad (6)$$

In testing the possible range of values for $K$, the criteria proposed by Bosdogan (1993) is applied in searching over the range from 1 to the smallest integer not exceeding $n^{0.3}$. The ICL information criteria can be used to choose a suitable model from amongst a wide range of mixing distributions. For example, if 4 different types of parametric (m) distributions are considered, then the number of models to choose from is ($n^{0.3}$x4). When all combinations of the different distributions are made possible, then a model implying say 7 different classes might contain, for example, a mixture of 2 normal distributions, 3 gamma distributions and 2 t-distributions. Clearly, the range of possibilities in this context widens tremendously. However, the model is much more tractable when the MDM is restricted to a single type of parametric distribution type (i.e. when all parametric distributions $g(.)$ belong to the same type).

In this case with regard to the classification of farms is would seem sensible to choose a single parametric distribution for use in the mixture of distributions model, which is consistent with any distributional assumptions imposed on the data during second stage modelling. With regard to how many characteristics are necessary to efficiently perform such a classification (i.e. the choice of $d$) it again would seem desirable to keep the number of characteristics to a minimum, otherwise the likelihood of contrived correlation in second stage analysis is increased. Ideally, the assumptions made in classifying farm types should not contradict the other assumptions used in subsequent stages of analysis.

An alternative selection criterion is the Normalised Entropy Criterion (NEC) (Celeux and Soromenho, 1996) which measures the ability of the mixture model to provide well-separated clusters. In doing so however, the NEC is essentially devoted to choosing the number of mixture components $K$, but not the model form $m$. Consequently, this criterion is not used here, but it is listed for completeness given that it may be useful when the number of components is of a primary interest.

*Data*

In this paper MDM techniques are used to classify dairy farms in Northern Ireland, based on the number of cows in each farm. The data for 5275 farms for 2000 was obtained from the Agricultural Census, which is carried out annually in Northern Ireland.

## 4. Application and Results

The analysis carried out is based on the assumption that the model generating the data is a mixture of normal distributions (restricting the *m* domain of potential candidate models). This is an arbitrary assumption, but if the classification of farms produced from the analysis is subsequently used in a linear programming study that assumes normal distributions then at least the assumption is consistent. In this application for purely illustrative purposes only one characteristic variable (i.e. $d$=1) is considered, namely, the number of dairy cows on each farm. Although, this is a simplistic approach it does serve to demonstrate the application of the MDM approach advocated in this paper. Normally, the choice of the number of variables, *d*, to include in the analysis should take in to account the purpose for which this classification is to be used.

The range of values for *K* was chosen based on the criteria proposed by Bosdogan (1993), which suggests searching over the range from 1 to the smallest integer not exceeding $n^{0.3}$ (which for $n$=5275 is 13). In this study, the EM algorithm is applied using both the BIC and the ICL criteria to chose the appropriate model. The normal distributions mixed are allowed to have different variances. The classification of each observation to any of the latent classes is carried out using the MAP principle.

The classification results based on the BIC criteria are given in Table 1, while the results based on the ICL criteria are given in Table 2. In each case the number of sub-groups are indicated in the first column, with the percentage of total farms in each sub-group indicated in the second column and the mean number of cows per farm in each sub-group indicated in the third column. It can be seen from these results given in Table 1 that the BIC criterion identifies six types of dairy farms according to their

size (measured in terms of herd number). The results presented in Table 2 indicate that the ICL criterion identifies five subgroups.

**Table 1 BIC results**

| Sub-Groups | Weight | Mean |
|---|---|---|
| **1** | 0.559003 | 14.34694 |
| **2** | 0.365583 | 56.88165 |
| **3** | 0.06875 | 113.2626 |
| **4** | 0.006474 | 171.3461 |
| **5** | 0.003986 | 225.6 |
| **6** | 0.00019 | 453.0 |

**Table 2 ICL results**

| Sub-Groups | Weight | Mean |
|---|---|---|
| **1** | 0.592038 | 17.33333 |
| **2** | 0.313555 | 56.88166 |
| **3** | 0.089479 | 111.2348 |
| **4** | 0.004739 | 223.6 |
| **5** | 0.00019 | 453.0 |

The final sub-group derived under both methods (with 453 cows) consists of a single farm. Reducing the sensitivity of the algorithm would reduce the number of sub-

groups to four using ICL and five using BIC. However, the quality of the density estimation worsens as a result. It is perhaps more practical to simply accept that there are 4 types (using ICL criterion) of dairy farms in Northern Ireland. An alternative approach is to curtail the dataset prior to the analysis by omitting extreme observations.

## 5. Conclusions

This paper demonstrates the potential use of the MDM in deriving model-based classification of farms. The main advantage of the proposed MDM approach as a method of cluster analysis is that it allows for a robust selection of the number of farm types (clusters) using transparent statistics based model selection criteria. Most methods of cluster analysis require subjective decisions to be made regarding the number of clusters. This paper demonstrates the MDM approach to cluster analysis by classifying dairy farms in Northern Ireland, based on the number of cows in each farm. The results indicate that four representative farms are needed, with a view to minimising aggregation bias, to describe the dairy sector in Northern Ireland.

The model-led application of farm classifications is of particular relevance to policy evaluation problems. Different policy measures impact on different farm characteristics and when evaluating the likely effect of such measures it is advisable to classify farms according to these characteristics. In this way the possibility for large errors stemming from aggregation of heterogeneous populations is avoided.

The use of the MDM in economic research is not novel. Most previous studies focus exclusively on the density estimation applications of the MDM. The idea of using

MDM for cluster analysis has been around for some time, but published applications of the technique are difficult to find. This paper helps fill that gap in the literature.

**References:**


Biernacki, C, Celeux,G. and Govaert,G. (2002) Assessing a Mixture Model for Clustering with and Integrated Completed Likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719-725.

Bozdogan, H. (1987) Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions, *Psychometrika*, 52 (3), 345-370.

Bozdogan, H. (1993) Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion on the Inverse-Fisher Information Matrix, in O. Opitz, B. Lausen and R. Klar (Eds.) *Information and Classification*, Heidelberg: Springer-Verlag, pp.40-54.

Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions, *Computational Statistics and Data Analysis*, 14, 315-332.

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

Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method? - Answers via model-based cluster analysis. *Technical Report No. 329*. Seattle: Department of Statistics, University of Washington.

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*, New York: Wiley.

Schwartz, G. (1978) estimating the Dimension of a Model, *Annals of Statistics*, 6, 461-464.