# One size does not fit all: an empirical investigation of the Romanian agriculture production function

Contributed Paper presented at the 91st UK AES Conference, 2017 Dublin.

Philip Kostov[1], Sophia Davidova[2], Alastair Bailey[3]

1 University of Central Lancashire, UK PKostov@uclan.co.uk

2 University of Kent, UK S.M.Davidova@kent.ac.uk (corresponding author)

3 University of Kent, UK A.Bailey@kent.ac.uk

Abstract

There are issues when researchers want to consider homogeneous, with regard to some functional relationship, groups. For example, in representative farm modelling analysts are interested in specifying groups of farms that have the same input/output relationship. This paper proposes to use the underlying functional relationship to derive such groupings. The paper employs finite regression mixture models to specify and estimate farm groups with regard to pre-specified functional relationship. The proposed approach is illustrated with regard to the aggregate production function of Romanian agriculture. The results point out to two farm clusters. The first one is more productive with a better use of capital and intermediate consumption. The second one makes a better use of land and labour. The calculated Shannon index shows that the second cluster is characterised by a higher level of land use diversity. The implications of the derived structure are discussed in light of two sets of policy – a production oriented and environment oriented one.

**Keywords:** finite mixture models; production function, Shannon index, land use diversity

JEL code: C21; Q14

# One size does not fit all: an empirical investigation of the Romanian agriculture production function

## Introduction

Economists and policy makers have always been interested in producers' responses to policies in order to achieve some national or sectoral objectives, e.g. growth, employment, food security. The way producers respond to policy depends on their production function. The assumption that all observed units have homogenous production function, thus a homogenous response is a very heroic one. If this assumption does not hold, production units' policy responses will be heterogeneous. This heterogeneity in responses is the focus of the present paper.

Modelling heterogeneous responses has a long tradition in economics, and in particular in agricultural economics in the area of the so called 'representative farm modelling'. The traditional approach splits the units of interest into relatively homogenous groups and models these separately. Often the purpose of such modelling is to use the results for mathematical programing models of these different homogeneous groups. The way these groups are derived can, however, be problematic. Often some form of factor analysis or principal components analysis is applied with regard to selected observable characteristics in order to identify the groups. The problem with this approach is that it yields groups which are similar with regard to the observable variables used in the analysis, but not necessarily with regard to the functional relationship which is of primary interest in such an approach.

This paper proposes to specify homogeneous groups with regard to the pre-specified functional relationship, in this case a production function. The groups are estimated using finite regression mixture models. We propose this method as the most adequate when the

issue in hand is either to investigate policy responses of groups of firms with similar production function, or to model their production function in the follow-up simulation model.

We illustrate the proposed approach with regard to the aggregate production function of Romanian agriculture. The implications of the derived structure are discussed in light of two sets of policy measures, namely a production oriented and an environmental one. The empirical analysis is focused on the aggregate agriculture production relationship. We provide a farm classification based on the production function which, in contrast to the farm characteristics generally used in the clustering approach, is not directly observable.

The empirical application is to farm structures in Romania. Examining such a classification for a specific country has distinct policy implications. Different groups of farms, identified using the proposed methodology, are expected to react differently to production incentives, since by definition they have different production functions. The aggregate reaction of the agricultural sector will be a weighted average of the responses of the different groups.

The results from the analysis point out to two farm clusters. What is striking in the obtained classification is that the relative shares of capital and labour are very similar across the clusters. And yet, the two groups make very different use of their capital and labour endowments in terms of the amount of output they manage to extract from each of these two production factors. The first group is more productive, but this productivity comes from capital and intermediate consumption, while the second group makes a better use of land and labour. Policies may have structural change effects changing the balance between the two clusters and induce different production responses. The first cluster has a productivity focus. Production oriented policies that aim at e.g. increasing food security or boosting exports through coupled output support, or conversely aim at increased competitiveness by removing price intervention, are easier to link to the first group of farms. Since land use is the ultimate

basis for agriculture, the paper adopts farmland use diversity as a crude measure to ascertain any structural effects of environmental policies. The second cluster is characterised by a higher level of land use diversity, which suggests that these farms are better suited to deliver environmental public goods and might be more affected by environmental policies and their reforms.

The remaining of the paper is structured as follows. The next section provides the motivation for the proposed approach in comparison to alternatives. The third section presents a short overview of Romanian agricultural sector. Sections 4 and 5 include methodology and data respectively. Section 6 discusses the results and section 7 concludes.

**Motivation**

Economic theory has a longstanding tradition of emphasising uniformity. The principle of the 'representative economic agent' is probably the best known theoretical abstraction in economics. Assuming such uniformity is very useful in deriving theoretical properties helping microeconomic models to be easily expressed into common sense logic. This approach has been very fruitful in producing logical outcomes based on sound principles of rationality. Furthermore, it has also provided a basis for statistical investigation. Since this concept is an abstraction and it is obtained by averaging the reactions of the actual economic agents, the representative agent responses can be obtained by averaging the observed responses of the actual agents. Hence, although directly unobservable, estimating a mean regression type of statistical model implicitly yields the response of the representative economic agent.

This uniformity principle, however useful, has its limitations and has been questioned. From a theoretical point of view, models of bounded rationality which combine two types of representative agents have been shown to be able to produce qualitatively different outcomes.

For example, De Long et al. (1990) present a model with rational agents and noise traders who behave randomly and interact with the rational agents. One of the surprising outcomes of this model is that the noise traders, who non-intentionally (i.e. randomly) make very risky investments, may under certain conditions end-up dominating the market. Kogan et al. (2006) further investigate this issue, which is now accepted in financial literature (see e.g. Cogley and Sargent, 2009; Le Baron, 2012; Luo, 2012).

This paper, however, is not concerned with the theoretical challenges to the uniformity principle, but rather with some empirical considerations. A major problem in empirical research is that the theory rarely prescribes the form of the functional relationship between the variables in question. It is essentially not possible to know beforehand the functional form of this relationship. Hence, the problem of 'representativeness', i.e. homogeneity in response, becomes interwined with the issue of the functional specification. There is a clear trade-off in this area. Using more flexible functional representation reduces this problem, but also makes the interpretation and inference more difficult, and in some cases impossible (as in the case of the curse of dimensionality problem). Using more restrictive functional representations results in more tractable models, but in this case the representativeness assumption is more likely to be violated simply because the used functional representation is inadequate. Therefore, the representativeness condition in empirical modelling is dependent on a given functional specification. In other words, the question of whether the units of analysis exhibit the same relationship is only meaningful with regard to the given functional form of this relationship.

To simplify the issue, the following discussion focuses on the production function, but our argument is equally applicable to other functional relationships. Grouping units of analysis with regard to their production function (or any other functional relationship of interest), as it is proposed in this paper, not only asks the relevant question (i.e. what different

functional relationships describe the data) directly, but also makes the classification issue explicitly dependent on the choice of the functional form. It provides a clear definition of the kind of representativeness the researcher is looking at. If the aim is to group farms with similar production function either because this is the characteristic of interest or because the intention is to model their production function in a follow–on simulation model, this is clearly the question that has to be asked. A clustering type of approach in contrast asks a very different question. It asks how similar the units appear to be with regard to some predefined observable characteristics. Such a question leaves the issue of 'representativeness' very vague. It also implicitly claims a kind of logically inconsistent universality. For example, one may use some set of 'relevant variables' to cluster units and then assume that the functional relationship is homogeneous within each cluster. However, the same approach could be applied to a wide range of relationships, such as e.g. cost, profit and production functions. Therefore, the units in the same cluster are assumed to have the same type of functional relationship for all of the above. This is a very unrealistic assumption.

Finally, there is another more practical consideration. Economic analysis is often based, as in this paper, on aggregate relationships, which undoubtedly contain unobserved heterogeneity. For example, when we look at the issue of production function, since technologies are very different for different farm typologies, it is reasonable to consider different production functions for different types of farming typologies, e.g. livestock, crop, vegetables etc. farms. Yet, doing so, results in a large number of underlying models without actually solving the problem of unobserved heterogeneity since even within a certain typology, different technologies could co-exists, based on characteristics that are not directly observable. Therefore, from a purely practical point of view, there is a trade-off: on the one hand, we would want a small number of functional relationships, but on the other, would want these relationships to encompass both the similarities and differences amongst the units

of interest. In other words, subject to the constraints defined by the choice of functional relationship, we want the best combination of such functional relationships that describe the data. Hence, in our application of the proposed method of classification the question becomes: how many distinct production functions can describe the output response of Romanian agriculture and what are their characteristics? In this way we not only provide a characterisation of an an economic sector (agriculture), but also simultaneously determine the behaviour of its production units.

Whenever the policies do not affect the structure of agriculture, i.e. they do not affect the balance (i.e. the weights) of the different groups, the proposed methodology will just provide an approximation to that response (i.e. production function). However useful such an approximation might be, there are alternatives that can achieve the same result (e.g. using a more flexible functional form). The real advantage is apparent when policies have structural effects and they affect the balance of the classified groups. In this case the structural change effects can be inferred by examining the differential production responses by different groups. To illustrate this, we consider two broad types of agricultural policies – production related and environmental.

**Agricultural sector in Romania**

Romania has long traditions in farming and currently is home of the largest number of farm holdings in the EU, 3.6 million, accounting for 33.5 per cent of all EU agricultural holdings (FSS, 2013). A major characteristic of Romanian farms is that they are biased towards small scale - about three quarters are small - measured in physical size they cultivate less than 2 hectares. At the same time, there are large farms playing a key role in agricultural production and productivity. According to agricultural census, farms larger than 50 ha cultivate 53 per cent of agricultural area. Popescu and al. (2016) calculated Gini coefficient for the size

distribution of farms in Romania. The value of Gini coefficient of 0.582 places Romania on the sixth place amongst the EU Member States (MS) according to the most unequal land size distribution.

Agriculture is also important for rural labour. Agricultural census points out that Romania engages the second largest number of farm workers in the EU (second only to Poland), equivalent to 1.5 million full-time equivalents measured in Annual Work Units (AWU) (Eurostat, Agricultural Census in Romania). Data also indicates a sharp decrease in labour between 2003 and 2010. However, Tocco et al. (2014) found little mobility of labour out of farming to other sectors of the economy. The major move out of farming was either to retirement or non-employment. To a great extent the decrease recorded by Eurostat might have been due to retirement, since during this period 46.4 per cent of farmers running farms with an area of less than 20 ha (the overwhelming majority of farms in Romania) were older than 65 years of age (Page and Popa, 2013).

Concerning capital, Romania has the lowest total asset value per farm in the EU (below Euro 40,000) due to low land prices, small farm sizes and less capital-intensive types of farming (EC, 2016). Fixed assets (land, farm and other buildings, forest capital, machinery and equipment, and breeding livestock) account on average to around 75 per cent of total assets. Buildings have the largest contribution to the fixed assets.

Although lagging behind farms in the other EU MS, Romanian small-farm landscape maintains rich agro-biodiversity. Page and Popa (2013) emphasise the provision of public goods by these types of farms, including sustainable land use and biodiversity conservation.

The inheritance of transition is likely to have influenced this quite heterogeneous farm structure and the unsatisfactory performance reflected in the gap with the other EU MS. Before the reforms of late 1980s - beginning of 1990s three types of farm structures were typical for Romanian agriculture – state farms, agricultural production cooperatives and

small-scale individual farming, particularly developed in mountainous regions. In 1989, individual farms accounted for 12 per cent of the total agricultural area, production cooperatives (with an average size of 2,400 ha) – to 59 per cent and the remainder was accounted for by the state farms (Rizov et al., 2001). Ten years later, individual farms managed 58.6 per cent of land and the remaining was cultivated by commercial companies, agricultural societies, farmers associations and other institutions. Rusu at al. (2002) classify the agricultural economy into two distinct structures, one which they call 'traditional', incorporating the majority of small-scale farms, and second, a sector tending to modernisation and productivity, i.e. agricultural commercial companies and agricultural associations. Strictly speaking, adapting pre-existing farming structures and the millions smallholders to a market economy carries forward a set of constraints that can restrict the possible production responses. Starting anew, on the other hand, does not imply such restrictions and could potentially result in different technological relationships and different production responses that are not alike those of the pre-existing farms. Furthermore, under the conditions of a rather turbulent transition period characterised by a series of shocks and a typical 'stop and go' approach, establishing a new farm could have been a quite different endeavour depending on when exactly the business was created, potentially resulting in even more diversity in underlying technologies.

**Methodology**

We employ finite regression mixture model to specify and estimate farm groups with regard to the pre-specified production function. It is assumed that, conditional on a set of covariates *X*, *y* arises from a probability distribution with the following density:

$$f\left(y|\theta, X\right) = \sum_{k=1}^{K} p_k g\left(y|\lambda_k, X\right) \quad (1)$$

where $p_k$ are the mixing proportions ($0<p_k<1$ for all $k$ and $\sum_{k=1}^{K} p_k = 1$), and $g_k\left(y\middle|\lambda_k\right)$ probability distribution, parameterised by $\lambda_k$. This means that $y$ can be viewed as drawn from $K$ different underlying (conditional) probability distributions. The parameters $\lambda_k$ specify a regression model, i.e. they include regression coefficients, as well as the distribution parameters. In this study, we use a linear regression specification (see De Sarbo and Cron, 1988; Wedel and Kamakura, 2001), but in principle any other parametric specification, could be used instead. The nature of the estimation algorithm is very general and allows for a wide range of specifications. Equation (1) states that the data-generating process for $y$, conditional on $X$, is a mixture of regressions. Thus if $y$ is the output and $X$ are the inputs, this expression states that the data comes from several distinct production functions.

One can obtain the maximum likelihood estimate for the parameters $\theta$ by using the Expectation Maximisation (EM) algorithm of Dempster et al. (1977) and then apply the 'maximum a-posteriori' (MAP) principle to assign observations to each of the underlying distributions. 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 observation $i$ belonging to $g_k(.)$ during the $m$-th iteration for all $i$ and $k$, is given by:

$$t_{ik}^{(m)} = t_k^{(m)}\left(y\middle|\theta^{(m-1)}, X\right) = \frac{p_k^{(m-1)} g_k\left(y\middle|\lambda_k^{(m-1)}, X\right)}{\sum_{l=1}^{K} p_l^{(m-1)} g_l\left(y\middle|\lambda_l^{(m-1)}, X\right)} \qquad (2)$$

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:

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

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 parametric specification and, therefore, no general formula can be given. The maximisation step is essentially the standard maximisation routine used to estimate the conditional model given some fixed, determined in the expectation step, mixing proportions. The generic equation (3) expresses calculating the log-likelihoods for each separate component and maximising the weighed likelihood with weights given by the posterior probabilities $p_k^{(m)}$. Thus, by adapting the maximisation step, a wide range of models could be fitted.

The above description assumes that we know the exact number of clusters. However, this is typically not the case. Choosing the appropriate number of mixing distributions (clusters) is essentially a model selection problem. One can estimate the regression mixture models for different number of clusters and then selects amongst these. A popular criterion in model selection problems is the Bayesian Information Criterion (BIC) (Schwarz, 1978).

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

where $m$ is any model (thus $m$ denotes the choice of the parametric (conditional) distributions $g(.)$ or any combination thereof), $K$ is the number of components, $L$ is the (maximised) complete log-likelihood and $v$ 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. The consecutive estimation strategy also ensures against the danger of over-fitting the statistical model (1).

We use the BIC as a main model choice criterion, although details on some alternatives are also provided. The BIC is based on an asymptotic approximation of the integrated log-likelihood valid under some regularity conditions. It has been proven that the BIC is consistent and efficient on practical grounds (e.g. Fraley and Raftery, 1998). Moreover, the whole class of penalised likelihood estimators, of which the BIC is a special class, are consistent (Keribin, 2000). The BIC is furthermore approximately equivalent to the popular in information theory Minimum Description Length (MDL) criterion.

If one needs to select of model where in addition to the model fit the ability to define well separated clusters is taken into account, the integrated complete likelihood (ICL) criterion can be used. The ICL can be expressed (Biernacki *et al*., 2002) as BIC with an additional entropy penalty term as follows:

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

where $z_{ik}$ are the cluster membership indicators. In the present application, we are not explicitly interested in the degree of separation of clusters. Nevertheless, applying the ICL can be used as an additional clustering criterion.

The mixture models with increasing number of components can be analysed in a nested models framework. Therefore, the Likelihood Ratio (LR) test can be readily applied to consecutively test for the number of components. In order to provide a valid small sample inference, the distribution of the LR tests statistic can be simulated via bootstrap. Such a bootstrap approach is however very expensive in computational terms. For this reason, we will only implement it to test the model selected by the information criteria.

The finite regression mixture approach describes the functional relationship as an hierarchical mixture model, where the data generation process generates each observation from a finite set of underlying sub-models, which define separate clusters. As explained in the motivation section, these clusters represent different functional relationships (i.e. different production functions). Hence, we define the representativeness condition directly with regard to the production function conditional on its functional form. An advantage of the finite mixture approach is the ease by which data observations can be attached to the different underlying production functions.

**Data and choice of functional form**

As explained, the approach to specify homogenous groups of observations based on underlying functional relationship is applied empirically to farms in Romania. Empirical estimations are based on data from the Farm Accountancy Data Network (FADN) for 2008. The implementation of the EU Common Agricultural Policy (CAP) creates methodological issues about how to treat the CAP the single area payments and other CAP subsidies, and by choosing the year immediately after the Romanian accession to the EU accession we hopefully avoid some of these issues.

In the empirical specification the farm output is specified as a function of four inputs, namely capital, labour, land and intermediate consumption (IC). Summary statistics for the data are presented in Table 1. Labour is measured in Annual Work Units (AWU) and Land in hectares, while all other variables are in monetary terms. Due to rounding, small numbers for the minimum values smaller than 0.5 appear as zeros in Table 1. Such relatively small farms mange to pass the FADN inclusion threshold, which in Romania is low in comparison to other EU MS in order to reflect the nature of Romanian farms structure. There seems to be

considerable heterogeneity in terms of all variable amongst the 870 farms included in the dataset. Since the mean values for all variables are closer to the minimum than the maximum values, there are more relatively smaller farming units and a very long right tail representing the smaller number of larger farms in the distribution for all considered variables. This distributional feature is not particularly surprising, but any such heterogeneity suggests that the functional relationships amongst these variables may also be heterogeneous. In particular, the considerable differences in terms of size that are evident in the data set could lend themselves to differences in the production relationship, since it is not unreasonable to expect that as farms grow larger, the organisation of their activities changes and therefor the input/output relationship might change too.

The key question in the paper is whether the Romanian farms can be described by the same production function. As already discussed, this question requires specifying the inputs and the functional form for the specific production function. There is extensive literature on the issue of the productions functions, and their theoretical and empirical properties (Griliches and Ringstad, 1971; Berndt and Christensen, 1973; Christensen and Lau, 1973). With regard to the problem in hand, it is advisable to employ a production function specification that is sufficiently flexible, since in a finite regression modelling framework there is a clear trade-off between flexibility and the potential number of homogenous groups, i.e. more flexible functional forms will reduce the number of groups. Here the translog functional specification is employed.

In the production function literature the term ‘flexible’ has a specific meaning. According to Diewert (1974), a functional form can be denoted as ‘flexible’ if its shape is only restricted by theoretical consistency. The translog functional specification can be restricted to satisfy the homotheticity, homogeneity or separability, but in this application no such restrictions have been applied. The main reason for this is that by avoiding restrictions

we can maintain its generality. Furthermore, as our previous argument demonstrates, there is a clear trade-off between flexibility and the potential number of clusters since flexible specifications would result in a smaller number of clusters. Therefore, since the question is whether a single production function specification is sufficient to describe the data, it makes sense to avoid imposing restrictions that could inflate the potential number of clusters.

Although in more recent studies the translog appears to have somewhat fallen out of favour with empirical researchers, it is still the most extensively investigated second order flexible functional form and surely the one with the most empirical applications as its empirical applicability in terms of statistical significance is outstanding (Feger, 2000). Furthermore, the fact that the translog function can be considered as a second order (Taylor series) approximation of a more general production function provides a sound justification in applying it here, since the uncertainty about the production function is a major justification for the present study.

An important reason for the choice of the translog specification is also that it is linear with regard to the parameters, which means that standard linear regression techniques can be used for estimation and testing purposes. In principle, estimating a finite regression model simply requires plugging in the M step an estimation routine for the underlying model, which creates tremendous flexibility since this means that the underlying model can be fully nonparametric. Linear specifications offer considerable savings in terms of computational costs.

One of the key issues in modern development is its sustainability (see e.g. Piorr, 2003; Waldhardt et al., 2003). In order to give insights into the sustainability of production and the possible implications of environmental policies biodiversity measures based on the land use are applied in addition to the production function. The underlying logic is that greater diversity in the land use will be associated with greater biodiversity. Two specific measures

are applied, namely richness and the equitability index based on the Shanon diversity index. The richness is simply the number of separate land uses found on a farm. The equitability index is a standardised Shannon index (divided by its maximum value, so that it fits the [0,1] interval).

More specifically, the Shannon index is $S = -\sum_{i=1}^{N} \alpha_i \, ln \, \alpha_i$, where $\alpha_i$ is the land area share allocated to the $i^{th}$ land use, while the equitability is defined as E=S/ Max(S).

**Results**

The model fitting BIC criterion indicates that a single common translog production function is not sufficient to describe the Romanian farms and points out to two clusters (Table 1). Furthermore, since ICL accounts for both model fit and cluster separation, the fact that the ICL also points out to a two-cluster model demonstrates that these two clusters are well separated. In practical terms, this means that at least some of the corresponding coefficients are significantly different, resulting in two quite different production functions, subject to the functional restriction of a translog functional form.

In order to confirm the above conclusion, LR bootstrap tests for 2 mixtures (clusters) were implemented. Since such tests are based on model fitting and do not take into account the cluster separation, they are only comparable to the BIC results. The probability levels for the bootstrap tests are shown in Table 2. The LR bootstrap tests agree with the information criteria that the Romanian farms can be split into two distinct clusters with regard to their underlying production function.

Table 3 presents the estimation results, while Table 4 shows the summary statistics for the used variables, both for the overall sample and by cluster. In order to facilitate the discussion, the summary statistics in Table 4 are for the raw variables rather than their

logarithmic transformation which is used in specification and estimation. The membership of Cluster 1 is smaller with 296 farms, while Cluster 2 consists of 574 observations. Cluster 1 contains bigger farms with regard to labour, capital and intermediate consumption. Comparing the means for the two clusters, the only input for which Cluster 2 has larger values is land. Hence, in general we can say the first cluster is characterised by larger farms. The larger average value of land in the input mix of the farms in the second cluster suggests that these might use a production technology that is much more land intensive, something that the estimation results might throw a more light on.

It is difficult to ascertain the differences between the cluster-wise production functions given in Table 3, due to their non-linear form. A reliable way to compare two non-linear functions is by comparing their partial correlation plots. This amounts to using the estimated models to predict the dependent variable and then plot the predicted values against the values for a given factor by keeping the other factors fixed at 'typical' values. In this way, one can visualise the effect of a given production factor when the rest of the inputs are kept fixed. The first issue is what would be the reasonable values for the fixed inputs. This would depend on the purpose of the above plot. If the interest is in average effects, using the average over the estimation sample values would be an easy way to achieve 'reasonable values'. Sometimes averaging would not be a reasonable strategy, in particular in the case of discrete values (see e.g. Kostov et al., 2008). In the present study all the inputs are continuous variables, therefore averaging over the estimation sample is a viable option.

The second issue concerns the need to create a prediction sample containing a range of values for input variable of interest, create the relevant (transformed) variables needed in the translog specification and predict from the estimated linear model. The only choice necessary is the range of values for the analysed input. We use a regular grid of 100 points defined over the range over which the input in question is observed. Since the two clusters

are quite different in their input mixes (see table 4), it is reasonable to produce separate ranges for each cluster. In this way the values for the variable of interest are actually observable within the estimation sample. The resulting plots show the range of values for each input by cluster and this facilitates the interpretation of the results. It also avoids the danger of predicting outside the range over which each of the two clusters is defined. As for the variables over which any such plot is conditioned upon (i.e. the other inputs), averaging over the whole sample is applied in order to ensure that the effects plotted for the two clusters are comparable (since all the rest is being equal). Since the summary statistics for both clusters exhibit considerable dispersion, it is easy to verify that such common 'typical' values lie comfortable within the range of observable values for each of the two clusters and therefore the synthetic observations created in order to produce the effects of interest are feasible.

Plotting the effects for each input can provide a useful overview of the differences between the corresponding production functions. However, the usefulness of such a comparison would be limited without information on how different statistically these are, which requires confidence intervals for such effects that can be obtained by bootstrapping the corresponding models. Here the nonparametric case bootstrap is used following Kostov et al. (2008).

The partial correlation plots for the inputs are presented in Figures 1-4. Both output and the input have been transformed back into the original units in order to facilitate a meaningful interpretation. Due to the non-linear nature of the model, the confidence intervals are asymmetric. The first noteworthy feature of these figures is that cluster 2 is considerably more homogenous in terms of the underlying production function, i.e. the confidence intervals for the effect of all four inputs are narrower than those for cluster 1. Although this on its own is not that surprising given the larger dispersion in the underlying inputs in cluster

1 (apart from land) as revealed by standard deviations in Table 4, the latter by no means guarantees a higher dispersion of the estimated effects. This difference in the homogeneity means that the farms in cluster 2 are much better characterised by their underlying production function than those in cluster 1. Taking into account that there are actually considerably more farms in cluster 2 and that cluster 1 farms are larger, it looks like that the growth in farm size could be responsible for farms moving away from a common production function. The other important results is that these differences in terms of different width of the corresponding confidence intervals, as well as in terms of underlying mean effects, are unevenly distributed amongst the different inputs.

In order to better explain such differences the own elasticities derivable form the estimated translog specifications for both clusters have been calculated for each farm and the mean values and their standard deviations are summarised in table 5. For comparative purposes Table 5 also includes elasticities calculated from a common (single cluster) translog applied to the full sample. Since the elasticities are in fact properties of the underlying production functions, they can be used to complement the partial correlation plots effects.

As mentioned previously, the effects are unevenly distributed amongst the inputs. Considering the capital input, cluster 1 employs more capital than cluster 2 and uses a wider range of capital inputs (Table 4 and Figure 1). Furthermore, cluster 1 is also more capital intensive in a sense that it manages to extract considerably more output from the capital it employs. This can be ascertained from the fact that the average contribution of capital to output is higher for cluster 1 over the whole range of capital values. Taking into account the associated confidence intervals, which do not overlap, the difference in these effects is statistically significant. It is also revealed that the confidence intervals for the effect of capital in cluster 2 are quite narrow indicating that cluster 2 consists of farms which are homogeneous with regard to the contribution of capital to their output. In contrast to this, the

corresponding confidence intervals for cluster 1 are considerably wider. The (own) elasticity of capital is higher in cluster 1 (Table 5), which is also visible from the figure 1 showing that the slope of its production curve is steeper for cluster 1. Yet, interestingly, both the mean values and standard deviation for the capital elasticity in cluster 1 coincide (subject to a rounding error) with those derived from a single cluster full sample estimation. Taking into account that there are a smaller number, although much larger farms in cluster 1, this shows that this cluster defines the role of capital in Romanian agriculture.

With regard to labour, again cluster 1 is characterised by larger farms employing both more labour and having a wider range of labour inputs (Table 4). However, on average the labour/capital mix is not that different between the two clusters which can be inferred by dividing the average values for labour and capital in Table 4 and comparing the ratios. However, contrary to the case for capital, the farms in cluster 2 make much better use of labour and they manage (except for the very small farms) to extract considerably more output per unit of labour employed (Figure 2). Hence, we can view the cluster 2 farms as more labour intensive. The average labour output elasticities for the two sectors are rather similar (Table 5). Once again the dispersion of the labour effects looks larger in cluster 1, but if we look at the width of the confidence interval at similar values for the labour input, and in particular for values observed over the larger farms in cluster 2, these are actually of similar magnitude. So, unlike any of the other inputs contributions, it cannot be claimed that labour effects are more homogenous in cluster 2.

Although cluster 1 in general uses less land (Table 4), the output from the two clusters with regard to land is not statistically different (Figure 3). While cluster 2 appears to be more land intensive in terms of both the slope of its partial effect, as well as its own elasticity (Table 5), this effect does not appear to be statistically significant, mainly due to the large dispersion of the land effect in cluster 1.

Finally, consider the effects of intermediate consumption (IC). These mirror the case of capital. Cluster 1 comprises of larger farms, which are relatively more productive, both in terms of the average output they can extract from IC, but also that this output effect is statistically larger than the one attributable to farms in cluster 2. Similarly to the case of capital, cluster 2 shows considerable homogeneity with regard to this effect. Furthermore, the examination of the cluster-wise and overall own elasticities shows that they mirror the case of the capital input – i.e. cluster 1 dominates in defining the total contribution of IC in Romanian agriculture.

To evaluate the biodiversity, expressed by diversity of land use within the two clusters, the histograms of the richness and equitability measures are plotted and presented in Figure 5 and 6 respectively. On both these figures the two histograms measure the relative frequencies and are overlaid over each other with semi-transparency. By choosing a light shade for cluster one and dark shade for cluster two, the intersection of the two histograms results in an intermediate shade. For each box the point of interest is which of the two clusters has greater probability, which can be easily established by looking at the shades for the top segnments - a lighter shade indicates that sector 1 has the higher probability and the darker shade corresponds to sector 2.

For both of these measures larger values signify more diversity. More specifically, for the richness measure larger values show more different land uses, which in general would be more amenable to environmental preservation and moving away from the monoculture type of agricultural system. The equitability measure, on the other hand, captures the extent to which these different forms of land use are evenly distributed – the value of zero correspond to the case where all land use is concentrated into a single use, while 1 reflects the most equitable distribution indicating that no type of land use dominates the others. It is to be

expected that more equitable land use distribution is more beneficial to the environment as it shows a higher farmed biodiversity.

Both measures demonstrate that cluster 2 is characterised by greater land use diversity. For both measures cluster two has more probability mass in the right part of the distribution (i.e. the larger values), while conversely cluster 1 has more probability mass in the left part (i.e. the smaller values). In particular, for richness, cluster 1 has higher relative probability only for the first histogram bin meaning that cluster 1 has disproportionally larger number of single land use farms. Therefore, from the point of view of the richness, cluster 2 has higher farmed biodiversity which is better for land fertility. The evidence for equitability is more mixed. Cluster 1 has greater relative probability in the first and the forth (values of 0.3 to 0.4) histogram bins. This still corroborates the result of higher land diversity use in cluster 2but the difference is not as dramatic as in the case of the richness measure. Due to these results we can define cluster 2 as more environmentally sustainable.

This leads to an interesting dichotomy in the derived classification. While cluster 1 appears to be more productive, it is a lot more concentrated in terms of land use and in conjunction with the more intensive capital use it might be characterised as less environmentally friendly than cluster 2. Hence, the balance between the two clusters might be affected by the objectives and measures of agricultural policies. Food production and trade oriented policies (such as increase of food security, boost of exports or imports substitution) may expand cluster 1, while a stronger focus on environment will benefit cluster 2.

## Conclusions

This paper proposes to use finite regression mixture models, based on an underlying relationship of interest for classification of heterogeneous units of analysis. The proposed

approach is applied to Romanian farms production function because due to the perceived heterogeneity resulting from a great extent from the legacy of transition.

The results suggest that there are at least two clusters with distinct production functions. The larger cluster 2 contains relatively smaller farms with respect to all factors of production except land. In addition to this, farms in this cluster are more labour intensive in a sense that they extract more output from their labour input. Cluster 2 is also characterised by a greater diversity in terms of land use, in other words with a higher farmed biodiversity which is beneficial to the environment. The empirical results support some more qualitative assertions that small farms are an important provider of environmental benefits (see e.g. Davidova et al., 2013; Page and Poppa, 2013).

Cluster1 consists of smaller number of relatively larger farms whose production function is more capital intensive and they manage to make a better use of their capital and intermediate consumption. This split alongside the capital-labour trade-off, and in particular the much greater heterogeneity that is observed with regard to the smaller capital-intensive cluster 1 suggests a possible explanation of traditional versus new farming technologies. In particular, this means that more traditional farming structures, most likely inheriting the technologies of the pre-transition era, are identifiable with the labour intensive sector. There is however also a new emerging capital-based agriculture. The latter is considerably more heterogeneous in terms of its production technology as farms may have been created at different stages of transition. It is, therefore, suggested that the differences between these two clusters might still bear the legacy of central planning and the emergence of new commercial farms during transition.

There are two important implications of the above farming structure. First, if the global food prices are high this would intensify the process of structural transformation exemplified by the emergence of capital-intensive farms. Since the aggregate production

function of Romanian agriculture can be viewed as a weighted average of the two underlying 'technologies', this essentially means a transition from the more labour intensive into the more capital intensive cluster. Such a transformation could perhaps surprisingly avoid the detrimental on overall employment due to the fact that it does not entail the classical 'substitution of capital for labour'. In terms of their input mix (i.e. the ratio of capital to labour) the two farm clusters are very similar which means that transformation of cluster 2 into cluster 1may not replace labour with capital, but essentially 'upgrade' the capital with a more productive one.

Second, it should not be forgotten that the values of European citizens have changed and they favour more environmentally friendly practices to more capital-intensive and productivist ones. Notwithstanding the design of CAP post-2020, it cannot be expected that the future policy will go back towards its origin and give strong incentives for farm intensification at the expense of environment. If strong environmentally focused policies are followed, these will benefit Cluster 2 which delivers more environmental goods. Such policies could constrain the future transformation of cluster 2 into cluster 1 (capital-intensive one) and hence avoid future degradation of the environmental sustainability of Romanian agriculture.

Under strongly environmentally oriented agricultural policies, farms in Cluster 1 may adjust their farming practice. A review of studies conducted by the Organisation of Economic Cooperation and Development (OECD) concluded that although there is an identified link between intensive production practices, implemented by larger farms, and a loss in farmed biodiversity, larger farms are more often signing contracts to enter agri-environmental schemes which entail a longer-term commitment (OECD, 2005). In the case of our results this may mean that some improvements in diversity of land use in Cluster 1 could be expected. However, in reality the decisions about farm practices are more nuanced. They

depend on the way producers view the relationship between output and environment protection - as a trade-off or as environment protection acting as a basis for future output sustainability. This calls for a more disaggregated study, where farm management data is augmented by detailed locational and attitudinal variables.

References:


Davidova, S., Bailey, A., Dwyer, J., Erjavec, E., Gorton, M., Thomson, K. (2013) Semi-Subsistence Farming – Value and Directions of Development, study prepared for the European Parliament Committee on Agriculture and Rural Development. Available at: http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495861/IPOL-AGRI_ET(2013)495861_EN.pdf

De Long, J. B. , Shleifer, A., Summers L. H., and R. J. Waldmann (1990) Noise Trader Risk in Financial Markets, *Journal of Political Economy* 98, 703-738.

Eurostat, Agricultural Census in Romania, statistics explained, available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_census_in_Romania#Labour_force

Farm structure survey 2013 - main results: Eurostat, statistics explained, available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Farm_structure_survey_2013_-_main_results

Kogan, L., S. Ross, J. Wang, and M. Westerfield *(*2006*)*, The Price Impact and Survival of Irrational Traders, *Journal of Finance*, 61, 195–229.

Luo, G. Y. (2012) Conservative traders, natural selection and market efficiency, *Journal of Economic Theory*, 147 (1), 310-335.

LeBaron, B., (2012) Heterogeneous gain learning and the dynamics of asset prices, *Journal of Economic Behavior & Organization,* 83 (3), 424-445.

Cogley, T., T. J. Sargent (2009) Diverse Beliefs, Survival and the Market Price of Risk, *The Economic Journal,* 119, 354-376.

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

Keribin, C. (2000) Consistent estimation of the order of mixture models, *Sankhya,* SeriesA 62(1), 49-66.

Köbrich,C., Rehman, T. and Khan, M.(2003). Typification of farming systems for constructing representative farm models: two illustrations of the application of multi-variate analyses in Chile and Pakistan, Agricultural Systems 76: 141–157.

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

OECD (2005) Farm structure and farm characteristics - Links to non-commodity outputs and externalities. OECD. Page N and Popa R (2013) Family Farming in Romania, Fundatia ADEPT Transilvania, available at: https://ec.europa.eu/agriculture/sites/agriculture/files/consultations/family-farming/contributions/adept_en.pdf

Piorr, H.P. (2003). Environmental policy, agri-environmental indicators and landscape indicators, Agriculture, Ecosystems and Environment 98: 17–33.

Popescu, A., Alecu, IN., Dinu, TA., Stoian, E., Condei, R. and Ciocan, H. (2016) Farm structure and land concentration in Romania and the European Union's agriculture. *Agriculture and Agricultural Science Procedia*, 10: 566-577.

Rizov, M., Gavrilescu, D., Gow, H., Mathijs, E. and Swinnen, J. (2001). Transition and enterprise restructuring: The development of individual farming in Romania. *World Development* 29 (7): 1257-1274.

Rusu, M., Florian, V., Popa, M., Marin, P. and Pamfil, V. (2002) Land Fragmentation and Land Consolidation in the Agricultural Sector – A Case Study from Romania , FAO (2002)

Schwarz, G., (1978) Estimating the dimension of a model, *Annals of Statistics*, 6(2), 461-464.

Straszheim, M.R. (1974) Hedonic estimation of housing market prices: A further comment. *The Review of Economics and Statistics* 56 (3): 404-406.

Straszheim, M.R. (1975) *An Econometric Analysis of the Urban Housing Market.* New York: National Bureau of Economic Research.

De Sarbo W.S. and W.L. Cron (1988) A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, 5, 249–282.

Tocco, B., Davidova, S., and Bailey, A. (2014) Labour Adjustment in Agriculture: evidence from Romania, *Studies in Agricultural Economics*, 116(2): 67-73.

Waldhardt, R., Simmering, D. and H. Albrecht (2003). Floristic diversity at the habitat scale in agricultural landscapes of Central Europe—summary, conclusions and perspectives, Agriculture, Ecosystems and Environment 98: 79–85.

Wedel, M. and W.A. Kamakura (2001). *Market Segmentation - Conceptual and Methodological Foundations*, International Series in Quantitative Marketing, Kluwer Academic Publishers, Boston, MA, 2nd edition.

Table 1. Data summary

| | Minimum | Maximum | Mean |
|---|---|---|---|
| Output | 439 | 506,142,700 | 714,715 |
| Capital | 0 | 37,216,478 | 299,879 |
| Labour (AWU) | 0 | 680 | 9 |
| Land (ha) | 0 | 21,565 | 273 |
| IC | 353 | 51,406,670 | 300,707 |

Table 2 Information Criteria Results for number of clusters

| Number of clusters | BIC | ICL |
|---|---|---|
| 1 | 2931.021 | NA |
| 2 | 2831.105 | 2931.105 |
| 3 | 2849.514 | 3783.277 |
| 4 | 2864.836 | 3931.209 |
| 5 | 2908.899 | 3906.923 |
| 6 | 2956.215 | 4279.229 |

Table 3. Bootstrapped LR test (5000 replications)

| Test | P value |
|---|---|
| 2 (NULL) vs 1 clusters | 0.72 |
| 2 (NULL) vs 3 clusters | 0.17 |

Table 4. Estimated translog for overall sample and clusters

| | All data | | Cluster 1 | | Cluster 2 | |
|---|---|---|---|---|---|---|
| | Coefficient | P-value | Coefficient | P-value | Coefficient | P-value |
| (Intercept) | 9.68 | 0.00 | 10.64 | 0.00 | 6.77 | 0.00 |
| capital | -0.21 | 0.01 | -0.33 | 0.06 | 0.08 | 0.20 |
| labour | 1.01 | 0.00 | 1.07 | 0.06 | 0.49 | 0.00 |
| land | 0.57 | 0.00 | 0.72 | 0.00 | 0.25 | 0.00 |
| ic | -0.37 | 0.00 | -0.45 | 0.10 | 0.08 | 0.18 |
| I(0.5 * capital$^2$) | 0.03 | 0.00 | 0.03 | 0.00 | 0.01 | 0.00 |
| I(0.5 * labour$^2$) | 0.05 | 0.16 | 0.00 | 0.26 | 0.31 | 0.00 |
| I(0.5 * land$^2$) | 0.13 | 0.00 | 0.14 | 0.00 | 0.16 | 0.00 |
| I(0.5 * ic$^2$) | 0.06 | 0.00 | 0.06 | 0.14 | 0.02 | 0.00 |
| capital*labour | -0.03 | 0.05 | -0.06 | 0.10 | 0.04 | 0.00 |
| capital*land | -0.02 | 0.00 | -0.03 | 0.01 | -0.01 | 0.00 |
| capital*ic | 0.01 | 0.07 | 0.02 | 0.16 | -0.01 | 0.12 |
| labour*land | -0.04 | 0.00 | 0.02 | 0.03 | -0.20 | 0.00 |
| labour*ic | -0.02 | 0.20 | -0.01 | 0.42 | -0.02 | 0.01 |
| land*ic | -0.05 | 0.00 | -0.07 | 0.00 | -0.01 | 0.06 |

Note: variable labels refer to variables in natural logarithms (i.e. capital is the natural logarithm of the capital variable)

Table 5 Summary statistics for the clusters

| | Cluster 1 | | | |
|---|---|---|---|---|
| | mean | sd | min | max |
| Output, 000s | 1,530 | 22,563 | 0 | 506,143 |
| capital, 000s | 472 | 2,219 | 0 | 37,216 |
| labour | 14 | 54 | 0 | 680 |
| land | 219 | 757 | 0 | 11,196 |
| ic, 000s | 763 | 9,502 | 0 | 212,143 |

| | Cluster 2 | | | |
|---|---|---|---|---|
| | mean | sd | min | max |
| Output 000s | 225 | 660 | 1 | 9,978 |
| capital, 000s | 197 | 732 | 0 | 15,334 |
| labour | 6 | 14 | 0 | 142 |
| land | 306 | 1,211 | 0 | 21,565 |
| ic, 000s | 182 | 980 | 0 | 23,479 |

Table 6 Elasticities (own)

| | capital | labour | land | ic |
|---|---|---|---|---|
| All data | | | | |
| Average | 0.13 | 0.44 | 0.26 | 0.18 |
| SD | 0.06 | 0.12 | 0.26 | 0.13 |
| Cluster1 | | | | |
| Average | 0.13 | 0.44 | 0.23 | 0.19 |
| SD | 0.06 | 0.12 | 0.25 | 0.13 |
| Cluster2 | | | | |
| Average | 0.10 | 0.44 | 0.40 | 0.10 |
| SD | 0.05 | 0.39 | 0.30 | 0.03 |

Figure 1. Effect of capital

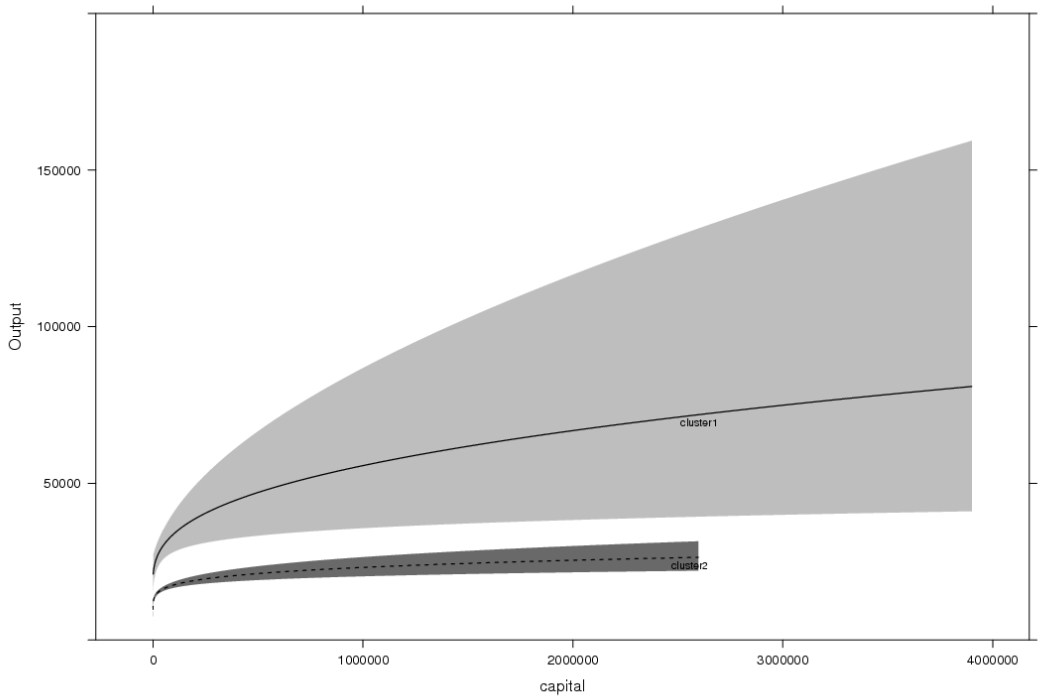


Figure 2. Effect of labour

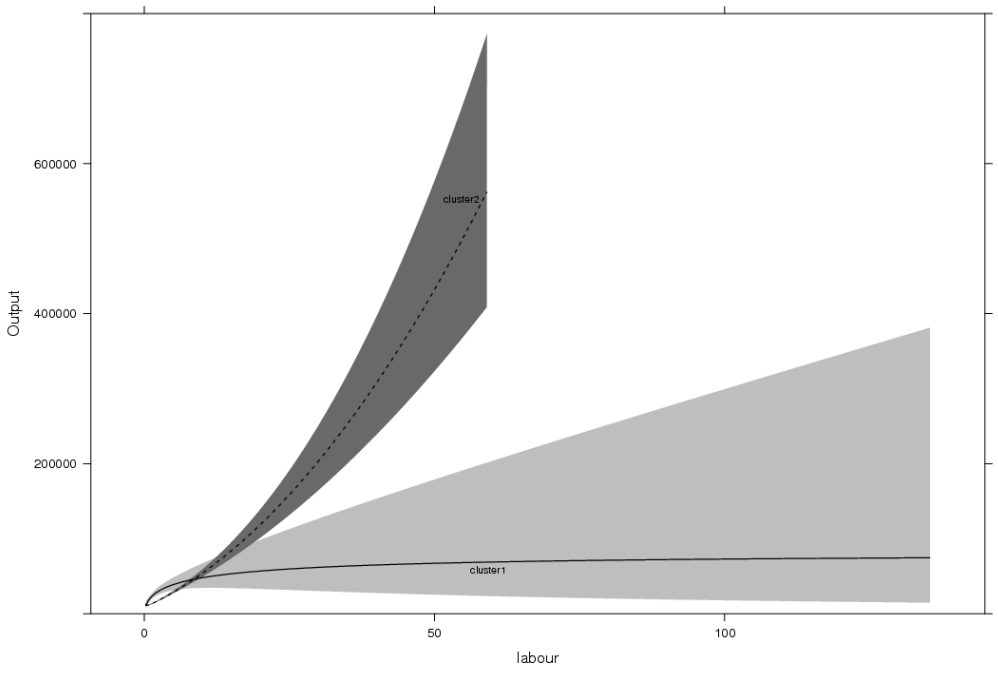

Figure 3 Effect of land

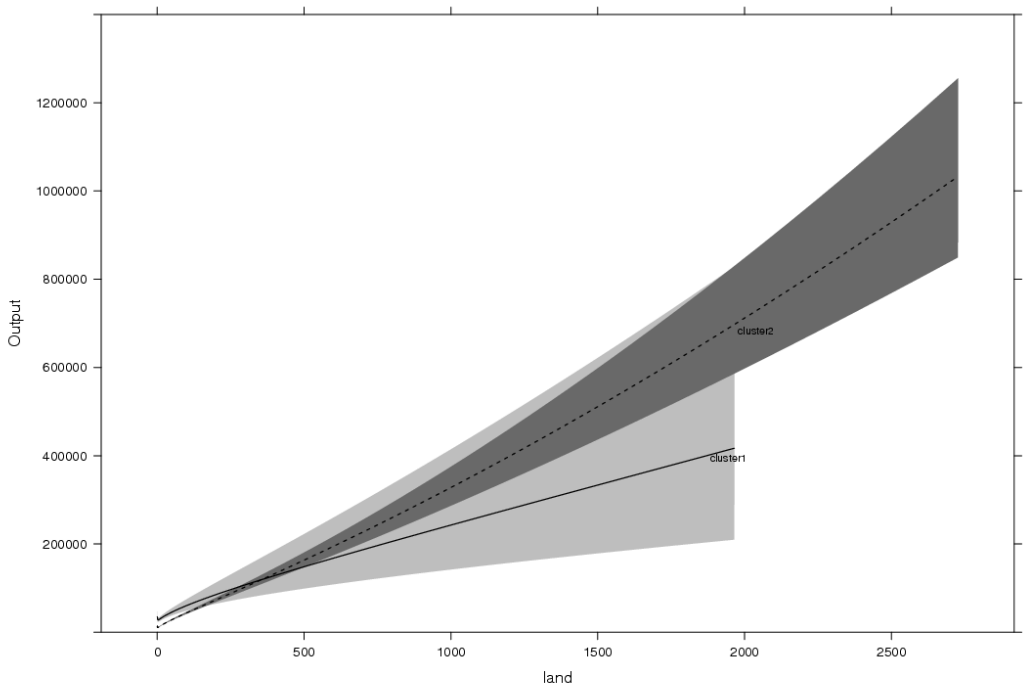


Figure 4. Effect of IC

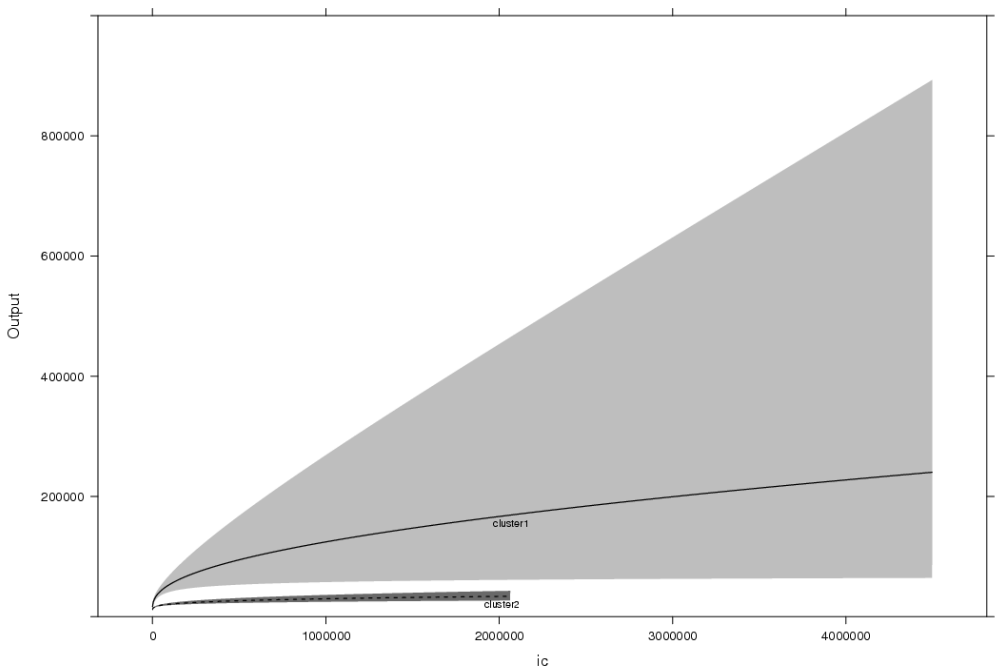

Figure 5. Richness distribution across the two clusters

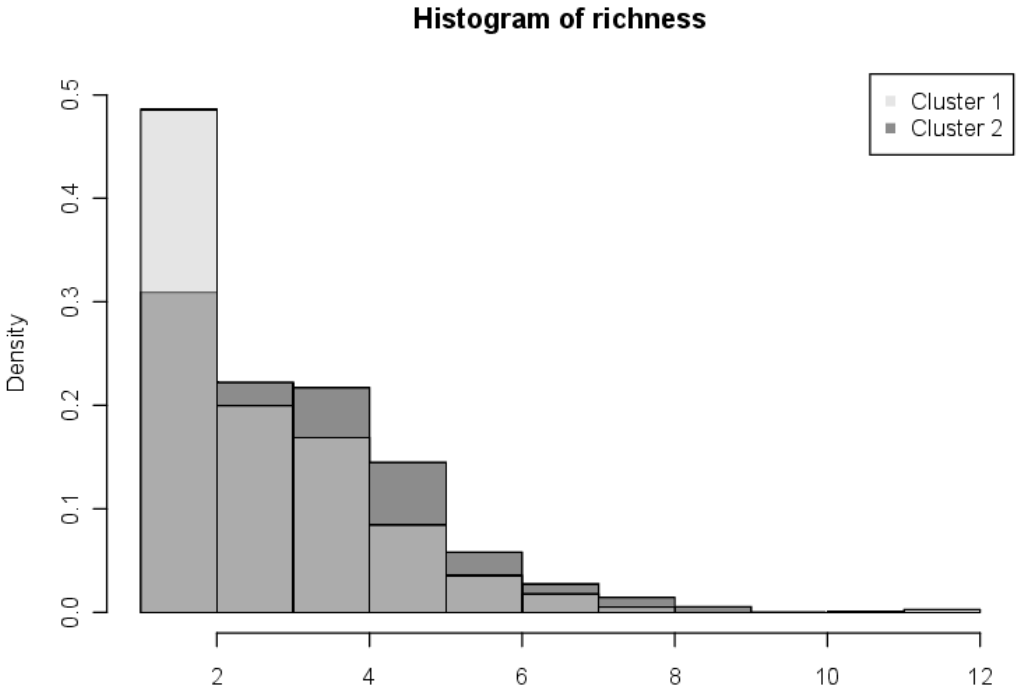

Figure 6. Equitability distribution across the two clusters