STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 615—638

A TYPOLOGY OF POLISH FARMS USING PROBABILISTIC D–CLUSTERING

Andrzej Młodak[1], Jan Kubacki[2]

ABSTRACT

The Agricultural Census conducted in Poland in 2010 was partially based on administrative sources. These data collection will be supplemented by sample survey of agricultural farm. This research is aimed at creation of an effective typology of Polish farms, which is necessary for proper sampling and reflection of many special types of agricultural activity, such as combining it with non–agricultural work. We propose some universal form of such typology constructed using data collected from administrative sources during the preliminary agricultural census conducted in autumn 2009. It is based on the especially prepared method of fuzzy clustering, i.e. probabilistic d–clustering adopted for interval data. For this reason, and because of an ambiguous impact of some key variables on classification, relevant criterions are presented as intervals. They are arbitrarily established, but also – as an alternative way – are generated endogenically, using an original optimization algorithm. For a better comparison, relevant classification for data collected "from nature" is provided.

Key words: agricultural census, probabilistic d–clustering, interval data.

1. Introduction

The Agricultural Census in Poland in 2010 was conducted according to significantly different rules than those used in its previous exercises. The main source of information gathered during the census was administrative databases. For instance, most of such data was collected from the farm registers maintained by local self – government authorities (such as Tax Register of Real Estates, Register of Lands and Buildings) or by the Agency for Restructuring and Modernization of Agriculture, which is engaged in services of applications of farmers concerning subventions from the European Union budget. The direct detailed survey (concerning mainly methods of agricultural production) was

[1] Statistical Office in Poznań, Urban Statistics Centre. E–mail: a.mlodak@stat.gov.pl.

[2] Statistical Office in Łódź, Centre of Realization of Statistical Surveys. E–mail: j.kubacki@stat.gov.pl.

planned to be conducted on 30% sample of farms with agricultural land area up to 1 ha and some farms with this area between 1 ha and 2 ha (the remaining were planned to be examined in an exhaustive way). The farms for which no or very few administrative data are available were also additionally interviewed (target survey).

Taking these expectations into account and following the fact that structure of Polish farms gradually changes (among others due to the accession of Poland to the European Union) and competitiveness on the market of agricultural products is ever-increasing, there is a necessity to construct a typology of farms which could allow these changes to be reflected also in future statistical agricultural surveys. Additionally, it is connected with the fact that the Agricultural Census will be conducted according to the rules of the Farm Structure Survey adopted within the EU (Regulation (EC) No 1166/2008 of the European Parliament and of the Council). It means that this survey shall be carried out in the form of a census, i.e. it should cover farms with area greater than 1 ha. In the Polish circumstances a necessity to examine smaller farms also occurs. They are not, however, a significant part of the overall population of farms, and their production is not especially significant, but due to the above mentioned special character of the Polish agriculture, a survey of them seems to be also required. Moreover, it is very important to reflect many special or mixed types of agricultural activity, such as combining it with non–agricultural work.

It is clear that a creation of a universal typology of agricultural holdings is very difficult. Some scientists argue even that, in practice, it is not feasible. Nevertheless, some – at least relatively efficient – typology is necessary to properly conduct the statistical surveys. We have undertaken a trial to construct such categorization on the basis of a dataset which was assumed to be available for all farms at the moment when the census started. It was burdened with some inconveniences, which we have tried to eliminate, although it was not always fully possible. However, the most serious of them seem to be minimized.

To satisfy these expectations we have constructed a basic and universal typology of farms using some fuzzy probabilistic d–clustering. It is a generalization of the proposal of A. Ben – Israel and C. Iyigun (2008) into an interval case. Because many key features determining a character of the farm are described by interval or ratio variables with continuous distribution of observation, the criterion of classification of a farm to a given class should be expressed by a set of intervals reflecting typical scope of values of relevant variables realized within this class. For any object (farm) we determine a class such that the probability of belonging of a farm to it is the greatest. The object will be assigned to this class. The criterion intervals can be established arbitrarily or determined using an endogenous optimization based on derivation of interval-valued function. Both approaches are here presented.

Our method differs significantly from many popular fuzzy classification algorithms, such as c-means Bezdek's approach (J. C. Bezdek (1973), R. J. Hathaway and J. C. Bezdek (1988)), Gustafson – Kessel method (D. E. Gustafson and W. C. Kessel (1979)), Gath – Geva probabilistic suggestion (I. Gath and, A. B. Geva (1989)) or even unsupervised FCM–NM algorithm based on normalized Mahalanobis distance (J. – M. Yih and S. – F. Huangh (2010)). All these proposals have a common feature – they are based on the fuzzy c–means procedure and are point–oriented. That is, the clusters are determined by their point centers (usually centroids). Moreover, some of them tend, for instance, to create spherical shaped clusters (Bezdek's method) or deform original diversification of objects, which is very important in multivariate analysis (FCM–NM). Our method reflects the common practical situation, when classes are defined by reference intervals of respective variables. The three area groups of farms (above 2 ha, 1–2 ha and 0–1 ha) can be the simplest and very good example in this context. Therefore, an original distance of a point from an interval was defined. The variations of diagnostic variables are kept. The computation seems to be also much faster, because in its basic part it is non–iterative.

The presented experiment is a test study using data collected during the preliminary (trial) census before the main Agriculture Census conducted in autumn 2009. It was conducted in all farms located in the following four rural gminas (Polish NUTS 5 territorial units):

- Gniezno (Wielkopolskie Voivodship – Polish NUTS 2 region),
- Kamień (Podkarpackie Voivodship),
- Kołobrzeg (Zachodniopomorskie Voivodship) ,
- Rutki (Podlaskie Voivodship).

As a result of this survey, two databases have been constructed. The first of them is called the 'master record' and contains all data collected from administrative sources. This file was the main basis of our classification, because on a similar file (but covering the whole population of farms) the sampling and other primarily census activities will have to be performed. The second ('gold record') consists of information received "from nature", i.e. directly from the farms using the modern interview techniques (such as CATI – Computer Assisted Telephone Interview, CAPI – Computer Assisted Personal Interview, etc.) and, of course, is much greater than the "master record", because the scope of information which can be gathered during individual contact with respondent is much more broader than collected in official registers. The "gold record" reflects then the information collected directly from respondents during the preliminary census whereas the "master record" – the data from administrative sources for the same respondents obtained directly during this census (and which were assumed to be collected also just before the main census). Therefore, to assess an efficiency of our construction, we have compared our results with those which can be obtained using this more detailed information contained in the "gold record" and using the same theoretical methods.

The paper is organized as follows. Firstly (chapter 2), we present our proposal of classification of farms with its justification. The chapter 3 contains a list of variables used to determine the classes of farms. Most of them are constructed especially for our investigation using the available information in the "master record" file. We explain exactly the methods of their computation. Next (chapter 4) the classification algorithm and method of endogenous optimization of criterion intervals is described. The empirical results of analysis and computation are given in chapter 5. Finally (chapter 6), main conclusions are formulated.

2. Main assumptions and proposal of a typology

The final form of the classification has been elaborated on the basis of many consultations with experts dealing with agricultural statistics. It was agreed that the classification should take into account three essential aspects of the analyzed problem. Firstly, in each of the size groups some farms conducting no agricultural activity can occur. Secondly, among farms conducting such activity it is worth to select these which specialize mainly in crop production and these in which animal production is prevailing direction of their activity. Theoretically, this division can be executable on the basis of results of Farm Structure Survey (FSS), but it is mainly sample (10% sample of individual farms). Because of this, it was assumed that the basis of full classification prepared before the census will be data coming from administrative sources. Of course, the scope of information available there is smaller that could be obtained from direct survey (such as FSS). This fact should be reflected in the main division.

Finally, we propose the following typology of farms:

1) Farms with the agricultural land area above 2 ha:
   a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C1A),
   b. Farms conducting productive agricultural activity with prevalence of the animal output (C1B),
   c. Farms which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C1C)
   d. Farms conducting no agricultural activity (C1D).
2) Farms with the agricultural land area between 1 ha and 2 ha:
   a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C2A),
   b. Farms conducting productive agricultural activity with prevalence of the animal output (C2B),
   c. Farms, which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C2C)
   d. Farms conducting no agricultural activity (C2D).
3) Farms with the agricultural land area below 1 ha:
   a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C3A),

b. Farms conducting productive agricultural activity with prevalence of the animal output (C3B),

c. Farms, which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C3C)

d. Farms conducting no agricultural activity (C3D).

This proposal has a introductory character. That is, when more data will be available it will be further developed by adding new subclasses defined by a nominal criterion based on values of a new variable. For example, if due to some local circumstances we would like to select within the C1A class, farms planting the flax, we should find all farms belonging to C1A, for which the sown area of flax is greater than 0.

3. Classification variables

Our approach is based on variables measured on interval or ratio scale. A restriction only to the nominal variables results in significant restriction of information on the size and structure of a given phenomenon, which could be quite obvious in practice. Therefore, the interval or ratio variables should be preferred. But, as we stated in the previous paragraph, the nominal variables could be used rather to create more detailed classification levels, what seems also to be easy for potential users of the classification. This assumption is, however, followed by many additional problems. For example, we have to decide, which farms should be classified to the group of farm where area of agricultural land used to the crop production is very small in relation to their total land area. It is also worth noting that in the case of interval or ratio variables, adherence of an object to a given class is usually expressed by some tolerance set of values (understood, in general, as a real interval). The collection of presently used variables satisfies these postulates.

Taking into account the scope of information available in administrative sources, we have proposed the following set of classification variables:

1) Total agricultural land area in ha (denoted further as *Land*)
2) Coefficient of intensity of crop production (in %) (*Crop*)
   It is defined as

$$Crop = \begin{cases} \frac{grc}{pzw+grc} \text{ if } pzw + grc \neq 0, \\ 0 \text{ if } pzw + grc = 0, \end{cases}$$

   where *grc* is the area of land under the agricultural activity used to the crop production in a farm, and *pzw* denotes the size of stocks in farms raising animals recalculated into main forage area.
3) Coefficient of intensity of animal production (in %) (*Animal*)
   It is defined as

$$Animal = \begin{cases} \frac{pzw}{pzw+grc} \text{ if } pzw + grc \neq 0, \\ 0 \text{ if } pzw + grc = 0, \end{cases}$$

where *grc* and *pzw* are as above.

4) Share of agricultural land maintained in good agricultural culture in the total land area of a farm (in %) (*Culture*)
5) Share of meadows and pastures in the total land area of a farm (%) (*Meadows*).

The *grc* will be computed as a sum of area of land under particular crops or used for agricultural production in another way (e.g. as orchards, tree and bush nurseries, fixed crops under cover – such as mushrooms, etc.).

The quantity *pzw* is computed as a number of farm animals being in a given farm recalculated per Livestock Units (LU; done by multiplication of number of particular animals by respective coefficients established in relevant EU and domestic authorities regulations – cf. e.g. A. Tonini (2007), A. Tonini and R. Jongeneel (2007) or H. Lipińska and J. Gajda (2006)) and divided by the average population of LU per 1 ha of Main Forage Area (MFA) in agricultural regions. In Poland, 4 large agricultural regions specified from the point of view of natural conditions and potential for agricultural development have been established. They have the following values of the MFA per one LU: 0.76 (Pomorze and Mazury – northern and north–western regions), 1.50 (Wielkopolska and Śląsk – western and south–western parts of Poland), 1.31 (Mazowsze and Podlasie – central and north – eastern regions) and 0.90 (Małopolska and Pogórze – southern part of the country). In the trial census each agricultural region was represented by one gmina.

The coefficients *Crop* and *Animal* show which type of production has the main importance in the agriculture. If one of them is greater than 50% then the latter must be smaller than 50%. If one of them amounts to zero and the latter does not, then the farm is regarded to be concentrated only on the production of type represented by non–zero index. If no production is conducted then both indices are equal to zero.

This set of variables is, of course, not ideal. Due to practical reasons, it had to be based, however, only on the database assumed to be collected for all farms directly before the main census (from administrative sources). As mentioned in Chapter 5, it was of non–satisfactory quality due to lack of a harmonization of various registers. Our trail to improve the quality of the "master record" has not, of course, eliminated all inconveniences, but minimized only most serious of them. On the other hand, it is commonly regarded that the specialization of production could be better assessed using the structure of marketable output or standard gross margin. These data are available only from The Farm Structure Survey (based only on a relatively small – 10% – sample of individual farms) but not from the administrative sources and therefore they cannot be used for the census purposes.

When using traditional classification method for the internal structure of a set of the farms described by a number of variables, the orders of magnitude of these variables need to be standardized to retain the uniform influence of the individual variables on the calculated distances. In our case, all the variables considered are "by definition" normalized on [0,1] (or, more precisely, on [0%, 100%]) because their values are presented in %. Therefore, no additional normalization seems to be necessary (as the basic characteristics of their distributions would remain practically unchanged after such transformation).

4. Classification algorithm

There exist many various methods of cluster analysis. Most of them generates, however, the classes in an endogenous way (e.g. by representatives, centroids or optimal thresholds of similarity), being exclusively a result of performance of the clustering process and properties of the model (cf. B. S. Everitt et al. (2001)). One can obtain then high–quality clusters, which are usually rather hard to interpret. In the analyzed case we have the classes being established arbitrarily due to some external circumstances (e.g. expectations of users). Therefore, the criterion of appurtenance also should be fixed "in advance". Due to variety of measurement methods and statistical properties of analyzed variables, the most reasonable solution seems to be a unique characterization of particular classes by intervals reflecting scopes of required, or typical realization of variables for farms belonging to these classes.

The classification method can be described as follows. Let $n \in \mathbb{N}$ denotes the number of objects (farms, in this case), and $m \in \mathbb{N}$ – the number of features (variables) characterizing these objects. Thus, we have at disposal m features $X_1, X_2, \dots, X_m$. Denote by x_{ij} a value of the feature X_j for i–th object, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$. Due to the properties of our specific model we assume that all observations are nonnegative. The set of all analyzed objects will be denoted as Γ. Each object belonging to Γ is uniquely represented by the vector $\gamma_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$. Assume that $k \in \mathbb{N}$, $1 \leq k \leq n$ is the fixed number of typological classes which the set Γ we would like to divide into. Our purpose is then to obtain a sequence of subsets $\Omega_r \subseteq \Gamma$, $r = 1,2,\dots,k$, such that $\Omega_r \cap \Omega_q = \emptyset$ for every $r, q = 1,2,\dots,k$, $r \neq q$ and $\bigcup_{r=1}^{k} \Omega_r = \Gamma$.

Each of the k proposed typological classes has to be described by unique criterions for allocation of a given object to it. For the feature X_j we determine then k_j ($k_j \in \mathbb{N}$, $2 \leq k_j \leq k$) of criterion intervals $\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{k_j j}$, such that $\bigcup_{q=1}^{k_j} \lambda_{qj} = \mathbb{R}_+ \cup \{0\} = [0,\infty)$, $j = 1,2,\dots,m$. The intervals are desired to be disjoint, although it is not absolutely necessary. According to these conditions, the interval λ_{rj} is of the form $\lambda_{qj} = [a_{qj}, b_{qj}) \subseteq \mathbb{R}_+$, where $a_{qj} < b_{qj}$, $q = 1,2,\dots,k_j - 1$, $j = 1,2,\dots,m$. Assume that any class Ω_r is determined by an interval vector $\Phi_r = (\varphi_{r1}, \varphi_{r2}, \dots, \varphi_{rm})$, where $\varphi_{rj} = [\alpha_{rj}, \beta_{rj}) \subseteq \mathbb{R}_+$ is the interval belonging to the set of criterion intervals of a given feature, i.e. $\varphi_{rj} \in \{\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{k_j j}\}$ (and therefore $\alpha_{rj} = a_{qj}$ and $\beta_{rj} = b_{qj}$ for some $q \in \{1,2,\dots,k_j\}$), and selected to establishment of a criterion characterizing this class, $r = 1,2,\dots,k, j = 1,2,\dots,m$.

For a better precision of further analysis, we have now to introduce a definition of a distance of a real number *y* from the interval $U = [u_1, u_2] \subseteq \mathbb{R}$, $u_1 \leq u_2$. We do this using the formula:

$$\delta(y,U) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } y \in U, \\ \min(|y-u_1|, |y-u_2|) & \text{if } y \notin U. \end{cases} \qquad (1)$$

Note that the definition (1) has a sense also if one of the limits of the interval U is infinite. In such case, if $y \notin U$, we assume as a distance the absolute value of a difference between *y* and the finite limit of *U*.

An aggregated distance of *i*–th object (represented by the vector γ_i) from the *r*–th typological class Ω_r described by the criterions Φ_r is defined to be a maximum of partial distances from particular criterion intervals, i.e.

$$d(\gamma_i, \Phi_r) \overset{\text{def}}{=} \max_{j=1,2,\dots,m} \delta(x_{ij}, \varphi_{rj}) \qquad (2)$$

$r = 1,2,\dots,k$, $i = 1,2,\dots,n$. That is, the distance from an object to a class is computed by calculation of the distances of data for variables describing a given object from respective intervals describing the class (expressed by (1)) and next a determination of maximum of them. This choice enables one to avoid a compensation of discrepancy in respect to some criterion by a similarity connected with other criterion, what is unfavorable from the point of view of the classification.

The purpose of our analysis is to determine optimum probabilities of assignment of an object represented by the vector γ_i to typological classes determined by the criterions $\Phi_1, \Phi_2, \dots, \Phi_k$. In this context, the key postulate is that this probability should be reversely proportional to a distance of the object from the given class. Therefore, it is proposed to apply the model of so–called probabilistic d–clustering (cf. e.g. A. Ben–Israel and C. Iyigun (2008)). It belongs to the tools of fuzzy classification. In the investigated case we would like to find numbers $p_k(\gamma_i)$, $r = 1,2,\dots,k$, $i = 1,2,\dots,n$, which minimize the value of the target function:

$$f(p_1(\gamma_i), p_2(\gamma_i), \dots, p_k(\gamma_i)) = \sum_{i=1}^{n} \sum_{r=1}^{k} d(\gamma_i, \Phi_r)\, p_r^2(\gamma_i) \quad (3)$$

with the conditions:

$$\sum_{r=1}^{k} p_r(\gamma_i) = 1$$

$$p_r(\gamma_i) \geq 0$$

for every $r = 1,2, \dots, k$, $i = 1,2, \dots, n$.

The results presented in the cited article can be applied also in this case. Assuming reasonable requirement that $p_r(\gamma_i)\, d(\gamma_i, \Phi_r) =$ const. (depending on γ_i) for every $r = 1,2, \dots, k$, $i = 1,2, \dots, n$, the optimum probability of assignment of an object represented by the vector γ_i to the class Ω_r described by Φ_r, is given by the formula:

$$p_r^*(\gamma_i) = \frac{\prod_{\substack{q=1,2,\dots,k \\ q \neq r}} d(\gamma_i, \Phi_q)}{\sum_{q=1}^{k} \prod_{\substack{u=1,2,\dots,k \\ u \neq q}} d(\gamma_i, \Phi_u)}, \qquad (4)$$

for every $r = 1,2, \dots, k$, $i = 1,2, \dots, n$. The object represented by the vector γ_i is assigned to such class for which the probability of assignment expressed by (4) is the greatest. When the optimal probabilities (4) for two or more classes are identical, then the classification of the object will be determined by maximum partial distance from the particular unit criterion. As we can conclude from the formula (4), a significantly important problem is such choice of limits of criterion intervals that enables to exclude a possibility of an occurrence of zero distances of an object from two or more different classes (what could result in the zero value of the denominator in (4)). Unlike many other classification algorithms, this approach is non-iterative, because the algorithm of assignment and – by the same token – the optimum class for a given farm is exactly computed using the formula (4) derived by mathematical methods. An iteration will be used to obtain the theoretical optimum classes, as we will describe in the next part of this chapter.

Some assessment of quality of obtained classification one could obtain by determination of k interval criterion vectors by an iterative algorithm, originally proposed using some ideas coming from papers by A. Ben–Israel and C. Iyigun (2008) and C. Iyigun (2007) and being a development of some concepts suggested by E. Weiszfeld (1937), adopted to our specific situation. To obtain an optimal criterion division $\Phi_1^*, \Phi_2^*, \dots, \Phi_k^*$ for the exercise (3), we have to consider the problem of differentiation of a function defined on a set of closed intervals contained in a real line. Let $\mathbb{IR} \overset{\text{def}}{=} \{[a,b]: a,b \in \mathbb{R}, a \leq b\}$ be this set. Of course, $\mathbb{R} \subseteq \mathbb{IR}$, because a real number is an interval itself (although of a thin form, i.e. with equal limits). Let $Y = [y_1, y_2] \in \mathbb{IR}$ will be non–trivial, i.e. $y_1 < y_2$. Consider the function $g: \mathbb{IR} \longrightarrow \mathbb{R}$ defined on its whole domain. Let h be some real number such that $\xi_{Y,h} \overset{\text{def}}{=} [y_1 + h, y_2 - h] \in \mathbb{IR}$. A lower and upper derivation of the function g at its argument Y are defined respectively as:

$$\underline{g'(Y)} = \lim_{h \to 0^-} \frac{g(\xi_{Y,h}) - g(Y)}{h} \text{ and } \overline{g'(Y)} = \lim_{h \to 0^+} \frac{g(\xi_{Y,h}) - g(Y)}{h}.$$

If $\underline{g'(Y_0)} = \overline{g'(Y_0)}$, then we say that the function g is differentiable at the argument Y_0, and its derivation at this argument will be denoted as $g'(Y_0)$ or $\left.\frac{\partial g}{\partial Y}\right|_{Y_0}$.

Let us come back to our distance (1), which – investigated as a function of intervals U – belongs to the analyzed family of interval functions. For its argument $V = [v_1, v_2] \in \mathbb{IR}$, we have then:

$$\left.\frac{\partial \delta(x,U)}{\partial U}\right|_V = \begin{cases} 0 \text{ if } (v_1, v_2) \ni x, \\ 1 \text{ if } v_2 < x \text{ or } v_1 > x. \end{cases}$$

Intervals of the form $[a, x], [x, b] \in \mathbb{IR}$ are places where the function δ is not differentiable. In both cases the lower derivation in such place equals to 0 but the upper amounts to 1.

Taking these observations into account, we can determine theoretical optimum criterion intervals. Our main purpose is to determine such limits of these intervals that the value of the function (3) was minimum. For any class Ω_r, $r = 1,2,\dots,k$, we consider two cases:

Case 1. The class Ω_r ($r \in \{1,2,\dots,k\}$) has such property that no object is strictly identifiable as belonging to it, i.e. $d(\gamma_i, \Phi_r) > 0$ for every $i = 1,2,\dots,n$. Then, the solution of these problems is to find such interval arguments for which the gradient of the function (3) restricted to this class amounts to zero. More formally, we would like for every $j = 1,2,\dots,m$ to find such intervals $\varphi_{rj} = [\alpha_{rj}, \beta_{rj}] \in \mathbb{IR}$, that

$$\frac{1}{m}\sum_{i=1}^{n} \frac{\partial \delta(x_{ij}, \varphi_{rj})}{\partial \varphi_{rj}} \frac{\delta(x_{ij}, \varphi_{rj})}{d(\gamma_i, \Phi_r)} p_r^2(\gamma_i) = 0. \tag{5}$$

Taking into account our conclusions about differentiation of the function δ, the equality (5) holds if and only if

$$\sum_{\substack{i=1,2,\dots,n: \\ x_{ij}<\alpha_{rj}}} \frac{(\alpha_{rj} - x_{ij}) p_r^2(\gamma_i)}{d(\gamma_i, \Phi_r)} + \sum_{\substack{i=1,2,\dots,n: \\ x_{ij}>\beta_{rj}}} \frac{(x_{ij} - \beta_{rj}) p_r^2(\gamma_i)}{d(\gamma_i, \Phi_r)} = 0 \tag{6}$$

for every $j = 1,2,\dots,m$.

Because both components of the sum on left–hand side of (6) are nonnegative, then the equality (6) holds only if each of them equals to zero. After relevant transformations we obtain the optimum limits of intervals of the form:

$$\alpha^*_{rj} = \sum_{\substack{i=1,2,\dots,n, \\ x_{ij}<\alpha^*_{rj}}} \frac{\frac{p_r^{*2}(\gamma_i)}{d(\gamma_i, \Phi^*_r)} x_{ij}}{\sum_{\substack{h=1,2,\dots,n, \\ x_{hj}<\alpha^*_{rj}}} \frac{p_r^{*2}(\gamma_h)}{d(\gamma_h, \Phi^*_r)}}, \tag{7a}$$

and

$$\beta^*_{rj} = \sum_{\substack{i=1,2,\dots,n,\\ x_{ij}>\beta^*_{rj}}} \frac{\frac{p_r^{*2}(\gamma_i)}{d(\gamma_i,\Phi^*_r)} x_{ij}}{\sum_{\substack{h=1,2,\dots,n,\\ x_{hj}>\beta^*_{rj}}} \frac{p_r^{*2}(\gamma_h)}{d(\gamma_h,\Phi^*_r)}}, \tag{7b}$$

$r = 1,2,\dots,k, j = 1,2,\dots,m.$

Case 2. Let Ω_r ($r \in \{1,2,\dots,k\}$) be a class such that there exist objects strictly identifiable as belonging to it, i.e. for some $i \in \{1,2,\dots,n\}$ we have $d(\gamma_i,\Phi_r) = 0$. Let Ξ_r be a set of objects strictly identifiable as belonging to the class Ω_r. Then the approximation of limits of optimum intervals can be obtained respectively as minimum and maximum values of respective features, i.e.

$$\alpha^*_{rj} = \min_{i:\in\Xi_r} x_{ij} \text{ and } \beta^*_{rj} = \max_{i\in\Xi_r} x_{ij} \tag{8}$$

for every $j = 1,2,\dots,m$.

The algorithm is now iterative. We start from arbitrarily fixed criterion intervals; the optimum probabilities (4) of appurtenance of objects to them are determined. Next, using the formulas (7a), (7b) or (8) and inserting to their right–hand sides all estimated values, we obtain first iteration of the optimum classes. Next, using them, we perform the second iteration and so on. We stop the procedure, when the distance between criterion structures of two successive iterations will be smaller than an arbitrarily established positive threshold ε. The distance of two criterion structures is calculated using the formula (where $\mathbf{\Phi} = (\Phi_1,\Phi_2,\dots,\Phi_k)$, $\mathbf{\Phi}' = (\Phi'_1,\Phi'_2,\dots,\Phi'_k)$ is assumed):

$$d_\#(\mathbf{\Phi},\mathbf{\Phi}') = \frac{1}{k}\sum_{r=1}^{k}\sqrt{\frac{1}{m}\sum_{j=1}^{m} d^2_{\mathcal{H}}(\varphi_{rj},\varphi'_{rj})},$$

where $d_{\mathcal{H}}$ is the Hausdorff distance between respective intervals. The Hausdorff distance between two intervals $U = [u_1,u_2], W = [w_1,w_2] \subseteq \mathbb{R}$, $u_1 < u_2$, $w_1 < w_2$, is defined as:

$$d_{\mathcal{H}}(U,W) = \max(|w_1 - u_1|,|w_2 - u_2|).$$

Therefore, the iteration is continued until $d_\#(\mathbf{\Phi},\mathbf{\Phi}') \leq \varepsilon$, where $\mathbf{\Phi}$, and $\mathbf{\Phi}'$ denote structures obtained in two subsequent iterations. Of course, the optimum collection of criterion intervals can contain for some variables also non–disjoint intervals.

It is worth noting that this method differs significantly from other well–known fuzzy classification algorithms. All of them are based on the following objective function.

$$f(p_1(\gamma_i), p_2(\gamma_i), \dots, p_k(\gamma_i), \bar{\varphi}_r) = \sum_{i=1}^{n} \sum_{r=1}^{k} d^2(\gamma_i, \bar{\varphi}_r)\, p_r^q(\gamma_i), \qquad (9)$$

where $\bar{\varphi}_r$ is the centroid of the group Φ_r, $q \in \mathbb{N}$ is fixed and $d^2(\gamma_i, \bar{\varphi}_r)$ is the distance of the object Γ_i from the centroid of respective group Φ_r ($i = 1,2,\dots,n$, $r = 1,2,\dots,k$), is defined in various ways. J. C. Bezdek (1973) and R. J. Hathaway and J. C. Bezdek (1988) define it to be the Euclidean norm on $\mathbb{R}^m$, D. E. Gustafson and W. C. Kessel (1979) as a modified Mahalanobis distance with preserved volume, in Gath–Geva approach (I. Gath and A. B. Geva (1989)) the distance is defined using the posterior probability function assuming that the normal distribution with expected variance and covariance matrix is chosen for generating a datum with prior distribution. Finally, the FCM – NM algorithm (J. – M. Yih and S. – F. Huangh (2010)) is based on the normalized Mahalanobis distance. All these algorithms are iterative and belong to the c–means clustering "family", i.e. they consist of iterations starting either with an initial guess for partitioning on prototype (centroid) vectors $\bar{\varphi}_r$ and is continued until the distances between two successive iterations are sufficiently small. That is, iteration stops with the first $\Phi^{(u)}$ such that $\left\|\Phi^{(u)} - \Phi^{(u-1)}\right\| < \varepsilon$, where ε is the arbitrarily established threshold of accuracy, $\Phi^{(u)}$ is the partition obtained at the u–th step, $u = 1,2,\dots$. Each of these concepts has its disadvantages: the Bezdek's method tends to create spherical clusters, in the Gustafson–Kessel method the added fuzzy covariance matrices in their distance measure are not directly described, in the Gath–Geva algorithm the assumption that the data are multivariate normally distributed can be inappropriate in practice. And, finally, the FCM–NM proposal deforms the original variation of the diagnostic variables.

Our method, although belonging to the fuzzy clustering tools (compare the objective functions (3) and (9)), seems to be much more practically useful. The typological classes are usually defined using the reference (or tolerance) intervals for particular variables and it satisfies this postulate. Moreover, the optimization is very simple and enables to compare practical criterions with their artificial but theoretically optimum equivalences with no significant influence of some inconvenient aspects, e.g. sphericality.

Here one can also compare the form of membership matrix (see for example formula (2) in I. Gath and A. B. Geva (1989)) and the probability of assignment (see formula (4) in our work), what reveals that some ideas are common in both approaches, but particular implementation of the algorithms may be different and sometimes leads to different results. More detailed analysis related to such comparison may be done in the future.

5. Results of classification

According to our assumptions, to perform the classification, one should define effective criterion intervals. They were established using main requirements concerning particular class or, if they are not specified, as typical observation for respective groups of farms occurring in other, but similar, statistical surveys (such as FSS) conducted in the near past. Our final choice is presented in Table 1. The upper element of each cell denotes the lower limit of the respective interval, and the lower one – its upper limit.

Table 1. Arbitrarily assumed criterion intervals

Variable	C1A	C1B	C1C	C1D	C2A	C2B	C2C	C2D	C3A	C3B	C3C	C3D
Land	2 10000	2 10000	2 10000	2 10000	1 1.999	1 1.999	1 1.999	1 1.999	0 0.999	0 0.999	0 0.999	0 0.999
Crop	50 100	0 49.9999	0 0	0 0	50 100	0 49.9999	0 0	0 0	50 100	0 49.9999	0 0	0 10
Animal	0 49.9999	50 100	0 0	0 0	0 49.9998	50 100	0 0	0 0	0 49.9999	50 100	0 0	0 0
Culture	50 100	0 49.9999	0.00001 100	0 0	50 100	0 49.9999	0.00001 100	0 0	50 100	0 49.9999	0.00001 100	0 0
Meadows	10 39.9999	40 100	0 9.9999	0 0	10 39.9999	40 100	0 9.9999	0 0	10 39.9999	40 100	0 9.9999	0 0

Source: Authors' elaboration.

The 'master record' is a file being a compilation of data from several various administrative sources, such as Tax Register of Real Estates or database maintained by the Agency for Restructuring and Modernization of Agriculture. Due to significant differences between these sources in terms of timeliness and scope of information, many contradictions within the data could be observed. The most important of them are:

- for 493 records the area of agricultural land maintained in good agricultural culture is larger than the total area of agricultural land,
- for 2 other records the area of meadows and pastures is larger than the total area of agricultural land,
- for other 112 records the area of arable land is larger than the total area of agricultural land,
- area of meadows and pastures is positive, but the area of agricultural land maintained in good agricultural culture amounts to 0 (next 790 records).

Summarizing, 1397 records (i.e. 37.7%) were defective and had to be removed from further analysis. Moreover, the 'master record' contains no data on neither sown area nor structure of the **basic** crops. One can find there only information on some "peripheral" crops, i.e. flax, hemp and hop. Therefore, we cannot also indicate farms where the agricultural production exceeds the thresholds adopted within EU. Moreover, the Agency for Restructuring and Modernization of Agriculture does not register the farms for which agricultural land area is smaller than 1 ha. These problems (taking into account also the fact that the Agency gathers no data on the total agricultural land, but only on the land maintained in the good agricultural culture) are the main difficulties in harmonization of the analyzed registers.

Due to lack of data on crops, a direct computation of the quantity *grc* necessary to determine the classification variables *Crop* and *Animal* is impossible. Therefore, estimation is needed. We have done it using the relevant data gathered during the Farm Structure Survey in 2007. That is, we have constructed a linear regression model with *grc* as explained variable and arable land area (*aland*) as explanatory variable. The regression function is of the form:

$$grc = 0.98455 \cdot aland - 0.03105. \qquad (10)$$

The value of the Student's t–test for intercept amounts to -3.02 (p=0.0025) and for the slope 5446.25 (p<0.0001). The analysis of variance and adjustment is presented in Table 2.

Table 2. Regression of *grc* according to arable land area – analysis of variance and assessment of adjustment

Analysis of variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	502043681	502043681	2.966E7	<.0001
Error	185159	3133940	16.9257		
Corrected Total	185160	505177621			

Adjustment of the model

Root MSE	4.1141	R-Square	0.9938
Dependent Mean	20.6096	Adj R-Sq	0.9938
Coeff Var	19.9620		

Source: Authors' elaboration using the SAS Enterprise Guide 4.2 environment.

On the basis of these results we can conclude that this model is well established and may be an effective tool for estimation. We have used it. If an estimate of *grc* obtained on the basis of regression function was negative (it is sometimes possible for very small farms), we have assumed it to be zero. The function (10) was finally used to estimate the variables *Crop* and *Animals*.

Now, we present the classification obtained using the "master record" set. The Table 3 contains specification of number of farms belonging to each class and average value of particular classification variables within it. To avoid some misclassification which is undesirable from the practical point of view, during the maximization of the probabilities (4) (with the distance of object from classes computed using (2)) , we have preferred those elements which are more strictly desired from the practical point of view. That is, we have minimized (4) only within the farms of the same size type (in terms of *Land* variable) as the classified object. That is, we have looked for such optimum class, for which the land area of the given farm belong to the respective land area interval describing this class. If necessary, this additional criterion was extended also to the variables *Crop* or *Animal*.

The quality of received clustering was assessed using three indices. The coefficient of homogeneity of clusters is given as

$$hm = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{n_r}\sum_{\substack{i=\{1,2,\ldots,n\}\\ \gamma_i \in \Omega_r}} d_e(\gamma_i, \bar{\gamma}_r),$$

and the coefficient of their heterogeneity, i.e. mutual separation level (assuming that *k*>1):

$$ht = \frac{1}{k(k-1)} \sum_{r=1}^{k} \sum_{\substack{s=1 \\ s \neq r}}^{k} d_e(\bar{\gamma}_r, \bar{\gamma}_s),$$

where $\bar{\gamma}_r$ is the centroid of the class Ω_r, i.e. the vector, which coordinates are arithmetic means of observations of respective variables for objects belonging to this class, r = 1, 2, …, k, and $d_e(\cdot,\cdot)$ denotes the Euclidean distance. The coefficient of correctness of clusters is a ratio of these two quantities (i.e. it equals to *hm*/*ht*). The closer to zero it is, the better the quality of clustering is.

Table 3. Classification of farms using the 'master record' data

Class	Number of farms in the class	Land	Crop	Animal	Culture	Meadows
C1A	911	16.2812	84.9414	15.0586	71.9838	20.1087
C1B	327	17.5099	33.2624	66.7376	87.6120	28.7586
C1C	17	6.1618	0	0	84.5360	0
C1D	31	9.4671	0	0	0	0
C2A	156	1.4583	98.9344	1.0656	45.2971	12.2058
C2B	13	1.5741	32.2028	67.7972	80.6275	32.5879
C2C	4	1.4850	0	0	90.0211	0
C2D	8	1.5625	0	0	0	0
C3A	190	0.4109	99.6811	0.3190	0	0
C3B	35	0.0479	1.5527	98.4473	0	0
C3C	0	0	0	0	0	0
C3D	438	0.0099	0	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

The coefficient of homogeneity amounts to 16.014, the coefficient of heterogeneity of clusters equals to 91.650 and hence the coefficient of correctness amounts to 0.1747. It is a very satisfactory result and therefore the division can be perceived as effective.

Using the procedure described in paragraph 4 we have determined the limits of classes in an econometrically optimum division. We have then obtained 11 non–trivial classes, which are uniquely described by the following intervals (for details, see Table 4).

Table 4. Optimum classes for the "master record"

Variable	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Land	2	2	2.020	1.030	0.490	1.100
	135.760	27.270	46.810	1.995	2.390	1.740
Crop	50.191	0	0	51.916	0	0
	100	0	0	100	0	0
Animal	0	0	0	0	0	0
	49.810	0	0	48.084	0	0
Culture	50.969	23.579	0	56.853	0	84.118
	100	100	0	100	85.057	100
Meadows	10	0	0	12.817	0	0
	39.877	0	0	39.288	0	0

Variable	Class 7	Class 8	Class 9	Class 10	Class 11
Land	1.010	1.320	0	0	0
	1.860	1.530	1.290	1.750	0.990
Crop	0	33.887	0	0	0
	0	35.401	0	72.930	0
Animal	0	64.599	0	0	0
	0	66.113	0	58.569	0
Culture	0	84.967	0	0	0
	0	86.364	0	0	0
Meadows	0	53.788	0	0	0
	0	57.516	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

These classes reflect better the actual structure within the analyzed database and seem to be rather easy to interpret. They are also similar to those presented in the Table 1. The one slightly more significant difference between them is that no optimum class is described by values of the *Animal* being above 67% and *Crop* below 33% (except for zero). Also values of the *Meadows* belonging to the interval (0, 10) were also omitted. The main probable reason of this phenomenon could be a fact that for most of farms with prevalence of animal production its domination over the crop production is not especially significant. On the other hand, the distance of the variable *Animal* from the interval (67,100] can be smaller than, e.g. a distance of *Meadows* from the thin interval [0,0]. For example, if for some farm the values of the classification variables are as follows: *Land*=1.50, *Crop*=25%, *Animal*=75%, *Culture*=85% and *Meadows* = 20%, then it will be classified to the class 8. Such situations affect the final results. The nature of such behavior can be explained also taking into consideration that there are relatively small numbers of farms with prevalence of animal production (what is evident for example for C2B and C2C classes), what – with some similarities for variables other than crop and animal in C1 and C2 classes – may cause that such farms were omitted in final classification. An advantage of these classes is the clear presentation of farms of various size and type which have conducted no agricultural production (classes 2, 3, 5, 6, 9, 11). It is confirmed by the comparative classification done using the new classes –1, 4, 8 or 10 with the average values of *Crop* and *Animal* amounting to, respectively, 86.88% and 15.12%, 99.14% and 0.86%, 34.64% and 65.36% and, finally, 22.61% and 17.39%.

For a better comparison, we will present now results of classification using the data collected "from nature", i.e. by direct interviewing the farmers (the "gold record" file). These data are, of course, much more detailed and therefore some classification variables computed using them are of higher quality than those determined using the "master record" database. The earlier established collection of criterion intervals (Table 1) remains without any change. The classification is presented in Table 5.

Table 5. Classification of farms on the basis of the 'gold record' (restricted to farms considered in Table 3.)

Class	Number of farms in the class	Land	Crop	Animal	Culture	Meadows
C1A	1021	12.6959	78.1635	21.8365	97.1076	27.1654
C1B	281	14.9352	40.4804	59.5196	97.9527	29.9669
C1C	19	11.8884	0	0	98.6842	0
C1D	25	4.1644	0	0	0	0
C2A	208	1.4361	87.1073	12.8927	96.7985	27.4768
C2B	30	1.4993	30.3071	69.6929	91.9563	21.9160
C2C	16	1.5250	0	0	94.9925	0
C2D	24	1.4688	0	0	0	0
C3A	138	0.4288	92.7006	7.2994	96.6787	13.5861
C3B	55	0.3685	24.8366	75.1634	83.2707	10.7372
C3C	12	0.5250	0	0	100	0
C3D	225	0.0547	0	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

The coefficient of homogeneity amounts to 20.4302, the coefficient of heterogeneity of clusters equals to 81.5134 and hence the coefficient of correctness amounts to 0.2417. It is a very good result. Comparing this structure of classification with the result obtained on the basis of the 'master record' and presented in Table 3, using the three most popular tests for location (i.e. for the hypothesis that the expected value of the distance between them equals zero), we can observe that they are consistent – Student's t statistics amounts to 0. 283698 (p=0.7819), sign test statistics equals to -2 (p=0.3877) and Wilcoxon signed rank statistics is also negative: -7.5 (p=0.5801). The calculations were conducted by UNIVARIATE SAS procedure using the difference between number of farms from Table 3 and Table 5. More details about this procedure can be found in Base SAS 9.2 Procedures Guide (2010), pp. 332–334. This consistency may be sometimes, however, not especially strong due to a fact that some data used to compute *grc* and *pzw* were much more detailed in 'gold record' (e.g. the additional categories of cattle for which special coefficients to calculate them per livestock units are used, were here presented).

Now, we present the limits of classes in an econometrically optimum division established using the procedure described in paragraph 4 and all records included in the 'gold record'. This way, we obtain 11 non–trivial classes, which are uniquely described by the following intervals (see Table 6).

Table 6. Optimum classes for the "gold record"

Variable	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Land	2	2.070	2	1	0.590	1
	922.660	141.230	11.080	1.990	2.490	1.930
Crop	50.064	0	0	50.171	0	0
	100	0	0	100	0	0
Animal	0	0	0	0	0	0
	49.936	0	0	49.829	0	0
Culture	50.523	35	0	60	0	50
	100	100	0	100	75	100
Meadows	10.062	0	0	11.333	0	0
	39.933	0	0	38.418	0	0

Variable	Class 7	Class 8	Class 9	Class 10	Class 11
Land	1	0.420	0	0.120	0
	1.980	0.990	1.380	0.990	0.960
Crop	0	51.123	0	0	0
	0	100	0	0	0
Animal	0	0	0	0	0
	0	48.877	0	0	0
Culture	0	50	0	50	0
	0	100	75	100	0
Meadows	0	10.101	0	0	0
	0	34.884	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

These classes are much more consistent with the structure of the analyzed data set than the arbitrarily fixed norms expressed in Table 1. The optimum classes are rather easy to interpret and correspond to the classes presented in Table 1. The only significant difference between both structures is that no class is described by the interval (50, 100] for the variable *Animal* and the interval (0,50] for *Crop*. The reason of this phenomenon is similar as in the case of the optimum classes for 'master record' but the situation observed here is slightly more difficult. The farms which have conducted no agricultural production are also well presented (classes 1, 3, 5, 6, 9, 10, 11). It is confirmed by the comparative classification done using the new classes – the farms conducting the agricultural production were classified only to the classes 1, 4 or 8, with the average values of *Crop* and *Animal* amounting to, respectively, 76.91% and 23.08%, 86.06% and 13.91% and, finally, 89.72% and 10.28%.

6. Conclusions

We have proposed an original method of classification of objects taking various characters of the used criterions into account. They may have less or more strict form, it means that they may be less or more fuzzy. Moreover, the projected groups of objects may be desired to have some arbitrarily fixed properties, resulting from commonly (of which legally) adopted norms. On the other hand, they should be, of course, optimum. Our proposal is a trial to satisfy all these expectations and to show how large is the distance between assumptions established "in advance" and obtained endogenically, only on the basis of the internal properties of used data basis.

Of course, the empirically examined collection of agricultural variables is here relatively small. It is a consequence of small scope of information contained in the "master record" file, which may serve as a classification features. In practice, it is possible to obtain much broader set containing many data that describe the character of a farm using "typical" intervals specific for it (e.g. number of poultry or ostrich units, area of fish ponds, area of mushrooms under cover, etc.). Also, sometimes economists expected to include in the analysis also variables characterizing the economic aspects of the farm activity, such as the standard gross margin (reflecting the value of marketable output), economical size (expressed in European Size Units), commodity output or employment in the farm, fulfillment of some production thresholds established by the EU regulations, etc. The proposed methods theoretically enable one to effectively involve all these postulates to the classification and also asses the internal structure of the data basis being at researcher's disposal. However, these economical variables were not available in our database and therefore they could not be used here. Our method gives the opportunity to introduce it to the model if it will be necessary in the future. It solves also the most difficult problem of usage of interval or ratio variables to the classification. In comparison with other fuzzy clustering methods this one is much more useful from the practical point of view, where classes are often defined using the reference intervals for particular characteristics. It is also more effective in context of the computational capacity.

The only inconvenience connected with this approach seems to be the necessity to establish some additional preferences during maximization of probability of appurtenance of object to particular classes. Despite using the "maximum" formula of distance of an object from a given class, the formula for probability measure (see equation (4)) does not exclude a possibility of compensation of discrepancy in respect to some criterion by a similarity connected with other criterion. The strong practical requirements enforce application of such correction.

However, in general, the proposed method can be assessed as useful in realization of important methodological tasks, such as preparatory works for the national censuses. This task may be realized in practice in any exercise of such type in the following way. Using the typology constructed by means of the proposed method, the basic area and profile groups are determined. On the basis of this division a survey methodology can be established. That is, it could be decided which groups of farms should be investigated by exhaustive survey and which by a sample survey (and in this case it can contribute to find an effective sample size). This enables one to rationalize the costs of statistical undertakings and optimize the quality of their results. The more diversified the used data set is, the more effective the final effects of its application should be. The obtained classification can be further developed by selecting in each class some subclasses by adding more nominal criteria, what is much easier than in data collection analyzed above and each user should be rather able to do it.

It could be also a good basis for a wider discussion on principles and efficiency of such classification method as well as on methods of its possible improvement. Of course, a critical view of our results is fully justifiable. The critics of it may recall in this context the argument that the economical quantities such as structure of agricultural output could be here better variables and the final results obtained using them may be different than the current product (e.g. taking into account that according to the Statistical Yearbook of Agriculture 2009 (GUS (2009)), the nationwide ratio of crop output to animal output is about 56:44 for gross output and 45:55 for market output, whereas in our case the respective relation between these two main types of farms was much more clear. One should remember, however, about two main features of our approach. Firstly, we were not able to use the strictly economical variables because they were not available in administrative sources used for the census. Secondly, due to the above reason we had to analyze the **physical** structure of production which could be significantly different from their **economical**, monetary value. Of course, if we had more information on the economical aspects at our disposal, the quality of the classification would be better.

REFERENCES

BEN – ISRAEL A., IYIGUN C. (2008) *Probabilistic d–Clustering*, Journal of Classification, vol. 25, pp. 5–26.

BEZDEK J, C. (1973) *Fuzzy Mathematics in Pattern Classification*, PhD Dissertation, Cornell University, Ithaca, New York.

Commission Regulation (EC) *No 1242/2008 of 8 December 2008 establishing a Community typology for agricultural holdings,* OJ L 335, 13.12.2008, pp. 3–24 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:335:0003:0024:EN:PDF

EVERITT, B. S., LANDAU, S., & LEESE, M. (2001) *Cluster analysis* (4th ed.). London: Arnold.

GATH I., GEVA, A. B. (1989) *Unsupervised optimal fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 11(7), pp 773–781.

GUS(2009) *Statistical Yearbook of Agriculture*, Central Statistical Office of Poland (GUS), Warszawa.

GUSTAFFSON D. E,, KESSEL W. C. (1979) *Fuzzy Clustering with a Fuzzy Covariance Matrix, Clustering with a Fuzzy Covariance matrix*, Proceedings of the IEEE Conference Decision Contribution, San Diego, CA, USA, p. 761–766.

HATHAWAY R. J., BEZDEK J. C. (1988) *Recent Convergence Results for the Fuzzy c-Means Clustering Algorithm*, Journal of Classification, vol. 5, p. 237–247.

IYIGUN C. (2007) *Probabilistic Distance Clustering*, A dissertation submitted to the Graduate School – New Brunswick Rutgers, The State University of New Jersey in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Graduate Program in Operations Research. Written under the direction of Professor Adi Ben–Israel, New Brunswick, New Jersey, November, 2007, http://www.benisrael.net/Iyigun-Thesis-Nov-07.pdf.

LIPIŃSKA H., GAJDA J. (2006) *Area of farms versus fodder base and cattle population in specialized dairy farms*, Annales Universitatis Mariae Curie-Skłodowska Lublin – Polonia, vol. LXI, Sectio E, pp. 225–236 (in Polish).

Regulation (EC) No 1166/2008 of the European Parliament and of the Council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation (EEC) No 571/88, OJ L 321, 1.12.2008, p. 14–34 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:321:0014:0034:EN:PDF.

SAS Institute Inc. (2010) *Base SAS® 9.2, Procedures Guide: Statistical Procedures*, Third Edition. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/procstat/63104/PDF/default/procstat.pdf.

TONINI A. (2007) *Agriculture and Dairy in Eastern Europe after Transition focused on Poland and Hungary*, PhD Thesis, Wageningen University, The Netherlands, http://library.wur.nl/wda/dissertations/dis4133.pdf.

Tonini A, Jongeneel R. (2007) *The Distribution of Dairy Farm Size in Poland: a Markov Approach Based on Information Theory*. Applied Economics, vol. 1, pp.1–15.

WEISZFELD E. (1937) Sur le point pour lequel les sommes des distances de n points donné et minimum, Tahoku Mathematical Journal, vol. 34, pp. 355–386.

YIH J. – M., HUANGH S. – F. (2010) *Unsupervised Clustering Algorithm Based on Normalized Mahalanobis Distance*, [in:] S. Chen and H. Wu (eds.) Proceedings of the 9th WSEAS Int. Conference on Applied Computer and Applied Computational Science, Electrical and Computed Engineering Series. A Series of Reference Books and Textbooks, WSEAS Press.