Polska Akademia Nauk • Instytut Badań Systemowych

AUTOMATYKA STEROWANIE ZARZĄDZANIE

Książka jubileuszowa
z okazji
70-lecia urodzin

PROFESORA KAZIMIERZA MAŃCZAKA

pod redakcją
Jakuba Gutenbauma

Polska Akademia Nauk • Instytut Badań Systemowych

AUTOMATYKA STEROWANIE ZARZĄDZANIE

Książka jubileuszowa
z okazji
70-lecia urodzin

PROFESORA KAZIMIERZA MAŃCZAKA

pod redakcją
Jakuba Gutenbauma

Warszawa 2002

Książka jubileuszowa z okazji
70-lecia urodzin
Profesora Kazimierza MAŃCZAKA

Redaktor
prof. dr hab. inż. Jakub Gutenbaum

Warszawa 2002

ISBN 83-85847-78-2

Wydawca: Instytut Badań Systemowych PAN
ul. Newelska 6 01-447 Warszawa
http://www.ibspan.waw.pl

Opracowanie składopisu: Anna Gostyńska, Jadwiga Hartman

Druk: KOMO-GRAF, Warszawa
nakład 200 egz., 34 ark. wyd., 31 ark. druk.

CLUSTER-WISE MODEL IDENTIFICATION: WHAT HAPPENED DURING THE LAST TWENTY YEARS?

Jan W. Owsiński
Systems Research Institute, Polish Academy of Sciences

Abstract: *The paper deals with one of the toughest problems in data analysis: given a set of observations we suppose they were generated by a set of different processes, to which different models correspond. We must, therefore, simultaneously split the set of observations into subsets corresponding to different models, and identify these models. In this formulation the problem has not found any satisfactory solution to date, except for the very special cases (single dimension or "brute force" applicability). The paper presents the questions encountered, some formulations and approaches used, and a positive proposal for at least one of the aspects of the overall problem, related to the objective function, which displays the feature of globality, that is – implication of a globally optimal solution to the problem. For the sake of shortness the respective problems and issues will just be signalled, with the detailed considerations left to more technical publications.*

Keywords: *model identification, cluster analysis, global optimum.*

1. Introduction: prerequisites

Assume we dispose of n observations x_i, $i \in I = \{1,...,n\}$, each one composed of $m+1$ values x_{ij}, corresponding to the same number of variables X_j, $j \in J = \{1,...,m+1\}$, serving to describe the observations.

We are asking for a model of the process generating these observations, in a static form. Without any loss to generality we will denote, rather traditionally, the single output variable from the model by Y, $Y = X_{m+1}$. Thus, in a classical case we will be looking at the model appearing through the form $Y = f(X_1,...,X_m) + \varepsilon_i$, this model being commonly identified – with respect to the form of $f(.)$ and the properties of ε – via a number of well-

known procedures, primarily of regression character. The variables $X_1,...,X_m$ span the space denoted E_X.

Yet, in this case we suspect that the observations are in fact produced not by a single model (process), but rather by a set of such – even if somewhat "similar" – models or processes. Let us denote the (unknown) number of such models p, these models being indexed by q, $q \in Q = \{1,...,p\}$. We know nothing, *a priori*, of these models, other than their assumed general form quoted above. This means, in particular, that we do not know what is the breakdown (partition) of the set I into subsets generated by and corresponding to the individual models q (the subsets being denoted A^q), nor what is the proper value of p (the number of models). We will initially only assume that $1 \leq p < n$, although, as it can be easily imagined, and demonstrated for definite cases, the assumption of $1 \leq p < m$ may also be in place, like in the standard regression models.

Thus, we look for models $f^q(.)$ – omitting the questions related to ε, at least for a while – along with their number and the subsets of observations, associated with them. Like in the single-model case, we will be trying to minimise a function of $\{\varepsilon_i\}$, assuming that $y_i = f^{q(i)}(x_{i1},...,x_{im}) + \varepsilon_i$, where $q(i)$ assigns a model q to an observation i according to the breakdown of I. Note that writing this equation defining y_i we assumed the breakdown (and the assignment $i \rightarrow q$) to be done, and so the error term ε not depending any longer upon the model choice (in particular, Mańczak 1979).

We will denote the breakdown, or partition, of I by P, $P = \{A^1,...,A^p\}$, with $\cup_q A^q = I$. The potential further properties of P, or implied by them, will be discussed in the paper later on.

In addition, we will assume that some definitions exist of distances and/or proximities in E_X, denoted, respectively, $d(.,.)$ and $s(.,.)$, and that these definitions can be appropriately extended to sets and to geometric structures in E_X. They will be denoted, when applied to sets, D and S, respectively. We will require of these only that they be positive, and in case distance and proximity are simultaneously defined, that the two display an opposite monotonicity, i.e. $d(x_1,x_2) \geq d(x_3,x_4) \Leftrightarrow s(x_1,x_2) \leq s(x_3,x_4) \quad \forall\, x_i \in E_X$, and likewise for D and S.

2. The problems

This very general formulation, first: leaves a lot of space void in terms of details, and, second: implies a number of different, often alternative formulations and potential solution forms. In order to proceed positively one has to fill the voids and make selection among the possible formulations.

There is quite a number of aspects, which intervene in the more detailed problem formulation and (the on-going) solution search, and it is, in fact, hardly possible to present them consistently within the frames of a short paper. Thus, we will just stop at some milestones along a winding road, and make comments related to them. A kind of a guide for such a road, though at a definitely earlier stage of respective developments, was offered in Owsiński (1989).

Let us first state that the problem outlined is in fact equivalent to the general one of cluster analysis ("*finding the subsets of observations that be internally possibly similar, while being possibly dissimilar between them*"). Thus, since the very problem of cluster analysis has hardly found a satisfactory solution, in theoretical as well as in algorithmic terms, no wonder the cluster-wise modelling problem has not. In fact, most of the issues we will be citing here apply in a very similar manner to the general clustering problem.

2.1. The nature of the model

We are looking for a genuine model, which can be used for forecasting, prediction, or design. Thus, $f(.)$ corresponds to a well-defined function (say, a linear regression function), assigning the values of y to those of $(x_{i1},...,x_{im})$. We are therefore not dealing with, for instance, the classical probability density function mixture problems, where primarily the density function parameters or other distribution characteristics were looked for (Bock 1996), for a very constructive overview of the relations between clustering and mixture-type or other probabilistic models). There are some more recent approaches involving mixture formulations that offer new vistas, also in cluster-wise modelling problem, and one of them will be commented upon further on. Nor are we interested – in the first place – in the problems of testing and validation of the otherwise identified clusterwise models (see, in particular, a recent article by Hennig (2000), treating the question of identifiability of such models). That is – we are looking for a constructive approach in determination of clusters and, simultaneously, models.

In particular, a discriminant function may, of course, play also a role of a model, which then ultimately just assigns a point to a class. In the extreme case, it is therefore admissible within this formulation to include sheer classification ("typology"), treating definite values of y, namely y^q, being the "types" of the clusters q, as the models.

2.2. Uniqueness of the overall model

Since the set I is broken down into internally coherent subsets, and little additional assumptions are made, it is possible that the subsets A_q are determined so that the corresponding f^q have overlapping argument domains. These domains may be defined as, for instance, the convex hulls of the $\{x_i\}$, $i \in A^q$, denoted $H(\{x_i\}^q)$. If the mappings $Y(H(\{x_i\}^q))$ for various q are overlapping, it may happen that more than one value of y correspond to an x_i or to some $x \in E_X$. Making of narrowing and thus simplifying assumptions concerning this issue may be inappropriate. It is customary to introduce an additional variable ($x(q)$) for the purpose of distinguishing the y's generated by different (overlapping) models (e.g., Nakamori and Ryoke 1994). The variables charged with such an assignment rarely, if ever, appear as natural phenomena, and certainly cannot be subject to the proper identification procedure.

2.3. The number of clusters and the monotonicity

For the sake of generality it is assumed here that the number of models – clusters – denoted p is not predefined. This is an essential aspect of the formulation, since none of the classical cluster analysis methods provides a convincing solution with this respect.

Thus, of all the clustering techniques the K-means-type ones are based upon a class of objective functions, which, as reflecting the sum of intra-cluster distances – or intra-cluster distances from the cluster-wise model – are inherently monotone in p. Thus, they get, generally, "better", as p increases, since ultimately (whenever applicable) *an object is a perfect model of itself (no error)*. Hence, additional criteria are required in order to select the "proper" value of p, when applying such approaches.

The agglomerative or divisive schemes do not omit this problem, neither. Since they provide a hierarchy of partitions, decision must be made as to the level of this hierarchy, which is to be retained as solution.

The "constructive" single-cluster-defining approaches (e.g., Mirkin 1996, for an excellent presentation) can lead to determination of certain A^q's as corresponding to "proper" clusters, but the overall set of anyhow thus determined subsets will in general not satisfy the condition $\cup_q A^q = I$. In order to satisfy it, either the "constructive" definition has to be broken, or additional "classification" performed, assigning the left-out i's to the already defined A^q's.

In any case, in a general formulation, an "external" criterion has to be applied, inherently alien to the original procedure, in order to determine the "proper" p.

2.4. Numerical issues

Side by side with the above issues, which are in practical terms translated into a lot of algorithmic, but also theoretical, details, there are also quite fundamental numerical issues in the solving of the problem considered. The multimodality, combinatorial character, NP-hardness, curse-of-dimensionality, etc., in addition to the definitional questions already alluded to, make out of it quite a playground for a multitude of often poorly justified heuristics. Let us just mention here the necessity of using multiple starting points (even up to the order of 10^6), not only in case of any of the K-means-type algorithms, but also many other methods, with very poor – if any – estimates on error bounds. Many of the mathematical programming tasks, which are formulated in this framework, are being solved through approximations, and/or with assumptions, which accordingly simplify the initially assumed model of the problem solved.

2.5. The nature of clusters

One of the approaches, which is used in both formulation and solution of the cluster-wise modelling problem is connected with the introduction of fuzziness. First of all, clusters A^q can be defined as fuzzy sets, that is, each x_i is assigned a number $\mu^q(x_i) \in [0,1]$, the membership coefficient, which corresponds to the degree, to which ith observation belongs to cluster q. Fuzziness is therefore naturally extended to the model $f^q(.)$ in that it will assign values to y on the basis of the respective membership functions. In addition, though, the very model can also take on a fuzzy form, which we shall not comment upon here. The membership coefficients can be required to satisfy the condition $\cup_q A_q = I$. On the other hand, this condition can also be satisfied by the overlapping, "crisp" clusters A^q, this situation giving rise

to a similar situation with respect to the determination of y as in the case of fuzzy clusters.

2.6. Models and procedures – the examples

A good illustration for the actually applied procedures is provided by the papers of Lau, Leung, Tse (1999), and Nakamori and Ryoke (1994). We will cite here the basic assumptions and the simplifications made in the two approaches. They differ considerably in terms of both concrete formulations of the problem and the solution methods applied thereto. One of them relates to the mixture formulation with the maximum likelihood function and so a mathematical programming framework (Lau, Leung, Tse 1999), with the cluster-wise linear regression of Späth (1979) as the original source of inspiration, and the papers by Celeux and Govaert (1993), as well as DeSarbo, Oliver and Rangaswamy (1989) as the essential points of reference. The other one is related to a fuzzy-set formulation regarding clusters and models, with an ellipsoidal model form, allowing for an eigenvalue-based sub-optimisation procedure (Nakamori and Ryoke 1994). This line of proceeding originates with the early papers on fuzzy clustering, like, first of all, Dunn (1974), and fuzzy linear regression – Jajuga (1986). Yet, the limitations of both, quite altogether complex procedures, are in many points similar:

(a) number of clusters: in both cases the number of clusters is largely assumed a priori (in the second case it can decrease from an initial number based on an external criterion); this is closely related to;

(b) the monotonicity of the (implied or explicit) objective function, which in both cases can be likened to the sum over clusters of the sums of errors with respect to the models sought and determined, that is – the more models, the smaller the sum of errors (down to the limit of capacity of determination of a model, i.e. the minimum cardinality of clusters);

(c) the starting points – some special procedures are applied in the two cases for generation of the proper starting points (random generation and Ward clustering), in view of both multimodality of the respective problems and the necessity of having a minimum cardinality of the initial clusters for determination of models;

(d) the dimensions of the problems treated: it is characteristic that in the two cases the dimensions, in terms of n, m, and p, of the examples shown, are quite small (n in dozens, m – a couple, p – similarly); this is,

in fact, the illustration of the limited numerical capacities of the methods and the technical algorithms involved;

(e) the optimisation algorithms: these are different (a variant of the E-M algorithm, eigenvalue-based assessment of the "matching" of observations and clusters, exchange algorithm, etc.); even, though, for the subproblems of the overall problem these algorithms do not provide a guarantee of obtaining a unique optimal solution; let us note at this point that fuzziness is often introduced into the clustering problems in order to secure a facility of computations (continuous, differentiable subproblems in place of the hard-to-treat combinatorial ones).

3. The question of the objective function and the algorithm

It appears that the formulation of the *objective function* is one of two essential issues in the formulation and solution of the problem here considered, side by side with the respective *optimisation algorithm*. Yet, it is obvious that the two are very closely related. The present paper focuses in its second part on the formulation of the objective function that would help in resolving the limitations related to monotonicity and the pre-defined number of clusters, but also provide a form that lends itself to a more effective and efficient optimisation. Thus, the objective function we look for should: (I) avoid monotonicity with respect to the number of clusters (and thereby provide the capacity of comparing essentially different partitions P and the corresponding models); (II) accommodate a possibly flexible formulation of the details of the problem (e.g. the distance/proximity definitions); (III) allow for a facile optimisation or at least sub-optimisation through either general or special procedures.

Now, let us introduce some notions and observations related to both the objective function formulations and the prerequisites for the design of algorithms.

Assume $D(f,A)$ assigns a real non-negative value to a model f and the set of observations indices A. Thus, $D(f^q,A^q)$ may denote the sum of error term for the model proper for the cluster q. We are definitely looking for the (exhaustive) partition P, incorporating the set of models $\{f^q\}_q$, for which the function $\Sigma_q\ D(f^q,A^q)$ attains minimum, like in virtually all the approaches used. Yet, it is exactly this formulation that entails the problem of monotonicity and of the determination of the number of clusters. The situation is, of course, the same, for the "dual" problem of max $\Sigma_q\ S(f^q,A^q)$, where cluster-wise similarities of the clusters and their models are summed.

It is quite natural to assume that a model f is uniquely determined (at least down to the precision of numerical algorithms) for the respective set of observations A. This is, for instance, the case of the classical LS and their non-orthodox variants. So, given the appropriate definitions, we can use the simplified notations $D(A^q)$ and $S(A^q)$, or even $D(q)$ and $S(q)$.

Definitely, most of the approaches refer to $D(A^q)$ or $S(A^q)$, or rather their respective sums, $\Sigma_q\ D(A^q)$ or $\Sigma_q\ S(A^q)$. Therefore, in view of monotonicity of such objective functions, the value of p is either defined a priori, or calculations are performed for a series of values of p, with an external criterion applied in order to determine the "best" value of p. In such a situation the search for a globally optimal P is replaced by the local search algorithms, with two techniques most widely applied, oftentimes in conjunction.

One is the "object exchange" technique, in which all of the individual objects $i \in I$ can be exchanged between the clusters A^q, if this leads to an improvement in the objective function $\Sigma_q\ D(A^q)$ or $\Sigma_q\ S(A^q)$. For this purpose, an increment function is used, $\Delta_i(q,q')$, reflecting the difference of value of the objective function resulting from moving of object i from cluster q to cluster q'. In many cases it can be shown that $\Delta_i(q,q')$ is a straightforward function of $D(i,q)$ and $D(i,q')$, or simply $D(i,q) - D(i,q')$ (and likewise for $S(.,.)$). Full iterations, in which entire I is successively analysed, are repeated until the change in the objective function gets small enough, or the P gets repeated (cycling), or a predefined number of full iterations have been performed. For several standard objective functions convergence to local extrema was proven.

The seminal algorithm of Späth (1979), from which a part of the title of this paper is derived, and the follow-up algorithmic varieties, used the "object exchange" technique.

The other one is the "centre-and-reallocate" technique, most popular in the K-means variety of the usual cluster analysis problem. Here, given that at the start the partition P is given, defined solely by the clusters A^q, first the cluster-proper models f^q are – locally – determined for these clusters, then, objects i are assigned in – again – a locally optimal manner to models f^q, forming new clusters, A^{*q}, and thus a new partition, P'. This is a complete iteration, after which new models, f^{*q}, will be determined.

Note that we have postulated very little of the models, distances-proximities, etc. Thus, we may deal with the least squares formulations, the

fuzzy-set theoretical settings, or even the likelihood functions resulting from mixture models. The general outline of the situation remains the same.

Given the sole possibility of performing local search, the entire procedure usually takes the following form: (i) some special algorithm is used to define the starting point for one of the above techniques, since a certain minimum cardinality of A^q is required in order to determine models (like in LS regression, for instance); these algorithms include Monte Carlo generation of initial clusters, classical progressive merger procedures (such as, for instance, the Ward technique in Nakamori and Ryoke 1994), and a number of custom-made procedures (like space-dividing ones); (ii) proceed with a local search technique for one or more of the pre-selected values of p; (iii) check with an external criterion (based, say, on $D(A^q,A^{q'})$) whether some clusters could not be merged, and, possibly, after the meger would have been performed, return to (ii); (iv) on the basis of (another) external criterion the local solution is retained, which has some special properties (e.g. the biggest drop in the otherwise monotone decreasing objective function for a given p). This, indeed, does not seem to be an internally consistent procedure.

Thus, even within the framework outline above, many of the methods and procedures applied have to somehow deal with the inter-cluster similarity or distance (e.g., Nakamori and Ryoke 1994, for the portion of the procedure where clusters are merged). Here, we can deal with distances determined through models: $D(f^{q'},f^{q''})$, through models and observations in different clusters: $D(f^{q'},A^{q''})$, or through observations in different clusters alone: $D(A^{q'},A^{q''})$. For simplicity we do not distinguish between these distance definitions, and generally denote them $D(q',q'')$. We wish to maximise $\Sigma_{q'}\Sigma_{q''}D(q',q'')$, since allowing for a small value of this "inter-cluster differentiation" measure might lead to indistinguishable clusters and models. Analogous definitions can be introduced for the similarities $S(.)$.

4. A general global objective function

We will now proceed to presentation of the principles of construction of the objective function that at least partly responds to challenges forwarded in the preceding section. Indeed, we have defined in Section 3 two elements of the general global objective function that we propose, in particular, for the cluster-wise model identification problem:

$$C^*(P) = C_D(P) + C^S(P) \rightarrow \min,$$

where $C_D(P) = \Sigma_q D(q)$, and $C^S(P) = \Sigma_{q'}\Sigma_{q''} S(q',q'')$, or

$$C^*(P) = C_S(P) + C^D(P) \rightarrow \max,$$

where $C_S(P) = \Sigma_q S(q)$, and $C^D(P) = \Sigma_{q'}\Sigma_{q''} D(q',q'')$.

It is natural that the elements of $C^*(P)$ and $C_*(P)$ (may) have opposite monotonicity in p (e.g. $C_D(P)$ decreasing with p, while $C^S(P)$ increasing with p). We will assume only here that this opposite monotonicity is of similar character as in case of $d(.,.)$ and $s(.,.)$ in Section 1.

By referring to either $C^*(P)$ or $C_*(P)$ we can avoid monotonicity and by solving the thus formulated minimisation or maximisation problem (if we are able to) obtain in a natural manner the number of clusters along with their composition and respective cluster-proper models.

Let us comment yet on two issues, which are related to the general formulation proposed.

First, it is obvious that it will quite often be so that the functions D and S will be closely related and one would be simply derived from the other. Yet, both for the sake of clarity of presentation (the distinct "two-sidedness" of the objective function), and in view of the fact that in many instances these two functions actually stem from different formulations (like $D(q)$ being the LS sum of errors, and $S(q',q'')$ representing correlations between model parameters), we insist on the distinction of the two elements. We will also see that this has a counterpart in the proposed algorithmic solutions.

Second, there is, obviously, quite a variety of feasible concrete formulations of the functions involved, satisfying the "opposite monotonicity" requirement, so that a high degree of flexibility is offered within the approach proposed. Within such a broad domain we may, in particular, deal with cases, where the overall objective function will be very close to monotonicity, implying global solutions with p close to n or to 1. This borders upon the frequent issue of explicit weights. Although, as we will see in the following section, a weight mechanism is being introduced into the general formulation, its purpose is not to influence the shape of the ultimate solution. The present author leaves the question of weights, whether implicit or explicit, to the discretion of the designer of the particular analytical exercise.

Let us emphasise at this point that the merits of the objective function proposed here are not merely related to the possibility of avoiding

monotonicity with respect to p (postulate I from the beginning of Section 3). There are, namely, several formulations of the objective function for the classical clustering problem, some of which can be transformed for the needs of the cluster-wise identification problem. Owsiński (1991, pp. 74-79), provides an overview of such formulations (see also Marcotorchino 1985, for a more formal, but narrower treatment of a similar problem). Beyond this somewhat dated review we can yet mention the functions proposed by Stanfel (1992), based on the information-theoretical considerations, or Fraley and Raftery (1998), following Schwarz (1978), the "Bayesian Information Criterion" applied to the mixture model-maximum likelihood formulation, similar to that of Lau, Leung, Tse (1999). Another known form that avoids monotonicity with respect to p is the pseudo-F-statistic.

Yet, in distinction to virtually all of these objective functions, the formulation proposed here has very important additional features: it is general enough to accommodate a lot of different concrete definitions at various levels of resolution (object descriptions, distances-proximities, clusters,...), including the possibility of appropriate expression of the cluster-wise identification problem (postulate II), and it suggests a definite algorithmic simplification, outlined in the subsequent section, under quite mild conditions (see, again, Owsiński 1991) – thus fulfilling postulate III.

5. An algorithmic suggestion

Although the general objective function proposed allows for avoiding the trap of monotonicity and for the search for the "proper" number of clusters without any additional criteria nor tests, the fundamental numerical difficulty remains, as attached to the concrete formulations of the functions involved. Thus, the algorithms used for these formulations (e.g. various exchange algorithms) will have to be used also in this framework. At this level of generality the sole facilitation – though not to be overlooked – is related to the possibility of making reasonable comparisons for various solutions, also those differing as to the value of p.

We have, however, assumed "opposite monotonicity", just in order to secure the fundamental properties of the objective function. This entails further – algorithmic – possibilities.

Denote, namely, by $\Delta^p C_D(P)$ the difference between the (optimum) values of $C_D(P)$ for a p and p-1. Likewise for $\Delta^p C^S(P)$. Unless they are equal zero, their signs differ. Their sum, $\Delta^p C^*(P) = \Delta^p C_D(P) + \Delta^p C^S(P)$, is the basis

for assessing whether it is worth to move up or down with p, at least locally. Moreover, this value can be used not just on the optimum partitions. Further, for quite a class of $C_D(P)$'s and $C^S(P)$'s it can be shown that the $\Delta^p C_D(P)$ and $\Delta^p C^S(P)$ are also (weakly) opposite monotonic in p, this fact being in direct connection with the cardinalities of respective (sub)sets involved (see Owsiński 1990). It is, generally, quite common to be able to establish definite regularities concerning dependence of $\Delta^p C_D(P)$ and $\Delta^p C^S(P)$ on p. If so, we can consider the formulation

$$C^*(P,r) = r\Delta^p C_D(P) + (1-r)\Delta^p C^S(P), \text{ with } r \in [0,1].$$

The proposed procedure would start from r=1, for which the global solution in terms of p is as close to n as the minimum size of clusters, necessary for identification of models, allows. As the value of r is decreased, the optimum value of p for $C^*(P,r)$ decreases as well, with the actual $P(p)$ being determined through step-by-step procedures. These procedures may, in particular, take the form of progressive mergers, like in the classical clustering schemes, or more complex procedures (e.g. split-and-merge), based on the analysis of the values of $\Delta^p C_D(P)$ and $\Delta^p C^S(P)$. The same, of course, applies to the objective function $C^*(P,r)$.

This algorithmic proposal leaves, of course, still a lot of questions open. Notwithstanding the simplification offered, numerical difficulties remain. They do not just apply to the optimisation procedure. Most of all – the determination of the starting point seems to be the essential difficulty in the cluster-wise model identification problem.

References

Bock H.-H. (1996) Probability models and hypotheses testing in partitioning cluster analysis. In: P. Arabie, L.J. Hubert, G. De Soete, eds. *Clustering and Classification*. World Scientific, River Edge (New Jersey), 377-453.

Celeux G., Govaert G. (1993) Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. of Statistical Computation and Simulation*, **47**, 127-146.

DeSarbo W.S., Oliver R.L., Rangaswamy A. (1989) A simulated annealing methodology for clusterwise linear regression. *Psychometrika*, **54**, 4, 707-736.

Dunn J. (1974) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics*, **3**, 32-57.

Fraley C., Raftery A.E. (1998) *How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. TR*, **329**, Department of Statistics, University of Washington, Seattle.

Hennig Ch. (2000) Identifiability of models for clusterwise linear regression. *J. of Classification*, **17**, 273-296.

Jajuga K. (1986) Linear fuzzy regression. *Fuzzy Sets and Systems*, **20**, 343-53.

Lau Kin-nam, Leung Pui-lam, Tse Ka-kit (1999) A mathematical programming approach to clusterwise regression model and its extensions. *EJOR.*, **116**, 640-652.

Mańczak K. (1979) *Metody identyfikacji wielowymiarowych obiektów sterowania*. WNT, Warszawa.

Marcotorchino F. (1985) *Maximal Association Theory*. F-091, IBM Centre Scientifique de Paris, Paris.

Mirkin B. (1996) *Mathematical Classification and Clustering*. Kluwer, Dordrecht.

Nakamori Y., Ryoke M. (1994) Identification of fuzzy prediction models through hyperellipsoidal clustering. *IEEE Transactions SME.*, **24**, 8, 1153-1173.

Owsiński J.W. (1989) On global optimality in cluster-wise regression. *Control and Cybernetics*, **18**, 1, 53-67.

Owsiński J.W. (1990) On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, **6**, 157-171.

Owsiński J.W. (1991) *Nowa metoda analizy skupień z globalna funkcją celu*. Rzoprawa doktorska (Ph.D. dissertation). Systems Research Institute, Polish Academy of Sciences, Warszawa.

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

Späth H. (1979) Clusterwise linear regression. *Computing*, **22**, 4, 367-373.

Stanfel L.E. (1992) *Contributions to a Theory of Clustering*. Typescript. Management Science and Statistics Department, The University of Alabama, Tuscaloosa.

ISBN 83-85847-78-2