k-Means, Ward and Probabilistic Distance-Based Clustering Methods with Contiguity Constraint

Andrzej Młodak^{1,2}

Published online: 26 August 2020
© The Author(s) 2020

Abstract

We analyze some possibilities of using contiguity (neighbourhood) matrix as a constraint in the clustering made by the *k*-means and Ward methods as well as by an approach based on distances and probabilistic assignments aimed at obtaining a solution of the multi-facility location problem (MFLP). That is, some special two-stage algorithms being the kinds of clustering with relational constraint are proposed. They optimize division of set of objects into clusters respecting the requirement that neighbours have to belong to the same cluster. In the case of the probabilistic *d*-clustering, relevant modification of its target function is suggested and studied. Versatile simulation study and empirical analysis verify the practical efficiency of these methods. The quality of clustering is assessed on the basis of indices of homogeneity, heterogeneity and correctness of clusters as well as the silhouette index. Using these tools and similarity indices (Rand, Peirce and Sokal and Sneath), it was shown that the probabilistic *d*-clustering can produce better results than Ward's algorithm. In comparison with the *k*-means approach, the probabilistic *d*-clustering—although gives rather similar results—is more robust to creation of trivial (of which empty) clusters and produces less diversified (in replications, in terms of correctness) results than *k*-means approach, i.e. is more predictable from the point of view of the clustering quality.

Keywords Contiguity constraint · *k*-Means method · Ward method · Probabilistic *d*-clustering · Correctness of clustering · Silhouette index · Rand index · Peirce index · Sokal and Sneath index

1 Introduction

The cluster analysis, aimed at efficient classifying given multidimensional objects (i.e. objects described by many statistical features reflecting a composite social or economic phenomenon,

✉ Andrzej Młodak
a.mlodak@stat.gov.pl

¹ Statistical Office in Poznań, Branch in Kalisz, ul. Piwoniczka 7–9, 62–800 Kalisz, Poland

² The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Inter-Faculty Department of Mathematics and Statistics, ul. Nowy Świat 4, 62–800 Kalisz, Poland

such as labour market, environmental protection, economic situation) to internally homogeneous and mutually heterogeneous classes, offers many interesting algorithms. They can be of various types, such as translational (when object can be translated from one predefined cluster to another until each object is located in cluster which centre is closest to it), hierarchical (when at each step the closest clusters are joined into one) and probabilistic (if the assignment of objects to clusters is defined by especially optimize probabilities). The first of the aforementioned classes is represented by the k -means method (see e.g. Ball and Hall (1967) or Everitt et al. (2011)), where k is the arbitrarily established number of clusters. A good example of the second type of clustering can be the Ward algorithm (cf. Ward (1963)), where the distance between clusters is computed as the Euclidean distance between their centroids.

Special attention will be paid, however, to the probabilistic d -clustering. This approach proposed by Ben-Israel and Iyigun (2008) and developed further by them (cf. Iyigun and Ben-Israel (2013)) is a special kind of the multi-facility location problem (MFLP). It consists of simultaneous optimization of probabilistic assignments of given objects to clusters and distances of such objects from centres of these clusters. This aim is achieved by minimization of the target function. This function is a summation of products of distances of objects from centres of clusters and a power of relevant probabilistic assignments. Such formulation of this problem allows an efficient and dexterous application of Weiszfeld's algorithm (Weiszfeld (1937)) to solve it and to obtain a high-quality solution. In relation to the interval data, this approach was studied also by Młodak and Kubacki (2010).

Using such method, one should, however, not forget about many important aspects of multivariate data analysis. One of them is the question of contiguity, called also the neighbourhood. It is usually associated with the close proximity of objects but this proximity can be understood in various ways. First of them and most natural is the physical dimension, which is based on existing common boundaries established as a result of the implementation of administrative decisions or political agreements. Such approach belongs to the fundamentals of statistical systems and surveys.

The second concept of neighbourhood is connected with proximity of objects according to some social or economical features. A good example of importance of such approach can be the construction of the Larger Urban Zones made within the panEuropean program of statistical monitoring of cities URBAN AUDIT, which are determined according to the intensity of journeys to work to the analyzed cities from gminas (the Polish NUTS 5 regions) for which such intensity was the greatest (i.e. where 20% of the employees living in a given NUTS 5 spatial unit usually commuted to the city—cf. Józefowski and Młodak (2009)). On the other hand, the sub-city districts—according to the assumptions of this project—are determined by joining spatially neighbouring areas within a city having similar structure of population and relevant compactness of buildings. Therefore, a clustering supported by information about the neighbourhood of simple units (e.g. statistical areas) is necessary (cf. Borys and Rogala (2008)).

The neighbourhood can be assessed also on the basis of a wider set of statistical variables, when the key aspect of interest has multivariate form. In this case, it is, however, very important that the variables being a basis of determination of such neighbourhood be different than those used for data estimation or clustering. The proximity of the objects is here a function of values of these variables for such objects.

Many types of neighbourhood were in the last years analyzed from various points of view. Kelejian and Prucha (1998) studied the cross-sectional autoregressive spatial model (CAS) which uses two different types of neighbourhood matrix for modelling spatial co-dependency

of statistical variables and proposed a three-step procedure to estimate parameters of such model. Wagner and Mantaj (2010) analyzed mathematical and practical properties of the neighbourhood matrix in the physical dimension and its applications in taxonomy. LeSage and Pace (2009) studied usefulness of the spatial econometric Bayesian model which takes the neighbourhood of areas into account. The same researchers (LeSage and Pace (2010)) verified the hypothesis about sensitiveness of the Spatial Regression Model (SAR) to individual specifications and investigated how to interpret its relevant parameters. Młodak (2013) turned his mind on some practical properties of both types of neighbourhood matrix, proposed an efficient method of their comparison and studied possibilities of estimation of variance of disturbances and spatial lag vectors and of effects and prediction correlation in CAS and SAR models. General theory of the physical neighbourhood was here also conceptualized.

Taking the above results into account, it should be not surprising that the importance of neighbourhood was appreciated also in the small area estimation. One can mention in this context a generalization of the Fay–Herriot model taking correlated random area effects between neighbouring areas into account modelled using the SAR process¹—cf. Singh et al. (2005), Petrucci and Salvati (2006) and Pratesi and Salvati (2008). Saefuddin et al. (2013) used the SAR and CAR (conditional autoregression) models to extend the hierarchical Bayes method towards improving quality of poverty estimation in East Java, Indonesia. And, last but not least, both small area estimation models with the sub-tools based on the neighbourhood of powiats (Polish NUTS 4 regions) were used by Kubacki and Jędrzejczak (2016) to estimate average per capita available income of households with the support of data from the Polish tax register (POLTAX) as auxiliary variables. It is worth recalling also that clustering of units is one of the factors deciding about efficiency of stratified sampling (cf. e.g. Zimmer et al. (2012)).

In this paper, we present and study efficiency of the k -means and Ward methods when contiguity constraint is imposed. That is, it is required that neighbours belong to the same cluster. To satisfy this condition, a type of clustering algorithm with relational constraint (cf. Ferligoj and Batagelj (1982), Basu et al. (2008)) is constructed. It consists of two steps. In the first of them, the pre-clusters consisting of objects having at least one neighbour are determined. Next, such pre-clusters and remaining objects (i.e. those which have no neighbour) are finally clustered. In the case of k -means method, the pre-clusters are represented by their centroids.

We make also an attempt to introduce the factor of neighbourhood (and therefore the additional information which it carries) to the probabilistic d -clustering. In the conventional approach formulated by Ben-Israel and Iyigun (2008), the target function of probabilistic assignments and centres of clusters is optimized by a specific iteration algorithm based on Weiszfeld's idea (cf. Weiszfeld (1937)). Therefore, this approach is a type of the multi-facility location problem (MFLP), which is aimed at simultaneous optimization of level of assignment and location of objects. Now, we propose a modified target function, where the matrix of distances of objects from the cluster centres is multiplied by the corrected neighbourhood matrix (i.e. taking also trivial neighbourhood—an object with itself—into account). We prove that many properties of the Ben-Israel and Iyigun model and its optimum solutions can be easily observed (in relevant adopted form) also in this case—especially in terms of asymptotics. Additionally, some original features concerning the relations of these modified methods with its conventional option and restrictions of some parameters are also revealed. In particular, the dual problem (which in this case becomes more complicated) is also shown.

¹ In these papers, the abbreviation SAR is explained as simultaneously autoregressive process.

On the basis of such modification, an algorithm of clustering with contiguity constraint is introduced. It consists also of two steps. In the first place, the objects having some neighbours are grouped using the modified optimization procedure. Each of such clusters is next represented by its centroid and these “new” objects and objects having no neighbours are clustered using the basic probabilistic d -clustering approach. One can emphasize that the proposed algorithms work over not the original objects but rather classes of contiguity equivalence relation. This method can be useful when predefinition of neighbourhood is necessary from some practical reasons.

The efficiency of introduced approaches and comparison of impact of the neighbourhood matrix on each of them was verified by a special simulation and empirical studies. In the simulation experiment, the number of clusters is arbitrarily established. It results from the pre-assumed groups of data drawn from relevant multivariate normal distribution. We analyze both balanced option of such sample (i.e. the sizes of subsamples generated from individual multivariate normal distributions are equal) and unbalanced realizations (the subsample sizes are drawn from the multinomial distribution). The impact of adding small and intensive noise to the data was also investigated. The neighbourhood is here established on the basis of the variables represented by some specific dimensions of such distributions (the others are regarded to be the model diagnostic variables, which are a basis of clustering). Special indices of homogeneity, heterogeneity and correctness of clusters and the silhouette index will be used to verify their quality. The conducted empirical study is based on current statistical data concerning the labour market in powiats (NUTS 4 regions) of the Wielkopolskie Voivodship (the Polish NUTS 2 region located in central-west part of the country). The criterion of neighbourhood of two powiats was here defined as the absolute difference between their population densities. Unlike the simulation study, the number of clusters is here established endogenously, by optimizing the index of correctness of clustering. The final set of clusters is that the value of such index is smallest. These two attempts show the utility of analyzed methods. Especially it was observed that the probabilistic d -clustering produces clusterings of better quality than Ward’s approach and is more robust to produce trivial and empty clusters and gives less diversified results in replications than the k -means approach (although in many cases the results of application of probabilistic d -clustering and k -means algorithms are similar). Moreover, the obtained results call into question the appropriateness of the silhouette index in this case.

The paper is organized as follows. Firstly (Section 2), we present the main assumptions of the concept of neighbourhood and algorithms for k -means and Ward methods which take the occurrence of contiguity constraint into account. Next (Section 3), the multi-facility location problem and probabilistic d -clustering are discussed. Section 4 contains the modification of the Ben-Israel and Iyigun model obtained by the introduction of a neighbourhood matrix is discussed, its properties and relevant dual problem are investigated. A relevant modified algorithm for this new approach is formulated. In Section 5, we describe the assumptions and results of conducted simulation study and in Section 6, the results of empirical analysis for the Wielkopolskie Voivodship in Poland. Finally, in Section 7, some major concluding remarks are formulated. Two appendices containing the proof of some mathematical results of the paper and results of additional computations are also included.

2 Preliminaries and Modified k -Means and Ward Algorithms

In this section, we recall the basic assumptions of cluster analysis and the concept of neighbourhood. We show also how one can modify the k -means and Ward algorithms when the contiguity constraint is imposed.

2.1 Basic Notions and Assumptions

Let n be a natural number and $\Gamma = \{\Gamma_1, \Gamma_2, \dots, \Gamma_n\}$ —the set of analyzed objects. We assume that these objects are characterized by m (where m is a natural number) features $X_1, X_2, \dots, X_m$. Any object Γ_i is then represented by a point $\gamma_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$, $i = 1, 2, \dots, n$. The aim of the clustering is to divide the set Γ into clusters $\Theta_1, \Theta_2, \dots, \Theta_s$, where $\Theta_j \subseteq \Gamma$, $\Theta_j \cap \Theta_k = \emptyset$ for $j \neq k$ and $\bigcup_{j=1}^s \Theta_j = \Gamma$, being maximally internally homogeneous and heterogeneous between themselves. The clusters can be defined e.g. using their centres $c_1, c_2, \dots, c_k$ (being often their centroids). Let $d(\mathbf{a}, \mathbf{b})$ be the distance between vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^m$.

In our study, we will use the model of *neighbourhood matrix* (called also the *contiguity matrix*). Such matrix of the form $\mathbf{W} = [w_{ij}]$ and size $n \times n$ is defined as

$$w_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } \Gamma_i \bowtie \Gamma_j, \\ 0 & \text{otherwise,} \end{cases}$$

for every $i, j = 1, 2, \dots, n$ (“ $\bowtie$ ” denotes the relation of *neighbourhood*—called also *adjacency*—of objects: the symbol $\Gamma_i \bowtie \Gamma_j$ means that Γ_i and Γ_j are neighbours). The matrix $\mathbf{W}$ is symmetric with zero diagonal entries. Moreover, the number of entries of $\mathbf{W}$ which are equal to 1 is exactly twice greater than the number of all pairs of neighbouring objects. In some options, instead of simple 1’s in the neighbourhood matrix levels of intensity of neighbourhood (belonging to $(0, 1]$) determined e.g. on the basis of distances of vectors describing such objects can be placed.² An object which is not adjacent to another (i.e. an object Γ_i for which $w_{ij} = 0$ for every $j \neq i, j = 1, 2, \dots, n$) is called *isolated*. If $w_{ij} > 0$ for any $j \neq i, j = 1, 2, \dots, n$ then Γ_i will be called *familiar*, because it is adjacent to each other.

If the objects are spatial areas, the neighbourhood can be defined in geographical manner. Two areas are then neighbours if they have a common border. However, the neighbourhood can be also established on the basis of some auxiliary features connected with some information associated with the main subject of clustering. Example 1 illustrates this idea.

Example 1. Assume that $n = 10$ and the neighbourhood will be determined on the basis of three auxiliary features Q_1, Q_2 and Q_3 . Their values for given objects are as follows:

Object	Q_1	Q_2	Q_3
Γ_1	8.3	0.5	−0.2
Γ_2	2.1	1.1	−1.3
Γ_3	−1.3	1.5	4.5
Γ_4	4.2	2.2	0.9
Γ_5	7.1	3.8	1.6
Γ_6	3.4	1.1	0.6
Γ_7	0.1	2.9	1.7

² For more details, see LeSage and Pace (2009) or Młodak (2013).

Object	Q_1	Q_2	Q_3
Γ_8	-0.8	1.4	-5.4
Γ_9	1.4	4.0	0.3
Γ_{10}	2.1	0.7	3.8

The features Q_1 , Q_2 and Q_3 are of sufficient variability (their coefficients of variation vary from 62.05 to 401.37%) and weakly correlated (the strongest correlation is observed between Q_2 and Q_3 —Pearson’s correlation coefficient amounts to 0.0865 in this case, the remaining ones are 0.0323 (for Q_1 and Q_3) and -0.0101 (Q_1 and Q_2). The neighbourhood is defined using the Euclidean distances between points γ_i , $i = 1, 2, \dots, n$. Hence, the matrix of distances is of the form:

Object	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6	Γ_7	Γ_8	Γ_9	Γ_{10}
Γ_1	0.0000	6.3253	10.7355	4.5727	3.9459	5.0010	8.7527	10.5195	7.7531	7.3811
Γ_2	6.3253	0.0000	6.7350	3.2342	6.3797	<i>2.3022</i>	4.0299	5.0309	3.3853	5.1157
Γ_3	10.7355	6.7350	0.0000	6.6106	9.1793	6.1205	3.4293	9.9131	5.5839	3.5623
Γ_4	4.5727	3.2342	6.6106	0.0000	3.3853	<i>1.3928</i>	4.2356	8.0827	3.3823	3.8820
Γ_5	3.9459	6.3797	9.1793	3.3853	0.0000	4.6883	7.0583	10.8245	5.8498	6.2809
Γ_6	5.0010	<i>2.3022</i>	6.1205	<i>1.3928</i>	4.6883	0.0000	3.9166	7.3301	3.5355	3.4771
Γ_7	8.7527	4.0299	3.4293	4.2356	7.0583	3.9166	0.0000	7.3123	<i>2.2045</i>	3.6401
Γ_8	10.5195	5.0309	9.9131	8.0827	10.8245	7.3301	7.3123	0.0000	6.6400	9.6716
Γ_9	7.7531	3.3853	5.5839	3.3823	5.8498	3.5355	<i>2.2045</i>	6.6400	0.0000	4.8611
Γ_{10}	7.3811	5.1157	3.5623	3.8820	6.2809	3.4771	3.6401	9.6716	4.8611	0.0000

We assume that objects Γ_h and Γ_i , $h \neq i$ are neighbours if the Euclidean distance between γ_h and γ_i is smaller than 3, $h, i = 1, 2, \dots, n$. Such distances were written italic in the aforementioned table. The neighbourhood matrix is then as follows:

Object	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6	Γ_7	Γ_8	Γ_9	Γ_{10}
Γ_1	0	0	0	0	0	0	0	0	0	0
Γ_2	0	0	0	0	0	1	0	0	0	0
Γ_3	0	0	0	0	0	0	0	0	0	0
Γ_4	0	0	0	0	0	1	0	0	0	0
Γ_5	0	0	0	0	0	0	0	0	0	0
Γ_6	0	1	0	1	0	0	0	0	0	0
Γ_7	0	0	0	0	0	0	0	0	1	0
Γ_8	0	0	0	0	0	0	0	0	0	0
Γ_9	0	0	0	0	0	0	1	0	0	0
Γ_{10}	0	0	0	0	0	0	0	0	0	0

Therefore, $\Gamma_2 \bowtie \Gamma_6$, $\Gamma_4 \bowtie \Gamma_6$ and $\Gamma_7 \bowtie \Gamma_9$.

From the practical point of view, two manners of normalization of the neighbourhood matrix are promoted in the literature. First, Keleşian and Prucha (1998) use the maximum absolute eigenvalue of the neighbourhood matrix. Let such value be denoted by τ . Thus, all entries of the spatial weight matrix are divided by τ . Hence, the normalized matrix has entries equal to 0 or $1/\tau$. Another approach, suggested by LeSage and Pace (2009)—called *row-standardization*—is based on standardization of rows of the neighbourhood matrix. Concretely, they normalize the matrix such that entries in each row with at least one non-zero entry adds up to one, i.e. each entry is divided by the sum of elements of the row which it belongs to. Such normalization becomes necessary in the case of analysis of the CAS models. It has to be performed in order to provide a parameter space for the spatial autoregressive parameters (cf. Młodak (2013)).

2.2 k -Means Algorithm with Contiguity Constraint

In many practical situations, the neighbourhood is defined on the basis of some additional information connected with the main data used to clustering. Thus, a natural expectation is that neighbours have to belong to the same cluster. In this way, a type of clustering with relational constraint, considered by Ferligoj and Batagelj (1982) and Basu et al. (2008) was constructed. Below we show how one can adopt k -means method to this situation.

The basis feature of this approach in these circumstances is the division of the set of objects into two disjoint subsets: $\Gamma_{(NB)}$, containing objects having at least one neighbour, and $\Gamma_{(NN)}$, consisting of objects which have no neighbour. Of course, $\Gamma_{(NB)} \cup \Gamma_{(NN)} = \Gamma$. Let now $\Gamma_{(NB)} = \{\Gamma_{(NB)1}, \Gamma_{(NB)2}, \dots, \Gamma_{(NB)n_{NB}}\}$ and $\Gamma_{(NN)} = \{\Gamma_{(NN)1}, \Gamma_{(NN)2}, \dots, \Gamma_{(NN)n_{NN}}\}$, where $n_{NB} + n_{NN} = n$, $\Gamma_{(NB)i} \in \Gamma$, $\Gamma_{(NN)k} \in \Gamma$, $i = 1, 2, \dots, n_{NB}$, $k = 1, 2, \dots, n_{NN}$. The modified k -means algorithm consists then of two parts: clustering the objects from $\Gamma_{(NB)}$ and clustering clusters obtained in the first part (represented by their centroids) and the objects from $\Gamma_{(NN)}$. This approach is therefore of the form indicated by Algorithm 1. In Part I, the objects having at least one neighbour are grouped according to the minimax criterion of distance of neighbouring objects from the centroids of clusters obtained in previous iteration (i.e. any two neighbours are included in such cluster which centroid is the closest to them, taking into account the greater of distances of these objects from it). In this way, any neighbours will belong to the same cluster. In Part II, the clusters obtained in Part I are represented by their centroids. This collection is extended by objects having no neighbour (and therefore being not considered in Part I) and k -means grouping is here conducted. For instance, if we have the objects A, B, C, D, E and F and $A \bowtie B$ and $C \bowtie D$, then we group first the objects A, B, C and D (so the structure of their clusters are $\{\{A, B\}, \{C, D\}\}$ or $\{\{A, B, C, D\}\}$) and next, the k -means clustering is conducted on the collection of these clusters and objects E, F.

Algorithm 1.

Part I

Step 1. Let $1 < s_{NB} < n_{NB}$ be the assumed number of clusters for $\Gamma_{(NB)}$ and $c_{(NB)1}^{(0)}, c_{(NB)2}^{(0)}, \dots, c_{(NB)s_{NB}}^{(0)}$ – their initial centroids. Determine the preliminary clusters (called also *the transitive closure*) $\Theta_{(NB)1}^{(1)}, \Theta_{(NB)2}^{(1)}, \dots, \Theta_{(NB)s_{NB}}^{(1)}$ as $\Gamma_{(NB)i}, \Gamma_{(NB)k} \in \Theta_{(NB)l}^{(1)}$ if $\Gamma_{(NB)i}$ and $\Gamma_{(NB)k}$ are neighbours and the distance of centroid $c_{(NB)l}^{(0)}$ of $\Theta_{(NB)l}^{(1)}$ from the farther points $\gamma_{(NB)i}, \gamma_{(NB)k}$ is the smallest among relevant distances from $c_{(NB)l}^{(0)}$ (centroids of $\Theta_{(NB)l}^{(1)}$), $l = 1, 2, \dots, s_{NB}$. Next, the centres of transitive closure are computed.

Step r -th ($r = 2, 3, \dots$). Determine clusters $\Theta_{(NB)1}^{(r)}, \Theta_{(NB)2}^{(r)}, \dots, \Theta_{(NB)s_{NB}}^{(r)}$ in the following way:

$$\Gamma_{(NB)i}, \Gamma_{(NB)k} \in \Theta_{(NB)l}^{(r)} \text{ if } \Gamma_{(NB)i} \bowtie \Gamma_{(NB)k} \text{ and}$$

$$\max \left\{ d\left(\gamma_{(NB)i}, c_{(NB)l}^{(r-1)}\right), d\left(\gamma_{(NB)k}, c_{(NB)l}^{(r-1)}\right) \right\} =$$

$$\min_{l=1,2,\dots,s_{NB}} \max \left\{ d\left(\gamma_{(NB)i}, c_{(NB)l}^{(r-1)}\right), d\left(\gamma_{(NB)k}, c_{(NB)l}^{(r-1)}\right) \right\},$$

for every $i, k \in \{1, 2, \dots, n_{NB}\}$. Compute centroids of such clusters $\mathbf{c}_{(NB)1}^{(r)}, \mathbf{c}_{(NB)2}^{(r)}, \dots, \mathbf{c}_{(NB)s_{NB}}^{(r)}$. If $\sum_{k=1}^{s_{NB}} d(\mathbf{c}_{(NB)k}^{(r)}, \mathbf{c}_{(NB)k}^{(r-1)}) < \varepsilon$ where $\varepsilon > 0$ is arbitrarily established level of precision then stop else go to step $r + 1$.

Part II

Step 1. Create $n^* \triangleq s_{NB} + n_{NN}$ objects $\Gamma_1^*, \Gamma_2^*, \dots, \Gamma_{n^*}^*$ such that $\gamma_i^* = \mathbf{c}_{(NB)i}^{(r)}$, where $\mathbf{c}_{(NB)i}^{(r)}$ is the centroid of i -th cluster $\Theta_{(NB)i}^{(r)}$ obtained finally in Part I, $i = 1, 2, \dots, s_{NB}$, and $\Gamma_i^* = \Gamma_{(NN)(n^*-i)}$, where $\Gamma_{(NN)(n^*-i)}$ is $n^* - i$ -th subsequent object which has no neighbour, $i = s_{NB} + 1, s_{NB} + 2, \dots, n^*$. Establish initial centroids $\mathbf{c}_1^{(1)}, \mathbf{c}_2^{(1)}, \dots, \mathbf{c}_{s^*}^{(1)}$.

Step r -th ($r = 2, 3, \dots$): determine clusters $\Theta_1^{(r)}, \Theta_2^{(r)}, \dots, \Theta_s^{(r)}$ in the following way:

$$\Gamma_i^* \in \Theta_l^{(r)} \quad \text{if} \quad d(\gamma_i^*, \mathbf{c}_l^{(r-1)}) = \min_{l=1,2,\dots,s_{NB}} d(\gamma_i^*, \mathbf{c}_l^{(r-1)}),$$

$i = 1, 2, \dots, n^*$. Compute centroids of such clusters $\mathbf{c}_1^{(r)}, \mathbf{c}_2^{(r)}, \dots, \mathbf{c}_{s^*}^{(r)}$. If $\sum_{k=1}^s d(\mathbf{c}_k^{(r)}, \mathbf{c}_k^{(r-1)}) < \varepsilon$ – where $\varepsilon > 0$ is the maximum acceptable level of imprecision – then stop else go to step $r + 1$.

We must prove that the relation expressed in the r -th step of Part I of Algorithm 1 is the equivalence relation. It is clear that it has the reflexive and symmetric properties. We will show now the transitivity. Let, for simplification, $u_l \triangleq d(\gamma_{(NB)l}^{(r-1)}, \mathbf{c}_{(NB)l}^{(r-1)})$, $v_l \triangleq d(\gamma_{(NB)h}^{(r-1)}, \mathbf{c}_{(NB)l}^{(r-1)})$ and $w_l \triangleq d(\gamma_{(NB)k}^{(r-1)}, \mathbf{c}_{(NB)l}^{(r-1)})$, $l = 1, 2, \dots, s_{NB}$. Assume that $\min_l (\max\{u_l, v_l\}) = \max\{u_{l^*}, v_{l^*}\}$ and $\min_l (\max\{v_l, w_l\}) = \max\{v_{l^*}, w_{l^*}\}$ for some $l^* \in \{1, 2, \dots, s_{NB}\}$. We have to show that $\min_l (\max\{u_l, w_l\}) = \max\{u_{l^*}, w_{l^*}\}$. By the definition, we have $\min_l (\max\{u_l, w_l\}) \leq \max\{u_{l^*}, w_{l^*}\}$. Note that $\min_l (\max\{u_l, w_l\}) \geq \min\left\{\min_l (\max\{u_l, v_l\}), \min_l (\max\{v_l, w_l\})\right\} = mnl$. Thus, $mnl = \min\{\max\{u_{l^*}, v_{l^*}\}, \max\{v_{l^*}, w_{l^*}\}\}$. If $v_{l^*} \geq \max\{u_{l^*}, w_{l^*}\}$, then $mnl = v_{l^*}$, what gives $\min_l (\max\{u_l, w_l\}) \geq \max\{u_{l^*}, w_{l^*}\}$ and, in consequence, $\min_l (\max\{u_l, w_l\}) = \max\{u_{l^*}, w_{l^*}\}$. Assume now that $v_{l^*} < \max\{u_{l^*}, w_{l^*}\}$. Let $\max\{u_{l^*}, w_{l^*}\} = w_{l^*}$. We have then $\min_l (\max\{v_l, w_l\}) = \max\{v_{l^*}, w_{l^*}\} = w_{l^*} = \max\{u_{l^*}, w_{l^*}\}$. Hence, $\min_l (\max\{u_l, w_l\}) = \min_l (\max\{u_l, v_l, w_l\}) \geq \min_l (\max\{v_l, w_l\}) = w_{l^*} = \max\{u_{l^*}, w_{l^*}\}$. Analogously, we can prove that if $\max\{u_{l^*}, w_{l^*}\} = u_{l^*}$, then $\min_l (\max\{u_l, w_l\}) = \min_l (\max\{u_l, v_l, w_l\}) \geq \min_l (\max\{u_l, v_l\}) = u_{l^*} = \max\{u_{l^*}, w_{l^*}\}$. In both cases, it leads to the conclusion that $\min_l (\max\{u_l, w_l\}) = \max\{u_{l^*}, w_{l^*}\}$.

Thus, in each step of any part of Algorithm 1, the given object or objects is/are assigned to the cluster defined by the nearest centroid. The nearest centroid is determined by minimization of distance between object(s) and all analyzed centroids. One can ask how to treat the situation when for the same object(s) two or more centroids are the nearest (i.e. there are more than one centroid which the distance from given object(s) is the same and the smallest). In such situation, such object(s) is/are assigned only to one of these clusters. In practice, if the algorithm is implemented to software programming language, it assigns such object(s) to such cluster which was firstly discovered as nearest. In consequence, the final clusters are disjoint.

The initial centroids can be defined randomly, taking the empirical distribution of diagnostic variables into account. The simulation and empirical studies (Sections 5 and 6) showed that it is a good solution, owing to applied maximum level of imprecision Algorithm 1 converges. Sometimes such convergence may denote, of course, that there will be no further changes after a finite number of steps.

The quality of such clustering can be assessed using the indices of homogeneity, heterogeneity and correctness of clusters (cf. e.g. Młodak (2006)). Especially useful in this context are the indices of this class based on the median of distances between objects—they are robust to incidental outliers inside clusters and between clusters. The index of homogeneity of l th cluster Θ_l is then of the form:

$$hm_l \stackrel{\text{def}}{=} \text{med}_{h: \Gamma_h, \Gamma_i \in \Theta_l, h \neq i} d(\gamma_h, \gamma_i),$$

$l = 1, 2, \dots, s$. Thus, the final aggregated index of homogeneity is defined as:

$$hm = \text{med}_{l=1,2,\dots,s} hm_l. \quad (1)$$

The smaller is the index hm then the more homogeneous is the internal structure of clusters.

The heterogeneity of cluster in relation to the remaining ones is given as

$$ht_l \stackrel{\text{def}}{=} \text{med}_{h: \Gamma_h \in \Theta_l, i: \Gamma_i \in \Gamma \setminus \Theta_l} d(\gamma_h, \gamma_i),$$

$l = 1, 2, \dots, s$. The aggregation of such indices leads to the final assessment of heterogeneity:

$$ht = \text{med}_{l=1,2,\dots,s} ht_l. \quad (2)$$

The greater is the index ht , the stronger heterogeneous are obtained clusters. The index of correctness reflects the ratio of homo- and heterogeneity of clusters:

$$cr = \frac{hm}{ht}. \quad (3)$$

The smaller values of cr inform about better correctness of clusters, i.e. that they are more homogeneous and more distinct from one another. Clustering can be regarded as efficient if $cr < 1$.

The construction of index (3) based on (1) and (2) is a proposal slightly than many indices used for evaluation of clustering algorithms. One can recall here the following known measures of quality of clustering.

The Dunn index (cf. Dunn (1974)) is based on the dissimilarity between clusters and maximum diameter of them, i.e.

$$DU = \min_{k=1,2,\dots,s} \left(\min_{l=1,2,\dots,s, l \neq k} \left(\frac{\text{diss}(\Theta_k, \Theta_l)}{\max_{r=1,2,\dots,s} \text{diam}(\Theta_r)} \right) \right),$$

where $\text{diss}(\Theta_k, \Theta_l) = \min_{h, i: \Gamma_h \in \Theta_k, \Gamma_i \in \Theta_l} \|\gamma_i - \gamma_j\|$, $\text{diam}(\Theta_r) = \max_{h, i: \Gamma_h \in \Theta_r} \|\gamma_i - \gamma_j\|$, $l, k, r = 1, 2, \dots, s$, and $\|\cdot\|$ is the Euclidean norm. The main disadvantage of DU is that—as observed e.g. by

Kamaruddin and Ravi (2020)—it is strongly affected by noisy data and outliers. The index cr overcomes these problems.

The Davies–Bouldin measure (cf. Davies and Bouldin (1979)) was created using the distance between centroids of clusters and sum of their diameters, i.e.

$$DB = \frac{1}{s} \sum_{k=1}^s \max_{l=1,2,\dots,s, l \neq k} \left(\frac{\text{diamc}(\Theta_k) + \text{diamc}(\Theta_l)}{\|c_k - c_l\|} \right),$$

where $\text{diamc}(\Theta_r) = \sqrt{\frac{1}{n_r} \sum_{i: \Gamma_i \in \Theta_r} \|\gamma_i - c_r\|^2}$ and n_r is the number of objects in Θ_r , $r = 1, 2, \dots, s$.

This measure assesses the differences between clusters only as the distance of their centroids, what sometimes can neglect the diversification of distances between individual objects belonging to various clusters. The index (3) to a large extent helps to avoid this inconvenience.

The silhouette index (cf. Kaufman and Rousseeuw (2005)) is a function of a_i —average distance of object Γ_i from all other objects in the same cluster—and b_i —the smallest the average dissimilarity between Γ_i and objects in any other clusters, which it does not belong to, $i = 1, 2, \dots, n$:

$$SI = \frac{1}{n} \sum_{i=1}^n \frac{b_i - a_i}{\max\{a_i, b_i\}}. \quad (4)$$

This index can be used to estimate only the first choice or the best partition and therefore is not useful in the case when intermediate clusters contained in the target one occur. The index (3) seems to be better from this point of view.

There are also some other interesting attempts in this context, e.g. the Maulik–Bandyopadhyay index (Maulik and Bandyopadhyay (2002)) is based on relation of intra-cluster distance for trivial clustering (i.e. when the set of objects Γ is the only cluster) for the current clustering and on the inter-cluster distance defined as maximal distance between centroids of clusters. However, its construction depends on some subjectively chosen parameter. On the other hand, Saitta et al. (2007) proposed a new index where between-class distance (being a sum of weighted distances between centroid of individual cluster and centroid of all the clusters) and within-class distance (sum of average distances of objects from centroids of relevant clusters) are used. In both cases, the inter-cluster distance may not fully reflect the diversification between clusters. Moreover, the intra-cluster variance is sensitive to incidental outliers within clusters and hence could be overestimated.

As one can see, the index (3) enables to avoid—or at least to substantially reduce—many aforementioned drawbacks of other indices. That is, it is based on distances between pairs of objects what enables to fully take internal diversification of clusters and differences between them into account. On the other hand, the usage of median of distances enables to robustify the quality assessment on incidental outliers in this respect. For a comparison, in the simulation studies described in Section 5, both (3) and (4) indices are used. Moreover, the proposal allows for easy separation the inter- and intra-cluster components (i.e. homo- and heterogeneity, respectively) of correctness of clustering and for clear recognition of contribution of each of these aspects to the total value of final index.

The members of clusters in each part of Algorithm 1, i.e. s_{NB} and s , can be established arbitrarily. If there is no other obstacle, it is comfortable to assume that $s_{NB} = s$. Such number can be, however, alternatively optimized endogenously. That is, we put $s = 2, 3, \dots$ and after finishing Algorithm 1, we compute the index of correctness (3). Finally, as the final number of

clusters, we assume such s , for which the value (3) is the smallest. The drawback of such approach is that it could be time and hard- and software capability consuming when basic dataset is large.

2.3 Contiguity Constraint in the Ward Method

The Ward method (cf. Ward (1963)) is a type of hierarchical clustering algorithm. In its conventional form, it starts from the analyzed objects as singletons, i.e. from the collection of clusters $\{\{\Gamma_1\}, \{\Gamma_2\}, \dots, \{\Gamma_n\}\}$. These n clusters are combined in successive steps to make one cluster containing all investigated objects. At each step, a pair of clusters whose merger minimizes the increase in the total sum of squares within-group error are merged. This increase is expressed as sum of squared differences between means of particular variables for these clusters divided by sum of reciprocal of their cardinalities. In practice, to obtain the optimal clustering, the procedure stops when the distance between merged clusters exceeds arbitrarily established threshold. When the Ward clustering is constrained by the restriction that neighbours should belong to the same cluster, it can be modified to the form presented in Algorithm 2.

Algorithm 2.

Step 1. Create the transitive closure $\Theta_{(NB)1}^{(1)}, \Theta_{(NB)2}^{(1)}, \dots, \Theta_{(NB)s_{NB}}^{(1)}$ in such way, that any two neighbours are included in the same cluster. For instance, if we have five objects A, B, C, D and E and $A \bowtie B$ and $C \bowtie D$ and $B \bowtie E$ then A, B and E will belong to one cluster and C and D to the second. In this way all objects between which there exists – direct or indirect – contiguity connection will belong to the same cluster. The transitive closure consists then of disjoint clusters.

Step 2. Set the starting collection $\Theta_1^{(2)}, \Theta_2^{(2)}, \dots, \Theta_{n^*}^{(2)}$ of clusters as consisting of the transitive closure obtained in Step 1 – i.e. $\Theta_k^{(2)} = \Theta_{(NB)k}^{(1)}$, $k = 1, 2, \dots, s_{NB}$ – and $\Theta_{s_{NB}+k}^{(2)} = \{\Gamma_{(NN)k}\}$, $k = 1, 2, \dots, n_{NN}$, i.e. objects having no neighbours treated as singleton clusters.

Step r -th ($r = 3, 4, \dots, n^* - 3$). Merge clusters $\Theta_{k^*}^{(r-1)}$ and $\Theta_{l^*}^{(r-1)}$ such that $k^*, l^* \in \{1, 2, \dots, n^* - r - 2\}$ and

$$d\left(c_{k^*}^{(r-1)}, c_{l^*}^{(r-1)}\right) = \min_{k, l \in \{1, 2, \dots, n^* - r - 2\}, k \neq l} d\left(c_k^{(r-1)}, c_l^{(r-1)}\right),$$

i.e. which are closest according to the distance between their centres; if $d\left(c_{k^*}^{(r-1)}, c_{l^*}^{(r-1)}\right) > d^*$, where $d^* > 0$ is arbitrarily established level of precision then the algorithm stops else go to step $r + 1$.

The distance d is here defined as

$$d\left(c_k^{(r)}, c_l^{(r)}\right) = \frac{\sum_{j=1}^m \left(\bar{x}_{jk}^{(r)} - \bar{x}_{jl}^{(r)}\right)^2}{\frac{1}{n_k^{(r)}} + \frac{1}{n_l^{(r)}}},$$

where $\bar{x}_{ju}^{(r)} = \frac{1}{n_u^{(r)}} \sum_{i: \Gamma_i \in \Theta_u^{(r)}} x_{ij}$ is the arithmetic mean of X_j in $\Theta_u^{(r)}$ and $n_u^{(r)}$ denotes the cardinality (i.e. the number of objects belonging to) of $\Theta_u^{(r)}$, $u = k, l, j = 1, 2, \dots, m, k, l = 1, 2, \dots, s, r = 1, 2, \dots$

Several issues connected with Algorithm 2 require more detailed explanation. One can observe the situation where all the objects belong to one cluster in transitive closure (step 1). However, this case occurs in practice very rarely. The assumptions of efficient cluster analysis (first of all, sufficiently high value of the coefficient of variation of diagnostic features) imposed clear diversification of analyzed objects from the point of view of the investigated composite phenomena. So, the combination of all the objects into one cluster is practically impossible. On the other hand, due to specific pre-clustering based on neighbourhood of objects according to some associated phenomenon, conducted in step 1 and hierarchical clustering made in steps 3, 4, ..., this method has a tendency to generate large final clusters. It can be observed e.g. for the results presented Section 6. So, Algorithm 2 does not “inherit” the inclination to generate clusters of similar size, which is one of immanent (and not always gainful) properties of the conventional Ward method (cf. e.g. Everitt et al. (2011)). Of course, it may happen that in step r the minimum distance will be achieved for two or more pairs of clusters from step $r - 1$. However, at each step only, one of these pairs is merged (the program usually does it with the first discovered optimal pair).

The threshold d^* can be established in various manners, e.g. quite exogenously or based on descriptive statistics of distances between clusters merged at particular stages over the whole algorithm (i.e. if steps 2, 3, ..., $n^* - 3$ leading to obtain complete set Γ are conducted). For instance, Mojena’s rule (Mojena (1977)) establishes such threshold as $d^* = \bar{d} + cs_d$, where $\bar{d}$ is the mean of minimum of distances, s_d —its standard deviation and c —some constant, usually $2.5 < c < 3$. The quality of clustering can be assessed using the indices (1), (2) and (3). The index (3) can be also used as optimization criterion of the algorithm—i.e. we regard as final such clustering for which the index (3) is the smallest.

3 Probabilistic d -Clustering

The probabilistic d -clustering,³ proposed by Ben-Israel and Iyigun (2008), is based on the optimization of probabilities of assignment of objects to clusters defined by their centroids, leading to obtain best clustering.

Formally, we will take into account $d(\gamma_i, c_k)$ the distance between the object Γ_i and centre c_k of cluster Θ_k and $p(\Gamma_i, \Theta_k)$ —the probability of assignment of the object Γ_i to the cluster Θ_k . For simplification, we will use in this context also shorter symbols d_{ik} and p_{ik} , $i = 1, 2, \dots, n, k = 1, 2, \dots, s$, respectively. According to the conventional definition of probability, two basic conditions hold:

$$0 \leq p_{ik} \leq 1, \text{ for every } i = 1, 2, \dots, n, k = 1, 2, \dots, s \quad (5)$$

and

³ The term “ d -clustering” is an abbreviation of *distance-clustering*.

$$\sum_{k=1}^s p_{ik} = 1 \text{ for every } i = 1, 2, \dots, n. \quad (6)$$

Let $\mathbf{p}_k = (p_{1k}, p_{2k}, \dots, p_{nk})$, $k = 1, 2, \dots, s$. Ben-Israel and Iyigun (2008) proposed clustering method aimed at minimization of the target function

$$f(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s) = \sum_{k=1}^s \sum_{i=1}^n d_{ik} p_{ik}^2 \quad (7)$$

with conditions (5) and (6) and a requirement of the probabilistic assignment to be reversely proportional to the distance of an object from the relevant given cluster,⁴ i.e.

$$p_{ik} = \frac{\alpha_k}{d_{ik}}, \quad (8)$$

where α_k is some constant depending only on k , $k = 1, 2, \dots, s$, $i = 1, 2, \dots, n$. It was proved that by these assumptions the optimal probability of assignment of Γ to cluster Θ_k will be of the form:

$$p_{ik} = \frac{\prod_{l=1, l \neq k}^s d_{il}}{\sum_{q=1}^s \prod_{l=1, l \neq q}^s d_{il}} \quad (9)$$

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

In the target function (7), instead of the squares of probabilities (i.e. p_{ik}^2), a generalized model based on their power of exponent $u \geq 1$ can be used:

$$p_{ik}^{(u)} = \frac{p_{ik}^u}{\sum_{q=1}^s p_{iq}^u} = \frac{\prod_{l=1, l \neq k}^s d_{il}^u}{\sum_{q=1}^s \prod_{l=1, l \neq q}^s d_{il}^u}, \quad (10)$$

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

On the basis of this result, Ben-Israel and Iyigun (2008)—inspired by Weiszfeld's approach—suggested an algorithm of iteration of centroids of clusters and optimal probabilistic assignments of objects to these clusters. Its scheme is given in Algorithm 3.

Algorithm 3.

Step 1. We assume arbitrarily the initial forms of centres $\mathbf{c}_1^{(0)}, \mathbf{c}_2^{(0)}, \dots, \mathbf{c}_s^{(0)}$ and the exponent $u_0 \geq 1$ as well as the increase of exponent, Δ . We put $r := 0$.

Step 2. Assuming that $\mathbf{c}_1^{(r)}, \mathbf{c}_2^{(r)}, \dots, \mathbf{c}_s^{(r)}$ (where $\mathbf{c}_k^{(r)}$ denotes the r -th iteration of $\mathbf{c}_k$, $k = 1, 2, \dots, s$) are given, we determine optimal probabilistic assignments according to the formula (10) and update the exponent: $u^+ := u + \Delta$.

Step 3. On the basis of results obtained in the Step 2, we determine optimal centres $\mathbf{c}_1^{(r+1)}, \mathbf{c}_2^{(r+1)}, \dots, \mathbf{c}_s^{(r+1)}$, for which the value of right – hand side of (7) is minimal. This minimization –

⁴ At the end of this section, we will show that this condition is in fact redundant. However, to present completely the results obtained by Ben-Israel and Iyigun (2008), we have placed it here at this moment.

according to the proposal expressed by Ben-Israel and Iyigun (2008) – is conducted with respect to $\mathbf{c}_k$ in $d(\gamma_i, \mathbf{c}_k) = d_{ik}$ when the values of p_{ik} are calculated according to (9) using $\mathbf{c}_1^{(r)}$, $\mathbf{c}_2^{(r)}$, ..., $\mathbf{c}_s^{(r)}$ and are assumed to be fixed. That is the centres are approximated using the formula

$$\mathbf{c}_k^{(r)} = \frac{\sum_{i=1}^n p_{ik}^2 \frac{\gamma_i}{\|\gamma_i - \mathbf{c}_k^{(r-1)}\|}}{\sum_{i=1}^n p_{ik}^2 \frac{1}{\|\gamma_i - \mathbf{c}_k^{(r-1)}\|}},$$

$k = 1, 2, \dots, s$, where $\|\cdot\|$ is the Euclidean norm.

Step 4. We repeat steps 2–3 putting $r = r + 1$ until $\sum_{k=1}^s d(\mathbf{c}_k^{(r)}, \mathbf{c}_k^{(r+1)}) < \varepsilon$, where ε is the arbitrarily established level of precision (e.g. $\varepsilon = 0.0001$).

Finally, each object is assigned to such cluster for which its probability of assignment is the greatest. In the case of ambiguity in assignment (i.e. if there are two or more clusters for which the probability of assignment are equally maximal), the object is assigned only to one of these clusters. The obtained clusters are the disjoint.

In fact, the subject of this study is so-called *multi-facility location problem* (MFLP), consisting of two following aspects:

- The assignment problem: having s centres $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_s$ assign each of the n objects to the cluster with the nearest centre,
- The location problem: having s clusters determined as a solution of the assignment problem, find centre—i.e. centroid—of each cluster, using relevant algorithm.

The probabilistic d -clustering resembles the fuzzy c -means method proposed by Dunn (1973) and Bezdek (1981). However, the target function is in this case given as

$$f(\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_s) = \sum_{k=1}^s \sum_{i=1}^n d_{ik}^2 q_{ik}^u, \quad (11)$$

where $\mathbf{q}_k = (q_{1k}, q_{2k}, \dots, q_{nk})$ and q_{ik} is the degree of membership of the object Γ_i in Θ_k ($i = 1, 2, \dots, n$, $k = 1, 2, \dots, s$) and $u \in [1, \infty)$. It is assumed that q_{ik} satisfy probability conditions analogous as (5) and (6). Function (11) achieves its minimum when

$$q_{ik} = \frac{d_{ik}^{\frac{2}{1-u}}}{\sum_{l=1}^s d_{il}^{\frac{2}{1-u}}}, \quad i = 1, 2, \dots, n, k = 1, 2, \dots, s.$$

The target function is then based on the squared distances of objects from centroids of clusters (instead of simple distances, as in (7)). Moreover, the degree of membership, q_{ik} , is here raised to some arbitrarily established power u . Algorithm 3 is applied here, respectively. It is worth noting that the optimization of q_{ik} is impossible when $u = 1$. Thus, in practice, $u \in [1, \infty)$. If $u = 3$, then $q_{ik} = p_{ik}$ for every $i = 1, 2, \dots, n$ and $k = 1, 2, \dots, s$. On the other hand, if $u = 2$, the Dunn–Bezdek target function (11) seems to be most similar to the Ben-Israel–Iyigun target function (7). The latter method seems to enable also to obtain better interpretable results as the former one.

However, the Ben-Israel and Iyigun approach has some practical drawbacks. Firstly, using squares of probabilities in (7) is in practical application hardly to interpret. Secondly, formula (8) assumes a dependency of p_{ik} only on the distance of i th object from k th cluster. But in

practice, explanatory variables describing a given social or economic composite phenomenon are mutually connected (independently of their formal connection such as e.g. correlation)—for instance, logically or by a type of information. Thus, the assignment of a given object to a given cluster depends to some extent on the assignment of remaining objects to this and other clusters. The conditions (5) and (6) seem to confirm such statement. Moreover, note that

$$\frac{\partial}{\partial p_{ik}} f(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s) = 2d_{ik}p_{ik} \quad (12)$$

for every $i = 1, 2, \dots, n$, $k = 1, 2, \dots, s$. From (6), we have $\sum_{k=1}^s p_{ik} = 1$, $i = 1, 2, \dots, n$. Using the Lagrange multipliers method and taking (12) into account, we obtain

$$2d_{ik}p_{ik} - \lambda_i = 0, \quad (13)$$

where λ_i are the Lagrange multipliers, $i = 1, 2, \dots, n$. Thus, from (13), we have $p_{ik} = \lambda_i / (2d_{ik})$ and from (6), $1 = \lambda_i \sum_{k=1}^s d_{ik}^{-1} / 2$. After determination of λ_i and placing it to the first of these equalities, we obtain (9). Thus, the Condition (8) is redundant and, in consequence, it is sufficient to restrict only to the conditions (5) and (6).

4 The Neighbourhood Matrix in Clustering Using the Ben-Israel–Iyigun Method

Now we will show how one can apply the additional information carried by the contiguity matrix in the probabilistic d -clustering. To introduce the neighbourhood of objects to the cluster analysis, we should remember that any object can be perceived as its own neighbour (what is called the *trivial neighbourhood*). It is important when distances between objects and centres of clusters are multiplied by the entries of the contiguity matrix—the distance of a given object from centres of clusters cannot be ignored. Therefore, a necessity of taking in the further study a trivial neighbourhood into account occurs. To do it, we will apply here an extended neighbourhood matrix, i.e. the $n \times n$ matrix $\mathbf{W}^* = [w_{ij}^*]$ with $w_{ii}^* = 1$ and $w_{ij}^* = w_{ij}$ for $i, j = 1, 2, \dots, n$, $i \neq j$.

In this case, our problem is to minimize the target function

$$f(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s) = \sum_{k=1}^s \sum_{i=1}^n p_{ik}^2 \sum_{h=1}^n w_{ih}^* d_{hk}, \quad (14)$$

by $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_s$ and relevant optimization of centroids of clusters. The optimal values of probabilities can be obtained analogously as for (6), i.e.

$$p_{ik} = \frac{\prod_{l=1, l \neq k}^s \sum_{h=1}^n w_{ih}^* d_{hl}}{\sum_{q=1}^s \prod_{l=1, l \neq q}^s \sum_{h=1}^n w_{ih}^* d_{hl}} \quad (15)$$

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

It is not difficult to see that if Γ_i is an isolated object, then p_{ik} given by (15) is not greater than the probabilistic assignment of this object to k th cluster in the conventional form (9). Additionally, p_{ik} in (15) is equal to 1 if and only if $\gamma_i = c_k$ and the object Γ_i is isolated.

In the case of optimization of centres of clusters when probabilistic assignments are given, we use the target function with the centres as its arguments, i.e.:

$$f(c_1, c_2, \dots, c_s) = \sum_{k=1}^s \sum_{i=1}^n p_{ik}^2 \sum_{h=1}^n w_{ih}^* d(\gamma_h, c_k). \quad (16)$$

Hence, for $k \in \{1, 2, \dots, s\}$

$$\frac{\partial f(c_1, c_2, \dots, c_s)}{\partial c_k} = \sum_{i=1}^n p_{ik}^2 \sum_{h=1}^n w_{ih}^* \frac{\partial d(\gamma_h, c_k)}{\partial c_k}.$$

If d is the Euclidean distance and $\|\cdot\|$ denotes the Euclidean norm, then

$$\frac{\partial f(c_1, c_2, \dots, c_s)}{\partial c_k} = - \sum_{i=1}^n p_{ik}^2 \sum_{h=1}^n w_{ih}^* \frac{\gamma_h - c_k}{\|\gamma_h - c_k\|}. \quad (17)$$

Assuming that $c_k \notin \{\gamma_1, \gamma_2, \dots, \gamma_n\}$ and taking the fact that the right-hand side of (18) should be zero, we obtain

$$c_k = \sum_{h=1}^n \eta_{hk} \gamma_h, \quad (18)$$

where

$$\eta_{hk} = \frac{\sum_{i=1}^n p_{ik}^2 w_{ih}^* \frac{1}{\|\gamma_h - c_k\|}}{\sum_{q=1}^n \sum_{i=1}^n p_{ik}^2 w_{iq}^* \frac{1}{\|\gamma_q - c_k\|}}, \quad (19)$$

$$h = 1, 2, \dots, n, k = 1, 2, \dots, s.$$

If in functions (14) and (16) the squares of probabilities are replaced with its any normalized power of exponent $u \geq 1$, i.e. with $p_{ik}^{(u)} = p_{ik}^u / \sum_{k=1}^s p_{ik}^u$, then formula (15) will be of the form

$$p_{ik}^{(u)} = \frac{\prod_{l=1, l \neq k}^s (\sum_{h=1}^n w_{ih}^* d_{hl})^u}{\sum_{q=1}^s \prod_{l=1, l \neq q}^s (\sum_{h=1}^n w_{ih}^* d_{hl})^u}, \text{ and formula (19) will be modified as } \eta_{hk} = \frac{\sum_{i=1}^n p_{ik}^{(u)} w_{ih}^* \frac{1}{\|\gamma_h - c_k\|}}{\sum_{q=1}^n \sum_{i=1}^n p_{ik}^{(u)} w_{iq}^* \frac{1}{\|\gamma_q - c_k\|}}, i = 1, 2, \dots, n, k = 1, 2, \dots, s.$$

Formula (19) can be a basis for iteration of centroids of clusters. The equalities (18) and (19) induce the mappings

$$T_k : c \rightarrow T_k(c),$$

where $T_k(c) := \sum_{i=1}^n \tilde{\eta}_{ik} \gamma_i$, and

$$\tilde{\eta}_{ik} = \frac{\sum_{h=1}^n p_{hk}^{(u)} w_{hi}^* \frac{1}{\|\gamma_i - c\|}}{\sum_{q=1}^n \sum_{h=1}^n p_{hk}^{(u)} w_{hq}^* \frac{1}{\|\gamma_q - c\|}} \quad (20)$$

for $c \notin \{\gamma_1, \gamma_2, \dots, \gamma_n\}$ and, by continuity, $T_k(\gamma_i) = \gamma_i$ for every $i = 1, 2, \dots, n$.

Gradient (17) is undefined if c_k is one of the points $\gamma_1, \gamma_2, \dots, \gamma_n$. To have a universal gradient, we will modify the definition given by Kuhn (1967, 1973) in the following way:

$$\frac{\partial^* f(c_1, c_2, \dots, c_s)}{\partial c_k} = \begin{cases} -R_k(c_k) & \text{if } c_k \notin \{\gamma_1, \gamma_2, \dots, \gamma_n\}, \\ R_k(\gamma_i), & \text{if } c_k = \gamma_i, \end{cases} \quad (21)$$

where

$$R_k(c_k) = \sum_{i=1}^n p_{ik}^{(u)} \sum_{h=1}^n w_{ih}^* \frac{\gamma_h - c_k}{\|\gamma_h - c_k\|}, \quad R_k(\gamma_i) = \max \left\{ \|R_{ik}\| - \sum_{h=1}^n p_{hk}^{(u)} w_{hi}^*, 0 \right\} \frac{R_{ik}}{\|R_{ik}\|} \text{ and} \\ R_{ik} = \sum_{q=1}^n p_{qk}^{(u)} \sum_{\substack{h=1, \\ h \neq i}}^n w_{qh}^* \frac{\gamma_h - \gamma_i}{\|\gamma_h - \gamma_i\|}, \quad i = 1, 2, \dots, n, k = 1, 2, \dots, s.$$

Formula (21) is equivalent to standard form (e.g. to (17) when $u=2$) if centroid does not coincide with any data point. Otherwise, the gradient is based on the prior expression where the sum is taken only by points different than point coinciding with the centroid. Such “trimmed” sum is next normalized by its length. The final value of gradient is a product of such value and difference between the aforementioned length and sum of probabilities of assignment of neighbours of the object represented by a point identical with centroid to a given cluster if such difference is positive or zero otherwise.

In the conventional case (cf. Iyigun and Ben-Israel (2013)), minimization of (7) is a probabilistic approximation of the problem of minimization of the function

$$g(c_1, c_2, \dots, c_s) = \sum_{k=1}^s \sum_{i: \Gamma_i \in \Theta_k} \omega_i d(\gamma_i, c_k), \quad (22)$$

where $\omega_i, i=1, 2, \dots, n$ are some weights of objects $\Gamma_1, \Gamma_2, \dots, \Gamma_n$, respectively. Thus, the minimal value of (22), denoted here by μ , is obtained by assigning the object Γ_i to a cluster with its nearest centre:

$$\mu(c_1, c_2, \dots, c_s) = \sum_{i=1}^n \omega_i \min_{k=1,2,\dots,s} d(\gamma_i, c_k). \quad (23)$$

In the analyzed situation when the neighbourhood matrix $\mathbf{W}^*$ is used, such minimization is an extension of (23) when such additional information is introduced. That is, we minimize the function

$$g_{\mathbf{W}^*}(c_1, c_2, \dots, c_s) = \sum_{k=1}^s \sum_{i: \Gamma_i \in \Theta_k} \sum_{h=1}^n w_{ih}^* d(\gamma_h, c_k)$$

and obtain its minimal value, $\mu_{\mathbf{W}^*}(c_1, c_2, \dots, c_s)$, in the following form:

$$\mu_{\mathbf{W}^*}(c_1, c_2, \dots, c_s) = \frac{1}{n} \sum_{i=1}^n \sum_{h=1}^n w_{ih}^* \min_{k=1,2,\dots,s} d(\gamma_h, c_k).$$

Fig. 1 Basic scheme of probabilistic d -clustering with contiguity constraint. Remark: C_1 and C_2 denote centroids of clusters. Source: Own elaboration based on Example 1.

Assume that the distance $d(\cdot, \cdot)$ is the convex function with respect to every argument. Thus, $\min_{k=1,2,\dots,s} d(\cdot, c_k)$ is also the convex function.⁵ Hence, from Jensen's inequality, we obtain:

$$\begin{aligned} \mu_{W^*}(c_1, c_2, \dots, c_s) &\geq \frac{1}{n} \sum_{i=1}^n \left(\sum_{h=1}^n w_{ih}^* \right) \min_{k=1,2,\dots,s} d \left(\frac{\sum_{h=1}^n w_{ih}^* \gamma_h}{\sum_{h=1}^n w_{ih}^*}, c_k \right) = \\ &= \frac{1}{n} \sum_{i=1}^n (a_i \xi_i + 1) \min_{k=1,2,\dots,s} d \left(\frac{\sum_{h=1}^n w_{ih}^* \gamma_h}{\sum_{h=1}^n w_{ih}^*}, c_k \right), \end{aligned} \quad (24)$$

where ξ_i denotes the number of neighbours of the object Γ_i , a_i —some constant depending on the way of normalization of the neighbourhood matrix. If $\mathbf{W}$ is not normalized, then $a_i = 1$. If the normalization is conducted using maximal absolute eigenvalue τ of this matrix, then $a_i = 1/\tau$. Finally, if $\mathbf{W}$ is row standardized, then $a_i = 1/\sum_{h=1}^n w_{ih}$, $i = 1, 2, \dots, n$.

So, the right-hand side of the inequality (24) is the value of the function μ , where $\omega_i = (a_i \xi_i + 1)/n$, and γ_i was replaced with a centroid of the set of points consisting of Γ_i and its neighbours, $i = 1, 2, \dots, n$.

On the other hand, using the Schwarz inequality, we have

⁵ See e.g. EECS (2014).

$$\begin{aligned}
\mu_{W^*}(c_1, c_2, \dots, c_s) &\leq \frac{1}{n} \sum_{i=1}^n \sqrt{\left(\sum_{h=1}^n w_{ih}^{*2} \left(\sum_{h=1, \substack{h=i \vee \Gamma_h \bowtie \Gamma_i}}^n \left(\min_{k=1,2,\dots,s} d(\gamma_h, c_k) \right)^2 \right) \right)} \\
&\leq \frac{1}{n} \sum_{i=1}^n \sqrt{(a_i^2 \xi_i + 1)} \sum_{h=1, \substack{h=i \vee \Gamma_h \bowtie \Gamma_i}}^n \min_{k=1,2,\dots,s} d(\gamma_h, c_k) \\
&= \frac{1}{n} \sum_{i=1}^n \sqrt{(a_i^2 \xi_i + 1)} \sum_{h=1}^n w_{ih}^* \min_{k=1,2,\dots,s} d(\gamma_h, c_k).
\end{aligned} \tag{25}$$

The last equality follows from the fact that summation performed by neighbours of Γ_i (and including Γ_i) is equivalent to the summation made by these h for which $w_{ih}^* = 1$ and—because $w_{ih}^* = 0$ if $i \neq h$ and Γ_i is not a neighbour of Γ_h —to the summation by all w_{ih}^* .

In this case, the right-hand side of this inequality is the value of the function μ with $\omega_i = \frac{1}{n} \sqrt{(a_i^2 \xi_i + 1)}$, increased by relevant value for neighbours of the object Γ_i , i.e. $\mu(c_1, c_2, \dots, c_s) + \frac{1}{n} \sum_{i=1}^n \sqrt{(a_i^2 \xi_i + 1)} \sum_{h=1, \substack{h=i \vee \Gamma_h \bowtie \Gamma_i}}^n \min_{k=1,2,\dots,s} d(\gamma_h, c_k)$.

The inequalities (24) and (25) present lower and upper bounds of minimum of the target function with respect to the number of neighbours of objects and minimal distances of relevant points (centroids of sets consisting of objects and its neighbours in the former and simple objects in the latter case) from the nearest centroid.

Assume that the distance $d(\cdot, \cdot)$ between points is calculated using the standard Euclidean formula. Similarly, as in the conventional case, the following theorem holds.

Theorem 1 a) For every set of points $c_1, c_2, \dots, c_s \in \mathbb{R}^m$, the condition

$$R_k(c_k) = \mathbf{0}$$

for every $k = 1, 2, \dots, s$ is sufficient and necessary for the points $c_1, c_2, \dots, c_s$ to minimize the function f defined by (16).

- b) The optimal centres of clusters $c_1^*, c_2^*, \dots, c_s^*$ are in the convex hull of the points $\gamma_1, \gamma_2, \dots, \gamma_n$.
- c) If c is the optimal centre of k th cluster, then it is the fixed point of the mapping T_k , i.e. $T_k(c) = c$ for some $k \in \{1, 2, \dots, s\}$.

This theorem shows that the version of probabilistic d -clustering with use of the neighbourhood matrix is in some sense a generalization of the conventional version of such approach described in Section 3. The proof of Theorem 1 is presented in Appendix 1.

The clustering is conducted according to the procedure of similar scheme as in the case of attempts described in subsections 2.2 and 2.3, but using Algorithm 3 and some properties of

this proposed modification based on the neighbourhood of objects. This procedure is described in Algorithm 4.

Algorithm 4.

Part I

Step 1. Let $\Theta_{(NB)1}^{(0)}, \Theta_{(NB)2}^{(0)}, \dots, \Theta_{(NB)s_{NB}}^{(0)}$ are starting clusters and $\mathbf{c}_{(NB)1}^{(0)}, \mathbf{c}_{(NB)2}^{(0)}, \dots, \mathbf{c}_{(NB)s_{NB}}^{(0)}$ — their centres. Determine the preliminary clusters $\Theta_{(NB)1}^{(1)}, \Theta_{(NB)2}^{(1)}, \dots, \Theta_{(NB)s_{NB}}^{(1)}$ such that $\Gamma_{(NB)i}, \Gamma_{(NB)k} \in \Theta_{(NB)l}^{(1)}$ if $\Gamma_{(NB)i}$ and $\Gamma_{(NB)k}$ are neighbours and the smaller probability of assignment of $\Gamma_{(NB)i}$ or $\Gamma_{(NB)k}$ to $\Theta_{(NB)l}^{(0)}$ is the greatest among all $\Theta_{(NB)l}^{(0)}, l = 1, 2, \dots, s_{NB}$. The probabilities of assignment are determined by (15). Next, the centres of preliminary clusters $\mathbf{c}_{(NB)1}^{(1)}, \mathbf{c}_{(NB)2}^{(1)}, \dots, \mathbf{c}_{(NB)s_{NB}}^{(1)}$ are computed by (18).

Step r -th ($r = 2, 3, \dots$). Determine clusters $\Theta_{(NB)1}^{(r)}, \Theta_{(NB)2}^{(r)}, \dots, \Theta_{(NB)s_{NB}}^{(r)}$ according to the rule:

$$\Gamma_{(NB)i}, \Gamma_{(NB)k} \in \Theta_{(NB)l}^{(r)} \text{ if } \Gamma_{(NB)i} \bowtie \Gamma_{(NB)k} \text{ and}$$

$$\begin{aligned} & \min \left\{ p \left(\Gamma_{(NB)i}, \Theta_{(NB)l}^{(r-1)} \right), p \left(\Gamma_{(NB)k}, \Theta_{(NB)l}^{(r-1)} \right) \right\} \\ & = \max_{l=1,2,\dots,s_{NB}} \min \left\{ p \left(\Gamma_{(NB)i}, \Theta_{(NB)l}^{(r-1)} \right), p \left(\Gamma_{(NB)k}, \Theta_{(NB)l}^{(r-1)} \right) \right\}, \end{aligned}$$

for every $i, k \in \{1, 2, \dots, n_{NB}\}$ (where probabilities $p \left(\Gamma_{(NB)i}, \Theta_{(NB)l}^{(r-1)} \right), i = 1, 2, \dots, n, l = 1, 2, \dots, s_{NB}$ are also computed by (15)) and determine centres of such clusters $\mathbf{c}_{(NB)1}^{(r)}, \mathbf{c}_{(NB)2}^{(r)}, \dots, \mathbf{c}_{(NB)s_{NB}}^{(r)}$ using (18), where in (19) $\mathbf{c}_k := \mathbf{c}_{(NB)k}^{(r-1)}$. If $\sum_{k=1}^{s_{NB}} d \left(\mathbf{c}_{(NB)k}^{(r)}, \mathbf{c}_{(NB)k}^{(r-1)} \right) < \varepsilon$ where $\varepsilon > 0$ is arbitrarily established level of precision then stop else go to step $r + 1$.

Part II

Step 1. Create new set of objects consisting of the clusters of neighbours determined in Part I and objects having no neighbours as singletons (cf. Step 1 of Part II of Algorithm

Step r -th ($r = 2, 3, \dots$): Compute centroids of the clusters established in Step 1 and conduct Algorithm in relation to them.

The relation expressed in r -th step of Part I of Algorithm 4 is the equivalence relation. One can prove it in a quite similar way as in the case of the relation used in the r -th step of Part I of Algorithm 1.

A defuzzification of clusters is similar as applied in Algorithm 3, i.e. each object (group of neighbouring objects) is assigned to such group for which the probability of assignment is the greatest. Of course, at each step of Algorithm 4, a situation when for the same object(s) two or more clusters are nearest according to the computed probability of assignment (i.e. for these clusters the probabilities are equal and maximal) is possible. Also in this case, the object(s) is (are) assigned only to one of these clusters.

The basic draft scheme of the idea of Algorithm 4 is presented in Fig. 1. The objects and the neighbourhood matrix are the same as in Example 1.

◀ **Fig. 2** Distributions of values of correctness and silhouette indices obtained in particular replications and best adjusted normal curves—balanced subsample sizes, no noise added. Source: Own computations using special algorithm written in the IML environment of the SAS Enterprise Guide 4.3. software.

The clustering is conducted using some diagnostic features different than those used to define the pairs of neighbours. The dashed lines denote the scope of current clusters.

5 Simulation Study

We will perform now an investigation of practical differences and similarities between analyzed clustering algorithms with contiguity constraint. Our study is based on simulated data combined from random samples produced using some types of multivariate normal distribution. That is, we have created a data set in $\mathbb{R}^7$ with $n=100$ objects. The data were simulated from four types of multivariate distribution $\mathcal{N}(\mu_k, \Sigma_k)$, $k=1, 2, 3, 4$, where

$$\begin{aligned} \mu_1 &= (2.5, -1.5, 3, 4, 0, 0, 0), \quad \Sigma_1 = \\ &\begin{bmatrix} 1 & 0 & 0 & 0 & -0.67 & -0.21 & -0.67 \\ 0 & 1 & 0 & 0 & -0.56 & 0.21 & 0.33 \\ 0 & 0 & 1 & 0 & -0.33 & -0.62 & 0.51 \\ 0 & 0 & 0 & 1 & -0.33 & 0.54 & 0.33 \\ -0.67 & -0.56 & -0.33 & -0.33 & 1 & 0 & 0 \\ -0.21 & 0.21 & -0.62 & 0.54 & 0 & 1 & 0 \\ -0.67 & 0.33 & 0.51 & 0.33 & 0 & 0 & 1 \end{bmatrix} \\ \mu_2 &= (-1.5, 4, 2, -3.5, 0, 0, 0), \quad \Sigma_2 = \\ &\begin{bmatrix} 1 & 0 & 0 & 0 & -0.66 & -0.21 & -0.67 \\ 0 & 1.3 & 0 & 0 & -0.67 & 0.34 & 0.31 \\ 0 & 0 & 1 & 0 & -0.33 & -0.62 & 0.51 \\ 0 & 0 & 0 & 1 & -0.33 & 0.54 & 0.33 \\ -0.66 & -0.67 & -0.33 & -0.33 & 1 & 0 & 0 \\ -0.21 & 0.34 & -0.62 & 0.54 & 0 & 1 & 0 \\ -0.67 & 0.31 & 0.51 & 0.33 & 0 & 0 & 1 \end{bmatrix} \\ \mu_3 &= (-3, -2, 4, 1.5, 0, 0, 0), \quad \Sigma_3 = \\ &\begin{bmatrix} 1 & 0 & 0 & 0 & -0.66 & -0.21 & -0.67 \\ 0 & 1 & 0 & 0 & -0.51 & 0.21 & 0.33 \\ 0 & 0 & 1.3 & 0 & -0.49 & -0.55 & 0.52 \\ 0 & 0 & 0 & 1 & -0.33 & 0.54 & 0.33 \\ -0.66 & -0.51 & -0.49 & -0.33 & 1 & 0 & 0 \\ -0.21 & 0.21 & -0.55 & 0.54 & 0 & 1 & 0 \\ -0.67 & 0.33 & 0.52 & 0.33 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

Thus, the analyzed data set was combined from four independent subsamples sampled from each of these distributions. This way, each object is described by seven ($m=7$) artificial variables. The choice of parameters of such distributions was motivated by several premises. Firstly, we have assumed that some of such dimensions are diagnostic variables being the basis of clustering (i.e. they can be regarded as describing some composite phenomenon) and the rest of them illustrate some associated—but different—field which is used to define the neighbourhood of analyzed objects. In our case, the first four dimensions reflect four

Table 1 Share of clusterings for which correctness and silhouette indices assume favourable and unfavourable values (in %)

Method	Used data		$cr > 1$ and	$cr < 1$ and	$cr > 1$ and	$cr < 1$ and
			$SI > 0$	$SI < 0$	$SI < 0$	$SI > 0$
k -means	Balanced	No noise	7.4	12.6	3.6	76.4
		Noise U(− 0.5,0.5)	7.8	12.3	2.0	77.9
		Noise U(− 2,2)	8.0	12.5	2.7	76.8
	Unbalanced	No noise	8.3	12.7	2.2	76.8
		Noise U(− 0.5,0.5)	8.2	10.7	2.2	78.9
		Noise U(− 2,2)	8.1	10.6	2.4	78.9
Ward	Balanced	No noise	54.7	4.1	9.4	31.8
		Noise U(− 0.5,0.5)	58.2	3.9	9.8	28.1
		Noise U(− 2,2)	53.3	4.2	7.2	35.3
	Unbalanced	No noise	52.9	3.7	8.7	34.7
		Noise U(− 0.5,0.5)	52.6	4.3	9.2	33.9
		Noise U(− 2,2)	50.5	4.2	6.5	38.8
d -clustering	Balanced	No noise	6.0	31.4	5.4	57.2
		Noise U(− 0.5,0.5)	5.4	33.9	6.2	54.5
		Noise U(− 2,2)	5.7	36.0	6.5	51.8
	Unbalanced	No noise	4.8	35.5	5.3	54.4
		Noise U(− 0.5,0.5)	5.3	35.8	4.0	54.9
		Noise U(− 2,2)	5.6	35.8	5.4	53.2

Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment)

diagnostic variables and the remaining three supporting variables used to determine the contiguity matrix. Hence, the variables from the former group are assumed to be correlated with variables belonging to the latter. Secondly, to ensure the sufficient variety of sampled data, the covariance matrices have high diagonal entries and zeros in the blocks restricted to only variables used to clustering. The same premise was a reason for defining the blocks for the variables used to determination of neighbourhoods as identity matrices. Moreover, the entries connected with the relation between these two types of variables are of high absolute value what enables to reflect the correlation between them.

Another important question is the number of objects sampled from each of the four aforementioned distributions. Of course, the easiest solution in this respect is to set the equal size of any subsample, i.e. 25. Such approach seems to be, however, slightly simplistic and hence, we have investigated also alternative choice of sizes of four subsamples. The new sizes were unbalanced because they were sampled from the multinomial distribution with $n = 100$ number of trials and equal event probabilities ($p_1 = p_2 = p_3 = p_4 = 0.25$).

Moreover, to assess how the analyzed algorithms are sensitive to disturbances, we have added to the simulated data two types of noise being generated from the uniform distribution on $[-0.5, 0.5]$ (small noise) and $[-2, 2]$ (intensive noise). These assumptions resulted in the production of six following data sets:

- Set 1: balanced subsample sizes, no noise added
- Set 2: balanced subsample sizes, added uniform noise from $[-0.5, 0.5]$
- Set 3: balanced subsample sizes, added uniform noise from $[-2, 2]$
- Set 4: unbalanced subsample sizes, no noise added
- Set 5: unbalanced subsample sizes, added uniform noise from $[-0.5, 0.5]$

- Set 6: unbalanced subsample sizes, added uniform noise from $[-2, 2]$.

The contiguity (neighbourhood) matrix was determined on the basis of first (lower) quartile of distances between objects in terms of the auxiliary variables (i.e. dimensions from 5 to 7 of the input data). That is, two objects are regarded to be neighbours if the Euclidean distance between two vectors representing these objects is not greater than $0.5q_1(d)$, where $q_1(d)$ is the first quartile of such distances across all pairs of objects (i.e. $q_1(d)$ is the lower quartile of $\{d(\gamma_h, \gamma_i), h, i = 1, 2, \dots, n, h < i\}$). Therefore, the neighbourhood information is incorporated into the model. Moreover, this choice ensures a rational proportion between the number of pairs of objects being neighbours and the number of pairs of non-neighbours.

The input data were normalized using the zero-unitarization formula, i.e. we have taken

$$z_{ij} = \frac{x_{ij} - \min_{k=1,2,\dots,n} x_{kj}}{\max_{k=1,2,\dots,n} x_{kj} - \min_{k=1,2,\dots,n} x_{kj}},$$

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

The clustering of analyzed objects was based on these normalized data and was constrained by the neighbourhood in the earlier described way. Three investigated methods of clustering with contiguity constraint were here used, i.e.:

- k -means algorithm (Algorithm 1), denoted by K
- Ward method (Algorithm 2), denoted by W
- Probabilistic d -clustering (Algorithm 4), denoted by S

According to the assumptions of the experiment, the number of clusters was arbitrarily assumed as $s = 4$. This choice is motivated by the fact that the data were simulated from four different distributions what defines four types of information. Therefore, one would expect that these predefined types were reflected by the obtained clusters. The initial centres of clusters are chosen randomly from the four aforementioned distributions (i.e. taking only their dimensions 1–4 into account). The level of precision (ε) below which the iteration is terminated was assumed as $\varepsilon = 0.001$ in any case.

The experiment was repeated 1000 times. The fundamental measure of efficiency and quality of clustering for each replication was the index of correctness defined by formula (3) as the ratio of indices of homogeneity (1) and heterogeneity (2). Alternatively, the silhouette index (4) was also computed. Of course, for each replication, the quality assessment concerns the optimal clustering of objects based on data generated in such replication. Thus, we obtain 1000 values of each of these two indices. All computations were conducted using the SAS Enterprise Guide 4.3. software.

Figure 2 shows the distributions of the index of correctness and silhouette index optimal clustering in each of 1000 replications of the experiment for balanced subsample sizes without a noise. Basic descriptive statistics of these distributions are also presented. Moreover, the best fitted normal curves are overlaid. The relevant figures for the remaining cases (i.e. balanced sizes with the $U(-0.5, 0.5)$ noise, balanced sizes with the $U(-2, 2)$ noise, unbalanced sizes with no noise, unbalanced sizes with the $U(-0.5, 0.5)$ noise and unbalanced sizes with the $U(-2, 2)$ noise) are in fact practically the same. The normal curve is placed here to show how the whole shape of empirical distributions (of which kurtosis, variation, tails, etc.) of analyzed indices is close to (or far from) the normal one. We can observe that the values of correctness index (3) are distributed much more close to the normal curve (around 1 or a value slightly smaller

Table 2 Basic descriptive statistics of distributions of similarity indices related to results of *k*-means and *d*-clustering algorithms

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)			U(− 2,2)	
Rand						
Mean	0.6499	0.6355	0.6236	0.6409	0.6509	0.6327
Maximum	1.0000	1.0000	1.0000	1.0000	1.0000	0.9978
Quantile 99%	0.9627	0.9807	0.9807	0.9795	0.9800	0.9707
Quantile 95%	0.9110	0.9230	0.9103	0.9077	0.9233	0.8990
Quantile 90%	0.8615	0.8583	0.8446	0.8600	0.8753	0.8514
Quantile 75% (Q3)	0.7549	0.7367	0.7044	0.7510	0.7555	0.7249
Quantile 50% (median)	0.6198	0.5958	0.5901	0.5990	0.6113	0.6006
Quantile 25% (Q1)	0.5412	0.5291	0.5246	0.5288	0.5380	0.5341
Quantile 10%	0.4915	0.4831	0.4745	0.4847	0.4949	0.4805
Quantile 5%	0.4624	0.4483	0.4382	0.4439	0.4604	0.4369
Quantile 1%	0.3851	0.3821	0.3746	0.3926	0.3835	0.3870
Minimum	0.3558	0.3081	0.3246	0.3402	0.3194	0.3184
Variance	0.0199	0.0208	0.0195	0.0213	0.0212	0.0193
Standard deviation	0.1411	0.1441	0.1397	0.1458	0.1457	0.1389
Skewness	0.4556	0.6369	0.7520	0.5671	0.5319	0.5574
Coefficient of variation (%)	21.7158	22.6764	22.3943	22.7556	22.3931	21.9533
Kurtosis	− 0.6560	− 0.4396	− 0.0690	− 0.6284	− 0.6333	− 0.3509
Peirce						
Mean	0.2829	0.2718	0.2582	0.2769	0.2962	0.2634
Maximum	1.0000	1.0000	1.0000	1.0000	1.0000	0.9886
Quantile 99%	0.8820	0.9645	0.9489	0.9527	0.9636	0.9049
Quantile 95%	0.7145	0.7163	0.6873	0.7155	0.7284	0.6669
Quantile 90%	0.5945	0.6013	0.5625	0.5992	0.6187	0.5521
Quantile 75% (Q3)	0.3920	0.3773	0.3450	0.3894	0.4126	0.3563
Quantile 50% (median)	0.2312	0.2117	0.2023	0.2226	0.2371	0.2196
Quantile 25% (Q1)	0.1276	0.1155	0.1200	0.1166	0.1390	0.1224
Quantile 10%	0.0667	0.0501	0.0542	0.0474	0.0687	0.0549
Quantile 5%	0.0289	0.0225	0.0208	0.0155	0.0369	0.0187
Quantile 1%	− 0.0235	− 0.0216	− 0.0254	− 0.0483	− 0.0216	− 0.0313
Minimum	− 0.0850	− 0.1183	− 0.1008	− 0.1081	− 0.1279	− 0.1117
Variance	0.0439	0.0470	0.0421	0.0477	0.0484	0.0399
Standard deviation	0.2094	0.2167	0.2053	0.2184	0.2199	0.1998
Skewness	1.0439	1.1476	1.3226	1.0756	1.0291	1.1522
Coefficient of variation (%)	74.0249	79.7313	79.5010	78.8646	74.2493	75.8517
Kurtosis	0.6972	0.9057	1.6591	0.8568	0.6412	1.2652
Sokal and Sneath						
Mean	0.6618	0.6565	0.6487	0.6583	0.6686	0.6529
Maximum	1.0000	1.0000	1.0000	1.0000	1.0000	0.9953
Quantile 99%	0.9491	0.9714	0.9700	0.9748	0.9724	0.9572
Quantile 95%	0.8765	0.8773	0.8688	0.8845	0.8838	0.8545
Quantile 90%	0.8189	0.8177	0.8065	0.8167	0.8361	0.8032
Quantile 75% (Q3)	0.7260	0.7172	0.6953	0.7227	0.7277	0.7108
Quantile 50% (median)	0.6407	0.6351	0.6276	0.6387	0.6474	0.6349
Quantile 25% (Q1)	0.5839	0.5752	0.5745	0.5782	0.5882	0.5770
Quantile 10%	0.5414	0.5334	0.5354	0.5308	0.5450	0.5354
Quantile 5%	0.5201	0.5169	0.5132	0.5090	0.5198	0.5143
Quantile 1%	0.4790	0.4783	0.4790	0.4647	0.4863	0.4779
Minimum	0.4485	0.4392	0.4304	0.4383	0.4435	0.4381
Variance	0.0114	0.0122	0.0111	0.0123	0.0123	0.0108
Standard deviation	0.1067	0.1103	0.1052	0.1110	0.1109	0.1041

Table 2 (continued)

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)	U(− 2,2)		U(− 0.5,0.5)	U(− 2,2)
Skewness	0.7649	0.8583	1.0114	0.7628	0.7842	0.8040
Coefficient of variation (%)	16.1168	16.8047	16.2104	16.8649	16.5915	15.9483
Kurtosis	0.1914	0.3268	0.8862	0.3011	0.1982	0.4300

Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment)

than 1) than the values of the silhouette index (4) which distribution is much more irregular and asymmetric.

Table 3 Values of diagnostic features of labour market in Poland in 2016

Powiat	X_1	X_2	X_3	X_4	X_5
chodzieski	10.4	14.8	15.7	37.9	3532.53
czarnkowsko-trzcianecki	7.0	14.4	20.1	31.1	3699.41
gnieźniński	7.8	17.1	17.8	16.1	3528.01
gostyński	7.3	13.8	17.9	32.6	3557.77
grodziski	5.1	20.3	19.4	37.7	3006.21
jarociński	5.8	22.3	22.3	12.7	3066.45
kaliski	4.2	13.8	14.9	14.5	3145.60
kępiski	2.2	26.9	26.6	7.8	2658.47
kolski	10.1	18.7	16.8	31.5	3808.90
koniński	14.2	9.8	12.0	142.6	3335.85
kościański	4.9	20.9	18.9	6.1	3491.93
krotoszyński	4.8	19.3	23.7	9.7	3202.47
leszczyński	4.3	14.0	16.8	23.6	3924.99
międzychodzki	5.5	17.2	23.7	25.7	3406.58
nowotomyski	3.1	22.6	20.2	3.9	3768.93
obornicki	6.1	16.5	11.2	14.4	3867.54
ostrowski	4.7	14.3	17.5	9.2	3576.57
ostrzeszowski	5.7	15.4	20.6	11.1	3509.40
piłski	6.9	21.8	20.0	13.4	3772.08
pleszewski	5.5	34.9	35.3	51.2	3399.21
poznański	2.3	13.6	11.0	7.4	3855.60
rawicki	5.9	16.7	24.1	18.0	3182.07
śłupecki	9.6	15.5	21.9	26.3	3317.41
szamotulski	4.7	15.8	11.8	6.6	4196.53
średzki	9.5	12.8	11.4	18.2	3785.23
śremski	4.0	20.3	15.4	18.5	3411.22
turecki	5.3	18.1	19.6	21.2	3167.19
wągrowiecki	8.7	17.2	20.3	39.8	3488.88
wolsztyński	2.6	25.3	16.3	7.0	3315.96
wrzeński	6.7	23.5	20.0	20.6	3607.50
złotowski	10.1	14.4	15.2	20.2	3221.67
Kalisz	4.6	14.1	14.9	3.2	3648.00
Konin	10.3	10.1	11.0	95.8	3785.02
Leszno	4.8	11.5	13.3	7.3	3406.02
Poznań	1.9	15.8	11.7	9.2	4770.94

Source: Own computations based on the resources of the Local Data Bank of the Statistics Poland (<http://bdl.stat.gov.pl>)

Table 4 Summary results for optimal clusters of powiats of the Wielkopolskie Voivodship

Method	Number of clusters	Index of		
		Homogeneity	Heterogeneity	Correctness
K	6	154.8827	328.8102	0.4710
W	2	421.1549	737.9376	0.5707
S	5	121.0170	313.2308	0.3864

Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment)

In general, one can observe that the probabilistic d -clustering provides, in general, better results than the Ward approach. The quartiles, mean, median and maximum of the cr index are clearly smaller for S than for W. Also skewness and kurtosis for S is less intensive. Between S and the k -means clustering (K), there is, in this context, greater similarity. It is worth noting, however, that S seems to be slightly more efficient in the cases where the analyzed index is high (i.e. when some problems with receiving distinguishable clusters occur). That is, the 95% and 99% percentiles and the maximum of both indices are usually greater for K than for S (e.g. 1.0468, 1.1440 and 1.5381 for K and 1.0458, 1.1022 and 1.3498 in the case of the correctness index for unbalanced sizes without a noise).

These values, however, are related to these data for which efficient clustering was especially difficult. Such cases were few. On the other hand, one can observe that the values of the index (3) for the results of d -clustering are least diverse and their distribution was similar to the normal one in terms of kurtosis. It shows that this approach is more stable in terms of quality than other analyzed algorithms. One more property strictly connected with this stability can be here underlined: for several replications in some cases, the k -means method has generated a half or more empty clusters. Thus, such replications had to be omitted in the analysis of final results (the relevant correctness index was extremely large). It is worth noting also that the same problem has occurred in some trials of repetition of the whole experiment, not only for the aforementioned options but also for the remaining ones. Nevertheless, in any case, the d -clustering had never produced so many empty clusters although the number of clusters is here also arbitrarily established.

In connection with these results, the following question arises: are the situations when the analyzed indices take extreme values ($cr > 1$ or $SI < 0$) occur at the same clusterings? Moreover, it would be interesting in how many clusterings the indices provide quite different results (e.g. $cr > 1$ and $SI > 0$). Table 1 contains percentages of clusterings for which such counterintuitive cases as well as quite consistent (i.e. when $cr < 1$ and $SI > 0$) are observed. One can see also that most clusterings obtained using the k -means and probabilistic d -clustering algorithms are of proper quality. However, in the case of the Ward method, for over 50% of clusterings, the correctness index shows that they are bad, whereas the silhouette index suggests that one of these indices their quality is quite satisfactory. It means that one of these indices is inappropriate in this case. Because SI is based on a sum of normalized average dissimilarities of objects within and inter clusters, these differences can reduce one to another and the final result could be then distorted. The fact that the Ward method is hierarchical and therefore generates internal clusters contained in the target ones seems to contribute to such situation. In the correctness index, such situation practically cannot appear and hence it can be perceived as better.

To investigate whether it really reflects the inconsistency of both clusterings, we will use three similarity indices belonging to the so-called α -family (cf. Albatineh (2010)). They are constructed as a function of the following quantities:

Fig. 3 Clusters of powiats of the Wielkopolskie Voivodship according to the condition of labour market obtained using the k -means method with contiguity constraint. Remark: The miniature of map placed in the left bottom corner shows the location of the Wielkopolskie Voivodship in Poland. Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment).

- a —number of pairs of objects belonging to the same clusters according to both analyzed methods
- b —number of pairs of objects belonging to the same cluster according to the first method and to different clusters according to the second approach

Fig. 4 Clusters of powiats of the Wielkopolskie Voivodship according to the condition of labour market obtained using the Ward method with contiguity constraint. Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment).

- c —number of pairs of objects belonging to different clusters according to the first method and to the same cluster according to the second approach
- d —number of pairs of objects belonging to different clusters according to both investigated methods

Fig. 5 Clusters of powiats of the Wielkopolskie Voivodship according to the condition of labour market obtained using probabilistic d -clustering method with contiguity constraint. Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment).

The three chosen options use optimally the information potential of all these binary counts and simultaneously are concentrated on some different—but important—properties of both clusterings⁶:

- Peirce index (Peirce (1884)):

$$PE = \frac{ad-bc}{(a+c)(b+d)},$$

- Rand index (Rand (1971)):

$$RD = \frac{a+d}{a+b+c+d},$$

- Sokal and Sneath index (Sokal and Sneath (1963)):

$$SS = \frac{1}{4} \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right).$$

The Rand and Sokal and Sneath indices take values from $[0,1]$, whereas the Peirce index has the range of $[-1,1]$. In all three cases, the minimum value denotes complete dissimilarity and the maximum—perfect similarity (i.e. identity) of both structures of clusters.

These indices will be used to investigate how the clusters obtained by the three analyzed methods recover the original clear-cut 4-cluster partition. The detailed table containing basic descriptive statistics for distributions of these indices over 1000 replications of the experiment for all its variants is presented in Appendix 2. Of course, due to the contiguity constraint, the level of reflection of original clusters cannot be very close to ideal. However, the situation, when in all variants in the Rand and Sokal and Sneath indices take values greater than 0.5 and Peirce—greater than 0 can be regarded here as satisfactory. It occurs in many clusterings generated by the three algorithms, but only in the case of probabilistic d -clustering approach both for balanced and unbalanced sizes and without and with a noise at least 50% replications of the experiment have given such favourable result. Moreover, maximum values for data with subsamples of balanced sizes are usually greatest in the case of probabilistic d -clustering.

Because the k -means and probabilistic d -clustering seem to be better solutions than the Ward approach, we will analyze now how close clusterings they produce. The basic descriptive statistics of distributions of these indices for the structures of clusters obtained in 1000 replications using k -means and d -clustering methods are collected in Table 2.

As we can see, the results show that in prevailing majority of simulated data, the objects are clustered similarly both if the k -means or d -clustering approach is used. The means and medians of Rand and Sokal and Sneath indices are greater than 0.59 and the Peirce index (which takes values from $[-1,1]$)—than 0.20. The skewness, variation and kurtosis of the indices (especially Rand and Sokal and Sneath) are also not large. Thus, one can be afraid that the silhouette index does often not properly reflect the quality of clustering in the analyzed situation.

⁶ These three indices belong to the so-called $\mathcal{L}$ -family of similarity indices, i.e. each of them is a linear function of sum of squared numbers of objects placed in one cluster according to one clustering and in other cluster according to the second compared clustering. Albatineh et al. (2006), Albatineh (2010) and Albatineh and Niewiadomska – Bugaj M. (2011) showed that any index φ from the $\mathcal{L}$ -family can be corrected as $\tilde{\varphi} \equiv \frac{\varphi - E(\varphi)}{1 - E(\varphi)}$, where $E(\varphi)$ is the expectation of φ conditional upon fixed sets of marginal counts of the matrix which entries are numbers of objects belonging to individual clusters according to compared clusterings. An example of such correction can be the Adjusted Rand Index (cf. Hubert and Arabie (1985)). However, in our studies, we use the conventional form of indices, because they are easier interpretable and comparable in this context.

6 Empirical Application

Now, we will present an example of practical application of the investigated clustering models and their properties in practical circumstances. We will analyze the condition of labour market in powiats (the Polish NUTS 4 regions) of the Wielkopolskie Voivodship in Poland in 2016. The Wielkopolskie Voivodship is the NUTS 2 unit of territorial division. It consists of 35 powiats. It is worth mentioning that four of them are the largest cities of the region (Poznań, Kalisz, Konin and Leszno) having the powiat status.

To realize this task, an initial set of variables describing the composite social and economic phenomenon—i.e. the labour market, was collected.⁷ The main criteria of their choice were data availability and logical and economical connections between them and with the situation on the labour market. All of them have the character of indices, i.e. they are ratios of some quantities expressed in absolute values. Such set was next verified. That is, variables having too small coefficient of variation—what implies negligible power of diversification—and variables too much correlated with others (using the method of reversed correlation matrix) were eliminated. More details about methods and concepts of these verifications can be found in the paper by Młodak (2014). Finally, the following set of diagnostic features has been obtained:

X_1 —registered unemployment rate in %

X_2 —coefficient of outflow from unemployment⁸ (in %)

X_3 —coefficient of inflow to unemployment⁹ (in %)

X_4 —number of unemployed persons per one job offer

X_5 —average monthly wage and salary in the national economy in zlotys (PLN)

The data concerning these variables are presented in Table 3.

Note that the variables X_1 , X_3 and X_4 are destimulants (which means that the higher is the value of the feature, the better is the situation of an object is in this context) whereas the remaining diagnostic features are stimulants (higher values are “better”). Therefore, to uniform performance of all of them, the destimulants were transformed into stimulants by taking their “bad” values with opposite signs.

Three analyzed algorithms of contiguity constrained clustering were used. Neighbourhood is here understood in traditional geographical sense. The main difference of the current exercise in comparison with the optimization methodology applied in the simulation study is that now the number of clusters was established endogenously. That is, from all divisions of the set of objects into clusters possible to obtain by Algorithms 1, 2, and 4, this one for which the correctness index (3) achieved its minimum was regarded to be final. All other computational proceedings were the same as in Section 5 (except for replications, noise addition and sampling, of course).

In Table 4, the summary statistics for optimal divisions into clusters obtained using the three analyzed methods are presented. This table contains the number of received clusters and indices of homogeneity, heterogeneity and correctness of such divisions.

We can observe then that the best results are obtained when the probabilistic d -clustering method was used. The optimal index of correctness takes here the smallest value from the three analyzed options. Therefore, it is the best method in this situation.

⁷ The source of the data was the Local Data Bank of the Central Statistical Office of Poland (<https://bdl.stat.gov.pl>).

⁸ Share of persons removed from the register of unemployment in a given period in the total number of unemployed persons.

⁹ Share of persons newly registered in the register of unemployment in a given period in the total number of unemployed persons.

The constrained Ward approach tends to generate relatively small number of cluster whereas the constrained k -means algorithm produces six clusters but less homogeneous.

As regards similarity of structures of clusters, the relevant indices Rand, Peirce and Sokal and Sneath (which amounted to 0.9950, 0.9848 and 0.9942, respectively) show that the results of the constrained k -means and probabilistic d -clustering methods are very close. The effects obtained by k -means and Ward approaches with contiguity constraint are slightly more distinct (the values of discussed indices are 0.7916, 0.7523 and 0.7739, respectively). The difference is, however, most considerable in the case of the constrained k -means and probabilistic d -clustering algorithms where the similarity indices taken the values 0.7933, 0.3451 and 0.7748, respectively. These conclusions coincide with regularities observed in the simulation study (Section 5).

On the other hand, if we compute the silhouette index, we can see that its value is very low (-0.8591 for K, -0.1267 for W and -0.4059 for S). This discrepancy (especially in comparison with the correctness index which signs good quality of clustering) is a premise to be afraid that the silhouette index could be inappropriate in this case.

Figures 3, 4 and 5 present territorial arrangement of individual clusters obtained by three analyzed methods. The constrained k -means method produces 6 clusters. The most numerous group of them (i.e. 1) consists of 29 powiats. The cluster 6 is created by two powiats located in the north part of the region, where the inflow of unemployed persons and their number per one job offer is especially high. The rest of powiats are singletons. They are far from their neighbours in the analyzed context and their labour markets have some specific feature distinguishing them from the others. For instance, the ostrzeszowski powiat is characterized by relatively low unemployment rate but also by high inflow of unemployed persons and high average monthly wage and salary.

The results of the probabilistic d -clustering taking neighbourhood into account are similar to those obtained by relevantly modified k -means approach. The main difference between them is that now the złotowski powiat is not a singleton but belongs to the cluster 3 together with the międzychodzki powiat. There exists also other two-element cluster consisting of the wągrowiecki powiat which was included by the constrained k -means algorithm in one cluster with the międzychodzki powiat with the średzki powiat (being a singleton in the constrained k -means clustering). The śremski powiat creates a single cluster both in k -means and probabilistic d -clustering with contiguity constraint. Also the greatest cluster of 29 remaining powiats is in both cases the same.

7 Concluding Remarks

The experiment and studies conducted in this paper showed that the contiguity constraint imposed on the k -means, Ward and probabilistic d -clustering methods offers the opportunity to construct efficient algorithms, where the neighbourhood plays an important role.

In the case when the final number of clusters is fixed (as in the simulation study), all these methods are, in general, robust both to unbalanced sizes of clusters and to noisy data. However, Ward's approach produces in any case worst results in comparison with the two remaining ones. On the other hand, the results of d -clustering are least diverse in replications and never contain a half or more of k empty cluster (whereas the k -means algorithm can sometimes produce such undesirable solutions).

Another question which sometimes can have an importance is taking the cardinalities of generated clusters in the algorithm into account. Although in used constraints no

special conditions of that type are imposed (cf. Wagstaff et al. (2001)),¹⁰ sometimes a special situation may occur, when e.g. in the course of iteration an empty cluster occurs and in consequence the final number of non-trivial clusters will be smaller than k . If such inconvenience is—due to some practical reasons—harmful, it can be avoided e.g. by imposing a minimum size of cluster and using method proposed by Bradley et al. (2000).

Most of these conclusions were confirmed when the number of clusters was established endogenously, on the basis of the optimization of the correctness index. The results provided by the probabilistic d -clustering are of highest quality and also similar to the k -means procedure results. It is worth noting, however, that from these three attempts, the probabilistic d -clustering has the smallest propensity to generate too small number of clusters or too many singletons, which can be perceived as one of its main advantages, reducing some inconveniences being sometimes a side effect of taking the neighbourhood into account.

By the way, we have composed the efficiency of the proposed correctness index and the silhouette index it turned out that despite two clusterings obtained using various methods are similar the values of the silhouette index for them can be quite different. This observation calls into question the utility of this latter index in such situation.

Appendix 1. Proof of Theorem 1

In some cases the proof can be perceived as an adaptation of similar theorem for conventional situation (cf. Iyigun and Ben-Israel (2013)).

a) If $\mathbf{c}_k \notin \{\gamma_1, \gamma_2, \dots, \gamma_n\}$ then $-\mathcal{R}_k(\mathbf{c}_k)$ is gradient (17) in $\mathbf{c}_k$ and the thesis of such item is, of course, true. Assume, that $\mathbf{c}_k = \gamma_i$ for some $i \in \{1, 2, \dots, n\}$. Consider the change of argument from γ_i to $\gamma_i + t\mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^m$ and $\|\mathbf{z}\| = 1$. Thus,

$$\begin{aligned} \left. \frac{\partial f_k(\gamma_i + t\mathbf{z})}{\partial t} \right|_{t=0} &= \frac{\partial}{\partial t} \sum_{q=1}^n p_{qk}^{(u)} \sum_{h=1}^n w_{qh}^* d(\gamma_h, \gamma_i + t\mathbf{z}) \Big|_{t=0} = \\ &= - \left(\sum_{q=1}^n p_{qk}^{(u)} \sum_{h=1, h \neq i}^n w_{qh}^* \frac{\gamma_h - \gamma_i}{\|\gamma_h - \gamma_i\|} \right) \mathbf{z} + \sum_{q=1}^n p_{qk}^{(u)} w_{qi}^* = -R_{ik}\mathbf{z} + \sum_{q=1}^n p_{qk}^{(u)} w_{qi}^* \end{aligned}$$

The greatest decrease of value of the function f_k is along R_{ik} , i.e. if $\mathbf{z} = R_{ik} / \|R_{ik}\|$. Then, $\mathbf{c}_k = \gamma_i$ if and only if $\sum_{q=1}^n p_{qk}^{(u)} w_{qi}^* - R_{ik} R_{ik} / \|R_{ik}\| \geq 0$, what is equivalent that $\|R_{ik}\| \leq \sum_{q=1}^n p_{qk}^{(u)} w_{qi}^*$.

- a) It follows from (18) and from the part a), i.e. from that if $\mathbf{c}_k^* = \gamma_i$ for some $i \in \{1, 2, \dots, n\}$, then $\mathcal{R}_k(\mathbf{c}_k^*) = 0$ ($k = 1, 2, \dots, s$).
- b) It follows from the part a).

¹⁰ The constraints are only defined using “must-link” (two objects have to be in the same cluster) or “cannot-link” (two objects must not be placed in the same cluster) rules and the centre of any cluster is updated only by averaging all of the points representing objects that have been assigned to it without any special restriction concerning its cardinality. However, in our case, such additional restrictions seem to be unnecessary. In Ward’s method, they are inappropriate because the number of clusters is here not fixed. In realization of the two remaining algorithms, problems of such type occur only incidentally and from the practical point of view, the decrease of the “true” number of clusters in relation to the assumed one has no essential meaning for inference. Moreover, any restriction in this context may lead to violation of the basic contiguity constraint.

Appendix 2

Table 5 Basic descriptive statistics of distributions of similarity indices of obtained clusterings in relation to the originally generated partition

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)			U(− 2,2)	
		Rand (K)				
Mean	0.3848	0.3772	0.3813	0.5221	0.5227	0.5235
Maximum	0.6123	0.6331	0.6107	0.6648	0.6608	0.6564
Quantile 99%	0.5673	0.5564	0.5528	0.6211	0.6265	0.6247
Quantile 95%	0.5289	0.5144	0.5261	0.5967	0.5945	0.5942
Quantile 90%	0.5048	0.4895	0.5027	0.5758	0.5760	0.5760
Quantile 75% (Q3)	0.4333	0.4188	0.4221	0.5484	0.5493	0.5479
Quantile 50% (median)	0.3675	0.3611	0.3638	0.5168	0.5187	0.5184
Quantile 25% (Q1)	0.3246	0.3240	0.3258	0.4921	0.4923	0.4954
Quantile 10%	0.3022	0.2964	0.3029	0.4754	0.4763	0.4772
Quantile 5%	0.2851	0.2838	0.2855	0.4674	0.4668	0.4673
Quantile 1%	0.2632	0.2632	0.2633	0.4454	0.4485	0.4507
Minimum	0.2424	0.2424	0.2424	0.4218	0.4311	0.4412
Variance	0.0058	0.0051	0.0053	0.0016	0.0016	0.0015
Standard deviation	0.0760	0.0711	0.0730	0.0399	0.0400	0.0392
Skewness	0.6520	0.7518	0.7268	0.5190	0.5358	0.5639
Coefficient of variation (%)	19.7399	18.8427	19.1530	7.6477	7.6480	7.4870
Kurtosis	− 0.3960	− 0.0987	− 0.2715	− 0.0303	− 0.0072	0.0995
		Peirce (K)				
Mean	0.0198	0.0194	0.0169	− 0.0336	− 0.0291	− 0.0267
Maximum	0.1155	0.1179	0.1029	0.3272	0.3415	0.3927
Quantile 99%	0.0682	0.0671	0.0698	0.2418	0.2385	0.2894
Quantile 95%	0.0517	0.0478	0.0430	0.1532	0.1396	0.1511
Quantile 90%	0.0416	0.0382	0.0346	0.0984	0.1017	0.0964
Quantile 75% (Q3)	0.0268	0.0272	0.0230	0.0185	0.0230	0.0234
Quantile 50% (median)	0.0168	0.0168	0.0142	− 0.0451	− 0.0390	− 0.0387
Quantile 25% (Q1)	0.0090	0.0088	0.0077	− 0.1093	− 0.0994	− 0.0964
Quantile 10%	0.0031	0.0027	0.0021	− 0.1385	− 0.1377	− 0.1352
Quantile 5%	0.0000	0.0002	− 0.0002	− 0.1590	− 0.1547	− 0.1490
Quantile 1%	− 0.0053	− 0.0044	− 0.0050	− 0.1765	− 0.1720	− 0.1730
Minimum	− 0.0121	− 0.0087	− 0.0111	− 0.1895	− 0.1836	− 0.1874
Variance	0.0002	0.0002	0.0002	0.0089	0.0085	0.0091
Standard deviation	0.0157	0.0149	0.0141	0.0943	0.0921	0.0954
Skewness	1.2532	1.2663	1.5224	0.8427	0.7877	1.0142
Coefficient of variation (%)	79.4035	76.9809	83.1956	− 280.6699	− 316.6357	− 356.8493
Kurtosis	2.5550	3.1992	4.1716	0.5748	0.5347	1.3974
		Sokal and Sneath (K)				
Mean	0.5101	0.5097	0.5085	0.4864	0.4883	0.4890
Maximum	0.5681	0.5658	0.5604	0.6321	0.6218	0.6319
Quantile 99%	0.5393	0.5379	0.5392	0.5937	0.5908	0.6102
Quantile 95%	0.5286	0.5256	0.5232	0.5623	0.5613	0.5611
Quantile 90%	0.5222	0.5205	0.5183	0.5394	0.5417	0.5418
Quantile 75% (Q3)	0.5134	0.5136	0.5115	0.5084	0.5102	0.5103
Quantile 50% (median)	0.5081	0.5080	0.5068	0.4797	0.4825	0.4823
Quantile 25% (Q1)	0.5039	0.5040	0.5034	0.4561	0.4585	0.4594
Quantile 10%	0.5012	0.5012	0.5009	0.4451	0.4462	0.4475
Quantile 5%	0.5000	0.5001	0.4999	0.4387	0.4410	0.4418

Table 5 (continued)

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)	U(− 2,2)		U(− 0.5,0.5)	U(− 2,2)
Quantile 1%	0.4976	0.4975	0.4970	0.4268	0.4319	0.4327
Minimum	0.4929	0.4950	0.4935	0.4148	0.4276	0.4260
Variance	0.0000	0.0001	0.0001	0.0014	0.0014	0.0014
Standard deviation	0.0089	0.0083	0.0078	0.0380	0.0372	0.0380
Skewness	1.5298	1.5653	1.8527	0.9028	0.8121	0.9779
Coefficient of variation (%)	1.7397	1.6276	1.5413	7.8215	7.6210	7.7662
Kurtosis	3.6013	4.2090	5.6383	0.6162	0.2747	0.8897
		Rand (W)				
Mean	0.4077	0.4118	0.3911	0.5286	0.5265	0.5322
Maximum	0.6123	0.5923	0.5558	0.6644	0.6465	0.6392
Quantile 99%	0.5582	0.5437	0.5278	0.6115	0.6130	0.6163
Quantile 95%	0.5257	0.5266	0.5061	0.5838	0.5859	0.5918
Quantile 90%	0.5082	0.5107	0.4895	0.5708	0.5743	0.5759
Quantile 75% (Q3)	0.4730	0.4760	0.4446	0.5510	0.5484	0.5537
Quantile 50% (median)	0.3990	0.4051	0.3822	0.5265	0.5232	0.5303
Quantile 25% (Q1)	0.3400	0.3497	0.3372	0.5025	0.5001	0.5088
Quantile 10%	0.3117	0.3140	0.3034	0.4880	0.4836	0.4899
Quantile 5%	0.3020	0.3020	0.3012	0.4809	0.4765	0.4791
Quantile 1%	0.2917	0.2878	0.2826	0.4647	0.4627	0.4608
Minimum	0.2826	0.2826	0.2727	0.4327	0.4259	0.4121
Variance	0.0055	0.0053	0.0045	0.0011	0.0012	0.0011
Standard deviation	0.0741	0.0726	0.0668	0.0330	0.0342	0.0336
Skewness	0.1993	0.1209	0.2900	0.3630	0.4010	0.1988
Coefficient of variation (%)	18.1697	17.6217	17.0660	6.2395	6.4982	6.3129
Kurtosis	− 1.0744	− 1.1570	− 0.9671	− 0.1238	− 0.0493	− 0.0578
		Peirce (W)				
Mean	− 0.0005	− 0.0005	− 0.0003	0.0126	0.0078	0.0114
Maximum	0.0399	0.0493	0.0292	0.2827	0.3500	0.4522
Quantile 99%	0.0209	0.0237	0.0197	0.1986	0.2108	0.2334
Quantile 95%	0.0126	0.0130	0.0125	0.1269	0.1227	0.1422
Quantile 90%	0.0086	0.0081	0.0081	0.0959	0.0922	0.1041
Quantile 75% (Q3)	0.0027	0.0023	0.0029	0.0469	0.0433	0.0483
Quantile 50% (median)	− 0.0018	− 0.0017	− 0.0014	0.0053	0.0009	0.0046
Quantile 25% (Q1)	− 0.0052	− 0.0050	− 0.0048	− 0.0293	− 0.0323	− 0.0341
Quantile 10%	− 0.0077	− 0.0076	− 0.0071	− 0.0590	− 0.0633	− 0.0690
Quantile 5%	− 0.0095	− 0.0088	− 0.0086	− 0.0753	− 0.0824	− 0.0987
Quantile 1%	− 0.0122	− 0.0119	− 0.0113	− 0.1287	− 0.1335	− 0.1586
Minimum	− 0.0147	− 0.0142	− 0.0147	− 0.1694	− 0.1773	− 0.1826
Variance	0.0000	0.0000	0.0000	0.0040	0.0043	0.0055
Standard deviation	0.0070	0.0069	0.0064	0.0631	0.0658	0.0743
Skewness	1.4027	1.5678	1.1115	0.6584	0.7528	0.7636
Coefficient of variation (%)	− 1286.2843	− 1495.7496	− 1968.3433	503.0074	844.4648	652.9588
Kurtosis	3.8909	4.9390	1.7978	1.0981	1.8125	2.3396
		Sokal and Sneath (W)				
Mean	0.4997	0.4998	0.4998	0.5055	0.5034	0.5049
Maximum	0.5235	0.5289	0.5172	0.6168	0.6135	0.6385
Quantile 99%	0.5121	0.5132	0.5109	0.5816	0.5840	0.5907
Quantile 95%	0.5070	0.5069	0.5065	0.5555	0.5541	0.5591
Quantile 90%	0.5048	0.5044	0.5041	0.5417	0.5391	0.5441

Table 5 (continued)

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)	U(− 2,2)		U(− 0.5,0.5)	U(− 2,2)
Quantile 75% (Q3)	0.5013	0.5013	0.5014	0.5211	0.5198	0.5219
Quantile 50% (median)	0.4991	0.4992	0.4993	0.5027	0.5004	0.5022
Quantile 25% (Q1)	0.4972	0.4974	0.4976	0.4866	0.4844	0.4852
Quantile 10%	0.4956	0.4958	0.4961	0.4735	0.4714	0.4697
Quantile 5%	0.4944	0.4949	0.4950	0.4669	0.4646	0.4591
Quantile 1%	0.4929	0.4930	0.4934	0.4525	0.4483	0.4418
Minimum	0.4914	0.4916	0.4916	0.4361	0.4188	0.4309
Variance	0.0000	0.0000	0.0000	0.0007	0.0008	0.0009
Standard deviation	0.0039	0.0039	0.0034	0.0271	0.0277	0.0302
Skewness	1.4794	1.6734	1.1808	0.5979	0.6610	0.5821
Coefficient of variation (%)	0.7810	0.7710	0.6902	5.3598	5.5020	5.9898
Kurtosis	4.7869	6.1910	2.6914	0.4540	0.0008	0.8380
		Rand (S)				
Mean	0.4806	0.4917	0.5012	0.5117	0.5128	0.5087
Maximum	0.6335	0.6422	0.6475	0.6446	0.6465	0.6537
Quantile 99%	0.6178	0.6246	0.6294	0.6100	0.6145	0.6125
Quantile 95%	0.5958	0.6023	0.6061	0.5811	0.5858	0.5757
Quantile 90%	0.5782	0.5857	0.5892	0.5637	0.5685	0.5584
Quantile 75% (Q3)	0.5548	0.5600	0.5625	0.5327	0.5330	0.5277
Quantile 50% (median)	0.5065	0.5206	0.5289	0.5055	0.5061	0.5026
Quantile 25% (Q1)	0.3946	0.4098	0.4382	0.4844	0.4855	0.4831
Quantile 10%	0.3558	0.3613	0.3676	0.4715	0.4715	0.4711
Quantile 5%	0.3424	0.3436	0.3439	0.4651	0.4658	0.4632
Quantile 1%	0.3143	0.3220	0.3222	0.4523	0.4559	0.4490
Minimum	0.2830	0.2727	0.2949	0.4257	0.4263	0.4349
Variance	0.0075	0.0074	0.0069	0.0013	0.0014	0.0012
Standard deviation	0.0866	0.0859	0.0833	0.0361	0.0340	0.0348
Skewness	− 0.3188	− 0.4905	− 0.6587	0.7947	0.8424	0.8728
Coefficient of variation (%)	18.0262	17.4682	16.6270	7.0468	7.2086	6.8430
Kurtosis	− 1.2667	− 1.0326	− 0.7406	0.3886	0.3486	0.7068
		Peirce (S)				
Mean	0.0280	0.0307	0.0290	0.0111	0.0119	0.0113
Maximum	0.1249	0.1569	0.1204	0.2886	0.3345	0.3062
Quantile 99%	0.1053	0.1024	0.0996	0.2035	0.1822	0.1855
Quantile 95%	0.0727	0.0775	0.0726	0.1154	0.1266	0.1240
Quantile 90%	0.0573	0.0649	0.0622	0.0946	0.0980	0.0876
Quantile 75% (Q3)	0.0390	0.0435	0.0404	0.0538	0.0516	0.0461
Quantile 50% (median)	0.0225	0.0249	0.0235	0.0075	0.0073	0.0061
Quantile 25% (Q1)	0.0119	0.0128	0.0125	− 0.0341	− 0.0294	− 0.0281
Quantile 10%	0.0054	0.0061	0.0050	− 0.0653	− 0.0656	− 0.0580
Quantile 5%	0.0024	0.0027	0.0022	− 0.0931	− 0.0887	− 0.0802
Quantile 1%	− 0.0026	− 0.0029	− 0.0024	− 0.1336	− 0.1388	− 0.1348
Minimum	− 0.0091	− 0.0097	− 0.0095	− 0.1584	− 0.1837	− 0.1890
Variance	0.0005	0.0006	0.0005	0.0043	0.0044	0.0039
Standard deviation	0.0224	0.0242	0.0226	0.0656	0.0663	0.0625
Skewness	1.3597	1.3407	1.1086	0.4242	0.4647	0.5134
Coefficient of variation (%)	79.8613	78.9507	78.1114	589.3300	554.6228	553.1354
Kurtosis	2.2612	2.6388	1.1076	0.8638	1.1230	1.3694
		Sokal and Sneath (S)				

Table 5 (continued)

Statistics	Balanced sizes			Unbalanced sizes		
	No noise	Noise		No noise	Noise	
		U(− 0.5,0.5)	U(− 2,2)		U(− 0.5,0.5)	U(− 2,2)
Mean	0.5158	0.5174	0.5165	0.5052	0.5056	0.5054
Maximum	0.5724	0.5909	0.5688	0.6175	0.6517	0.6234
Quantile 99%	0.5619	0.5572	0.5567	0.5895	0.5838	0.5835
Quantile 95%	0.5422	0.5446	0.5422	0.5546	0.5585	0.5577
Quantile 90%	0.5327	0.5371	0.5357	0.5445	0.5448	0.5430
Quantile 75% (Q3)	0.5224	0.5251	0.5233	0.5256	0.5246	0.5226
Quantile 50% (median)	0.5125	0.5138	0.5132	0.5035	0.5034	0.5029
Quantile 25% (Q1)	0.5062	0.5068	0.5067	0.4834	0.4857	0.4863
Quantile 10%	0.5029	0.5032	0.5026	0.4675	0.4685	0.4723
Quantile 5%	0.5012	0.5016	0.5012	0.4575	0.4594	0.4615
Quantile 1%	0.4986	0.4986	0.4986	0.4437	0.4414	0.4409
Minimum	0.4947	0.4943	0.4945	0.4322	0.4279	0.4262
Variance	0.0002	0.0002	0.0002	0.0009	0.0009	0.0008
Standard deviation	0.0131	0.0141	0.0131	0.0301	0.0305	0.0290
Skewness	1.3865	1.3310	1.0880	0.3761	0.4518	0.4997
Coefficient of variation (%)	2.5472	2.7325	2.5329	5.9534	6.0378	5.7365
Kurtosis	2.2754	2.5655	0.9867	0.3060	0.6974	0.8861

Source: Own computations using the SAS Enterprise Guide 4.3. software (of which its IML environment)

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Albatineh, A. N. (2010). Means and variances for a family of similarity indices used in cluster analysis. *Journal of Statistical Planning and Inference*, 140, 2828–2838.
- Albatineh, A. N., & Niewiadomska – Bugaj M. (2011). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification*, 5, 179–200.
- Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, 23, 301–313.
- Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science*, 12, 153–155.
- Basu, S., Davidson, I., & Wagstaff, K. (Eds.). (2008). *Constrained clustering: advances in algorithms, theory, and applications*. Boca Raton: CRC Press.
- Ben-Israel, A., & Iyigun, C. (2008). Probabilistic d-clustering. *Journal of Classification*, 25, 5–26.
- Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.
- Borys T. and Rogala P. (eds.) (2008). *Quality of live on local level. An indicators – based study*, United Nations Development Programme (UNDP) in Poland, Warsaw.

- Bradley P. S., Bennett K. P., & Demiriz A. (2000). *Constrained k-means clustering* (technical report MSR–TR–2000–65). Redmond: Microsoft.
- Davies D. L. and Bouldin D. W. (1979), *A cluster separation measure*, IEEE (Institute of Electrical and Electronics Engineers) transactions on pattern analysis and machine intelligence, PAMI-1(2):224–227.
- Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3, 32–57.
- Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 95–104.
- EECS (2014), *Convex functions*, Electrical Engineering and Computer Sciences, Instructional and Electronics Support, University of California, Berkeley, U.S.A., http://inst.eecs.berkeley.edu/~ee227a/fa10/login/l_cvx_fcns.html [as of 08.03.2019].
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*, *Wiley Series in Probability and Statistics* (5th ed.). Chichester: Wiley.
- Ferligoj, A., & Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, 47(4), 413–426.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, vol. 2(1), 193–218.
- Iyigun C. and Ben-Israel A. (2013), *The multi-facility location problem: a probabilistic decomposition method*, Computational Optimization and Applications, www.optimization-online.org/DB_FILE/2012/08/3555.pdf.
- Józefowski, T., & Młodak, A. (2009). *Observation of flows of population in Polish statistics—problems and challenges*. In E. Elsner & H. Michel (Eds.), *Assistance for the younger generation. Statistics and planning in big agglomerations* (pp. 61–76). Berlin: Institut für Angewandte Demographie IFAD.
- Kamaruddin S. and Ravi V. (2020). GRNN++: a parallel and distributed version of GRNN under apache spark for big data regression. In: Sharma, N., Chakrabarti, A., Balas V. E. (eds.), “Data management, analytics and innovation, proceedings of ICDMAI 2019, volume 1”, Springer Nature, Singapore, Pte. Ltd., pp. 215–227.
- Kaufman L. and Rousseeuw P. (2005) Finding groups in data. *An introduction to cluster analysis*, Series: Wiley Series in Probability and Statistics, Wiley, Hoboken.
- Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17, 99–121.
- Kubacki, J., & Jędrzejczak, A. (2016). Small area estimation of income under spatial SAR model. *Statistics in Transition new series*, 17, 365–390.
- Kuhn H. W. (1967). *On a pair of dual nonlinear programs*, [in:] J. Abadie (ed.), *Methods of nonlinear programming*, Amsterdam, North-Holland, pp. 38–54.
- Kuhn, H. W. (1973). A note on Fermat’s problem. *Mathematical Programming*, 4, 98–107.
- LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. Boca Raton: Taylor & Francis Group.
- LeSage J. P. and Pace R. K. (2010), *The biggest myth in spatial econometrics*, Working Paper, Available at SSRN: <http://ssrn.com/abstract=1725503>.
- Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE (Institute of Electrical and Electronics Engineers) Transactions on Pattern Analysis and Machine Intelligence*, 24(12), 1650–1654.
- Młodak A. (2006), *Taxonomic analysis in regional statistics*, Advisory and Information Centre DIFIN, Warsaw, Poland (in Polish).
- Młodak, A. (2013). Neighbourhood of spatial areas in the physical and socio-economical context. *Computational Statistics*, 28, 2379–2414.
- Młodak, A. (2014). On the construction of an aggregated measure of the development of interval data. *Computational Statistics*, 29, 895–929.
- Młodak, A., & Kubacki, J. (2010). A typology of Polish farms using some fuzzy classification method. *Statistics in Transition – new series*, 11, 615–638.
- Mojena, R. (1977). Hierarchical grouping methods and stopping rules: and evaluation. *Computer Journal*, 20, 359–363.
- Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, 4, 453–454.
- Petrucci, A., & Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 169–182.
- Pratesi, M., & Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17, 113–141.
- Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.
- Saeffuddin A., Wigena A. H., Nuyartono N. and Kusumaningrum D. (2013), *Development and application of Bayesian spatial analysis on poverty data in East Java, Indonesia*, paper presented on the 53rd ERSA (European Regional Science Association) Congress: Regional integration: Europe, the Mediterranean and the World Economy, 27th – 31st August 2013, Palermo, Italy, http://www-sre.wu.ac.at/ersa/ersaconfs/ersa13/ERSA2013_paper_01043.pdf.

- Saitta S., Raphael B. and Smith I. F. C. (2007). A bounded index for cluster validity, [in:] Perner P. (eds.) “Machine learning and data mining in pattern recognition”, MLDM, Lecture Notes in Computer Science, Springer Verlag, vol. 4571, pp. 174–187.
- Singh, B. B., Shukla, K., & Kandu, D. (2005). Spatial temporal models in small area estimation. *Survey Methodology*, 31, 183–195.
- Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.
- Wagner, W., & Mantaj, A. (2010). Contiguity matrix of spatial units and its properties on example of land districts of Podkarpacie voivodship. *Statistics in Transition – new series*, 11, 187–203.
- Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). *Constrained K-means clustering with background knowledge*, *Proceedings of the Eighteenth International Conference on Machine Learning*, June 28 – July 01, 2001 (pp. 577–584). San Francisco: Morgan Kaufmann Publishers, Inc..
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Weiszfeld, E. (1937). Sur le point par lequel la somme des distances de n points donnés est minimum. *The Tôhoku Mathematical Journal*, 43, 355–386.
- Zimmer S., Kim J. and Nusser A. (2012). A hierarchical clustering algorithm for multivariate stratification in stratified sampling, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, available at https://www.amstat.org/sections/srms/proceedings/y2012/files/306416_76476.pdf.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.