# Finding Groups in Ordinal Data – an Examination of Some Clustering Procedures

Marek Walesiak[1] and Andrzej Dudek[2]

[1,2] Wrocław University of Economics, Nowowiejska 3, 58-500 Jelenia Góra,
marek.walesiak@ue.wroc.pl, andrzej.dudek@ue.wroc.pl

**Summary.** The article evaluates, based on ordinal data simulated with cluster.Gen function of clusterSim package working in **R** environment, some cluster analysis procedures containing GDM distance for ordinal data (see [4, 18, 19]), nine clustering methods and eight internal cluster quality indices for determining the number of clusters. Seventy two clustering procedures are evaluated based on simulated data originating from a variety of models. Models contain the known structure of clusters and differ in the number of true dimensions, the number of categories for each variable, the density and shape of clusters, the number of true clusters, the number of noisy variables. Each clustering result was compared with the known cluster structure from models applying Hubert and Arabie's [2] corrected Rand index.

**Key words:** Clustering, clusterSim, Ordinal data, Simulation models.

## 1 Introduction

Four basic scales are distinguished in the theory of measurement: nominal, ordinal, interval and ratio scale. Among these four scales of measurement the nominal is considered the lowest. It is followed by the ordinal, the interval, and the ratio one which is the highest. They were introduced by Stevens [15].

Systematics of scales refers to transformations which retain relations of the respective scale. These results are well-known and presented e.g. in the paper [3], p. 106. Any strictly increasing functions are the only permissible transformations within the ordinal scale. The main characteristics of ordinal scale are summarised in Table 1.

## 2 Clustering Procedures for Ordinal Data

Major steps in cluster analysis procedure for ordinal data include (see e.g. [10], pp. 341-343): the selection of objects and variables, the selection of a distance measure, the selection of clustering method, determining the number

**Table 1.** Rules for ordinal scale of measurement

| Scale | Basic empirical operations | Allowed mathematical transformations | Allowed arithmetic operations |
|---|---|---|---|
| Ordinal | equal to, greater than, smaller than | any strictly increasing functions | counting of events (numbers of relations equal to, greater than, smaller than) |

Source: Adapted from [15], pp. 25, 27

of clusters, cluster validation, describing and profiling clusters. Variable normalization step is omitted while performing comparisons with cluster analysis procedure for metric data. The purpose of normalization is to adjust the size and the relative weighting of input variables (see e.g. [11], p. 182). Normalization is used when variables are measured with metric data. Normalization is not necessary with regard to ordinal scale, because only the relations: equal to, greater than, smaller than are permitted with ordinal values.

The construction of distance measure for ordinal data should take these relations into account and should be based on relations between the two analyzed objects and the other objects (context distance measure). In statistical data analysis literature few distance measures for variables measured with ordinal data were suggested. Only GDM distance measure $d_{ik}$ proposed by Walesiak [18], pp. 44-45 satisfies ordinal scale conditions (see Table 1):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{\left[ \sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2 \right]^{\frac{1}{2}}}, \tag{1}$$

$$a_{ipj}(b_{krj}) = \begin{cases} 1 \text{ if } x_{ij} > x_{pj}(x_{kj} > x_{rj}) \\ 0 \text{ if } x_{ij} = x_{pj}(x_{kj} = x_{rj}) \\ -1 \text{ if } x_{ij} < x_{pj}(x_{kj} < x_{rj}) \end{cases} \text{ for } p = k, l; r = i, l \ , \tag{2}$$

where: $i, k, l = 1, \ldots, n$ – the number of objects,
$j = 1, \ldots, m$ – the number of variables,
$x_{ij}(x_{kj}, x_{lj})$ – $i$-th ($k$-th, $l$-th) observation on the $j$-th variable.

Article [4] discusses the properties of GDM distance measure.

Other proposals (e.g. Kendall distance measure [7], p. 181; Gordon distance [5], p. 19; Podani distance [12]) imply the assumption that the ranks are measured with at least, the interval scale (when the differences can be calculated). It is also worth mentioning the following argument, presented by Kaufman and Rousseeuw [6], p. 30: "Therefore, most authors advice treating the ranks as interval-scaled and applying the usual formulas for obtaining dissimilarities (like the Euclidean or Manhattan distance)".

The selected clustering procedures included in the article are as follows:

1. GDM distance measure for ordinal data – GDM2 distance in clusterSim package.

2. The selected methods of cluster analysis (stats and cluster packages):

– $k$-medoids (pam);

– seven hierarchical agglomerative algorithms: single link (single), complete link (complete), group average link (average), weighted average link (mcquitty), incremental sum of squares (ward), centroid (centroid), median (median). The Ward, centroid and median methods are easy to implement with distance matrix for only squared Euclidean distance. These methods could be used with any distance measure, however, the results would lack useful interpretation (see [1], pp. 141, 145);

– hierarchical divisive method by Macnaughton-Smith et. al. [8] – diana.

3. The selected internal cluster quality indices for determining clusters' number (all formulas and references for indices you can find in pdf files of clusterSim package [20]): Davies-Bouldin – index.DB, Calinski-Harabasz – index.G1, Baker & Hubert – index.G2, Hubert & Levine – index.G3, gap – index.Gap, Hartigan – index.H, Krzanowski & Lai – index.KL, Silhouette – index.S.

For Davies-Bouldin, Calinski-Harabasz, gap, Hartigan, and Krzanowski & Lai indices medoids of clusters (representative objects of clusters) are used instead of centroids of clusters.

## 3 Simulation Experiment Characteristics

Data sets are generated in nine different scenarios (see Table 2). Models contain the known structure of clusters. Simulation models differ in the number of true dimensions (variables), the number of categories for each variable, the density and shape of clusters, the number of true clusters, the number of noisy (irrelevant) variables. The noisy variables are simulated independently, based on uniform distribution. Variations of noisy variables, in the generated data, are required to be similar to non-noisy ones (see [9], [13], p. 322).

The clusters in models presented in Table 2 contain continuous observations (metric data). Discretization process is performed on each variable in order to obtain ordinal data (see [20]). The number of categories $k_j$ for categorical variable $X_j$ determines the width of each class intervals $\left[\max_i\{x_{ij}) - \min_i\{x_{ij}\}\right] \Big/ k_j$. Each class interval receives category $1, \ldots, k_j$ independently for each variable and the actual value of variable $x_{ij}$ is replaced by these categories. The number of categories may be different for each variable. The example of dicretization process is shown in Fig. 1.

The next step was to perform one out of seventy two clustering procedures (containing GDM distance for ordinal data, nine clustering methods and eight internal cluster quality indices for determining the number of clusters) with

**Table 2.** Experimental factors for simulation models

| $m$ | $v$ | $nk$ | $cl$ | $lo$ | Centroid of clusters | Covariance matrix $\sum$ | $ks$ |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 4, 6 | 3 | 60, 30, 30 | (0; 0), (1.5; 7), (3; 14) | $\sigma_{jj} = 1, \quad \sigma_{jl} = -0.9$ | 1 |
| 2 | 3 | 7 | 3 | 45 | (1.5; 6, – 3), (3; 12; –6) (4.5; 18; –9) | $\sigma_{jj} = 1$ $(1 \leq j \leq 3)$, $\sigma_{12} = \sigma_{13} = -0.9, \quad \sigma_{23} = 0.9$ | 1 |
| 3 | 2 | 5, 7 | 5 | 50, 20, 25, 25, 20 | (5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5) | $\sigma_{jj} = 1, \quad \sigma_{jl} = 0.9$ | 2 |
| 4 | 3 | 5, 7, 5 | 5 | 25 | (5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5) | $\sigma_{jj} = 1$ $(1 \leq j \leq 3)$, $\sigma_{jl} = 0.9$ $(1 \leq j \neq l \leq 3)$ | 2 |
| 5 | 2 | 5 | 5 | 20, 45, 15, 25, 35 | (0; 0), (0; 10), (5; 5), (10; 0), (10; 10) | $\sigma_{jj} = 1, \quad \sigma_{jl} = 0$ | 3 |
| 6 | 2 | 6, 8 | 4 | 35 | (–4; 5), (5; 14), (14; 5), (5; –4) | $\sigma_{jj} = 1, \quad \sigma_{jl} = 0$ | 3 |
| 7 | 3 | 6 | 4 | 25, 25, 40, 30 | (–4; 5; –4), (5; 14; 5), (14; 5; 14), (5; –4; 5), | a | 4 |
| 8 | 3 | 5, 6, 7 | 5 | 35, 25, 25, 20, 20 | (5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5) | b | 4 |
| 9 | 2 | 7 | 3 | 40 | (0; 4), (4; 8), (8; 12) | c | 4 |

$m$ – model, $v$ – number of variables, $nk$ – number of categories (one number means the same number of categories for each variable); $cl$ – number of clusters; $lo$ – number of objects in each cluster (one number means that clusters contain the same number of objects); $ks$ – shape of clusters (1 – elongated, 2 – elongated and not well separated, 3 – normal, 4 – different for each cluster);

$$\text{a: } \sum\nolimits_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \sum\nolimits_2 = \begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{bmatrix}, \sum\nolimits_3 = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix},$$
$$\sum\nolimits_4 = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix};$$

$$\text{b: } \sum\nolimits_1 = \begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{bmatrix}, \sum\nolimits_2 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \sum\nolimits_3 = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix},$$
$$\sum\nolimits_4 = \begin{bmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{bmatrix}, \sum\nolimits_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

$$\text{c: } \sum\nolimits_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \sum\nolimits_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \sum\nolimits_3 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Source: authors' compilation with clusterSim package (see [20])

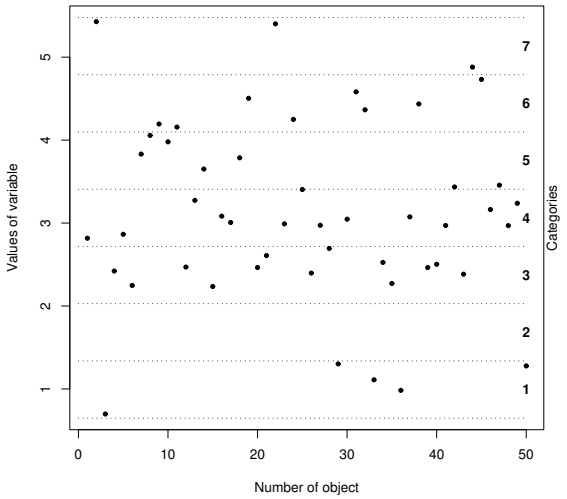


**Fig. 1.** The example of discretization process
Source: authors' compilation

each model. The analysis reffered only to clustering results from 2 to 10 clusters. Next each clustering result was compared with the cluster structure known from models applying Hubert and Arabie's [2] corrected Rand index. The maximum value of corrected Rand index is 1 for identical partitions and its expected value is zero when the partitions are selected at random. Fifty realizations were generated from each setting.

## 4 Discussion on Simulation Results

In table 3 nine clustering methods are ranked, based on adjusted Rand index mean values for nine models and eight internal cluster quality indices (with 50 simulations).

The following conclusions can be drawn from the results presented in Table 3:

– group average method is definitely the best, while single link method is the worst for clustering ordinal data,

– Ward method ensures better results in clustering ordinal data with noisy variables.

Table 4 presents internal quality indices of clustering results ranking based on adjusted Rand index mean values for nine models and nine clustering methods (with 50 simulations).

Based on the results in Table 4 the following conclusions can be drawn:

**Table 3.** Clustering methods ranking based on adjusted Rand index mean values

| Method | Mean | | Shape of clusters | | | | | | | | No. of noisy variables | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | | 2 | | 3 | | 4 | | 0 | | 2 | | 4 | |
| average | 0.545 | 1 | 0.514 | 1 | 0.509 | 2 | 0.494 | 1 | 0.625 | 1 | 0.739 | 1 | 0.508 | 1 | 0.388 | 1 |
| ward | 0.512 | 2 | 0.473 | 3 | 0.479 | 3 | 0.465 | 2 | 0.591 | 3 | 0.680 | 7 | 0.482 | 2 | 0.373 | 2 |
| mcquitty | 0.506 | 3 | 0.450 | 4 | 0.473 | 4 | 0.445 | 3 | 0.606 | 2 | 0.706 | 3 | 0.463 | 3 | 0.350 | 4 |
| diana | 0.499 | 4 | 0.477 | 2 | 0.532 | 1 | 0.388 | 6 | 0.565 | 5 | 0.704 | 4 | 0.428 | 6 | 0.364 | 3 |
| complete | 0.484 | 5 | 0.433 | 5 | 0.466 | 5 | 0.418 | 5 | 0.573 | 4 | 0.700 | 5 | 0.436 | 5 | 0.315 | 5 |
| pam | 0.465 | 6 | 0.415 | 6 | 0.446 | 6 | 0.425 | 4 | 0.539 | 6 | 0.664 | 8 | 0.422 | 7 | 0.310 | 6 |
| centroid | 0.408 | 7 | 0.384 | 7 | 0.362 | 7 | 0.370 | 7 | 0.479 | 8 | 0.721 | 2 | 0.451 | 4 | 0.051 | 8 |
| median | 0.402 | 8 | 0.343 | 8 | 0.362 | 8 | 0.341 | 8 | 0.510 | 7 | 0.690 | 6 | 0.381 | 8 | 0.136 | 7 |
| single | 0.312 | 9 | 0.324 | 9 | 0.238 | 9 | 0.256 | 9 | 0.390 | 9 | 0.613 | 9 | 0.291 | 9 | 0.032 | 9 |

Shape of clusters: 1 – elongated, 2 – elongated and not well separated, 3 – normal, 4 – different for each cluster

**Table 4.** Internal quality indices of clustering results ranking based on adjusted Rand index mean values

| Index | Mean | | Shape of clusters | | | | | | | | No. of noisy variables | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | | 2 | | 3 | | 4 | | 0 | | 2 | | 4 | |
| KL | 0.472 | 1 | 0.424 | 2 | 0.432 | 1 | 0.440 | 1 | 0.553 | 1 | 0.722 | 1 | 0.442 | 1 | 0.254 | 2 |
| G1 | 0.430 | 2 | 0.422 | 3 | 0.406 | 4 | 0.352 | 5 | 0.503 | 3 | 0.616 | 4 | 0.423 | 2 | 0.250 | 3 |
| Gap | 0.414 | 3 | 0.440 | 1 | 0.323 | 8 | 0.341 | 6 | 0.505 | 2 | 0.687 | 2 | 0.346 | 7 | 0.208 | 8 |
| G3 | 0.408 | 4 | 0.359 | 6 | 0.421 | 2 | 0.353 | 4 | 0.469 | 6 | 0.559 | 8 | 0.408 | 3 | 0.257 | 1 |
| S | 0.404 | 5 | 0.381 | 4 | 0.373 | 5 | 0.339 | 7 | 0.482 | 4 | 0.585 | 6 | 0.399 | 4 | 0.226 | 5 |
| H | 0.397 | 6 | 0.368 | 5 | 0.370 | 6 | 0.327 | 8 | 0.479 | 5 | 0.594 | 5 | 0.361 | 6 | 0.234 | 4 |
| G2 | 0.391 | 7 | 0.313 | 8 | 0.406 | 3 | 0.358 | 3 | 0.456 | 7 | 0.583 | 7 | 0.373 | 5 | 0.218 | 6 |
| DB | 0.391 | 8 | 0.343 | 7 | 0.362 | 7 | 0.373 | 2 | 0.454 | 8 | 0.628 | 3 | 0.337 | 8 | 0.208 | 7 |

KL – Krzanowski & Lai, G1 – Caliski-Harabasz, Gap – gap, G3 – Hubert & Levine, S – Silhouette, H – Hartigan, G2 – Baker & Hubert, DB – Davies-Bouldin

– Krzanowski & Lai and Calinski & Harabasz indices present the best results in searching for optimal number of clusters in ordinal data,

– gap and Davies-Bouldin indices definitely show worse results in searching for optimal number of clusters in ordinal data containing noisy variables.

Table 5 presents the ranking of seventy two clustering procedures based on adjusted Rand index mean values for nine models and 50 simulations.

With reference to the aggregated results of simulations illustrated in Table 5 the following conclusions can be made:

– clustering with group average link algorithm turns out to be the most efficient way for the simulation experiment, while applying Krzanowski & Lai index. This method, combined with Gap, Hartigan, Calinski-Harabasz and Davies-Bouldin indices, was ranked respectively at the fourth, sixth, seventh and ninth position,

– the second and the third positions were taken by Ward method, along with applying Krzanowski & Lai and Gap indices,

**Table 5.** Clustering procedures ranking based on adjusted Rand index mean values (the selected results)

| Rank | Method | Mean | Index | Shape of clusters | | | | | | | | No. of noisy variables | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 | | 2 | | 3 | | 4 | | 0 | | 2 | | 4 | |
| 1 | average | 0.623 | KL | 0.553 | 7 | 0.577 | 1 | 0.608 | 1 | 0.710 | 1 | 0.853 | 3 | 0.590 | 1 | 0.426 | 1 |
| 2 | ward | 0.610 | KL | 0.537 | 9 | 0.550 | 5 | 0.596 | 2 | 0.708 | 2 | 0.852 | 4 | 0.571 | 2 | 0.407 | 4 |
| 3 | ward | 0.578 | Gap | 0.648 | 2 | 0.447 | 39 | 0.495 | 7 | 0.673 | 3 | 0.857 | 2 | 0.502 | 11 | 0.375 | 14 |
| 4 | average | 0.573 | Gap | 0.649 | 1 | 0.440 | 46 | 0.496 | 6 | 0.662 | 4 | 0.883 | 1 | 0.481 | 18 | 0.354 | 24 |
| 5 | mcquitty | 0.565 | KL | 0.488 | 16 | 0.528 | 8 | 0.533 | 4 | 0.662 | 5 | 0.801 | 9 | 0.512 | 9 | 0.381 | 13 |
| 6 | average | 0.564 | H | 0.556 | 6 | 0.531 | 7 | 0.471 | 12 | 0.654 | 6 | 0.726 | 19 | 0.544 | 3 | 0.423 | 2 |
| 7 | average | 0.558 | G1 | 0.565 | 4 | 0.518 | 10 | 0.476 | 11 | 0.634 | 10 | 0.735 | 16 | 0.543 | 4 | 0.395 | 8 |
| 8 | pam | 0.553 | KL | 0.476 | 21 | 0.508 | 13 | 0.534 | 3 | 0.647 | 7 | 0.845 | 5 | 0.478 | 19 | 0.336 | 30 |
| 9 | average | 0.538 | DB | 0.486 | 17 | 0.502 | 16 | 0.530 | 5 | 0.601 | 18 | 0.772 | 14 | 0.474 | 20 | 0.367 | 18 |
| 10 | diana | 0.535 | KL | 0.466 | 23 | 0.571 | 3 | 0.457 | 16 | 0.609 | 16 | 0.780 | 12 | 0.458 | 28 | 0.367 | 17 |
| – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| 68 | median | 0.334 | DB | 0.267 | 69 | 0.288 | 65 | 0.313 | 60 | 0.425 | 66 | 0.678 | 35 | 0.266 | 68 | 0.059 | 61 |
| 69 | single | 0.292 | S | 0.302 | 67 | 0.247 | 69 | 0.228 | 70 | 0.358 | 69 | 0.618 | 60 | 0.250 | 69 | 0.008 | 66 |
| 70 | single | 0.269 | DB | 0.253 | 72 | 0.200 | 70 | 0.246 | 69 | 0.342 | 70 | 0.614 | 61 | 0.182 | 70 | 0.012 | 65 |
| 71 | single | 0.243 | Gap | 0.259 | 70 | 0.132 | 72 | 0.205 | 71 | 0.331 | 71 | 0.571 | 71 | 0.150 | 71 | 0.007 | 67 |
| 72 | single | 0.235 | H | 0.254 | 71 | 0.137 | 71 | 0.181 | 72 | 0.322 | 72 | 0.551 | 72 | 0.146 | 72 | 0.007 | 69 |

– the single-link algorithm, combined with Hartigan, Gap and Davies-Bouldin indices, is the least efficient method for ordinal data clustering.

## 5 Limitations

In our analysis the random generation of data set comes from multivariate normal distribution in which clusters' locations and the homogeneity of shapes are defined by means (centroids) and covariance matrices (distortion of objects). Such approach is typical for many other simulation studies, presented e.g. in papers [14, 16, 17]. The infinite number of cluster shapes for any number of dimensions becomes the main problem regarding data generation with known cluster structure. It seems substantiated to consider other distributions and copula functions in data generation process for data with non-standard cluster shapes. This task poses substantial difficulties, especially in case of ordinal data.

In our simulation study we do not take into account such methods like as spectral clustering for ordinal data and non-distance based methods (e.g. Latent Class Analysis for ordinal data).

## References

1. M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, San Francisco, London, 1973.

2. L.J. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193-218, 1985.
3. K. Jajuga and M. Walesiak. Standardisation of data set under different measurement scales. In R. Decker and W. Gaul, editors, *Classification and Information Processing at the Turn of the Millennium*, pages 105-112, Springer-Verlag, Berlin, Heidelberg, 2000.
4. K. Jajuga, M. Walesiak, and A. Bąk. On the general distance measure. In M. Schwaiger and O. Opitz, editors, *Exploratory Data Analysis in Empirical Research*, pages 104-109, Springer-Verlag, Berlin, Heidelberg, 2003.
5. A.D. Gordon. *Classification.* Chapman & Hall/CRC, London, 1999.
6. L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley, New York, 1990, 2005.
7. M.G. Kendall. Discrimination and classification. In P.R. Krishnaiah, editor, *Multivariate Analysis I*, pages 165-185, Academic Press, New York, 1966.
8. P. Macnaughton-Smith, W.T. Williams, M.B. Dale, and L.G. Mockett. Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature*, June, 202:1034-1035, 1964.
9. G.W. Milligan. An algorithm for generating artificial test clusters. *Psychometrika*, 50(1):123-127, 1985.
10. G.W. Milligan. Clustering validation: results and implications for applied analyses. In P. Arabie, L.J. Hubert, and G. de Soete, editors, *Clustering and Classification*, pages 341-375, World Scientific, Singapore, 1996.
11. G.W.Milligan and M.C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, 5(2):181-204, 1988.
12. J. Podani. Extending Gowers general coefficient of similarity to ordinal characters. *Nature*, 48:331-340, 1999.
13. W. Qiu and H. Joe. Generation of random clusters with specified degree of separation. *Journal of Classification*, 23(2):315-334, 2006.
14. G. Soffritti. Identifying multiple cluster structures in a data matrix. *Communications in Statistics. Simulation and Computation*, 32(4):1151-1177, 2003.
15. S.S. Stevens. Measurement, psychophysics and utility. In C.W. Churchman and P. Ratooch, editors, *Measurement. Definitions and Theories*, pages 18-63, New York, 1959.
16. R. Tibshirani and G. Walther. Cluster validation by predicting strength. *Journal of Computational and Graphical Statistics*, 14(3):511-528, 2005.
17. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, ser. B, 63(2):411-423, 2001.
18. M. Walesiak. *Statystyczna analiza wielowymiarowa w badaniach marketingowych [Multivariate Statistical Analysis in Marketing Research].* Wrocław University of Economics, Research Papers no. 654, 1993.
19. M. Walesiak. *Uogólniona miara odległości w statystycznej analizie wielowymiarowej [The Generalised Distance Measure in Multivariate Statistical Analysis].* Wydawnictwo AE, Wrocław, 2006.
20. M. Walesiak and A. Dudek. *clusterSim package*, URL http://www.R-project.org/, 2009.