pattern.GDM1(clusterSim)

An application of GDM1 distance for metric data to compute the distances of objects from the upper (ideal point co-ordinates) or lower (anti-ideal point co-ordinates) pattern object

The main goal of the linear ordering methods is to identify the order of the objects with respect to predetermined criterion. Usually the synthetic measure, which aggregates the partial information contained in the variables, is used.

The GDM1 distance measure can be applied for computing distances from the pattern object in the linear ordering methods. Here:

1. We start with data matrix $[x_{ij}]$, where x_{ij} denotes *i*-th observation on *j*-th variable.

Table 1. Data matrix (17 objects and 10 variables)

Voivodships	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Poland	98.5	94.4	92.1	73.9	84.5	3.5	2.67	0.76	61.9	23.2
Dolnoslaskie	99.1	91.8	89.0	85.2	80.1	3.41	2.63	0.77	60.6	23.0
Kujawsko-pomorskie	99.0	94.7	90.3	72.9	82.5	3.48	2.74	0.79	58.7	21.4
Lubelskie	97.3	94.1	92.5	69.0	88.7	3.69	2.82	0.76	63.7	22.6
Lubuskie	99.1	94.2	91.4	75.6	81.7	3.59	2.77	0.77	62.8	22.7
Lodzkie	96.2	89.5	86.1	61.4	79.9	3.31	2.46	0.74	57.7	23.4
Malopolskie	98.2	96.2	95.0	79.0	85.0	3.46	2.70	0.78	62.3	23.0
Mazowieckie	97.8	95.7	93.7	77.2	90.1	3.35	2.45	0.73	61.2	25.0
Opolskie	99.3	95.2	93.3	77.2	83.8	3.61	2.80	0.77	64.5	23.1
Podkarpackie	97.8	95.3	94.6	88.2	87.5	3.77	3.04	0.81	66.7	21.9
Podlaskie	97.9	94.9	94.1	45.3	89.9	3.80	2.79	0.73	64.1	23.0
Pomorskie	99.7	97.6	94.4	75.8	86.0	3.53	2.70	0.76	62.2	23.0
Slaskie	99.1	93.2	91.4	71.1	79.5	3.44	2.65	0.77	61.4	23.2
Swiętokrzyskie	96.5	92.6	91.5	69.7	88.9	3.48	2.73	0.78	60.8	22.3
Warminsko-mazurskie	99.8	97.7	94.8	72.8	88.2	3.60	2.83	0.79	60.0	21.2
Wielkopolskie	99.3	95.6	93.1	70.0	83.1	3.73	2.83	0.76	68.5	24.2
Zachodniopomorskie	99.7	95.7	93.3	80.6	87.3	3.57	2.72	0.76	61.8	22.8

Source: Voivodships Statistical Yearbook, Poland 2008.

Data on the Polish voivodships, owing to the conditions of the population living in cities in 2007. The analysis includes the following variables:

x1 – dwellings in % fitted with water-line system,
x2 – dwellings in % fitted with lavatory,
x3 – dwellings in % fitted with bathroom,
x4 – dwellings in % fitted with gas-line system,
x5 – dwellings in % fitted with central heating,
x6 – average number of rooms per dwelling,
x7 – average number of persons per dwelling,
x8 – average number of persons per room,
x9 – usable floor space in square meter per dwelling,
x10 – usable floor space in square meter per person.

2. Three types of performance variables are distinguished:

– stimulants – where higher value means better performance,
– destimulants – where low values indicate better performance,

– nominants – where the best value is implied. Object performance is positively assessed if the measure has implied value.

Types of performance variables in example:

x1 – x6, x9, x10 – stimulants,

x7, x8 – destimulants.

In `performanceVariable` we give information about performance variables – `c("s","s","s","s","s","s",d","d","s","s")`.

In `nomOptValues` vector we give information about nominal values of nominants.

3. Researcher determine whether the variables:

a) are measured with the `ratio` (`"r"`) scale,

b) are measured with the `interval` (`"i"`) scale,

c) are mixed (vector with `r/i` values): some are measured with the ratio scale and some are measured with the interval scale.

In our example: `scaleType<-"r"`

4. Nominants are converted into stimulants using formula:

a) `quotient` (for variables measured on ratio scale only): $x_{ij} = \frac{\min\{nom_j; x_{ij}^N\}}{\max\{nom_j; x_{ij}^N\}}$,

where: x_{ij}^N – *i*-th observation on *j*-th nominant variable,

nom_j – nominal value of the *j*-th nominant variable;

b) `difference` (for variables measured on ratio and interval scale): $x_{ij} = -\left|x_{ij}^N - nom_j\right|$.

The scale level of nominant variables	Method of transformation	Transformed variable scale level
a) ratio	quotient	ratio
	difference	interval
b) interval	difference	interval
c) mixed:		
– for variables measured on ratio scale	quotient	ratio
– for variables measured on interval scale	difference	interval
– all variables (ratio and interval)	difference	interval

5. Decision concerning variable normalisation. After normalisation we receive normalised matrix data $[z_{ij}]$, where z_{ij} denotes normalised value of the *j*-th variable for the *i*-th object.

Table 2. Allowed normalisation formulas for metric data

Variable scale level	data matrix $[x_{ij}]$		
	ratio	ratio	interval or ratio/interval
Selection of variable normalization formula	n6 – quotient transformation (x/sd) n6a – positional quotient transformation (x/mad) n7 – quotient transformation (x/range) n8 – quotient transformation (x/max) n9 – quotient transformation (x/mean) n9a – positional quotient transformation (x/median) n10 – quotient transformation (x/sum) n11 – quotient transformation x/sqrt(SSQ)	n1 – standardization n2 – positional standardization n3 – unitization n3a – positional unitization n4 – unitization with zero minimum n5 – normalization in range [–1, 1] n5a – positional normalization in range [–1, 1] n12 – normalization n12a – positional normalization	n1 – standardization n2 – positional standardization n3 – unitization n3a – positional unitization n4 – unitization with zero minimum n5 – normalization in range [–1, 1] n5a – positional normalization in range [–1, 1] n12 – normalization n12a – positional normalization
Transformed variable scale level	ratio	interval	interval

For details see in data.Normalization_details pdf file.

6. The co-ordinates of pattern object consist of the best variables' values (for ideal point co-ordinates) or consist of the worst variables' values (for anti-ideal point co-ordinates).

7. Upper pattern – ideal point co-ordinates consists of the best variables' values. Two types of construction `upper` pattern are distinguished in `patternType`:

a) `"dataBounds"` – pattern should be calculated as following: maximum for stimulants, minimum for destimulants,
b) `"manual"` – pattern should be given in `patternManual` and pattern co-ordinates contain:
   – real numbers,
   – "min" – for minimal value of variable (for destimulants),
   – "max" – for maximal value of variable (for stimulants).

8. Lower pattern – anti-ideal point co-ordinates consists of the worst variables' values. Two types of construction `lower` pattern are distinguished in `patternType`:

a) `"dataBounds"` – pattern should be calculated as following: minimum for stimulants, maximum for destimulants,
b) `"manual"` – pattern should be given in `patternManual` and pattern co-ordinates contain:
   – real numbers,
   – "min" – for minimal value of variable (for stimulants),
   – "max" – for maximal value of variable (for destimulants).

9. If the weights are not equal in GDM1 it is necessary to give the weights w_j in `weights`:

a) `"different1"` – vector of different weights should satisfy conditions: each weight takes value from interval [0; 1] and sum of weights equals one – e.g. `c(0.14,0.16,0.1,0.05,0.2,0.12,0.05,0.08, 0.04,0.06)`,
b) `"different2"` – vector of different weights should satisfy conditions: each weight takes value from interval [0; m] and sum of weights equals *m* (*m* – the number of variables) – (e.g. `c(0.5,1.5,1.6,0.8,0.8,2.0,0.2,0.6,1.5,0.5)`.

10. For each object the GDM1 distance from the pattern object (ideal or anti-ideal point) is determined.

11. We sort the objects based on ascending order of GDM1 distances from pattern object (upper pattern) or descending order (lower pattern).

12. Graphical presentation of results.

References

Jajuga, K., Walesiak, M. (2000), *Standardisation of data set under different measurement scales*, In: R. Decker, W. Gaul (Eds.*), Classification and information processing at the turn of the millennium*, Springer-Verlag, Berlin, Heidelberg, 105-112.

Jajuga, K., Walesiak, M., Bak, A. (2003), *On the general distance measure*, In: M. Schwaiger, O. Opitz (Eds.), *Exploratory data analysis in empirical research*, Springer-Verlag, Berlin, Heidelberg, 104-109.

Walesiak, M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych* [*Multivariate Statistical Analysis in Marketing Research*], Wroclaw University of Economics, Research Papers no. 654.

Walesiak, M. (1999), *Distance Measure for Ordinal Data*, Argumenta Oeconomica, No. 2 (8), 167-173.

Walesiak, M. (2002), *Propozycja uogolnionej miary odleglosci w statystycznej analizie wielowymiarowej*, In: J. Paradysz (Ed.), *Statystyka regionalna w służbie samorzadu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza, Centrum Statystyki Regionalnej, AE w Poznaniu, Poznan, 115-121.

Walesiak, M. (2006), *Uogolniona miara odległosci w statystycznej analizie wielowymiarowej* [*The Generalized Distance Measure in multivariate statistical analysis]*, Wydanie drugie rozszerzone, Wydawnictwo AE, Wroclaw.

Walesiak, M. (2008), *Cluster analysis with clusterSim computer program and R environment*, Acta Universitatis Lodziensis. Folia Oeconomica, No. 216, 303-311.

Walesiak, M. (2009), *Analiza skupien [Cluster analysis]*, In: M. Walesiak, E. Gatnar (Eds.), *Statystyczna analiza danych z wykorzystaniem programu R [Statistical data analysis with R program]*, WN PWN, Warszawa, 407-433.

Walesiak, M. (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R* [*The Generalized Distance Measure GDM in multivariate statistical analysis with R]*, Wydawnictwo UE, Wroclaw.

Walesiak, M., Dziechciarz, J., Bąk, A. (1998), *Ordinal variables in the segmentation of advertisement receivers*, In: A. Rizzi, N. Vichi, H.H. Bock (Eds.), *Advances in Data Science and Classification*, Springer, Heidelberg, 655-662.