# MODIFICATIONS OF THE GOWER SIMILARITY COEFFICIENT

**ZDENĚK ŠULC, JIŘÍ PROCHÁZKA, MARTIN MATĚJKA**
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
e-mail: zdenek.sulc@vse.cz, jiri.prochazka@vse.cz, martin.matejka@vse.cz

## Abstract

*In economic datasets, e.g. in questionnaire surveys data, mixed-type variables often occur. For some tasks, e.g. market segmentation, cluster analysis is usually needed. When clustering objects are characterized by mixed-type variables, there are two basic choices: latent class analysis or cluster analysis based on similarity measures. The paper focuses on similarity measures for mixed data which can be applied in hierarchical cluster analysis. We propose several modifications of the Gower similarity coefficient. Instead of the simple matching approach in its part determined for nominal variables, the modifications considering variability or frequency distribution of a certain variable are used. The clustering results based on proposed modifications are compared with results obtained with the original Gower coefficient and further, with alternative methods to mixed-data clustering; namely, two-step cluster analysis and latent class analysis. The comparison is performed on several economic datasets. Moreover, the classification performance of the introduced measures is examined in a subsequent analysis.*

***Key words:*** *Gower similarity coefficient, modifications, mixed-type variables, hierarchical cluster analysis.*

## 1. Introduction

In questionnaire surveys, there occur mostly categorical variables as predefined answers for given questions. Often, they are accompanied by several quantitative variables, such as a respondent’s salary or a number of family members. When performing a market segmentation on such a mixed-type dataset, hierarchical cluster analysis is often used. Then, a similarity measure for mixed-type data must be used. One of the most popular measures is the Gower similarity coefficient (see Gower, 1971). Since its introduction, new approaches to similarity determination, especially for nominal data, were developed. The nominal part of the Gower coefficient treats similarity between two categories by one, if the categories match and zero otherwise. This is a very simplistic approach. Therefore, this paper introduces three modifications of the Gower similarity coefficient. They differ in their nominal parts, which are based on some of recent findings in this area; for example, they take into account variability of categories in a variable, or frequencies distribution of categories (see Boriah et al., 2008; Šulc and Řezanková, 2014).

The Gower similarity coefficient and its modifications are compared and evaluated from a point of view of their clustering performance in hierarchical cluster analysis (HCA). Moreover, these results are further compared with two alternative methods to mixed-data clustering, namely, two-step cluster analysis and latent class analysis. By the evaluation, an emphasis is put on substantial interpretation. For a comparison, four economic datasets are used. The subsequent analysis examines the modifications of the Gower coefficient in respect to their classification performance.

The paper is organized as follows. Section 2 is divided into tree subsections. The first one introduces the Gower dissimilarity coefficient and its modifications, the second one presents alternative methods for mixed-data clustering, and in the third one, an external evaluation criterion for an accuracy of a cluster assignment is presented. Applications to real datasets is presented in Section 3. The results are summarized in Conclusion.

## 2. Modifications of the Gower coefficient and Alternative Methods for Mixed Data

### 2.1. Modifications of the Gower coefficient

The original Gower coefficient (see Gower, 1971), was introduced as a similarity measure. However, for purposes of HCA, it is usually expressed as a dissimilarity measure. Let us assume the data matrix **X** = [$x_{ic}$], where $i$ = 1, 2, …, $n$ ($n$ is the total number of objects) and $c$ = 1, 2, …, $m$ ($m$ is the total number of variables). Then, dissimilarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$, which are characterized by values of mixed-type variables, is expressed using the formula

$$d_G(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} w_{ijc} d_{ijc}}{\sum_{c=1}^{m} w_{ijc}}, \tag{1}$$

where $d_{ijc}$ is a dissimilarity measure between the $i$-th and $j$-th objects by the $c$-th variable ($c$ = 1, ..., $m$), and $w_{ijc}$ takes the value zero, if either the $i$-th or the $j$-th object by the $c$-th variable is missing; otherwise, it takes the value one.

If the $c$-th variable is nominal (or alternative), dissimilarity between its two categories is treated as zero for matches of categories, and as one otherwise.

If the $c$-th variable is numeric, dissimilarity is expressed by the formula

$$d_{ijc} = \frac{|x_{ic} - x_{jc}|}{\max(x_c) - \min(x_c)}. \tag{2}$$

If the $c$-th variable is ordinal, all the categories are transformed according to the formula

$$x_{ic} = \frac{r_{ic} - 1}{R_c - 1}, \tag{3}$$

where $r_{ic}$ is the rank number of the $i$-th ordinal category ($r$ = 1, ..., $R_c$), and the $R_c$ is the maximal rank number of the $c$-th variable. After this transformation, the outcome values can be used in Equation (2) for numeric variables.

In this paper, three modifications of the Gower dissimilarity coefficient are introduced: Gower_VE, Gower_IOF and Gower_LIN. All the modifications try to improve the nominal part of this coefficient. They are based on similarity measures for nominal data, which use additional characteristics about a nominal-scale variable in comparison to the classic simple matching approach used in the Gower coefficient.

The first modification, Gower_VE measure, is based on the Variable Entropy (VE) similarity measure introduced by the authors[1]. It assigns higher weights to the matches in

[1] Currently, the paper under the name "Novel similarity measures for categorical data based on mutability and entropy" is in a review process.

variables with high variability, because they are rarer, than to the matches in low-variability variables. It can be expressed by the formula:

$$d_{ijc} = \begin{cases} 1 + \dfrac{1}{\ln K_c} \displaystyle\sum_{u=1}^{K_c} p_u \ln p_u & \text{if } x_{ic} = x_{jc}, \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

where $u$ ($u$ = 1, ..., $K_c$) is a category in the $c$-th variable, and $p_u$ is a relative frequency of the $u$-th category in the $c$-th variable. The dissimilarity measure takes values from zero to one. Values closer to zero suits for lower dissimilarity between examined categories.

The second modification, the Gower_IOF dissimilarity measure, is based on the Inverse Occurrence Frequency (IOF) measure, which was originally introduced in (Sparck-Jones, 1972) as a similarity measure for information retrieval. This measure assigns higher weights to less frequent mismatches as it is expressed by the formula

$$d_{ijc} = \begin{cases} 0 & \text{if } x_{ic} = x_{jc}, \\ 1 - \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})}, & \textit{otherwise}, \end{cases} \tag{5}$$

where $f(x_{ic})$ is an absolute frequency of the value $x_{ic}$ in the $c$-th variable. The dissimilarity measure takes the value zero in case of match of categories, and the values from zero to the number $(1 - 1/(1 + \ln(n/2)^2)$ otherwise, which with an increasing dataset size converges to one.

The Gower_LIN measure is inspired by the LIN measure (see Lin, 1998). It assigns higher weights to mismatches to less frequent categories, and it has the formula:

$$d_{ijc} = \begin{cases} 0 & \text{if } x_{ic} = x_{jc}, \\ 1 - \dfrac{2 \cdot \ln(p(x_{ic}) + p(x_{jc}))}{\ln p(x_{ic}) + \ln p(x_{jc})}, & \text{otherwise,} \end{cases} \tag{6}$$

where $p(x_{ic})$ is a relative frequency of the value $x_{ic}$ in the $c$-th variable. It takes values from zero to one; the value one is obtained if there are only two categories in a variable, and its limit is close to zero if both the observed categories have very small relative frequencies.

## 2.2. Alternative Methods to Mixed-Data Clustering

In this subsection, two alternative approaches to hierarchical mixed-data clustering are presented; namely, two-step cluster analysis (2STEP) and latent class analysis (LCA).

The 2STEP method (see SPSS, 2014), can be found in the IBM SPSS software. It consists of two steps. In the first one, preliminary clusters are created sequentially; in the second one, hierarchical clustering is applied to the preliminary clusters. The 2STEP method is based on the modified BIRCH algorithm (see Zhang et al., 1996), which enables to work both with numeric and categorical data. This algorithm decides for each object whether it will be assigned into one of the previously created clusters, or into a new one. This way, a cluster feature (CF) tree is created. It stores only collective information for all records falling into the same entry. Thus, there are ensured speed and low memory requirements for an analysis. A disadvantage of the CF approach is its dependence on an initial order of objects. Thus, it is recommended to order objects randomly before an analysis.

The second step of the 2STEP method performs hierarchical clustering on preliminary clusters (leaves of the CF tree). An agglomerative clustering is used, because it enables to compute a sequence of cluster solutions. For determining a distance between clusters, the log-likelihood distance is used. This measure deals with both numeric and categorical variables. It assumes that numeric variables are normally distributed, and categorical variables have a multinomial distribution. The distance can be interpreted as a decrease of log-likelihood when merging two clusters into a new one.

The LCA method (see Vermunt and Magidson, 2004), represents a model-based approach to object clustering. It can be found in various software, e.g. LatentGold. It assumes that the clustered objects are generated by a mixture of several probability distributions, where the number of components is corresponding to the number of clusters. The method deals with both numeric and categorical variables. The latent class model consists of a class variable, and a set of two or more observed (manifest) variables, which are mutually independent. The best model is chosen according to the lowest value of the AIC or BIC criterion. The classification of objects into clusters is based on prevalence, which is probability that a given object belongs to a given cluster. Usually, an object is assigned to a cluster with the highest prevalence. More information about LCA can be found in e.g. Berlin et al. (2014a, 2014b).

### 2.3. External Evaluation Criterion

For purposes of this paper, the Rand index (see Thalamuthu et al., 2006), is used. In comparison to the purity index, which is also known as the accuracy, it considers all possible combinations of each of $n \times (n-1)/2$ pairs of objects in a dataset. There are four possible types of decisions, which can be obtained. A true positive (TP) state occurs if two similar objects are assigned into the same cluster, and a true negative (TN) state is obtained if two dissimilar objects are assigned into two different clusters. Two types of errors can occur. A false positive (FP) state arises if two dissimilar objects are assigned into the same cluster, and a false negative (FN) state appears if two similar objects are assigned into two different clusters. Then, the Rand index can be expressed by the formula:

$$Rand = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \tag{7}$$

It represents a ratio of correctly assigned objects, both positively and negatively, out of all possible pairs. It takes values from zero to one. Values closer to one represent better accuracy in cluster assignment.

## 3. Experimental Application on Real Datasets

An experimental part of this paper consists of two analyses. In the first one, clustering performance of the Gower dissimilarity coefficient modifications is evaluated on four different economic datasets; the first three ones come from questionnaire surveys, clustering of which represent one of the most typical way of their use, the fourth one describes properties of economic goods, which is also a very frequent task. The evaluation is performed from an aspect of substantial interpretation, i.e. meaningfulness of created clusters. The results are also compared with outputs of the 2STEP and LCA methods. In the second analysis, classification performance of the modifications is evaluated on four real economic datasets from the UCI Machine Learning Repository (see Frank and Asuntion, 2010).

In the first analysis, the first dataset comes from the Study of Economic Expectations, which was conducted by the University of Wisconsin Survey Center, and which was held during 1994 to 2002 (see Dominitz and Manski, 2004). In the dataset, which is named as Expectations, 2,409 respondents from the 9th to 15th wave are included after the data cleaning. For the application of clustering, set of seven variables regarding a respondent's profile is included. There are two numeric variables, respondent's age and the number of adults in a household, and four categorical ones: party affiliation [democrat, republican, independent, no preference, other party], marital status [married – spouse present, separated, divorced, widowed, never married, married – spouse absent], high school diploma [yes, no], and working last week [yes, no].

The second dataset comes from the EU-SILC (European Union - Statistics on Income and Living Conditions), which was held in 2011, and it deals with flats properties, where respondents live. It consists of three numeric variables: number of habitable rooms, area of apartment in m$^2$, and monthly costs. Moreover, there are six dichotomous variables concerning the presence or absence of of various negative influences: humidity, dark apartment, small apartment, noise, dirt, high criminality. The dataset contains 1,162 objects.

The third dataset deals with work sectors in USA. It is derived from the Adult dataset (see Frank and Asuntion, 2010). It contains three quantitative variables: age, years of education, and working hours per week. There two categorical variables: work class [government, private, self-employed], and gender [male, female]. Totally, the dataset contains 1,538 objects.

The fourth dataset describes properties of round cut diamonds (see Chu, 2001). There are seven numeric variables: weight in carat units, price in Singapore dollars, length in mm, width in mm, depth in mm, depth in %, and table in %. There are also three categorical variables: cut quality [fair, good, very good, premium, ideal], and color [D, E, F, G, H, I, J], clarity [I1, SI1, SI2, VS1, VS2, VVS1, VVS2]. The original dataset contained 53,940 observations. Since the hierarchical clustering used in this paper is very time-demanding, the number of observations was reduced to 1,611.

The analysis is mainly based on substantial interpretation of cluster solutions produced by different similarity measures and alternative methods to mixed-data clustering. This kind of analysis depends mainly on researcher's experience.

For the second analysis, four datasets from the UCI Machine Learning Repository were chosen. All selected datasets contain mixed-type variables, and a class membership variable. Their basic properties are displayed in Table 1.

Table 1: Basic properties of datasets used for the second analysis (after data cleaning).

| Dataset | Number of objects | Number of numerical variables | Number of categorical variables | Number of classes |
|---|---|---|---|---|
| Hepatitis | 80 | 6 | 13 | 2 |
| Contraceptive Method Choice | 659 | 4 | 9 | 3 |
| Credit Approval | 1473 | 2 | 7 | 2 |
| Post-Operative | 87 | 1 | 7 | 3 |

Source: the authors.

Except for the Credit Approval dataset, where the variables A14 and A15 were deleted, the structures of other datasets were unchanged. Since the examined similarity measures cannot deal with missing observations, the listwise deletion method was performed on datasets containing missing values.**Results of the First Analysis**

In the Expectations dataset, the examined similarity measures and methods varied considerably. The best results are provided by the Gower_LIN measure in the three-cluster solution, which divides respondents to people who work and have a high school, people who do not have a high school (there is 48% people without a job), and people who do not work and do not have a high school. This measure is closely followed by the three-cluster solution of the 2STEP method, which clusters depend mostly on the working last week and age variables. Its first cluster contains only employed people with a high school, who are mostly married (97%). There is a prevalence of republicans (37%). The second cluster contains younger employed people with a high school, who are mostly never married (53%). The third cluster contains mostly unemployed people (87%). For the rest of similarity measures, only the two-cluster solutions are meaningful. The original Gower measure, Gower_VE and the LCA method divide objects according to the number of adult people into larger and smaller households. The Gower_IOF measure, on the contrary, divides people into working and jobless people.

In the EU-SILC dataset, better clusters were obtained by the measures and methods which rely more on numerical variables, i.e. 2STEP, LCA, Gower and Gower_VE. The best clusters were gotten using the 2STEP method, which separated the objects not only according to numerical, but to categorical variables as well. Its three-cluster solution can be interpreted in the following way. The first cluster comprises larger flats, which have on average 4.8 rooms, the area of 106 m$^2$, and the costs of 5,001 CZK per month. These flats have almost none of negative properties, such as noise, dirt, etc. These negative properties are concentrated in the second cluster consisting of flats with a higher ratio of negative properties, which varies from 13% (dark flat) to 46% (dirt). The third cluster comprises rather smaller flats (avg. no. of habitable rooms is 2.6, and the flat area is 61 m$^2$) with the lower average costs of 4,585 CZK, and without any negative properties. The next advantage of clusters provided by 2STEP is their similar size, which is not usual by other methods. The measures Gower and Gower_VE and the LCA method also provided very good clusters dividing objects according to the number of habitable rooms and their flat area. However, they did not take into account categorical variables in comparison to the 2STEP method. The Gower_IOF and Gower_LIN dissimilarity measures provide substandard clusters in this dataset.

In the Adult dataset, the best results were provided by the three-cluster solution of the 2STEP method, which preferred the influence of categorical variables. The clusters are as follows. The first one contains men in a private sector, the second one contains women in a private sector, and the third one consists of people working in a government sector or as self-employees. This distribution of clusters is logical, and moreover, it provides information about working hours, age, and years of education for each group of people. For instance, it was found out that men in a private sector work about six hours per week more than women. Clusters provided by the Gower_IOF and Gower_LIN measures, which are the same, divide people first according to their gender, and second according to their work class. Thus, this is exactly the opposite order in comparison with the 2STEP method. Unfortunately, this setting produces clusters with lower substantive interpretation of created clusters. Good clusters were also provided by the three-cluster solution of the Gower_VE measure. The clusters are divided according to the age, years of education, and working hours in the last week; thus, they illustrate the differences in education and working hours among different generations. The first cluster consists of highly educated people whose average age is is 44 years, who work about 45 hours weekly. The second one consists of people whose average age is 36 years, who have studied 10 years, and they work 40 hours per week on average. The third cluster contains older people with the average age of 61 years, six years of education, who work 32 hours per week on average. The original Gower dissimilarity measure provided clusters in a similar way, but less strictly in comparison to the Gower_VE measure. The worst clusters were provided by the LCA method.

In the Diamonds dataset, all the examined similarity measures and the LCA method prefer three-cluster solutions, except for the 2STEP method, which allowed to use only two-cluster solution, since the higher-cluster solutions lose their substantial interpretation. The best clusters are provided by the LCA method, which divides the diamonds into three groups mainly according to the numeric variables carat and price. The first cluster contains diamonds for the average price 785 Singapore dollars (0.34 carats). Their cut is in 52% considered as ideal. The second cluster comprises diamonds of a medium size; their average price is 2,154 Singapore dollars (0.63 carats). The third cluster contains larger diamonds for the average price of 7,451 Singapore dollars (1.23 carats). The LCA method is closely followed by the examined dissimilarity measures in the following order: Gower, Gower_VE, Gower_LIN and Gower_IOF. All of these dissimilarity measures sort the diamonds to three groups according to their size; therefore, they differ only in cluster sizes.

### 3.2. Results of the Second Analysis

Values of the Rand index for the original Gower measure and its modifications are expressed in Table 2.

Table 2: Values of the Rand index for the Gower measure and its modifications.

| Dataset | Gower | Gower_IOF | Gower_LIN | Gower_VE |
|---|---|---|---|---|
| Hepatitis | 0.497 | 0.708 | 0.708 | 0.494 |
| Contraceptive Method Choice | 0.529 | 0.513 | 0.504 | 0.549 |
| Credit Approval | 0.504 | 0.580 | 0.504 | 0.506 |
| Post-Operative | 0.465 | 0.478 | 0.486 | 0.497 |

Source: the authors.

Overall, the highest values of the Rand index are gotten by the Gower_IOF dissimilarity measure. In comparison to the other examined measures, it performs much better in the Hepatitis (together with Gower_LIN) and the Credit Approval datasets. The second best results are provided by the Gower_VE dissimilarity measure, which performs very well apart from the Hepatitis dataset. It is followed by the Gower_LIN measure, and further, by the original Gower dissimilarity measure, which classification results were never the best ones. It is worth mentioning that the dissimilarity measures, which put higher importance to numerical variables (Gower, Gower_VE), perform well on the same datasets. The similar situation occurs by measures, which put higher importance to categorical variables (Gower_IOF, Gower_LIN).

## 4. Conclusion

This paper introduced three modifications of the Gower dissimilarity coefficient, and examined their clustering performance in hierarchical cluster analysis (HCA) on economic datasets with mixed-type variables, which results were further compared with two-step cluster analysis (2STEP) and latent class analysis (LCA). Moreover, the examined dissimilarity measures were also evaluated from a point of view of their classification performance.

The introduced dissimilarity measures could be divided into the Gower_VE measure, which put higher importance to numeric variables, and to Gower_IOF and Gower_LIN measures, which put higher importance to categorical variables. The Gower_VE measure behaved in a similar way as the original Gower coefficient, but often it created more separated (and thus unbalanced) clusters. Generally, the Gower and Gower_VE dissimilarity measures seemed to create more meaningful clusters in economic datasets. On the other hand, the dissimilarity measures Gower_IOF and Gower_LIN performed very well in a classification task on non-economic datasets.

Both alternative methods to mixed-data clustering performed pretty well. Except for the Diamonds dataset, the 2STEP method provided some of the best clusters. This method is fast to perform as well. Unfortunately, it can be found only in commercial software IBM SPSS; and thus, it is not freely available. The LCA method provided average results in used datasets. In two datasets were its clusters among the best ones, and in other two they belonged to the worst ones.

Due to a limited number of the used datasets in the analyses, there cannot be made any final judgments about the introduced measures and their clustering performance. In the next research, we are going to examine their properties from a different point of view on a higher number of generated datasets.

## Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/41/2016.

## References

[1] BERLIN, K. S., WILLIAMS, N. A., PARRA, G. R. 2014a. An introduction to latent variable mixture modeling (part 1) : Overview and cross-sectional latent class and latent profile analyses. In Journal of Pediatric Psychology, 2014, vol. 39, iss. 2, pp. 174-187.

[2] BERLIN, K. S., PARRA, G. R., WILLIAMS, N. A. 2014b. An introduction to latent variable mixture modeling (part 2) : Longitudinal latent class growth analysis and growth mixture models. In Journal of Pediatric Psychology, 2014, vol. 39, iss. 2, pp. 188-203.

[3] BORIAH, S., CHANDOLA, V., KUMAR, V. 2008. Similarity measures for categorical data : A comparative evaluation. In Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, 2008, pp. 243-254.

[4] CHU, S. 2001. Pricing the C's of Diamond Stones. In Journal of Statistics Education, 2001, vol. 9, iss. 2.

[5] DOMINITZ, J., MANSKI, C. 2004. The survey of economic expectations. Technical report, http://faculty.wcas.northwestern.edu/~cfm754/see_introduction.pdf.

[6] FRANK, A., ASUNCION, A. 2010. UCI machine learning repository. University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml.

[7] GOWER, J. C. 1971. A general coefficient of similarity and some of its properties. In Biometrics, 1971, vol. 28, iss. 4, pp. 857-871.

[8] LIN, D. 1998. An information-theoretic definition of similarity. In ICML '98: Proceedings of the 15th International Conference on Machine Learning. San Francisco : Morgan Kaufmann Publishers, 1998, pp. 296-304.

[9] SPARCK-JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, 1972, vol. 28, iss. 1, pp. 11-21.

[10] SPSS. 2014. Help. Chicago : SPSS, 2014.

[11] ŠULC, Z., ŘEZANKOVÁ, H. 2014. Evaluation of recent similarity measures for categorical data. In: Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wrocław : Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 2014. pp. 249-258.

[12] THALAMUTHU, A. et al. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. In Bioinformatics, 2006, vol. 22, iss. 19, pp. 2405-2412.

[13] VERMUNT, J. K., MAGIDSON, J. 2004. Latent class analysis. In The Sage encyclopedia of social sciences research methods, 2004, pp. 549-553.

[14] ZHANG, T., RAMAKRISHNAN, R., ANDLIVNY, M. 1996. Birch : An efficient data clustering method for very large databases. SIGMOD Rec., 1996, vol. 25, iss. 2, pp. 103-114.