The Stata Journal (*submitted*) *vv*, Number *ii*, pp. 1–10

simcadi: Similarity Indices for Categorical Distributions

Stephan Huber[1]
University of Regensburg
Regensburg, Germany
stephan.huber@wiwi.uni-regensburg.de

Abstract. In this article we introduce the command `simcadi`. It helps to calculate indicators for the similarity of categorically ordered variables and distribution, respectively. In particular, it permits the calculation of the Cosine index, and indices introduced by Finger and Kreinin (1979), Bray and Curtis (1957), Dice (1945), Sørenson (1948), Jaccard (1912), Grubel and Lloyd (1971), Ružička (1958), and Gower (1971). Moreover, it allows us to compute the development of a distribution over time. The command offers various options for an efficient handling of datasets, because it permits the calculation of benchmarks of comparison automatically, and the incoporation of complex weighting schemes.

Note The `simcadi` command was written for the Statistical Software Stata (©StataCorp). It is available upon request, or can be downloaded here:
www.uni-regensburg.de/wirtschaftswissenschaften/vwl-moeller/forschung/index.html

1 Introduction

Measuring the relationship between two variables is essential for empirical work. If variables contain categorically ordered observations, the preferred method for measuring the relationship between variables is similarity indices and dissimilarity indices[2], respectively. The most well-known indices were introduced by Finger and Kreinin (1979), Bray and Curtis (1957), Dice (1945), Sørenson (1948), Jaccard (1912), Grubel and Lloyd (1971), Ružička (1958), and Gower (1971). These indices are used extensively in economic and ecological research, and their mathematical features were studied intensively in Boyce et al. (1995), Boriah et al. (2008), Egghe (2010), and Goshtasby (2012), for example.

In this article, we introduce the `simcadi` command. It facilitates the organization of a dataset in such a way that it is easy to calculate a series of similarity indices for categorical variables and distributions. Although there is great diversity in the definition of similarity indices, the indices which are implemented in `simcadi` can all handle categorically ordered samples, and range from zero to one, with zero indicating no similarity, and one indicating maximum similarity. The indices are the result of a function of scores resulting from the comparison of the value measured for an attribute

2. One is the logical complement of the other: a similarity index measures 'how close' two samples are, whereas a dissimilarity index measures the reverse.

in one sample with the value measured for the same attribute in a second sample (cf. Johnston 1976, p.11f).

Basically, `simcadi` offers three main features: First, `simcadi` can handle datasets that are organized differently (wide vs. long). That means, the comparison of two distributions is not only possible for distributions that are organized in two variables (wide), but also if two, or many distributions are captured in one single variable (long). Second, the command permits the choice of one, or more distribution(s) to which all other distributions should be compared. Alternatively, it allows the incorporation of a weighting scheme which is used to calculate a reference of comparison for each distribution. This is useful, for example, if all distributions should be compared to a weighted subset of distributions rather than to a specific distribution. To give an example, assume you want to compare the disaggregated exports of country A (B,C,...) to only those countries with which country A (B,C,...) shares a common border. Using our command, you only need the disaggregated (categorized) export flows and a contiguity matrix of all the exporting countries. Then, the command uses both to compare the export structure of country A (B,C,...) to the average export structure of all countries that share a common border with country A (B,C,...). Moreover, the command helps to implement other user-written indicators, because the `simcadi` command optionally permits the possibility of saving the variables of comparison. Third, in addition to the indices mentioned above, the command permits the indication of the change in a distribution over time, as was done by De Benedictis and Tajoli (2007b,a), for example.

Similar to metric distance measurements, these similarity indices for categorically ordered variables can also be used to quantify the difference between samples. However, the similarity indices usually do not satisfy the axiom of triangle inequality like distance measurements.[3] In contrast to the calculation and analysis of metric distance measurements, which is already well implemented in Stata (see Fenty 2004), we are not aware of a Stata command to calculate similarity indices.[4] Nevertheless, it has to be mentioned that some similarity indices can be used optionally to execute a cluster analysis in Stata (see *measure_option*), and that the `matrix dissimilarity` command does in fact calculate some similarity measurements. However, it does so for matrices only, incorporates no options for handling the dataset or calculating distributions to work as a benchmark of comparison. Moreover, it is not specifically built to measure similarities of categorical distributions.

In the following sections we introduce the calculation of the indices, explain the syntax, and offer a series of examples to show the flexibility of the command.

3. According to this axiom, the distance between object a and b is shorter than the sum of the distances via another point, c.
4. The R-Project community offers the vegdist command. This is part of the vegan package (see https://github.com/vegandevs/vegan) and enables some dissimilarity indices to be computed. However, the options for handling the dataset to calculate an individual benchmark of comparison are limited.

2 The simcadi command

`simcadi` implements seven indices all of which are defined between zero and one. We overview the formulae of calculation in section 1, where $s_{Atk} = x_{Atk}/X_{At}$ is the value shares of distribution A in category k at time t, with $X_{At} = \sum_k x_{Atk}$. The value shares of distribution B, to which distribution A is compared, are denoted with s_{Btk}, and M_A is the number of columns (excluding missing values).

Please note that the naming conventions vary. The Finger-Kreinin Index, for example, was reinvented several times and is also known under various names as found in works by researchers like Czekanowski (1909), Dice (1945), and Sørenson (1948). Another example is the quantitative version of the Jaccard index, which is sometimes named after Ružička (1958). To give a further example, the Bray-Curtis Index which is probably the most well known index, is sometimes also called the Finger-Kreinin Index, or vice versa, because when using shares—as `simcadi` does by default—both indicators are algebraically identical.

In a nutshell, all measurements consider the overlap of two distributions in some way, whereby the Finger-Kreinin Index (FKI) does so the most directly. The main advantage of the Finger-Kreinin Index is that it is "not influenced by the relative sizes or scales"(Finger and Kreinin 1979, p.906) of the category, and is invariant with respect to proportional sub-classifications, which is shown with an axiomatic approach in Sun and Ng (2000). It should be considered that some indices may give misleading values if data entries are not strictly positive.

The Change in Distribution Index differs from the indices mentioned previously, because it does not compare two distributions but uses one of the indices mentioned previously to calculate the change of one distribution over time. Thus, it is a function of the distribution in two points in time: t and t'. We exemplify its calculation with the Finger-Kreinin Index in section 1. However, any other similarity index can be used for calculation. Consequently, it is also defined between zero and one, with zero indicating that the distribution did not change, and one indicating that the distribution in t' has nothing in common with the distribution in t.

2.1 Syntax

There are three ways to use the command: The first permits the comparison of two variables (*variable_1 variable_2*).

`simcadi` *variable_1 variable_2* [*if*] [*in*] , `class(`*varname*`)` [*options*]

The second can be used to compare a number of distributions (specified with `id()`) that are captured in one single variable (*variable_1*) to the distributions specified in `wcountry()`.

`simcadi` *variable_1* [*if*] [*in*] , `class(`*varname*`)` `id(`*varname*`)` `wcountry(`*name*`)`

Table 1: Formulae for Calculating Similarity Indices

Indicator	Description
Finger-Kreinin Index (FKI)	$FKI_{At} = \sum_k min(s_{Atk}, s_{Btk})$
The Bray-Curtis Index (BCI)	$BCI_{At} = 1 - \frac{\sum_k \lvert s_{Atk} - s_{Btk} \rvert}{\sum_k (s_{Atk} + s_{Btk})}$
Gower Index (GOW)	$GOW_{it} = 1/M_A \cdot \sum_k \left(\frac{\lvert s_{Atk} - s_{Btk} \rvert}{(\max(s_{At}) - \min(s_{At}))} \right)$
Ružička Index (RUZ)	$RUZ_{At} = \frac{\sum_k min(s_{Atk}, s_{Btk})}{\sum_k max(s_{Atk}, s_{Btk})}$
Jaccard Index (JAC)	$JAC_{At} = \frac{\sum_k (s_{Atk} \cdot s_{Btk})}{\left(\sum_k s_{Atk}^2 + \sum_k s_{Btk}^2 - \sum_k s_{Atk} \cdot s_{Btk}\right)}$
Grubel-Lloyd Index (GLI)	$GLI_{At} = \frac{1}{k} \sum_k \frac{(s_{Atk} + s_{Btk*}) - \lvert s_{Atk} - s_{Btk} \rvert}{(s_{Atk} + s_{Btk})}$
Cosine Index (COS)	$COS_{At} = \frac{\sum_k (s_{Atk} \cdot s_{Btk})}{\sqrt{\sum_k s_{Atk}^2} + \sqrt{\sum_k s_{Btk}^2}}$
Change in Distribution (CID)	$CID_{At,t'}^{\text{Similarity Index}}(s_{Atk}, s_{A,t',k})$
Example: $CID_{At,t'}^{\text{FKI}} = 1 - \sum_k \min(s_{Atk}, s_{A,t=t-t',k})$	

[*options*]

The third should be used to incorporate an external weighting scheme set by `using` *filename*. `simcadi` uses this scheme to calculate a benchmark distribution of comparison for each distribution.

`simcadi` *variable_1* `using` *filename* [*if*] [*in*], `class(`*varname*`)` `id(`*varname*`)`
[*options*]

2.2 Options

`class(varname)` specifies the variable containing the categories of the distributions (e.g. the trade classification).

`id(varname)` specifies the variable that identifies the distributions (e.g. the countries).

`wcountry(name)` sets the distribution with which each distribution should be compared. Note: if more than one distribution is set, the unweighted average of the distributions is calculated to work as the distribution of comparison.

`wvarname(varname)` specifies the name of the variable that contains the weights (only applicable if a weighting matrix is assigned).

`varpartner(varname)` specifies the name of the second variable in the weighting scheme file (e.g. the trading partner). Only applicable only if a weighting matrix is assigned.

`cid(#)` calculates the Change in Distribution index (for each indicator chosen). # sets a period to which each 'id(varname)' is compared.

`realvalues` use the real values instead of the shares. Note: Applying real values follows that some indices (Gower) are defined between zero and one. (The default uses shares.)

`time(#)` sets the time period (e.g. 2012).

`timevar(varname)` specifies the name of the time variable.

`savecomp(filename` stores the file with the distributions/variables that are compared (Note: The user can use this file to calculate further indices.)

`saveresult(filename)` stores the calculated indices under the name filename.

`detail` offers a detailed description of the calculation process.

`braycurtis` calculates the Bray-Curtis Index (only allowed if the variable of interest contains positive values only).

`cosine` calculates the Cosine Index.

`finger` calculates the Finger-Kreinin Index.

`gower` calculates the Gower Index.

`grubel` calculates the Grubel-Lloyd Index.

`jaccard` calculates the Jaccard Index.

`ruzicka` calculates the Ruzicka Index.

3 Examples...

...how `simcadi` can handle 'wide' data

Assume you have consumption data for three people (Adam, Brittany, Charlie) who consume three goods (beer, bread, water) at two different periods (1, 2) in the wide format. The following Stata output shows how to compare the consumption profile of Adam with that of Charlie:

```
. use example_wide, clear
. list

  time    good   Adam   Brittany   Charlie
```

```
  1.       1    beer     18          9          0
  2.       1   bread     18         15         10
  3.       1   water     18         13         10
  4.       2    beer     19          7          1
  5.       2   bread     18         11         12

  6.       2   water     17         16         12
```

```
. simcadi Charlie Adam, class(good) timevar(time) realvalues
Calculation of the indices...
Merge the results...
Results were not saved
. list

  1.   time   finger~d   jaccar~d   braycu~d   gowerR~d   cosine~d   grubel~d
          1         20   .4433497   .5405405   1.133333   .8164966   .4761905

                ruzick~d                                 compare2
                .3703704                        Charlie with Adam

  2.   time   finger~d   jaccar~d   braycu~d   gowerR~d   cosine~d   grubel~d
          2         25    .532767   .6329114   .8787879   .8274392   .5758621

                ruzick~d                                 compare2
                 .462963                        Charlie with Adam
```

If you are only interested in the development of the distribution from period one to period two, you can use the `cid` option together with `time`:

```
. simcadi Charlie Adam, class(good) time(2) timevar(time) cid(1) finger cosine
Calculation of the indices...
Merge the results...
Results were not saved
. list

       time   CIDfin~d   finger~d   CIDcos~d   cosine~d            compare2

  1.      2        .04   .6881481   .0017316   .8274392   Charlie with Adam
```

...how `simcadi` can handle 'long' data

`simcadi` is especially built to handle datasets in the long format. The Stata excerpt below shows how to measure the similarity of Adam's consumption with that of Brittany and Charlie using the `wcountry` option if all the consumption data are captured within a single variable.

```
. list
```

```
       name   time    good   consume
  1.     Adam      1    beer        18
  2.     Adam      1   bread        18
  3.     Adam      1   water        18
  4. Brittany      1    beer         9
  5. Brittany      1   bread        15

  6. Brittany      1   water        13
  7.  Charlie      1    beer         0
  8.  Charlie      1   bread        10
  9.  Charlie      1   water        10
 10.     Adam      2    beer        19

 11.     Adam      2   bread        18
 12.     Adam      2   water        17
 13. Brittany      2    beer         7
 14. Brittany      2   bread        11
 15. Brittany      2   water        16

 16.  Charlie      2    beer         1
 17.  Charlie      2   bread        12
 18.  Charlie      2   water        12
```

```
. simcadi consume , class(good) id(name) wcountry(Adam) ///
>         time(1) timevar(time) savecomp(filename, replace) ///
>         finger gower
----------------------
A: No obvious errors.
----------------------
Q: Did the user specify a weight dataset?
A: NO. The exports should be compared to the equally weighted average of the following c
> ountries (Adam).
Automatic calculation of the weighting scheme...
...DONE
----------------------
file filename.dta saved
Calculation of the indices...
Warning: The Gower Index cannot be calculated properly. Please check your data and the p
> rerequisites of the Gower Index.
Merge the results...
...DONE
----------------------
Results were not saved
. list
```

```
        name   time   finger~d   gowerg~d
  1.     Adam      1          1          .
  2. Brittany      1   .9099099   .3703704
  3.  Charlie      1   .6666667   .4444444
```

```
. use filename, clear
. list
```

```
        name   time    good      comp2      comp1
  1.     Adam      1    beer   .3333333   .3333333
```

2.	Adam	1	bread	.3333333	.3333333
3.	Adam	1	water	.3333333	.3333333
4.	Brittany	1	beer	.3333333	.2432432
5.	Brittany	1	bread	.3333333	.4054054
6.	Brittany	1	water	.3333333	.3513514
7.	Charlie	1	beer	.3333333	0
8.	Charlie	1	bread	.3333333	.5
9.	Charlie	1	water	.3333333	.5

A warning note shows that the Gower Index cannot be calculated properly. Here, it is easy to show what went wrong by considering the formula of the Gower Index, and by looking at the content of the file 'filename', which was saved by the option `savecomp()`: The denominator of the Gower Index is zero for the comparison of Adam with Adam, which is undefined. We want to highlight that the saved file can also be used to calculate a user-specified index. Moreover, the `wcountry` option also allows for more than one distribution to be included. `wcountry(Adam Brittany Charlie)`, for example, would yield a benchmark of comparison that contains the average shares of consumption for all three persons. The next section explains how to use more sophisticated weighting schemes.

...how `simcadi` can handle user-specified weighting schemes

Sometimes the user wishes to specify a more complex weighting scheme. To give an example, trade economist are often interested in the export composition of countries, but do not wish to compare the exports of all countries to a single export composition, rather than comparing the exports of each country to only those countries that are on the same continent, or to countries that share a common border. What is even more complex, but quite common in spatial science, is that economists wish to weight the countries in the comparison differently with the inverse distance between countries, for example. In the following, we explain how this can be done using `simcadi`.

Assume you want to compare Brittany with all the others, but the men should only be compared to other men. Additionally, for men, you want to put more weight on other men's distributions. This means that you do not want to compare each distribution to a single distribution, but aim to compare each distribution with a specific distribution. The following Stata excerpt shows how this can be done. We add the `detail` option to give a deeper insight into what the command is doing: First, the command checks whether or not obvious errors exist. These could be incorrectly named variables, or weighting data that were named or specified incorrectly, for example. Second, `simcadi` computes a distribution of comparison for each person as follows:

$$s_{i*tk} = \sum_{j;j\neq i} \left(w_{ij} \cdot s_{jtk}\right), \tag{1}$$

where w_{ij} denotes the weight between subjects i and j. Our command automatically ensures that the weights add up to one for each subject, $\sum_j w_{ij} = 1 \; \forall i$. Depending on the weighting matrix, the resulting benchmark may be different for each subject. For more details, please see the following Stata Output excerpt:

```
. use weightdata,clear

. list

     +-----------------------------+
     |     name      name2   weight |
     |-----------------------------|
  1. |     Adam       Adam        1 |
  2. |     Adam   Brittany        0 |
  3. |     Adam    Charlie        2 |
  4. | Brittany       Adam        1 |
  5. | Brittany   Brittany        1 |
     |-----------------------------|
  6. | Brittany    Charlie        1 |
  7. |  Charlie       Adam        2 |
  8. |  Charlie   Brittany        0 |
  9. |  Charlie    Charlie        1 |
     +-----------------------------+

. use example_consume, clear

. simcadi consume using weightdata, class(good) id(name) time(1) timevar(time) ///
>         varpartner(name2) detail finger grubel savecomp(compfile, replace)
----------------------
Q: Any obvious errors?
Value variable exists                    ---> consume
Time variable exists                     ---> time
id variable exists                       ---> name
Weight variable (in weight dataset) exists       ---> weight
Categorization exists                    ---> good
Variable name exists in the weight dataset
Variable name2 exists in the weight dataset
A: No obvious errors.
----------------------
Q: Which index should be calculated?
A: The Grubel-Lloyd Index.
A: The Finger-Kreinin Index.
A: The indicator(s) are calculated using shares.
3 = # of distinct values of name
----------------------
Q: Did the user specify a weight dataset?
A: YES ---> weightdata
----------------------
----------------------
Q: Do the weights add up to one for each name (in the raw weight data)
A: NO, we need to adjust the weight...
   ...weights successfully adjusted
----------------------
Calculation of the s_i*kt...
               3 = # of distinct values of good
               27 = # of observations in the bilateral dataset
Merge the weighting scheme with trade data...
----------------------
Q: Are all countries in the weighting scheme matched successfully?
Note: If this is not the case, the weights do not add up to one for each name
A: YES
----------------------
file compfile.dta saved
Calculation of the indices...
```

```
Calculation of the Grubel-Lloyd index...
...DONE
Calculation of the Finger-Kreinin index...
...DONE
Merge the results...
...DONE
──────────────────────
Results were not saved

. list

     +------------------------------------------+
     |     name   time    finger~d    grubel~d  |
     |------------------------------------------|
  1. |     Adam      1    .7777778    .7380952  |
  2. | Brittany      1    .9489489    .9384114  |
  3. |  Charlie      1    .7777778    .5833333  |
     +------------------------------------------+

. use compfile, clear

. list

     +--------------------------------------------------+
     |     name   time    good       comp2       comp1  |
     |--------------------------------------------------|
  1. |     Adam      1    beer    .1111111    .3333333  |
  2. |     Adam      1   bread    .4444445    .3333333  |
  3. |     Adam      1   water    .4444445    .3333333  |
  4. | Brittany      1    beer    .1921922    .2432432  |
  5. | Brittany      1   bread    .4129129    .4054054  |
     |--------------------------------------------------|
  6. | Brittany      1   water    .3948949    .3513514  |
  7. |  Charlie      1    beer    .2222222           0  |
  8. |  Charlie      1   bread    .3888889          .5  |
  9. |  Charlie      1   water    .3888889          .5  |
     +--------------------------------------------------+
```

4 References

Boriah, S., V. Chandola, and V. Kumar. 2008. Similarity Measures for Categorical Data: a Comparative Evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, ed. K. W. Chid Apte, Haesun Park and M. J. Zaki, chap. 21, 243–254.

Boyce, B. R., C. T. Meadow, and D. H. Kraft. 1995. *Measurement in Information Science*. New York: Academic Press.

Bray, J. R., and J. T. Curtis. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological monographs* 27(4): 325–349.

Czekanowski, J. 1909. Zur Differentialdiagnose der Neandertalgruppe. *Korrespondenzblatt der deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte* 40: 44–47.

De Benedictis, L., and L. Tajoli. 2007a. Openness, Similarity in Export Composition, and Income Dynamics. *Journal of International Trade and Economic Development* 16(1): 93–116.

———. 2007b. Economic Integration and Similarity in Trade Structures. *Empirica* 34(2): 117–137.

Dice, L. R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297–302.

Egghe, L. 2010. Good Properties of Similarity Measures and Their Complementarity. *Journal of the American Society for Information Science and Technology* 61(10): 2151–2160.

Fenty, J. 2004. Analyzing Distances. *Stata Journal* 4(1): 1–26.

Finger, J. M., and M. E. Kreinin. 1979. A Measure of 'Export Similarity' and Its Possible Uses. *Economic Journal* 89(356): 905–12.

Goshtasby, A. A. 2012. *Image Registration: Principles, Tools and Methods*, chap. 2, 7–61. London: Springer Science & Business Media.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* (27): 857–871.

Grubel, H. G., and P. J. Lloyd. 1971. The Empirical Measurement of Intra-Industry Trade. *Economic record* 47(4): 494–517.

Jaccard, P. 1912. The Distribution of the Flora in the Alpine Zone. *New Phytologist* 11(2): 37–50.

Johnston, J. W. 1976. Similarity Indices I: What Do They Measure? Paper prepared for the Nuclear Regulatory Commission BNWL-2152, Pacific Northwest Laboratory.

Ružička, M. 1958. Anwendung Mathematisch-Statisticher Methoden in Der Geobotanik (Synthetische Bearbeitung Von Aufnahmen). *Biologia, Bratisl* 13: 647–661.

Sørenson, T. 1948. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content. *Kongelige Danske Videnskabernes Selskab* 5(1-34): 4–7.

Sun, G.-Z., and Y.-K. Ng. 2000. The Measurement of Structural Differences Between Economies: an Axiomatic Characterization. *Economic Theory* 16(2): 313–321.

About the authors

Stephan Huber works at the chair of Joachim Möller at the University of Regensburg and he is doctoral candidate at the University of Trier. His thesis is about disaggregated international bilateral trade flows, and the impact of FDI and international trade on economic development.