MultiDendrograms 5.2

Manual

Sergio Gómez, Alberto Fernández
Universitat Rovira i Virgili, Tarragona (Catalonia)
https://webs-deim.urv.cat/~sergio.gomez/multidendrograms.php

Contents

1. Introduction

MultiDendrograms is a simple yet powerful program to make the Hierarchical Clustering of real data, distributed under an Open Source license. Starting from a proximities (distances or similarities) matrix, *MultiDendrograms* calculates its dendrogram using the most common Agglomerative Hierarchical Clustering algorithms (e.g., Single Linkage, Complete Linkage, Average Linkage and Ward), but also new and more general algorithms (e.g., Versatile Linkage). Additionally, it allows the tuning of many of the graphical representation parameters, and the results may be easily exported to file.

MultiDendrograms implements the variable-group algorithms in [1] to solve the non-uniqueness problem found in the standard pair-group algorithms and implementations. This problem arises when two or more minimum distances between different clusters are equal during the amalgamation process. The standard approach consists in choosing a pair, breaking the ties between distances, and proceeds in the same way until the final hierarchical classification is obtained. However, different clusterings are possible depending on the criterion used to break the ties (usually a pair is just chosen at random!), and the user is unaware of this problem.

The variable-group algorithms group more than two clusters at the same time when ties occur, given rise to a graphical representation called *multidendrogram*. Their main properties are:

- When there are no ties, the variable-group algorithms give the same results as the standard pair-group ones.
- They always give a uniquely determined solution.
- In the multidendrogram representation for the results one can explicitly observe the occurrence of ties during the agglomerative process. Furthermore, the height of any fusion interval (the *bands* in the program) indicates the degree of heterogeneity inside the corresponding cluster.

The main characteristics of *MultiDendrograms* are:

- Multiplatform, runs in Windows, Linux and MacOS.
- Graphical user interface.
- Also command-line direct calculation without graphical user interface.
- Implementation of variable-group algorithms for Agglomerative Hierarchical Clustering.
- Works with positive and negative distances and similarities.
- Many parameters for the customization of the dendrogram layout.
- Navigation through the dendrogram information in a folder-like window.
- Calculation of the corresponding ultrametric matrix.
- Calculation of dendrogram measures such as the cophenetic correlation coefficient.
- Save dendrogram details in text, Newick and JSON formats.
- Save dendrogram plot as JPG, PNG and EPS.

MultiDendrograms page: https://webs-deim.urv.cat/~sergio.gomez/multidendrograms.php
MultiDendrograms in GitHub: https://github.com/sergio-gomez/MultiDendrograms

Please cite [1] if you use *MultiDendrograms* in your publications.

[1] Alberto Fernández and Sergio Gómez: *Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms,* Journal of Classification **25** (2008) 43–65. DOI: 10.1007/s00357-008-9004-x

2. Hierarchical clustering algorithms

MultiDendrograms includes the most commonly used hierarchical clustering algorithms, e.g., Single Linkage, Complete Linkage, Ward and Arithmetic Linkage (also known as UPGMA or Average Linkage). You can also choose between the unweighted and the weighted versions of the algorithms, e.g., Unweighted Centroid and Weighted Centroid. Importantly, it includes a new parameterized algorithm known as Versatile Linkage, which includes Single Linkage, Complete Linkage and Average Linkage as particular cases, and which naturally defines two new algorithms, Geometric Linkage and Harmonic Linkage (hence the convenience to rename UPGMA as Arithmetic Linkage, to emphasize the existence of different types of averages).

The difference between the available hierarchical clustering methods rests in the way the proximity between clusters is defined. During the agglomeration process, the data items are iteratively joined to form clusters, and the idea is just to merge always the clusters which are at the minimum distance (or maximum similarity). However, given two clusters, each one formed by several data observations, there exist many ways of defining the proximity between the clusters from the proximities between their constituent observations. Among these linkage methods, we have the following:

- *Single Linkage*: the distance between clusters equals the minimum distance between all pairs of individuals.
- *Complete Linkage*: the distance between clusters equals the maximum distance between all pairs of individuals.
- *Average* or *Arithmetic Linkage*: the distance between clusters equals the arithmetic mean distance between all pairs of individuals.
- *Versatile Linkage*: the distance between clusters equals the generalized power mean distance between all pairs of individuals. It depends on a parameter $p \in (-\infty, +\infty)$, with the following particular cases:
  - $p \to +\infty$: recovers Complete Linkage.
  - $p = +1$: defines Arithmetic Linkage, which is equivalent to Average Linkage (UPGMA).
  - $p \to 0$: defines Geometric Linkage.
  - $p = -1$: defines Harmonic Linkage.
  - $p \to -\infty$: recovers Single Linkage.
- *Centroid*: the distance between clusters equals the square of the Euclidean distance between the centroids of each cluster.
- *Ward*: the distance between clusters is a weighted squared Euclidean distance between the centroids of each cluster.
- *Beta Flexible*: the distance between clusters is a weighted sum of the distances between clusters in the previous iteration. It depends on a parameter $\beta \in [-1, +1]$, which recovers the Average Linkage when $\beta = 0$.

MultiDendrograms implements all the linkage methods for both distances and similarities, except Centroid and Ward, which are only available for distances.

Please cite [2] if you use *Versatile Linkage* in your publications.

[2] Alberto Fernández and Sergio Gómez: *Versatile linkage: a family of space-conserving strategies for agglomerative hierarchical clustering,* Journal of Classification **37** (2020) 584–597. DOI: 10.1007/s00357-019-09339-z

3. Input data

The data file must represent a proximities (distances or similarities) matrix, like the one in the following table:

	a	b	c	d
a	0	2	4	7
b	2	0	2	5
c	4	2	0	3
d	7	5	3	0

Centroid and Ward methods assume the use of squared Euclidean distances. However, input data in *MultiDendrograms* are not expected to be squared Euclidean distances, and the user can enter non-squared Euclidean distances as with any of the other methods. When Centroid or Ward methods are used, *MultiDendrograms* internally works with squared Euclidean distances and, after the calculations are finished, the square root of the heights in the dendrogram are given as result.

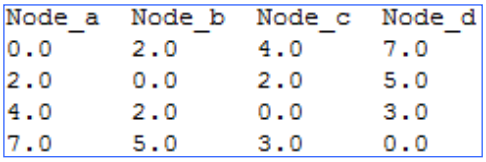
There are three different arrangements these data can be stored in a text file such that *MultiDendrograms* may accept them: matrix, triangular and list formats.

Matrix-like file format

Each line in the text file contains a data matrix row. The characteristics of these files are:

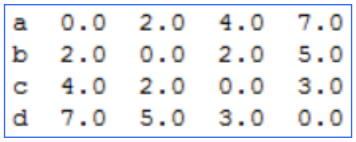
- The matrix must be symmetric
- Within each row, the elements are separated by: spaces (' '), tab character, semicolon (';'), comma (',') or vertical bar ('|').
- It is possible to include the names in an additional first row or column, but not in both.
- If present, the labels of the nodes can not contain any of the previous separators.
- The diagonal elements are used in the non-uniform origin option.
- Blank lines and lines starting with the comment symbol ('#') are discarded.

Some different representations for the previous data could be:



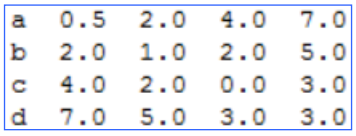
Node_a	Node_b	Node_c	Node_d
0.0	2.0	4.0	7.0
2.0	0.0	2.0	5.0
4.0	2.0	0.0	3.0
7.0	5.0	3.0	0.0

Matrix-like with node labels in first row, data separated by tabs



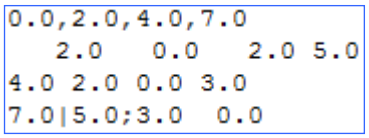
a	0.0	2.0	4.0	7.0
b	2.0	0.0	2.0	5.0
c	4.0	2.0	0.0	3.0
d	7.0	5.0	3.0	0.0

Matrix-like with node labels in the first column, data separated by spaces



a	0.5	2.0	4.0	7.0
b	2.0	1.0	2.0	5.0
c	4.0	2.0	0.0	3.0
d	7.0	5.0	3.0	3.0

Matrix-like with node labels in the first column, data separated by spaces, and non-zero diagonal values



```
0.0,2.0,4.0,7.0
   2.0   0.0   2.0 5.0
4.0 2.0 0.0 3.0
7.0|5.0;3.0  0.0
```

Matrix-like file without node labels, data separated by all kind of separators

Triangular-like file format

The text file contains the lower triangular data of the distances or similarities matrix. The characteristics of these files are:

- Within each row, the elements are separated by: spaces (‘ ’), tab character, semicolon (‘;’), comma (‘,’) or vertical bar (‘|’).
- It is possible to include the names in an additional first row or column, but not in both.
- If present, the labels of the nodes can not contain any of the previous separators.
- The diagonal elements are used in the non-uniform origin option.
- Blank lines and lines starting with the comment symbol (‘#’) are discarded.

Some different representations for the previous data could be:

```
Alice Bob Carol Dave
0
2 0
4 2 0
7 5 3 0
```

Triangular-like with node labels in first row, data separated by tabs

```
a 0
b 2 0
c 4 2 0
d 7 5 3 0
```

Triangular-like with node labels in the first column, data separated by spaces

```
0
2 0
4 2 0
7 5 3 0
```

Triangular-like file without node labels

```
0
2 1
4 2 0
7 5 3 3
```

Triangular-like file without node labels, and non-zero diagonal values

List-like file format

Each line in the text file contains three elements, which represent the labels of two nodes and the distance (or similarity) between them. The characteristics of these files are:

- The separators between the three elements may be: spaces (' '), tab character, semicolon (';'), comma (',') or vertical bar ('|').
- The labels of the nodes can not contain any of the previous separators.
- *MultiDendrograms* accepts either the presence or absence of the symmetric data elements, i.e., if the distance between nodes `a` and `b` is `2.0`, it is possible to include in the list the line "`a b 2.0`", or both "`a b 2.0`" and "`b a 2.0`". If both are present, the program checks if the values are equal.
- Diagonal elements, i.e., values from an element to itself (e.g., "`a a 1.5`"), if present, are used in the non-uniform origin option.
- If there are missing pairs (other than diagonal elements), the default value in the `md.ini` configuration file is assigned.
- Blank lines and lines starting with the comment symbol ('#') are discarded.

For example, three list-like files for the previous data could be:

```
a b 2
a c 4
a d 7
b c 2
b d 5
c d 3
```

Simple list

```
a b 2
a c 4
a d 7
b a 2
b c 2
b d 5
c a 4
c b 2
c d 3
d a 7
d b 5
d c 3
```

Complete list

```
a a 0
a b 2
a c 4
a d 7
b a 2
b b 0
b c 2
b d 5
c a 4
c b 2
c c 0
c d 3
d a 7
d b 5
d c 3
d d 0
```

```
a b 2
a c 4
a d 7
b c 2
b d 5
c d 3
b b 1
d d 3
```

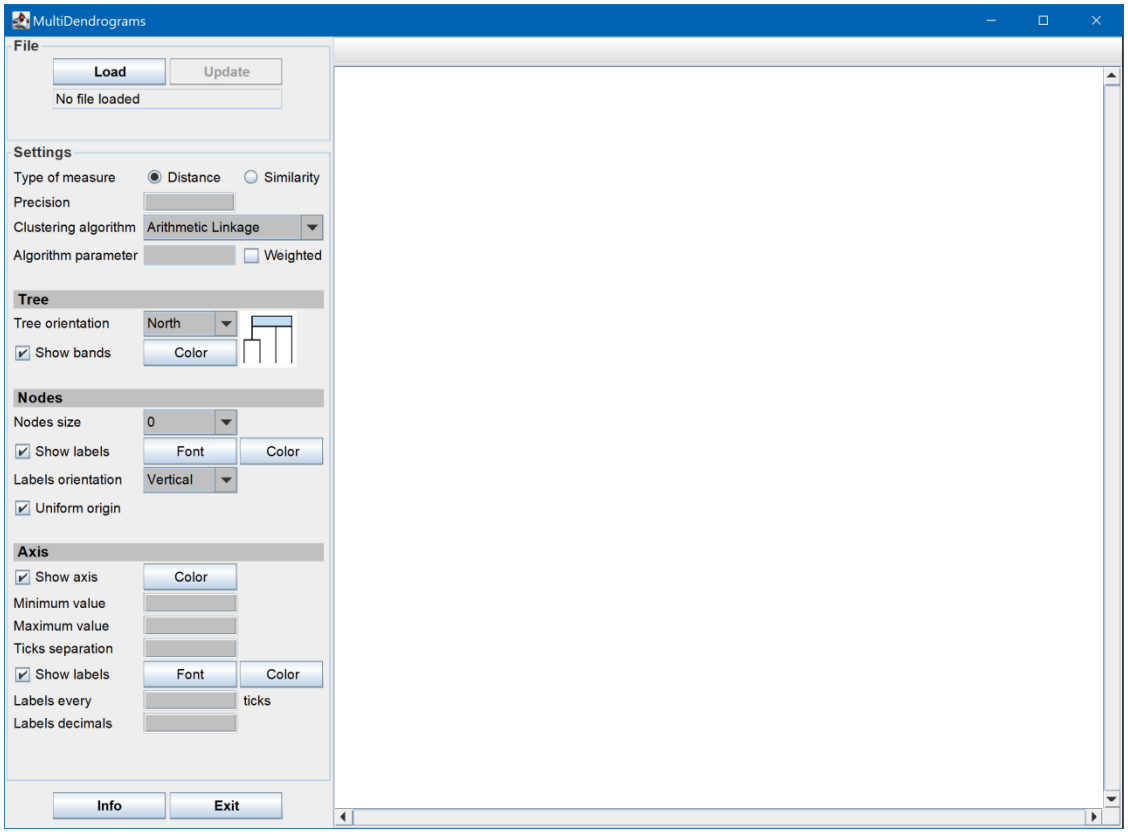
Lists with diagonal elements

4. Loading data

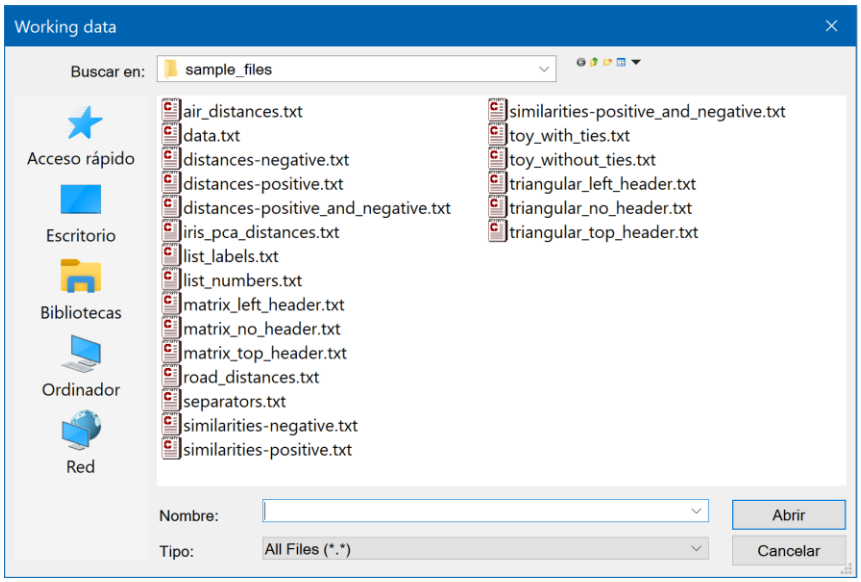
Once we have our data in a compatible format, we can load them into *MultiDendrograms*.

1. Choose the desired settings, mainly the Type of measure and the Clustering algorithm. These settings will be explained in detail in the next sections.

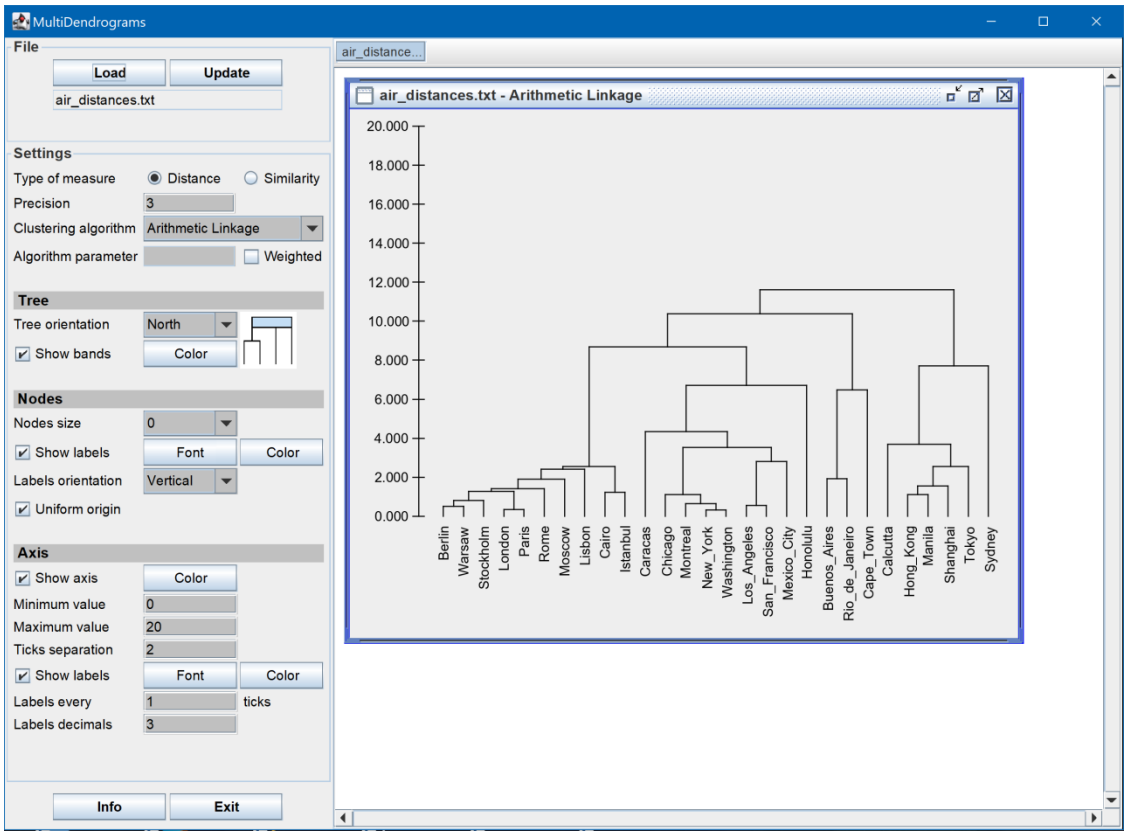
2. Click on the 'Load' button:



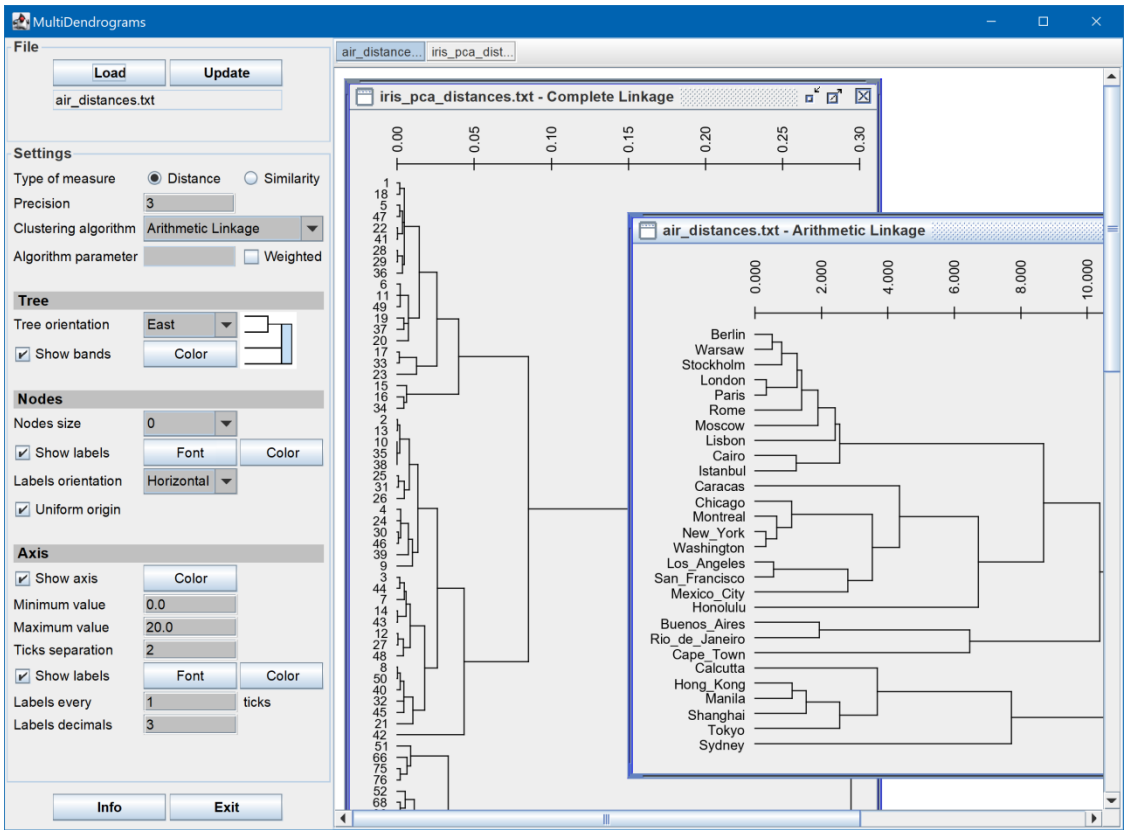
3. Select the file to open and then click on the 'Open' button:



4. Now the data is loaded and its dendrogram representation is shown:

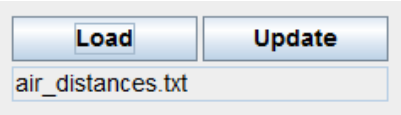


5. Take advantage of the right and bottom scrollbars to handle large dendrograms:

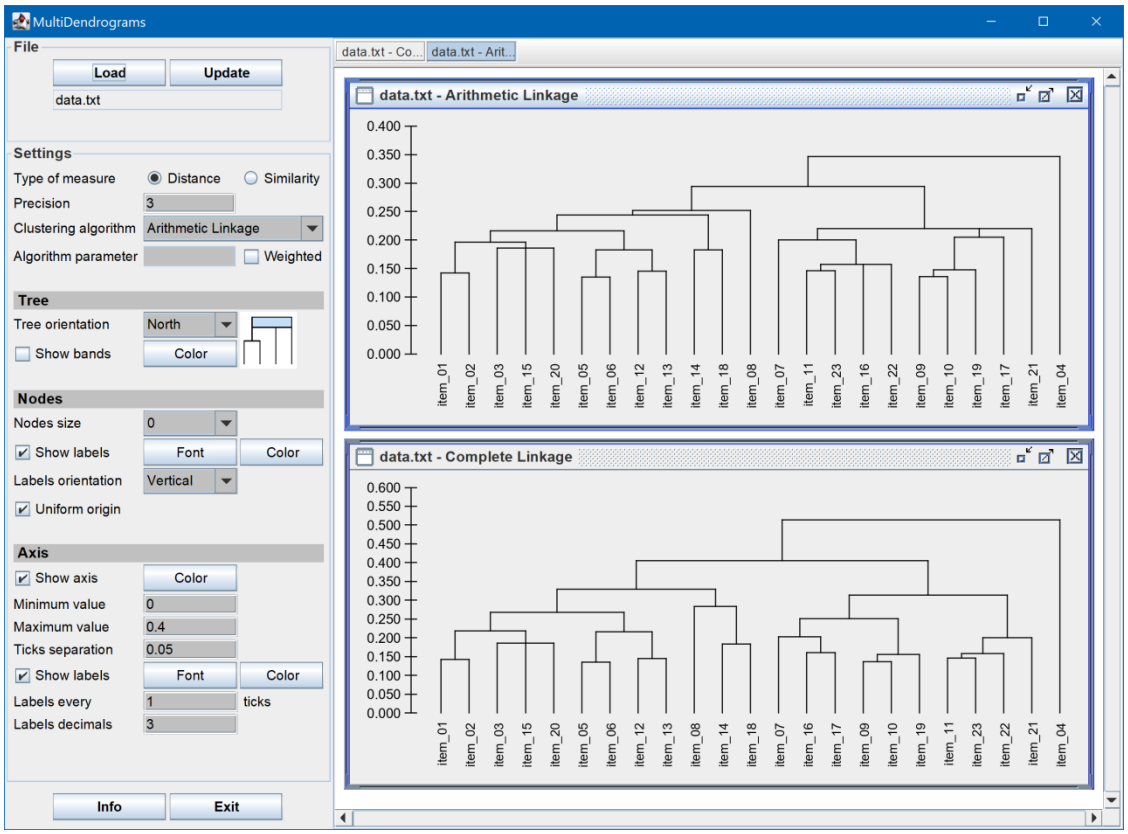


5. Actions

MultiDendrograms only has two action buttons, **Load** and **Update**. **Load** is used to read data from a file and create a new window for the dendrogram, using the current values of the parameters, while **Update** is needed for the actualization of the active dendrogram when one or more parameters are changed. Below these buttons it is shown the name of the data file of the active dendrogram.



It is possible to load the same data file several times, to compare the dendrogram appearance for different parameters settings.

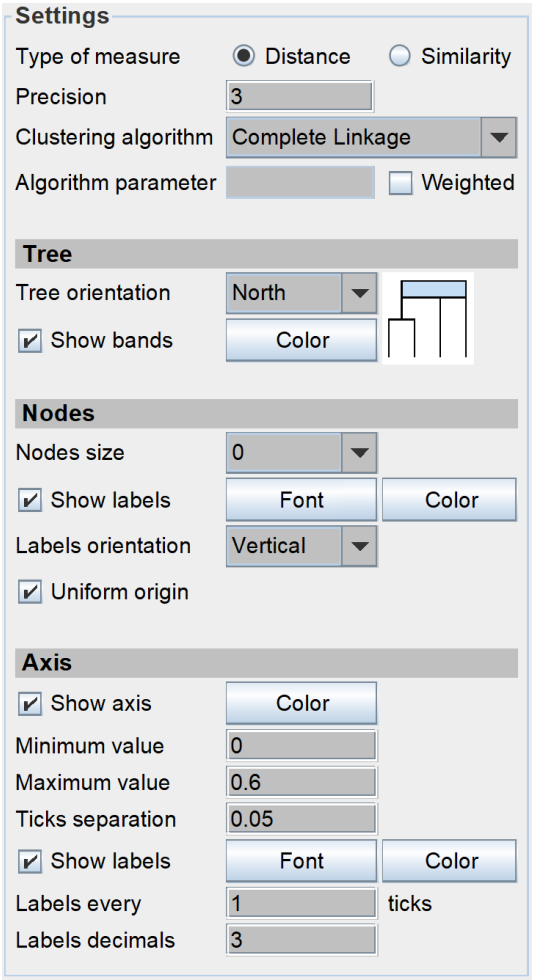


The parameters shown always correspond to the active (selected) dendrogram window.

Finally, there are two additional buttons, **Info** to show information of the program, and **Exit** to quit the application.

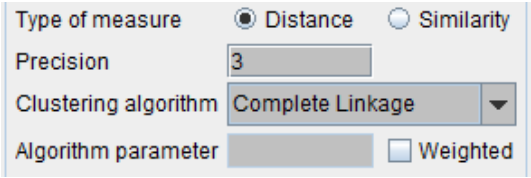
6. Settings

The program automatically applies default values to the parameters depending on the data loaded, which should be adjusted as desired. The following figure shows the settings tab, with four different areas corresponding to the main data representation, tree, nodes and axis settings respectively:

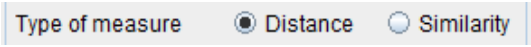


Changes in the main data representation parameters affect the structure of the dendrogram tree, thus it needs to be fully recalculated, operation which may take several seconds, even minutes (depending on the data size and the computer speed). On the other hand, changes in the tree, nodes and axis settings only modify the visual representation of the dendrogram, which are much faster to update.

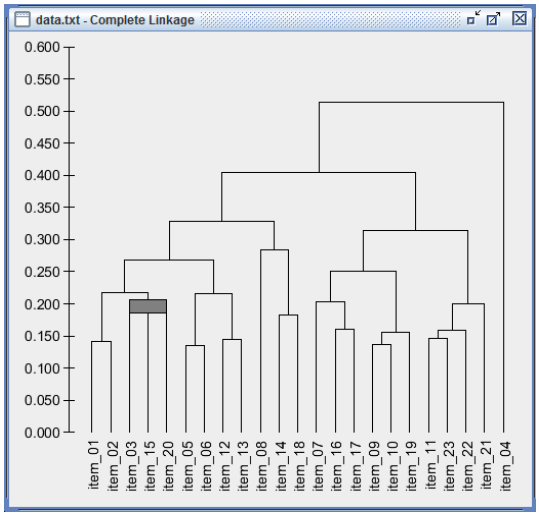
Main data representation settings



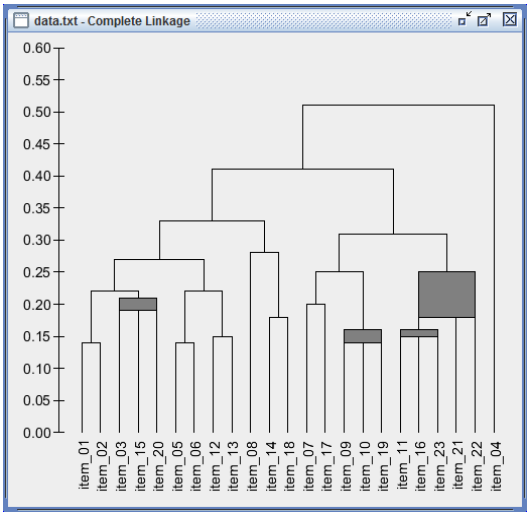
- **Type of measure**: It allows choosing between two kinds of measures, ***distance*** and ***similarity***. Choose between them according to the meaning of the loaded data. With ***distances***, the closer the elements the lower their distance. On the contrary, with ***similarity***, the closer the elements the larger their weight. By default, ***distance*** is selected.



- **Precision**: Number of decimal significant digits of the data and for the calculations. This is a very important parameter, since equal distances at a certain precision may become different by increasing its value. Thus, it may be responsible of the existence of tied distances. The rule should be not to use a precision larger than the resolution given by the experimental setup which has generated the data. By default, the precision is set to that of the data value with the largest number of significant decimal digits.

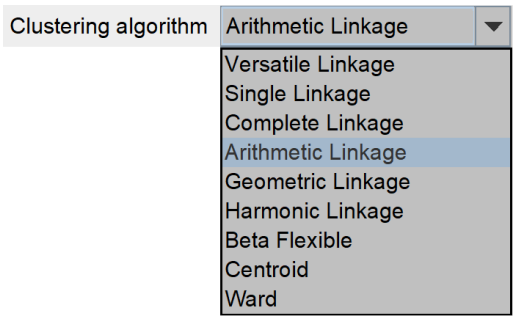


Precision = 3

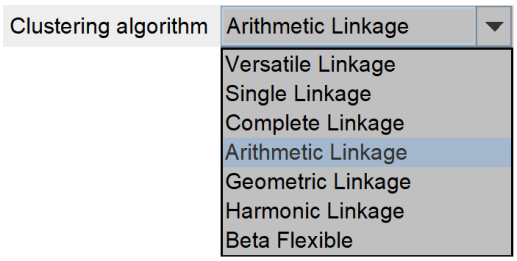


Precision = 2

- **Clustering algorithm**: Nine clustering algorithms are available, ***versatile linkage***, ***single linkage***, ***complete linkage***, ***arithmetic linkage***, ***geometric linkage***, ***harmonic linkage***, ***beta flexible, centroid*** and ***ward***. From them, ***centroid*** and ***ward*** are only available for distances, not for similarities. By default, ***arithmetic linkage*** is selected.

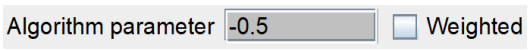


Clustering algorithms for Distance

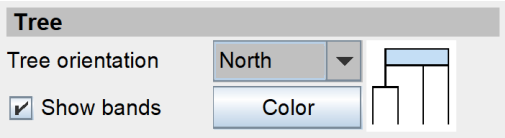


Clustering algorithms for Similarity

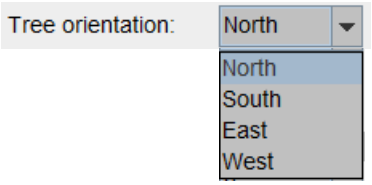
The clustering algorithms ***versatile linkage*** and ***beta flexible*** require the specification of an ***algorithm parameter***, which can take any real value in the case of ***versatile linkage*** and any real value in the range [-1, 1] in the case of ***beta flexible***. Additionally, you can choose between the ***unweighted*** and ***weighted*** variants of all the clustering algorithms.

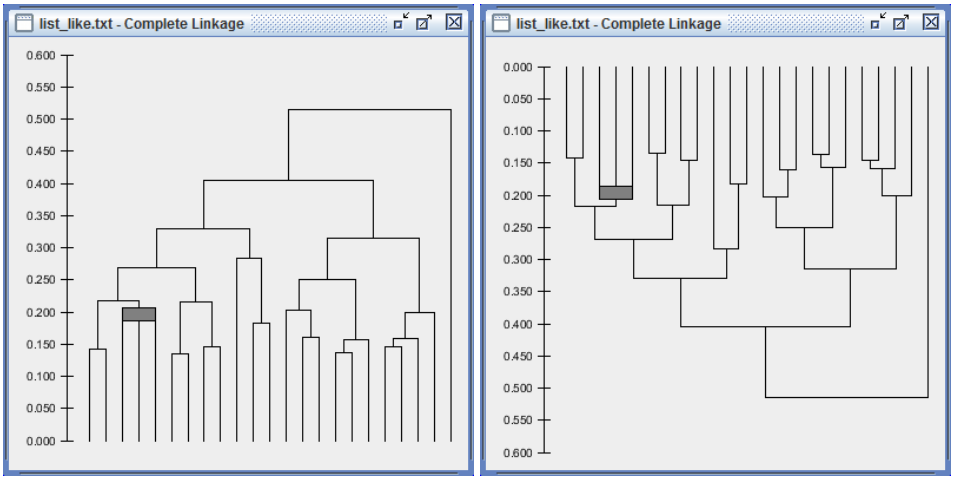


Tree settings



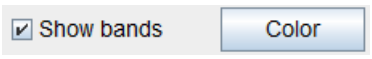
- **Tree orientation**: Four orientations are available, ***north***, ***south***, ***east*** and ***west***, which refer to the relative position of the root of the tree. By default, ***north*** is selected.

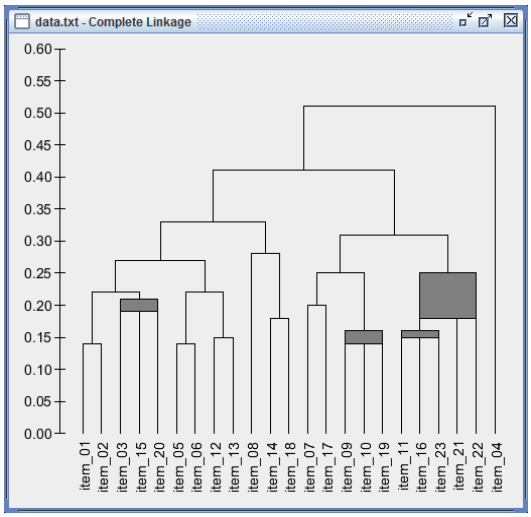




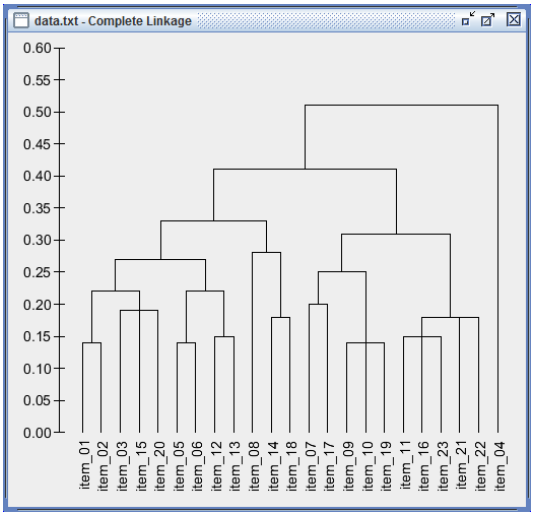
North South West East

- **Show bands**: It allows showing a band or not in case of tied minimum distances between three or more elements, and selecting the color of the band. If selected, the bands show the heterogeneity of all the distances between the clustered elements. Otherwise, the elements are grouped at their minimum distance. By default, ***show bands*** is selected, and its default color is ***light gray***.





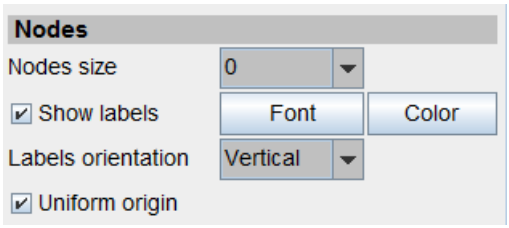
Show bands



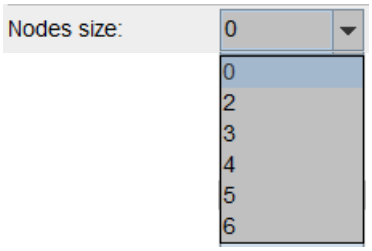
Do not show bands

Let us explain the meaning of the bands. In *MultiDendrograms*, if several pairs of elements share the same minimal distance, they are clustered together in one step. For instance, suppose that the minimal distance is 0.4, and that they correspond to the tied pairs (A,B) and (B,C). *MultiDendrograms* puts them together in the same cluster (A,B,C) at height 0.4. However, if the distance (A,C) is 0.5, it is possible to represent the cluster (A,B,C) as a rectangle which spans between heights 0.4 and 0.5, thus showing the heterogeneity of the clustered elements.

Nodes settings



- **Nodes size**: Six different node sizes are available. By default, 0 is selected (i.e., nodes not shown):

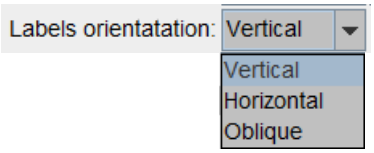


Node sizes from 0 to 6

- **Show labels**: It allows showing or not the labels of the nodes, and selecting their color and font. By default, ***show labels*** is selected, the font is ***Arial*** and the color is ***black***:

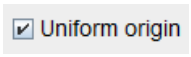
Show labels Do not show labels

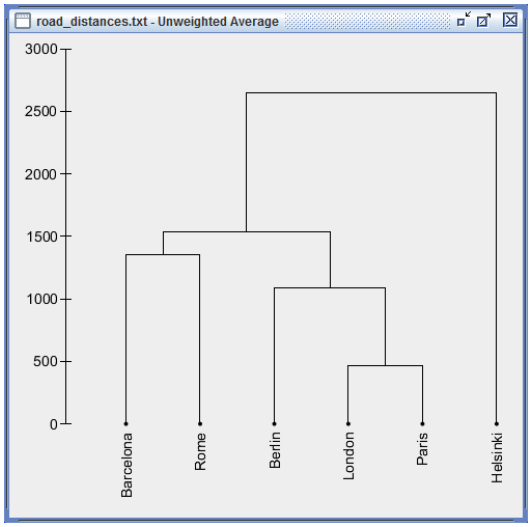
- **Labels orientation**: Three orientations are available: ***vertical***, ***horizontal*** and ***oblique***. By default, ***vertical*** is selected:



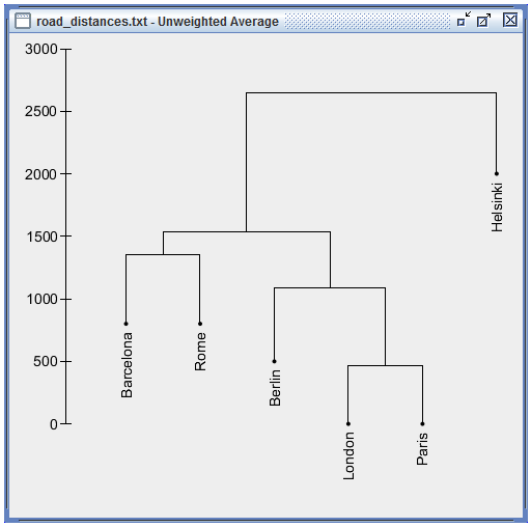
Horizontal Oblique Vertical

- **Uniform origin**: It allows choosing between uniform and non-uniform origins of the nodes. By default, ***uniform origin*** is selected. In the non-uniform mode, the height at which the nodes are drawn is taken from the data file, in the form of values (distances or similarities) between an element and itself (e.g., "`a a 3.5`" in list form, or diagonal values in matrix and triangular forms). For missing diagonal values, height infinity is chosen (minus infinity in the case of distances, plus infinity in the case of similarities):





Uniform origin

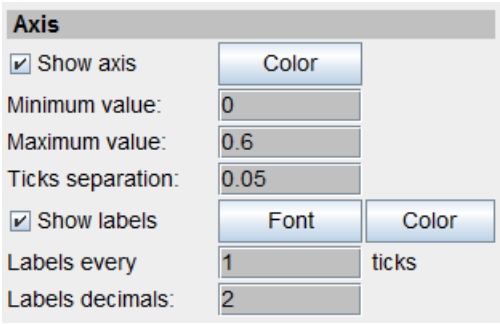


Non-uniform origin

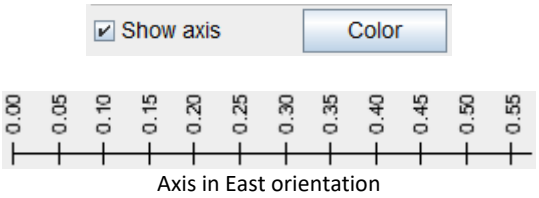
```
Barcelona 800   1864  3448  1494  1036  1355
Berlin    1864  500   1620  1099  1068  1505
Helsinki  3448  1620  2000  2680  2636  2864
London    1494  1099  2680  0     462   1875
Paris     1036  1068  2636  462   0     1424
Rome      1355  1505  2864  1875  1424  800
```

Note the values in the diagonal of the matrix

Axis settings

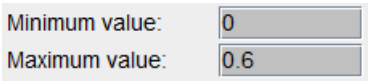


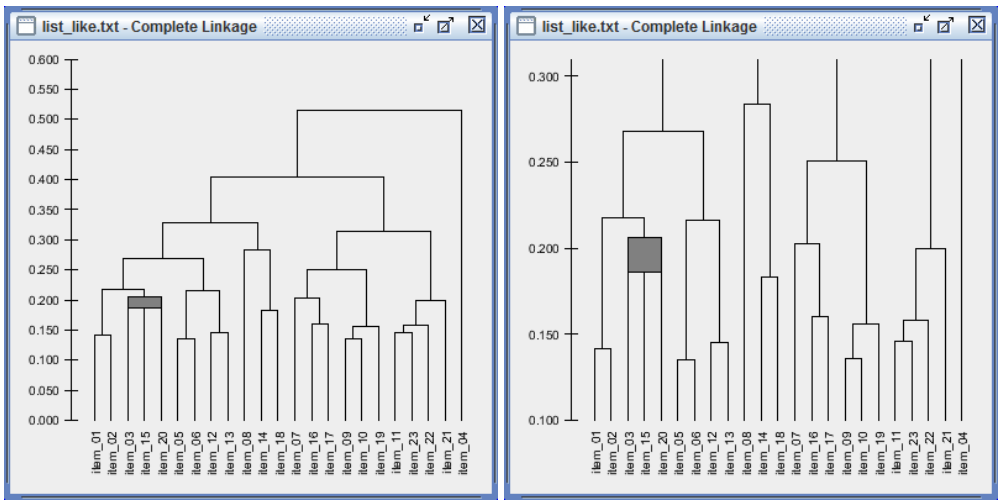
- **Show axis**: It allows showing or not the axis, and selecting its color. By default, ***show axis*** is selected and the selected color is ***black***.



Axis in East orientation

- **Minimum value / Maximum value**: They allow choosing the minimum and maximum value of the axis, respectively. They also affect the view of the dendrogram. The default values are calculated from the data.

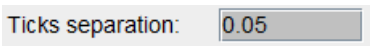


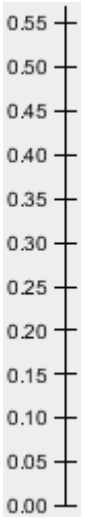


Minimum = 0, Maximum = 0.6

Minimum = 0.10, Maximum = 0.31

- **Ticks separation**: It allows choosing the separation between consecutive ticks of the axis. The default value is calculated from the data:

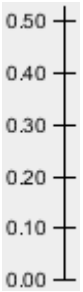




Ticks separation = 0.05

Ticks separation = 0.10

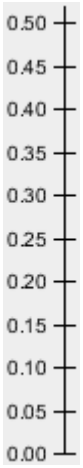
- **Show labels**: It allows showing or not the labels of the axis, and selecting their color and font. By default, ***show labels*** is selected, the font is ***Arial*** and the color is ***black***.



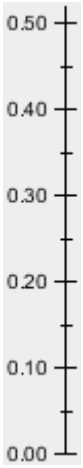
Show labels

Do not show labels

- **Labels every … ticks**: Number of consecutive ticks to find the next labeled tick. By default is set to 1.

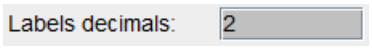


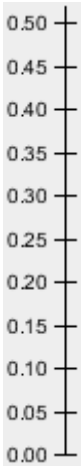
Labels every 1 tick



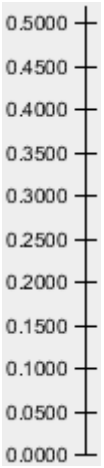
Labels every 2 ticks

- **Labels decimals**: Number of decimal digits of the tick labels. By default it is set equal to the ***precision*** parameter.





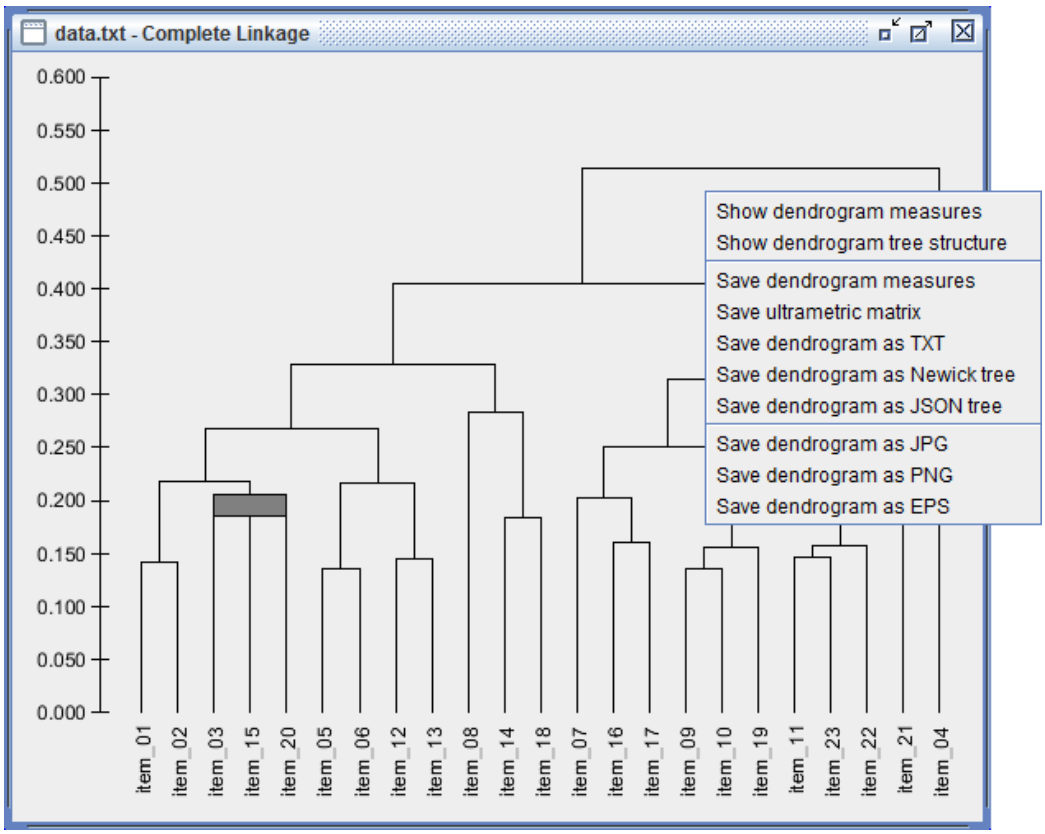
Labels decimals = 2



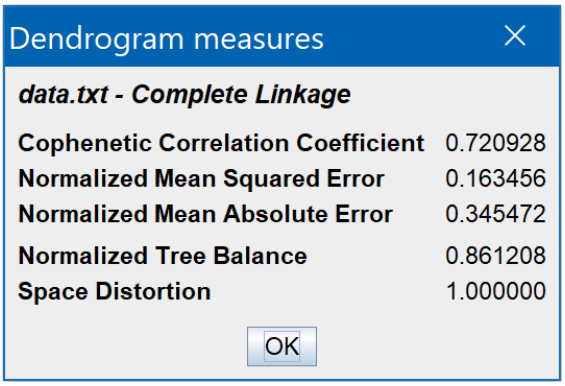
Labels decimals = 4

7. Analyzing and exporting results

The contextual menu, available by right-clicking the dendrogram windows, gives access to several options for analyzing and exporting the results to file.

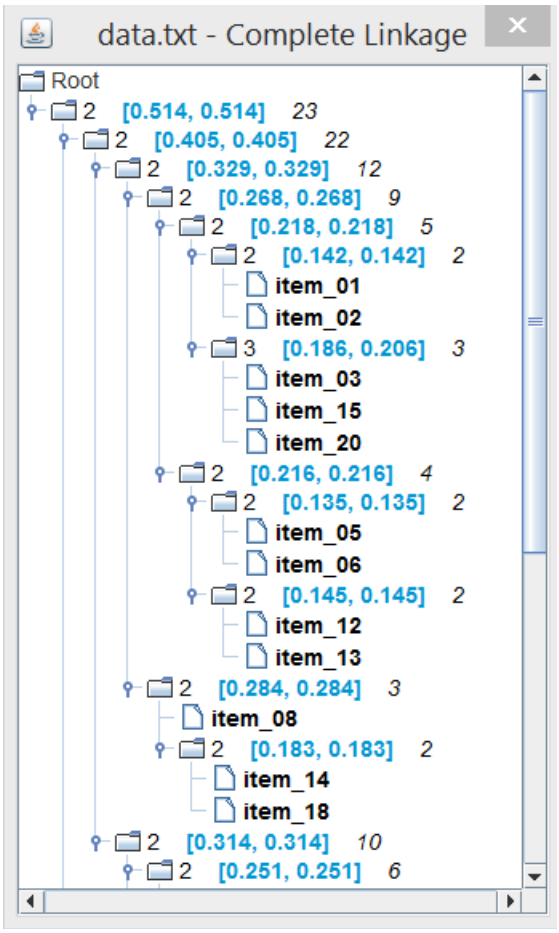


- **Show dendrogram measures**: Calculates several mesures of the active dendrogram: the ***cophenetic correlation coefficient***, the ***normalized mean squared error***, the ***normalized mean absolute error***, the ***normalized tree balance*** and the ***space distortion***.



Dendrogram measures

- **Show dendrogram tree structure:** Opens a window which contains all the information of the dendrogram in a navigable folder-like structure:



The available information in the details window is:

- Number of children of each interior node of the dendrogram. The interior nodes in the dendrogram representation correspond to the clusters found during the agglomeration process, and the children may be other interior nodes or data items.
- Minimum and maximum distances at which the children of an interior node are joined to form a new cluster. These values may only be different in case of tied distances, which become a band in the multidendrogram representation.
- Number of data items (leaves of the tree) under each interior node of the dendrogram.
- List of children of each interior node, which may be either interior nodes or data items.

- **Save dendrogram measures**: Saves the dendrogram measures to a text file.

```
Cophenetic Correlation Coefficient : 0.720928
Normalized Mean Squared Error      : 0.163456
Normalized Mean Absolute Error     : 0.345472
Normalized Tree Balance            : 0.861208
Space Distortion                   : 1.000000
```

Dendrogram measures

The information provided is the same as in ***Show dendrogram measures***.

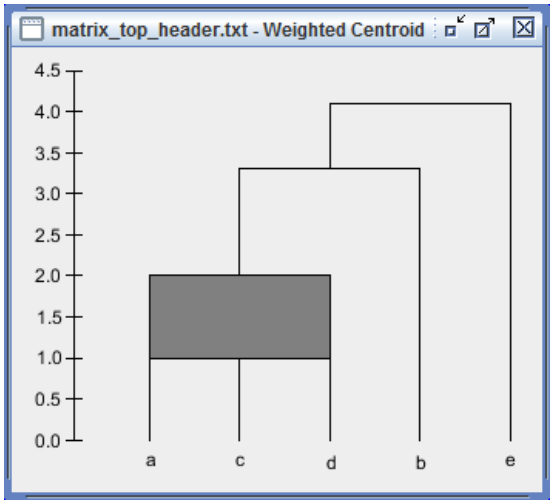
- **Save ultrametric matrix**: Calculates the ultrametric matrix corresponding to the loaded data and saves it to a text file in matrix form, with the nodes labels in the first row. This text file can then be easily loaded into any text editor or spreadsheet application (e.g., Microsoft Excel).

a	b	c	d	e
0.0	1.2	1.0	2.0	5.0
1.2	0.0	3.0	7.0	4.0
1.0	3.0	0.0	1.0	8.0
2.0	7.0	1.0	0.0	6.0
5.0	4.0	8.0	6.0	0.0

Original distances matrix

a	c	d	b	e
0.0	1.0	1.0	3.3	4.1
1.0	0.0	1.0	3.3	4.1
1.0	1.0	0.0	3.3	4.1
3.3	3.3	3.3	0.0	4.1
4.1	4.1	4.1	4.1	0.0

Ultrametric matrix



- **Save dendrogram as TXT**: Saves the dendrogram details to a text file.

```
+ 2  [4.1, 4.1]  5
  + 2  [3.3, 3.3]  4
    + 3  [1.0, 2.0]  3
      *  a
      *  c
      *  d
    *  b
  *  e
```

Dendrogram in text format

The information provided in this format is equivalent to the one in ***Show dendrogram details***, with symbol '+' marking interior nodes of the dendrogram and '*' marking the data items (leaves).

- **Save dendrogram as Newick tree**: Saves the dendrogram details in Newick tree format (see https://en.wikipedia.org/wiki/Newick_format).

```
(((a:1.0,c:1.0,d:1.0):2.3,b:3.3):0.8,e:4.1);
```

Dendrogram in Newick format

In this format only names and branches' lengths are used, and the information given by the bands is lost. However, it has the advantage that it is a standard format used in many other applications, thus allowing their use to generate other graphical representations.

- **Save dendrogram as JSON tree**: Saves the dendrogram details in JSON format.

```
{
 "name": "", "height": 4.1, "margin": 0.0, "length": 0.0,
 "children": [
  {
   "name": "", "height": 3.3, "margin": 0.0, "length": 0.8,
   "children": [
    {
     "name": "", "height": 1.0, "margin": 1.0, "length": 2.3,
     "children": [
      {"name": "a", "height": 0.0, "margin": 0.0, "length": 1.0, "size": 1},
      {"name": "c", "height": 0.0, "margin": 0.0, "length": 1.0, "size": 1},
      {"name": "d", "height": 0.0, "margin": 0.0, "length": 1.0, "size": 1}
     ]
    },
    {"name": "b", "height": 0.0, "margin": 0.0, "length": 3.3, "size": 1}
   ]
  },
  {"name": "e", "height": 0.0, "margin": 0.0, "length": 4.1, "size": 1}
 ]
}
```

Dendrogram in JSON format

JSON is a language-independent open standard format especially suitable to the exchange of information between different applications, e.g., between browsers and servers (see https://en.wikipedia.org/wiki/JSON). In this format all the information to build the dendrogam is available: names of data items (leaves), heights of data and clusters, branches' lengths (redundant, since they can be calculated from heights, but convenient for representations which only make use of the lengths) and margins (the size of the bands).

- **Save dendrogram as JPG, PNG, EPS**: It is also possible to save the image of the dendrogram in three different formats (JPG, PNG and EPS) using their corresponding ***Save dendrogram as*** context menu items. The image reproduces the dendrogram as it is represented in the corresponding window, preserving the current aspect ratio and size.

8. Command-line direct calculation

It is possible to use *MultiDendrograms* in command-line mode to calculate the dendrogram without the graphical interface. This is useful in several situations:

- To automate the generation of many dendrograms using scripts.
- When there is no need of a plot of the dendrogram.
- When the plot of the dendrogram is to be performed with a different program.
- When the number of elements is too large to allow a graphical representation.
- To be able to call *MultiDendrograms* from a different application.

The input parameters of a command-line call are:

- The name of the input data file, in matrix, triangular or list format.
- The proximity type: distances or similarities.
- The precision, i.e., the number of decimal significant digits of the data and for the calculations. This parameter is optional, if not given it is calculated from the data. However, the rule should be not to use a precision larger than the resolution given by the experimental setup which has generated the data.
- The clustering method: versatile linkage, single linkage, complete linkage, arithmetic linkage, geometric linkage, harmonic linkage, centroid, ward or beta flexible.
- The parameter of the method (only for versatile linkage and beta flexible).
- Choose between weighted or unweighted method.
- The type of origin: uniform or non-uniform. In uniform type the diagonal values of the data are discarded and all elements are assigned the same height, otherwise the diagonal values define the heights of the elements.

The output results are:

- A file with the dendrogram measures.
- A file with the dendrogram tree in text format.
- A file with the dendrogram in Newick format.
- A file with the dendrogram in JSON format.
- A file with the ultrametric matrix.
- The parameters and the dendrogram measures are also printed to standard output.

The syntax of a command-line direct calculation is:

java -jar multidendrograms.jar -direct PARAMETERS

where the PARAMETERS are:

FILE_NAME PROX_TYPE [PRECISION] METHOD [METHOD_P] [WEIGHTED] [ORIGIN]

The details are given in Appendix A.

APPENDIX A. Requirements, installation and execution

Requirements

To run *MultiDendrograms* it is necessary to have installed a recent version of the Java Runtime Environment (JRE) or the Java Development Kit (JDK). The minimum version of Java needed is Java 8, but it is recommended to use Java 12 or higher. From Java 9 above, the application is scaled using the system scaling, thus avoiding problems with small fonts or windows when using very high resolution (e.g., 4K) screens. You can check if Java is already in your computer and its version following these steps:

1. Open a console shell or command prompt (in Windows: Win+R -> cmd -> Enter)
2. Type: java -version

We recommend the installation of OpenJDK instead of the Oracle versions of Java, due to the changes in license. There are several options to install OpenJDK 12:

- OpenJDK: https://jdk.java.net/12/, requires manual installation (ZIP and TAR.GZ files)
- Zulu: https://www.azul.com/downloads/zulu, provides installers for the main platforms (MSI for Windows, DMG for MacOS, RPM and DEB for Linux)
- OpenJDKBuilds: https://github.com/ojdkbuild/ojdkbuild, installers for Windows
- AdoptOpenJDK: https://adoptopenjdk.net, installers for Windows

Installation

MultiDendrograms does not require installation. Just unpack the main ZIP file into a folder using any unzip program, e.g., 7-zip, iZarc, WinRAR or WinZip.

Basic execution

- Windows: double-click **multidendrograms.bat** or **multidendrograms.jar**
- Linux: run **multidendrograms.sh** or **multidendrograms.jar**
- MacOS: double-click **multidendrograms.jar** or run **multidendrograms.sh**

Advanced execution

In the command-line:

java -jar multidendrograms.jar [options]

The program accepts these options:

```
-h | -help
        Syntax help

-loglevel  LEVEL
        Sets the verbosity level of the logger
                LEVEL: verbosity level, one of
                        OFF, SEVERE, WARNING, INFO, CONFIG,
                        FINE, FINER, FINEST, ALL
                        Default value of LEVEL: WARNING
```

-direct NAME PROX_TYPE [PRECISION] METHOD [METHOD_P] [WEIGHTED] [ORIGIN]
Direct calculation of the multidendrogram without graphic interface
NAME: name of the data file
PROX_TYPE: proximity type, one of
D, DIST, DISTANCE, DISTANCES
S, SIM, SIMILARITY, SIMILARITIES
PRECISION: number of decimal significant digits, auto if missing value
METHOD: agglomeration type, one of
VL, VERSATILE_LINKAGE
SL, SINGLE_LINKAGE
CL, COMPLETE_LINKAGE
AL, ARITHMETIC_LINKAGE
GL, GEOMETRIC_LINKAGE
HL, HARMONIC_LINKAGE
CD, CENTROID
WD, WARD
BF, BETA_FLEXIBLE
METHOD_P: method parameter necessary for
VL, VERSATILE_LINKAGE
BF, BETA_FLEXIBLE, between -1 and +1
Default value for METHOD_P: 0
WEIGHTED: weighted method, one of
W, WEIGHTED
UW, UNWEIGHTED
Default value for WEIGHTED: UNWEIGHTED
ORIGIN: origin type, one of
UO, UNIFORM_ORIGIN
NUO, NON_UNIFORM_ORIGIN
Default value: UNIFORM_ORIGIN

There are sample script files to show some direct calculations:
- Windows: **multidendrograms-cmd.bat**
- Linux and MacOS: **multidendrograms-cmd.sh**

Examples:

java -jar multidendrograms.jar
java -jar multidendrograms.jar -loglevel OFF
java -jar multidendrograms.jar -direct data.txt DISTANCES 3 Ward
java -jar multidendrograms.jar -direct data.txt SIMILARITIES 3 Complete_Linkage
java -jar multidendrograms.jar -direct data.txt D CL
java -jar multidendrograms.jar -direct data.txt D 3 CL
java -jar multidendrograms.jar -direct data.txt D 3 Versatile_Linkage +1
java -jar multidendrograms.jar -direct data.txt D 3 VL -1 W
java -jar multidendrograms.jar -direct data.txt D CL UO
java -jar multidendrograms.jar -direct data.txt D 3 CL NUO

APPENDIX B. Customization of the graphical user interface

In the `ini` folder, under the installation directory, there is a configuration file, `md.ini`, which is used to define many of the characteristics of the graphical user interface (GUI) of *MultiDendrograms*. It is a text file that you may edit at your convenience. After any modification, the changes will take effect the next time you start *MultiDendrograms*. The most important uses are the selection of the language and changing the font sizes, styles and colors of the text in the application.

Language

By default, the graphical user interface of *MultiDendrograms* is shown in English. Currently, it is possible to choose between the following languages:

- English
- Catalan
- Spanish
- German

There is a language file (e.g., `lang_english.l`) in the `ini` folder for each of the languages available. The selection of the language is made in the first lines of the configuration file `md.ini`. To change the selected language, just open `md.ini` in an editor, uncomment the desired language and comment the rest (comments start with a '#' character). For example, to choose English the contents should be:

```
# Language file selection
language = ini/lang_english.l
#language = ini/lang_catalan.l
#language = ini/lang_spanish.l
#language = ini/lang_deutsch.l
```

To translate *MultiDendrograms* to other languages, just create a new language file `lang_xxx.l`, containing the translation of all the lines in any other language file, and add the corresponding line to the configuration file `md.ini`. If you send us your new language file, we can include it in future versions of *MultiDendrograms*.

Size

MultiDendrograms was initially configured to be used in screens with resolutions starting from 1024x768 pixels. In the last years, with the advent of high-resolution screens, the text in *MultiDendrograms* may appear very small, being difficult to be read. To alleviate this problem, we have prepared configuration files for three different font sizes: 10, 12 and 14 pt:

- `md_size10.ini`
- `md_size12.ini`
- `md_size14.ini`

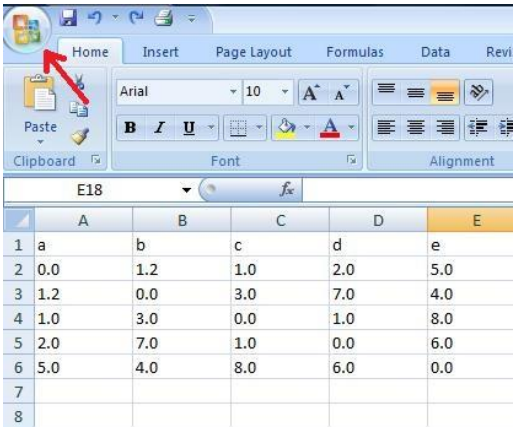
To select one of them, e.g., that of size 14pt, simply rename `md.ini` to `md_old.ini` and copy the file `md_size14.ini` into `md.ini`. Of course, you may edit it as you desire to adapt any parameter you want.

You may also modify the size of the interface by changing the `scalingFactor` parameter in `md.ini`. *MultiDendrograms* detects the presence of high-resolution screens and scales its interface accordingly, but the best option is to use Java 9 or above, as previously explained.

APPENDIX C. Preparing input data with Microsoft Excel

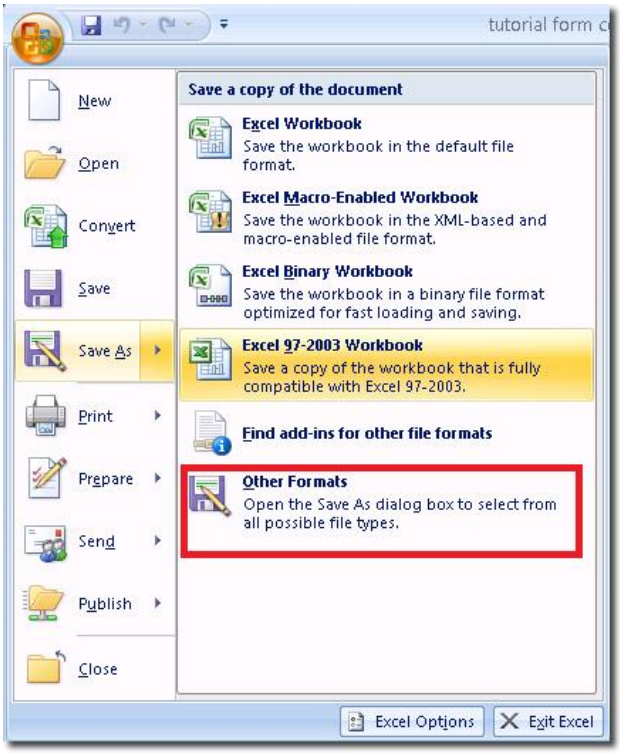
MultiDendrograms cannot load data directly from a Microsoft Excel (or similar) file, we first need to save our data in a compatible format. We will assume you have Microsoft Excel 2007, but similar procedures apply to other versions and similar programs.

1. Click on the button with the Microsoft Office logo:

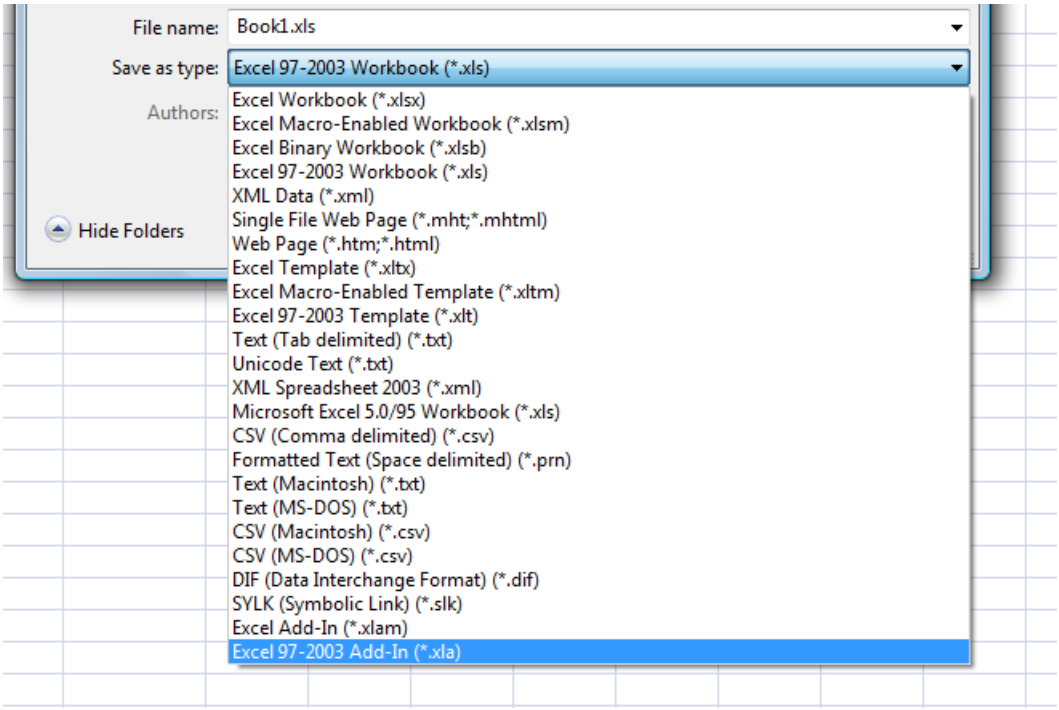


	A	B	C	D	E
1	a	b	c	d	e
2	0.0	1.2	1.0	2.0	5.0
3	1.2	0.0	3.0	7.0	4.0
4	1.0	3.0	0.0	1.0	8.0
5	2.0	7.0	1.0	0.0	6.0
6	5.0	4.0	8.0	6.0	0.0
7					
8					

2. Select the option ‘Save As’ and then ‘Other Formats’:



3. Now choose the file format to save the file:

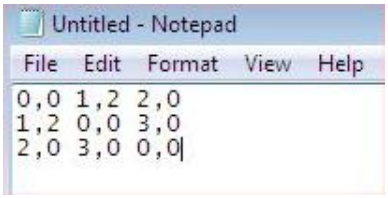


The compatible formats are the following:

- Text (Tab delimited)
- Unicode Text
- CSV (Comma delimited)
- Formatted text (Space delimited)
- Text (Macintosh)
- Text (MS-DOS)
- CSV (Macintosh)
- CSV (MS-DOS)

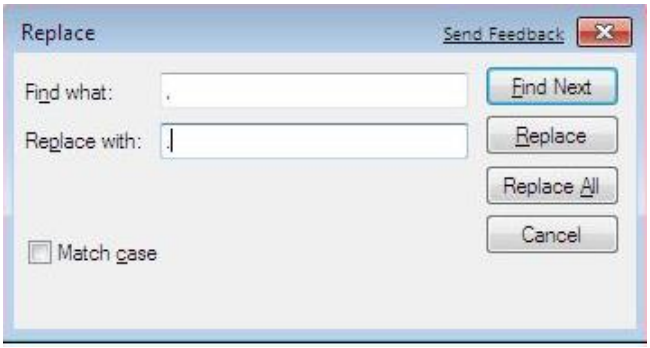
4. *MultiDendrograms* needs that decimal numbers use the character '**.**' as the decimal symbol, e.g., "3.1416". Unfortunately, some regional system configurations use different decimal symbols, e.g., in Spanish it is '**,**' as in "3,1416". In these cases, the previously exported file has to be edited to change the decimal symbol to '**.**':

   - Open the file with Notepad (or any other file editor):

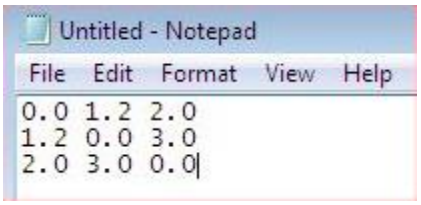


- Go to the ‘Edit’ menu and select ‘Replace’:

- Replace all appearances of ‘,’ with ‘.’:



- Save the modified file:



APPENDIX D. History of changes

MultiDendrograms 5.2
- Input of the power value for Versatile Linkage directly through the algorithm parameter
- Sign of the algorithm parameter unchanged for similarity data

MultiDendrograms 5.1
- Automatic scaling of graphical user interface
- Upgraded to Java 8; OpenJDK 12 or above recommended

MultiDendrograms 5.0
- Reorganization of clustering algorithms
- New parameterized Versatile Linkage and Beta Flexible clustering algorithms
- New Geometric Linkage and Harmonic Linkage clustering algorithms
- Calculation of tree balance and space distortion
- Save dendrogram measures to file

MultiDendrograms 4.1
- Export dendrograms to JSON format

MultiDendrograms 4.0
- Graphical user interface at different sizes
- Positive and negative distances and similarities
- Uniform and non-uniform origin of nodes
- Improved configuration file
- Translation to German
- Improved performance

MultiDendrograms 3.2
- New format for dendrogram navigation and save as text file

MultiDendrograms 3.1
- Data in triangular form

MultiDendrograms 3.0
- Scrollbars in dendrograms panel
- Command-line direct calculation of multidendrogram
- Ward hierarchical clustering
- Check if new version is available
- Confirmation before closing
- Improved performance
- Major source code refactoring

MultiDendrograms 2.1
- Export dendrograms to Newick format
- Show calculation progress
- Improved GUI
- Improved performance

MultiDendrograms 2.0

- Completely new multiplatform (Windows, Linux, MacOS, etc.) application
- Added Graphical User Interface (GUI)
- Control of the dendrogram appearance
- Navigation through the dendrogram details
- Accepts distance and similarity matrices
- Export dendrograms to JPG, PNG and EPS
- Calculation of ultrametric deviation measures

MultiDendrograms 1.0

- Windows command-line application to compute multidendrograms
- Windows command-line application to compute ultrametric matrices
- Windows command-line application to generate EPS plots

APPENDIX E. Request, comments, bugs and acknowledgements

If you have any comment about *MultiDendrograms*, e.g., to request some functionality in future versions, or if you find a bug, please send us an email to any of the following addresses:

- Sergio Gómez: sergio.gomez@urv.cat
- Alberto Fernández: alberto.fernandez@urv.cat

In the case of bugs, please send us all the information needed to reproduce it: a detailed description, a sample data file, the version of the program, the operating system, the Java version, the parameters, the series of actions which result in the bug, the expected result, snapshots, etc. Don't worry if you cannot supply all this information from the beginning, let's just start a conversation and try to find the solution.

Finally, we want to acknowledge all the people who have contributed to the development of this application: Justo Montiel and David Torres, who developed the first versions of *MultiDendrograms* as part of their studies in Computer Science at Universitat Rovira i Virgili (Tarragona) under our supervision; Luce Prignano, for being one of the first users and providing useful comments; Roger Gómez and Mireia Gómez for the translation to German; Franco Lancia for his interest in *MultiDendrograms*, which has resulted in an incorporation of our algorithms in his T-Lab software for text analysis. We also want to thank all the people who have trusted in *MultiDendrograms* for their research, even citing in their publications the program and our papers appeared in *Journal of Classification*.

APPENDIX F. License

MultiDendrograms is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 2.1 of the License, or (at your option) any later version.

MultiDendrograms is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with *MultiDendrograms*; if not, see http://www.gnu.org/licenses/lgpl.html.