User guide
Toolbox for clustering analysis of seismic data

Juan Flórez*, Diana Perdomo*, Hernán Benítez*[1] , Alberto Benavides*,
Olga Baquero**, Jiber Quintero**, Elkin Salcedo**

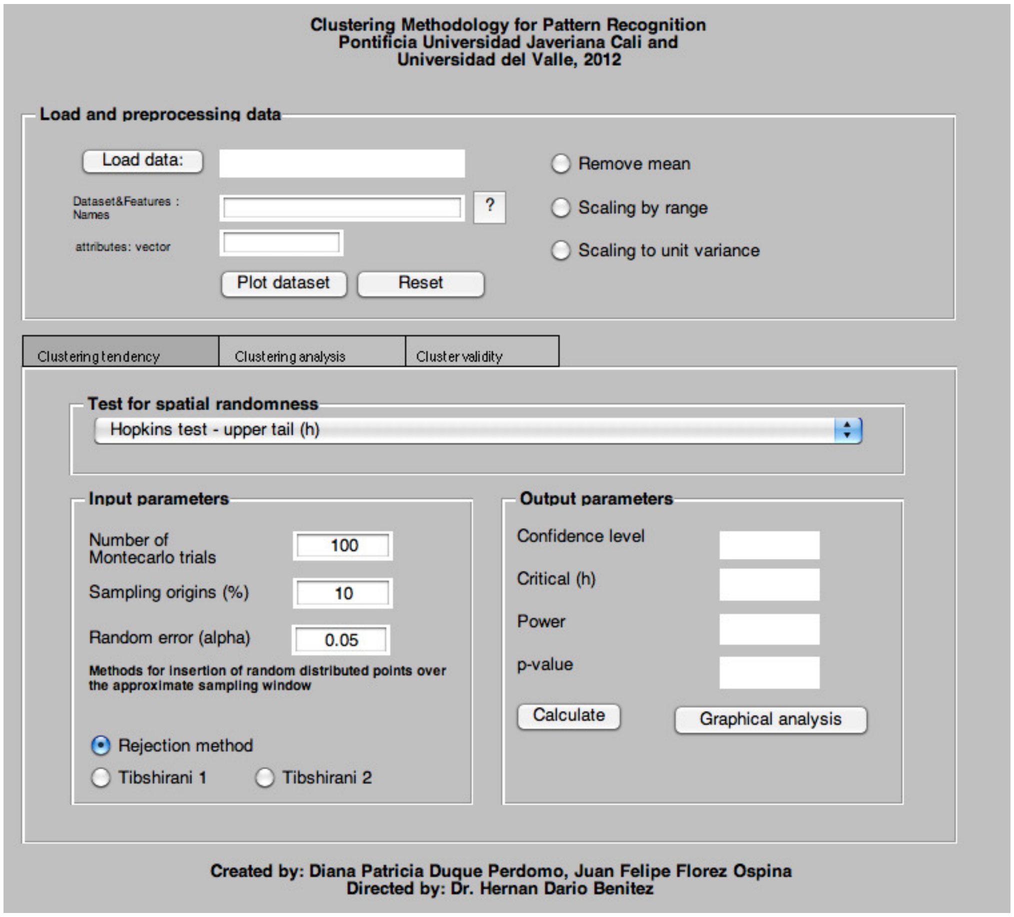
*[1]Corresponding author: hbenitez at javerianacali.edu.co
*Department of Electronics and Computer Sciences, Pontificia Universidad Javeriana
Calle 18 No 118-250-Cali, Colombia

**Universidad del Valle, Observatorio Sismológico y Geofísico del Suroccidente Colombiano
Edificio 331 - oficina 3010. Ciudad Universitaria Meléndez-Cali, Colombia

August 21, 2012

0.1 User guide

This is a user guide of clustering analysis of seismic data toolbox. This toolbox is created to facilitate the analysis of seismic data of Colombia's South West seismic catalog. We implemented this toolbox in Matlab 7.10 and user gets a graphical interface by executing archive clustering_pattern_recognition_swc.m. This toolbox comprises 4 blocks: preprocessing, clustering tendency evaluation, clustering, and clustering validity. Figure 1 shows the main graphical environment. Next sections show toolbox blocks by using as example IRIS dataset obtained from UCI repository [1].



Clustering Methodology for Pattern Recognition
Pontificia Universidad Javeriana Cali and
Universidad del Valle, 2012

Load and preprocessing data

Test for spatial randomness: Hopkins test - upper tail (h)

Input parameters

Number of Montecarlo trials: 100

Sampling origins (%): 10

Random error (alpha): 0.05

Methods for insertion of random distributed points over the approximate sampling window

● Rejection method
○ Tibshirani 1
○ Tibshirani 2

Output parameters

Confidence level:

Critical (h):

Power:

p-value:

Created by: Diana Patricia Duque Perdomo, Juan Felipe Florez Ospina
Directed by: Dr. Hernan Dario Benitez

Figure 1: Graphical interface for data clustering in R^n space

0.2 IRIS clustering analysis

This sections shows IRIS dataset analysis by using IRIS dataset.

Load and preprocessing of data

- Load dataset
- Parameter configuration
- 2D Visualization

Load and preprocessing of data

Data set is loaded from file 'iris.mat'. IRIS dataset and it comprises a random sample of three iris flower species: setosa, versicolor, and virginica. Each species is represented by 50 observations with features length, sepal width, petal length, and petal width. To load the data user must select **Load data button**, search archive .mat or .csv and open it. Loaded dataset must be a $N \times l$ matrix where N is the number of observations and l is the number of attributes. Columns must not have headers, they must have only numerical values.

Parameter configuration

Figure 2.a shows more input parameters besides Load data. These are:

- **Dataset&Features Names**: User inputs features and dataset names. Ex: { sepal length, sepal width, petal length, petal width, Iris Dataset }
- attributes vector: user defines attributes in dataset. Ex: All attributes in dataset is default option. Features can be selected as: $[2, 3]$. If this feature selections needs to be eliminated we just remove this vector.
- Plot dataset: user visualizes dataset that shows window presented in Figure 3.a. This option also allows 3D visualization.
- Reset: reset all fields.

Content - Clustering tendency verification

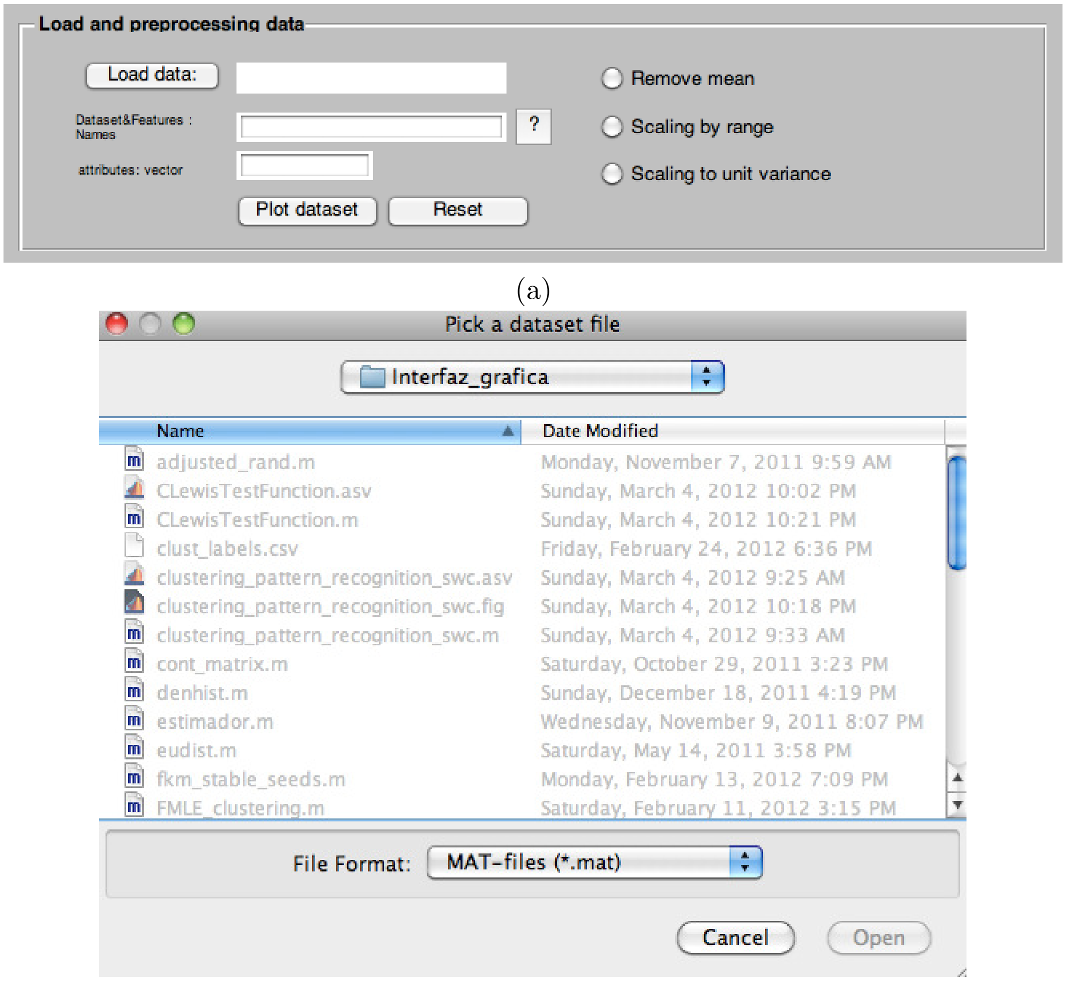
- Selection and parameter configuration
- Spatial randomness test.
- Visual analysis

Selection and parameter configuration

Figure 4 shows the visual environment of clustering tendency verification. This is obtained by pressing tab textbfTest for spatial randomness. The visual environment of clustering tendency verification provides Hopkins and Cox-Lewis tests. As input parameters we have:

- **Number of Montecarlo trials:** Number of datasets randomly generated to determine probability density function (pdf) of statistica q given null hypothesis H_0. By default it is 100.

  Sampling origins (%): This is the percentage of dataset observations used to generate sampling origins. By default it is 10 %.
- **Random error (α):** probability of rejecting null hypothesis even if it is true. Default value is 0.05.
- **Methods for insertion:** methods to insert random patterns in approximated sampling window. Default method is **Rejection method**. Methods Tibshirani1 and Tibshirani2 are presented [2]. It is important to mention that the acceptance or rejection of null hypothesis depends on these insertion methods hence user must define which of them represent better the null hypothesis of random structure.



(a)

Name	Date Modified
adjusted_rand.m	Monday, November 7, 2011 9:59 AM
CLewisTestFunction.asv	Sunday, March 4, 2012 10:02 PM
CLewisTestFunction.m	Sunday, March 4, 2012 10:21 PM
clust_labels.csv	Friday, February 24, 2012 6:36 PM
clustering_pattern_recognition_swc.asv	Sunday, March 4, 2012 9:25 AM
clustering_pattern_recognition_swc.fig	Sunday, March 4, 2012 10:18 PM
clustering_pattern_recognition_swc.m	Sunday, March 4, 2012 9:33 AM
cont_matrix.m	Saturday, October 29, 2011 3:23 PM
denhist.m	Sunday, December 18, 2011 4:19 PM
estimador.m	Wednesday, November 9, 2011 8:07 PM
eudist.m	Saturday, May 14, 2011 3:58 PM
fkm_stable_seeds.m	Monday, February 13, 2012 7:09 PM
FMLE_clustering.m	Saturday, February 11, 2012 3:15 PM

(b)

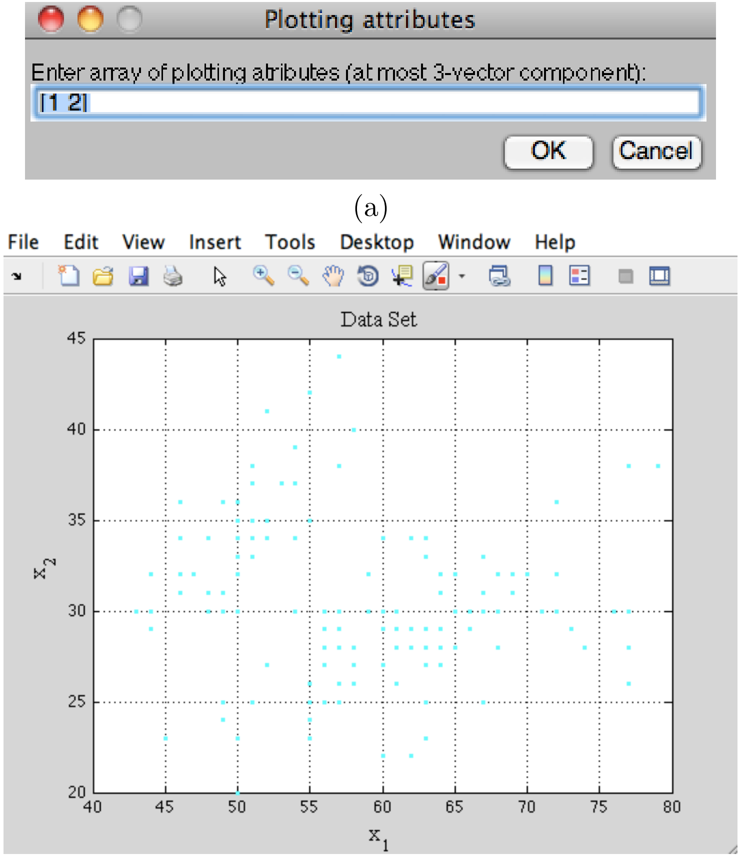
Figure 2: **(a)**Data loading and preprocessing **(b)** Window to load data

Spatial randomness test

Once the user inputs parameters for clustering tendency verification, she/he must press button **Calculate** seen in low right part at Figure 4. After pressing this button a pop-up window compels the user to subtract mean from dataset. This is necessary since each statistical test inserts random patterns for comparison in a spherical neighborhood at the center of feature space. **Remove mean**, in the visual environment of clustering tendency verification, subtracts mean from dataset. By pressing **Calculate** button again results are obtained and shown as in Figure 6. Since p-value = 0, null hypothesis is rejected and alternative hypothesis of clustered data is accepted hence IRIS dataset has cluster tendency.

Visual analysis

An alternative way of analyzing the results obtained is pressing button **Graphical analysis** to visualize statistic pdf under hypothesis H_0 and H_a estimated from histograms density. Figure **??** shows results after evaluating IRIS dataset. Hopkins and Cox-Lewis statistics values are close to



Plotting attributes

Enter array of plotting atributes (at most 3-vector component):

[1 2]

(a)

(b)

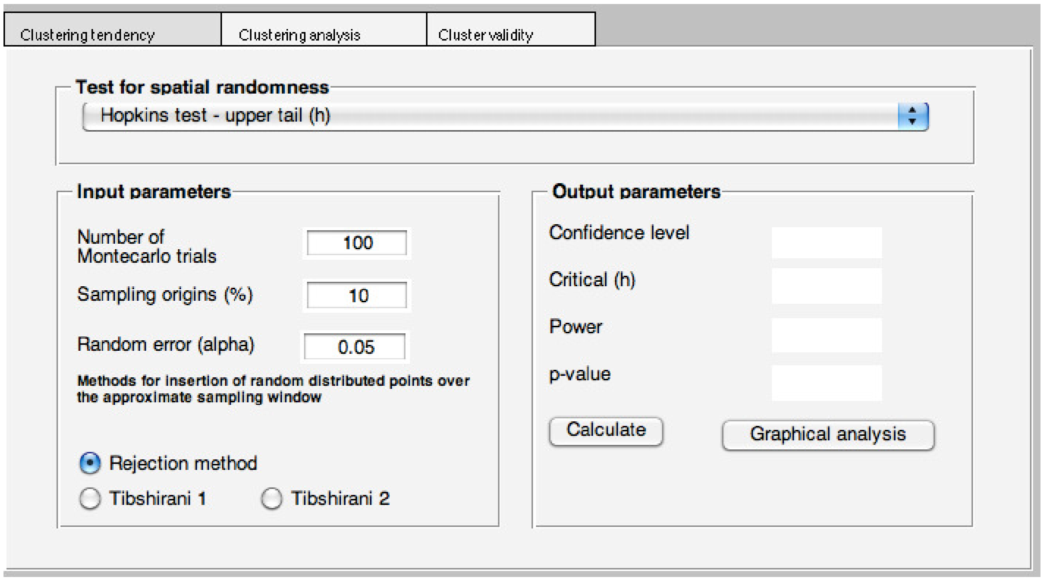
Figure 3: **(a)** Window to select attributes to be plotted. **(b)** Plot of attribute 4 vs attribute 1 in IRIS dataset, default value is [1,2].

0.75.

Clustering analysis

This block classifies the dataset according to the following input parameters:

- **Number of classes:** Number of classes k.
- **Initialization method:** This places initial positions of centroids. These are the methods available: **Random seeds**, **Refining starting seeds** (apt for k-means and fuzzy k-means) and **fuzzy k-means stable seeds** (proposed procedure to initialize Gath and Geva algorithm).
- **Tolerance:** Maximum variation allowed between current and past iteration. This also known as stopping criteria. Its defaults value is 1e-5.
- **Fuzzifier:** This parameter is only used for fuzzy clustering algorithms and it can be seen as fuzzines degree of classification. Its interval is [2,2.3].



Clustering tendency | Clustering analysis | Cluster validity

Test for spatial randomness

Hopkins test - upper tail (h)

Input parameters

Number of Montecarlo trials: 100

Sampling origins (%): 10

Random error (alpha): 0.05

Methods for insertion of random distributed points over the approximate sampling window

● Rejection method

○ Tibshirani 1 ○ Tibshirani 2

Output parameters

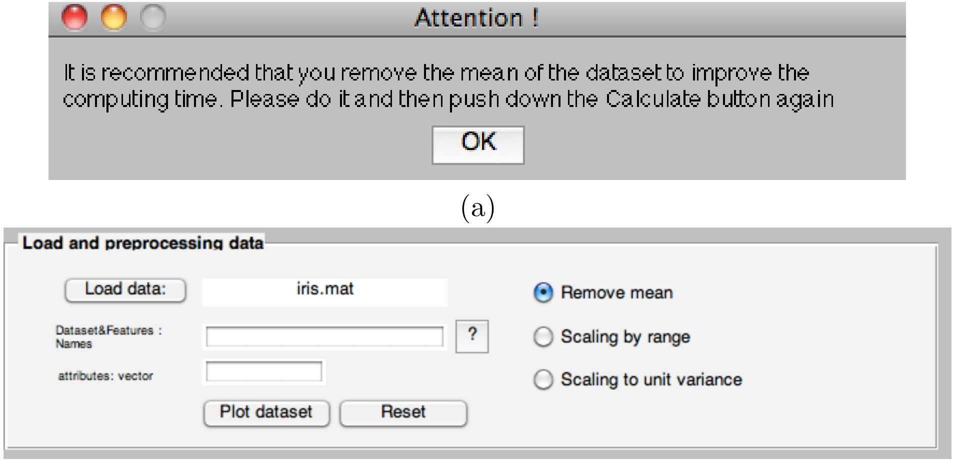
Confidence level

Critical (h)

Power

p-value

Figure 4: Visual environment of clustering tendency verification, the tab exhibits Hopkins and Cox-Lewis tests for clustering tendency verification



Attention !

It is recommended that you remove the mean of the dataset to improve the computing time. Please do it and then push down the Calculate button again

(a)

Load and preprocessing data

Load data: iris.mat

Dataset&Features : Names

attributes: vector

● Remove mean

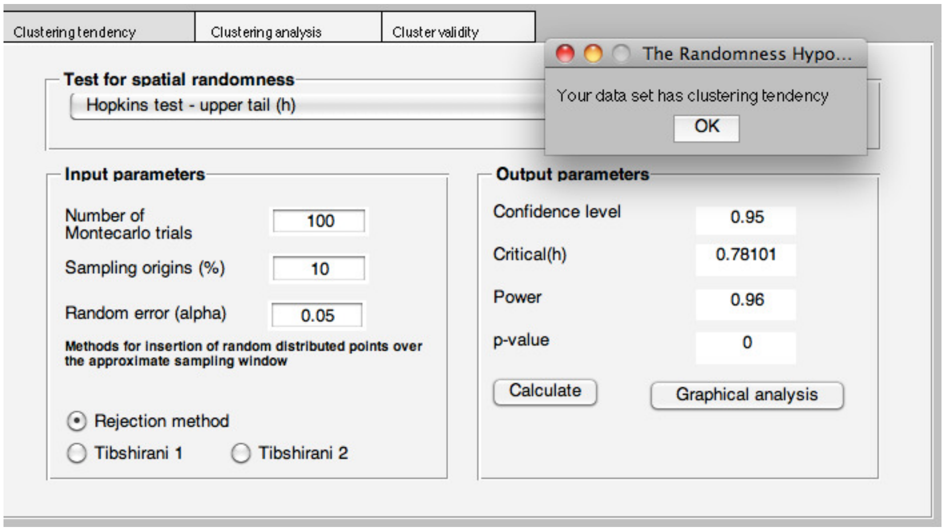
○ Scaling by range

○ Scaling to unit variance

(b)

Figure 5: **(a)** Pop-up showing that clustering tendency test requires mean subtraction from dataset. **(b)** Selection of **Remove mean** subtracts mean in dataset

To cluster a dataset, user selects clustering algorithm as shown in Figure 8.a and its initialization method Figure 8.b. Then, number of classes k is stablished and button **Clustering** is pressed in **options**. Results of data clustering are visualized in Figure 8.c for Gath and Geva algorithm. User can store labels from clustered dataset in format .csv by pressing button **Import data labels**. Labels can take values from dataset $\{2, ..., k\}$ as shown in Figura 8.d.



The Randomness Hypo...

Your data set has clustering tendency

Test for spatial randomness: Hopkins test - upper tail (h)

Input parameters

Number of Montecarlo trials: 100

Sampling origins (%): 10

Random error (alpha): 0.05

Methods for insertion of random distributed points over the approximate sampling window

● Rejection method
○ Tibshirani 1
○ Tibshirani 2

Output parameters

Confidence level: 0.95

Critical(h): 0.78101

Power: 0.96

p-value: 0

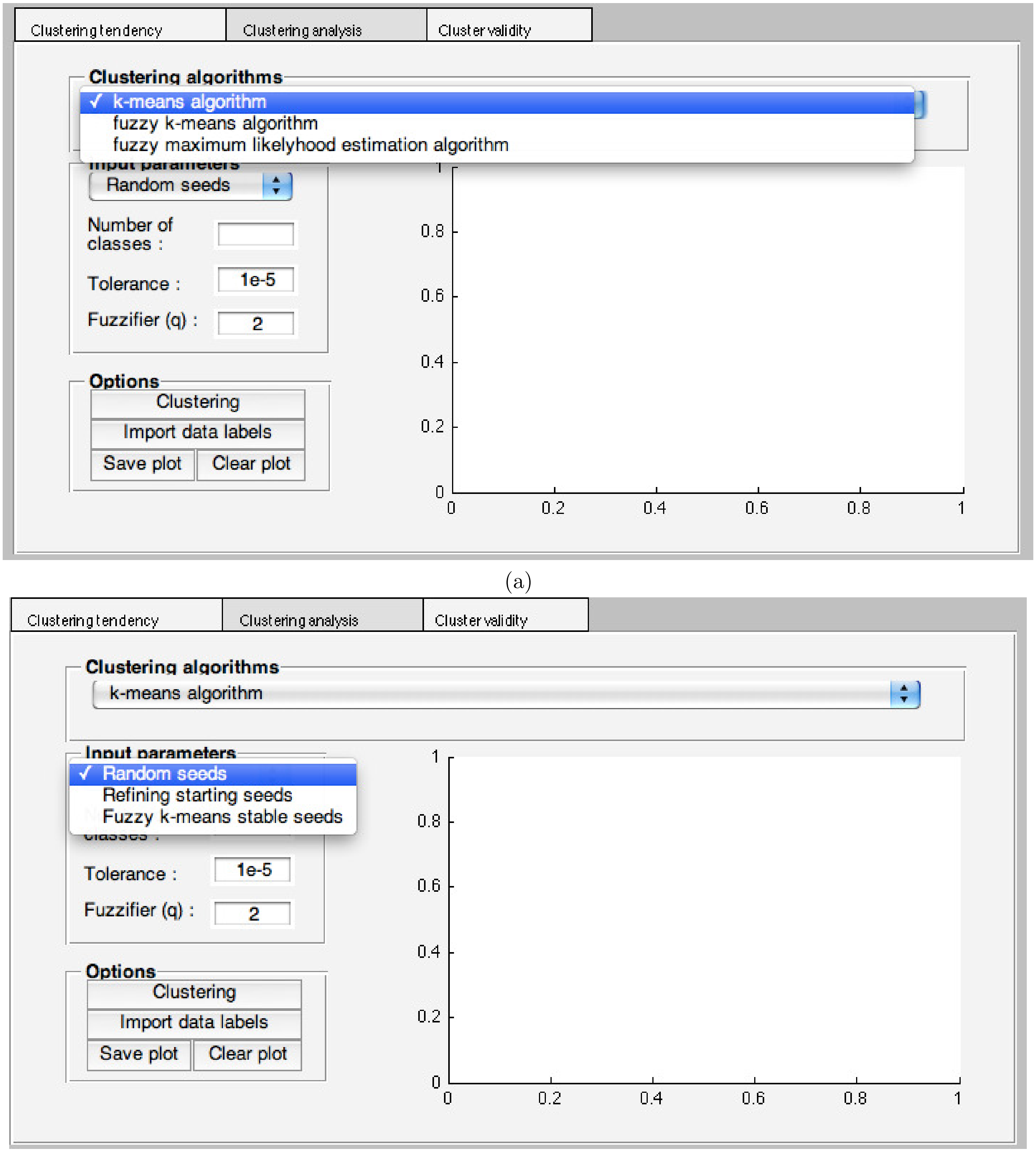
Figure 6: Hopkins test results after pressing button **Calculate**, p-value = 0 then null hypothesis H_0 is rejected

Clustering validity

The objective of clustering validity is to quantitatively evaluate the results of a clustering algorithm. User can use Dunn index [3] to assess partitions generated by algorithms k-means and fuzzy k-means. On the other hand, metrics fuzzy hyper-volume F_{HV} and probability density P_D evaluates the results of algorithm Gath and Geva [4]. To verify cluster tendency the user must press on tab Cluster validity. Cluster validity displays:

- **Lower limit:** Minimum number of classes to evaluate.
- **Upper limit:** Maximum number of classes to evaluate.
- **Tolerance:** It is the maximum difference allowed between current iteration and next iteration. This difference is used as stopping criteria. Its default value is 1e-5.
- **Fuzzifier:** Fuzzy classifications uses this parameter and it can be seen as the degree of fuzziness in the classification. Its interval is [2,2.3].
- **Number of runs:** Number of times that validation is executed to visualize the variation of validation index. This variation is observed with error bars around the mean index value.

To start validation process the user should select the algorithm and its initialization mode as shown in Figura 10.a and .b. After inputting the parameters user must select button **To Calculate**. Results of validating k-means partitions in the validation interval for k $\{2, ..., 7\}$ shows that the best partition for Iris dataset is k=2. In contrast, indexes for Gath and Geva algorithm, evaluated for the same k interval and dataset, shows that the best partition for Iris dataset is k=7. It is important to note that partition validation depends on the indexes and it is common to find differences in the results provided [5].



Tolerance : 1e-5

Fuzzifier (q) : 2

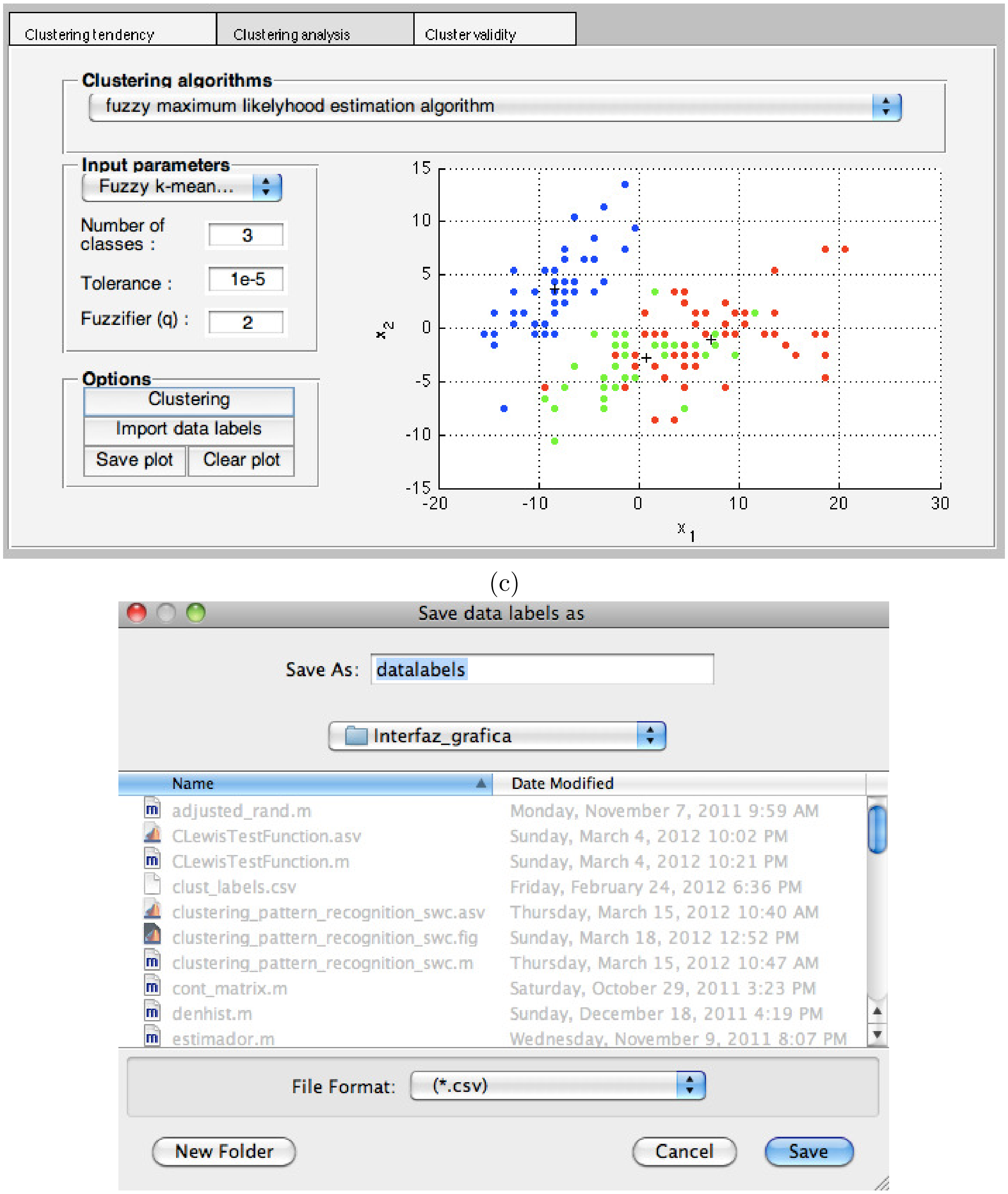
(a)

Tolerance : 1e-5

Fuzzifier (q) : 2

(b)

Figure 7: **(a)** Clustering algorithm selection **(b)** Initialization method selection



Clustering algorithms: fuzzy maximum likelyhood estimation algorithm

Input parameters: Fuzzy k-mean...

Number of classes : 3

Tolerance : 1e-5

Fuzzifier (q) : 2

(c)

Save data labels as

Save As: datalabels

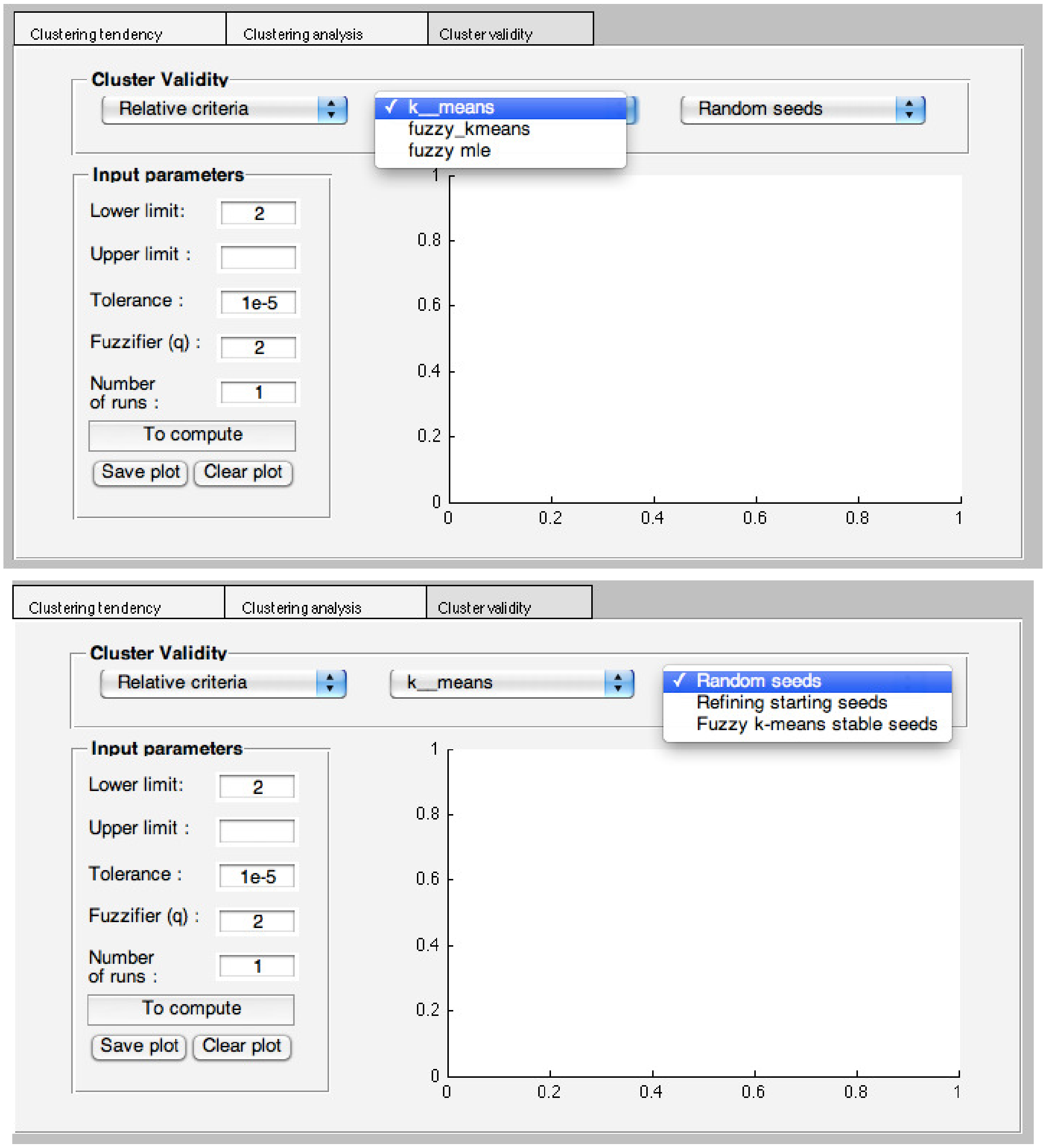
Interfaz_grafica

Name	Date Modified
adjusted_rand.m	Monday, November 7, 2011 9:59 AM
CLewisTestFunction.asv	Sunday, March 4, 2012 10:02 PM
CLewisTestFunction.m	Sunday, March 4, 2012 10:21 PM
clust_labels.csv	Friday, February 24, 2012 6:36 PM
clustering_pattern_recognition_swc.asv	Thursday, March 15, 2012 10:40 AM
clustering_pattern_recognition_swc.fig	Sunday, March 18, 2012 12:52 PM
clustering_pattern_recognition_swc.m	Thursday, March 15, 2012 10:47 AM
cont_matrix.m	Saturday, October 29, 2011 3:23 PM
denhist.m	Sunday, December 18, 2011 4:19 PM
estimador.m	Wednesday, November 9, 2011 8:07 PM

File Format: (*.csv)

(d)

Figure 8: **(c)** Results of algorithm execution **(d)** Classification labels exportation



Lower limit: 2

Upper limit :

Tolerance : 1e-5

Fuzzifier (q) : 2

Number of runs : 1

Lower limit: 2

Upper limit :

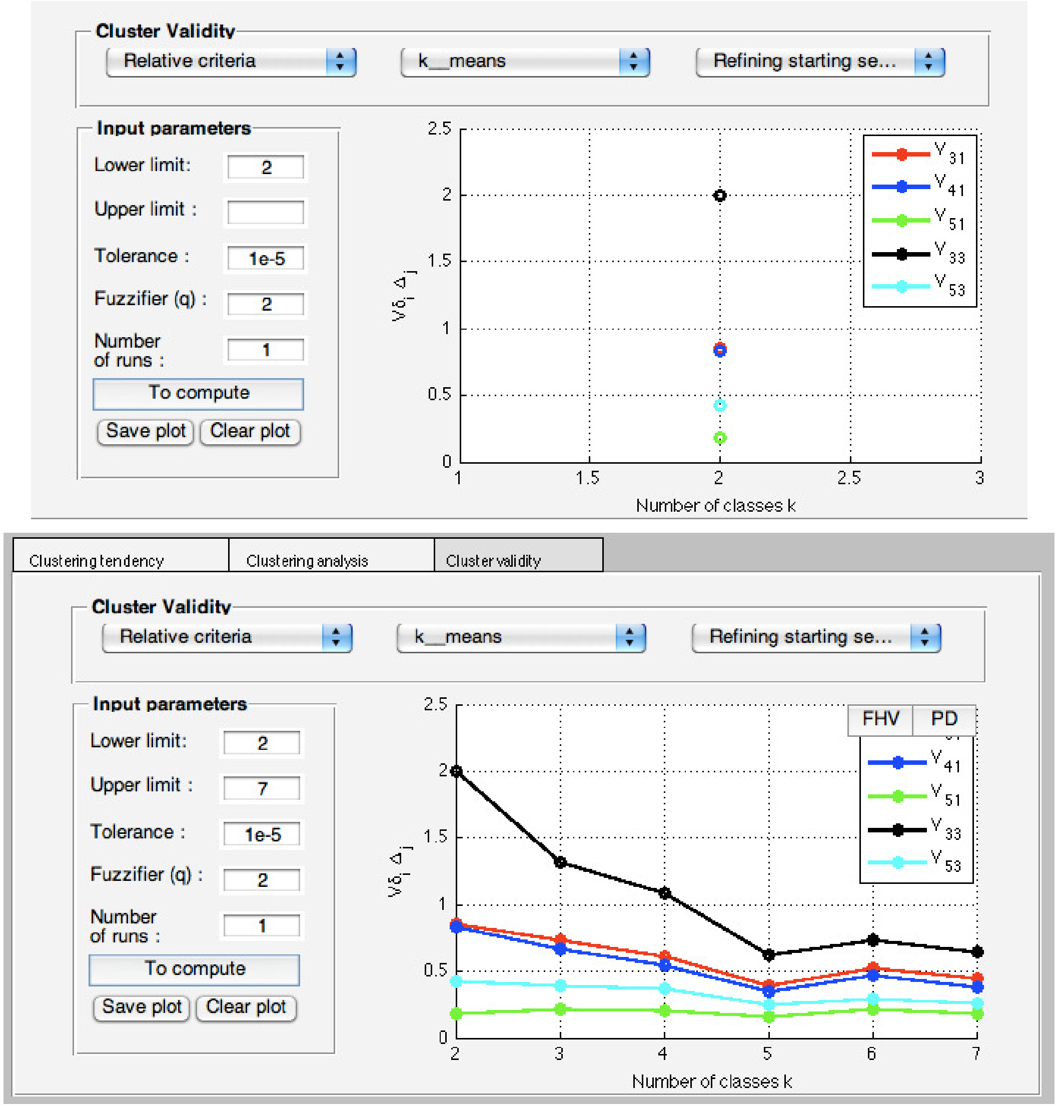
Tolerance : 1e-5

Fuzzifier (q) : 2

Number of runs : 1

(a)
(b)

Figure 9: **(a)** Clustering algorithm selection **(b)**Initialization mode selection



Cluster Validity

Relative criteria | k__means | Refining starting se...

Input parameters

Lower limit: 2

Upper limit :

Tolerance : 1e-5

Fuzzifier (q) : 2

Number of runs : 1

Clustering tendency | Clustering analysis | Cluster validity

Cluster Validity

Relative criteria | k__means | Refining starting se...

Input parameters

Lower limit: 2

Upper limit : 7

Tolerance : 1e-5

Fuzzifier (q) : 2

Number of runs : 1

(c)
(d)

Figure 10: **(c)** Clusters validity in **K-means** with initialization **Refining Starting Seeds**. **(d)** Clustering validity for Gath and Geva algorithm with initialization **Fuzzy k-means stable seeds**.

Bibliography

[1] A. Frank and A. Asuncion. Uci machine learning repository, 2010.

[2] R Tibshirani, G Walther, and T Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.

[3] J. Bezdek and N Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cyber- netics, Part B: Cybernetics.*, 28:301–315, 1998.

[4] I. Gath and A.B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:773 - 780, 1989.

[5] J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva, and N.R. Pal. will the real iris data please stand up?. *IEEE Transactions on Fuzzy Systems*, 7:368–369, 1999.