SERA
Seismology and Earthquake Engineering
Research Infrastructure Alliance for Europe

DELIVERABLE 23.2

D23.2 Statistical tools characterizing seismicity clustering and magnitude distribution

Work package	WP23
Lead	IGPAS
Authors	Konstantinos Leptokaropoulos, Monika Sobiesiak, Piotr Sałek, Stanisław Lasocki - IGPAS
Reviewers	N/A
Approval	Management Board
Status	Final
Dissemination level	Public
Delivery deadline	30.04.2019
Submission date	30.04.2019
Intranet path	DOCUMENTS/DELIVERABLES/SERA_D23.2_Statistical_Tools.zip

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730900.

Table of Contents

Summary

Deliverable D23.2 comprises computer programs (Applications) which were developed and compiled in three versions in order to allow different interactivity level with the user and to support alternative ways of importing input data and parameters. These applications are statistical tools grouped in two individual Toolboxes, one for analysing seismicity clustering and one for characterizing the complexity of magnitude distribution, respectively. In total 4 applications, written in Matlab were developed, and they are compatible with Matlab Version 2017b or later. These applications are:

- Toolbox 1 - "Clustering/Transformation to Equivalent Dimensions Toolbox": Clustering of seismic events in the phase (i.e. parameter) space built by any group of parameters unequivocally associated to a set of either seismic or any other events (Parameters transformed to Equivalent Dimensions; Lasocki, 2014):
    - Application 1A: *"T2ED"* – Transformation to Equivalent Dimensions (ED)
    - Application 1B: *"Clustering"* – cluster analysis of transformed to ED data obtained from Application 1A (T2ED)
- Toolbox 2 - "Magnitude Complexity Toolbox": Tests for investigating complexity of a random variable (i.e. magnitude) distribution. *[This Toolbox is mainly, but not exclusively focused on magnitude distribution. One can analyse any other random variable]*:
    - Application 2A: "ADTestMag" – Anderson-Darling (AD) Test for testing the hypothesis that a given sample data (i.e. sample of magnitudes) has been drawn from the exponential or Weibull distribution (Anderson and Darling, 1954)
    - Application 2B: "MM_MB" – Testing the existence of multi-modes/ multi-bumps in the distribution of a given random variable (i.e. magnitude) (Lasocki and Papadimitriou, 2006)

A short description of the 4 Applications is provided in the Appendix. These Applications are available in different versions in order to ensure flexibility and convenience for the Users.

- Standalone Version V1. Interactive standalone versions. Input is interactively introduced by the users step-by-step through a Graphical User Interface *[GUI – i.e. graphically selected, from pop-up windows, by typing etc]*. The input data are generally imported from ASCII files and the Input Data requirements are specified in the applications documentation. Each Version V1 application pack includes the following material:
    - Main Application *[*.m Matlab Program]*
    - Application Description Documentation *[READ_ME*.docx file]*
    - Input data Directories(s) with sample data for testing
- Standalone Version V2. Wrapper standalone versions. Input is set within "Wrapper Scripts" (used for data selection and parameter setting). Once the parameters are defined within the wrapper scripts and the User runs them, the application is performed without any interruption. Each Version V2 application pack includes the following material:
    - Main Application *[*.m Matlab Function]*
    - Wrapper script for data loading and input parameters setting *[*.m Matlab Program]*
    - Visualization and other Auxiliary scripts *[*.m Matlab Programs & Functions]*
    - Application Description Documentation *[READ_ME*.docx file]*
    - Input data Directories(s) with sample data for testing

- IS-EPOS Platform Version V3. On-line versions. These are modified version V2 scripts which are currently under integration process for on-line usage within the IS-EPOS Platform - https://tcs.ah-epos.eu/.

The standalone versions (V1, V2) are ready and fully operational. However, they are continuously upgraded in order to ensure an efficient and user friendly workflow as well as to fix bugs and for other issues. A detailed step-by-step description of the applications' use, as well as explanation of input parameters and output results is included within the scripts (code comments) and also in separate documents accompanying the scripts (individual documents for each application version). A summary of the characteristics of these diverse versions is demonstrated in Table 1. The main features of the available versions are:

- Input data: In V2 applications the input files (and corresponding paths) are specified within the wrapper scripts. The files are read by the wrapper scripts and the corresponding parameters are used as input for the main application functions. For the V1 applications the following cases are distinguished:
    - Application A1 ("T2ED") uses data in ASCII format, specified within the documentation of the Application. The input data files must be located in specified directories and they are interactively selected by the User.
    - Application 1B ("Clustering") uses only data in the output format of Application 1A ("T2ED"). The cluster analysis can only be performed in the Equivalent Dimensions parameter space. Therefore "T2ED" must run before "Clustering", and the output file (matlab structure) of "T2ED" must be moved to the corresponding directory for "Clustering". The input data files are interactively selected by the User.
    - Toolbox 2 applications use single vectors for input data, which can be uploaded e.g. from an ASCII file.
- Interactivity: The Version V2 applications can be executed as classical Matlab functions without any interactivity, by just setting the values of the input arguments in the wrapper scripts. The User can interfere within these scripts in order to change the input data as well as the input parameters values, before executing the main application function(s). In such way, no interactivity is introduced after running the function, however the User can previously define all the arguments within the auxiliary wrapper script.

  On the other hand, in the Version V1 applications the user may interactively (graphically or manually, from lists, plots or by typing) select step-by-step the input arguments. In in the special case of ToolBox 2, the applications (i.e. "ADTestMag_V2_8" and "MM_MB_V2_8" scripts), have a dual-mode behaviour, where the User may enable/disable interactivity. If all input arguments are set, then the applications operate as a function. However, if only the input data file is introduced, the applications switch to interactive mode.
- Visualization: In the V2 applications, within the wrapper script the User can activate or comment the plotting option, such that output visualizations connected to the application are created as well.

Please note that scripts comments and documentation, will be continuously updated in order to improve the applications' performance and interaction with the Users. As soon as approved by the EC, toolboxes will be available through the SERA website at www.sera-eu.org.

Table 1. Characteristics of the diverse Application Standalone versions available.

	APPLICATION VERSION	INPUT FORMAT	INTERACTIVITY	WRAPPER SCRIPT	VISUALIZATION
APPLICATION 1A – "T2ED"	T2ED_V1_8	ASCII files*	Yes	No	Yes
	T2ED_V2_8	ASCII files*	No	Yes	Yes/No
APPLICATION 1B – "CLUSTERING"	Clustering_V1_8	T2ED output	Yes	No	Yes
	Clustering_V2_8	T2ED output	No	Yes	Yes/No
APPLICATION 2A – "ADTESTMAG"	ADTestMag_V1_8	ASCII file	Yes/No	No	Yes
	ADTestMag_V2_8	Vector	No	Yes	Yes/No
APPLICATION 2B – "MM_MB"	MM_MB_V1_8	ASCII File	Yes/No	No	Yes
	MM_MB_V2_8	Vector	No	Yes	Yes/No

*The input data format requirements are thoroughly specified within the Application Description Documentation

References

Anderson T. W., and D. A. Darling, (1954), "A test of goodness of fit", *J. Amer. Stat. Assoc.*, 49, 765-769, doi:10.1080/01621459.1954.10501232.

Cox, D. R., (1966), Notes on the analysis of mixed frequency distributions, Br. J. Math. Stat. Psychol., 19, 39-47, doi.org/10.1111/j.2044-8317.1966.tb00353.x.

Efron, B., and R. J. Tibshirani (1993), An Introduction to the Bootstrap, CRC Press, Boca Raton, Fla.

Lasocki S. and E. E, Papadimitriou (2006), "Magnitude distribution complexity revealed in seismicity from Greece", *J. Geophys. Res.*, 111, B11309, doi:10.1029/2005JB003794.

Lasocki S. (2014), Transformation to equivalent dimension - a new methodology to study earthquake clustering, *Geophys. J. Int.*, 197, 1224-1235,doi:10.1093/gji/ggu062.

Marsaglia, G. and J. Marsaglia (2004), Evaluating the Anderson-Darling distribution, *J. Stat. Soft.*, 9, 1-5.

Silverman, B. W. (1986), Density estimation for statistics and data analysis, CRC press, 175 pp.

Appendix – Applications Short Description

Toolbox 1 – Clustering:

- Application A1 – "T2ED": This function constitutes the innovative part of the "Clustering " Toolbox which is performed for multi-parameter space. The function takes as input seismic or/and operational data parameters and transform them into their equivalent dimensions following the methodology introduced by Lasocki (2014).
- Application A2 – "Clustering": The function is actually a compilation of existing and well-known clustering algorithms available within the MATLAB libraries, therefore the corresponding functions, descriptions information and references can be retrieved from the Matlab help. The functions used are "kmeans", "linkage", "cluster" & "fcm". This Application takes as input the

output file created after executing Application A1 "T2ED", therefore all analyses are performed in the Equivalent Dimension parameter space.

Toolbox 2 – Magnitude Complexity:

- Application B1 – "ADTestMag": This function performs the Anderson-Darling test (e.g. Anderson & Darling, 1954; Marsaglia & Marsaglia, 2004) for testing the Null Hypothesis, H_0, that a given set of data (e.g. magnitudes), follows the exponential or the Weibull distribution. This is accomplished as a function of minimum parameter cut-off (i.e. completeness magnitude), therefore, multiple results are produced (iteration process). The corresponding p-values for the H_0 is the main output of the program. Before applying the AD test, the input parameter values are randomized within their round-off interval, following the formula introduced by Lasocki and Papadimitriou (2006).
- Application B2 – "MM_MB": The function studies a time series (e.g. magnitude) distribution complexity by means of the Mulitmodality Test (Silverman, 1986; Efron and Tibshirani, 1993). Two null hypotheses (H0s) are tested:
    - H01 - multimodality: The input parameter PDF is unimodal
    - H02 - multi-bump: The input parameter PDF has one bump to the right of the mode.

  A mode is a local maximum of probability density and a bump is an interval [a,b] such that the probability density is concave over [a,b] but not over any larger interval (Silverman, 1986). The importance of modes and bumps relies on the fact that multiple occurrences of these features in a PDF indicate, for most standard densities, a mixing of components (e.g. Cox, 1966).