
DELIVERABLE

D24.4 Test-bed validation of tools and resulting high level products: software toolbox, validation methodologies, demonstration report

Work package	WP24: Characterizing the activity rates of induced and natural earthquakes
Lead	ETH
Authors	Shyam Nandan, ETH; Laurentiu Danciu, ETH; Stefan Hiemer*, former at ETH; Celso Guillermo Reyes, ETH; Konstantinos Leptokaropoulos, Stanisław Lasocki, IGPAS; Monika Sobiesiak, IGPAS; Piotr Sałek, IGPAS ; Danijel Schorlemmer, GFZ-Potsdam; Fabrice Cotton, GFZ-Potsdam; Stefan Wiemer, ETH Zurich; Domenico Giardini, ETH]; Antonio_Petruccelli, ETH
Reviewers	-
Approval	Management Board
Status	Final
Dissemination level	Public
Delivery deadline	30.04.2020
Submission date	28.04.2020
Intranet path	DOCUMENTS/DELIVERABLES/SERA_D24.4_Test-bed_validation_tools_products.pdf



Table of Contents

Summary.....	3
1 Chapter 1: Open-Source Toolbox For Seismicity Analysis	5
1.1 Summary.....	5
1.2 ZMAP 7.0	5
1.3 Jupyter Notebooks.....	7
1.4 PyMap.....	8
1.5 References.....	12
2 Chapter 2: An Objective Method For Estimating Time Variation Of Magnitude Of Completeness With Application To Eshm20 Earthquake Catalog.....	13
2.1 Summary.....	13
2.2 ESHM20 Earthquake Catalog and Completeness Super Zones.	14
2.3 Global Magnitude of Completeness, b-value, and a-value.....	15
2.4 Time Series of Magnitude of Completeness.....	18
2.4.1 Algorithm.....	18
2.4.2 Explanation of the Algorithm and Description of the Important Hyper-Parameters.	19
2.4.3 Demonstartion Report: Aplicability to the ESHM20 Unified Earthquake Catalog.....	20
2.5 Outlook	23
2.6 References.....	23
3 Chapter 3: Time-Dependent Induced-Seismicity Models, Their Dependency On Time-Varying Operational Parameters And Guidelines For Time-Dependent Hazard Evaluation	24
3.1 Summary.....	24
3.2 Demonstration Report: Time-and-Technology dependent Induced Seismicity.....	24
3.3 Methodology.....	26
3.4 Software Toolbox: SHAPE package.....	29
3.5 Demonstartion Report: Case study	34
3.6 References.....	39
4 Chapter 4: Earthquake sequences and background seismicity.....	41
4.1 Summary.....	41
4.2 Introduction	41
4.2.1 Statistical and deterministic approaches for earthquake forecasting.....	41
4.2.2 Temporal variations in earthquake-size distribution.....	42
4.3 Methodology and application.....	43
4.3.1 MATLAB® code implementation.....	43
4.3.2 Demonstration Report: Retrospective analysis for aftershocks forecasting.....	45
4.3.3 Verification and benchmarking.....	45
4.4 Future developments	48
4.5 References.....	49

5	Chapter 4: ESHM20 Python Toolkit for development of regional seismic hazard models	50
5.1	Summary	50
5.2	Introducing the ESHM20 Toolkit	50
5.3	ESHM20 Toolkit: General Methodology for Code Development	51
5.4	ESHM20 Toolkit: OpenQuake Changes in support of the toolkit	51
5.4.1	Additions to openquake.hmtk.seismicity.occurrence	51
5.4.2	openquake.hmtk.seismicity.occurrence.stromeyer_gruenthal	52
5.4.3	Additions to openquake.hmtk.seismicity.declusterer	53
5.4.4	Additions to openquake.hmtk.seismicity.completeness	53
5.5	ESHM20 Toolkit: Seismic Hazard Model Development Processing Flow:	54
5.5.1	Part 1: From Raw catalogs to Declustered Catalogs divided by depth	54
5.5.2	Part 2: Magnitude of Completeness	54
5.5.3	Part 3: Calculating the Gutenberg-Richter Parameters	56
5.5.4	ESHM20 Toolkit: Requirements	59
5.6	ESHM20 Toolkit: Interface	59
5.6.1	Python Integrated Development Environment (IDE) (ex. PyCharm)	59
5.6.2	Jupyter Notebook interface	60

Summary

In this work package, we aimed to update and develop new statistical tools for better estimation of seismic activity rates for long and short term seismic hazard assessment as well as for the improved evaluation of the other earthquake triggering phenomena, for instance in areas with anthropogenic seismicity.

As such, we make five crucial contributions in this direction, which are presented in following four independent chapters aligned with the main tasks of the deliverable: test-bed validation of tools and resulting high level products: software toolbox, validation methodologies, demonstration report.

Chapter 1, summarizes the state of development of two open-source software toolboxes for seismicity analysis. Chapter 2, presents an objective method for estimating time variation of the magnitude of completeness and its application to ESHM20 earthquake catalog; which can be seen as a demonstration report for large regional scale of seismic hazard assessment. Note, that the completeness method illustrated in Chapter 2 was derived within the JRA2 and JRA3 joint efforts, and it is used in the development of the 2020 updates of the European Seismic Hazard Model (ESHM20, in SERA JRA3). In

Chapter 3, we describe some time-dependent induced-seismicity models and their dependency on time-varying operational parameters as well as guidelines for time-dependent hazard evaluation. This effort summarizes the applicability and demonstration report to local scale seismic analysis, i.e. the induced seismicity.

Next, Chapter 4 deals with the problem of characterizing earthquake sequences and tectonic background activity rates using an innovative approach, which involves temporal variations in earthquake-size distribution.

Finally, Chapter 5, ESHM20-toolbox, provides an overview of the toolkit used to compute the activity rates for the 2020 European Seismic Hazard Model developed within the SERA JRA3. Each chapter is complete with its references and accompanying appendices.

1 Chapter 1: Open-Source Toolbox For Seismicity Analysis

Authors: Danijel Schorlemmer, GFZ-Potsdam; Fabrice Cotton, GFZ-Potsdam; Celso reyes, ETH Zurich; Stefan Wiemer, ETH Zurich

1.1 Summary

Two seismicity toolboxes, ZMAP 7.0 and PyMap, were developed. ZMAP 7.0 is based on recent MATLAB versions while PyMap is a Python library. ZMAP 7.0 resembles the functionality of previous ZMAP versions and is updated to run with newer MATLAB versions.

PyMap is a new development and its framework include abstraction layers for coordinate systems (earthquakes, mining events, laboratory acoustic emissions) and for data grids (Cartesian, triangular and hexagonal spherical tessellation). The framework is designed such that all codes using the abstraction layer can be agnostic about the coordinate system and the grid.

The framework and first basic functions are accompanied by a full 4D-viewer that allows local to global visualizations of earthquake catalogs, earthquake data, coastlines, topography, and data grids. PyMap development is continued in collaboration with the Collaboratory for the Study of Earthquake Predictability (<https://www.scec.org/research/projects/CSEP/scec3.html>) and the pyrocko open-source toolbox (<https://pyrocko.org/>).

1.2 ZMAP 7.0

We developed a new version of the ZMAP toolbox published many years ago (Wiemer, 2001). This new Version 7 (Reyer and Wiemer 2019) is compatible with versions of MATLAB from 2018 onward. Current ZMAP developments facilitate the implementation and transfer of existing modules to PyMap and allow crosschecking of PyMap and ZMAP tools.

Earthquake catalogs are probably the most fundamental products of seismology and remain arguably the most useful for tectonic studies. Modern seismograph networks can locate upwards of a hundred thousand earthquakes annually, providing a continuous and sometime overwhelming stream of earthquake locations. ZMAP is a set of tools driven by a graphical user interface (GUI), designed to help seismologists analyze catalog data.

ZMAP is primarily a research tool suited to the evaluation of catalog quality and to addressing specific hypotheses; however, it can also be useful in routine network operations. ZMAP was first published in 1994, with the last major release, version 6.0, in 2001.

ZMAP 7 depends upon MathWorks MATLAB® R2018a or higher, and will work on Windows, MacOSX and Linux operating systems. Additionally, the following MATLAB toolboxes must be installed:

- Mapping Toolbox
- Statistics and Machine Learning Toolbox
- Parallel Computing Toolbox [optional, enables parallel computing]

ZMAP is currently hosted on GitHub at: <https://github.com/CelsoReyes/zmap7>, with links to the download available from the main SED website <http://www.seismo.ethz.ch/en/research-and-teaching/products-software/software/ZMAP/>.

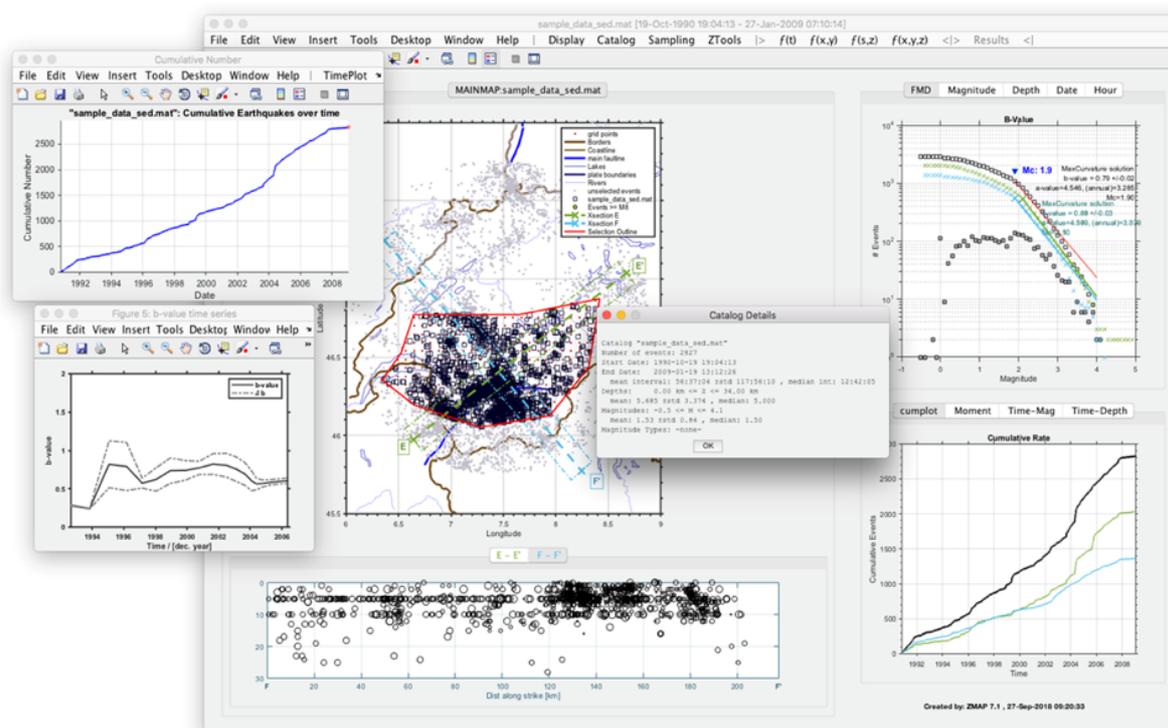


Figure 1: Sample ZMAP 7 interface

ZMAP 7 represents a major reworking of ZMAP 6.0. Every aspect of ZMAP has been modified—from the user interface through the data representations within the program—according the following goals:

- Make ZMAP compatible with modern MATLAB installations. MATLAB has evolved far beyond the version for which ZMAP 6.0 was designed and in several cases broke backwards compatibility. In the intervening years, new, robust techniques for performing object oriented design have driven changes to the graphics system underpinning the user interface, as well as to the language itself.
- Make it easier to add additional functionality. By leveraging functions and classes, future users inherit a consistent interface that allows the easy addition powerful analysis routines with very little code duplication.
- Make the user interfaces more consistent and interactive. Frequently recurring user-interfaces (e.g. dialog boxes) were once generated at a low level in each routine and differed wildly between routines. Now, the most common of these have been consolidated and a method for consistently generating them has been added.
- Make code more robust. Originally, ZMAP code consisted of a large selection of scripts that operated on global variables and made assumptions about the GUI's state. Callbacks, a primary component of GUIs, were string-based scripts that were “invisible” to the MATLAB syntax checker. Now, scripts have been extracted into individual functions, allowing for better code reuse and allowing the languages validation tools to efficiently function.
- Make existing code more readable and maintainable. This has involved reducing code duplication through the use of consolidated helper functions and classes. Home-grown functionality has been removed in favor of standard toolbox functions, further reducing the need for maintenance. All entities (classes, functions, and variables) are being renamed to reduce the cognitive burden of maintainers to follow.

Simplify access to event catalogues. ZMAP can load catalog data directly from FDSN Event Webservice web sites in addition to a variety of file formats or local variables.

While optimizing for speed was not a goal in itself, the previous points all contribute to the ease of finding and fixing inefficiencies resulting from naïve algorithm choice or inefficient porting from other languages (esp. FORTRAN). Additionally, several algorithms may take advantage of parallel processing capabilities.

ZMAP 7 is currently in alpha release stage while its functionality continues to evolve. However, the basic functionality is in place and already allows one to easily explore earthquake catalogs. When users encounter bugs or user-unfriendly behavior, they are encouraged to report them to the ZMAP developer(s) via the GitHub issue reporting system, conveniently accessible from within the ZMAP help menu. These issues are visible by both the program maintainer(s) and the community and becomes a touchstone for understanding which aspects of ZMAP are important to the community.

1.3 Jupyter Notebooks

The Jupyter Notebook is an open-source web application/interface that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Jupyter notebooks are developed open-source software, open-standards, and services for interactive computing across many programming languages.

Table 1		
Learning Goals and Their Associated Jupyter Notebooks		
	Learning Goal	Notebook Title
1	Python data types	Introduction to Python
2	Import and export data	Introduction to reading data and plotting using Pandas
		Introduction to plotting data as a heat map
3	Basic figure making	Introduction to reading data and plotting using Pandas
		Introduction to plotting data as a heat map
		Introduction to scatter plots and histograms
4	Histograms	Introduction to scatter plots and histograms
5	Map making	Creating a map in cartopy and plotting data on it
		Plotting focal mechanisms on a cartopy map
		Plotting a pretty map using cartopy
		Plotting heat map data on a map using cartopy
6	Exploratory data analysis	Maps for unshaped data using cartopy
		Introduction to plotting data as a heat map
		Introduction to scatter plots and histograms
		Creating a map in cartopy and plotting data on it
7	Jupyter Notebooks	Plotting heat map data on a map using cartopy
		All notebooks

Instead of having a linear progression (notebook 1, notebook 2, notebook 3, etc.), we assign Notebooks to learning goals to aid in the modularity of the materials.

Figure 2: Table 1 from Aiken et al. (2018), summarizing the main learning goals and associated Jupyter notebooks of their Python library for earthquake statistics computation and data visualization

1.4 PyMap

PyMap code development has started with the development of Python Jupyter Notebooks, focusing on the statistical seismology computational chain from input data (earthquake catalogs) to data quality analysis to computation of earthquake-catalog statistical parameters and results visualization. Tools to compute statistical parameters such as seismicity rates and b-values were developed as a Python library by Aiken et al. (2018). The corresponding Jupyter notebooks, which are freely available, provide tutorials for statistical seismicity computations intended for use by seismology students.

Figure 2 provides a list of the notebooks and their associated learning goals, which focus on data analysis using the Pandas library and visualization of earthquake statistics such as spatiotemporal b-value variations. The notebooks were successfully applied in a Master's level PSHA course at the University of Potsdam, providing exposure to map-making and statistical analysis using Python to students with little to no programming experience.

Because PyMap is designed as a computational toolbox developed in Python for statistical seismology, we found the approach using only Jupyter notebooks to not be sufficient for the projected tasks of PyMap. In particular, we did not want to implement the different coordinate systems and the abstract grid backend into a distribution of only notebooks. Therefore, we decided to first develop a pure toolbox implementation with higher abstraction to be later used in Jupyter notebooks (as a black box).

PyMap's default input dataset are earthquake catalogs and it comprises tools for data quality analysis, statistical parameters of earthquake catalogs and provides a viewer for data and result visualization. Except for the viewer, PyMap deliberately does not provide any user-interface tools as it is meant to be a toolbox (or function library) to be used in Python scripts, Jupyter notebooks, or included in own software projects.

The development in Python ensures largest possible flexibility regarding its use on different operating systems. While PyMap is developed under Linux, the toolbox functionality will work without modifications under MacOS and Windows. Because Python itself, the Python libraries (e.g. numpy, matplotlib, and PyMap) are all released under an open-source license, no license fee or permission is necessary for using and further developing PyMap.

PyMap addresses a common difficulty of earthquake statistics tools (e.g. ZMAP): correctly representing all key values over a large range of orders of magnitude and operating on different coordinate systems. Earthquakes or similar events are in general represented in three different coordinate systems:

- Earthquake hypocenter locations are in most cases represented in latitude, longitude, and depth. This classical system uses in most cases degrees for latitude and longitude and kilometers for depth. This representation is compatible with the standardized representation in QuakeML (Schorlemmer et al., 2004).
- Small-earthquake hypocenters in mines or induced-seismicity events hypocenters in injection sites are often represented with their northing, easting, and depth relative to a fixed location.
- Acoustic-emission locations in laboratory experiments are represented purely in relative coordinates to a fixed point of the sample (itself a relative coordinate).

PyMap unifies these three coordinate systems into a single one and provides the necessary library function to operate with the new coordinate system such that all further functions on PyMap are agnostic of the coordinate system. This approach has been chosen to avoid the common rewriting of functions to adapt to a different coordinate system. Any function written using PyMap's abstraction of the coordinate system will work on data from any of the three coordinate systems or even combinations thereof.

The newly introduced coordinate system of PyMap is called 5C (for five-element coordinates). 5C contains the five elements latitude, longitude, depth, northing offset, easting offset. The aforementioned coordinate systems are represented as

- Earthquake hypocenters are given as latitude, longitude, depth with the offsets set to zero.
- Events defined in relation to a fixed point use the latitude and longitude of the fixed point and store the lateral offset in the offset component. The depth represented the offset depth combined with the depth of the fixed point.
- Acoustic-emission hypocenters use any value (always the same for each sample) for the reference latitude and longitude and store the relative location in depth and the offsets.

PyMap defines various distance measures that operate on the 5C system so that all functions of the toolbox can use this abstraction layer to ensure full operational capabilities independent of the coordinate system used in the input data. Already implemented are an Euclidian distance and the distance along a great circle. Moreover, this system allows to combine regular earthquake catalogs with catalogs of induced seismicity (according to the second type) and process and visualize them together. Likewise, PyMap uses a large-range floating point variable for the focal time of the events to allow to store simultaneously thousands of years for (historic) earthquake catalogs as well as microseconds for laboratory experiments.

One of the standard features of such toolboxes like PyMap or ZMAP is the computation of the spatially varying parameters of earthquake catalogs. For simplicity, these are usually represented in a Cartesian grid. For regional analyses, the Cartesian grids are very convenient and simple to use. However, the larger the region, the greater the range of area covered by each grid cell; the closer a cell to the poles, the (relatively) smaller the cell (if defined in fractions of degrees).

To accommodate the need for equal-sized grid cells on a regional to global scale, PyMap offers grids based on triangular and hexagonal spherical tessellation. Similar to the solution for the different coordinate systems, PyMap provides a grid abstraction layer that allows any function using any type of grid to be grid-agnostic. This layer allows to query whether or not a point is within a cell, on a cell boundary and if so, to which cell the boundary belongs.

For importing earthquake catalogs into PyMap, we developed an interface to the QuakePy package. QuakePy is the reference implementation for QuakeML and provides import and export filter for various earthquake catalog formats. QuakePy simplifies complex earthquake catalogs for the use in PyMap.

PyMap at the current stage of development provides basic statistics tools (e.g. earthquake activity rates, b-values, etc.). The framework is designed in such a way that further statistical functions can be easily implemented given the abstraction layers for coordinate systems and grids.

As mentioned previously, PyMap comes with one graphical user interface tool. This is the 4D-viewer Sparrow. Sparrow operates by default on the globe and allows to zoom into any area of interest. It can display earthquake catalogs color-coded in full 3D, see Figures 3 & 4.

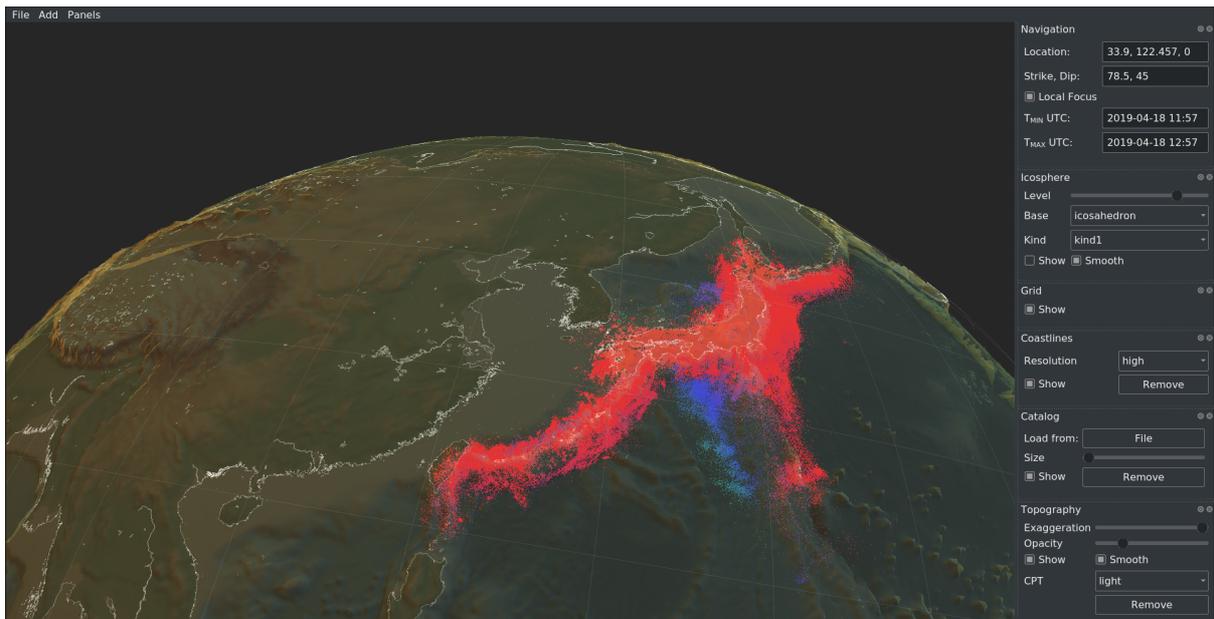


Figure 3: Sparrow display of the earthquake catalog of the Japan Meteorological Agency (JMA) together with coastlines, topography, and bathymetry.

The display can be augmented with coastlines, 3D-topography, and with seismological data, e.g. faults and source models of larger earthquakes, see Figure 5. Sparrow offers the user to find locations by latitude/longitude or by name searches (e.g. “Tokyo”).

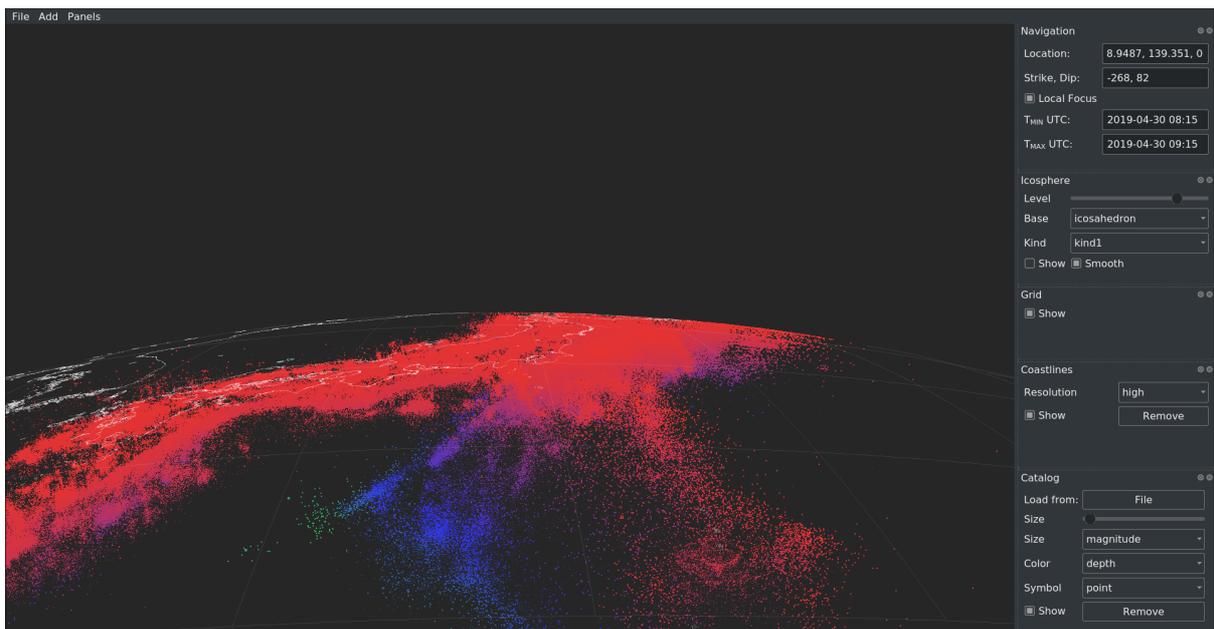


Figure 4: Details of the 3D view of the catalog of the Japan Meteorological Agency in Japan. Events are color-coded according to depths. The subducting slab is nicely visible.

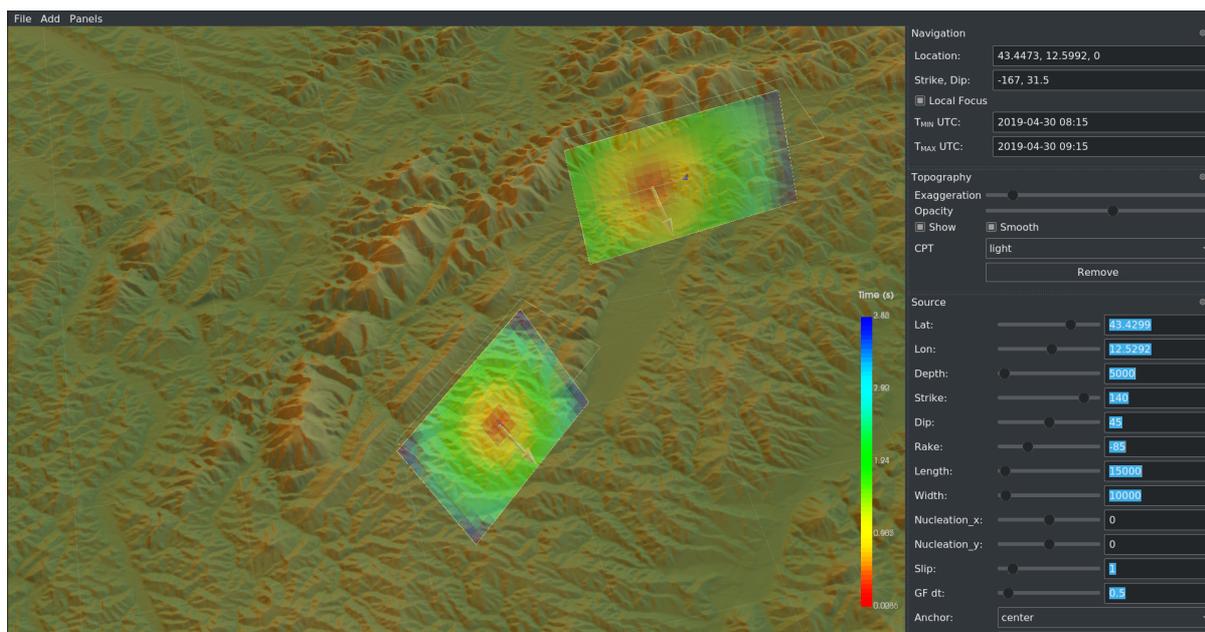


Figure 5: Sparrow display of two scenario earthquake sources. Color-coded is the rupture time. The sources are embedded into a topography display.

The fourth dimension in Sparrow is the time. Sparrow can be used to display the evolution of seismicity as an animation to help the user better identify features of interest.

Any computed value represented in any of the supported grid types can be displayed in Sparrow, see Figure 6. This feature, as often used in ZMAP on Cartesian grids, is one key element of PyMap. The new grid abstraction layer allows for the different supported grid types to be used for computations and to be displayed without the user having to develop the rather complex visualization procedure for the complex grids.

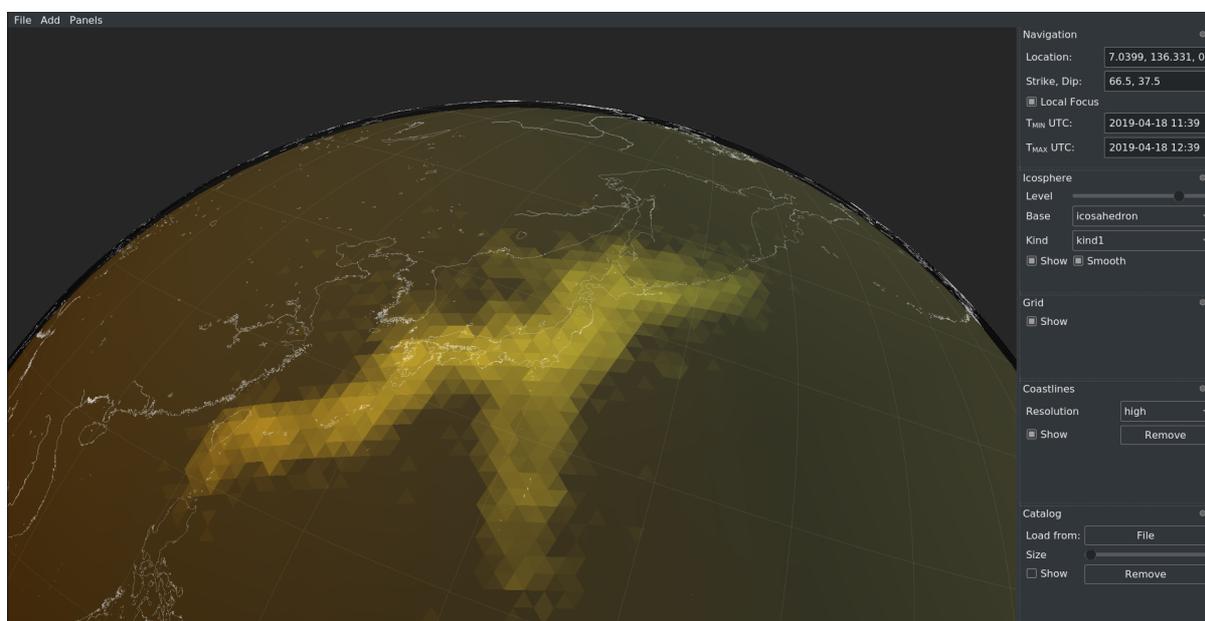


Figure 6: Sparrow display of the earthquake density of the earthquake catalog of the Japan Meteorological Agency (JMA) on a triangular grid over Japan.

Overall, the development of PyMap has not finished. The main task to develop the framework for PyMap is completed. In the near future, we will continue to fill PyMap with useful functions that will allow first users to quickly assemble statistical analyses of seismicity.

One particular development is very important. The Collaboratory for the Study of Earthquake Predictability is redesigning their earthquake forecast testing center software (Schorlemmer & Gerstenberger, 2007; Zechar et al., 2010) and, given the overlap in personnel, the PyMap and CSEP developments will be merged for all parts that are related to earthquake catalog manipulation and statistical seismology, i.e. the core part of PyMap. CSEP's system is also developed in Python and thus integration of PyMap tools will be relatively easy.

Further synergies stem from the collaboration with Pyrocko (Heimann et al., 2017), an open-source seismology toolbox and library, written in Python. Pyrocko is designed for performing a variety of geophysical tasks, like seismological data processing and analysis, modelling of InSAR, GPS data and dynamic waveforms, or for seismic source characterization. PyMap is available from the pyrocko repository. Please refer to pyrocko.org for further information about downloading pyrocko and PyMap.

1.5 References

Aiken, J. M., C. Aiken, and F. Cotton: A Python library for teaching computation to seismology students, *Seismological Research Letters*, Vol. 89, No. 3, p. 1165–1171, 2018.

Heimann, S., M. Kriegerowski, M. Isken, S. Cesca, S. Daout, F. Grigoli, C. Juretzek, T. Megies, N. Nooshiri, A. Steinberg, H. Sudhaus, H. Vasyura-Bathke, T. Willey and T. Dahm: Pyrocko - An open-source seismology toolbox and library. V. 0.3., GFZ Data Services, 2017, 10.5880/GFZ.2.1.2017.001

Reyes, C. and Wiemer, S. ZMAP7, a refreshed software package to analyze seismicity. In *Geophysical Research Abstracts* (Vol. 21), 2019.

Schorlemmer, D., A. Wyss, S. Maraini, S. Wiemer and M. Baer: QuakeML—An XML schema for seismology. *ORFEUS Newsletter*, 6(2), 9, 2004.

Schorlemmer, D. and M. Gerstenberger: RELM Testing Center. *Seismol. Res. Lett.*, 78(1), 30-36, 2007.

Wiemer, S.: A software package to analyze seismicity: ZMAP. *Seismological Research Letters*, 72(3), pp.373-382, 2001, 10.1785/gssrl.72.3.373

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling and T. H. Jordan: The Collaboratory for the Study of Earthquake Predictability – perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience*, 22(12), 1836-1847, 2010, 10.1002/cpe.1519

2 Chapter 2: An Objective Method For Estimating Time Variation Of Magnitude Of Completeness With Application To Eshm20 Earthquake Catalog

Authors: Shyam Nandan, ETH; Laurentiu Danciu, ETH; Stefan Hiemer, RichterX; Celso Guillermo Reyes, ETH; Stefan Wiemer, ETH; Domenico Giardini, ETH

This method and results presented in this chapter are soon to be submitted for peer review.

2.1 Summary

The magnitude of completeness (M_c) of earthquake catalogs is an indispensable parameter that has critical implications for probabilistic seismic hazard assessment, primarily because of the estimate of the exponent (b-value) of Gutenberg Richter (GR) law hinges on the reliable evaluation of M_c . The reliability of M_c estimation depends on the correct specification of the space-time partitions in which the assumption of constant M_c holds.

We have developed an automatic method that:

- a) detects the changes in the reporting rates of earthquakes with time for a given area
- b) uses it to define the time partitions with uniform reporting rates
- c) estimates M_c within each time partitions using maximum curvature method [Wiemer and Wyss, 2000].

We apply this method to the ESHM20 earthquake catalog and obtain the M_c time steps in predefined “Completeness Super Zones” (CSZs). Using the estimated M_c time steps, we evaluate the Gutenberg-Richter (GR) parameters: a-value, b-value, and corresponding 95% confidence intervals in all the CSZs. Finally, we report the sensitivity of the estimated time partitions, corresponding M_c , and the GR pars to the choices of critical hyperparameters.

The study region is divided into 51 shallow and 5 deep CSZs (see figures 1a and b). These spatial zonations are done based on the homogeneity of earthquake completeness history, i.e., the earthquake reporting rates are thought to be spatially homogenous within each of the CSZs. The shallow and the deep CSZs contain 56,283 and 3,673 earthquakes with *depth* $< 60\text{ km}$ and $\geq 60\text{ km}$, respectively. In tables 1 and 2 (also see figures 1c and d), we report the number of earthquakes in the shallow and deep CSZs individually.

2.2 ESHM20 Earthquake Catalog and Completeness Super Zones.

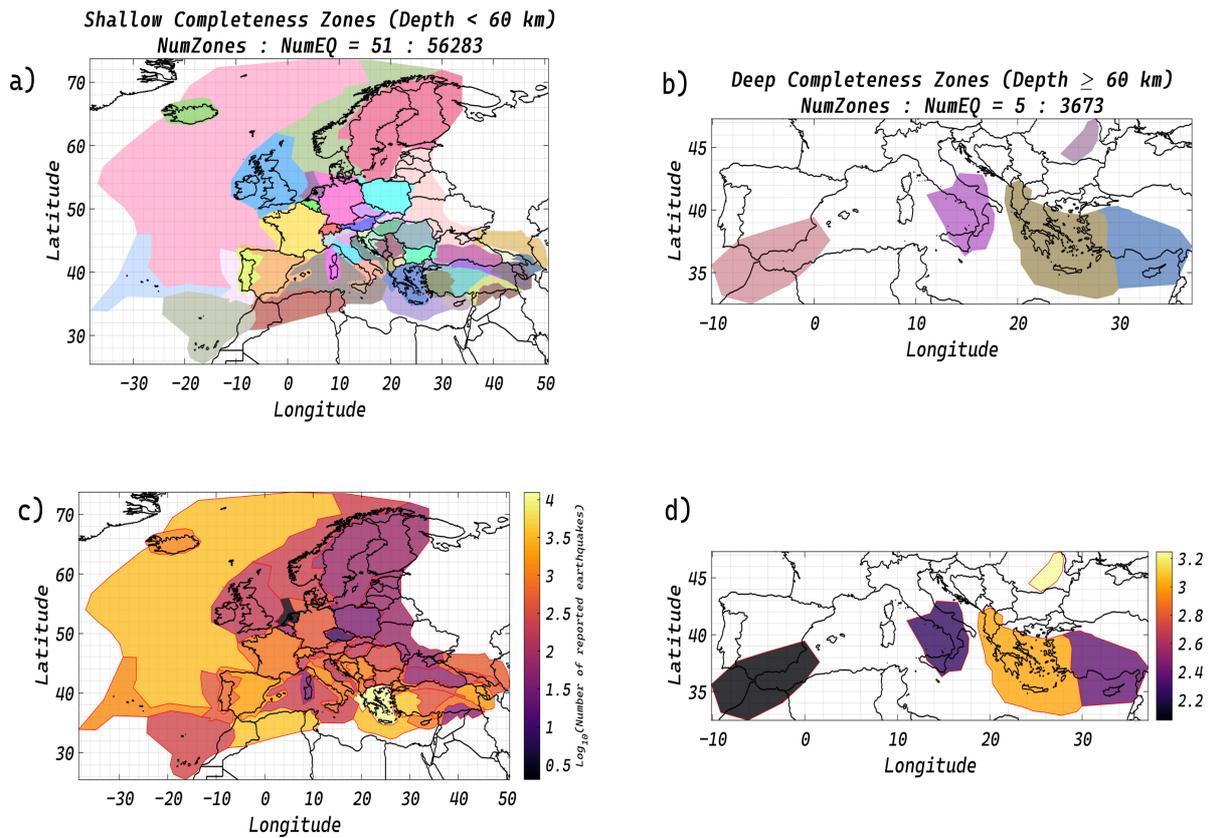


Figure 1: Completeness Super Zones (CSZs) and the total number of reported earthquakes; (a) 51 shallow CSZs containing earthquakes with *depth* < 60 km; (b) 5 deep CSZs containing earthquakes with *depth* ≥ 60 km; (c) the number of earthquakes in the shallow CSZs shown as a colormap; (d) the number of earthquakes in the deep CSZs.

Table 1: List of all shallow CSZs, their ids, and the number of earthquakes reported in each of them.

<i>CSZ_NAME</i>	<i>CSZ_ID</i>	<i>Number_of_Earthquakes</i>
<i>CSZ_AT</i>	SZ01	306
<i>CSZ_CH</i>	SZ02	1176
<i>CSZ_BL</i>	SZ03	79
<i>CSZ_DE</i>	SZ04	493
<i>CSZ_FR</i>	SZ05	933
<i>CSZ_UK</i>	SZ06	175
<i>CSZ_HR</i>	SZ07	182
<i>CSZ_IT1</i>	SZ08	575
<i>CSZ_NL</i>	SZ09	2
<i>CSZ_PL</i>	SZ10	37
<i>CSZ_RO</i>	SZ11	324
<i>CSZ_SK</i>	SZ12	128
<i>CSZ_SLO</i>	SZ13	658
<i>CSZ_CZ</i>	SZ14	9
<i>CSZ_SC</i>	SZ15	77
<i>CSZ_AL</i>	SZ16	986
<i>CSZ_BG</i>	SZ17	486
<i>CSZ_SRB</i>	SZ18	1298
<i>CSZ_ESP</i>	SZ19	2073
<i>CSZ_PT</i>	SZ20	681
<i>CSZ_TR2</i>	SZ21	4591
<i>CSZ_GR</i>	SZ22	12521
<i>CSZ_MK</i>	SZ23	327
<i>CSZ_CAUS</i>	SZ24	486
<i>CSZ_CY</i>	SZ25	1080

<i>CSZ_NAME</i>	<i>CSZ_ID</i>	<i>Number_of_Earthquakes</i>
<i>CSZ_SG</i>	SZ32	4015
<i>CSZ_BS</i>	SZ33	57
<i>CSZ_BIH</i>	SZ34	904
<i>CSZ_IT5</i>	SZ35	42
<i>CSZ_IT4</i>	SZ36	1528
<i>CSZ_IT2</i>	SZ37	870
<i>CSZ_SM</i>	SZ38	236
<i>CSZ_MM</i>	SZ39	141
<i>CSZ_IT6</i>	SZ40	666
<i>CSZ_TR7</i>	SZ41	1179
<i>CSZ_TR1</i>	SZ42	878
<i>CSZ_TR4</i>	SZ43	439
<i>CSZ_TR5</i>	SZ44	49
<i>CSZ_TR3</i>	SZ45	1829
<i>CSZ_TR6</i>	SZ46	383
<i>CSZ_IT3</i>	SZ47	247
<i>CSZ_CAR</i>	SZ48	558
<i>CSZ_NO</i>	SZ49	255

Table 2: List of all deep CSZs, their ids, and the number of earthquakes reported in each of them

<i>CSZ_NAME</i>	<i>CSZ_ID</i>	<i>Number_of_Earthquakes</i>
<i>CSZ_ES</i>	DZ01	114
<i>CSZ_IT</i>	DZ02	178
<i>CSZ_GR</i>	DZ03	1083
<i>CSZ_TR</i>	DZ04	241
<i>CSZ_RO</i>	DZ05	1768

2.3 Global Magnitude of Completeness, b-value, and a-value

In Figure 2(a), we show the empirical magnitude distributions (EMDs) for the shallow ESHM20 catalog for the three choices of magnitude discretizations (Δm). To obtain these EMDs, we first discretize the magnitudes reported in the catalog using either $\Delta m = 0.1, 0.2$, or 0.3 . We then count the number of times the different discrete magnitudes appear in the discretized catalog.

In figures 2(b) and 2(c), we show the dependence of the estimates the b and a values on the choice of the magnitude of completeness (M_c) for the three Δm 's. The estimates of b values are obtained using the following formula, which is valid for the magnitudes discretized at Δm intervals [Tinti and Mulargia, 1987; Marzocchi and Sandri, 2009].

$$b = \ln \left(1 + \frac{\Delta m \times N}{\sum_{i=1}^N (m_i - M_c)} \right) \times \frac{1}{\Delta m} \times \frac{1}{\ln 10} \quad (1)$$

In Equation 1, the m_i 's are N discrete magnitudes ($\geq M_c$) reported in the catalog.

Having estimated the b values, the a values can be estimated using the following formula:

$$a = \log_{10} N - \log_{10} \left(\sum_{j=1}^K 10^{-b(m_j - \frac{\Delta m}{2})} \right) - \log_{10} (1 - 10^{-b\Delta m}) \quad (2)$$

Note that, the a -values reported in Figure 2 have to be normalized by the duration of the catalog to obtain the usually reported 'yearly' a -values.

It is evident from Figure 2b and 2c that the choice of M_c has a major impact on the estimated of both b and a values, thus justifying the need for its proper assessment.

We find that as we increase the value of assumed M_c , the estimates of b and a increase. However, at $M_c \approx 4.8$, the estimates of both these parameters attain stability and do not change significantly. We use the stability in the estimates of b and a values as a proxy for the completeness of the catalog [Cao and Gao, 2002]. Based on this analysis, we can consider the shallow ESHM20 catalog to be complete for $M \geq 4.8$. A similar analysis for the deep ESHM20 catalog leads to the same outcome for the magnitude of completeness (Figure 3). We also find that the estimates of overall M_c are insensitive to the choice of Δm .

Using $M_c = 4.8$, we find that the global estimates of (a, b) values for the shallow ESHM20 catalog are (8.47, 0.92), (8.14, 0.86), and (8.46, 0.92) for the three choices of Δm 's respectively. For the deep ESHM20 catalog, the three options for Δm 's, in combination with $M_c = 4.8$ yield (6.11, 0.67), (5.84, 0.62), (6.2, 0.69) respectively as the estimates of (a, b) values.

Indeed, in making the global estimates of M_c , a and b values, we implicitly assume that these parameters do not vary in space and time. While there is no consensus on the Spatio-temporal variation of a and especially b values, no contention exists on the variation of M_c in space and time. The Spatio-temporal variation of M_c has mainly anthropogenic origins: as the coverage of the seismic networks varies in space and time, so does the magnitude of completeness. The estimation of this Spatio-temporal variation forms the subject of the next section.

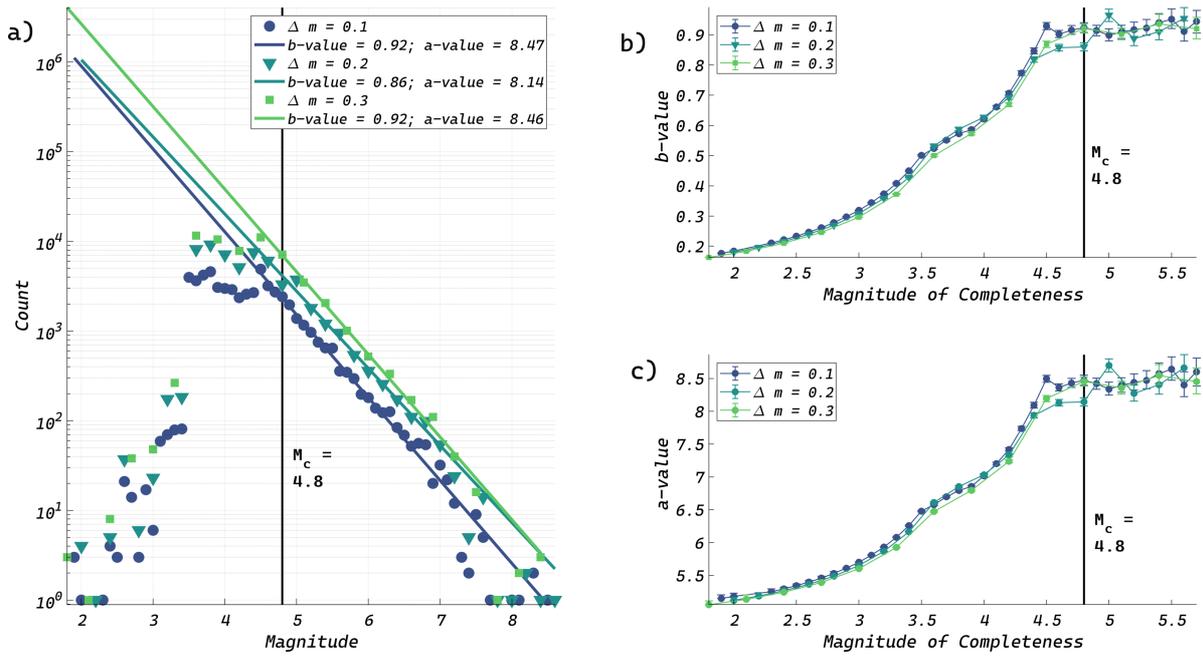


Figure 2: (a) Overall empirical magnitude distribution of the shallow earthquake catalog using three different magnitude discretizations ($\Delta m = 0.1, 0.2$ and 0.3); the estimated b-value and a-value, assuming $M_c = 4.8$, for the three different Δm 's are reported in the legend; (b-c) the estimates of b-value and a-value and their 95% confidence interval as a function of varying choice of M_c for the three different Δm 's; both b-value and a-value seem to stabilize for $M_c \approx 4.8$, thus justifying the decision of choosing $M_c = 4.8$ as the global estimate.

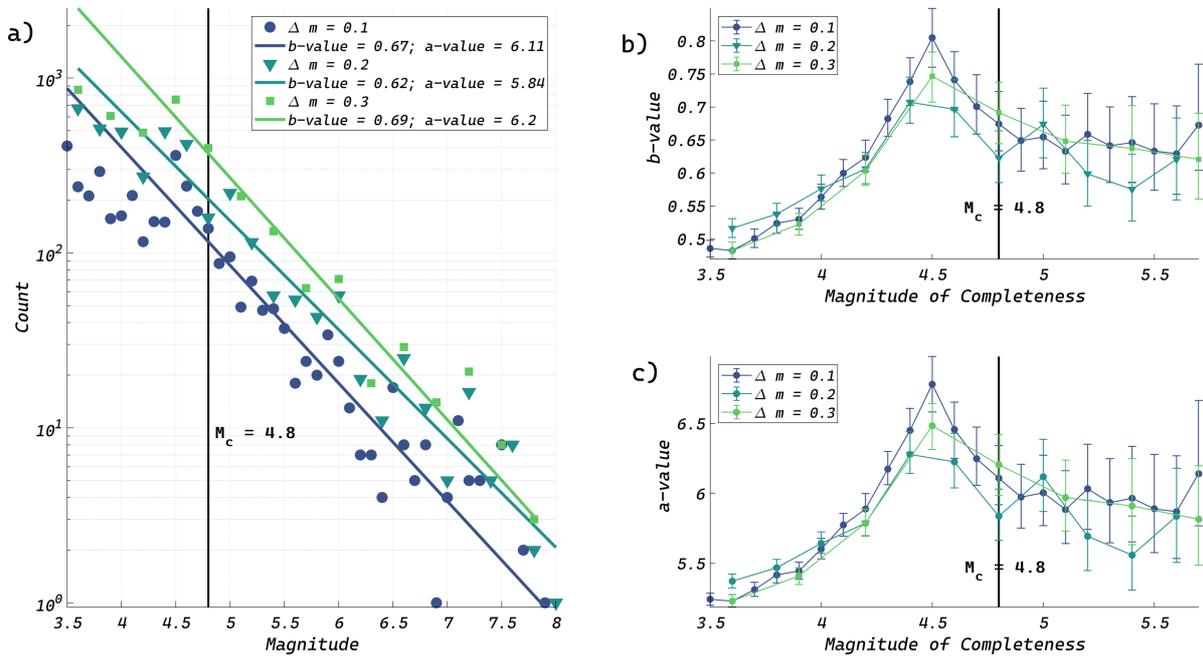


Figure 3: Same as Figure 2 but for the deep catalog

2.4 Time Series of Magnitude of Completeness

As mentioned in Section 1, we use the spatial partitions (or CSZs) provided by the experts and only restrict ourselves to inferring the time partitions and corresponding magnitude of completeness. We assume M_c to be piecewise constant within each of the time partitions and estimate it using the non-parametric Maximum-Curvature method. In the following: we first define the algorithm used to infer the magnitude of completeness time series (MCTS), explain some of the concepts and hyper-parameters used in the algorithm, and then illustrate its working on some selected CSZs. We present the result of the application of the algorithm on all the other CSZs in Appendix #.

2.4.1 Algorithm

1. Bin the reported magnitudes at Δm intervals, to obtain the discretized magnitudes m^Δ . Make a list of unique magnitudes (U) from the list of discretized magnitudes. Let the list be composed of n individual elements $\{m_1 \dots m_n\}$, such that $m_1 > m_2 > \dots > m_n$.
2. Initialize:
 - a. $t_{\text{current}} = t_1$, $m_{\text{current}} = m_1$, where t_1 is the time of the first event in the catalog and m_1 is the first entry of the list U .
 - b. $i = 1$ and $j = 1$.
 - c. Set the magnitude of completeness time series, MCTS, as an empty list.
3. If $t_{\text{current}} \geq t_{\text{last}} - T_{\text{gap}}$ or $i > n$, go to step 12. Note that, t_{last} is the time of the last event in the catalog and T_{gap} is the minimum time duration allowed for a completeness period.
4. Create a list L_{current} , whose entries are the times (t), the discretized magnitudes (m^Δ) of those earthquakes which satisfy the condition: $t \geq t_{\text{current}}$ and $m^\Delta \geq m_{\text{current}}$. Let the list be composed of N earthquakes.
5. If $N < N_t$, set $i = i + 1$, $m_{\text{current}} = m_i$ and go to step 3.
6. Compute the empirical normalized cumulative number time series ($\text{NCNTS}_{\text{emp}}$) on a predefined time grid ($t_{\text{cur}}: dt: t_{\text{last}}$). $\text{NCNTS}_{\text{emp}}$ at any time t is defined as:

$$\text{NCNTS}_{\text{emp}}(t) = \frac{N(\leq t)}{N} \quad (3)$$

where, $N(\leq t)$ is the number of earthquakes in L_{current} with times at least t . Define distance, d_{emp} , as:

$$d_{\text{emp}} = \max[\text{NCNTS}_{\text{theo}}(t) - \text{NCNTS}_{\text{emp}}(t)] \quad (4)$$

where $\text{NCNTS}_{\text{theo}}(t) = t \times \frac{N}{(t_{\text{last}} - t_{\text{current}})}$ is the theoretical normalized cumulative number time series.

7. Simulate N times, uniformly at random, within the time interval $[t_{\text{current}}, t_{\text{last}}]$. Compute the $\text{NCNTS}_{\text{sim}}$ and d_{sim} as defined in equations 3 and 4.
8. Repeat step 6, several times (say 1000) and make a list of d_{sim} estimated from each simulation. From this list, estimate D_{sim}^q corresponding to the q^{th} quantile of the distribution of d_{sim} .
9. If $D_{\text{sim}}^q < d_{\text{emp}}$, set $t_{\text{current}} = t_{\text{current}} + dt$ and go to step 3.
10. Using the magnitudes in the list L_{current} and the Maximum-Curvature method, estimate M_c .
11. Update the MCTS using the pair $(t_{\text{current}}, M_c)$ based on any of the following applicable conditions. Let the last entry of the MCTS be $(T_{\text{last}}^{\text{last}}, M_c^{\text{last}})$
 - a. If $t_{\text{current}} > T_{\text{last}}^{\text{last}} + T_{\text{gap}}$ and $M_c < M_c^{\text{last}}$ append $(t_{\text{current}}, M_c)$ to the MCTS.
 - b. If $t_{\text{current}} > T_{\text{last}}^{\text{last}} + T_{\text{gap}}$ and $M_c \geq M_c^{\text{last}}$, leave MCTS unchanged.

- c. If $t_{\text{current}} \geq T^{\text{last}} + T_{\text{gap}}$ and $M_c < M_c^{\text{last}}$, change the last entry of MCTS to (T^{last}, M_c) .
 - d. If $t_{\text{current}} \geq T^{\text{last}} + T_{\text{gap}}$ and $M_c \geq M_c^{\text{last}}$, leave MCTS unchanged.
 - e. If MCTS does not contain any entries, append $(t_{\text{current}}, M_c)$ to the MCTS.
- Set $t_{\text{current}} + dt$, $i = i + 1$, $m_{\text{current}} = m_i$ and go to step 3.

12. Return MCTS.

2.4.2 Explanation of the Algorithm and Description of the Important Hyper-Parameters.

The algorithm described in section 3.1 is the result of the following intuitions:

1. Over time, more and smaller earthquakes are recorded, as a result M_c can be assumed to decrease with time.
2. Above M_c , the times of reported earthquakes are expected to uniformly distributed if the catalog is properly declustered.

To test if a set of data points are uniformly distributed, we perform a statistical test of uniformity as described in steps 6-9. We computed the distance between the cumulative density function (CDF) of the observed empirical distribution and the theoretical CDF expected from a perfectly uniformly distributed dataset. We then compare this distance to the distribution of distances computed for random data simulated from a uniform distribution. We reject/accept the uniformity hypothesis based on the condition specified in Step 9. The distance threshold required to reject the uniformity hypothesis depends on the value of hyper-parameter q , which has to be prespecified.

It is important to note that the distance defined in Equation 3 is a modified form of Kolmogorov-Smirnov (KS) distance. The modification is done to account for the imperfections of the declustering methods, which sometime fail to remove the aftershocks from the catalog. As a result, the observed normalized cumulative number time series (NCNTS) often shows a concave bulge relative to the theoretical NCNTS expected from an adequately declustered catalog. Using the modified KS distance (Equation 3), we automatically get rid of this problem, as those deviations in the observed NCNTS get automatically ignored. As the proposed takes declustered catalogs as its input, the choice of the declustering method naturally becomes one of the vital hyper-parameters.

In the proposed algorithm, we use the Maximum-Curvature method [Wiemer and Wyss, 2000] to estimate the magnitude of completeness for any given completeness period. We primarily made this choice as the Maximum-Curvature is non-parametric and very simple to use. However, there are more sophisticated methods for the estimation of the magnitude of completeness [Mignan and Woessner, 2012]. Thus, the choice of the method for the evaluation of the magnitude of completeness is the third important hyper-parameter.

Last but not least, the hyper-parameters Δm and T_{gap} are respectively used to discretize the reported magnitudes and to ensure that the minimum time duration of any completeness period is at least T_{gap} .

In summary, the proposed algorithm has 5 hyper-parameters: q , Δm , T_{gap} , choice of the declustering method, selection of the M_c estimation method. As there is no obvious way to optimize these parameters, we should study the sensitivity of M_c estimates to these hyper-parameters.

2.4.3 Demonstration Report: Applicability to the ESHM20 Unified Earthquake Catalog

In this section, we first present the completeness time series for the 5 selected CSZs. These selected CSZs include: CSZ_CH, CSZ_DE, CSZ_FR, CSZ_UK, and CSZ_GR, which respectively contain 1176, 493, 973, 175 and 12,521 earthquakes. The results for all other CSZs are presented in the Appendix.

For obtaining these time series, we make the following choices for the 5 hyper-parameters: $q = 95\%$ ile, $\Delta m = 0.2$, $T_{\text{gap}} = 10$ years, window based declustering method (Gardner and Knopoff, 1974) with space-time windows specified by Gruenthal (personal communication) for declustering the catalog (winGT), and Maximum-Curvature method for estimation of M_c .

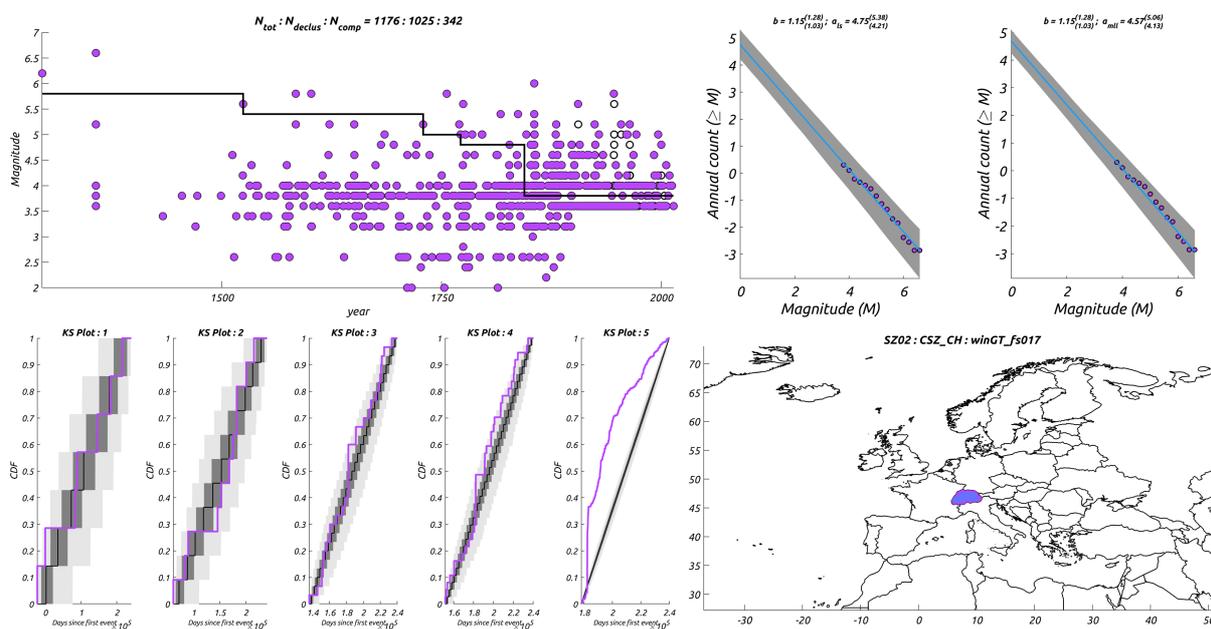


Figure 4: Time series of the magnitude of completeness, KS plots for the resulting completeness periods, annualized empirical and best fit magnitude distribution given the completeness periods and corresponding M_c for the CSZ_CH completeness superzone; (top left panel) N_{tot} , N_{declus} , and N_{comp} represent the total number of earthquakes reported in the CSZ, number of earthquakes remaining after application of winGT and number of declustered earthquakes above completeness; Purple (filled) circles show the declustered earthquakes, empty circles show the earthquakes labelled as aftershocks (or foreshocks) by the declustering method; Solid black line indicates the piece-wise constant M_c time series; (bottom left panels) Theoretical NCNTS (solid black line) and its 95% confidence bounds (shaded light grey region) and observed NCNTS (purple line) for the (from earliest to latest) inferred completeness periods; (top right panels) annualized empirical (purple circles) and best Gutenberg-Richter (solid blue line) magnitude distribution; Both magnitude distributions have been obtained using the data above the inferred magnitude of completeness; The estimates of the parameters of GR distribution are indicated at the top of the panel along with their 2.5%ile(subscript) and 97.5%ile (superscript), which are estimated using bootstrapping; The b-values are calculated using the Equation 1, while the a-values (with ml subscript) are estimated using either Equation 2 and a-values (with ls subscript) are estimated using least square fitting of the empirical magnitude distribution. Both a-values are estimated on the basis of the same b-value.

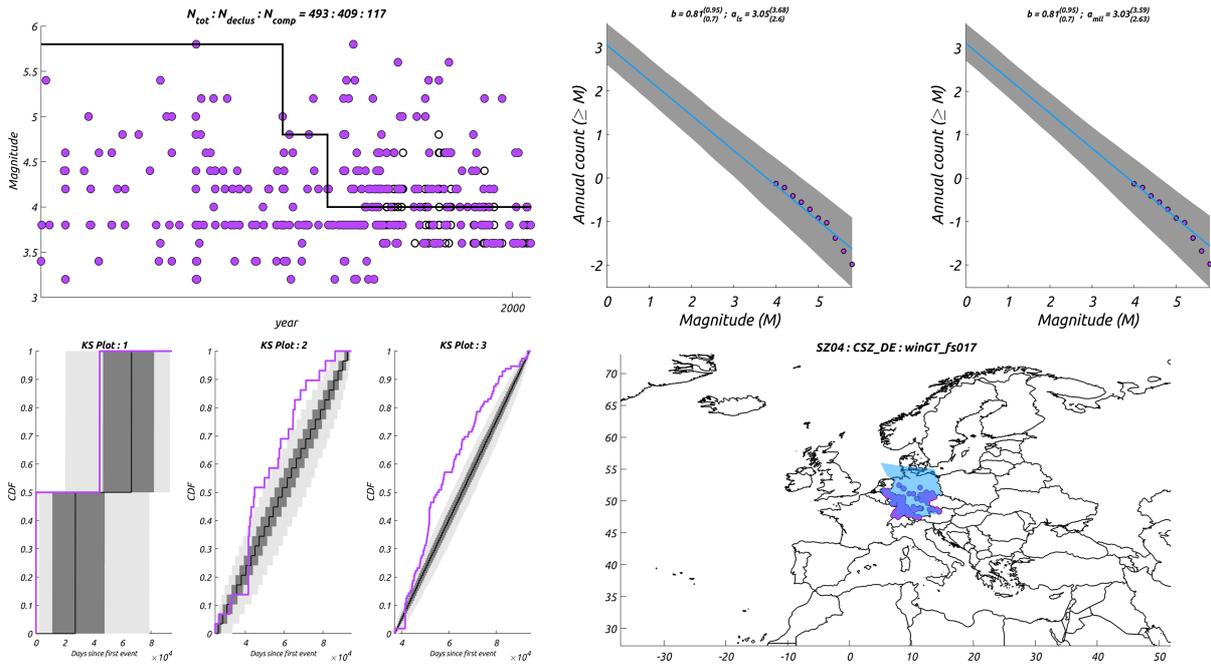


Figure 5: Same as Figure 4 but for CSZ_DE

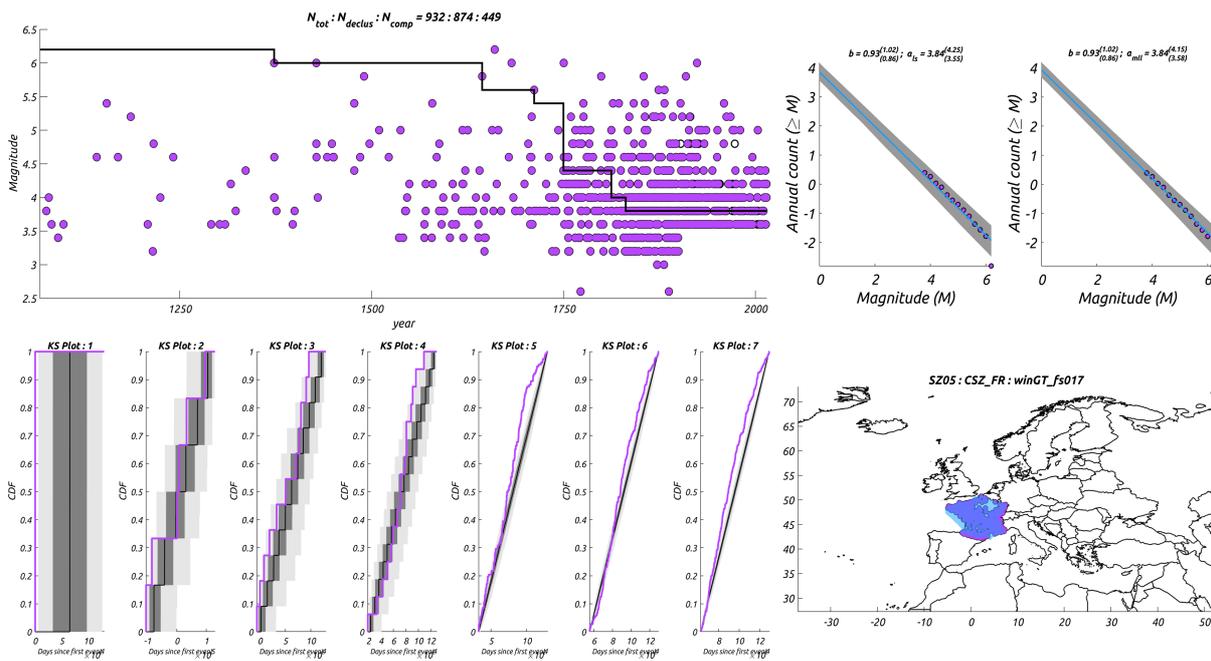


Figure 6: Same as Figure 4 but for CSZ_FR

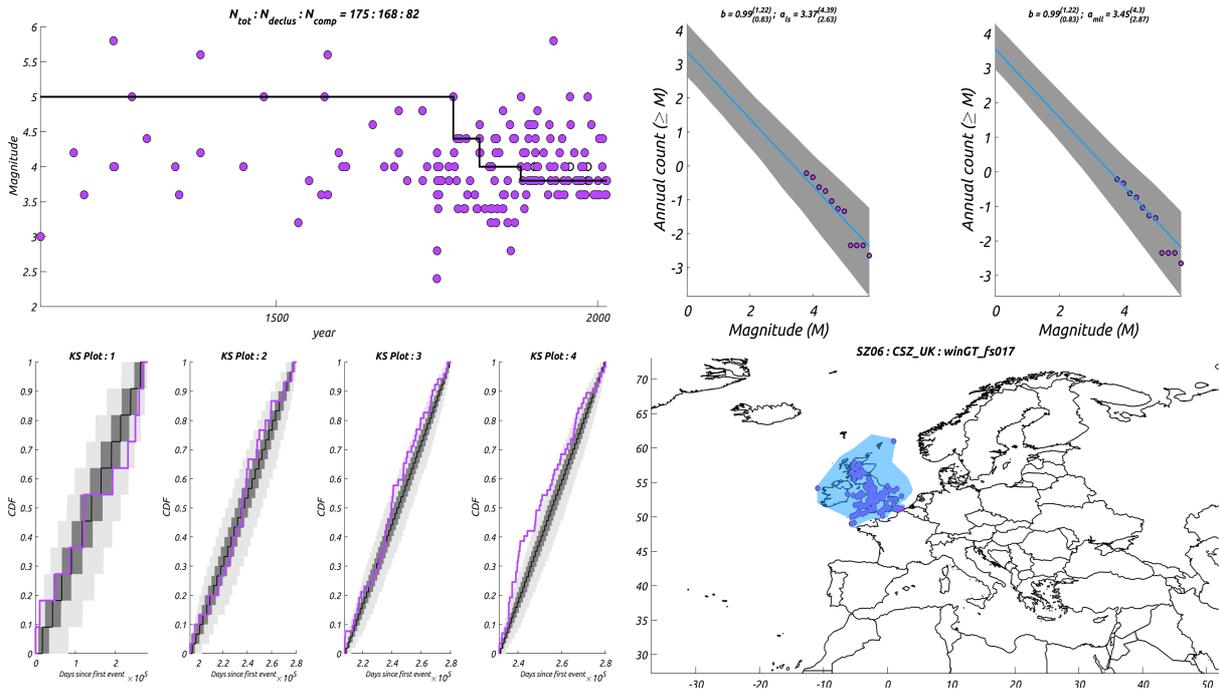


Figure 7: Same as Figure 4 but for CSZ_UK

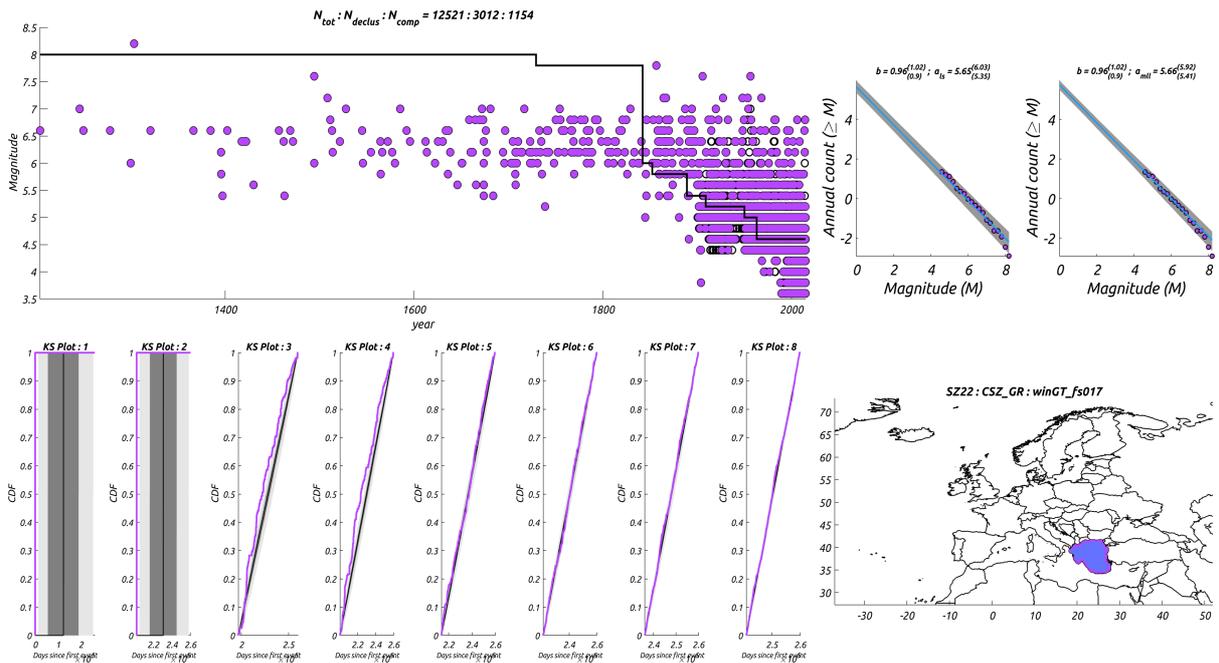


Figure 8: Same as Figure 4 but for CSZ_GR

2.5 Outlook

The analysis presented in this study primarily lacks in the following regards, which need to be addressed in complementary future studies:

1. The proposed method has been directly applied to the ESHM20 unified earthquake catalog, on which the ground truth about regarding the magnitude of completeness and the resulting a and b -values are not known. Thus, the reliability of the method cannot be fully ascertained. To properly assess the reliability of the method, we have to apply it on the synthetic catalog on which the underlying ground truth is known.
2. As the method relies on the knowledge of the hyper-parameters and spatial partitions, which have to be pre-specified by the experts, the technique remains partly expert-driven. To make the method more objective, we need to find ways to optimize the choices of these hyper-parameters and the spatial partitions.

Finally, the estimated completeness time series present new opportunities for improvement of seismicity based forecasting models, such as Epidemic Type Aftershock Sequence models (ETAS) [Nandan et al., 2019a,b,c].

2.6 References

- Cao, A., & Gao, S. S. (2002). Temporal variation of seismic b -values beneath northeastern Japan island arc. *Geophysical research letters*, 29(9), 48-1.
- Marzocchi, W., & Sandri, L. (2009). A review and new insights on the estimation of the b -value and its uncertainty. *Annals of geophysics*, 46(6).
- Mignan, A., & Woessner, J. (2012). Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*, 1-45.
- Nandan, S., Ouillon, G., & Sornette, D. (2019). Magnitude of Earthquakes Controls the Size Distribution of Their Triggered Events. *Journal of Geophysical Research: Solid Earth*, 124(3), 2762-2780.
- Nandan, S., Ouillon, G., Sornette, D., & Wiemer, S. (2019). Forecasting the Full Distribution of Earthquake Numbers Is Fair, Robust, and Better. *Seismological Research Letters*, 90(4), 1650-1659.
- Nandan, S., Ouillon, G., Sornette, D., & Wiemer, S. (2019). Forecasting the rates of future aftershocks of all generations is essential to develop better earthquake forecast models. *Journal of Geophysical Research: Solid Earth*, 124(8), 8404-8425.
- Tinti, S., & Mulargia, F. (1987). Confidence intervals of b values for grouped magnitudes. *Bulletin of the Seismological Society of America*, 77(6), 2125-2134.
- Wiemer, S., & Wyss, M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the western United States, and Japan. *Bulletin of the Seismological Society of America*, 90(4), 859-869.

3 Chapter 3: Time-Dependent Induced-Seismicity Models, Their Dependency On Time-Varying Operational Parameters And Guidelines For Time-Dependent Hazard Evaluation

Authors: Konstantinos Leptokarpoulos, Stanisław Lasocki, Monika Sobiesiak, Piotr Sałek, - IGPAS

3.1 Summary

Many seismic processes, specifically those induced or triggered by activities for exploitation of georesources, are time-dependent. As a result, the corresponding seismic hazard posed by such processes is time-dependent as well. Within Task 24.3, a software tool for short term, time-dependent hazard analysis has been developed and implemented. This tool is SHAPE (Seismic HAZard Parameters Evaluation). It is particularly, though not exclusively, relevant for anthropogenic seismicity investigation. SHAPE enables an assessment of time-dependent hazard quantified by the Mean Return Period (MPR) of a given magnitude and the Exceedance Probability (EP) of a given magnitude within a predefined time period. Hence SHAPE estimates the time-dependent source component of seismic hazard. The variation in time of this component originates from the time-variability of industrial factors driving seismic activity. SHAPE is therefore useful to monitor the changes of seismic response to technological operations and to control the effectiveness of the undertaken hazard mitigation strategies. Nevertheless, SHAPE can be evenly applied to non-anthropogenic seismicity cases without any limitations.

In section 3.2 an overview of induced seismicity and its characteristic features are presented. Those features are closely related to anthropogenic activities, therefore induced seismicity and the corresponding hazard demonstrate a significant time-dependency.

In **section 3.3** the proposed methodology for dealing with time-dependent anthropogenic seismic hazard on industrial sites is described. In **section 3.4** the SHAPE software package is presented. In **section 3.5** an application of SHAPE package to the data from the north western part of The Geysers (TG) geothermal field, California, is demonstrated.

Parts of this material have been gathered as the manuscript Leptokarpoulos, K. and S. Lasocki: "SHAPE: A MATLAB software package for time-dependent seismic hazard analysis", which has been accepted for publication in Seismological Research Letters and presently is under publishing process.

3.2 Demonstration Report: Time-and-Technology dependent Induced Seismicity

Due to its significant socio-economic impact, seismicity induced or triggered by exploitation of georesources and the related hazards receive increasing scientific and public interest. Anthropogenic activities such as fluid injection and extraction, mining operations and water reservoir impoundment perturb stresses and lead to the occurrence of considerable seismic activity, even in areas previously characterized as aseismic. All of these activities alter the rock matrix equilibrium by applying complex mechanical, hydraulic, thermal and chemical

interactions. The combination of the aforementioned phenomena causes seismic activity which may potentially lead to events strong enough to threaten the integrity of the infrastructure and cause problems to the production process. Occasional, stronger anthropogenic events may even result to casualties and damages in extended areas leading to remarkable public concern. The well-known case of Basel 2006 earthquake led to project cancelation (insurance claims reached ~7 million CHF, Gischig and Wiemer, 2013) and the recent Pohang Mw 5.5 earthquake (Ellsworth et al., 2019) in Korea (135 injuries, 1700 people displaced from their homes, ~\$300 million total damage, Lee et al., 2019) are only two of the mostly known cases of seismicity associated with georesources exploitation. The vast economic impact as well as the vicinity of the epicentres to urbanized areas strengthens the need of accurate hazard assessment in the areas surrounding industrial sites. It is therefore of paramount importance to develop analytical tools, which could lead to implementation of risk mitigation measures. The problem is intrinsically time-dependent because the anthropogenic seismic processes are tightly linked to the inducing, time-variable technological operations.

Seismic hazard, determined as the level of shaking at a given point, caused by an earthquake, consists of three components, which are symbolically referred to as source, path and site. The source component includes the properties of seismicity, the path component refers to the properties of seismic waves propagation from the source to the receiving point and the site component includes properties of the medium at the receiving point, which have an impact on the resulting ground motion. Out of these three components, only the source component changes in time in most of the anthropogenic seismicity cases.

In this connection, time-changes of the probabilistic parameters of seismicity representing the source component of hazard, indicate the trends of induced seismic hazard development, as well as the effectiveness of the undertaken hazard mitigation actions.

Anthropogenic seismicity demonstrates some characteristic features which differentiate it from tectonic seismicity. Among those features, some of them are particularly relevant for seismic hazard evaluation purposes:

First, induced seismic events are usually limited to a specified volume in the vicinity of the inducing technological activities. Second, the induced seismicity energy release is, in the vast majority of the reported cases, lower than the natural one. Third, there is a close (yet not always straightforward) relation between seismic activity and production/ operational parameters (e.g. mined out mass/volume, injection rate/pressure, reservoir water level etc), therefore the process is intrinsically non-stationary and time-dependent.

Following these attributes, it is preferred to study and interpret changes of specified hazard parameters rather than their absolute values. For example, at a given site, e.g. at the vicinity of an injection well, where events with $M \geq 3.0$ occur once per year, then the corresponding exceedance probability of $M \geq 3.0$ within a time period of $dt=1\text{day}$ is very low. Nevertheless, given the fact that seismic hazard is strongly time-dependent, such an exceedance probability should be compared with the corresponding values estimated for previous time periods.

A significant change of this parameter in comparison to its previous values, would provide strong indication of a remarkable change of seismic hazard, regardless of the parameter

absolute value. In addition, because seismic hazard parameters change fast due to the technological activities variation, a selection of a long time horizon, dt , would have no practical meaning.

Earthquake magnitude distribution is routinely considered as exponential, modelled by the well-known Gutenberg-Richter (GR) law, parameterized by the so-called b -value, which quantifies the relative likelihood of stronger earthquakes. As generally stated in literature, b -values close to one are typical for tectonic seismicity (e.g. El-Isa and Eaton, 2014 and references therein), whereas higher b -values (close to or even higher than 2) are often evident in anthropogenic seismicity (e.g. Wyss, 1972). Nevertheless, complexity and rapid changeability of technological factors inducing seismicity may result in significant deviations of the observed magnitude distributions from the GR law (Lasocki, 2017).

It has been shown (Lasocki, 2001; Urban et al., 2016) that the GR law may be not appropriate to model the magnitude distribution in anthropogenic seismic hazard analysis. Preliminary results (Lasocki et al., 2017) indicated a complex (i.e. non-exponential) magnitude distribution in TG as well as in Oklahoma region and propose the adaptation of a non-parametric approach for modelling the events size distribution for seismic hazard purposes.

Leptokaropoulos (2020) showed that the entire magnitude distribution at the north western part of TG geothermal field is definitely complex and non-exponential with the shape of the distribution demonstrating at least 2 bumps (the shape of distribution subsequently changes from convex to concave). The same author performed a spatio-temporal seismicity analysis and specified 10 seismic clusters: 3 of these clusters demonstrate b -values ~ 1.10 and correspond to low injection rates; 3 other clusters demonstrate b -values ~ 1.40 and correspond to medium injection rates; The rest 4 clusters are generally associated with high injection rates and the magnitude exponentiality hypothesis is rejected by the Anderson-Darling test at 0.01 significance. This deviation from exponentiality, may be caused by rapid changes of stress values and orientations due to changes in pore pressure as a result of fluid injection. Moreover, thermal stresses and chemical effects change material properties and dynamic response (e.g. Majer et al., 2007; Izadi and Elsworth, 2015) introducing additional complexity to the process. Such effects have not yet been sufficiently examined and deeply understood, however, they definitely result to time-dependent seismic hazard, tightly connected with anthropogenic activities.

Due to these facts, there is a need for alternative modelling of magnitude distribution to apply when GR model is clearly inadequate. For this reason non-parametric (data-driven) approaches are also implemented in SHAPE software. The detailed description of magnitude distribution models, together with the methodology followed for time-dependent seismic hazard assessment is presented in the following sections.

3.3 Methodology

The methodology of time-dependent hazard analysis is based on the works of Lasocki (e.g. 1993a, 1993b; 2017). In his approach the dependence on time of seismic hazard is modelled by the dependence on time of parameters of the stationary distribution models.

Next, it is assumed that hazard changes in time are slow enough to be approximated by stationarity in time intervals whose length allows for estimating the distribution model parameters. In result, the hazard estimates are assessed through successive estimations in a sliding time window, under the assumption that the seismic process within each window is stationary.

When the seismic process is stationary, thence when this process is studied in an individual time window, the source component of seismic hazard is characterized by the distribution of number of event occurrences in the prescribed time period, by the magnitude distribution of events and by the distribution of epicentre or hypocentre location, all distributions being independent of time. Because, as mentioned, anthropogenic seismic sources occur over a limited volume, the distribution of source location is often not modelled and it is assumed that the hazard values are the same within the whole engaged part of the rock mass. Such an approach is used in SHAPE.

Usually, also in SHAPE, it is assumed that the seismic process is Poissonian. Then the event occurrences are fully characterized by the seismic activity rate, λ , which is estimated by the number of events that occurred in the time window divided by the time window length.

Two parameters related to seismic hazard are estimated in SHAPE. The first is the Mean Return Period (MRP) of a given magnitude, M_1 , defined as the mean time elapsed between successive events of $M \geq M_1$:

$$MRP = \frac{1}{\lambda(1-F_m(M_1))} \quad (1)$$

where, λ , is the seismic activity rate of the events with magnitude greater than or equal to the catalog completeness level, M_c , and F_m is the Cumulative Distribution Function (CDF) of magnitude.

The second parameter estimated is the Exceedance Probability (EP) defined as the occurrence probability of the earthquake of magnitude M_1 , within a time period, dt:

$$EP = 1 - e^{-\lambda dt(1-F_m(M_1))} \quad (2)$$

Four different methods of F_m estimation are supported by SHAPE, two assuming the validity of the Gutenberg-Richter (GR) law and two Non-Parametric (NP) approaches:

Unbounded GR model (GRU)

The assumption that earthquake magnitudes statistically follow the GR relation, such that $\log N = a - bM$, where N is the number of earthquakes with magnitude $\geq M$, and that there is no limit for earthquake magnitude, leads to the negative exponential distribution of magnitude with the Probability Density Function (PDF), f_m , given as:

$$f_m(M) = \beta e^{-\beta(M-M_c+\frac{\Delta M}{2})}, M \geq M_c \quad (3)$$

where ΔM is the magnitude round-off interval (reporting accuracy) and the parameter β is connected to the GR law b -value as $\beta = \ln 10 \cdot b$.

The corresponding CDF reads:

$$F_m(M) = 1 - e^{-\beta(M-M_C+\frac{\Delta M}{2})}, M \geq M_C \quad (4)$$

The maximum likelihood estimate of β , for grouped magnitude values within their round-off interval is given by (Bender, 1983):

$$\frac{1}{\beta} = \langle M \rangle - M_C + \frac{\Delta M}{2} \quad (5)$$

where $\langle M \rangle$ is the arithmetic mean of magnitudes of events with $M \geq M_C$.

Truncated GR model (GRT)

Assuming a hard end point of the magnitude distribution, M_{max} , the magnitude PDF for $M_C \leq M \leq M_{max}$ reads (Page, 1968):

$$f_m(M) = \frac{\beta e^{-\beta(M-M_C+\frac{\Delta M}{2})}}{1 - e^{-\beta(M_{max}-M_C+\frac{\Delta M}{2})}} \quad (6)$$

The corresponding CDF reads:

$$F_m(M) = \begin{cases} 0 & \text{for } M < M_C \\ \frac{1 - e^{-\beta(M-M_C+\frac{\Delta M}{2})}}{1 - e^{-\beta(M_{max}-M_C+\frac{\Delta M}{2})}} & \text{for } M_C \leq M \leq M_{max} \\ 1 & \text{for } M > M_{max} \end{cases} \quad (7)$$

The maximum likelihood estimate of β is given as the root of the equation (Page, 1968):

$$\frac{1}{\beta} + \frac{\hat{M}_{max} - M_C + \frac{\Delta M}{2}}{1 - e^{\beta(\hat{M}_{max} - M_C + \frac{\Delta M}{2})}} - \langle M \rangle - M_C + \frac{\Delta M}{2} = 0 \quad (8)$$

In SHAPE the upper bound of magnitude distribution, M_{max} , is estimated using the generic formula of Kijko and Sellevoll (1989):

$$\hat{M}_{max} = M_{maxobs} + \int_{M_C}^{M_{maxobs}} [F_M(M)]^k dM \quad (9)$$

where M_{maxobs} is the largest magnitude value in the k-element sample of the observed magnitudes with $M \geq M_C$.

The estimation of the distribution parameters, β , M_{max} requires numerical solving of the system of two equations, (8) and (9). If this process fails to reach convergence, then the simplified formula of Robson and Whitlock (1964) is applied:

$$\hat{M}_{max} = 2M_{maxobs} - M_{max2obs} \quad (10)$$

where $M_{max2obs}$ is the second largest magnitude value.

Unbounded NP (NPU) and Truncated NP (NPT) models

In addition to the aforementioned two parametric approaches, non-parametric (model-free) estimators of magnitude distribution are implemented in SHAPE as well. These non-parametric approaches have been introduced and adapted to estimation of magnitude distribution by

Lasocki et al. (2000) and Kijko et al. (2001), further developed in Orlecka-Sikora and Lasocki (2005) and Lasocki and Orlecka-Sikora (2008). These approaches are based on the kernel density estimator which sums the symmetric probability densities (kernels), individually associated with the observations (Silverman, 1986):

$$\widehat{f}_m(M|\{M_i\}, h) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{M-M_i}{h}\right) \quad (11)$$

where, h , is the non-negative smoothing parameter (bandwidth), n stands for the number of observations, $\{M_i, i = 1, \dots, n\}$, are the magnitudes, and $K(\bullet)$, is the kernel function. The Gaussian kernel is used in SHAPE and the estimators of magnitude PDF and CDF read, respectively:

$$\widehat{f}_m(M) = \frac{\sum_{i=1}^N \left[\left(\frac{1}{\alpha_i h}\right) e^{-\frac{1}{2} \left(\frac{M-M_i}{\alpha_i h}\right)^2} \right]}{(\sqrt{2\pi}) \sum_{i=1}^N \left[\Phi\left(\frac{M_{\max}-M_i}{\alpha_i h}\right) - \Phi\left(\frac{M_c-M_i}{\alpha_i h}\right) \right]} \quad (12)$$

$$\widehat{F}_m(M) = \frac{\sum_{i=1}^N \left[\Phi\left(\frac{M-M_i}{\alpha_i h}\right) - \Phi\left(\frac{M_c-M_i}{\alpha_i h}\right) \right]}{\sum_{i=1}^N \left[\Phi\left(\frac{M_{\max}-M_i}{\alpha_i h}\right) - \Phi\left(\frac{M_c-M_i}{\alpha_i h}\right) \right]} \quad (13)$$

where $\Phi(\bullet)$ denotes the standard Gaussian CDF, and h is calculated from the equation (Kijko, et al., 2001):

$$\sum_{i,j} \left\{ \left[\frac{(M_i-M_j)^2}{2h^2} - 1 \right] \exp\left[-\frac{(M_i-M_j)^2}{4h^2}\right] - 2 \left[\frac{(M_i-M_j)^2}{h^2} - 1 \right] \exp\left[-\frac{(M_i-M_j)^2}{2h^2}\right] \right\} / \sqrt{2} = 2n \quad (14)$$

$\alpha_i = \sqrt{g/\tilde{f}(M_i)}$, $i = 1, \dots, N$ are local bandwidth factors, which cause that the smoothing factor adapts to uneven data density along the magnitude range,

$$\tilde{f}(M_i) = \frac{1}{N h \sqrt{2\pi}} \sum_{k=1}^N \exp\left[-0.5 \left(\frac{M_i-M_k}{h}\right)^2\right], \quad g = \left[\prod_{i=1}^N \tilde{f}(M_i) \right]^{\frac{1}{N}} \quad (15)$$

As in the parametric approach, setting $M_{\max} \rightarrow \infty$, leads to the NPU, whereas a finite value of M_{\max} leads to the NPT model.

The non-parametric approaches to seismic hazard estimation showed that they provide results with tolerable, limited errors regardless of whether the actual magnitude distribution follows the Gutenberg-Richter relation or it is complex (Kijko et al., 2001). The drawback of these approaches is that they need considerably numerous magnitude data samples, the best $N \geq 50$ elements, which can be difficult to obtain from short time windows in the presented time-dependent hazard estimation.

3.4 Software Toolbox: SHAPE package

SHAPE (Leptokaropoulos and Lasocki, 2020) facilitates a time-dependent hazard analysis by estimating the activity rate, b -value, mean magnitude and hazard parameters: MRP and EP in sliding time windows, according to the methodology presented above. The development of SHAPE resulted from a combination of different relevant applications which are already implemented within the IS-EPOS platform (Orlecka-Sikora, et al., 2020) and are freely available for on-line usage (<https://tcs.ah-epos.eu/>). Within the IS-EPOS platform various tools can be

found for the use in stationary and time-dependent probabilistic seismic hazard assessment problems. Based on the aforementioned on-line applications available in IS-EPOS, the SHAPE package presented here embodies additional features and constitutes a generalized, stand-alone Matlab software.

Two SHAPE versions are available: SHAPE_ver1 is a stand-alone version in which the hazard analysis is performed within a series of steps, allowing a high interactivity level with the User (Figure 1). This version supports a GUI in order to allow the User interactively select the options and parameter values needed for the calculations.

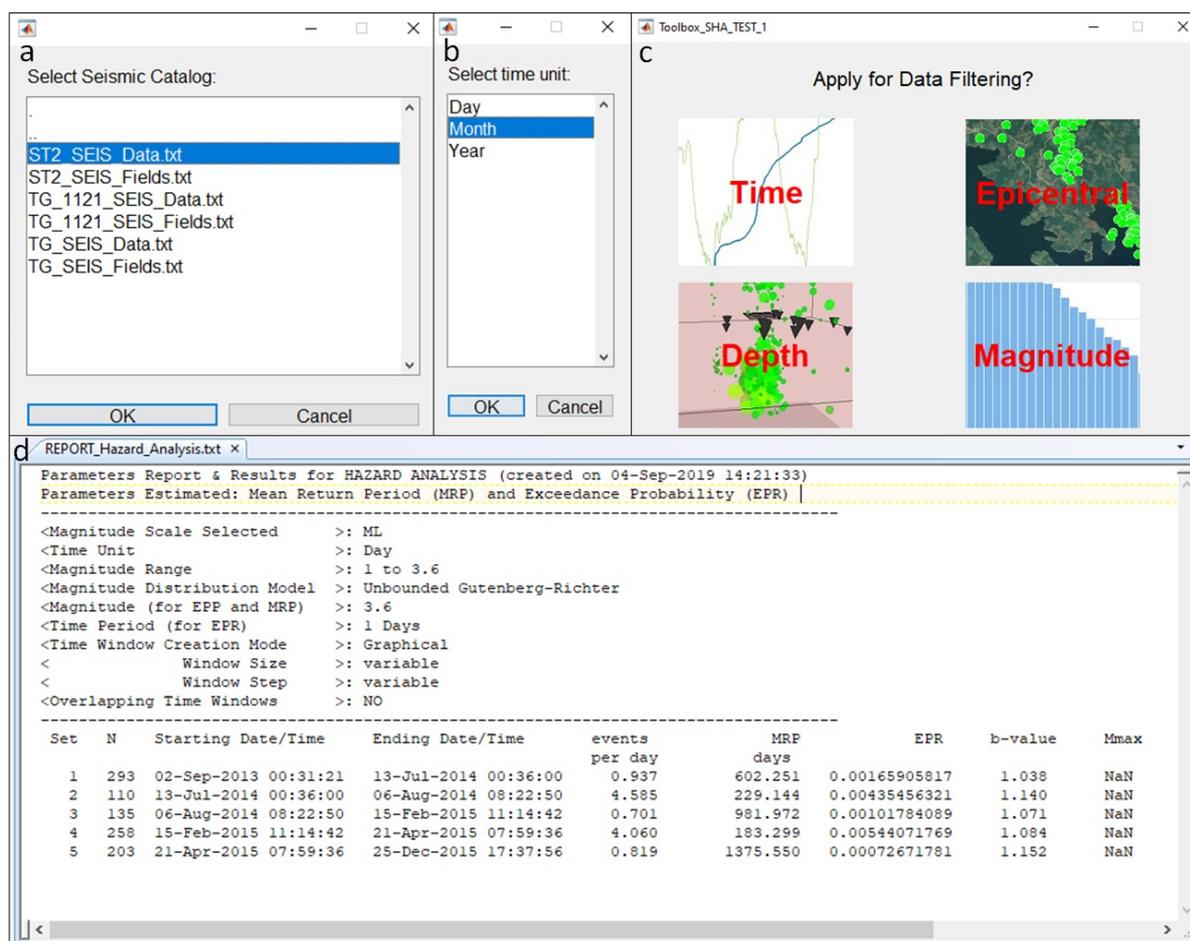


Figure 1. Snapshots from application of SHAPE_ver1 interactive version. (a) Seismic Catalog selection window, (b) time unit selection window, (c) filtering routine selection window and (d) output ASCII file.

SHAPE_ver2 is performed internally by the system as a series of steps and the input arguments are defined by the User in a so-called, Wrapper script. Once these parameters are set and the Wrapper script runs, the Application is performed without any further interruption. The input files in both versions must be in ASCII format (e.g. *.txt). For both SHAPE versions, three Input Directories must be available, one mandatory, containing the seismic catalog and two optional including production data and the parameters for time windows for the (time-dependent) analysis. In addition to the hazard estimates the package offers visualization of the results and

generation of a report summarizing the input parameter values and the output results (Figure 1d).

The package can be downloaded and implemented under GNU General Public Licence and is freely available to all users. Both versions comprise functions and auxiliary scripts written in Matlab and they are compatible with Matlab Version 2017b or later. They also require the 'Statistics and Machine Learning' Matlab Toolbox. SHAPE_ver1 also requires the 'Image Processing' Matlab Toolbox to support the GUI environment.

The source codes together with the accompanying material (data sample files, relative documents and complete User Guide documentation for SHAPE_ver1 and SHAPE_ver2) describing step by-step the implementation process, acceptable data formats, description of the parameters etc, can be found in the following repository:

https://git.plgrid.pl/projects/EA/repos/sera-applications/browse/SHAPE_Package.

The workflow of SHAPE is described by the flowchart shown in Figure 2. These steps are interactively performed in SHAPE_ver1, which, in addition facilitates a data filtering routine and graphical selection of the time windows (these steps are indicated as dark grey boxes in Figure 2). For this reason it is recommended that the user applies SHAPE_ver1 at the first place for a bulk investigation of a dataset in order to find potential sub-sets as well as appropriate parameter values leading to substantial results.

Once data and input options have been approximately constrained then the user may switch to SHAPE_ver2, for fast iterations, allowing fine tuning of the parameter values and comparison of the results obtained by diverse inputs. The workflow of SHAPE_ver1 is summarized below (the numbers in the circles shown in Figure 2 correspond to the numbers of the steps of analysis):

Step 1. Mode selection, between "Seismic Data" (i.e. mode 1) and "Seismic and Production Data" (i.e. mode 2). Although only seismic catalog is needed for calculation of hazard parameters there is an option to upload operational data as well. As mentioned in the previous section, anthropogenic seismicity properties and seismic hazard are well established as being directly connected with the inducing technological activities. For this reason operational parameters can be considered within SHAPE for the selection of appropriate time windows as well as for the visualization of the output in order to facilitate results interpretation. Nevertheless, mode 2 can be disregarded when natural (tectonic) seismicity is studied.

Step 2. The Data selection is done from pop-up windows. The user selects a seismic catalog and the corresponding catalog fields file from the "CATALOGS" directory. If Mode 2 is selected (see: Step 1), the user also selects a technological activity data file and the corresponding data fields file from the "PRODUCTION DATA" directory. In this latter case, the user is further requested to specify a particular parameter (e.g. reservoir water level, gas production volume, wellhead pressure, etc.) to be displayed in time filtering (see: Step 4), in interactive time window selection (see: Step 5) and in the output visualization (see: Step 7).

Step 3. The program reads from the input file the available magnitude scales in the uploaded catalog and requests from the user to select one of them. All the calculations from this point and on will be performed considering this magnitude scale.

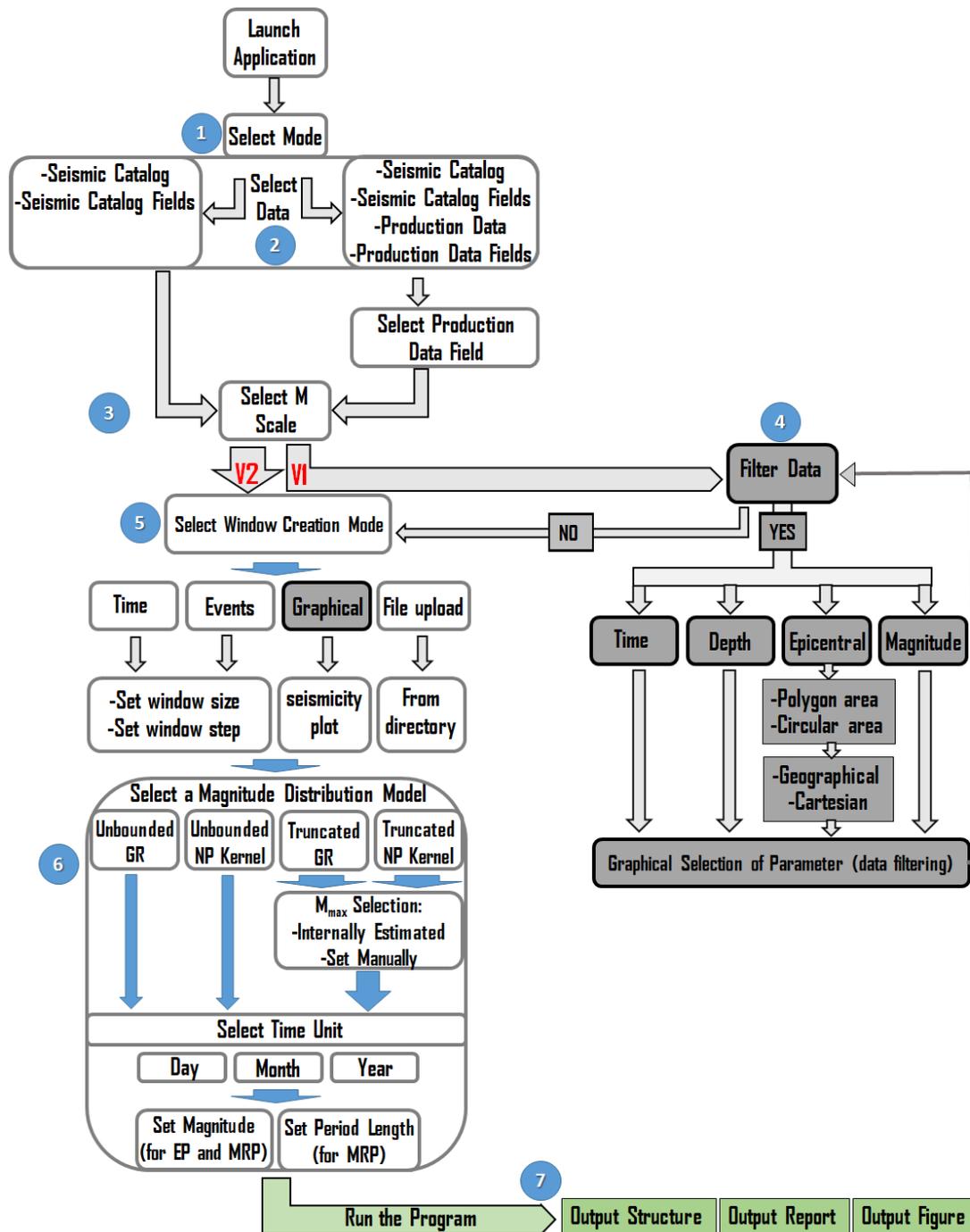


Figure 2. Flowchart with SHAPE basic processing workflow. The numbers within the circles correspond to the steps of the process as described in the main text. Dark boxes show the operations performed only in SHAPE_ver1. V1 and V2 refer to SHAPE_ver1 and SHAPE_ver2, respectively.

Step 4. This step consists of an iterative process which takes place only in SHAPE_ver1. The user has the chance to constrain the uploaded catalog data in terms of four filters, selected from a pop-up window. The user initially selects whether he/she wishes to perform filtering, thus the same filter can be applied many time as long as the user selects ‘yes’ in the corresponding dialog box. Each time a filter has been applied, a message appears in the screen showing the remaining number of events in the filtered catalog and the program requests from the user further data filtering. Once user’s choice is ‘no’ the program proceeds to step 5. The available filters are:

- a. Time filtering: The user is requested to select a starting and an ending time point from a graph showing the cumulative number of events in time, in order to constrain the period of analysis between these two points. If production data have been uploaded (see: Step 2) then the corresponding time-series of the selected technological parameter is also plotted in the second vertical axis of the same graph.
- b. Epicentral location filtering: This filtering can be applied in either geographical or Cartesian coordinates, if they are included in the uploaded dataset (seismic catalog). After selecting the coordinate system from a pop-up window, the user is provided two additional options, either a polygonal or a circular area, for constraining the events to be considered for seismic hazard analysis. In both cases, the user graphically specifies and adjusts the area.
- c. Depth filtering: The user can inspect the vertical distribution of the events as well as a histogram of events from a figure generated in a pop-up window. The user can change the number of bins to update the histogram and finally selects the depth range to be considered for the analysis.
- d. Magnitude filtering: The User is here requested to choose graphically the minimum magnitude threshold (essentially corresponding to the catalog completeness level), from a histogram representing the frequency magnitude distribution of the events in the uploaded dataset.

In SHAPE_ver2 no data filtering takes place, except the magnitude filtering by defining a minimum magnitude threshold within the Wrapper script. Therefore the data should be already filtered according to the user specifications either from SHAPE_ver1 implementation or externally (by own means).

Step 5. The remaining data after filtering is now divided in windows defined by the user by means of 4 different modes, i.e. 'Time', 'Events', 'Graphical' and 'File'. In SHAPE_ver2, the 'Graphical' option is not available. If 'Time' mode is selected, the User has to define the window size and window step (in days), by typing values in the corresponding fields in a pop-up window appearing in the screen. If 'Events' mode is selected the User has to define the window size (in events) and the window step (in days). Alternatively, the User can select 'graphical', for interactive graphical selection of subsequent points from a plot. Finally if 'File' is selected the program browses the "TIME_WINDOWS" directory and the user selects from there a file with the starting and ending points of the time windows to be considered.

Step 6. The user now selects the magnitude distribution model and other input parameters for seismic hazard analysis. The magnitude distribution model is selected among the four provided by SHAPE (GRU, GRT, NPU, and NPT, see: Methodology section). In addition, the target magnitude for EP and MRP calculation and the target time period for EP calculation are set as well. For the truncated distribution models (GRT and NPT) an option is available, in which the maximum magnitude is calculated by SHAPE together with its bias (Lasocki and Urban, 2011). SHAPE offers also the option to manually set a predefined M_{max} , which can be estimated independently e.g. from historical records of tectonic earthquakes, McGarr (2014) method or seismogenic index (Shapiro et al, 2010) in the case of induced seismicity, theoretical scaling relations (Galis et al., 2017), etc. Finally, the time unit (day, month or year) in which the final results (I and MRP) are calculated is selected at this step.

Step 7. Outputs. There are three outputs produced by the program and saved in "Outputs_SHA" directory:

1. A matlab structure “SHA.mat” containing fields with inputs and output parameters and information on the time windows. The structure has as many cells as the number of time windows generated.
2. A report, ‘REPORT_Hazard_Analysis.txt’ is generated and stored, including a summary of the input parameters and data considered, as well as the results obtained from the analysis.
3. A figure in .mat as well as in .jpg format is stored. This is automatically generated in SHAPE_ver1 and optionally generated in SHAPE_ver2 (Figure 3).

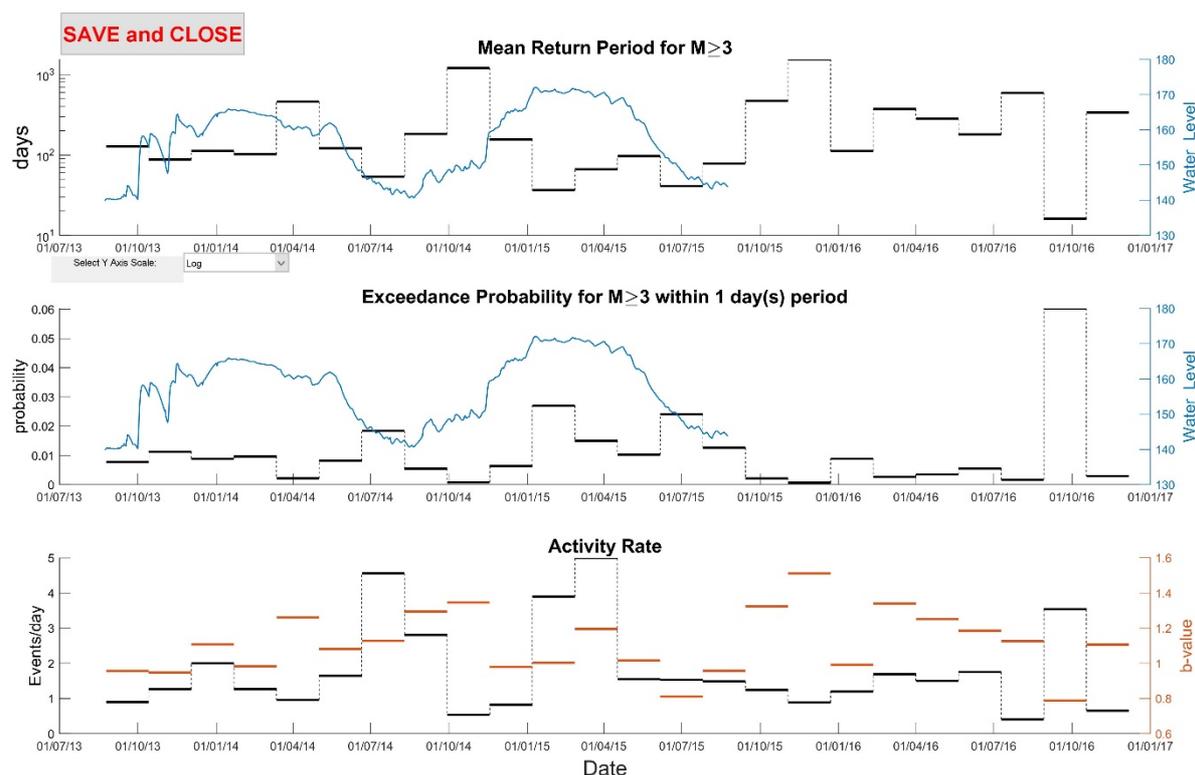


Figure 3. Example of an output figure produced by SHAPE, considering equal time windows.

The subsequent horizontal bars indicate: Upper frame - the mean return periods of events with $M \geq 3.0$ calculated for each one of the 30 days long time windows (notice the option to switch between linear/logarithmic y-scale), Middle frame- the exceedance probabilities of the $M \geq 3.0$ within a selected target period ($dt=1$ day in this case) and Lower frame - the mean activity rate for each time window. The blue curves in upper and middle frames denote the daily fluctuation of the production parameter (water level in the reservoir in this example). The brown horizontal bars in the lower frame show the b-values for each one of the time-windows.

3.5 Demonstration Report: Case study

The selected site in this example is The Geysers (TG) geothermal field, California, the largest geothermal system in the world, operating since the 1960's. Seismicity with $M > 2.0$ started after 1969 and in 1982 an event with $M=4.6$ occurred, which is the second largest event ever occurred in a geothermal site (as of March 2020). The analysis presented here is conducted on an isolated seismic cluster located at the North-Western part of TG (data can be found at: [https://tcs.ah-epos.eu/#episode:THE GEYSERS Prati 9 and Prati 29 cluster](https://tcs.ah-epos.eu/#episode:THE_GEYSERS_Prati_9_and_Prati_29_cluster)). A relocated seismic catalog comes from Martínez-Garzón et al. (2014) and Kwiatek et al. (2015), whereas several studies have been conducted already for this particular area dealing with association

of seismicity properties with injection activities (e.g. Staszek et al., 2017; Garcia-Aristizabal, 2018; Leptokaropoulos et al., 2018; Orlecka-Sikora and Cielesta, 2020; Lasocki and Orlecka-Sikora, 2020).

In this application, SHAPE is used for estimating hazard source parameters considering both parametric and non-parametric magnitude distribution models. Convertito et al. (2012) suggested that due to the limited dimension of the seismogenic volume in anthropogenic seismicity case studies, a truncated magnitude distribution (bounded between M_c and M_{max}) must be preferred. For this reason the truncated distributions (GRT, NPT) are chosen to be tested and compared with each other in this case study. A cluster of 1121 seismic events located in the close vicinity (<600m) from Prati-9 injection well is analysed. The maximum magnitude was set equal to $M_{max}=3.2$, as resulted from the application of the Kijko-Sellevoll generic formula (9) and the truncated Gutenberg-Richter magnitude distribution model (GRT, Equations 6 and 7) to the entire population of 1121 events in the selected area (the maximum observed magnitude was equal to 3.16). The target magnitude for MRP and EP was set equal to 2.75 (10 events with magnitude greater than or equal to 2.75 occurred within the entire study period). The target period length for EP was set to $dt=1$ day. The time windows considered for the analysis are related to injection rate values and magnitude distribution properties, as derived by Leptokaropoulos (2020). The time window selection criterion was chosen to be the p-value of the Anderson-Darling (AD) test of exponentiality, under the null hypothesis, H , that the magnitudes of a dataset (within a selected time window) follow the exponential distribution. A trade-off between window size and number of events included in each window was necessary in order to achieve robust results. In such way, after examining and combining the results for different window widths (25-150 events) optimal time windows were further sought semi-manually, based on the periods recognized from the earlier steps to demonstrate significantly high and significantly low p-values derived from the AD test. Eventually, 10 time windows were defined.

Figure 4 shows the results obtained by both the exponential (GR law) as well as the non-parametric approaches. It is shown that the non-parametric approach provides smaller probabilities of exceedance for $M>2.75$ in all but 2 cases (the inverse stands for the MRP). In order to test and quantify the efficiency of each method, the actual (observed) number of events ($M>2.75$) for each time window is plotted together with the expected number of events ($M>2.75$) considering GRT and NPT approaches, respectively. This expected events number can be derived as the duration of each time window divided by the corresponding MRP. These results are presented in Figure 5. It is shown that in almost all time windows, the NPT approach gives a number closer to the actual one than the GRT approach does. It is also noteworthy that for the first time window both approaches severely overestimate the $M>2.75$ events (0 actual versus 4 from NPT and 7 from GRT). The empirical and modelled CDFs by GRT and NPT approaches are demonstrated in Figure 6.

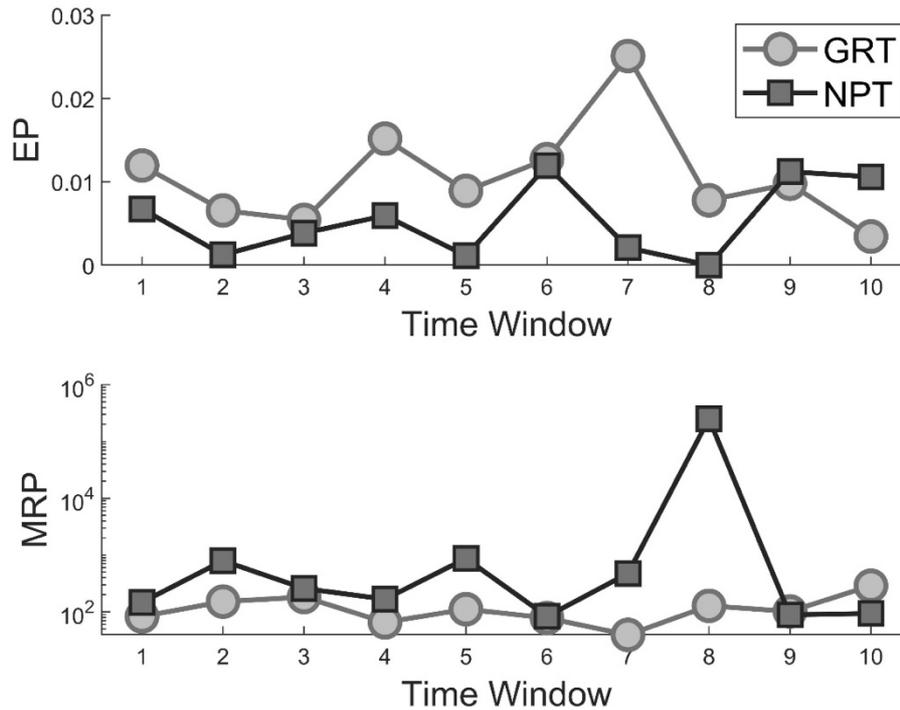


Figure 4. Exceedance Probability (upper frame) and Mean Return Period (lower frame) at TG (Prati-9 site), for 10 non-overlapping time windows. The light circles indicate the Truncated Gutenberg Richter model, whereas the dark squares indicate the Truncated Non Parametric model.

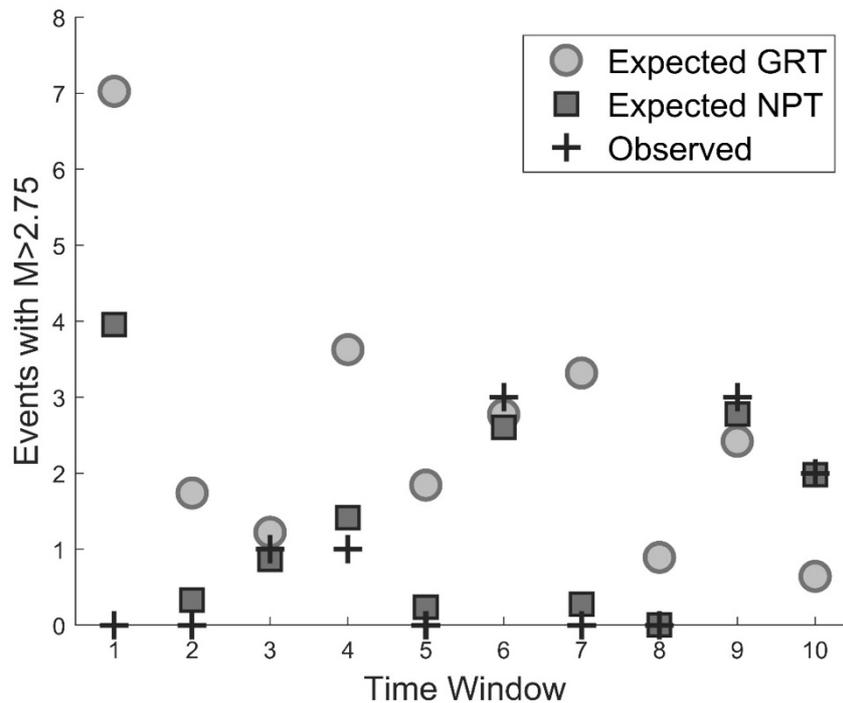


Figure 5. Observed (crosses) and expected number of events with $M > 2.75$ found during each one of the 10 time windows. The light grey circles correspond to GRT prediction whereas the dark grey squares indicate the NPT predicted values.

The performance of both approaches is quantified by means of the Spearman rank correlation coefficient (R^2), calculated for all 10 time windows and also excluding the first window (as an outlier). For the GRT the $R^2 = -0.08$ ($p\text{-value} = 0.83$) for all time windows and $R^2 = 0.10$ ($p\text{-value} = 0.80$) for the 9 latter time windows. This clearly indicates that the GR model is not

appropriate to describe the magnitude distribution at the particular site, failing to achieve agreement with the observed values.

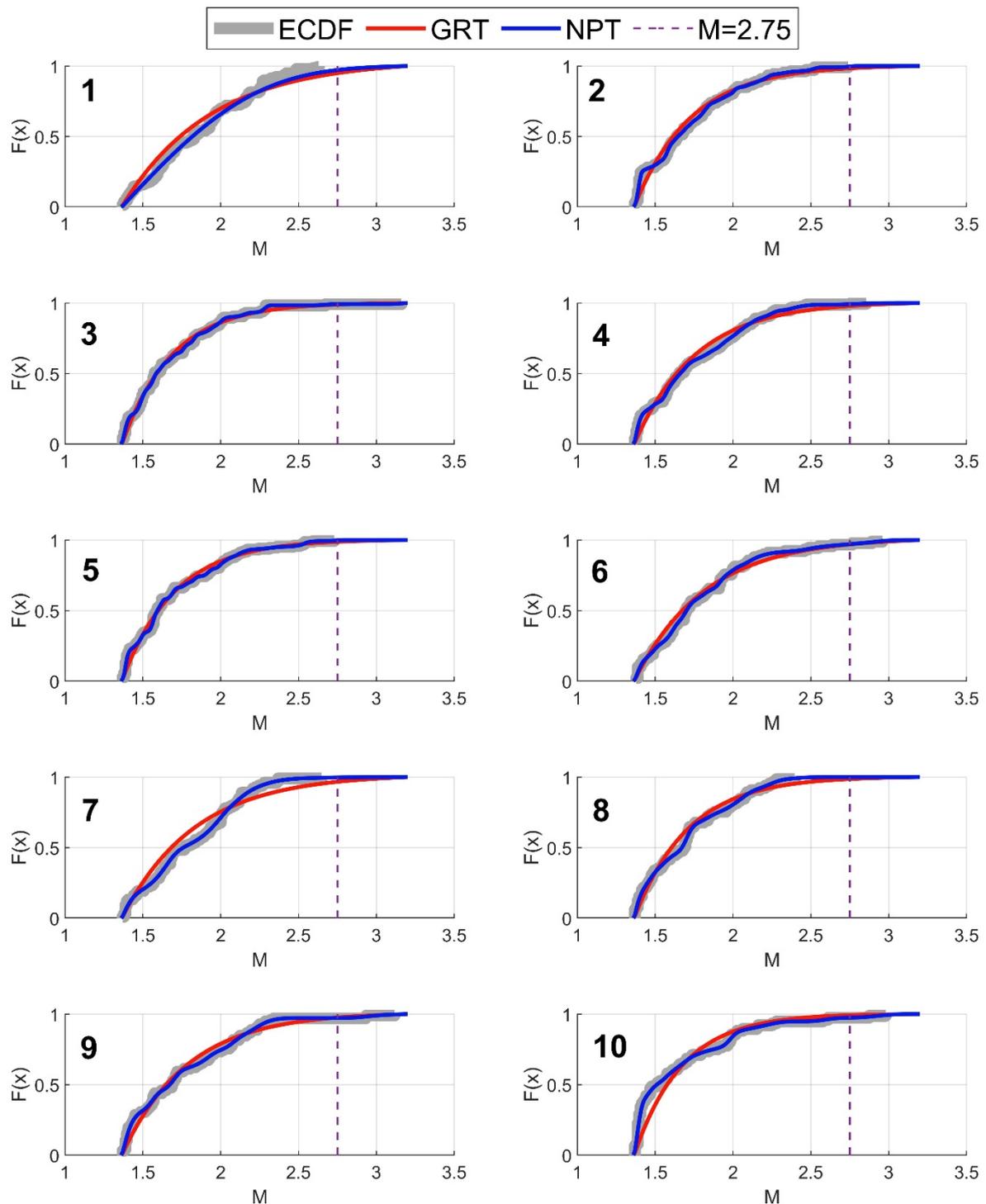


Figure 6. CDF plot, $F(x)$, of Empirical magnitude (grey curve), GRT model (red curve) and NPT model (blue curve). The vertical dashed line indicates the $M=2.75$. Numbers 1-10 correspond to the 10 studied time windows.

On the other hand, the NPT approach results to $R^2=0.61$ (p -value=0.06) for all time windows and $R^2=0.95$ (p -value=0.0005) for the 9 latter time windows, suggesting that the corresponding results are in much better accordance with the actual observations. Also note that the total number of events with $M>2.75$ predicted in total for all the 10 windows by the GRT are 25.5 (18.5 for the nine latter windows) and by the NPT are 14.4 (10.5 for the nine latter

windows). Given the actual 10 events observed, it seems that both models overestimate the number of events (equivalently the occurrence probabilities), however, the NPT approach is definitely more appropriate than the GRT model. This stands in agreement with the findings of Urban et al. (2016), who showed a violation of GR law in several cases studies of man-made seismicity. Finally, during time windows 1, 4, 7 and 10 there are very large differences between actual and expected number of events assuming the GRT, which is in agreement with Leptokaropoulos (2020) who showed that magnitude distribution significantly deviates from exponential during the aforementioned periods.

3.6 References

- Bender, B. (1983). Maximum likelihood estimation of the b-values for magnitude grouped data, *Bull. Seismol. Soc. Am.*, 73, 3, 831-851.
- Convertito, V., N. Maercklin, N. Sharma, and A. Zollo (2012). From induced seismicity to direct time-dependent seismic hazard, *Bull. Seismol. Soc. Am.*, 102, No. 6, 2563-2573, doi: 10.1785/0120120036.
- El-Isa, Z.H. and D. W. Eaton (2014). Spatiotemporal variations in the b-value of earthquake magnitude-frequency distributions: Classification and causes, *Tectonophysics*, 615–616, 1–11.
- Ellsworth, W. L., D. Giardini, J. Townend, S. Ge, and T. Shimamoto (2019). Triggering of the Pohang, Korea, earthquake (M_w 5.5) by enhanced geothermal system stimulation, *Seismol. Res. Lett.*, doi: 10.1785/0220190102.
- Galis, M., J. P. Ampuero, P. M. Mai and F. Cappa (2017). Induced seismicity provides insight into why earthquake ruptures stop, *Sci. Adv.*, 3, eaap7528.
- García-Aristizabal, A. (2018). Modelling fluid-induced seismicity rates associated with fluid injections: examples related to fracture stimulations in geothermal areas, *Geophys. J. Int.*, 215, 471-493.
- Gischig, V. S., and S. Wiemer (2013). A stochastic model for induced seismicity based on non-linear pressure diffusion and irreversible permeability enhancement, *Geophys. J. Int.*, 194, no. 2, 1229-1249, doi: 10.1093/gji/ggt164.
- Izadi, G. and D. D. Elsworth (2015). The influence of thermal-hydraulic-mechanical-and chemical effects on the evolution of permeability, seismicity and heat production in geothermal reservoirs, *Geothermics*, 53, 385–395.
- Kijko, A., and M. A. Sellevoll (1989). Estimation of earthquake hazard parameters from incomplete data files. Part I. Utilization of extreme and complete catalogs with different threshold magnitudes, *Bull. Seismol. Soc. Am.*, 79, no. 3, 645–654.
- Kijko, A., S. Lasocki, and G. Graham (2001). Nonparametric seismic hazard analysis in mines. *Pure Appl. Geophys.*, 158, No. 9-10, 1655–1676, doi: 10.1007/PL00001238.
- Kwiattek, G., P. Martínez-Garzón, G. Dresen, M. Bohnhoff, H. Sone, and C. Hartline (2015). Effects of long-term fluid injection on induced seismicity parameters and maximum magnitude in northwestern part of the Geysers geothermal field, *J. Geophys. Res.*, 120, doi:10.1002/2015JB012362.
- Lasocki, S. (1993a). Statistical prediction of strong mining tremors. *Acta Geophys. Pol.*, 41, no.3, 197-234.
- Lasocki, S. (1993b). Statistical short-term prediction in mining induced seismicity. *Rockbursts and Seismicity in Mines 93*, Proc. 3rd Int. Symp. on Rockbursts and Seismicity in Mines Kingston, Canada, 16-18 Aug. 1993 (Young, R.P., ed.) Balkema, Rotterdam, pp. 211-216.
- Lasocki, S. (2001). Quantitative evidences of complexity of magnitude distribution in mining-induced seismicity: Implications for hazard evaluation. In: *Rockbursts and Seismicity in Mine: Dynamic rock mass response to mining* (van Aswegen, G., Durrheim, R.J., Ortlepp, W.D. eds.) SAIMM S27, Johannesburg, 543-550.
- Lasocki, S. (2017). Probabilistic Assessment of Mining-Induced Time-Dependent Seismic Hazards, Chapter 11.3 in *Rockburst Mechanisms, Monitoring, Warning, and Mitigation* (Xia-Ting Feng, ed.), Butterworth-Heinemann (Elsevier), Oxford, United Kingdom, pp. 366-380.
- Lasocki, S., A. Kijko, and G. Graham (2000). Model-free seismic hazard estimation, H. Gokcekus (Editor), Proc. Int. Conf. Earthquake Hazard and Risk in the Mediterranean Region, EHRMR'99, Educational Foundation of Near East University, Lefkosa, Northern Cyprus, 503–508.
- Lasocki, S., and B. Orlecka-Sikora (2008). Seismic hazard assessment under complex source size distribution of mining-induced seismicity, *Tectonophysics*, 456, No. 1-2, 28–37, doi: 10.1016/j.tecto.2006.08.013.

- Lasocki, S., and B. Orlecka-Sikora (2020). High injection rates counteract formation of far-reaching fluid migration pathways at The Geysers geothermal field, *Geophys. Res. Lett.*, 47, e2019GL086212. <https://doi.org/10.1029/2019GL086212>.
- Lasocki, S., and P. Urban (2011). Bias, variance and computational properties of Kijko's estimators of the upper limit of magnitude distribution, M_{max} , *Acta Geophys.*, 59, 659-673, doi: 10.2478/s11600-010-0049-y.
- Lasocki, S., P. Urban, M. Staszek, B. Orlecka-Sikora, and A. Lisowska (2017). Report on modelling the probability distribution use in seismic hazard analysis. EU SHEER (SHale gas Exploration and Exploitation induced Risk) project, Deliverable D4.2, retrieved from <http://www.sheerproject.eu/images/deliverables/SHEER-Deliverable-4.2.pdf>.
- Lee, K.-K., W. L. Ellsworth, D. Giardini, J. Townend, S. Ge, T. Shimamoto, I.-W. Yeo, T.-S. Kang, J. Rhie, D.-H. Sheen, C. Chang, J.-U. Woo and C. Langenbruch (2019). Managing injection-induced seismic risks, *Science*, 364, 730-732, doi: 10.1126/science.aax1878.
- Leptokarpoulos, K. (2020). Magnitude distribution complexity and variation at The Geysers geothermal field, *Geophys. J. Int.*, Submitted Manuscript.
- Leptokarpoulos, K., and S. Lasocki (2020). SHAPE: A MATLAB software package for time-dependent seismic hazard analysis, *Seismol. Res. Lett.*, accepted for publication.
- Majer, E. L., R. Baria, M. Stark, S. Oates, J. Bommer, B. Smith, and H. Asanuma (2007). Induced seismicity associated with Enhanced Geothermal Systems, *Geothermics*, 36, 185-222.
- Martínez-Garzón, P., G. Kwiatek, H. Sone, M. Bohnhoff, G. Dresen, and C. Hartline (2014). Spatiotemporal changes, faulting regimes, and source parameters of induced seismicity: A case study from The Geysers geothermal field, *J. Geophys. Res.*, 119, 8378–8396.
- McGarr, A. (2014). Maximum magnitude earthquakes induced by fluid injection, *J. Geophys. Res.*, 119, 1008-1019, doi:10.002/2013JB010597.
- Orlecka-Sikora, B. and S. Lasocki (2005). Nonparametric characterization of mining induced seismic sources. Proc. Sixth Int. Symp. on Rockburst and Seismicity in Mines 9-11 March 2005, Australia (Potvin, Y., Hudyma, M., eds.), Australian Centre for Geomechanics, Nedlands, pp. 555-560.
- Orlecka-Sikora, B. and S. Cielesta (2020) Evidence for subcritical rupture of injection-induced earthquakes, *Sci. Rep* 10., 4016. doi: 10.1038/s41598-020-60928-0.
- Orlecka-Sikora, B., S. Lasocki, J. Kocot, T. Szepieniec, J.-R. Grasso, A. Garcia-Aristizabal, M. Schaming, P. Urban, G. Jones, I. Stimpson, S. Dineva, P. Sałek, K. Leptokarpoulos, G. Lizurek, D. Olszewska, J. Schmittbuhl, G. Kwiatek, A. Blanke, G. Saccarotti, K. Chodzińska, Ł. Rudziński, I. Dobrzycka, G. Mutke, A. Barański, A. Pierzyna, E. Kozlovskaya, J. Nevalainen, J. Kinscher, J. Sileny, M. Sterzel, S. Cielesta, and T. Fischer (2020). An open data infrastructure for the study of anthropogenic hazards linked to georesource exploitation. *Sci. Data*, 7, 89. doi:10.1038/s41597-020-0429-3.
- Page, R. (1968). Aftershocks and microaftershocks., *Bull. Seismol. Soc. Am.*, 58, 1131–1168.
- Robson, D.S., and J. H. Whitlock (1964). Estimation of a truncation point, *Biometrika*, 51, no 1–2, 33–39. doi:10.1093/biomet/51.1-2.33.
- Shapiro, S. A., C. Dinske, and C. Langenbruch (2010). Seismogenic index and magnitude probability of earthquakes induced during reservoir fluid stimulation, *The Leading Edge* 29, no. 3, 304–309, doi: 10.1190/1.3353727.
- Silverman, B. W. (1986). *Density Estimation for Statistic and Data Analysis*, Chapman and Hall, London, 175 pp.
- Staszek, M., B. Orlecka-Sikora, K. Leptokarpoulos, P. Martínez-Garzón, and G. Kwiatek (2017). Temporal static stress drop variations in relation to technological activity at The Geysers geothermal field, California, *Geophys. Res. Lett.*, 44, 7168-7176, doi: 10.1002/2017GL073929.
- Urban, P., S. Lasocki, P. Blascheck, A. F. do Nascimento, N. V. Giang, and G. Kwiatek (2016). Violations of Gutenberg–Richter in anthropogenic seismicity, *Pure Appl. Geophys.*, 173, No. 5, 1517–1537, doi: 10.1007/s00024-015-1188-5.
- Wyss, M. (1972). Towards a physical understanding of the earthquake frequency distribution, *Geophys. J. Roy. Astron. Soc.*, doi.org/10.1111/j.1365-246X.1973.tb06506.x.

4 Chapter 4: Earthquake sequences and background seismicity

Author: Antonio Petruccelli, ETH Zürich

4.1 Summary

In this deliverable, the problem of characterizing earthquake sequences and tectonic background activity rates is explored using a trending innovative approach, which involves temporal variations in earthquake-size distribution. In fact, variations in earthquake-size distributions are often referred to changes in crustal stress conditions. The developed algorithm automatically subdivides the temporal domain into time-subsets, by estimating for each of them the magnitude decay scaling parameter (*b*-value) which quantifies the relative amount of more/less hazardous events, i.e. high/less stressed zones. As an example of application, a retrospective test of such methodology is performed for a seismic sequence recently occurred in Switzerland. Thereby, the importance of monitoring the variability of earthquake occurrences for subsequent time windows using *b*-value could improve the consistency and the reliability of next hazard models, as well as to track in real time the evolution of seismic sequences and stress conditions.

4.2 Introduction

4.2.1 Statistical and deterministic approaches for earthquake forecasting

Analyzing, characterizing and understanding earthquakes sequences, in addition to background-seismicity rates, constitute some of the biggest challenges that seismology and earthquake engineering have to face with. Earthquakes tend to cluster both in space, along preferential directions which remark local tectonics, and in time.

Threatening earthquakes with detectable moderate/large magnitudes are followed by cascade of events, named aftershocks, which typically last from days to months or even years. On the other side, seismic activity could also increase with time prior to a relative strong shock, i.e. foreshocks, with rates that sometimes are monotonic.

However, during an ongoing earthquake sequence, it is very hard to separate the boundaries between the approaching end of the foreshocks sequence and the upcoming mainshock, which denotes the starting of aftershocks sequence or even of new foreshocks. This decision is even more crucial when it has to be given prospectively: in fact, a systematic decay in the aftershock rate over time is one of the milestones in statistical seismology (Omori Law).

Based on such empirical observation, statistical methodologies, such as the Epistemic Type Aftershock Sequence (ETAS, Ogata, 1998) or similar (Marzocchi et al., 2017, and references therein), are widely used across the community, to describe the branching process of

aftershocks generation. According to ETAS, then, there is no difference between foreshocks, mainshock and aftershocks. The rupture process is epidemic, where any shocks is able to generate its own descendance. Despite the successfulness and reliability of such tools, the scientific answer they can then provide is purely stochastic.

On the contrary, from a deterministic, physical-based point of view, the problem has a quite clear understanding, but it is poorly constrained, and most of all it is rather complex to solve. Tectonic loading drives and forces earthquake to interact with each other during the sequences, continuously changing stress conditions within the surrounding medium, and possibly influencing nearby faulting systems. Static and dynamic stress changes transfer due to sudden displacements decay with distance, both encouraging or inhibiting rupture processes.

According to this scheme, then, a foreshocks sequence represents a precursory process, i.e. a response to precursory slip, while an aftershock sequence would represent both decrease or increase (and thereby new foreshocks) in stress conditions or interactions.

Most of physical conditions have high degree of uncertainties, reasons why most likely physics-based approach hadn't had any better success with respect to the abovementioned statistical approaches. Apart from controlled and reproducible laboratory experiments, stress inferences are only possible using indirect measures and thereby assuming raw precisions.

4.2.2 Temporal variations in earthquake-size distribution

However, stress changes do not impact only activity rates but also the frequency-size (or magnitude) distribution (FMD) of the subsequent earthquakes, which means how is the proportion between big and small events. The frequency rate is described by an empirical log-linear law (Gutenberg and Richter, 1956, from now on GR), which describes the number of events above a certain threshold magnitude m $\log N(m)=a-bm$.

The a -value denotes the seismic productivity while the b -value acts as a factor scale (~ 1) of the law by quantifying the relative proportion of big, more hazardous events with respect to small, less hazardous events. Statistical models (such as ETAS) of seismicity rates assume for the sake of simplicity temporal invariance of the b -value, although variations, sometimes marked as random fluctuations, were firstly documented back to 40 years ago.

However, repeated laboratory experiments (Amitrano, 2003) have provided and confirmed that b -value is linearly inversely related to stress conditions (Scholz, 1968), meaning that high-stressed samples (or zones) should suffer a decrease of b -value (increase of high-magnitude event proportion) while low-stressed samples (or zones) should exhibit an increase of b -value (increase of low-magnitude event proportion).

Despite initial skepticism about the nature of such variations, the continuous increase of publications and scientific evidence in favor of this view are pushing the seismological community towards new insights of this fundamental parameter and its predictive power. In particular, given also the abundance of data available right after the mainshock occurring (with thousands of aftershocks or possible foreshocks), a real-time monitoring system of the transient change of b -value during seismic sequences can be implemented (Gulia and Wiemer, 2019).

According to such strategy, the monitored, transient b -value signal would be nothing more than a translation of the earthquake cycle: a high b -value (above the reference, usually around 1) would then be an index of decreasing/relaxation of stress conditions, while a low b -value might indicate an increase in stress condition. In such view, foreshocks of an upcoming big event are expected to exhibit progressively low b -values, while aftershocks decaying should show an increase of the b -value.

4.3 Methodology and application

4.3.1 MATLAB® code implementation

The code of the algorithm was entirely developed in MATLAB®, partially taking advantage of some routines already available in the Zmap package for statistical seismology (Wiemer, 2001).

The code is subdivided in 2 main parts:

1. data selection;
2. construction and analysis of the b -value time series.

PART 1: Data Selection

The algorithm takes as input a tabulated catalog, subdivided into columns which are longitude (degrees), latitude (degrees), origin time (year, month, day, hour, minute and seconds), magnitude (preferably of the same type, M_l , M_w , etc), hypocentral depth (km) (Figure 1).

The first filter on data regards the amount of data that are going to be used to construct the time series (see Part 2). In order to do so, we refer to this data, and its relative parameters, as the “overall”. This requires a time domain, i.e. starting and ending time period (in decimal year), a geographic space domain, i.e. a box in which filter earthquake locations according to longitude, latitude and hypocentres, and an overall magnitude of completeness.

The magnitude of completeness M_c is the minimum magnitude above which the log-linear decay in number of events with magnitude is observed, i.e. it is the minimum magnitude in the GR model. This threshold divides the incomplete part of the catalog (below M_c) from the complete part of the catalog (above M_c), for which, by definition, the b -value can be estimated. The algorithm uses the so-called Maximum Curvature Method (Wiemer and Wyss, 2000) which estimates M_c as the magnitude bin with the highest occurrence of earthquakes, plus an additive positive correction factor to avoid underestimation effects.

The b -value, instead, once that M_c has been established, can be computed using a maximum likelihood approach. That means that the estimated b -value is the one that maximize the loglikelihood of each observable (complete magnitudes above M_c), given the GR probability density function. The maximization can be performed both numerically and analytically.

The algorithm prefers the latter approach by using the formula of Aki (1965), with a correction due to magnitudes being discrete values (Bender, 1983).

The b -value is the linear slope of the decay in number of events and it ranges from 0.6-0.7 to 1.5 to 2. It critically depends on the choice of the assumed completeness and its value might

vary significantly for different values of completeness: a recommendable approach is making sure that the b -value would approximately lie constant (meaning that linearity in the decay in number of the events is preserved) within a certain range of possible completeness (linearity check).

The b -value coming from the overall dataset is referred as the reference b -value for the time series. As additional test, the algorithm also performs a “quality measure” on the linear fit, by computing a goodness-of-fit test. The goodness of fit test is computed as the complementary percentage residual of data with respect to the GR linear model.

A goodness-of-fit typically higher than 80-85% indicates a good measure of linearity of data. Once that the starting overall dataset has been filtered, it is ready to be subdivided.

PART 2: construction and analysis of the b -value time series

To construct the time series (Figure 2), reference values for the b -values and for the M_c are established and used as central values with respect to evaluate the variations. A reasonable choice, as explained above, could be to pick the values estimated for the overall starting dataset, but other approaches (mean or median values) are also accepted and reasonable.

For retrospective analyses (see 4.3..2), the overall dataset is then divided in a pre- and post-mainshock subsets. The pre-mainshock dataset usually contains lower seismicity rates with respect to the subsequent post- dataset, so it is recommended to enlarge the initial window as much as possible to cover a sufficient number of events (usually of the order of hundred events). In fact, seismicity rates nose up right after the mainshock, so the post-mainshock period have higher density of events. It spans the time period until seismicity rates recovers the primitive values or until a new big shock is detected, so a further subdivision is then necessary.

The time series is effectively computed if more than a user-defined minimum number of events N_{\min} is available. This parameter acts as window length-meter as it governs the resolution of the time-series: meaning that high values in N_{\min} risk to smooth too much the signal, making impossible to detect transient variations, while a little value might lead to underestimations in the M_c and in the subsequent b -value, as they critically depend on the amount of data used to estimate them.

Then, starting from the N_{\min}^{th} earthquake, time-subsets are backward sampled by selecting the previous N_{\min} events, and the window is then moved forward by one event: in this way, each temporal estimation is relative to the previous N_{\min} events.

This first step is necessary to assess the temporal trend of M_c , always using the maximum curvature method and the proper correction. In order to guarantee statistical robustness, M_c estimation is bootstrapped hundred times, and the final value correspond to the mean value of the M_c distribution, opportunely rounded to the magnitude bin.

The b -value with its uncertainty is then estimated, from the average M_c , using the analytical maximum likelihood approach. As additional quality assurance step, it is always suggested to check the linearity test of the fit, as explained above, once that a proxy b -value is estimated.

If the test, for some reason, is not passed, correction can be gradually adjusted in order to obtain a goodness-of-fit estimation, and then accept the proxy b -value as definitive estimation.

Other similar approaches can be implemented, for example by requiring a minimum number of events (usually 50 or 25) for the b -estimation, as well.

4.3.2 Demonstration Report: Retrospective analysis for aftershocks forecasting

The 2019 Valais sequence

It is here summarized a retrospective test of this algorithm to the Sion earthquake sequence (M3.3), occurred in the Valais Canton (VS, Switzerland) back to the end of 2019.

Located close to the borders with France and Italy, Valais is the most seismically active region in Switzerland, where also the last destructive historic earthquake (M5.8, on January 25th, 1958) took place. The sequence occurred in one of the best-known areas of activity (Fig. 1), with two earthquakes of M3.3 on November 5th followed by three aftershocks of M3.2 and M3. Shocks were widely felt by the populations with hundreds of notifications to the Swiss Seismological Center. Right before the starting of the sequence, only about 200 earthquakes were recorded over the 2019. The rapid increase in number of events is all concentrated in the month of November. Parallely, the frequency-magnitude distribution for the 2019 denotes a relative low b -value with a magnitude of completeness of 0.6. In the FMD, linearity of b -value is also preserved for a wide range of different completeness cut-off (from about 0.5 to 2) with goodness-of-fits well above the 80% minimum required by the algorithm.

The Valais sequence was characterized by low seismicity rates before November 5th, with apparently no foreshocks before this date. After the M3.3, seismicity rates increased and remained relatively high during the following days, where aftershocks were recorded. Computing reliable time series of the b and M_c in aftershocks zones is difficult, mostly because the completeness is strongly affected by limitations in network detection. In fact, the first hours (or sometimes days) are usually removed from the analysis, as incompleteness tends inevitably to increase.

However, for the Valais sequence, by not considering the early hours after the M3.3, where the b -value is high (possibly as an effect of high M_c), the b -value suffers a significant drop right before the Nov 5th 07:18:42 M3.0 aftershock, and it remained low afterwards. This could have been considered a possible “red-alarm” for a new shock coming, which effectively occurred after a few hours (Nov 5th 19:51:13 M3.2). After that b -value rose up, well below the reference value. On the contrary, none b -value drop is recorded before the Nov 7th 18:35:38 M3.2.

4.3.3 Verification and benchmarking

These procedures follow up most of the methodological guidelines already proposed in literature, where other seismic sequences had been analysed (Gulia et al., 2017). Immediately after the mainshock, the b -value is observed to increase of about 10%-20%. This increase is statistically significant and was observed by stacking the temporal trends of several sequences (Gulia et al., 2017; Gulia and Wiemer, 2019). In fact, elastic rebound theory predicts that after a mainshock, a certain amount of time is needed to recover the strain released by the mainshock. This might turn into a possible decrease of the seismic risk, i.e. no big event coming

shortly. Even in the Valais sequence, right after the end of the last strong aftershocks, the b -values lie well above (~ 1) the reference value (~ 0.75) assessed by the algorithm, by confirming this assertion. Under certain conditions, then, the amplitude and the static/dynamic changes might inhibit some portions of the fracture, by locking them and by increasing b -value.

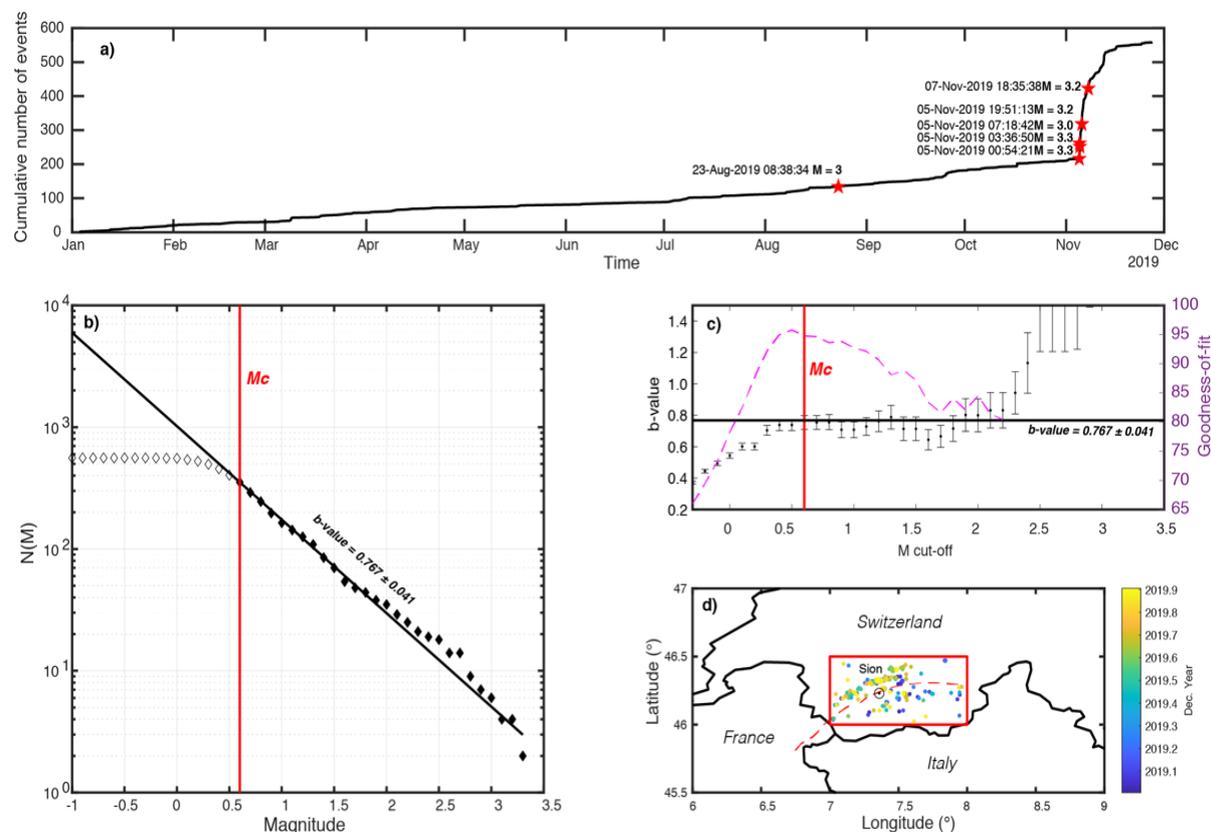


Figure 1: Earthquakes occurred in Switzerland during 2019 (Swiss Seismological Service catalog). **a)** Cumulative number of earthquakes (red star denotes events with magnitude above 3.0). **b)** Frequency-magnitude distribution of data in a) (complete data with filled mark). b -value (black) and magnitude of completeness (0.6) are also reported. **c)** Possible b -values as a function of different completeness threshold. In purple the correspondent goodness-of-fit (%), computed as $g_{fit} = 1 - \frac{\sum_i |N(M_i) - f(M_i)|}{\sum_j N(M_j)}$, $f(M_i) = 10^{a-b(M_i-M_c)}$. **d)** Geographical distribution of the events.

On the contrary, a transient b -value well below the reference might be the proof that the perturbation of stress conditions has been critical and causing then a rapid increase in the proportion of the more hazardous events (Gulia and Wiemer, 2019), like showed in Fig. 2. However, in order to provide a more comprehensive description, other factors such as the distance from the modeled fault or the tectonic regime should have taken into consideration.

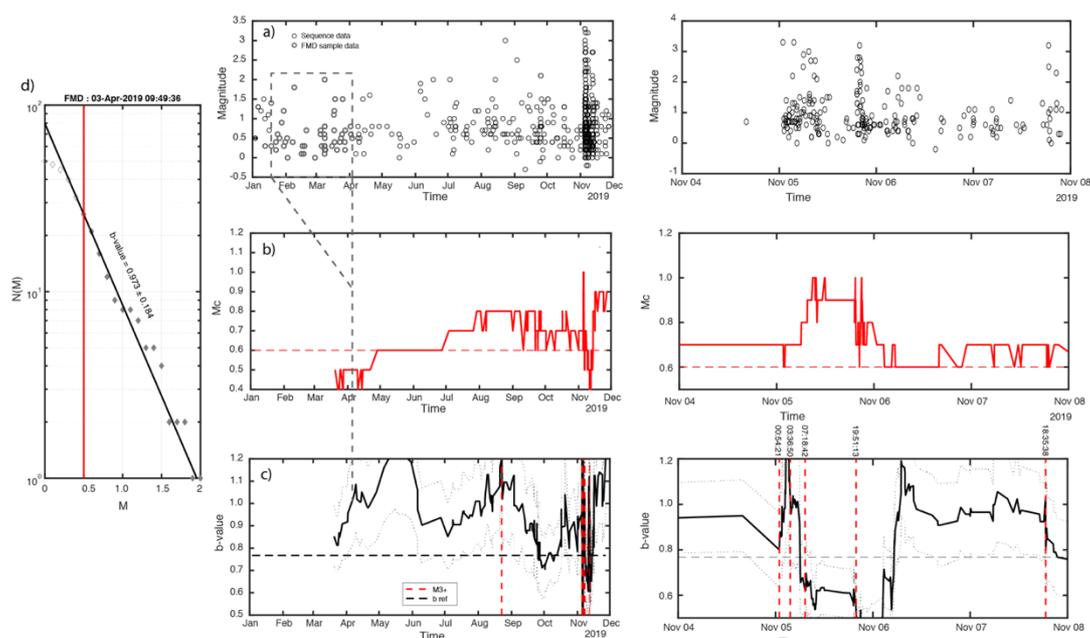


Figure 2: Temporal analysis of b -value and M_c for the Valais sequence (left panels refer to the entire 2019, right panels to the sequence early days). **a)** Magnitude-time plot (grey circles refer to a sample FMD, see d)). **b)** M_c -time plot (reference M_c -value of Fig. 1b horizontal dashed line) **c)** b -value – time plot (dotted lines denote the confidence intervals). **d)** Sample FMD of a).

However, benchmarking tests are always recommended to re-analyze previous sequences (i.e., L’Aquila 2009, Amatrice-Norcia 2017 earthquakes), in the light of the new methods proposed, like the one here exposed. In fact, improvements in seismic network detections, which usually guarantee data completeness increase, ensure catalogs to be updated and give the possibility to test both the consistency and the time/completeness-independencies of the results. Moreover, updating of national hazard maps or addition of new seismogenic source models might result into additional precious ingredients to be coupled with b -value variability hypothesis, in order to test and possibly enrich the detail level of the next generation of seismic risk models.

Although these methodologies are still under development, the ideas and the application here exposed suggest that the evolution of b -values, considered a proxy of the crustal stress conditions, can act as a 1st order discriminator between normal aftershocks and precursory events. Moreover, the results here discussed have been inspired and match recent advances made in the topic. However, despite the lack of physics-based understandings and predictive modeling, this hypothesis presents a new angle of view in which aftershocks sequence can be modelled and understood.

4.4 Future developments

Earthquake forecasting sciences have as ultimate targets the full, or partially full, knowledge of the whole seismic process (enucleation and time-space development). It is then crucial and essential to have new models that are able to forecast upcoming events, in order to mitigate as much as possible the related risk. Such models, in addition to be scientifically reliable, have also to be run simultaneously to a certain seismic sequence going on, i.e. in a prospective sense. In fact, the capability to provide to public institutions and to media valid forecast statements is the last and very important component of any decision-maker program for earthquake risk.

In such sense, the algorithm here presented seems to be one potential candidate to be part of such programs. Temporal b -value analyses are currently going towards a real-time risk mitigation.

Tools like a traffic light system for the b -value (Gulia and Wiemer, 2019), where risk is quantified by the relative amount (green) or decrease (red) of the transient b with respect to the reference (yellow), was recently proposed and have successfully forecasted retrospectively two important aftershocks sequences (Amatrice-Norcia 2017 and Tohoku 2011).

A yellow status indicates no significant change, or difficulty in determining the variation, i.e. no information gain. On the contrary, the capability of the $b(\text{time})$ algorithm to declare a green/red status might constitute a big step forward of the seismological community towards earthquake forecasting and really pone a milestone for short-term hazard assessment.

In fact, in the short term, reliable forecasts are only available whereas aftershocks take place and where network coverage is sufficient. For the upcoming future, then, the methodology has to be refined on the basis of additional data, also heterogeneous, which after homogenization processes need to be put together in a broader context, becoming then a standard for the community. Moreover, the methods have to be enriched and coupled with cost-benefits analyses and uncertainty quantification theory. Additionally, spatial variations in b -value can be considered for a more comprehensive picture. The imprinting of such modifications on b -forecasting might be then significant and worthwhile for investing time and resources. However, the predictive picture of b -value is not fully exhaustive as, for example, the forecast of big events ($\sim M6$ and above) is really tough, as such events are quite rare in nature.

Nevertheless, a common effort of all the community in thinking, implementing and testing forecasting models is needed and would possibly lead seismology to reduce the gap from other forecasting sciences.

4.5 References

- Aki, K. (1965), Maximum likelihood estimate of b in the formula $\log N=a-bM$ and its confidence limits, *Bull. Earthq. Res. Inst., Univ. Tokyo*, 43, 237–239. 7
- Amitrano, D. (2003). Brittle-ductile transition and associated seismicity: Experimental and numerical studies and relationship with the b value. *Journal of Geophysical Research*, 108(B1), 2044.
- Bender, B.; Maximum likelihood estimation of b values for magnitude grouped data. *Bulletin of the Seismological Society of America* ; 73 (3): 831–851. doi: <https://doi.org/>
- Gulia, L., Rinaldi, A. P., Tormann, T., Vannucci, G., Enescu, B., & Wiemer, S. (2018). The effect of a mainshock on the size distribution of the aftershocks. *Geophysical Research Letters*, 45, 13,277–13,287. <https://doi.org/10.1029/2018GL080619>
- Gulia, L., Wiemer, S. Real-time discrimination of earthquake foreshocks and aftershocks. *Nature* 574, 193–199 (2019). <https://doi.org/10.1038/s41586-019-1606-4>
- Gutenberg, B., Richter, C. F. , Earthquake magnitude, intensity, energy and acceleration. *Bull. Seismol. Soc. Am.* 46, 105–145 (1956).
- Marzocchi, Taroni, Falcone, *Sci. Adv.* 2017;3: e1701239
- Ogata, Y., Space-time point-process models for earthquake occurrences. *Ann. Inst. Stat. Math.* 50, 379–402 (1998).
- Scholz, C. H. (1968). The frequency-magnitude relation of microfracturing in rock and its relation to earthquakes. *Bulletin of the Seismological Society of America*, 58, 399– 415.
- Wiemer, S., 2001. A software package to analyze seismicity: ZMAP. *Seismological Research Letters*, 72(3), pp.373-382
- Wiemer, S., and M. Wyss (2000), Minimum magnitude of complete reporting in earthquake catalogs: examples from alaska, the western united states, and japan, *Bull. Seismol. Soc. Am.*, 90, 859–869.

5 Chapter 4: ESHM20 Python Toolkit for development of regional seismic hazard models

Authors: Celso Reyes, Laurentiu Danciu ETH Zürich

5.1 Summary

In this deliverable, the problem of characterizing earthquake sequences and tectonic is addressed through the creation and application of the ESHM20 Python Toolkit, which is a collection of Python routines that can synthesize a declustered catalog with geospatial information and associated magnitude of completeness through time, to derive Gutenberg-Richterbased models to describe the seismic occurrence rates.

5.2 Introducing the ESHM20 Python Toolkit

The ESHM20 Toolkit allows the user to develop seismic hazard models from a catalog, and series of defined geospatial areas. The tools created for this work package take the provided historical and instrumental catalogs, along with details from a spatial database at three differing regional zoom levels. From largest to smallest, these are the completeness super-zones (CSZ), the TECTO-zonation (TECTO) and the area sources (ASZ) – for information on the definition of these sources we refer to the SERA D25.3 – ESHM20 - Seismogenic sources update M36. This toolbox is currently under development and is used in data analysis, model development and hazard calculation input file generation for the 2020 European Seismic Hazard Model – SERA JRA3.

The main features of the toolbox are tailored to provide an end to end processing flow from input datasets to input files for hazard calculation. The processing flow assumes prerequisite of input datasets: earthquake catalogue, active faults, subduction sources, area sources; all inputs must be curated and error free.

The main modules of the toolkit consist of:

- Statistical analysis of the earthquake catalogue: declustering and completeness
- Statistical fitting of seismogenic source models; this module can handle the area sources and the active faults characterization.
- Visualization of various components of the model (plots of magnitude frequency distribution for every source, comparison plots between different model components)

This toolbox is implemented in Python 3.6, but portions of the toolbox has been informed by previously existing MATLAB, R, and FORTRAN codes as well as publications (eg. Stromeyer-Gruenthal, 2015) The toolbox uses some libraries of the GEM's OpenQuake Modeler Toolkit (<https://github.com/GEMScienceTools/oq-mbtk>).

5.3 ESHM20 Toolkit: General Methodology for Code Development

The rapid pace of prototyping throughout this phase has made the creation of tests to cover all aspects of this toolbox infeasible, although this has been mitigated through a few design and/or workflow choices:

- *Object Oriented (OO) Approach* : Heavily used, mission-critical, and error-prone sections of code are encapsulated into classes that provide intuitive interfaces to the rest of the program. This helps to ensure the stability of each section of code while providing well-defined units for testing. An additional benefit that cannot be overstated is the increased readability of the resulting code, which makes errors easier to spot, and makes modifying the code during the frequently required changes easier to provide.
- *Foregoing Notebooks during development*: Jupyter (ipython) notebooks have been eschewed in favor of working with python files in a full featured IDE (PyCharm). While notebooks are excellent for prototyping simple systems, running existing workflows, and sharing results, they provide inadequate tools for code development. They provide no easy way to analyze errors, debug, perform stack traces, nor perform static analysis. Their persistent states provide opportunities for insidious bugs that are state dependent. Notebooks were used to prototype the workflow, and iterate on some logic, but the code was then entered into python modules for debugging and reliability purposes. Additionally, notebooks provide easy access to the final deliverable.
- *Static Analysis* : Extensive use of static analysis and rejecting Python “duck typing” in favor of stricter type hints, helps to ensure that the code pieces fit well together, and allows the IDE to easily point out inconsistencies without having to run the code.
- *Git Repository* : The value of an online versioned repository system for tracking changes, discovering introduced errors, and sharing code is incalculable.
<https://gitlab.seismo.ethz.ch/efehr/eshm20>

All codes were tested against original source codes or publication; i.e. the MATLAB implementation of the Stomeyer and Grunthal 2015. Otherwise, the test coverage was developed on the fly.

5.4 ESHM20 Toolkit: OpenQuake changes in support of the toolkit

This toolbox has been created with the intention of merging into OpenQuake as part of the hazard modelers toolkit (<https://github.com/GEMScienceTools/oq-mbtk>). It leans heavily on existing OQ-MBTK libraries, but additional methods have been added as described below.

5.4.1 Additions to `openquake.hmtk.seismicity.occurrence`

The class, *OccurrenceObservationDetails*, provides a single-point of contact for the earthquake observations, including the magnitude bin centers, number of observations within each bin and timespan for each bin’s observations. These values all derive from the regionally-filtered and declustered earthquake catalog and associated completeness table. This class can provide observed rates and cumulative observed rates without requiring the formulas to be repeated throughout the codebase. This class also allows all the associated data to be passed around within the engine without having to track all the individual information pieces. This is used extensively within the ESHM20 python toolkit, as well as in the Stromeyer-Gruenthal based occurrence calculations.

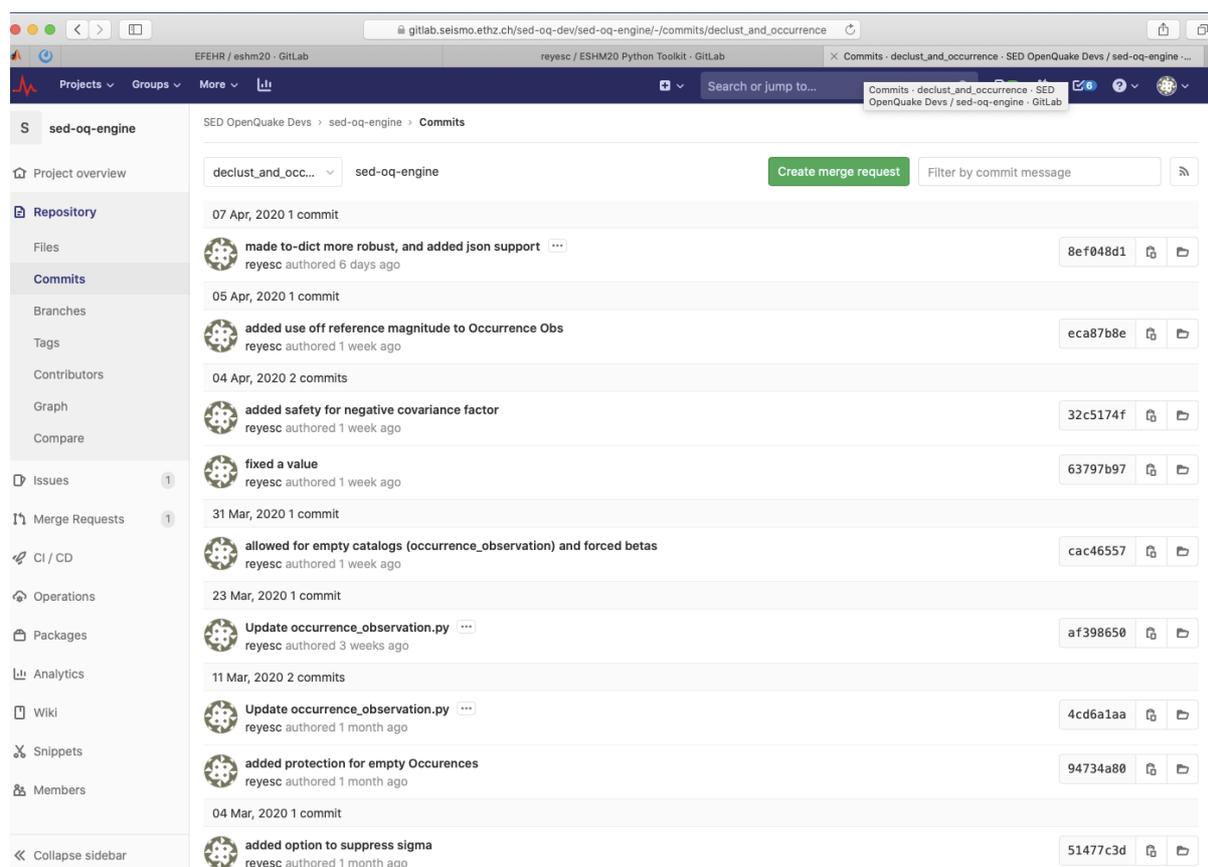


Figure 3 Screenshot of the ESHM20 Python Toolbox GitLab page, showing recent commits.

5.4.2 openquake.hmtk.seismicity.occurrence.stromeyer_gruenthal

In Stromeyer and Gruenthal's 2015 paper, they describe methods for calculating a_{GR} and b_{GR} parameters based on the observed earthquake occurrence. Not only do they implement Weichert's method, but they also provide details for implementing common beta and prior beta versions. Each method is implemented with two approximations, the small approximation and the finite approximation. The small approximation should be used when the bin width is less than 0.5. Above this threshold, the finite approximations should be used. All six of these methods have been implemented according to the paper, with their results verified against the GFZ implementations.

The *WeichertSmall* and *WeichertFinite* classes accept *OccurrenceObservationDetails* for one zone, and calculate the a_{GR} , b_{GR} , and covariance from the dataset. Meanwhile, the *CommonBetaSmall* and *CommonBetaFinite* classes accept multiple *OccurrenceObservationDetails* to determine a beta that matches the group. The *PriorBetaSmall* and *PriorBetaFinite* functions also accept multiple zones' worth of *OccurrenceObservationDetails*, but are evaluated in the context of a known beta. All these routines are vectorized, resulting in excellent performance, as well as the ability to map functions directly to the paper's equations without additional cognitive overhead.

5.4.3 Additions to `openquake.hmtk.seismicity.declusterer`

Routines were added for both Reasenber (1985) and Zaliapin (Zaliapin et al, 2008) methods of declustering. These supplemental declustering methods were examined during the evaluation phase, but not in the final beta model product, which relies upon the Gruenthal declustering window (already contained within OpenQuake).

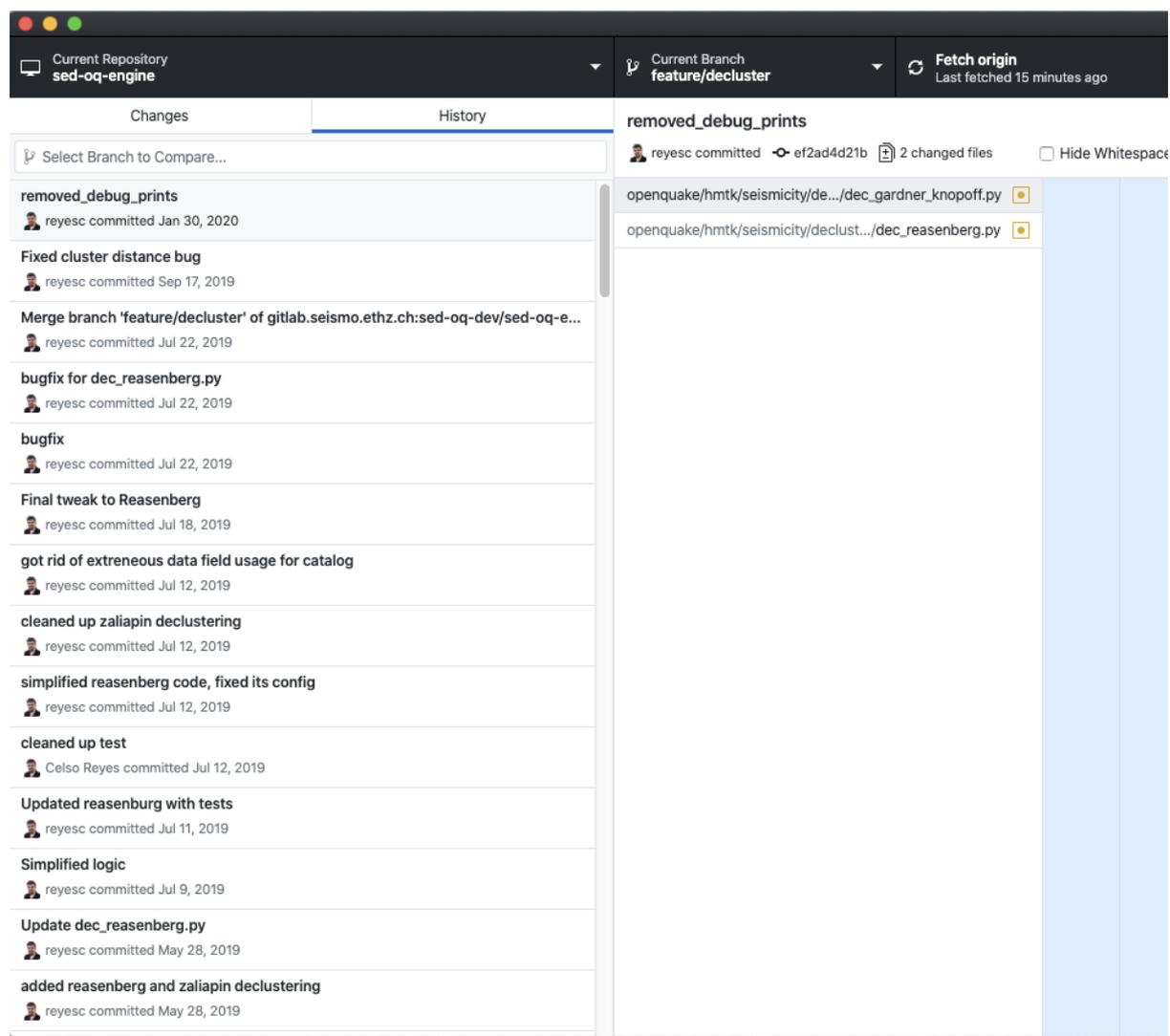


Figure 4 Screenshot of GIT showing recent changes to the occurrence and declustering portions of the oq-engine.

5.4.4 Additions to `openquake.hmtk.seismicity.completeness`

At the time of this deliverable the completeness technique described in Chapter 3 is ongoing to be translated into this toolbox. The original method was developed in MATLAB with subroutines from ZMAP 7.

ESHM20 Toolkit: Seismic Hazard Model Development Processing Flow:

5.4.5 Part 1: From Raw catalogs to Declustered Catalogs divided by depth

The provided catalogs EMEC_FINAL_20190218.csv and SERAcatalogue_1000-1899_v1.0.csv. Catalogs are cleaned, merged, and then split into shallow (<60km) and deep catalogs. Catalogs are associated with their enclosing tectonic region, and then resampled to fill in unknown depths. The catalogs are then declustered by selecting various declustering techniques: i.e. Windowed declustering, such as Gardiner-Knopoff, Gruenthal, or Uhrhammer, as well as Reasenberg (1985) and Zailapin (2008)

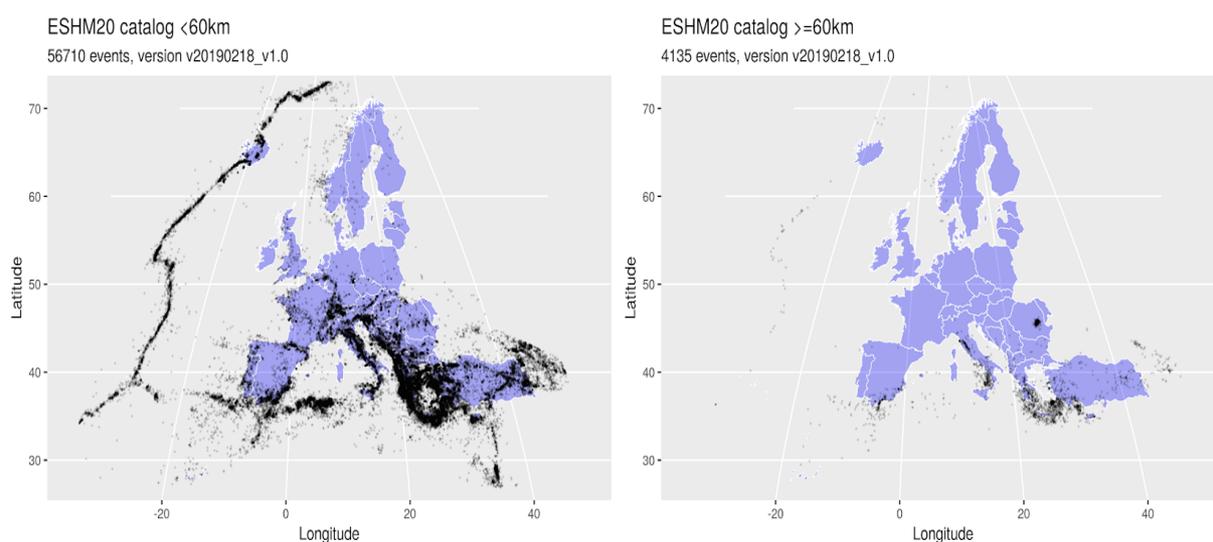


Figure 5 Map view of unified earthquake catalogues ESHM20 shallow (left) and deep(right).

5.4.6 Part 2: Magnitude of Completeness

To be converted to Python, and treated elsewhere in this document. Declustered catalogs are used to calculate the M_c through time separately for each declustering method and for the shallow and deep catalogs. The result is a completeness table expressing the M_c through time for each completeness zone (see all CSZ in Figure 4).



Figure 6: Completeness super zones for estimating the magnitude of completeness of the earthquake catalogues, and testing the forecast of the declustering analysis.

```
In[13]: print(cc.data.get_group('SZ02'))
```

	CSZ_NAME	CSZ	Mc	Time	Id
6	CSZ_CH	SZ02-CSZ_CH	3.7	2015	2
7	CSZ_CH	SZ02-CSZ_CH	3.7	1852	2
8	CSZ_CH	SZ02-CSZ_CH	4.3	1812	2
9	CSZ_CH	SZ02-CSZ_CH	4.5	1771	2
10	CSZ_CH	SZ02-CSZ_CH	4.9	1685	2
11	CSZ_CH	SZ02-CSZ_CH	5.1	1524	2
12	CSZ_CH	SZ02-CSZ_CH	5.9	1295	2

Figure 7 Table of completeness for Superzone SZ02. Of primary importance is the Mc and Time columns, which are used in filtering catalogs according to each zone.

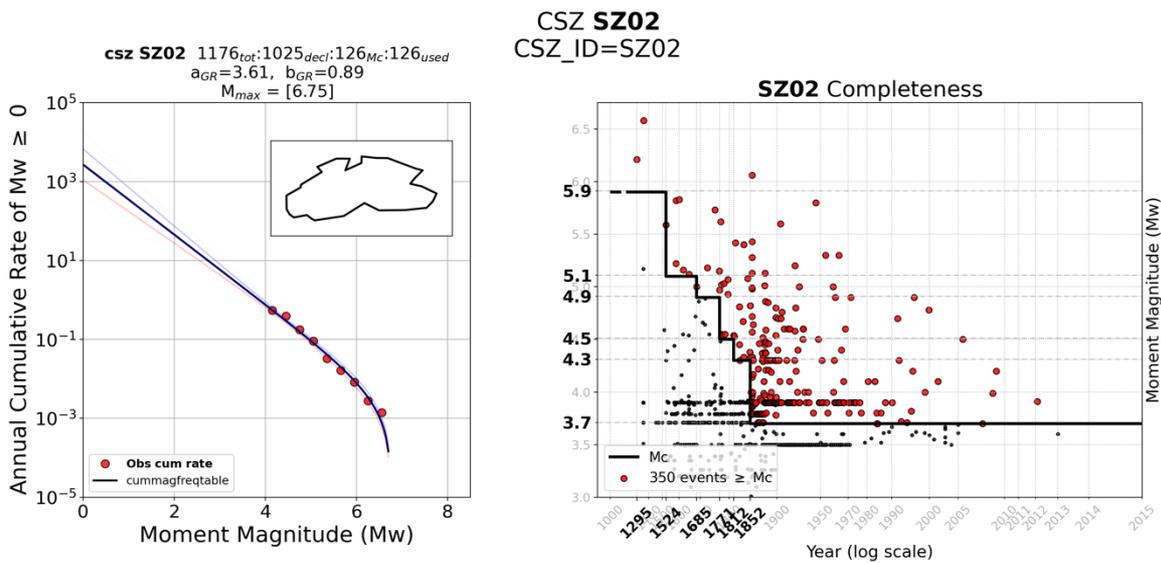


Figure 8 Tecto zones that lie within SZ02 (note that this results might be updated in the D25.3 section 1.3 Completeness)

5.4.7 Part 3: Calculating the Gutenberg-Richter Parameters

The modeled activity obeys two assumptions: that seismicity follows a Poisson process, and secondly, that seismicity follows a Gutenberg- Richter power law model according to equation: $\log N = a_{GR} - b_{GR} * M$, where N is the cumulative number of earthquakes per year equal to or greater than magnitude M, and a_{GR} and b_{GR} are constants.

The activity rate (a_{GR} -value) represents the total seismic productivity of a given source ($=\log N(M)$, when $M = 0$), or the log of number of events ($M > M_0$); the b_{GR} -value is the negative slope of the recurrence curve expressing average ratio of exponentially distributed small and large magnitude earthquakes.

Earthquake recurrence parameters of each area source are derived from the declustered earthquake catalogue (see D25.3 Section 1.3) using the completeness intervals. This methodology may apply to any geospatial polygons, eg. ASZ, TECTO or CSZ-zonation.

The core calculation kernel is the maximum likelihood procedure of Weichert (1980) as interpreted through Stromeyer and Gruenthal (2015).

The activity parameters are estimated accounting for the uncertainties of the a-, b-GR parameters described by a covariance matrix (Stromeyer and Grunthal 2015).

ESHM20 Toolkit: Resource Managing Classes

ZoneManager : Upon starting the ESHM20 Engine, the files containing geographic and attributes for each zone is read by the ZoneManager. This acts as the intermediary between the details of each zones and the main program, and provides an interface that allows the zones to be queried in a unified manner, even though the attributes vary from file to file. This is the class that handles the relationships between parent and child zones. Maintaining the relationships between zones and the regions to which the zone belongs is critical for being able to compare seismicity at the various scales, and therefore judge the validity of modeled seismicity. It is also the relationship between zones at each level and the CSZ that dictates how the magnitude of completeness-through-time gets applied.

CompletenessManager: is a simple class that reads and maintains the master Completeness Table. It allows adjustments to be applied at one place, and propagate throughout the engine.

ColorManager: Used for visualizations, this class allows colors to map to dictionary keys or other values, allowing for a consistent color interface across multiple plots.

ESHM20 Toolkit: General Processing Flow for Source Characterization

During catalog preprocessing, each event is assigned to its CSZ zone, and an initial pass is run to determine and mark events that meet that zone's Magnitude of Completion criteria. This is important for handling zones that span multiple CSZ's, where a unified M_c cannot exist, owing to differing completeness windows and magnitudes, and other inconsistencies that arise from merging disparate seismic networks

After preprocessing, the determination of Gutenberg-Richter parameters follows a generally similar scheme. Except where noted in the sections "Processing CSZ", "Processing TECTO", and "Processing ASZ"

The declustered catalog of earthquakes within each zone is filtered with the appropriate CSZ-mapped Completeness table. The filtered catalog is then binned according to magnitude, and a duration is assigned to each bin. These are used to determine the Gutenberg-Richter parameters.

Of note, a TECTO or ASZ zone may span multiple CSZ zones. If so, then it is determined which zone dominates, by having the most locally M_c -Complete events within that zone. Then, that winning zone's

completeness is applied to the entire catalog within the zone and is used for further calculation. If a zone is assigned to a new CSZ, then the ZoneManager is updated appropriately.

The a_{GR} and b_{GR} values are calculated according to Weichert using the observed occurrences. If the b -value lies outside a predetermined range, in this case, 1.2 and 0.55, then the occurrence bin may be decreased by one (the lowest magnitude bin is removed), and the GR is recalculated. If there is an error during the calculation, the zone is assigned pre-determined values. The number of events used in the calculation, the number of events originally observed, and the number of dropped bins are all stored and can be accessed later.

ESHM20 Toolkit: Visualization

Visualization provides opportunities to scrutinize the results to validate their validity and determine if the model parameters fit reality. Some of the example figures are listed below, along with selected examples:

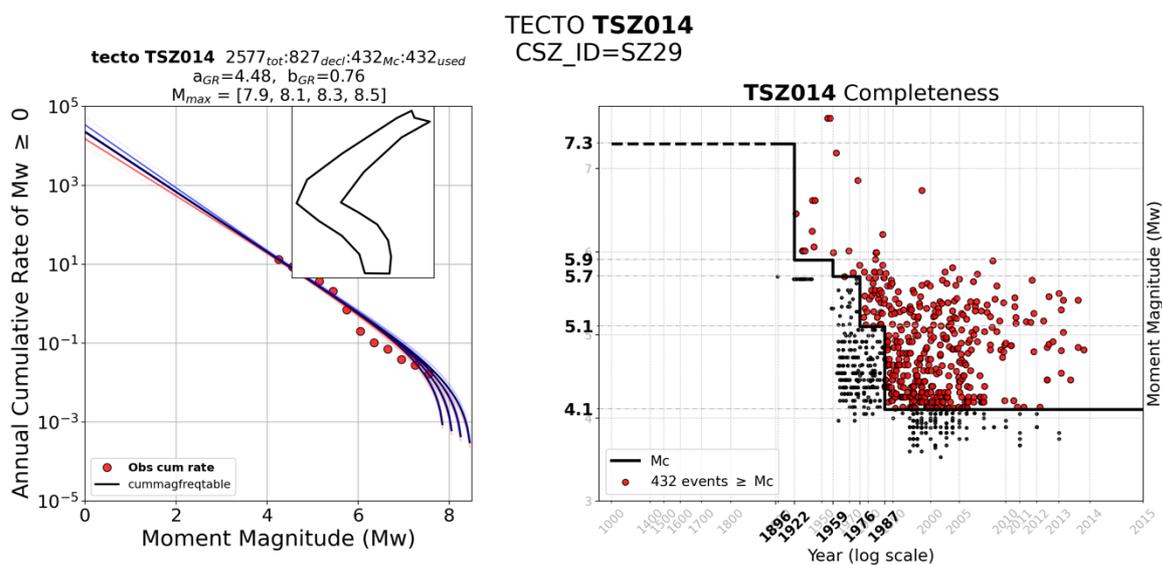


Figure 9 : Seismicity details within a single zone.

Zone Plots (Figure 9) will all have the same general features. The left-hand plot shows the cumulative observed rate (red circles), along with the predicted cumulative rate using the calculated a_{GR} and b_{GR} . Often, several (weighted) maximum magnitudes are provided, and these are each shown with their own solid line. The blue and red lines express the covariance, as calculated by Weichert.

Depending on the quality of the calculation, the covariance may be a fixed value. The inset will show the outline of the zone of interest, to help orient the viewer.

The right-hand plot shows the catalog that was the basis of the aforementioned plot. Notice the logarithmic x-scale, this emphasizes the more recent events, where the most data exist.

The Magnitude of completeness through time is represented as a solid black line, with events appearing as red circles ($M \geq M_c$) or gray dots ($M < M_c$). Dates and magnitudes from the completeness table are in bold.

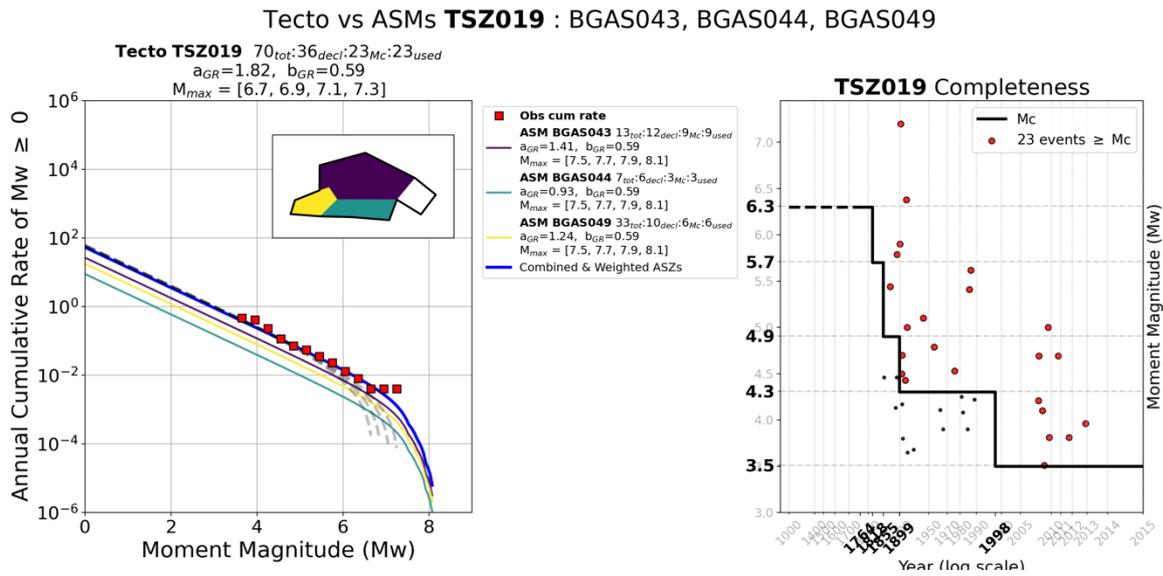


Figure 10: Tecto vs ASZ plots showing the modeled relationship between the Tecto and ASZ modeled rates.

Tecto vs ASZ plots (Figure 10) and CSZ vs Tecto plots are similar to the zone plots, but have additional detail. The parent zone’s outline is plotted in Black. The predicted cumulative occurrence rates appear as dashed lines whose intensity is dependent upon the weighting factor for each. Each zone is plotted in a unique color, along with its weighted cumulative occurrence.

The total combined weighted observations for all sub-zones is plotted as a solid blue line. Details for each zone are provided in the legend. In this case, the parent zone is TECTO zone TSZ019, while the subordinate zones BGAS043, BGAS044, BGAS049 are plotted as filled areas within the parent zone. When the blue line and dashed line meet, then the model is in excellent agreement.

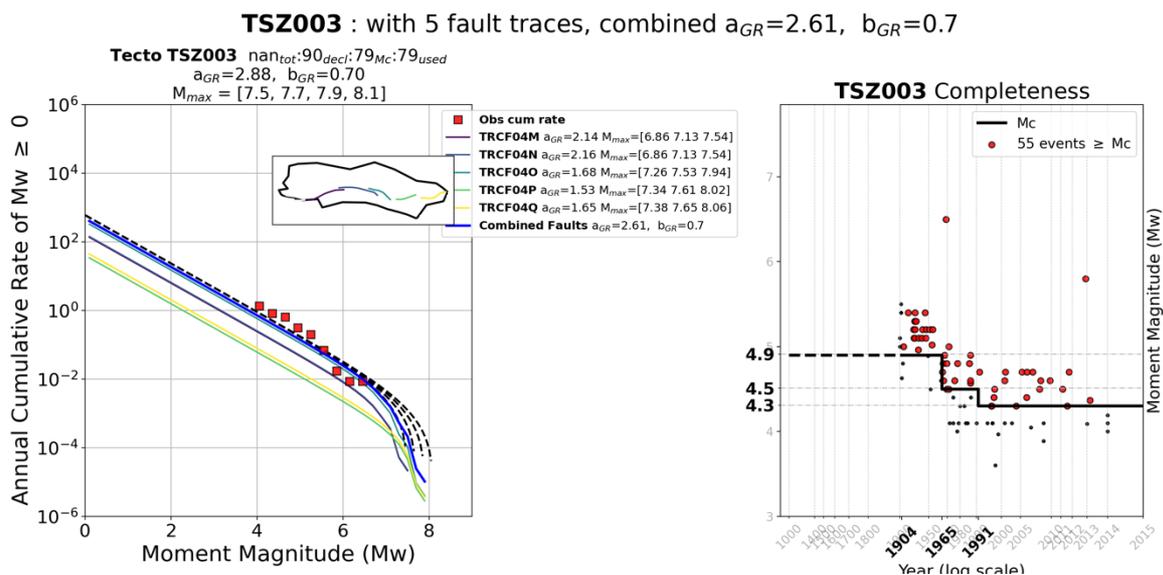


Figure 11 : TECTO zone cumulative annual occurrence plotted against total occurrences produced by faults within that zone.

Similarly, Faults can be plotted against the TECTO zone (Figure 11) and their traces are plotted on the map. Otherwise, this is similar to the TECTO vs ASZ plot.

ESHM20 Toolkit: NRML input files for use with OpenQuake

OQ-engine introduced new standards that allow a clear definition of the input and output of a seismic hazard models (Pagani et al 2014). The input/output standards, called Natural hazard Risk Markup Language (NRML) are variants of XML (eXtensible Markup Language). XML is an open-source standard is an exchange data format, which is both human and machine-readable.

The current NRML-version is available online at <https://github.com/gem/oq-engine/blob/master/openquake/hazardlib/nrml.py> and allows the description of seismic sources, ground-motion models, as well as logic-tree structures. For further details of the NRML standard definition, we recommend the online resources of the OQ-engine available at

<http://www.globalquakemodel.org/openquake/support/documentation/>.

5.4.8 ESHM20 Toolkit: Requirements

Primary requirements for the model fitting engine includes: Python \geq 3.6, GeoPandas (working with shape files and geospatial relationships), pandas (workhorse for dealing with table data), matplotlib (for visualizations), OpenQuake (for occurrences, g-r calculations, declustering, and a myriad of other functions)

5.5 ESHM20 Toolkit: Interface

Most of the parameters used within this toolkit are provided in the definitions.py file. After modifying these parameters, the user may continue the process within jupyter notebooks or in a python development environment.

5.5.1 Python Integrated Development Environment (IDE) (ex. PyCharm)

The functions from within ESHM toolkit can be run from any python prompt, but for active development, it is recommended that an IDE be used (Figure 12). Here one can fine-tune all the parameters and examine the results throughout the process. Although in principle this may be similar to running in a notebook, powerful introspection, debugging, and syntax checking capabilities provide for a robust and efficient experience. A full example of this usage can be found in `run_the_gr_engine.py`. The details applied at each zone level is found in `level_independent_processing_routines.py`.

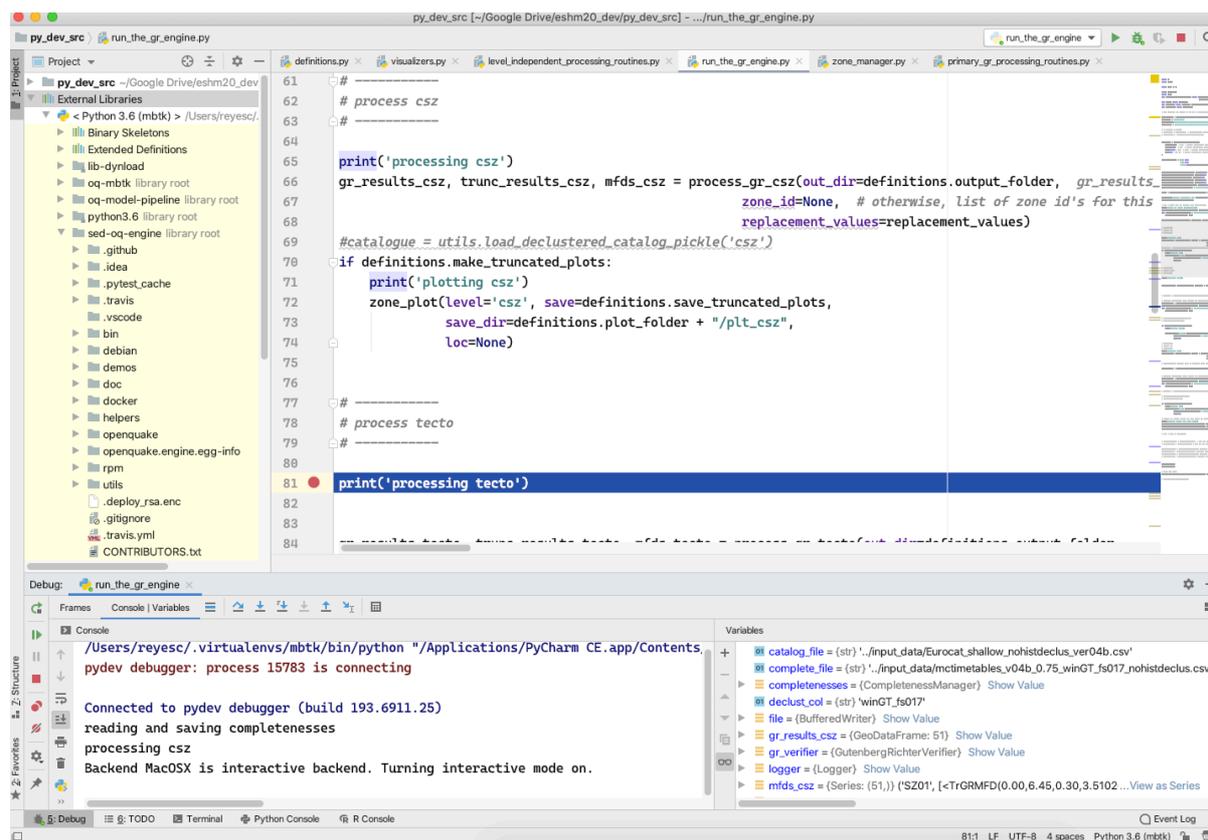


Figure 12 Running the ESHM within an integrated development environment (PyCharm)

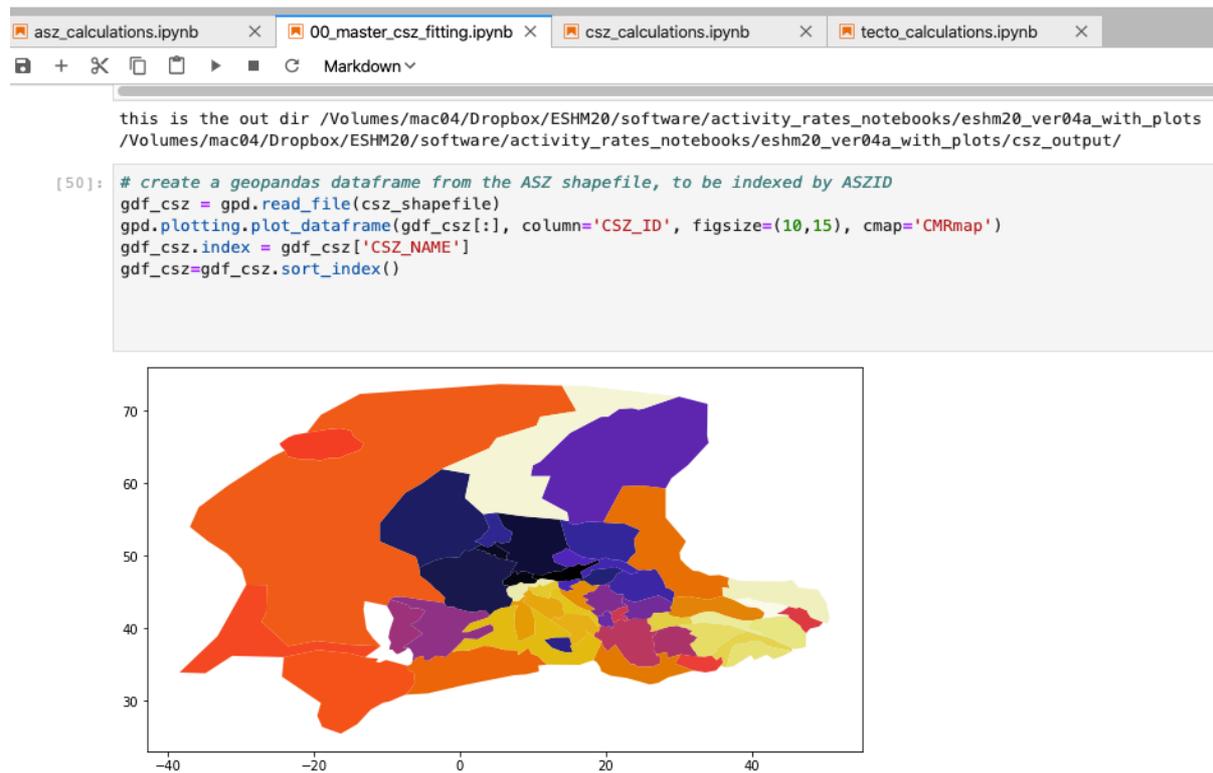
Processing Workflow, as expressed through Jupyter Notebooks

The toolkit can be implemented as a workflow from a series of jupyter notebooks:

1. 01_configuration.ipynb : prepares directories, preferences and data for further processing
2. 02_csz_fitting.ipynb : run the Gutenberg-Richter (GR) fitting for superzones (CSZ zones)
3. 03_tecto_fitting.ipynb: run the GR fitting for tectonic regions (TECTO zones)
4. 04_asz_fitting.ipynb: Run the GR fitting for area seismicity (ASZ zones)
5. 05_supplimentary_asz_fitting.ipynb : Handle inadequately fitted ASZ regions common-beta and prior-beta fitting.

5.5.2 Jupyter Notebook interface

The Jupyter Notebook is an open-source web application/interface that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Jupyter notebooks are developed open-source software, open-standards, and services for interactive computing across many programming languages.



load the completeness, and associate it to each zone

add following fields to the source_model dataframe

column	Description
completeness	contains [[yr, mag_comp],[...]] in descending year order

Figure 13 Example notebook screenshot showing CSZs as different colors

5.6 References

Reasenber, Paul. "Second-order moment of central California seismicity, 1969–1982." *Journal of Geophysical Research: Solid Earth* 90, no. B7 (1985): 5479-5495.

Stromeyer, D., and G. Grünthal. "Capturing the uncertainty of seismic activity rates in probabilistic seismic-hazard assessments." *Bulletin of the Seismological Society of America* 105, no. 2A (2015): 580-589.

Zaliapin, Ilya, Andrei Gabrielov, Vladimir Keilis-Borok, and Henry Wong. "Clustering analysis of seismicity and aftershock identification." *Physical review letters* 101, no. 1 (2008): 018501.

Liability claim

The European Commission is not responsible for any use that may be made of the information contained in this document. Also, responsibility for the information and views expressed in this document lies entirely with the author(s).