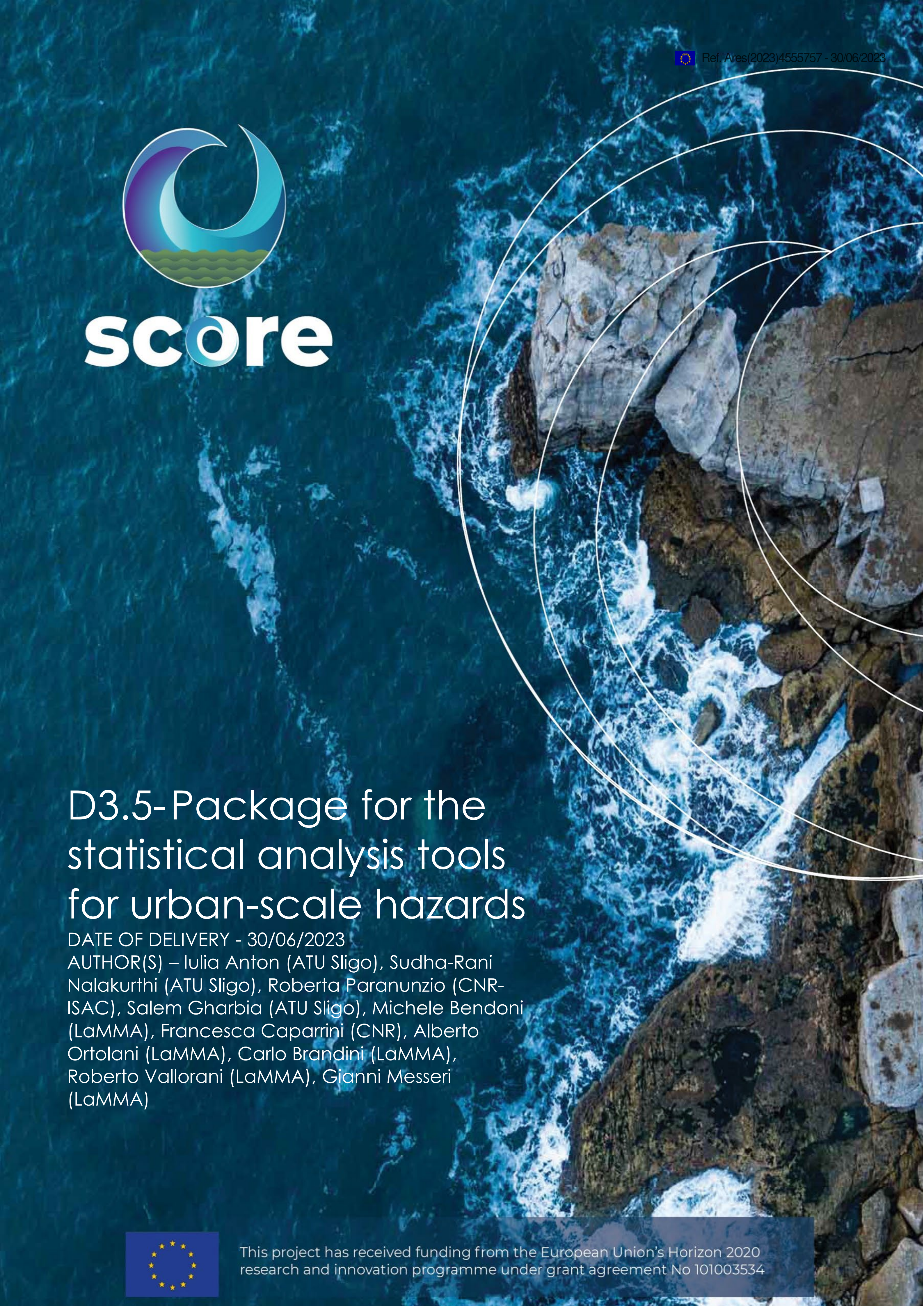

D3.5-Package for the statistical analysis tools for urban-scale hazards

DATE OF DELIVERY - 30/06/2023

AUTHOR(S) – Iulia Anton (ATU Sligo), Sudha-Rani Nalakurthi (ATU Sligo), Roberta Paranunzio (CNR-ISAC), Salem Gharbia (ATU Sligo), Michele Bendoni (LaMMA), Francesca Caparrini (CNR), Alberto Ortolani (LaMMA), Carlo Brandini (LaMMA), Roberto Vallorani (LaMMA), Gianni Messeri (LaMMA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003534

DOCUMENT TRACKS DETAILS

Project acronym	SCORE
Project title	Smart Control of the Climate Resilience in European Coastal Cities
Starting date	01.07.2021
Duration	48 months
Call identifier	H2020-LC-CLA-2020-2
Grant Agreement No	101003534

Deliverable Information	
Deliverable number	D3.5
Work package number	WP3
Deliverable title	Package for the statistical analysis tools for urban-scale hazards
Lead beneficiary	ATU Sligo
Author(s)	Iulia Anton (ATU Sligo), Sudha-Rani Nalakurthi (ATU Sligo), Roberta Paranunzio (CNR-ISAC), Salem Gharbia (ATU Sligo), Michele Bendoni (LaMMA), Francesca Caparrini (CNR), Alberto Ortolani (LaMMA), Carlo Brandini (LaMMA), Gianni Messeri (LaMMA), Roberto Vallorani (LaMMA)
Due date	30/06/2023
Actual submission date	30/06/2023
Type of deliverable	Demonstrator
Dissemination level	Public

VERSION MANAGEMENT

Revision table			
Version	Name	Date	Description
V 0.1	Iulia Anton (ATU Sligo), Sudha-Rani Nalakurthi (ATU Sligo), Roberta Paranunzio (CNR-ISAC), Salem Gharbia (ATU Sligo), Michele Bendoni (LaMMA), Francesca Caparrini (CNR), Alberto Ortolani (LaMMA), Carlo Brandini (LaMMA), Roberto Vallorani (LaMMA), Gianni Messeri (LaMMA)	23/05/2025	First draft
V 0.2	Asier Undabeitia (Naider), Giovanni Serafino (MBI)	22/06/2023	Updated draft internally reviewed
V 0.3	Iulia Anton (ATU Sligo), Sudha-Rani Nalakurthi (ATU Sligo), Roberta Paranunzio (CNR-ISAC), Salem Gharbia (ATU Sligo), Michele Bendoni (LaMMA), Francesca Caparrini (CNR), Alberto Ortolani (LaMMA), Carlo Brandini (LaMMA), Roberto Vallorani (LaMMA), Gianni Messeri (LaMMA)	28/06/2023	Updated draft after contribution from partners
V1.0	Iulia Anton (ATU Sligo) Salem Gharbia (ATU Sligo)	30/06/2023	Final version

All information in this document only reflects the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

LIST OF ACRONYMS AND ABBREVIATIONS

Acronym / Abbreviation	Meaning / Full text
CCLL	Coastal City Living Lab
EBA	Ecosystem Based Approach
EU	European Union
RPO	Research Performing Organisation
SME	Small and Medium-sized Enterprises

BACKGROUND: ABOUT THE SCORE PROJECT

SCORE is a four-year EU-funded project aiming to increase climate resilience in European coastal cities.

The intensification of extreme weather events, coastal erosion and sea-level rise are major challenges to be urgently addressed by European coastal cities. The science behind these disruptive phenomena is complex, and advancing climate resilience requires progress in data acquisition, forecasting, and understanding of the potential risks and impacts for real-scenario interventions. The Ecosystem-Based Approach (EBA) supported by smart technologies has potential to increase climate resilience of European coastal cities; however, it is not yet adequately understood and coordinated at European level.

SCORE outlines a co-creation strategy, developed via a network of 10 coastal city 'living labs' (CCLLs), to rapidly, equitably and sustainably enhance coastal city climate resilience through EBAs and sophisticated digital technologies.

The 10 coastal city living labs involved in the project are: Sligo and Dublin, Ireland; Barcelona/Vilanova i la Geltrú, Benidorm and Basque Country, Spain; Oeiras, Portugal; Massa, Italy; Piran, Slovenia; Gdansk, Poland; Samsun, Turkey.

SCORE will establish an integrated coastal zone management framework for strengthening EBA and smart coastal city policies, creating European leadership in coastal city climate change adaptation in line with The Paris Agreement. It will provide innovative platforms to empower stakeholders' deployment of EBAs to increase climate resilience, business opportunities and financial sustainability of coastal cities.

The SCORE interdisciplinary team consists of 28 world-leading organisations from academia, local authorities, RPOs, and SMEs encompassing a wide range of skills including environmental science and policy, climate modelling, citizen and social science, data management, coastal management and engineering, security and technological aspects of smart sensing research.

EXECUTIVE SUMMARY

This document is a deliverable of the SCORE project, funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003534.

The aim of this document is to develop a set of local-scale analysis tools to help analyse historical data and projections related to coastal flooding, extreme weather events, and other hazards. These tools will be used to estimate the trends in parameters of interest, identify suitable distributions and return periods and group homogeneous conditions. A detailed explanation of how to use the tools and data produced by D3.5 is provided by D3.6.

LINKS WITH OTHER PROJECT ACTIVITIES

The subsequent project activities and deliverables considered in the preparation of D3.5 include the input data from Task 3.1 and the time-series of downscaled data from Task 3.2. These data were used to produce the statistical parameters necessary for Task 3.4, such as return periods and data classification. The output of D3.5 will be used in D3.6 to produce a document describing the processing tools and data of D3.5, including a description of the tools used and the description of the produced datasets. Further, the D3.6 outcomes will be used to derive the hazard maps for WP6 and will also be integrated in the DT to take into account future weather and hydraulic scenarios.

TABLE OF CONTENT

INDEX OF TABLES

1. INTRODUCTION

1.1. Scope of the deliverable

The purpose of D3.5 is to provide a set of statistical analysis tools that can be used to analyze historical data and projections of urban-scale flood scenarios and elaborations produced in Task 3.3. The tools can be used to estimate trends in parameters of interest (sea level, waves, wind, rainfall, sea temperature, etc.), identify suitable distributions and return periods, perform data clustering through multivariate analysis. The output of D3.5 includes statistical parameters such as return periods and data classification which can be used in Task 3.4.

The D3.3 deliverable is available in the next link: https://doi.org/10.5281/zenodo.8034107 .

1.2. Structure and content of deliverable

The deliverable is organised in several sections related to the statistical analysis tools of different parameters, especially for significant wave heights and river discharge. Each section provides a short introductory description of the content.

Procedures are written in R language (R Core Team, 2021). The choice of R is based on its extensive collection of statistical libraries and packages, which provide comprehensive data analysis tools. Combined with its intuitive data manipulation tools, R has robust data visualization capabilities that make it a preferred tool for exploring and presenting statistical data. R is also an open-source language that is accessible to all users, making it a cost-effective and widely adopted statistical tool.

The following scripts (Table 1) are documented within them.

Table 1: Files available under this package

Aim	File Name	File Description
Time Series Analysis	TimeSeriesAnalysis.r	This script analyses the time series inputted by the user and include different plots and a summary statistics
EVA analysis for different parameters	General_EVA.r	This script calculates the extreme value analysis for multiple parameters. In this file EVA is provided for significant wave height.
EVA analysis for river discharge	River_discharge_EVA.r	This script calculates the extreme value analysis for river discharge.
Find the optimal threshold	Calculate_threshold.r	This script includes multiple methods to calculate the threshold (Thompson, Anderson-Darling, Solari, Graphical).
Data Clustering	DataClustering.r	This script includes data clustering analysis (partitioning clustering and hierarchical clustering).

2. TIME SERIES ANALYSIS

The file name for the timeseries analysis is “TimeSeriesAnalysis.r”.

This script includes:

- *Data collection and analysis:* create a data frame of a NetCDF file and calculates the number of values for each year. It contains also numerous plots to investigate the timeseries (yearly maxima, monthly maxima, daily mean
- *Trend analysis*: used to test for a monotonic trend in a univariate data set. The result of this script includes the Kendall's tau statistic and the p-value.
- *Summary statistics*: generate a concise summary of a given data set. It provides a quick overview of the data, including the number of observations, the mean, median, minimum, and maximum values, and other descriptive statistics. Additionally, this summary can be used to describe the distribution of the data, calculate the correlation between variables, and create graphical representations of the data.

3. EXTREME VALUE ANALYSIS

For the extreme value analysis, this package includes three files: General_EVA.r, River_discharge_EVA.r and Calculate_threshold.r.

The file General_EVA.r contains the methods used to calculate the threshold and return levels, based on the Peak Over Threshold Method and Block Maxima methods. In addition to a result table with all the relevant parameters (scale, shape, location, negative maximum likelihood estimation, AIC and BIC), the code also includes multiple plots (return levels, histograms, density plots, QQs, and QQ2) required in choosing the appropriate distribution/methodology. For this file, the threshold is calculated only using the Thompson test (Thompson et al., 2009) from the “TEA” R package (Ossberger, 2022) or the Anderson-Darling test from the “EVA” R package (Bader et al., 2018). For calculating the return levels using POT method, the user will need to input the value for the threshold either from one of the two methods from this file or from all the other methods from Calculate_threshold.r file. As an example, in this file, the threshold was chosen based on the results from the Anderson-Darlin test. More information on this selection can be found in D3.6.

The River_discharge_EVA.r file is similar to General_EVA.r, except it applies to a specific parameter, River discharge.

In the file Calculate_threshold.r., the threshold to use for the Peak Over Threshold method can be determined using the Graphical approach, Thompson methodology (Thompson et al., 2009), Anderson-Darling Test (Bader et al., 2018) or Solari method (Solari et al., 2017). A script is also included in the file for creating a table that shows the return levels and thresholds for all the methods.

4. DATA CLUSTERING

The file name for clustering analysis is "DataClustering.r"

This file DataClustering.r includes a script to compute clustering analysis exploiting common clustering algorithms, methods to find the best number of clusters and validation approaches. Clustering algorithms find the structure in the observations so that elements of the same group (cluster) are more similar to each other than to those from different clusters. The script includes:

- Functions to compute *partitioning cluster*, specifically, k-means (MacQueen, 1967) and related plots (e.g., cluster plot).
- As alternative, functions to perform *hierarchical cluster* (Johnson 1967), specifically, exploiting the agglomerative approach and relative plots (e.g., dendrogram).
- Functions to compute the *elbow method* to assess the right number of clusters for the k-means analysis (Thorndike 1953) and related plot.
- A *silhouette analysis* to validate clusters (Rousseeuw, 1987) and related plot (e.g., clusters silhouette plot).

5. ANALYSIS OF WEATHER CIRCULATION TYPES

For what concerns the weather circulation types, the files that contain class centroids are provided. They are obtained from ERA5 (1979-2005) data.

The names of the files are the following:

- Centroidi_PCT-z500_COLy
- Centroidi_PCT-mslp_COLy
- Centroidi_SAN-z500_COLy
- Centroidi_SAN-mslp_COLy

The file name, SAN or PCT indicates the classification methods. PCT is the classification by obliquely rotated Principal Components in T-mode; SAN is a non-hierarchical cluster analysis method like k-means. 500 hPa geopotential height (z500) and Mean Sea Level Pressure (mslp) are the variables used for the classification.

These files are in ASCII format and consist of 9 columns, one for each type of circulation. Each column consists of 61x61 values (3721 rows or grid points) covering a domain of latitude 35°N-50°N and longitude 5°E-20°E latitude, regularly spaced at 0.25 degrees. The first row (of any of the 9 columns) gives the values of z500 or mslp in meter (m) and hectopascal (hPa) respectively, for the southernmost latitude and the westernmost longitude (i.e., 35°N-5°E). The second row is for the second longitude from west still for the southernmost latitude (i.e., 35°N-5.25°E) and so on up to the 61st row (corresponding to 35°N-20°E). After 61 rows in fact also the latitude progresses (and longitude starts again from the first value), it means that the 62nd row is for 35.25°N-5°E, the 123rd row is for 35.5°N-5°E, etc. The last row is for the easternmost and northernmost grid point (i.e., 50°N-20°E).

A second set of files is the following:

- SAN-mslp_COLy.cla
- SAN-z500_COLy.cla
- PCT-mslp_COLy.cla
- PCT-z500_COLy.cla

It contains the type of daily circulation from Jan. 1, 1979 to Dec. 31, 2005. The file name specifies the classification method (PCT or SAN) and the variable used (mslp or z500).

The files are in ASCII format. The first column is a progressive integer from 1 to 9862 corresponding to the progressive number of days from the 1st of January 1979 up to 31st of December 2005. The second column is the number of the type of circulation associated with the day (i.e., values from 1 to 9).

This information is included also in a README file provided with the dataset.

6. CONCLUSIONS

In conclusion, D3.5 provides a set of statistical analysis tools to analyze historical data and urban-scale flood scenarios. It includes an R package with numerous files for time series analysis, data clustering, and extreme value analysis, which will be used to estimate trends in parameters of interest, identify distributions and return periods, and calculate thresholds. Centroids and daily classification types are also provided for a first analysis domain regarding the weather circulation types. The output of D3.5 will be used in Task 3.4, providing statistical parameters such as return periods and data classification.

REFERENCES

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Annals of Applied Statistics*, *12*(1). https://doi.org/10.1214/17-AOAS1092

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241-254.

Ossberger, J. (2022). *Package "tea": Threshold Estimation Approaches*. Cran R Package. https://cran.r-project.org/web/packages/tea/tea.pdf

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: URL https://www.R-project.org/. Accessed on 31st May 2023.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Solari, S., Egüen, M., Polo, M. J., & Losada, M. A. (2017). Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, *53*(4). https://doi.org/10.1002/2016WR019426

Thompson, P., Cai, Y., Reeve, D., & Stander, J. (2009). Automated threshold selection methods for extreme wave analysis. *Coastal Engineering*, *56*(10). https://doi.org/10.1016/j.coastaleng.2009.06.003

Thorndike, R.L. (1953). "Who Belongs in the Family?". Psychometrika. 18 (4): 267–276. doi:10.1007/BF02289263. S2CID 120467216.