



European
Commission

JRC SCIENTIFIC INFORMATION SYSTEMS AND DATABASES REPORT

WebAriadne 3.0.0 User's Manual [DRAFT VERSION]

Emmanuele Sordini



This publication is a Scientific Information Systems and Databases report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

EU Science Hub

<https://ec.europa.eu/jrc>

JRCXXXXXX

EUR XXXXX XX

PDF	ISBN XXX-XX-XX-XXXXX-X	ISSN XXXX-XXXX	doi:XX.XXXX/XXXXXX
Print	ISBN XXX-XX-XX-XXXXX-X	ISSN XXXX-XXXX	doi:XX.XXXX/XXXXXX

Ispra: Publications Office of the European Union, 2020

© European Union 2020



The reuse policy of the European Commission is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Except otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated. For any use or reproduction of photos or other material that is not owned by the EU, permission must be sought directly from the copyright holders.

All content © European Union, 2020

How to cite this report: Author(s), *Title*, EUR (where available), Publisher, Publisher City, Year of Publication, ISBN 978-92-79-XXXXX-X (where available), doi:10.2760/XXXXX (where available), JRCXXXXXX.

Contents

1	Introduction	5
2	Scope and prerequisites	6
3	Access	7
3.1	System requirements and supported browsers.....	7
3.2	Login	7
4	Supported data formats	1
4.1	CSV Format	1
4.2	COMEXT Format	1
4.3	"NON-COMEXT" formats	2
5	Main page	3
5.1	WebAriadne's dashboard	3
5.2	Main menu bar.....	4
5.2.1	"Dataset" menu	5
5.2.2	"Applications" menu	5
5.2.3	"Results" menu	5
5.2.4	"User" Menu.....	6
6	Workflow.....	7
7	User notification	8
8	Importing and managing data sets	9
8.1	Importing datasets: WebAriadne's import wizard.....	9
8.1.1	Column data formats	13
8.2	Managing existing datasets	14
9	Running statistical procedures	17
9.1	Introduction	17
9.2	Generalized Benford's law tool.....	17
9.3	Benford's law for customs	19
9.3.1	Worked example of the Benford's law for customs	21
9.4	Robust Regression Outliers	23
9.4.1	Parameters for Robust Regression Outliers	24
9.4.2	Additional variables	24
9.4.3	A worked example for Robust Regression Outliers	24
10	Viewing and exporting results.....	29
10.1	Main result view	29
10.2	Detailed result view for the Generalized Benford's law tool	30
10.3	Detailed result set view for Benford's law for customs.....	31
10.4	Detailed result view for Robust Regression Outliers.....	35
11	Glossary	38

1 Introduction

ARIADNE is an application for the detection of statistical anomalies and underlying structures in large-scale data. It allows selected users to import, pre-process and analyze their data with standard SAS procedures, produce user-chosen descriptive statistics for insight into and exploration of data, and to run SITAF - developed statistical procedures.

WebARIADNE is the web-based version of ARIADNE. WebARIADNE allows selected users to run SITAF-developed procedures for insight into and exploration of data sets of interest. WebARIADNE consists of a rich-client JavaScript user interface and a Java-based backend running on Apache Tomcat that acts as the main hub for all client requests and calls to the statistical procedures. WebAriadne supports multiple statistical and mathematical platforms, namely: SAS, MATLAB and R; each statistical procedure available in WebAriadne is written in one of these three languages. WebAriadne makes it possible to manage user-supplied raw data sets, use them as input to selected statistical procedures, and retrieve the results of processing runs.

Please note that WebAriadne is currently undergoing important changes. New features are being added on a regular basis, therefore the content of this manual might not be up to date with the latest version of the application. Thank you for your understanding.

2 Scope and prerequisites

The content of this manual is applicable only to the version of WebAriadne indicated on the cover. The purpose of this manual is to provide the user with a guide on how the application can be accessed and used, and contains a minimum amount of technical information. Additionally, it is assumed that the user already has appropriate working knowledge of the following concepts:

- Some basic understanding of statistics, such as: outliers, linear regression, multivariate data, numerical and categorical variables. **These topics are not covered in this user manual.**
- Statistical procedures available in WebAriadne (purpose, domain of application, parameters, results)
- CSV (short for "comma-separated values") file format

If you do not meet these prerequisites, it is strongly recommended you fill in this knowledge gap before you start using WebAriadne. Hands-on courses on WebAriadne and related concepts might be offered in the future, but the planning is not known at the time of this writing.

For more details on the implementation or the architecture of WebAriadne and on the statistical procedures accessible through WebAriadne, please refer to the relevant documentation.

3 Access

3.1 System requirements and supported browsers

Here are the minimum system requirements for WebAriadne:

- Screen resolution of at least 1280 x 720 pixels. However, a resolution of 1280 x 1024 pixels or better is recommended.
- 24-bit color depth

The following browsers are currently supported by WebAriadne:

- Internet Explorer, version 11 (older versions not supported)
- Microsoft Edge (any version)
- Mozilla Firefox, version 70 or later
- Google Chrome version 75 or later
- WebAriadne will most likely work with other browsers such as Opera, but full compatibility is not guaranteed.

For the time being, mobile devices are not supported. The user interface will likely not work on smartphones with screen 6" wide or less, while it *should* work on 7-plus-inch tablets.

3.2 Login

WebAriadne is available at the following web address:

<https://webariadne.jrc.ec.europa.eu>

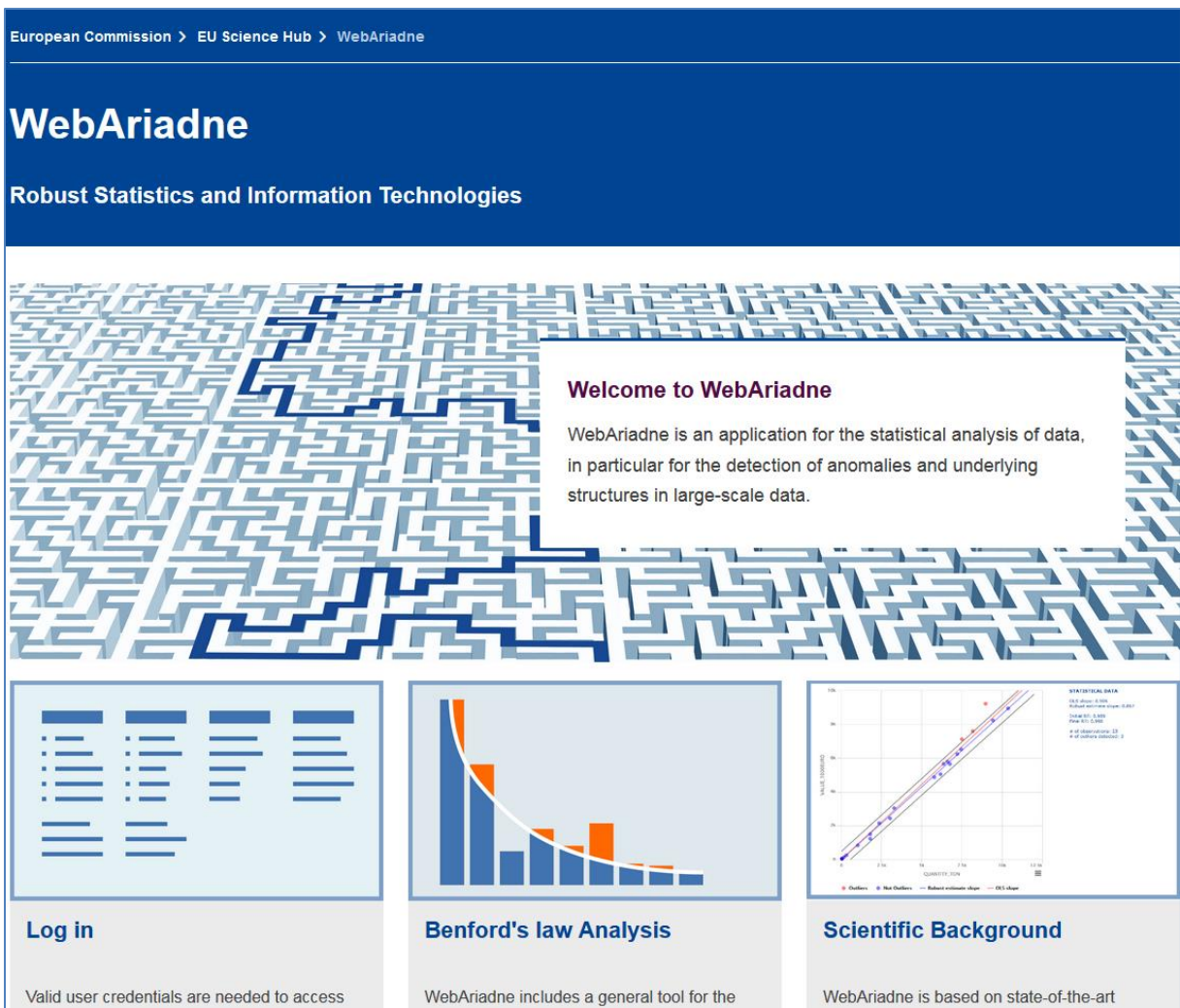


Figure 1 – Upper portion of WebAriadne’s landing page (homepage)

All the content of the WebAriadne website is private except for the homepage, a brief explanation page and the disclaimer and privacy statement page. A valid account is needed in order to be able to access WebAriadne. Please contact the webmaster if you do not have one.

Once logged in, the user will be taken to the homepage. A “logout” link is available on the upper right.

Warning: users will be logged out automatically after thirty minutes of idle time.

Important note: each user is associated to one profile containing the list of applications they are allowed to use. The content of each user profile has been set and/or agreed upon based on the user’s needs, body of affiliation, privacy issues or separate negotiations. Any profile change request will have to be addressed in writing to the relevant hierarchy on the basis of solid grounds. Any such requests addressed directly to the webmaster will be disregarded.

4 Supported data formats

4.1 CSV Format

WebAriadne features a data import wizard allowing import of user data. The format supported is the very popular CSV format (CSV: short for “Comma Separated Values”). The CSV format has the following properties:

- Each data set is available in the form of a text file;
- Each observation (record) is available in a separate line;
- Data fields are separated by a special character, such as tab, comma and semicolon. The field separator must be the same throughout a single file;
- The order and name of the fields are fixed and all records in a file must have the same format;
- Optionally, a header row with column names (separated by the same field separator) can be present.

In order to be imported with WebAriadne, raw text files must meet the following constraints:

- **Size:** max. 10 Mb over the Internet; 50, max. 100 Mb for EU institutions and JRC external sites; 1 Gb for JRC Ispra local connections
- **Valid field separators (delimiters):** comma (“,”), colon (“:”), semicolon (“;”), “pipe” (“|”), tab
- **Minimum number of fields (columns):** 2 (two). Single-column files cannot be imported.
- Maximum number of fields (columns): 300
- **Decimal separator:** decimal point (“.”) is allowed. The comma can only be used as a field delimiter.
- Thousands separator: not accepted;
- **String quote:** string fields should be enclosed in double quotations (“”). Alternatively, the user must make sure that fields do not contain any reserved chars or blank spaces¹;
- **Header row:** recommended but not compulsory. If not present in the CSV file, the application will choose default column names (e.g. “Variable1”, “Variable2”, etc.), which can be individually changed by the user;
- **Record (row) separator:** Unix or Windows text file formatting (“\n” or “\n\r”).

As a general rule, the format will have to be consistent throughout the file, because data interpretation and duplicate-key errors will cause the import to fail. These errors could also be treated as simple warnings, but then all records causing such errors will be discarded. So, please make sure your data is high quality before importing it.

Important note: all data imported must be completely anonymized, i.e. it must not contain any kind of personal data (e.g. name, last name, personal ID number, legal entity ID, VAT ID, etc.). Failure to comply with this requirement will result in immediate deactivation of your account.

4.2 COMEXT Format

Some statistical procedures available in WebAriadne only support the COMEXT format, a CSV format type for trade data of the COMEXT database, published on a regular basis by Eurostat. The COMEXT format features a set of fields with predefined formats and names:

Column name	Format	Example	Notes
DECLARANT	2-letter ISO country code	“GB”	
PARTNER	2-letter ISO country code	“CL”	
PRODUCT	8-letter NC product code	“08062010”	

¹ For example: strings must not contain commas if the comma is chosen as the field separator, otherwise the import will fail. The file must be carefully inspected prior to importing.

Column name	Format	Example	Notes
FLOW	Integer value	1	
STAT_REGIME	Integer value	1	
PERIOD	Year and month, YYYYMM	"200805"	Data must have monthly frequency
QUANTITY_TON	Floating-point value	198.10	
VALUE_1000EURO	Floating-point value	232.17	
SUP_QUANTITY	Integer value	0	

Below is a small excerpt of a comma-separated COMEXT raw text file with a heading line. It is advisable to enclose strings in quotation marks ("").

"DECLARANT","PARTNER","PRODUCT","FLOW","STAT_REGIME","PERIOD","QUANTITY_TON","VALUE_1000EURO","SUP_QUANTITY"

"GB","CL","08062010",1,1,"200805",198.10,232.17,0

"GB","CL","08062010",1,1,"200806",236.50,253.32,0

"GB","CL","08062010",1,1,"200807",45.80,33.94,0

"GB","CL","08062010",1,1,"200808",146.90,123.33,0

"GB","CL","08062010",1,1,"200811",20.50,24.18,0

"GB","CL","08062010",1,1,"200903",41.90,46.68,0

"GB","CL","08062010",1,1,"200904",22.80,25.54,0

"GB","CL","08062010",1,1,"200905",189.00,220.03,0

"GB","CL","08062010",1,1,"200906",89.50,88.98,0

"GB","CL","08062010",1,1,"200907",261.70,279.67,0

For a description of the name and meaning of each field, please refer to the COMEXT documentation.

Any kind of trade-related data will also do, provided it contains the fields described above. The data can be aggregated at any level (e.g. month or day), or it may as well contain the individual transactions.

4.3 "NON-COMEXT" formats

WebAriadne's import wizard is a general-purpose tool, therefore it can handle all kinds of CSV files, not just COMEXT or trade-related data, provided it is compliant with the format requirements described in section 4.1. Some statistical procedures available in WebAriadne can process non-COMEXT (or trade data): just make sure the data you import can be handled by the statistical procedure(s) you intend to use.

TODO: data type vs statistical procedure compatibility table

5 Main page

WebAriadne's main landing page (Figure 1) contains a link to the login popup dialog (Figure 2 and Figure 3).

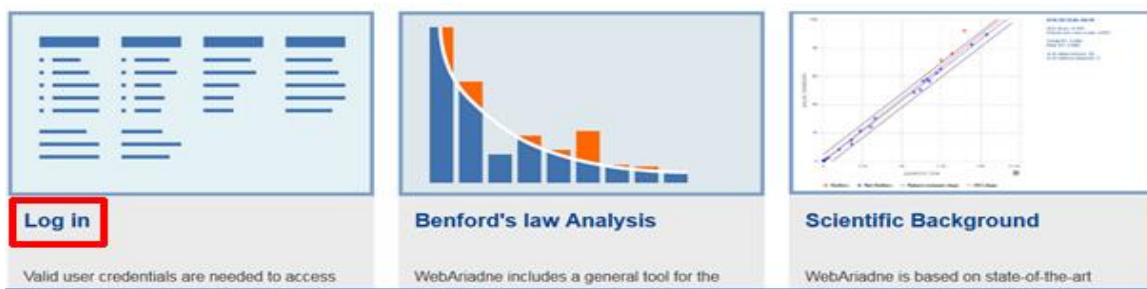


Figure 2 –WebAriadne's login link

Figure 3 –WebAriadne's login pop-up dialog box

Upon successful login, you will be taken to the application's main page. The main page is made up of a set of tiles depending on the current version of the application and user profile. The leftmost tile is a link to the application dashboard, which provides an overview of past and recent activity, and from which all the application's main features can be accessed.

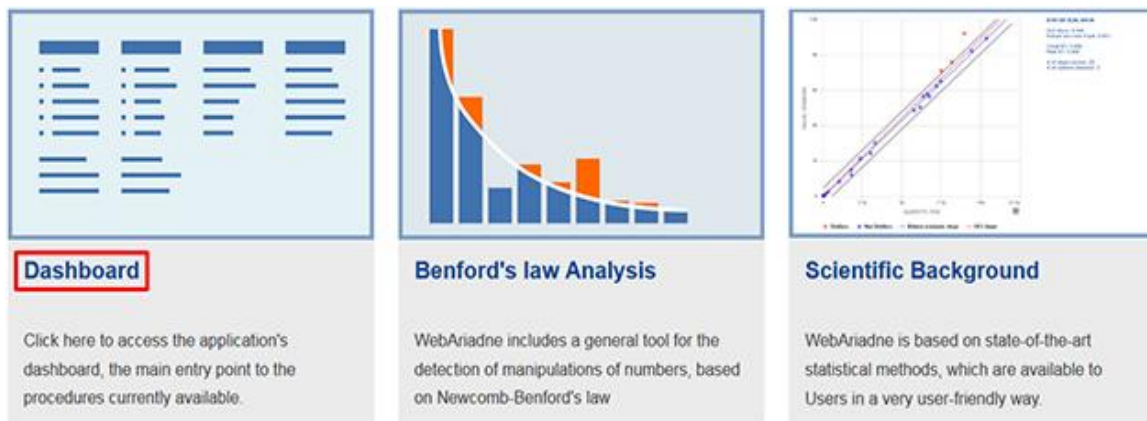


Figure 4 –WebAriadne's homepage after successful login with the link to the dashboard

5.1 WebAriadne's dashboard

The application dashboard is divided into three main sections from top to bottom:

1. Main menu bar with several menu items and links, such as Home, Dashboard, Dataset, etc.
2. Application wizard (inactive at the moment)
3. Activity overlook panel, taking up about $\frac{3}{4}$ of the total area.

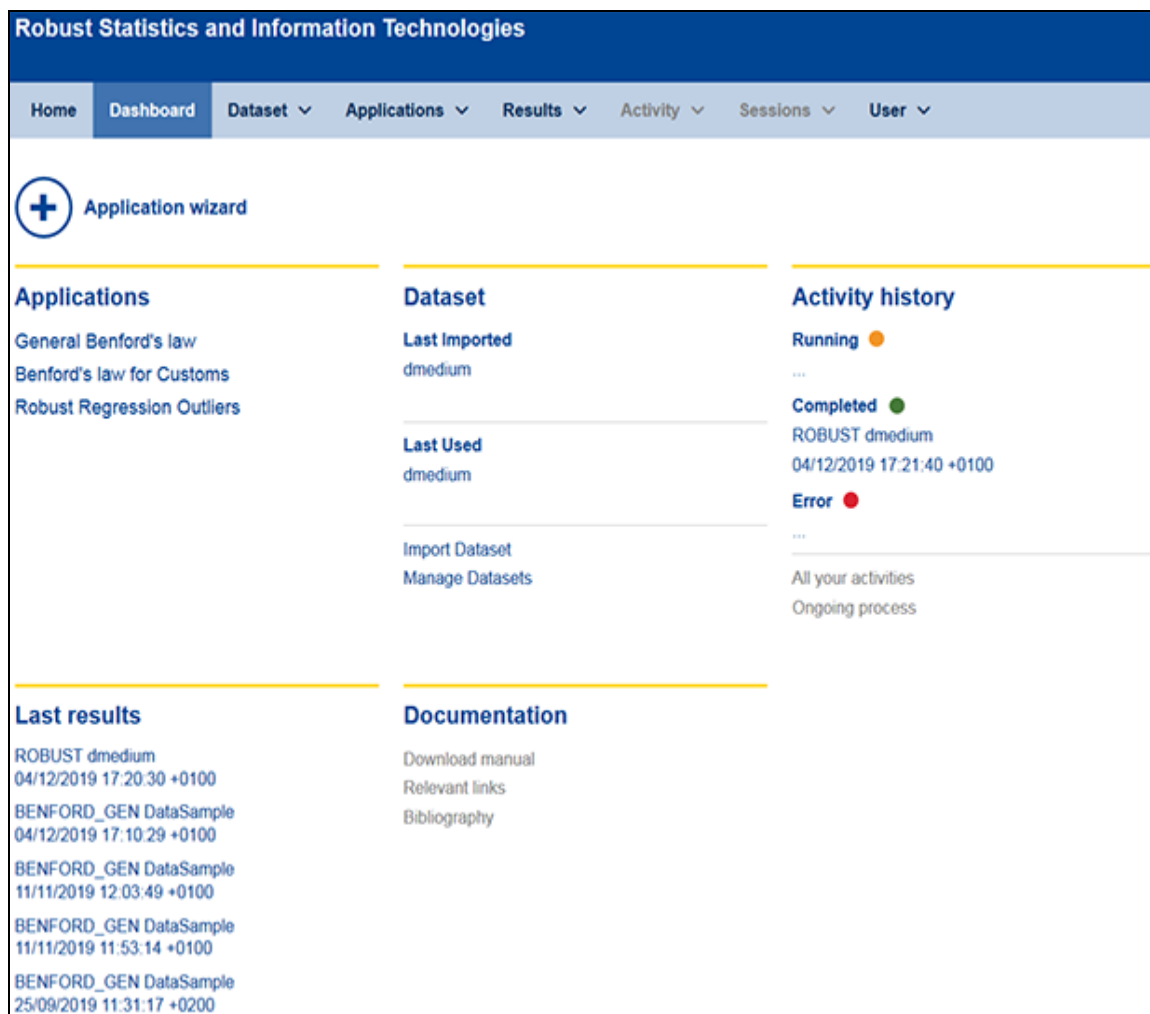


Figure 5 –WebAriadne’s dashboard view (cropped version)

The application dashboard is made up of the following sections (from left to right and from top to bottom):

- *Applications* section, containing links to all statistical applications currently available in WebAriadne;
- *Dataset* section, listing the names of the last imported and last used datasets, plus a link to the *Import Dataset* page (i.e. the import wizard) and a link to the *Manage Datasets* page
- *Activity history* section, containing the list of the currently running, completed and failed processing runs;
- *Last results* sections, containing direct links to the results of the last five processing runs, listed in reverse chronological order;
- *Documentation* section, currently disabled.

All user-specific information such as data sets, activity (usage sessions, processing history), result sets obtained from processing runs are private for each user, **and as such are not accessible to other users**.

5.2 Main menu bar

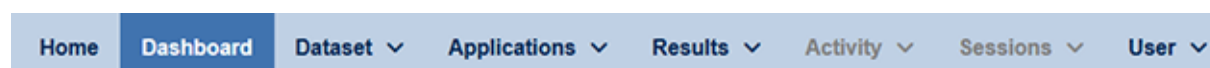


Figure 6 –WebAriadne’s menu bar

The application’s main menu bar is depicted in Figure 6. The menu bar contains the following links/drop-down menus from left to right:

1. Home (link);

2. Dashboard (link);
3. Dataset: drop-down menu containing all the dataset-related items
4. Applications: drop-down menu through which all the available statistical applications available can be accessed;
5. Results: drop-down menu providing access to all result viewing/presentation related features

5.2.1 “Dataset” menu

Home	Dashboard	Dataset ^	Applications v	Results v	Activity v	Sessions v	User v
Manage datasets			Requirements			Documentation	
Import dataset			How to			Example	
Aggregate							

Figure 7 – “Dataset” menu

The “Dataset” menu (Figure 7) currently contains two active items:

- *Manage datasets*: selecting this item will take you to the dataset management page (.....)
- *Import dataset*: selecting this item will bring up the dataset import wizard (.....)

5.2.2 “Applications” menu

Home	Dashboard	Dataset v	Applications ^	Results v	Activity v	Sessions v	User v
General Benford's law					Documentation		
Benford's law for Customs					Guideline		
Robust Regression Outliers					Setting		
					Credits		

Figure 8 – “Applications” menu

The “Applications” menu contains three active items, corresponding to the three statistical procedures available in WebAriadne:

- *General Benford's law*: selecting this item will take you to the generalized Benford's law page (.....)
- *Benford's law for customs*: selecting this item will take you to the Benford's law for Customs page (.....)
- *Robust Regression Outliers*: selecting this item will take you to the Robust Regression Outliers page (.....)

5.2.3 “Results” menu

Home	Dashboard	Dataset v	Applications v	Results ^	Activity v	Sessions v	User v
All results					Documentation		
Last result					Guideline		
Results by dataset					Setting		
Results by session					Credits		

Figure 9 – “Results” menu

The “Results” menu currently contains two active items:

- All results: selecting this item will open the overall result view page (.....)
- Last result: selecting this item will bring up the detailed result view for the processing run last performed by the user (.....)

5.2.4 “User” Menu

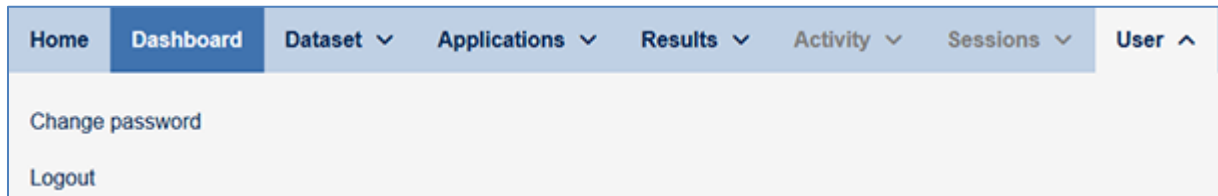


Figure 10 – “User” menu

The “User” menu contains two items:

- *Change password*: selecting this item will open the password change dialog box (.....).
- *Logout*: click here to log out from WebAriadne.

6 Workflow

WebAriadne is an application meant for managing and storing user-supplied data sets and using them for processing with ARIADNE statistical procedures. A key concept to understanding WebAriadne and making the best possible use of it is the **application workflow**. The workflow is the sequence of steps that must be taken in order to:

1. Import a data set;
2. Pre-process the data set (if needed);
3. Use it as an input to any of the available statistical procedures;
4. Get the results of the processing.

The actual processing time can vary widely up to several orders of magnitude (from a few minutes to many hours), depending on a number of factors, such as data set size, statistical procedure, statistical procedure parameters, current server load, etc. Since it is practically impossible to reliably predict the duration of a processing run, WebAriadne adopts an asynchronous execution model. This means that the user launches a statistical processing, but they do not have to stay idle waiting for the result. Rather, any statistical processing job is put on the application's execution pipeline, and the user will be notified of the outcome by e-mail.

WebAriadne offers a standard workflow consisting of the following main steps:

1. Pick a statistical procedure;
2. Select an existing data set or import a new one through the import wizard;
3. Set the parameters for the desired statistical procedure²;
4. Click the "Run procedure button";
5. Wait for the notification e-mail. Meanwhile, you can either log out or continue using the application;
6. If the run was successful, go to the result presentation section where you can view (and possibly download) the result of your processing run.

WebAriadne also offers a data set management page, where the user can preview or delete existing data sets. The import wizard is also accessible directly from the application's main menu.

² Many parameters have default values.

7 User notification

WebAriadne offers several ways of informing the user of the progress and outcome of processing runs:

1. Status dialog boxes (success or error) providing immediate feedback;
2. E-mail notification providing some basic feedback on the outcome of processing runs, following WebAriadne's asynchronous job scheduling model;
3. Progress of terminated and currently running tasks through the dedicated section in the application dashboard; "Running tasks" tool;
4. Visualization of results of processing runs and plotting of on-the-fly charts.

8 Importing and managing data sets

8.1 Importing datasets: WebAriadne's import wizard

Users can import their own data by uploading plain text files in the CSV format (see section 4.1 for details). WebAriadne offers an import wizard that can be accessed through the application dashboard, the main menu ("Import dataset" menu item) or directly from the user interface individual statistical procedure.

Import dataset

Upload file

Browse 1 **File selected:** 2

Dataset name

3

Excerpt of file

4

Delimiter **String quotes** **Variable name in the first row**

5

Preview of imported data

<input type="checkbox"/> Name	Type	Sample values
6		

7 **Import**

Figure 11 – WebAriadne's import wizard

The import wizard (see Figure 11) contains the following controls (from left to right and from top to bottom):

1. "Browse" button: clicking this button will bring up a file browser dialog box, which can be used for browsing local drive units (both fixed and removable) to select data sets (i.e. CSV files) for import. Once the local file is selected, it gets immediately uploaded onto the server;
2. *Selected file* area: this is where the name of the local dataset selected with the file browser will be displayed;
3. *Data set name*: a text field which will be populated with the name of the file selected minus any extensions (e.g. "myfile.csv" will become "myfile"). This text field is editable, so you can change the dataset name to your liking³. This is the name the data set will have in the application once imported;
4. *File excerpt area*: small, non-editable text area showing the first few lines of the file being imported;
5. *Control row*: group of controls for setting the field delimiter, the string quote, and to specify whether a header row is available.
6. *Import area*: area showing a preview of the formatting of the imported data. Here the user can choose which columns to import, and individually set the name and data type of each column;
7. *Import button*, which becomes enabled when all import wizard parameters have been correctly set. Clicking this button will import the dataset into WebAriadne.

³ Only letters A-Z, a-z and numbers from 1 to 9 are allowed in data set names. Blank spaces and other special characters not belonging to the UTF-8 format are not allowed.

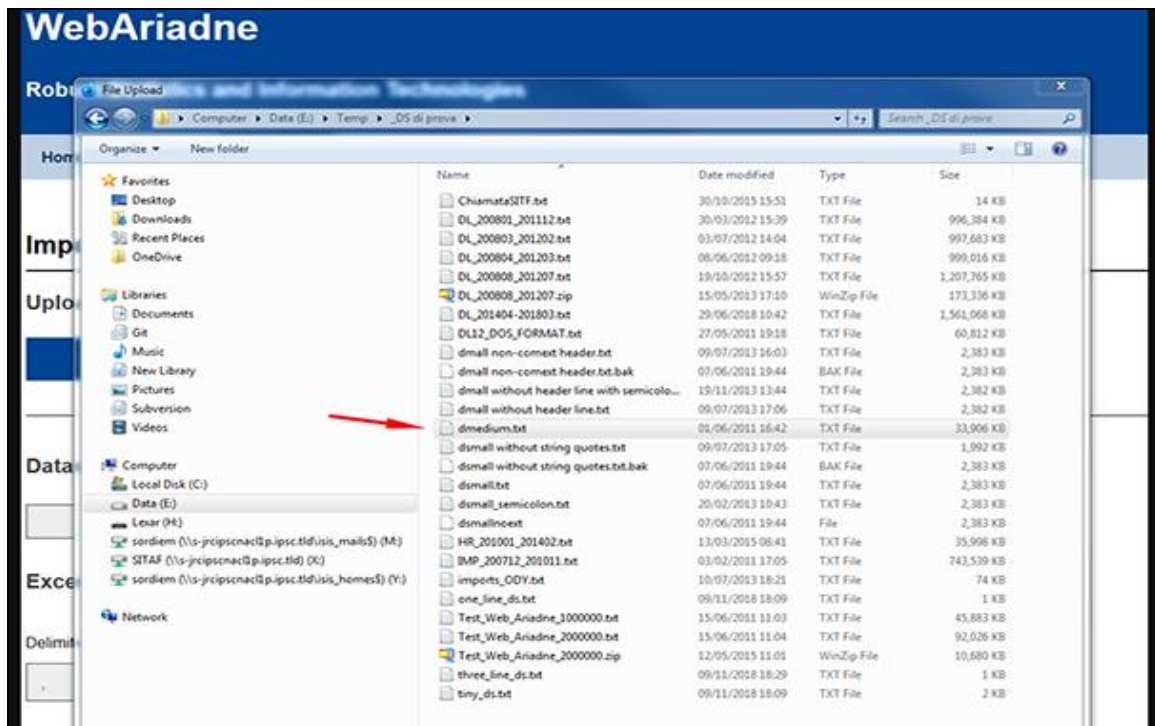


Figure 12 – Importing a dataset – Local file browser

Below are the detailed steps for importing a new data set. The example uses a COMEXT data set, but the steps are virtually the same with any kind of CSV file:

1. Click the “Browse” button. This will bring up the local file browser dialog. Browse your local PC and select a file to upload (say, “dmedium.txt”, Figure 12), double click it or select it and click “Open”.
2. The file will start uploading. The actual upload time will basically depend on the size of the file and on the local internet connection bandwidth. Some pop-up messages will be displayed informing the user of the current status. If everything goes smoothly, the user will typically see the following messages: “Uploading raw data file...”, “File successfully uploaded...” and “Updating preview...” (Figure 13);

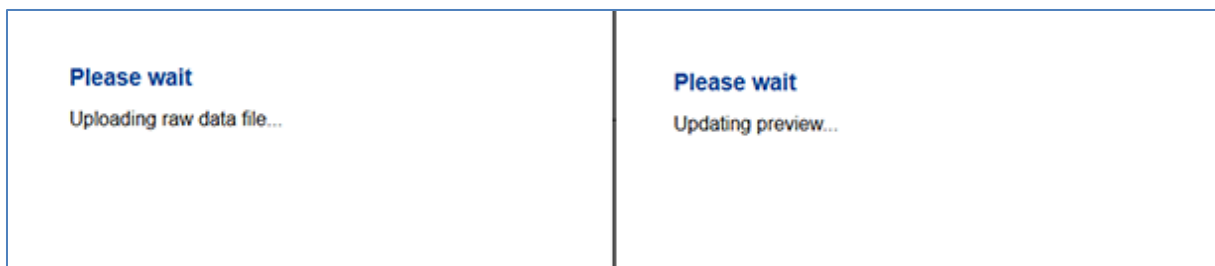


Figure 13 – Importing a dataset – Typical import status messages.

3. If the upload process was successful, the import wizard will be populated with a preview of the data being imported.

Browse

File selected: dmedium.txt

Dataset name

dmedium

Excerpt of file

"DECLARANT","PARTNER","PRODUCT","FLOW","STAT_REGIME","PERIOD","QUANTITY_TON","VALUE_1000EURO","SUP_QUANTITY"
"GB","CL","08062010",1,1,"200805",198.10,232.17,0
"GB","CL","08062010",1,1,"200806",236.50,253.32,0
"GB","CL","08062010",1,1,"200807",45.80,33.94,0
"GB","CL","08062010",1,1,"200808",146.90,123.33,0
"GB","CL","08062010",1,1,"200811",20.50,24.18,0
"GB","CL","08062010",1,1,"200903",41.90,46.68,0

Delimiter

String quotes

Variable name in the first row

,

"

☒

Preview of imported data

<input checked="" type="checkbox"/>	Name	Type		Sample values
<input checked="" type="checkbox"/>	DECLARANT	Custom string[2]	<input type="checkbox"/>	GB GB GB GB
<input checked="" type="checkbox"/>	PARTNER	Custom string[2]	<input type="checkbox"/>	CL CL CL CL
<input checked="" type="checkbox"/>	PRODUCT	Custom string[8]	<input type="checkbox"/>	08062010 08062010 08062010 08062010
<input checked="" type="checkbox"/>	FLOW	Custom string[1]	<input type="checkbox"/>	1 1 1 1
<input checked="" type="checkbox"/>	STAT_REGIME	Custom string[1]	<input type="checkbox"/>	1 1 1 1

Import

Figure 14 – Importing a dataset – Typical import status messages.

- Change the data set name if needed by typing in the field.
- Select the field delimiter (Figure 15). If the wrong column separator is selected, the application will duly inform the user (Figure 16)⁴.

Delimiter

,

▼

,

:

;

TAB

:

|

Figure 15 – Importing a dataset – Selecting "field delimiter" (or "column separator").

⁴ This error message will also show up after uploading the CSV file in case the separator is different from the default value (comma, ","). Just pick another separator char in the "Delimiter" drop-down box..

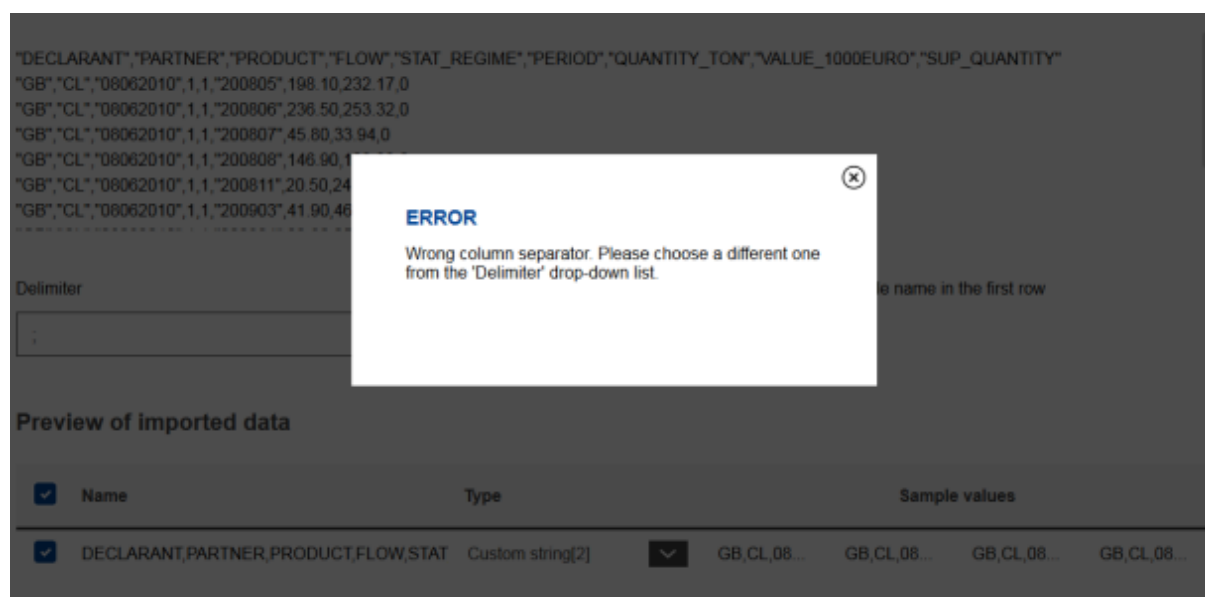


Figure 16 – Importing a dataset – Wrong column separator message.

6. Choose the columns to be imported by ticking the relevant checkboxes. All columns are selected (ticked) by default (Figure 17, #1). The checkbox in the preview header will toggle all columns on/off (Figure 17, #1)⁵;

Preview of imported data							
<input checked="" type="checkbox"/>	Name	Type	Sample values				
<input checked="" type="checkbox"/>	DECLARANT	Custom string[2]	GB	GB	GB	GB	
<input checked="" type="checkbox"/>	PARTNER	Custom string[2]	CL	CL	CL	CL	
<input checked="" type="checkbox"/>	PRODUCT	Custom string[8]	08062010	08062010	08062010	08062010	
<input checked="" type="checkbox"/>	FLOW	Custom string[1]	1	1	1	1	

Figure 17 – Importing a dataset – Toggling column import and changing column names.

7. If needed, change the individual column names by clicking on the column name: this will open a text field where the user can edit the column name (Figure 17, #3)⁶.
8. Check thoroughly the data type proposed for each column by the import wizard, and change it if needed (Figure 18). This is a **crucial step**, because choosing the proper format is essential to ensure correct handling of the data set (see section 8.1.1 for details);

⁵ Please note that at least two columns must be imported.

⁶ Only letters (A-Z, a-z) and numbers are allowed in column names. No other special chars are allowed, including blank spaces.

<input checked="" type="checkbox"/> DECLARANT	Custom string[2]	GB	GB	GB	GB
<input checked="" type="checkbox"/> PARTNER	Custom string	CL	CL	CL	CL
<input checked="" type="checkbox"/> PRODUCT	Date(dd-MM-yyyy)	08062010	08062010	08062010	08062010
<input checked="" type="checkbox"/> FLOW	Date(dd/MM/yyyy)	1	1	1	1
<input checked="" type="checkbox"/> STAT DE/IME	Date(yyMM)				
	Date/yyyy)				
	Date/yyyy-MM-dd 00:...				
	Date/yyyyMM)				
	Decimal				
	Integer				
	String[255]				
	String[2]				
	String[50]				

Figure 18 – Importing a dataset – Selecting the proper data type for each column.

- Carefully review your settings and click the import button. A confirmation message (Figure 19) will show up asking the user to double-check some parameters (column names in particular). Click “Yes” if you are sure all the settings have been verified.

Confirm

Please note: some statistical procedures have limits on the length of categorical (i.e. string) variable names, which must not exceed 8 (eight) chars in length. Please make the necessary adjustments before importing.

Are you sure you want to import this DataSet?

Figure 19 – Importing a dataset – Import confirmation pop-up dialog.

- The actual import will now take place. The actual import time will depend on the size and complexity of your data set. If the import is successful, a message will be displayed accordingly (Figure 20).

<p>Please wait</p> <p>Importing data set...</p>	<p>Status</p> <p>File successfully imported.</p> <div style="text-align: right; margin-top: -20px;"> X </div>
--	---

Figure 20 – Importing a dataset – Import status messages.

- When the “File successfully imported” pop-up box is finally displayed, close it down by clicking the “X” button on the top right (Figure 20, right). The newly imported data set will be now available for use with every statistical procedure available in WebAriadne.

Important note: data sets are immutable once imported. No data can be added to or removed from the dataset, nor can column names or data types be changed.

8.1.1 Column data formats

WebAriadne offers twelve different data formats:

- One for integer numbers ("Integer");
- One for decimal (floating-point) numbers ("Decimal");
- Four string formats, three of which have predefined lengths (2, 50 and 255 chars: "String[2]", "String[50]" and "String[255]") and one custom-length string ("Custom String");
- A total of six date formats: "dd-MM-yyyy", "dd/MM/yyyy", "yyMM", "yyyy", "yyyy-MM-dd 00:00:00", "yyyyMM"), where "y" is one year digit, "M" one month digit, and "d" one day digit. "yyyy" means a four-digit year, "MM" a two-digit month, etc. Please note that only the date info will be stored, while the time of day (hours, minutes and seconds) is discarded.

Utmost care should be taken in setting each column type, which could affect the way the data set is processed by each statistical procedure. Therefore, the user must have some knowledge of the meaning and type of the data being imported.

Generally speaking, strings represent either free-text or, most frequently, *categorical* variables, which can take any value from a more or less predefined set of string values. Two examples of categorical variables for COMEXT data are:

- Origin and destination countries, whose values are taken from the ISO3166-1 two-char codes;
- Product codes, whose values are picked from the 8- or 10-digit Eurostat TARIC product nomenclature.

Two number types (integer and decimal) are sufficient to cover virtually all needs, but this is not the case with strings, which can widely vary in length. Too short of a string length will result in data truncation, whereas too long of a string could result in unnecessary space occupation and possibly slower performance.

Since putting too many individual entries with different string lengths in the list would have resulted in a "messy" drop-down box, WebAriadne features a custom string option that can be used to set the length of a string column as shown in the figure below:

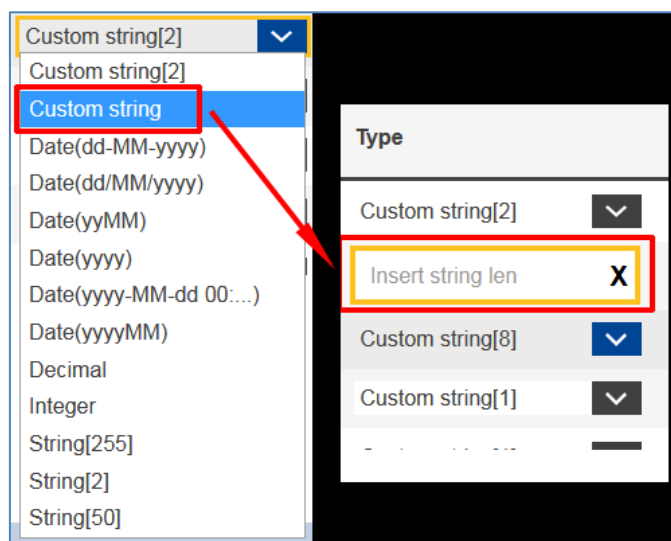


Figure 21 – Importing a dataset – Setting custom string length.

If "Custom string" type is picked from the drop-down list, a textfield will appear in the column type allowing the user to type in the desired string length. Valid lengths are between 1 and 1024 chars. Please use high values sparingly as this could significantly affect performance. After the selection is confirmed, the value chosen will show up in the drop-down box in the corresponding column as "Custom string[chosen length]", e.g. "Custom string[125]".

8.2 Managing existing datasets

WebAriadne offers a dataset management list view ("Your datasets", Figure 22) which is accessible through the "Dataset">"Manage datasets" menu item.




Your datasets			
Name	Import date	Records	
dmedium2	19/12/2019-12:14:42	714695	
DataSample	18/09/2019-14:50:31	487	
dmedium	04/12/2019-16:57:46	714695	
<div>Delete</div>			

Figure 22 – Importing a dataset – Dataset management view.

For each imported data set the following information is displayed:

- Name of dataset (“Dataset name” field in the import wizard);
- Import date and time;
- Number of records in the data set (also known as “lines” or “observations”).

Each line in the list view contains an information (“i”) icon on the far right. Clicking this icon will bring up the dataset preview box (Figure 23), showing the name, type and some sample values for each column of the dataset. This view is especially useful to recall the structure of the dataset prior to using it with any procedure available in WebAriadne.

Preview: dmedium2					
Column name	Type	Sample values			
DECLARANT	Custom string[2]	GB	GB	GB	GB
PARTNER	Custom string[2]	CL	CL	CL	CL
PRODUCT	Custom string[8]	08062010	08062010	08062010	08062010
FLOW	Custom string[1]	1	1	1	1
STAT_REGIME	Custom string[1]	1	1	1	1
PERIOD	Date(yyyyMM)	2008-05-01	2008-06-01	2008-07-01	2008-08-01
QUANTITY_TON	Decimal	400.4	300.5	45.0	440.0

Figure 23 – Managing datasets – Dataset preview dialog box.





Your datasets			
Name	Import date	Records	
dmedium2	19/12/2019-12:14:42	714695	
DataSet	18/09/2019-14:50:31	487	
dmedium	04/12/2019-16:57:46	714695	
			

Figure 24 – Managing datasets: selecting a data set for deletion.

Whenever a dataset row is selected, the “Delete” button will become enabled, as shown in Figure 24. Clicking the “Delete” button will prompt the user for confirmation (Figure 25). If deletion is successful, the user will be duly notified, as shown in Figure 26.

Confirm

Please note that this operation cannot be undone Are you still sure you want to delete the selected dataset?

Yes

No

Figure 25 – Managing datasets: deleting a data set.

Status

Dataset deleted

Figure 26 – Managing datasets: successful deletion message box.

Warning: since there is a direct link at application level between a data set and all result sets produced by running that dataset through any available procedure, **deleting a data set will wipe out all result sets deriving from it. Deleted data sets and result sets thereof cannot be recovered in any way.**

9 Running statistical procedures

9.1 Introduction

WebAriadne offers a dedicated user interface for each available statistical procedure. The use workflow is essentially the same for all statistical procedures (see section 6 for details).

WebAriadne has two types of statistical procedure interfaces:

12. One “wizard-like” interface type, where the user selects (or imports) a dataset, and then sets all parameters through a series of guided screens, each having a handful to set and a “Next”/“Back” button pair for easy navigation. Once all parameters have been set, the user can launch the application by clicking the “Run” button.
13. One “old-fashioned”, legacy interface type, where the user can still select (or import) the dataset to be processed, but all the parameters (up to a few dozens) and the “Run” button are displayed in one single view.

Type 1 interfaces are more user-friendly and more intuitive to use, because pages with too many parameters and settings can easily become cluttered up and confusing. WebAriadne has a handful of legacy type 2 interfaces, but all recently added statistical procedures and all procedures that will be added in the future have (or will have) a type 1 interface.

9.2 Generalized Benford’s law tool

The Generalized Benford’s law tool can be accessed via the “Applications” menu or the link on the dashboard. The generalized Benford’s law tool GUI has three “pages”, or “screens”.

In first page (Figure 27), the user can select an existing dataset through the “Select” drop-down box, or import a new one by clicking the “Import” button. The “Import” button will bring up the import wizard, which can be used as described in section 8.1. The only difference being the presence of a “Back” button on the top left (Figure 28).

General Benford's law analysis

1/3 Load dataset

Select an existing dataset or upload a new one.

Select*

Select a dataset ▼ Import

Date created:
Size records:
Source:
Preview:

Next

Figure 27 – Generalized Benford’s tool: first page

Figure 28 – The import wizard showing a “Back” button when invoked from the individual statistical procedures

Figure 29 – Generalized Benford’s tool: first page with a data set selected

The “Back” button takes the user back to the Benford tool, with the newly imported data set already selected. Additionally, some info on the dataset is displayed on the right-hand side of the page, with a link (“Preview: open”) that will bring up the dataset preview dialog box (Figure 23). The “Next” button also becomes enabled.

Figure 30 – Generalized Benford’s tool: second page

Clicking the “Next” button takes to the Benford's law tool's second page (Figure 30). This page contains two drop-down lists (“Grouping categories” and “Target variables”) which are set to “None” and “All” by default. These two drop-down are meant for grouping variables, which will be added in a future version. For the moment, just click “Next” to come to the tool's third page.

The third page contains three controls:

1. "Applicability checks": drop-down list which is pre-set to "None" and will be implemented in a future version;
2. "Statistical tests": drop-down list through which the user can select the type of Benford compliance tests to be performed on the data set's variables:
 - (a) "Standard test for first and second digits": standard Benford's law compliance test;
 - (b) "Standard test for first and second digits": two-stage Benford's law compliance test with simulations. The latter is more accurate but takes much longer to compute (by more than an order of magnitude)
3. "Confidence level": drop-down list for setting the confidence level. Four predefined values are available: 0.01, 0.05, 0.10 and 0.20.

Finally, once all the parameters have been set, press "Run" to run the Benford's law tool.

After hitting the "Run" button, a dialog box will show up asking the user whether they wish to go back to page 1 to run the Benford's law tool on another data set, or they prefer to be taken to the application dashboard, as shown below:

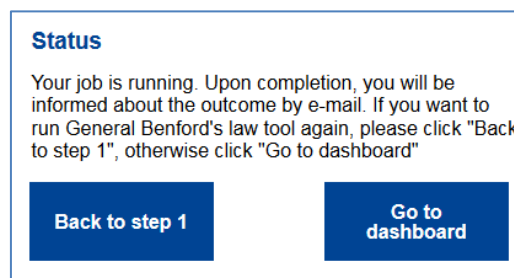


Figure 31 – Generalized Benford's tool: option pop-up dialog box after running the statistical procedure.

9.3 Benford's law for customs

WebAriadne has another Benford's law tool targeted at analyzing trade data: therefore, this tool is only meant for analysts and specialists in customs and trade fraud and prevention.

This version of the Benford's law tool consists of one single view, where the data set and all the parameters must be selected prior to running the application (Figure 32).

Benford's law for Customs

Dataset

Select*

Select a value

Import

Date created:
Size records:
Source:
Preview:

For combination of

Product*

Trader/Destination*

Value*

Weight*

Supplementary unit variable
☐

Supplementary quantity

Advance setting (optional)

Max traders

100000

Max transition for trader (N)

1000

Max products for trader(M)

500

Run

Figure 32 – Benford’s law for customs: user interface page

With reference to Figure 32, here is a more detailed description of all available parameters (from left to right and from top to bottom):

- “Data set” selection or import: select an existing data set or import a new one;
- “Product”: drop-down list for selecting the variable representing the product in the data set;
- “Trader/Destination”: drop-down list for selecting the variable representing the trader **or** the destination;
- “Value”: drop-down list for selecting the variable representing the **value** of transactions;
- “Weight”: drop-down list for selecting the variable representing the weight of traded goods;
- “Supplementary unit variable”: tick this box if the figures goods traded are expressed in terms of units traded instead of weight;
- “Supplementary quantity”: this drop-down list is normally grayed out, and will be enabled only if the “Supplementary unit variable” checkbox is ticked. Use this drop-down list to pick the variable representing the supplementary quantity of goods traded;

All the controls marked by an asterisk are mandatory, and the same variable cannot be selected twice.

The product, trader/destination, supplementary unit variables are categorical (i.e. string) variables, while value, weight and supplementary quantity are numbers (either decimal or integer).

Three advanced parameters are also available:

- “Max traders”: threshold on the number of traders processed. That is, the procedure will process the first N traders (default: 100000).
- “Max transactions per trader (N)”: discard all traders with more than N transactions (default: 1000). This can be useful for discarding traders with an excessive number of transactions (e.g. big e-commerce companies such as Amazon);
- “Max products per trader (M)”: discard all traders with more than M distinct products traded (default: 500). Can be used to selectively exclude some traders which are not significant despite having a high number of transactions or individual products traded.

Just change the value of the three advanced parameters by typing in the new value. These parameters do not normally need to be changed, unless in very special cases.

The “Run” button will become enabled when all the mandatory parameters have been set.

9.3.1 Worked example of the Benford’s law for customs

Let’s consider a dataset containing the following variables:

- “product”: code of the traded product (TARIC CN or other alphanumeric code)
- “id_trader”: the id of the trader involved in the transaction. The code can be any numerical or alphanumerical unique id for a trader, provided that it does not allow the user to derive any of the trader’s personal data (e.g. name or company name, VAT ID, etc). **Use of non-anonymized data in WebAriadne is prohibited by the applications terms and conditions of use.**
- “value”: value of transaction (in euros or other currency)
- “net_mass”: weight of the units. Supplementary units, if provided, can also be used.

We decide to leave the other parameters untouched. After choosing our settings., the user interface will look like the following figure:

Benford's law for Customs

Dataset

Select*

2014_50K

Import

Date created: 31/01/2020 16:25:41
Size records: 47439
Source: N/A
Preview: [open](#)

For combination of

Product*

product

Trader/Destination*

id_trader

Value*

value

Weight*

net_mass

Supplementary unit variable
☐

Supplementary quantity

Select a value

Advance setting (optional)

Max traders

100000

Max transition for trader (N)

1000

Max products for trader(M)

500

Run

Figure 33 – Benford's law for customs: user interface page with parameters selected

Once the user is happy with the parameter chosen, they can run the statistical procedure pressing the "Run" button.

9.4 Robust Regression Outliers

Robust Regression Outliers

Dataset

Select*

Select a value

▼

Import

Date created:

Size records:

Source:

Preview:

For combination of

Variable 1

▼

☒ Each
☐ All

Variable 2

▼

☒ Each
☐ All

Variable 3

▼

☒ Each
☐ All

Independent variable*

▼

Dependent variable*

▼

Trade data

☐

Time variable

▼

Calculation method*

Select a value

▼

Alpha

0.10

Intercept

☐

Select

▼

Run

Figure 34 – Robust Regression Outliers: user interface page

Robust Regression Outliers (Figure 34) is a procedure for performing regression between two variables (“X”, or independent, and “Y”, or dependent) using robust statistical techniques. “Robust” means that these algorithms can single out anomalous values in a population and discard them for performing calculations, thus yielding better quality results than traditional techniques.

Robust Regression Outliers is a procedure with many controls all in one page. These controls are divided in different sections:

- Dataset area, where the user can select an existing data set or import a new one (will call the import wizard);
- “For combination of” area, where all the main parameters can be set;
- “Additional variables” section, where the user must specify the so-called *additional variables*.

Although targeted primarily at trade data, Robust Regression Outliers can be used to analyze any type of data having two numerical variables (X, the so-called "independent variable") and Y (the so-called "dependent variable") and one or more grouping categorical variables.

9.4.1 Parameters for Robust Regression Outliers

Here are the main controls for the Robust Outliers procedure:

- "Variable 1", "Variable 2" and "Variable 3": these are the so-called grouping variables. Depending on the user's selection, the data set will be divided into subgroups of different size. A robust regression will be performed for each subgroup. One can choose from no grouping variable by leaving the three drop-down boxes unset) up to three grouping variables. Each drop-down box will be populated with the names of the categorical (i.e. string) variables from the currently selected dataset. The more the grouping variables, the smaller the subsets.
- "Each/all" buttons: these buttons will determine the way the Bonferroni corrections are calculated.
- "Independent variable": the "X" variable used for performing the robust regressions. This drop-down box will be populated with the list of purely numerical variables available in the currently selected data set.
- "Dependent variable": the "Y" variable used for performing the robust regressions. This drop-down box will be populated with the list of purely numerical variables available in the currently selected data set.
- "Trade data": tick this box when analyzing trade data. This will turn on some calculations specific for trade data analysis and will enable the "Time variable" drop-down box (see below).
- "Time variable": this drop-down box will be enabled only if the "Trade data" checkbox is ticked. Use this drop-down list to pick the time variable.
- "Calculation Method": pick the robust regression algorithm to be used for performing the robust regressions. The following options are available: "Backward Search", "Least Mean of Squares", "Least Trimmed Squares", "S (Rousseeuw-Yohai)", "M: Huber (1981)", "MM: Yohai (1987)", "Forward Search". Each robust regression algorithm has its advantages and disadvantages. Explaining the details is beyond the scope of this user's manual: please refer to the relevant documentation and literature.
- "Alpha": significance level. Decimal number between 0 and 1. Defaults to 0.10, typical values are 0.20, 0.10, 0.05, 0.01.
- "Intercept": tick this box if you want the intercept to be calculated for the regressions.

9.4.2 Additional variables

This section contains a multiple-selection drop-down list through which the user can select multiple categorical variables from the ones making up the data set. Additional variables **must** be set, and all those categorical variables not used as grouping variables must be selected as additional variables. Since the current version allows up to three grouping variables, we have four possible cases:

1. If no (zero) grouping variables were set, then three additional variables must be set
2. If one grouping variable was set, then two additional variables must be set
3. If two grouping variables were set, then one additional variable must be set
4. If three grouping variables were set, then no (zero) additional variables must be set

Each data set variable can be selected only once, either as a grouping variable or as an additional variable.

A worked use case will illustrate the use of grouping variables and additional variables.

9.4.3 A worked example for Robust Regression Outliers

Let's have a dataset in the COMEXT CSV format (see section 4.2 for details) having about 50000 "observations" (synonym for "lines" or "records") and a header row. Let's also suppose we pick only one grouping variable, "PRODUCT" and we select all default/typical values for the other parameters. The Robust Regression Outliers execution interface **just prior to setting the additional variables** will look like the following:

Robust Regression Outliers

Dataset

Select*

dsmall

▼

Import

Date created: 20/12/2019 12:45:58
Size records: 49999
Source: N/A
Preview: [open](#)

For combination of

Variable 1

PRODUCT

▼

☒ Each
☐ All

Variable 2

Select a value

▼

☒ Each
☐ All

Variable 3

Select a value

▼

☒ Each
☐ All

Independent variable*

QUANTITY_TON

▼

Dependent variable*

VALUE_1000EURO

▼

Trade data
☒

Time variable

PERIOD

▼

Calculation method*

LTS: Least Trimmed Squares

▼

Alpha

0.10

Intercept
☐

Additional variables

Select

Select a value

▼

Run

Figure 35 - Robust Regression Outliers' GUI before picking the additional variables

There is no default or typical setting for the calculation method: the selection must be made knowing beforehand the advantages or disadvantages of each method. In this case we chose LTS because it is reasonably fast, but not the most accurate one.

Since we picked one grouping variable, "PRODUCT", but the data set has two other variables ("PARTNER" and "DECLARANT") that could be used for dividing the dataset into subgroups. We chose not to use these two variables as grouping, but we **must** select them as additional variables. The additional variables drop-down list will look like the following:

Figure 36 - Drop-down list for selecting the additional variables.

The drop-down list will show the list of all variables from the selected dataset, with the ones that have already been selected in the interface being greyed out. Let's pick

"PARTNER" and "DECLARANT" as additional variables. The interface will now show the list of additional variables selected:

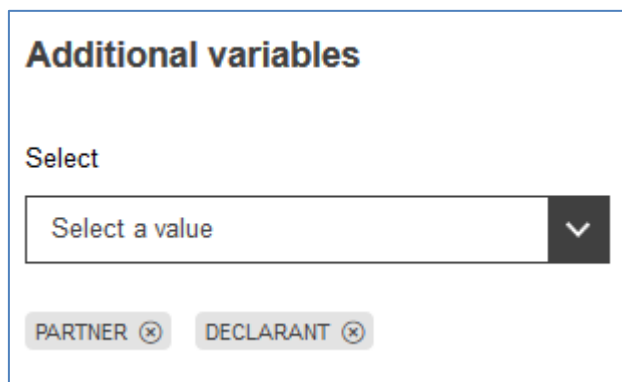


Figure 37 – Additional variables selection interface after picking two variables

The selected variable will be listed just below the selection box. The two variables selected will be now greyed out (and therefore no longer selectable) in the drop-down list:

EMBED Photoshop.Image.14 \s

Figure 38 – The additional variables selected are greyed out in the drop-down list

To deselect an additional variable, click the "X" that is located to the right, next to each selected variable name:

Figure 39 – Click the "X" icon to deselect an additional variable

Please note: the selection of grouping variables takes precedence over the selection of additional variables. Therefore, if an already selected additional variable is selected as a grouping variable, it will be automatically removed from the list of selected additional variables.

Once you are happy with your selection, the user interface will look like depicted in Figure 40, and you can finally click the yellow "Run" button to run the procedure.

After pressing the "Run" button a status pop-up message will appear such as the one shown in Figure 41.

Robust Regression Outliers

Dataset

Select*

dsmall

Import

Date created: 20/12/2019 12:45:58
Size records: 49999
Source: N/A
Preview: [open](#)

For combination of

Variable 1

PRODUCT

☒ Each
☐ All

Variable 2

Select a value

☒ Each
☐ All

Variable 3

Select a value

☒ Each
☐ All

Independent variable*

QUANTITY TON

Dependent variable*

VALUE 1000EURO

Trade data
☒

Time variable

PERIOD

Calculation method*

LMS: Least Mean of Squares

Alpha

0.10

Intercept
☐

Additional variables

Select

Select a value

PARTNER ✕ DECLARANT ✕

Run

Figure 40 – Robust Outliers user interface with all the parameters selected

✕

Status

Your data are being processed.
You may logout now or do another analysis.
As soon as your results are ready you will be notified by e-mail.

Figure 41 – Pop-up message shown after pressing the "Run" button

Since the processing time is not predictable, the user will be notified by e-mail. WebAriadne's dashboard has a section showing the status (running, completed, error) of processing jobs for the currently authenticated user.

When the job is completed, the user will receive an e-mail message akin to the one shown in Figure 42 below. Actual content might vary depending on the application version and on the e-mail client used.

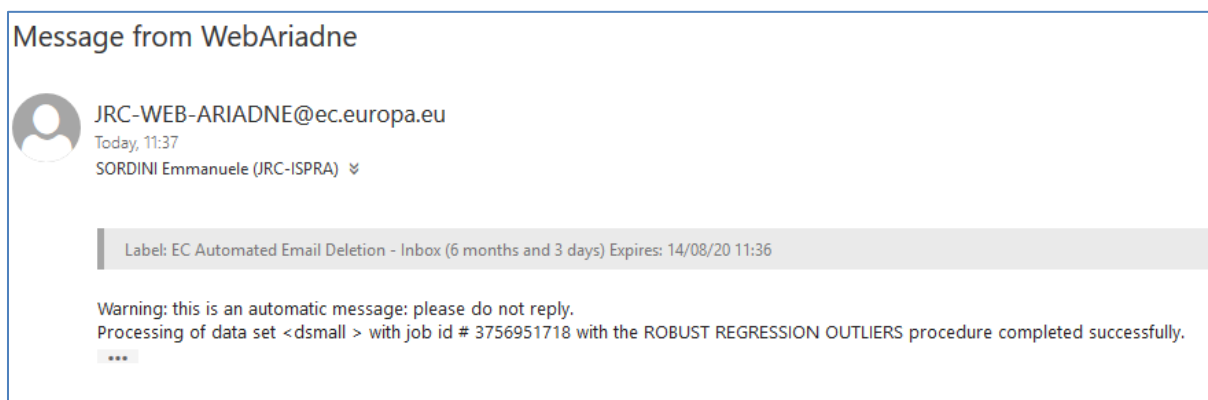


Figure 42 – E-mail message informing the user of the successful completion of a job

10 Viewing and exporting results

10.1 Main result view

WebAriadne offers a result viewing and exporting section. This section can be accessed through the application menu ("Results" > "All results"). The application menu also offers a shortcut to the result produced by the last completed job ("Results" > "Last result").

In addition to the drop-down menus, the application dashboard will show in the bottom right section the last five results, listed in reverse chronological order. Each entry is a link which will open directly the detailed view for that result set.

The main result set view is shown in Figure 43.











All results					
Job ID	Date created	Application	Methods	Dataset	
3756951718	13/02/2020 11:36:11 +0100	ROBUST	LTS	dsmall	
30544134382	20/12/2019 12:28:54 +0100	BENFORD_GEN	N/A	DataSample	
29179113478	04/12/2019 17:20:30 +0100	ROBUST	LTS	dmedium	
29178602306	04/12/2019 17:10:29 +0100	BENFORD_GEN	N/A	DataSample	
27172913243	11/11/2019 12:03:49 +0100	BENFORD_GEN	N/A	DataSample	
27172392009	11/11/2019 11:53:14 +0100	BENFORD_GEN	N/A	DataSample	
23106675947	25/09/2019 11:31:17 +0200	BENFORD_GEN	N/A	DataSample	
22513846252	18/09/2019 14:50:46 +0200	BENFORD_GEN	N/A	DataSample	
<div> < Page <input type="text" value="1"/> of 1 >  1 - 8 Results of 8</div> <div>Import as new dataset Delete</div>					

Figure 43 – Main result set view

The main result view contains the list of all available result sets produced by the current user. Each row refers to a single result set and contains the following information (from left to right):

- Job id (the same displayed in the notification e-mail);
- Date and time of creation of the result set, which is equivalent to the date and time of job completion;
- Statistical procedure/application with which the dataset was processed;
- Methods/algorithms used to process the dataset (applies to some statistical procedure only)
- Name of the dataset used for processing;
- "i" icon/button. Clicking this button will open the detail view for that result set. Each detail view will look differently depending on the statistical application.

The result set view is sortable by clicking on each header name ("Job ID", "Date created", etc.). Clicking on the header name will toggle between ascending and descending order. The view can be sorted according to one column at a time.

Should the number of result sets exceed the maximum number of entries displayable in the list, the view will be paged. Pagination controls are available on the bottom left corner of the view.

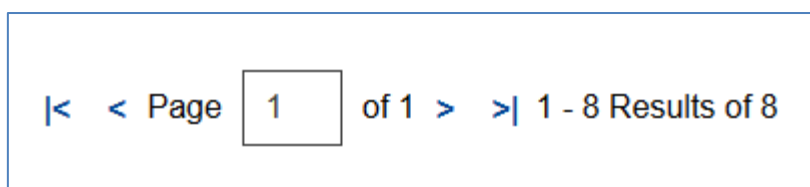


Figure 44 – Pagination controls for the main result set view

The pagination control group for the main result set view (Figure 44) contain the following controls from left to right:

- Arrow button for jumping to the first page;
- Arrow button for going back to the previous page;
- Text field for jumping to a specific page (type in the page number + press Enter);
- Arrow button for advancing to the next page;
- Arrow button for advancing to the last page.

10.2 Detailed result view for the Generalized Benford’s law tool



All entries in the main result set view with the “BENFORD_GEN” tag are produced running a data set through the Generalized Benford’s law tool. Clicking the “I” icon for any of these entries will open the detailed view for that job. The result presentation section is made up of two sections:

- Top section, containing the list of the source dataset’s variables with an overview of Benford’s law compliance
- Bottom section, containing the statistical details for the selected variable from the source data set.

General Benford's law analysis			
<p>Top table: for each variable the result of 3 Benford's law tests. Bottom table: for the selected variable in the top table, some details on the test selected in the dropdown (First digit, second, first and second)</p>			
<input type="text" value="Search column"/>			
Column name	First digit	Second digit	First and second digits
rmua	×	×	×
sba	✓	×	×
sbab	×	✓	×
sbadc	✓	×	×
sbaii	✓	✓	✓
sbbi	✓	×	×
sbc	×	×	×
sbc2	×	×	×
sbc3	×	×	×
sbcco	✓	×	×
sbcl	✓	×	×
sbda	✓	×	×
sbdai	×	×	×
sbdco	✓	×	×
Descriptive information about the analysis of rmua			

Figure 45 – Generalized Benford’s law: source data set column view

The top section (Figure 45) contains the list of the variables (columns) from the source data set. For each of them, the Benford’s compliance for the first significant digit, the second digit and the first and second digit

together is displayed. A green tick sign  indicates that the corresponding column is Benford-compliant, while a red “X”  denotes Benford non-compliance.

A “search column” textbox in the upper section makes it possible to filter the view by typing in some column name substring, while the “Results” button on the upper left will take back to the result list view.

If any variable row is selected (see Figure 45 with “mua” column selected) it will be highlighted in yellow, and the bottom half of the view will be populated with the details for that variable, like in the following figure:

Descriptive information about the analysis of rmua		
First digit	Second digit	First and second digits
Number of observations used in the analysis n		50254
Number of missing value		0
Number of zeroes		4104
Number of negative values		4
Minimum value		0.001
Maximum value		29676888
Data range (log10 scale)		10.472418358
χ^2 statistic		1371.4491000452613
Threshold for χ^2 statistic		20.09023502966324
P-value of the χ^2 statistic		0

Figure 46 – Generalized Benford’s law: source data set column view

The descriptive information section contains several indicators⁷, such as:

- Number of observations used in the analysis;
- Number of missing values;
- Number of negative values;
- Minimum and maximum, and a data range in Log10 scale;
- Value of the X^2 statistic with the relevant threshold and p-value.

Three separate tabs are available, respectively providing the above indicators for (from left to right):

- The first significant digit;
- The second significant digit;
- The combination of the first and second digit.

10.3 Detailed result set view for Benford’s law for customs

All entries in the main result set view with the “BENFORD” tag are produced running a data set through the Benford’s law for customs. Clicking the “I” icon for any of these entries will open the detailed view for that job.

The detailed view for the Generalized Benford’s is divided into two sections:

- Top section containing the list of traders filtered by the statistical procedure (Figure 47)
- Bottom section, containing the detailed view for the selected trader (Figure 48).

⁷ Please note that a detailed description of these indicators is beyond the scope of this user’s manual. Please refer to the relevant documentation and literature.

Results

RUN 5340064164

Result of 02/03/2020 19:21:11 - Procedure: BENFORD - Source DS:sample_50K_2

12345678910

Trader ID	Transactions	Products	Value Range	Repeated values	Risk Index	Second digit
6	432	80	5.3844	1.16	10	
9	509	22	4.2012	1.18	9	
13	135	68	4.1659	0.00	4	
17	58	44	2.1967	0.00	3	
8	130	25	3.5657	0.00	2	

Figure 47 – Benford’s law for customs: top section

The top section provides a list view containing the most statistically relevant traders, which are listed in descending order of risk index⁸. Risk indexes range from 1 to 10 and are color-coded, with red being the highest (10) and violet the lowest (1). The following information is displayed for each trader:

1. Trader’s unique id;
2. Total number of transactions for trader in the current data set;
3. Total number of distinct products traded by the trader;
4. Transaction value range for that trader⁹;
5. Proportion (%) of repeated values for that trader;
6. Risk index (as described above);
7. Second-digit flag. An asterisk marks a second-digit anomaly (non-Benford compliance) in transaction values.

The “Search trader” field at the top of the dialog box makes it possible to narrow the search down by typing in some trader name/id substring. Clicking the “Results” button on the top left will take back to the main result list view.

Each row can be selected to open the detail for that trader which is displayed in the bottom section of the page (Figure 48). The background of the currently selected row will turn yellow.

⁸ The risk index is based on the p-value. The lower the p-value, the higher the risk index.

⁹Calculated on a logarithmic scale, i.e. $\text{Log10}[\text{max}/\text{min}]$

Descriptive information about the trader: 6

Product ID	Number of transaction	Estimated trader price	Estimated market price	Price deviation	Market share	
8509900000	81	15.41	15.41	0.00	100.00	
8516900099	50	12.35	12.35	0.00	100.00	
8509400000	48	6.29	6.29	0.00	100.00	
8516400000	29	9.76	9.76	0.00	100.00	
8516797090	26	5.23	5.23	0.00	100.00	
8516710000	25	6.82	6.82	0.00	100.00	
8509800000	19	13.90	13.90	0.00	100.00	
8508110000	10	4.71	4.71	0.00	100.00	
8516299100	10	4.97	4.97	0.00	100.00	
8516609000	7	4.15	4.15	0.00	100.00	
8415900090	7	22.04	23.22	5.09	99.98	
3926909790	6	11.88	30.48	61.04	96.49	
8508700090	6	14.58	14.58	0.00	100.00	
8501310099	5	5.24	5.24	0.00	100.00	

Download in Excel format

Figure 48 – Benford’s law for customs: bottom section (trader detail view)

The following information is displayed in the trader detail view. In this context, the “market” means the data set which produced the signals.

1. Product id;
2. Number of transactions in which that product was traded by the trader under consideration;
3. Estimated trader price (per unit of weight or supplementary unit, depending on the case);
4. Estimated “market” price per unit of weight or supplementary unit. The “market” refers to the current data set
5. Trader vs. market price deviation, i.e. $100 \times (\text{market_price} - \text{trader_price}) / \text{market_price}$;
6. Market share (%) for the currently selected trader and product;
7. Button for the transaction detail plot.

The view can be sorted according to columns 1-6, one column at a time by clicking on the column header. Repeated clicks will toggle between ascending and descending order.

Users can download the data for the trader under consideration by clicking the “Download in Excel format” button.

Product ID	Number of transaction	Estimated trader price	Estimated market price	Price deviation	Market share	
8509900000	81	15.41	15.41	0.00	100.00	Plot icon
8516900099	50	12.35	12.35	0.00	100.00	Plot icon
8509400000	48	6.29	6.29	0.00	100.00	Plot icon
8516400000	29	9.76	9.76	0.00	100.00	Plot icon
8516797090	26	5.23	5.23	0.00	100.00	Plot icon
8516710000	25	6.82	6.82	0.00	100.00	Plot icon
8509800000	19	13.90	13.90	0.00	100.00	Plot icon
8508110000	10	4.71	4.71	0.00	100.00	Plot icon
8516299100	10	4.97	4.97	0.00	100.00	Plot icon
8516609000	7	4.15	4.15	0.00	100.00	Plot icon
8415900090	7	22.04	23.22	5.09	99.98	Plot icon
3926909790	6	11.88	30.48	61.04	96.49	Plot icon
8508700090	6	14.58	14.58	0.00	100.00	Plot icon
8508190000	5	2.82	2.82	0.00	100.00	Plot icon

Figure 49 – Benford’s law for customs: trader detail view with plot button column highlighted

The scatter plot for each product can be opened by clicking the light blue plot icon at the far right of each row (Figure 49). A plot will be displayed containing all the transactions for that product in the “market” (i.e. data set) at hand: please refer to Figure 50. Quantities are displayed along the X axis, and values along the Y axis. Transactions performed by the currently selected trader are marked by red dots, while transactions performed by other traders are marked by blue dots. If the current trader’s maker share is 100%, the plot will only contain red dots. The plot window can be closed down by clicking the “X” icon at the top right corner.

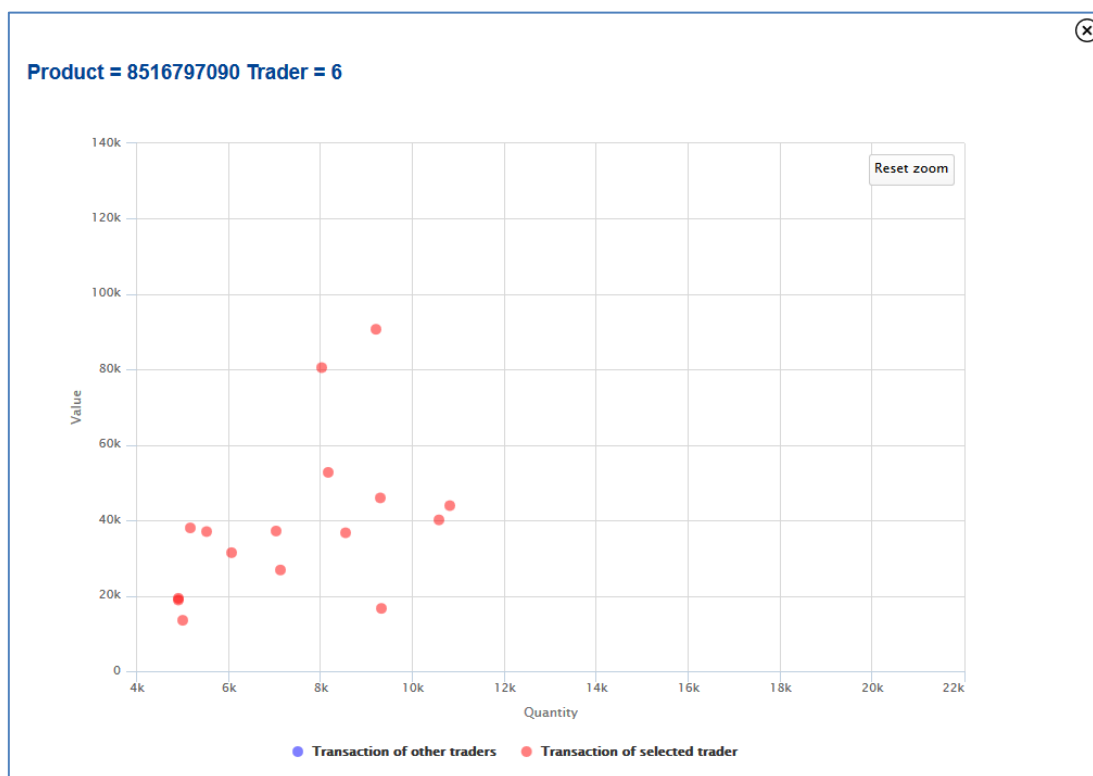


Figure 50 – Benford’s law for customs: selected trader/product scatter plot in zoom mode

The plot offers some degree of interaction:

- Hovering with the mouse on each data point will display the current values (quantity and value);
- Clicking and dragging the mouse on the plot makes it possible to zoom in any portion of the area. When in zoom mode, a “Reset zoom” button will show up on the top right.

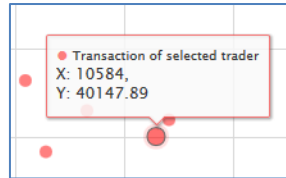


Figure 51 – Benford’s law for customs: data point mouseover tooltip

10.4 Detailed result view for Robust Regression Outliers

Go to the result list view and search for your Robust Regression Outliers result set with the desired Job id, then click the “I” icon. This will bring up the result detail view (Figure 52):

Results

RUN 3756951718

Result of 13/02/2020 11:36:11 - Procedure: ROBUST - Source DS:dsmall

PRODUCT	PARTNER	DECLARA...	QUANTIT...	VALUE_1...	PERIOD	NUMGOOD...	FINAL_B...	FINAL_R...	YRESID_...	STDR_FL...	YSTUDEN...	LOG_PVA...
0806209...	CL	GB	747.5	1115.87	2008-11...	34.0	1.19313...	0.99207...	224.002...	49.9270...	4.48660...	-10.089...
0806209...	CL	GB	580.0	952.91	2009-02...	34.0	1.19313...	0.99207...	260.892...	49.1535...	5.30770...	-12.501...
0808108...	CL	GB	8185.9	7598.24	2008-05...	20.0	0.86069...	0.99910...	552.656...	145.976...	3.78593...	-7.3783...
0808108...	CL	GB	7518.2	7118.64	2008-06...	20.0	0.86069...	0.99910...	647.744...	144.709...	4.47617...	-8.9525...
0808108...	CL	GB	8990.1	9224.21	2008-07...	20.0	0.86069...	0.99910...	1486.45...	147.628...	10.0688...	-19.866...
0809209...	CL	GB	503.4	2364.06	2007-12...	13.0	2.56073...	0.97974...	1074.98...	244.863...	4.39015...	-7.7281...
0809209...	CL	GB	421.2	2187.55	2009-12...	13.0	2.56073...	0.97974...	1108.96...	243.104...	4.56170...	-8.0273...
0809301...	CL	GB	487.0	998.28	2008-01...	11.0	1.21842...	0.99719...	404.905...	41.0693...	9.85907...	-13.915...
0809301...	CL	GB	1380.2	2410.82	2008-02...	11.0	1.21842...	0.99719...	729.145...	48.8163...	14.9365...	-17.821...
0809301...	CL	GB	946.3	1472.71	2008-03...	11.0	1.21842...	0.99719...	319.711...	44.2894...	7.21869...	-11.154...
0809301...	CL	GB	935.3	1536.13	2010-02...	11.0	1.21842...	0.99719...	396.534...	44.1917...	8.97304...	-13.061...
0809309...	CL	GB	171.9	350.73	2008-01...	11.0	1.48352...	0.93336...	95.7124...	33.7315...	2.83747...	-4.7316...
0809309...	CL	GB	236.7	508.22	2008-02...	11.0	1.48352...	0.93336...	157.070...	39.4191...	3.98461...	-6.6524...
0809400...	CL	GB	4134.4	4371.77	2008-04...	13.0	1.15719...	0.99658...	-412.54...	132.164...	-3.1214...	-5.4226...
0809400...	CL	GB	3195.5	2779.0	2009-03...	13.0	1.15719...	0.99658...	-918.82...	121.807...	-7.5432...	-12.587...

< < Page 1 of 192 > >

Display 1 - 20 of 3828

Download results

Figure 52 – Robust Regression Outliers: result set detail view

The result detail view for the Robust Regression Outliers procedure contains a list of all records that were identified as outliers (also called “signals”) by the statistical procedure. For each outlier the following information are displayed in the table:

- The values of the grouping variables or additional variables set by the user (“PRODUCT”, “PARTNER” and “DECLARANT” in this example);
- The values of the independent variable (“X”) and the dependent variable (“Y”) selected by the user (“QUANTITY_TON” and “VALUE_1000EURO” in this example);
- The time variable if set (“PERIOD” in this example);
- “NUMGOODOBS”, i.e. the number of good data points (also called “observations”) in the set used for performing the linear regression;
- A number of additional statistical indicators (e.g. “final beta”, “final R²”, etc.)¹⁰, that can be used to assess the quality of the results.

The result detail view also offers the following controls:

¹⁰ Again, it is beyond the scope of this user’s manual to explain the meaning and use of such indicators.

- “Results” button to close the detail view and go back to the results list;
- “Download results” to download the result set as an Excel or CSV file depending on the result set size¹¹. The exported file will contain many more fields and indicators than displayed in the detail view;
- A set of controls (arrows, text field) to navigate the pages, akin to those available in Figure 44.

Each row is associated to a scatter plot that can be opened up by clicking the light blue plot icon than can be found on the far right.

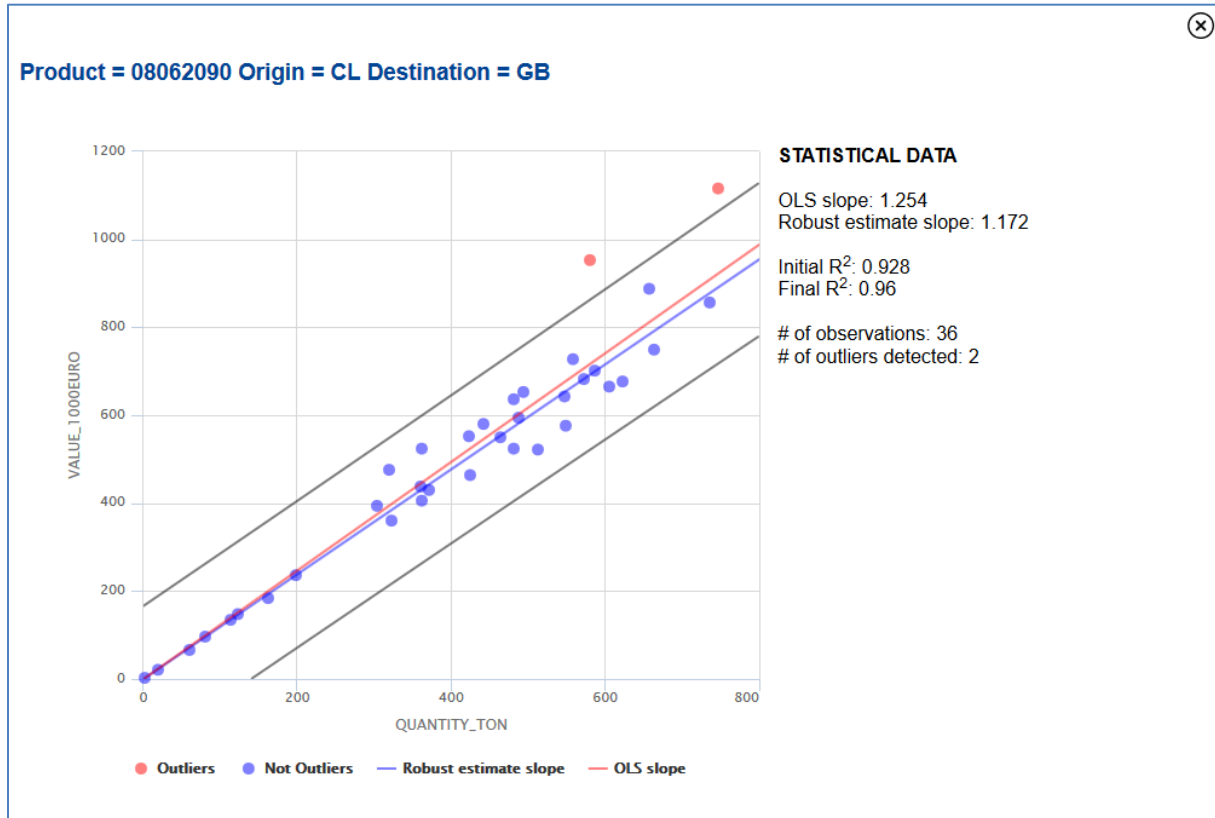


Figure 53 – Robust Regression Outliers: scatter plot

The outlier scatter plot (Figure 53) contains a lot of information:

- All observations of the group to which the selected row belongs. The size and composition of the group is determined by the grouping variables selected upon execution.
- “Good” observations (i.e. “non-outliers”) are marked by violet dots, while anomalies (“outliers”) are marked by red dots.
- OLS¹² interpolation line as a red line;
- Robust interpolation line (calculated after removing the outliers) as a violet line;
- Confidence bands (black lines)
- Some statistical indicators: OLS and robust estimate slope, initial and final R^2 , total number of observations in the group, number of outliers.

The plot offers some interaction features:

- Hovering the mouse pointer on a dot will show the values of the independent variable (“X”) and independent variable (“Y”) and the type (“Outlier”/“Non-outlier”) for that point;

¹¹ Excel format for small result sets, CSV format for big result sets.

¹² OLS = Ordinary Least Squares, a non-robust interpolation method.

- The view can be zoomed in by clicking and dragging on an area of the plot. When in zoom mode, a “Reset zoom” button will appear.
- The plot windows can be closed down at any time by clicking on the “X” icon at the top right corner.

11 Glossary

Observation, Row, Record	One individual entry in a dataset, a “tuple” of data making up a data set. Since data sets are stored as database tables in WebAriadne, one observation is precisely one row in the table.
Field, Column, Variable	One attribute of an individual observation/row/record. Can be of different type: string, integer number, decimal number, etc.
Categorical variable	String variable that can take on one of a limited, and usually fixed, number of possible values (called “categories”). Values for a categorical variable can also be numerical, but numbers are treated as strings and not as actual numbers. One such example are CN goods product codes. More info available at https://en.wikipedia.org/wiki/Categorical_variable
Signal	Observation/Row/Record that has been marked as an anomaly according to the criterion applied in a given statistical procedure. For example, signals found by running a dataset through the Robust Regression Outliers procedure are the records marked as outliers by that procedure.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: <https://publications.europa.eu/en/publications>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub
ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



EU Science, Research and Innovation



EU Science Hub



Publications Office
of the European Union