Integrated Tool for AI-assisted Exploratory Data Analysis and Quality Control of Heterogeneous Statistical Data

Eugenia Koblents and Ana Esteban

Statistics Department, Banco de España

Abstract

The BELab Data Science team at Banco de España has developed an integrated tool that standardizes Exploratory Data Analysis (EDA) and Data Quality Management (DQM) across the heterogeneous microdata sets hosted in the laboratory. Despite differences in size, structure, and confidentiality, the shared tabular format of these datasets enables a unified analytical workflow suitable for diverse users, including data producers, lab technicians, analysts, and researchers. The tool provides automated exploratory dashboards, multilevel data inspection, and an AI-powered interface for querying both aggregated and micro-level information. It also evaluates key DAMA data-quality dimensions—completeness, uniqueness, validity, and consistency—detects structural and metadata-related issues, and incorporates multivariate analysis and anomaly-detection methods to support robust and efficient data assessment.

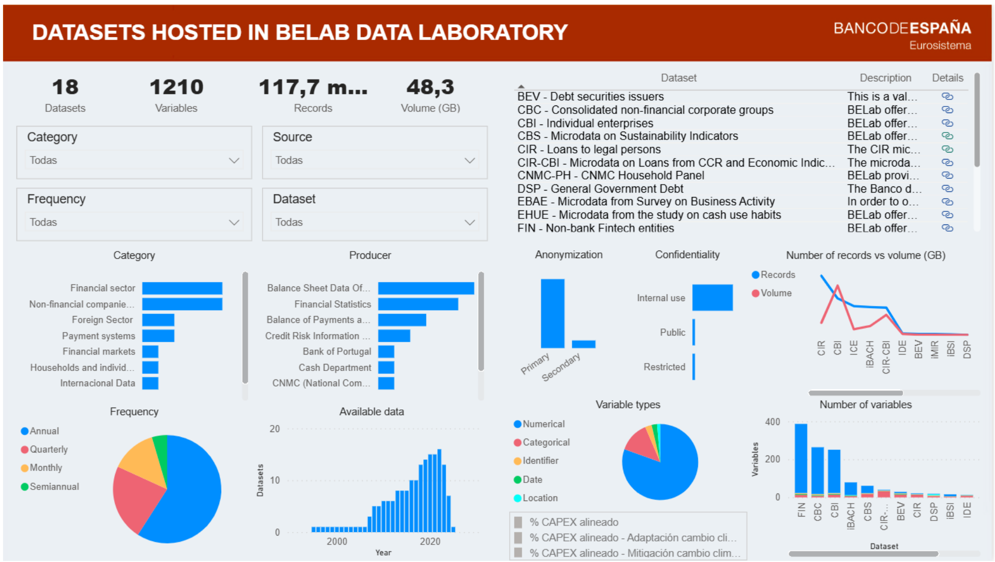
Keywords: Data Quality Management, Exploratory Data Analysis, DAMA framework.

Contents

1. Introduction

BELab [1] is the data laboratory of Banco de España, a secure environment that provides external researchers with access to highly confidential and diverse microdata originating from both internal departments and external institutions. It offers controlled workspaces, standardized quality-control procedures, and regulated output-review processes to ensure that all analyses comply with strict confidentiality, accuracy, and governance standards. By centralizing heterogeneous datasets, BELab supports advanced analytical work and enables reliable, reproducible research within a robust data-management framework.

BELab currently hosts 18 microdata collections originating from a wide range of sources. Figure 1 presents an interactive dashboard summarizing their main characteristics. As shown, these datasets differ substantially in their nature, size, number and types of variables, update frequency, and confidentiality level. They are produced by different providers and therefore follow heterogeneous generation processes and quality-assurance standards.



DATASETS HOSTED IN BELAB DATA LABORATORY

BANCODEESPAÑA
Eurosistema

18	1210	117,7 m...	48,3
Datasets	Variables	Records	Volume (GB)

Category: Todas
Source: Todas
Frequency: Todas
Dataset: Todas

Dataset	Description	Details
BEV - Debt securities issuers	This is a val...	🔗
CBC - Consolidated non-financial corporate groups	BELab offer...	🔗
CBI - Individual enterprises	BELab offer...	🔗
CBS - Microdata on Sustainability Indicators	BELab offer...	🔗
CIR - Loans to legal persons	The CIR mic...	🔗
CIR-CBI - Microdata on Loans from CCR and Economic Indic...	The microda...	🔗
CNMC-PH - CNMC Household Panel	BELab provi...	🔗
DSP - General Government Debt	The Banco d...	🔗
EBAE - Microdata from Survey on Business Activity	In order to o...	🔗
EHUE - Microdata from the study on cash use habits	BELab offer...	🔗
FIN - Non-bank Fintech entities	BELab offer...	🔗

Figure 1 Power BI dashboard summarizing the microdata collections currently hosted by the BELab data laboratory of Banco de España.

To ensure consistency and efficiency in the exploratory data analysis (EDA) and data quality management (DQM) of such heterogeneous microdata, BELab has recently developed a generic, integrated tool designed to support a wide range of users—including data producers, lab technicians, data analysts, and researchers—in performing EDA and DQM on the datasets hosted in the laboratory.

Despite significant variations in the size, nature, and confidentiality levels of these datasets, they share a common structure that enables a high degree of standardization and generalization in EDA and DQM processes.

For EDA, the tool supports the standardized exploration of highly heterogeneous tabular datasets. Users can examine data at various levels of detail, including both aggregated information and the underlying microdata, depending on their access rights and specific interests. The tool also enables the automatic creation of interactive exploratory dashboards for large collections of tabular datasets and can serve as a data catalog to showcase available data. Additionally, it incorporates an AI-powered interface that allows users to ask questions about both aggregated and micro-level data.

In terms of data quality, the tool automatically evaluates several dimensions of the DAMA (Data Management Association) framework for DQM—namely, completeness, uniqueness, validity, and consistency. It checks whether the data align with the associated metadata, identifies structural and formatting issues, and detects duplicate records and invalid entries. Its EDA capabilities also support the identification and analysis of inconsistencies. Furthermore, the tool includes machine learning–based multivariate analysis and anomaly detection functionalities. This paper provides an overview of the tool and its main features.

2. BELab data and metadata

Each BELab dataset is composed of one or multiple CSV files, all of which share the same reading and extraction parameters (separator, encoding, etc.). Each file corresponds to a specific time interval—typically reporting years. All BELab data files adopt a standard tabular structure in which each row represents an observation or record and each column corresponds to a variable. A date or timestamp column is generally included, enabling the direct representation of time-series information within the table. The datasets follow a wide format, meaning that each variable is stored in its own column rather than encoded in a long or stacked layout. This wide structure provides a consistent and flexible foundation for both exploratory analysis and data-quality assessments. The tool can also be adapted to handle datasets stored in other file formats or relational databases.

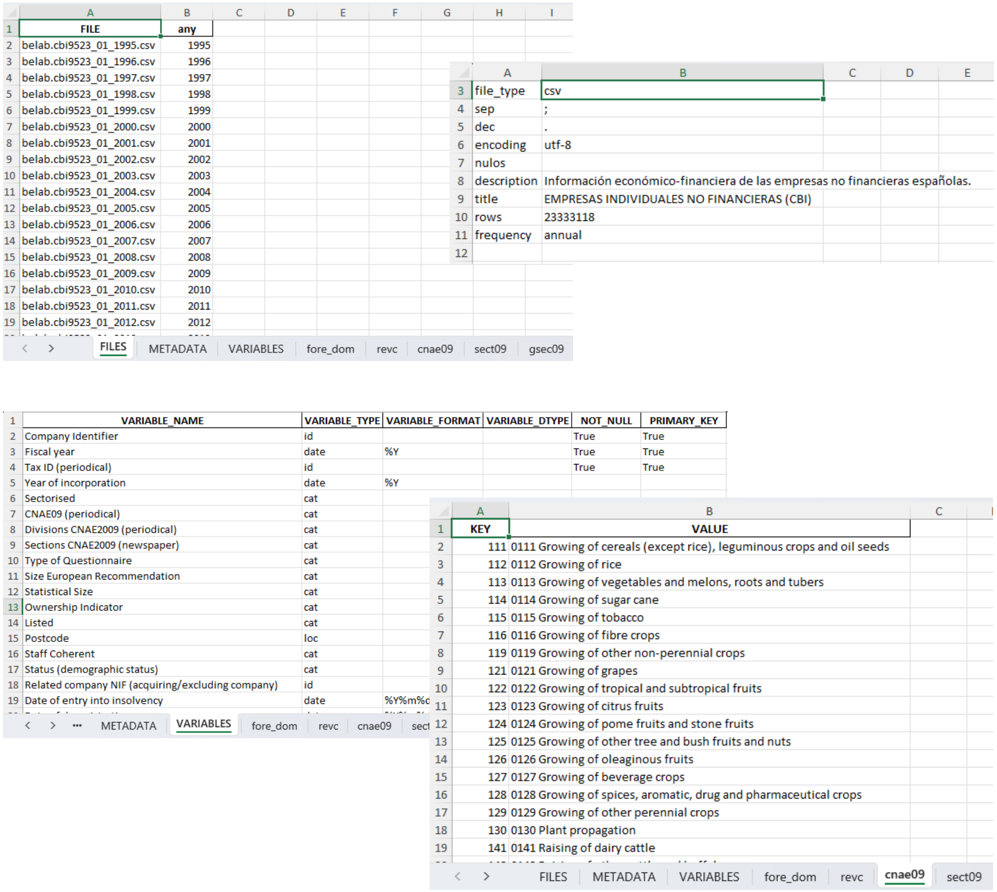
A standardized metadata structure has been defined to enable a generic tool design, so that datasets with very different volumes, numbers of variables, or variable types can be processed in a uniform way. These metadata files facilitate the standardization of data visualization, analysis, and quality checks, providing a unified interface for all datasets. In BELab, metadata is currently stored in Excel workbooks containing several standardized sheets, the most relevant of which are FILES, METADATA, and VARIABLES:

- The FILES sheet reports the time interval covered by each data file.
- The METADATA sheet includes general dataset-level information such as file type, separator, encoding, and dataset description.
- The VARIABLES sheet contains the most relevant information for generalizing the tool design. It lists all variables in the dataset along with their variable type, description, expected format, and other characteristics. The key element here is the variable type, as it enables the implementation of standardized procedures. Six variable types have been defined: identifier (id), date (date), numerical (num), categorical (cat), geographical (geo), and text (text). These types refer to the interpretation of the variable rather than its physical data type. For example, categorical variables may be stored as integers, floats, or strings, but all are treated as categorical by the tool. Formats for selected variable types—such as dates or identifiers—are also included. Additional attributes may be specified, such as NOT_NULL or PRIMARY_KEY, indicating

which variables cannot contain missing values or form part of the dataset's primary key.

Finally, the metadata file includes an additional sheet for each categorical variable, detailing all its possible categories and providing definitions for each category.

Examples of these metadata sheets for one of the BELab datasets (CBI, Individual enterprises [2]) are shown in Figure 2. This metadata structure enables standardized data exploration and quality-control procedures and allows the same analyses to be applied consistently across all variables of the same type (numerical, categorical, etc.).



FILE	any
belab.cbi9523_01_1995.csv	1995
belab.cbi9523_01_1996.csv	1996
belab.cbi9523_01_1997.csv	1997
belab.cbi9523_01_1998.csv	1998
belab.cbi9523_01_1999.csv	1999
belab.cbi9523_01_2000.csv	2000
belab.cbi9523_01_2001.csv	2001
belab.cbi9523_01_2002.csv	2002
belab.cbi9523_01_2003.csv	2003
belab.cbi9523_01_2004.csv	2004
belab.cbi9523_01_2005.csv	2005
belab.cbi9523_01_2006.csv	2006
belab.cbi9523_01_2007.csv	2007
belab.cbi9523_01_2008.csv	2008
belab.cbi9523_01_2009.csv	2009
belab.cbi9523_01_2010.csv	2010
belab.cbi9523_01_2011.csv	2011
belab.cbi9523_01_2012.csv	2012

file_type	csv
sep	;
dec	.
encoding	utf-8
nulos	
description	Información económico-financiera de las empresas no financieras españolas.
title	EMPRESAS INDIVIDUALES NO FINANCIERAS (CBI)
rows	23333118
frequency	annual

VARIABLE_NAME	VARIABLE_TYPE	VARIABLE_FORMAT	VARIABLE_DTYPE	NOT_NULL	PRIMARY_KEY
Company Identifier	id			True	True
Fiscal year	date	%Y		True	True
Tax ID (periodical)	id			True	True
Year of incorporation	date	%Y			
Sectorised	cat				
CNAE09 (periodical)	cat				
Divisions CNAE2009 (periodical)	cat				
Sections CNAE2009 (newspaper)	cat				
Type of Questionnaire	cat				
Size European Recommendation	cat				
Statistical Size	cat				
Ownership Indicator	cat				
Listed	cat				
Postcode	loc				
Staff Coherent	cat				
Status (demographic status)	cat				
Related company NIF (acquiring/excluding company)	id				
Date of entry into insolvency	date	%Y%m%d			

KEY	VALUE
111	0111 Growing of cereals (except rice), leguminous crops and oil seeds
112	0112 Growing of rice
113	0113 Growing of vegetables and melons, roots and tubers
114	0114 Growing of sugar cane
115	0115 Growing of tobacco
116	0116 Growing of fibre crops
119	0119 Growing of other non-perennial crops
121	0121 Growing of grapes
122	0122 Growing of tropical and subtropical fruits
123	0123 Growing of citrus fruits
124	0124 Growing of pome fruits and stone fruits
125	0125 Growing of other tree and bush fruits and nuts
126	0126 Growing of oleaginous fruits
127	0127 Growing of beverage crops
128	0128 Growing of spices, aromatic, drug and pharmaceutical crops
129	0129 Growing of other perennial crops
130	0130 Plant propagation
141	0141 Raising of dairy cattle

Figure 2 Example of metadata file for the CBI dataset.

3. Modular tool design

The implemented tool follows a modular design that allows users to apply only the processing steps required for each dataset and facilitates the seamless integration of new functionalities. The entire system is developed in Python and includes a graphical user interface built with Plotly Dash, enabling non-technical users to analyse diverse datasets without needing any programming skills. The interface is currently annotated in Spanish.

The tool is designed to accommodate datasets with widely varying volumes and characteristics. Its inputs consist of the microdata files and the associated metadata file. Before processing, the microdata are converted into Parquet format to enable efficient column-level extraction from large files. To ensure scalability and streamline the analysis of large datasets, the tool's workflow is organized into three main stages (Figure 3):

1. Configuration (Step 1): The user completes an interactive web form specifying the processing tasks to be performed (Figure 4). This form is currently implemented in Streamlit, with a planned migration to Plotly Dash. Users select the dataset and files to be processed, configure sampling, filtering, and aggregation options, and choose the DAMA quality-framework parameters. The selected configuration is saved to a configuration file in JSON format, which serves as input for the next stage.

2. Offline processing (Step 2): The main data processing is executed offline, and the results are stored in multiple output files. The configuration file drives this phase, during which sampling, filtering, and aggregation are performed according to the user-defined settings. A large collection of time-series aggregated indicators is precomputed to ensure fast rendering in Step 3. DAMA quality-control tests are executed, and multivariate analysis and anomaly detection are performed. Additionally, the dataset documentation is indexed to support the AI-assisted interface. This stage can be computationally intensive depending on dataset size. All outputs generated here are saved for later exploration.

3. Interactive exploration (Step 3). In this final stage, the user examines the processed outputs through an interactive web interface enhanced with AI-based capabilities (Figure 5). The interface permits the selection of any available dataset, together with the corresponding configuration and results files generated in the previous step. It comprises several dedicated sections for variable-type-specific exploration, DAMA-aligned data-quality diagnostics, multivariate anomaly-detection results, and an interactive summary dashboard. Users may also compute custom aggregates directly within the interface. The integrated AI assistant enables natural-language queries regarding dataset documentation, metadata, microdata, and aggregated statistics, thereby facilitating a more intuitive and efficient navigation of the analytical environment. Since all computationally intensive processes are executed offline in Step 2, the exploration phase remains fast, fluid, and responsive.

The tool outputs a collection of result files generated in Step 2 and an HTML or PDF report summarizing the main findings produced in Step 3. This three-stage architecture enables efficient processing of large datasets while maintaining a smooth user experience.

The tool is currently hosted on a workstation and accessed locally by users. A migration to a production server at Banco de España is planned in the coming months.

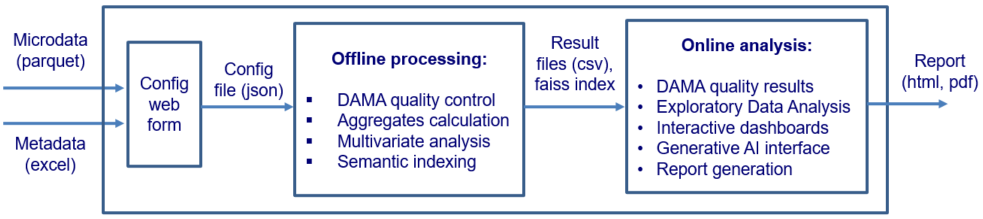
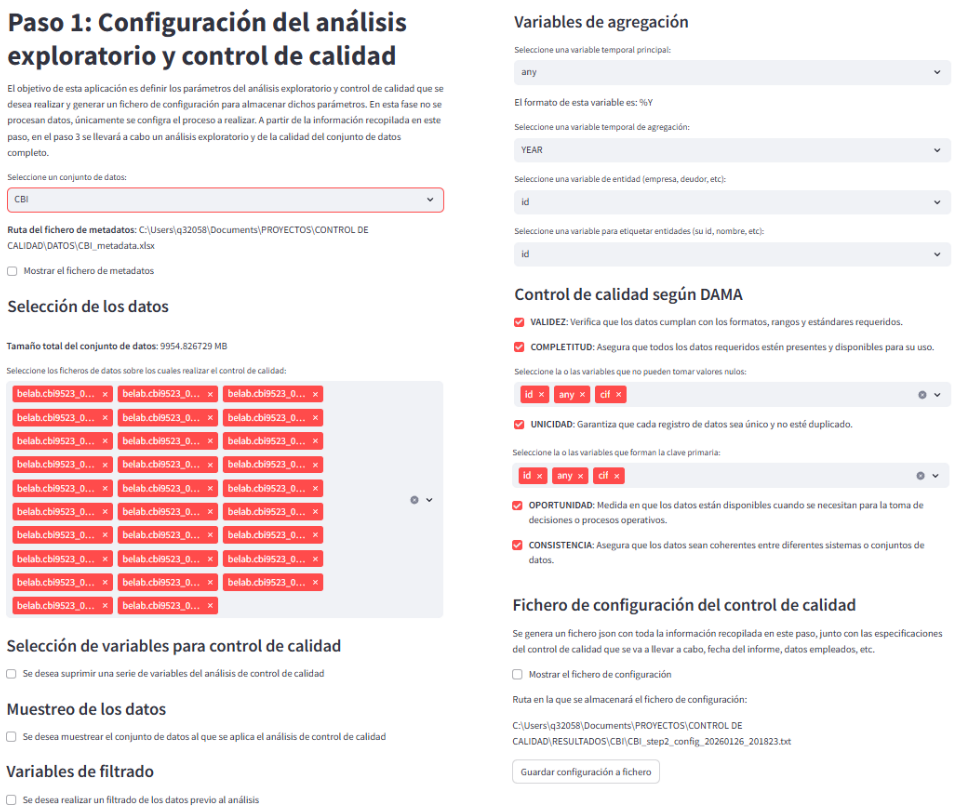


Figure 3 Summarized workflow of the developed EDA and DQM tool.



Paso 1: Configuración del análisis exploratorio y control de calidad

El objetivo de esta aplicación es definir los parámetros del análisis exploratorio y control de calidad que se desea realizar y generar un fichero de configuración para almacenar dichos parámetros. En esta fase no se procesan datos, únicamente se configra el proceso a realizar. A partir de la información recopilada en este paso, en el paso 3 se llevará a cabo un análisis exploratorio y de la calidad del conjunto de datos completo.

Seleccione un conjunto de datos:

CBI

Ruta del fichero de metadatos: C:\Users\q32058\Documents\PROYECTOS\CONTROL DE CALIDAD\DATOS\CBI_metadata.xlsx

□ Mostrar el fichero de metadatos

Selección de los datos

Tamaño total del conjunto de datos: 9954.826729 MB

Seleccione los ficheros de datos sobre los cuales realizar el control de calidad:

belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... × belab.cbi9523_0... ×
belab.cbi9523_0... × belab.cbi9523_0... ×

Selección de variables para control de calidad

□ Se desea suprimir una serie de variables del análisis de control de calidad

Muestreo de los datos

□ Se desea muestrear el conjunto de datos al que se aplica el análisis de control de calidad

Variables de filtrado

□ Se desea realizar un filtrado de los datos previo al análisis

Variables de agregación

Seleccione una variable temporal principal:

any

El formato de esta variable es: %Y

Seleccione una variable temporal de agregación:

YEAR

Seleccione una variable de entidad (empresa, deudor, etc):

id

Seleccione una variable para etiquetar entidades (su id, nombre, etc):

id

Control de calidad según DAMA

☑ **VALIDEZ:** Verifica que los datos cumplan con los formatos, rangos y estándares requeridos.

☑ **COMPLETITUD:** Asegura que todos los datos requeridos estén presentes y disponibles para su uso.

Seleccione la o las variables que no pueden tomar valores nulos:

id × any × cif ×

☑ **UNICIDAD:** Garantiza que cada registro de datos sea único y no esté duplicado.

Seleccione la o las variables que forman la clave primaria:

id × any × cif ×

☑ **OPORTUNIDAD:** Medida en que los datos están disponibles cuando se necesitan para la toma de decisiones o procesos operativos.

☑ **CONSISTENCIA:** Asegura que los datos sean coherentes entre diferentes sistemas o conjuntos de datos.

Fichero de configuración del control de calidad

Se genera un fichero json con toda la información recopilada en este paso, junto con las especificaciones del control de calidad que se va a llevar a cabo, fecha del informe, datos empleados, etc.

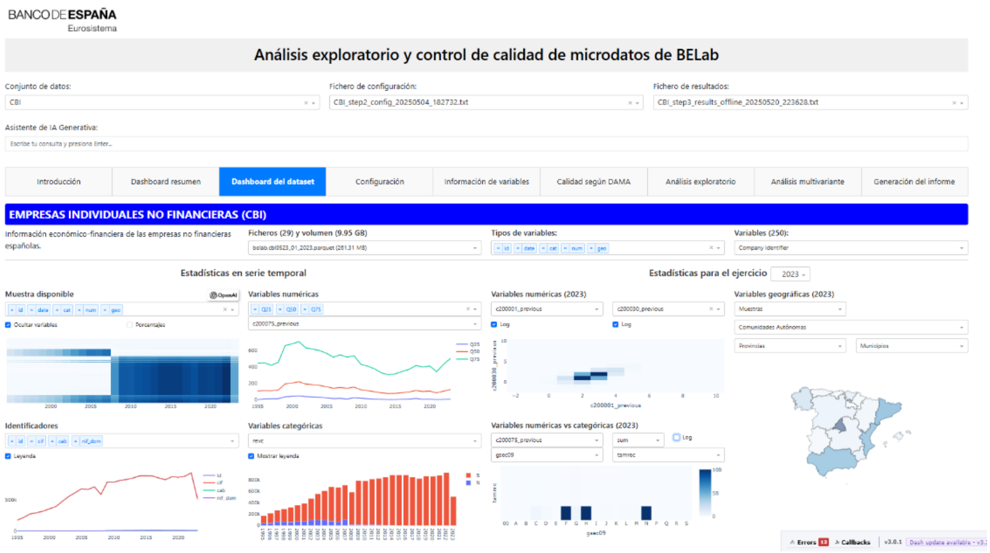
□ Mostrar el fichero de configuración

Ruta en la que se almacenará el fichero de configuración:

C:\Users\q32058\Documents\PROYECTOS\CONTROL DE CALIDAD\RESULTADOS\CBI\CBI_step2_config_20260126_201823.txt

Guardar configuración a fichero

Figure 4 Interactive configuration web form specifying the processing to be performed (temporarily implemented in Streamlit).



BANCO DE ESPAÑA
Eurosistema

Análisis exploratorio y control de calidad de microdatos de BELab

Conjunto de datos: CBI

Fichero de configuración: CBI_step2_config_20250504_182732.txt

Fichero de resultados: CBI_step3_results_offline_20250520_223628.txt

Asistente de IA Generativa: Escribe tu consulta y presiona Enter...

EMPRESAS INDIVIDUALES NO FINANCIERAS (CBI)

Información económico-financiera de las empresas no financieras españolas.

Ficheros (29) y volumen (9.95 GB): belab.cbi9523_01_2023.parquet (261.31 MB)

Tipos de variables: id, date, cat, num, geo

Variables (250): Company Identifier

Estadísticas en serie temporal

Estadísticas para el ejercicio 2023

Figure 5 EDA and DQM application user interface, showing the automatic dashboard generation section.

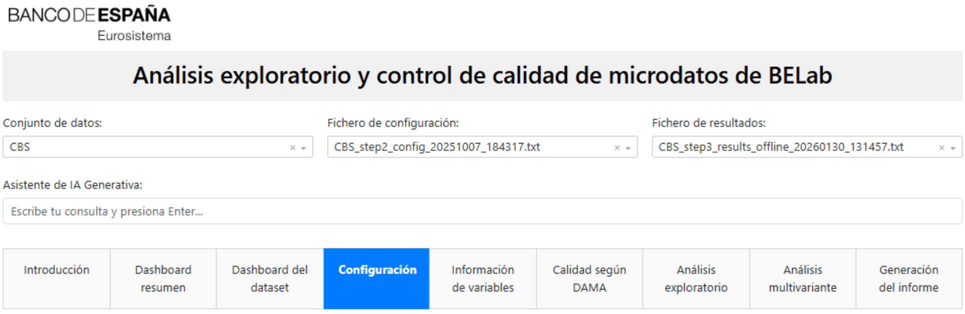
4. Current features of the tool

The tool follows a modular design, and new functionalities are continuously incorporated. Figure 6 presents the current layout of the graphical user-interface header and its main sections. At the top of the interface, the user may select a dataset from a dropdown menu, as well as any of the available configuration and results files. A generative-AI prompt is also provided, enabling natural-language queries about the microdata, aggregated data, metadata, and all associated documentation; this functionality is described in Section 4.5.

Below the header, the following sections are accessible through dedicated tabs:

- **Introduction:** offers a brief description of the tool and its main integrated modules.
- **Summarizing dashboard:** displays the interactive dashboard that summarises all available datasets, as shown in Figure 1.
- **Dataset dashboard:** presents the automatically generated interactive dashboard for the selected dataset, as described in Section 4.4.
- **Configuration:** summarises the main configuration parameters specified during the setup stage, including the list of processed data files, sampling, filtering and aggregation parameters, DAMA test settings, the list of modules executed, and the resulting processing times (Figure 7).
- **Variable information:** provides an overview of the dataset's variables and their main characteristics (Figure 8), listing the variables by type and displaying the metadata provided in the *VARIABLES* tab of the metadata file.
- **DQM tests (DAMA dimensions):** presents the results of the data-quality tests performed during offline processing, as described in Section 4.1.
- **Exploratory data analysis:** provides comprehensive functionality for examining all variables in the dataset, organised by variable type, as detailed in Section 4.2. This section enables systematic inspection of distributions, temporal patterns, and potential data-quality issues.
- **Multivariate analysis and anomaly detection:** presents the multivariate analysis outputs produced during the offline processing stage, as described in Section 4.3. This module supports the exploration of complex interdependencies among variables and the identification of anomalous observations using machine-learning–based techniques.
- **Report generation:** reserved for the generation of final reports (currently under development).

The tool's main capabilities are detailed in the subsequent sections.



Análisis exploratorio y control de calidad de microdatos de BELab

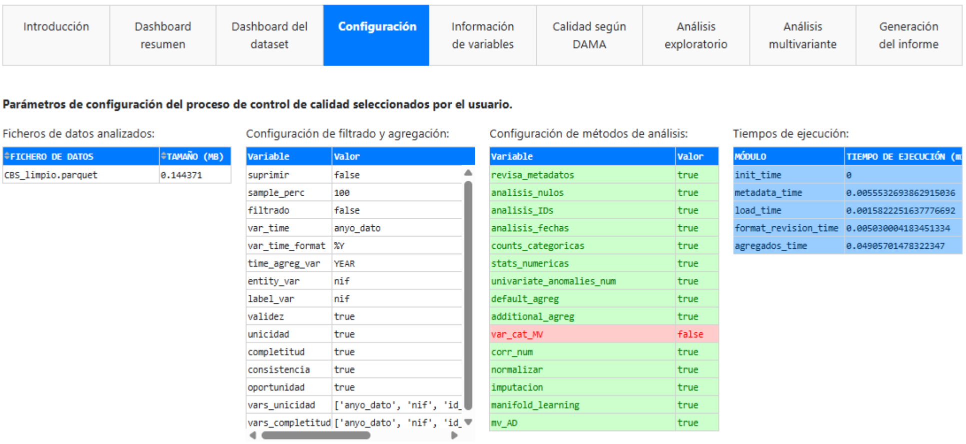
Conjunto de datos: CBS

Fichero de configuración: CBS_step2_config_20251007_184317.txt

Fichero de resultados: CBS_step3_results_offline_20260130_131457.txt

Asistente de IA Generativa: Escribe tu consulta y presiona Enter...

Figure 6 Tool header and main sections.



Parámetros de configuración del proceso de control de calidad seleccionados por el usuario.

Ficheros de datos analizados:

FICHERO DE DATOS	TAMAÑO (MB)
CBS_limpio.parquet	0.144371

Configuración de filtrado y agregación:

Variable	Valor
suprimir	false
sample_perc	100
filtrado	false
var_time	anyo_dato
var_time_format	%Y
time_agreg_var	YEAR
entity_var	nif
label_var	nif
validez	true
unicidad	true
completitud	true
consistencia	true
oportunidad	true
vars_unicidad	['anyo_dato', 'nif', 'id_
vars_completitud	['anyo_dato', 'nif', 'id_

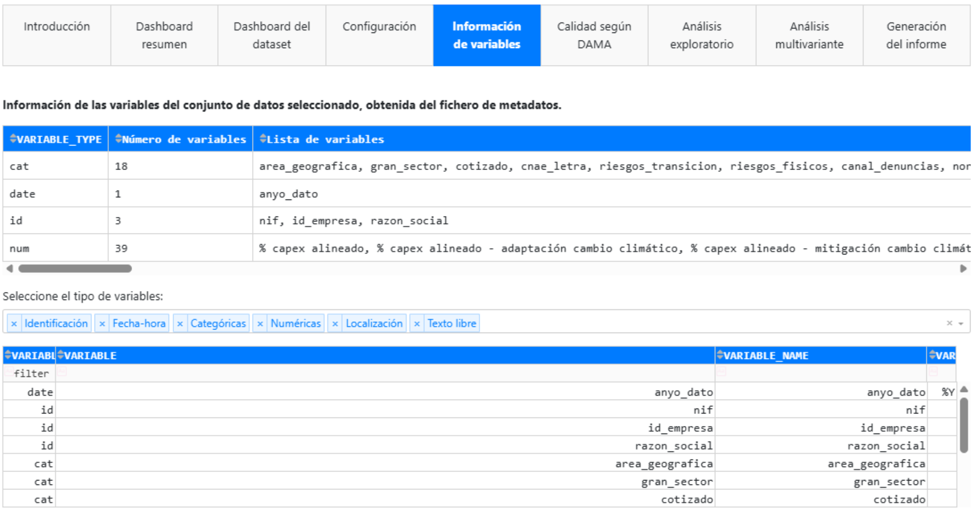
Configuración de métodos de análisis:

Variable	Valor
revisa_metadatos	true
analisis_nulos	true
analisis_IDs	true
analisis_fechas	true
counts_categoricas	true
stats_numericas	true
univariate_anomalies_num	true
default_agreg	true
additional_agreg	true
var_cat_MV	false
corr_num	true
normalizar	true
imputacion	true
manifold_learning	true
mv_AD	true

Tiempos de ejecución:

MÓDULO	TIEMPO DE EJECUCIÓN (m
init_time	0
metadata_time	0.0055532693862915036
load_time	0.0015822251637776692
format_revision_time	0.005030004183451334
agregados_time	0.04905701478322347

Figure 7 Configuration parameters section.



Información de las variables del conjunto de datos seleccionado, obtenida del fichero de metadatos.

VARIABLE_TYPE	Número de variables	Lista de variables
cat	18	area_geografica, gran_sector, cotizado, cnae_letra, riesgos_transicion, riesgos_fisicos, canal_denuncias, nor
date	1	anyo_dato
id	3	nif, id_empresa, razon_social
num	39	% capex alineado, % capex alineado - adaptación cambio climático, % capex alineado - mitigación cambio climát

Seleccione el tipo de variables: × Identificación × Fecha-hora × Categóricas × Numéricas × Localización × Texto libre

VARIABL	VARIABLE	VARIABLE_NAME	VAR
filter			
date	anyo_dato	anyo_dato	%Y
id	nif	nif	
id	id_empresa	id_empresa	
id	razon_social	razon_social	
cat	area_geografica	area_geografica	
cat	gran_sector	gran_sector	
cat	cotizado	cotizado	

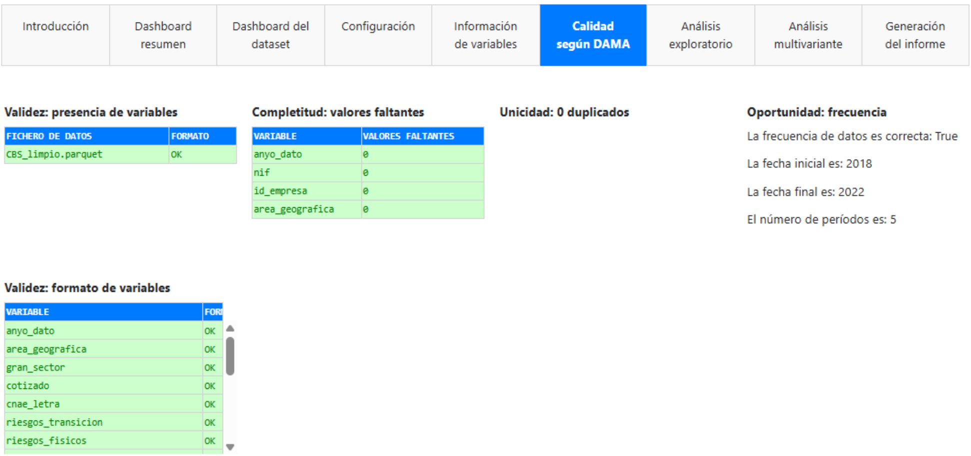
Figure 8 Variable information section.

4.1. DQM tests according to DAMA quality framework

The DAMA (Data Management Association) framework defines the key dimensions of data quality and provides a foundation for assessing and improving data quality within organizations. The tool automatically evaluates the following DAMA dimensions as part of its DQM functionalities:

- Completeness: Ensures that all required data is present and available. The tool verifies that all variables defined in the metadata are present in the data files and checks that variables labelled as NOT_NULL contain no missing values.
- Uniqueness: Ensures that each record is unique and not duplicated. The tool tests whether duplicate records exist by checking combinations of variables labelled as PRIMARY_KEY in the metadata.
- Validity: Ensures that data conforms to required formats and standards. The tool verifies that each variable matches the format specified in the metadata. In particular, date variables are checked against the expected date format, and categorical variables are validated to ensure they contain only predefined categories.
- Timeliness: Ensures that data is up-to-date and available when needed. The tool checks the frequency of reporting, the number of periods, and the correctness of start and end dates.

In summary, the tool assesses whether the data conforms to the associated metadata, identifies structural and formatting issues, and detects duplicates and invalid entries according to the configuration parameters selected in the first processing step. Results are presented in the web interface using visual summaries to facilitate interpretation (Figure 9).



Validez: presencia de variables

FICHERO DE DATOS	FORMATO
CBS_limpio.parquet	OK

Completitud: valores faltantes

VARIABLE	VALORES FALTANTES
anyo_dato	0
nif	0
id_empresa	0
area_geografica	0

Unicidad: 0 duplicados

Oportunidad: frecuencia

La frecuencia de datos es correcta: True

La fecha inicial es: 2018

La fecha final es: 2022

El número de períodos es: 5

Validez: formato de variables

VARIABLE	FOR
anyo_dato	OK
area_geografica	OK
gran_sector	OK
cotizado	OK
cnae_letra	OK
riesgos_transicion	OK
riesgos_fisicos	OK

Figure 9 DQM test results based on the DAMA dimensions validity, completeness, uniqueness and timeliness.

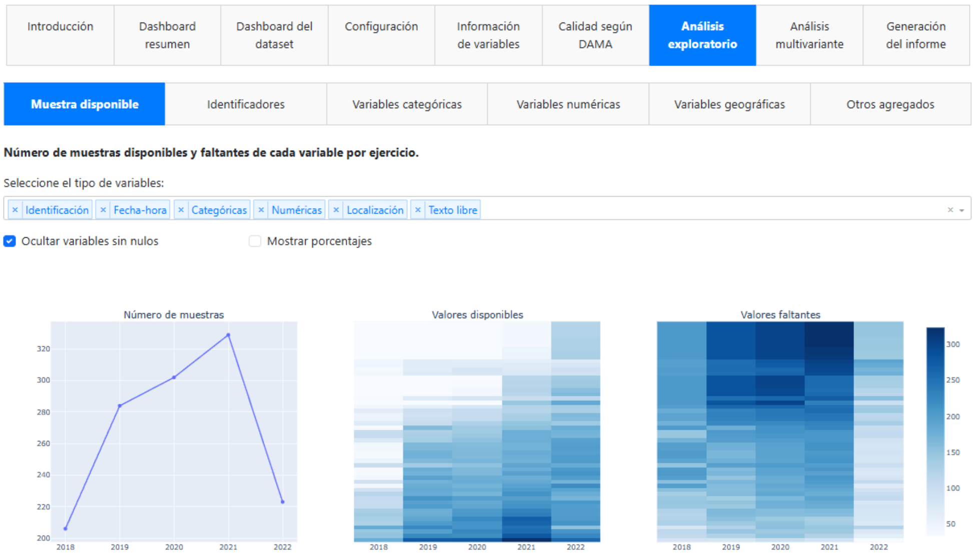
4.2. Exploratory data analysis

The EDA section of the tool enables users to comprehensively explore all variables in each dataset—organised by variable type and available at multiple levels of detail, from aggregated information to the underlying microdata. The tool relies on the variable categorisation provided in the metadata file to standardise the exploration workflow across variable types. The EDA section is structured into the following subsections:

- Available sample
- Identifying variables
- Categorical variables
- Numerical variables
- Geographical variables
- Custom aggregates

4.2.1. Available sample

This section enables users to examine the presence of data for each variable over time. It allows to select the variable types of interest, and it shows the overall number of samples, the number of samples of each variable, and the number of missing values of each variable per reporting period. This information is graphically presented on the user interface (Figure 10).



Número de muestras disponibles y faltantes de cada variable por ejercicio.

Seleccione el tipo de variables:

× Identificación × Fecha-hora × Categóricas × Numéricas × Localización × Texto libre

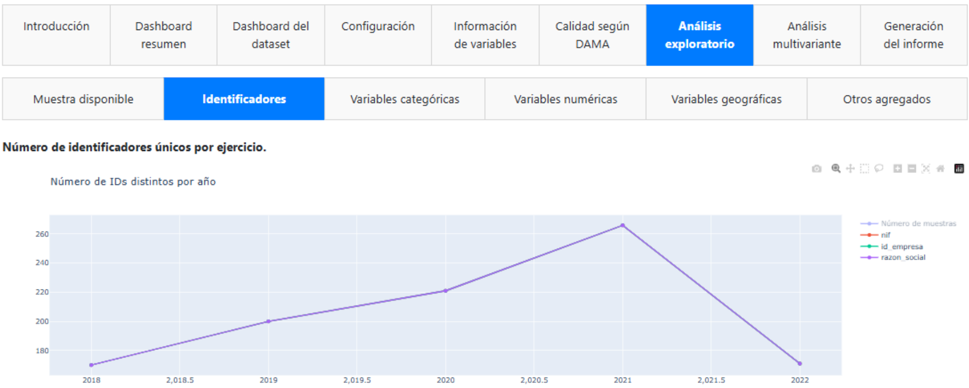
☑ Ocultar variables sin nulos   □ Mostrar porcentajes

Figure 10 EDA section of the tool: exploration of the available sample.

4.2.2. Identifying variables

This subsection reports the number of distinct entities per reporting period represented in the dataset. Users may select the identifying variables of interest and assess the consistency of their relationships. When multiple identifiers refer to the same underlying entity, they are expected to exhibit parallel behaviour across time—specifically, the number of distinct entities associated with each identifier should coincide. Deviations from this pattern, such as one identifier yielding a larger number of entities than another within the same period, may signal inconsistencies or underlying data-quality issues.

Figure 11 illustrates this functionality by displaying the yearly number of distinct entities according to several identifiers, all of which show perfectly aligned counts. In addition, an automatic consistency-checking procedure for identifier relationships will be incorporated into the DQM section of the tool, enabling systematic validation of these constraints.



Número de identificadores únicos por ejercicio.

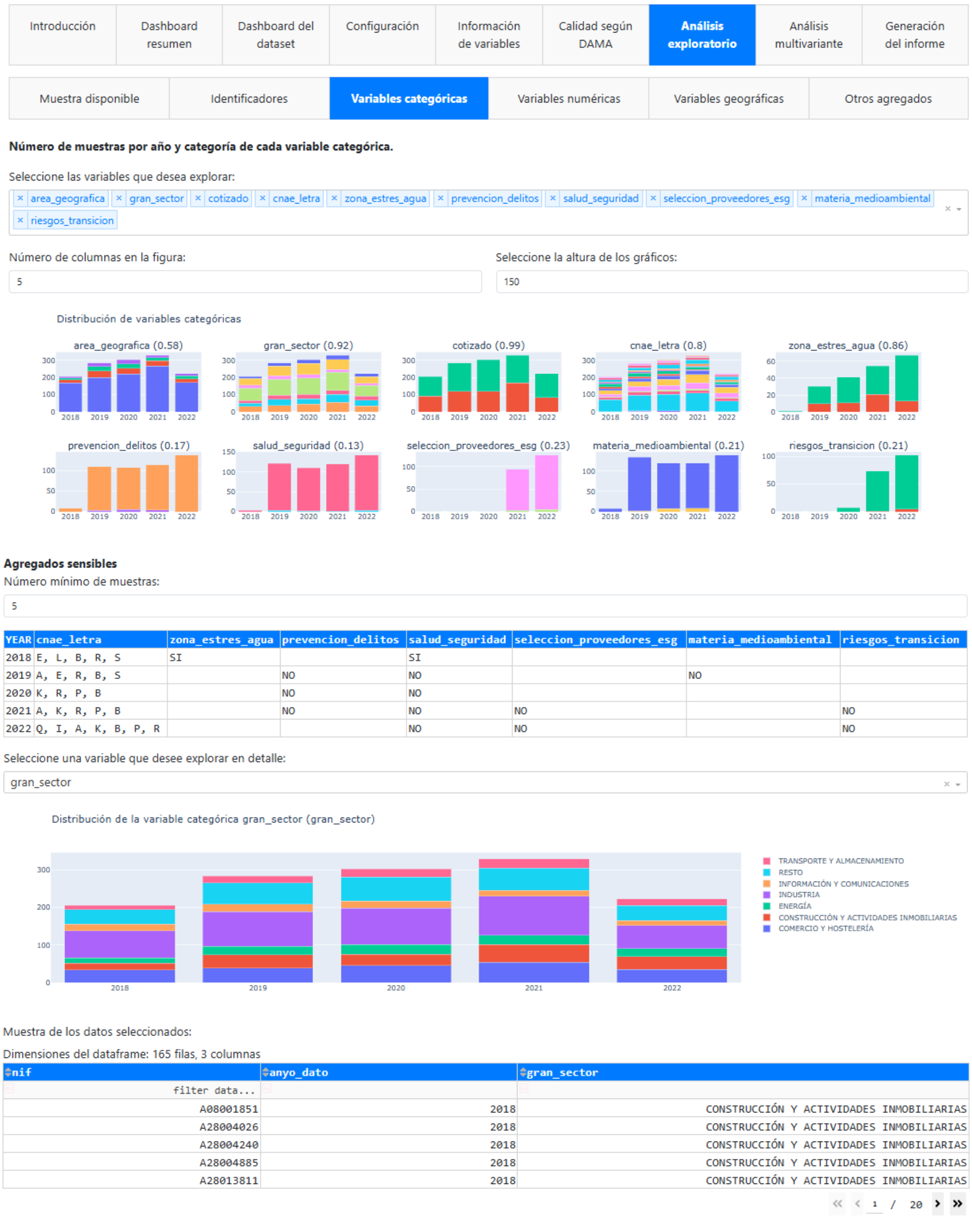
Figure 11 EDA section of the tool: exploration of identifying variables.

4.2.3. Categorical variables

This subsection enables users to examine the distribution of all categorical variables over time simultaneously and to identify potential inconsistencies (see Figure 12). All visualisations are generated dynamically, even for datasets containing a large number of categorical variables, leveraging counts pre-computed offline to ensure efficient and scalable rendering. Users may specify the variables of interest and customise the layout of the visualisations, including the number of columns and the height of the plots.

The tool also provides a table that flags any aggregated plots that may raise confidentiality concerns because the number of observations underlying the corresponding aggregate falls below a predefined threshold. Specifically, the tool indicates which categories, for each variable and reporting period, contain fewer observations than the specified minimum. For example, in Figure 12 the tool highlights potential confidentiality issues in 2018 for categories E, L, B, R and S of the CNAE variable, as these aggregates are based on fewer than five observations.

If users identify any issue or an aspect they want to investigate further, they can click on the graph of interest or select the variable from a dropdown menu to display a more detailed view. When a user clicks on the detailed graph, the tool also shows a sample of the underlying data for the selected category—provided the user has the necessary access rights to the microdata.



Número de muestras por año y categoría de cada variable categórica.

Seleccione las variables que desea explorar:

× area_geografica × gran_sector × cotizado × cnae_letra × zona_estres_agua × prevencion_delitos × salud_seguridad × seleccion_proveedores_esg × materia_medioambiental × riesgos_transicion

Número de columnas en la figura: 5

Seleccione la altura de los gráficos: 150

Agregados sensibles

Número mínimo de muestras: 5

YEAR	cnae_letra	zona_estres_agua	prevencion_delitos	salud_seguridad	seleccion_proveedores_esg	materia_medioambiental	riesgos_transicion
2018	E, L, B, R, S	SI		SI			
2019	A, E, R, B, S		NO	NO		NO	
2020	K, R, P, B		NO	NO			
2021	A, K, R, P, B		NO	NO	NO		NO
2022	Q, I, A, K, B, P, R			NO	NO		NO

Seleccione una variable que desee explorar en detalle: gran_sector

Muestra de los datos seleccionados:

Dimensiones del dataframe: 165 filas, 3 columnas

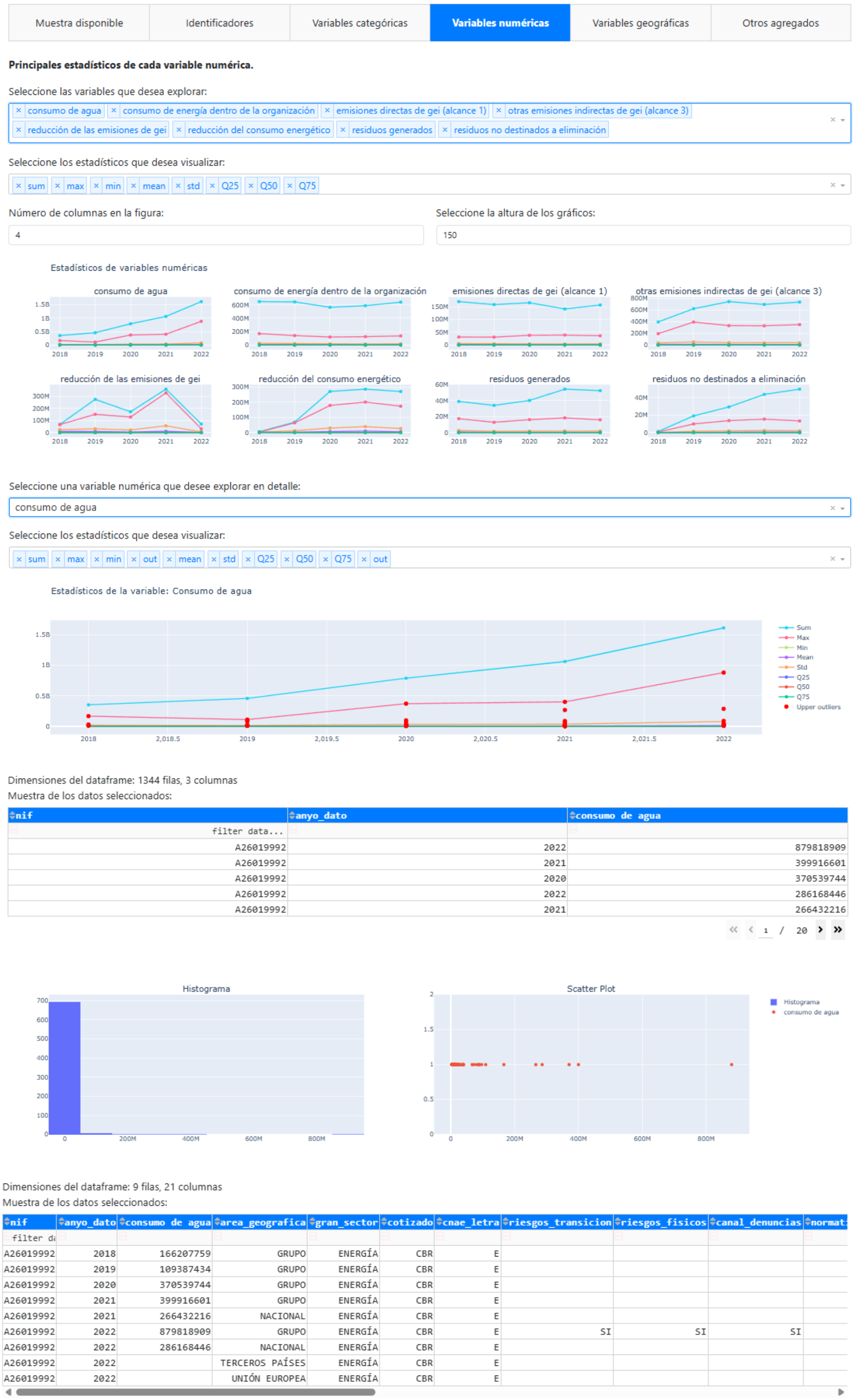
nif	anyo_dato	gran_sector
filter data...		
A08001851	2018	CONSTRUCCIÓN Y ACTIVIDADES INMOBILIARIAS
A28004026	2018	CONSTRUCCIÓN Y ACTIVIDADES INMOBILIARIAS
A28004240	2018	CONSTRUCCIÓN Y ACTIVIDADES INMOBILIARIAS
A28004885	2018	CONSTRUCCIÓN Y ACTIVIDADES INMOBILIARIAS
A28013811	2018	CONSTRUCCIÓN Y ACTIVIDADES INMOBILIARIAS

« ‹ 1 / 20 › »

Figure 12 EDA section of the tool: exploration of categorical variables.

4.2.4. Numerical variables

This subsection allows users to examine multiple descriptive statistics for all numerical variables simultaneously, thereby facilitating the detection of potential issues or inconsistencies in the data (see Figure 13). Users may select the variables of interest—by default, all variables are included—and the tool is able to render even very large numbers of plots efficiently. Some datasets hosted in BELab contain hundreds of numerical variables; nevertheless, all visualisations are generated instantaneously because the underlying aggregated statistics are pre-computed offline. This design choice enables seamless exploration of large-scale datasets.



Principales estadísticos de cada variable numérica.

Seleccione las variables que desea explorar:

× consumo de agua × consumo de energía dentro de la organización × emisiones directas de gei (alcance 1) × otras emisiones indirectas de gei (alcance 3) × reducción de las emisiones de gei × reducción del consumo energético × residuos generados × residuos no destinados a eliminación

Seleccione los estadísticos que desea visualizar:

× sum × max × min × mean × std × Q25 × Q50 × Q75

Número de columnas en la figura: 4

Seleccione la altura de los gráficos: 150

Estadísticos de variables numéricas
consumo de agua
consumo de energía dentro de la organización
emisiones directas de gei (alcance 1)
otras emisiones indirectas de gei (alcance 3)
reducción de las emisiones de gei
reducción del consumo energético
residuos generados
residuos no destinados a eliminación
2018
2019
2020
2021
2022

Seleccione una variable numérica que desee explorar en detalle:

consumo de agua

Seleccione los estadísticos que desea visualizar:

× sum × max × min × out × mean × std × Q25 × Q50 × Q75 × out

Dimensiones del dataframe: 1344 filas, 3 columnas

Muestra de los datos seleccionados:

nif	anyo_dato	consumo de agua
filter data...		
A26019992	2022	879818909
A26019992	2021	399916601
A26019992	2020	370539744
A26019992	2022	286168446
A26019992	2021	266432216

« ‹ 1 / 20 › »

Dimensiones del dataframe: 9 filas, 21 columnas

Muestra de los datos seleccionados:

nif	anyo_dato	consumo de agua	area_geografica	gran_sector	cotizado	cnae_letra	riesgos_transicion	riesgos_fisicos	canal_denuncias	normat
filter da										
A26019992	2018	166207759	GRUPO	ENERGÍA	CBR	E				
A26019992	2019	109387434	GRUPO	ENERGÍA	CBR	E				
A26019992	2020	370539744	GRUPO	ENERGÍA	CBR	E				
A26019992	2021	399916601	GRUPO	ENERGÍA	CBR	E				
A26019992	2021	266432216	NACIONAL	ENERGÍA	CBR	E				
A26019992	2022	879818909	GRUPO	ENERGÍA	CBR	E	SI	SI	SI	
A26019992	2022	286168446	NACIONAL	ENERGÍA	CBR	E				
A26019992	2022		TERCEROS PAÍSES	ENERGÍA	CBR	E				
A26019992	2022		UNIÓN EUROPEA	ENERGÍA	CBR	E				

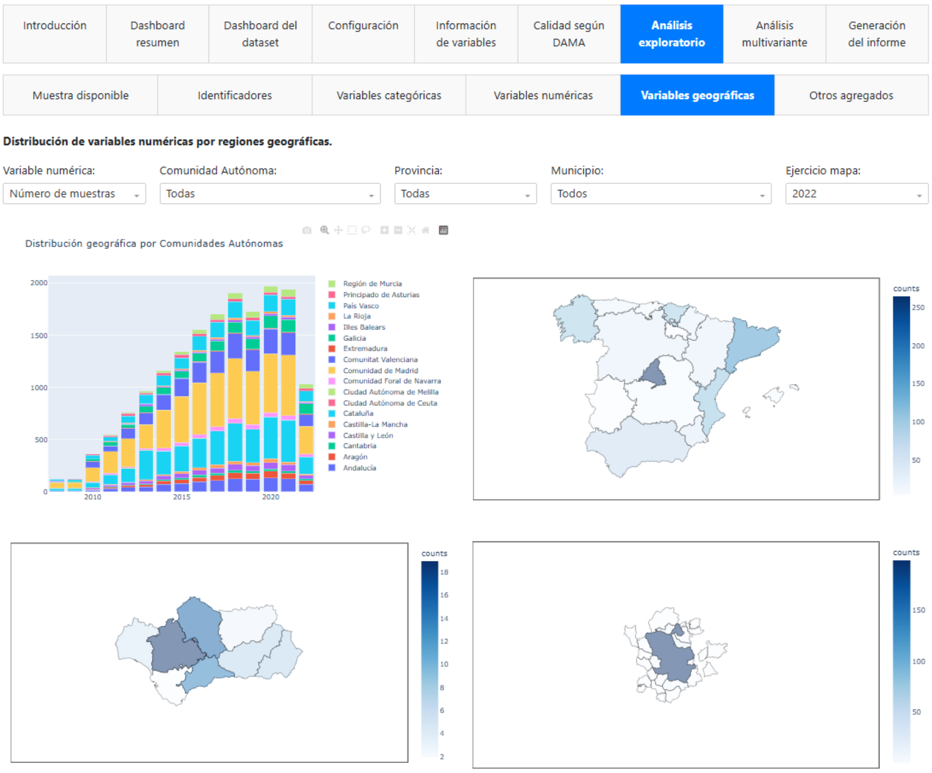
Figure 13 EDA section of the tool: exploration of numerical variables.

Users may also specify which aggregated statistics are displayed (e.g., sum, maximum, minimum, mean, standard deviation, quantiles, and outliers). In particular, the tool can visualise outliers for all numerical variables, identified using the standard boxplot rule. When the user clicks on one of the plots or selects a variable from the dropdown menu, the corresponding variable is shown in greater detail in the lower panel. The user may again select the statistics of interest and interact with the resulting visualisation in greater depth.

If the user clicks on the detailed plot, a sample of the underlying microdata is displayed—conditional on the user having the appropriate access rights. A histogram and a scatter plot are also provided, allowing potential anomalies or data-quality issues to be identified rapidly. Finally, clicking on any data point reveals the corresponding underlying microdata in the panel below.

4.2.5. Geographical variables

If the dataset under analysis includes geographical information—such as postal codes or standard regional classifications (e.g., municipality, province, autonomous community)—the tool provides dedicated functionality for exploring these variables (see Figure 14). Users may choose the numerical measure to be displayed (e.g., the number of observations per region or any other numerical variable). The selected variable is then visualised both as a bar plot and on a map, and users may further refine the exploration by filtering results by region or reporting period.



Distribución de variables numéricas por regiones geográficas.

Variable numérica: Número de muestras

Comunidad Autónoma: Todas

Provincia: Todas

Municipio: Todos

Ejercicio mapa: 2022

Figure 14 EDA section of the tool: exploration of geographical information.

When a finer geographical level is selected (e.g., province, municipality, or other lower-level administrative division), the tool automatically updates the map to display the corresponding regional boundaries. This enables users to inspect geographical patterns at the desired level of detail and to detect potential anomalies or inconsistencies in the data.

4.2.6. Custom aggregates

Finally, the tool allows users to compute and visualise custom aggregates directly from the dataset. In particular, users may examine the distribution of one or two numerical variables for any selected reporting period by choosing them from dropdown menus, with the option to display the values on a logarithmic scale. If a single numerical variable is selected, its distribution is presented as a histogram; if two numerical variables are selected, the tool displays a heatmap illustrating their joint distribution.

The tool also facilitates the exploration of relationships between numerical and categorical variables for a given reporting period. The user first selects a numerical variable, an aggregation function (e.g., sum, mean, maximum), and optionally a logarithmic transformation. One or two categorical variables may then be chosen to compute and visualise the aggregated numerical variable across the selected categorical dimensions. When a single categorical variable is selected, the tool produces a bar plot showing the aggregated values by category. When two categorical variables are selected, a heatmap is generated to represent the joint distribution of the numerical variable across the cross-classification of the two categorical variables.

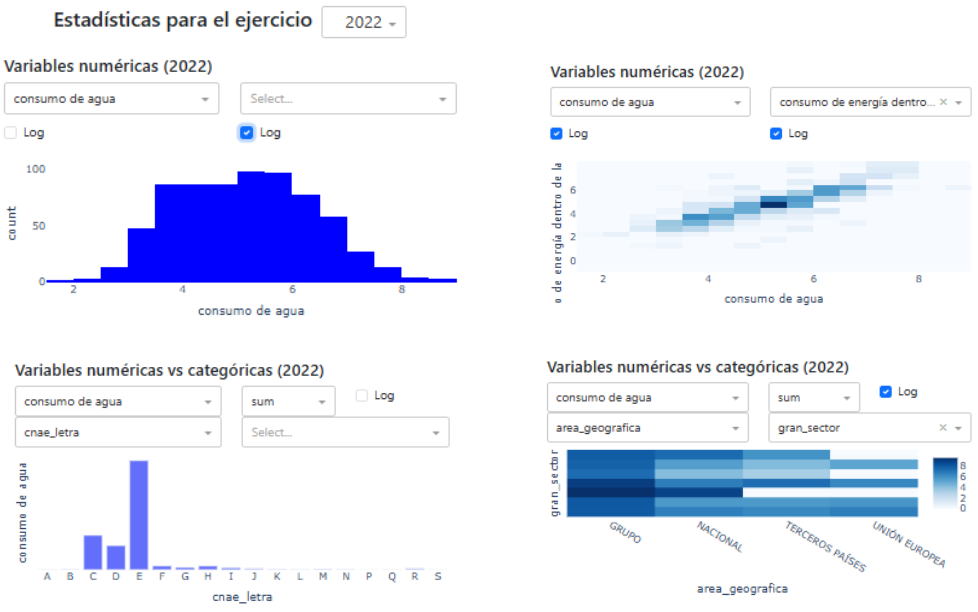
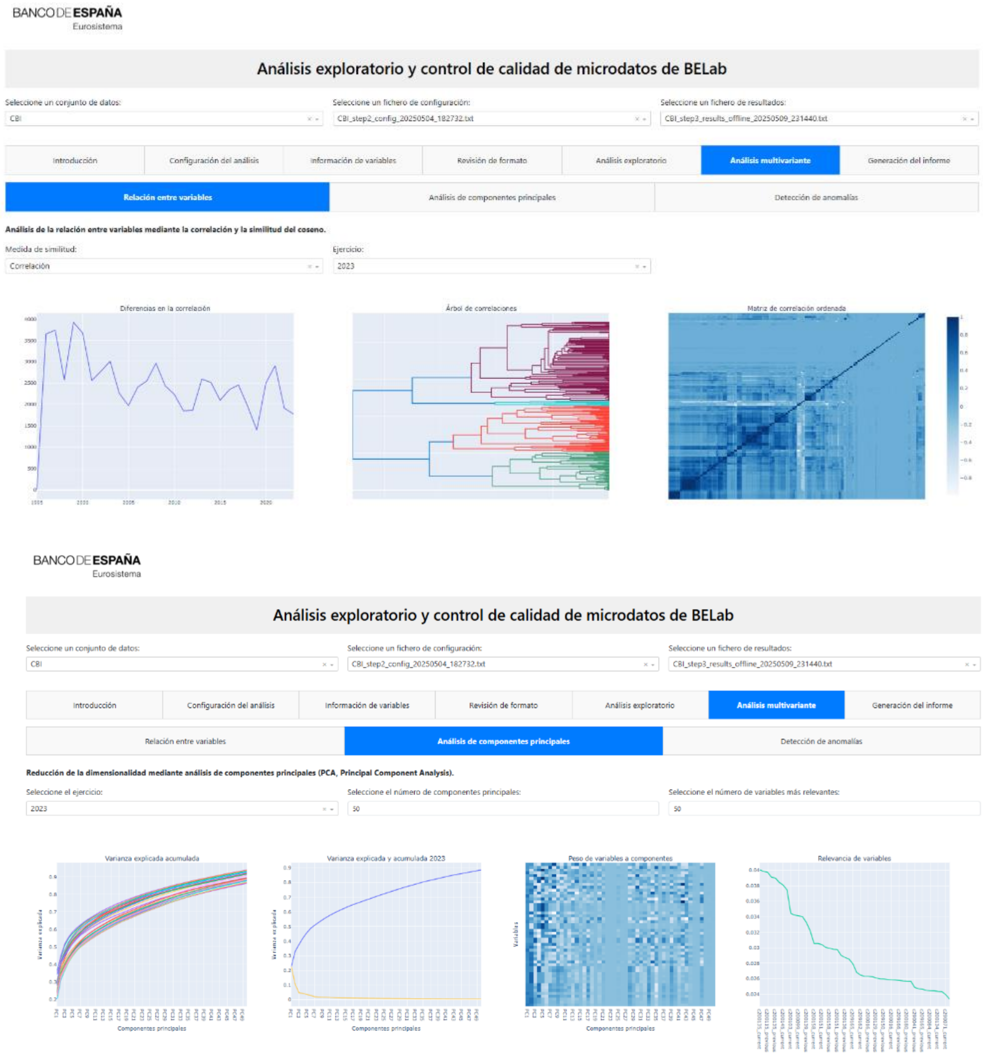


Figure 15 EDA section of the tool: exploration of custom aggregates.

These custom aggregates are computed on demand rather than precomputed, since the number of possible combinations of numerical and categorical variables grows quickly with dataset dimensionality. However, because the underlying data are stored in a columnar format (Parquet), reading only the required variables is highly efficient, enabling the tool to generate these customised visualisations rapidly even for large datasets.

4.3. Multivariate anomaly detection

The tool also incorporates a set of multivariate analysis and anomaly-detection functionalities based on the numerical variables of the dataset, implemented through machine-learning methods (see Figure 16 and Figure 17). It currently includes the following modules:



Análisis exploratorio y control de calidad de microdatos de BELab

Seleccione un conjunto de datos: CBI

Seleccione un fichero de configuración: CBI_step2_config_20250504_182732.txt

Seleccione un fichero de resultados: CBI_step3_results_offline_20250509_231440.txt

Análisis de la relación entre variables mediante la correlación y la similitud del coseno.

Medida de similitud: Correlación

Ejercicio: 2023

Análisis exploratorio y control de calidad de microdatos de BELab

Seleccione un conjunto de datos: CBI

Seleccione un fichero de configuración: CBI_step2_config_20250504_182732.txt

Seleccione un fichero de resultados: CBI_step3_results_offline_20250509_231440.txt

Reducción de la dimensionalidad mediante análisis de componentes principales (PCA, Principal Component Analysis).

Seleccione el ejercicio: 2023

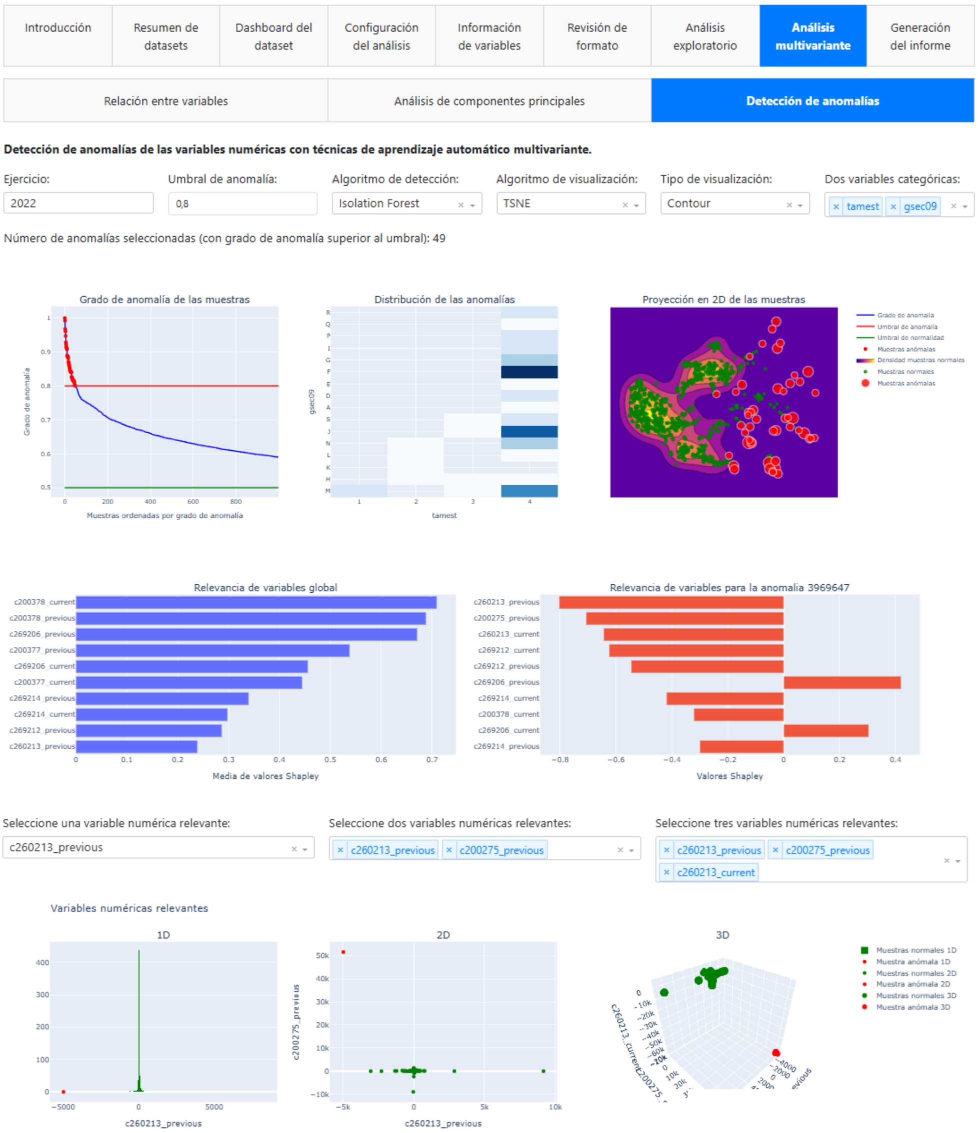
Seleccione el número de componentes principales: 50

Seleccione el número de variables más relevantes: 50

Figure 16 Multivariate analysis (similarity among variables and PCA).

- **Similarity analysis.** This subsection enables users to examine pairwise relationships among numerical variables by computing cosine similarity and Pearson correlation. The tool displays the temporal evolution of similarities, allowing users to detect potential inconsistencies or structural changes over time. In addition, a dendrogram derived from the similarity matrix is presented as a hierarchical clustering tree, providing an interpretable representation of variable groupings and highlighting clusters as well as variables that deviate markedly from the rest. A heatmap of the similarity matrix, with variables ordered according to the dendrogram, further facilitates the identification of coherent clusters, redundancies, and unexpected associations.

Principal Component Analysis (PCA). This subsection offers an interactive visualisation of the PCA results generated during the offline processing step. Users may inspect the proportion of variance explained by each principal component and assess the potential for dimensionality reduction or the presence of latent structure in the data. The tool also provides a heatmap depicting the contribution of each variable to each principal component in each reporting period, thereby enabling a detailed assessment of variable loadings. Finally, the interface highlights the variables that contribute most strongly to the principal components, assisting users in identifying the dimensions that drive the main patterns of variability in the dataset.



Detección de anomalías de las variables numéricas con técnicas de aprendizaje automático multivariante.

Ejercicio: 2022

Umbral de anomalía: 0,8

Algoritmo de detección: Isolation Forest

Algoritmo de visualización: TSNE

Tipo de visualización: Contour

Dos variables categóricas: tamest, gsec09

Número de anomalías seleccionadas (con grado de anomalía superior al umbral): 49

Seleccione una variable numérica relevante: c260213_previous

Seleccione dos variables numéricas relevantes: c260213_previous, c200275_previous

Seleccione tres variables numéricas relevantes: c260213_previous, c200275_previous, c260213_current

Figure 17 Explainable multivariate anomaly detection.

- **Multivariate anomaly detection.** This subsection displays the results of various anomaly-detection algorithms applied to the numerical variables. Users may select the reporting period, the anomaly-score threshold, the detection method (currently Isolation Forest, with additional algorithms easily integrable), a manifold-learning visualisation technique (e.g., t-SNE, MDS, UMAP), the type of plot (contour or surface), and up to two categorical variables to analyse the distribution of anomalies. The tool reports the highest anomaly scores, provides a heatmap of anomaly incidence across the selected categorical dimensions, and presents the chosen manifold-learning embedding, where anomalous observations are shown in red against a two- or three-dimensional kernel density estimate of the nominal population.

Clicking on an anomalous point in the manifold-learning plot displays a bar chart highlighting the variables that contribute most strongly to the anomaly score. For comparison, an equivalent bar chart showing the most influential variables at the global level is also provided. In addition, the tool presents 1D, 2D, and 3D visualisations for combinations of the three most influential variables, enabling users to examine in detail the multivariate patterns that led to the classification of the observation as anomalous.

4.4. Automatic dashboard generation

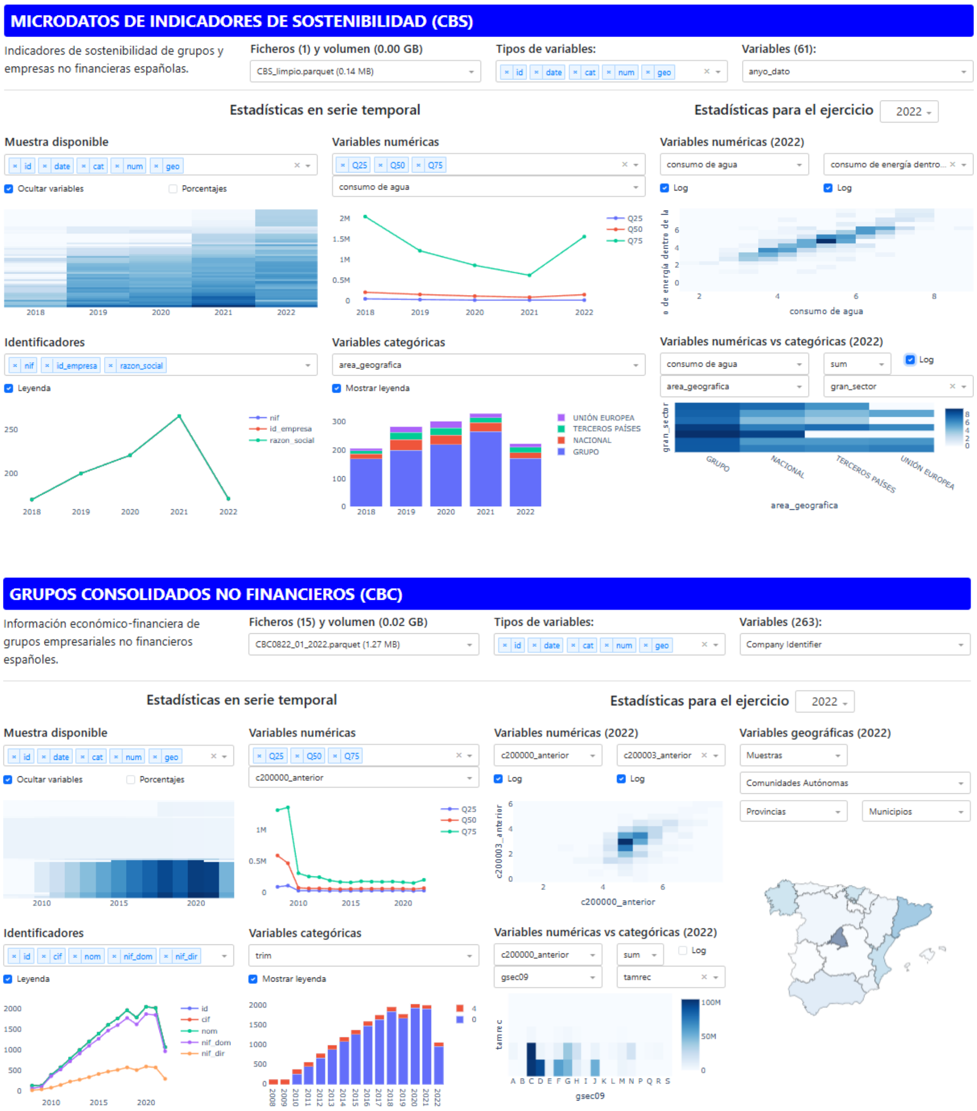
The tool automatically generates interactive dashboards for each dataset, providing a comprehensive overview of the available information. The dashboard layout has been designed in a generic and dataset-agnostic manner, allowing all collections to be explored effectively. This is possible because most datasets include variables of the standard types—identifiers, dates, numerical variables, and categorical variables—while sections corresponding to less common variable types (such as geographical or text variables) are displayed only when relevant.

At the top of each dashboard, the tool presents a description of the selected dataset, together with the list of files and their sizes, as well as a summary of the variables grouped by type. The dashboard offers a condensed representation of the information contained in the EDA section. For example, one section is dedicated to the available sample, showing the number of observations per variable and reporting period, as described in Section 4.2.1. Another section presents the identifying variables, mirroring the information in Section 4.2.2. Additional sections summarise the distribution of categorical variables and the main descriptive statistics for numerical variables, following the structure of Sections 4.2.3 and 4.2.4. All these visualisations adopt a time-series perspective, using the selected time-aggregation variable along the horizontal axis.

On the righthand side of the dashboard, users may choose a specific reporting period and examine the relationships among numerical and categorical variables, in a manner consistent with Section 4.2.6. When the dataset contains geographical information, an additional section appears, displaying the corresponding variables on a map, following the approach described in Section 4.2.5.

Once the generic layout has been defined, the generation of dashboards for any additional dataset entails virtually no additional cost, as all dashboards adhere to the same structure and are generated automatically. This feature is particularly valuable when showcasing large collections of datasets for which manual dashboard construction would be impractical.

Figure 18 provides two examples of automatically generated dashboards for the CBS ([4]) and CBC ([3]) BELab datasets. As illustrated, the dashboard design adapts dynamically to the presence or absence of geographical information, while retaining an identical organisation for all other components.



MICRODATOS DE INDICADORES DE SOSTENIBILIDAD (CBS)

Indicadores de sostenibilidad de grupos y empresas no financieras españolas.

Ficheros (1) y volumen (0.00 GB): CBS_limpio.parquet (0.14 MB)

Tipos de variables: id, date, cat, num, geo

Variables (61): anyo_dato

Estadísticas en serie temporal

Estadísticas para el ejercicio: 2022

Muestra disponible: id, date, cat, num, geo

☑ Ocultar variables □ Porcentajes

Variables numéricas: Q25, Q50, Q75; consumo de agua

Variables numéricas (2022): consumo de agua ☑ Log; consumo de energía dentro... ☑ Log

Identificadores: nif, id_empresa, razon_social

☑ Leyenda

Variables categóricas: area_geografica

☑ Mostrar leyenda

Variables numéricas vs categóricas (2022): consumo de agua; sum; ☑ Log; area_geografica; gran_sector

GRUPOS CONSOLIDADOS NO FINANCIEROS (CBC)

Información económico-financiera de grupos empresariales no financieros españoles.

Ficheros (15) y volumen (0.02 GB): CBC0822_01_2022.parquet (1.27 MB)

Tipos de variables: id, date, cat, num, geo

Variables (263): Company Identifier

Estadísticas en serie temporal

Estadísticas para el ejercicio: 2022

Muestra disponible: id, date, cat, num, geo

☑ Ocultar variables □ Porcentajes

Variables numéricas: Q25, Q50, Q75; c200000_anterior

Variables numéricas (2022): c200000_anterior ☑ Log; c200003_anterior ☑ Log

Variables geográficas (2022): Muestras; Comunidades Autónomas; Provincias; Municipios

Identificadores: id, cif, nom, nif_dom, nif_dir

☑ Leyenda

Variables categóricas: trim

☑ Mostrar leyenda

Variables numéricas vs categóricas (2022): c200000_anterior; sum; □ Log; gsec09; tamrec

Figure 18 Automatic dashboard generation (CBS and CBC, respectively).

4.5. Generative AI interface

The tool integrates a generative-AI interface that enables users to query all available information—including documentation, aggregated data, microdata, and metadata—using natural language. The implemented AI interface follows the workflow depicted in Figure 19. When a user submits a query concerning BELab's documentation or data, a first AI agent (implemented via Azure OpenAI) classifies the query into one of two categories:

1. The query can be answered using BELab documentation or metadata, or
2. The query requires access to data (microdata or aggregated data).

If the query is classified as documentation- or metadata-based, the upper workflow is executed. In this case, relevant passages from BELab's documentation are retrieved using a Retrieval-Augmented Generation (RAG) architecture. All BELab documentation is indexed during the offline processing stage using a hybrid search strategy that combines a traditional index—generated through Azure Search—with a semantic index built using open-source Hugging Face embedding models and a FAISS-based vector store. Azure OpenAI is then used to generate the final natural-language response.

Conversely, if the query is classified as data-based, the lower workflow is executed. In this case, an Azure OpenAI instance generates Python code tailored to the user's request and informed by the metadata of the selected dataset. This code is then executed locally, accessing the data stored in Parquet files and ensuring that no underlying data are transmitted to external AI providers. A final natural-language response is subsequently produced using Azure OpenAI. Because this last step may require processing sensitive information, an on-premises AI deployment would be necessary to comply with confidentiality requirements. At present, such an on-premises deployment is not yet available; therefore, this functionality has been tested only on non-confidential datasets.

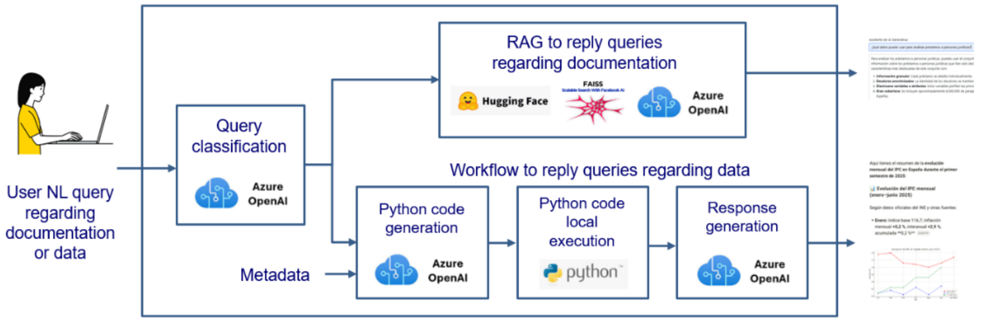

User NL query regarding documentation or data
Query classification
Azure OpenAI
RAG to reply queries regarding documentation
Hugging Face
FAISS
Scalable Search With Facebook AI
Azure OpenAI
Workflow to reply queries regarding data
Metadata
Python code generation
Azure OpenAI
Python code local execution
python
Response generation
Azure OpenAI
Asistente de IA Generativa
¿Qué datos puedo usar para analizar préstamos a personas jurídicas?
Aquí tienes el resumen de la evolución mensual del IPC en España durante el primer semestre de 2025:
Evolución del IPC mensual (enero–junio 2025)
Según datos oficiales del INE y otras fuentes:
Enero: índice base 116,7; inflación mensual +0,2 %, interanual +2,9 %, acumulada **0,2 %**

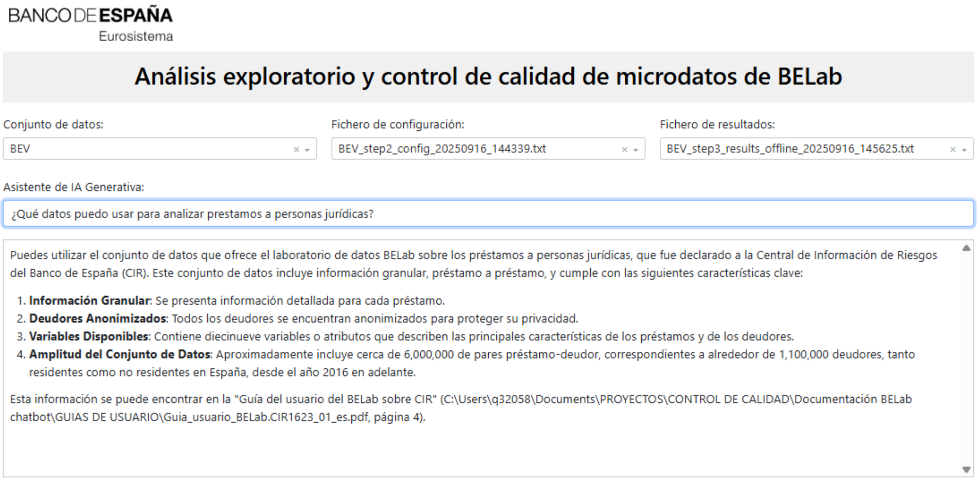
Figure 19 Generative AI interface workflow.

The data-based workflow described above represents a simplified variant of broader information-extraction initiatives developed by international organizations. A prominent example is StatGPT ([5]), an artificial-intelligence platform designed to operate on official statistical data. StatGPT enables users to query, transform, analyse, visualise and interpret data through a natural-language interface, and incorporates user-interaction steps after each processing block to validate the intermediate results produced by the different AI agents.

Developed in collaboration with the International Monetary Fund (IMF), StatGPT relies on an AI assistant that generates queries following the SDMX (Statistical Data and Metadata eXchange) standard. The platform is currently in a testing phase, and its continued development is expected to be shared with additional institutions in the near future.

Figure 20 and Figure 21 illustrate two examples of the operation of the Generative AI interface in the two scenarios described above: documentation-based queries and data-based queries. In both cases, the system correctly interprets the user's request and produces an appropriate response based on the relevant sources of information.

In the first example (Figure 20), the user poses the following query in Spanish: "¿Qué datos puedo utilizar para analizar los préstamos a personas jurídicas?" ("What data can I use to analyze loans to legal entities?"). The system classifies this query as documentation-based and retrieves the most relevant information from BELab's indexed documentation. It then generates a natural-language response describing the appropriate dataset and provides the corresponding document name, file path, and page number. Importantly, in this scenario the response is not restricted to the dataset selected in the interface, but draws upon the entire collection of indexed documentation and metadata.



BANCO DE ESPAÑA
Eurosistema

Análisis exploratorio y control de calidad de microdatos de BELab

Conjunto de datos: BEV

Fichero de configuración: BEV_step2_config_20250916_144339.txt

Fichero de resultados: BEV_step3_results_offline_20250916_145625.txt

Asistente de IA Generativa:

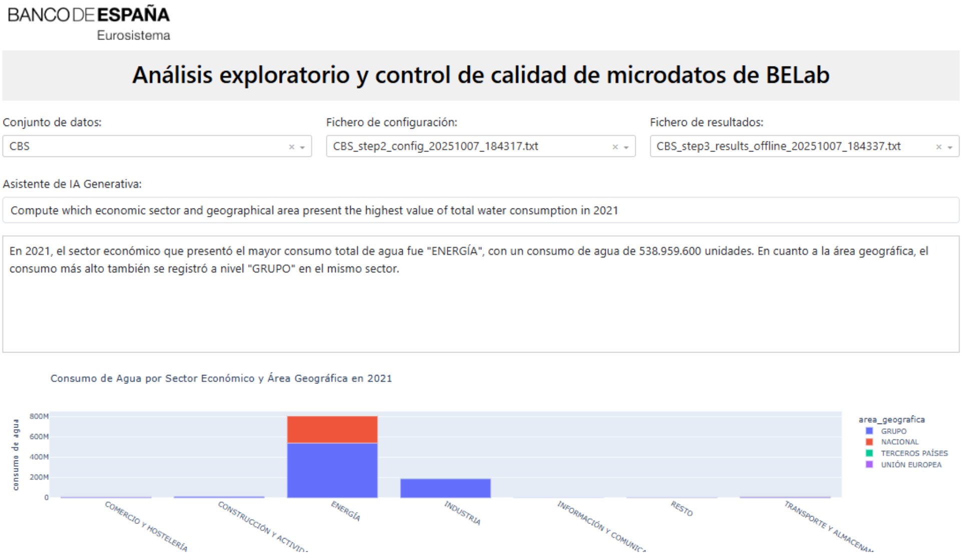
¿Qué datos puedo usar para analizar prestamos a personas jurídicas?

Puedes utilizar el conjunto de datos que ofrece el laboratorio de datos BELab sobre los préstamos a personas jurídicas, que fue declarado a la Central de Información de Riesgos del Banco de España (CIR). Este conjunto de datos incluye información granular, préstamo a préstamo, y cumple con las siguientes características clave:

1. **Información Granular**: Se presenta información detallada para cada préstamo.
2. **Deudores Anonimizados**: Todos los deudores se encuentran anonimizados para proteger su privacidad.
3. **Variables Disponibles**: Contiene diecinueve variables o atributos que describen las principales características de los préstamos y de los deudores.
4. **Amplitud del Conjunto de Datos**: Aproximadamente incluye cerca de 6,000,000 de pares préstamo-deudor, correspondientes a alrededor de 1,100,000 deudores, tanto residentes como no residentes en España, desde el año 2016 en adelante.

Esta información se puede encontrar en la "Guía del usuario del BELab sobre CIR" (C:\Users\q32058\Documents\PROYECTOS\CONTROL DE CALIDAD\Documentación BELab chatbot\GUIAS DE USUARIO\Guia_usuario_BELab.CIR1623_01_es.pdf, página 4).

Figure 20 Generative AI interface example: documentation- and metadata-based query.



BANCO DE ESPAÑA
Eurosistema

Análisis exploratorio y control de calidad de microdatos de BELab

Conjunto de datos: CBS

Fichero de configuración: CBS_step2_config_20251007_184317.txt

Fichero de resultados: CBS_step3_results_offline_20251007_184337.txt

Asistente de IA Generativa:

Compute which economic sector and geographical area present the highest value of total water consumption in 2021

En 2021, el sector económico que presentó el mayor consumo total de agua fue "ENERGÍA", con un consumo de agua de 538.959.600 unidades. En cuanto a la área geográfica, el consumo más alto también se registró a nivel "GRUPO" en el mismo sector.

Figure 21 Generative AI interface example: data-based query.

In the second example (Figure 21), the user poses a data-oriented query in English: "Compute which economic sector and geographical area present the highest value of total water consumption in 2021." The system correctly classifies the query as data-based, uses the AI agent to generate the Python code required to compute the answer, and executes this code locally, ensuring that no sensitive data are transmitted to external AI providers. The system then returns a natural-language explanation accompanied by an interactive plot supporting the result. The user may subsequently verify the correctness of the output by consulting the corresponding visualisations in the EDA section or in the dataset dashboard.

5. Conclusions and next steps

To ensure consistency and efficiency in exploratory data analysis (EDA) and data quality management (DQM) for the heterogeneous microdata hosted in BELab, a generic tool has recently been designed and developed. Implemented entirely in Python and equipped with an extensive, user-friendly, and interactive web interface, the tool is intended for a broad range of business users—including data producers, lab technicians, analysts, and researchers—without requiring programming expertise.

The tool is particularly well suited for analysing and assessing the quality of time-series microdata sets. It incorporates a wide set of standardised tests and visualisations that facilitate efficient exploration and quality assessment across diverse collections of microdata, enabling users to rapidly identify key data characteristics and potential issues. It can also automatically generate standardised interactive dashboards for all datasets, thereby eliminating the need for manual dashboard construction. In addition, the tool integrates a generative-AI assistant that allows users to query microdata, metadata, aggregated statistics, and associated documentation using natural language. Modularity and ease of reuse were fundamental design principles, supporting collaborative development and the integration of new functionalities.

The main challenge encountered during development involved defining a generic framework applicable to datasets with highly heterogeneous nature, volume, and confidentiality levels, compounded by limited standardisation in data and metadata structures, which complicates generalisation.

The tool has already enabled the identification of several categories of errors and inconsistencies in the data and metadata hosted in BELab, including:

- inconsistencies between data and metadata, as well as formatting errors;
- incoherent temporal patterns in numerical and categorical variables;
- anomalies in relationships between variables, detected through multivariate analysis and anomaly-detection techniques.

Planned next steps include:

- automating report generation and initiating systematic error reporting to data producers;
- expanding the set of tests aligned with the DAMA data-quality dimensions;

- parallelising data-processing workflows within the Banco de España's data lake.
- deploying the tool on the corporate web server, taking into consideration confidentiality requirements and restrictions on code sharing.

References

1. Banco de España. *BELab Data Laboratory*. https://www.bde.es/wbe/en/para-ciudadano/servicios/belab/
2. Banco de España. *Microdatos de Empresas Individuales (CBI)*. https://www.bde.es/wbe/en/para-ciudadano/servicios/belab/contenido/microdatos-disponibles/microdatos-empresas-individuales-cbi.html
3. Banco de España. *Microdatos de Grupos Empresariales No Financieros Consolidados (CBC)*. https://www.bde.es/wbe/en/para-ciudadano/servicios/belab/contenido/microdatos-disponibles/microdatos-grupos-empresariales-no-financieros-consolidados-cbc.html
4. Banco de España. *Microdatos de Indicadores de Sostenibilidad (CBS)*. https://www.bde.es/wbe/en/para-ciudadano/servicios/belab/contenido/microdatos-disponibles/microdatos-de-indicadores-de-sostenibilidad--cbs-.html
5. StatGPT. *AI Platform for Official Statistical Data*. https://statgpt.dialx.ai/