



TÜRKİYE CUMHURİYET
MERKEZ BANKASI

SCALING DATA TRUST IN CENTRAL BANKING

A distributed data quality framework for statistical excellence

Elif Beyza Çeltik
Data Governance and Statistics Department
Central Bank of the Republic of Türkiye

BASEL • AUGUST 2026



WHY DATA TRUST MATTERS

Central banks operate as a source of truth. As data expands, manual SQL checks become a bottleneck and a publication risk. Therefore a validation framework is needed to ensure the quality of the data being processed and published.

THE DESIGN OBJECTIVE

Place a systematic quality checkpoint between incoming data and policy or public reporting.

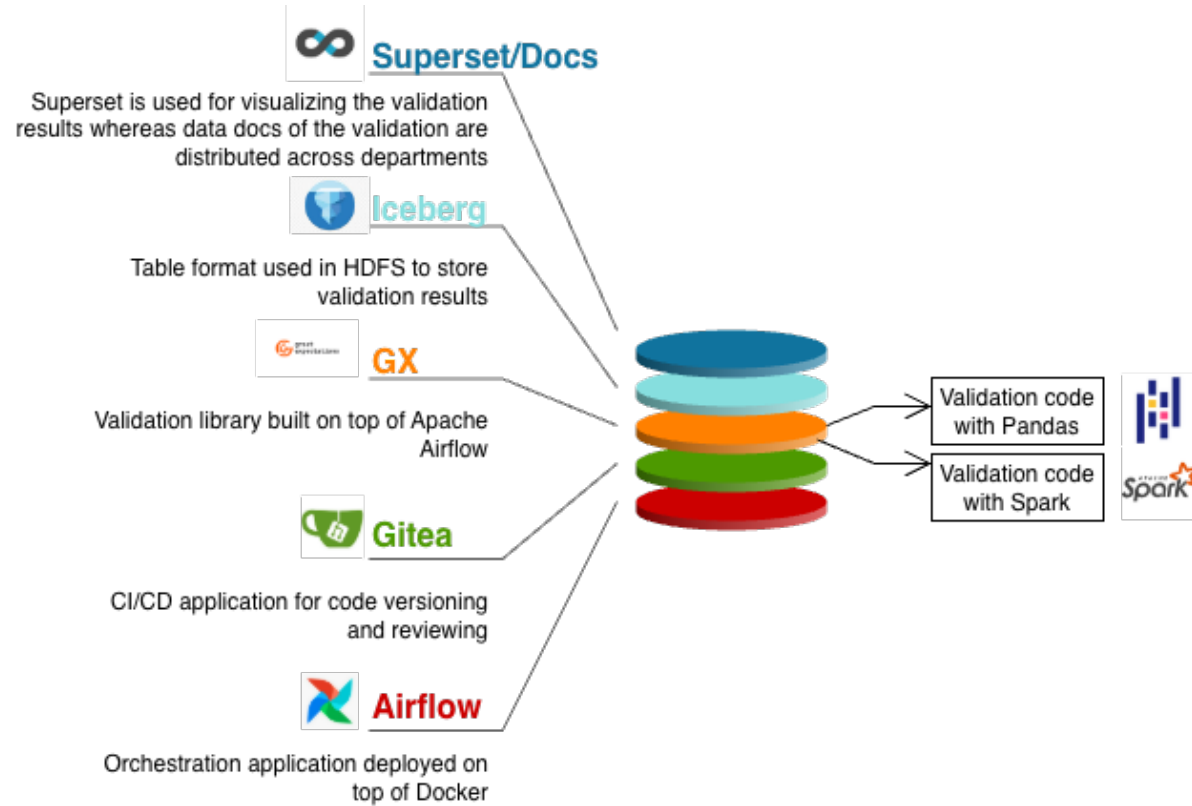
Accuracy · Reliability · Timeliness · Consistency

DATA

TRUST

Validated before
it becomes a fact

ARCHITECTURE



Superset visualizes

Iceberg preserves results

GX validates

Gitea versions

Airflow orchestrates

VALIDATION LOGIC

While GX Built-in rules may be used more frequently and easily, custom SQL controls can also be registered to the framework and execute as if they are GX rules.

BUILT-IN EXPECTATIONS

- **Completeness · NULL constraints**
- **Uniqueness · composite keys**
- **Validity · sets and regex**
- **Timeliness · dynamic freshness**
- **Consistency · paired-column checks**

CUSTOM SQL EXPECTATIONS

Cross-field dependency

If one amount exists, its related amount must also exist.

Data mapping

Country, bank and other relationships remain consistent.

Economic plausibility

Amounts stay within exchange-rate tolerance ranges.

The same suite can run by different distribution groups and periods through configuration.

PROCESSING

Pandas keeps smaller slices simple; Spark distributes large scale data across the cluster.

PANDAS

≈500,000

rows · 9 expectations

5.57 minutes

APACHE SPARK

≈5,000,000

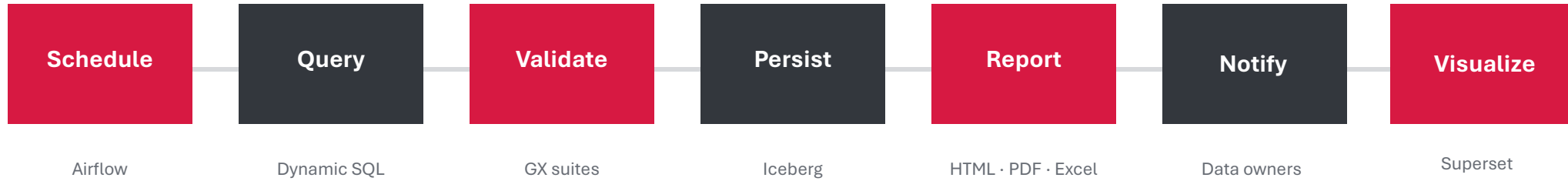
rows · 9 expectations

2.42 minutes

4 executors · 8 GB memory · 4 cores each

CHECKPOINT LIFECYCLE

In built-in expectations validate step is simpler and query step is irrelevant while in custom SQL we use queries first then GX validates. Results are written into Iceberg tables to later be used for dashboards.



Failure → notify data owners with failure alert

OBSERVED OPERATIONAL IMPACT

METRIC	BEFORE AUTOMATION	AFTER FRAMEWORK DEPLOYMENT
First-pass validation	Several hours · manual SQL	Minutes · automated GX
Coverage	Limited · ad hoc	Hundreds of expectations
Scalability	Single-node	Distributed Spark
Auditability	Logs	Iceberg + reports
Publication risk	Human-dependent	Systematic alerts

Automation improves speed, coverage and accountability.
The framework shifts data quality from downstream inspection to a governed production capability.

WHAT THIS CHANGES

Data quality becomes infrastructure for trust

- | | | |
|----|---|---|
| 01 | Scale validation without losing control | Pandas–Spark execution supports both focused and large-scale workloads. |
| 02 | Make every result traceable | Versioned rules, Iceberg results and generated reports create an auditable chain. |
| 03 | Move quality checks before dissemination | Critical expectations can protect statistical releases before errors propagate. |
-

NEXT **Kubernetes-native scheduling · AI-assisted expectation discovery · anomaly and drift detection · OpenMetadata access**



From Public Good to Public Input: Extending Data Governance in the Age of AI Intermediation

Archana Dilip
Assistant Adviser
Reserve Bank of India

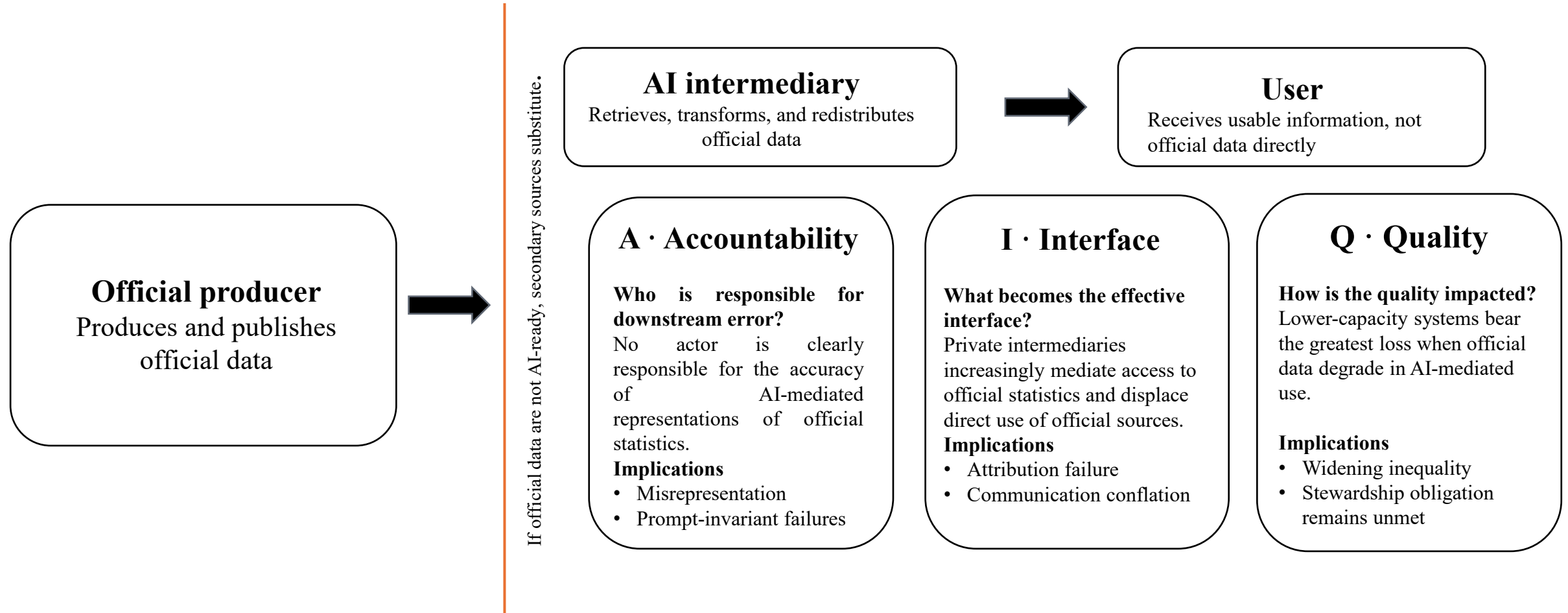
The Last-Mile Problem in Official Data

Official data are produced correctly, but the bridge to actual user understanding is fragile and fragmented.



The concern for central banks: official statistics may be produced well, yet fail to function as a public good in practice when the last mile is broken.

Gaps manifested by the last mile problem



Why AI misrepresentation matters?

Monetary policy operates through expectations



Financial stability surveillance depends on accurate cross-country data

International economic governance becomes miscalibrated



AI misrepresentation threatens central bank credibility

Overview of the cross-country audit

Datasets considered		
S.No.	Basket A: Internationally redistributed series	Basket B: Nationally anchored series
1	Policy interest rate — current official rate	Overnight and standing facility operations by instrument
2	Central bank total assets	Real estate sector non- performing loan ratio
3	Aggregate current account balance	Retail payment transactions by instrument
4	Headline consumer price index (CPI) and core CPI	Outstanding banking system liquidity
5	Broad money as a per cent of gross domestic product (GDP)	Active macroprudential instrument parameter

Central banks considered	
S.No.	Central bank
1	Banco Central de Bolivia
2	Banco de México
3	Bank Indonesia
4	Bank of England
5	Bank of Ghana
6	Bank of Japan
7	Centrale Bank van Suriname
8	European Central Bank
9	Federal Reserve
10	Nepal Rastra Bank
11	Office des Changes Morocco
12	Reserve Bank of India

LLMs considered: GPT-5.5, Gemini 3.0 Flash, Claude 4.6 Sonnet.

Common failure types encountered

Open data gap	Accountability	Users are pushed away from official channels
Metadata gap	Quality	Misinterprets units, period, method, or revision status
Standards gap	Quality	Assumes the wrong statistical framework
Language gap	Interface	Substitute toward secondary English-language sources
Machine readability gap	Interface	Secondary intermediaries take over
Vintage & revision gap	Quality	Historically correct but currently wrong information
Attribution gap	Accountability	Official producers lose visibility and cannot verify authority

Extending data governance

Publish key series in AI-ready formats

Dissemination

Open data, Standards, Metadata

Build revision and provenance architecture into the data system

Production, Dissemination

Attribution, Standards, Metadata

Conduct periodic AI-readiness and representation audits

All

Audit source displacement and communication conflation

Attribution, Vintage and revision

Use web analytics, server logs, and API monitoring

Open data, Attribution

Introduce provenance tracing tools where feasible

Attribution, Vintage and revision

Establish correction and incident reporting channels

Monitoring
&
Feedback

Recommendations



International Organisations

Develop an AI-readiness dissemination standard

Develop an official statistics provenance framework for AI

Develop international standards for AI intermediaries

Release more AI ready datasets on behalf of national agencies



Public-Private Partnership

Create public interest AI access agreements

Establish attribution and fidelity expectations for AI intermediaries

Require disclosure for derivative AI products

Use licensing and API registration to create accountability

Establish correction, escalation, and feedback mechanisms with AI providers

Some cost-effective recommendations

Add structured metadata markup to key statistical release pages

Publish a simple machine-readable file alongside every PDF or HTML release.

Publish machine-readable advance release calendars and update notices

Adopt an AI communication policy distinguishing statistics from narrative communication

Create a collective monitoring platform at the national level

Commission independent third-party audits

Prioritising AI-readiness for important datasets not captured by international organisations

Thank you!



Central Bank of Mauritania

Data governance and the use of alternative data for policy-making

13th Biennial IFC Conference, BIS, Basel - Switzerland

August 2026

Data Governance - Modernisation under way

A decade of financial-infrastructure reform now feeds one governed data platform

Real-time payment data since 2023 in place

The MACSS national payment system provides real-time data on interbank flows, strengthening the traceability of transactions.

A modernised financial infrastructure in place

QCBS core banking, ATS-CSD, the BSA supervision platform, Refinitiv and Swift Scope 2. Each system produces granular data, standardised at source.

Risk oversight across the sector in place

Supervision of commercial banks and microfinance institutions, with prudential reporting automated through the BSA platform.

Data governed end to end in place

API collection from the banks, quality scoring of every submission, centralisation in a data warehouse paired with a business intelligence platform, systematic cross-checking of sources and decision dashboards for the general directorates.

Digital money data under way

Card payments (GIMTEL) and e-wallet data are being integrated, together with the balance of payments and ANSADE statistics.

Next: eight more sources

CSD · QCBS · Refinitiv · Swift · BSA · Bloomberg · IMF · World Bank

Each one a source the Bank will no longer have to reconcile by hand.

Data Governance - Some figures — Data collected from the banks

One collection channel, one referential, one governed history since January 2024

78 M+

records collected

Every declaration filed by the primary banks since the platform went live, centralised and version-controlled.

17

banks in scope

14 reporting regularly

10

declaration APIs

the E07 to E11 series

Jan 2024

platform live

Collected through the API gateway, quality-scored on arrival, then centralised into a single governed history.

49,500

accounts mapped

to the Bank’s referentials

144

mappings produced

covering 16 of the 17 banks

WHAT THE COLLECTION COVERS

E07 FS · E10 FE

Daily outgoing and incoming foreign-currency flows

E08 OCD · E09 PE

Documentary credit openings and maturity forecasts

E11 RDCC

Correspondent account statements and closing balances

Base Clientèle

Clients’ income and profession, firms’ turnover and business group — 44 new fields

Data Governance - Some figures — Digital money and inclusion

Two years of wallet activity, and what it says about financial inclusion

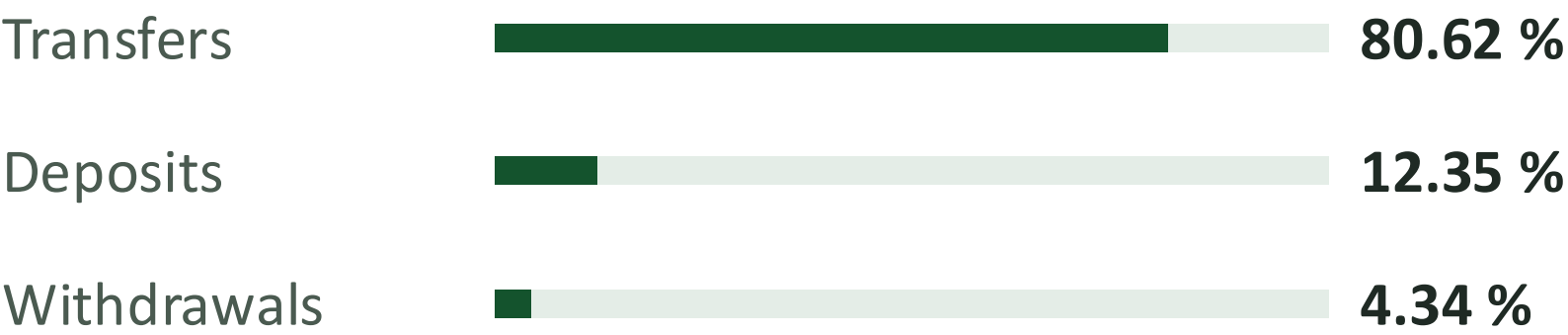
E-WALLETS · JULY 2024 TO JULY 2026

4.57 T

MRU exchanged through seven wallets

1.76 billion transactions over the period — 2.23 million a day.

Share of value by transaction type



Transfers dominate: the wallets are used to move money between people, not only to cash in and out.

Seven wallets in the market

bankily · masrvi · sedad · bim-bank · click · amanty · bci-pay — bankily carries the bulk of the value.

FINANCIAL INCLUSION

5.00 M

accounts held by 1.91 million clients

3.204 million active, 1.785 million dormant — 2.62 accounts per client.

Global inclusion score



One account in three lies dormant — the next gain in inclusion is in usage, not in opening more accounts.

Where it comes from

Real-time payment data since 2023 — the inclusion rate rose from 20.8 % in 2022 to 55 % in 2025.

Data Governance - Data cross-validation

Rebuilding our own reports from raw data, then comparing them with declarations

Exercise	Banks processed	States rebuilt	Match with the banks' declarations	
Credit registry Rebuilt from the monthly detailed balance sheets Declarations match the rebuilt registries almost perfectly: the collected data is reliable enough to reproduce a regulatory report.	3	18	99 %	
Monetary market Weekly deposit positions rebuilt from raw balances, automated through a BI report Here the gap is wide — and that is the finding. Each divergence points to a mapping, a practice or an account classification to correct with the bank.	12	+200	45 %	

Cross-checking is the only way to tell a data gap from a reporting gap. The contrast between the two exercises says precisely where declarations diverge from the source data, and where remediation with the banks should start

Data Governance - Data cleaning with AI

A progressive cascade: each stage removes the cases the next one need not see

01

Data cleaning

Format normalisation and standardisation — spaces, case, variants

02

Direct mapping

Deterministic correspondence through the reference table

03

Vector embedding

Semantic inference on abbreviations, misspellings and language variants

04

LLM classifier

Automatic classification of the residual complex cases only

CLEAN DATA RATE BY FIELD, ONCE THE CASCADE HAS RUN

Currency	E09 PE · E10 FE	7.67 M		99.99 %
Country	E10 FE	2.97 M		99.98 %
Country	E07 FS	960 K		99.84 %
Named fields	E08 OCD	12 K		98.81 %
Named fields	RDCC	321 K		91.16 %
Economic nature	incoming flows	2.99 M		89.58 %
Economic nature	outgoing flows	1.00 M		55.53 %

THE FIELD THAT PROVED THE POINT

Country is a free-text field. On arrival,

92.92 %

of values needed cleaning. After the cascade, more than

99.8 %

are clean — with no manual intervention.

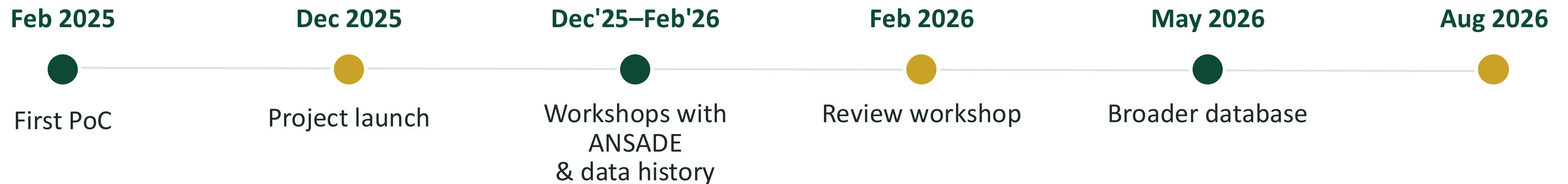
Run entirely in-house

Python · multilingual sentence transformer (mpnet) ·
Qwen 2.5 : 14B · Ollama

The model runs locally on the Bank’s own infrastructure — no banking record ever leaves it.

Inflation nowcasting — Filling data gaps with alternative data

An AI-based nowcasting model for inflation, combining earth observation with international data



Turning free earth-observation data into policy-grade macroeconomic signals

Inflation nowcasting — A homegrown data infrastructure

Earth-observation proxies, engineered locally, combined with international signals

Earth-Observation Proxies

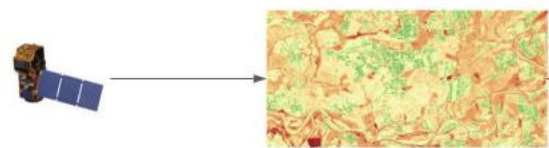
-  **Night-time lights** – Local economic and industrial activity
-  **SAR ship detection** – Fishing and port trade activity
-  **Cement thermal signal** – Cement and real-estate activity
-  **Vegetation indices** – Agricultural output
-  **Built-up indices** – Urban growth
-  **CO₂ / NO₂ emission signals** – Transport and industrial output
-  **Drought indices** – Climate stress on livestock

International Connectors

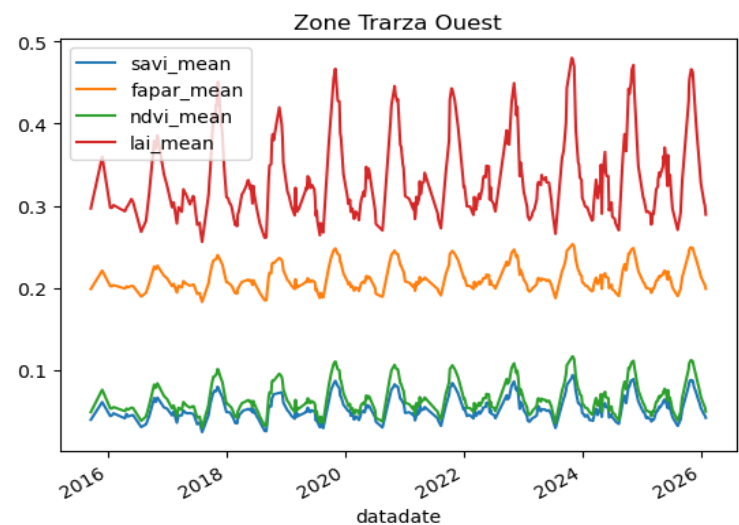
-  **5 partner-country trade links** – France, Spain, Germany, Belgium, Italy
-  **US macro indicators** – CPI, PPI, M2, Federal Funds Rate
-  **European statistics (Eurostat)** – Trade and macro indicators of partner economies
-  **Global equity indices** – S&P500, CAC40, DAX, NIKKEI
-  **Global commodity prices** – Brent oil, gold, silver, copper

Inflation nowcasting — From raw signals to indicators

From raw earth-observation signals to locally engineered indicators



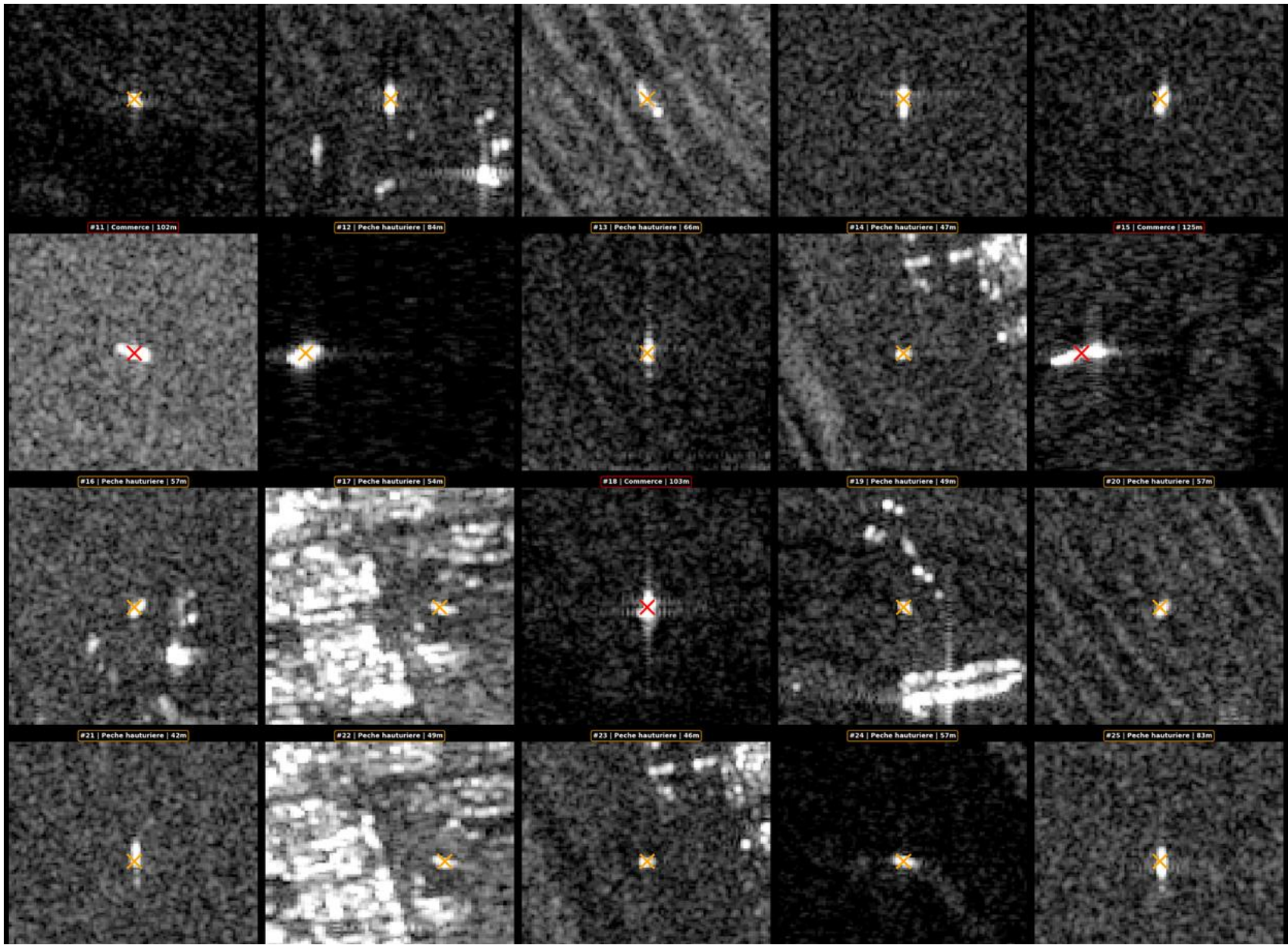
Satellite-to-signal processing pipeline



Vegetation indices — Trarza Ouest



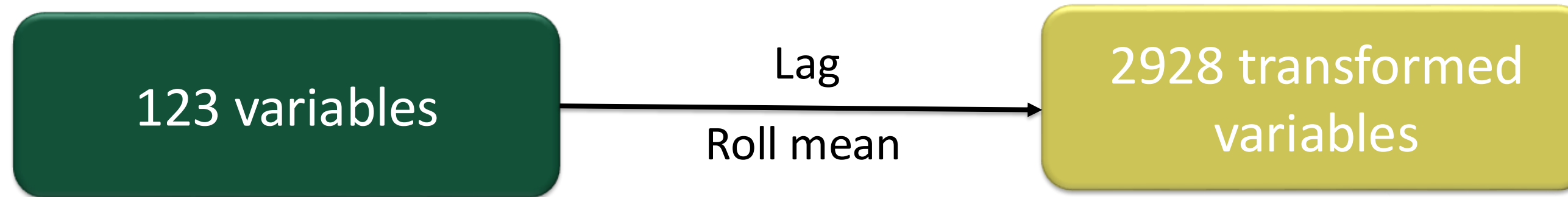
Cement thermal signal — Industrial site



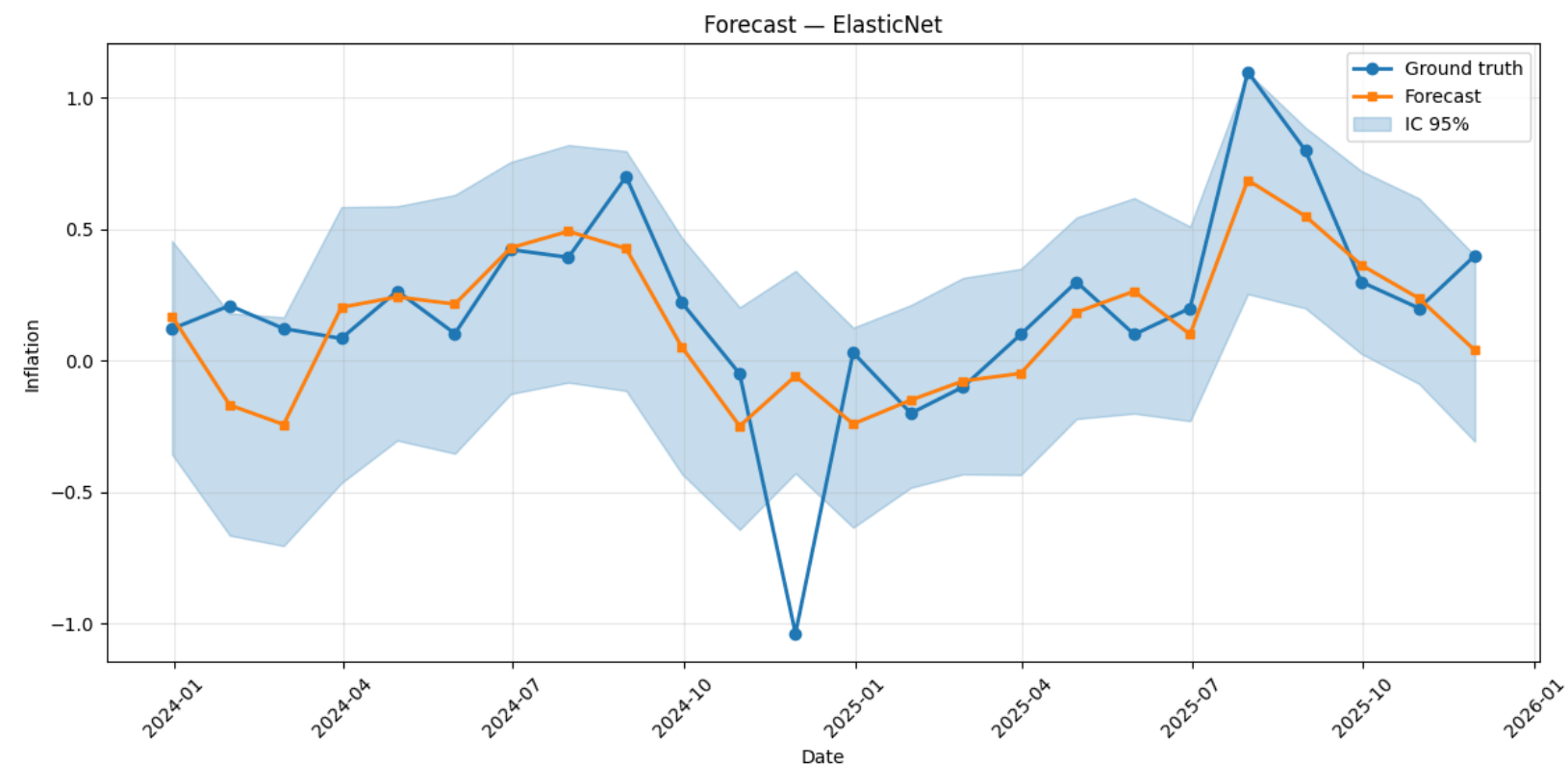
SAR Ship Detection — Fishing & Port Trade Activity

Inflation nowcasting — Two feature-selection pipelines

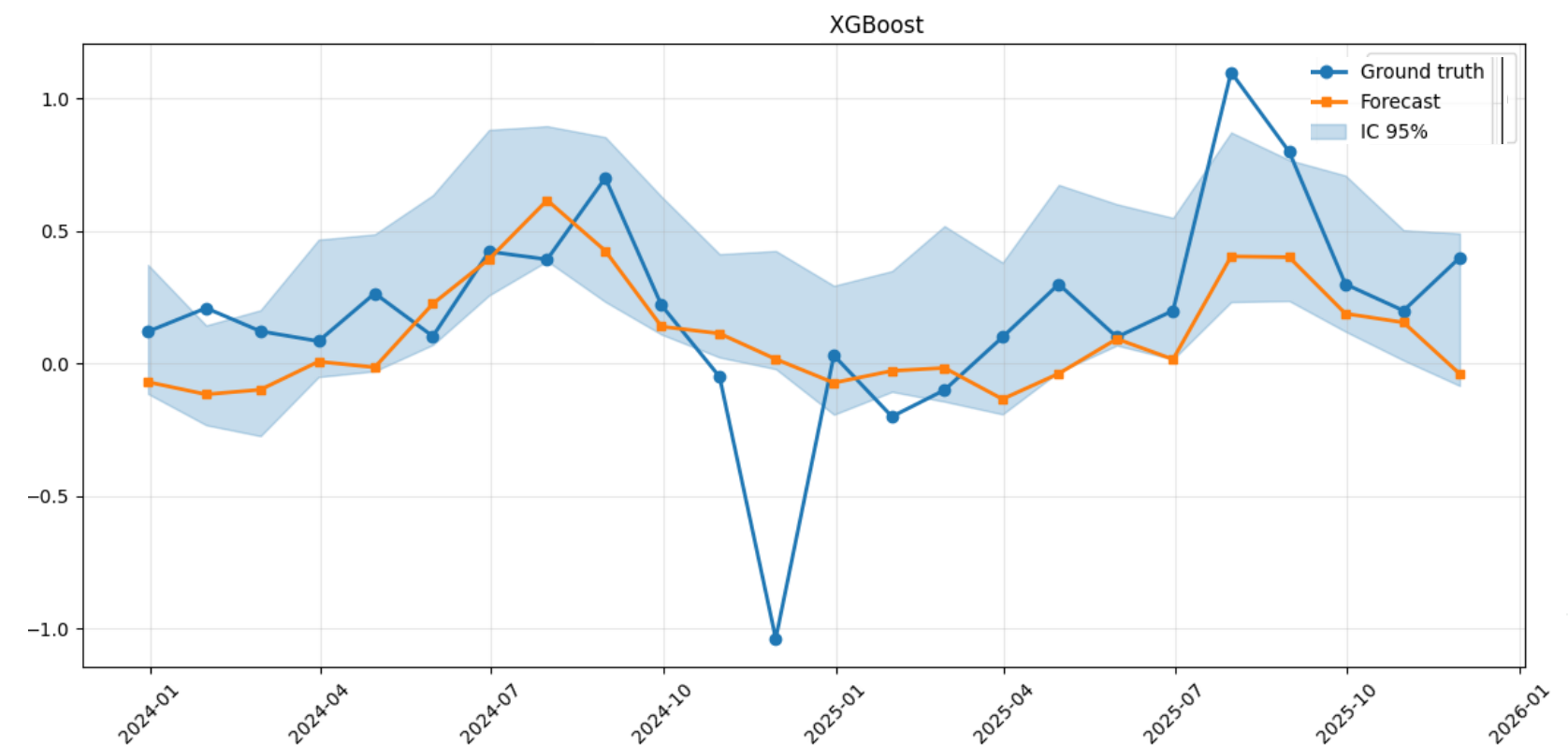
Eight model families tested (deep learning, machine learning, statistical) — ML performs best



*Selection based on **directional temporal relationships** : 23 vars*



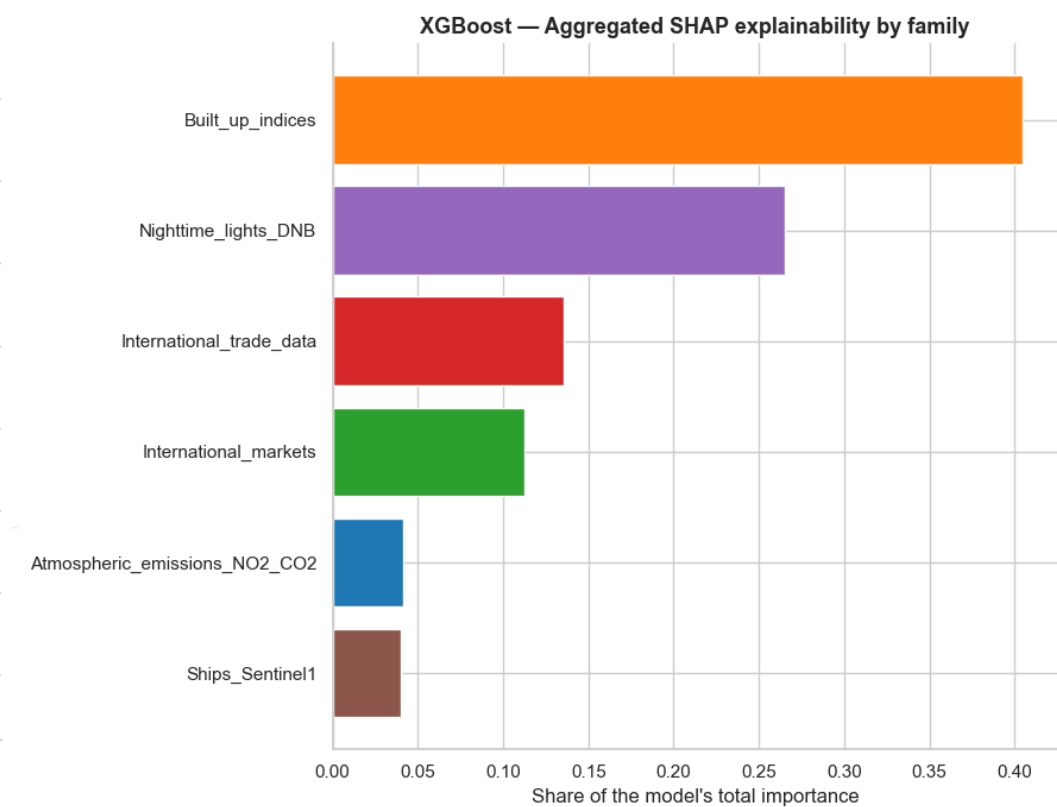
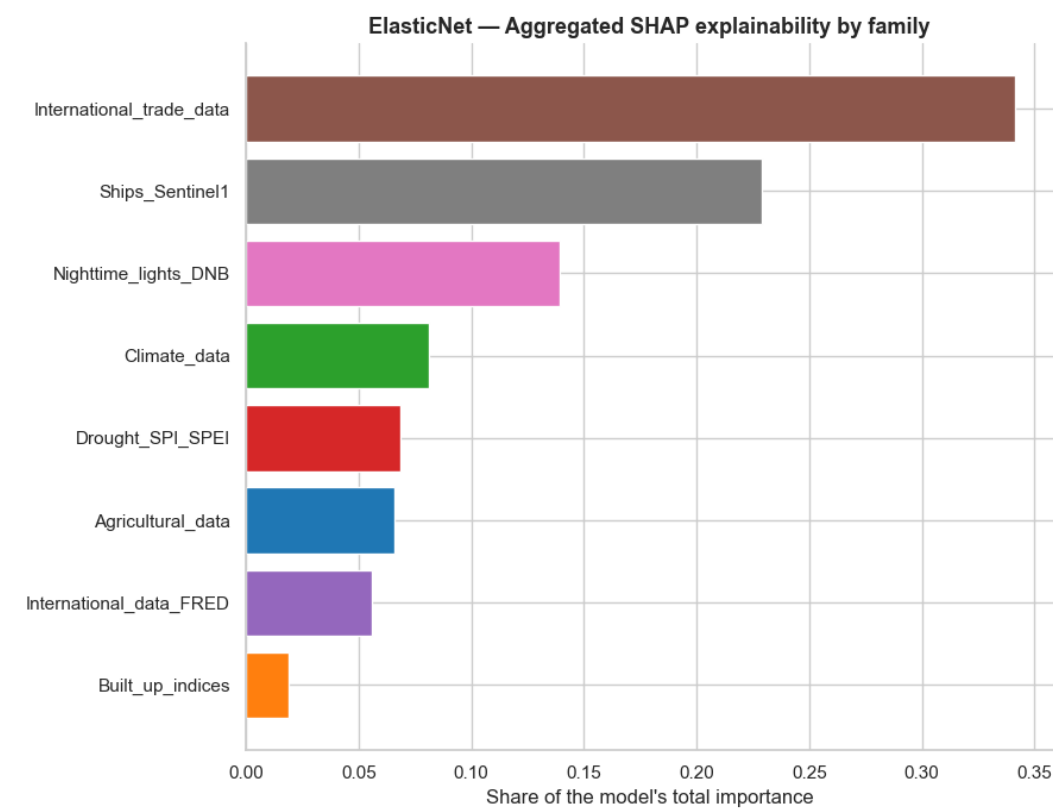
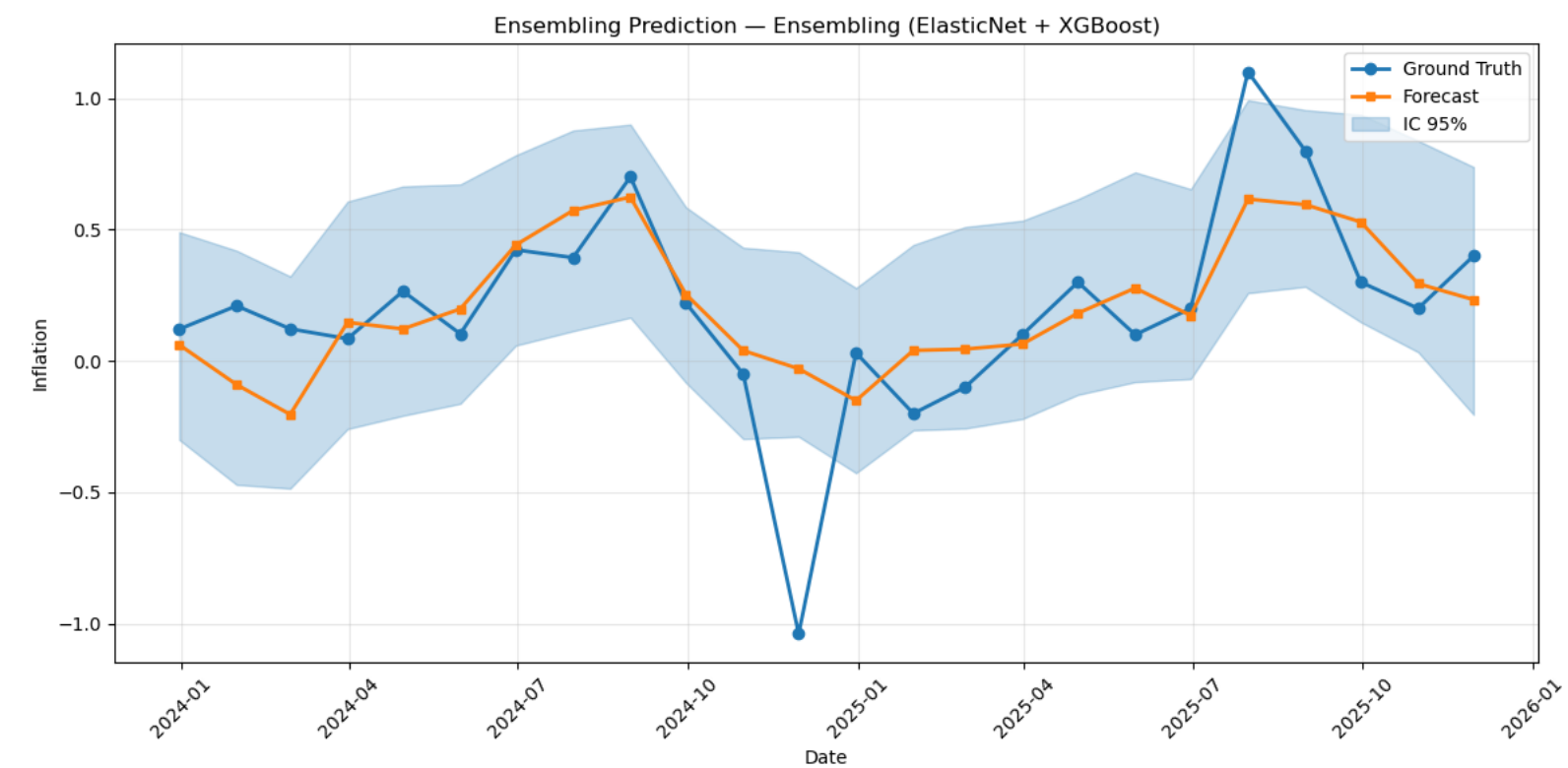
*Selection based on **statistical dependency** : 12 vars*



Inflation nowcasting — Ensembling and explainability

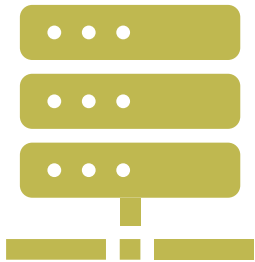
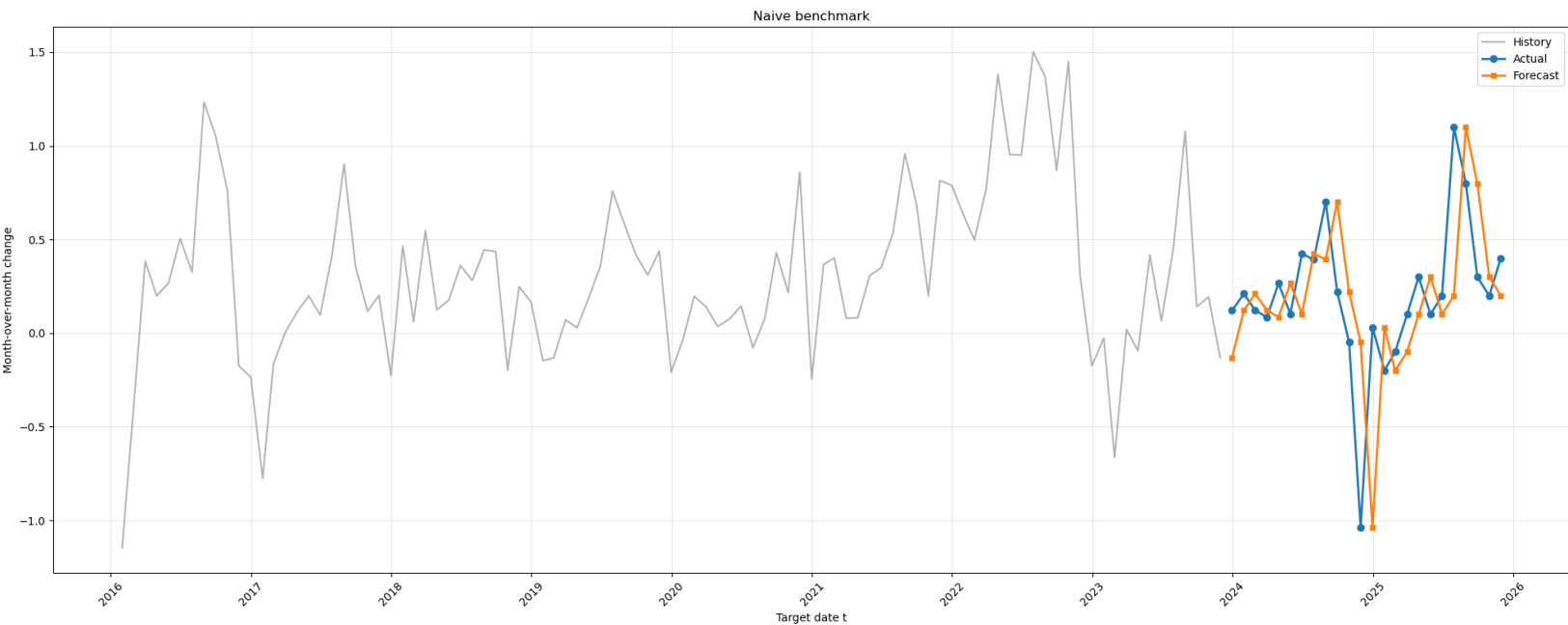
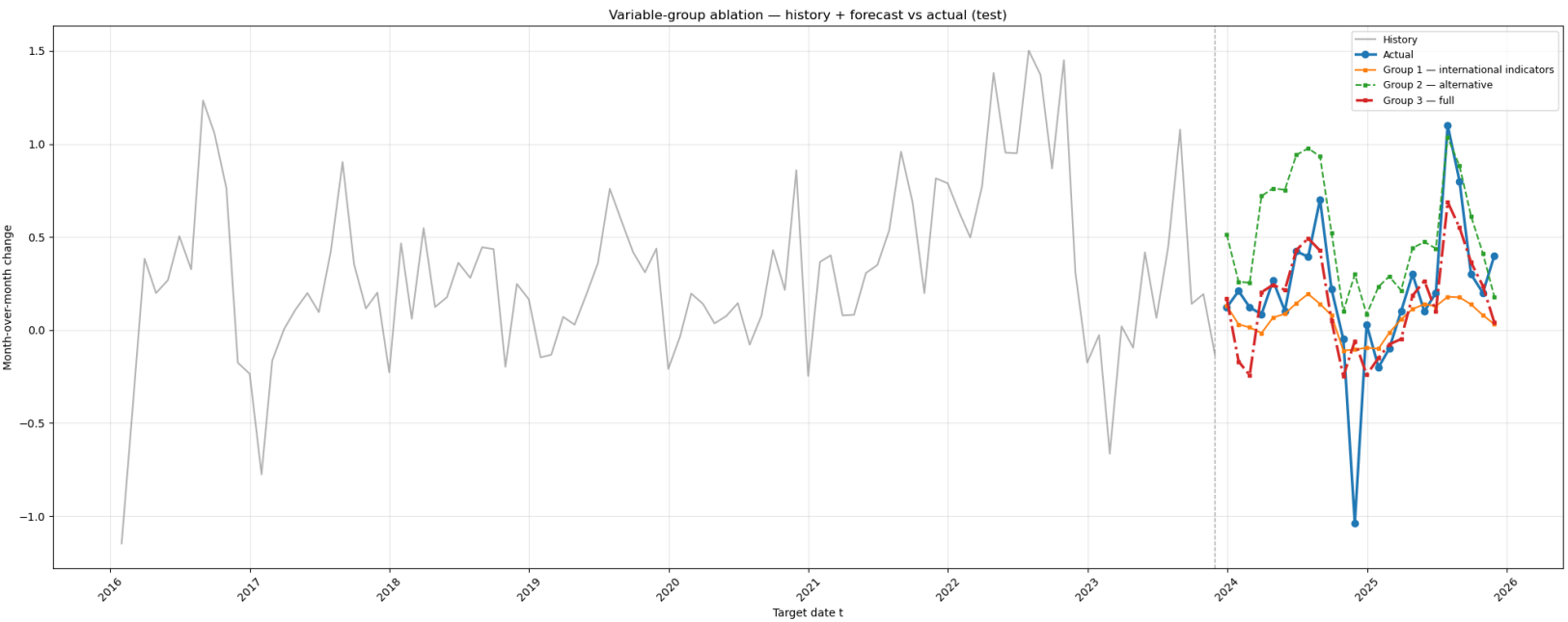
Two pipelines combined: the ensemble beats each model taken alone

	MAE	RMSE	MASE
Pipeline 1 - ElasticNet	0.1988	0.2841	0.6482
Pipeline 2 - XGBoost	0.2331	0.3182	0.7602
Ensembling	0.1872	0.2756	0.6109

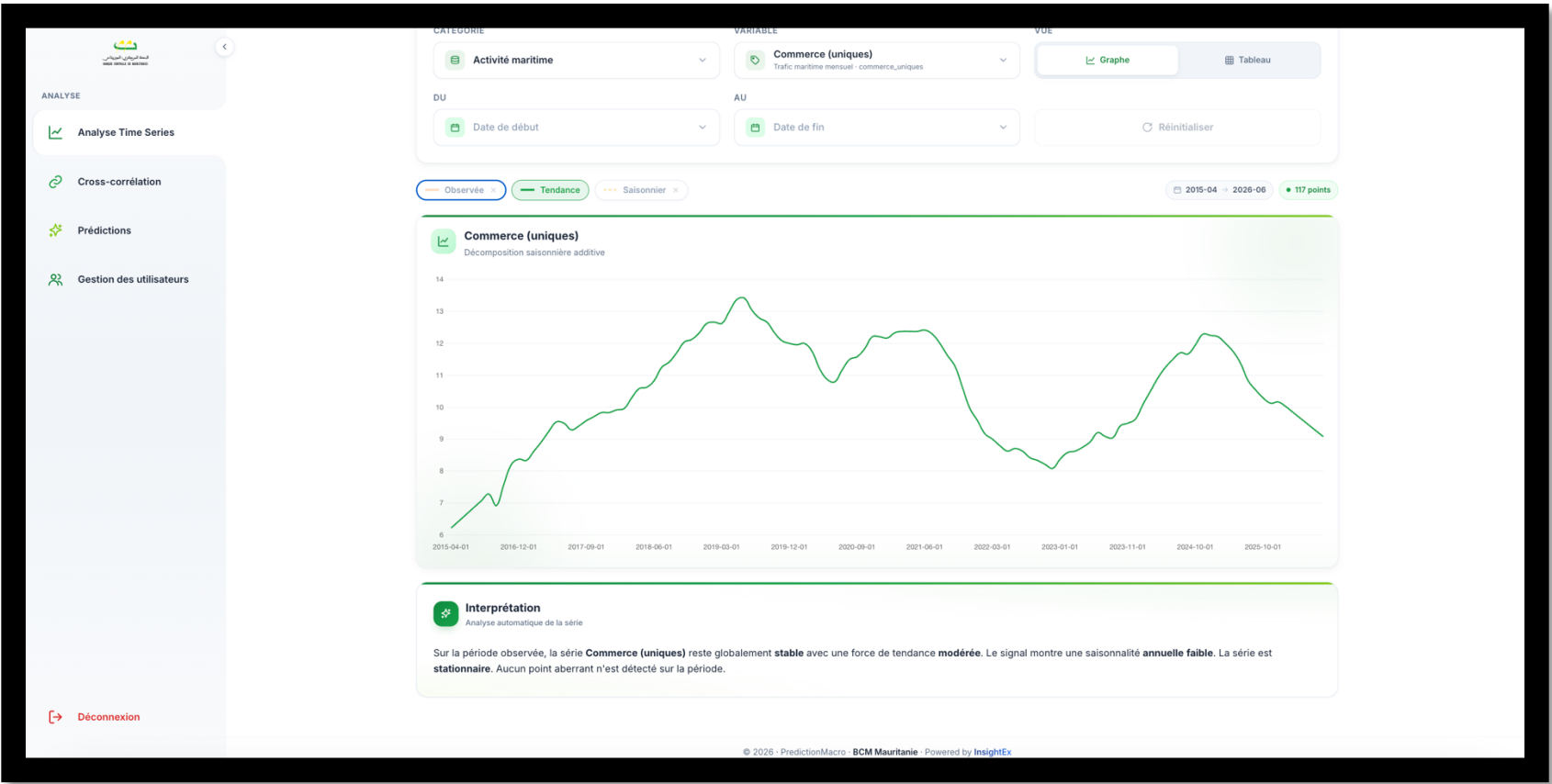


Inflation nowcasting — Robustness tests and deployment

Satellite data significantly improve the prediction of large inflation moves



A web service collecting data 7/7



A frontend for end-users

13TH BIENNIAL IFC CONFERENCE

"FILLING INFORMATION GAPS IN CENTRAL BANKING: EFFECTIVELY
LEVERAGING DATA AND INNOVATIVE TOOLS"

SESSION 8 – THE QUEST FOR BETTER, MORE AND FASTER DATA

AI-BASED TOOL FOR AUTOMATED EXTRACTION OF HETEROGENEOUS STATISTICAL DATA

EUGENIA KOBLENTS, STATISTICS DEPARTMENT

DATA SCIENCE EXPERT

21 AUGUST 2026

BANCODE **ESPAÑA**
Eurosistema





ÍNDICE

1. Motivation and goal
2. Previous work
3. Architecture and main modules
4. Results exploration and validation
5. Conclusions and future work

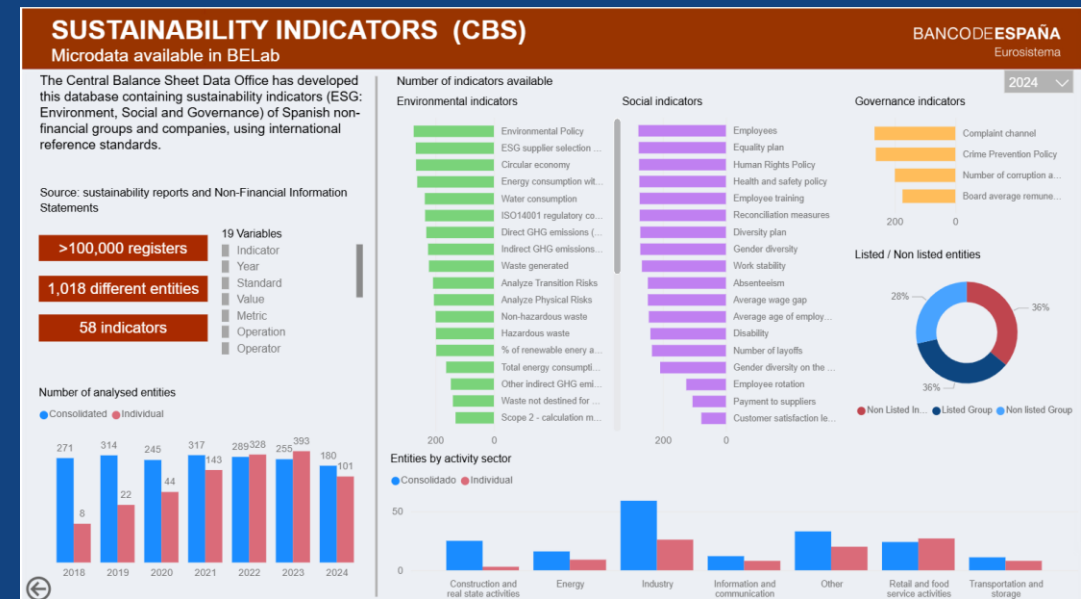
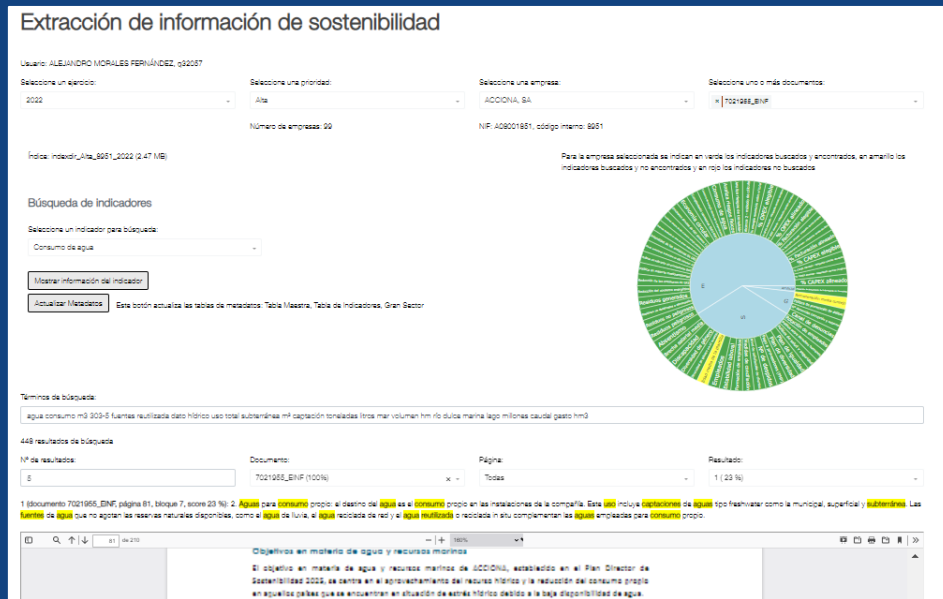
MOTIVATION AND GOAL

- ❑ The objective of this project is the design and development of a **generic tool** for the **automated extraction of heterogeneous statistical data** from large **collections of documents**, leveraging **AI** techniques, built-in quality controls, and support for human validation of the extracted information.

- ❑ This is a strategic project for the Statistics Department, and **multiple use cases** have been identified:
 - The initial use case focuses on the **extraction of 60 sustainability indicators** from non-financial statements published by Spanish non-financial corporations.
 - Extraction of **financial and accounting information** for the Central Balance Sheet Database
 - Automated **economic activity and institutional sector** classification of companies, including holdings and financial headquarters

PREVIOUS WORK

- ❑ Between 2021 and 2023, the team built a **web application for semi-automated information extraction and storage** to support business users in the retrieval of sustainability indicators.
- ❑ As a result, a **structured micro-dataset** with more than 130,000 data points was created. For each data point, both the **extracted value** and its **exact location** within the source document are available.
- ❑ The resulting dataset is available to researchers through the Banco de España's **BE Lab** data laboratory.

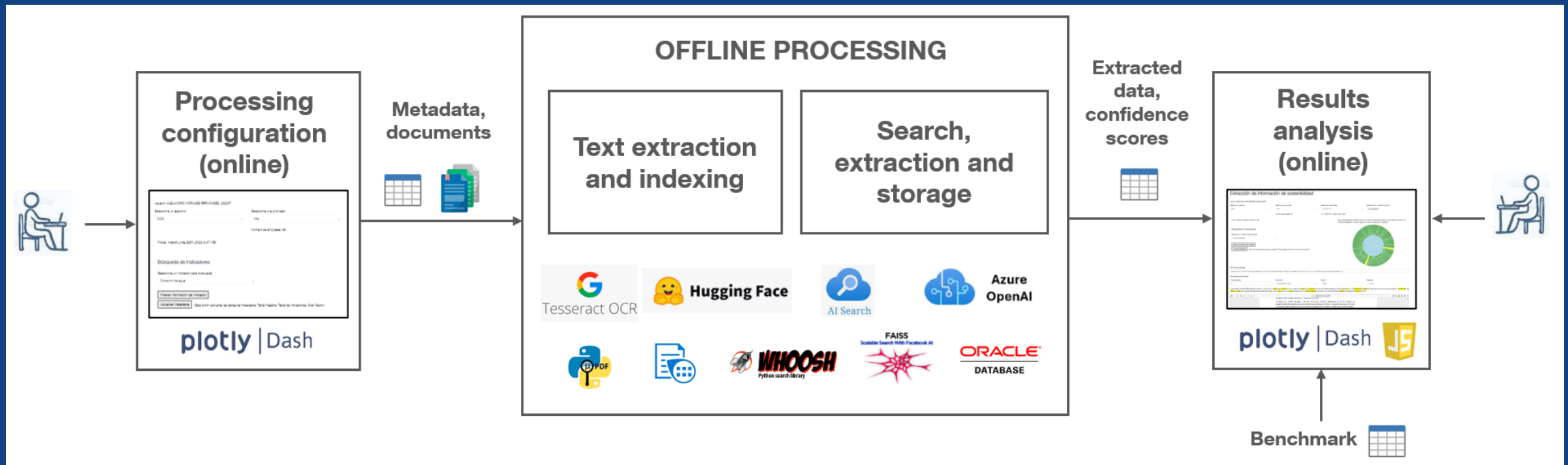


- Creation of a structured sustainability database from company reports – a web application prototype for information retrieval and storage
- Micro-database for sustainability (ESG) indicators developed at the Banco de España

NEW AI-BASED TOOL: ARCHITECTURE AND MAIN MODULES

□ The implemented workflow is organized into three main stages (Python):

1. **Processing configuration:** the user selects the relevant metadata, documents, and configuration parameters.
2. **Offline processing:** the system automatically extracts the data of interest from the documents.
3. **Results analysis:** the user reviews, validates, and, if necessary, corrects the extracted data.



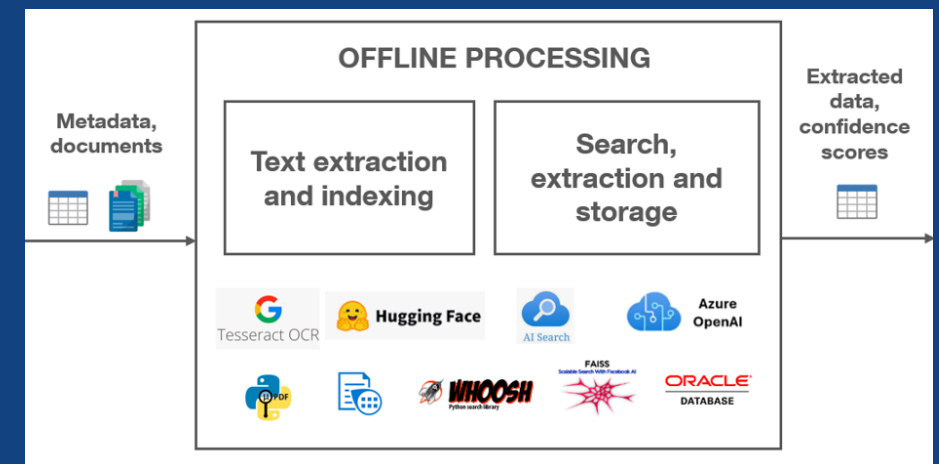
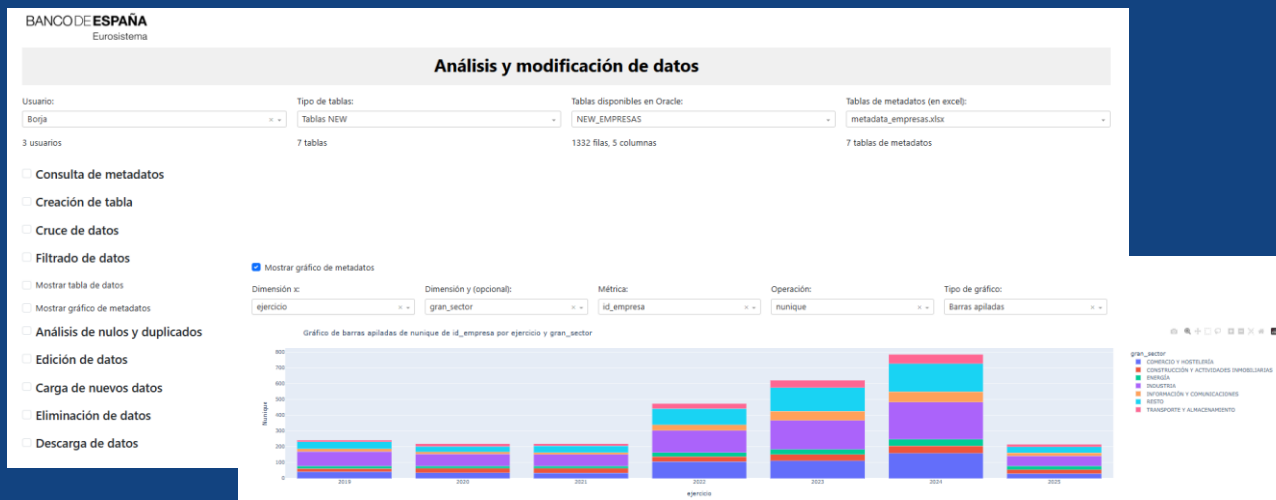
NEW AI-BASED TOOL: ARCHITECTURE AND MAIN MODULES

❑ Processing configuration stage:

1. Data and metadata consultation
2. Data merging and filtering
3. Data visualization and exploration
4. Data editing and deletion
5. Data upload and download
6. Processing launch and monitoring

❑ Offline processing stage:

1. Digital and scanned text extraction (OCR, pymupdf)
2. Indexing and search:
 - Traditional keyword-based search (whoosh).
 - Planned migration to Azure Search.
3. Tables processing with Azure Document Intelligence
4. Data extraction with Azure OpenAI (gpt-5-mini)
5. Storage to Oracle database



RESULTS EXPLORATION AND VALIDATION

- ❑ An **interactive web-based user interface** has been implemented to support users in the analysis and validation of the extracted data.
- ❑ The tool presents a summary of the extraction results together with key quality and diagnostic metrics, such as the number of extracted values, agreement with ground truth, total cost, and latency.
- ❑ It also allows users to identify and select non-matching values for detailed inspection and review.

Made with AI

BANCODE **ESPAÑA**
Eurosystem

Document Information Extraction

File:
sos_ia_comparison_gpt5_diagnostics_revision.csv

Period: All Company type: All Company: All Indicator type: All Indicator: All

Text quality: All Prompt quality: All Metadata quality: All Data source: All Calculation type: All

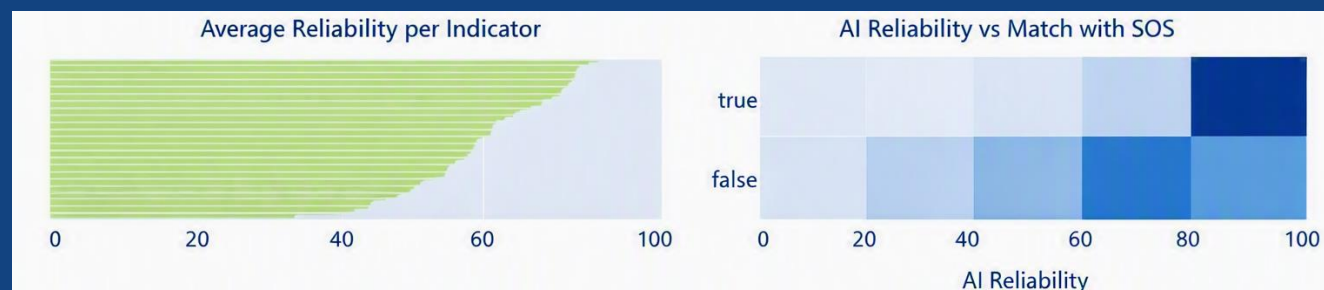
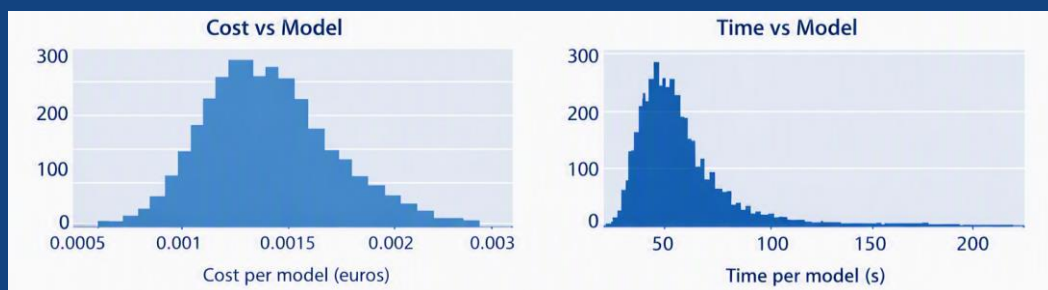
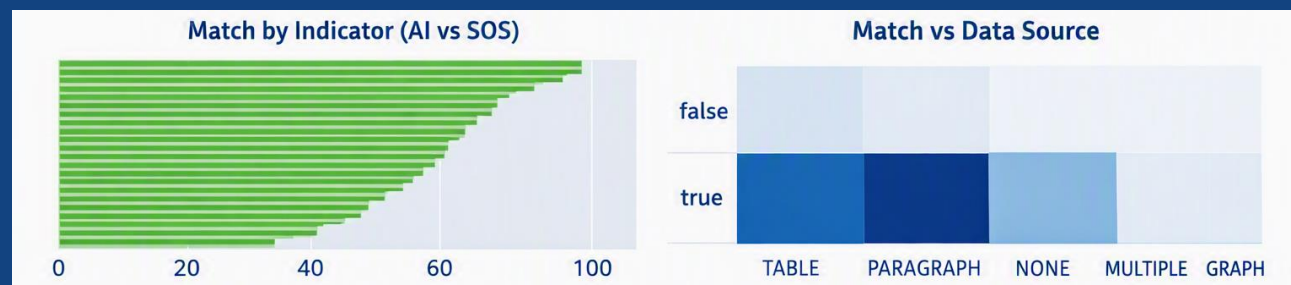
Model: All Value match: Both Reviewed: All ☐ Filter missing ☐ Filter categoricals

Reviewed: All ☒ Soft match ☒

Nº of periods: 2	Nº of companies: 86	Nº of compenis: 86	Nº of documents: 88	Nº of values: : 4985
Found SOS: 74%	Found AI: 69%	Found OC: 74%	Nº of indicators: 58	Nº values: 4985
Total cost: 8.29 €	Total time: 34.19 h	Total time: 34.19 h	Filler match: 80%	Match all: 29%

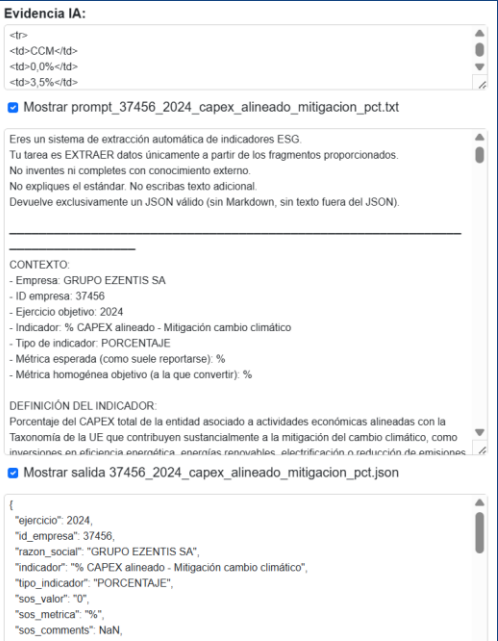
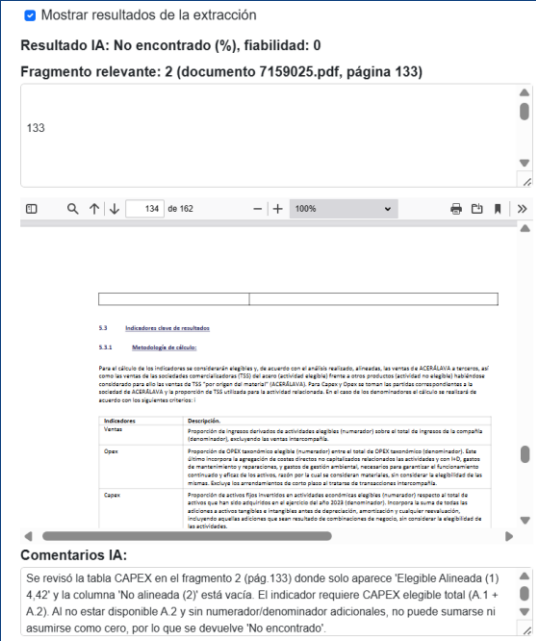
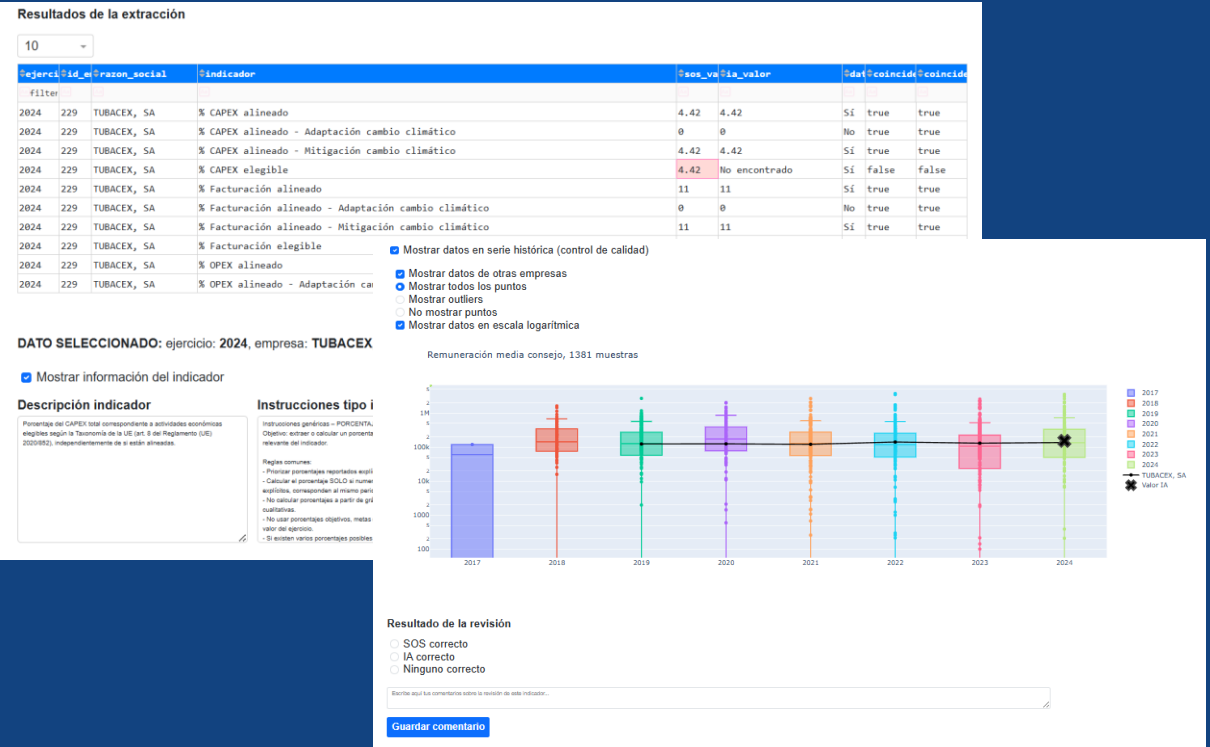
RESULTS EXPLORATION AND VALIDATION

- ❑ The tool offers diagnostic visualizations of document-extraction quality and consistency.
 - **Quality metrics:** text, prompt, and metadata quality; data origin; calculation types; indicators needing review, etc.
 - **Coincidence analysis:** rates by company and indicator; correlations with origin, calculation type, and AI confidence; distributions of pages, length, cost, and time.



RESULTS EXPLORATION AND VALIDATION

- ❑ Enables full contextual review of each data point, including metadata, PDFs, prompts, and LLM outputs.
- ❑ Allows to detect both human and AI-generated extraction errors, supporting iterative refinement of ground truth, metadata, and prompts.
- ❑ Provides QC visualizations, showing indicator trends over time and comparing manual and AI-extracted values for validation.



CONCLUSIONS AND FUTURE WORK

□ Summary and challenges:

- An MVP has been developed and tested using 5,000 manually annotated data points.
- An agreement of 80% between AI and ground truth has been achieved, with a target of > 90%.
- One of the main challenges faced has been the metadata and prompt refinement.
- The tool is used to support quality control of human-annotated data.

□ Future work:

- Improve core **extraction reliability**
- Strengthen **validation and review** workflows
- Validate at scale and **optimize quality**

□ International collaboration

- **Leveraging the IFC community** to foster collaboration among institutions with similar needs, promoting the exchange of experiences, best practices and joint developments.



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

Thank you for your attention!