ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Social Statistics Section

Data Integration for Official Statistics: Traditional and Big Data Sources.

22 January 2026

Istat

Speakers

15:00–15:10 (Italy) / 2:00–2:10 (UK)
Data integration in Istat
Elena Grimaccia (Senior Research Fellow, Central Directorate of Demographic Statistics and Population Census (DCDC), Istat)

15:10 –15:30 (Italy) / 2:10 –2:30 (UK)
Italian Population Census: Register-Based Population Size Estimation
Fabrizio Solari (Senior Research Fellow, Central Directorate of Demographic Statistics and Population Census (DCDC), Istat)

15:30–15:50 (Italy) / 2:30–2:50 (UK)
Non-Traditional Data Sources for Official Statistics: Insights from AIS Data
Marco Di Zio (Research Manager, Central Directorate for Methodology and Statistical Process Design (DCME), Istat)

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Social Statistics Section

Data integration in Istat

Elena Grimaccia

22 January 2026

Istat

Background

Growing surveys' costs

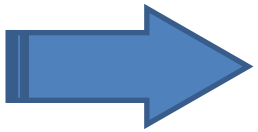
Declining response rates

Increasing demand for more timely and detailed statistics.

Need to ensure quality and sustainability of official statistics

Proliferation of new data sources

Growing capacity to store, process, and analyse ever-growing amounts of data



From direct observation to an integrated production system

Why Data Integration Matters

Data integration allows to:

- Combine multiple administrative registers and survey data to reduce survey burden while enhancing coverage.
- Produce timely, granular statistics for regional, social, and economic monitoring.
- Contain statistical burden on respondents and reduce the overall costs of official statistical production.

Integration improves statistical reliability and consistency while maintaining confidentiality standards (Solari, 2023; Righi, 2018).

Istat production process

Istat data production process has 3 main components:

- Sample surveys
- System of integrated registers (administrative data).
- Non-traditional data sources (big data).

System of integrated registers: Extended Population Register (EPR)

A Register-Based Framework for Population Statistics

Project Overview

- Integration of Population Register (PR) with multiple administrative data sources
- Foundation of the Permanent Population and Housing Census (PPHC)
- Annual population counts replacing decennial census

Key Innovation

Signs of Life (SoL) Methodology

Administrative data converted into simplified statistical indicators tracking individual presence in Italy across multiple sources

Audit surveys validate classification accuracy

Impact & Benefits

- Reduced costs and response burden
- Improved timeliness with annual data
- Enhanced data quality through integration
- Foundation for other statistical registers

Commuting Matrix for Work Purposes -1

New Estimation Methodology Based on Data Integration: From Census Counts to Statistical Modelling

Matrix Definition: Number of individuals moving between or within municipalities, at least 3 days/week, from residence to usual workplace
Use: Definition of Local Labour Systems 2021

Methodological Evolution
2011: Exhaustive Census (Complete population surveyed)
2021: Integrated Approach (Sample survey + administrative data + statistical modelling)

Borrowing Strength Approach
Since only a sample is directly observed, statistical models "borrow strength" from multiple information sources to estimate the complete matrix

Commuting Matrix for Work Purposes - 2

Integrated Data Sources

- **Base Register of Individuals:** demographics, employment status
- **2021 Census Sample Survey:** direct observations
- **Administrative work data:** INPS, tax records (UC, AT)
- **2011 Commuting Matrix:** historical patterns
- **Spatial variables:** distance, travel time, urban typology

Models account for:

- Distance between origin/destination
- Economic activity sector
- Municipal characteristics
- Historical patterns

Results

✓ **Achieved** Consistency with 2011 matrix and observed data

✓ **Achieved** Coherent Local Labour Systems definition

⚠ **Limitation**

Abroad commuting not estimated (insufficient data)

Trusted Smart Statistics in Istat

International Framework

European Statistical System (ESS) Milestones

- **Scheveningen Memorandum (2013)**
  - Launched the structured experimentation with Big Data within the ESS.
  - Focus on harnessing new data sources for more timely and reliable official statistics.
- **Bucharest Memorandum (2018)**
  - Introduced the forward-looking concept of Trusted Smart Statistics.
  - Established the Trusted Smart Surveys methodology, integrating traditional and new data sources.

United Nations

- **UN Committee of Experts on Big Data and Data Science for Official Statistics**
  - Develops **international standards and methodological guides** (e.g., for using mobile phone data in tourism, mobility, and migration statistics).
- **UN Statistical Framework for Measuring Sustainable Tourism (2024)**
  - Integrates economic, environmental, and social data across territories.

Quality Principles

- Privacy by design and by default (GDPR)
- European Statistics Code of Practice
- Transparency and verifiability
- Statistical confidentiality guarantee

Istat

Big Data Applications: From Experimentation to Official Statistics

OFFICIAL STATISTICS - In Production:
Consumer Price Indices
- **Scanner data from supermarket barcodes (since 2018)**
- **Revolutionized sampling strategy**
- **Real-time price monitoring**

Social Issues:
- **Social media analysis: Gender violence and hate speech detection**

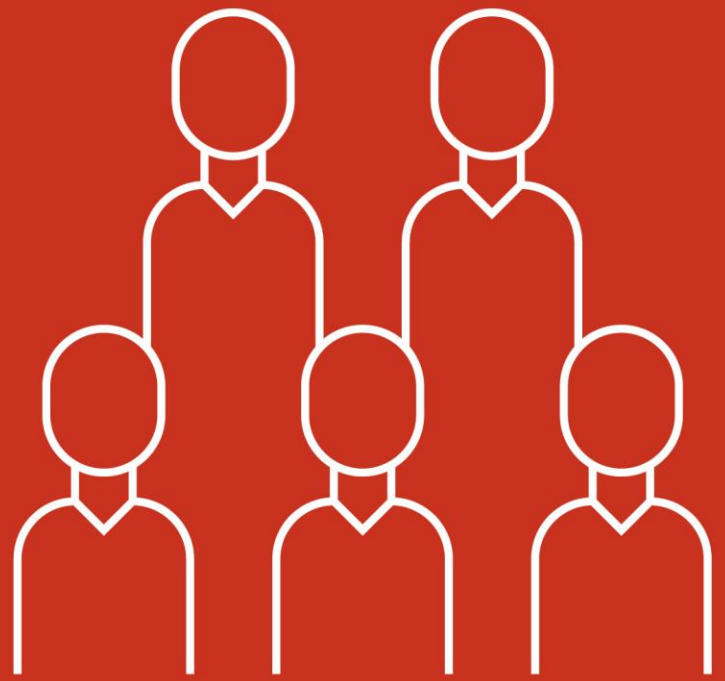
Social & Economic:
Social Mood on Economy Index: Daily sentiment from Twitter data (from Feb 2016)
Ukraine war impact analysis: Specific sentiment monitoring
Job market indicators: Web scraping from job portals and websites
Business web presence: Analysis of enterprise websites (10+ employees)

Territorial & Environmental:
- **Road safety indicators**: OpenStreetMap data for accident analysis
- **Urban green areas**: Satellite imagery for quantifying urban vegetation
- **Agricultural land use**: Satellite images for soil classification

Mobility & Transport:
- **Mobile Network Operators (MNO) data**: Population movements and tourism statistics
- **AIS (Automatic Identification System)**: Maritime transport statistics
- **Mobile phone data**: Commuting patterns, occasional movements, tourism flows

Challenges of big data in official statistics

- Access & Availability
- Technical Capacity
- Data Quality & Bias
- Methodological Gaps
- Legal & Ethical Issues
- Cultural and organizational barriers to innovation
- Sustainability



CENSIMENTI PERMANENTI
POPOLAZIONE
E ABITAZIONI

Italian Population Census: Register-Based Population Size Estimation

A. Bernardini, N. Cibella, F. Solari
Istat, Population and Housing Census Division

Outline

- Fully Register-Based Estimation Process
- Quality Assessment Framework
  - Extended Population Register and Usually Resident Population Estimation
  - Audit Survey
  - Quality Assessment
- Comparison with the Population Count Estimation in Ireland

Fully Register-Based Estimation Process

- Many countries, including Italy, are shifting from traditional surveys to administrative-data-based censuses, a trend accelerated by the COVID-19 interruption of fieldwork.

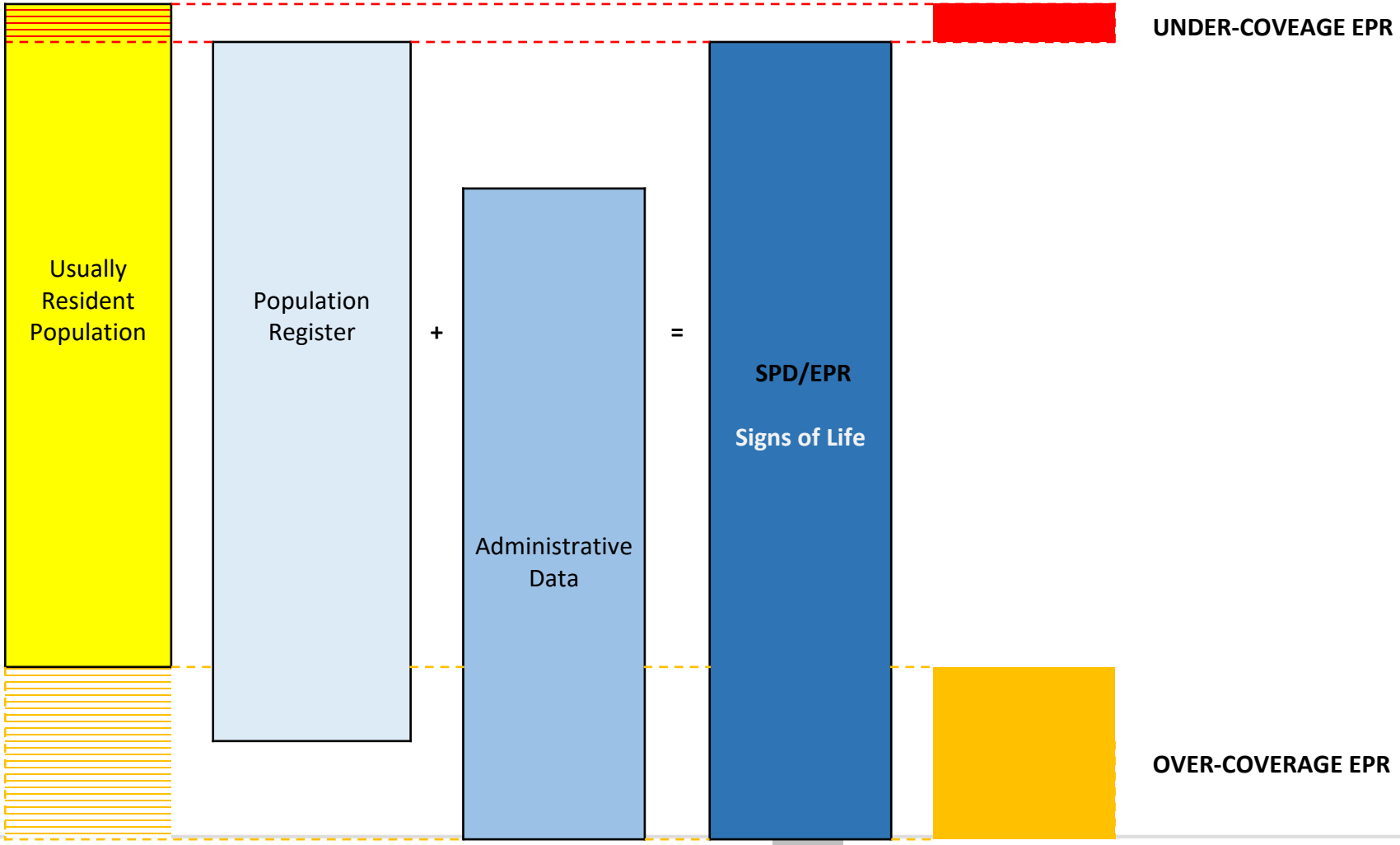
- In Italy from 2020, a Statistical Population Dataset (SPD) is built from the Population Register by integrating it with multiple administrative archives (labour, education, taxes, earnings, pensions, benefits, permits) to detect individual “Signs of Life” (SoL).

- SoL refer to individual activities denoting potential usual residence. They are classified according to type of source and duration patterns.

Fully Register-Based Estimation Process

- The EPR consists of pairs of individual-address.
- In terms of individuals, the EPR is assumed not to suffer from under-coverage or at most to be affected only by negligible under-coverage compared with over-coverage.
- A set of SoL driven deterministic rules is applied in order to: 1) assign each person a prevalent residence; 2) remove from the EPR the individuals considered not to belong to the usually resident population.
- This choice can be seen as a dichotomisation approach of fractional counting proposed by Zhang (2021).

Usually resident Population estimation



Quality Assessment

- In order to provide quality assessment for the usually resident population size estimates, the audit sampling approach is followed, i.e. the survey is carried out to provide quality measures of the register based population size estimate.

- The Audit Survey aims to pursue two goals:

  a) providing a measure of the error of register based population size estimates;

  b) collecting data to be used to improve the definition of the SoL driven deterministic rules.

Audit Survey

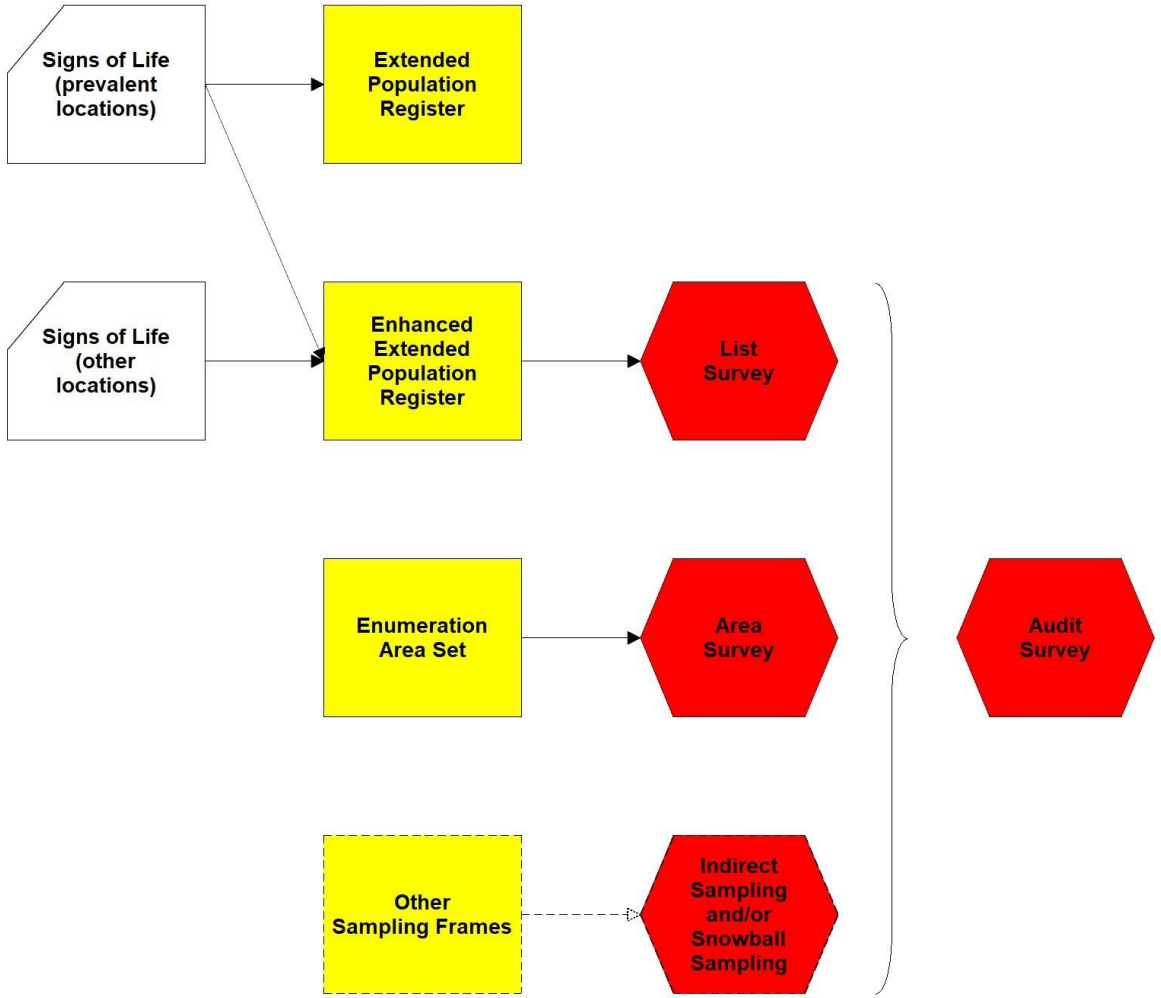
- A multiple frame approach is adopted for the Audit Survey.
- Two components are planned to be included in the Audit Survey at the moment:

  a) a list survey to evaluate the EPR over-coverage;

  b) a small area survey to appraise potential under-coverage of the EPR;

  c) additional components may be included to reduce survey under-enumeration for hard-to-survey sub-populations, e.g., indirect sampling or snowball sampling components.

Audit Survey Components



Audit Survey Sampling Design

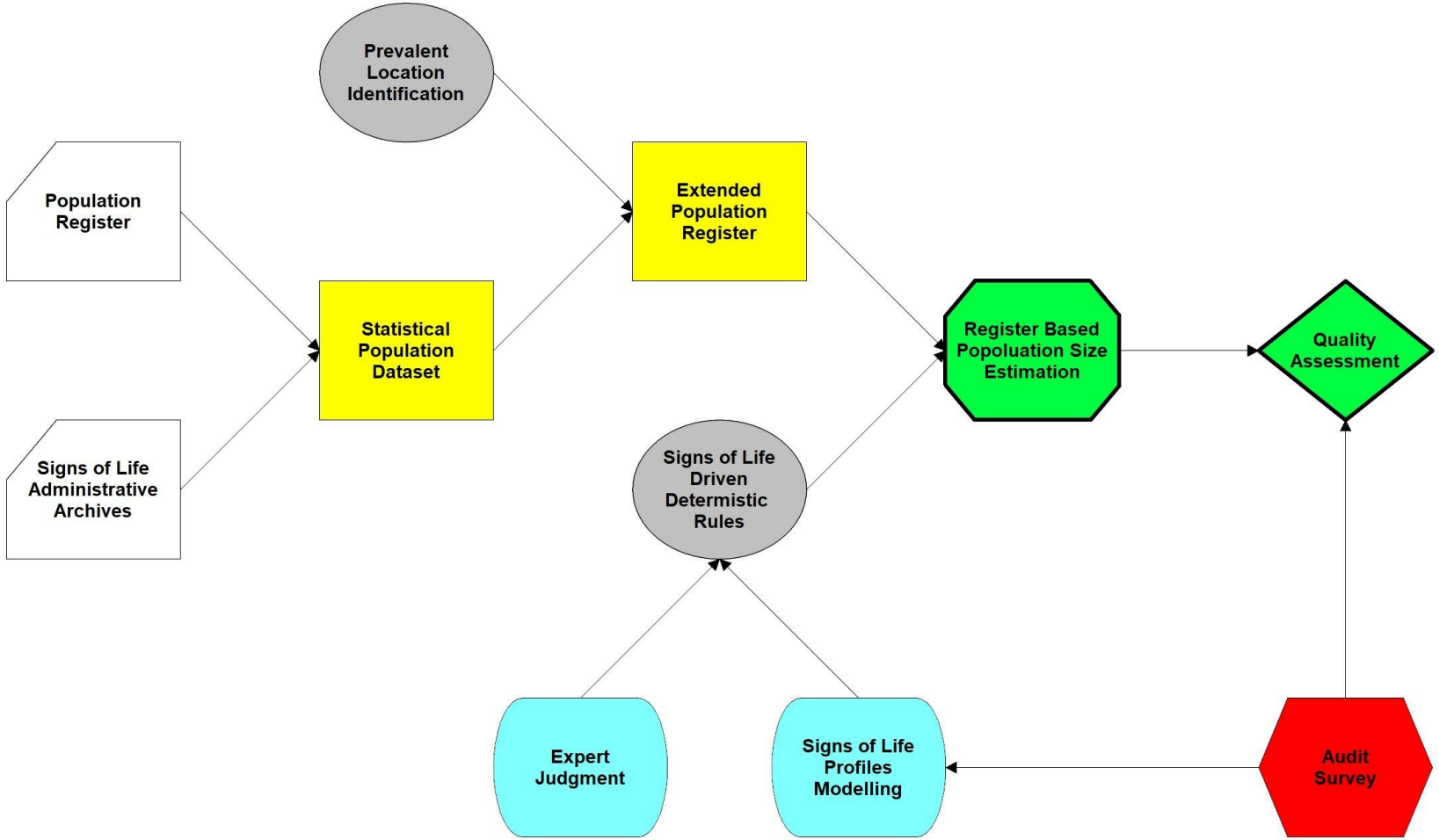
- The Audit Survey sampling allocation allows to fulfill both the aims described in a) and b).
- The sampling allocation is provided in Solari, Bernardini & Cibella (2023).
- The sampling design is described in Solari (2023) and it is based on graph sampling theory.

Sampling Design based on Graph Sampling

- Graph sampling (Zhang, 2023) is a general conceptual framework in which standard sampling theory appears as a special case.
- Graph sampling applies when population units are connected to sampling frames through graph-structured relationships.
- Incidence Weighting Estimator (IWE) produces an estimate $\widehat{N}$ of *N*.
- An estimate of the *MSE* of the fully registered based estimate N_R can be computed as follows:

$$MSE(N_R) = \left(\widehat{N} - N_R\right)^2 - Var(\widehat{N})$$

Population size estimation process



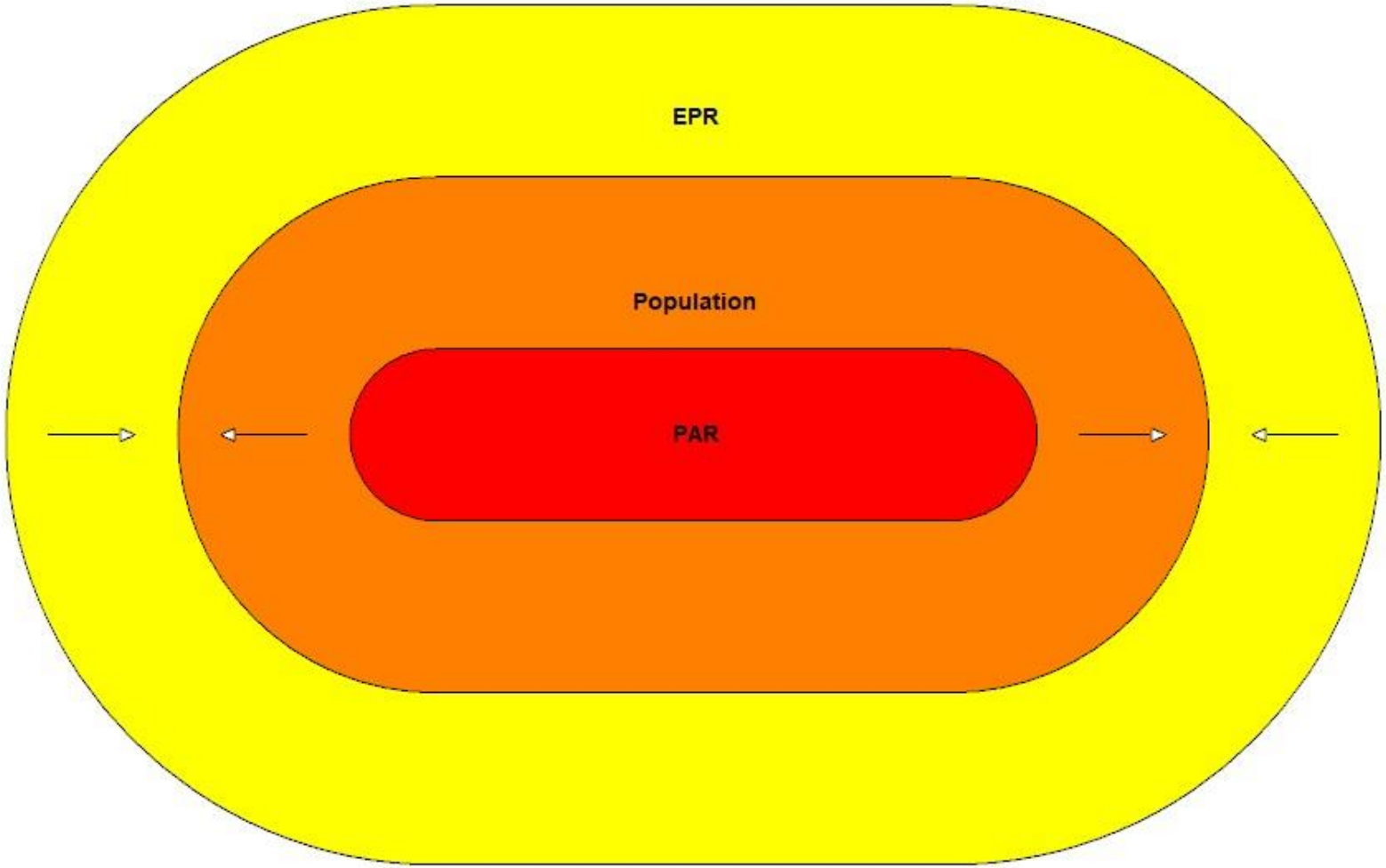
The Irish Approach (Dunne & Zhang, 2024)

- CSO built an SPD from available administrative data sources. The SPD is called the Person Activity Register (PAR).
- The PAR takes a Signs-of-Life approach and tends to include persons where there is evidence that they have engaged with the State and therefore that they live in Ireland for a given reference year.
- The PAR is expected to suffer from undercoverage but not from overcoverage.

Comparison

- The rationale under the EPR is symmetric with respect to the choice adopted for the PAR.
- The target population is investigated from two opposite directions, specifically adjusting for over-coverage in the EPR and for under-coverage in the PAR.
- The choice adopted by Istat reflects the availability of the PR. Although it is not considered to be enough consistent to produce population size estimates directly, the PR is the more relevant source of information on usually resident population in Italy.

The Irish Approach (Dunne & Zhang, 2024)



References

- Bernardini, A., Chieppa, A., Cibella, N., Gallo, G., Solari, F., and Zindato, D. (2022). Evolution of the Italian Permanent Population Census: lessons learned from the first cycle and the design of the Permanent Census beyond 2021, ECE/CES/GE.41/2022/4, UNECE Conference of European Statisticians - Group of Experts on Population and Housing Censuses, Geneva, 21-23 September 2022.
- Dunne J, Zhang L.-C. (2024). A system of population estimates compiled from administrative data only, Journal of the Royal Statistical Society Series A: Statistics in Society, 187, 3–21.

References

- Solari, F., Bernardini, A., and Cibella, N. (2023). Statistical framework for fully register based population counts, METRON, 81, 109-129.
- Solari, F. (2023). A graph sampling approach for audit survey sampling design supporting register based population size estimation, Technical Report, Istat Advisory Committee on Statistical Methods.
- Zhang, L.-C. (2021). On provision of UK neighbourhood population statistics beyond 2021, Report for ONS.
- Zhang, L.-C. (2021). Graph Sampling. Chapman & Hall/CRC: Boca Raton, FL.

Non-Traditional Data Sources for Official Statistics: Insights from AIS Data

Marco Di Zio, Istat

Webinar Royal Statistical Society, 22 January 2026

Istat

AIS data for maritime statistics

Project started officially August 2023

Arosio F.M., Pappagallo A., Salamone N., Valentino L., Amato F., Caspanello M., Massacci G., Ortame F., Pugliese F., Riccobono C. ,Sisti F., Talice S.

Istat

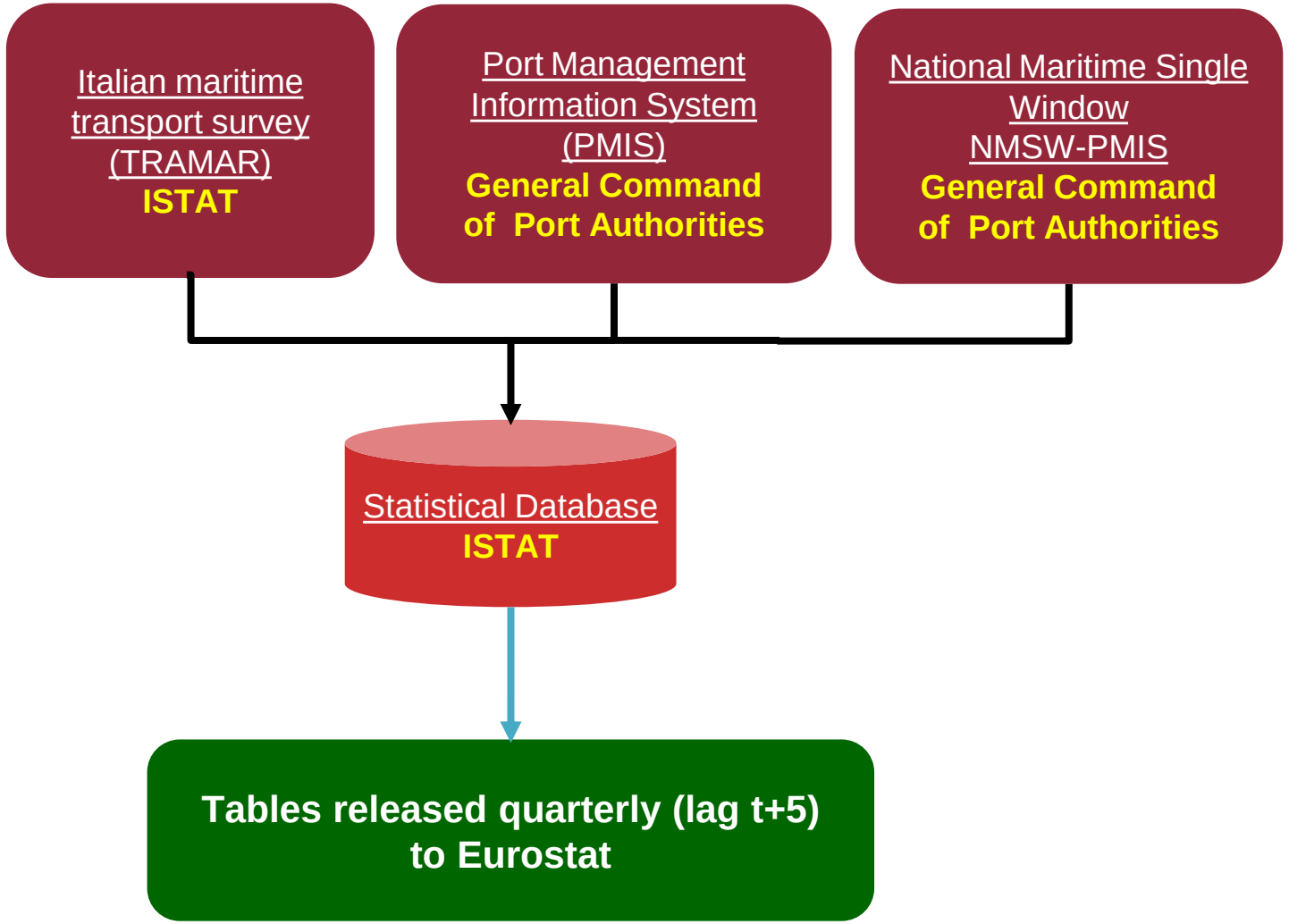
Istat Maritime Transport Survey

The Istat survey on maritime transport aims to provide statistics on the transport of goods and passengers carried out by vessels engaged in commercial activities.

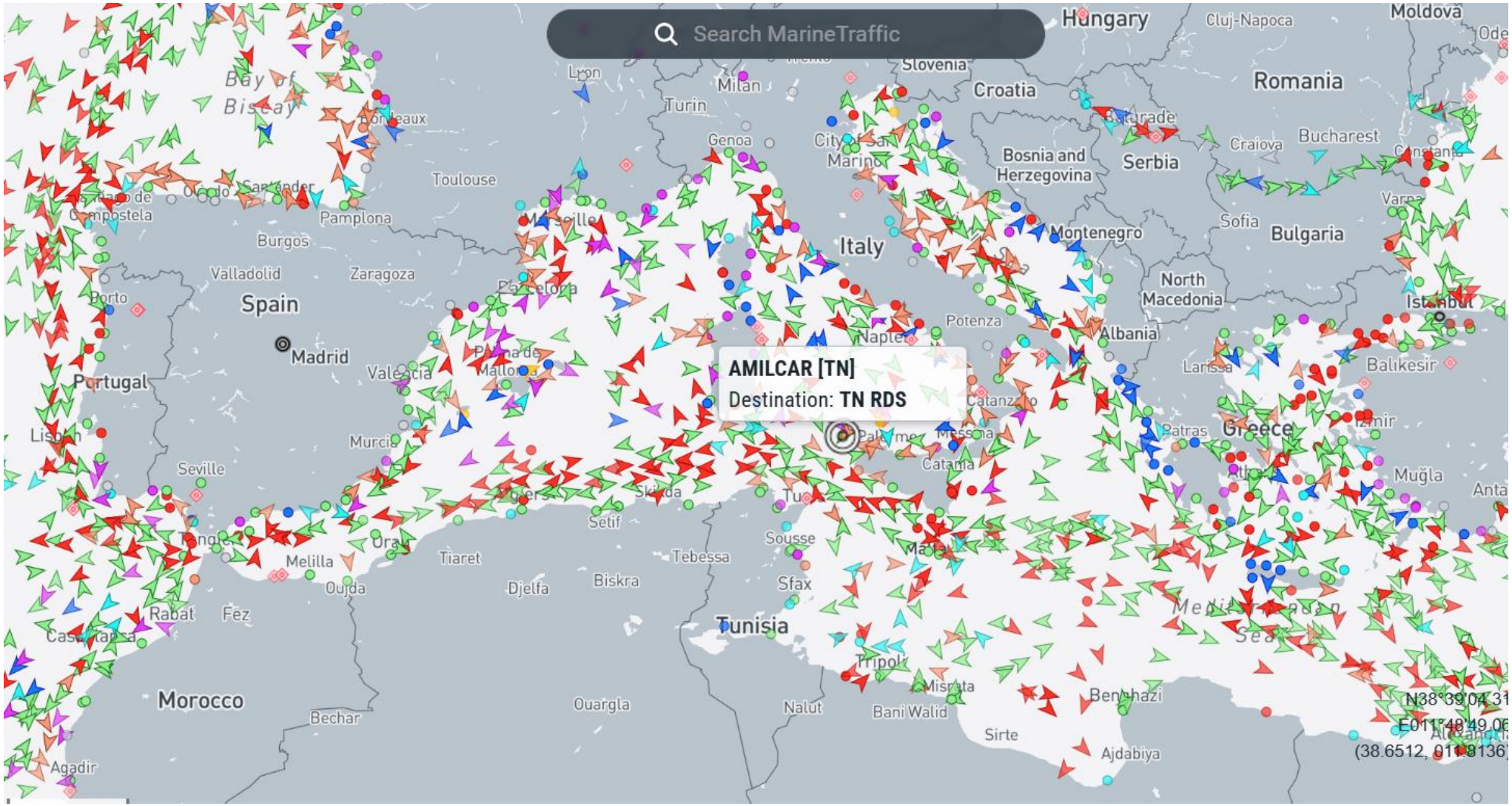
Census of shipping agencies/owners of ships and admin data from Port authorities. It aims to survey all arrivals and departures in Italian ports of vessels with a gross tonnage of at least 100 tonnes.

The survey is a fulfilment of EU obligations.

Eurostat request: improving timeliness and quality



Data on the internet (using AIS)

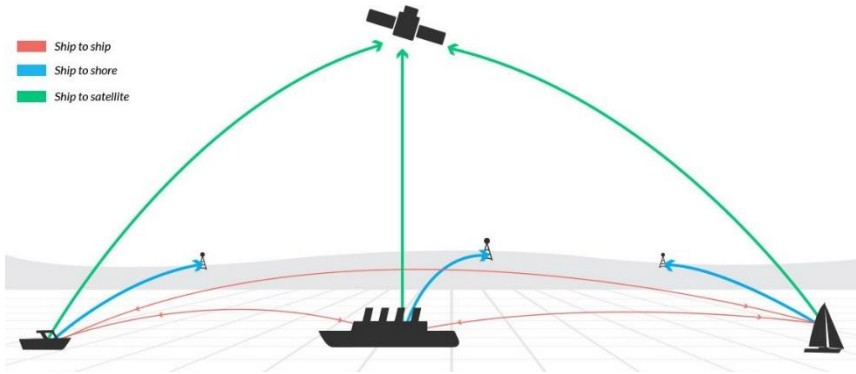


AIS data

DATA transmitted (every 6 minutes):

1.Static Information

MMSI number. IMO number, Name and Call Sign,
Length and Beam, Type of ship, Location of position fixing antenna



2. Dynamic Information

Ship's position, Position timestamp (in UTC), Course Over Ground (COG)

- Mandated for all vessels ≥ 300 GT engaged on international voyages, and all passenger ships
- AIS cannot be switched off

Common feeling: it is easy!

In practice: **it is not easy**

Problems.

- ✓Access & Availability
- ✓Data Quality & Bias
- ✓Methodological Gaps
- ✓Legal & Ethical Issues
- ✓Technical Capacity
- ✓Cultural and organizational barriers to innovation

Sustainability

Issues to deal with

✓Data availability & legal issues. Data are not free.

Fortunately, recently, UN made it available for free (for members of Committee of Experts on Big Data).

✓Data management

is not straightforward for technical capacity: data in cloud, not possible to download, a huge amount of data,..

Now we made experience, and we are able to work with AIS data

next step: transform for (our) statistical purposes

Even with data at hand, it is not easy to produce (our) statistics

They are signals, need transformation to have “routes”, for instance for the count of docked vessels for each port

First elements to compute:

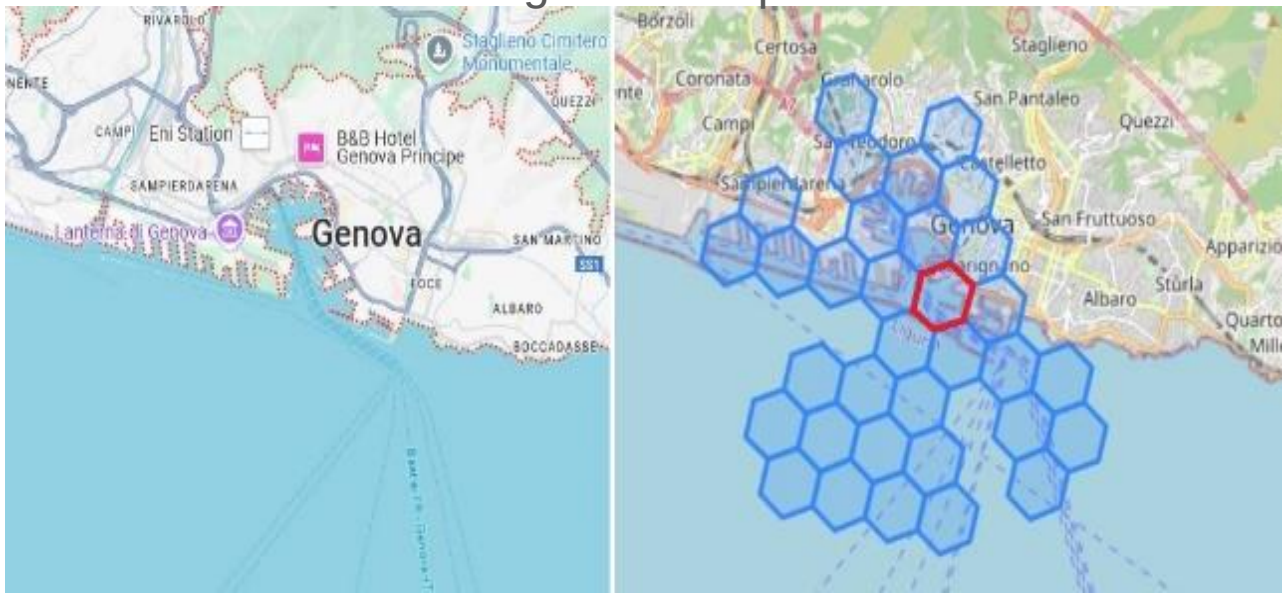
1. *Delimit an area defining a port*
2. *From signals, define when the vessel is docked.*

Defining a port

Definining each port area as a geo-referenced polygon (use H3 system).

Hexagon corresponding to port coordinates augmented with hexagons containing potential positions of the stationary ships (estimated through data)

Fig. Genova port



From signals to route

Static data					Dinamic data					Travel related data	
IMO	MMSI	CALLSIGN	VESSEL NAME	VESSEL TYPE	TIME	COORDINATES	NAVIG. STATUS	SPEED	COURSE	DESTINATION	DRAFT
8401561	20110115	ZAD4L	FROJDI II	Cargo	04/06/2023 19:45	41.1323 16.8530	MOORED	0	258	Ravenna	null

IMO	VESSEL TYPE	DEPARTURE PORT	ARRIVAL PORT	DEPARTURE DATE	ARRIVAL DATE
8401561	Cargo	ITBRI (Bari)	ITRAN (Ravenna)	04/09/2023	05/09/2023
9483712	Passenger	ITGOA (Genova)	ILOLB (Olbia)	05/09/2023	06/09/2023

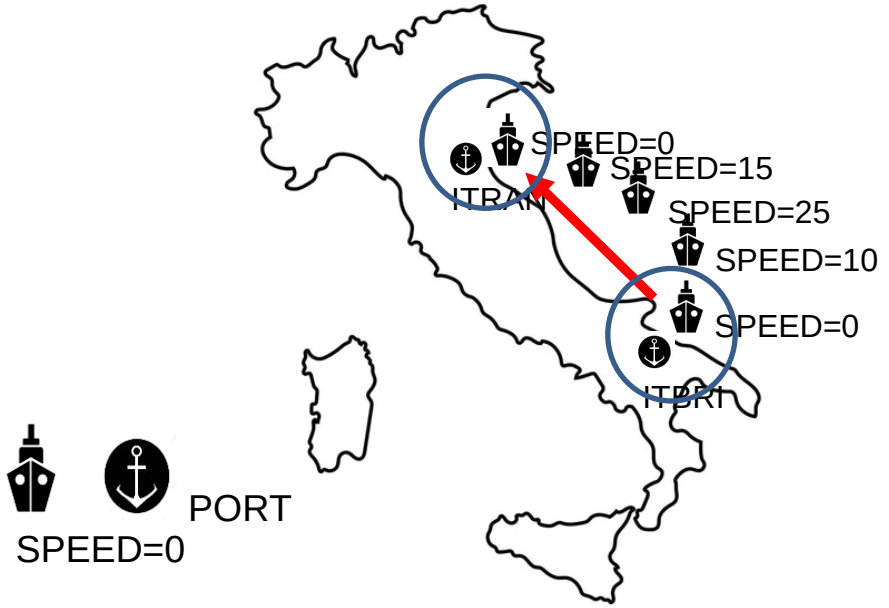
Reconstruct all maritime voyages involving ships departing or arriving at an Italian port.

A voyage is defined by two key events: departure and arrival.

In AIS these events are reflected in records that meet a specific criterion:

- Speed = 0 in an hexagon of a port

Routes built through the sequence of signals, ports and speed



Quality of transformed data

Data for (our) statistical use are now available.

- *Are there any gaps?*
- *What is the quality?*

Quality of transformed data

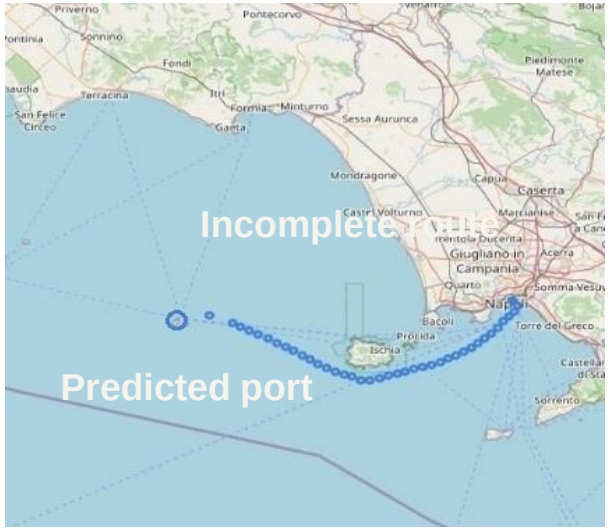
1. Macroediting. Comparison of aggregated statistical figures
2. Record linkage for micro comparison

(some) problems:

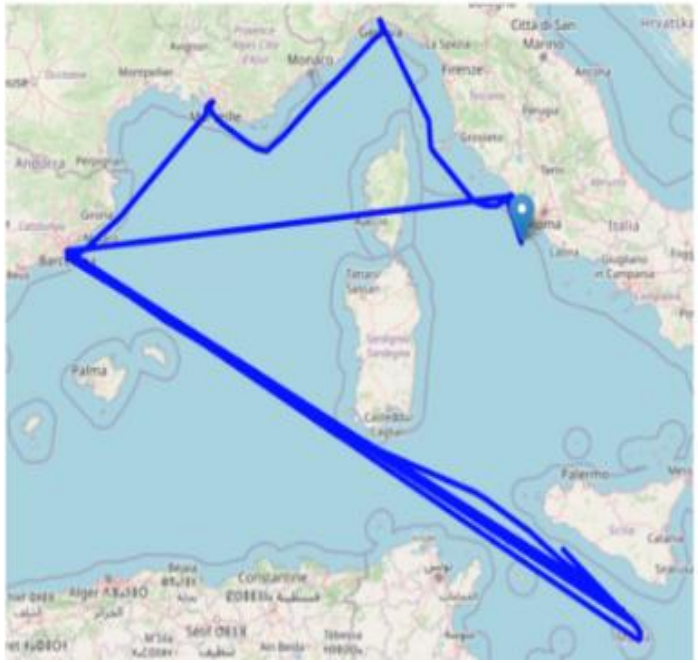
- Short and frequent routes are under-counted
- Interruption of signals
- Anomalous routes

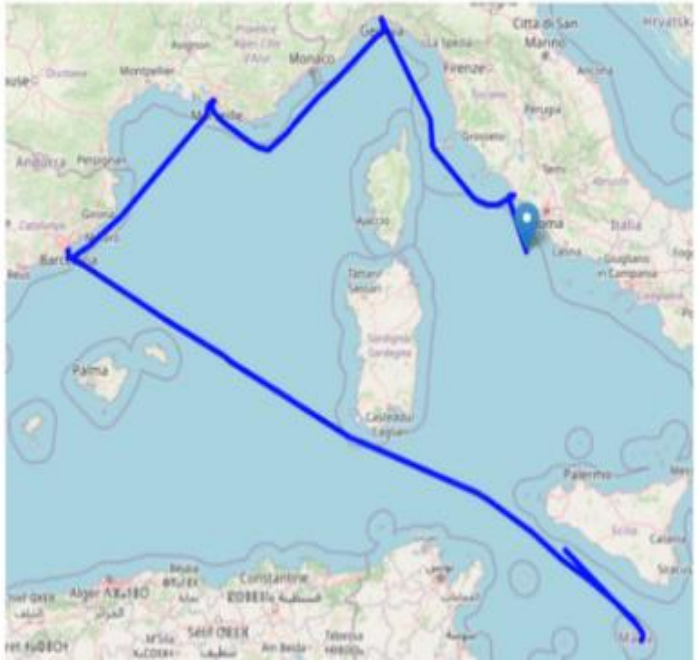
Imputation of incomplete routes

- Developed deterministic and deep learning imputation methods



Methods for Anomalous routes - outliers

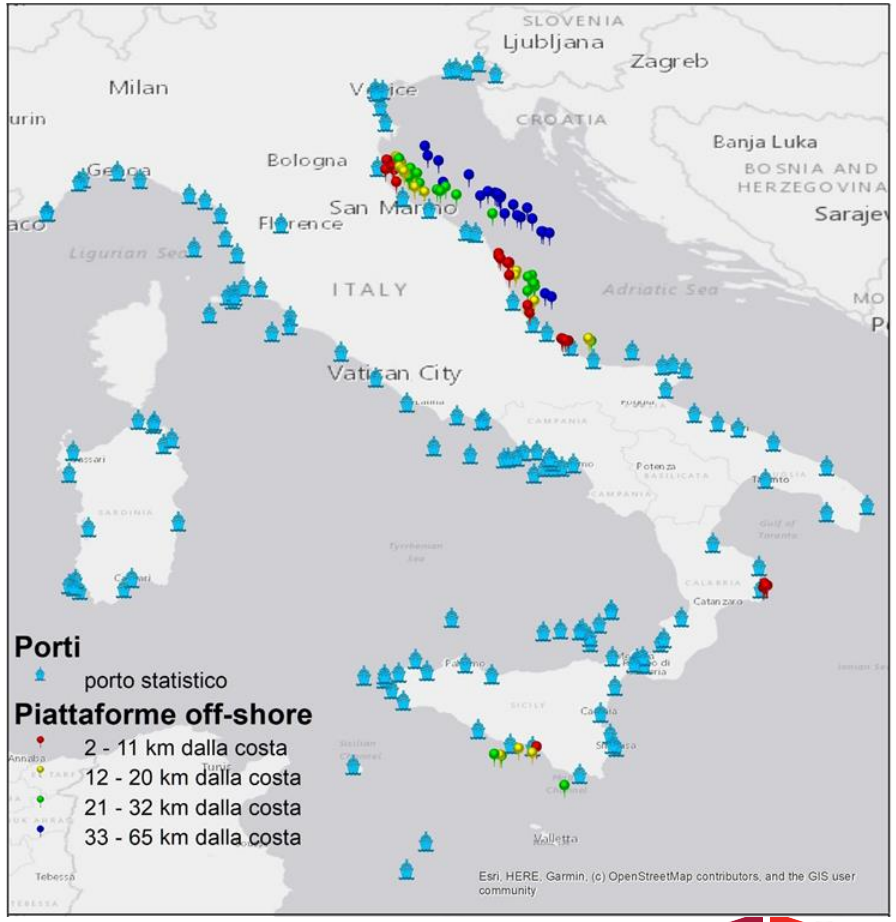




Now we can use AIS data, next question is how

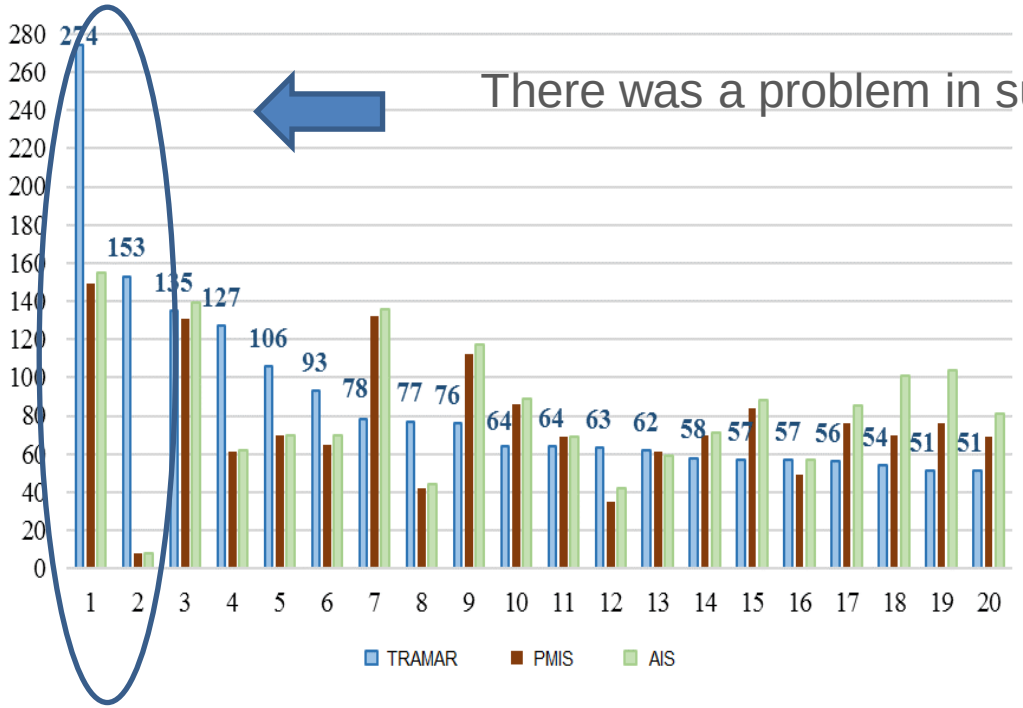
Improve some statistical figures, for instance the offshore platforms

The location of the offshore platforms has allowed the identification in AIS, for 2022, of almost 5,000 arrivals, the official procedure slightly more than 230.



AIS for improving quality

Arrivals in TRAMAR survey (blue), in admin data PMIS (red) and in AIS (green) for each cruise ship (id ship replaced by numbers 1 to 20).



Improving statistics on ‘Cruise’ vessels

Comparison of raw data

	N. IMO (id vessels)	N. Voyages	N. ports
AIS	162	5122	66
TRAMAR	158	4428	52

Analysis of discrepancies

Improving statistics on ‘Cruise’ vessels

The analysis of discrepancies leads to some corrections in the process.

	N. IMO (id vessels)	N. Voyages	N. ports
AIS	162	5122	66
TRAMAR	~~158~~ 167	~~4428~~ 5250	~~52~~ 60

Checks are still in progress

Final remarks

Now, we have AIS data available for the the statistical process.

Currently, AIS are used as auxiliary information:

To check , validate and improve official procedure

Further analysis are needed for introducing AIS data into the statistical production process

Problems: Methodological Gaps

- How to integrate the sources to obtain “the count of…”
- How to use AIS for more timely statistics?
- Statistics on goods and passengers, information not in AIS
- How to compute the degree of uncertainty of these counts?

Finally, listen to big data…

Think different…,

- AIS as functional data
- Network analysis for the routes
- …

Questions and Answers

MARCO DI ZIO dizio@istat.it
FABRIZIO SOLARI solari@istat.it
ELENA GRIMACCIA elena.grimaccia@gmail.com; elgrimac@istat.it

22 January 2026

Istat