Big data–based national statistical production

Alessandra Righi and Monica Scannapieco

Workshop on “Big Data & Machine Learning
Applications for Central Banks”
Bank of Italy, Rome, 21-22 October 2019

Istat

Overview

1. Istat ongoing BD activities
2. ML techniques within the ongoing activities
   - ✓ Piloting
   - ✓ Experimental statistics
   - ✓ Official statistics
3. Quality issues

Istat ongoing BD activities

- Since 2013 Istat is moving along the path of the production of statistical information almost in real-time using big data sources
- New indicators / information are added to the traditional ones derived from surveys, continuing to ensure the quality of Official statistics
- The ongoing process is aimed at widening and deepening the output, innovating methods and providing timely data

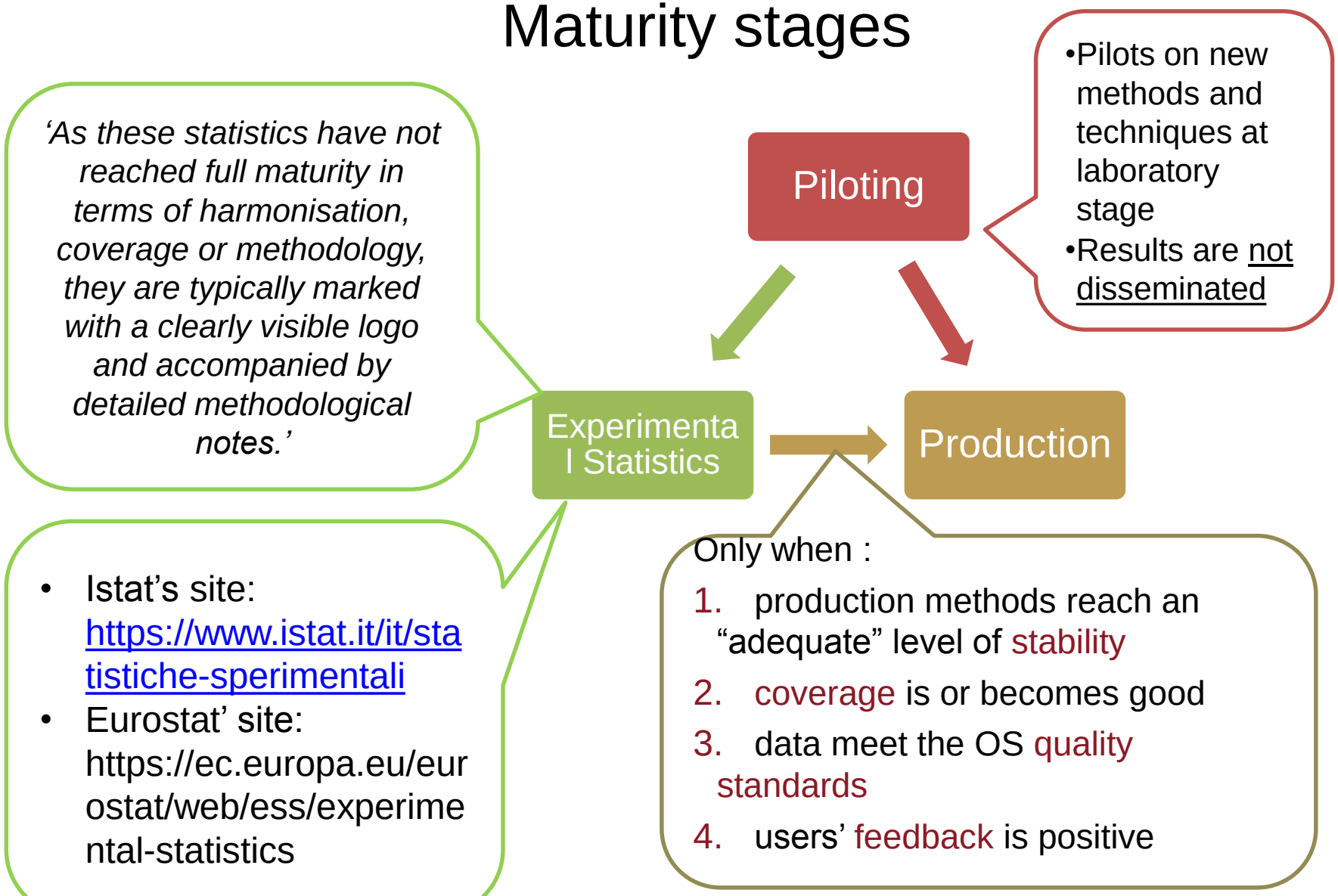
Istat ongoing BD activities

- As the production of new OS takes time:
  - ✓ to develop new methodologies
  - ✓ to translate them into technological and organisational solutions
  - ✓ to comply with quality requirements and harmonisation rules

- Istat experiments with new sources and new methodologies in the data production and offers the results to the public for evaluation as Experimental Statistics aimed at :
  1. filling knowledge gaps
  2. shortening the time for data production providing timely evidence to policymaking
  3. fostering new analyses

Istat ongoing BD activities

Maturity stages



Istat ongoing BD activities

Source	Outcome	Maturity
Scanner data	CPI/HICP	Official statistics
Internet data (web scraping)	ICT in enterprises	Experimental statistics
	Business register	Piloting
Social Media (Twitter)	Social mood on Economy index	Experimental statistics
Satellite Images	Land-use / Agriculture	Piloting
Mobile phone	Mobility / Turism	Piloting

ML techniques in the ongoing activities:

Piloting

Land Cover and Land Use

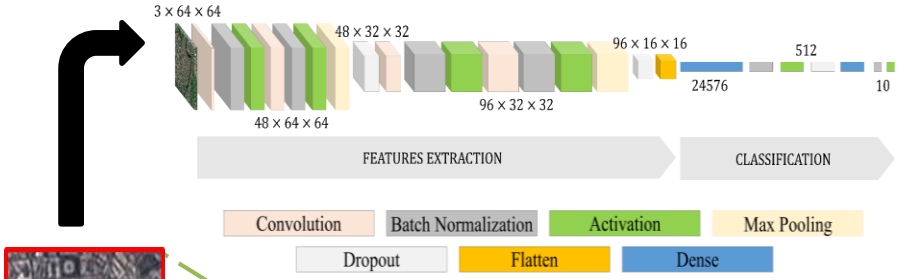
- Eurostat has been carrying out the LUCAS survey every 3 years since 2006 in order to estimate the Land Cover (LC) and Land Use (LU) within the EU up to NUTS-2 :
  - A 2-phase area sample survey of the whole EU territory
    - 1st phase: Master Sample of ~1.1 million points in a square grid of (2 km x 2 km) cells
    - 2nd phase: ~330,000 random points from the Master Sample
  - Direct data collection, mainly on the ground (~70% of 2nd phase points), the rest by clerical photo-interpretation

- Deep Learning **+ Satellite Imagery** data (e.g. Sentinel-2) are used for LC estimation
  - **Classify-and-Count approach**
    - Train an image classification algorithm to predict the LC class of a satellite image tile from Eurosat
    - Divide the satellite images covering a target area into tiles and use the trained algorithm to predict LC classes
    - Obtain LC statistics for the target area by simply computing the relative frequencies of predicted LC classes

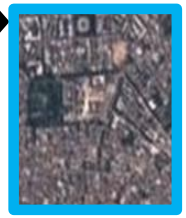
- **Pros**: (i) dramatic reduction of data collection costs/burden, (ii) more timely statistics, (iii) supplying LC statistics beyond the NUTS-2 level, (iv) producing (moderate resolution) maps of the whole territory

- **Aim**: **Can a fully automated approach provide LC estimates of satisfactory accuracy? We are valuating external accuracy without an available benchmark**

Trained Deep Neural Network

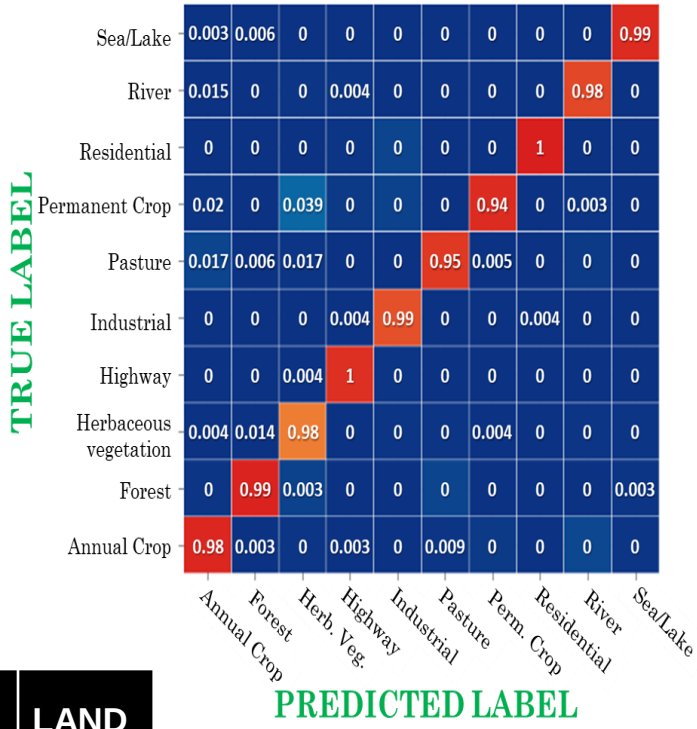


RESIDENTIAL



Satellite image- Sentinel 2 (Puglia crop)

CLASS	LAND COVER
…	…
Residential	14.8 %
…	…



Internal accuracy in Eurosat

ML techniques in the ongoing activities:

Experimental statistics

ICT Usage by Enterprises from Web Scraped Data

- The annual Survey on ICT Usage in Enterprises ('ICT survey') collects data on the usage of Information and Communication Technologies, the Internet, e-business and e-commerce in enterprises
  - ✓ Target population: enterprises with 10 or more employees (184,000 enterprises in 2017)
  - ✓ Sample size: planned ~32,000 – respondents ~21,000
- **Fact**: ~70% of the target population owns a web site (~130,000 enterprises)
- **Idea**: for enterprises having a website some ICT data could be gathered directly from the web → The IaD (Internet As a Data Source) paradigm
- **Project Outline**:
  1) **Search URLs** (website addresses) for the largest possible part of the population
  2) **Scrape** the textual content of identified **websites**, $\mathbf{X}$
  3) **Train a Machine Learner** (ML) to predict a set of ICT variables, $\mathbf{Y}$, from the scraped text, using as training set survey and scraped data jointly
  4) **Use** the trained **ML to predict values**, $\mathbf{Y}^{\mathbf{PRED}}$, for all the enterprises that were not observed by the survey but whose websites have been scraped
  5) **Use** survey data, $\mathbf{Y}^{\mathbf{OBS}}$, and **Big Data**, $\mathbf{Y}^{\mathbf{PRED}}$, **to compute Experimental Statistics**

ICT Usage by Enterprises from Web Scraped Data

Main phases of the process (1/2)

- **Target Variables**
  - ✓ Three ICT dichotomous variables were selected as targets: (1) Web Ordering (akin to e-commerce), (2) Job Advertisements, (3) Links to Social Media
- **URL Retrieval**
  - ✓ About 100,000 URLs were identified. Only a small portion from the ICT survey and admin data. The rest from a complex procedure: (i) send batch queries to a search engine, (ii) get returned URLs and scrape them, (iii) score the URLs according to the retrieved information, (iv) select the most likely URL according to a fitted model
- **Web Scraping**
  - ✓ The texts of about 85,000 enterprise websites were collected
- **Feature Extraction**
  - ✓ Using jointly survey data and scraped texts, Natural Language Processing methods were employed to identify word sequences (n-grams) with the highest predictive power for the different target variables
- **Machine Learning**
  - ✓ ML models were defined to predict the target variables given the extracted features (n-grams). The subset of about 21,000 enterprises for which both survey data and scraped texts were available was used to train, test and validate different ML algorithms

ICT Usage by Enterprises from Web Scraped Data

Main phases of the process (2/2)

- **Comparative Evaluation of ML**
  - ✓ For variable Web Ordering the Random Forest emerged as the best ML candidate
  - ✓ Note that as the dataset is imbalanced in Web Ordering (20% YES / 80% NO) the F1-score is the most relevant performance measure

ML	Accuracy	Recall	Precision	F1-score
Logistic	0.88	0.64	0.66	0.65
SVM	0.90	0.62	0.76	0.68
Random Forest	**0.90**	**0.72**	**0.74**	**0.73**

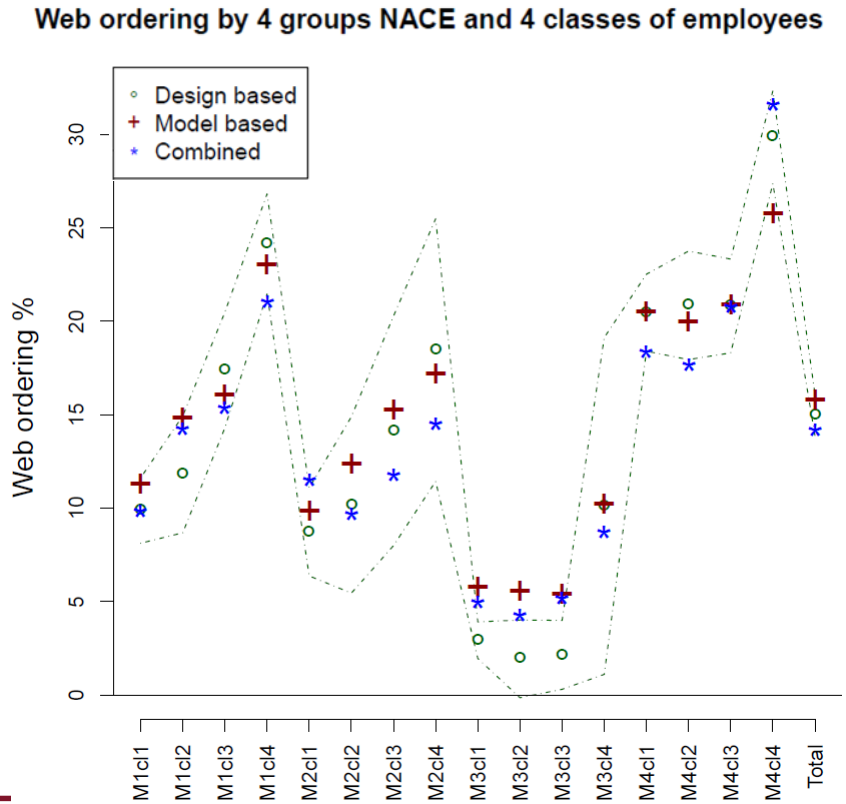
- **Computation of Experimental Estimates**
  - ✓ Predicted values of target variables (obtained by ML) only available for the reached subpopulation $\mathbf{U^2}$ (i.e. enterprises owning a website that was successfully scraped)
  - ✓ Target population $\mathbf{U^1}$ is larger (i.e. enterprises owning a website), but unbiased estimates of population totals are available for it from the ICT survey: $\hat{Z}^{U^1} = \sum_{k \in (s \cap U^1)} w_k \, z_k$
  - ✓ Perform a pseudo-calibration to make $\mathbf{U^2}$ representative of $\mathbf{U^1}$: $\sum_{j \in U^2} \tilde{w}_j \, z_j = \hat{Z}^{U^1}$
  - ✓ Compute experimental estimates using pseudo-calibration weights: $\hat{Y}_{exp} = \sum_{j \in U^2} \tilde{w}_j \, y_j$

ICT Usage by Enterprises from Web Scraped Data

Results and possible impact

- Overall, estimates produced using web scraped data have been found plausible
- No large discrepancies observed w.r.t. survey estimates at population level
  - ✓ *Big Data estimates often fall within survey based 95% confidence limits*
  - ✓ *Unsurprisingly, domain estimates show larger deviations than in the survey*
- **Possible Impact**
  - ➡ The Internet As a Data Source paradigm will allow Istat to algorithmically enrich the Business Register (ASIA)
  - ➡ Eurostat could delegate the observation of some ICT variables to web scraping and shrink the yearly ICT questionnaire accordingly (direct observation would be needed only to update training sets, say once every three years)

Enterprise Characteristics	Survey Estimate	Confidence Interval		Big Data Estimate
		Lower Bound	Upper Bound	
WEB ORDERING	14.97	13.81	16.13	15.51
JOB ADVERTISEMENTS	10.78	10.02	11.53	13.91
SOCIAL MEDIA	31.25	29.90	32.60	36.68



Istat

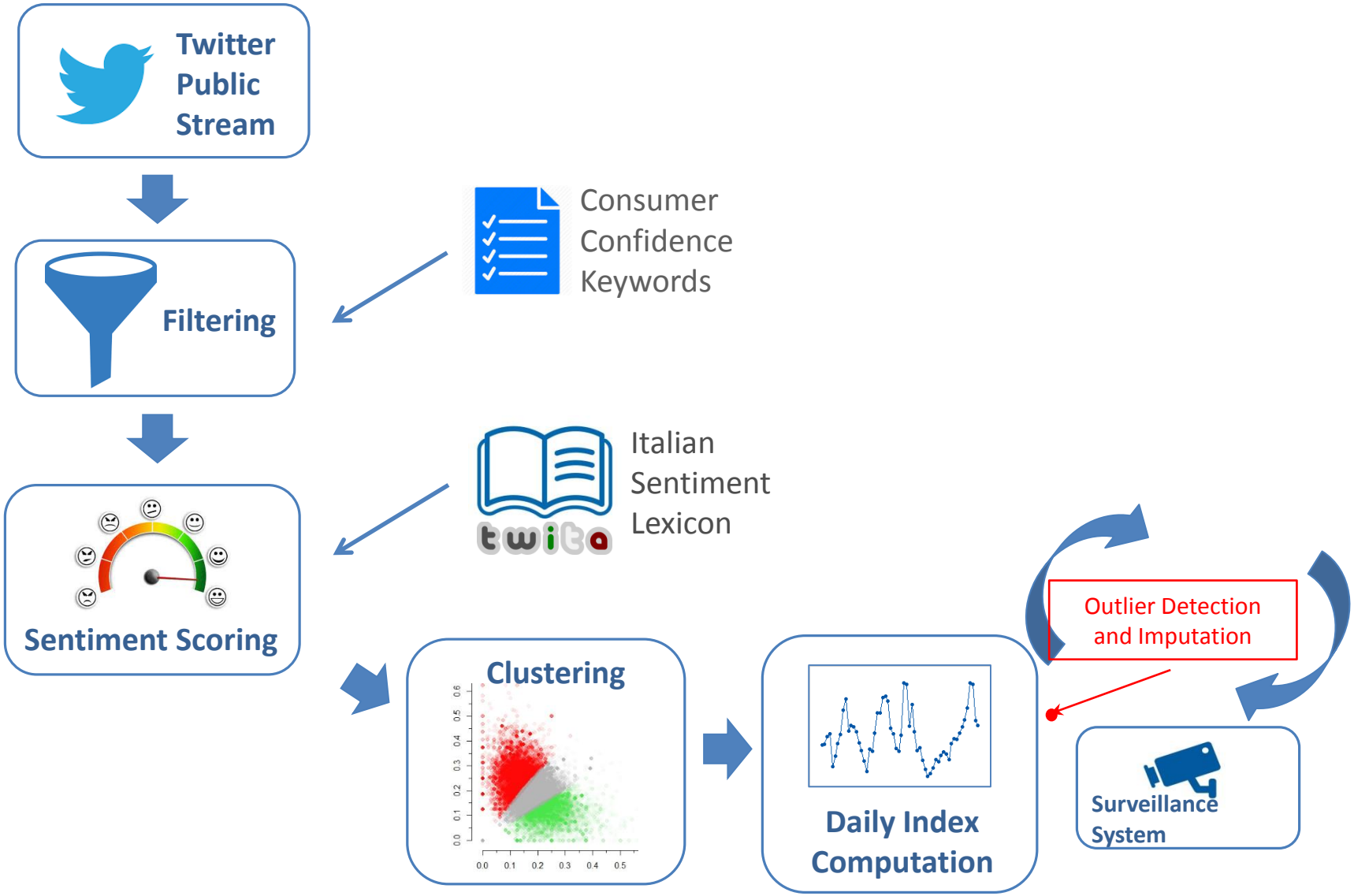
Social Mood on Economy Index

***Domain-specific* sentiment index to assess the mood about the economic situation of the Italian-speaking Twitter users**

- **Procedure**
  - ✓ Download of about 40,000 daily filtered messages by streaming API
  - ✓ Cleaning and normalization phase
  - ✓ A **Sentiment analysis procedure** calculates positive and negative sentiment scores for each tweet using an unsupervised, lexicon-based approach and matching message words against entries of an Italian Sentix Lexicon
  - ✓ Atomic scores of matched words are averaged to yield tweet-level scores
  - ✓ Subsequently, tweets are clustered according to their sentiment scores into three mutually exclusive classes: Positive, Negative and Neutral tweets using the k-means method

- The daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes

Sentiment Index da Twitter: Pipeline



ML techniques in the ongoing activities:

Official statistics

Scanner data to compile Italian CPI

«Transaction data obtained from retail chains containing data on turnover and quantities per item code from which unit value prices (the average of prices paid by purchasers) *can be derived at item code level»*

- In 2018 Istat has officially introduced scanner data of grocery products (excluding fresh food) in the CPI estimation to replace price collection for 79 indices of products' aggregate belonging to five ECOICOP Divisions (01, 02, 05, 09 and 12)
- In 2019 scanner data for 2,146 outlets, including 534 hypermarkets and 1,612 supermarkets of the main 16 RTCs covering the entire national territory are monthly collected by Istat on a weekly basis at item code level
- This is done in agreement with Retail trade chains (RTCs), with the collaboration of the Association of modern distribution and Nielsen

Scanner data to compile Italian CPI

- A static approach (similar to the traditional data collection method) is adopted for the sampling of items:
  - ✓ A cut off sample of barcodes (GTINs) within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover)
  - ✓ The GTINs selected in December are kept fixed during the following year
- Monthly prices are derived by arithmetic averages of weekly prices weighted by quantities
- No ML techniques are actually used in the production process

Future developments

Extension of the price collection to discount, points with small surface, pharmacies, clothing stores, home electronics

Dynamic approach for sampling to increase the number of scanner data for the calculation of inflation

Quality issues

- **Different prevalence** of the use of ML techniques used in the ongoing activities according to their maturity level
- Actual use gradually decreases moving from piloting activities to the OS production, because ML techniques require **a NEW approach to quality issues,** not so familiar to official statisticians
- The quality of the TS is important for accuracy of ML methods:
  - ✓ When the training set is a non-probability sample of the target population, the **accuracy may be worse** when applied to previously unseen data and the predictions could not be trusted alone
- ML techniques are especially useful for data types that are not easily processable with traditional methods, for instance, textual data and images

Quality issues

- The predictive ML-based approach still needs to be fully thought out and how it can be combined with the traditional approaches (designed-based and model-based) is currently under investigation

Consequently,

- Quality of the TS should be governed
- posterior analysis are needed to diminish any risk and to assess the use of a specific ML technique even with humans in the loop

However:

- ML is a viable choice in several cases (e.g. text sources, images, big size…) and in some of them can be the only viable choice

THANKS

righi@istat.it
scannapi@istat.it

https://www.istat.it/it/statistiche-sperimentali/sperimentazioni-su-big-data

For more details :

Enterprise characteristics (scannapi@istat.it)
Social Mood (zardetto@istat.it)
Land Cover (frpuglie@istat.it)
Scanner Data (polidoro@istat.it)