

Artificial Intelligence as a driver for Innovation in Official Statistics

Francesco Pugliese, Central Directorate for Methodology
and Design of Statistical Processes, **Istat**

ARTIFICIAL INTELLIGENCE AND OFFICIAL STATISTICS: THE NEW FRONTIER OF INNOVATION

Caserta | 12 November 2025 | University of Campania “Luigi Vanvitelli” | Department of Mathematics
and Physics | Abramo Lincoln Street, 5

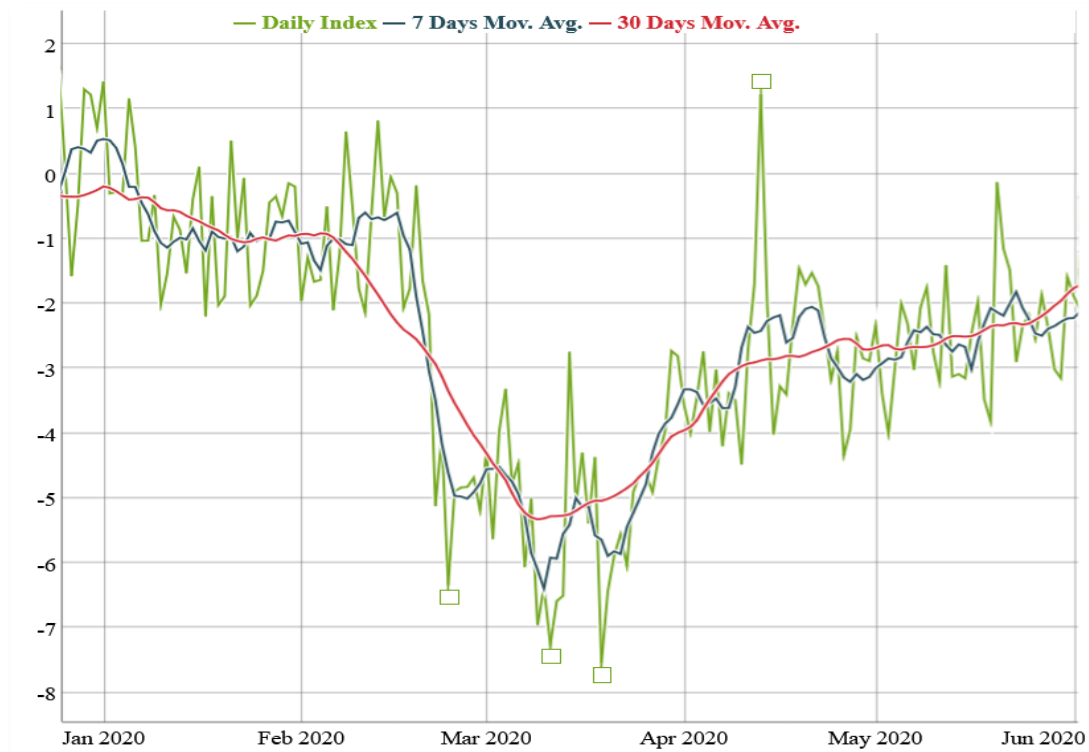
QUINDICESIMA GIORNATA ITALIANA DELLA STATISTICA



Use Case 1: AI for Sentiment Analysis

- The Social Mood on Economy Index (SMEI) is developed by ISTAT
- Measures the Italian sentiment on economy
- Extracts the sentiment from Twitter data
- The sentiment analysis is conducted using a lexicon-based, unsupervised, approach

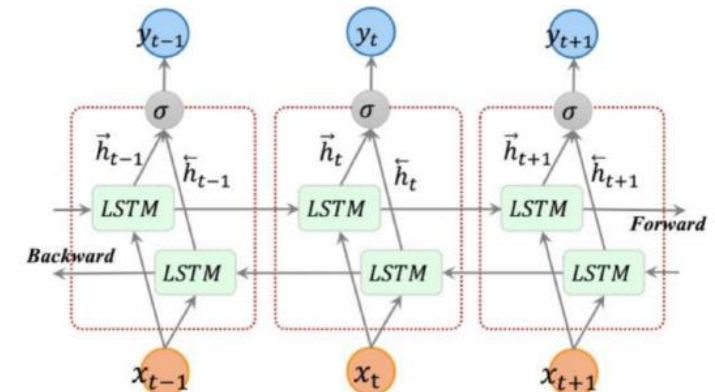
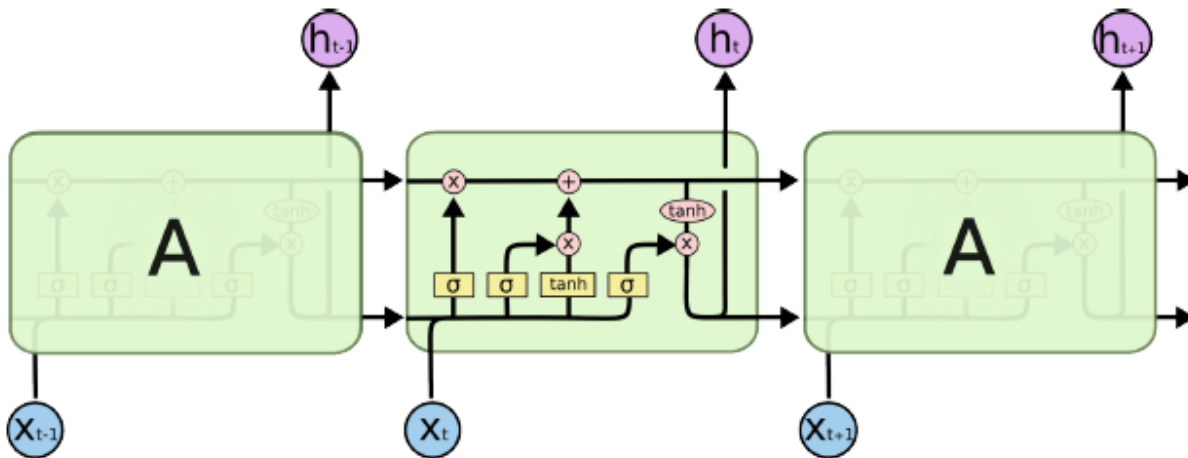
ISTAT's SMEI January-May 2020: the Covid outbreak in Italy



Use Case 1: AI for Sentiment Analysis

Bidirectional LSTM

- Our classification model is a Bidirectional Long short-term memory (BiLSTM). LSTMs are a type of RNN able to process long sequences of data whereas BiLSTM are able to understand the left context and right context of sequences.
- A LSTM memory cell is composed of 4 units exactly: an input gate, an output gate, a forget gate and a self-recurrent neuron.



Use Case 1: AI for Sentiment Analysis

FastText

- Our **Word Embeddings Model** (*Catanese, E., Bruno, M., Stefanelli, L., & Pugliese, F., 2023, September - Winner of the Best Paper Award - 1st Place - CARMA 2023*) is **FastText** which is an evolution of **Word2Vec**.
- **FastText** enables to quickly train models on large corpora
- **FastText** make use of hierarchical soft-max based which is able to noticeably drop computational complexity and can be trained on more than one billion words in less than ten minutes using a standard multicore **CPU**.



Use Case 1: AI for Sentiment Analysis

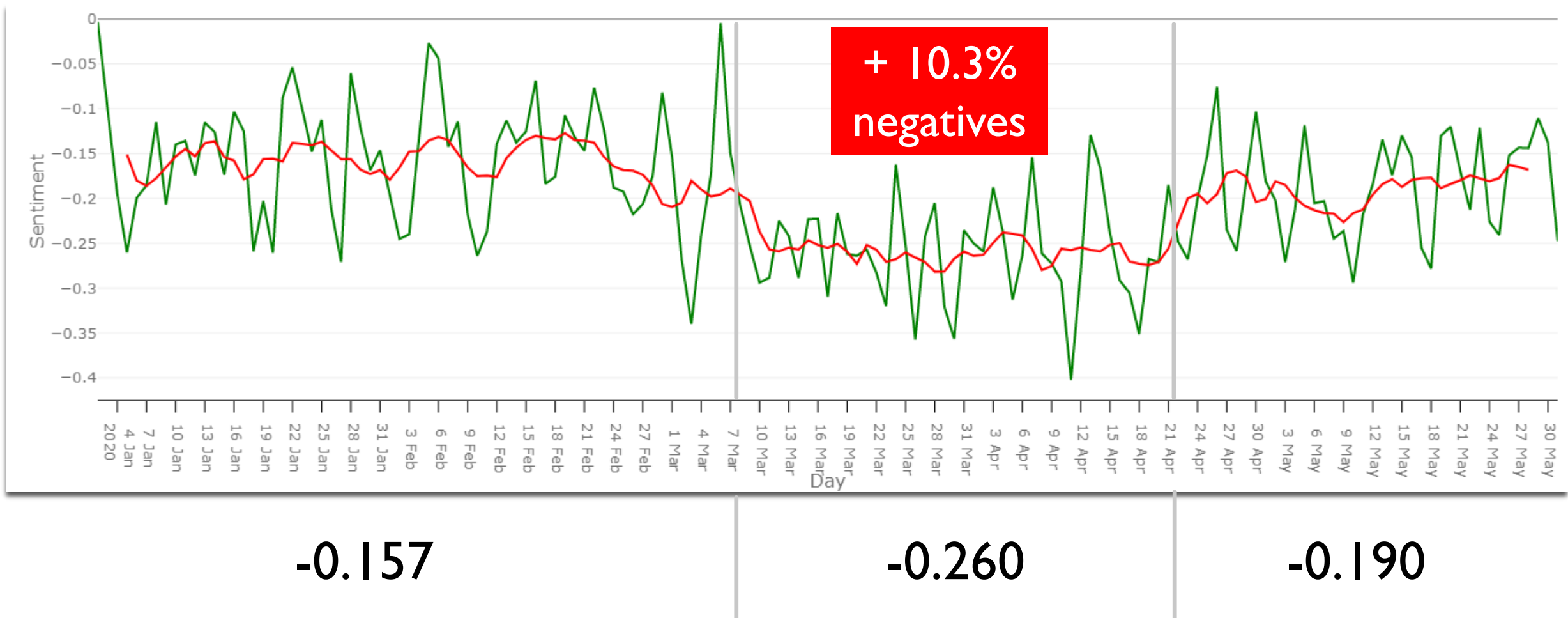
Why a new Index?

- ISTAT's Index (SMEI) needed a revision to deal with new events (Covid pandemic)
 - The proposed model may have a higher adaptability to new information
 - The proposed model is a neural network that has been trained on a set of labelled Tweets that do not contain references to the pandemic
- Computed daily using the sentiment predicted by the model on a set of tweets
 - The sentiment is predicted using a neural network
 - The set of tweets is the same used by ISTAT for the SMEI
 - Computed for the first half of 2020

$$DL - SMEI = \frac{N_P - N_N}{N_P + N_N}$$

Use Case 1: AI for Sentiment Analysis

The proposed Index (DL-SMEI)



Use Case 1: AI for Sentiment Analysis

1st January – 7th March



8th March – 21st April



22nd April – 31st May



Use Case 1: AI for Sentiment Analysis

Focus on second period (8th March – 21st April)

○ Positive words

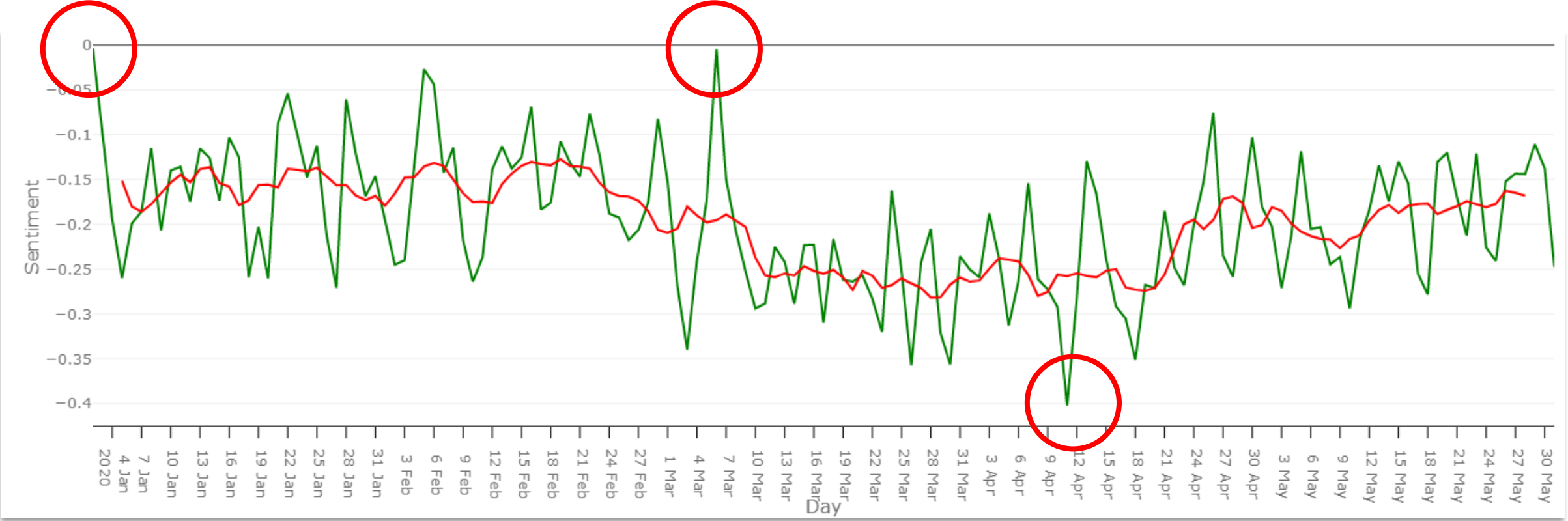


Negative words



Use Case 1: AI for Sentiment Analysis

Maximum and Minimum points



Use Case 1: AI for Sentiment Analysis

Second Maximum on the 6th March

○ Positive words



Negative words



Use Case 2: AI for Land Cover

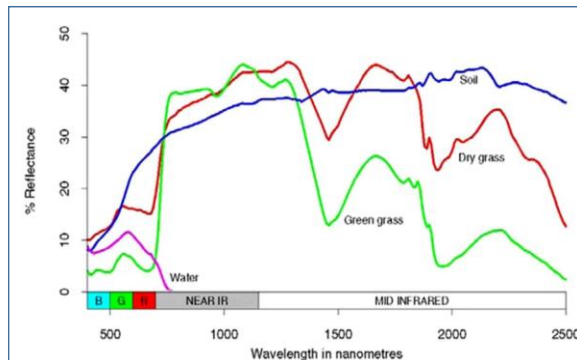
GOALS

Land Cover (LC) statistics (*Bernasconi, E., De Fausti, F., Pugliese, F., Scannapieco, M., & Zardetto, D., 2022*) and maps are a very important statistical product. As they require a big effort to be created, the idea is to build an automatic system that processes satellite

- Automatic Land Cover Estimates
- Automatic Land Cover Maps

ML Approaches to LC from Images

Standard approach: Spectral Signature



- Different LC classes have different reflectance spectra
 - ✓ Variation of reflectance with EM frequency can be used to predict LC class
 - ✓ Trained ML algo predicts the LC class of image pixels independently
 - ✓ Decision on each pixel does not depend on neighboring pixels

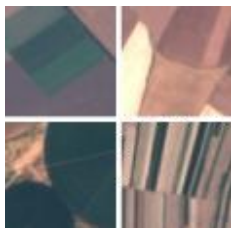
New approach: Computer Vision (Deep Learning)



- Different LC classes have different visual/spatial patterns
 - ✓ Variation of visual/spatial patterns can be used to predict LC class
 - ✓ Trained ML algo (CNN/U-net) predicts LC class of image pixels based on information from neighboring pixels
 - ✓ Decision on each pixel depends on the whole sub-image (tile) the pixel belongs to

Use Case 2: AI for Land Cover (

ANNUAL CROP



RIVER



FOREST



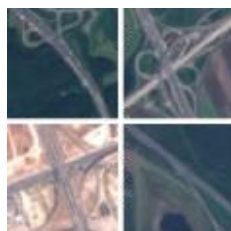
RESIDENTIAL



INDUSTRIAL



HIGHWAY



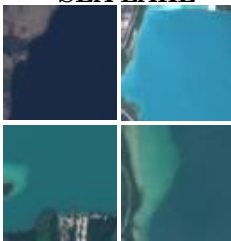
PASTURE



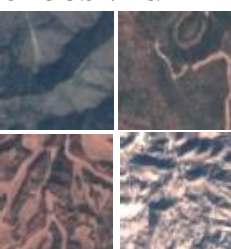
PERMANENT CROP



SEA LAKE



HERBACEOUS VEGETATION



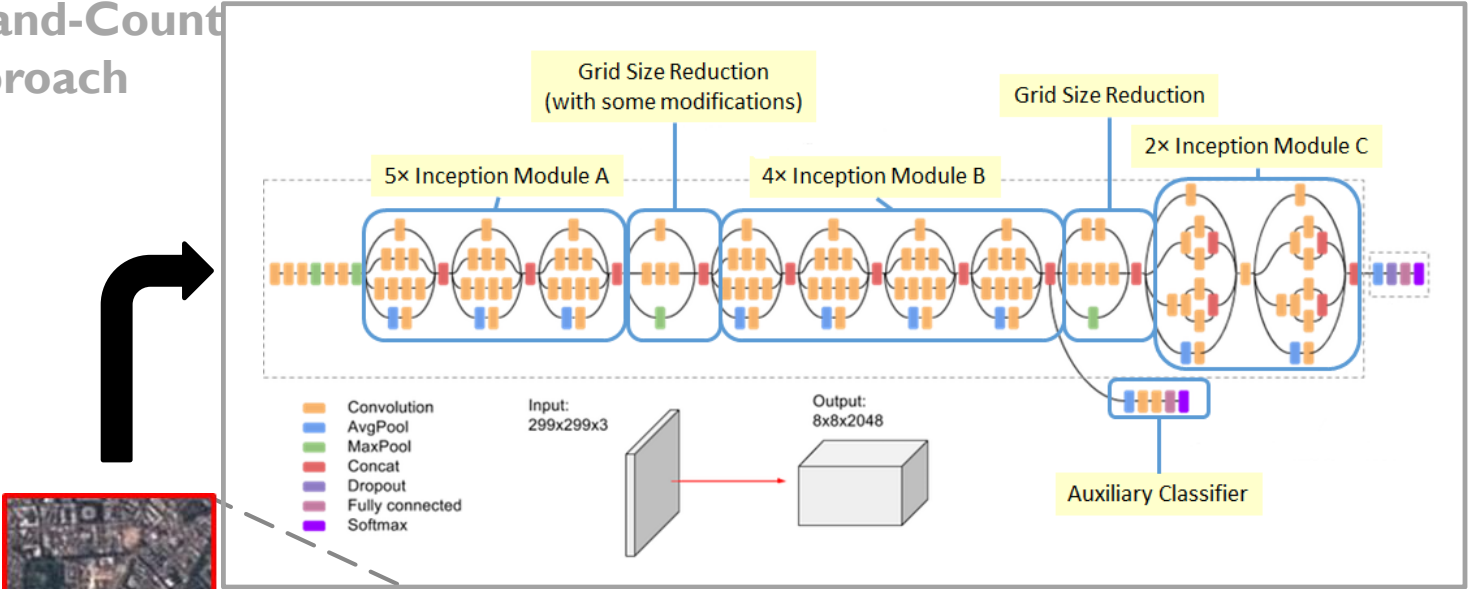
EuroSAT dataset

(<https://github.com/phelber/eurosat>):

- Based on Sentinel-2 satellite images
- 27000 geo-referenced and labeled image patches (each one of 64x64 pixels)
- 10 different Land Use and Land Cover classes, with 2000-3000 images per class
- RGB (8-bit) and Multi-Spectral (13 spectral bands, 16-bit) versions available

Use Case 2: AI for Land Cover

Classify-and-Count
Approach

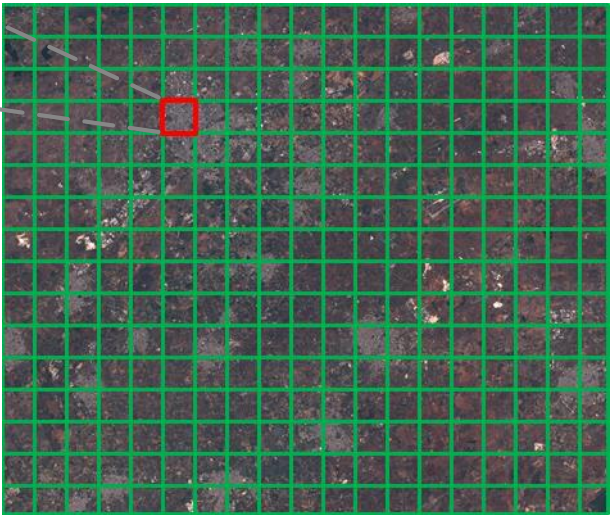


CNN: Inception-V3 Archite

RESIDENTIAL

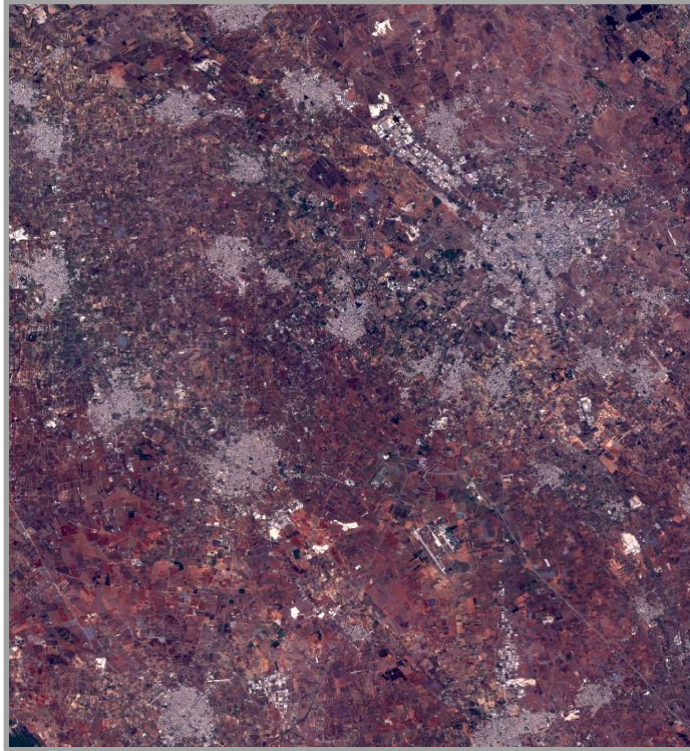


Input Satellite
Image



LAND COVER CLASS	AREA SHARE
...	...
RESIDENTIAL	$\frac{45}{16 * 19} \cong 15\%$
...	...

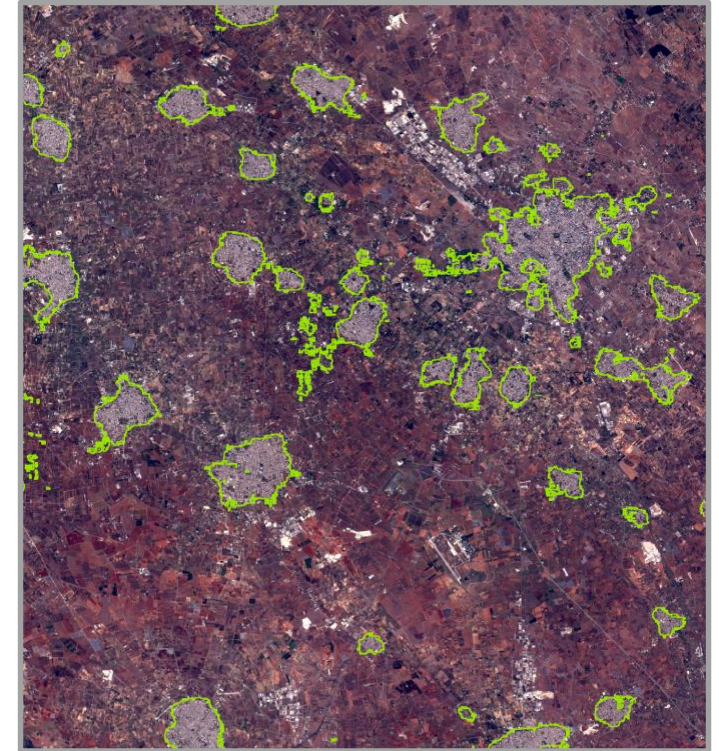
Use Case 2: AI for Land Cover



[A]
The ‘Lecce image’
(751 km²)



[B]
Automated LC map derived
from the ‘Lecce image’



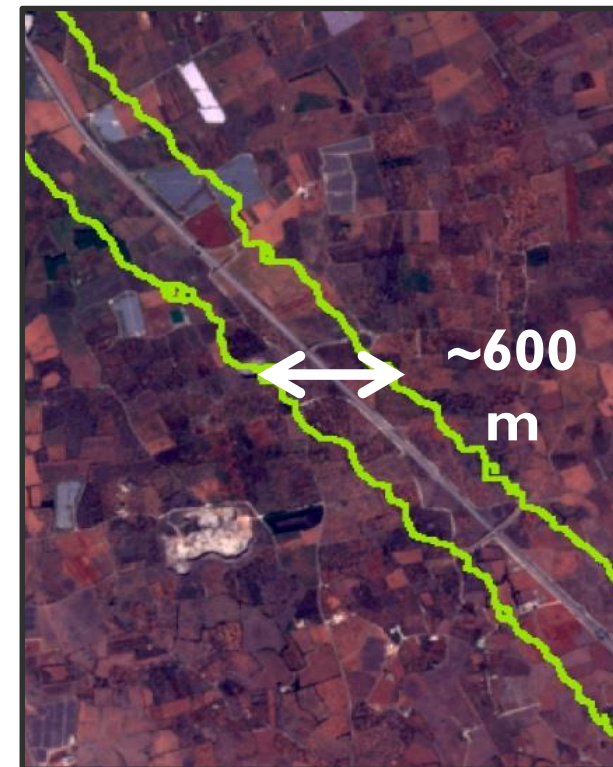
[C]
Edge line of the ‘Residential’
class derived from [B] overlaid
on [A]

Use Case 2: AI for Land Cover



[D]

A detailed view of the course of the Arno River (cropped from the **'Pisa image'**, 443 km²) overlaid with a semitransparent version of the corresponding automated LC map

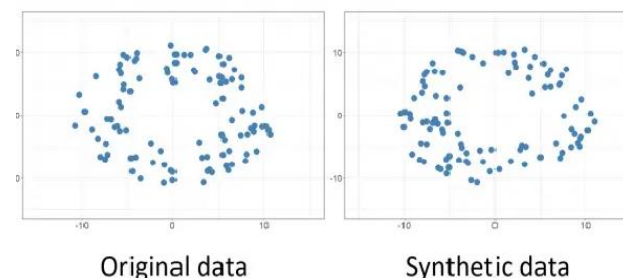


[E]

A highway fragment from the **'Lecce image'** overlaid with the edge line of the 'Highway' class

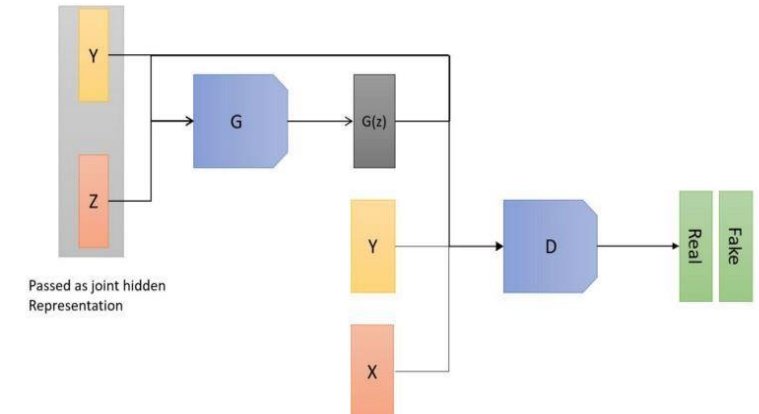
Use Case 3: AI for the generation of Synthetic Data

- ✓ With the **digitalization of information** and the increasing accessibility of **administrative data**, the amount of data to be handled has grown substantially in recent years. This raises significant concerns about **data protection** and **privacy** since the disclosure of sensitive information can pose serious risks to individuals, institutions, and public administrations.
- ✓ **Synthetic data** (*Pugliese, F., Pappagallo, A., & De Cubellis, M., 2024, June*) are artificially created datasets intended to **replicate** the statistical characteristics and structure of real data, while preventing the exposure of **sensitive or personally identifiable** information.



Use Case 3: AI for the generation of Synthetic Data

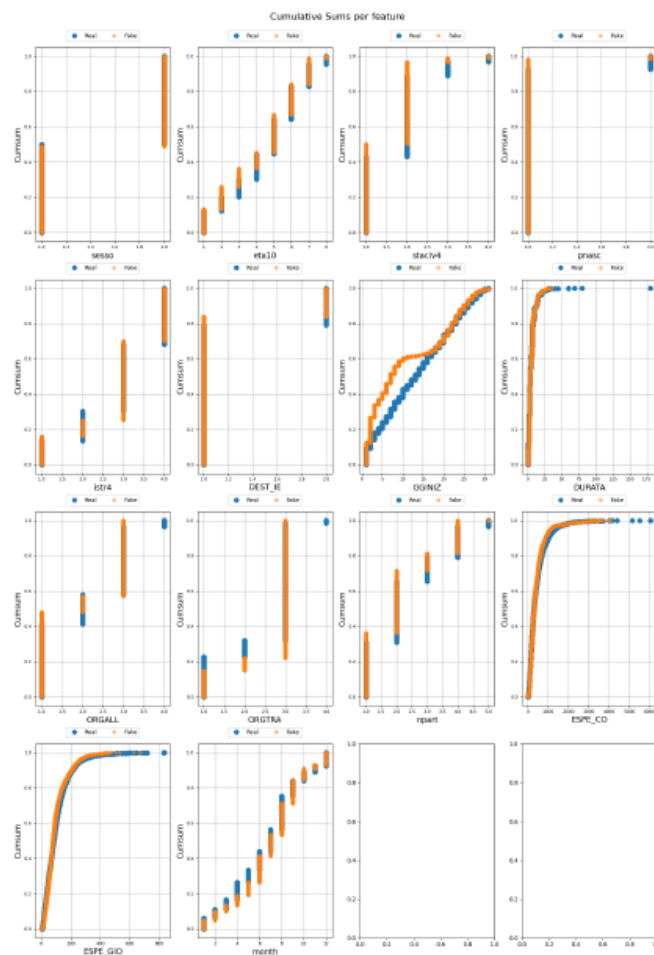
- ✓ **Keep in mind!** Synthetic data do not necessarily reproduce the atomic data (categorical, numerical, etc) from the original source. **Synthesis** capability might rely rather on the **relationships** between different kind of **entities**, such as people and objects, school and neighborhoods, or users and cellular antennas.
- ✓ In this study, the **analyzed data** relate to the frequency and attributes of trips and vacations undertaken by residents of Italy. They originate from **the Istat Trips and Holidays survey**, as a module of the **Household Budget Survey (HBS)** which collects information on the tourism flows of residents. This information includes journeys made for leisure or work, both domestically and internationally.
- ✓ In the WPI3's Istat PoC of AI/ML Essnet project, we compared different AI methods: **CT-GAN, VAEs, DP-CTGAN, SMOTENC, Random Forest, XGBoost**



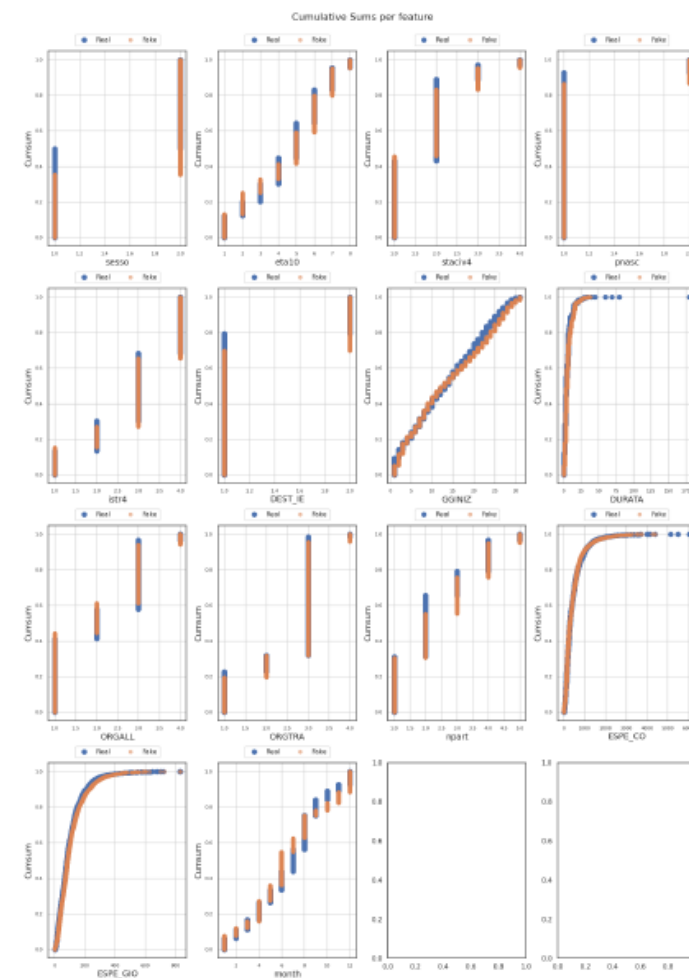
Use Case 3: AI for the generation of Synthetic Data

- ✓ Note: In the **original distribution itself**, the values are *not* sums of previous values. Only in the **cumulative sum version** do the values accumulate.
- ✓ It is **evident** from the **cumulative** distributions that the less effective methods, such as **Random Forest** or **XGBoost**, are **less** efficient on categorical variables, where they often **fail** to reproduce the categories, i.e. the domain values of these variables.

Variational Autoencoder (VAE)



CT-GAN



Use Case 3: AI for the generation of Synthetic Data

- ✓ As we can observe, apart from **SMOTENC** which tends to reproduce the data exactly, the best methods seem to be the deep learning-based ones such as CTGAN and VAE, as they achieve high accuracy but not identical to the original dataset, as expected. In fact, during the synthetic data generation process something is always **lost** in terms of the **properties** of the data being reproduced.

Model	Accuracy	F1-Score	Recall
Original Data	0.964516848	0.964516848	0.964516848
RF	0.660685592	0.660685592	0.660685592
XGB	4.46349e-05	4.46349e-05	4.46349e-05
SMOTENC	0.992813783	0.992813783	0.992813783
DPCTGAN	4.46349e-05	4.46349e-05	4.46349e-05
VAE	0.892385288	0.892385288	0.892385288
CTGAN	0.896134619	0.896134619	0.896134619

Conclusions and Next Steps

Conclusions

References

- [1] Catanese, E., Bruno, M., Stefanelli, L., & Pugliese, F. (2023, September). Measuring Social Mood on Economy during Covid times: effects of retraining Supervised Deep Neural Networks. In *5th International Conference on Advanced Research Methods and Analytics (CARMA 2023)* (pp. 139-147). Editorial Universitat Politècnica de València.

- [2] Bernasconi, E., De Fausti, F., Pugliese, F., Scannapieco, M., & Zardetto, D. (2022). Automatic extraction of land cover statistics from satellite imagery by deep learning. *Statistical Journal of the IAOS*, 38(1), 183-199.

- [3] Pugliese, F., Pappagallo, A., & De Cubellis, M. (2024, June). Generation of Synthetic Data from Mobile Network Operators (MNO) Data Through Generative Adversarial Networks (GANs). In *Scientific Meeting of the Italian Statistical Society* (pp. 69-74). Cham: Springer Nature Switzerland.

Thanks

FRANCESCO PUGLIESE | francesco.pugliese@istat.it