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

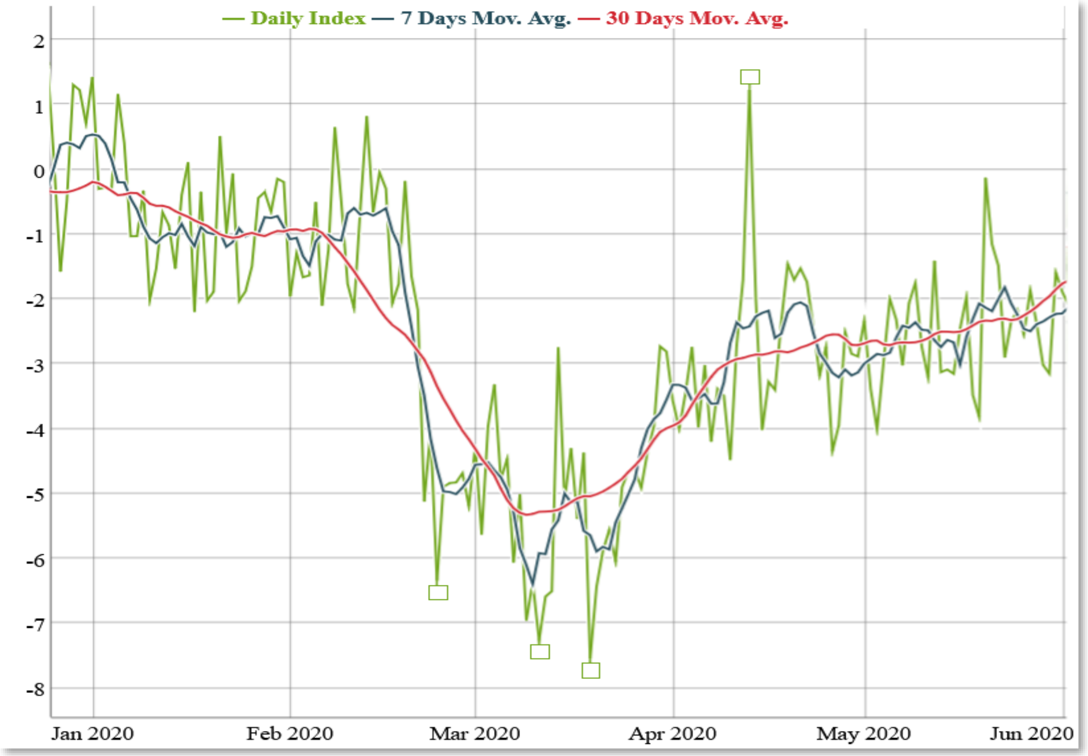
Caserta | 12 November 2025 | University of Campania “Luigi Vanvitelli” | Department of Mathematics and Physics | Abramo Lincoln Street, 5

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

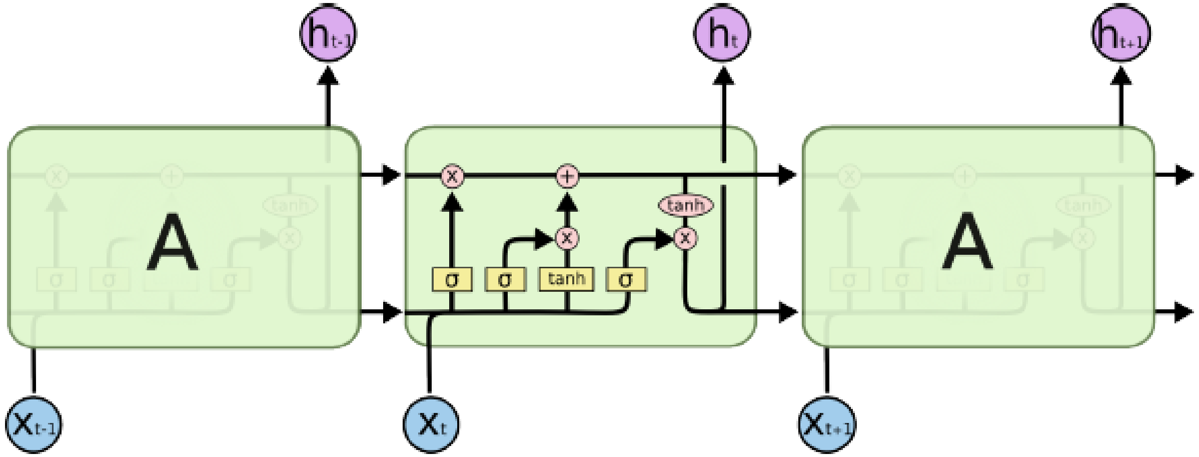


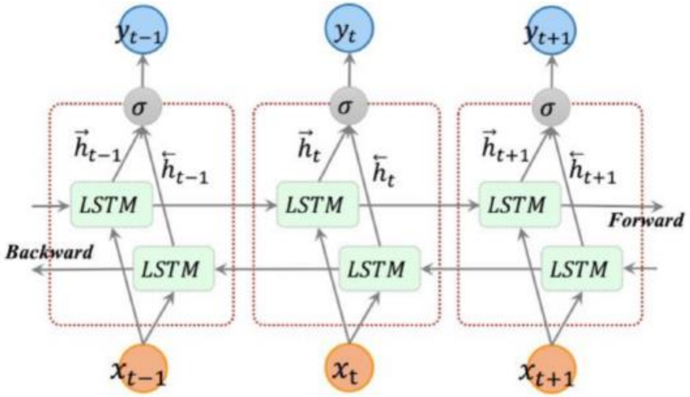
Use Case 1: AI for Sentiment Analysis

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?

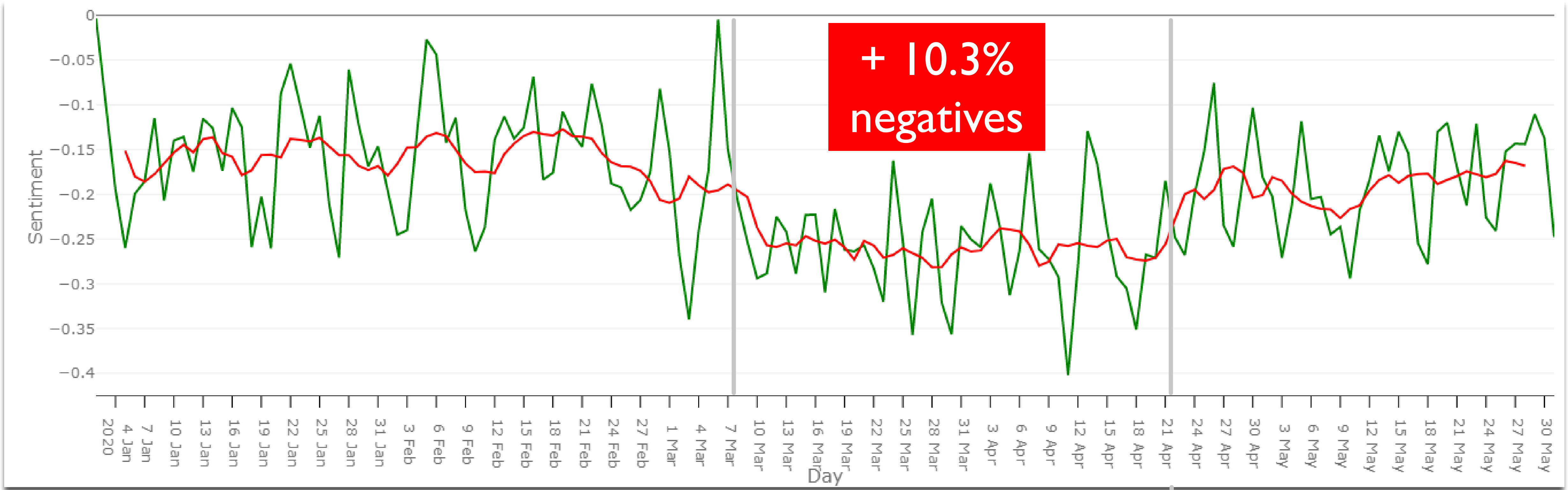
- ISTAT’s Index (SMEI) needed a revision to deal with new events (Covid pandemic)
- The proposed model may have a higher adaptability to new information
- The proposed model is a neural network that has been trained on a set of labelled Tweets that do not contain references to the pandemic

- 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)



+ 10.3% negatives

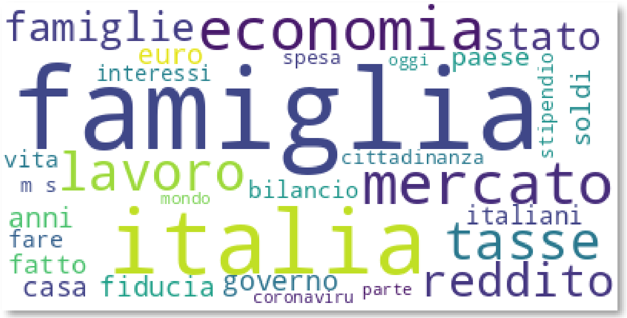
-0.157

-0.260

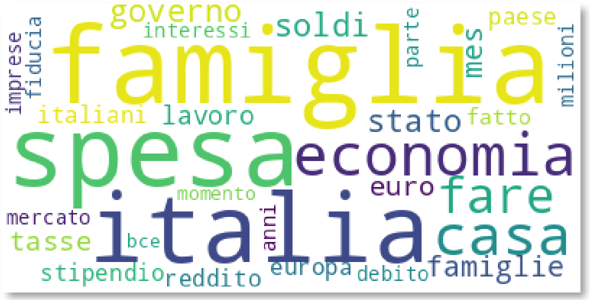
-0.190

Use Case 1: AI for Sentiment Analysis

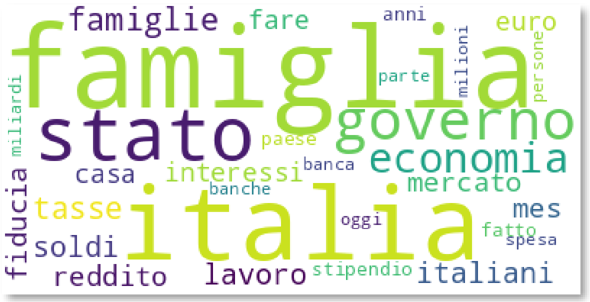
1st January – 7th March



8th March – 21st April



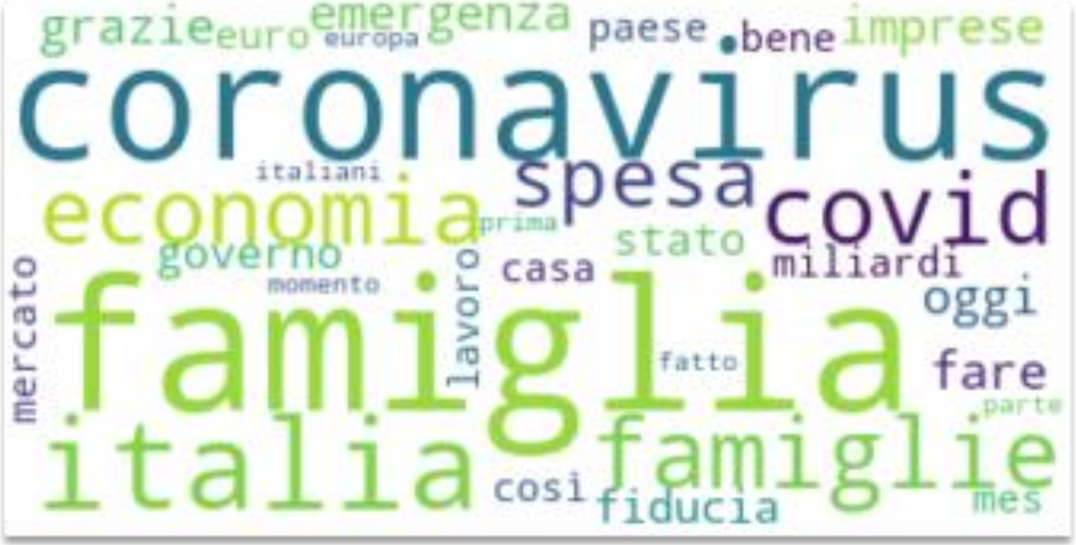
22nd April – 31st May



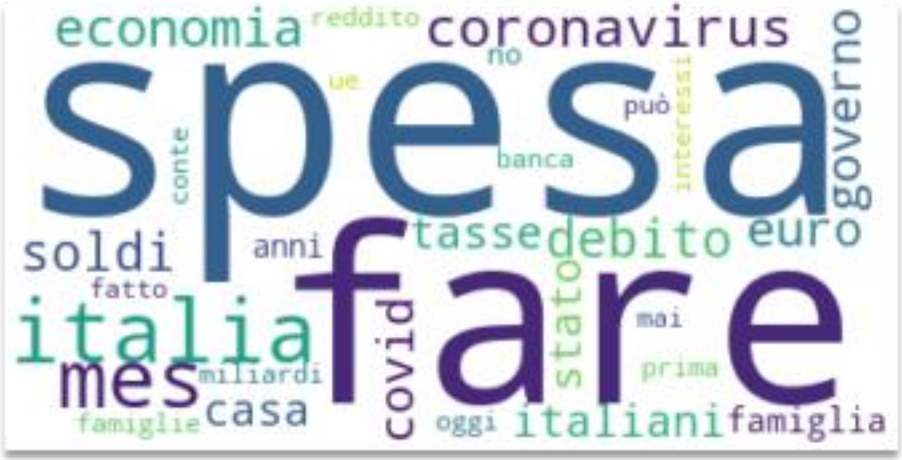
Use Case 1: AI for Sentiment Analysis

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

○Positive words

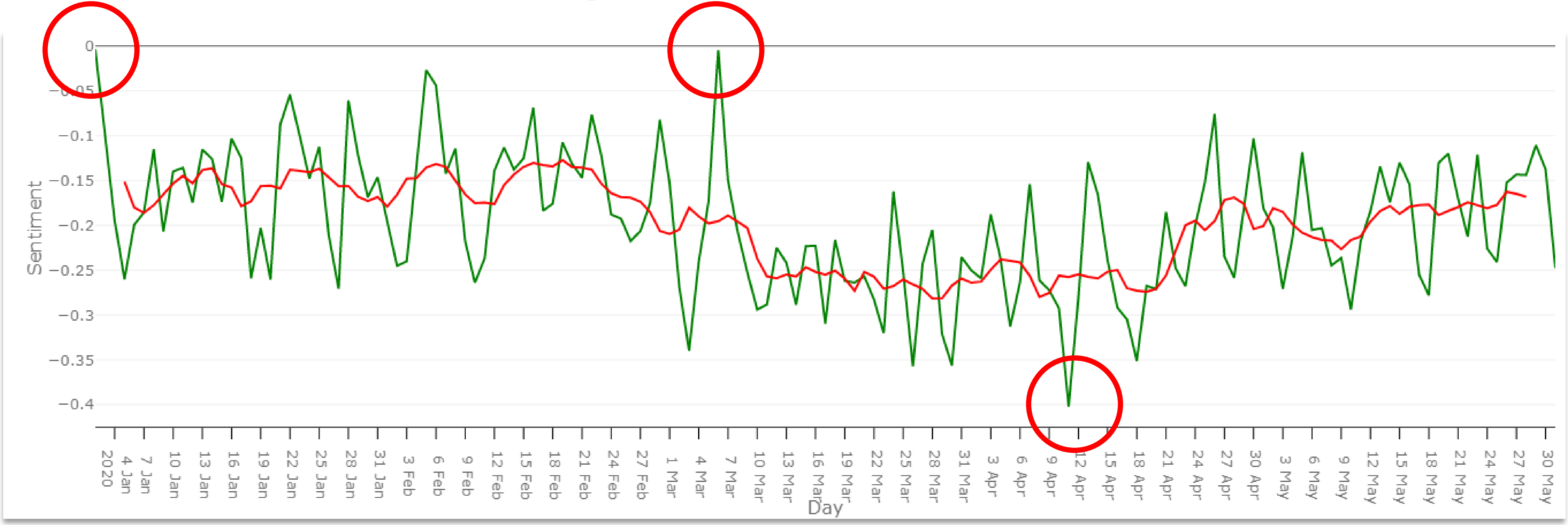


Negative words



Use Case 1: AI for Sentiment Analysis

Maximum and Minimum points

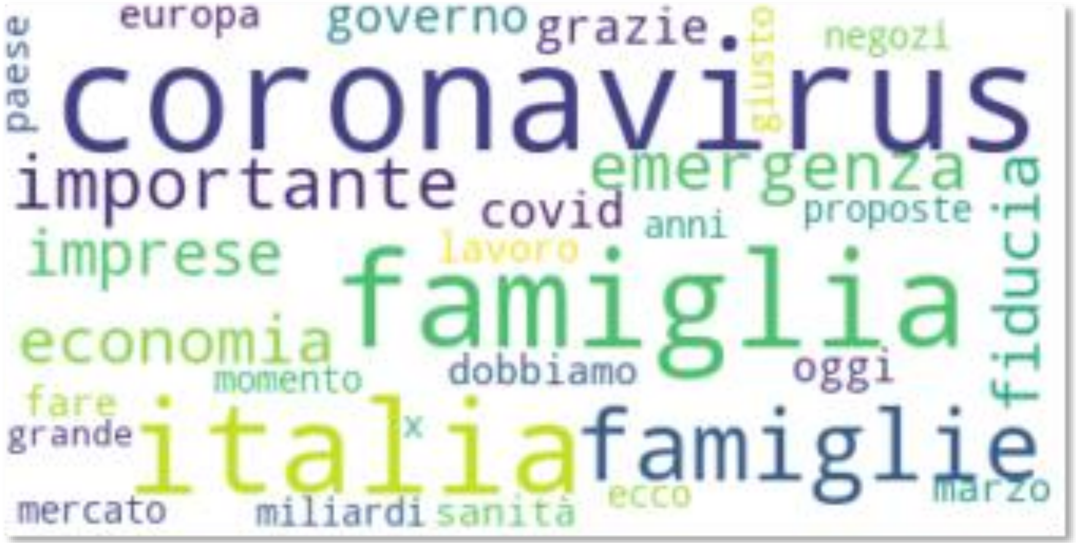


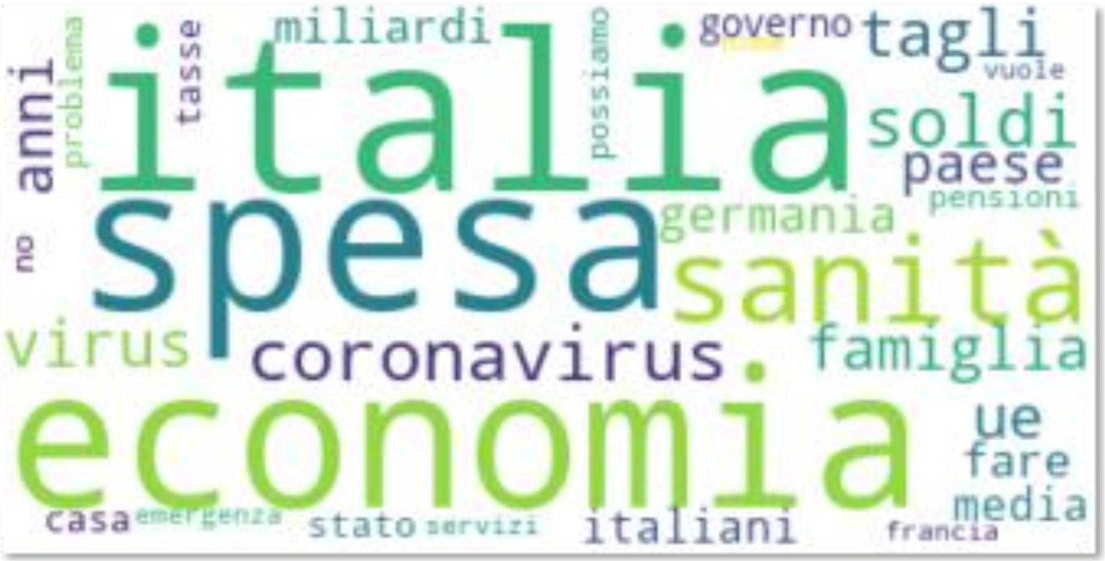
Use Case 1: AI for Sentiment Analysis

Second Maximum on the 6th March

○Positive words

Negative words





Use Case 2: AI for Land Cover

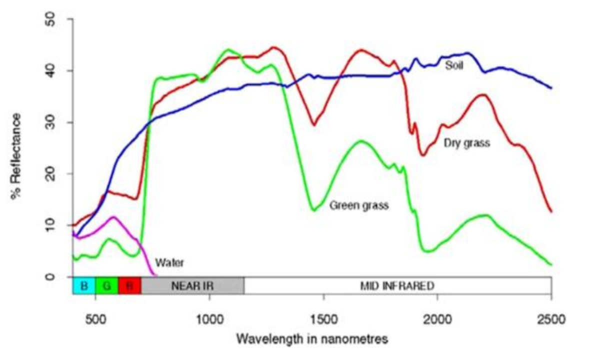
GOALS

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 sate

- Automatic Land Cover Estimates
- Automatic Land Cover Maps

ML Approaches to LC from Images

Standard approach: Spectral Signature

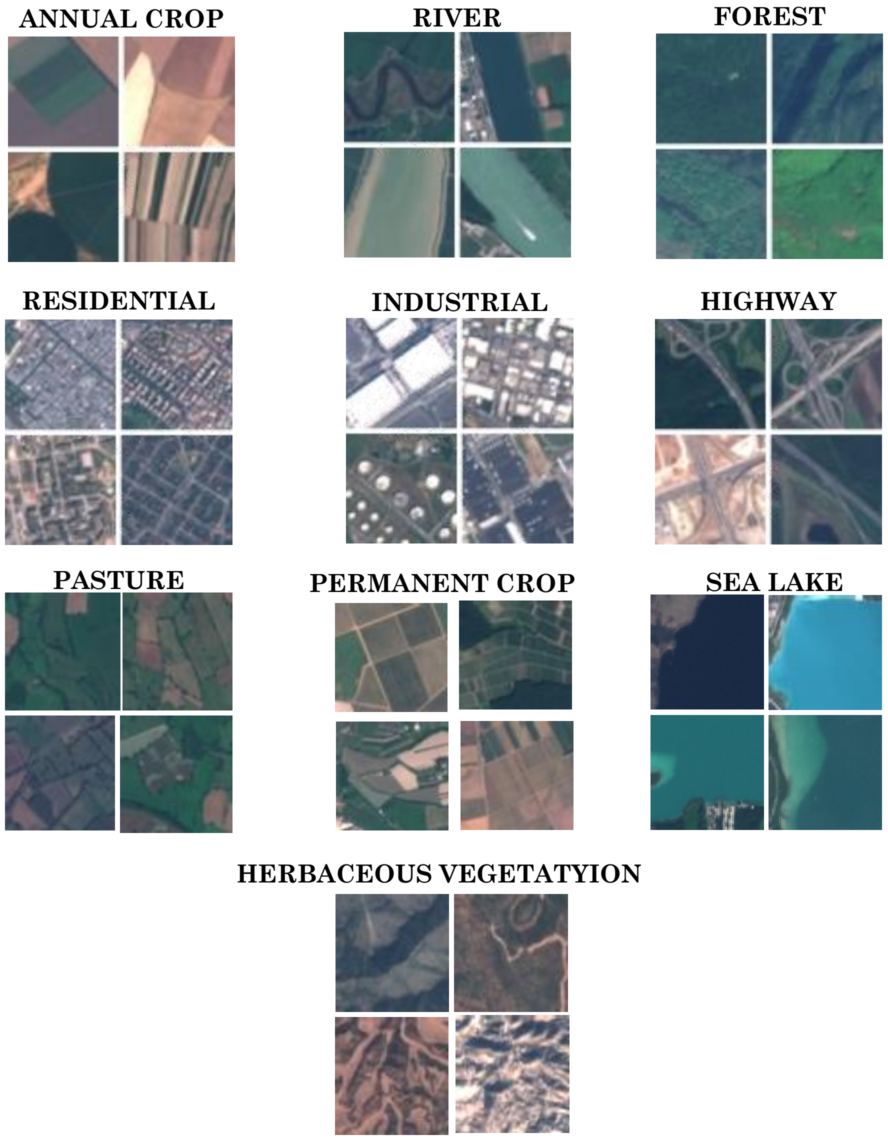


- Different LC classes have different reflectance spectra
  - ✓ Variation of reflectance with EM frequency can be used to predict LC class
  - ✓ Trained ML algo predicts the LC class of image pixels independently
  - ✓ Decision on each pixel does not depend on neighboring pixels

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 (



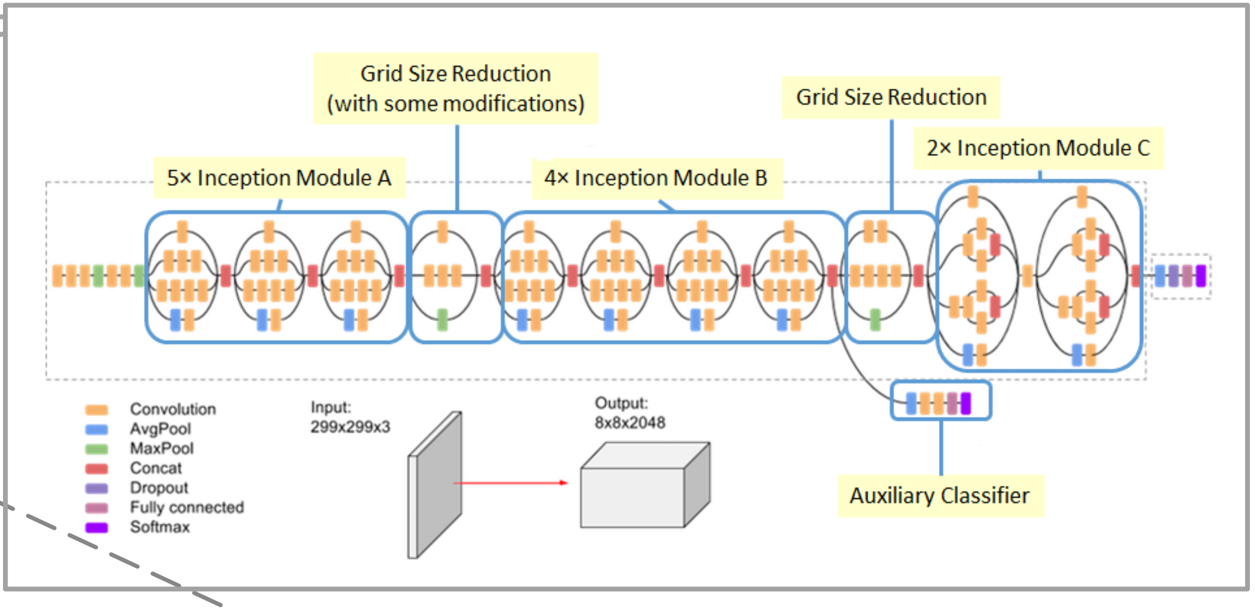
EuroSAT dataset

(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

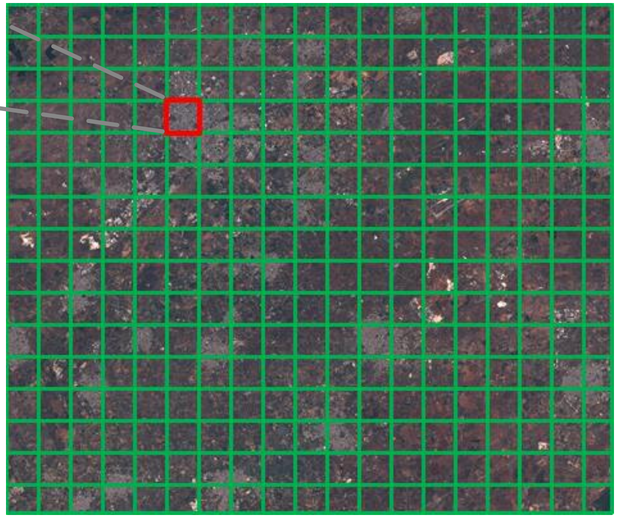
Classify-and-Count Approach



CNN: Inception-V3 Archite

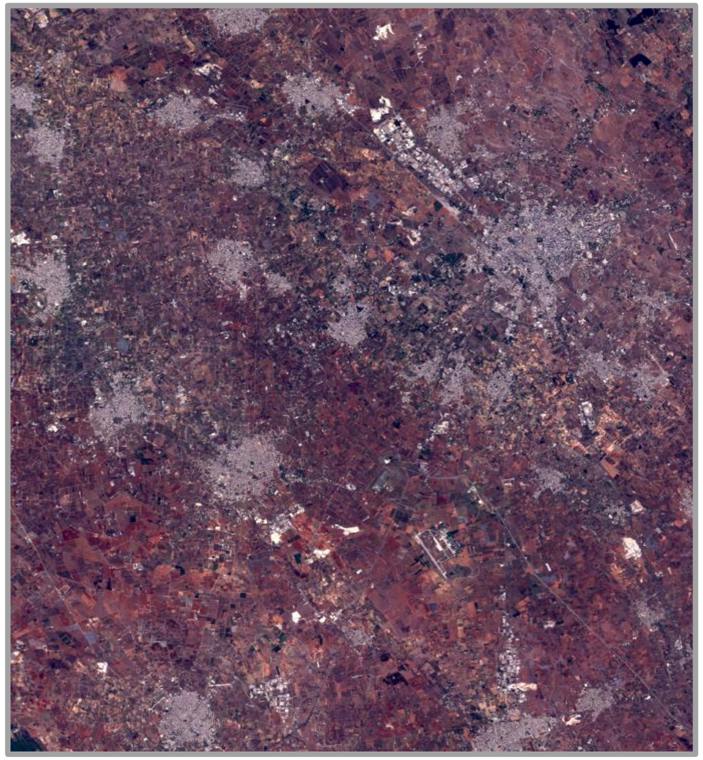
RESIDENTIAL

Input Satellite Image

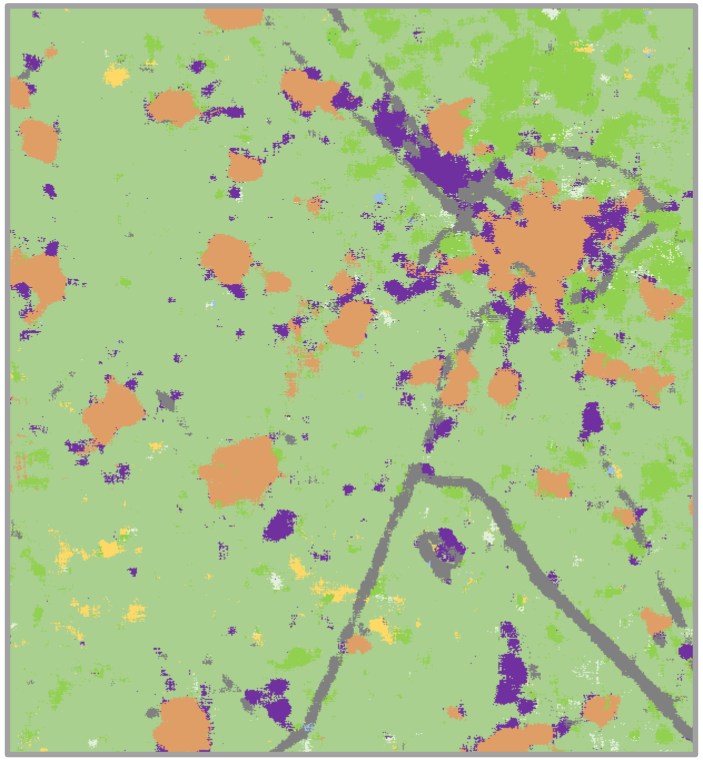


LAND COVER CLASS	AREA SHARE
…	…
RESIDENTIAL	$\frac{45}{16 * 19} \cong 15\%$
…	…

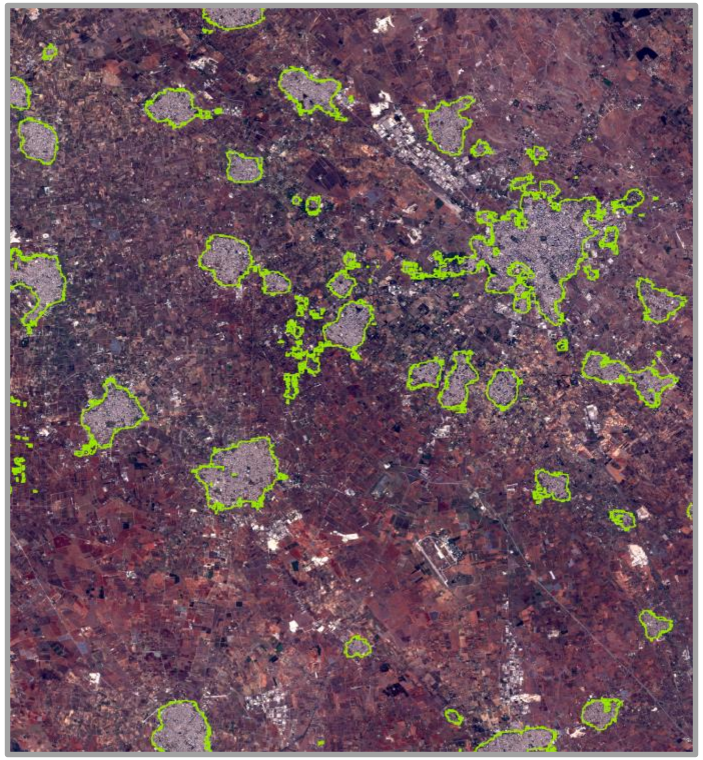
Use Case 2: AI for Land Cover



[A]
The ‘Lecce image’
(751 km^2)

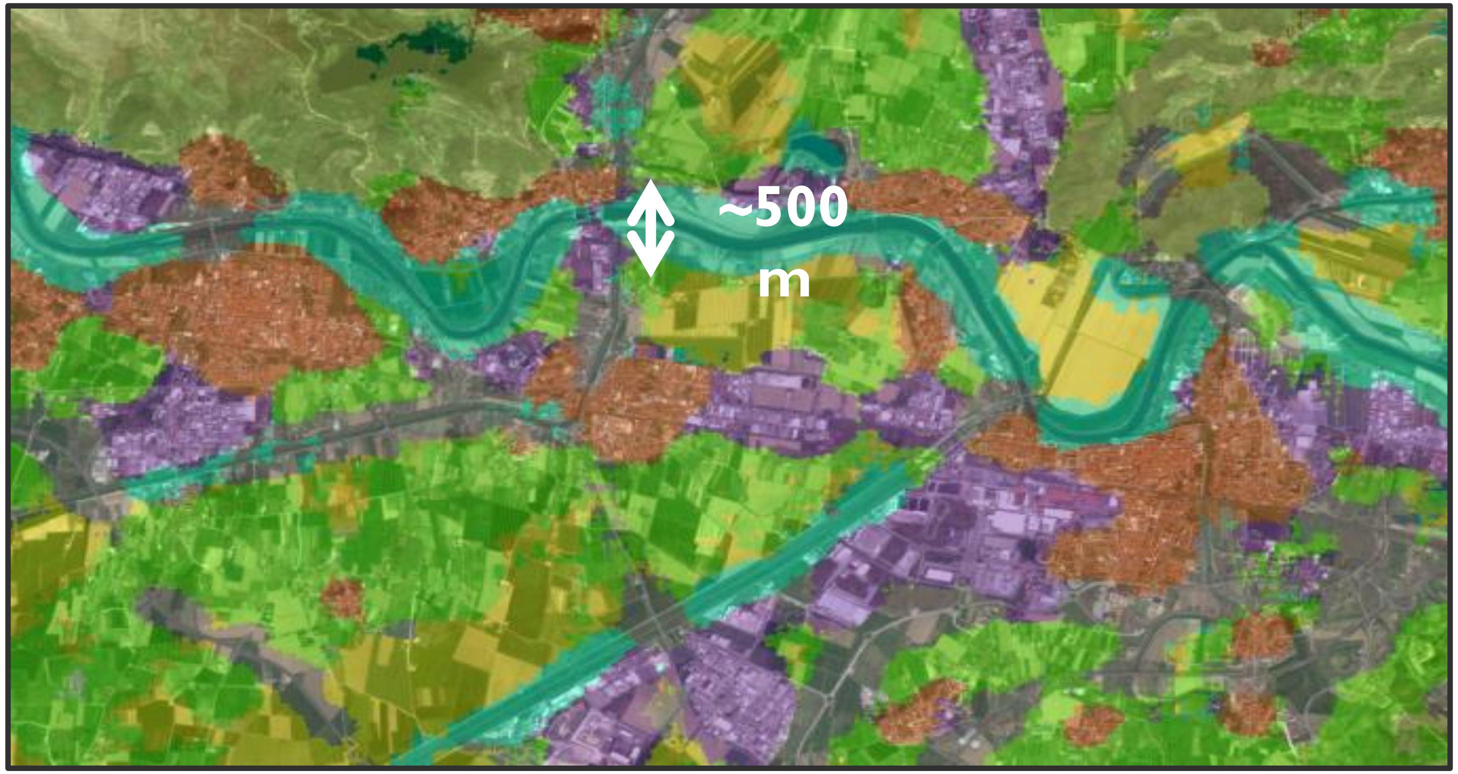


[B]
Automated LC map derived from the ‘Lecce image’

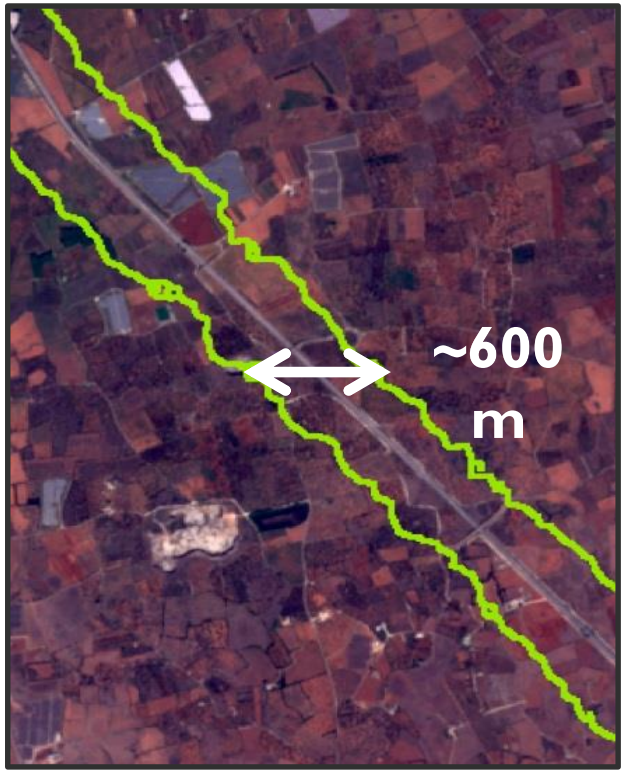


[C]
Edge line of the ‘Residential’ class derived from [B] overlaid on [A]

Use Case 2: AI for Land Cover



[D]
A detailed view of the course of the Arno River (cropped from the **'Pisa image'**, 443 km^2) overlaid with a semitransparent version of the corresponding automated LC map

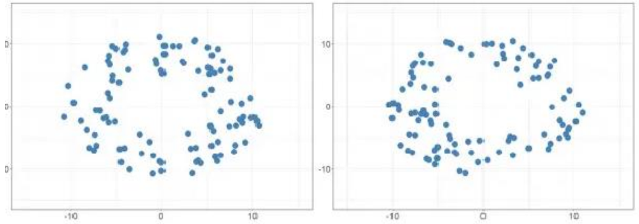


[E]
A highway fragment from the **'Lecce image'** overlaid with the edge line of the 'Highway' class

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.



Original data

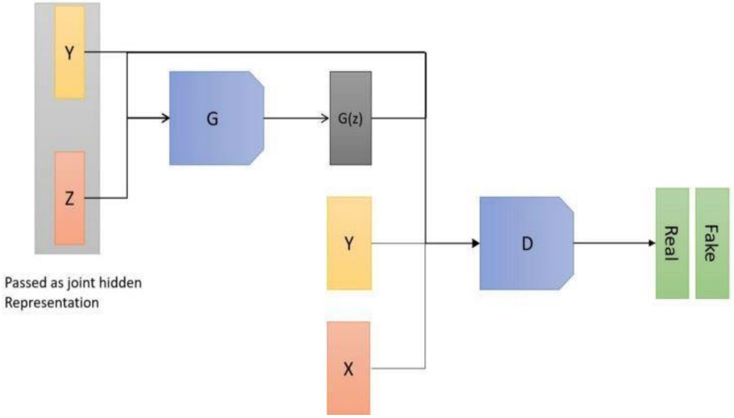
Synthetic data

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 WP13's Istat PoC of AI/ML Essnet project, we compared different AI methods: **CT-GAN, VAEs, DP-CTGAN, SMOTENC, Random Forest, XGBoost**

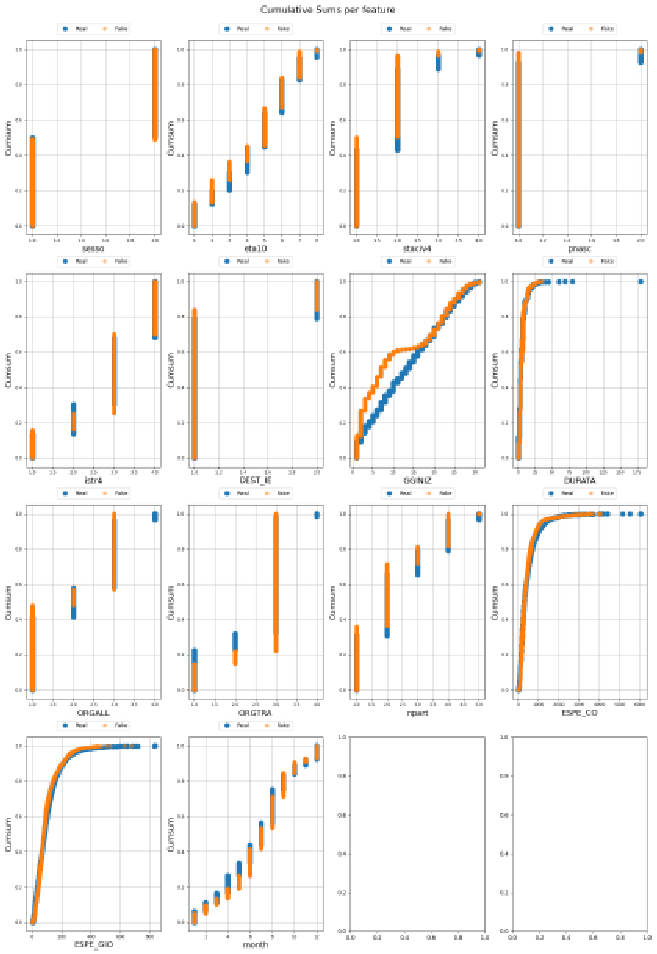


Use Case 3: AI for the generation of Synthetic Data

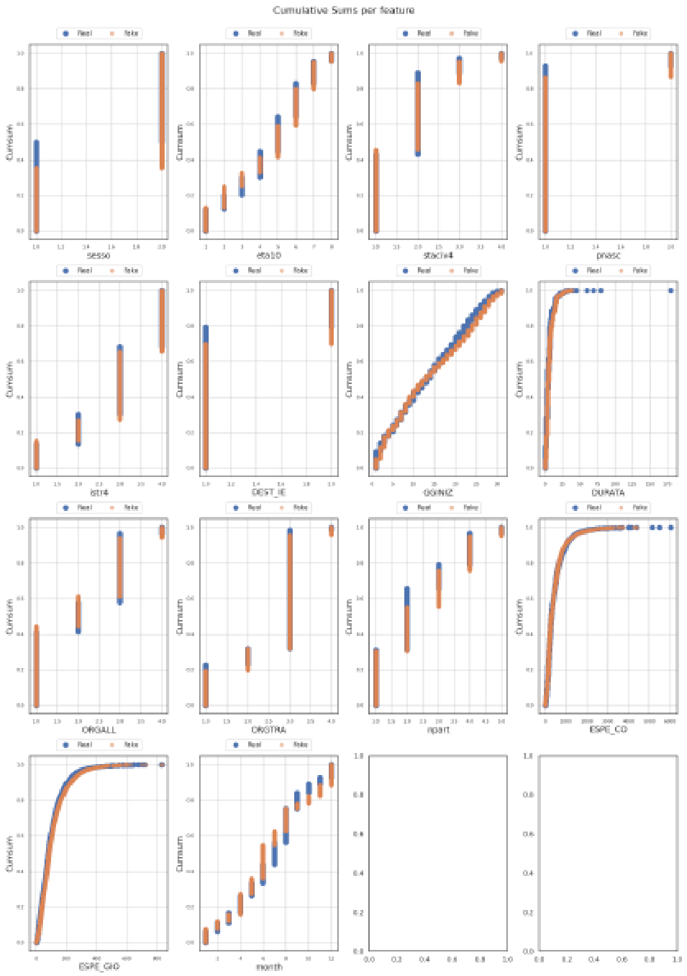
- ✓ Note: In the **original distribution itself**, the values are *not* sums of previous values. Only in the **cumulative sum version** do the values accumulate.

- ✓ 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



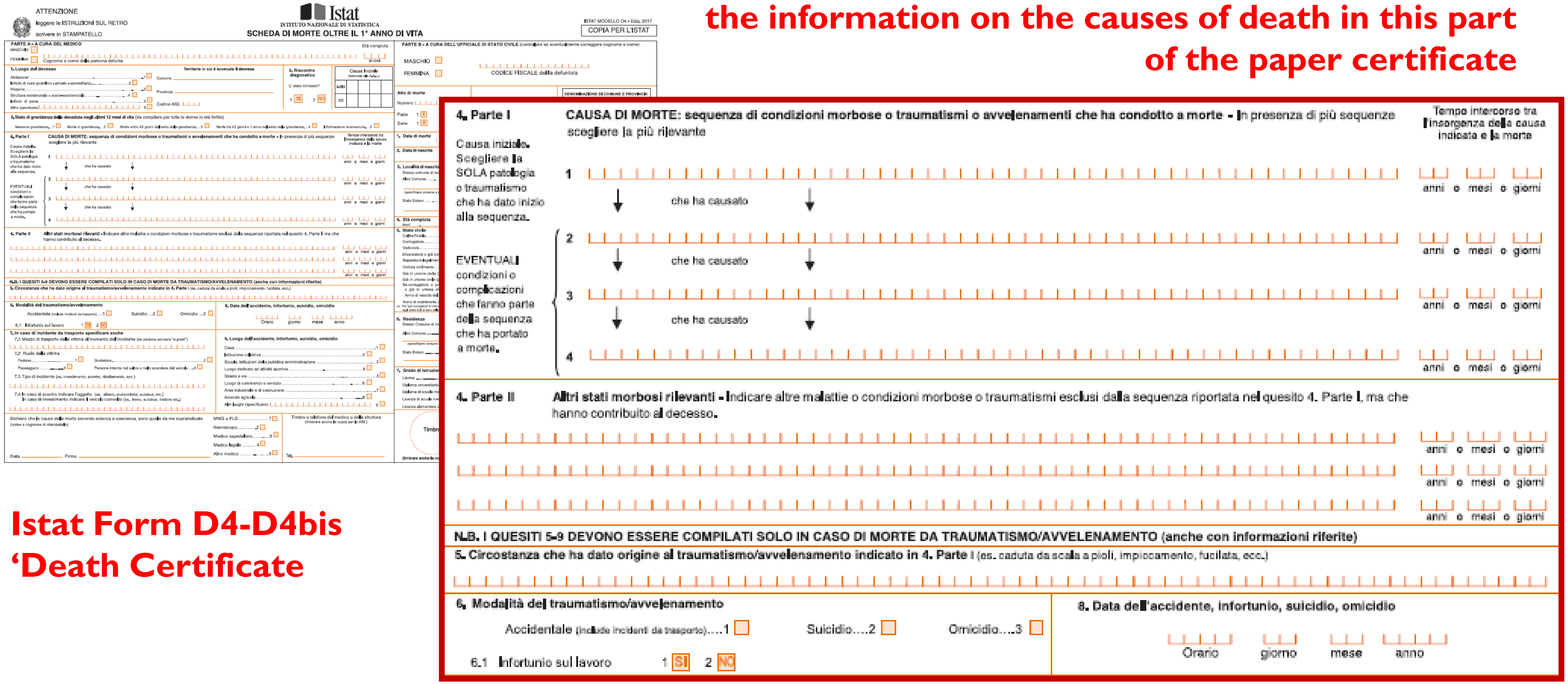
Use Case 3: AI for the generation of Synthetic Data

✓ 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

Use Case 4: AI for the automatic classification of Causes of Death

Detection Model



ATTENZIONE
leggere le ISTRUZIONI SUL RETRO
scrivere in STAMPATELLO

Istat
ISTITUTO NAZIONALE DI STATISTICA
SCHEDA DI MORTE OLTRE IL 1° ANNO DI VITA

ISTAT MODELLO D4 - Ediz. 2017
COPIA PER L'ISTAT

PARTE A - A CURA DEL MEDICO

MASCHIO □
FEMMINA □ Cognome e nome della persona defunta

Età compiuta (IN ANNI)

1. Luogo del decesso

Abitazione ....1 □
Istituto di cura (pubblico - privato - accreditato) ....2 □
Hospice ....3 □
Struttura residenziale o socio-assistenziale ....4 □
Istituto di pena ....5 □
Altro (specificare) ....9 □

Territorio in cui è avvenuto il decesso

Comune ____
Provincia ____
Codice ASL

2. Riscontro diagnostico

E' stato richiesto?

1 SI 2 NO

Causa Iniziale (riservato alla A.S.L.)

I-XIX				
XX				

3. Stato di gravidanza della deceduta negli ultimi 12 mesi di vita (da compilare per tutte le donne in età fertile)

Nessuna gravidanza....1 □ Morte in gravidanza....2 □ Morte entro 42 giorni dall'esito della gravidanza....3 □ Morte tra 43 giorni e 1 anno dall'esito della gravidanza....4 □ Informazione sconosciuta....5 □

4. Parte I **CAUSA DI MORTE: sequenza di condizioni morbose o traumatismi o avvelenamenti che ha condotto a morte** - In presenza di più sequenze scegliere la più rilevante

Tempo intercorso tra l'insorgenza della causa indicata e la morte

Causa iniziale. Scegliere la SOLA patologia o traumatismo che ha dato inizio alla sequenza.

1 ____ anni o mesi o giorni

che ha causato

EVENTUALI condizioni o complicazioni che fanno parte della sequenza che ha portato a morte.

2 ____ anni o mesi o giorni

che ha causato

3 ____ anni o mesi o giorni

che ha causato

4 ____ anni o mesi o giorni

4. Parte II **Altri stati morbosi rilevanti** - Indicare altre malattie o condizioni morbose o traumatismi esclusi dalla sequenza riportata nel quesito 4. Parte I, ma che hanno contribuito al decesso.

____ anni o mesi o giorni
____ anni o mesi o giorni
____ anni o mesi o giorni

N.B. I QUESITI 5-9 DEVONO ESSERE COMPILATI SOLO IN CASO DI MORTE DA TRAUMATISMO/AVVELENAMENTO (anche con informazioni riferite)

5. Circostanza che ha dato origine al traumatismo/avvelenamento indicato in 4. Parte I (es. caduta da scala a pioli, impiccamento, fucilata, ecc.)

6. Modalità del traumatismo/avvelenamento

Accidentale (include incidenti da trasporto)....1 □ Suicidio....2 □ Omicidio....3 □

6.1 Infortunio sul lavoro 1 SI 2 NO

8. Data dell'accidente, infortunio, suicidio, omicidio

Orario giorno mese anno

7. In caso di incidente da trasporto specificare anche

7.1 Mezzo di trasporto della vittima al momento dell'incidente (se pedone scrivere "a piedi")

7.2 Ruolo della vittima

Pedone ....1 □ Guidatore ....2 □
Passeggero ....3 □ Persona intenta nel salire o nello scendere dal veicolo....4 □

7.3 Tipo di incidente (es. investimento, scontro, ribaltamento, ecc.)

7.4 In caso di scontro indicare l'oggetto (es. albero, motocicletta, autobus, etc.)
In caso di investimento indicare il veicolo coinvolto (es. treno, autobus, trattore etc.)

9. Luogo dell'accidente, infortunio, suicidio, omicidio

Casa ....1 □
Istituzione collettiva ....2 □
Scuola, istituzioni della pubblica amministrazione ....3 □
Luogo dedicato ad attività sportive ....4 □
Strade e vie ....5 □
Luogo di commercio e servizio ....6 □
Area industriale e di costruzione ....7 □
Azienda agricola ....8 □
Altri luoghi (specificare) ____ 9 □

Dichiaro che le cause della morte secondo scienza e coscienza, sono quelle da me sopraindicate (nome e cognome in stampatello)

Data ____ Firma ____

MMG o PLS....1 □
Necroscopo....2 □
Medico ospedaliero....3 □
Medico legale ....4 □
Altro medico ....5 □

Timbro e telefono del medico o della struttura (timbrare anche la copia per la ASL)

Tel. ____

PARTE B - A CURA DELL'UFFICIALE DI STATO CIVILE (controllare ed eventualmente correggere cognome e nome)

MASCHIO □
FEMMINA □

CODICE FISCALE della defunto/a

Atto di morte

Numero

Parte 1 I
Serie 1 B

DENOMINAZIONE DI COMUNE E PROVINCIA

1. Data di morte

2. Data di nascita

3. Località di nascita

(specificare comune e

Stato Estero

4. Età compiuta

Anni

5. Stato civile

Celibe/Nubile
Coniugato/a
Vedovo/a
Unito/a civilmente

6. Residenza

Stato Estero

7. Grado di istruzione

Laurea
Diploma universitario

Timbro

4. Parte I **CAUSA DI MORTE: sequenza di condizioni morbose o traumatismi o avvelenamenti che ha condotto a morte** - In presenza di più sequenze scegliere la più rilevante

Tempo intercorso tra l'insorgenza della causa indicata e la morte

Causa iniziale. Scegliere la SOLA patologia o traumatismo che ha dato inizio alla sequenza.

1 ____ anni o mesi o giorni

che ha causato

EVENTUALI condizioni o complicazioni che fanno parte della sequenza che ha portato a morte.

2 ____ anni o mesi o giorni

che ha causato

3 ____ anni o mesi o giorni

che ha causato

4 ____ anni o mesi o giorni

4. Parte II **Altri stati morbosi rilevanti** - Indicare altre malattie o condizioni morbose o traumatismi esclusi dalla sequenza riportata nel quesito 4. Parte I, ma che hanno contribuito al decesso.

____ anni o mesi o giorni
____ anni o mesi o giorni
____ anni o mesi o giorni

N.B. I QUESITI 5-9 DEVONO ESSERE COMPILATI SOLO IN CASO DI MORTE DA TRAUMATISMO/AVVELENAMENTO (anche con informazioni riferite)

5. Circostanza che ha dato origine al traumatismo/avvelenamento indicato in 4. Parte I (es. caduta da scala a pioli, impiccamento, fucilata, ecc.)

6. Modalità del traumatismo/avvelenamento

Accidentale (include incidenti da trasporto)....1 □ Suicidio....2 □ Omicidio....3 □

6.1 Infortunio sul lavoro 1 SI 2 NO

8. Data dell'accidente, infortunio, suicidio, omicidio

Orario giorno mese anno

For each individual death occurring within the national territory, the certifying physician records the information on the causes of death in this part of the paper certificate

**Istat Form D4-D4bis
'Death Certificate**

Use Case 4: AI for the automatic classification of Causes of Death

The project aims to:

- identify the most suitable AI methods for coding causes of death;
- create an algorithm for coding causes of death by applying these methods;
- evaluate the performance of automatic coding and the quality of the data by applying this algorithm.

The main expected results are:

- a reduction in coding variability among coders and in errors, resulting in improved quality;
- in view of the imminent replacement of the paper form with an electronic model, AI methods will be useful to maintain high performance in recognizing medical text, even in the absence of the preprocessing currently applied during the registration of paper forms.
- an increase in the number of death certificates coded completely automatically, without human intervention;

Use Case 4: AI for the automatic classification of Causes of Death

The information on the causes is recorded by physicians on a paper death certificate, digitized, and coded according to the **tenth revision of the International Classification of Diseases (ICD-10)** of the **World Health Organization (WHO)**.

The application of the ICD is a complex process that involves:

- first, assigning a code to all the causes certified by physicians on the forms (multiple causes).
- then, selecting only one code for each death: the **underlying cause**.

Both phases are carried out according to the rules described in the ICD, and both the multiple causes and the underlying cause are used for mortality statistics.

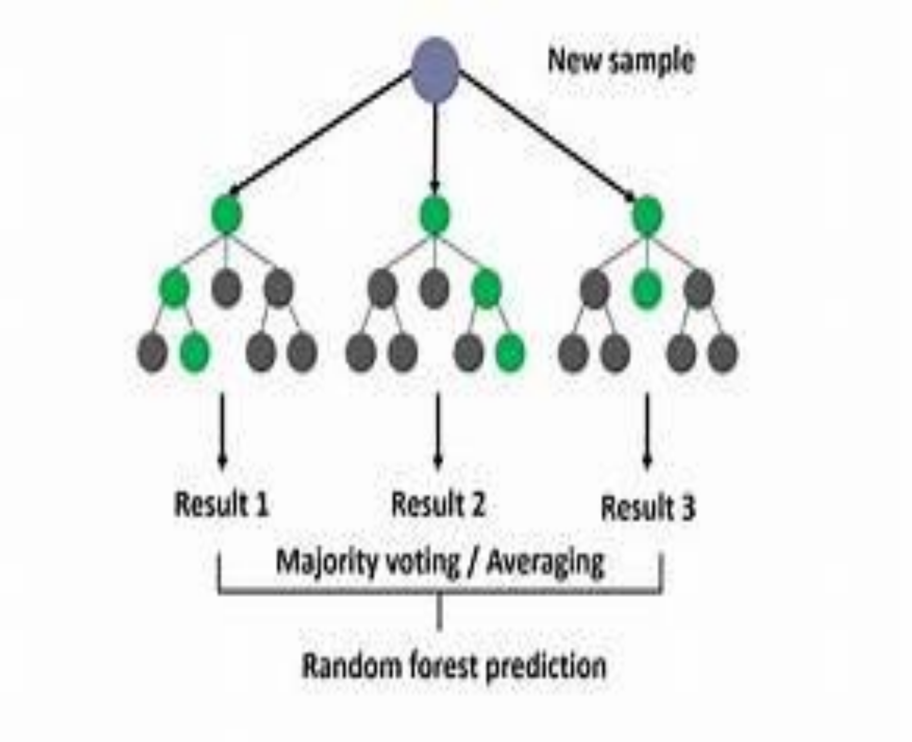
Use Case 4: AI for the automatic classification of Causes of Death

Training set 22_23: about 4 million texts.
53,876 classes after class cleaning and drops with cutoff.
Too much data for training: we sample progressively (starting from 300,000 texts).
DS_Gold_standard (external): 951 texts assigned a code whose correctness has been verified.
TEST_08_09 (test_set_blind): 11,909 texts to be labeled.

Valid ICD codes
Flag (A, C, CON, P, etc). Example:
Separators (';', '/', etc). Example: *CACHESSIA TERMINALE CON IRA IN GRANDE ANZIANO* coded as **N179 / R54**.
(Note: "Cachessia terminale con IRA in grande anziano" = "Terminal cachexia with acute renal failure in a very elderly patient.")

Use Case 4: AI for the automatic classification of Causes of Death

TF-IDF on the dataset to obtain the normalized term-document matrix; **Random Forest** is used for classification (which works on SL).
We take a sample of **300,000 texts** for training.
Initially **54,767 classes**, later reduced to **27,364**, after non-semantic splitting (in cases where the number of classes equals the number of texts).
We use a **cutoff of 2**, meaning we consider only classes with at least two instances.
There are about **17,666 classes** that appear only once, and these are removed.
In the end, we are left with **9,698 classes**.



Use Case 4: AI for the automatic classification of Causes of Death

On training set of 300,000:
Accuracy: 0.9761%
F1-score (macro-weighted): 0.9745%
On test set (20%):
Accuracy: 0.9467%
F1-score (macro-weighted): 0.9422%

Accuracy on test set of about 65,161 texts:

- #codes = #texts (size = 63,636, 97.66% of the test dataset): Accuracy: 0.9528% – F1-score (macro-weighted): 0.9505%
- #codes > #texts (size = 1,525, 2.34% of the test dataset): Accuracy: 0.69% – F1-score (macro-weighted): 0.71% → 1,060 correct, 465 incorrect (outputs exported to Excel and CSV for sharing)
- #texts > #codes (size = 0)

(Note: The "%" signs seem to be used instead of decimals — e.g., accuracy 0.9761 likely means 97.61%.)

Use Case 4: AI for the automatic classification of Causes of Death

Accuracy on test set of about 65,161 texts:
#codes > #texts (size = 1,525, 2.34% of the test dataset): Accuracy: 0.69% – F1-score (macro-weighted): 0.71% → 1,060 correct, 465 incorrect.

The correct cases are mostly those in which one text corresponds to two codes.
We check what percentage of the texts corresponding to the correct cases are present in the training set: **915 out of 1,050 (86%)**.
We check what percentage of the texts corresponding to the incorrect cases are present in the training set: **76 out of 475 (16%)**.
Theory: Random Forest performs well when it finds the same words combined in the same way in the training set.

Use Case 4: AI for the automatic classification of Causes of Death

Accuracy on test set (sample of 950 "gold_standard" texts):
For this dataset, the initial number of classes is **876**. After non-semantic splitting (in cases where the number of classes equals the number of texts — i.e., for 197 cases), we obtain **847 classes**, and the number of texts to be classified becomes **1,154**.
The classes are reduced to **716** after removing the time period.

Accuracy on "gold_standard" test set: overall **0.41% F1-score** → problem: only **24% of the gold standard texts** are present in the training set, while **62% of the classes** are present in the training set.
•#codes = #texts (size = 948, 82.15% of the gold standard dataset): Accuracy: 0.55% – F1-score (macro-weighted): 0.52%
•#codes > #texts (size = 206, 17.85% of the gold standard dataset): Accuracy: 0.033% – F1-score (macro-weighted): 0.033% → 7 correct, 199 incorrect (outputs exported to Excel and CSV for sharing)
•#texts > #codes (size = 0)

OBSERVATION: for the label encoding of the classes, we include both the classes present in the training set and those in the gold standard, to prevent cases where classes in the gold standard are missing from the training data.
In the previous case, this issue did not arise because we had obtained it using a **stratified train-test split**, which ensures that each class has a sample in both the training and test sets.

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

Istat | Istituto Nazionale di Statistica