# INNOVARE LA PRODUZIONE CENSUARIA CON TECNICHE DI INTELLIGENZA ARTIFICIALE

Conferenza Nazionale di Statistica

**ANTONELLA BERNARDINI-** ISTAT - anbernar@istat.it | **NICOLETTA CIBELLA-** ISTAT - cibella@istat.it | **ANTONIO LAUREATI PALMA -** ISTAT - lauretip@istat.it |

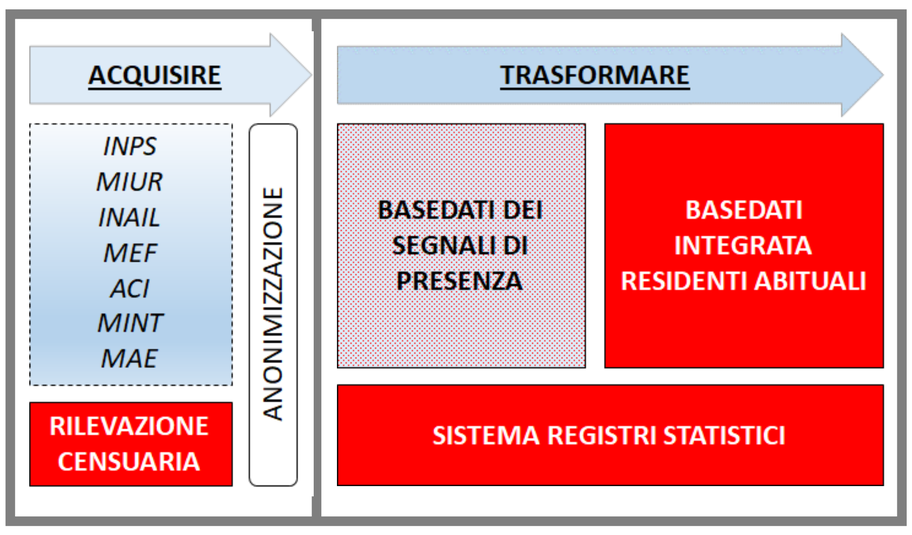
Gli Istituti Nazionali di Statistica (INS) storicamente producono statistiche basate su modelli prevalentemente probabilistici. Solo negli ultimi anni le opportunità fornite dai processi di digitalizzazione della PA dei dati amministrativi consentono importanti innovazioni di processo. In questo contesto **le tecniche di Intelligenza Artificiale possono permettere agli INS di fornire nuovi e più tempestivi prodotti statistici**.

Nel nuovo percorso censuario, intrapreso dal 2020, l'uso congiunto di dati amministrativi e di indagine permette di stimare la popolazione abitualmente dimorante e valutarne la corretta collocazione sul territorio. **Il nuovo processo del Censimento Permanente della Popolazione richiede una architettura produttiva di nuova generazione** in grado di sostenere il cambiamento epistemologico nella produzione delle informazioni statistiche.

Il nuovo processo di produzione statistica è articolato su una architettura concettuale a tre livelli:

I. livello organizzativo - è necessaria la compartecipazione attiva di più esperti nel processo di estrazione, comprensione e valorizzazione della conoscenza dei dati;
II. livello di processo - un cambio paradigmatico nella definizione del ciclo di vita del dato, proprio del *Data Mining*;
III. livello infrastrutturale - supportare la gestione di grandi volumi di dati, l'elaborazione efficace degli algoritmi di Machine Learning e gli strumenti per il data analytics.

Le tecniche di Machine Learning sui dati amministrativi, trasformati in segnali di presenza degli individui nella fonte, 'segnali di vita', utilizzano la rilevazione campionaria censuaria per l'addestramento del modello di Machine Learning




Il **ciclo di vita del dato** si articola in quattro fasi:

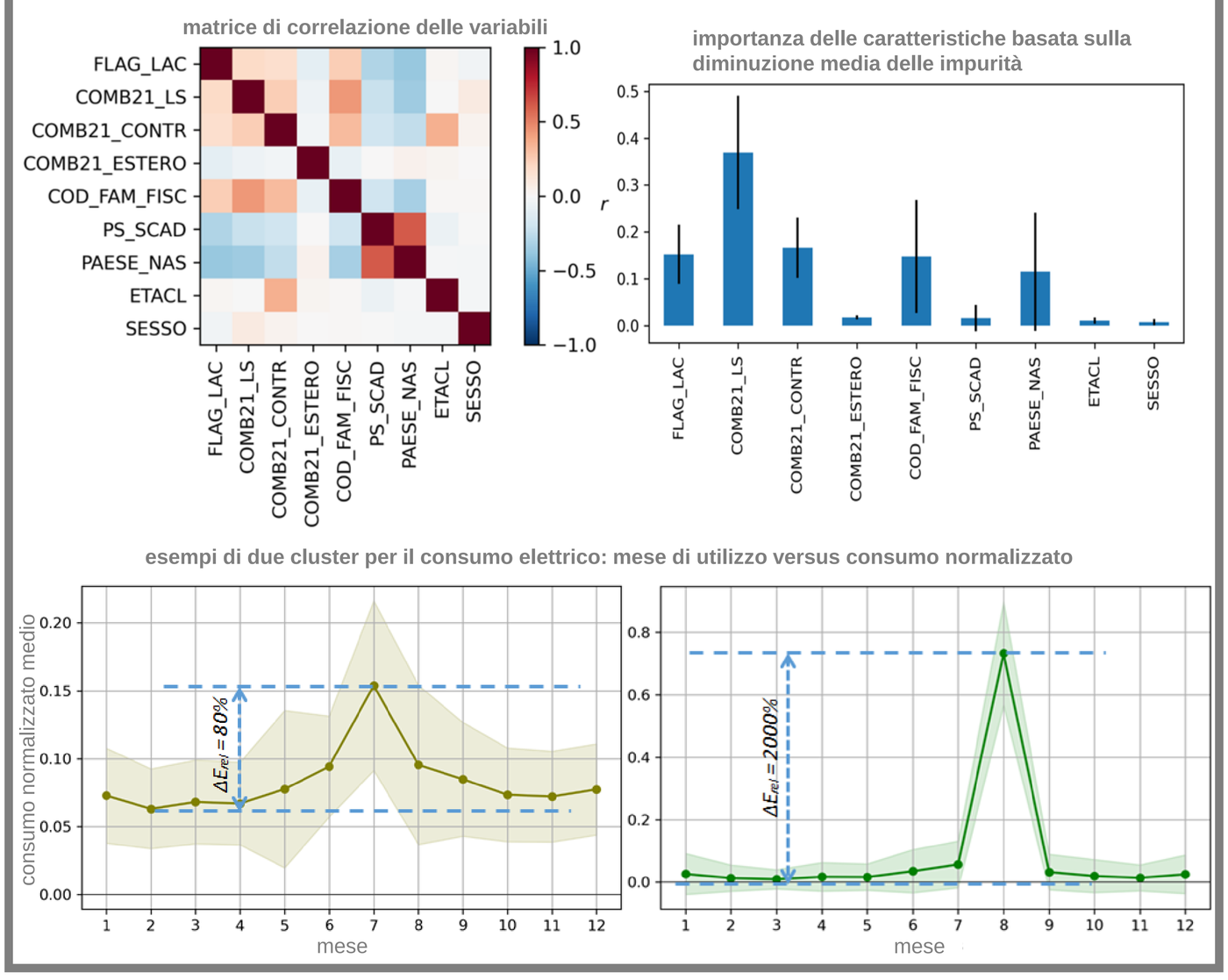
- ACQUISIRE: **si integrano più fonti amministrative con i dati della rilevazione sul campo**, condotta ogni anno su circa 2.530 comuni e oltre 4.000.000 di individui. Tutti i dati sono anonimizzati attraverso il sistema interno dell'Istituto.

- TRASFORMARE: **i dati amministrativi sono riorganizzati in segnali di presenza e monitorati rispetto alle fonti disponibili annualmente**. I segnali sono categorizzati in modo ordinale a seconda della rilevanza della fonte o attraverso trasformazioni *ad hoc*. Ad esempio, i consumi energetici sono clusterizzati e classificati a seconda del profilo di utilizzo, uso residenziale e non.

- CONOSCERE: in questa fase diverse competenze professionali sono coinvolte nell'attività di definizione del processo. **Sono individuate le variabili d'interesse nelle fonti disponibili e le diverse competenze tematiche interagiscono per lo sviluppo del piano di coerenza dei dati**. Il piano è finalizzato alla costruzione dei set di addestramento degli algoritmi di Machine Learning per la classificazione supervisionata del conteggio della popolazione.

- INFERIRE: la fase inferenziale delinea l'identificazione dei modelli di Machine Learning più adatti allo scopo. Nel nostro contesto è stato utilizzato il *Support-Vector Machines* con *Radial Basis Function kernel* (SVM-RBF). Sono stati confrontati diversi modelli; **si è scelto il SVM per le ottime performance di accuratezza riscontrate e soprattutto perché consente un facile utilizzo bilanciato delle classi**, 'dimoranti abitualmente' e 'non dimoranti abitualmente' considerando che il loro rapporto è dell'ordine del 3%.


## COMPRENDERE




**RISULTATI**

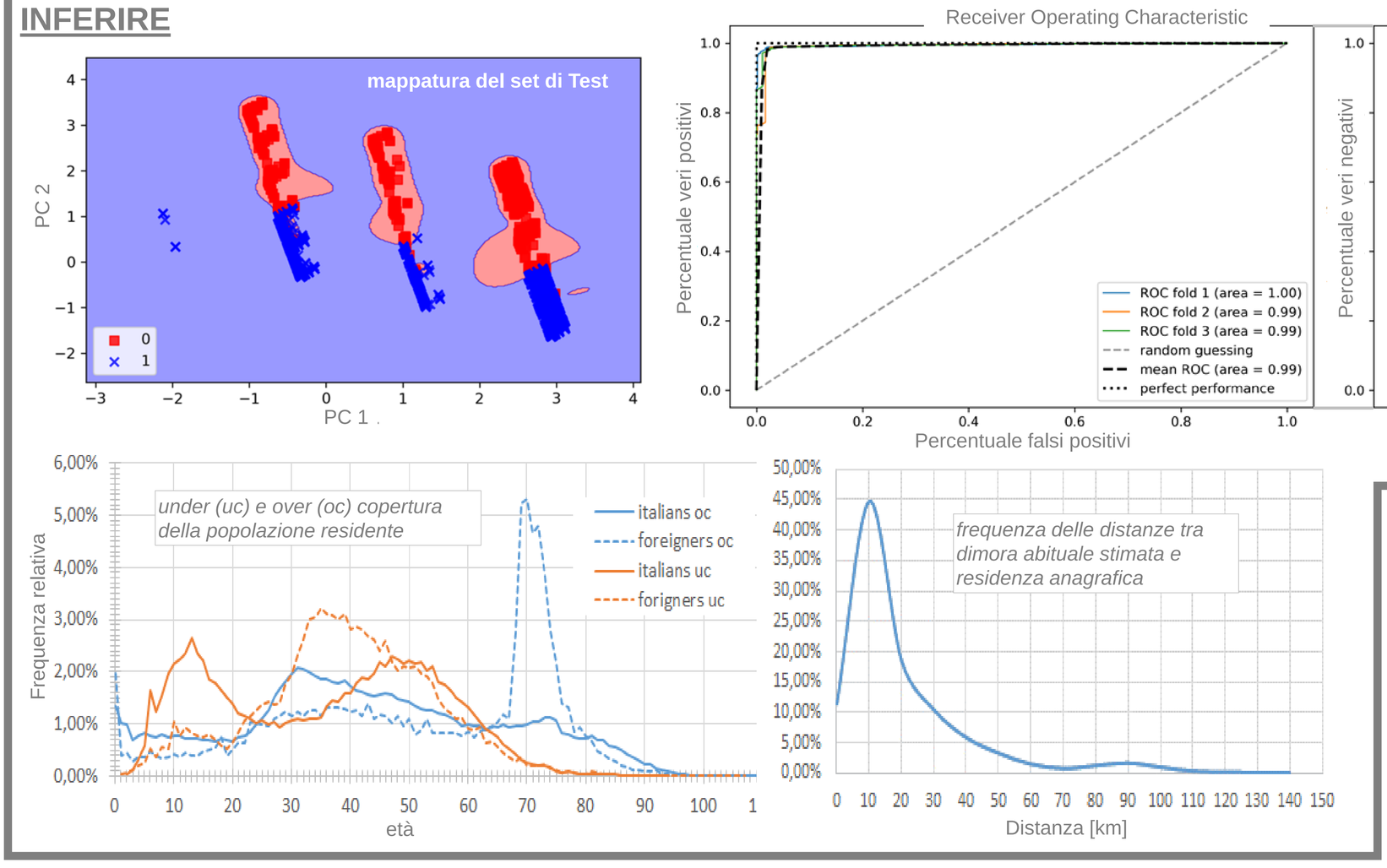
Il processo di trasformazione che integra i diversi archivi amministrativi ha consentito di costruire un database dei segnali di presenza per 60.511.100 record per l'anno 2021, ciascun record è stato classificato attraverso il Machine Learning in 'dimorante abitualmente' e 'non dimorante abitualmente' in Italia.

**Il training set, utilizzato per addestrare i modelli di classificazione, è stato costruito a partire dai dati delle indagini campionarie 2021 del Censimento Permanente della Popolazione.** In particolare sono stati utilizzati circa 1.300.000 record, rilevati dall'indagine Areale, per definire la classe 'dimoranti abitualmente', e circa 46.000 record di non rispondenti all'indagine da Lista, con i relative esiti di non presenza, per la classe dei 'non dimoranti abitualmente'.

**In questo contesto è, quindi, necessario affrontare la sfida di gestire una classificazione sbilanciata**: la popolazione dei 'dimoranti' (circa il 98% di tutti i 'segnali di vita') è molto più grande di quella dei 'non dimoranti' (circa il 3% dei dati).

I risultati ottenuti sono stati confrontati con le informazioni presenti nel registro anagrafico. **E' stata riscontrata una concordanza tra modello di Machine Learning e Registro anagrafico del 99,6% e una discordanza di circa lo 0,3% per la sovra copertura e di circa lo 0,1% per la sotto copertura**. Le performance delle due classificazioni sono mostrate indicativamente nella mappatura dei domini del set di test e quantitativamente nei due grafici relativi alla curva del *Receiver Operating Characteristic* (ROC).
Nel grafico in basso a sinistra sono mostrati i risultati ottenuti per la popolazione italiana e straniera, per la sovra e sotto coperta, in funzione dell'età. Nel grafico in basso a destra sono riportate le frequenze della distanza tra dimora abituale stimata da modello e residenza anagrafica.

## INFERIRE




**CONCLUSIONI**

Il conteggio della popolazione abitualmente dimorante attraverso le tecniche di ML è basato su un nuovo processo epistemologico per la produzione statistica (*data science*), dove il ciclo di vita del dato facilita la multidisciplinarietà. Infatti, alla base del processo risiede l'individuazione del miglior set di addestramento in grado di rappresentare e consentire la classificazione sbilanciata della popolazione dimorante abitualmente e non.

In questo lavoro viene descritto ed applicato un processo produttivo basato sulle tecniche di Intelligenza Artificiale nel contesto della statistica ufficiale. **I risultati ottenuti dimostrano l'adeguatezza del percorso intercorso per un miglioramento della qualità complessiva del processo produttivo, sia rispetto agli output statistici che di performance**.

**Referenze:**

- A. Bernardini, N. Cibella, et al (2024), "Improving the design of the Italian permanent population and housing census: a transition towards a massive use of administrative data" METRON, Springer, Sapienza Università di Roma, vol. 82(1), pages 5-17, April
- Zuppardo, M., Calian, V., Harðarson, Ó. , (2022) Machine learning estimation of the resident Population, Statistical Journal of the IAOS
- UNECE, (2021).Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses, United Nations, New York