CAMBIAMENTO CLIMATICO E VULNERABILITÀ DEMOGRAFICA:

"Un'analisi delle unità territoriali in Emilia-Romagna"

Botticelli Tommaso, Cesari Jacopo, Zecchini Giovanni

TECHNICAL REPORT

1. Introduzione

Questo report tecnico illustra i metodi e le tecniche impiegate per la pulizia, elaborazione e analisi del dataset MEMOREC, contenente informazioni su fattori ambientali, sociali, demografici e climatici nella regione Emilia-Romagna.

L'analisi è stata condotta al fine di rispondere a specifiche domande di ricerca riguardanti la relazione tra condizioni ambientali ed eventi climatici con la struttura demografica.

L'intera analisi è stata implementato in Python, per gestire grandi quantità di dati ed applicare tecniche avanzate di analisi statistica e machine learning, come cluster analysis e modelli di regressione lineare e logistica. Il report descrive in dettaglio le fasi del processo, dalla pulizia e imputazione dei dati all'aggregazione e costruzione di variabili derivate, fino all'analisi e interpretazione dei risultati.

2. Pulizia ed elaborazione dei dati

2.1 Trattamento delle mensilità

Sono state adottate le seguenti procedure per la gestione dei dati mancanti e delle anomalie nella variabile 'Mese':

- Le osservazioni con valori mancanti (NA) nella variabile 'Mese' sono state imputate con il valore 0, al fine di preservare l'integrità del dataset e garantire la coerenza dell'analisi.
- Le osservazioni con 'Mese' pari a 13, rappresentative di valori aggregati su base annuale, sono state rimosse, in quanto non coerenti con l'unità temporale di riferimento dell'analisi mensile.

2.2 Trattamento delle Unità Territoriali Aggregatesi nel Tempo

Un aspetto critico è stata gestione delle unità territoriali soggette a modifiche amministrative, come fusioni comunali o cambiamenti di confini. Per garantire la coerenza dei dati nel tempo, è

stata implementata una mappatura dei codici e dei nomi delle unità territoriali, unificando i dati dei comuni che hanno subito modifiche amministrative. Sono stati inoltre esclusi i comuni che hanno cambiato regione e non fanno più parte dell'Emilia-Romagna.

Le variabili d'interesse sono state classificate in due categorie principali a seconda del processo di aggregazione desiderato, e sono state aggregate per somma o per media ponderata.

2.3 Rimodellazione del Dataset

Il dataset è stato trasformato dalla forma long, in cui gli indicatori erano riportati nella stessa colonna, alla forma wide, con le variabili distribuite lungo le colonne. Questa trasformazione è stata effettuata per ottimizzare l'analisi e garantire uniformità nella copertura temporale delle variabili annuali e mensili. Per le variabili annuali, i dati sono stati replicati su base mensile, in modo da preservare la struttura temporale del dataset.

2.4 Imputazione dei Valori Mancanti

I valori mancanti sono stati trattati con diverse tecniche in base alla natura della variabile. Per le variabili geografiche, come l'altitudine e la superficie territoriale, è stata utilizzata l'imputazione con il valore valido più vicino per garantire coerenza. Per le variabili continue, come temperature medie e percentuali, è stata adottata la media comunale.

Per i decessi suddivisi per fasce di età, l'imputazione è stata effettuata sulla base della distribuzione dei decessi totali. In particolare, quando mancavano i dati sui decessi per fasce di età ma era disponibile il totale, questi sono stati stimati applicando la proporzione media osservata. Viceversa, quando il totale risultava mancante ma erano presenti i decessi di una specifica fascia, il valore complessivo è stato ricostruito in modo coerente con la distribuzione osservata.

2.5 Creazione di Variabili Derivate

Sono state calcolate nuove variabili utili per migliorare l'analisi ed altre già presenti sono state ricalcolate in forma percentuale per permettere un confronto tra comuni. Sono stati inoltre calcolati indicatori climatici, e tassi demografici unità territoriale.

3. Analisi esplorativa

3.1 Preparazione alle analisi

Sono stati creati due ulteriori dataset a partire da quello principale: uno aggregato a livello annuale e uno a livello comunale, contenente le medie temporali delle variabili. Questo approccio ha consentito un'analisi sia longitudinale (per l'identificazione dei trend temporali) sia trasversale (per l'analisi delle differenze territoriali). Inoltre, le principali variabili sono state suddivise in due categorie: ambientali-climatiche e socio-demografiche, al fine di semplificare e focalizzare l'analisi su ciascun tipo di variabile.

3.2 Cluster Analysis

La cluster analysis è stata eseguita dopo aver standardizzato le variabili. Il numero ottimale di cluster è stato determinato utilizzando il metodo del gomito (Elbow Method) e successivamente validato tramite l'indice di Silhouette. Questa analisi ha permesso di individuare gruppi di comuni con caratteristiche ambientali simili, facilitando l'identificazione di aree con specifiche peculiarità territoriali.

Per valutare la significatività delle differenze tra i cluster in relazione alle variabili ambientali e demografiche, è stato effettuato un test ANOVA. Inoltre, sono state calcolate statistiche descrittive e creati grafici radar per un confronto visivo delle caratteristiche dei gruppi.

4. Analisi empirica

L'analisi del dataset è stata guidata da una serie di domande di ricerca finalizzate a comprendere l'influenza dei fattori ambientali e demografici sulla popolazione e sulla vulnerabilità territoriale della regione Emilia-Romagna.

4.1 Quali fattori ambientali caratterizzano le aree soggette a spopolamento?

4.1.1 Modello di Regressione Lineare Semplice

È stato costruito un modello di regressione lineare per esaminare l'impatto delle variabili ambientali sullo spopolamento, utilizzando i dati sulle percentuali di spopolamento tra il 2004 e il 2023. Le variabili ambientali sono state utilizzate come covariate, con l'aggiunta di una costante.

4.1.2 Modello Ridotto

Le variabili non significative sono state eliminate dal modello, e il Variance Inflation Factor (VIF) è stato calcolato per verificare la collinearità tra le covariate.

4.2 In che modo l'età anziana della popolazione si lega allo stato degli edifici e alla vulnerabilità ambientale?

4.2.1 Correlazione lineare

Sono state calcolate le correlazioni tra la percentuale di popolazione anziana e variabili come lo stato degli edifici e la vulnerabilità ambientale nel 2018. La matrice di correlazione è stata visualizzata tramite un heatmap.

4.2.2 Modello di Regressione

Un modello di regressione lineare ha analizzato la relazione tra la popolazione anziana e variabili ambientali e socio-demografiche. Sono state effettuate trasformazioni logaritmiche e a radice quadrata della variabile dipendente.

4.2.3 Modello Ridotto

Sono state eliminate le variabili non significative e aggiunti termini quadratici per migliorare la qualità del modello.

4.3 Quali sono i principali fattori ambientali che influenzano la mortalità nella popolazione anziana (over 65)?

4.3.1 Percentili di Mortalità

Sono stati calcolati il 98° e il 2° percentile del tasso di mortalità della popolazione anziana (65+), per identificare i comuni con tassi di mortalità estremi (alti e bassi).

4.3.2 Modelli Logistici

Sono stati costruiti modelli Logit per analizzare i fattori ambientali che influenzano il tasso di mortalità alta (≥ 98° percentile) e bassa (≤ 2° percentile) nella popolazione anziana.

- Modello Logit per Mortalità Alta (98° Percentile): La variabile dipendente è stata codificata come binaria (alta mortalità o meno). Sono stati eliminati i predittori meno significativi, e sono stati calcolati gli Odds Ratios (OR) per interpretare l'effetto delle variabili.
- Modello Logit per Mortalità Bassa (2° Percentile): Un modello simile è stato applicato per analizzare i fattori che influenzano la mortalità bassa.

4.4 Quali comuni sono maggiormente esposti ai rischi naturali?

Per identificare i comuni più esposti ai rischi naturali, sono state selezionate variabili ambientali e calcolati gli Z-score per ciascuna. Gli Z-score negativi sono stati trasformati in zero, e successivamente sono stati sommati per ottenere uno "score di rischio ambientale". I comuni con i punteggi più alti sono quelli maggiormente esposti ai rischi. Infine, è stata creata una heatmap per visualizzare i Z-score delle variabili per i comuni a maggior rischio.

4.5 Diagnostica del Modello

Per ogni domanda di ricerca sono state effettuate analisi diagnostiche per verificare le assunzioni del modello, tra cui:

- Calcolo del VIF
- Distribuzione dei residui
- Confronto tra predizioni e valori reali
- Analisi dei residui vs valori predetti
- Scale-Location Plot
- Identificazione di Outliers (distanza di Cook) e Punti di Leverage
- Verifica della Normalità dei Residui (Q-Q plot e il test di Shapiro-Wilk)
- Tracciamento della curva ROC con calcolo del punteggio AUC (solo per 4.3)