UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Partecipazione elettorale in Italia:
un'analisi ecologica a livello comunale

Relatore
Prof. Livio Finos
Dipartimento di Scienze Statistiche

Correlatore
Prof.ssa Venera Tomaselli
Università di Catania

Laureando
Nicola Menegon
Matricola 2067184

ANNO ACCADEMICO 2024/25

Indice

1 Introduzione **1**

1.1 Il fenomeno 1

1.2 Il dataset 3

2 Analisi esplorative **5**

2.1 Mappe di affluenza 6

2.2 Correlazioni 6

2.3 PCA 9

3 Modelli per la selezione delle variabili **13**

3.1 Modelli lineari 13

3.1.1 Best subset selection 14

3.1.2 Selezione stepwise 16

3.2 LASSO 17

3.3 Alberi di regressione 22

3.4 Confronto e verifica delle assunzioni 24

3.5 Modelli per dati spazialmente correlati 27

4 La fallacia ecologica **33**

5 Conclusioni **38**

Bibilografia **39**

Appendice A **41**

Ringraziamenti **42**

1 Introduzione

La questione della bassa affluenza elettorale è tornata al centro del dibattito pubblico a seguito delle elezioni europee dell'8-9 giugno 2024, in occasione delle quali meno di metà degli aventi diritto si è recata alle urne. Ciò ha di fatto sancito la prima occasione nella storia della Repubblica in cui la maggioranza degli italiani ha scelto di non scegliere. In questa analisi non ci soffermeremo particolarmente sul trend decrescente osservato negli ultimi due decenni, piuttosto cercheremo di capire, in media, quali elementi favoriscano o meno l'affluenza all'interno dei singoli comuni. Per farlo, utilizzeremo un dataset raccolto e messo a disposizione dalla prof.ssa Tomaselli, che ringrazio. Dopo una breve discussione del fenomeno, cercheremo di evidenziare alcune interessanti relazioni tra variabili, per poi costruire un modello statistico che possa fornire qualche utile spunto interpretativo.

1.1 Il fenomeno

L'interpretazione delle motivazioni che spingono gli elettori a scegliere di non votare è complessa. L'astensione può essere causata da fattori molto diversi gli uni dagli altri, che tuttavia risultano indistinguibili nei dati elettorali aggregati. Cause del tutto eterogenee tra loro, infatti, sfociano nello stesso risultato osservabile. In letteratura il tema è ampiamente affrontato, sia in Italia che all'estero, dove si rivelano trend differenti da paese a paese. Benché ogni elezione sia unica nelle modalità e nel contesto in cui si svolge, è bene notare che alcuni effetti sociali perdurano tra elezioni di tipo e periodo differenti (Leighley, Nagler 1992).

Nell'interpretare il fenomeno, è bene distinguere perlomeno tra astensionismo volontario e involontario. Come evidenziato in Ferrara et al. 2023, quest'ultimo tipo si rivela in crescita negli ultimi anni e riguarda all'incirca il 4% degli aventi diritto, con maggiore prevalenza nelle province del Mezzogiorno. Si tratta di anziani con ridotte capacità motorie, studenti e lavoratori fuori sede impossibilitati a rientrare, o semplicemente persone in viaggio o vacanza lontane dal loro comune di residenza (in quest'ultimo caso particolarmente per elezioni minori o convocate prima della scadenza naturale della legislatura). D'altro canto, gli autori notano che una parte maggioritaria dell'astensionismo

è dovuta a cause politiche di diversa natura. Anche qui, gli elettori possono essere mossi da molteplici motivazioni, come disinteresse generale, assenza di rappresentanza partitica o astensione di protesta. Spesso è interessante osservare anche il numero di schede bianche o nulle, che possono essere indice di un voto di dissenso. Chiaramente, non è possibile fare inferenza diretta sulle cause a partire dai risultati cumulati.

Per quanto riguarda gli elementi socio-economici, è ampiamente accettato che disuguaglianze sociali e povertà incidano negativamente sull'affluenza. Bloise et al. 2021 evidenziano in particolare che l'aumento del non voto è associato a una maggiore disuguaglianza e alla precarietà. Ciò non ha effetto solo sull'affluenza ma anche sui risultati elettorali. Infatti, i partiti "tradizionali" (centro-destra e centro-sinistra) mantengono il consenso nelle regioni con maggiore ricchezza e perdono voti dove la disuguaglianza e la precarietà aumentano. Al contrario, risultati migliori per la Lega correlano con classi medie più povere, mentre per il Movimento 5 Stelle questo è più correlato a situazioni di povertà assoluta.

Diverse altre variabili entrano in gioco. Ad esempio, Harka, Rocco 2022 notano che una maggiore istruzione porta ad un effetto negativo sull'affluenza. Questo risultato è in contrasto con i trend rilevati nel resto d'Europa e può essere in parte spiegato da un aumento dell'astensione di protesta nella popolazione più istruita, ipotesi che trova riscontro nel contemporaneo aumento della proporzione di schede bianche. Gli autori notano inoltre che questa tendenza è più debole se si analizzano i dati ottenuti da sondaggi rispetto a quella rilevata dai dati aggregati dei risultati elettorali. Questo potrebbe essere dovuto ad un *reporting bias*, ovvero al fatto che i soggetti più istruiti siano più propensi a riportare la loro propensione al voto. Questo aspetto rappresenta un argomento a favore dell'utilizzo di dati ecologici, come sarà discusso in seguito.

Al netto delle questioni politiche di per sé, è importante notare che i diversi elementi socio-demografici e le differenti condizioni economiche non hanno effetto solo sulla probabilità con cui un elettore si rechi alle urne, ma si riflettono anche sui risultati finali. Il tema, quindi, si presta ad analisi ed interpretazioni ben più approfondite, che richiedono una visione d'insieme del fenomeno e non si possono esaurire in una breve discussione o in un determinato modello statistico.

1.2 Il dataset

Il dataset a disposizione contiene informazioni su 7896 comuni italiani, ovvero la totalità al 2024. Poiché il nome del comune non è necessariamente unico, questi sono univocamente identificati dal loro codice ISTAT. Ciò inoltre semplifica l'utilizzo di dataset complementari, anche in altre lingue, eliminando eventuali problemi di traduzione dei nomi.

Per ogni osservazione sono rilevate 307 variabili raccolte in blocco da diverse indagini ISTAT, di cui una decina sono di descrizione geografica del comune, e ulteriori 91 su aventi diritto al voto, affluenza, schede bianche e nulle nelle ultime tre elezioni politiche ed europee. Da queste, sono stati costruiti circa 100 ulteriori indicatori. Molte delle variabili riportate sono rilevazioni annuali di una stessa variabile, dunque abbiamo a disposizione diverse serie storiche rilevate a partire dal 2014 e fino al 2021. Riguardo a queste, alcune considerazioni sono necessarie. In primo luogo, va notato che non tutte le variabili sono state rilevate nello stesso periodo temporale e, per questo motivo, risulta complessa l'analisi della loro correlazione con l'affluenza nel tempo. Inoltre, l'analisi della serie storica in sé non è di nostro diretto interesse. Piuttosto, siamo interessati a valutare l'influenza delle caratteristiche strutturali di ogni comune sull'affluenza media. L'utilizzo della media dei valori rilevati nel periodo 2014-2021 permette di limitare l'impatto di possibili fluttuazioni annuali. Pertanto, nel seguito utilizzeremo la media dei valori rilevati.

Poiché i dati sono raccolti in modo aggregato comune per comune, l'analisi che ci apprestiamo a svolgere è di tipo *ecologico*, ovvero non farà riferimento ai comportamenti individuali dei soggetti, ma al modo in cui l'ambiente in cui questi vivono ne influenzi la probabilità di recarsi alle urne.

Una prima scrematura delle variabili è stata fatta a mano, eliminandone alcune di difficile interpretazione. Quelle conservate sono di diversi tipi:

- *Descrizione del comune:* codice ISTAT, nome, provincia e regione, classificazione Strategia Nazionale Aree Interne (da "polo" a "ultra periferico")

- *Caratteristiche anagrafiche:* popolazione totale, percentuale di popolazione nelle fasce d'età 19-29, 30-44, 75+ anni, divisi per sesso, percentuale di maschi e femmine sul totale

- *Dati ISTAT:* variabili socio-economiche e demografiche, come percentuale di studenti, indice di fragilità, tasso di occupazione, natalità, mortalità, immigrazione ecc.

- *Affluenza elettorale:* affluenza alle elezioni Europee (media 2014/2019/2024), affluenza alle elezioni politiche (media 2013/2018/2022), divise per sesso.

Per l'elenco completo delle variabili si veda l'appendice **??**. Il dataset non è privo di dati mancanti. Questi possono essere dovuti a diversi fattori, tra cui il fatto che alcuni comuni, negli anni, si fondono o dividono[1]. I comuni della Valle d'Aosta non riportano dati sull'affluenza per le elezioni politiche, mentre hanno dati completi per le elezioni europee. Ciò è dovuto al modo in cui il Ministero dell'Interno raccoglie e centralizza i dati. Aggiungere manualmente questi dati sarebbe teoricamente possibile grazie alla piattaforma Eligendo, tuttavia questi non riportano il codice ISTAT, dunque l'unione dei due dataset andrebbe effettuata utilizzando come chiave il nome del comune; questa opzione si rivela poco praticabile a causa dell'eterogeneità delle denominazioni, dovuta anche al bilinguismo della regione. Inclusi questi, 234 comuni riportano almeno un dato mancante tra le variabili selezionate. Questi comuni sono di piccole dimensioni e distribuiti in modo sufficientemente omogeneo tra le regioni geografiche, possiamo quindi escluderli dalle nostre analisi al costo di una piccola perdita di informazione. Dunque, dove non sia diversamente specificato, il numero di osservazioni considerato è $n = 7662$. Il codice R utilizzato per tutte le analisi è disponibile a `https://github.com/Menegon28/`.

[1]Si pensi che nel 2013 il numero di comuni era 8090, dunque è sceso di quasi 200 unità in un decennio.

2 Analisi esplorative

Il dataset si presta a diverse analisi introduttive piuttosto semplici. È opportuno innanzitutto valutare la distribuzione marginale della variabile risposta. L'unità statistica è il singolo comune.

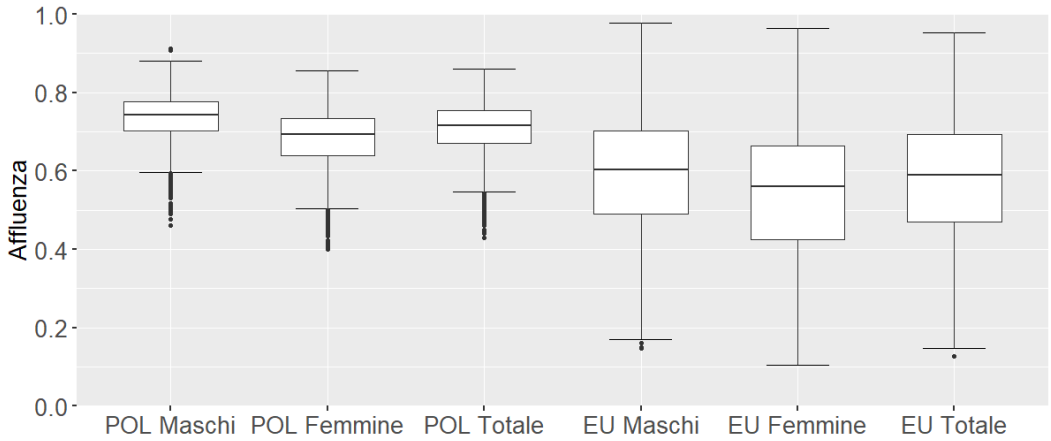

Figura 1: Boxplot delle distribuzioni dell'affluenza media di elezioni politiche ed europee, per sesso

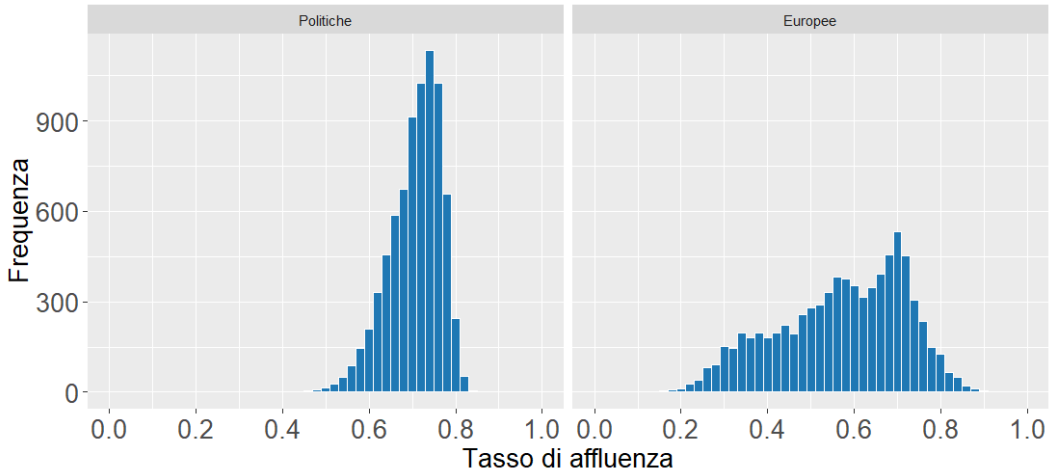

Figura 2: Distribuzioni dell'affluenza media di elezioni politiche ed europee

Marginalmente, si nota che la variabilità dell'affluenza a livello di comune è decisamente maggiore nelle elezioni europee rispetto alle politiche. Difatti, la varianza dell'affluenza a livello di comune alle europee risulta circa 6 volte quella delle politiche. Inoltre, almeno a livello descrittivo, l'affluenza sembra essere maggiore tra i maschi che tra le femmine, in entrambe le consultazioni. Ciò è confermato anche dalle frequenze marginali rapportate alla popolazione del comune.

2.1 Mappe di affluenza

Vediamo alcune mappe sull'affluenza alle elezioni politiche e europee. Questi grafici sono realizzati con la libreria Vega-Altair di Python, utilizzando un file topojson[2].

Figura 3: Affluenza per comune alle elezioni politiche - media 2013, 2018, 2022

Dalla Figura 3 e Figura 4 emerge che nelle elezioni politiche l'affluenza appare più elevata ed uniforme su tutto il territorio nazionale, mentre le europee vedono una notevole disomogeneità. In entrambi i casi, si evidenzia una minore partecipazione nelle regioni del sud e nelle isole, con un crollo particolare per le elezioni europee.

La Figura 5 evidenzia che in quasi tutti i comuni, esclusi alcuni nella Provincia Autonoma di Bolzano, l'affluenza degli elettori di sesso maschile è risultata in media maggiore di quella degli elettori di sesso femminile.

2.2 Correlazioni

Un primo aspetto interessante è valutare le correlazioni tra le variabili rilevate nei diversi comuni. Si noti che questo procedimento assegna lo stesso peso ad ogni comune, senza tener conto della popolazione complessiva.

[2]Disponibile liberamente a `https://github.com/openpolis/geojson-italy`

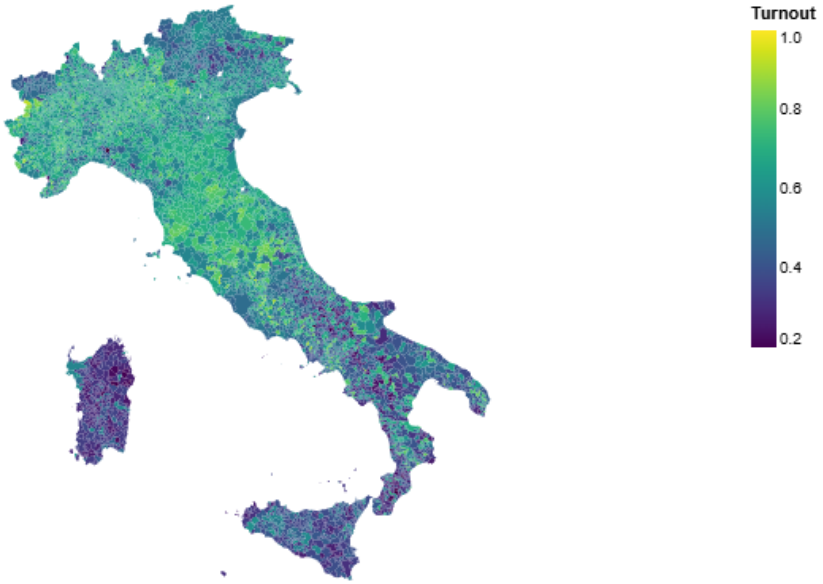

Figura 4: Affluenza per comune alle elezioni europee - media 2014, 2019, 2024

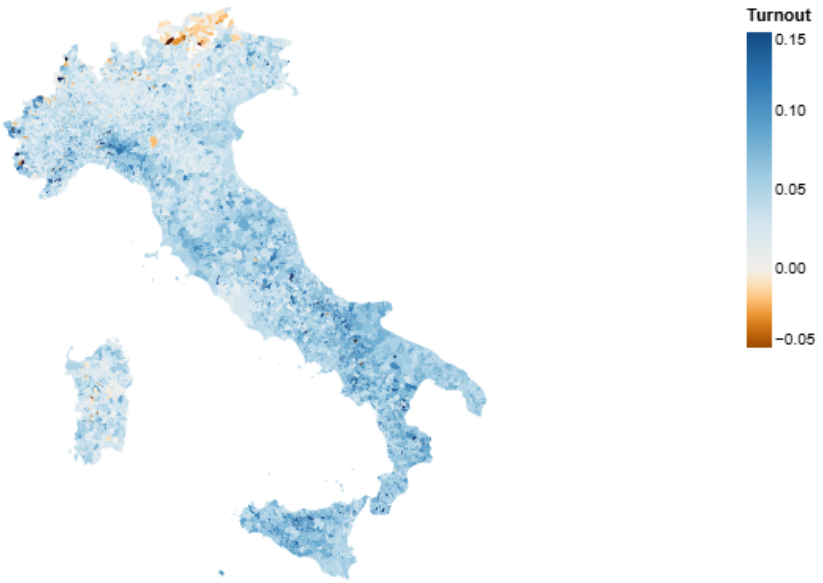

Figura 5: Differenza dell'affluenza per sesso - media politiche 2013, 2018, 2022. Valori positivi (verso il blu) indicano maggiore affluenza degli elettori di sesso maschile, valori negativi (verso l'arancione) viceversa

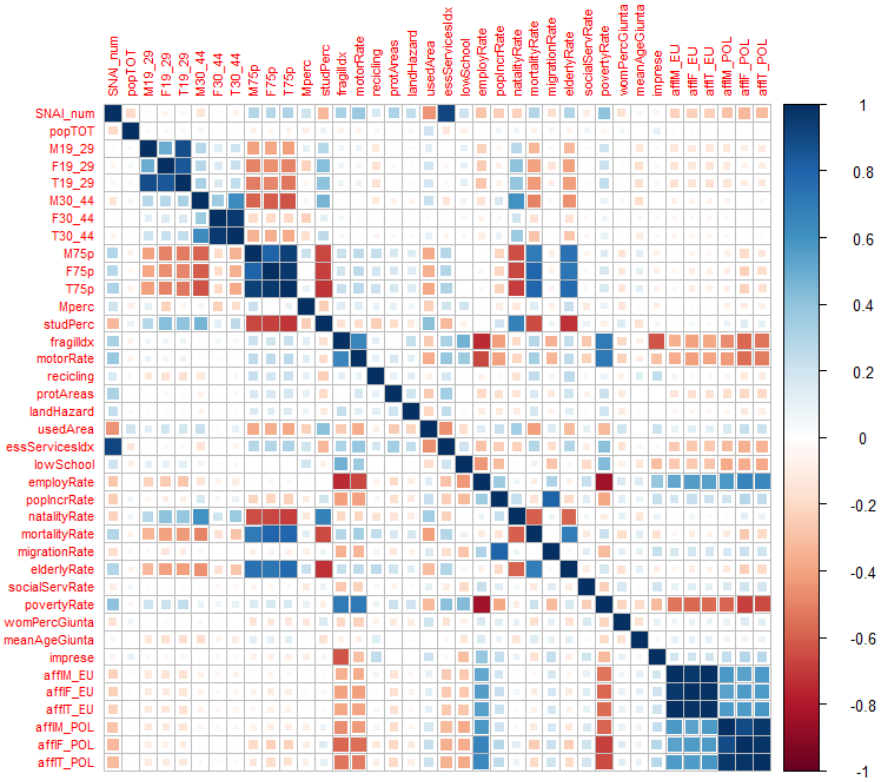

Figura 6: Visualizzazione grafica delle correlazioni tra variabili

In Figura 6 osserviamo già qualche dato interessante. Tra le correlazioni tra variabili non elettorali, notiamo, ad esempio, che l'indice di fragilità cresce con la proporzione di popolazione con bassi livelli di istruzione e con l'indice di povertà, mentre è correlato negativamente con il tasso di occupazione. Il tasso di motorizzazione appare fortemente correlato positivamente con il tasso di povertà e negativamente con il tasso di occupazione. È chiaro che sarebbe sbagliato inferire da ciò che chi lavora tenda a possedere meno automobili. Piuttosto, una lettura possibile è quella per cui i comuni più periferici tendano ad avere tassi di occupazione più bassi e contemporaneamente tassi di motorizzazione più alti, mentre il contrario è in genere vero per le grandi città. Ciò trova riscontro nelle correlazioni con le classificazioni *SNAI*. Da questo esempio possiamo trarre un'importante valutazione sul modo in cui è opportuno leggere questi dati. Come vedremo approfonditamente nel Capitolo 4, inferire sui comportamenti individuali a partire da dati raggruppati è un errore. Per questo motivo i dati qui presentati vanno interpretati solo a livello di comune.

Per quanto riguarda l'affluenza, le variabili con cui questa appare essere

correlata più positivamente sono il tasso di incremento della popolazione e il tasso di occupazione, mentre negativamente con indice di fragilità e tasso di povertà. Notiamo altresì che affluenza alle elezioni europee e politiche hanno una correlazione pari a 0.58.

2.3 PCA

La *Principal Component Analysis* è una tecnica di riduzione della dimensionalità che trasforma la matrice dei dati in componenti ortogonali, cercando di spiegare la maggior parte della variabilità possibile con un piccolo numero di dimensioni.

Sia X vettore aleatorio p-dimensionale, assumiamo

$$X \sim \mathrm{f}_p(\mu, \Sigma),$$

dove $\mathrm{f}_p(\cdot)$ è una qualsiasi distribuzione p-variata, con

- $\mu = E(X)$ il vettore delle medie,
- $\Sigma = \mathrm{Var}(X)$ la matrice di varianze e covarianze,
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ gli autovalori di Σ,
- $e_1, e_2, \ldots, e_p$ i corrispondenti autovettori di Σ.

Dati $a_1, a_2, \ldots, a_p$ vettori reali, si considerino le p combinazioni lineari

$$\begin{aligned}
Y_1 &= a_1^T X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= a_2^T X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= a_p^T X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}$$

Le componenti principali sono quelle $Y_1, \ldots, Y_p$ che sono tra loro incorrelate e hanno varianza massima. Il problema dunque è trovare $a_i, i = 1, \ldots, p$, tali da massimizzare $\mathrm{Var}(a_i^T X)$ sotto il vincolo di incorrelazione $\mathrm{Cov}(a_i^T X, a_j^T X) = 0$ per $j < i$ e quello di identificabilità $a_i^T a_i = 1$.

Si dimostra che ciò succede in corrispondenza di $a_i = e_i$ e dunque

$$Y_i = e_i^T X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1, \ldots, p$$

Si ha, inoltre, che

$$\mathrm{Var}(Y_i) = e_i^T \Sigma e_i = \lambda_i, \quad i = 1, 2, \ldots, p,$$

$$\mathrm{Cov}(Y_i, Y_k) = e_i^T \Sigma e_k = 0, \quad i \neq k.$$

Pertanto la proporzione di varianza spiegata dall'i-esima componente principale è $\lambda_i / tr(\Sigma)$.

Per quanto riguarda il corrispettivo campionario, sia X la matrice dei dati, con matrice di covarianze stimata S, siano

$$\left(\hat{\lambda}_1, \hat{e}_1\right), \left(\hat{\lambda}_2, \hat{e}_2\right), \ldots, \left(\hat{\lambda}_p, \hat{e}_p\right)$$

le coppie degli autovalori-autovettori di S, dove

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0.$$

Allora, la i-esima componente principale è

$$\hat{y}_i = X\hat{e}_i = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \quad i = 1, 2, \ldots, p,$$

dove x_j è la j-esima colonna di X. Le proprietà viste in precedenza valgono per i corrispettivi campionari. (R. Johnson, Wichern 2013, cap. 8)

Applicazione

Le variabili hanno scale diverse tra loro, dunque la standardizzazione della matrice dei dati è necessaria. Utilizzando `prcomp` otteniamo

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.6146	2.0372	1.33216	1.23932	1.13710	1.08556
Proportion of Variance	0.2442	0.1482	0.06338	0.05485	0.04618	0.04209
Cumulative Proportion	0.2442	0.3924	0.45575	0.51060	0.55678	0.59887

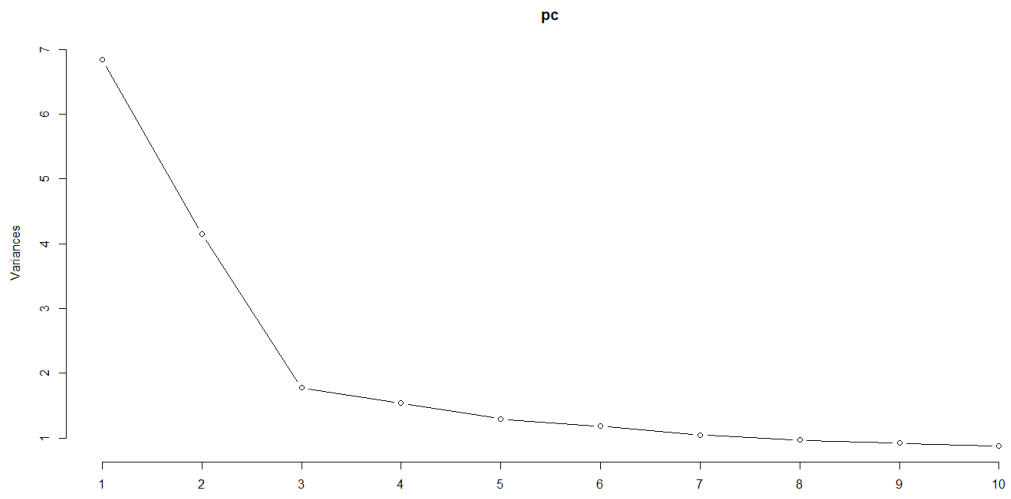

Figura 7: Screeplot

Lo screeplot visualizza graficamente la quantità di varianza spiegata da ogni componente. Le prime due componenti principali spiegano già quasi il 40% della varianza dei dati, dunque uno scatterplot bidimensionale può essere utilizzato per fornire una visualizzazione intuitiva.

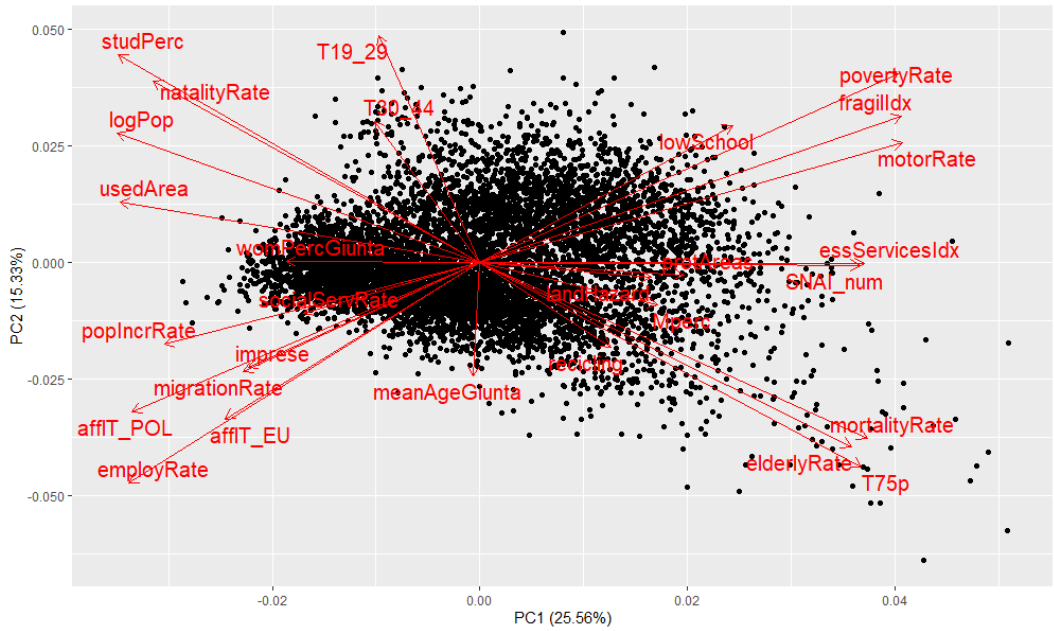

Figura 8: Biplot delle prime due componenti principali

Dal biplot in Figura 8 emergono nuovamente relazioni tra le variabili già viste in precedenza. La prima componente principale sembra rappresentare una valutazione della grandezza del comune. Infatti, valori alti di PC1 indicano

valori della classificazione SNAI più alti (ovvero comuni più piccoli e periferici), così come valori maggiori per tasso di fragilità e di mortalità. Valori bassi di PC1, al contrario, indicano popolazione e suo tasso di crescita maggiori, così come maggiore affluenza media. La seconda componente principale non è di facile interpretazione, ma sembrerebbe essere correlata con la distribuzione delle fasce d'età, con valori alti ad indicare una popolazione più giovane. Da questa analisi si evince un effetto positivo sull'affluenza del tasso di occupazione, di migrazione e di aumento della popolazione, oltre a un effetto negativo del tasso di povertà e dell'indice di povertà.

3 Modelli per la selezione delle variabili

Siamo interessati a modellare l'andamento dell'affluenza nei comuni rispetto alle variabili socio-demografiche rilevate, che considereremo come regressori. In particolare, si vorrebbero selezionare quelle variabili che mostrano un effetto rilevante sulla risposta e quantificarne direzione e intensità. Come si è visto, elezioni politiche ed europee hanno trend di voto chiaramente differenti. Per questo motivo, ci limiteremo all'analisi delle elezioni politiche, per brevità.

3.1 Modelli lineari

Una prima strada percorribile è quella della costruzione di un modello lineare normale, definito come

$$Y = X\beta + \varepsilon, \qquad \text{con } \varepsilon \sim N_n(0, \sigma^2 I)$$

ignorando per ora eventuali correlazioni tra comuni spazialmente vicini. Il modello completo include 26 variabili più la risposta. Alcune questioni vanno dibattute prima dell'adattamento del modello.

In primis, va valutato se l'utilizzo di una procedura di stima OLS sia corretto o se sia più appropriato assegnare alle osservazioni un peso proporzionale alla popolazione del comune. In generale, utilizzando dati ecologici non vi è una risposta valida in assoluto. Per dati raccolti da un campionamento stratificato, ad esempio, l'utilizzo di stime WLS è preferibile per ridurre il bias degli stimatori (Korn, Graubard 1995), in particolare se si intendono trarre conclusioni a livello individuale e non aggregato. Tuttavia, se l'elemento di interesse sono le relazioni a livello ecologico, come nel nostro caso, Frohlich et al. 2001 rilevano che non è necessario pesare per la popolazione, in particolare quando le aggregazioni hanno un effettivo significato. Quest'ultimo sarà dunque l'approccio adottato per questa analisi.

In secondo luogo, è bene chiarire quale sia il fine principale del modello: in questo caso, l'interpretazione ha certamente la priorità sulle capacità predittive. Non è di nostro particolare interesse, infatti, prevedere l'affluenza di un comune immaginario fissate delle specifiche caratteristiche socio-demografiche,

nonostante questo rimanga comunque un utilizzo possibile[3]. Piuttosto, il nostro scopo è quello di comprendere quali caratteristiche dei comuni influenzano l'affluenza finale.

3.1.1 Best subset selection

Se non ammettiamo interazioni tra le esplicative, il numero di regressori è sufficientemente piccolo da permettere di valutare ogni possibile sottogruppo. Per scegliere il modello migliore, applichiamo una penalizzazione secondo il numero di parametri da stimare. Per garantire un modello parsimonioso, appare opportuno utilizzare il Criterio di informazione Bayesiano, definito come

$$BIC = k\ln(n) - 2\ln(\hat{L})$$

dove $\hat{L}$ indica il valore massimizzato della funzione di verosimiglianza per il modello stimato e k il numero di parametri stimati, che nel caso di un modello lineare corrisponde al numero p di variabili nel modello, più intercetta e varianza degli errori.

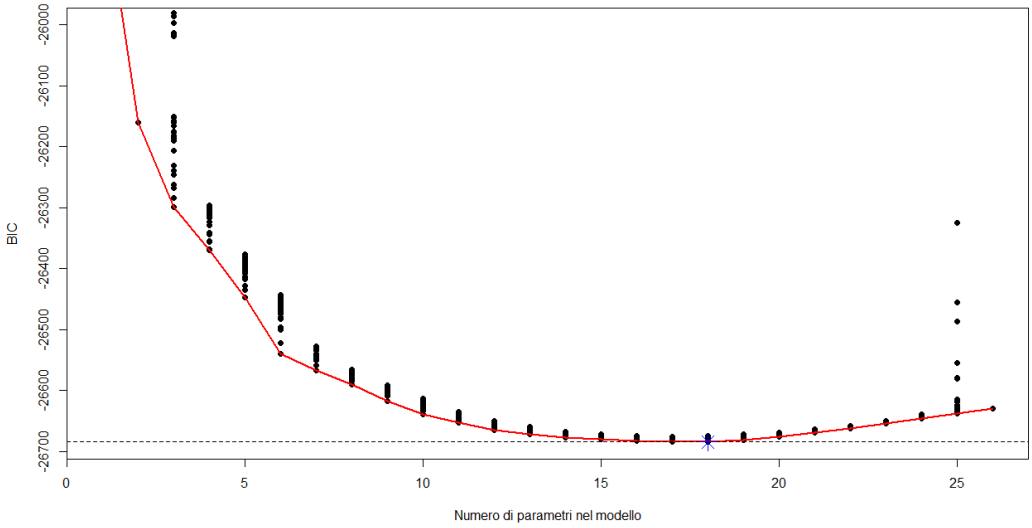

Figura 9: Andamento del BIC dei 30 migliori modelli per ogni numero di variabili

[3]Si noti, però, che questa formulazione del modello non costringe i valori predetti nell'intervallo [0,1] e sarebbe dunque possibile, per qualche fantasiosa ma lecita combinazione di valori dei regressori, ottenere una previsione per l'affluenza negativa o superiore al 100%.

Variabile	Stima	Std. Error	p-value
(Intercept)	0.84782	0.02127	6.07e-316
povertyRate	-0.00188	0.00010	3.15e-73
employRate	0.00165	0.00012	3.15e-42
logPop	-0.00686	0.00054	7.41e-37
lowSchool	-0.00067	0.00007	6.59e-21
natalityRate	-0.00306	0.00040	2.03e-14
essServicesIdx	-0.00052	0.00007	5.02e-14
T75p	-0.15958	0.02464	9.95e-11
popIncrRate	-0.00006	0.00001	4.01e-07
meanAgeGiunta	-0.00043	0.00009	5.50e-07
protAreas	-0.00011	0.00002	1.13e-06
landHazard	0.00021	0.00004	1.62e-06
studPerc	0.12449	0.02786	7.98e-06
Mperc	-0.11206	0.02997	1.86e-04
elderlyRate	-0.00002	0.00001	2.11e-04
fragilIdx	-0.00137	0.00038	3.37e-04
imprese	0.00012	0.00003	3.63e-04
SNAI_num	0.00392	0.00122	1.27e-03
usedArea	0.00020	0.00006	2.11e-03

Tabella 1: BSS: coefficienti di regressione per le variabili selezionate

Si noti che, fissato il numero di parametri inclusi, tutti i modelli considerati hanno uguale penalizzazione BIC, dunque la selezione può avvenire anche in base a R^2. Pertanto, non si verifica nessun trade-off tra capacità previsive e complessità e il modello migliore è identificato univocamente (Miller 2002). La scelta finale del modello richiede però un confronto tra i modelli ottimali per ciascun numero di parametri. Tale valutazione deve necessariamente basarsi su un criterio che ne penalizzi la complessità: la scelta tra AIC, BIC o altri criteri dipende, in ultima analisi, da un compromesso tra accuratezza previsiva e parsimonia.

I valori stimati dei parametri sono riportati in Tabella 1. Risulta $R^2 = 0.5091$. Il segno dei coefficienti $\hat{\beta}$ indica se un aumento unitario della variabile corrispondente porta a un aumento o una diminuzione media dell'affluenza. Si noti che i p-values riportati sono distorti dalla procedura di Best Subset Selection. Infatti, gli stessi dati sono in questo caso utilizzati sia per la selezione del modello che per l'inferenza, il che viola l'assunto che il modello sia stato fissato *a priori*. Per questo motivo, il rischio commettere un errore di I tipo è aumentato rispetto a quello nominale, per via del fatto che la procedura seleziona automaticamente i regressori più correlati con la risposta. Tra i risultati più interessanti emerge l'effetto positivo del tasso di occupazione e gli effetti negativi di tasso di povertà, popolazione, percentuale di over 75, percentuale di residenti a bassa scolarizzazione e tasso di natalità.

3.1.2 Selezione stepwise

Se si ammettono interazioni a due, il processo di ricerca esaustivo diventa computazionalmente troppo complesso da eseguire. Infatti, il numero di subset da valutare è $2^{p+\binom{p}{2}}$, con p numero di variabili nel modello completo.

Pertanto, è necessario ricorrere a procedure iterative: una via percorribile è quella della selezione stepwise. Nella formulazione classica di questo approccio, ad ogni passo si valuta l'inclusione di una nuova variabile esplicativa e contestualmente si eliminano quelle variabili già presenti il cui contributo sia diventato non significativo, ad un livello prefissato, dopo il nuovo inserimento. Si raggiunge il modello finale quando nessuna delle variabili escluse nel modello supera il test di ingresso e nessuna delle variabili incluse risulta dare un contributo non significativo. (Salvan et al. 2020, par 2.5) Tuttavia, l'uti-

lizzo del p-value come criterio di entrata e uscita delle variabili è ampiamente considerato errato, in quanto porta a inflazione di falsi positivi e overfitting del modello (Kuhn, K. Johnson 2019). Vi si preferisce, ancora una volta, un criterio di informazione come BIC.

Con questo procedimento, otteniamo quanto riportato in Tabella 2.

Si ha $R^2 = 0.5585$, leggermente maggiore del modello precedente. Anche in questo caso i p-values vanno interpretati con cautela. I coefficienti di interazione possono essere di interesse nel valutare gli effetti congiunti delle variabili. Tuttavia, questo modello risulta di difficile interpretazione vista la grande quantità di parametri ancora presenti nel modello selezionato. Ciò può far preferire il modello più semplice che, seppur meno flessibile, restituisce risultati più comprensibili.

3.2 LASSO

I metodi di selezione delle variabili visti fin'ora si basano su un confronto tra diversi modelli effettuato tramite un qualche criterio. Un metodo di selezione (e regolarizzazione) alternativo è la regressione *lasso*. Questa procedura aggiunge una penalità per la dimensione dei parametri del modello direttamente nel processo di stima. In particolare, questo metodo utilizza una penalità l_1[4]. In questo modo, si deve adattare un solo modello. La stima dei parametri è la soluzione di

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{soggetto a} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

Nella sua forma lagrangiana, ciò equivale a

$$\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

ovvero, si minimizza una quantità data dalla somma dei quadrati degli scarti ordinaria e una penalità dipendente dalla magnitudo dei β e da un parametro

[4] A differenza della regressione Ridge, ciò costringe le stime dei parametri delle variabili meno importanti ad essere esattamente uguali a 0, in modo da effettuarne una selezione.

Variabile	Stima	Std. Error	p-value
(Intercept)	0.9429970	0.1013983	1.80e-20
povertyRate	0.0027917	0.0004113	1.23e-11
employRate	0.0003434	0.0008245	6.77e-01
essServicesIdx	0.0040645	0.0004898	1.24e-16
lowSchool	-0.0063900	0.0004324	9.49e-49
logPop	-0.0002788	0.0120884	9.82e-01
meanAgeGiunta	0.0006759	0.0003263	3.84e-02
imprese	-0.0007023	0.0001463	1.62e-06
mortalityRate	0.0113276	0.0014961	4.13e-14
natalityRate	0.0475289	0.0051786	5.57e-20
studPerc	-3.6452064	0.4795428	3.28e-14
T75p	-1.6962450	0.1774516	1.57e-21
Mperc	-0.0881956	0.1308228	5.00e-01
SNAI_num	0.0128751	0.0047716	6.99e-03
motorRate	-0.0022578	0.0005906	1.33e-04
popIncrRate	-0.0000753	0.0000130	7.44e-09
protAreas	-0.0002448	0.0000590	3.37e-05
employRate:lowSchool	0.0000525	0.0000069	4.15e-14
employRate:essServicesIdx	-0.0000358	0.0000052	7.41e-12
povertyRate:essServicesIdx	-0.0000260	0.0000055	2.54e-06
employRate:logPop	0.0004949	0.0000558	8.96e-19
lowSchool:natalityRate	0.0002781	0.0000317	2.34e-18
imprese:studPerc	0.0038386	0.0006410	2.22e-09
essServicesIdx:natalityRate	-0.0000777	0.0000154	4.89e-07
employRate:T75p	0.0192188	0.0026967	1.12e-12
essServicesIdx:SNAI_num	-0.0001625	0.0000264	8.08e-10
employRate:mortalityRate	-0.0001818	0.0000227	1.25e-15
meanAgeGiunta:SNAI_num	-0.0002718	0.0000830	1.06e-03
povertyRate:motorRate	-0.0000260	0.0000064	4.68e-05
employRate:natalityRate	-0.0004914	0.0000508	4.93e-22
povertyRate:natalityRate	-0.0004113	0.0000405	4.50e-24
popIncrRate:protAreas	0.0000016	0.0000004	1.52e-05
motorRate:protAreas	0.0000080	0.0000022	2.12e-04
logPop:natalityRate	-0.0018069	0.0002533	1.06e-12
logPop:studPerc	0.1140523	0.0182008	3.90e-10
studPerc:T75p	1.8849231	0.3012816	4.15e-10
studPerc:Mperc	4.9121687	0.8434229	5.98e-09
logPop:Mperc	-0.0965068	0.0231393	3.07e-05
logPop:motorRate	0.0002250	0.0000551	4.49e-05
SNAI_num:motorRate	0.0002488	0.0000782	1.47e-03
lowSchool:imprese	0.0000092	0.0000029	1.30e-03

Tabella 2: Stepwise: coefficienti di regressione delle variabili selezionate

di tuning λ. In pratica, la funzione da minizzare è

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

I parametri λ e s hanno una corrispondenza biunivoca e sono inversamente proporzionali. (James et al. 2021)

Questo metodo è sensibile alla scelta di λ: se $\lambda = 0$ si ottengono le usuali stime OLS, mentre se $\lambda \to \infty$ tutti i $\hat{\beta}$ tendono a 0. Si noti dunque che ogni valore di λ produce un diverso set di coefficienti $\hat{\beta}_\lambda$. Un approccio comunemente utilizzato per la scelta di un valore consono è quello della cross-validation, che prevede di dividere il dataset in k parti di simile numerosità, allenare il modello su $k-1$ di queste e testarlo nella rimanente parte, ad esempio calcolandone il Mean Square Error. Si ripete il processo k volte e il valore stimato risulta dalla media dei k valori di MSE calcolati. Per la scelta di λ ciò viene ripetuto per diversi valori del parametro. Si può scegliere il valore di λ per cui il MSE è minimo oppure quello corrispondente all'estremo superiore dell'intervallo di confidenza dell'MSE, se si vuole ottenere un modello più parsimonioso.

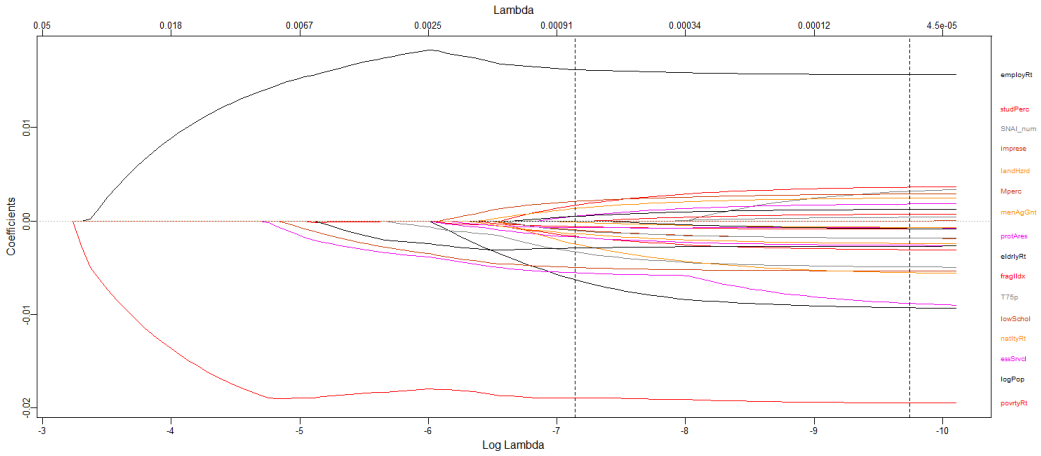

Figura 10: Stima dei parametri di regressione al variare del valore di λ

Applicando questo metodo ai nostri dati, otteniamo quanto riportato in Figura 10. Come si può vedere, all'aumentare del valore assegnato a λ un numero crescente di parametri viene costretto al valore 0. Più il valore di λ per cui questo succede è alto più la variabile è da ritenersi rilevante ai fini del modello.

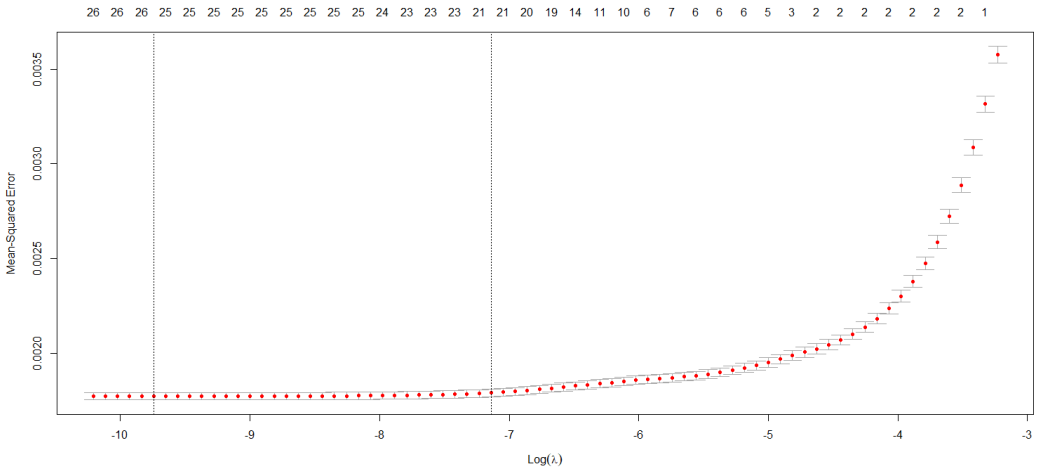

Figura 11: Cross-validation per la scelta del valore di λ

Il valore per cui il MSE risulta minimo è evidenziato in Figura 12 dalla linea tratteggiata di sinistra, mentre il più grande λ per cui non si rifiuta il test di uguaglianza degli MSE è identificato dalla linea di destra. Scegliere un valore più grande porta a un modello più parsimonioso, come si vede dal numero di variabili ancora incluse nel modello, riportato in alto. Si noti che i risultati qui ottenuti sono aleatori in quanto dipendenti dalla suddivisione nei k folds (e dal valore di k, scelto arbitrariamente). Nel nostro caso, tuttavia, i risultati qui mostrati sono robusti rispetto a variazioni del valore di k e a diverse suddivisioni delle osservazioni nelle parti.

La scelta tra i due valori possibili può basarsi su un approccio training/test. Si adatta il modello e si effettua la cross-validation su circa il 75% delle osservazioni e lo si testa sulle rimanenti, secondo una divisione casuale. Il procedimento viene ripetuto diverse volte per ottenere una stima robusta del criterio di test scelto. Nel nostro caso, sia RMSE che MAE indicano una capacità predittiva leggermente migliore per il valore di λ più piccolo. Tuttavia, tale differenza è di piccola entità, come si vede in Figura 12. Per questo motivo il valore ritenuto preferibile è quello maggiore, in quanto porta a una semplificazione notevole del modello migliorandone così l'interpretabilità ed evitando overfitting.

Le stime dei parametri ottenute per il valore scelto di λ sono riportati in Tabella 3.

Si noti che il modello *lasso* è anche un metodo di regolarizzazione, pertanto

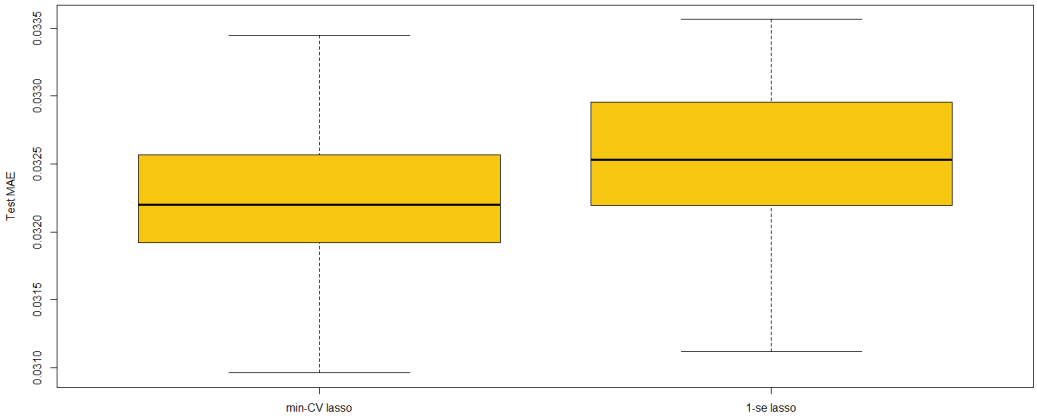

Figura 12: Capacità predittive del modello scegliendo il λ che minimizza il MSE di cross-validation (sinistra) o il più alto ancora compreso nell'intervallo di confidenza (destra)

i parametri risultano direttamente confrontabili, per cui valori più grandi in valore assoluto indicano maggiore effetto della corrispondente variabile. Infatti, la matrice delle covariate dev'essere precedentemente standardizzata, in modo che la penalità risulti omogenea per tutte le variabili. Inoltre, questo modello è adatto anche in caso di multicollinearità: in presenza di variabili molto correlate il modello lasso tende a selezionarne solo una e a costringere a zero i coefficienti delle altre[5]. Tuttavia, va notato che la scelta di quale mantenere è instabile e può dipendere da piccole variazioni nel dataset (Chan et al. 2022). Per questo motivo, è d'obbligo ricordare che questo è solo uno dei modelli che si sarebbero potuti ottenere e che l'inclusione di una variabile non esclude l'effetto di altre che vi siano correlate.

Diverse stime sono esattamente pari a zero, pertanto le corrispondenti variabili sono escluse dal modello. Il segno della stima indica se all'aumentare della variabile l'affluenza, in media, aumenta o diminuisce. Dalla Figura 10 e dalla relativa Tabella 3 emerge chiaramente l'effetto positivo del tasso di occupazione e quello negativo del tasso di povertà, entrambi con magnitudo ben superiore a quello di tutte le altre variabili. Tra gli effetti negativi appaiono notevoli quelli dell'indice dei servizi essenziali e la porzione di popolazione con bassi livelli di istruzione, mentre tra quelli positivi la percentuale di studenti e

[5] A differenza della regressione Ridge, che, al contrario, ripartisce equamente il peso tra variabili collineari.

Variable	Estimate	Variable	Estimate
(Intercept)	0.70840	essServicesIdx	-0.00527
SNAI_num	.	lowSchool	-0.00475
logPop	-0.00463	employRate	0.01654
T19_29	.	popIncrRate	-0.00048
T30_44	.	natalityRate	-0.00108
T75p	-0.00233	mortalityRate	-0.00063
Mperc	-0.00049	migrationRate	.
studPerc	0.00071	elderlyRate	-0.00301
fragilIdx	-0.00097	socialServRate	0.00014
motorRate	.	povertyRate	-0.01885
recicling	-0.00045	womPercGiunta	.
protAreas	-0.00130	meanAgeGiunta	-0.00090
landHazard	0.00075	imprese	0.00167
usedArea	.		

Tabella 3: Lasso: coefficienti di regressione delle variabili selezionate

la presenza di imprese. Questi risultati sono compatibili con quelli del modello BSS.

3.3 Alberi di regressione

Un'altra via è quella di utilizzare un albero di regressione. Questa tecnica ha il pregio di essere facilmente interpretabile e non richiede nemmeno l'assunzione di linearità della relazione tra esplicative e risposta. Il dominio p-dimensionale delle esplicative viene partizionato tramite tagli paralleli agli assi delle variabili e il valore assunto dalla risposta in una determinata partizione è stimato dalla media campionaria delle rilevazioni ivi contenute. Il problema è quindi decidere come effettuare questi tagli e quando fermarsi. Adottiamo un metodo di ricerca *greedy*, effettuando i tagli consecutivamente in modo che ogni taglio minimizzi una certa funzione di perdita, ad esempio la devianza totale. Tuttavia, è necessario definire un guadagno minimo necessario (detto *complexity parameter*, *cp*) perché un taglio venga effettuato, altrimenti l'albero interpolerebbe i dati. Questo tipo di modello, inoltre, è particolarmente prono ad overfitting. Anche qui sono possibili approcci basati su train/test o cross-validation.

Un primo albero, basato su un basso valore di cp, è riportato in Figura 13. L'albero risultante è piuttosto complesso. La Figura 14 mostra l'evoluzione dell'errore di cross-validation rispetto alla complessità dell'albero, che dipende

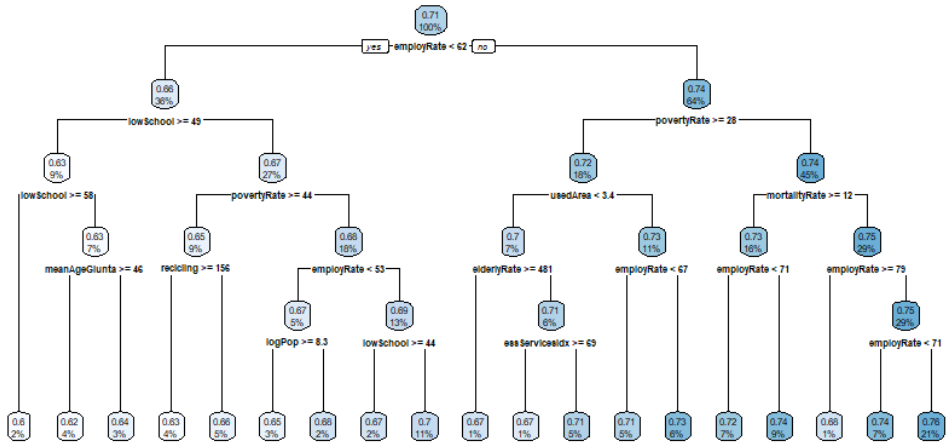

Figura 13: Un primo albero di regressione

dalla scelta di *cp*. Una buona scelta è normalmente il massimo valore di *cp* per cui l'errore di cross-validation stimato sia entro un errore standard dall'errore minimo. Ciò significa scegliere il valore di *cp* più a sinistra tra quelli che comportano un errore che sta sotto la linea orizzontale. Nel nostro caso ciò comporta un albero di dimensione 11.

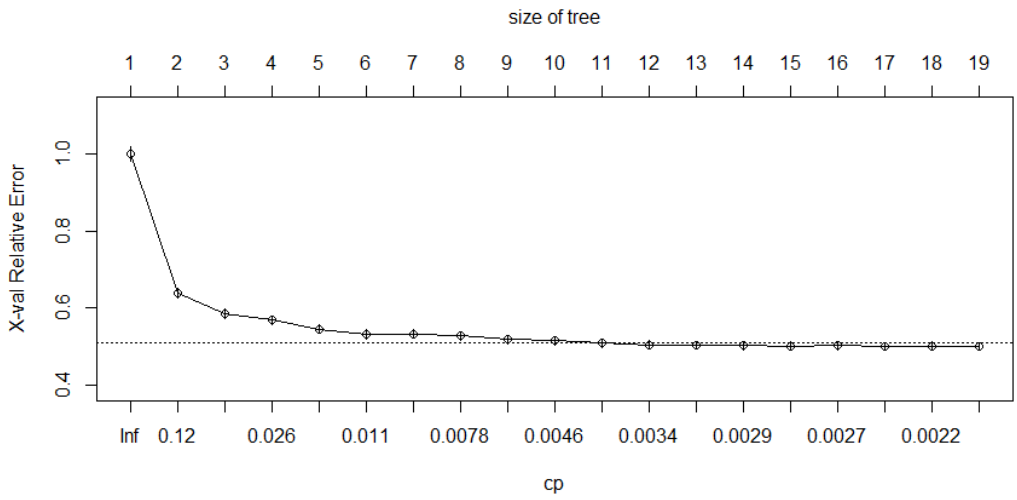

Figura 14: Errore di cross-validation per diverse dimensioni dell'albero

Potando l'albero come descritto, si ottiene quello riportato in Figura 15. Risultano evidenti i ruoli di tasso di occupazione, scolarizzazione, tasso di povertà e mortalità. Le loro interpretazioni sono coerenti con quanto visto nei modelli precedenti. Partendo dalla radice dell'albero, si assegna un comune alla partizione determinata di valori che questo assume per le variabili utilizzate nei

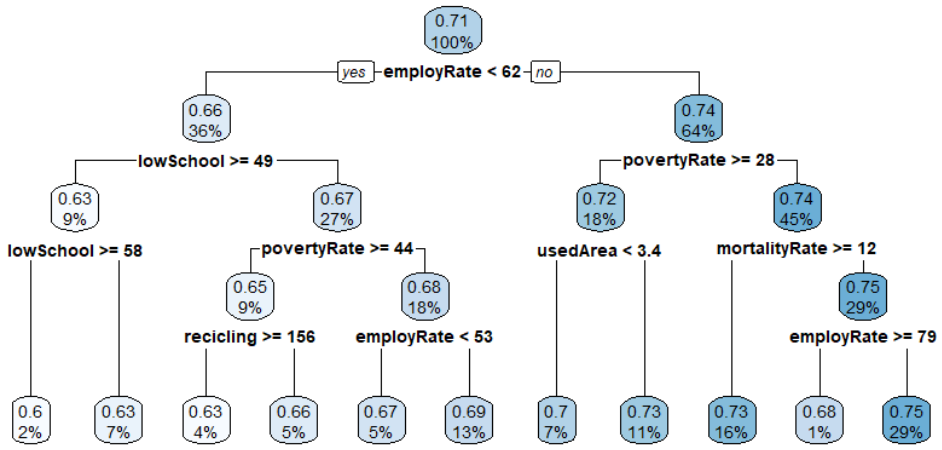

Figura 15: Albero potato

tagli. Si noti che si è effettuata implicitamente una selezione delle variabili. Tuttavia, questo modello è molto sensibile a variazioni nel dataset, per cui piccoli cambiamenti potrebbero modificare parti notevoli dell'albero e dunque la selezione delle variabili è da intendersi come indicazione di quali di queste abbiano un effetto più forte, senza escludere necessariamente l'effetto delle altre. In effetti, gli alberi di regressione sono più adatti ad un utilizzo in chiave di previsione piuttosto che di selezione delle variabili.

3.4 Confronto e verifica delle assunzioni

Si sono esplorati diversi modelli compatibili con i dati a disposizione, prestando particolare attenzione nella loro costruzione ad evitare overfitting, con l'obiettivo di ottenere soluzioni il più possibile parsimoniose.

Le tre diverse tecniche di selezione delle variabili adottate (BSS, LASSO e alberi di regressione) convergono, pur partendo da assunzioni differenti, nell'identificare un ristretto gruppo di variabili come predittori chiave. In tutti i modelli, infatti, emergono l'effetto positivo del tasso di occupazione e quello negativo del tasso di povertà e della bassa scolarizzazione. In questo senso, la coerenza rilevata rafforza la validità delle conclusioni tratte dai singoli modelli.

È opportuno ora valutare l'adeguatezza dei modelli proposti.

Il modello ottenuto tramite best subset selection prevede le usuali assunzioni del modello lineare. In particolare, media nulla, omoschedasticità, normalità e indipendenza degli errori, oltre a linearità nel rapporto tra le esplicative e

la risposta. L'assunzione di normalità degli errori è certamente violata, tuttavia ciò non rappresenta un problema vista l'alta numerosità considerata, per cui vale la normalità asintotica degli stimatori, grazie al teorema del limite centrale.

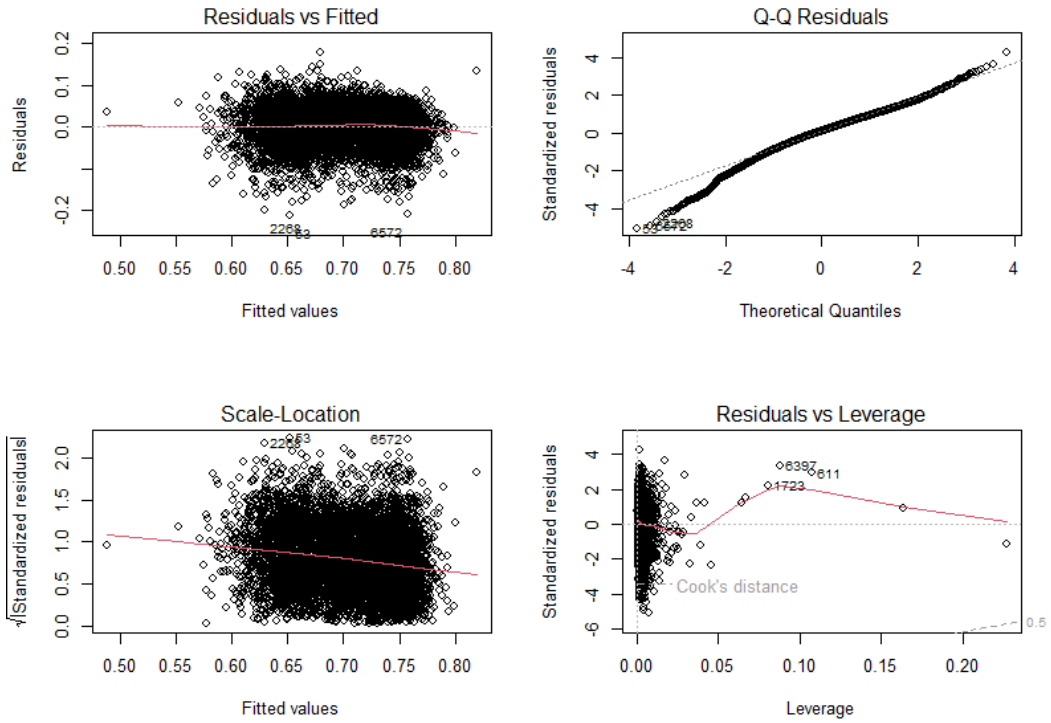

Figura 16: Analisi dei residui del modello ottenuto tramite best subset selection

L'analisi dei residui in Figura 16 è soddisfacente e mostra che l'assunzione di omoschedasticità è sufficientemente adeguata. Per quanto riguarda l'inferenza, sia per questo modello che per il *lasso* è necessario ricorrere a procedure bootstrap o post-selezione. Come si è già evidenziato, infatti, i p-values del modello BSS risultano più piccoli di quanto dovrebbero.

Sia il modello lineare che il modello *lasso* assumono linearità. La verifica di questa assunzione non è banale in presenza di molteplici regressori per via del fatto che la relazione tra ognuna delle esplicative e la risposta va valutata al netto delle altre.

Una via possibile è quella del *component and component-plus-residual plot*, descritta ad esempio in Fox 2015. Definiamo i residui parziali per la j-esima esplicativa come

$$e^{(j)} = e + \hat{\beta}_j X_j$$

dove e indica i residui globali del modello, X_j il valore assunto dalla j-esima esplicativa e $\hat{\beta}_j$ la stima del relativo parametro nel modello. Si valuta il grafico $(X_j, e^{(j)})$, in particolare confrontando la retta $\hat{\beta}_j X_j$ (componente lineare, tratteggiata in blu nel grafico) con una curva di smoothing dei residui parziali (in viola). Una buona concordanza tra le due indica che non vi sono elementi per dubitare dell'assunzione di linearità.

Applicando questo procedimento alle variabili del modello BSS si ottiene quanto riportato in Figura 17, che per brevità riporta solo le 6 variabili più significative. È stato rimossa una osservazione con valore particolarmente elevato di *essServicesIdx*, che rappresentava un punto leva. Da questi risultati l'assunzione di linearità appare verificata per tutte le variabili. Tuttavia, va notato che questo metodo ha potenza ridotta nel caso di esplicative fortemente correlate tra loro, com'è nel nostro caso. Ciononostante, non si rilevano deviazioni evidenti dalla linearità, dunque l'assunzione è ragionevolmente accettabile, almeno approssimativamente.

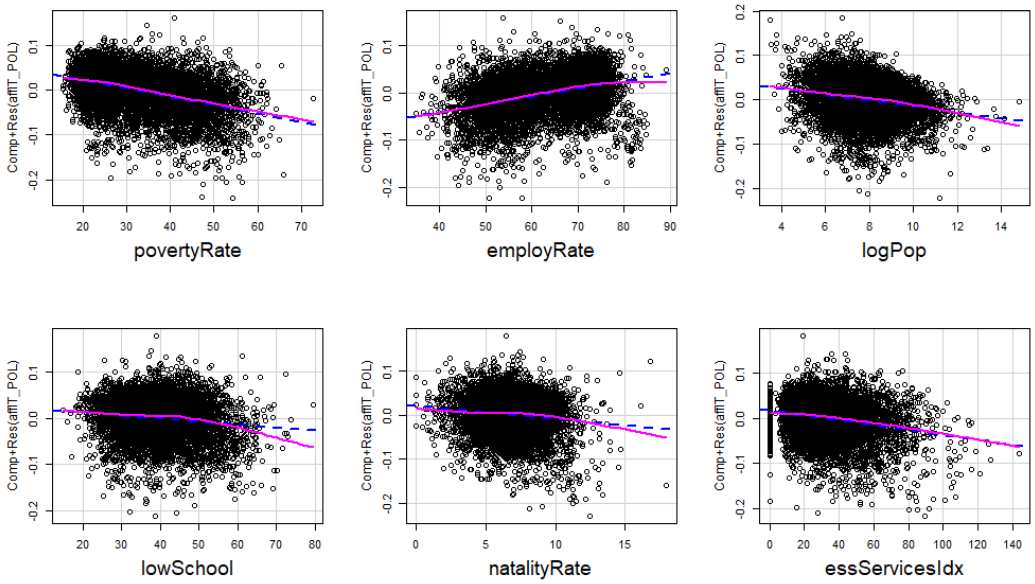

Figura 17: CCPR plot per alcune delle variabili del modello BSS

Una assunzione dubbia è quella di indipendenza degli errori del modello lineare. È plausibile, infatti, che comuni geograficamente vicini abbiano risultati simili tra loro, anche tenendo conto dei valori assunti dalle relative variabili esplicative. In effetti, in Figura 18 emergono dei cluster territoriali in cui i residui sono particolarmente elevati, in particolare in Trentino-Alto Adi-

ge, Sardegna e Calabria. Ciò sembra indicare presenza di correlazione spaziale residua. Esploriamo ora più approfonditamente questa ipotesi.

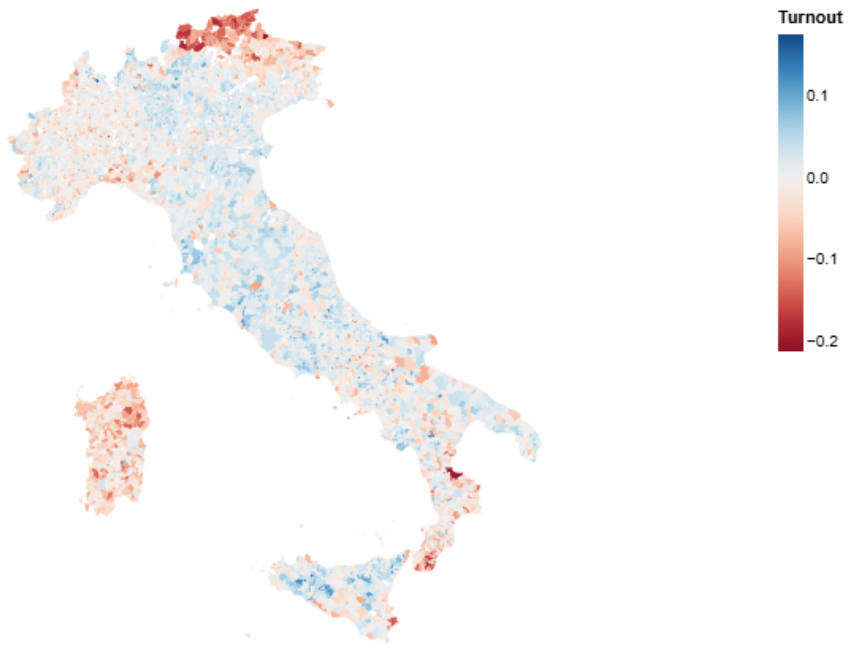

Figura 18: Mappa dei residui del modello BSS

3.5 Modelli per dati spazialmente correlati

I modelli geostatistici tengono conto, nella stima dei parametri, della presenza di correlazione negli errori. L'assunzione di base è quella che osservazioni vicine nello spazio tendano ad avere valori più simili tra loro, anche una volta che si sia tenuto conto dei valori assunti dalle variabili esplicative. A tal fine si assume una certa struttura per la funzione di correlazione spaziale.

Sia $D \subset \mathbb{R}^2$ il dominio spaziale che rappresenta l'insieme dei centroidi dei comuni italiani. Definiamo il relativo processo geostatistico come una collezione di variabili aleatorie $\{Z(s) : s \in D\}$, dove $Z(s)$ rappresenta l'affluenza nel comune il cui centroide è s.

Per descrivere la dipendenza spaziale tra le osservazioni si utilizza il semivariogramma, definito come

$$\gamma(h) = \frac{1}{2} \operatorname{Var}\big(Z(s) - Z(s+h)\big)$$

dove h rappresenta il vettore di spostamento spaziale. Assumiamo che il processo sia isotropo, ovvero che la dipendenza spaziale sia funzione della sola

distanza tra due punti e non dalla direzione dello spostamento: h può allora semplicemente indicare la distanza euclidea tra due unità.

Vogliamo verificare due differenti ipotesi. La prima è valutare se vi sia evidenza di correlazione spaziale, ovvero se comuni vicini abbiano risultati più simili tra loro anche tenendo conto dei valori assunti dalle relative variabili esplicative. La seconda, tenuto conto del risultato della precedente ipotesi, è valutare se le coordinate di per sé siano una variabile significativa - ovvero se, a parità di condizioni, un comune del sud abbia in media affluenza minore rispetto ad un uguale comune del nord, oppure se queste differenze siano imputabili completamente alle variabili in esame.

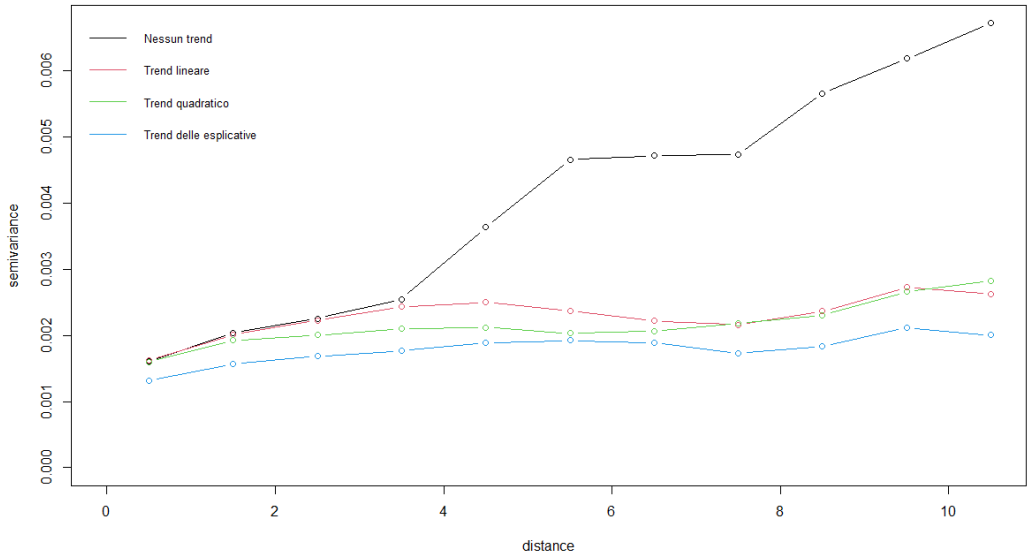

Figura 19: Semivariogramma

La Figura 19 mostra l'andamento marginale del semivariogramma per i dati a nostra disposizione, in quattro diversi casi. La distanza è misurata in gradi di latitudine e longitudine; un grado corrisponde a poco più di 100 chilometri. La linea nera evidenzia la presenza di un trend spaziale, in quanto il semivariogramma cresce con la distanza, indicando maggiore eterogeneità tra comuni spazialmente lontani. Le linee rossa e verde rappresentano il semivariogramma dei residui rispetto a un trend, rispettivamente, lineare o quadratico della media del processo rispetto alle coordinate. Il trend quadratico non sembra migliore di quello lineare, che dunque risulta preferibile. Ciò indicherebbe

che una parte della variabilità sia da attribuirsi alla posizione geografica del comune.

Tuttavia, va notato che l'apparente presenza di correlazione spaziale potrebbe essere causata da una differenza dei valori assunti in media dalle variabili concomitanti in diverse regioni geografiche, com'è lecito attendersi. Si veda, sempre in Figura 19, la linea azzurra, calcolata sui residui rispetto ad un trend dato dalle esplicative selezionate dal modello modello BSS. Come si vede, questo modello per la media è migliore del trend lineare sulle coordinate e sembrerebbe indicare assenza, o effetto molto debole, della componente spaziale.

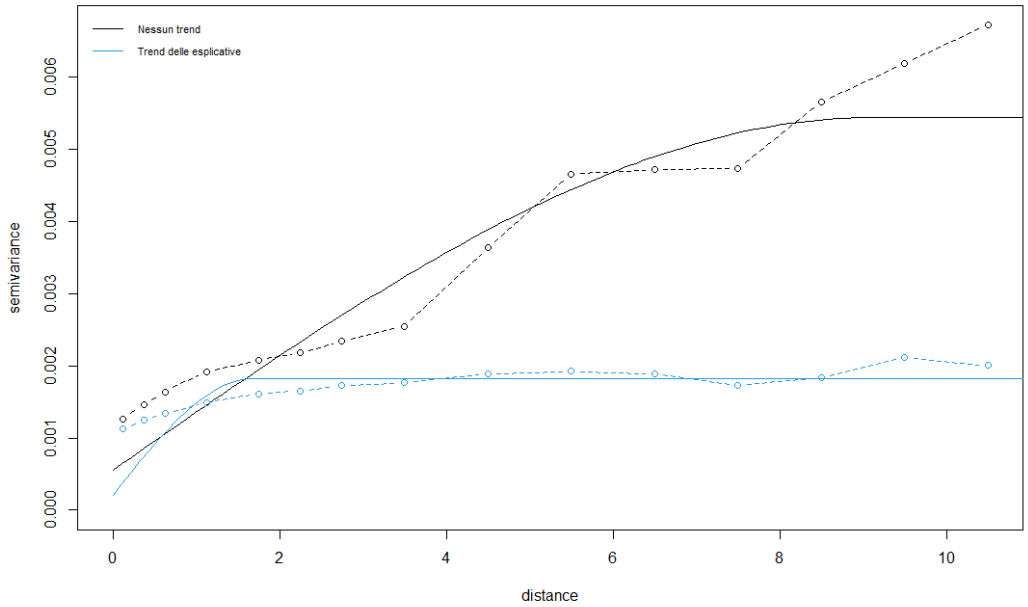

Figura 20: Modelli parametrici per il semivariogramma

Le linee continue in Figura 20 mostrano l'adattamento di un modello parametrico per il variogramma per i dati grezzi (in nero) e sui residui rispetto al trend determinato dalle variabili esplicative del modello BSS (in azzurro). Il modello utilizzato è definito come

$$\gamma(h) = c_0 + c\big(1 - \rho(h;\phi)\big)$$

dove c_0 è l'effetto nugget, c il partial sill e $\rho(h;\phi)$ la funzione di correlazione spaziale. Per poter stimare il modello è dunque necessaria un'assunzione pa-

rametrica per $\rho(h;\phi)$. Una scelta possibile, coerente con la tipologia di dati in oggetto, è la famiglia sferica, che prevede una correlazione decrescente per distanze fino a ϕ e nulla per distanze maggiori. In particolare, è definita come

$$\rho(h) = \begin{cases} 1 - \dfrac{3h}{2\phi} + \dfrac{h^3}{2\phi^3} & \text{se } 0 \leq h \leq \phi \\ 0 & \text{se } h > \phi \end{cases}$$

dove ϕ rappresenta il parametro di scala.

Per il modello rappresentato dalla linea azzurra si ottiene $\hat{\phi} = 1.68$, che indica presenza di correlazione spaziale tra comuni i cui centroidi distano meno di 1.68°, ovvero circa 180 chilometri.

Per verificare la seconda ipotesi e per adattare il modello vero e proprio è ora necessario ricorrere ad una assunzione parametrica sui residui rispetto al trend della media: utilizziamo un processo stocastico gaussiano.

Un processo stocastico $\{Z(s) : s \in D\}$ è detto gaussiano se, per ogni insieme finito di localizzazioni $s_1, \ldots, s_n \in D$, il vettore aleatorio $Z = \big(Z(s_1), \ldots, Z(s_n)\big)^T$ ha distribuzione normale multivariata, ovvero

$$Z \sim \mathcal{N}_n(\mu, \Sigma),$$

dove $\mu = \mathbb{E}[Z] \in \mathbb{R}^n$ e $\Sigma_{ij} = \mathrm{Cov}(Z(s_i), Z(s_j))$.

Nonostante si sia in precedenza rifiutata l'ipotesi di normalità sui residui del modello BSS, va ricordato che la normalità si riferisce in questo contesto alla distribuzione congiunta del processo. Per questo motivo, l'assunzione di normalità è plausibile, almeno come ipotesi di lavoro.

La stima di questo modello è computazionalmente onerosa, con una complessità dell'ordine di $O(n^3)$. Per questo motivo, il modello è valutato su $n = 2000$ comuni scelti casualmente, stratificati per macro-area geografica e popolazione. Inoltre, per lo stesso motivo, non effettuiamo una selezione stepwise ma ci limitiamo a mantenere nel modello le variabili esplicative che risultano immediatamente significative. Il primo modello valutato è quello che include le stesse variabili del modello BSS, che dà i risultati in Tabella 4.

A livello $\alpha = 0.01$ risultano significative le prime 9 variabili, le quali rimangono tutte significative anche quando considerate da sole in un modello

ridotto, il quale risulta dunque il modello preferibile. Le stime dei coefficienti e relativi standard error sono riportati in Tabella 5. L'aggiunta di latitudine e longitudine, utilizzate come esplicative, non apporta alcun contributo: entrambe risultano non significative, con p-value maggiore di 0.5. Ciò suggerisce che la minore affluenza media nei comuni del sud e delle isole sia da attribuirsi esclusivamente ai diversi valori assunti in media dalle rilevanti variabili socio-demografiche ed economiche. Ovvero, il modello esclude la presenza di un "effetto territoriale", imputando invece le differenze di affluenza solamente alle disuguaglianze strutturali tra macro-aree.

Questo è il modello più parsimonioso ottenuto fin'ora. Le indicazioni che fornisce sono compatibili con quelle dei modelli precedenti, con effetto positivo di tasso di occupazione e percentuale di studenti, e negativo per popolazione, percentuale di residenti con bassa scolarizzazione, tasso di povertà, percentuale di over 75, tasso di natalità e di crescita della popolazione.

	Stima	Std. Error	p-value
Intercept	0.81443	0.04526	0.00e+00
logPop	-0.00697	0.00105	3.28e-11
lowSchool	-0.00075	0.00017	1.27e-05
employRate	0.00115	0.00027	2.37e-05
povertyRate	-0.00091	0.00023	6.44e-05
studPerc	0.22885	0.05879	9.92e-05
popIncrRate	-0.00007	0.00002	2.07e-04
natalityRate	-0.00274	0.00075	2.35e-04
T75p	-0.13701	0.04504	2.35e-03
Mperc	-0.14878	0.04907	2.43e-03
protAreas	-0.00011	0.00004	1.08e-02
fragilIdx	-0.00152	0.00068	2.66e-02
meanAgeGiunta	-0.00024	0.00014	9.13e-02
elderlyRate	-0.00001	0.00001	3.47e-01
essServicesIdx	-0.00013	0.00014	3.71e-01
usedArea	0.00008	0.00014	5.74e-01
SNAI_num	-0.00091	0.00229	6.90e-01
imprese	0.00002	0.00006	7.47e-01
landHazard	0.00003	0.00009	7.57e-01

Tabella 4: Coefficienti per il primo modello

	Stima	Std. Error	p-value
Intercept	0.76005	0.05734	0.00e+00
logPop	-0.00516	0.00090	9.76e-09
employRate	0.00135	0.00026	2.50e-07
lowSchool	-0.00084	0.00017	3.95e-07
studPerc	0.25775	0.05539	3.26e-06
povertyRate	-0.00099	0.00023	1.20e-05
T75p	-0.17708	0.04101	1.58e-05
Mperc	-0.17605	0.04855	2.88e-04
natalityRate	-0.00256	0.00075	6.10e-04
popIncrRate	-0.00006	0.00002	1.60e-03

Tabella 5: Coefficienti del modello scelto

4 La fallacia ecologica

Nei capitoli precedenti si è analizzata una serie di relazioni tra variabili a livello di comune. È necessario fare qualche appunto sul modo in cui questi risultati vadano interpretati, onde evitare di trarre conclusioni errate.

La *fallacia ecologica* è l'errore inferenziale per cui si attribuiscono a livello individuale relazioni osservate su dati aggregati. Il punto di riferimento in merito è Robinson 1950, nel quale viene messa in dubbio per la prima volta una metodologia fin lì ampiamente applicata, per cui l'utilizzo di dati aggregati era spesso ingiustificato. Un esempio, divenuto celebre, può evidenziare tale incongruenza. Si considerino i dati del censimento del 1930, per i residenti di almeno 10 anni di età, in tutti i 48 stati degli Stati Uniti: da questi si estraggono, per ogni stato, la percentuale di nati all'estero e la percentuale di alfabetizzati. Tra queste 48 coppie di valori, la correlazione rilevata è pari a 0.526. Dunque, l'indicazione che si ottiene da queste rilevazioni aggregate è che la maggior presenza di nati all'estero sia indicazione di maggiore alfabetizzazione. L'implicazione che questo valga anche a livello individuale, ovvero che i nati all'estero abbiano tassi di alfabetizzazione maggiore dei nati negli Stati Uniti, è tuttavia errata. Infatti, dagli stessi dati si ottiene, se si analizzano a livello individuale, una correlazione pari a -0.118. Questo risultato è notevole perché evidenzia, in un procedimento che può apparire altrimenti valido, la presenza di rischi concreti nell'inferenza, al punto che la correlazione considerata non è solamente sovrastimata, ma è addirittura stimata nella direzione opposta rispetto a quella corretta.

L'errore logico avviene nel passaggio dalle relazioni aggregate a quelle individuali. Infatti, questo salto impone implicitamente una assunzione di omogeneità interna ai gruppi e tra le aree, detta *constancy assumption*. Ad esempio, nel caso sopra citato stiamo in pratica assumendo che la percentuale di alfabetizzazione sia costante all'interno di ognuno dei due gruppi in ogni stato. D'altronde, come nota Freedman 1999, l'effetto dovuto a variabili di confondimento è presente in ogni studio osservazionale e dunque, una volta che si sia tenuto conto degli effetti contestuali, l'assunzione è meno stringente, limitandosi ad imporre che l'alfabetizzazione interna ai gruppi non dipenda dalle percentuali che questi gruppi rappresentano in un dato stato. Se ciò fosse

soddisfatto, la correlazione varrebbe anche a livello individuale. Spesso però questa assunzione è violata e siamo dunque in presenza di *bias di aggregazione*. Anche nel nostro esempio, infatti, l'assunzione non regge: l'immigrazione verso gli Stati Uniti tende a concentrarsi negli stati più ricchi e alfabetizzati, dunque la proporzione di nati all'estero non è indipendente dal tasso di alfabetizzazione.

In letteratura sono riportati innumerevoli altri esempi, anche riguardanti dati elettorali simili a quelli a nostra disposizione. Klein, Freedman 1993 mostrano come regressioni ecologiche su dati elettorali siano state utilizzate in una causa legale riguardante il *Voting Rights Act*, il quale protegge le minoranze da una costruzione dei collegi elettorali che li svantaggi appositamente (il cosiddetto *gerrymandering*). Il modello utilizzato restituiva risultati per il candidato ispanico fondamentalmente diversi dividendo la popolazione in due parti (Ispanici - Non Ispanici, 33%) o in tre parti (Bianchi - Neri - Ispanici, 17%), evidenziando la presenza della fallacia, in quanto dai dati aggregati sulla composizione dei collegi si intendevano trarre conclusioni sulla tendenza di voto dei singoli gruppi. Anche in questo caso, la *constancy assumption* non regge: il voto è determinato da innumerevoli altri fattori, ed è illogico supporre che la distribuzione dei voti tra gli ispanici sia uguale in tutti i distretti.

Vediamo un caso che esemplifica la gravità potenziale di questa fallacia, ripreso da Piantadosi et al. 1988. Si considerino N individui, divisi in r gruppi di numerosità $n = k$, per cui vengono rilevate due variabili, X e Y. Possiamo immaginare che i gruppi rappresentino le aree geografiche entro cui questi sono contenuti. Per ogni gruppo g, le unità assumono per X valori interi consecutivi da $k(g-1)+1$ a kg, mentre per Y assumono valori gli stessi valori in ordine inverso, da kg a $k(g-1)+1$.

Questo disegno è riportato, per $N = 9$ e $k = 3$, in Tabella 6. Per questi dati la correlazione a livello individuale è $\rho = 0.6$. Se invece si aggregano i dati secondo il gruppo, tutti i gruppi hanno uguale media per le due variabili, dunque la correlazione ecologica è $\rho_e = 1$. Tuttavia, all'interno del singolo gruppo la correlazione tra gli individui è perfettamente inversa, con $\rho_w = -1$. Ciò è dovuto al fatto che la covarianza tra gli individui può essere scomposta come

$$Cov(X,Y) = Cov_W(X,Y) + Cov_B(X,Y)$$

X	Y	Gruppo
1	3	1
2	2	1
3	1	1
3	5	2
4	4	2
5	3	2
5	7	3
6	6	3
7	5	3

Tabella 6: Dati per illustrare la fallacia ecologica

ovvero la somma di covarianza nei gruppi e tra i gruppi. Assumere che la correlazione a livello ecologico corrisponda a quella a livello individuale corrisponde ad ignorare la covarianza interna ai gruppi. Lo stesso concetto si estende al caso di regressione. Si consideri il modello

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \ldots, n.$$

Il corrispondente modello per dati aggregati è

$$\bar{Y}_g = \beta_0^* + \beta_1^* \bar{X}_g + \varepsilon_g \quad g = 1, \ldots, r$$

dove $\bar{Y}_g$ e $\bar{X}_g$ rappresentano la media della corrispondente variabile nel gruppo g, oppure, in caso di variabile dicotomica, la proporzione di individui che assumono una fissata modalità entro il gruppo. In questa formulazione è possibile che l'esplicativa e l'errore siano correlati a livello aggregato. Inoltre, non vi è alcuna garanzia che $\hat{\beta}_1^* \simeq \hat{\beta}_1$.

L'utilizzo di dati aggregati non è tuttavia da scartare a priori. In effetti, la principale critica di Robinson 1950 è che gli studi ecologici sono spesso utilizzati, esplicitamente o implicitamente, per trarre conclusioni a livello individuale. Se l'elemento di interesse sono le relazioni ecologiche o gli effetti dovuti all'appartenenza ad un gruppo, le conclusioni ottenute da questi metodi sono perfettamente valide. Ad esempio, un utilizzo della regressione ecologica nel valutare la bontà dell'effetto di una nuova politica applicata in una certa area geografica è lecito. Inoltre, i dati ecologici sono in molti casi più facilmente ottenibili rispetto a quelli a livello individuale; in alcuni contesti i dati a livello

individuale non sono ottenibili - si pensi ai risultati elettorali, in cui il vantaggio dei dati aggregati, completi e facilmente disponibili, è notevole sui dati individuali, costosi e soggetti ad errore campionario. L'utilizzo di dati ecologici può anche essere finalizzato all'individuazione di pattern macroscopici, in particolare a livello geografico, che potrebbero sfuggire ai dati individuali. Un ulteriore utilizzo è quello di formulazione di ipotesi esplorative, che vanno poi confermate tramite l'utilizzo di modelli più affidabili. Di contro, va notato che a causa del *bias* di aggregazione la scelta del raggruppamento utilizzato influenza necessariamente i risultati. Pertanto, è necessario scegliere gruppi che abbiano un significato esplicito. Inoltre, non è possibile in questo contesto controllare per variabili di confondimento individuali, ma solamente per la loro versione aggregata, il ché può risultare in una omissione di fattori di rilievo.

È importante notare che non esiste un livello unico su cui i dati vadano analizzati. Infatti, le relazioni individuali possono interagire con quelle a livello aggregato producendo effetti differenti in aree o gruppi diversi. Per questo, anche l'utilizzo di dati a livello individuale non è privo di rischi: la *fallacia atomistica* è l'errore che si commette quando, partendo da dati a livello individuale, si traggono conclusioni a livello aggregato. Questa rappresenta l'inverso della fallacia ecologica. Infatti, i pattern che si rilevano a livello individuale possono variare notevolmente tra gruppi, poiché le unità appartenenti a uno stesso cluster tendono a condividere caratteristiche comuni.

Per dati di questo tipo è necessario, se si vogliono trarre conclusioni su tutti i livelli dell'analisi, utilizzare *modelli multilivello* (anche detti *gerarchici*). Questi modelli non assumono indipendenza tra gli individui ma tengono conto della correlazione tra unità in uno stesso gruppo tramite l'utilizzo di effetti fissi e effetti casuali. Questi ultimi modellano la variabilità tra gruppi rendendo le stime dei parametri dipendenti dal gruppo di appartenenza degli individui, facendo ricorso a una qualche assunzione parametrica sulla forma della matrice di correlazione dei cluster.

Nel caso da noi analizzato l'unità ecologica è il comune, mentre l'unità individuale è rappresentata dal singolo elettore. L'interpretazione dei coefficienti dei modelli, come si è già detto, è da intendersi a livello aggregato. Si prenda ad esempio il modello geostatistico stimato in Tabella 5. Il coefficiente *Mperc* indica che un comune composto da soli elettori di sesso femminile re-

gistrerebbe in media una affluenza del 17.6% superiore rispetto a quella di un comune composto da soli maschi. Benché sia un'ipotesi suggestiva, concludere da ciò che le elettrici femmine abbiano affluenza media maggiore è un errore inferenziale[6]. Piuttosto, è corretto affermare che l'affluenza in un dato comune tende a crescere con la percentuale di maschi. Come nell'esempio sui nati all'estero e l'analfabetismo, questo può essere dovuto ad altre cause non immediatamente evidenti. Questa specifica variabile non è di particolare interesse a livello ecologico, ma, se ci soffermiamo su altre come *employRate*, *povertyRate* o *lowSchool*, possiamo trarre delle conclusioni notevoli.

L'interpretazione suggerita dal modello è quella per cui i comuni più sviluppati economicamente e socialmente, ovvero quelli con tasso di occupazione, di studenti e di scolarizzazione maggiore e tasso di povertà minore, siano quelli in cui l'affluenza è maggiore. Questa interpretazione dei risultati non presenta caratteristiche riconducibili alla fallacia ecologica, se si è attenti a non cadere nella tentazione di dedurre da questo che, ad esempio, chi ha un'occupazione sia più propenso al voto[7]. Un naturale sviluppo di questa analisi è l'impiego di modelli multilivello al fine di ottenere stime più affidabili e inferenze a livello individuale. Questo richiederebbe l'utilizzo di dati individuali, ottenuti ad esempio da sondaggi stratificati per territorio, che non abbiamo al momento a disposizione.

[6]Questo è un chiaro caso di fallacia ecologica: l'affluenza complessiva degli elettori maschi è stata maggiore in tutte le elezioni considerate.

[7]Semmai, questa può essere interpretata come indicazione per una ipotesi esplorativa.

5 Conclusioni

In questa tesi è stata condotta un'analisi sui dati, aggregati a livello comunale, delle elezioni politiche del 2013, 2018 e 2022 e delle elezioni europee del 2014, 2019 e 2024. L'obiettivo primario è stato l'identificazione di quali caratteristiche socio-demografiche ed economiche influenzino l'affluenza, basandosi su un'analisi di tipo ecologico, ovvero focalizzandoci sulle relazioni a livello di comune e non di singolo elettore.

Dopo aver raccolto e pre-processato i dati, sono stati adattati diversi modelli, con l'obiettivo di effettuare una selezione delle variabili. In primo luogo si è seguito l'approccio del modello lineare, utilizzando tecniche come la best subset selection e la selezione stepwise. Dopodiché, si sono esplorati modelli più flessibili, come lasso e alberi di regressione. Da ultimo, si è utilizzato un modello geostatistico basato su un processo gaussiano, al fine di tener conto della potenziale correlazione spaziale tra comuni vicini.

Un risultato notevole è la coerenza delle conclusioni emerse dai diversi modelli. Tutti i modelli costruiti, pur partendo da assunzioni differenti, sono concordi nell'evidenziare alcuni punti principali: emerge chiaramente un effetto positivo sull'affluenza di tasso di occupazione e percentuale di studenti, mentre negativo per tasso di povertà e percentuale di residenti con bassi livelli di istruzione. Ciò suggerisce che, a livello comunale, contesti economicamente e socialmente più sviluppati tendano a mostrare maggiore partecipazione elettorale.

Il modello geostatistico è risultato utile per interpretare la dimensione spaziale del fenomeno. Questo, infatti, indica presenza di correlazione spaziale per comuni vicini, in particolare fino a 180 chilometri di distanza. Inoltre, questo non rileva un impatto significativo delle coordinate (latitudine e longitudine) sull'affluenza, una volta che si sia tenuto conto delle altre variabili. Ciò indica l'assenza di un "effetto territoriale", ovvero suggerisce che le differenze di affluenza tra macro-aree geografiche che emergono dalle analisi esplorative siano da attribuire alle disuguaglianze strutturali che le caratterizzano.

Infine, si è posta particolare attenzione ai limiti interpretativi dell'analisi, discutendo alcune note di cautela che si rendono necessarie onde evitare inferenze errate dovute alla fallacia ecologica.

Bibilografia

Bloise, Francesco et al. (2021). «Inequality and voting in Italy's regions». In: *Territory, Politics, Governance* 9.3, pp. 365–390. DOI: 10.1080/21622671.2020.1837219.

Chan, Jireh Yi-Le et al. (2022). «Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review». In: *Mathematics* 10.8. ISSN: 2227-7390. DOI: 10.3390/math10081283.

Ferrara, Alessandra et al. (2023). «L'Italia che non vota: dinamiche e propagazione spazio-temporale dell'astensionismo». In: *Italian Journal of Electoral Studies (IJES)* 86.2, pp. 35–51. DOI: 10.36253/qoe-14083.

Fox, John (2015). *Applied regression analysis and generalized linear models.* Sage Publications. ISBN: 9781452205663.

Freedman, David A. (1999). «Ecological inference and the ecological fallacy». In: *International Encyclopedia of the social & Behavioral sciences* 6.4027-4030, pp. 1–7.

Frohlich, N et al. (2001). «Assessing socioeconomic effects on different sized populations: To weight or not to weight?» In: *Journal of Epidemiology & Community Health* 55.12, pp. 913–920. DOI: 10.1136/jech.55.12.913.

Harka, Elona, Lorenzo Rocco (2022). «Studying more to vote less. Education and voter turnout in Italy». In: *European Journal of Political Economy* 72, p. 102115. DOI: 10.1016/j.ejpoleco.2021.102115.

James, Gareth et al. (2021). *An introduction to statistical learning.* 2ª ed. Springer texts in statistics. New York, NY: Springer. ISBN: 978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1.

Johnson, R.A., D.W. Wichern (2013). *Applied Multivariate Statistical Analysis: Pearson New International Edition.* Pearson Education. ISBN: 9781292037578.

Klein, Stephen P., David A. Freedman (1993). «Ecological Regression in Voting Rights Cases». In: *CHANCE* 6.3, pp. 38–43. DOI: 10.1080/09332480.1993.10542376.

Korn, Edward L., Barry I. Graubard (1995). «Examples of Differing Weighted and Unweighted Estimates from a Sample Survey». In: *The American Statistician* 49, pp. 291–295. DOI: 10.1080/00031305.1995.10476167.

Kuhn, Max, Kjell Johnson (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. 1st. Chapman e Hall/CRC. ISBN: 9781315108230. DOI: 10.1201/9781315108230.

Leighley, Jan E., Jonathan Nagler (1992). «Socioeconomic Class Bias in Turnout, 1964–1988: The Voters Remain the Same». In: *American Political Science Review* 86.3, pp. 725–736. DOI: 10.2307/1964134.

Miller, A. (2002). *Subset Selection in Regression*. 2nd. Chapman e Hall/CRC. ISBN: 9780367396220. DOI: 10.1201/9781420035933.

Piantadosi, Steven et al. (1988). «The ecological fallacy». In: *American journal of epidemiology* 127.5, pp. 893–904.

Robinson, W. S. (1950). «Ecological Correlations and the Behavior of Individuals». In: *American Sociological Review* 15.3, pp. 351–357. DOI: 10.2307/2087176.

Salvan, Alessandra et al. (2020). *Modelli lineari generalizzati*. Milano: Springer Milan. ISBN: 978-88-470-4002-1. DOI: 10.1007/978-88-470-4002-1_2.

Appendice A

Variabile	Descrizione	Fonte
ISTAT_CODE	Codice ISTAT - Comune	ISTAT
Municipality	Comune	ISTAT
Macroareaterr	Macro-area territoriale ISTAT	ISTAT
Cod_Reg	Codice ISTAT - Regione	ISTAT
Regione	Regione	ISTAT
Cod_Prov	Codice ISTAT - Provincia	ISTAT
Provincia	Provincia	ISTAT
Sigla_Prov	Sigla Provincia	ISTAT
SNAI_21y27y	Strategia Nazionale Aree Interne (2021-2027)	SNAI
ComDest-Prev_21y27y	Comune di destinazione Strategia Nazionale Aree Interne (2021-2027)	SNAI
SNAI_num	Conversione numerica di SNAI_21y27y, da polo = 1 a ultraperiferico = 6	SNAI
studPerc	Studenti a distanza fino a 4 h dal Comune di residenza. 2022	ISTAT
fragilIdx	Indice di Fragilità Comunale. Mediana 2018-2021 (decili)	ISTAT
motorRate	Tasso di motorizzazione ad alta emissione per 100 abitanti. Media 2018-2021	ISTAT (su dati ACI)
recicling	Raccolta indifferenziata dei rifiuti urbani per abitante. Media 2018-2021	ISTAT - ISPRA (Catasto nazionale dei rifiuti)
protAreas	Superficie comunale coperta da aree naturali protette. Media 2018-2021	ISTAT - Ispra su dati del Ministero dell'Ambiente e della Sicurezza Energetica
landHazard	Superficie aree con pericolosità da frane elevata e molto elevata. Media 2018-2021	ISTAT - ISPRA
usedArea	Superficie suolo consumato. Media 2018-2021	ISTAT - ISPRA
essServicesIdx	Indice di accessibilità ai servizi essenziali. Media 2018-2021	ISTAT
lowSchool	Popolazione 25-64 anni con bassi livelli di istruzione. Media 2018-2021	ISTAT - Censimento permanente
employRate	Tasso di occupazione 20-64 anni. Media 2018-2021	ISTAT - Censimento permanente
popIncrRate	Tasso di incremento della popolazione. Media 2018-2021	ISTAT - Censimento permanente
natalityRate	Tasso di natalità. Media 2014-2020	ISTAT - Movimento Pop. Resid.
mortalityRate	Tasso di mortalità. Media 2014-2020	ISTAT - Movimento Pop. Resid.
migrationRate	Tasso migratorio. Media 2014-2020	ISTAT - Movimento Pop. Resid.
elderlyRate	Indice di vecchiaia. Media 2014-2020	ISTAT - Popolazione Residente
socialServRate	Bambini presi in carico dai servizi comunali per l'infanzia. Media 2014-2020	ISTAT - Indagine censuaria sugli interventi e servizi sociali dei Comuni
povertyRate	Contribuenti IRPEF con reddito complessivo inferiore a 10.000 €. Media 2014-2020	ISTAT - Elaborazione su dati Ministero dell'Economia e delle Finanze
womPercGiunta	Donne negli organi decisionali – Giunte comunali. Media 2014-2020	ISTAT - Anagrafe degli Amministratori locali e regionali
meanAgeGiunta	Età media amministratori locali. Media 2014-2020	ISTAT - Anagrafe degli Amministratori locali e regionali
imprese	Tasso di imprenditorialità. Media 2014-2020	ISTAT - Elaborazione su Registro statistico delle imprese attive (ASIA - Imprese)
affl M/F/T_POL/EU	Affluenza per sesso o totale, media delle ultime tre elezioni politiche o europee	Ministero dell'Interno

Tabella 7: Tabella delle variabili

Ringraziamenti

Desidero innanzitutto ringraziare il prof. Livio Finos per la disponibilità dimostrata, per l'umanità e per gli spunti di riflessione offerti nel corso della stesura. Allo stesso modo, ringrazio la prof.ssa Venera Tomaselli per la raccolta dei dati, senza i quali questa tesi non sarebbe potuta esistere, e per la fiducia dimostrata nel metterli a disposizione.

Ringrazio la mia famiglia, in particolare i miei genitori Vania e Maico, i nonni Nazzarena, Milena e Gianni e lo zio Vanni, per il sostegno incondizionato che mi avete da sempre dimostrato, per la pazienza e per l'incoraggiamento. Grazie per avermi permesso di affrontare questo percorso con serenità.

Un ringraziamento speciale a Bea, Boci, Francesca, Giulia e Karin, per esserci stati dall'inizio alla fine, per essermi stati sempre accanto, nonostante tutto. Grazie per le risate, per le esperienze passate insieme, per i discorsi poco seri e per quelli più profondi, per essere sempre un punto di riferimento. Grazie per essere stati la costante quando tutto cambiava, compresi noi stessi, per il supporto e per l'affetto che mi avete sempre dimostrato. Grazie per tutto quello che mi avete insegnato, per avermi aiutato a crescere e a conoscere meglio anche me stesso. Spero di essere riuscito anch'io, allo stesso modo, a lasciare qualcosa a voi.

Grazie a Andrea, Bazan e Elisa, per la compagnia durante questi tre anni di lezioni, per le risate, per il tempo passato a studiare e per quello passato a far nulla, per aver condiviso l'ansia prima di tanti esami e il senso di libertà dopo. Grazie per esserci stati ogni giorno e per aver reso questo percorso più facile da affrontare.

Vi voglio bene. Grazie.

Nicola Menegon

Padova, settembre 2025