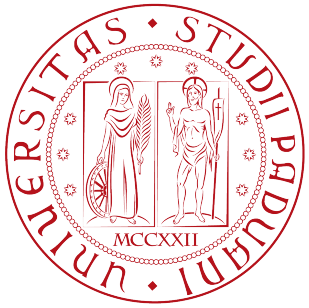
Università degli studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche



Tesi di Laurea

Analisi delle curve di mortalità: stima non parametrica della densità con approccio bayesiano

Relatore: Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Correlatore: Prof. Tommaso Rigon
Dipartimento di Economia, Università di Milano-Bicocca

Laureando: Davide Agnoletto
Matricola N. 1236932

Anno Accademico 2020/2021

Indice

Introduzione

La curva di mortalità per età consiste nella distribuzione dei decessi annuali ripartiti in classi di età. Essa rappresenta lo strumento principale per lo studio della mortalità per età e la sua modellazione è una sfida che attrae gli studiosi da sempre. Questa curva non è costante e nel corso degli anni ha mutato forma e caratteristiche, per questo motivo riveste un ruolo primario come indicatore del progresso di una società.
L'obiettivo di questa tesi è la modellazione di tale funzione tramite modelli che garantiscano una elevata flessibilità ma la cui stima non sia gravosa dal punto di vista computazionale. In questo senso, si discute un modello appartenente alla classe dei modelli bayesiani non parametrici per la stima di una funzione ed una sua estensione nel caso le curve da stimare siano molteplici. Inoltre, tale estensione fornisce come effetto secondario l'individuazione di *clusters* nella popolazione di curve. L'utilità di questo tipo di approccio consiste nel fatto che, a differenza di modelli tradizionali, rende possibile la modellazione di curve di mortalità anche molto variabili e dalle forme irregolari. Ciò è possibile in quanto la stima per ciascuna curva di mortalità tiene in considerazione l'informazione contenuta nelle curve contenute nel medesimo gruppo.
Vengono trattate sia la formulazione teorica del modello che la struttura degli algoritmi di stima.
I modelli vengono poi applicati prima alle curve di mortalità della popolazione italiana dal 1872 al 2018 e successivamente a quelle per ciascun comune italiano relativamente all'anno 2020. Lo scopo di quest'ultima applicazione

è quello capire il comportamento del modello anche in situazioni in cui le curve sono particolarmente irregolari per motivi che possono essere legati al fatto che molti comuni sono scarsamente abitati e agli effetti della pandemia di COVID-19 che ha colpito l'Italia in quell'anno.

I Capitoli 1 e 2 presentano gli strumenti demografici e probabilistici necessari alla comprensione dell'analisi sviluppata nel seguito. Nei Capitoli 3 e 4 si discutono gli aspetti teorici e l'implementazione del modello per la stima di una curva di mortalità e la sua estensione, commentando poi le rispettive applicazioni. Successivamente nel Capitolo 5 si discutono i risultati dell'adattamento del modello ai dati comunali relativi alla mortalità da COVID-19. Infine vengono riportate le conclusioni del lavoro svolto, valutando possibili limiti ed estensioni future.

Capitolo 1

La curva di mortalità

In questo capitolo vengono esposte brevemente esposti alcuni concetti legati alla curva di mortalità. Si inizia con una veloce definizione delle caratteristiche principali della curva di mortalità per età. Successivamente si presentano alcuni modelli parametrici tra quelli maggiormente utilizzati per lo studio di queste curve indicandone i punti di forza e di debolezza al fine di avere più chiaro il contesto nel quale si inserisce l'approccio che viene adottato nel seguito. Infine si descrivono i dati utilizzati nelle analisi.

1.1 Le principali caratteristiche

La tavola di mortalità venne introdotta per la prima volta da John Graunt in una pubblicazione del 1662 (ristampa: Graunt, 1977) e venne poi perfezionata nel corso del tempo, acquisendo una struttura matematica più rigorosa. In ambito demografico rappresenta un importante strumento per l'analisi statistica dei decessi nella popolazione e grazie ad essa è possibile ottenere indicazioni sul livello del progresso di una popolazione sulla base della longevità di quest'ultima (Livi Bacci, 1999).
Vengono elencate di seguito alcune delle quantità necessarie per la costruzione della tavola di mortalità, in particolare quelle che poi saranno rilevanti nel prosieguo (per maggiori dettagli si veda Livi Bacci, 1999). Si considerano

intervalli di età unitari da 0 a 110 anni ed un intervallo finale aperto $[110, \infty)$. Per $x = 0, \ldots, 110$ si ha:

- m_x, tasso di mortalità specifico per età, ottenuto come il rapporto tra il numero dei decessi ed il numero di individui esposti al rischio di morte per età ed anno;

- a_x, numero medio di anni vissuti nell'intervallo tra l'eta x e $x+1$ per coloro che muoiono in tale intervallo. Questa quantità è ottenuta come $a_x = 1/2$ per $x = 0, \ldots, 109$ e $a_{110} = 1/m_{110}$ per $x = 110$;

- q_x, probabilità di morte all'età x per un individuo arrivato in vita all'età x che viene calcolata come

$$q_x = \frac{m_x}{1 + (1 - a_x) \cdot m_x}$$

  per $x = 0, \ldots, 109$ e $q_{110} = 1$ per $x = 110$;

- l_x, numero di sopravviventi all'età $x > 0$ partendo da una popolazione fittizia di $l_0 = 10^5$ persone, che vengono ottenuti come $l_x = l_0 \cdot \prod_{i=0}^{x-1}(1 - q_i)$;

- d_x, distribuzione del numero di decessi per età ottenuti come $d_x = l_x \cdot q_x$ per $x = 0, \ldots, 109$ e $d_{110} = l_{110}$ per $x = 110$. Si avrà quindi che $\sum_{x=0}^{110} d_x = l_0$;

- d_x^*, distribuzione normalizzata del numero di decessi per età, ottenuti come $d_x^* = d_x / \sum_{x=0}^{110} d_x$ e perciò si ha che $\sum_{x=0}^{110} d_x^* = 1$.

Esistono delle precise relazioni che permettono il passaggio da una quantità ad un'altra in maniera biunivoca, per questo motivo i modelli elaborati per lo studio della mortalità nel corso degli anni utilizzano come quantità di interesse il tasso di mortalità o la probabilità di morte o il numero dei decessi. Sebbene le altre funzioni contengano la medesima informazione, può rivelarsi utile lavorare con la distribuzione del numero di decessi per età poiché queste

evidenziano meglio alcune caratteristiche della mortalità sulla popolazione, come ad esempio l'età modale alla morte, il suo spostamento in avanti e la compressione della curva attorno ad essa. Ci si riferisce alla distribuzione normalizzata dei decessi per età con il termine "curva di mortalità". Questa possiede tipicamente due o tre punti di massimo. Dall'esempio riportato in Figura 1.1 si possono cogliere tre caratteristiche fondamentali della curva e, più in generale, del fenomeno della mortalità:

1. la mortalità infantile e nei primi anni di età, identificata con un picco iniziale della curva;

2. la mortalità prematura, identificata con l'emergere di una gobba nella curva a partire dai 15 anni.

3. la mortalità adulta, identificata con il massimo in corrispondenza di età avanzate.

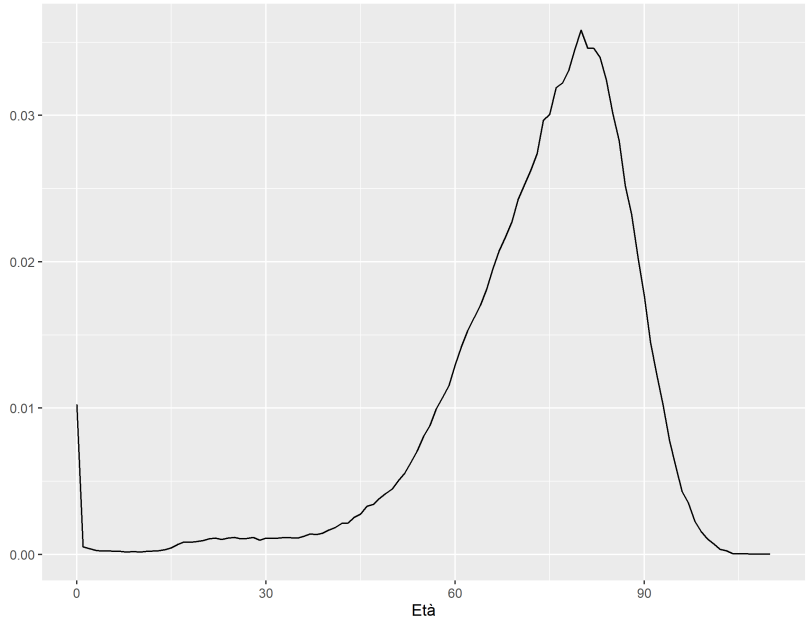


Figura 1.1: Curve di mortalità per la popolazione maschile italiana nell'anno 1988.

1.2 Alcuni modelli parametrici

Lo sviluppo di modelli per la mortalità umana nel corso degli anni ha rappresentato una sfida sia per gli statistici che per studiosi di altre discipline, perciò si dispone in letteratura di una vasta gamma di alternative.

Il modello di proposto da Siler (1979) rappresenta una generalizzazione del modello di Gompertz-Makeham (Makeham, 1860). Quest'ultimo prendeva in considerazione solamente la mortalità ad età avanzata, mentre il modello di Siler permette di cogliere sia la mortalità ad età avanzata che la mortalità infantile e nei primi anni di vita. Tale modello rappresenta il tasso di mortalità all'età x come segue:

$$m_x = a_1 \exp\{-b_1 x\} + a_2 + a_3 \exp\{b_3 x\},$$

dove a_1 è l'intensità della mortalità infantile, b_1 è la velocità con cui la mortalità infantile diminuisce all'aumentare dell'età, a_3 è l'intensità della mortalità ad età avanzata e b_3 è il tasso di incremento di questa all'aumentare dell'età.

Un'alternativa al modello di Siler è il modello proposto da Heligman e Pollard (1980) che tiene conto di tutte e tre le componenti della mortalità. In questo caso vengono modellate le quote delle probabilità di morte all'età x:

$$\frac{q_x}{1-q_x} = A^{(x-B)^C} + D \exp\left\{ - E \left(\log \frac{x}{F} \right)^2 \right\} + GH^x.$$

Si può notare come la mortalità viene decomposta in tre parti, ed ognuno dei parametri ha anche una propria interpretazione. La prima parte si riferisce alla mortalità infantile e nei primi anni di vita, con A che rappresenta il tasso di mortalità infantile, B che rappresenta il tasso di mortalità per i bambini che hanno compiuto un anno di vita e C che è relativo al declino del tasso di mortalità dopo il primo anno di età. La seconda parte è associata con la mortalità prematura, con D che ne rappresenta la severità, E la diffusione e F la collocazione tra le età. La terza ed ultima parte è associata alla mortalità in età avanzata, con G che rappresenta il livello base di tale componente della

mortalità e H il tasso di incremento della mortalità all'aumentare dell'età. Infine, un terzo modello preso in esame è quello proposto da Mazzuco, Scarpa e Zanotto (2018) che, a differenza dei due modelli parametrici visti in precedenza, modella direttamente i decessi al variare dell'età x. Infatti la curva di d_x^* può essere vista come una funzione di densità, quindi l'obiettivo diventa quello di trovare una funzione di densità che abbia un buon adattamento ai decessi presenti nella tavola di mortalità. Questa funzione di densità dovrà avere almeno due mode, una alla nascita ed una in età avanzata, a cui se ne può aggiungere una terza che permette di catturare la mortalità prematura. Si utilizza perciò una densità mistura le cui due componenti sono una *Half-Normal* ed una Normale asimmetrica bimodale:

$$f(x) = m \cdot \phi\left(\frac{x}{\sigma}\right) + (1-m) \cdot 2\omega^{-1} \left(\frac{1+\alpha((x-\xi)/\omega)^2}{1+\alpha}\right) \phi\left(\frac{x-\xi}{\omega}\right) \Phi\left\{\lambda\left(\frac{x-\xi}{\omega}\right)\right\}.$$

Poiché la densità di una Normale asimmetrica bimodale può essere riscritta a sua volta come una mistura di una Normale asimmetrica di una Normale asimmetrica modificata, la densità della distribuzione proposta risulta essere una mistura di tre componenti:

1. una *Half-Normal* per cogliere la mortalità infantile;
2. una Normale asimmetrica modificata per cogliere le morti premature;
3. una Normale asimmetrica per cogliere la mortalità ad età avanzata.

L'idea alla base di tale approccio tornerà utile nel Capitolo successivo.

1.3 I dati analizzati

1.3.1 La mortalità in Italia

Un insieme di dati presi in esame sono le curve di mortalità per età della popolazione maschile e femminile italiana per tutti gli anni dal 1872 al 2018.

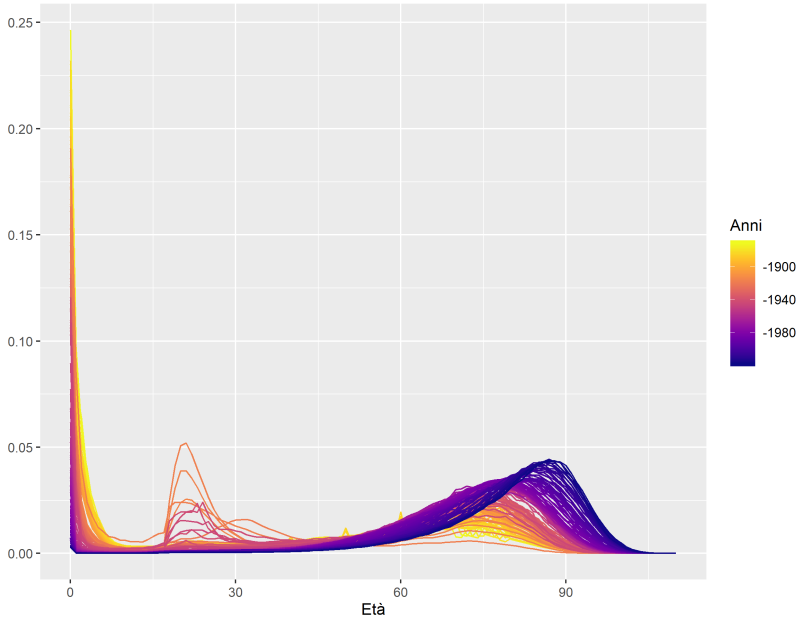


Figura 1.2: Curve di mortalità per la popolazione maschile italiana dal 1872 al 2018.

Questi dati sono stati ottenuti dallo *Human Mortality Database* (HMD, University of California, Berkeley e Max Plank Institute for Demographic Research, 2013), una base di dati creata per fornire dati dettagliati riguardanti le principali caratteristiche della mortalità in diversi paesi, ponendosi in particolar modo l'obiettivo di agevolare la ricerca su questi ambiti.

Le curve di mortalità per età sono state ottenute a partire dai decessi per classe di età contenuti nelle tavole di mortalità per ciascun anno. Ogni curva dei decessi aveva radice 10^5, quindi si è divisa ciascuna curva per tale valore ottenendo così delle curve di mortalità per età che sommano ad 1. Le distribuzioni che non sommavano precisamente ad 1 (ma a numeri di poco inferiori o di poco superiori a causa di arrotondamenti) sono state riscalate in modo tale da sommare esattamente ad 1. Le curve ottenute quindi rappresentano la proporzione di persone decedute nelle 111 classi di età previste, da $[0, 1)$ a $[110, \infty)$, nell'arco di un anno, cioè quelle che precedentemente sono state definite come distribuzioni del numero di decessi per età normalizzate.

Si considerano separatamente le curve per gli uomini e quelle per le donne in

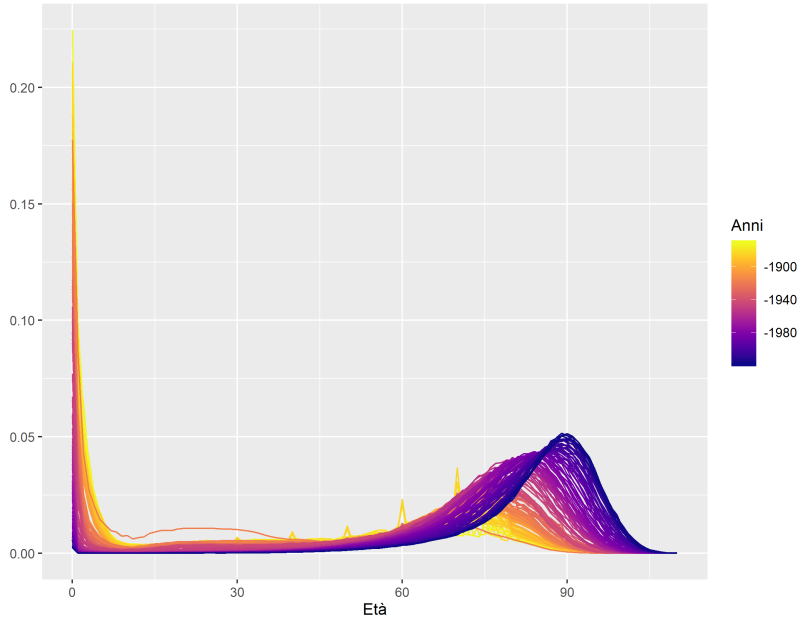


Figura 1.3: Curve di mortalità per la popolazione femminile italiana dal 1872 al 2018.

quanto vi sono sostanziali differenze di forma a parità di anno. Per quanto riguarda la popolazione maschile è immediato vedere dalla Figura 1.2 come all'inizio del periodo considerato la mortalità infantile arrivasse anche a valori come 0.25, mentre col passare degli anni questa ha iniziato progressivamente a decrescere e l'età modale alla morte si è sempre più spostata in avanti. Ciò è dovuto al miglioramento delle condizioni igienico-sanitarie, ai progressi scientifici che hanno avuto risvolti nella vita comune delle persone e al cambiamento dei regimi alimentari. Una diminuzione ha riguardato nel corso degli anni anche la mortalità accidentale tipica della popolazione maschile, ad eccezione degli della Prima e della Seconda Guerra Mondiale.

Per la popolazione femminile, i dati in Figura 1.3 rispecchiano grossomodo la stessa dinamica descritta nel caso maschile. In questo caso però si raggiunge un'età modale più elevata, che arriva fino quasi ai 90 anni nel 2018. Ciò è conseguenza del fenomeno, noto come "supermortalità" maschile, per il quale a quasi tutte le età la mortalità maschile risulta maggiore di quella femminile. È possibile notare anche delle curve decisamente frastagliate (in giallo nella

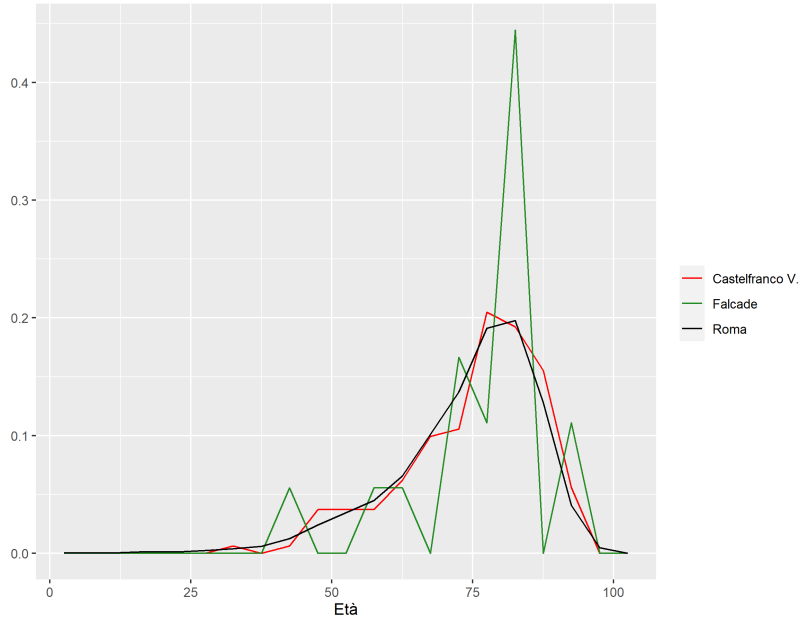


Figura 1.4: Curve di mortalità per la popolazione maschile residente nei comuni di Roma (in nero), Castelfranco Veneto (in rosso) e Falcade (in verde) per l'anno 2020.

Figura 1.2) che corrispondono agli anni 1883 e 1884. La loro forma potrebbe essere causata dal processo di raccolta dei dati.

1.3.2 La mortalità per comune nel 2020

Un secondo insieme di dati utilizzati sono le curve di mortalità per età della popolazione maschile e femminile nell'anno 2020 per ciascuno dei 7903 comuni italiani esistenti. Questi dati sono stati ottenuti a partire dai dataset con i decessi giornalieri in ogni singolo comune di residenza per sesso e classi di età quinquennali pubblicati dall'Istituto nazionale di Statistica (2021) nel proprio sito web. Le curve ottenute quindi rappresentano il nuemro di persone decedute nelle 21 classi di età quinquennali, da $[0, 5)$ a $[100, \infty)$, nell'arco dell'anno 2020 in ciascun comune italiano.

A differenza delle precedenti, le curve in questione, avendo classi di età quinquennali, risultano nel complesso meno lisce. Tali curve vengono prese in

considerazione principalmente perché contengono le informazioni relative ai decessi dovuti alla pandemia di COVID-19 che ha colpito l'Italia nel 2020. Inoltre sono su base comunale, quindi si avrà che curve di mortalità appartenenti a comuni differenti possono avere forma e caratteristiche molto diverse fra loro. Ad esempio possiamo vedere in Figura 1.4 come la curva di mortalità per la popolazione maschile di un comune molto popoloso come quello di Roma risulti liscia e molto simile alle curve di mortalità dell'intera popolazione maschile italiana. La curva per un comune non capoluogo del Veneto come Castelfranco Veneto (TV) avente più di 33000 abitanti (Istituto nazionale di Statistica, 2020) invece risulta ancora abbastanza regolare, anche se con qualche increspatura in più, mentre la curva di mortalità per un comune montano con meno di 2000 abitanti (Istituto nazionale di Statistica, 2020) come Falcade (BL) è decisamente irregolare. Questo tipo di irregolarità nei dati sono attribuibili al caso e non sembra sensato attribuire loro un significato strutturale.

Capitolo 2

Il processo di Dirichlet

In questo capitolo si definisce il processo di Dirichlet e se ne descrivono le principali proprietà. In ambito bayesiano, la distribuzione di Dirichlet come distribuzione a priori gode della proprietà di coniugazione con i parametri della distribuzione multinomiale. A partire da questa proprietà, Ferguson (1973) ha introdotto il processo di Dirichlet, ossia una distribuzione di probabilità nello spazio delle misure di probabilità che induce una distribuzione di Dirichlet finito-dimensionale quando i dati sono raggruppati. L'attenzione a tale processo è dovuta al fatto che è uno dei principali e più semplici strumenti bayesiani non parametrici per misure di probabilità casuali e sarà alla base dei modelli proposti nei capitoli successivi.

2.1 Definizione e proprietà

Definizione 1. Sia $(\Omega, \mathcal{B}, P)$ uno spazio di probabilità, dove $\mathcal{B}$ indica l'insieme di tutti i possibili sottoinsiemi dello spazio campionario Ω. La misura di probabilità P segue un processo di Dirichlet se per ogni partizione finita $\{B_1, \ldots, B_k\}$ di Ω la distribuzione congiunta di $P(B_1), \ldots, P(B_k)$ corrisponde ad una distribuzione di Dirichlet k-dimensionale con parametri

$\alpha P_0(B_1), \ldots, \alpha P_0(B_k)$, cioè

$$P(B_1), \ldots, P(B_k) \sim \text{Dir}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)) \tag{2.1}$$

dove P_0 indica la misura di base e $\alpha > 0$ è il parametro di precisione. Allora si scrive $P \sim \text{DP}(\alpha, P_0)$.

Per comprendere meglio il significato della misura di base e del parametro di precisione si possono considerare i momenti e la distribuzione marginale del processo di Dirichlet. Per un qualsiasi $B \in \mathcal{B}$ la distribuzione marginale di $P(B)$ segue una distribuzione Beta

$$P(B) \sim \text{Beta}(\alpha P_0(B),\ \alpha(1 - P_0(B))) \tag{2.2}$$

dalla quale seguono direttamente il valore atteso di P

$$\mathbb{E}(P(B)) = P_0(B) \tag{2.3}$$

e la varianza di P

$$\mathbb{V}(P(B)) = \frac{P_0(B)(1 - P_0(B))}{1 + \alpha}. \tag{2.4}$$

Dunque, il processo di Dirichlet è centrato nella misura di base, che infatti corrisponde al suo valore atteso, mentre il parametro di precisione entra nella varianza del processo, determinando quanto questo sia concentrato intorno alla misura di base.

Una definizione alternativa che può rendere più efficace la comprensione di come funziona il processo è quella proposta da Sethuraman (1994) nota come rappresentazione *stick-breaking*. Sia δ_y una distribuzione con unico punto di massa in y. Si ha allora che

$$P = \sum_{h=1}^{\infty} v_h \delta_{y_h} \tag{2.5}$$

segue un $\mathrm{DP}(\alpha, P_0)$ se $y_h \overset{\text{iid}}{\sim} P_0$ per $h = 1, \ldots, \infty$, e $v_h = z_h \prod_{j=1}^{h-1}(1 - z_j)$ con $z_h \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ per $h = 1, \ldots, \infty$. Si può vedere questa rappresentazione come l'azione di spezzettare un bastoncino (*stick*) di lunghezza unitaria. Si inizia spezzando il bastoncino nel punto $z_1 \sim \mathrm{Beta}(1, \alpha)$ e si assegna massa z_1 ad un punto casuale $y_1 \sim P_0$. La rimanente massa $(1 - z_1)$ varrà quindi divisa in corrispondenza della proporzione $z_2 \sim \mathrm{Beta}(1, \alpha)$ e la relativa massa $(1 - z_1)z_2$ è assegnata al punto casuale $y_2 \sim P_0$. Questi passaggi si ripetono infinite volte. Tale rappresentazione evidenzia la natura discreta del processo di Dirichlet e può risultare molto utile (con un opportuno troncamento) per generare il processo in algoritmi MCMC. Inoltre fa emergere ancora una volta il ruolo del parametro di precisione α: valori piccoli di α fanno rimanere ad ogni passo poca porzione di bastoncino per il passo successivo, ottenendo così una distribuzione più concentrata. Viceversa, valori grandi di α hanno l'effetto opposto, quindi conducono ad una distribuzione meno concentrata. Infine, tramite questa rappresentazione si riesce bene a cogliere il fatto che il processo di Dirichlet è discreto. Un'ulteriore proprietà consiste nel fatto che la distribuzione predittiva per una nuova osservazione proveniente da $P \sim \mathrm{DP}(\alpha, P_0)$ può essere scritta sequenzialmente:

$$y_1 \sim P_0$$

$$y_2 | P, y_1 \sim P$$

con

$$P | y_1 \sim \mathrm{DP}\left(\alpha + 1, \frac{\alpha}{\alpha + 1} P_0 + \frac{1}{\alpha + 1} \delta_{y_1}\right). \tag{2.6}$$

Ciò implica che, marginalizzando rispetto a P la distribuzione di una nuova osservazione condizionata alla precedente è

$$y_2 | y_1 \sim \frac{\alpha}{\alpha + 1} P_0(y_2) + \frac{1}{\alpha + 1} \delta_{y_1},$$

e ciò significa che y_2 assumerà lo stesso valore di y_1 con probabilità $\frac{1}{\alpha+1}$, mentre verrà generato casualmente da P_0 con probabilità $\frac{\alpha}{\alpha+1}$. Ripetendo

questa logica per ogni osservazione estratta da P si ottiene la procedura proposta da Blackwell, MacQueen et al. (1973) nota come generalizzazione dello schema ad urne di Polya:

$$y_n|y_1,\dots,y_{n-1} \sim \begin{cases} \delta_{y_j^*} & \text{con probabilità } \frac{n_j}{\alpha+n-1} \quad j=1,\dots,k \\ P_0 & \text{con probabilità } \frac{\alpha}{\alpha+n-1}, \end{cases} \tag{2.7}$$

dove k è il numero di valori distinti presenti in $y_1,\dots,y_{n-1}$, con $y_1^*,\dots,y_k^*$ questi valori distinti e $n_1,\dots,n_k$ le relative frequenze osservate. Si ha dunque che ogni osservazione y_i sarà uguale ad un valore precedentemente osservato y_j^* con probabilità proporzionale alla frequenza n_j con cui tale valore è stato osservato nel passato, mentre corrisponderà ad una nuova estrazione da P_0 con probabilità proporzionale ad α. Inoltre, poiché $y_1,\dots,y_n$ sono scambiabili (Bernardo, 1994), si osserva che la stessa descrizione può essere applicata a qualsiasi y_i date le $y_{-i} = (y_1,\dots,y_{i-1},y_{i+1},\dots,y_n)$, ottenendo

$$y_i|y_{-i} \sim \left(\frac{\alpha}{\alpha+n-1}\right)P_0(y_i) + \sum_{h=1}^{k^{(-i)}} \left(\frac{n_h^{(-i)}}{\alpha+n-1}\right)\delta_{y_h^{*(-i)}}$$

per $i=1,\dots,n$, dove $y_h^{*(-i)}$, $h=1,\dots,k^{(-i)}$, sono i distinti valori assunti da y_{-i} e $n_h^{(-i)} = \sum_{j\neq i} \mathrm{I}(y_j = y_h^{*(-i)})$.

Questa caratteristica di P può risultare molto utile per affrontare problemi di raggruppamento (*clustering*). Infatti, come conseguenza di questa struttura ad urne, il numero totale di nuove estrazioni da P_0 (inclusa la prima) è generalmente molto più piccolo di n. Ciò significa quindi che il numero di gruppi è molto minore di n. Le probabilità di nuove estrazioni ai passi $1,2,\dots,n$ sono rispettivamente $1, \frac{\alpha}{\alpha+1},\dots,\frac{\alpha}{\alpha+n-1}$ e perciò il numero di distinti valori, equivalente al numero di gruppi, K_n atteso è (Hjort et al., 2010, Sezione 2.2.2):

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} \simeq \alpha \log\frac{n}{\alpha} \quad \text{per } n\to\infty. \tag{2.8}$$

Si può inoltre ricavare la distribuzione esatta di K_n e le sue approssimazioni normali e Poisson (Antoniak, 1974).

2.2 Utilizzo in ambito bayesiano

Così come la distribuzione di Dirichlet finito-dimensionale usata come distribuzione a priori è coniugata con la verosimiglianza multinomiale, il processo di Dirichlet utilizzato come distribuzione a priori gode anch'esso di una proprietà di coniugazione per la stima della distribuzione ignota di osservazioni indipendenti e identicamente distribuite (i.i.d.). Più precisamente, se $y_i|P \overset{\text{iid}}{\sim} P,\ i = 1, \dots, n$, e viene assegnata come distribuzione a priori per P la misura di probabilità indotta dal processo di Dirichlet, indicata con $P \sim \text{DP}(\alpha, P_0)$, allora la distribuzione a posteriori di $P|y_1, \dots, y_n$ è (Hjort et al., 2010, Sezione 2.2.3)

$$P(B_1), \dots, P(B_k)|y_1, \dots, y_n \sim \text{Dir}(\alpha P_0(B_1) + n_1,\ \dots,\ \alpha P_0(B_k) + n_k) \quad (2.9)$$

con $B_1, \dots, B_k$ partizione dello spazio campionario e $n_h = \sum_{i=1}^{n} 1_{y_i \in B_h}$ per $h = 1, \dots, k$, che corrisponde ad un processo di Dirichlet:

$$P|y_1, \dots, y_n \sim \text{DP}\left(\alpha + n,\ \frac{\alpha P_0 + \sum_{i=1}^{n} \delta_{y_i}}{\alpha + n}\right).$$

In questo senso si parla di approccio bayesiano non parametrico in quanto si sta facendo inferenza non solamente su uno o più parametri, ma su un'intera distribuzione. Il valore atteso a posteriori risulta

$$\mathbb{E}(P(B)|y_1, \dots, y_n) = \frac{\alpha}{\alpha + n} P_0(B) + \frac{n}{\alpha + n} \sum_{i=1}^{n} \frac{1}{n} \delta_{y_i}.$$

Dunque il valore atteso a posteriori può essere interpretato come una media pesata tra la media a priori P_0, a cui è assegnato peso proporzionale ad α, e la distribuzione empirica $\sum_{i=1}^{n} \frac{1}{n} \delta_{y_i}$, alla quale è assegnato peso proporzionale a n.

In particolare, una situazione in cui il processo di Dirichlet può risultare molto utile è quando si hanno delle osservazioni x_i indipendenti ed identicamente distribuite (iid) provenienti da una variabile casuale continua avente densità f e l'obiettivo è quello di ottenere una stima bayesiana di tale densità. Partendo dal presupposto che il modo più semplice per stimare una densità è quello di usare un istogramma, Gelman et al. (2013) descrivono come si può sviluppare una versione più flessibile di istogramma che giustifica l'adozione di un metodo bayesiano non parametrico per la stima della densità.

Si assume l'esistenza di nodi $l = (l_0, l_1, \ldots, l_k),\ l_0 < l_1 < \ldots < l_{k-1} < l_k$ e tali che $x_i \in [l_0, l_k]$. Un modello per la stima della densità che è analogo ad usare un istogramma è il seguente:

$$f(x) = \sum_{h=1}^{k} 1_{x \in (l_{h-1}, l_h]} \ \frac{\pi_h}{l_h - l_{h-1}}, \quad x \in \mathbb{R}$$

con $\pi = (\pi_1, \ldots, \pi_k)$ ignoto vettore di probabilità. A questo punto, si potrebbe completare la specificazione bayesiana del modello indicando come distribuzione a priori per il parametro π una distribuzione di Dirichlet con iperparametri $(a_1, \ldots, a_k)$. Sfruttando la proprietà di coniugazione tra la distribuzione di Dirichlet a priori ed il nucleo della densità di $f(x)$ proporzionale a quello di una distribuzione multinomiale, si otterrebbe una distribuzione a posteriori per π che segue ancora una distribuzione di Dirichlet con parametri aggiornati. Questo stimatore riesce ad approssimare in maniera adeguata la vera densità con una notevole semplicità computazionale dovuta alla coniugazione.

Tuttavia, tale approccio ha lo svantaggio che i risultati dipendono dal numero e dalla posizione dei nodi l. Una soluzione a questo problema è quella di utilizzare come distribuzione a priori per π un processo di Dirichlet, corrispondente alla (2.1), che permette di specificare la distribuzione a priori per qualsiasi partizione $B_1, \ldots, B_k$ di $\mathbb{R}$. Così facendo, si specifica solamente che a priori all'intervallo B_k è assegnata probabilità $P(B_k)$ e non si specifica come tale probabilità sia distribuita all'interno dell'intervallo. L'idea alla base

è quindi quella di eliminare la sensibilità dovuta alla scelta della partizione (cioè al numero e alla posizione dei nodi) ed invece indurre una distribuzione a priori valida per tutte le possibili partizioni $B_1, \ldots, B_k$ di $\mathbb{R}$ e per qualsiasi k. Grazie alla proprietà di coniugazione (2.9) si ottiene in maniera automatica anche la distribuzione a posteriori per π. Vista la natura dei dati, questo approccio può rivelarsi molto appropriato per problemi di stima bayesiana di curve di mortalità.

Capitolo 3

Modello per la stima della densità

In questo capitolo si espone un modello basato sull'idea di considerare la distribuzione del numero di decessi per età $d_x,\ x = 0, \ldots, 110$, descritta nella Sezione 1.1 come una realizzazione proveniente da una variabile casuale con distribuzione multinomiale. Tale assunzione permette di utilizzare come distribuzione a priori per il parametro di tale densità una misura di probabilità indotta da un processo di Dirichlet. Il modello proposto gode quindi di una certa flessibilità che gli consente di adattarsi bene a curve di mortalità con forme molto diverse o particolarmente irregolari. Si descrive poi l'algoritmo di stima di tale modello ed infine si commentano i risultati della sua applicazione su dati reali.

3.1 Il modello

Siano $x_i, i = 1, \ldots, n$, le età esatte alla morte del soggetto i-esimo. Allora il modello che si vuole proporre è il seguente

$$x_i|\tilde{p} \overset{\text{iid}}{\sim} \tilde{p} \tag{3.1}$$

e la conoscenza a priori su $\tilde{p}$ è espressa come

$$\tilde{p}|\theta \sim \text{DP}(\alpha, P_0(\cdot;\theta)). \tag{3.2}$$

Si sta quindi affermando che le età esatte alla morte di ciascun individuo provengono dalla misura di probabilità indotta da un processo di Dirichlet $\tilde{p}$ con misura di base che dipende dal parametro θ.
Tuttavia, è bene sottolineare come per "età esatta alla morte" si intende una informazione "ideale", più dettagliata rispetto all'età alla morte misurata in anni che misura l'esatto istante della morte di ciascun individuo. Si tratta perciò di una partizione più fine dell'informazione contenuta nelle tradizionali tavole di mortalità e di conseguenza nelle curve di mortalità, le quali invece contengono invece la categorizzazione delle x_i in classi di età sottoforma di conteggi.
Questo modello, dunque, non può essere utilizzato direttamente per curve di mortalità. Occorre formulare un modello per dati aggregati. Consideriamo allora i conteggi $d_0, \ldots, d_{110}$ definiti come

$$d_j = \sum_{i=1}^{n} \mathrm{I}(x_i \in [j, j+1))$$

per $j = 0, \ldots, 109$ e $d_{110} = \sum_{i=1}^{n} \mathrm{I}(x_i \in [110, \infty))$ con $n = \sum_{j=0}^{110} d_j$ e dove $\mathrm{I}(\cdot)$ è la funzione indicatrice. Questi dati aggregati corrispondono alla distribuzione del numero di decessi per età descritta nella Sezione 1.1. Si avrà quindi che n è pari alla numerosità della popolazione, eventualmente fittizia, iniziale.
Sulla base di tale aggregazione, è possibile considerare $d_0, \ldots, d_{110}|\tilde{p}$ come una realizzazione di una variabile casuale con distribuzione multinomiale:

$$d_0, \ldots, d_{110}|\tilde{p} \sim \mathrm{Multinom}(n, \pi) \tag{3.3}$$

con $\pi = (\pi_0, \ldots, \pi_{110})$, $\sum_{j=0}^{110} \pi_j = 1$, $\pi_j \in [0, 1]$, vettore di parametri ignoti che indicano le probabilità di morte alle varie classi di età. Più precisamente, π_j, $j = 0, \ldots, 109$, rappresenta la probabilità che la morte di un individuo avvenga tra l'età j e l'età $j + 1$, mentre π_{110} rappresenta la probabilità di morte dopo il centodecimo anno di età.

A partire dalla distribuzione a priori definita per $\tilde{p}$ nella (3.2) è possibile ottenere in maniera diretta la distribuzione a priori per π. Infatti la conoscenza a priori su ciascun $\pi_j,\ j = 0, \ldots, 110$, corrisponde alla probabilità assegnata dalla misura di probabilità indotta da $\tilde{p}$ nell'intervallo $[j, j+1)$, esprimibile con la seguente notazione:

$$\pi_j = \tilde{P}([j, j+1)) \tag{3.4}$$

per $j = 0, \ldots, 109$, dove $\tilde{P}(\cdot)$ rappresenta la misura di probabilità indotta da un processo di Dirichlet $\tilde{p}$, e $\pi_{110} = \tilde{P}([110, \infty))$. Con questo modello, avendo dei dati con meno informazione dei dati "ideali", si andrà a fare inferenza solamente su π non su $\tilde{p}$. La conoscenza a priori su π sarà quindi espressa come

$$\pi_0, \ldots, \pi_{110} | \theta \sim \text{Dir}(\alpha P_0([0,1);\theta), \ldots, \alpha P_0([110, \infty);\theta))$$

con misura di base

$$P_0([j, j+1);\theta) = \int_j^{j+1} f(z;\theta)dz$$

per $j = 0, \ldots, 109$ e $P_0([110, \infty);\theta) = \int_{110}^{\infty} f(z;\theta)dz$. In questo caso, come $f(\cdot;\theta)$ si sceglie di utilizzare una funzione di densità che abbia caratteristiche simili a quelle di alcuni modelli utilizzati per lo studio delle curve di mortalità. Sulla base di Aliverti, Mazzuco e Scarpa (2021) e Mazzuco, Scarpa e Zanotto (2018), coerentemente con quanto spiegato nella Sezione 1.2, si decide di usare una densità corrispondente ad una mistura di tre componenti:

1. una *Half-Normal* per cogliere la componente di mortalità infantile;
2. una Normale per cogliere le morti premature;
3. una Normale asimmetrica per cogliere la mortalità ad età anziana.

La densità di f sarà perciò:

$$f(x;\theta) = \psi_0 2\phi\left(\frac{x}{\gamma}\right) + \psi_1 \phi\left(\frac{x-\mu}{\sigma}\right) + \psi_2 \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\frac{x-\xi}{\omega}\right) \tag{3.5}$$

per $x \geq 0$, dove $\phi(\cdot)$ e $\Phi(\cdot)$ indicano rispettivamente la funzione di densità e di ripartizione di una normale standard, $\psi_0 = 1 - \psi_1 - \psi_2$ e

$$\theta = (\psi_1, \psi_2, \gamma, \mu, \sigma, \xi, \omega, \lambda)$$

con $\psi_1, \psi_2 \in [0,1]$ tali che $\sum_{j=0}^{2} \pi_j = 1,\ \sigma, \omega > 0,\ \mu, \xi, \lambda \in \mathbb{R}$. Poiché la presenza di questi vincoli può causare difficoltà nell'implementazione dell'algoritmo MCMC, si decide di utilizzare una riparametrizzazione che elimina le restrizioni allo spazio parametrico. Si definisce allora

$$\tau = \left(\log \frac{\pi_1}{1 - \pi_1 - \pi_2}, \log \frac{\pi_2}{1 - \pi_1 - \pi_2}, \log \gamma, \mu, \log \sigma, \xi, \log \omega, \lambda \right).$$

Si decide infine di porre una distribuzione a priori su τ che sia poco informativa. Per ogni elemento di τ si utilizza come distribuzione a priori una normale con deviazione standard pari a 10. Le otto distribuzioni a priori vengono considerate tra loro indipendenti. La scelta conservativa di utilizzare delle distribuzioni a priori poco informative è dovuta al fatto che si vuole che questo algoritmo produca dei risultati validi per curve che hanno forme molto diverse fra loro e che l'inferenza sia determinata dai dati osservati.

3.2 L'algoritmo

Per la stima dei parametri π e θ del modello si utilizza un algoritmo MCMC ibrido, in quanto alterna un passo di *Gibbs-sampling* per la stima di π ad uno di *Metropolis-Hastings* per quella di τ. Sia

$$L(\pi_0, \ldots, \pi_{110}) \propto \prod_{j=0}^{110} \pi_j^{d_j} \tag{3.6}$$

la funzione di verosimiglianza per una curva di mortalità che segue il modello definito in (3.3), sia

$$\pi | \theta(\tau) \sim \mathrm{Dir}(\alpha P_0([0,1); \theta(\tau)), \ldots, \alpha P_0([110, \infty); \theta(\tau))) \tag{3.7}$$

con $\theta(\tau)$ che indica il parametro θ scritto nella riparametrizzazione τ e sia $h(\tau)$ la distribuzione a priori per τ definita come descritto in precedenza. I passi dell'algoritmo sono allora i seguenti:

1. **Passo di *Gibbs-sampling* per la stima di $\boldsymbol{\pi}$.**

$$\pi|- \sim \text{Dir}(d_0 + \alpha P_0([0,1);\theta(\tau)), \ldots, d_{110} + \alpha P_0([110,\infty);\theta(\tau))) \quad (3.8)$$

ricavata grazie alla proprietà di coniugazione tra (3.7) e (3.6). Da questa relazione è possibile evidenziare una caratteristica chiave di questo modello: la distribuzione a posteriori su π è una distribuzione di Dirichlet i cui parametri dipendono dalla misura di probabilità indotta dalla misura di base e dai valori $d_0, \ldots, d_{110}$ osservati.

2. **Passo di *Metropolis-Hastings* per la stima di $\boldsymbol{\tau}$.**

$$p(\tau|-) \propto g(\pi|\theta(\tau))h(\tau)$$

dove $g(\pi|\theta(\tau))$ è la densità della distribuzione a priori per $\pi|\theta(\tau)$ definita in (3.7) che corrisponde a

$$g(\pi|\theta(\tau)) = \frac{\Gamma(\sum_{j=0}^{110} a_j)}{\prod_{j=0}^{110}\Gamma(a_j)} \prod_{j=0}^{110} \pi_j^{a_j - 1}$$

con $a_j = \alpha P_0([j, j+1);\theta(\tau))$ e dove $\Gamma(\cdot)$ è la funzione Gamma. Si noti che in questo caso $\theta(\tau)$ entra anche nel rapporto tra funzioni Gamma, quindi questo non può essere semplificato.

3.3 Alcuni commenti

Una volta espressa la distribuzione a posteriori marginale per π (3.8), è possibile fare qualche considerazione su come, grazie alla struttura di tale distribuzione, il modello gode di una certa flessibilità. Infatti a seconda del

valore che assume il parametro di precisione α il modello dà maggiore o minore importanza all'informazione che proviene dalle osservazioni. Il vantaggio principale di utilizzare il modello proposto anziché stimare direttamente il modello parametrico utilizzato come misura di base definito nella (3.5) sta proprio in questa maggiore flessibilità. In questo contesto si ha che per valori piccoli di α la distribuzione a posteriori di π dipende maggiormente dalla curva osservata piuttosto che dalla misura di base e ciò potrebbe portare il modello a sovradattarsi alla curva che si sta stimando. Al contrario, per valori molto grandi di α la (3.8) sarebbe più spostata verso la misura di base parametrica anziché verso le informazioni provenienti dalle osservazioni, con il rischio che così il modello non riesca a cogliere particolarità nelle forme delle curve. Tuttavia, quest'ultima caratteristica garantisce al modello la possibilità di cogliere andamenti peculiari della curva osservata senza però seguire troppo il rumore presente nelle osservazioni. In questo caso il valore del parametro di precisione α è stato fissato pari a 10^5, cioè pari a n. Tale scelta è da interpretare quindi come cautelativa nei confronti di entrambe queste situazioni estreme, infatti in questo modo viene assegnato uguale peso all'informazione proveniente dalla curva osservata e dalla misura di base.

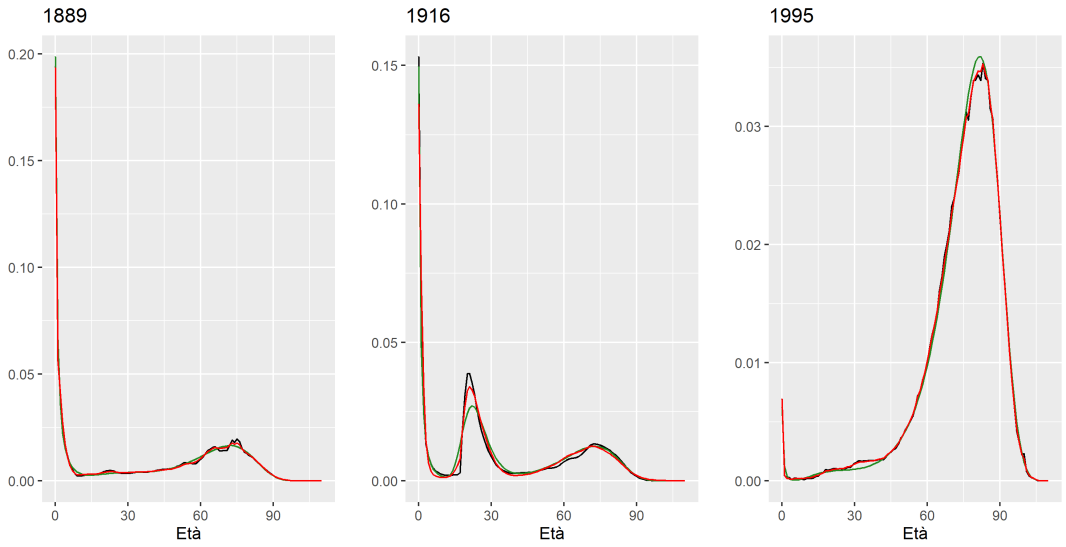


Figura 3.1: Curve di mortalità per la popolazione maschile italiana degli anni 1889, 1916 e 1995 (in nero) con la curva stimata dal modello proposto (in rosso) e quella stimata dal modello di Heligman-Pollard (in verde).

3.4 Applicazione: alcune curve di mortalità

Una volta implementato l'algoritmo descritto tramite il software `R`, si decide di provare il modello su alcune curve di mortalità. Sono scelte le curve degli anni 1889, 1916 e 1995 per la popolazione italiana maschile e quelle degli anni 1880, 1935 e 1997 per la popolazione italiana femminile. La scelta di tre diversi anni per ogni sesso è dovuto è voluta per poter valutare le prestazioni del modello con curve aventi forme molto diverse tra loro. Per le curve relative alla popolazione maschile l'algoritmo è arrivato a convergenza dopo rispettivamente 10000, 38000 e 15000 iterazioni per gli anni 1880, 1916 e 1995, mentre per la popolazione femminile la convergenza è stata raggiunta dopo 20000 iterazioni per il 1880 e 10000 iterazioni per 1935 e 1995. In tutte e sei le situazioni le prima metà di iterazioni sono state scartate in quanto considerate periodo di *burn-in*. I *traceplot* per le stime degli elementi del parametro τ sono riportati nella Appendice A.

A queste sei curve viene adattato anche il modello di Heligman-Pollard uti-

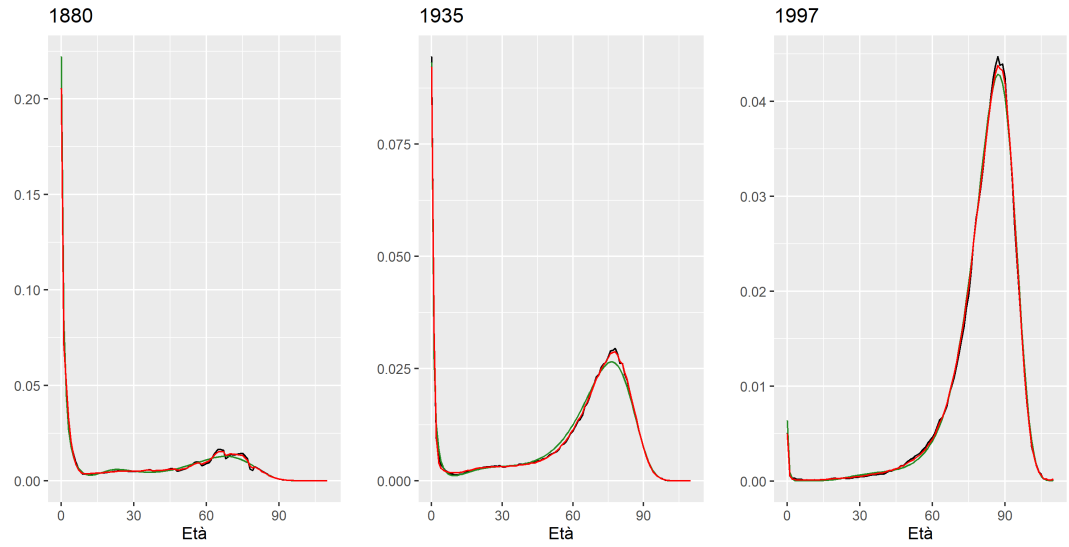


Figura 3.2: Curve di mortalità per la popolazione femminile italiana degli anni 1880, 1935 e 1997 (in nero) con la curva stimata dal modello proposto (in rosso) e quella stimata dal modello di Heligman-Pollard (in verde).

lizzando la libreria `HPbayes` (Sharrow, 2015) del software `R` seguendo quanto fatto da Emilidha e Danardono (2017). Le stime ottenute con questi due modelli vengono poi confrontate tramite errore quadratico medio.

I risultati per la popolazione maschile sono rappresentati in Figura 3.1. Possiamo subito notare come il modello proposto sembra adattarsi molto bene per gli anni 1889 e 1995, mentre l'adattamento è meno buono per l'anno 1916, in particolare per la moda in corrispondenza dei 20 anni di età dovuta alla morte dei soldati nella Prima Guerra Mondiale. Tuttavia il modello proposto, grazie alla sua flessibilità, sembra cogliere meglio del modello di Heligman-Pollard questo sbalzo. I termini di prestazioni previsive, come riportato nella Tabella 3.1, in tutti e tre i casi il modello proposto risulta migliore del modello di Heligman-Pollard in termini di errore quadratico medio. Nella Figura 3.2 sono rappresentate le stime ottenute per la popolazione femminile. Anche in questo caso il modello proposto sembra avere un adattamento migliore sia graficamente che in termini di errore quadratico medio (Tabella 3.1) in tutte e tre le situazioni.

Si può dunque concludere che il modello proposto in tutte le situazioni prese

	Modello proposto	**Heligman-Pollard**
Maschi - 1889	2756.6	19949.7
Maschi - 1916	69062.2	88755.6
Maschi - 1995	278.2	3093.3
Femmine - 1880	4139.5	52884.2
Femmine - 1935	1462.5	9619.9
Femmine - 1997	723.6	3360.3

Tabella 3.1: Errore quadratico medio per le stime dei d_x (per una popolazione fittizia di 10^5) ottenute con il modello proposto e con quello di Heligman-Pollard nei sei casi considerati.

in esame conduca a dei risultati migliori in termini di adattamento rispetto al modello di Heligman-Pollard. Va sottolineato che queste prestazioni sono da attribuire alla maggiore flessibilità del modello proposto dovuta alla sua struttura non parametrica. Tale caratteristica permette al modello di cogliere andamenti particolari ed irregolari specifici della curva. Dall'altra parte la stima dei parametri della misura di base ed il valore del parametro di precisione α pari a 10^5, che attribuisce pari peso all'informazione proveniente dai dati osservati e a quella proveniente dal modello parametrico scelto come misura di base del processo di Dirichlet, fa in modo che le stime ottenute dal modello non tendano a seguire andamenti troppo irregolari della curva dovuti al rumore presente nei dati.

Capitolo 4

Estensione del modello nel caso di più curve

Si può estendere il modello esposto nel Capitolo 3 al caso in cui non si stia considerando solamente una curva di mortalità bensì un insieme di queste. Tale estensione permette di ottenere una stima per ciascuna curva ma, come effetto accessorio, anche quello di individuare nella popolazione dei gruppi di curve con caratteristiche comuni. Curve appartenenti allo stesso gruppo avranno perciò la stessa stima. Dopo aver specificato il modello, se ne descrive l'algoritmo di stima ed infine si commentano i risultati dell'applicazione sulle curve di mortalità italiane dal 1872 al 2018.

4.1 Il modello

Si considerino J curve di mortalità per età rappresentate nella matrice D:

$$D = \begin{bmatrix} d_0^{(1)} & \dots & d_{110}^{(1)} \\ \vdots & & \\ d_0^{(J)} & \dots & d_{110}^{(J)} \end{bmatrix}$$

nella quale l'elemento $d_l^{(j)}$ indica il numero di morti tra l'età l e $l+1$ per la j-esima curva di mortalità, $j = 1, \ldots, J$. Si ha quindi che ciascuna riga della matrice D corrisponde ad una distribuzione del numero di decessi per età, seguendo la definizione data nella Sezione 1.1.
Il modello che si vuole adattare a questi dati è un'estensione di quello esposto nel Capitolo 3 valida nel caso in cui si stia considerando più di una curva di mortalità. Anche in questo caso l'obiettivo primario rimane quello di fornire una stima della densità delle curve, ma a questo si aggiunge come sottoprodotto l'individuazione di gruppi di curve che hanno caratteristiche simili fra loro. Sembra ragionevole infatti che curve aventi forme simili possano avere una stima della densità comune.
L'idea alla base è quella di pensare che all'interno della popolazione formata dalle J curve di mortalità siano presenti al più H gruppi latenti. In fase di stima dal modello, per H viene scelto un valore molto grande, maggiore del numero di gruppi che ci si aspetta di trovare nella popolazione. Ogni curva di mortalità può intendersi come generata da uno di questi gruppi e curve provenienti dallo stesso gruppo avranno caratteristiche simili. La popolazione di curve può dunque essere rappresentata come una mistura di distribuzioni multinomiali:

$$d_0^{(j)}, \ldots, d_{110}^{(j)} \overset{\text{iid}}{\sim} \sum_{h=1}^{H} w_h M_h \tag{4.1}$$

per $j = 1, \ldots, J$, con

$$M_h \sim \text{Multinom}(\pi_{0h}, \ldots, \pi_{110h})$$

per $h = 1, \ldots, H$. I parametri $w_1, \ldots, w_H$ corrispondono alle probabilità della mistura perciò si ha che $w_h \in [0, 1],\ h = 1, \ldots, H,\ \sum_{h=1}^{H} w_h = 1$.

Il modello in (4.1) può essere riscritto come:

$$d_0^{(j)}, \ldots, d_{110}^{(j)} | G_j = h \sim \text{Multinom}(\pi_{0h}, \ldots, \pi_{110h}) \tag{4.2}$$

$$G_j \sim \text{Cat}(w_1, \ldots, w_H) \tag{4.3}$$

per $j = 1, \ldots, J$, con $G_j = h,\ h = 1, \ldots, H$, che indica che la curva j appartiene al gruppo h. Si ha quindi che ogni curva dato il gruppo a cui appartiene si distribuisce come una multinomiale con parametri che sono uguali per tutte le curve di quel gruppo. Con la notazione Cat della (4.3) si intende che G_j, che indica il gruppo a cui appartiene al curva j, è una variabile aleatoria categoriale che può assumere valori $1, \ldots, H$ con probabilità rispettivamente $w_1, \ldots, w_H$.
A questo punto, per il parametro $w = (w_1, \ldots, w_H)$ viene utilizzata come distribuzione a priori una distribuzione di Dirichlet non informativa

$$w_1, \ldots, w_H \sim \text{Dir}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right) \tag{4.4}$$

come suggerito da Rousseau e Mengersen (2011). Infine, come nel modello per singola curva presentato in precedenza, i parametri della distribuzione multinomiale per ciascun cluster $\pi_h = (\pi_{0h}, \ldots, \pi_{110h}),\ h = 1, \ldots, H$, vengono intesi come derivanti dalla relazione (3.4) con la misura di probabilità indotta da un processo di Dirichlet che genera dei dati "ideali" (cioè quelli definiti nella Sezione 3.1). Si imposta quindi

$$\pi_{0h}, \ldots, \pi_{110h} \sim \text{Dir}(a_0, \ldots, a_{110}) \tag{4.5}$$

per $h = 1, \ldots, H$, con $a_x = \alpha \int_x^{x+1} f(t|\theta)dt$ dove α rappresenta il parametro di precisione del processo di Dirichlet e $f(\cdot)$ è la misura di base definita nella (3.5). In questo caso, però, per semplicità computazionale $a_0, \ldots, a_{110}$ vengono supposti noti ed uguali per tutti i gruppi.

4.2 La distribuzioni a posteriori e l'algoritmo

Una volta specificato il modello si può ricavare la distribuzione a posteriori per i parametri $w = (w_1, \ldots, w_H)$ e

$$\pi = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_H \end{bmatrix} = \begin{bmatrix} \pi_{0,1} & \ldots & \pi_{110,1} \\ \vdots & & \\ \pi_{0,H} & \ldots & \pi_{110,H} \end{bmatrix}. \tag{4.6}$$

Dalla (4.2) si ricava la seguente funzione di verosimiglianza

$$L(w, \pi) \propto w_1^{m_1} \cdot \ldots \cdot w_H^{m_H} \prod_{h=1}^{H} \left(\pi_{0h}^{s_{0h}} \cdot \ldots \cdot \pi_{110h}^{s_{110h}} \right)$$

con $s_{xh} = \sum_{j:G_j=h} d_x^j,\ x = 0, \ldots, 110$, e $m_h = \sum_{j=1}^{J} \mathrm{I}(G_j = h),\ h = 1, \ldots, H$, dove con la notazione $\{j : G_j = h\}$ si intendono tutte le curve appartenenti al gruppo h, per $j = 1, \ldots, J$. Con questa e le densità delle distribuzioni a priori definite in (4.4) e (4.5) si ottiene il nucleo della distribuzione a posteriori per w e π:

$$p(w, \pi|D) \propto w_1^{m_1 + \frac{1}{H} - 1} \cdot \ldots \cdot w_H^{m_H + \frac{1}{H} - 1} \prod_{h=1}^{H} \left(\pi_{0h}^{s_{0h} + a_0 - 1} \cdot \ldots \cdot \pi_{110h}^{s_{110h} + a_{110} - 1} \right).$$

Per la stima di w e π si utilizza un algoritmo MCMC che alterna tre passi di *Gibbs-sampling*:

1. **Aggiornamento della composizione dei gruppi**.
   Ad ogni curva j viene assegnato un gruppo G_j di appartenenza. Le probabilità che G_j assuma valori $1, \ldots, H$ sono aggiornate come segue:

   $$\Pr(G_j = h|-) \propto w_h \cdot \pi_{0h}^{d_0^{(j)}} \cdot \ldots \cdot \pi_{110h}^{d_{110}^{(j)}}$$

   per $h = 1, \ldots, H$ e $j = 1, \ldots, J$. L'idea è che la probabilità che la curva j appartenga al gruppo h è elevata se le frequenze osservate per quella

curva "concordano" con le probabilità che quel gruppo assegna ad ogni classe di età.

2. **Aggiornamento della stima di $\boldsymbol{w}$.**
   L'aggiornamento dei parametri della mistura w avviene sfruttando la proprietà di coniugazione di cui gode la distribuzione di Dirichlet con il nucleo di una verosimiglianza multinomiale:

$$w_1, \ldots, w_H|- \sim \text{Dir}\left(\frac{1}{H} + m_1, \ldots, \frac{1}{H} + m_H\right).$$

   Ciò significa che maggiore è il numero di curve appartenenti ad un gruppo, maggiore sarà il peso della rispettiva funzione di densità multinomiale all'interno della densità mistura per l'intera popolazione di curve.

3. **Aggiornamento della stima di $\boldsymbol{\pi}$.**
   Le probabilità di morte alle diverse classi di età vengono aggiornate per ogni gruppo sfruttando la stessa coniugazione del passo precedente:

$$\pi_{0h}, \ldots, \pi_{110h}|- \sim \text{Dir}(s_{0h} + a_0, \ldots, s_{110h} + a_{110})$$

   per $h = 1, \ldots, H$. Quindi per ogni gruppo i parametri della distribuzione di Dirichlet dipenderanno dal numero e dal tipo di curve contenute in esso. Maggiore è il numero di curve, più le stime di π_h penderanno verso i valori osservati delle curve piuttosto che verso quelli degli iperparametri comuni fissati a priori.

4.3 Applicazione: la mortalità in Italia

Dopo aver implementato nel software `R` l'algoritmo descritto, il modello è stato adattato separatamente ai dati relativi alla popolazione maschile e femminile italiana descritti nel Capitolo 1.3.1. In entrambe le situazioni si è utilizzato come troncamento al limite superiore al numero di gruppi $H = 12$.

Ai parametri $w_1, \ldots, w_H$ è stato assegnato valore iniziale pari a 1/12 per ognuno degli H elementi, mentre per ogni riga della matrice π definita in (4.6) si utilizza come valore di partenza il vettore della media del numero di decessi normalizzato alle varie classi di età osservate. Il valore dell'iperparametro α, dopo alcune prove, è stato scelto essere 30000. Come valori noti ed uguali tra tutti i gruppi $a_0, \ldots, a_{110}$ si utilizza il vettore della media del numero di decessi normalizzato nelle 111 classi di età osservate moltiplicato per α. In entrambi i casi vengono effettuate dall'algoritmo MCMC 10000 iterazioni delle quali vengono scartate le prime 2500 in quanto considerate periodo di *burn-in*.
I risultati ottenuti dall'adattamento del modello sono poi confrontati con il tradizionale metodo di *clustering* k-medie (MacQueen et al., 1967).

4.3.1 Maschi

Per la popolazione maschile le curve di mortalità a disposizione sono rappresentate in Figura 1.2. In questo caso il modello individua otto gruppi, un numero inferiore al limite superiore fissato a dodici. Risulta evidente dalla Tabella 4.1 come ciascun gruppo individuato corrisponda ad una fase temporale ben definita e ciò mostra nitidamente come la distribuzione del numero di decessi per età cambi la sua forma nel corso del tempo per via del miglioramento dello stile di vita dovuto ai benefici del progresso tecnologico. L'unica eccezione è rappresentata dal gruppo 3 che comprende le curve corrispondenti agli anni delle due Guerre Mondiali, nelle quali la proporzione di morti fra i 20 e i 35 anni di età risulta nettamente superiore rispetto al resto del periodo preso in considerazione. Inoltre in questo gruppo è inserito anche l'anno 1919, che non è un anno di guerra, ma nel quale si verificò l'epidemia di influenza Spagnola (1918-19) che provocò un aumento della mortalità infantile che si aggiunse agli effetti della Prima Guerra Mondiale (Livi Bacci, 2020).
Per ogni gruppo viene inoltre fornita una stima della rispettiva riga della matrice π definita nella (4.6). Questi risultati sono presentati nella Figura 4.2

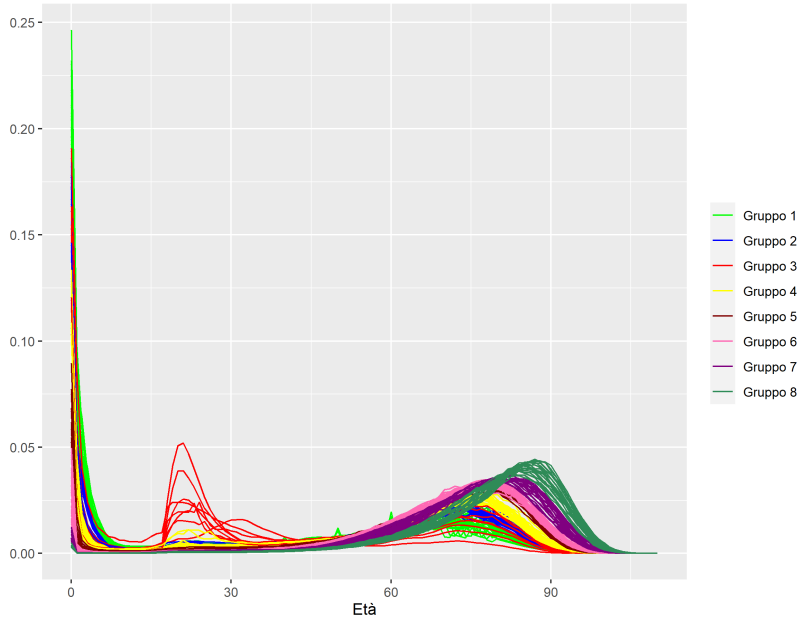


Figura 4.1: Distribuzioni del numero di decessi per età per la popolazione maschile italiana dal 1872 al 2018 ripartite negli otto gruppi individuati dal modello.

ed evidenziano un buon adattamento delle curve stimate per ciascun gruppo. I risultati ottenuti da questo con il modello proposto sono stati confrontati con quelli ottenuti tramite l'algoritmo di *clustering* k-medie. Poiché que-

	Anni
Gruppo 1	1872-95
Gruppo 2	1896-1913, 1920.21
Gruppo 3	1915-19, 1942-44
Gruppo 4	1914, 1922-41, 1945-46
Gruppo 5	1947-58
Gruppo 6	1959-83
Gruppo 7	1984-99
Gruppo 8	2000-18

Tabella 4.1: Composizione dei gruppi di curve di mortalità per la popolazione italiana maschile ottenuta con il modello proposto.

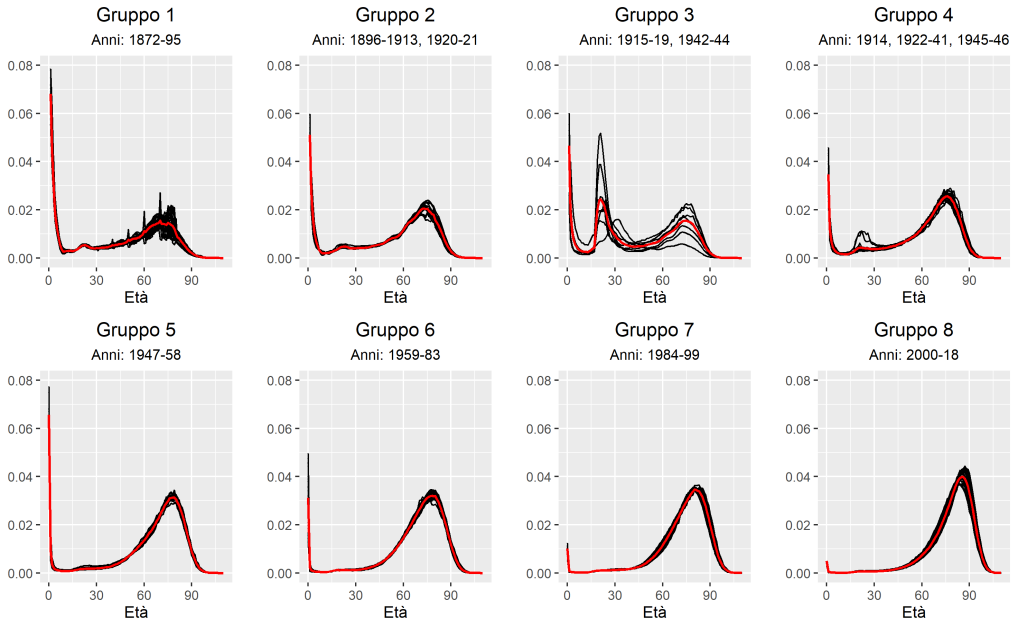


Figura 4.2: Distribuzioni del numero di decessi per età normalizzate (in nero) ripartite nei sei gruppi individuati con (in rosso) le stime di π_h per ciascun gruppo.

sto algoritmo necessita di conoscere già in partenza il numero di gruppi da identificare, è stato deciso che tale valore corrispondesse al numero di gruppi non vuoti individuati dal modello proposto, quindi otto in questo caso. k-medie individua anch'esso dei gruppi che possono essere identificati con dei precisi periodi temporali. Inoltre, tramite tale algoritmo è possibile ottenere un centroide per ogni gruppo, che in questo caso può intendersi come stima non parametrica delle curve all'interno di ogni gruppo. Vengono quindi confrontate le stime delle righe di π corrispondenti ai gruppi non vuoti con quelle ottenute facendo un clustering k-medie alla popolazione di curve di mortalità standardizzate. Il confronto tra il gruppo a cui ogni curva viene assegnata con i due diversi algoritmi è riportato in Tabella 4.2 ed evidenzia che i risultati sono tra loro coerenti. Le stime fornite dal modello proposto risultano avere un errore quadratico medio *in-sample* lievemente superiore (0.00042 per k-medie contro 0.00049 per il modello proposto).

Modello	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6	Gr. 7	Gr. 8
Gr. 1	19	5	0	0	0	0	0	0
Gr. 2	0	17	0	3	0	0	0	0
Gr. 3	0	1	4	3	0	0	0	0
Gr. 4	0	0	0	23	0	0	0	0
Gr. 5	0	0	0	0	12	0	0	0
Gr. 6	0	0	0	0	5	20	0	0
Gr. 7	0	0	0	0	0	3	13	0
Gr. 8	0	0	0	0	0	0	4	15

Tabella 4.2: Differenze nella composizione dei gruppi di curve di mortalità per la popolazione italiana femminile tra il modello proposto (righe) e l'algoritmo k-medie (colonne).

4.3.2 Femmine

Per la popolazione femminile le curve di mortalità a disposizione sono rappresentate in Figura 1.3. Vengono individuati dal modello sei gruppi, anche questa volta meno del limite superiore fissato, che sono rappresentati in Figura 4.3. Anche in questo caso come nel precedente ciascun gruppo individuato corrisponde ad un preciso arco temporale (Tabella 4.3).
Per ogni gruppo viene fornita una stima della densità (Figura 4.4) e tutte sembrano avere un buon adattamento. I risultati ottenuti da questo con il

	Anni
Gruppo 1	1872-96, 1918
Gruppo 2	1897-1917, 1919-22
Gruppo 3	1923-47
Gruppo 4	1948-65, 1968
Gruppo 5	1966-67, 1969-88
Gruppo 6	1989-2018

Tabella 4.3: Composizione dei gruppi di curve di mortalità per la popolazione italiana femminile ottenuta con il modello proposto.

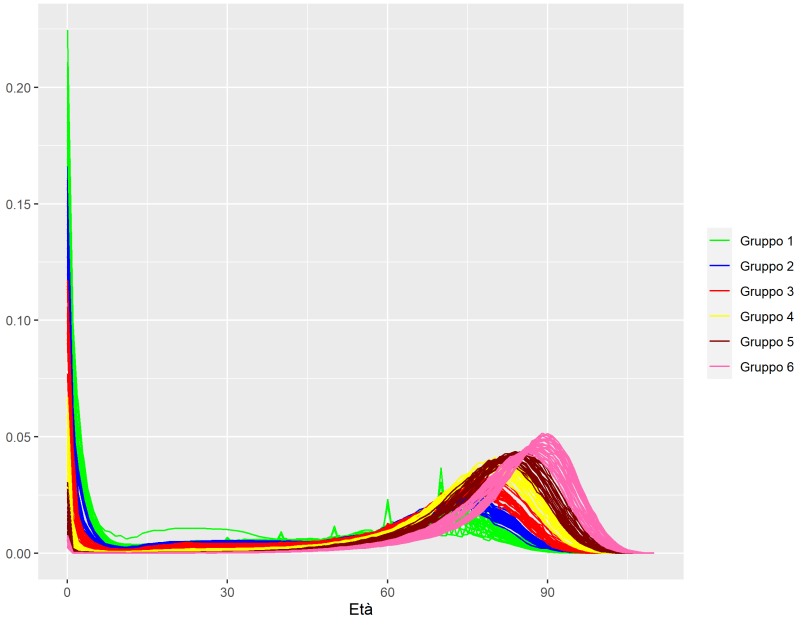


Figura 4.3: Distribuzioni del numero di decessi per età per la popolazione femminile italiana dal 1872 al 2018 ripartite nei 6 gruppi individuati dal modello.

modello proposto sono stati confrontati con quelli ottenuti tramite il tradizionale algoritmo di *clustering* k-medie. Ancora una volta il numero di gruppi per questo algoritmo è stato fissato pari al numero di gruppi non vuoti individuati dal modello proposto, sei in questo caso. La composizione dei gruppi identificati da questo algoritmo sembra coerente con i risultati del modello proposto, come mostrato dalla Tabella 4.4. Confrontando i centroidi ottenuti dall'algoritmo k-medie applicato alla popolazione di curve standardizzate con le stime di π fornite dal modello proposto, queste ultime forniscono un errore quadratico medio lievemente inferiore (0.00040 per k-medie contro 0.00038 per il modello proposto).

4.3.3 Commenti

A fronte dei risultati ottenuti è possibile effettuare qualche considerazione:

1. sotto l'aspetto del *clustering* i due metodi hanno portato ad avere risul-

Modello	Gr. 1	Gr. 2	Gr. 3	Gr. 4	Gr. 5	Gr. 6
Gr. 1	25	1	0	0	0	0
Gr. 2	0	24	1	0	0	0
Gr. 3	0	0	25	0	0	0
Gr. 4	0	0	0	19	0	0
Gr. 5	0	0	0	4	18	0
Gr. 6	0	0	0	0	4	26

Tabella 4.4: Differenze nella composizione dei gruppi di curve di mortalità per la popolazione italiana femminile tra il modello proposto (righe) e l'algoritmo k-medie (colonne).

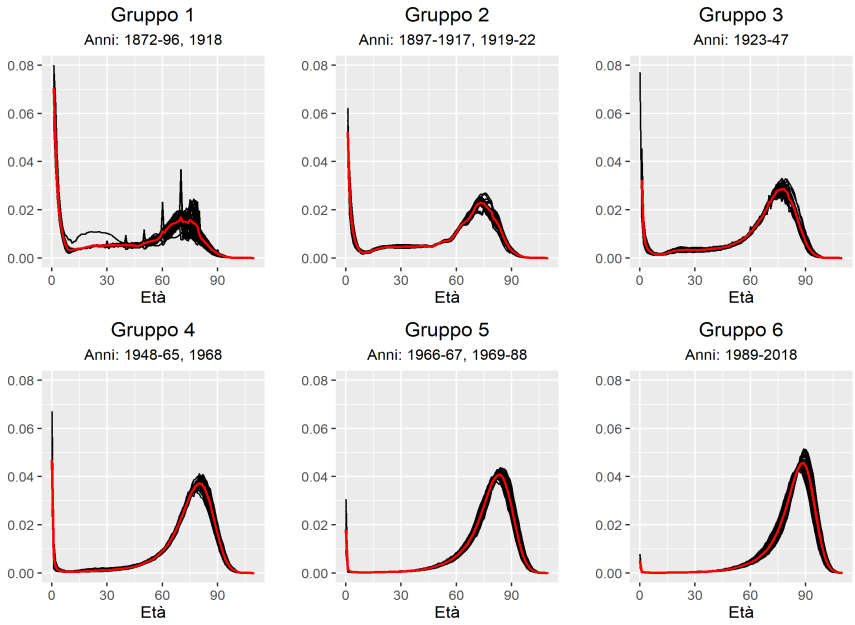


Figura 4.4: Distribuzioni del numero di decessi per età normalizzate (in nero) ripartite nei sei gruppi individuati con (in rosso) le stime di π_h per ciascun gruppo.

tati coerenti fra di loro in entrambe le situazioni ed in linea con quanto ci si poteva attendere (ad esempio nel caso della popolazione maschile, le curve degli anni delle due Guerre Mondiali sono racchiuse nello stesso gruppo, che comprende esclusivamente quelle). Tuttavia vale la

pena di sottolineare il modello proposto è in grado di individuare autonomamente il numero di cluster presenti nella popolazione delle curve, mentre l'algoritmo tradizionale k-medie necessita della specificazione a priori del numero di cluster da identificare;

2. in entrambe le situazioni, il confronto in termini di errore quadratico medio delle stime fornite per ogni cluster dai due metodi è risultato essere molto simile. È bene però sottolineare che le stime della densità fornite dal modello proposto hanno alla base la formulazione teorica esposta nei Capitoli 3 e 4, mentre i centroidi ottenuti con il metodo k-medie consistono semplicemente in una media fra le osservazioni;

3. entrambi i metodi nelle situazioni in esame considerano ogni curva come indipendente dalle altre. In particolare, il modello proposto assume le curve come osservazioni indipendenti ed identicamente distribuite provenienti da una mistura di distribuzioni multinomiali (4.1). Entrambi i metodi quindi non tengono conto dell'autocorrelazione presente nei dati, che invece sembra lecito aspettarsi.

Capitolo 5

Analisi della mortalità per i comuni italiani nel 2020

Una volta implementato il modello proposto e visti i risultati della sua applicazione a curve di mortalità riguardanti l'intera popolazione italiana, sembra interessante tentare un'analisi delle curve di mortalità a livello comunale relative all'anno 2020. Ciò risulta stimolante in quanto una grossa parte di queste curve presentano irregolarità che possono essere dovute dovute principalmente a due motivi: il fatto che un comune sia scarsamente abitato e l'effetto che può aver avuto la pandemia di COVID-19 che ha colpito l'Italia in quell'anno. La flessibilità garantita dal modello proposto permette di cogliere eventuali andamenti dovuti agli effetti della pandemia, mentre le stime per le curve dei comuni meno abitati non risentono dell'irregolarità delle curve di mortalità in quanto tali stime si basano anche sull'informazione apportata dalle altre curve contenute nello stesso gruppo. Inoltre, il fatto che il modello può essere anche uno strumento per il *clustering* consente di poter individuare comuni che hanno avuto un impatto analogo della pandemia dal punto di vista della mortalità.

5.1 Utilizzo del modello e dell'algoritmo

Per l'analisi delle curve di mortalità a livello comunale relative all'anno 2020 si utilizzano i dati forniti dall'Istituto nazionale di Statistica (2021) e descritti nella Sezione 1.3.2. Questi consistono nelle curve di mortalità della popolazione italiana maschile e femminile nell'anno 2020 per ciascuno dei 7903 comuni italiani esistenti. Per questa analisi per semplicità si utilizzano esclusivamente i dati relativi alla popolazione maschile. A differenza di quelli utilizzati nelle applicazioni precedenti, questi dati utilizzano classi di età quinquennali che vanno da $[0, 5)$ a $[100, \infty)$.
A questo insieme di curve di mortalità si vuole adattare il modello proposto nel Capitolo 4. Per via della presenza di classi quinquennali anziché annuali nelle curve osservate è stato necessario apportare qualche modifica al modello specificato precedentemente, tuttavia tali modifiche non hanno modificato né la struttura del modello né l'algoritmo di stima. Ciò è possibile grazie al fatto che la distribuzione a priori per ciascuna riga del parametro π (4.6) è una misura di probabilità indotta da un processo di Dirichlet, come mostrato nella (3.4). Questo permette di specificare tale distribuzione a priori per qualsiasi partizione dell'asse reale (Gelman et al., 2013), consentendo di utilizzare le classi quinquennali mantenendo la validità del modello e la correttezza della sua specificazione.
Per ciascuna curva, si consideri $x_i,\ i = 1, \dots, n$, l'età esatta alla morte del soggetto i-esimo, come definito nella Sezione 3.1. Allora definiamo ${}_5d_j$ come il numero di decessi avvenuti tra l'età j e l'età $j + 5$, cioè

$$ {}_5d_j = \sum_{i=1}^{n} \mathrm{I}(x_i \in [j, j + 5)) $$

per $j = 0, 5, \dots, 95$ e ${}_5d_{100} = \sum_{i=1}^{n} \mathrm{I}(x_i \in [100, \infty))$. Una distribuzione di decessi per classe di età sarà perciò rappresentata come ${}_5d_0, \dots, {}_5d_{100}$.

È possibile rappresentare J curve di mortalità nella matrice ${}_5D$:

$$ {}_5D = \begin{bmatrix} {}_5d_0^{(1)} & \dots & {}_5d_{100}^{(1)} \\ \vdots & & \\ {}_5d_0^{(J)} & \dots & {}_5d_{100}^{(J)} \end{bmatrix} $$

nella quale l'elemento ${}_5d_l^{(j)}$ indica il numero di morti tra l'età l e $l+5$ per la j-esima curva di mortalità, $j = 1, \dots, J$. Si ha quindi che ciascuna riga della matrice ${}_5D$ corrisponde ad una distribuzione del numero di decessi per classi di cinque anni di età.
Il modello diventa dunque

$$ {}_5d_0^{(j)}, \dots, {}_5d_{100}^{(j)} \overset{\text{iid}}{\sim} \sum_{h=1}^{H} w_h M_h $$

per $j = 1, \dots, J$ con

$$ M_h \sim \text{Multinom}(\pi_{0h}, \dots, \pi_{100h}) $$

per $h = 1, \dots, H$, dove H corrisponde a quello descritto nella Sezione 4.1 ed i parametri $w_1, \dots, w_H$ corrispondono alle probabilità della mistura quindi $w_h \in [0,1]$ $h = 1, \dots, H$, $\sum_{h=1}^{H} w_h = 1$. A questo punto, le distribuzioni a priori per i parametri w e π rimangono le stesse della specificazione precedente:

$$ w_1, \dots, w_H \sim \text{Dir}\left(\frac{1}{H}, \dots, \frac{1}{H}\right) $$

$$ \pi_{0h}, \dots, \pi_{100h} \sim \text{Dir}(a_0, \dots, a_{100}) $$

per $h = 1, \dots, H$ con $a_x = \alpha \int_x^{x+5} f(t|\theta)dt$ dove α rappresenta il parametro di precisione del processo di Dirichlet e $f(\cdot)$ è la misura di base definita nella (3.5). Anche in questo caso per semplicità computazionale $a_0, \dots, a_{100}$ vengono supposti noti ed uguali per tutti i gruppi.
A questo punto anche l'algoritmo MCMC per la stima del modello descritto

nella Sezione 4.2 mantiene invariata la sua struttura basata sull'alternanza di tre passi di *Gibbs-sampling*.

5.2 Decessi calcolati su una popolazione fittizia

Una prima analisi che si decide di svolgere è quella che prevede l'utilizzo dei decessi calcolati su una popolazione fittizia di grandezza 10^5 uguale per tutte le curve. In questa maniera ogni curva influisce allo stesso modo nella composizione dei gruppi e nella stima dei parametri w e π, indipendentemente dal numero di morti totale reale del comune. Così facendo, di fatto, si decide di fare *clustering* non considerando il totale dei morti che ciascun comune ha fatto registrare bensì basandosi esclusivamente sulla forma delle curve di mortalità. Questa analisi permette quindi di individuare effetti comuni della struttura per età della mortalità tra i comuni. Si è prima svolta una analisi su tutti i comuni italiani e successivamente si sono presi in considerazione solamente i comuni campani.

5.2.1 Italia

L'analisi svolta per semplicità ha considerato solamente le curve per la popolazione maschile ed si è utilizzato sempre il software `R`. In questa situazione, dopo diverse prove, si è scelto come valore dell'iperparametro $\alpha = 3 \times 10^8$ e dunque come valori noti ed uguali tra tutti i gruppi $a_0, \ldots, a_{100}$ si utilizza il vettore della media del numero di decessi fittizio normalizzato nelle 21 classi di età osservate moltiplicato per α. Come limite superiore al numero di gruppi si utilizza $H = 20$. Ai parametri $w_1, \ldots, w_H$ viene assegnato come valore iniziale $1/20$ per ognuno degli H elementi, mentre per ogni riga della matrice π si utilizza come valore iniziale il vettore della media del numero di decessi normalizzato alle varie classi di età. L'algoritmo MCMC arriva a convergenza dopo 10000 iterazioni delle quali vengono scartate le prime 5000 considerate periodo di *burn-in*. Si vuole infine sottolineare come questo adattamento del modello a quasi ottomila curve risulta ben più oneroso computazionalmente

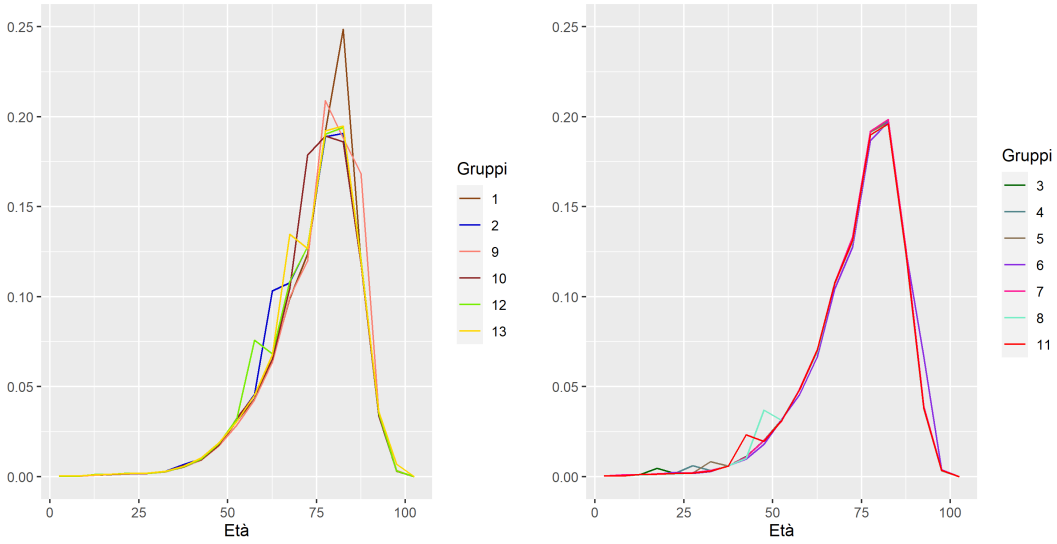


Figura 5.1: Confronto fra le stime di π_h di ciascun gruppo ottenute utilizzando il numero di decessi fittizio. Per rendere più semplice la comprensione nel grafico a sinistra sono riportate le stime per i gruppi che presentano picchi nella curva dopo i 50 anni, mentre a destra tutti le altre.

rispetto all'applicazione vista al Capitolo precedente. Infatti, a parità di valore di H e di macchina, in questo caso si impiega circa 25 volte di più per effettuare lo stesso numero di iterazioni.

I risultati dell'analisi sono riportati nella Figura 5.2, dove è rappresentato il confronto tra d_x^* osservati e le stime di π_h per ciascun gruppo, e nella Figura 5.1, dove è rappresentato il confronto tra le stime di π_h di ciascun gruppo. Le mappe che mostrano la ripartizione geografica dei gruppi sono contenute nell'Appendice B.

Vengono individuati 13 gruppi che sembrano essere principalmente caratterizzati dalla forma della curva, in particolare risulta evidente come vengano assegnate allo stesso gruppo curve che hanno dei picchi considerevoli in corrispondenza di una certa classe di età. In questo senso, ad esempio, il gruppo 1 sembra contenere le curve aventi un picco di decessi in proporzione al totale in corrispondenza della classe di età $[80 - 85)$. Di questo gruppo fanno parte, tra gli altri, i comuni di Bergamo, Fombio e Somaglia, questi ultimi

inseriti nella prima "zona rossa" nel febbraio 2020. Si notano inoltre alcuni gruppi la cui curva stimata risulta avere delle gobbe in diverse classi di età tra i 50 e i 75 anni, come ad esempio i gruppi 2, 10, 12, 13. In sostanza quindi, il modello riesce bene ad identificare le curve di mortalità che hanno caratteristiche comuni.
A livello geografico sembra che i gruppi 10 e 13, le cui curve stimate evidenziano una gobba in corrispondenza delle classi di età $[70-75)$ e $[65-70)$ rispettivamente, siano più presenti tra i comuni dell'Italia Settentrionale, mentre molti comuni di Emilia-Romagna, Toscana, Umbria e Marche appartengono ai gruppi 6 e 9. Tuttavia in generale non sembrano esserci forti legami tra la posizione geografica del comune e la sua appartenenza ad un determinato gruppo.

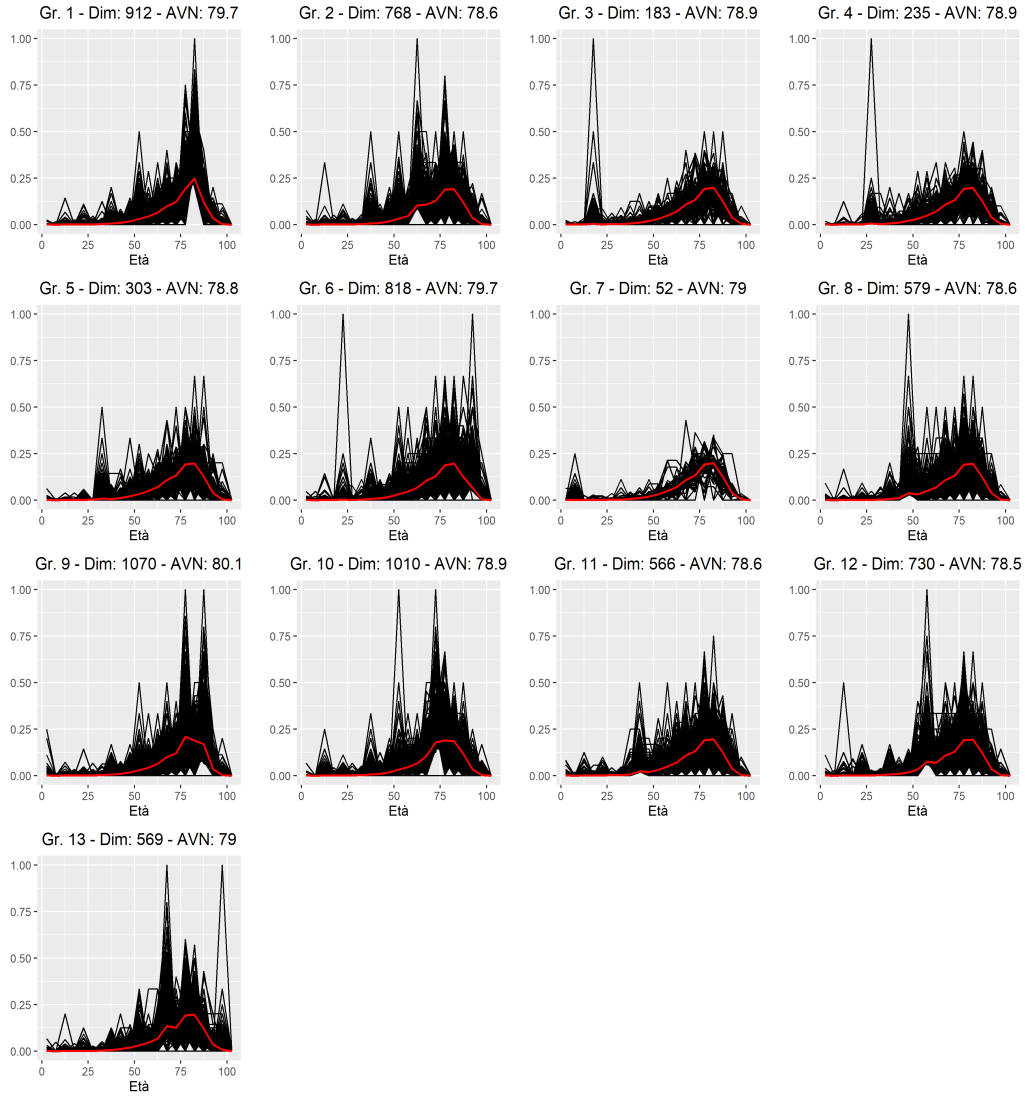


Figura 5.2: Distribuzioni del numero di decessi per età normalizzate (in nero) ripartite nei tredici gruppi individuati con (in rosso) le stime di π_h per ciascun gruppo ottenute utilizzando i decessi fittizi. Vengono riportate anche dimensione del gruppo (Dim) e aspettativa di vita alla nascita stimata (AVN).

5.2.2 Campania

Dopo aver applicato il modello alle curve di mortalità per la popolazione maschile di tutti i comuni italiani, si è rivolta l'attenzione ai comuni della Campania.

In questo caso come limite superiore per il numero di gruppi si è utilizzato $H = 20$, quindi viene assegnato $1/20$ come valore iniziale a ciascun elemento del parametro w. Per ogni riga della matrice π si utilizza come valore iniziale il vettore della media del numero di decessi normalizzato alle varie classi di età. Dopo diverse prove si è scelto come valore dell'iperparametro $\alpha = 1 \times 10^7$, dunque come valori noti ed uguali tra tutti i gruppi $a_0, \ldots, a_{100}$ si utilizza il vettore della media del numero di decessi fittizio normalizzato nelle 21 classi di età osservate moltiplicato per α. L'algoritmo MCMC arriva a convergenza dopo 10000 iterazioni, delle quali vengono scartate le prime

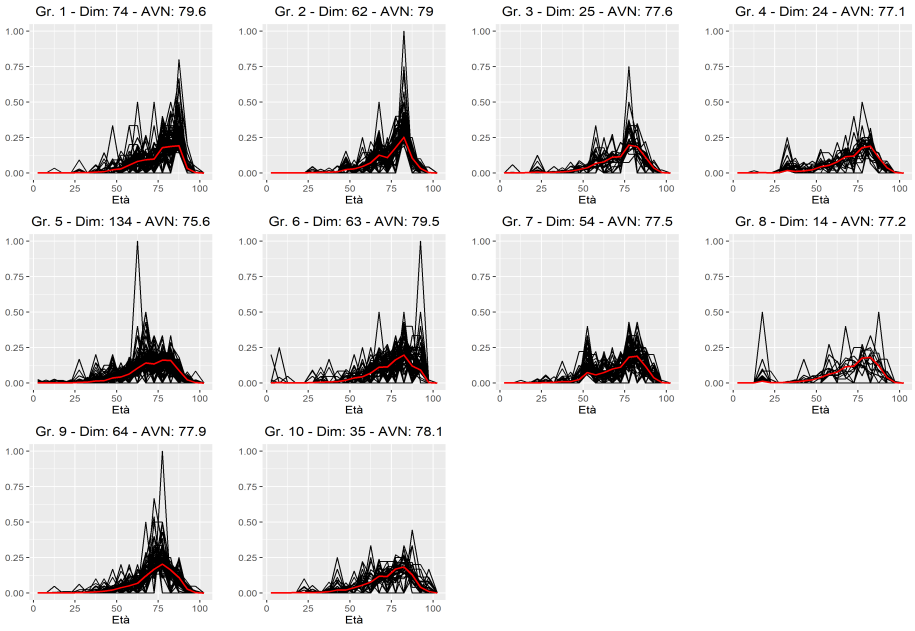


Figura 5.3: Distribuzioni del numero di decessi per età normalizzate (in nero) ripartite nei dieci gruppi individuati con (in rosso) le stime di π_h per ciascun gruppo ottenute utilizzando i decessi fittizi per i comuni della Campania. Vengono riportate anche dimensione del gruppo (Dim) e aspettativa di vita alla nascita stimata (AVN).

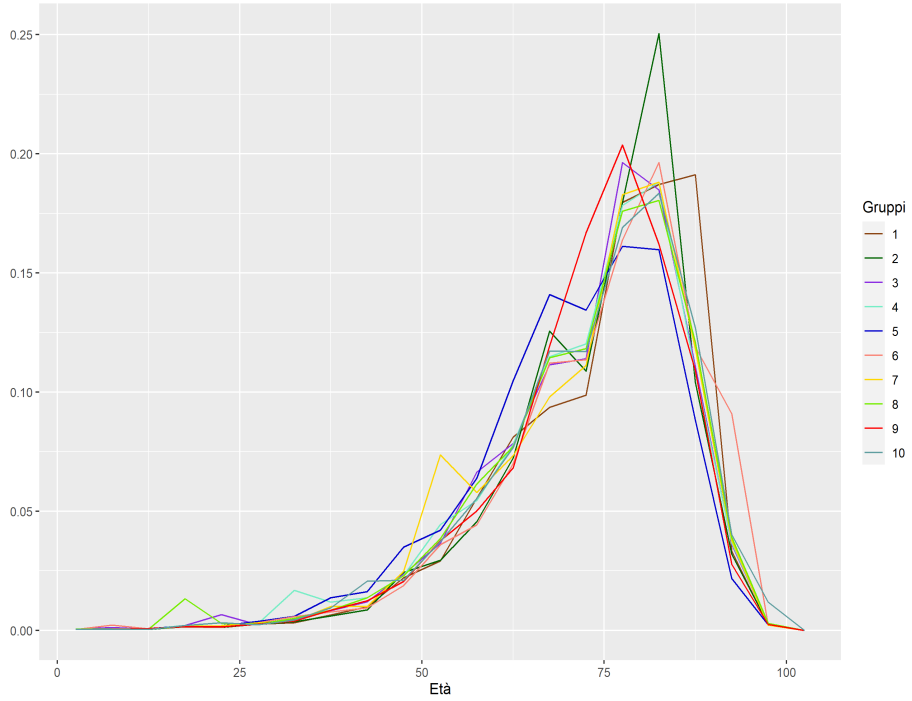


Figura 5.4: Confronto fra le stime di π_h di ciascun gruppo ottenute utilizzando il numero di decessi fittizio per i comuni della Campania.

5000 considerate periodo di *burn-in*.

I risultati dell'analisi sono riportati nella Figura 5.3, dove si confrontano d_x^* osservati e stime di π_h per ciascun gruppo, e nella Figura 5.4, dove si confrontano le stime di π_h di ciascun gruppo.

Vengono individuati 10 gruppi che sembrano essere caratterizzati soprattutto dalla forma della curva, come riscontrato anche nel caso dei comuni italiani. Sembra evidente come il gruppo 5 racchiuda le distribuzioni dei decessi dei comuni che riscontrano una proporzione di morti più elevata degli altri per le classi di età $[55-60)$ e $[60-65)$. Inoltre il gruppo 9 racchiude le curve di mortalità che sembrano avere un anticipo dell'età a cui si verifica il picco della proporzione di decessi rispetto agli altri gruppi.

A questo punto, guardando i risultati in Figura 5.5, si nota come ci sia forte caratterizzazione geografica del gruppo 5 in quanto racchiude tutti i comuni dell'*hinterland* di Napoli, quelli che compongono la cosiddetta "Terra dei

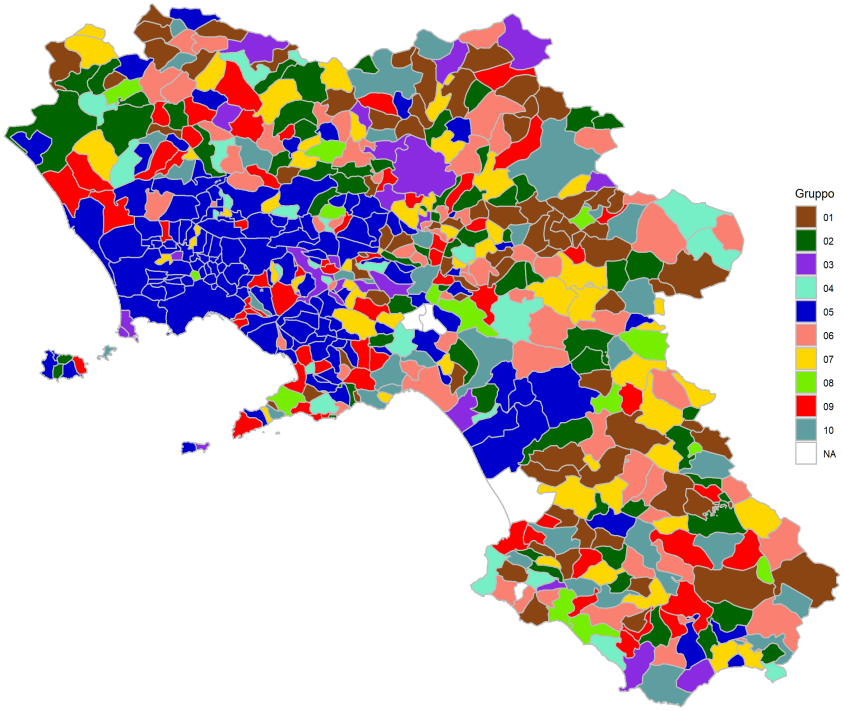


Figura 5.5: Gruppo di appartenenza di ciascun comune per la regione Campania (ottenuti utilizzando i decessi fittizi).

Fuochi". Con questo appellativo ci si riferisce al territorio dei novanta comuni compresi tra la provincia di Napoli e quella di Caserta interessato dal fenomeno delle discariche abusive e dall'abbandono incontrollato di rifiuti urbani e speciali, associato spesso alla combustione degli stessi (ARPA Campania, 2021). I roghi dei rifiuti hanno destato preoccupazione a causa dei fumi che si sprigionano e delle sostanze inquinanti riversate sui terreni agricoli che possono mettere rischio la salute della popolazione locale, come dimostra uno studio condotto da Istituto Superiore di Sanità e Procura della Repubblica di Napoli Nord (2020) che individua l'esistenza di una relazione causale tra la presenza di rifiuti in questo territorio e la formazione di tumori nella popolazione. Nella Figura 5.6 sono rappresentati i novanta comuni della "Terra dei Fuochi" e il rispettivo gruppo di appartenenza individuato dal modello. Risulta evidente come l'appartenenza al gruppo 5 sia predominante (ben 63

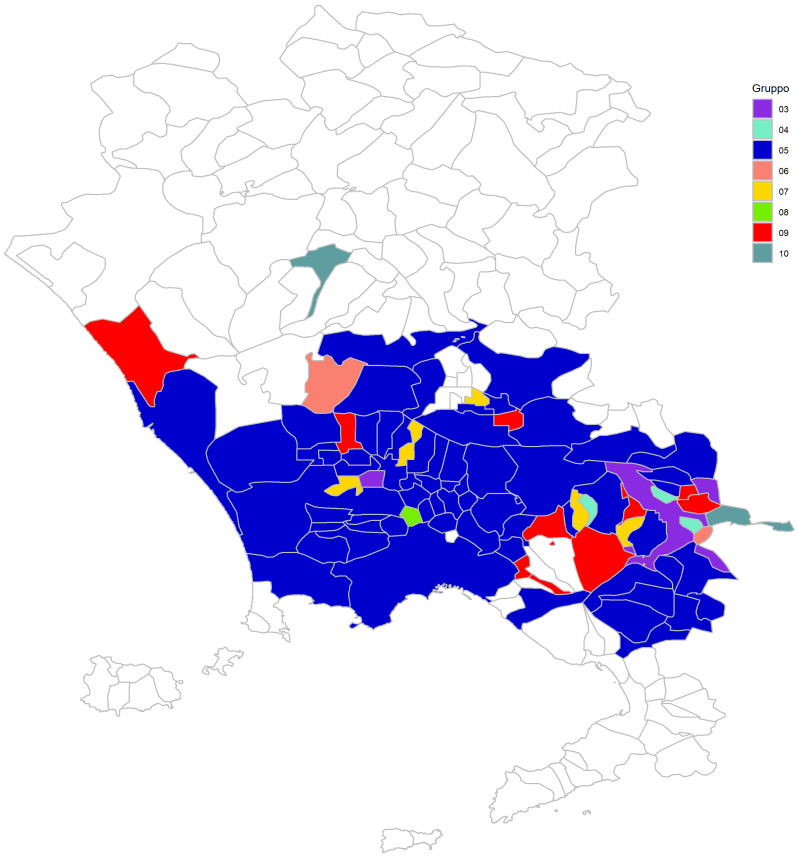


Figura 5.6: Gruppo di appartenenza di ciascun comune appartenente alla cosiddetta "Terra dei Fuochi" (ottenuti utilizzando i decessi fittizi).

comuni su 90 ne fanno parte) ed anche il gruppo 9 è rappresentato da 10 comuni.

5.3 Decessi reali

Un'alternativa all'analisi svolta precedentemente è quella di utilizzare il numero reale dei decessi per ciascun comune, anziché calcolarli su una popolazione fittizia. In questa maniera i comuni con più decessi, che verosimilmente sono i più popolosi, avranno molto più peso rispetto a quelli con qualche cen-

	Dimensione	AVN	NMD
Gruppo 1	1018	80.8	53.2
Gruppo 2	4326	79.1	31.2
Gruppo 3	1011	77.6	53.3
Gruppo 4	366	78.3	71.4
Gruppo 5	516	80.2	118.9
Gruppo 6	558	75.2	57.4

Tabella 5.1: Dimensione, aspettativa di vita alla nascita (AVN), numero medio di decessi totali nell'anno (NMD) per ciascun gruppo individuato utilizzando i decessi reali.

tinaio di abitanti che nell'arco dell'anno hanno fatto registrare un manciata di decessi nella stima di w e π ed anche nella costruzione dei gruppi. L'idea alla base è quella che, considerando come segnale la presenza di un effetto dovuto alla pandemia di COVID-19, la proporzione di rumore è molto più alta nelle curve di comuni con poco popolosi. In questa modo il modello effettua una sorta di *borrowing of information*, infatti la stima della curva per comuni poco popolosi terrà conto, in maniera proporzionale al numero totale dei decessi, anche dell'informazione proveniente dai comuni più popolosi che sono assegnati al medesimo gruppo.

Per questa analisi, svolta con `R`, dopo alcune prove è stato scelto $\alpha = 10$, valore considerevolmente più basso rispetto alle situazioni precedente poiché, utilizzando il numero reale dei decessi, il totale dei decessi per ciascun comune era ampiamente inferiore rispetto alla popolazione fittizia iniziale (10^5) utilizzata solitamente. Come valori per $a_0, \ldots, a_{100}$ si utilizza il vettore della media del numero di decessi reali normalizzato nelle 21 classi di età osservate moltiplicato per α. Si imposta poi $H = 20$ ed ai parametri $w_1, \ldots, w_H$ viene assegnato valore iniziale $1/20$ per ognuno degli H elementi. Per ogni riga della matrice π si utilizza come valore iniziale il vettore della media del numero di decessi normalizzato alle varie classi di età. L'algoritmo MCMC effettua 20000 iterazioni delle quali vengono scartate le prime 10000 in quanto

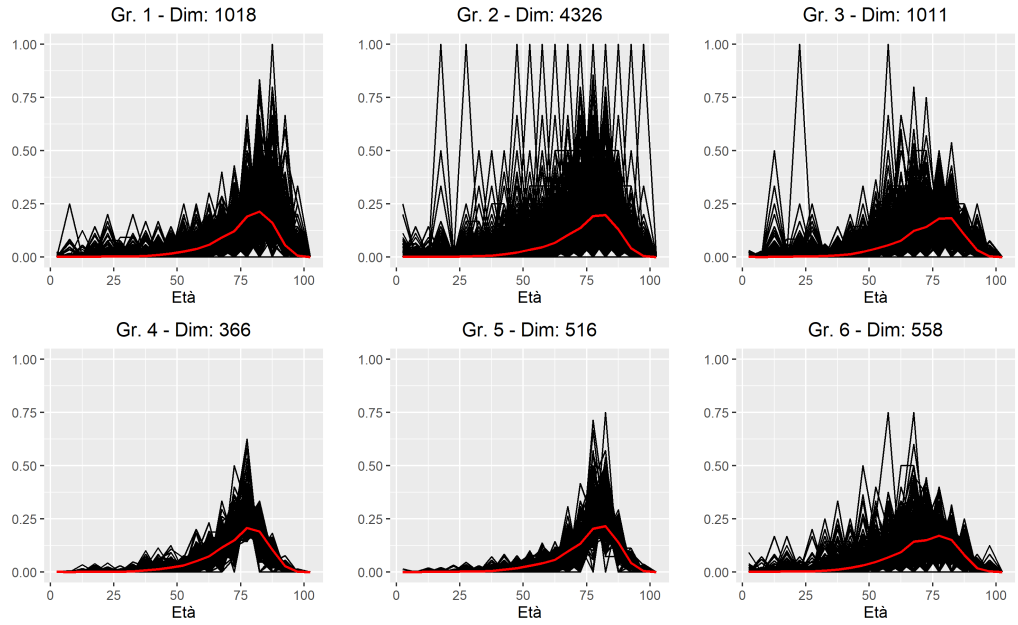


Figura 5.7: Distribuzioni del numero di decessi per età normalizzate (in nero) ripartite nei sei gruppi individuati con (in rosso) le stime di π_h per ciascun gruppo ottenute utilizzando i decessi reali.

vengono considerate periodo di *burn-in*.

Sono stati individuati sei gruppi, notevolmente meno rispetto al caso con i decessi calcolati su una posizione fittizia. Nella Tabella 5.1 sono riportate delle informazioni relative ai gruppi che aiutano a comprendere la loro composizione; nella Figura 5.7 è rappresentato il confronto tra d_x^* osservati e le stime di π_h per ciascun gruppo mentre nella Figura 5.8 è rappresentato il confronto tra le stime di π_h di ciascun gruppo. Le mappe che mostrano la ripartizione geografica dei gruppi sono contenute nell'Appendice C.

Spicca immediatamente all'occhio come più della metà dei comuni vengono assegnati al gruppo 2. Questo perché in questo gruppo sono presenti sia il comune di Roma che moltissimi altri comuni con un numero molto bassi di decessi nell'anno (31 in media) e ciò è dovuto al fatto che comuni poco popolosi hanno poco peso nella stima dei parametri rispetto a grandi comuni capoluogo. Infatti a dispetto della presenza di molte curve frastagliate, la curva delle stime di $\pi_{0,2}, \ldots, \pi_{100,2}$ risulta decisamente liscia. Tale risultato

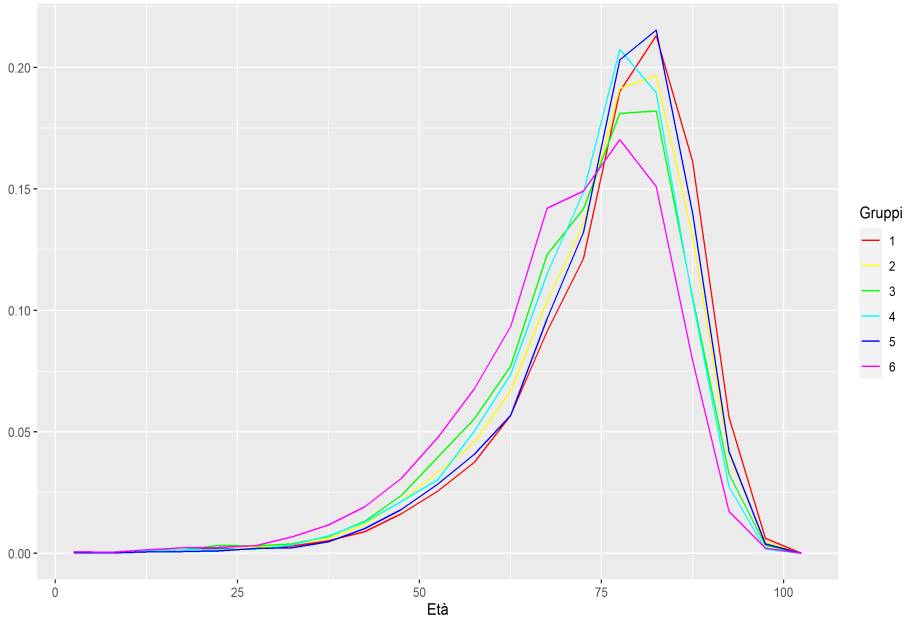


Figura 5.8: Confronto fra le stime di π_h di ciascun gruppo ottenute utilizzando il numero di decessi reale.

porta ragionevolmente ad affermare che l'irregolarità della curva di mortalità per comuni con pochi abitanti non è dovuta ad una effettiva presenza di questa irregolarità nella popolazione di curve, bensì alla scarsa popolazione residente in quel comune.

Si può fare invece il discorso in senso opposto per il gruppo 5. Questo gruppo infatti è composto in media da comuni che fanno registrare quasi 119 morti e racchiude al suo interno molti comuni capoluogo del Nord Italia come Milano, Venezia, Genova, Torino, Brescia e Bergamo.

Il gruppo 1 sembra essere quello che racchiude i comuni con le distribuzioni dei decessi per età più spostate verso età anziane in quanto è quello con aspettativa di vita alla nascita più elevata (80.8 anni). È formato principalmente da comuni di Emilia-Romagna e Centro Italia e racchiude anche i comuni capoluogo veneti di Padova, Treviso, Verona e Vicenza.

Il gruppo 4 è composto da comuni non capoluogo di medie dimensioni prevalentemente lombardi e la sua aspettativa di vita risulta essere la terza più bassa tra quelle individuate. Al suo interno sembra esserci una differenza di

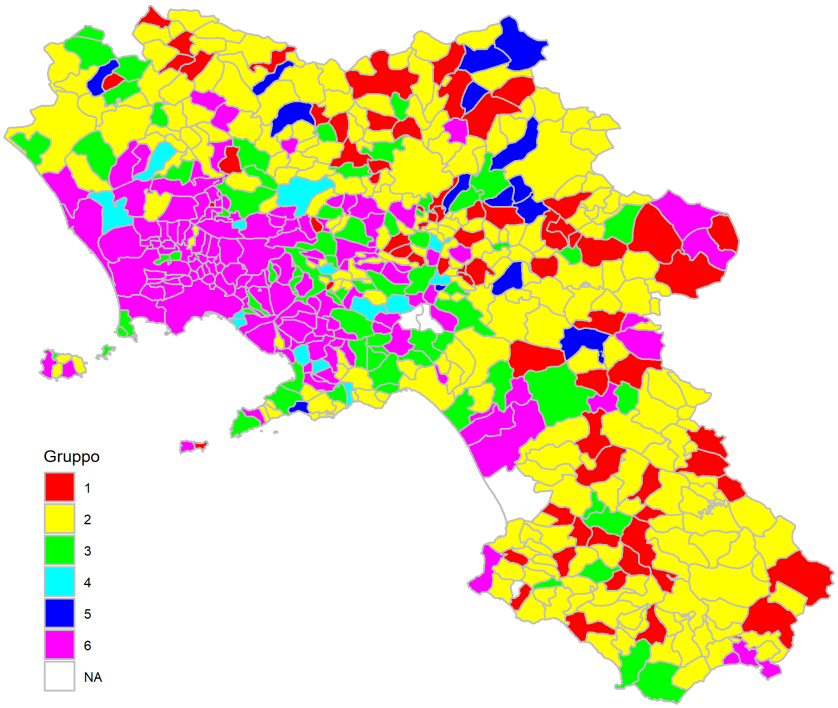


Figura 5.9: Gruppo di appartenenza di ciascun comune per la regione Campania.

forma delle curve di mortalità meno marcata rispetto ad altri gruppi.
Il gruppo 3 ha anch'esso una aspettativa di vita bassa rispetto agli altri (77.6 anni) e racchiude comuni principalmente appartenenti a Lombardia, Veneto e regioni del Sud Italia. La curva delle stime di $\pi_{0,3}, \ldots, \pi_{100,3}$ fa emergere una gobba per classi di età precedenti alla classe di età modale. A questo gruppo appartiene il comune di Lodi.
Il gruppo 6 è quello che sembra racchiudere i comuni con distribuzioni di decessi per età maggiormente spostate verso classi di età più giovani come dimostrato dall'aspettativa di vita alla nascita (75.2, per distacco la più bassa) e dalla forma delle stime di $\pi_{0,6}, \ldots, \pi_{100,6}$. Inoltre questo gruppo ha anche una forte caratterizzazione geografica in quanto racchiude tutti i comuni dell'*hinterland* di Napoli (Figura 5.9), molti comuni lombardi ed è anche molto presente in Sardegna.

Si fa infine notare che dei dieci comuni lodigiani inclusi nella prima cosiddetta "zona rossa" nel febbraio 2020 (Codogno, Castiglione d'Adda, Casalpusterlengo, San Fiorano, Bertonico, Fombio, Terranova dei Passerini, Somaglia, Maleo e Castelgerundo) sette di essi appartengono ai gruppi 3 e 6, cioè quelli con le curve di mortalità più spostate verso età più giovani.

Conclusioni

L'obiettivo principale di questa tesi era quello di discutere un modello flessibile in grado di ottenere delle stime di curve di mortalità per età aventi forme anche molto diverse fra loro e numerosità ridotta. È stato allora introdotto un modello bayesiano non parametrico basato sul processo di Dirichlet per la stima di una sola curva di mortalità ed una estensione valida nel caso le curve di mortalità da stimare siano molteplici. Tale estensione porta come risultato accessorio anche la possibilità di effettuare *clustering* senza che sia necessario definire in partenza il numero di gruppi che si vuole ottenere, in quanto saranno i dati ed il parametro di precisione α ad individuare il numero ottimale di gruppi.

Il modello per la stima di una singola curva fornisce buoni risultati, riuscendo a cogliere bene anche i più diversi andamenti grazie alla sua parte non parametrica. I risultati dell'adattamento di tale modello confrontati con quelli ottenuti tramite il modello parametrico proposto da Heligman e Pollard (1980) risultano migliori secondo la metrica dell'errore quadratico medio *in-sample*.

L'estensione al caso di più curve viene utilizzata per modellare le curve di mortalità della popolazione italiana maschile e femminile dal 1872 al 2018. I risultati sono buoni, infatti i gruppi individuati dal modello sembrano ragionevoli e la loro composizione è in linea con quella ottenuta tramite un tradizionale metodo di *clustering k-means*, tuttavia quest'ultimo non permette l'individuazione automatica del numero di gruppi. Le curve stimate per ciascun gruppo si adattano bene alle curve osservate. In questo caso però

occorre sottolineare come il modello non tenga conto della dipendenza che ragionevolmente si ritiene esserci tra curve di mortalità di anni vicini, perciò questa può essere una strada da percorrere per migliorare il modello.
Successivamente il modello è stato adattato ai dati relativi ai decessi comunali per l'anno 2020 riferiti alla popolazione maschile, che quindi racchiudevano anche i morti dovuti alla pandemia di COVID-19. Si sono seguiti due approcci differenti tra loro: nel primo si utilizzava la distribuzione dei decessi calcolata su una popolazione fittizia uguale per tutti i comuni, nel secondo si utilizzavano i decessi reali. Le due strade derivano da due diversi punti di vista del problema. Infatti, considerando i decessi fittizi si vuole analizzare come varia la struttura per età della mortalità tra i diversi comuni, mentre con i decessi reali anche il totale dei decessi per ciascun comune influisce sulla composizione dei gruppi individuati e sulle stime dei parametri. Con il primo approccio il modello riesce bene a cogliere le differenze di forma tra le diverse curve. Non sembra essere presente una forte correlazione spaziale tra i gruppi, tuttavia adattando il modello ai dati comunali della regione Campania si può individuare una marcata caratterizzazione dei comuni della cosiddetta "Terra dei Fuochi". Anche i risultati ottenuti con il secondo approccio sono ragionevoli e coerenti con quanto ci si potesse aspettare, inoltre sembra essere presente anche qualche particolarità sulla distribuzione dei gruppi a livello geografico. In questo caso è possibile pensare ad alcune estensioni da poter sviluppare per tenere conto della possibile presenza di correlazione spaziale tra comuni vicini oppure per l'inserimento di covariate nel modello.

Appendice A

Alcuni *traceplot*

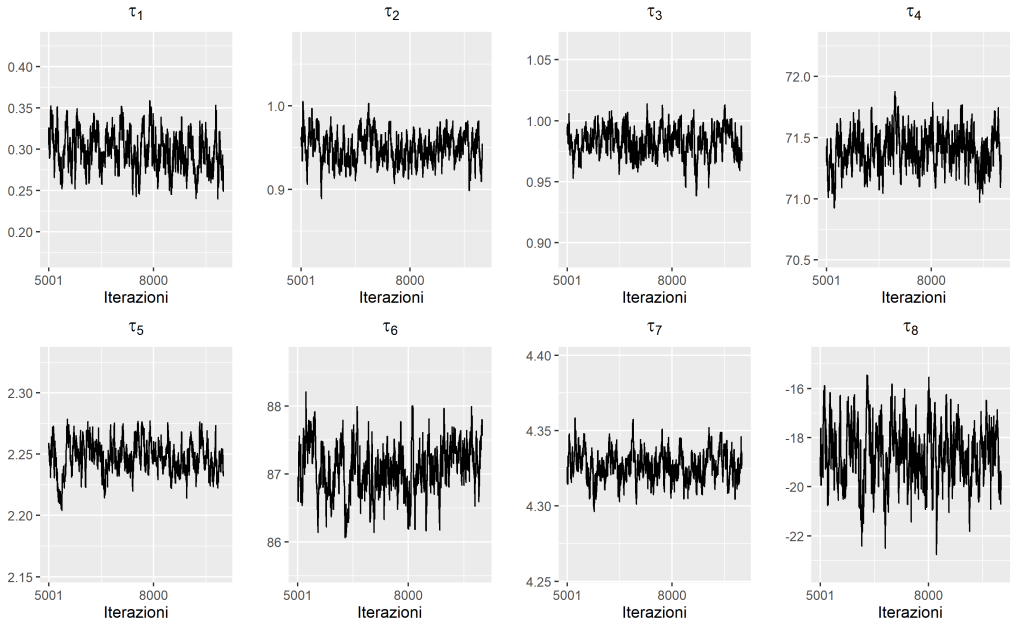


Figura A.1: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana maschile del 1889 dopo il periodo di *burn-in* (Sezione 3.4).

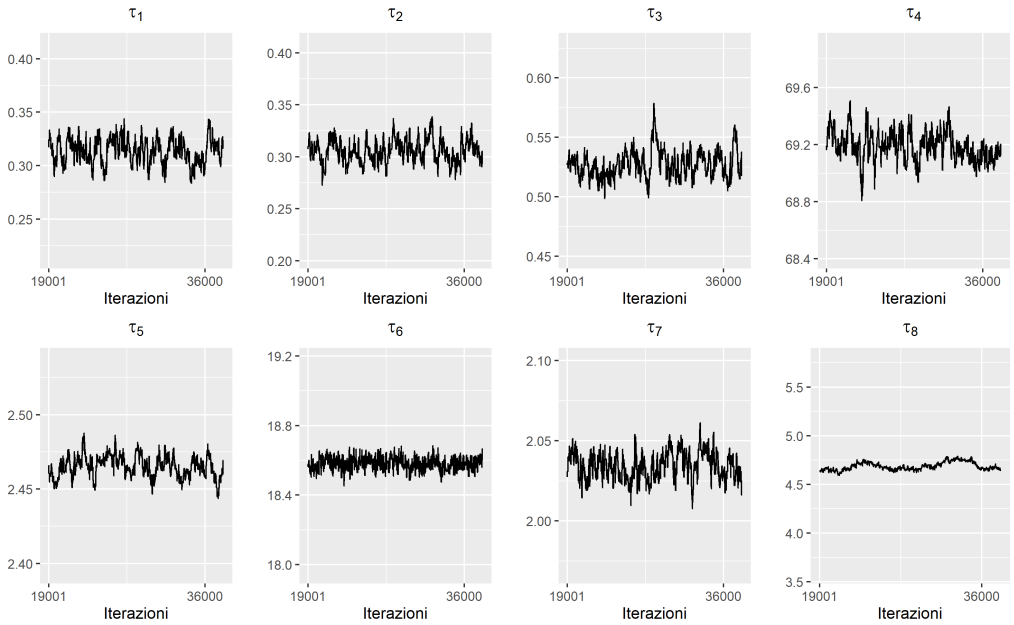


Figura A.2: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana maschile del 1916 dopo il periodo di *burn-in* (Sezione 3.4).

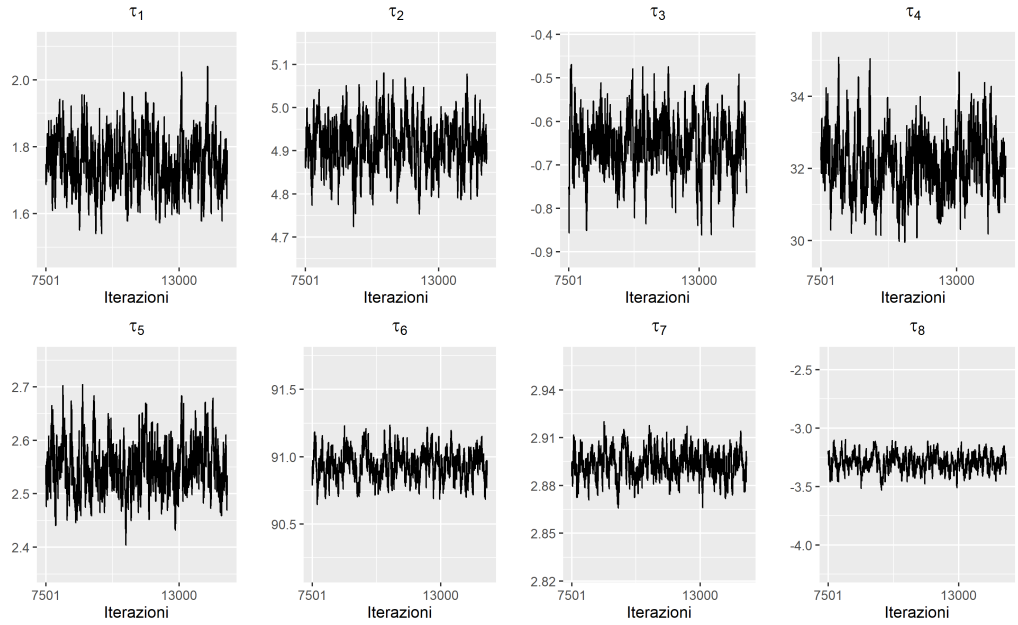


Figura A.3: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana maschile del 1995 dopo il periodo di *burn-in* (Sezione 3.4).

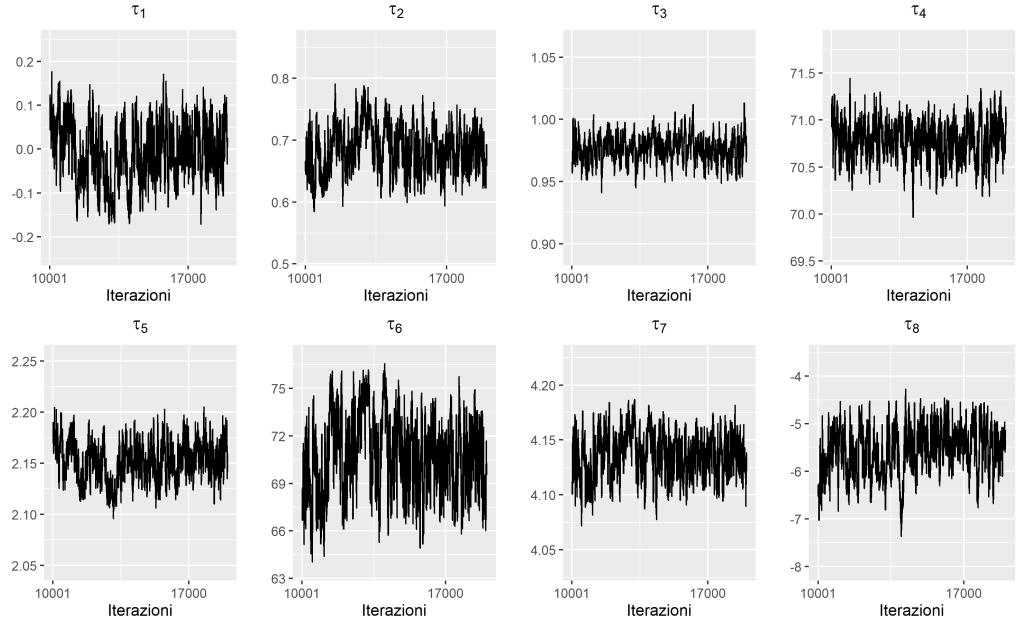


Figura A.4: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana femminile del 1880 dopo il periodo di *burn-in* (Sezione 3.4).

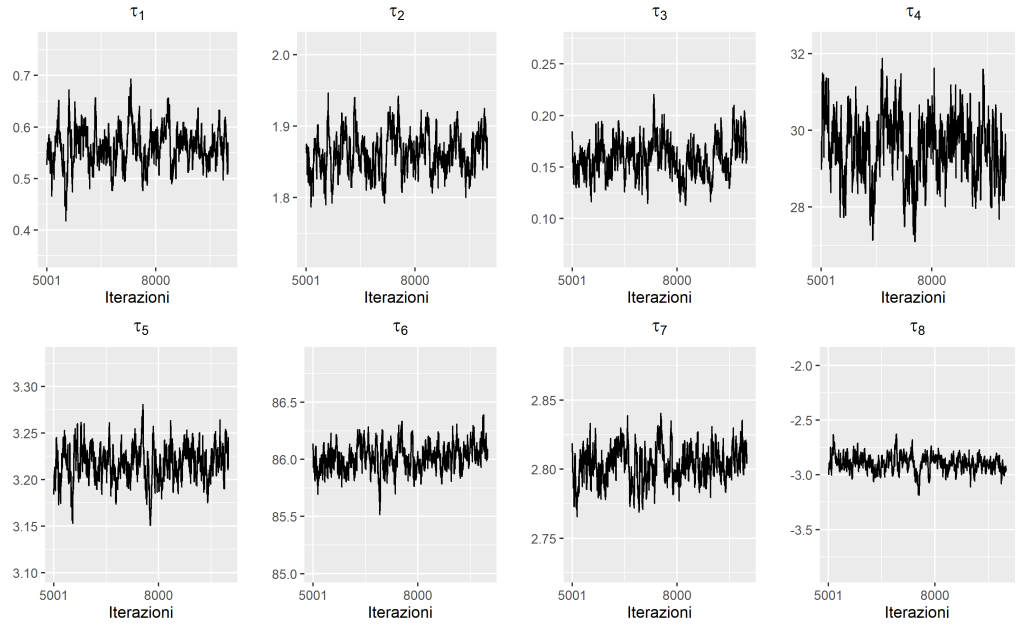


Figura A.5: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana femminile del 1935 dopo il periodo di *burn-in* (Sezione 3.4).

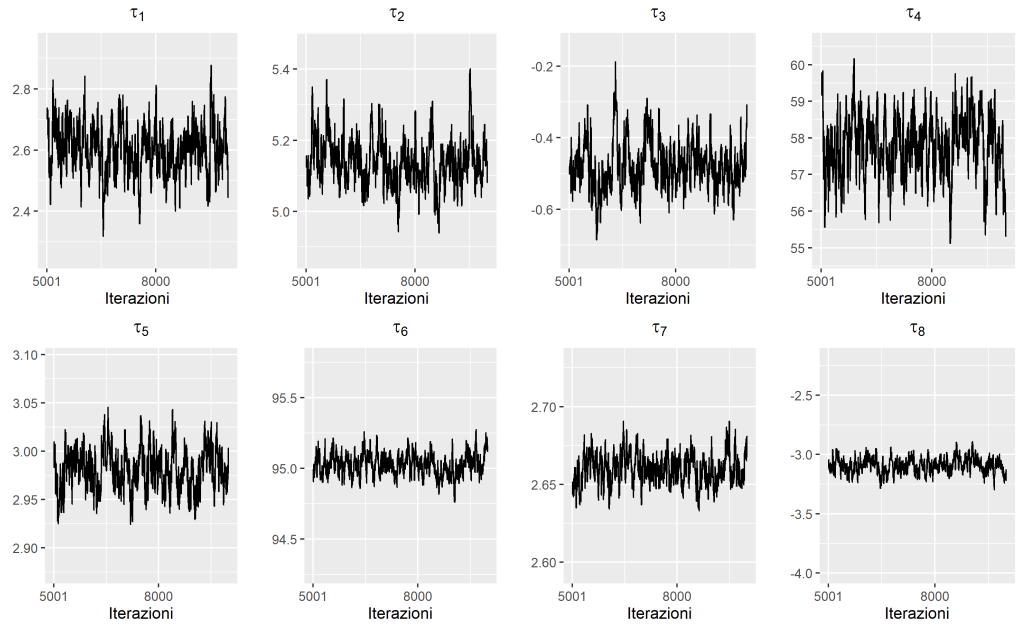


Figura A.6: *Traceplot* delle stime degli elementi del parametro τ ottenute dall'algoritmo MCMC per la curva di mortalità della popolazione italiana femminile del 1997 dopo il periodo di *burn-in* (Sezione 3.4).

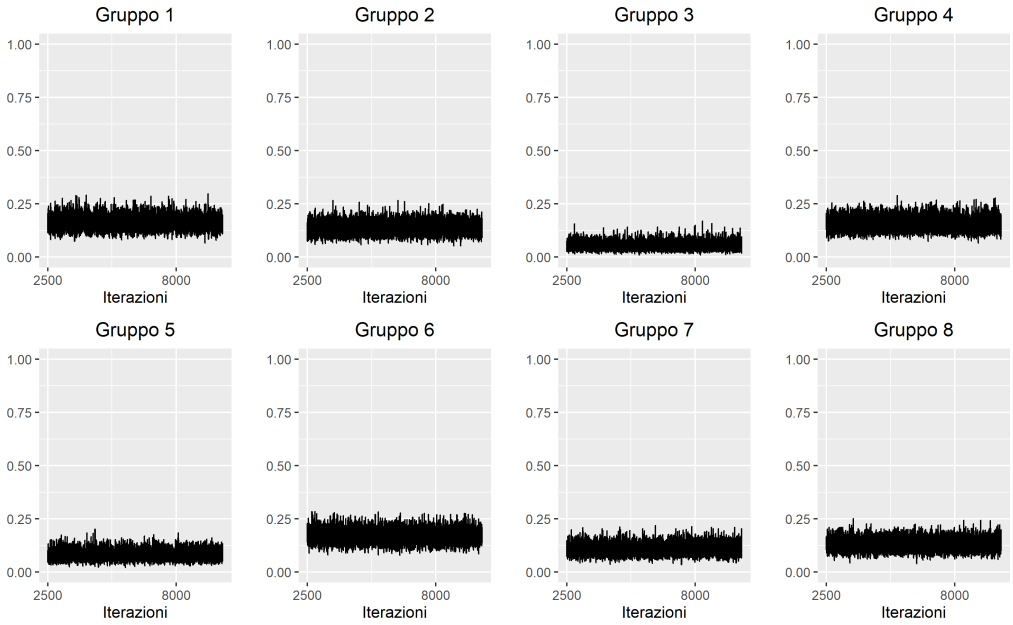


Figura A.7: *Traceplot* delle stime dei parametri w_h per i gruppi non vuoti ottenute dall'algoritmo MCMC dopo il periodo di *burn-in* relativamente all'analisi della popolazione maschile (Sezione 4.3.1).

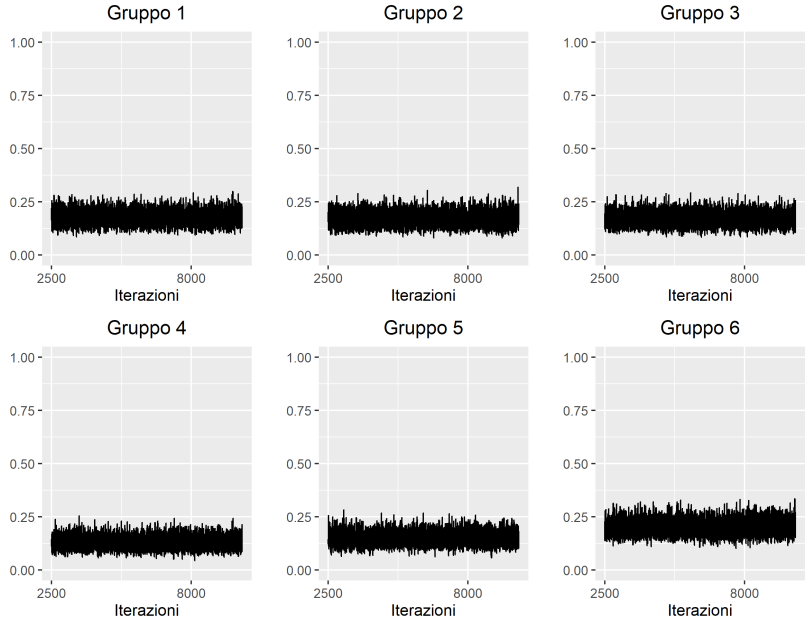


Figura A.8: *Traceplot* delle stime dei parametri w_h per i gruppi non vuoti ottenute dall'algoritmo MCMC dopo il periodo di *burn-in* relativamente all'analisi della popolazione femminile (Sezione 4.3.2).

Appendice B

Analisi della mortalità per comune nel 2020 con decessi fittizi: mappe

Vengono di seguito riportati i grafici che mostrano il gruppo a cui ogni comune viene assegnato a seguito dell'analisi della mortalità per comune nel 2020 utilizzando i decessi calcolati a partire da una popolazione fittizia di grandezza 10^5.
I comuni che nei grafici appaiono come `NA` sono generalmente piccoli comuni che non hanno fatto registrare alcun decesso durante l'anno oppure si sono recentemente fusi con degli altri e le mappe `GADM` utilizzate non erano aggiornate.

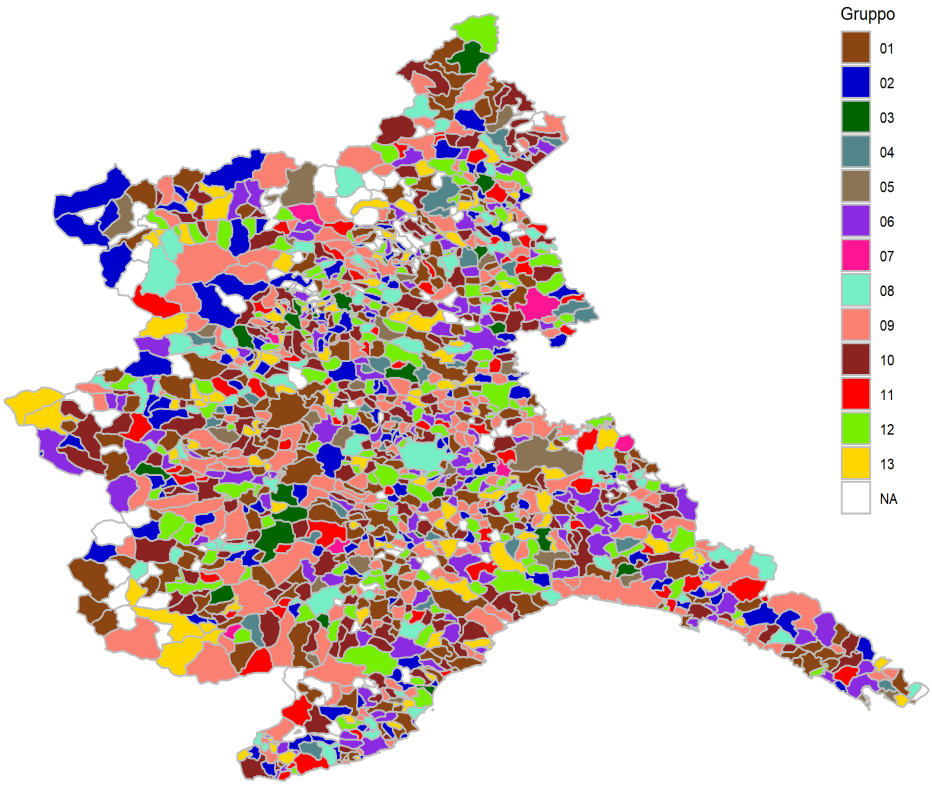


Figura B.1: Gruppo di appartenenza di ciascun comune per le regioni Valle d'Aosta, Piemonte e Liguria.

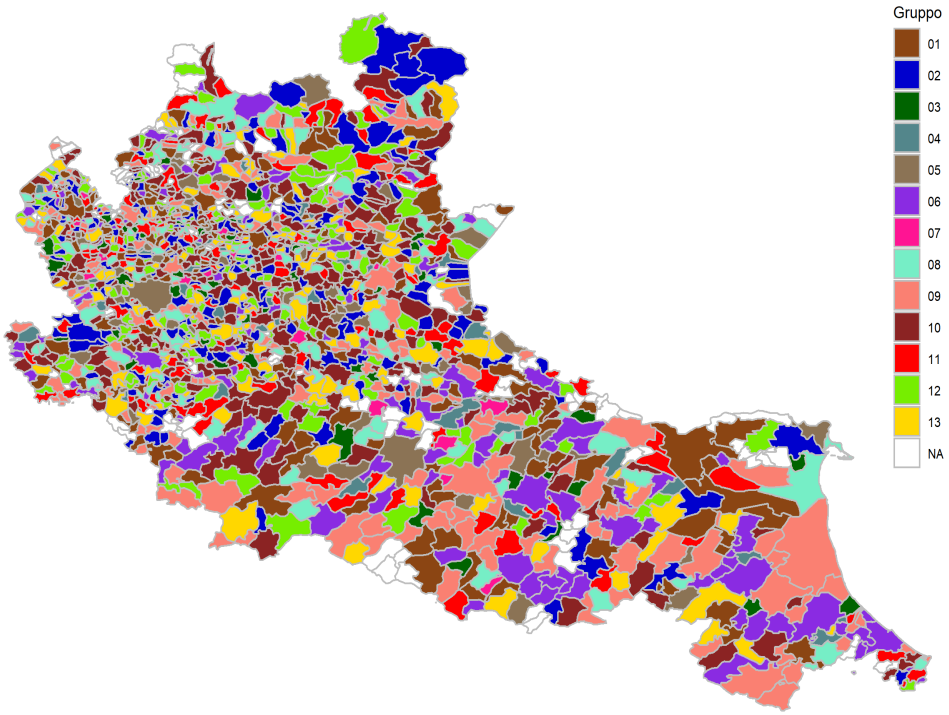


Figura B.2: Gruppo di appartenenza di ciascun comune per le regioni Lombardia ed Emilia-Romagna.

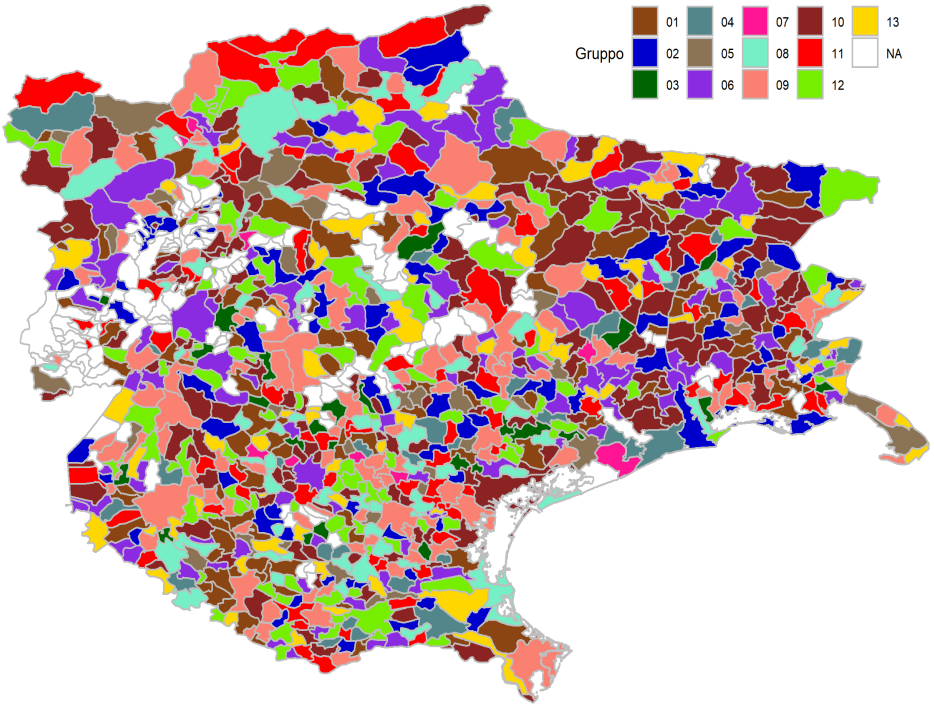


Figura B.3: Gruppo di appartenenza di ciascun comune per le regioni Trentino-Alto Adige, Veneto e Friuli-Venezia Giulia.

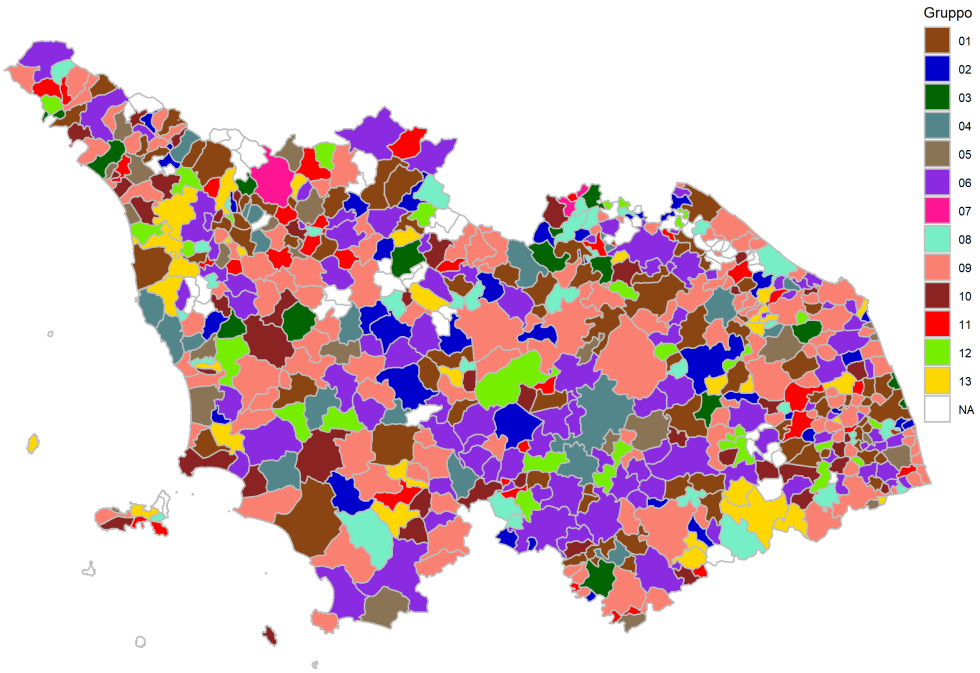


Figura B.4: Gruppo di appartenenza di ciascun comune per le regioni Toscana, Umbria e Marche.

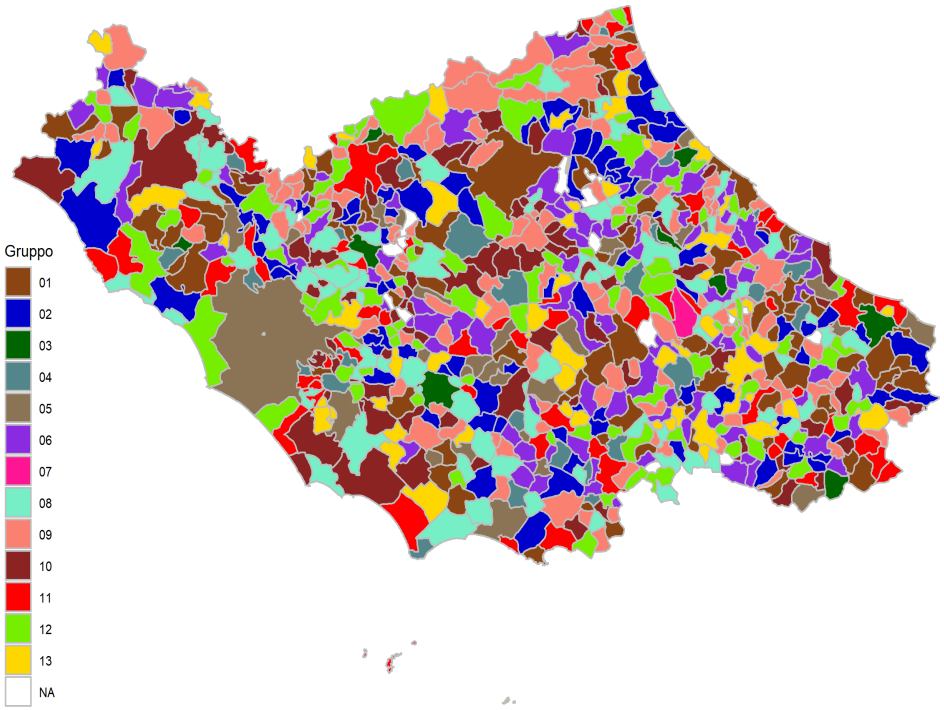


Figura B.5: Gruppo di appartenenza di ciascun comune per le regioni Lazio, Abruzzo e Molise.

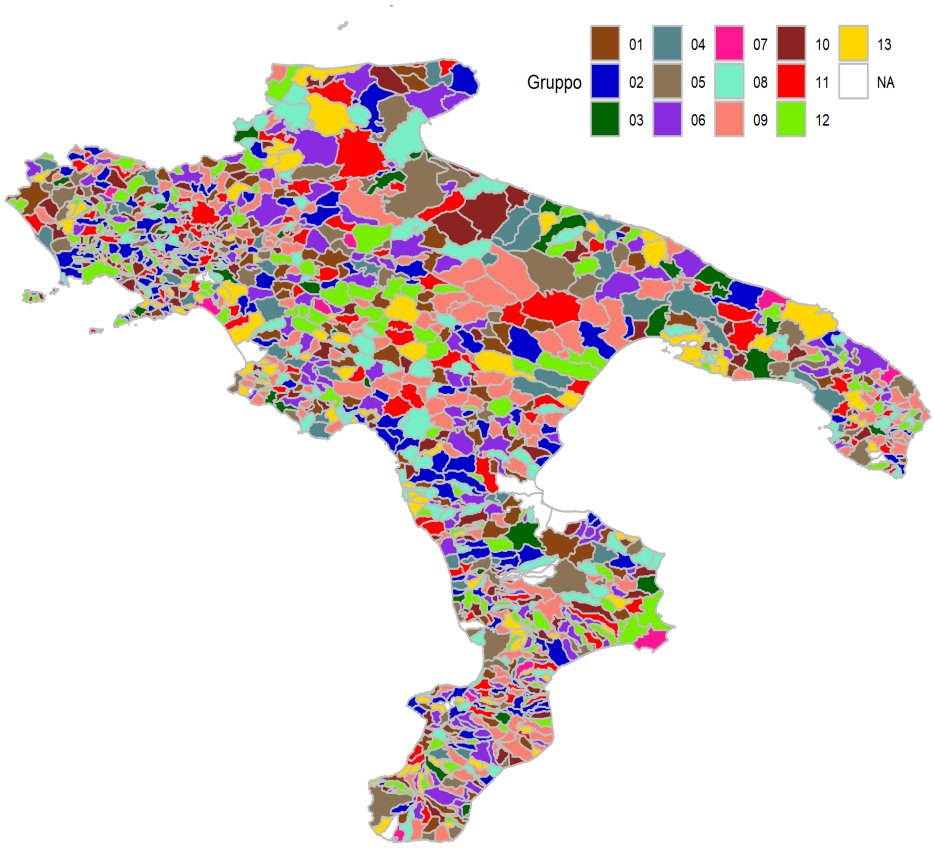


Figura B.6: Gruppo di appartenenza di ciascun comune per le regioni Campania, Puglia, Basilicata e Calabria.

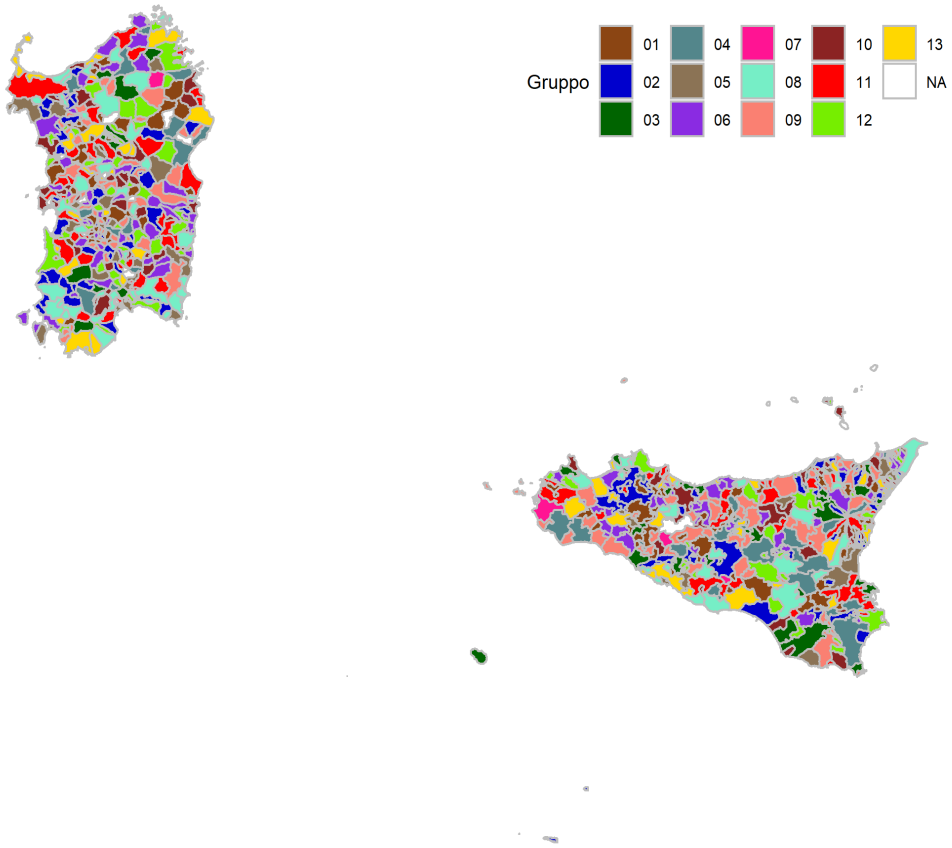


Figura B.7: Gruppo di appartenenza di ciascun comune per le regioni Sicilia e Sardegna.

Appendice C

Analisi della mortalità per comune nel 2020 con decessi reali: mappe

Vengono di seguito riportati i grafici che mostrano il gruppo a cui ogni comune viene assegnato a seguito dell'analisi della mortalità per comune nel 2020 utilizzando i decessi reali.
I comuni che nei grafici appaiono come `NA` sono generalmente piccoli comuni che non hanno fatto registrare alcun decesso durante l'anno oppure si sono recentemente fusi con degli altri e le mappe `GADM` utilizzate non erano aggiornate.

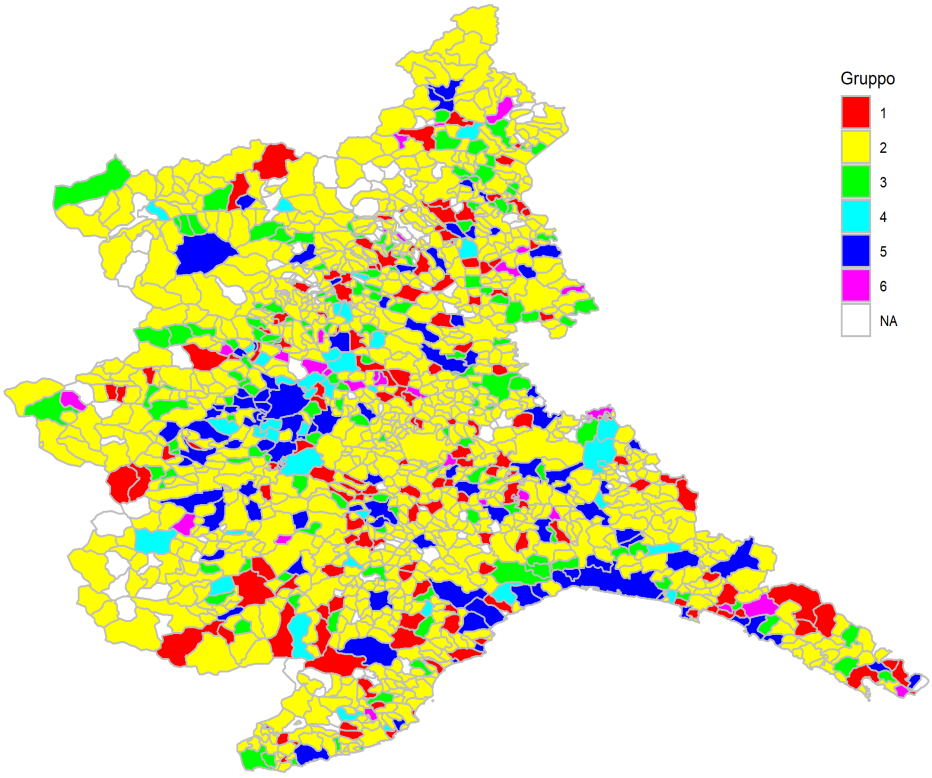


Figura C.1: Gruppo di appartenenza di ciascun comune per le regioni Valle d'Aosta, Piemonte e Liguria.

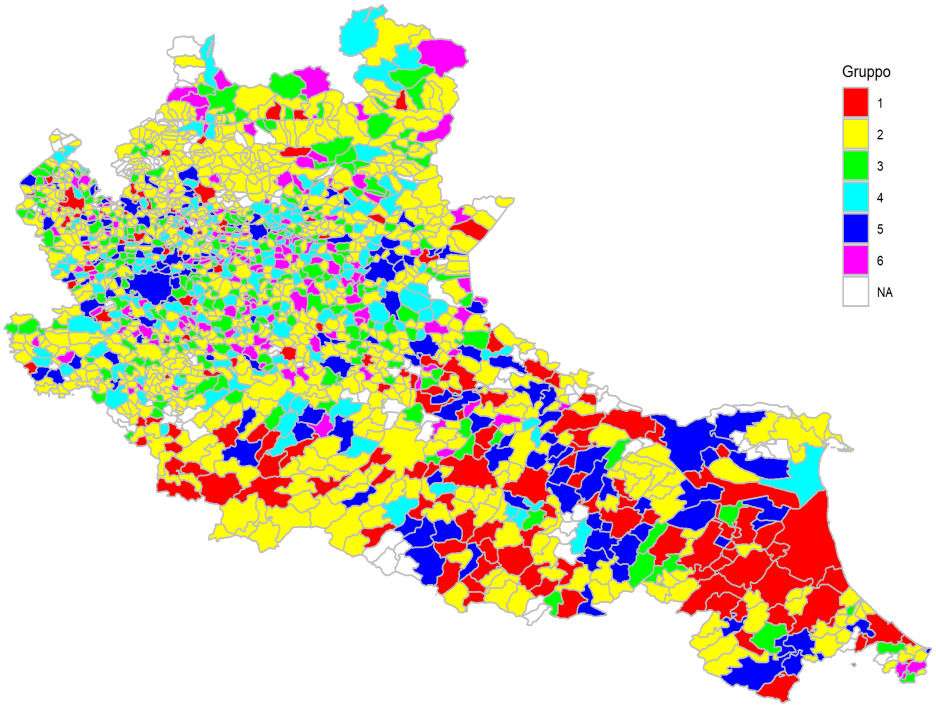


Figura C.2: Gruppo di appartenenza di ciascun comune per le regioni Lombardia ed Emilia-Romagna.

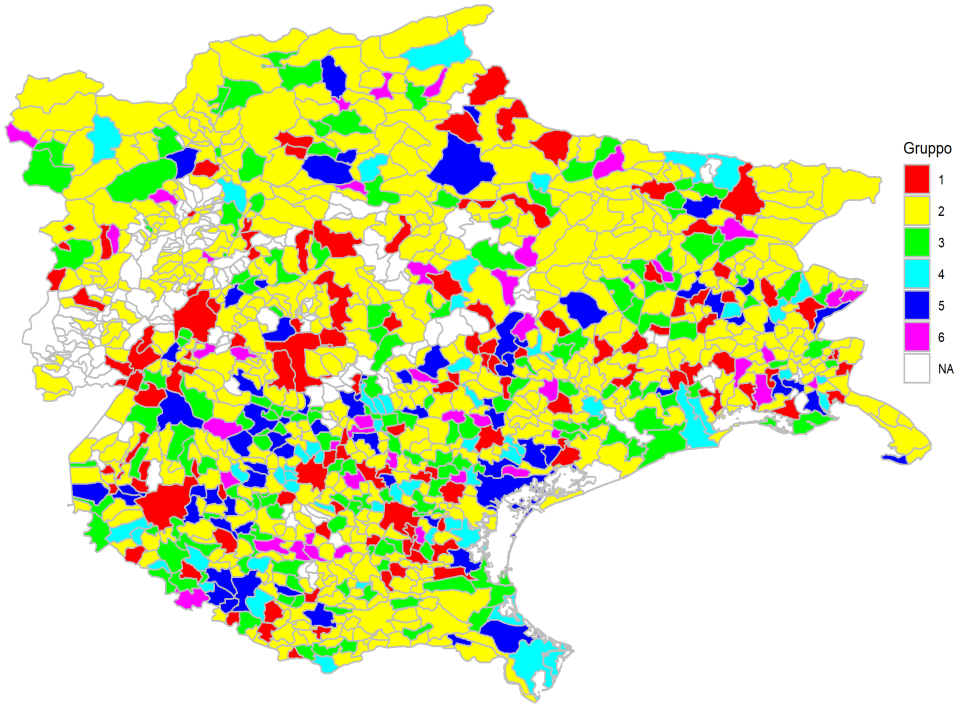


Figura C.3: Gruppo di appartenenza di ciascun comune per le regioni Trentino-Alto Adige, Veneto e Friuli-Venezia Giulia.

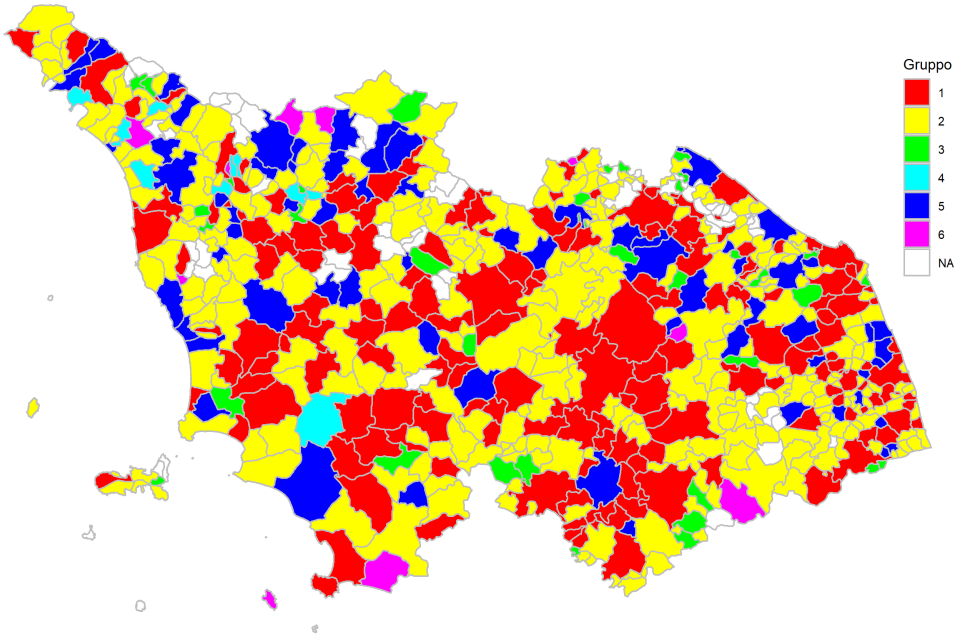


Figura C.4: Gruppo di appartenenza di ciascun comune per le regioni Toscana, Umbria e Marche.

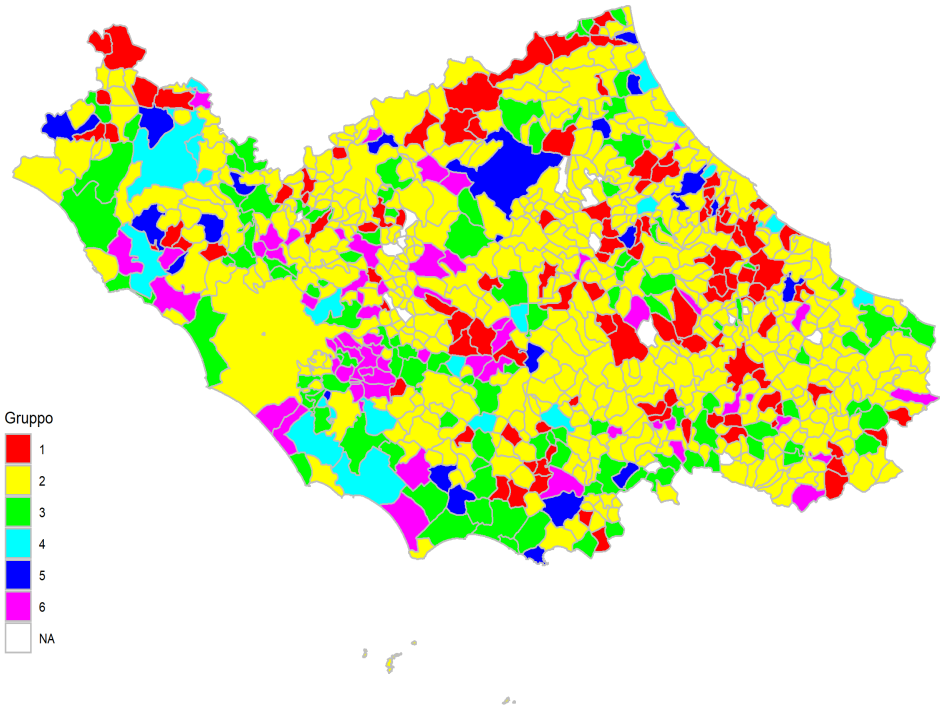


Figura C.5: Gruppo di appartenenza di ciascun comune per le regioni Lazio, Abruzzo e Molise.

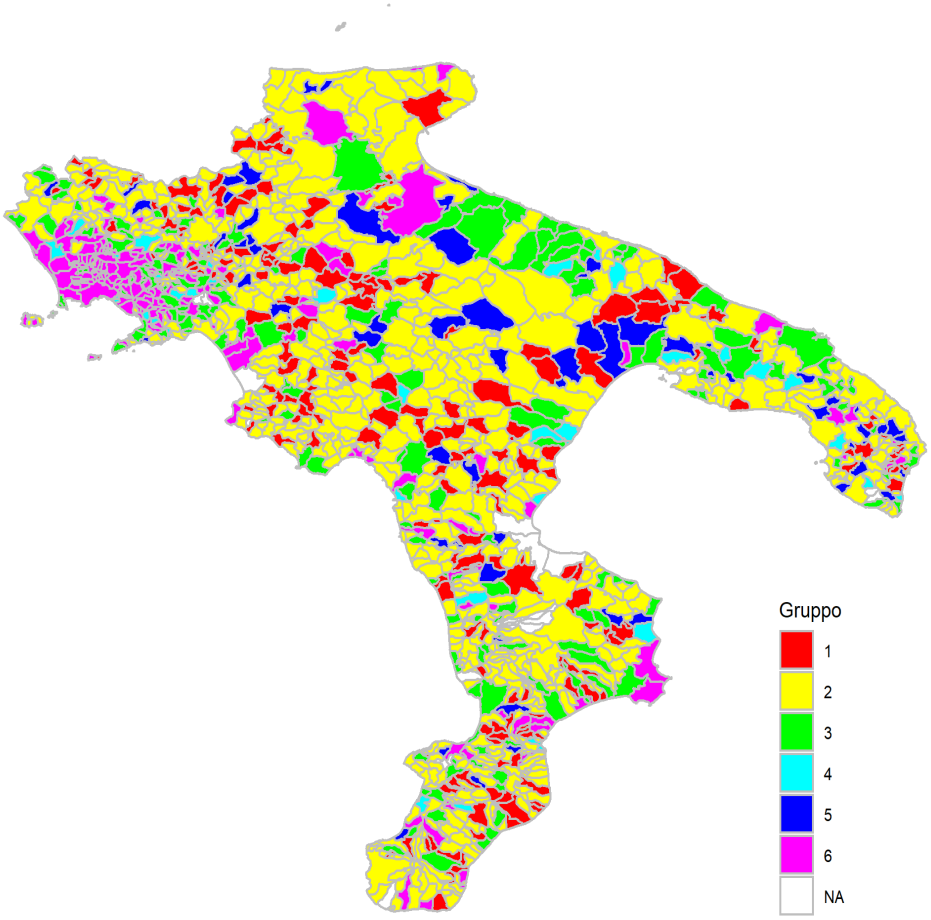


Figura C.6: Gruppo di appartenenza di ciascun comune per le regioni Campania, Puglia, Basilicata e Calabria.

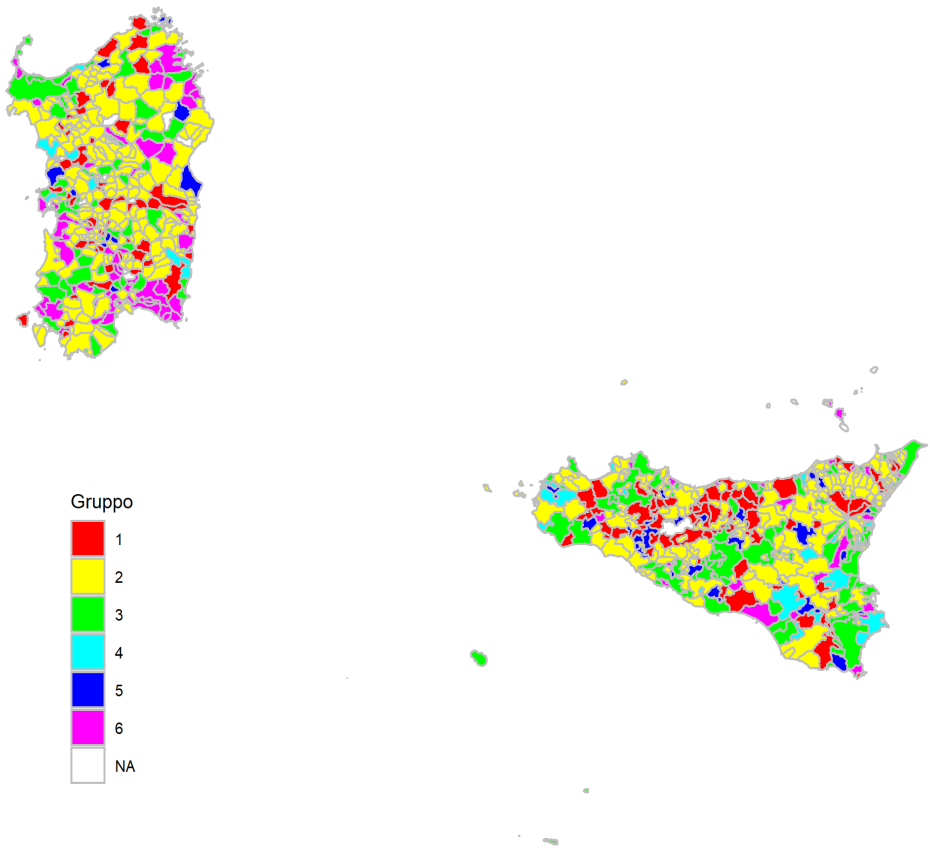


Figura C.7: Gruppo di appartenenza di ciascun comune per le regioni Sicilia e Sardegna.

Bibliografia

Aliverti, E., Mazzuco, S. e Scarpa, B. (2021). «Dynamic modeling of mortality via mixtures of skewed distribution functions». *arXiv preprint arXiv:2102.01599*.

Antoniak, C. E. (1974). «Mixtures of dirichlet processes with applications to bayesian nonparametric problems». *The Annals of Statistics*, pp. 1152–1174.

ARPA Campania (2021). *L'ARPAC e la Terra dei Fuochi*. URL: https://www.arpacampania.it/terra-dei-fuochi. Ultima visita il 29/08/2021.

Azzalini, A. (2013). *The skew-normal and related families*. Cambridge University Press.

Bernardo, J.M. (1994). «Bayesian Statistics». *Probability and Statistics* 2, pp. 345–407.

Blackwell, D., MacQueen, J. B. et al. (1973). «Ferguson distributions via Pólya urn schemes». *The Annals of Statistics* 1, pp. 353–355.

Curzio, S. (2020). «Analisi funzionale di curve di mortalità con un approccio bayesiano non parametrico». Tesi di Laurea Magistrale. Università degli Studi di Padova.

Durante, D., Dunson, D. B. e Vogelstein, J. T. (2017). «Nonparametric bayes modeling of populations of networks». *Journal of the American Statistical Association* 112, pp. 1516–1530.

Emilidha, W. P. e Danardono (2017). «Modelling hospital mortality data using the Heligman-Pollard model with R HPBayes». *AIP Conference Proceedings*. Vol. 1827. AIP Publishing LLC, p. 020025.

Ferguson, T. S. (1973). «A bayesian analysis of some nonparametric problems». *The Annals of Statistics*, pp. 209–230.

Gelman, A. et al. (2013). *Bayesian data analysis*. CRC press.

Graunt, J. (1977). «Natural and political observations mentioned in a following index, and made upon the bills of mortality». *Mathematical Demography*, pp. 11–20.

Heligman, L. e Pollard, J. H. (1980). «The age pattern of mortality». *Journal of the Institute of Actuaries* 107, pp. 49–80.

Hjort, N. L. et al. (2010). *Bayesian nonparametrics*. Cambridge University Press.

Istituto nazionale di Statistica (2020). *Popolazione residente al 1° gennaio 2020*. URL: http://demo.istat.it/popres/index.php?anno=2020&lingua=ita. Ultima visita il 06/08/2021.

– (2021). *Decessi e cause di morte: cosa produce l'Istat*. URL: www.istat.it/it/archivio/240401. Ultima visita il 15/07/2021.

Istituto Superiore di Sanità e Procura della Repubblica di Napoli Nord (2020). *Studio sull'impatto sanitario degli smaltimenti controllati ed abusivi di rifiuti nei 38 comuni del circondario della Procura della Repubblica di Napoli Nord*. URL: https://www.procuranapolinord.it/allegatinews/A_42658.pdf.

Livi Bacci, M. (1999). *Introduzione alla demografia*. Torino: Loescher Editore.

– (2020). «Le conseguenze demografiche della Prima Guerra Mondiale». *Geodemografia 2019. 15 scritti per meglio comprendere il mondo*.

MacQueen, J. et al. (1967). «Some methods for classification and analysis of multivariate observations». *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. Oakland, CA, USA, pp. 281–297.

Makeham, W. M. (1860). «On the law of mortality and the construction of annuity tables». *Journal of the Institute of Actuaries* 8, pp. 301–310.

Mazzuco, S., Scarpa, B. e Zanotto, L. (2018). «A mortality model based on a mixture distribution function». *Population studies* 72, pp. 191–200.

Müller, P. et al. (2015). *Bayesian nonparametric data analysis*. Springer.

Rousseau, J. e Mengersen, K. (2011). «Asymptotic behaviour of the posterior distribution in overfitted mixture models». *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, pp. 689–710.

Sethuraman, J. (1994). «A constructive definition of Dirichlet priors». *Statistica sinica*, pp. 639–650.

Sharrow, D. J. (2015). *Package* `HPbayes`. URL: https://cran.r-project.org/web/packages/HPbayes.

Siler, W. (1979). «A competing-risk model for animal mortality». *Ecology* 60, pp. 750–757.

– (1983). «Parameters of mortality in human populations with widely varying life spans». *Statistics in medicine* 2, pp. 373–380.

University of California, Berkeley e Max Plank Institute for Demographic Research (2013). *Human Mortality Database*. URL: www.mortality.org. Ultima visita il 19/05/2021.

Ringraziamenti

Giunto alla conclusione di questi cinque anni ritengo doveroso ringraziare chi mi ha aiutato in maniera particolare durante questo percorso.

Innanzitutto vorrei ringraziare il Prof. Scarpa e il Prof. Rigon che mi hanno seguito con grande disponibilità durante la stesura dell'elaborato e mi hanno fatto scoprire ed appassionare ad un ambito della Statistica non trattato nei corsi di laurea magistrale.

È d'obbligo un ringraziamento particolare a Andrea, Danny e Matteo, compagni conosciuti all'inizio del percorso di studi con i quali è nato un legame di amicizia cementato nel corso degli anni anche al di fuori dell'Università.

Desidero poi ringraziare Federica per essermi stata accanto durante quest'ultimo anno ed avermi aiutato nei momenti di difficoltà.

Ringrazio mia sorella Silvia, la persona con la quale litigo più spesso ma sulla quale so che posso sempre contare.

Infine, un ringraziamento speciale va ai miei genitori Simonetta e Dario, dai quali tanto ho imparato e tanto ho ancora da imparare. Spero di aver ripagato una parte dei sacrifici fatti per avermi permesso di studiare.