

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in Statistica per le tecnologie e le
scienze



RELAZIONE FINALE

**STIMA DEL TASSO DI FECONDITÀ TOTALE DEI COMUNI
ITALIANI A PARTIRE DALLA STRUTTURA PER ETÀ**

Relatore Prof. Stefano Mazzuco
Dipartimento di Scienze Statistiche

Laureando Chiara Degan
Matricola N 1167278

Anno Accademico 2019/2020

Indice

Introduzione	5
1 I metodi di stima	7
1.1 iTFR: Implied Total Fertility Rate	11
1.2 xTFR: Extended Total Fertility Rate	11
1.3 iTFR ⁺ e xTFR ⁺	12
1.4 bTFR: Bayesian Total Fertility Rate	12
2 Descrizione dei dati	17
3 Analisi dei dati	23
3.1 iTFR e xTFR	23
3.2 iTFR ⁺ e xTFR ⁺	28
3.3 bTFR	30
Conclusioni	35
Bibliografia	43
A Codice R utilizzato	45

Introduzione

Per analizzare le dinamiche di una popolazione è necessario studiare non solo l'andamento della mortalità e della migratorietà, ma anche della fecondità. È infatti noto che le variazioni all'interno delle popolazioni possono essere ricondotte o a fenomeni migratori o a oscillazioni del numero di nascite e di morti. Di seguito viene prestata particolare attenzione proprio alla fecondità, essendo la causa preliminare della crescita o della decrescita della popolazione. È infatti un argomento che sta entrando a far parte sempre di più di discussioni politiche ed economiche, principalmente perché i livelli di fecondità continuano a decrescere di anno in anno, soprattutto in Europa. Ciò sta portando a un'invecchiamento repentino della popolazione e quindi a una diminuzione della stessa. In particolare ci si focalizzerà sullo studio dei dati italiani, essendo lo stato con indici di fecondità più bassi in Europa e nel mondo.

Nonostante si stiano perfezionando le tecniche di raccolta dati, rendendole sempre più attendibili e accurate, in numerose situazioni risulta ancora complicato calcolare gli indici di fecondità con metodi tradizionali. Primo fra tutti è il Tasso di Fecondità Totale (TFT), il quale, anche se molto semplice e intuitivo, richiede la conoscenza del numero di nati per classe di età della madre. Per piccole estensioni territoriali, come i comuni italiani, questi dati disaggregati non sono disponibili e perciò, fino ad ora, non è ancora stato possibile stimare il numero medio di figli per donna. Per anni sono stati proposti metodi alternativi per calcolare il TFT e da uno studio recente, pubblicato a gennaio dell'anno corrente (Schmertmann & Hauer, [2020](#)), vengono suggeriti

cinque nuovi metodi. Tra questi, tre sono inediti, due invece sono stati ricavati da studi precedenti (Hauer et al., 2013; Schmertmann & Hauer, 2019) e perfezionati successivamente. L'innovazione di tali indici sta nel fatto che sono basati non più sui tassi specifici di fecondità, ma sulla struttura per età della popolazione considerata. In questo modo risulta possibile fornire una misura della fecondità anche per popolazioni ridotte o per cui i soli dati disponibili sono quelli per sesso ed età. Tali indici verranno esposti nel dettaglio nei capitoli più avanti.

In seguito si è deciso di applicare questi nuovi metodi di stima per ricavare informazioni riguardo la fecondità dei comuni italiani. In generale è risaputo che il Tasso di Fecondità Totale in Italia sta diminuendo sempre di più, tanto da essere lo stato europeo con indice più basso. In particolare, nel 2019, il TFT dell'intero paese è risultato essere pari a 1.29, molto al di sotto della soglia di rimpiazzo e pari solamente alla Spagna. Inoltre, con i dati resi disponibili dall'Istat, è anche possibile effettuare un'analisi più fine a livello territoriale in riferimento alle ripartizioni Nord, Centro e Sud, alle regioni e alle province. Per esempio, sempre relativamente al 2019, si è constatato che la fecondità più alta si manifesta nel settentrione¹ (TFT pari a 1.36), con anche cospicue differenze nelle varie province (Istat, 2019a).

Ciò che invece non è possibile fare è una descrizione analitica dei comuni perchè continuano a essere inaccessibili i tassi specifici di fecondità. Avendo però a disposizione i dati suddivisi per sesso ed età, è stato possibile applicare i suddetti metodi e ricavare delle stime dell'intensità della fecondità. In particolare ci si focalizzerà sull'elaborazione dei dati relativi al 2019 essendo la raccolta disponibile più recente. Tutto questo con lo scopo di capire le diverse dinamiche demografiche del paese e sottolineare se c'è differenza o meno tra comuni di varie regioni e province. Non ci si sofferma quindi sull'interpretazione sociale e storica dei risultati, perchè ambiti non di competenza, ma solamente sul calcolo e l'interpretazione numerica.

¹Regioni del Nord-Est (Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige, Veneto) e del Nord-ovest (Liguria, Lombardia, Piemonte, Valle d'Aosta)

Capitolo 1

I metodi di stima

Come già anticipato nell'introduzione, i metodi di stima che saranno presentati in questo capitolo dipendono dalla struttura per età della popolazione d'interesse. Essi fanno parte di una grande famiglia di stimatori che ricavano informazioni sfruttando le relazioni esistenti tra le varie quantità demografiche conosciute, facendo riformulazioni algebriche di indici generalmente utilizzati o proponendo metodi alternativi per incrementare dati incompleti. In letteratura si dice che tali stimatori utilizzano metodi indiretti di stima e sono particolarmente adoperati quando l'informazione disponibile è scadente o insufficiente.

Per quanto riguarda il calcolo del Tasso di fecondità Totale, si sa che è necessario conoscere i tassi specifici per età. Quest'ultimi però non sono sempre disponibili soprattutto in paesi o regioni con registi di nascite poco dettagliati, in popolazioni che per loro tradizione non raccolgono dati (e.g. gli aborigeni australiani, o le popolazioni indigene in Africa e altri stati in via di sviluppo) o in zone geografiche talmente piccole per cui i dati vengono presentati solamente in forma aggregata. Per codesto motivo, negli anni, sono stati proposti numerosi metodi indiretti di stima del TFT, ma anche con taluni non sono mancati i problemi. Spesso infatti, questi stimatori, si basano sull'informazione derivante da altre variabili, come la percentuale di donne sposate e l'età media al matrimonio. Noto è il modello di regressione

Bogue-Palmore (Bogue & Palmore, 1964; Palmore, 1968) che stima il TFT sfruttando la sua relazione con il numero di bambini di 0-4 anni, con i tassi di mortalità e con il numero di donne sposate per fascia di età. Il problema principale di tale metodo è che, con il passare del tempo, le relazioni cambiano velocemente e di conseguenza diminuisce la validità dell'equazione di regressione ottenuta da Bogue-Palmore. Un altro esempio è il modello coniato da Rele nel 1967 (Arriga et al., 1994), nel quale si assume che, a livelli fissati di mortalità, esiste una relazione lineare tra il TFT e la proporzione di bambini di 0-4 anni per numero di donne tra 15-49 anni. In questo caso intercetta e coefficiente angolare sono già definiti e cambiano a seconda del valore della speranza di vita alla nascita (considerata tra 20 e 70 anni con intervalli di dieci anni).

In entrambi i casi vengono quindi sfruttate informazioni che mutano molto nel tempo e nello spazio, rendendo così gli stimatori utili solamente nei specifici contesti in cui sono stati pensati. Tale motivo li rende più soggetti a errore rispetto ai nuovi metodi di stima proposti da Schmertmann e Hauer nel 2020, per i quali è stata dimostrata la robustezza e, per alcuni di loro, anche la maggiore correttezza (Hauer et al., 2013). Essi per l'appunto si basano, come già menzionato, principalmente sulla relazione che sussiste tra TFT e struttura per età della popolazione. Qui di seguito viene descritta tale relazione in termini matematici per poter poi comprendere meglio la natura degli stimatori.

Prima di tutto è necessario fare alcune considerazioni che riguardano il comportamento generale di una popolazione. È importante tenere presente che la mortalità infantile in genere è molto elevata soprattutto nel primo anno di vita, perciò non è sempre detto che un bambino nato nell'anno x sia ancora vivo alla fine del periodo di conteggio. Allo stesso modo è possibile che anche la madre, che ha partorito nell'anno x , muoia prima della fine dell'anno o della classe pluriennale che si sta considerando. Inoltre è ragionevole pensare che una donna stia all'interno dell' n -esimo intervallo solo per una frazione di tempo, e che poi, se sopravvissuta, passi alla classe di età successiva.

Si assuma infine che l'intervallo di età fertile della donna vada dai 15 ai 49 anni e che le classi di età considerate siano quinquennali, allora il numero medio di figli sopravvissuti al di sotto dei 5 anni, per numero di donne sopravvissute appartenenti alla classe di età a , è

$$\begin{aligned}
 K_a &= \left[\frac{L_{a-5}}{L_a} \cdot F_{a-5} + F_a \right] \frac{L_0}{2} \\
 &= TFT \cdot \frac{L_0}{5} \cdot \frac{1}{2} \left(\frac{L_{a-5}}{L_a} \cdot \phi_{a-5} + \phi_a \right) \\
 &= TFT \cdot s \cdot p_a
 \end{aligned} \tag{1.1}$$

Tale formula deriva da un riarrangiamento della prima riga della matrice di Leslie e i suoi elementi sono:

- F_a tassi specifici di fecondità relativi alle classi di età $[a, a + 5)$;
- $TFT = \sum_a 5F_a$ Tasso di Fecondità Totale;
- $\phi_a = \frac{5F_a}{TFT}$ proporzione di nati da donne in età a rispetto al totale. Per le assunzioni fatte precedentemente $\phi_{10} = 0$ nel calcolo di K_{15} . Inoltre, anche se $\phi_{45} > 0$, si presuppone che $K_{50} = 0$;
- L_a tempo di esposizione al rischio per le donne nella classe di età a in una tavola con radice $l_0 = 1$;
- $s = \frac{L_0}{5}$ probabilità di sopravvivenza tra i bambini di 0-4 anni. Tale valore è identico per tutte le classi di età a della madre;
- p_a media della fecondità tra $a - 5$ e a con peso leggermente maggiore per i primi anni in modo da tenere conto della possibile morte della madre. Tale fattore quindi varia a seconda dell'età, è circa zero per $a = 15$ e $a = 45$, ed ha valori più elevati per $a = 20$ e $a = 30$;

Sapendo che W_a è il numero di donne nella classe di età $[a, a + 5)$ e quindi che $W = \sum_a W_a$ è il numero totale di donne di 15-49 anni, il numero totale

di bambini sopravvissuti tra gli 0-4 anni è

$$\begin{aligned}
 C &= \sum_a W_a K_a \\
 &= TFT \cdot s \cdot \left(\sum_a W_a p_a \right) \\
 &= TFT \cdot s \cdot p \cdot W
 \end{aligned} \tag{1.2}$$

Dove $p = \sum W_a p_a / W$ è una media pesata dei tassi specifici p_a i cui pesi sono la struttura per età W_a / W . Di conseguenza il Tasso di Fecondità Totale può essere calcolato moltiplicando il tasso generico di fecondità (C/W) per un *child mortality multiplier* ($1/s$) e un *age-structure multiplier* ($1/p$):

$$TFT = \frac{1}{s} \cdot \frac{1}{p} \cdot \frac{C}{W} \tag{1.3}$$

Si assume cioè che il numero medio di nati in una piramide di età può essere approssimato con il numero di bambini tra i 0-4 anni, nati da donne appartenenti alla stessa piramide, tenendo conto del valore della mortalità infantile. Essendo C e W derivanti dalla stessa piramide d'età, allora qualsiasi cambiamento dovuto a migrazioni è poco rilevante. Perciò la relazione descritta dalla formula 1.3 continua a valere anche per popolazioni aperte.

Qui di seguito verranno descritti più nel dettaglio i cinque metodi di stima utilizzati nelle analisi per i comuni italiani: $iTFR$, $xTFR$, $iTFR^+$, $xTFR^x$ e $bTFR$ (in ordine dal più semplice al più elaborato). Avendo a disposizione solamente i dati riguardanti la struttura per età è possibile calcolare i primi due, per gli altri invece è necessario ricavarsi i valori di ${}_5q_0$.

È stato inoltre dimostrato (Schmertmann & Hauer, 2020) che, all'aumentare della complessità dello stimatore, diminuisce l'errore assoluto. Questo è dovuto al fatto che si aggiungono dettagli riguardanti sia la struttura per età sia la mortalità che fanno in modo che le stime siano più precise. Sottoponendoli poi a varie situazioni demografiche, e utilizzando dati ricavati da diverse fonti, si è verificata anche la loro robustezza.

1.1 iTFR: Implied Total Fertility Rate

Il primo stimatore, ed il più semplice, è chiamato *Implied Total Fertility Rate* dal fatto che si ricava implicitamente dalla struttura per età. Esso era già stato presentato in un altro studio (Hauer et al., 2013), nel quale veniva ricavato attraverso un riarrangiamento algebrico del Tasso Generale di Fecondità (GFT).

Basandosi invece sull'equazione 1.3 risulta essere pari a:

$$\text{iTFR} = 7 \cdot \frac{C}{W} \quad (1.4)$$

In tale stimatore si assume che la mortalità infantile sia nulla, e quindi che $s \approx 1$, e che le donne in età fertile siano uniformemente distribuite all'interno delle classi di età. Considerando classi quinquennali all'interno dell'intervallo di età fertile $[15, 50)$ si ottiene che $p = 1/7$.

1.2 xTFR: Extended Total Fertility Rate

Con il secondo stimatore si allenta l'assunzione di equidistribuzione delle donne nelle classi di età. È importante tenere conto non solamente della suddivisione entro le classi, ma anche del numero di donne presenti nella classe di età più fertile $[25, 34]$. Si apporta quindi un aggiustamento riguardo il parametro $1/p$ in modo tale che sia più elevato se le donne tra i 25-34 anni sono poche, e viceversa.

Da un calcolo empirico su dati ricavati dallo Human Fertility Database (HFD), condotto in precedenza dai due autori dell'articolo (Schmertmann & Hauer, 2019), lo stimatore risulta essere pari a:

$$\text{xTFR} = (10.65 - 12.55 \pi_{25-34}) \cdot \frac{C}{W} \quad (1.5)$$

dove π_{25-34} è la proporzione di donne in età 25-34 sul totale di donne presenti

in età fertile. Ciò di cui bisogna fare attenzione è che i coefficienti 10.65 e -12.55 sono attendibili solamente se l'intervallo di età fertile considerato è $[15, 50)$.

1.3 $iTFR^+$ e $xTFR^+$

Nei due stimatori trattati precedentemente è stata assunta mortalità circa pari a zero, questo significa che quando i livelli di mortalità sono elevati il TFT viene sottostimato e aumenta l'errore di stima.

Negli stimatori descritti di seguito si fa un aggiustamento per ovviare tale problema:

$$iTFR^+ = \left(\frac{7}{1 - 0.75 {}_5q_0} \right) \cdot \frac{C}{W} \quad (1.6)$$

$$xTFR^+ = \left(\frac{10.65 - 12.55 \pi_{25-34}}{1 - 0.75 {}_5q_0} \right) \cdot \frac{C}{W} \quad (1.7)$$

dove ${}_5q_0$ è la probabilità di morte entro i primi 5 anni di vita, dalla cui applicazione risulta che il parametro $1/s \approx 1/(1 - 0.75 {}_5q_0)$. Tale approssimazione è stata ipotizzata prima di tutto a partire dalla relazione logica esistente tra gli elementi della tavola di mortalità. In seguito poi è stata verificata attraverso un calcolo empirico basato sui dati della Svezia, essendo lo stato con la più lunga serie storica (266 anni) disponibile all'interno dello Human Mortality Database (HMD).

1.4 bTFR: Bayesian Total Fertility Rate

Diversamente dai quattro stimatori precedenti, che sono deterministici, il quinto è derivato da un modello statistico bayesiano. Anch'esso si basa sulla struttura per età della popolazione, infatti fornisce una stima del TFT partendo dalla sua relazione con i parametri C (numero di bambini tra 0-4 anni - formula 1.2) e $W = (W_{15}, \dots, W_{45})'$. Una volta definita la distribuzione

a posteriori del TFT condizionato ai valori di C e W , lo stimatore bTFR viene poi definito come valore mediano di tale distribuzione.

Per definizione, la distribuzione a posteriori in un modello bayesiano è proporzionale al prodotto tra la funzione di verosimiglianza e le distribuzioni a priori dei parametri di interesse:

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

Nel nostro caso il modello ha cinque parametri: TFT , $\beta = (\beta_1, \beta_2)$, ${}_5q_0$, k . I primi tre fanno riferimento alla fecondità, gli ultimi due invece alla mortalità. Di seguito viene descritto come sono state elicitate le distribuzioni a priori per ogni singolo parametro.

Parametri di fecondità

Si consideri il vettore di parametri $\phi = (\phi_{10}, \phi_{15}, \dots, \phi_{45})$ dell'equazione (1.1): essendo proporzioni la loro somma sarà pari a uno e perciò possono essere tutte riscritte come quote $\gamma_a = \ln\left(\frac{\phi_a}{\phi_{15}}\right)$. Riuscendo quindi a stimare i parametri γ_a si riescono a ricavare le relative probabilità $\phi_a(\gamma) = \frac{\exp(\gamma_a)}{\sum_z \exp(\gamma_z)}$. Utilizzando dati ricavati da HFD, sempre in maniera empirica, si ricava il modello $\gamma = m + X\beta$ (Schmertmann & Hauer, 2019) dove $m \in \mathbb{R}^7$ e $X \in \mathbb{R}^{7 \times 2}$ sono costanti e $\beta \in \mathbb{R}^2$ sono i coefficienti che riscalano le colonne della matrice X . Quest'ultimi vengono assunti a media e covarianza nulle, a varianza unitaria e limitati all'intervallo di valori $[-2, +2]$.

In conclusione quindi si ha:

$$X = \begin{pmatrix} 0 & 0.27 & 0.54 & 0.73 & 0.88 & 1.04 & 1.52 \\ 0 & 0.32 & 0.51 & 0.51 & 0.35 & 0.05 & -0.72 \end{pmatrix}'$$

$$m = \begin{pmatrix} 0 & 1.39 & 1.59 & 1.23 & 0.45 & 89 - 0. & -3.44 \end{pmatrix}'$$

$$\beta \sim N(0, I_2)$$

Si noti che le righe del vettore m in corrispondenza delle classi di età [20, 25) e [25, 30) hanno dei valori più elevati perchè corrispondono alle classi di età le cui donne sono più fertili. Inoltre, essendoci nella prima colonna della matrice X valori crescenti, si può dire che il primo elemento del vettore β influenza l'età media al parto, e perciò a valori alti di β_1 corrisponde un posticipo delle nascite. Valori elevati di β_2 , invece, implicano la presenza di una maggiore concentrazione di nascite negli anni 20-29. Anche qui si noti che nella seconda colonna della matrice X si hanno costanti più grandi in corrispondenza delle classi di età più fertili.

Infine, per quanto riguarda il parametro TFT si assume che si distribuisca come una uniforme su intervallo $[0, 20]$. Tale assunzione non influisce affatto sul risultato finale, anzi così facendo si fa in modo che il risultato della stima sia completamente determinato dai dati e non da assunzioni a priori.

Parametri di mortalità

Per modellare la mortalità delle donne e dei bambini si adotta un modello relazionale (Wilmoth et al., 2012) che stima i tassi di mortalità basandosi su due parametri indipendenti: la probabilità di morire prima del compimento dei 5 anni ${}_5q_0$ e un parametro di forma k . Inoltre nel modello sono presenti 4 costanti che variano a seconda del sesso a_x, b_x, c_x, v_x , ricavate sempre empiricamente dall'utilizzo di dati dello HMD.

Tale modello è scritto nella forma

$$\ln \mu_x({}_5q_0, k) = a_x + b_x[\ln {}_5q_0] + c_x[\ln {}_5q_0]^2 + v_x k \quad x = 0, 1, 5, \dots, 45. \quad (1.8)$$

Una volta stimati i tassi di mortalità specifici per età, si sfrutta la loro relazione con il numero di sopravvivententi l_x , per riuscire a ricavare poi i tempi di esposizione al rischio L_x . In particolare si utilizza la definizione della funzione di sopravvivenza

$$l_{x+n} = l_x \cdot e^{-\int_x^{x+n} \mu(x) dx}$$

da cui deriva che $l_0 = 1$, $l_1 = e^{-\mu_0}$, $l_{x+5} = l_x \cdot e^{-n\mu_x}$.

I valori di L_x invece si ottengono attraverso il metodo dell'approssimazione

trapezoidale da cui risulta che $L_0 = \frac{1}{2}(l_0 + l_1) + \frac{4}{2}(l_1 + l_5)$ e $L_x = \frac{5}{2}(l_x + l_{x+n})$ con $x = 5, \dots, 45$ (Schmertmann e Hauer, 2019).

Come distribuzione a priori per le probabilità si assume che ${}_5q_0 \sim \text{Beta}(a, b)$, dove a e b sono scelti in maniera tale che $P[{}_5q_0 < 1/2\min({}_5\hat{q}_0)] = 0.05$ e $P[{}_5q_0 < 2\max({}_5\hat{q}_0)] = 0.95$. In questo modo si ha il 90% di probabilità che il vero valore del parametro sia compreso tra metà del minimo valore stimato e il doppio del massimo valore stimato.

Per il parametro k invece si assume che sia limitato entro l'intervallo $[-2, +2]$ e che si distribuisca come una normale standard trocata $k \sim N(0, 1)$.

Modello completo

Ciò che manca per completare il modello è la funzione di verosimiglianza. Per poterla calcolare è necessario determinare la famiglia di distribuzioni di cui fa parte il campione, nel nostro caso quindi bisogna trovare una distribuzione di probabilità per il numero totale di bambini sopravvissuti (C). Si assuma che $C_a = W_a K_a$ si distribuisca come una Poisson e che i valori delle diverse classi, condizionati ai tassi di fecondità e di mortalità, siano statisticamente indipendenti. Di conseguenza si può affermare che anche $C = \sum_a C_a$ si distribuisce come una Poisson con parametro $\sum_a W_a K_a$:

$$C|W, TFT, \beta, {}_5q_0, k \sim \text{Poisson}[\sum_a W_a K_a(TFT, \beta, {}_5q_0, k)]$$

La distribuzione di Poisson esprime la probabilità che si verifichi un certo numero di eventi in uno specifico dominio spazio-tempo, sapendo che mediamente se ne verifica un numero pari a λ . Risulta essere quindi perfetta per l'evento qui considerato perchè nel caso in cui $\sum_a W_a K_a$ sia piccola, viene assegnata una maggiore probabilità a numeri più vicini allo zero, e viceversa. In letteratura è infatti conosciuta come "distribuzione degli eventi rari".

Definiti tutti gli elementi utili per determinare la distribuzione a posteriori del TFT condizionato ai dati ricavati dalla struttura per età, risulta che:

$$P(TFT|C) \propto \int L(C|TFT, \beta, {}_5q_0, k) f_\beta(\beta) f_q({}_5q_0) f_k(k) d\beta d{}_5q_0 dk \quad (1.9)$$

Una volta definita la distribuzione del TFT, nell'articolo (Schmertmann & Hauer, 2020) propongono di procedere con il campionamento applicando l'algoritmo Markov Chain Monte Carlo (MCMC) ma, essendo un argomento non ancora trattato, si è deciso di proseguire in un'altra maniera. Si è utilizzata la procedura veloce suggerita da Schmertmann nel 2019 (si veda il codice scaricabile all'indirizzo <https://github.com/mathewhauer/iTFR-replication>) che permette di ricavare in modo empirico la distribuzione a posteriori senza utilizzare Stan (l'intero codice della funzione è riportato anche in Appendice A.2).

Capitolo 2

Descrizione dei dati

Il territorio italiano è ripartito in 20 regioni, ognuna delle quali è suddivisa in province e città metropolitane che comprendono a loro volta una serie di unità territoriali ancora più piccole, i comuni. Quest'ultimi variano di anno in anno perchè vengono approvate nuove fusioni o incorporazioni con lo scopo di aumentare l'efficienza all'interno delle unità amministrative e ridurre i costi. Dal 1991 infatti il numero è diminuito sempre di più fino a raggiungere un totale di 7903 comuni il 1° luglio 2020. Per quanto riguarda invece province e città metropolitane il numero rimane costante, rispettivamente pari a 93 e 14.

È quindi importante tenere conto del periodo temporale in cui sono stati rilevati i diversi dati, in maniera tale che si faccia riferimento alle stesse numerosità e si evitino errori. In particolare, per tutte le analisi effettuate nel documento, sono stati considerati i dati rilevati al 1° gennaio 2019, data in cui erano presenti 7926 comuni. Tali dati sono stati scaricati da <http://demo.istat.it/>, un sito reso disponibile dall'Istat contenente tutti i dati anagrafici più recenti riguardanti la popolazione italiana. Al suo interno è possibile trovare tabelle di dati relativi alla popolazione straniera residente suddivisi per età e sesso o anche il Bilancio demografico mensile. Ci sono inoltre informazioni più dettagliate riguardo i fenomeni demografici, come i tassi di natalità, la tavola di mortalità, l'indice di vecchiaia o l'età media al

parto.

Qui di seguito verranno descritti più nel dettaglio i dati utilizzati e verranno giustificate le scelte che si sono fatte per la loro selezione. Prima di tutto si è considerata la raccolta dati relativa alla popolazione residente in ogni comune, suddivisa per sesso, età e stato civile. Questi dati provengono da un'indagine che l'Istat conduce dal 1992 presso le Anagrafi dei comuni italiani, il quale precisa che i "dati per età, a livello comunale, sono oggetto di una procedura di validazione statistica qualitativa e quantitativa, atta a superare le incongruenze che a volte si manifestano in talune anagrafi comunali, a causa del non perfetto allineamento delle stesse con le risultanze derivanti dall'ultimo censimento". Tale avvertenza sarà utile poi per un corretto utilizzo dei dati e una giusta interpretazione dei risultati. Il dataset in questione quindi non contiene solamente dati relativi alla struttura per età ma anche alle modalità dello stato civile (uniti civilmente, coniugati, divorziati, vedovi, nubili/celibi). Non essendo però variabili d'interesse si decide di eliminarle e di semplificare la raccolta in modo da rendere più leggero e comprensibile il file. In conclusione si ottiene un documento con cinque variabili: la denominazione del comune, l'età dell'individuo (classi annuali), il numero di maschi e di femmine per fascia di età e il codice alfanumerico del comune. A ogni unità territoriale, quindi anche a province e regioni, l'Istat ha infatti attribuito un codice che viene aggiornato in seguito ad ogni minima variazione amministrativa. Essi sono così definiti (si veda "Elenco codici statistici e denominazioni gennaio 2019" - Istat, [2020a](#)):

- regione: codice di due caratteri alfanumerici con range 01-20;
- provincia: codice di tre caratteri in formato alfanumerico con range 001-111. Per comodità e per fini statistici si sono mantenuti gli stessi codici anche con l'entrata in vigore delle città metropolitane. Per lo stesso motivo, non sono stati eliminati i codici relativi alle quattro province del Friuli-Venezia Giulia sopresse nel 2016. Sono invece state escluse le province dal 104 al 107 poichè sono state incorporate ad altre;

- comune: il codice si ottiene concatenando il codice della provincia con un numero progressivo che parte da 001 all'interno della provincia stessa. Dato che negli anni il numero di comuni è diminuito sempre di più, alcuni codici sono stati eliminati e perciò non è possibile stabilire un intervallo continuo dei valori.

Ricapitolando quindi, nel primo dataset considerato, sono disponibili i dati disaggregati per sesso e fascia di età. Essi perciò possono essere utilizzati per calcolare il numero totale di bambini di età compresa tra gli 0-4 anni e il numero di donne appartenenti all'intervallo di età fertile [15, 50).

Avendo a disposizione solamente questi dati è possibile determinare i primi due stimatori $iTFR$ e $xTFR$ nei quali si assume che la mortalità infantile sia nulla. Per poter misurare i successivi due, $iTFR^+$ e $xTFR^+$, è invece necessario conoscere la probabilità di morte tra gli 0-4 anni ${}_5q_0$. Tale valore non è già disponibile tra i dati forniti dall'Istat quindi verrà stimato, ma per giustificare le scelte fatte è necessario fare prima delle considerazioni.

Come già specificato, l'Italia è il paese europeo con indici di fecondità più bassi. Al contrario, il numero di decessi sta diminuendo sempre di più tanto che nel 2019 è stato registrato un saldo naturale pari a -212mila unità (Istat, 2019a). Questo vuol dire che per ogni 100 morti, il ricambio è assicurato da appena 67 nascite. La speranza di vita alla nascita sta via via crescendo tanto da essere pari a 85,3 anni per le donne e 81 anni per gli uomini, come anche l'età media della popolazione uguale a 45,7 anni al 1° gennaio 2020. Ciò è dovuto principalmente all'innalzamento del livello di istruzione, all'aumento dell'efficienza dei servizi sanitari e ad un miglioramento delle condizioni di vita. Tutto questo comporta infatti una maggiore probabilità di sopravvivenza soprattutto nelle prime fasce di età dove generalmente si registra un numero elevato di decessi. Ormai da tempo infatti il livello di mortalità infantile in Italia è tra i più bassi non solamente in Europa ma in tutto il mondo.

Per tale motivo ci si aspetta che la probabilità di morte a livello provinciale possa costituire una ragionevole approssimazione dei valori comunali. È inoltre importante tenere conto che l'estensione di alcuni comuni è talmente

piccola che il numero di bambini è decisamente inferiore alla media. Questo comporterebbe una maggiore variabilità delle stime e quindi una minore efficienza. Si è voluto comunque verificare codesta assunzione sfruttando i dati giornalieri sui decessi, suddivisi per regione, provincia e comune, resi disponibili dall'Istat (si veda "Dati giornalieri sulla mortalità dei comuni al 31 maggio 2020" - Istat, 2020c). All'interno di tale raccolta si trova il numero totale di morti, divisi per sesso e classi di età, relativi ai giorni degli eventi per il 93.1% dei comuni italiani. Sfruttando tali dati si è calcolato il numero di decessi totali nei comuni facendo una distinzione tra età 0 e classe di età [1, 4]. Si è subito notato che la maggiorparte di essi erano nulli, e che altri invece risultavano esageratamente elevati. Questo sicuramente, come detto prima, avrebbe portato a delle distorsioni considerevoli perciò si è deciso di utilizzare i valori provinciali di mortalità.

Quest'ultimi sono stati ricavati dalle tavole di mortalità scaricabili dal sito <http://demo.istat.it/>. In esse, oltre ai valori della probabilità di morte ${}_5q_0$ (moltiplicati per mille), sono disponibili anche il numero di sopravvissuti l_x e di decessi d_x , il numero di anni vissuti L_x e la speranza di vita alla nascita e_x . Tutte queste quantità in realtà non verranno direttamente utilizzate perché verranno stimate in maniera differente al momento del calcolo di $bTFR$. Saranno solamente impiegate le probabilità per stimare i valori di $iTFR^+$ e $xTFR^+$.

Dalla figura 2.1 si può vedere la distribuzione delle probabilità in tutto il territorio italiano. Si nota facilmente che la probabilità nelle regioni del Sud Italia, eccetto la Sardegna, sembra essere maggiore rispetto a quella del Nord e del Centro. Ci sono però delle eccezioni come le province di Mantova, Cremona e Bergamo in Lombardia o Biella, Alessandria e Novara in Piemonte, che si distinguono rispetto alle altre per il loro colore più chiaro. Nonostante questo i valori di tutte le province sono comunque molto bassi, esattamente come ci si aspettava.

Anche per disegnare tale mappa si sono utilizzati dati ricavati dal sito dell'Istat (i comandi utilizzati in R sono invece disponibili in Appendice A.1). In

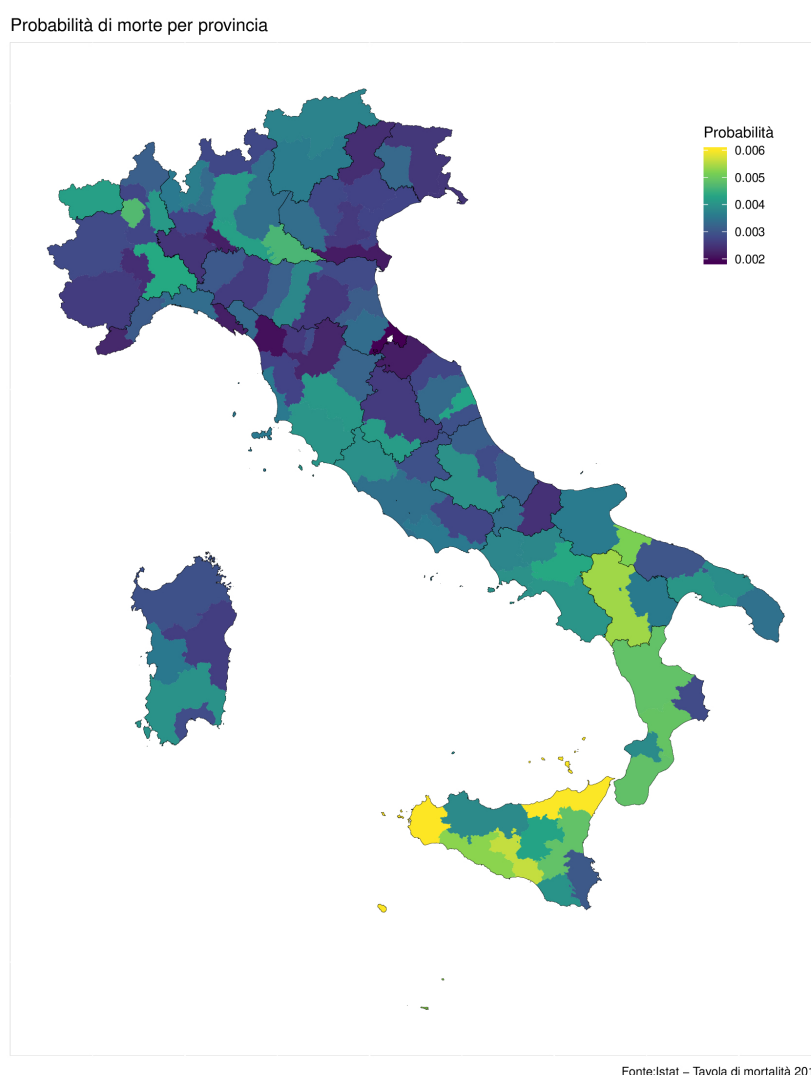


Figura 2.1: Mappa dei valori della probabilità di morire tra gli 0 e 4 anni nelle province e nelle città metropolitane italiane nel 2019.

quanto Istituto Nazionale di Statistica infatti, ogni anno, oltre a pubblicare l'archivio dei codici e delle denominazioni territoriali, mette a disposizione un file con i confini amministrativi aggiornati il 1° gennaio dell'anno corrente. Il file in questione è costituito da dati in formato shapefile, distinti per ripartizione geografica, regione, provincia e comune. Essi comprendono, oltre ai relativi codici alfanumerici e le rispettive denominazioni, anche informazioni

riguardo latitudine e longitudine, vastità dell'area considerata e lunghezza del territorio (per maggiori informazioni si veda il documento di "Descrizione dei dati geografici dei confini delle unità amministrative a fini statistici" - Istat, [2020b](#)).

In conclusione quindi, per il calcolo degli stimatori, è stato necessario scaricare solamente due dataset, uno che ci fornisce i dati riguardanti la struttura per età della popolazione, l'altro che viene utilizzato per ricavare i valori delle probabilità a livello provinciale. Nel capitolo seguente verranno spiegati i procedimenti di calcolo e verrà descritta la maniera in cui sono stati applicati i metodi esposti nel primo capitolo.

Capitolo 3

Analisi dei dati

Le analisi presenti nell'elaborato sono state svolte tramite l'uso del software statistico Rstudio. Prima di tutto si sono calcolati ad uno ad uno tutti gli stimatori. Poi, per agevolare l'analisi, i valori trovati sono stati rappresentati in cinque mappe cromatiche distinte.

Nel seguito verranno esposte le modalità di calcolo degli stimatori e i risultati ottenuti. I primi quattro, $iTFR$, $xTFR$, $iTFR^+$, $xTFR^+$, verranno trattati a coppie, l'ultimo invece sarà illustrato a parte.

3.1 $iTFR$ e $xTFR$

Sia per calcolare $iTFR$ (Formula 1.4), sia per $xTFR$ (Formula 1.5) è necessario determinare il numero di bambini tra gli 0-4 anni (C) e il numero di donne (W) di età [15-49] presenti in ogni comune. Avendo a disposizione i dati disaggregati per sesso e fasce di età annuali, essi sono stati semplicemente ricavati come sommatoria dei totali di maschi e di femmine per le età interessate. Una volta determinato il tasso generico di fecondità C/W si sono quindi ottenute facilmente le stime.

Si ricordi che ciò che distingue i due stimatori è la presenza o meno dell'assunzione di equidistribuzione delle donne nelle diverse fasce di età. Tale assunzione, in un paese sviluppato come l'Italia, è poco realistica dato

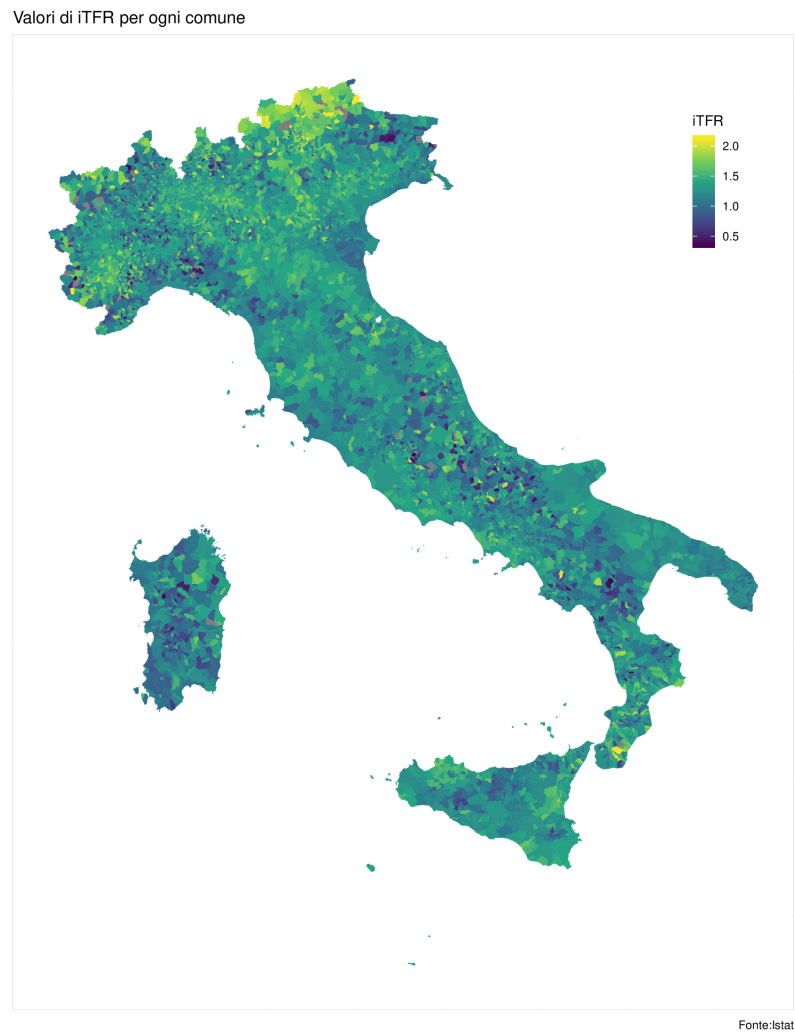


Figura 3.1: Mappa dei valori di $iTFR$ nei comuni italiani nel 2019.

che ormai da anni il numero di donne (come anche quello degli uomini) sta continuando a decrescere. Il calo drastico delle nascite, iniziato negli anni 70 fino agli anni 90, ha infatti portato a una progressiva riduzione della popolazione femminile. In particolare stanno diminuendo sempre di più le donne in età fertile perchè coloro che sono nate nella seconda metà degli anni 60, che sono le più numerose, stanno uscendo dalla fase riproduttiva (Istat, 2019b). Nel 2019, ad esempio, si stima che la maggior parte della

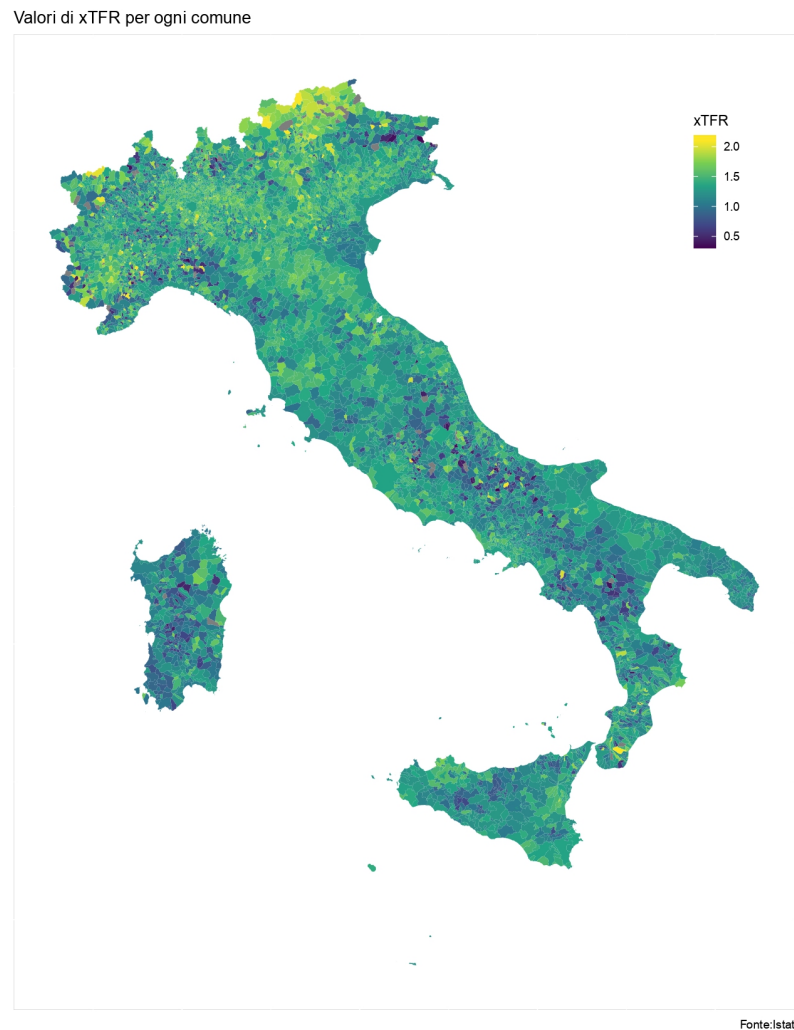


Figura 3.2: Mappa dei valori di $xTFR$ nei comuni italiani nel 2019.

popolazione femminile avesse 50-54 anni, e che il numero di donne in età fertile sia diminuito di circa 180 mila unità rispetto ai due anni precedenti (Istat, [2019a](#)).

Di conseguenza l'*age-structure multiplier* $1/p$ dell'equazione (1.5) sarà più grande per comuni aventi una proporzione di donne tra i 25 e i 34 anni minore, e viceversa. Con tale parametro si tiene quindi conto della diversa distribuzione delle donne e anche dell'eterogeneo livello di fertilità nelle varie

classi di età. Per questo motivo si ritiene che lo stimatore $xTFR$, rispetto a $iTFR$, abbia una maggiore prestazione. Possiamo affermare ciò anche alla luce dei risultati riportati nel documento di Schmertmann e Hauer 2020, dove si specifica che tale rapporto tra le quantità vale solo nel caso in cui ci siano livelli di mortalità relativamente bassi. Codesta regola infatti non vale quando si considerano paesi con livelli alti di mortalità e di fecondità (condizioni che in genere si riscontrano in paesi in via di sviluppo).

Per poter confrontare più dettagliatamente i due stimatori, i valori calcolati sono stati rappresentati in due mappe cromatiche (Figure 3.1 - 3.2). Ciò che si può notare è il lieve aumento di $xTFR$ rispetto a $iTFR$ soprattutto nei comuni del Nord Italia ma anche in Toscana, Marche e Umbria, in cui le sfumature sono più chiare e tendenti al giallo. Per quanto riguarda invece la macro-regione del Mezzogiorno¹ e il Lazio, i valori per la maggior parte risultano essere molto simili.

Questo porta a pensare, secondo quanto detto precedentemente, che la distribuzione delle donne in età fertile sia molto più equa al Sud e in alcune regioni del Centro rispetto al resto del paese. In generale comunque non sembrano esserci particolari distinzioni tra i due stimatori, anzi nel complesso le distribuzioni dei valori sembrano essere piuttosto simili.

Come si nota da entrambe le figure, il range dei valori stimati è stato troncato a $[0.35, 2.15]$. Inizialmente l'intervallo comprendeva valori poco realistici che andavano da 0 a 5.25 per $iTFR$ e 5.35 per $xTFR$. Controllando il dataset originale si è notato che, per i 25 comuni² i cui Tassi di Fecondità

¹Da qui in poi si farà riferimento alla ripartizione geografica adottata dall'Istat. Il Nord comprende le regioni Liguria, Lombardia, Piemonte, Valle d'Aosta, Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige, Veneto. Il Centro comprende le regioni Lazio, Marche, Toscana ed Umbria. E infine l'area del Mezzogiorno comprende le regioni dell'Italia meridionale (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia) e le isole

²Moncenisio, Noasca, Ribordone, Cervatto, Bellino, Briga Alta, Igliano, Marmora, Rittana, Roaschia, Rocca Cigliè, San Giorgio Scarampi, Serole, Rhêmes-Notre-Dame, Fascia, Gorreto, Rondanina, Cassiglio, Oliva Gessi, Drenchia, Zerba, Colledimacine, Molise, Modolo, Rosazza

Totali risultavano pari a zero, le numerosità totali di maschi e femmine erano nulle per la maggior parte delle classi di età d'interesse. In particolare, essendo tutti comuni poco popolati, risultava nulla la stima del numero di bambini C (Formula 1.2), e il numero di donne W tra 15-49 anni si rivelava essere molto basso (anche pari a 5). Mancavano cioè di intere generazioni che portavano ad ottenere un rapporto C/W pari a zero.

Per quanto riguarda invece l'estremo superiore dell'intervallo, l'unico comune con $iTFR$ pari a 5.25 è risultato essere Ingria, un comune della città metropolitana di Torino che conta in totale 44 abitanti. Per lo stesso comune, $xTFR$ è risultato essere pari a 3.28, quindi in ogni caso è stato sovrastimato. Si rileva quindi la situazione opposta a quella esposta precedentemente. Il numero di bambini di 0-4 anni e il totale di donne W anche qui sono molto bassi ma non nulli. Di conseguenza il rapporto tra queste due quantità non risulta essere pari a zero, anzi appare esageratamente elevato. Questo è un errore molto comune quando si hanno popolazioni ridotte che porta a una sovrastima o una sottostima del vero parametro.

Un altro esempio di sovrastima si ottiene analizzando l'estremo superiore dell'intervallo di $xTFR$: il comune con valore pari a 5.35 è Pedesina, in provincia di Sondrio, avente un totale di 38 abitanti. Tali risultati sono dovuti al fatto che ci sono ancora comuni molto piccoli per cui è difficile ottenere stime corrette a causa dell'errore campionario. Si è deciso quindi di limitare i valori in maniera tale che fossero più credibili, cercando contemporaneamente di eliminare un esiguo numero di stime. Si sono voluti comunque mantenere i comuni con numero di figli per donna pari a 0.35 poichè non è difficile credere che ci siano unità territoriali con numero di bambini talmente basso da tendere a zero. Al contrario, si sono eliminati tutti i comuni con TFT maggiore di 2.15 perchè è ormai da anni che l'Italia non presenta un livello di fecondità così alto.

3.2 $iTFR^+$ e $xTFR^+$

In questi due stimatori viene fatto un aggiustamento al parametro $1/s$ per rimediare all'assunzione di nullità della mortalità infantile (Formule 1.6 - 1.7). Come già anticipato, per stimare ${}_5q_0$, è stata utilizzata la probabilità di morte tra i 0-4 anni relativa alle province. Sfruttando i codici alfanumerici

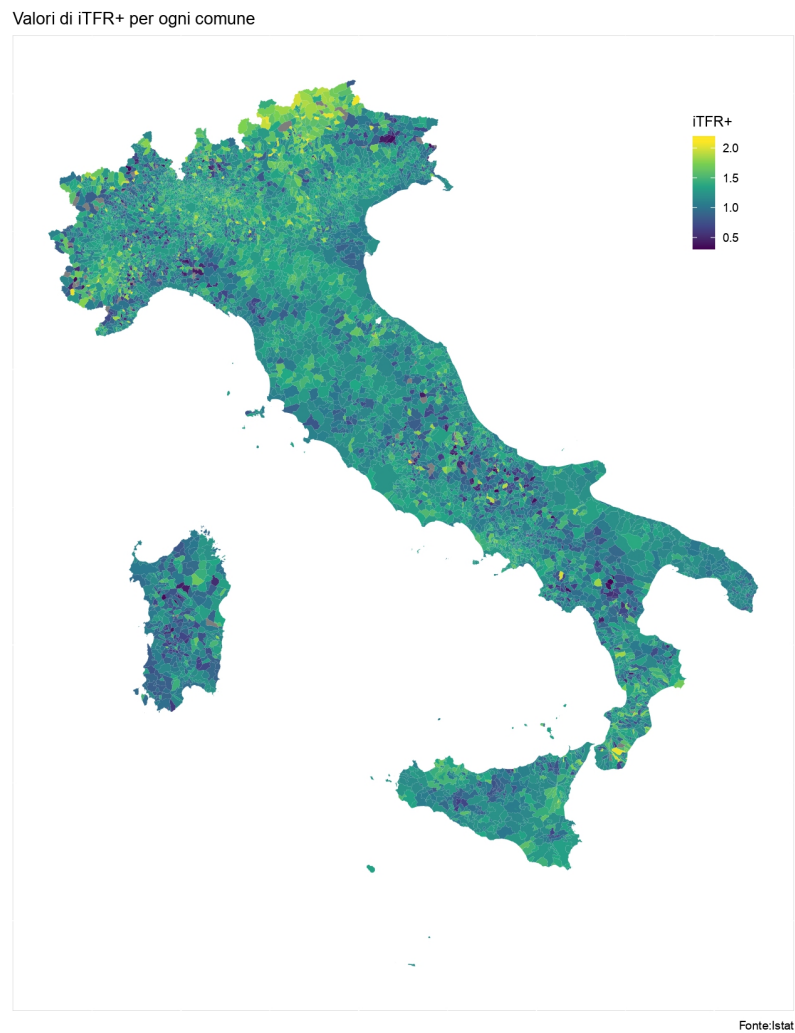


Figura 3.3: Mappa dei valori di $iTFR^+$ nei comuni italiani nel 2019.

forniti dall'Istat (Istat, [2020a](#)) e conoscendo la numerosità dei comuni per

ogni provincia, è stato possibile attribuire ogni valore al rispettivo comune.

Esattamente come ci si aspettava, il confronto tra le distribuzioni dei due stimatori porta alle stesse conclusioni precedenti. I valori di $xTFR^+$ (Figura 3.4) nel Nord del paese, e in alcune regioni del Centro, sono leggermente superiori a $iTFR^+$ (Figura 3.3), mentre risultano essere praticamente uguali nel resto della penisola. Si nota inoltre che i valori di $iTFR$ e $xTFR$ sono molto

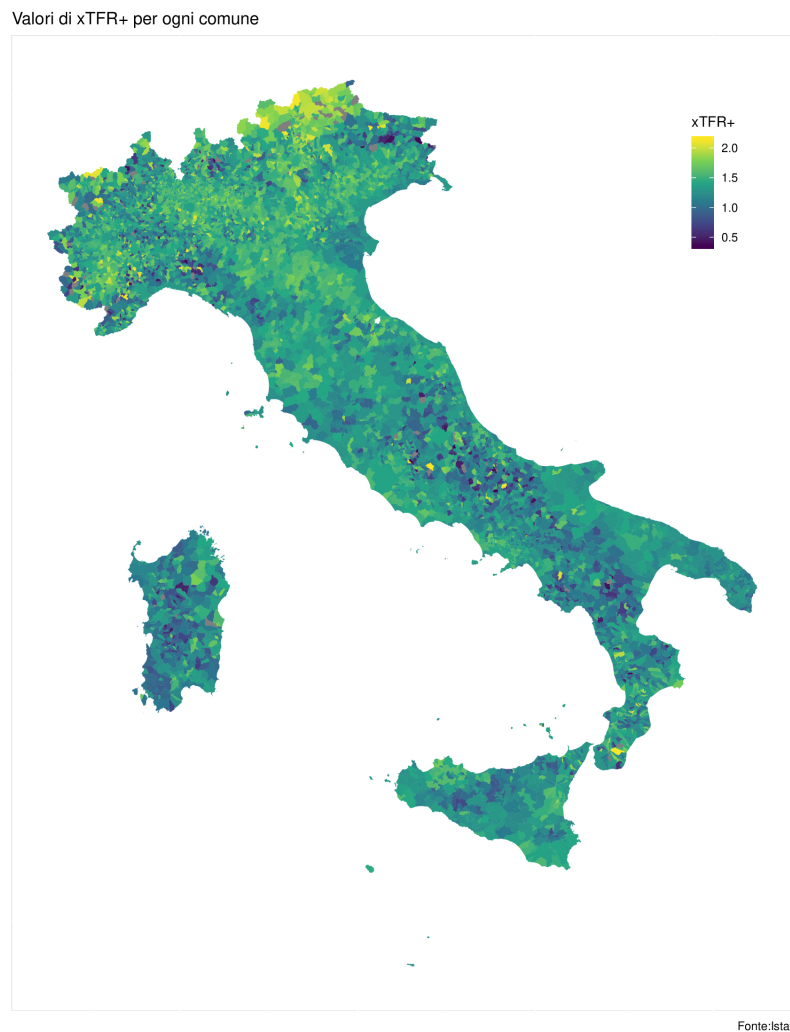


Figura 3.4: Mappa dei valori di $xTFR^+$ nei comuni italiani nel 2019.

simili ai valori di $iTFR^+$ e $xTFR^+$. Questo è dovuto al fatto che la mortalità

infantile in Italia è molto bassa, tanto che le probabilità ${}_5q_0$ tendono a zero. I diversi livelli di fecondità tra Nord e Sud del paese possono essere anche in parte spiegati dalla diversa distribuzione della probabilità di morte (Figura 2.1): valori più elevati della mortalità infantile, implicano valori di fecondità più bassi.

Anche qui l'intervallo di valori è stato troncato allo stesso modo perchè il range originale era $[0, 5.24]$ per $iTFR^+$ e $[0, 5.31]$ per $xTFR^+$. Come nel caso precedente, si rileva di nuovo la presenza di 25 comuni aventi stime del TFT pari a zero causate dalla nullità del parametro C. Inoltre, analizzando l'estremo superiore degli intervalli, si ottiene di nuovo che il comune con valore più elevato di $xTFR^+$ è Pedesina e quello di $iTFR^+$ è Ingria. Si nota quindi che l'aggiunta di informazione riguardo la mortalità infantile non ha contribuito alla riduzione dell'errore campionario. Ciò è esattamente come ci si aspettava, poichè che i due stimatori precedenti, $iTFR$ e $xTFR$, sono rispettivamente uguali a $iTFR^+$ e $xTFR^+$ a meno di un parametro $1/s$ (Formula 1.6) quasi pari a uno. Alla luce di tali risultati si sono quindi riarrangiati gli intervalli in maniera tale che i valori stimati fossero compresi tra $[0.35, 2.25]$. In questo modo i valori sono più credibili ed è possibile fare un confronto tra tutte le cartine.

3.3 bTFR

Per stimare bTFR è stata utilizzata un'unica funzione al cui interno vengono calcolate empiricamente una ad una tutte le quantità descritte nel primo capitolo. Inizialmente riceve come argomenti i dati ricavati dalla struttura per età della popolazione, quindi il numero di bambini di 0-4 anni e il numero di donne per ogni fascia di età W_a , e alla fine restituisce una lista di elementi contenenti i valori campionati della distribuzione e la relativa mediana (Appendice A.2).

La distribuzione dei valori di $bTFR$ (Figura 3.5) sembra essere la stessa di $xTFR^+$ ma in realtà in alcune zone del Nord-Ovest, come la Valle d'Aosta e il

Piemonte, è più elevato. Tale differenza potrebbe derivare dal fatto che non si considera più C (numero di bambini di 0-4 anni) come una costante, ma viene modellata come una variabile casuale di Poisson. Per il resto le differenze sono talmente minime che nel complesso neanche si notano. Più numerosi sono invece i comuni colorati di grigio, cioè coloro a cui è stato assegnato il valore NA. Il range dei valori di $bTFR$ inizialmente andava da 0.16 a 7.87,

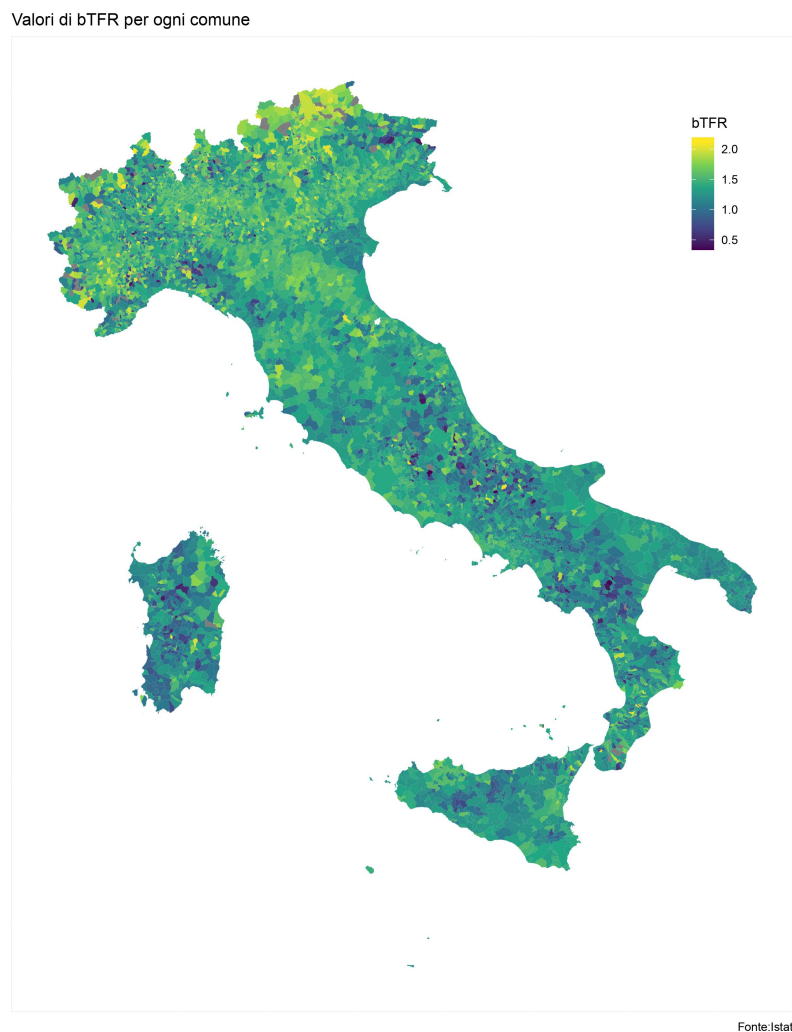


Figura 3.5: Mappa dei valori di bTFR nei comuni italiani nel 2019.

cioè comprendeva numeri esageratamente elevati o esageratamente bassi che

sicuramente non si verificano in uno stato sviluppato e specialmente in Italia. È quindi stato riarrangiato in maniera tale che i valori fossero compresi tra 0.35 e 2.15 come negli altri casi. Si noti che con tale metodo non si ottengono valori nulli del TFT. Questo perché il numero totale di bambini tra 0-4 anni non si ricava più da un conteggio dei dati, ma viene considerato come variabile casuale. Specificamente si assegna a C una distribuzione di Poisson, che, per definizione, è particolarmente appropriata quando si considerano eventi rari. Nella Tabella 3.1 si può vedere come tale assunzione, e le altre relative ai parametri di mortalità e fecondità, hanno cambiato il valore del TFT nei 25 comuni in cui risultava nullo.

Come ci si aspettava, la maggior parte dei comuni presenta valori bassi (al di sotto di 1) ma comunque più credibili di quelli stimati fin'ora. Dieci tra questi vengono esclusi dall'analisi geografica perché sono al di fuori dell'intervallo considerato. Per i restanti si ritiene che le stime siano più attendibili e che rispecchino meglio la realtà. Fanno eccezione i comuni di Moncenisio, Ribordone, Briga Alta e Gorreto che hanno valori esageratamente alti per dei comuni così poco abitati.

Nella Tabella 3.2 sono stati confrontati anche alcuni valori dello *Standard Error*. Si sono riportati i valori dei primi cinque e gli ultimi cinque comuni ordinati per popolazione perché sono quelli che presentano differenze più significative. Risulta quindi che avere una popolazione più elevata comporta delle stime più precise, infatti i valori di *Standard Error* più bassi si verificano nei primi comuni della lista. Ciò è in linea con quanto detto fin'ora perché è inevitabile che si ottengano stime più precise avendo un campione disponibile più grande.

Confrontando invece comuni di diversa superficie si ottengono stime dell'errore più simili e per questo non sono state riportate. Ciò che condiziona maggiormente la correttezza delle stime non è tanto la grandezza del comune, ma più la sua densità.

Si nota che tra i comuni più scarsamente popolati della Tabella 3.2 rientrano anche Briga Alta e Moncenisio, due tra i quattro comuni che nella

Nome del comune	bTFR
Moncenisio	1.50
Noasca	0.34
Ribordone	1.50
Cervatto	0.77
Bellino	0.34
Briga Alta	1.46
Igliano	0.79
Marmora	0.86
Rittana	0.25
Roaschia	0.51
Rocca Cigliè	0.30
San Giorgio Scarampi	0.31
Serole	0.30
Rhêmes-Notre-Dame	0.56
Fascia	0.98
Gorreto	1.43
Rondanina	0.64
Cassiglio	0.26
Oliva Gessi	0.27
Drenchia	0.73
Zerba	0.60
Colledimacine	0.48
Molise	0.30
Modolo	0.16
Rosazza	0.68

Tabella 3.1: Valore di bTFR dei 25 comuni in cui le stime ottenute dagli altri quattro metodi erano nulle.

Tabella 3.1 presentano un valore di TFT maggiore di uno. Anche se il relati-

vo *Standard Error* non è esageratamente alto, messo a confronto con quello dei comuni più popolosi, fa pensare che tali stime non siano molto attendibili esattamente come si era già ipotizzato precedentemente.

Nella Tabella 3.2 spicca inoltre il valore di *bTFR* relativo al comune di Pedesina. Tale comune è situato a nord della Lombardia ed è appartenente alla provincia di Sondrio. Ha un totale di 44 abitanti tra cui 2 bambini di 0-4 anni e 4 donne tra i 15 e i 49 anni. Se poi si considerano le madri suddivise per fascia di età W_a si nota che nessuna si trova nella fascia di età più fertile 25-34, e che tutte e quattro hanno età maggiore di 35 anni. Inoltre la probabilità di morire prima del compimento dei 5 anni ${}_5q_0$ è tra i valori più bassi della penisola (0.003). Queste considerazioni fanno intuire che i valori estremi che si sono ottenuti, anche con i metodi *xTFR* e *xTFR*⁺, sono sovrastimati e per niente affidabili anche se valore dello *Standard Error* non sembra essere esageratamente alto. Analizzando più nello specifico i singoli stimatori si può capire il perché di tali risultati. Per calcolare *xTFR* è necessario prima determinare *age-structure multiplier* $1/p$ utilizzando il parametro π_{25-34} (proporzione di donne tra i 25-34 anni sul totale). Quest'ultimo, per quanto detto precedentemente, risulta essere pari a zero, e di conseguenza si ottiene un valore di $1/p$ molto elevato (pari a 10.65). Moltiplicandolo poi per il rapporto C/W si ottiene appunto un valore di *xTFR* pari a 5.35. Anche la stima di *bTFR* è condizionata dalla struttura per età della popolazione. Il fatto che ci siano così poche donne, e che tre di esse siano concentrate nella fascia di età [40,44], influisce in maniera significativa sul valore finale del TFT. In particolare condiziona la quantità C , avente un ruolo importante all'interno nel modello, che viene stimata a partire dal vettore di valori $W_a K_a$ (dove K_a si ricava dalla Formula 1.1). In generale quindi si è riusciti a delineare un comportamento comune tra tutti i metodi di stima che abbiamo trattato. A parità di comune considerato, si sono ottenute delle stime abbastanza simili anche se alcune sono ritenute più attendibili di altre. Un problema di questi stimatori è la loro estrema sensibilità ai livelli di densità della popolazione nel comune. In tutti i casi infatti è stato necessario riadattare l'intervallo

Nome del comune	bTFR	SE
Roma	1.45	0.0138
Milano	1.43	0.0145
Napoli	1.41	0.0040
Torino	1.41	0.0123
Palermo	1.48	0.0062
Torresina	2.94	0.5904
Briga Alta	1.46	0.5607
Pedesina	7.87	0.9713
Moncenisio	1.50	0.5339
Monterone	3.76	0.7799

Tabella 3.2: Standard Error di $bTFR$ di alcuni comuni. I primi cinque sono i comuni più popolosi d'Italia, gli ultimi cinque i più scarsamente popolati.

di valori calcolati per poter trarre delle conclusioni più realistiche possibili. Il metodo di $bTFR$ è stato particolarmente utile perchè si sono potute verificare in maniera più analitica le proprie tesi. Inoltre con tale metodo è stato possibile ricavare una stima del valore del TFT anche per quei comuni i cui soli dati anagrafici non erano sufficienti.

Conclusioni

L'Italia, essendo un paese abbastanza esteso, presenta dei livelli di fecondità molto variabili a seconda delle zone geografiche. Da anni ormai si registra un Tasso di Fecondità più elevato nel Nord Italia rispetto al resto della penisola. In particolare nel 2019 (la cui media nazionale era di 1.29 figli per donna) il TFT è risultato essere pari a 1.39 nelle regioni settentrionali, ben superiore a quello registrato nel Mezzogiorno (1.26) e nel Centro (1.25). Le differenze poi si estendono anche a livello provinciale, per esempio al Nord i massimi valori si registrano nelle province di Trento (1.43) e di Bolzano (1.69) (Istat, [2019a](#)). Tale tendenza si rispecchia anche a livello comunale ed è particolarmente visibile dalle cartine sopra descritte (Figure 3.1-3.5). Come già specificato, i livelli di fecondità più elevati interessano tutte le regioni del Nord Italia e del Centro, ad eccezione del Lazio. Entrando più nel dettaglio si nota che nei comuni del Trentino-Alto Adige e nella zona più a Sud-Ovest del Piemonte si registrano valori del TFT pari, o addirittura superiori a 2. Al contrario, sembrano esserci comuni con valori particolarmente bassi (0.5) nelle zone al confine tra Liguria, Piemonte e Emilia-Romagna o a Nord del Friuli-Venezia Giulia. In più ci sono anche piccoli comuni a confine tra Piemonte e Lombardia, che a causa della loro estensione limitata non risaltano propriamente nella cartina.

Scendendo lungo la penisola, nelle regioni centrali, i valori del TFT hanno una distribuzione più uniforme, intorno all' 1.50. Questi indici sembrano essere particolarmente elevati rispetto alla media nazionale, ma sono utili per comprendere più nel dettaglio le specifiche tendenze territoriali.

A partire dal Lazio e l'Abruzzo si registrano valori più bassi, soprattutto lungo le zone dell'Appennino meridionale (zone centrali di Abruzzo, Molise e Campania, più l'intera Basilicata). Qui spiccano molti più comuni che hanno valori del TFT intorno all'estremo inferiore 0.35, inoltre ce ne sono veramente pochi aventi un numero medio di figlio per donna maggiore di 1.5. Valori così bassi possono essere particolarmente credibili per comuni aventi una scarsa densità abitativa i cui abitanti sono per lo più anziani e i cui giovani si sono trasferiti in altre zone dell'Italia.

In merito alle isole invece, è difficile trovare un andamento comune. In Sicilia si registrano alcuni valori più bassi al confine tra le province di Agrigento e Palermo, per il resto il TFT sembra essere pari a 1.25, con qualche valore più alto nella provincia di Catania e nella zona della provincia di Palermo confinante con il mare. In Sardegna invece ci sono valori più bassi nell'entroterra e nella parte più a ovest dell'isola.

In conclusione quindi, i valori stimati sembrano rispecchiare l'andamento generale del paese. Alcune quantità sono superiori rispetto a quanto ci si aspetterebbe, ma allo stesso tempo, ci sono altrettanti valori molto al di sotto dell'indice regionale o provinciale. Questo è normale perché, il fatto che ci siano numerose discrepanze all'interno delle stesse regioni o province, fa in modo che i diversi valori si compensino e portino al Tasso di fecondità totale già conosciuto.

Il valore più alto di TFT al Nord può essere spiegato dal fatto che sempre più giovani stanno emigrando dalla zona del mezzogiorno al settentrione per ragioni lavorative o anche familiari. Questo comporta una maggiore concentrazione delle donne nelle regioni del Nord e quindi un numero più alto di figli per donna. I piccoli comuni del Centro e del Sud invece vengono spopolati tanto che vengono a mancare intere generazioni di donne e di bambini. Inoltre come abbiamo visto (Figura 2.1) la probabilità di morte infantile è più elevata nella parte bassa della penisola e ciò inevitabilmente comporta un valore delle stime minore.

Per quanto riguarda i metodi utilizzati, in linea di massima sembrano con-

durre tutti alle stesse conclusioni. Nel nostro caso i valori di $iTFR$ e $iTFR^+$ sono i meno attendibili, soprattutto nelle zone in cui il numero di donne per fascia di età fertile si discosta di molto dalla equidistribuzione. Riguardo il metodo di stima $xTFR$, per quanto semplice, sembra essere comunque affidabile. Essendoci una mortalità infantile molto bassa non risulta neanche necessario porre una modifica al parametro $1/s$ per ottenere $xTFR^+$. Il metodo probabilistico $bTFR$ però è risultato essere il più utile. Anche se la distribuzione di valori stimati era molto simile a quella di $xTFR^+$, questo metodo ha permesso di fare delle considerazioni anche riguardo la precisione delle stime. Inoltre consente di risolvere parzialmente il problema della nullità dei dati perchè considera il numero di bambini C (Formula 1.2) come una variabile casuale. Il problema principale di tali metodi è che risultano essere estremamente sensibili alla densità dell'area geografica considerata, dato che si basano tutti sulla struttura per età della popolazione. Ma conoscendo l'accuratezza dei dati forniti dall'Istat, si pensa che i valori ottenuti siano per la maggior parte attendibili.

Nella tabella sottostante (Tabella 3.3) vengono riportati alcuni valori del TFT ottenuti con i vari metodi. Si prendono come esempio gli stessi comuni approfonditi nella Tabella 3.2.

La prima cosa che si nota è che in tutti i casi gli stimatori $iTFR$ e $iTFR^+$ sono uguali, come anche la coppia di valori $xTFR$ e $iTFR^+$. Ciò si era già precisato commentando le cartine cromatiche e facendo delle considerazioni sui dati a disposizione. Se invece si confrontano le coppie di valori di $iTFR$ e $xTFR$ o $iTFR^+$ e $iTFR^+$ si nota che persiste una differenza non banale tra le stime. Anche questo si era notato nel confronto generale ma non ci si era soffermati sul paragone tra i singoli comuni. Prestando particolare attenzione agli ultimi cinque comuni della tabella si vede che le differenze sono ancora più sostanziali, probabilmente perchè, in comuni poco popolati, l'assunzione di equidistribuzione è più azzardata rispetto che negli altri. Ovviamente per i comuni di Briga Alta e Moncenisio, le prime quattro stime risultano tutte uguali a zero perchè si basano sullo stesso rapporto C/W .

Nome del comune	iTFR	iTFR ⁺	xTFR	xTFR ⁺	bTFR
Roma	1.29	1.28	1.39	1.40	1.45
Milano	1.34	1.33	1.38	1.38	1.43
Napoli	1.33	1.32	1.38	1.38	1.41
Torino	1.29	1.29	1.35	1.35	1.41
Palermo	1.41	1.40	1.45	1.45	1.48
Torresina	1.17	1.16	1.43	1.42	2.94
Briga Alta	0	0	0	0	1.46
Pedesina	3.59	3.49	5.35	5.31	7.87
Moncenisio	0	0	0	0	1.50
Monterone	1.17	1.16	1.78	1.77	3.76

Tabella 3.3: Valori del TFT ricavati con i vari metodi messi a confronto. I primi cinque sono i comuni più popolosi d'Italia, gli ultimi cinque i più scarsamente popolati.

Considerando invece i valori di $bTFR$ si nota che sono tutti più elevati rispetto a quelli ottenuti con gli altri metodi. Nei comuni più popolosi le differenze con $xTFR^+$ sono meno accentuate. Al contrario, nella seconda parte della tabella si rilevano stime esageratamente alte e che, come si è già visto (Tabella 3.2), hanno una variabilità maggiore. Tali considerazioni rispecchiano l'ipotesi secondo cui tutti gli stimatori sono molto influenzati dalla numerosità del comune studiato.

In conclusione, non essendoci mai stata possibilità di stimare il valore del TFT nei comuni, non è possibile fare un confronto con altri metodi. Da quelli considerati in questa tesi sembra comunque di aver ottenuto delle buone approssimazioni anche se con alcune eccezioni. Nonostante questo, sulla base dei risultati ottenuti, non è possibile dire quale metodo di stima sia più adatto. Per capire quale metodo è più appropriato, è necessario prima di tutto fare delle considerazioni riguardo la numerosità dei bambini e delle donne, e alla mortalità infantile nel comune d'interesse. Fatta un'analisi accurata della

struttura per età, è possibile poi procedere alla selezione del metodo di stima più opportuno.

Bibliografia

- Arriga, E.E., P.D. Johnson e E. Jamison (1994). *Population analysis with microcomputers*. Vol. 1. United States Census Bureau.
- Bogue, D. e J. Palmore (1964). «Some empirical and analytic relations among demographic fertility measures, with regression models for fertility estimation». In: *Demography*, pp. 316–338.
- Bureau, U.S Census e USAID (n.d.). *Workshop of Demographic Analysis and Evaluation: Indirect estimation based on age structure. Rele's Method*. United States Census Bureau.
- Hauer, M., J. Baker e W. Brown (2013). «Indirect estimates of total fertility rate using child woman/ratio: A comparison with the Bogue-Palmore method». In: *PLoS One*. URL: <https://doi.org/10.1371/journal.pone.0067226>.
- Istat (2019a). *Indicatori demografici*. Istat. URL: https://www.istat.it/it/files/2020/02/Indicatori-demografici_2019.pdf.
- (2019b). «Tendenze demografiche e percorsi di vita». In: *Rapporto annuale*. Cap. 3. URL: <https://www.istat.it/storage/rapporto-annuale/2019/capitolo3.pdf>.
- (2020a). *Codici statistici delle unità amministrative territoriali: comuni, città metropolitane, province e regioni*. URL: <https://www.istat.it/it/archivio/6789>.
- (2020b). *Confini delle unità amministrative a fini statistici al 1° gennaio 2020*. URL: <https://www.istat.it/it/archivio/222527>.

- Istat (2020c). *Decessi e cause di morte: cosa produce l'Istat*. URL: <https://www.istat.it/it/archivio/240401>.
- Palamuleni, M.E. (2013). «Estimates of Fertility for South Africa based on Rele Method, 1996-2011». In: *Journal of Human Ecology*, pp. 257–265.
- Palmore, J. (1968). *Regression estimates of changes in fertility, 1955-60 to 1965-75, for most major nations and territories*. Honolulu, Hawaii: East-West Population Institute.
- Preston, H.S., P. Heuveline e M. Guillot (2001). *Demography: measuring and modeling population processes*. Blackwell Publishing.
- Research Methods, National Centre for (n.d.). *An introduction to MCMC methods and Bayesian Statistics*. University of Bristol.
- Schmertmann, C.P. e M.E. Hauer (2019). «Bayesian estimation of total fertility from a population's age-sex structure». In: *Statistical Modelling*, pp. 225–247.
- (2020). «Population Pyramids Yield Accurate Estimates of Total Fertility Rates». In: *Demography*, pp. 221–241.
- Tuchfeld, B.S., L.L. Guess e D.W. Hastings (1974). «The Bogue-Palmore technique for estimating direct fertility measures from indirect indicators as applied to Tennessee counties, 1960 and 1970». In: *Demography* Vol. 11.2.
- Wilmoth, J. et al. (2012). «A flexible two-dimensional mortality model for use in indirect estimation». In: *Population Studies*, pp. 1–28.

Appendice A

Codice R utilizzato

Codice A.1: Comandi per ottenere la mappa cromatica

```
library(ggplot2)
library(ggmap)
library(rgdal)
library(sf)
library(broom)
library(plyr)
library(maptools)

#creo un nuovo dataframe avente come variabili il codice del comune e il
#rispettivo iTFR
cod<-unique(dati$codice)
grafico<-data.frame(cod, iTFR)

#apro il shapefile contenente i dati sui confini dei comuni italiani
shapecom<-readOGR(dsn=".",layer="Com01012019_WGS84", stringsAsFactors=
FALSE)
shapecom@data$id<-rownames(shapecom@data)
myshape3<-fortify(shapecom)
myshapecom<-join(myshape3, shapecom@data, by="id")
```

```
#creo nuovo dataframe in cui ad ogni singolo comune, oltre al suo codice,
  vengono associati i valori di latitudine, longitudine e iTFR
prova1<-merge(myshapecom, grafico, by.x="PRO_COM", by.y="cod")
attach(prova1)
prova1<-prova1[order(group, order),]

ggplot() + geom_polygon(data=prova1, aes(x=long, y = lat, group = group,
  fill=prova1[,21])) +
  geom_path(data=prova1, aes(x=long, y=lat, group=group), color="
    trasparente")+
  geom_path(data=myshape, aes(x=long, y=lat, group=group), color="
    trasparente", size=0.05)+
  scale_fill_viridis_c()+
  labs(fill="iTFR", title="Valori di iTFR per ogni comune", caption=c("
    Fonte: Istat"))+
  theme(title=element_text(face="plain"), legend.position=c(0.90,0.85),
    panel.background = element_rect(fill="white", color="grey", size
      =0.1),
    axis.text.x= element_blank(), axis.text.y= element_blank(), axis.
      ticks = element_blank(),
    axis.title = element_blank())
```

Codice A.2: Funzione per ottenere bTFR senza l'utilizzo di Stan

```

library(LearnBayes)

requirefast_bTFR = function(C , W , q5_est ,
                             LTFR = 0, HTFR = 20, delta_TFR=.02,
                             Lq5 = .0001, Hq5 =.0999,
                             rownames=NULL) {
  require(assertthat)

  n = length(C) #nel nostro caso quindi equivale al numero di comuni
  W_matrix = matrix(W, nrow=n, byrow=FALSE)
  q5_matrix = matrix(q5_est, nrow=n, byrow=FALSE)

  if (n==1) {
    assert_that( all(length(W_matrix)==7),
                 msg='There is only one C value, but W does not have 7
                     elements')
  } else if (n>1) {
    assert_that( all(nrow(W_matrix)==n, ncol(W_matrix)==7),msg='C and W
                        are incompatible')
    if (class(q5_matrix)=='matrix') {
      assert_that( all(nrow(q5_matrix)==n),msg='C and matrix of q5_est are
                        incompatible')
    }
    if (class(q5_matrix)=='numeric') {
      assert_that( all(length(q5_matrix)==n),msg='C and vector of q5_est
                        are incompatible')
    }
  }

  result = vector('list',n)
  if (!is.null(rownames)) names(result) = rownames

```

```

#parametri di fertilit
#si definiscono prima le costanti
m = c(0, 1.39212, 1.59237, 1.22909, 0.4523, -0.88676, -3.43722)
names(m) = seq(15,45,5)

X = matrix( c(0, 0.27443, 0.54078, 0.73193, 0.87739, 1.04049, 1.51719,
              0, 0.31782, 0.51079, 0.51072, 0.34831, 0.05289, -0.72364),
            nrow=7,ncol=2)
rownames(X) = seq(15,45,5)

#si ricavano phi10,...,phi45 sfruttando la loro relazione con gamma=m+X*
  beta
#phi10 = 0 perch si ipotizza che la fertilit prima dei 15 anni sia
  nulla
phi = function(beta1,beta2) {

  gamma = as.vector( m + X %*% c(beta1,beta2))
  phi   = exp(gamma) / sum(exp(gamma))

  result = c(0, phi)
  names(result) = seq(10,45,5)

  return(result)

}

#parametri di mortalit
bigL = function(q5,k) {
  #af, bf, cf, vf sono costanti e sono state ricavate dallo studio di
  wilmoth(si prendono in considerazione le costanti relative al
  sesso femminile)

```

```

#sono 11 valori, uno per ogni classe 0, 1, 5, 10...
af = c(-0.6619, -99, -2.5608, -3.2435, -3.1099, -2.9789, -3.0185,
      -3.0201, -3.1487, -3.269, -3.5202)
bf = c(0.7684, -99, 1.7937, 1.6653, 1.5797, 1.5053, 1.3729, 1.2879,
      1.1071, 0.9339, 0.6642)
cf = c(-0.0277, -99, 0.1082, 0.1088, 0.1147, 0.1011, 0.0815, 0.0778,
      0.0637, 0.0533, 0.0289)
vf = c(0, -99, 0.2788, 0.3423, 0.4007, 0.4133, 0.3884, 0.3391, 0.2829,
      0.2246, 0.1774)

#log tassi di mortalit per classi di et x=0,1,5,10,...,45
logmx = af + bf * log(q5) + cf * (log(q5))^2 + vf * k
names(logmx) = c(0,1,seq(5,45,5))

#tassi di mortalit
#il parametro relativo alla classe di et [1,5) si calcola in maniera
  differente perch , essendo q5 un parametro del modello, non sono
  ricavabili le costanti a1, b1, c1, v1
mx = exp(logmx)
mx['1'] = -1/4 * (mx['0'] + log(1-q5))

#numero di sopravvivenenti
n = c(1,4,rep(5,9))
little_l = c(1, exp( -cumsum( n*mx )))
names(little_l) = c(0,1,seq(5,50,5))

#tempi di esposizione al rischio stimati attraverso approssimazione
  trapezoidale
L = n/2 * (head(little_l, -1) + tail(little_l, -1))
result = c(L['0'] + L['1'], tail(L,-2))
return(result)
}

```

```

#K_a con TFT=1
eta_a= function(q5,k, beta1, beta2) {

  this.phi = phi(beta1, beta2)
  this.L    = bigL(q5,k)

  #probabilit di sopravvivenza(di arrivare alla classe successiva) ->
    L(a)/L(a+5)
  surv = this.L[paste(seq(10,40,5))]/this.L[paste(seq(15,45,5))]

  #formula 1.1
  eta_a = this.L['0']/5 * 1/2 * (surv * this.phi[paste(seq(10,40,5))] +
    this.phi[paste(seq(15,45,5))])

  names(eta_a) = seq(15,45,5)

  return(eta_a)
}

q5_vals = Lq5 + (Hq5-Lq5)*1:4/5
k_vals = seq(-2, +2, length=5)

#creo un data frame con ogni combinazione disponibili di q5, k, beta1 e
  beta2
G = expand.grid(q5=q5_vals, k=k_vals, b1=k_vals, b2=k_vals)

#calcolo K_a per ogni combinazione dei parametri sempre ipotizzando TFR
  =1
this.eta = t(apply(G, 1, function(x) eta_a(x[1], x[2], x[3], x[4])))

```

```

colnames(this.eta) = paste0('eta', seq(15,45,5))

#definisco le distribuzioni a priori k e beta
#k e i beta si distribuiscono come una normale standard troncata a
  [-2,2]
G$fkbeta = apply(G,1,function(x) prod(dnorm(x[2:4])))
G$fq5     = NA
G$fprior  = NA

#aggiungo le colonne per ogni valore k_a
G = data.frame(G, this.eta)

#calcolo per ogni comune
for (i in 1:n) {

  this.C = C[i]
  this.W = W_matrix[i,]
  this.q5 = q5_matrix[i,]

  #numero totale di bambini di 0-4 anni C_a con TFR=1
  this.Cstar1 = as.matrix( G[, paste0('eta', seq(15,45,5))] ) %*% this.W

  #q5 ~ Beta(a,b)
  ab = LearnBayes::beta.select( list(x= min(q5_matrix[i,])/2, p=.05),
                                list(x= 2*max(q5_matrix[i,]), p=.95))

  #distribuzione a priori dei q5
  G$fq5 = dbeta( G$q5, ab[1],ab[2])

  #distribuzione a priori completa
  G$fprior = G$fkbeta * G$fq5

```

```

#definisco la distribuzione a posteriori
posterior_prob = function(TFR) {
  Cstar = TFR * this.Cstar1
  L      = dpois(this.C, Cstar)
  return ( sum( L * G$fprior))
}

#creo dei valori di TFR come se fosse una uniforme [0,20]
TFR_vals = seq(from = LTFR+delta_TFR/2,
               to   = HTFR-delta_TFR/2,
               by   = delta_TFR)

pp = sapply(TFR_vals, posterior_prob)

#interpolazione lineare tra i punti con coordinate x e y
inv_cdf = approxfun(x=cumsum(pp)/sum(pp), y=TFR_vals + delta_TFR/2,
                    ties=mean)
#funzione approxfun restituisce a sua volta una funzione

#stimo la densit del campione (con metodo kernel) di valori che
#trovo facendo l'interpolazione
simdens = density(inv_cdf(seq(.005,.995,.01)), adjust=1.5, na.rm=TRUE)

#ritorna una lista in cui ogni elemento comprende i valori x e y della
#densit e il valore mediano
result[[i]] = list(dens=simdens, median=inv_cdf(.50))

}

if (n==1) return(result[[1]])
if (n>1) return(result)
}

```