

**STIMA DELL'INDICE DI POVERTÀ O
ESCLUSIONE SOCIALE
E DELLE SUE COMPONENTI
PER I DISTRETTI SOCIOSANITARI
E LE PROVINCE DELL'EMILIA-ROMAGNA**

sulla base dei dati tratti dalla
Indagine EU-SILC, anni di riferimento 2019 e 2020,
mediante l'impiego di tecniche di Stima per Piccole Aree

Maria Rosaria Ferrante

Dipartimento di Statistica "P. Fortunati", Università di Bologna

Con la collaborazione di

Enrico Fabrizi

dell'Università Cattolica – sede di Piacenza

Giugno 2023

Indice

1.	Introduzione	p.4
2.	Povert� ed esclusione sociale, disparit� territoriali e definizione delle politiche	p.6
3.	Alcune note a corredo	p.9
4.	I domini di studio	p.11
5.	La base dei dati. L'indagine EU-SILC	p.13
6.	La strategia di stima adottata. Le stime dirette e la valutazione della loro affidabilit�	p.14
7.	I modelli per "piccole aree"	p.18
8.	Le stime da modello	p.19

Appendice A - L'indagine EU-SILC, la produzione delle stime dirette e della loro varianza	p.59
--	-------------

A1. Indagine EU-SILC. Caratteristiche generali	p.60
---	-------------

A1.1 Popolazione obiettivo	p.61
A1.2 Periodo di rilevazione e periodo di riferimento	p.61
A1.3 Piano di campionamento	p.61
A1.4 Pesi campionari	p.63
A1.5 Domini pianificati e domini di interesse	p.67
A2. Reddito disponibile equivalente	p.68
A2.1 Scala di equivalenza	p.70
A3. Definizione degli indicatori di interesse	p.71
A4. Definizione degli stimatori per la produzione delle stime dirette	p.75
A5. Stima della varianza delle stime dirette	p.77
A6. Stime dirette e relativi coefficienti di variazione	p.82

Appendice B - Produzione delle stime da modello e valutazione della loro affidabilità	p.88
B1. Modelli per piccole aree. Introduzione	p.89
B2. Inclusione delle provincie delle rimanenti regioni italiane nel processo di stima	p.90
B3. Analisi e selezione dell'informazione ausiliaria	p.91
B4. I modelli di stima per piccole aree	p.93
B4.1 Modello per i tassi	p.93
B4.2 Specificazione delle distribuzioni a priori per gli iperparametri	p.97
B5. Analisi e selezione dell'informazione ausiliaria	p.97
B6. Simulazione delle distribuzioni a posteriori	p.101
B7. La selezione dei modelli.	p.102
B8. Il benchmarking delle stime da modello	p. 103
Riferimenti bibliografici	p.106

1. Introduzione

Obiettivo di questa ricerca è la stima dei seguenti indicatori riferiti ai distretti sociosanitari ed alle province dell'Emilia-Romagna:

1. indicatore di Rischio di povertà o esclusione sociale (anche detto AROPE, At Risk Of Poverty or Social Exclusion), dato dalla quota di individui che sono caratterizzati da almeno una delle condizioni individuate dagli indicatori 2, 3 e 4 seguenti.
2. indicatore di Rischio di povertà (anche detto AROP, At Risk Of Poverty), dato dalla quota degli individui che dispongono di un reddito equivalente inferiore ad una soglia (di povertà) individuata nel 60% del reddito mediano nazionale equivalente
3. indicatore di Deprivazione materiale grave (anche detto SMD, Severe Material Deprivation) dato dalla quota di individui che vivono in famiglie che sperimentano almeno quattro tra i seguenti nove sintomi di disagio:
 - a) non poter riscaldare adeguatamente l'abitazione;
 - b) non poter sostenere una spesa imprevista (il cui importo, in un dato anno, è pari a 1/12 del valore della soglia di povertà rilevata nei due anni precedenti);
 - c) non potersi permettere un pasto proteico (carne, pesce o equivalente vegetariano) almeno una volta ogni due giorni;
 - d) non potersi permettere una settimana di ferie all'anno lontano da casa;
 - e) non potersi permettere un televisore a colori;
 - f) non potersi permettere una lavatrice;
 - g) non potersi permettere un'automobile;
 - h) non potersi permettere un telefono;
 - i) essere in arretrato nel pagamento di bollette, affitto, mutuo o altro tipo di prestito.
4. indicatore di Bassa intensità lavorativa (anche detto QJ, Quasi-Joblessness), dato dalla quota di individui di età compresa tra 0 e 59 anni, che vivono in famiglie ad intensità di lavoro molto bassa, inferiore al 20%. L'intensità di lavoro è definita come rapporto fra il numero totale di mesi lavorati dai componenti della famiglia durante l'anno di riferimento dei redditi (quello precedente all'anno di rilevazione) e il numero totale di mesi teoricamente disponibili per attività lavorative. Nel calcolo di tale rapporto si considerano gli adulti (di età 18–59 anni,

ad esclusione degli studenti nella fascia di età 18–24). La soglia riferita all'intensità di lavoro è, come detto sopra, pari al 20%.

In sintesi, l'indicatore AROPE è dato dalla quota di individui che sono caratterizzati da almeno una delle seguenti tre condizioni: i) vivono in famiglie a bassa intensità di lavoro, ii) vivono in famiglie a rischio di povertà, iii) vivono in famiglie in condizioni di severa deprivazione materiale. La misura della povertà ed esclusione sociale è pensata, in altre parole, secondo tre linee principali:

- la “povertà monetaria”, basata sul reddito disponibile familiare come variabile obiettivo per la valutazione delle risorse disponibili
- la deprivazione misurata quale impossibilità degli individui di raggiungere i propri obiettivi (“funzionamenti”) a causa dello stato di ristrettezze in cui vivono
- l'esclusione sociale valutata in chiave prospettica, che caratterizza quella parte di popolazione che, pur in assenza di problemi economici, è da considerarsi a forte rischio di esclusione sociale in quanto vive in una famiglia in cui nessuno lavora o i cui componenti sono occupati in attività lavorative in modo discontinuo e poco intenso.

Una definizione dettagliata di questi parametri e dei relativi stimatori è riportata nell'Appendice A3.

Come dominio oggetto di stima sono stati scelti i distretti, per il ruolo che hanno assunto negli ultimi anni nell'ambito delle politiche regionali volte alla promozione dell'integrazione sociale e sanitaria, e le province.

Le stime sono state ottenute utilizzando i dati dell'indagine Survey on Income and Living Condition (EU-SILC) 2019 (reddito 2018) e 2020 (reddito 2019) e mediante metodi di stima “per piccole aree”. L'impiego di tali metodi di stima cosiddetti “da modello” si rende necessario in quanto il numero delle unità campionate sia nei distretti sia nelle province è troppo piccolo per ottenere stime affidabili con i metodi di stima basati sul disegno campionario che Eurostat utilizza correntemente per stimare questi parametri.

Una nota definitoria. Con riferimento ai parametri reddituali usiamo la dicitura “indicatore riferito all'anno x ” per denotare l'indicatore calcolato sulla base dei dati di reddito raccolti nell'indagine dell'anno x , riferiti all'anno $x - 1$ (ad esempio il Rischio di povertà riferito al 2017 è calcolato sui redditi 2016). Questa convenzione è usata dall'Istat nelle proprie pubblicazioni. Per uniformità con le pubblicazioni Eurostat e Istat, l'indicatore AROPE viene riferito all'anno di indagine: si tenga conto che il rischio di povertà è calcolato sui redditi dell'anno precedente e la bassa intensità di lavoro è

calcolata sul numero totale di mesi lavorati dai componenti della famiglia durante l'anno precedente l'indagine.

2. Povertà ed esclusione sociale, disparità territoriali e definizione delle politiche

Il Parlamento europeo, il Consiglio e la Commissione hanno proclamato nell'Agenda 2030 il Pilastro europeo dei diritti sociali. Il pilastro, che stabilisce 20 principi fondamentali che dovranno guidarci verso un'Europa sociale forte, equa, inclusiva e ricca di opportunità nel XXI secolo, propone tre obiettivi da raggiungere entro il 2030 nei settori dell'occupazione, delle competenze e della protezione sociale. Uno degli obiettivi più importanti riguarda il tema della povertà ed esclusione sociale: entro il 2030 il numero di persone a rischio di povertà o esclusione sociale dovrebbe ridursi di almeno 15 milioni e, di queste, almeno 5 milioni dovrebbero essere bambini. Anche la strategia Europa 2020 per una crescita intelligente, sostenibile e inclusiva comprendeva un obiettivo in materia di povertà ed esclusione sociale: ridurre tra il 2010 ed il 2020, nell'UE, di almeno 20 milioni il numero di persone a rischio di povertà o esclusione sociale. Al termine del periodo interessato dalla strategia Europa 2020, dobbiamo riconoscere che i progressi verso questo importante obiettivo non sono stati chiaramente così importanti come dichiarato negli obiettivi. Più in dettaglio, l'obiettivo era che il numero di persone AROPE fosse ridotto da 116,1 milioni (stima da dati EU-SILC 2008) a 96,1 milioni. purtroppo, la stima del numero di individui AROPE ottenuta sulla base delle informazioni EU-SILC 2019 è notevolmente superiore a questo obiettivo: 107,5 milioni (21,4%).

Più in dettaglio, in Italia, nel 2019, il numero di individui a rischio di povertà o esclusione sociale ammonta a circa 15 milioni e 400 mila, più o meno un quarto della popolazione del paese (ARPE stimato al 25,6%). Anche se in leggera diminuzione rispetto all'anno precedente (27,3%), il livello italiano rimane comunque superiore a quello europeo, decisamente più elevato di Francia (17,9%) e Germania (17,4%). I paesi per cui emerge un AROPE superiore a quello italiano sono Bulgaria (32,8%), Romania (31,2%), Grecia (30,0%), Lituania (26,3%).

Se guardiamo ai 3 indicatori che sono alla base dell'ARPE emerge che le stime di rischio di povertà, grave deprivazione materiale e quota di chi vive in famiglie con una intensità di lavoro molto bassa sono rispettivamente pari al 20,1%, 7,4% e 10,1%. Tali valori testimoniano che i tre indicatori colgono aspetti diversi della povertà, intesa in senso più ampio rispetto a quella di tipo economico.

Sulla base alla strategia Europa 2020, il nostro Paese avrebbe dovuto raggiungere l'obiettivo di avere una popolazione a rischio di povertà o esclusione sociale di poco meno di 13 milioni di persone entro il 2020. Con i livelli registrati nel 2019, l'obiettivo appare ancora molto lontano e difficile da ottenere nel breve periodo.

I risultati relativi alla povertà e all'esclusione sociale a livello di UE, calcolati come media ponderata dei risultati nazionali, nascondono notevoli differenze tra gli Stati membri. Come già segnalato, nel 2019 per gli stati membri che soffrono maggiormente di tale problema (Bulgaria, Romania, Grecia e Lituania) circa un terzo della popolazione era a rischio di povertà o di esclusione sociale. I livelli più bassi si sono invece registrati nella Repubblica ceca (10,1%), in Finlandia (11,6%) e in Slovacchia (11,9%)

Accanto alle disparità tra i paesi europei, emerge una certa eterogeneità territoriale anche all'interno dei paesi europei. Con riferimento al nostro paese le disparità regionali nell'AROPE, e negli indicatori in cui l'AROPE si articola, sono molto ampie e nel mezzogiorno, come atteso, le situazioni di povertà ed esclusione sociale sono decisamente più importanti.

Se si guarda, infatti, al divario Nord-sud con riferimento all'AROPE, emerge che quasi la metà degli individui nel Sud del paese sono a rischio di povertà ed esclusione sociale, contro circa un individuo ogni cinque nel Nord. Più in dettaglio, nel Nord-est il Rischio di Povertà ed Esclusione Sociale si attesta al 13,2%, circa un terzo di quanto si osserva a Sud e nelle Isole (rispettivamente 41,6% e 43,6%).

Con riferimento all'AROP, più di un terzo circa degli individui nel mezzogiorno sono a Rischio di povertà (33,7% per il Sud e 36,8% per le Isole), contro un individuo ogni dieci nel Nord-est (9,5%). Il divario si mostra ancora più importante se si considerano le stime dell'AROP ottenute a livello di NUTS2 (Nomenclature of Territorial Units for Statistics) per cui si va da valori che si attestano intorno al 40% di Sicilia e Campania a valori pari al 6,1% in Valle D'Aosta, 8,0% per la Provincia Autonoma di Trento, 8,4% per Friuli-Venezia-Giulia. Tale divario è presente anche per l'indicatore di grave deprivazione materiale (SMD): il valore stimato dell'indicatore è pari a 2,9% nel Nord-est e raggiunge il 12,7% nel Sud e il 15,4% nelle Isole.

Questi risultati confermano un grado particolarmente elevato di disparità spaziale in Italia, dove le regioni più ricche hanno un AROPE di oltre 3 volte l'entità di quelle delle regioni povere del sud del paese.

Gli indicatori sociali, e nello specifico l'AROPE e gli indicatori AROP, SMD e QJ, sulla base dei quali l'AROPE è costruito, potrebbero essere di grande interesse se ottenuti per un dettaglio territoriale maggiore rispetto a quello regionale (NUTS2). Avere consapevolezza di dove vivono i poveri o gli individui a rischio di esclusione sociale può aiutare ad aumentare l'efficacia delle politiche che mirano a contenere questi fenomeni. La disaggregazione regionale è particolarmente rilevante per quelle politiche che hanno come attori principali le autorità regionali o locali (si pensi ad esempio alle politiche sull'alloggio, sull'assistenza all'infanzia e sull'assistenza sociale).

Oggi la UE considera principalmente quale misura del benessere ambiti sub-nazionali il PIL pro-capite. Tuttavia, come ormai universalmente riconosciuto, il PIL pro-capite non riflette la povertà individuale o la disuguaglianza economica. Le politiche di

coesione fanno, in altre parole, perno sul livello territoriale definito dalle Regioni Amministrative (NUTS2) data la mancanza di informazioni statistiche ad un livello di maggior dettaglio territoriale. Tuttavia, avere informazioni sulla eterogeneità dentro le regioni potrebbe essere utile per ripartire i finanziamenti europei in maniera più efficiente, ad esempio distinguendo dentro le regioni aree più o meno densamente popolate, aree rurali e metropolitane, aree distinte per grado di urbanizzazione, ecc. Al fine di valutare la convergenza spaziale riferita a parametri di povertà e disuguaglianza, ci si potrebbe, ad esempio chiedere se il nostro paese, notoriamente caratterizzato da elevata eterogeneità territoriale, presenti un rischio di povertà ed esclusione sociale maggiore nelle aree urbane o nelle aree rurali di regioni appartenenti alle macroaree del Nord o del Sud. Ancora, ci sono paesi in cui il modello di sviluppo economico è policentrico e quindi riferito a città popolate distribuite sul territorio mentre altri paesi in cui il modello di sviluppo è monocentrico, concentrato intorno alla capitale. E cosa dire dell'interesse di un confronto tra centri metropolitani europei in termini di rischio di povertà ed esclusione sociale al fine di valutare quali sono le aree metropolitane che necessitano di intervento dell'EU per migliorare il livello di benessere economico e sociale delle città?

Un'ulteriore considerazione a favore della necessità di disporre, per una migliore definizione delle politiche, di stime riferite ad un maggiore livello di dettaglio territoriale rispetto al NUTS2. Il livello NUTS2 in EU fa riferimento a regioni di ampiezza molto diversa e ad un numero di aree NUTS2 diverso da paese a paese e questo determina problemi di comparabilità. L'utilizzo di un dettaglio territoriale di livello NUTS3 ridurrebbe al minimo la differenza tra le dimensioni regionali e consentirebbe una valutazione delle disparità regionali maggiormente comparabile tra paesi ed assicurerebbe la possibilità di formulare soluzioni locali e globali che siano tra loro coerenti.

Infine, sarebbe molto importante che le stime ottenute ad un certo dettaglio territoriale includessero anche la stima del loro standard error. L'assenza di questa importante informazione rende non conclusiva qualsiasi considerazione sulle differenze tra aree.

Ad oggi nel nostro paese l'AROE viene reso disponibile (stimato a partire dai dati EU-SILC) a livello di NUTS2 (Regioni amministrative). La numerosità campionaria relativa all'indagine EU-SILC non consente, infatti, di ottenere stime affidabili di indicatori di povertà e disuguaglianza economica ad un livello di dettaglio subregionale e distinguendo le singole aree metropolitane o rurali. Sappiamo quindi che negli stati membri dell'occidente EU il rischio di povertà ed esclusione sociale tende ad essere più alto nelle città mentre nell'est Europa è generalmente più elevato nelle zone rurali, ma non sappiamo valutare se questo rischio risulta eterogeneo dentro ogni area subnazionale.

In presenza di questo tipo di limitazioni, per ottenere stime affidabili ad un buon livello di dettaglio territoriale, si possono impiegare metodi di stima "da modello", cosiddetti

“per piccole aree”. Con l’espressione gergale “da modello” si intende che ci siamo allontanati dal paradigma classico dell’inferenza per quantità descrittive delle popolazioni finite (in cui l’unica sorgente di casualità, in linea di principio, è quella del campionamento probabilistico governato da chi conduce l’indagine) per specificare un modello probabilistico descrittivo della variabile obiettivo.

3. Alcune note a corredo

Il reddito disponibile

La definizione di reddito a cui fanno riferimento gli indicatori presi in considerazione in questa ricerca è quella di reddito disponibile.

Il reddito disponibile è dato dalla somma del reddito imponibile netto e del reddito esente. Si tratta di un aggregato che cerca di approssimare al meglio il concetto di risorse disponibili per l’individuo o per la famiglia. Una sua definizione dettagliata, quella adottata dall’indagine EU-SILC secondo le linee guida del documento di Camberra, è fornita nell’Appendice A2 di questo rapporto.

La famiglia (household) come unità di indagine

Nello studio della povertà e dell’esclusione sociale si assume che le risorse disponibili all’interno della famiglia vengano messe in comune e utilizzate per soddisfare i bisogni e incrementare le ‘capabilities’ di tutti i suoi membri in ugual misura.

La definizione di famiglia adottata è la stessa che ISTAT adotta nelle indagini di argomento sociale: un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti o aventi dimora abituale nello stesso comune. Si tratta quindi della ‘famiglia di fatto’ o volendo utilizzare la terminologia inglese della ‘household’.

Considerando come unità di indagine la famiglia, il reddito disponibile rilevante per la definizione e lo studio della povertà è quello “complessivo familiare”, ossia il reddito disponibile di tutti i membri della famiglia, che si suppone venga messo in comune per il benessere dei membri. L’assunzione implicita è che la ripartizione delle risorse tra i membri della famiglia sia egualitaria. Si tratta di un’ipotesi piuttosto forte, che può avere l’effetto di portare a sottostime anche rilevanti del livello di disuguaglianza nella distribuzione delle risorse all’interno della popolazione, qualora l’egualitarismo all’interno delle famiglie venga meno.

La variabile utilizzata per descrivere le risorse economiche disponibili a livello individuale è data dal “reddito disponibile equivalente” calcolato a partire dal reddito

complessivo della famiglia. Ad ogni membro della stessa famiglia viene quindi attribuito lo stesso valore del reddito equivalente.

La scala di equivalenza

Il reddito complessivo familiare non consente di confrontare il livello di risorse individuali effettivamente disponibile per membri di famiglie aventi dimensione e composizione differenti. Una semplice divisione per il numero di componenti (reddito pro-capite) non terrebbe conto delle economie di scala che la messa in comune delle risorse comporta. Poiché queste economie di scala dipendono oltre che dalla dimensione anche dalla composizione del nucleo, una metodologia molto utilizzata per “standardizzare” i nuclei familiare consiste nel calcolarne una “dimensione equivalente” secondo un’opportuna scala. Ciò comporta la divisione del reddito disponibile complessivo familiare per questo coefficiente e l’attribuzione del reddito equivalente risultante a ciascun membro della famiglia in ugual misura.

In questa ricerca è stata adottata la scala di equivalenza OCSE modificata, adottata da Eurostat e consigliata nelle linee guida di analisi di EU-SILC. Questa scala è utilizzata anche dalla Banca d’Italia (Banca d’Italia, 2006). Essa si basa sulla seguente definizione di dimensione equivalente:

$$EQ_{SIZE} = 1 + 0.5(CF_{14+} - 1) + 0.3CF_{13-}$$

Dove CF_{14+} è il numero di membri della famiglia di età maggiore o uguale a 14 mentre CF_{13-} è il numero dei rimanenti.

Va sottolineato, come questa scala, rappresenti da un punto di vista scientifico, una scelta tra diverse possibilità e possa essere quindi oggetto di critica. Una scala alternativa, molto popolare in Italia, per la sua applicazione alle misure della povertà basate sull’indagine sui Consumi delle Famiglie condotta da ISTAT è la “scala Carbonaro” (Carbonaro, 1985). Va notato tuttavia, come questa scala sia funzione monotona (crescente a tassi decrescenti) del numero di membri della famiglia ma non consideri la composizione per età. Nello stesso spirito, alcune pubblicazioni OCSE, con riferimento principalmente ai confronti internazionali, fanno riferimento ad una dimensione equivalente definita semplicemente come radice quadrata del numero di membri.

Alcuni riferimenti alla letteratura internazionale sul problema delle scale di equivalenza possono essere Atkinson (1998), Atkinson et al. (1995). Aaberge e Melby (2005) discutono invece più accuratamente il problema della sensitività delle stime dei parametri della distribuzione del reddito alla scelta delle scale di equivalenza.

4. I domini di studio

I distretti

Come già anticipato nel paragrafo introduttivo, la scelta dei distretti sanitari dell'Emilia-Romagna come ambito territoriale di riferimento per le stime si deve al ruolo che i distretti hanno assunto nella definizione delle politiche regionali per l'integrazione sociale e sanitaria. Il distretto si è venuto a configurare come una macro-struttura aziendale fortemente legata alla comunità locale, che si pone fra l'azienda sanitaria, spesso ampia e lontana dai singoli cittadini, e gli enti locali, per realizzare i programmi per l'integrazione sociosanitaria (Balduzzi e Moirano, 2011).

Tale ruolo viene stabilito dalla normativa sin dalle riforme degli anni '90 (d.lgs. n. 502 del 1992 come modificato dal d.lgs. n. 517 del 1993) con le quali il distretto diviene centro di governance di una complessa rete di servizi territoriali. Con la riforma sanitaria del 1999 (d. lgs. N. 229/1999) poi, una serie di disposizioni stabiliscono i principi volti a potenziare il distretto, delineandone le funzioni, gli aspetti programmatici ed una responsabilità legata agli obiettivi di salute, nel rispetto delle autonomie regionali che sono lasciate libere di disegnare il modello organizzativo meglio confacente alle caratteristiche dei singoli Servizi sanitari regionali (Balduzzi e Moirano, 2011).

Con la riforma del Titolo V, Parte seconda della Costituzione, ad opera della legge costituzionale n. 3 del 2001, l'ascrizione della tutela della salute nel novero delle materie di legislazione corrente (e cioè regionale) e la valorizzazione della potestà amministrativa hanno rafforzato le responsabilità regionali nella definizione e nelle scelte sul distretto (Balduzzi e Moirano, 2011).

L'Agenas (Agenzia nazionale per i servizi sanitari regionali) ha effettuato negli anni 2005-2006 e nel 2010 un'indagine sulle modalità di funzionamento dei distretti, al fine di comprendere meglio l'evoluzione nel tempo del distretto. Nel 2010 l'indagine si è focalizzata sulla raccolta di informazioni relative a: la funzione del distretto come punto di riferimento per il cittadino e gli strumenti di comunicazione con gli utenti, il coordinamento della rete di servizi e la presa in carico dei problemi sociosanitari complessi, le modalità di integrazione relazione tra i diversi attori dell'assistenza primaria (Balduzzi e Moirano, 2011). Dei 711 distretti italiani, 681 (96%) hanno partecipato all'indagine. I risultati mostrano un sistema in cui generalmente il distretto programma le sue attività assieme agli enti locali ed in cui alla figura del Direttore vengono affidate complesse responsabilità (Balduzzi e Moirano, 2011). Emergono anche elementi di eterogeneità fra i distretti rispetto agli specifici compiti svolti. Non tutti i distretti, ad esempio, partecipano alla definizione e al monitoraggio del Piano di Zona (il documento di programmazione territoriale con il quale comuni e ULSS mettono a punto le politiche sociosanitarie rivolte alla popolazione). Lo fa circa l'84% dei distretti e con modalità diverse. Il profilo professionale e la specifica funzione del Direttore del

distretto poi, non sono omogenee nei distretti, anche perché la normativa stessa non definisce un profilo e una qualificazione accademica ideale per il Direttore.

In Emilia-Romagna i distretti attivi sono 38.

La Regione Emilia-Romagna, per far fronte ai mutamenti in atto nella struttura della popolazione e nella composizione delle famiglie, che vedono un progressivo incremento dell'incidenza delle fasce di popolazione di età avanzata e molto avanzata e l'aumento delle condizioni di solitudine, ha istituito (art. 51, L.R. n. 27/2004) il Fondo regionale per la non autosufficienza (Frna). Il programma per l'impiego dell'Frna è basato su linee strategiche del Piano sociale e sanitario che riguardano: lo sviluppo dell'assistenza domiciliare integrata (sanitaria e sociale), l'aumento dell'assegno di cura per anziano non autosufficienti (con un'attenzione specifica per coloro che non percepiscono l'indennità di accompagnamento), la qualificazione delle assistenze familiari, i servizi di telesoccorso e teleassistenza, i ricoveri temporanei di sollievo, il sostegno alle reti informali di solidarietà sociale (dal "portierato" al "custode sociale") (Curcetti, 2011). Tale programma viene attuato attraverso il piano annuale di attività elaborato dal Comitato di distretto (Sindaci) e dal Direttore del distretto. Le risorse vengono ripartite fra i distretti dell'Azienda USL, in quanto luoghi deputati alla programmazione e al governo dei servizi necessari alla popolazione di riferimento. Il 50% circa delle risorse assegnate attraverso il Frna e utilizzate per la domiciliarità viene utilizzato per finanziare gli "assegni di cura".

Le province

Il riferimento alle province (NUTS3) è importante nell'ottica di offrire informazioni statistiche territorialmente disaggregate ai policy makers locali, più aderenti rispetto a quelle regionali o addirittura nazionali alle esigenze informative connesse ai processi di governance delle aree provinciali o metropolitana. Le autorità pubbliche locali hanno la possibilità di delineare azioni politiche in risposta ai bisogni economici, sociali e ambientali della comunità locale e le loro scelte possono avere impatti significativi sulle condizioni di contesto sottostanti al benessere sociale del territorio. Per realizzare tali azioni hanno bisogno di informazioni statistiche riferite a "piccole aree" che, se provenienti da indagini campionarie, la statistica ufficiale non rende usualmente disponibili in quanto le indagini ufficiali non sono costruite per fornire stime affidabili di tale dettaglio. Con riferimento a parametri di povertà e disuguaglianza, ad esempio, se si considera la pubblicazione "Il Benessere Equo e Sostenibile delle Province – 2015", nella sezione degli indicatori relativa al Benessere Economico mancano informazioni riferite a indicatori di rischio di povertà, disuguaglianza e deprivazione (stimati a partire dai dati raccolti mediante indagini campionarie quali EU-SILC e Indagine sulle spese delle famiglie) mentre tali indicatori sono disponibili con riferimento nazionale. Tale situazione appare al momento non superabile, visto che sono pochissime le indagini campionarie presso le famiglie in grado di restituire informazioni statisticamente

significative al di sotto del livello regionale. Infine, la scelta del riferimento provinciale vuole seguire le indicazioni che OCSE ha fornito nell'ambito dell'iniziativa "Better life", per promuovere la disseminazione delle *best practices* di utilizzo degli indicatori di benessere a sostegno delle politiche locali, e le linee guida per l'attuazione a livello locale della strategia "Europa 2020 per le città e i territori".

Le stime da noi prodotte a livello provinciale potrebbero quindi completare ed arricchire il set di indicatori del benessere già prodotto per le province da ISTAT-SISTAN con un insieme di indicatori riferiti al benessere economico locale, caratterizzati da un adeguato grado di affidabilità, che sappiano cogliere le specificità locali e che siano utili per la programmazione locale.

5. La base dei dati. L'indagine EU-SILC

Le stime che proponiamo sono basate sui dati dell'indagine EU-SILC 2019 ed EU-SILC 2020. Di EU-SILC abbiamo già anticipato che si tratta di un progetto comunitario finalizzato alla produzione, secondo metodologie comuni e condivise, di una serie di indicatori relativi alla distribuzione del reddito, alla povertà, all'istruzione, al mercato del lavoro e alle condizioni di vita delle famiglie con la finalità principale di valutare empiricamente i progressi in tema di integrazione e coesione sociale.

In questo paragrafo ci limitiamo a mettere in luce solo alcune caratteristiche tecniche "essenziali" dell'indagine EU-SILC, rimandando una discussione più dettagliata delle caratteristiche del disegno campionario all'appendice A.

- L'indagine viene ripetuta ogni anno a partire dal 2004. Il periodo di riferimento del reddito (rilevato mediante l'indagine) è l'anno solare precedente a quello di rilevazione.
- Il campione dell'indagine ha una struttura a panel ruotato. Ogni anno solo un quarto del campione viene rinnovato; ciascun individuo che entri a far parte del campione viene quindi intervistato per 4 anni consecutivi, sia che rimanga all'interno della stessa famiglia, sia che ne formi una nuova. Questo permette di studiare, accanto alle dinamiche "macro" anche quelle microeconomiche relative alle tematiche coperte dall'indagine, cercando di limitare i problemi di "attrito" e la necessaria limitatezza nel tempo implicata dai "panel puri".

6. La strategia di stima adottata. Le stime dirette e la valutazione della loro affidabilità

Le stime dirette

Le stime pubblicate da ISTAT per l'indagine EU-SILC, così come per tutte le altre indagini condotte dall'Istituto sono basate su stimatori "da disegno". Si tratta di stimatori ottimizzati rispetto alle proprietà probabilistiche relative alla casualità dovuta alla selezione casuale del campione dalla popolazione. Non viene assunto nessun modello probabilistico che spieghi il modo in cui i dati di popolazione relativi ad una certa variabile siano generati. Questa scelta, adottata anche da tutti gli altri maggiori Istituti Nazionali di Statistica, è motivata dal tentativo di ridurre al minimo le ipotesi necessarie per giungere alle stime, in modo tale che queste possano essere viste, dal pubblico generale a cui sono indirizzate, come il più possibile oggettive, o per lo meno "neutrali".

In particolare, l'ISTAT utilizza stimatori per calibrazione (Deville e Särndal, 1992). Questi stimatori permettono di sfruttare informazioni ausiliarie disponibili da fonte amministrativa mantenendo per medie e totali la struttura a somma pesata, che rende particolarmente semplice per gli utenti riprodurre autonomamente le stime prodotte dall'istituto.

Se facciamo riferimento al dominio di studio (sottopopolazione) d , ad esempio, lo stimatore del totale Y di una data variabile y potrà essere espresso come:

$$\hat{Y}_d = \sum_{i=1}^{n_d} w_i y_i$$

dove w_i , con $i = 1, \dots, n_d$, è il peso campionario o fattore di riporto all'universo ed n_d indica il numero di unità campionarie relative al dominio d . Il peso per l'unità statistica i -sima è definito come l'inverso delle probabilità di inclusione moltiplicato per un fattore di correzione (calibrazione) che permette di correggere, almeno in parte, gli effetti della non-risposta. In gergo questi stimatori vengono detti "diretti" e "dirette" le stime ottenute sulla loro base.

Premesso questo, la strategia generale di stima che proponiamo per i parametri descritti nei paragrafi precedenti è la seguente:

- produrre le stime dirette dei parametri per tutti i domini di interesse;
- valutare l'affidabilità statistica degli stimatori diretti;
- sostituire le stime dirette con stime ottenute con metodologie più complesse ("da modello") per i parametri le cui stime dirette non risultano sufficientemente affidabili.

Tutte e tre le operazioni appena elencate sono piuttosto complesse.

Le stime dirette ottenute applicando i pesi pubblicati da ISTAT insieme ai dati di EU-SILC per l'Italia presenta aspetti problematici. I domini rappresentati dai distretti non sono infatti "domini pianificati" e sommando i pesi relativi alle unità del campione incluse nel dominio non si ottiene la dimensione esatta della popolazione (noto attraverso le anagrafi); inoltre se si considerano sotto-campioni classificando per sesso, età o altre variabili, il loro peso relativo in termini di somma dei pesi può essere diverso da quello reale nella popolazione del distretto. La calibrazione è infatti effettuata a livello regionale, il livello più dettagliato per cui ISTAT pubblica ufficialmente le stime.

Al fine di ottenere un sistema di pesi che, all'interno di ciascun dominio riporti alla effettiva dimensione della popolazione, si fa in modo che la somma dei pesi all'interno di ciascun dominio sommi alla dimensione della sua popolazione così come risulta dalle anagrafi comunali. Nella stessa direzione si può richiedere che i pesi dei sotto-campioni di distretto ottenuti classificando per sesso, classi d'età e nazionalità dei rispondenti sommino alla dimensione che queste sottopopolazioni hanno a livello di distretto. Questa idea, attraente in linea di principio, richiede in pratica che i campioni per cui si ricalibrano i pesi siano grandi. Una ricalibrazione a livello di piccoli domini può porre i pesi "sotto stress", aumentandone in modo anche cospicuo la variabilità. L'obiettivo di stimatori con una potenziale minore distorsione viene così ottenuto al prezzo di un forte incremento della variabilità degli stessi, secondo un "effetto disegno" ben noto in letteratura.

Una soluzione di compromesso, considerata in una versione preliminare di questo rapporto, può essere quella di utilizzare le province come livello di calibrazione, al posto dei distretti. Tecnicamente, ciò potrebbe portare ad un sistema di pesi che riporti ad una popolazione più "vicina" a quella del distretto, limitando tuttavia l'effetto di inflazione della varianza dovuto alla calibrazione di (sotto-)campioni numericamente molto piccoli. Un'analisi accurata ha rilevato tuttavia che questa soluzione non è soddisfacente: le differenze di struttura nelle popolazioni sono in genere più forti tra i distretti della stessa provincia che non tra una provincia e l'altra, per cui si finisce con l'ottenere pesi simili per struttura a quelli ricalibrati a livello regionale ma più variabili e che comunque non riportano ai totali di distretto e non rappresentano accuratamente la struttura della loro popolazione.

La decisione finale è stata quindi quella di calcolare le stime dirette utilizzando i pesi "ufficiali" prodotti da ISTAT, calibrati a livello regionale. Per gli stimatori, come i tassi, che hanno una struttura di media aritmetica, sono stati utilizzati stimatori diretti di tipo Hajék

$$\hat{Y}_d = \frac{\sum_{i=1}^{n_d} w_i y_i}{\sum_{i=1}^{n_d} w_i}$$

la cui dipendenza dalla mancata “calibrazione” dei pesi a livello locale è molto ridotta rispetto a stimatori che assumono nota la dimensione del dominio N_d : $\hat{Y}_d^* = N_d^{-1} \sum_{i=1}^{n_d} w_i y_i$

Infine, la valutazione dell’affidabilità degli stimatori, in altre parole la stima delle loro varianze, non può essere effettuata con metodi elementari: il disegno campionario è complesso e i parametri oggetto di stima sono funzioni non lineari dei dati. Per questo si è fatto ricorso ad un metodo di stima della varianza degli stimatori diretti basato sulla linearizzazione, adattando alle caratteristiche peculiari dell’indagine EU-SILC per l’Italia quanto precedentemente noto in letteratura.

I dettagli relativi alla stima della varianza delle stime dirette sono descritti in dettaglio nel paragrafo A5 dell’Appendice A.

Valutazione dell’affidabilità delle stime dirette

Il campione EU-SILC per l’Italia è un campione di grandi dimensioni. Le stime dei parametri descrittivi della povertà e della disuguaglianza distributiva del reddito a livello nazionale, ripartizionale e regionale sono in generale caratterizzate da un livello di affidabilità sufficiente. L’adeguatezza della precisione di queste stime è coerente con l’obiettivo dichiarato di EU-SILC di produrre stime utili per lo studio e il monitoraggio della coesione sociale a livello regionale.

Come anticipato, nel caso di parametri le cui stime fossero sufficientemente affidabili l’analisi può essere condotta sulle stime “dirette”, mentre occorre introdurre stime “da modello” per gli altri.

Per valutare il grado di affidabilità delle stime, è stato utilizzato il coefficiente di variazione al fine di rendere più agevoli i confronti fra la variabilità delle stime dei diversi parametri.

I risultati relativi all’affidabilità delle stime dirette riferite al 2019, valutata attraverso il CV %, sono riportati in dettaglio nelle Tabella A2 ed A4 del par. A6 (Appendice A), rispettivamente per i distretti e per le province dell’Emilia-Romagna.

In sintesi, in quasi tutte le province il CV % riferito alla stima dell’AROP supera il 25% con un massimo pari al 49.25%. Con riferimento alla stima del SMD in 7 province su 9 il CV % supera il 50%. Il CV % riferito alla stima del QJ supera sempre il 25%. E, infine, per 4 province su 9 il CV% della stima dell’AROPE è superiore al 25%.

Con riferimento ai distretti, in numero superiore alle province, per aiutare le considerazioni di sintesi, riportiamo nella Tab. 1 alcune statistiche sintetiche relative all'affidabilità delle stime dirette. In particolare, vengono riportati, per ciascun parametro oggetto di stima, i valori minimo, mediano, medio e massimo del coefficiente di variazione per distretto. Si noti che non è stato possibile calcolare le stime dirette per i 9 distretti in cui non si aveva nessuna famiglia campionata. In linea di principio non è possibile calcolare le misure di povertà (e stimare le loro varianze) anche nel caso in cui vi siano famiglie campionate nel distretto, ma nessuna tra loro risulti povera. In particolare, gli algoritmi di lisciamento delle varianze (di cui tratteremo nel dettaglio nel par. A5 dell'Appendice A e B4.1 dell'Appendice B) permettono di assegnare un ragionevole livello di incertezza anche alle stime il cui coefficiente di variazione stimato sulla base dell'algoritmo diretto risulti uguale a 0.

Nel calcolo dei coefficienti di variazione sono stati esclusi i casi in cui esso è uguale a 0. Dalla Tabella 1 è facile notare come le stime dell'AROP e dell'AROE siano caratterizzati, per un numero pari circa al 50% dei distretti, da coefficienti di variazione molto elevati e superiori alla soglia del 25% adottata da ISTAT. La scelta della soglia del 25% è in linea con altri istituti nazionali di statistica, per discriminare tra le stime "pubblicabili" e quelle non sufficientemente affidabili per essere pubblicate. Il caso dei tassi di deprivazione materiale severa e di bassa intensità lavorativa appaiono caratterizzati da un grado minore di affidabilità in quanto circa il 50% dei distretti ha un CV% superiore al 50%. I valori massimi dei CV% sono per tutti gli indicatori molto elevati e intorno al 100%.

Tabella 1. Statistiche descrittive relative alla distribuzione dei coefficienti di variazione delle stime dirette per distretto

Statistiche	Coefficiente di Variazione %			
	rischio di povertà o esclusione sociale AROE	rischio di povertà AROP	deprivazione materiale severa SMD	bassa intensità lavorativa QJ
minimo	5.7	7.2	11.8	13.5
mediana	26.0	30.4	51.2	45.9
media	29.3	35.4	55.0	49.7
massimo	98.3	102.8	103.7	105.7

La numerosità riferita al campione estratto per l'Italia nell'indagine EU-SILC 2020 è, purtroppo, calata drasticamente: si passa da un campione formato da 43400 individui ad una numerosità pari a 29252. Tale riduzione interessa, come è ovvio, anche la regione

Emilia-Romagna che vede passare la numerosità dai 3177 per il 2019 ai 2163 individui per il 2020, con una riduzione pari al 32% circa.

7. I modelli per “piccole aree”

Al fine di risolvere il problema dell'inefficienza delle stime dirette di alcuni parametri, dovuta alla scarsa numerosità campionaria all'interno dei distretti e delle province, abbiamo proceduto “sostituendole” con stime ottenute con metodi per piccole aree, seguendo un'impostazione “da modello”. Nel contesto della stima “da modello” si suppone che tra i parametri di interesse esista una “relazione strutturale” e lo sfruttamento di questa relazione strutturale permette di migliorare la precisione della stima in un dominio, utilizzando l'informazione proveniente da tutti gli altri domini. La relazione strutturale tra i parametri di interesse tipicamente coinvolge informazioni ausiliarie, utili per spiegare la variabilità sottostante ai parametri stessi.

In questa sede non vogliamo addentrarci nei dettagli relativi alla specificazione e alla stima dei modelli per piccole aree utilizzati, e al modo in cui tecnicamente le stime dei parametri descrittivi della povertà ed esclusione sociale sono ottenute sulla base di questi modelli (si rinvia per questo all'Appendice B). Ci limiteremo, schematicamente ad evidenziare qui alcune scelte che caratterizzano il lavoro che abbiamo svolto.

- Tipo di modelli; i modelli per piccole aree possono essere classificati in due grandi famiglie: “area level” e “unit level” a seconda che il modello probabilistico venga specificato al livello delle stime di dominio oppure di unità statistica (si veda Rao, 2003, cap. 5). In questa ricerca sono stati considerati solo modelli di tipo “area level”, poiché permettono una più agevole specificazione di distribuzioni statistiche in linea con la diversa natura dei parametri e di sfruttare informazioni statistiche ausiliarie specifiche per area, in mancanza di informazioni ausiliarie specifiche per unità.
- Informazione ausiliaria; la disponibilità di informazioni di carattere fiscale, come il reddito imponibile ed il numero delle dichiarazioni dei redditi a livello comunale, ha consentito di calcolare per distretto tassi importanti come l'imponibile pro-capite, il numero delle dichiarazioni per residente di 15 o più anni ed il reddito imponibile dichiarato in media. Altre informazioni considerate a livello di distretto sono di tipo anagrafico, come la quota di stranieri residenti e la densità abitativa.
- I criteri utilizzati per selezionare queste informazioni sono stati, oltre all'efficacia statistica, la possibilità di ottenere queste informazioni su base regolare anno per anno e l'interpretabilità, anche economica, dei modelli specificati. In generale le informazioni di origine fiscale si sono rivelate quelle con maggiore capacità esplicativa nei confronti delle misure considerate. Per

maggiori dettagli sulle informazioni ausiliarie utilizzate per ogni parametro oggetto di stima si veda l'Appendice B.

- L'impostazione seguita nella stima è quella Bayesiana. Le ragioni dietro questa scelta sono molteplici: i) i recenti progressi nei metodi computazionali, in particolare dei metodi Markov Chain Monte Carlo, permettono la specificazione di modelli complessi senza che ciò comporti un aggravio dei problemi di stima. In particolare è agevole specificare e analizzare modelli che non facciano ricorso all'ipotesi di normalità come in questa ricerca; ii) la specificazione e l'analisi di modelli Bayesiani gerarchici permette di tenere adeguatamente in conto tutte le fonti di incertezza riguardanti i dati e quindi di produrre stime di cui è possibile valutare in modo esaustivo l'affidabilità; nel caso specifico della stima per piccole aree i dati con struttura informativa gerarchica si riferiscono a più livelli di osservazione/appartenenza, ad es. individuale e territoriale (ma anche individuale e sociale). Nel caso specifico della stima per piccole aree si utilizzano naturalmente modelli gerarchici perché le osservazioni individuali sono naturalmente raggruppate (e quindi hanno una struttura gerarchica) in domini; iii) al di là delle differenze concettuali di fondo, stime "frequentiste" e bayesiane, finiscono nei casi più semplici per essere sostanzialmente analoghe. Nel caso dell'analisi di modelli relativamente complessi, come quelli considerati in questa ricerca, i predittori puntuali ottenuti come sintesi delle distribuzioni a posteriori, mostrano in generale buone proprietà frequentiste, così come le varianze a posteriori costituiscono buone approssimazioni degli errori quadratici medi frequentisti.
- Al fine di sfruttare la relazione spaziale tra le stime, si specifica un modello con effetti di area correlati spazialmente. Se ci fosse una relazione spaziale tra le stime la si potrebbe sfruttare e "prendere a prestito" informazione dalle aree spazialmente "vicine" in modo da sfruttare la relazione eventualmente presente tra fenomeni in differenti aree. Si ritiene, in altre parole, che ciò che si osserva in un'area è determinato, almeno in parte, da ciò che accade in altre aree a questa vicine.
- Al fine di sfruttare la relazione tra le stime, si specifica un modello multivariato in cui si ipotizza una struttura di dipendenza tra le stime dei quattro tassi. Non è detto che tale correlazione sia presente ma, nei casi in cui lo fosse, la specificazione di un modello multivariato (questo termine è usato con riferimento alle variabili dipendenti) consente, nella stima di uno dei tassi, di "prendere a prestito informazione" dai restanti.

8. Le stime da modello

In questo paragrafo ci occuperemo della descrizione dei principali risultati ottenuti con le procedure di stima descritte nei paragrafi precedenti e nelle Appendici A e B, sia in

termini di validazione dei modelli utilizzati per ottenere le stime da modello, sia con riferimento alla rappresentazione dei fenomeni oggetto di studio.

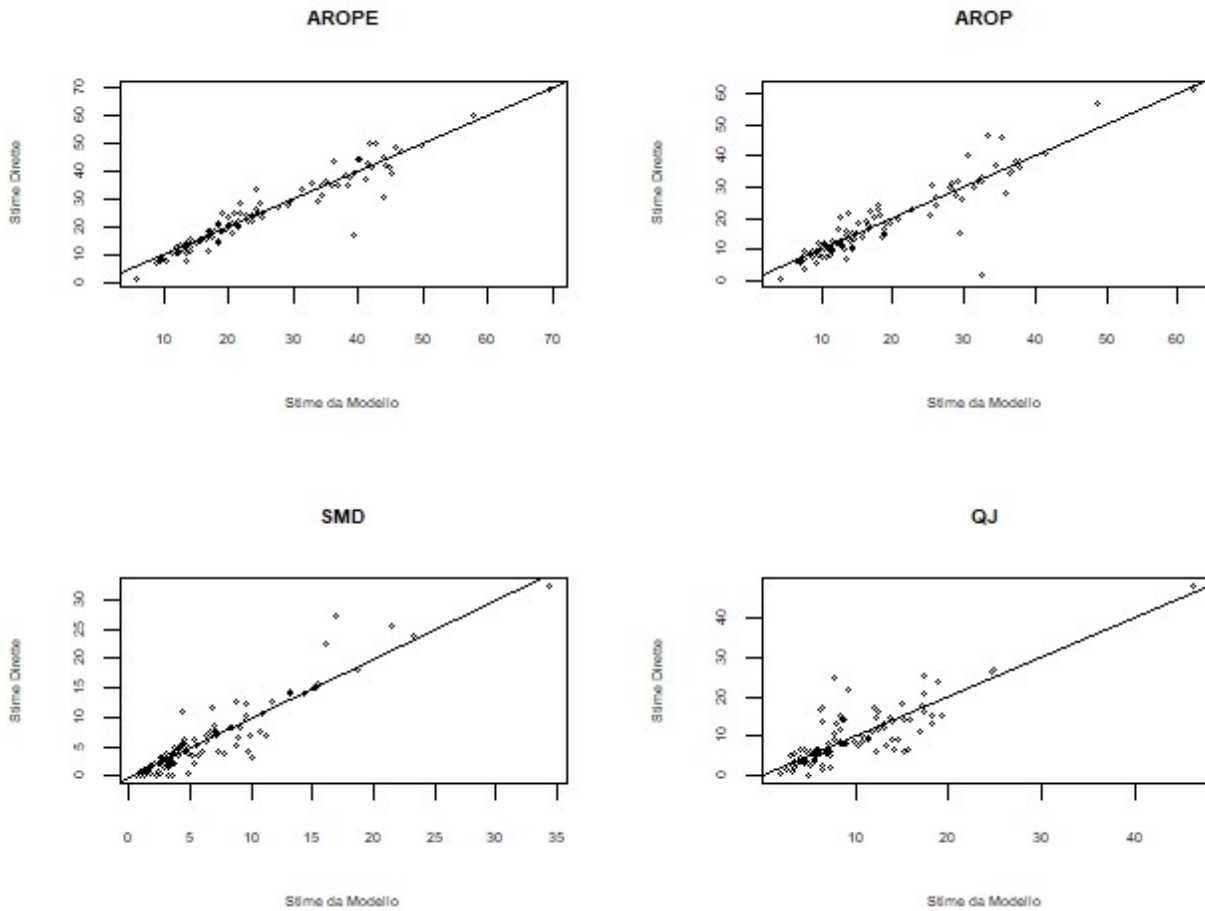
La verifica e validazione dei modelli viene condotta con riferimento al set di dati 2019, che costituisce, appunto, il primo insieme di informazioni sulla base del quale viene individuata e validata la strategia di stima che poi sarà usata anche negli anni successivi.

Con riferimento alla rappresentazione del fenomeno, gli aspetti che metteremo in evidenza non esauriscono tuttavia le informazioni contenute nei dati e altri aspetti potrebbero essere colti da analisti diversi. Si preferisce, inoltre, non avanzare possibili interpretazioni relative alle cause delle diverse ed eterogenee situazioni che si delineano sulla base delle stime o delle relazioni che potrebbero essere ravvisate tra indicatori ed altri fenomeni che caratterizzano il sistema economico. Infine, al fine di non appesantire la trattazione, non saranno commentati i risultati relativi agli intervalli di confidenza rappresentati nei grafici seguenti dai segmenti sovrapposti alle barre che riferiscono delle stime puntuali. Il lettore può ovviamente valutare per i casi di interesse l'eventuale grado di sovrapposizione degli intervalli (si tenga tuttavia in conto che una moderata sovrapposizione è compatibile con una differenza "significativa" tra le stime).

Nel seguito si riportano innanzitutto alcune statistiche descrittive e grafici utili a valutare la performance del metodo di stima impiegato. Tale analisi è condotta solo con riferimento ai dati 2019, in quanto la performance del metodo di stima viene valutato solo sui dati riferiti al primo anno di stima. In particolare si considereranno i) la relazione tra stime dirette e stime da modello: poiché le stime dirette, benché non affidabili, sono non distorte, è importante che le stime da modello siano nel complesso quanto più possibile vicine a quelle dirette, ii) la relazione tra CV delle stime dirette e CV delle stime da modello: è auspicabile che i secondi siano apprezzabilmente più piccoli dei primi, iii) la riduzione del CV delle stime da modello rispetto al CV delle stime dirette in relazione alla numerosità campionaria: ci si attende che tale riduzione sia orientativamente più elevata per aree con numerosità campionaria piccola. L'analisi sarà effettuata per provincia, dato il numero non elevato di distretti che possono essere considerati.

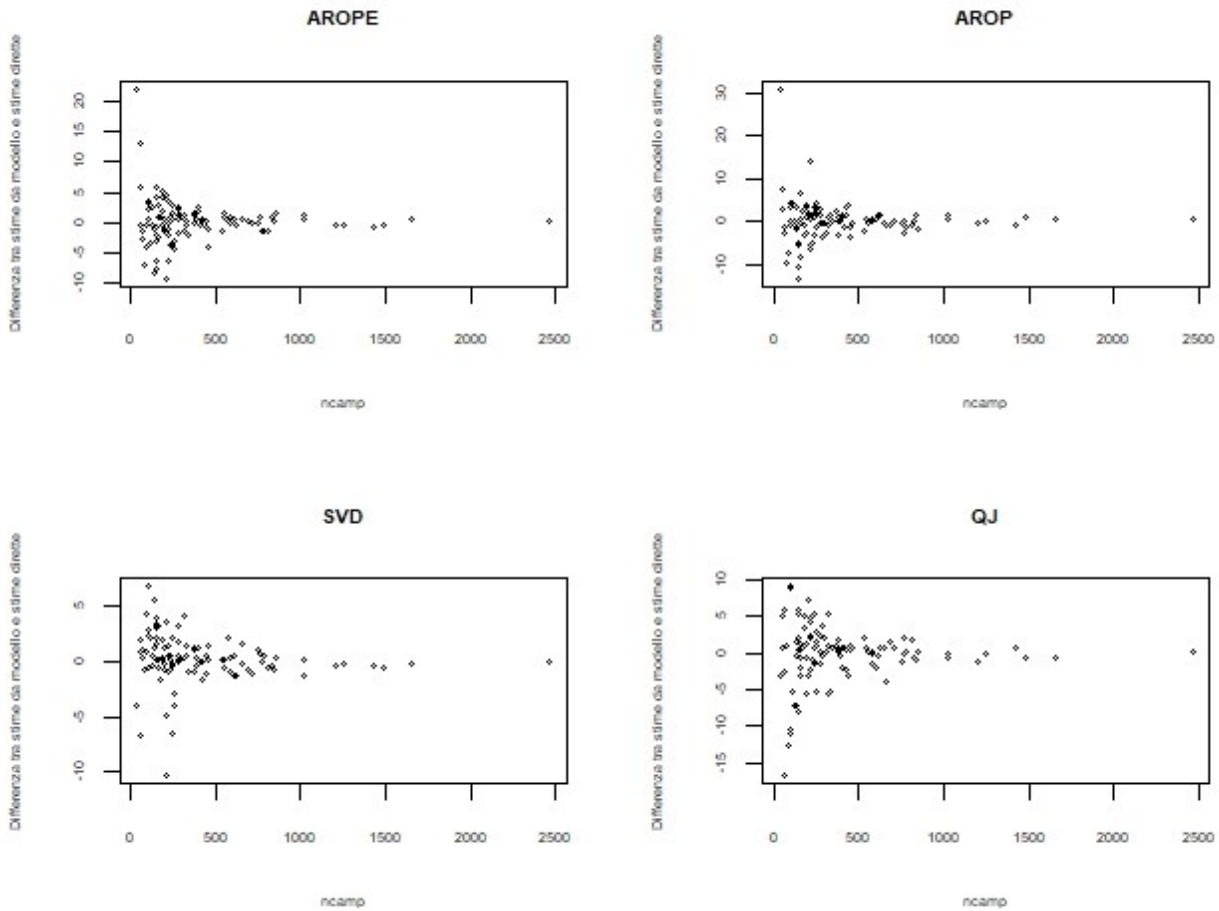
Dal Gr. A emerge che le stime dirette e quelle da modello, per tutti i parametri considerati, sono nella gran parte sulla bisettrice di primo e terzo quadrante, e sono quindi molto vicine. La distanza relativa a queste stime per alcune province si interpreta sia considerando l'effetto delle covariate, sia con riferimento all'uso di un modello con struttura di relazione spaziale che ha "lasciato" le stime dirette sulla base delle relazioni tra aree contigue. I coefficienti di correlazione tra le stime dirette e le stime da modello sono, rispettivamente per AROPE, AROP, SVD e QJ, pari a 0.96, 0.92, 0.93, 0.83.

Gr. A – Stime dirette vs Stime da modello per le province



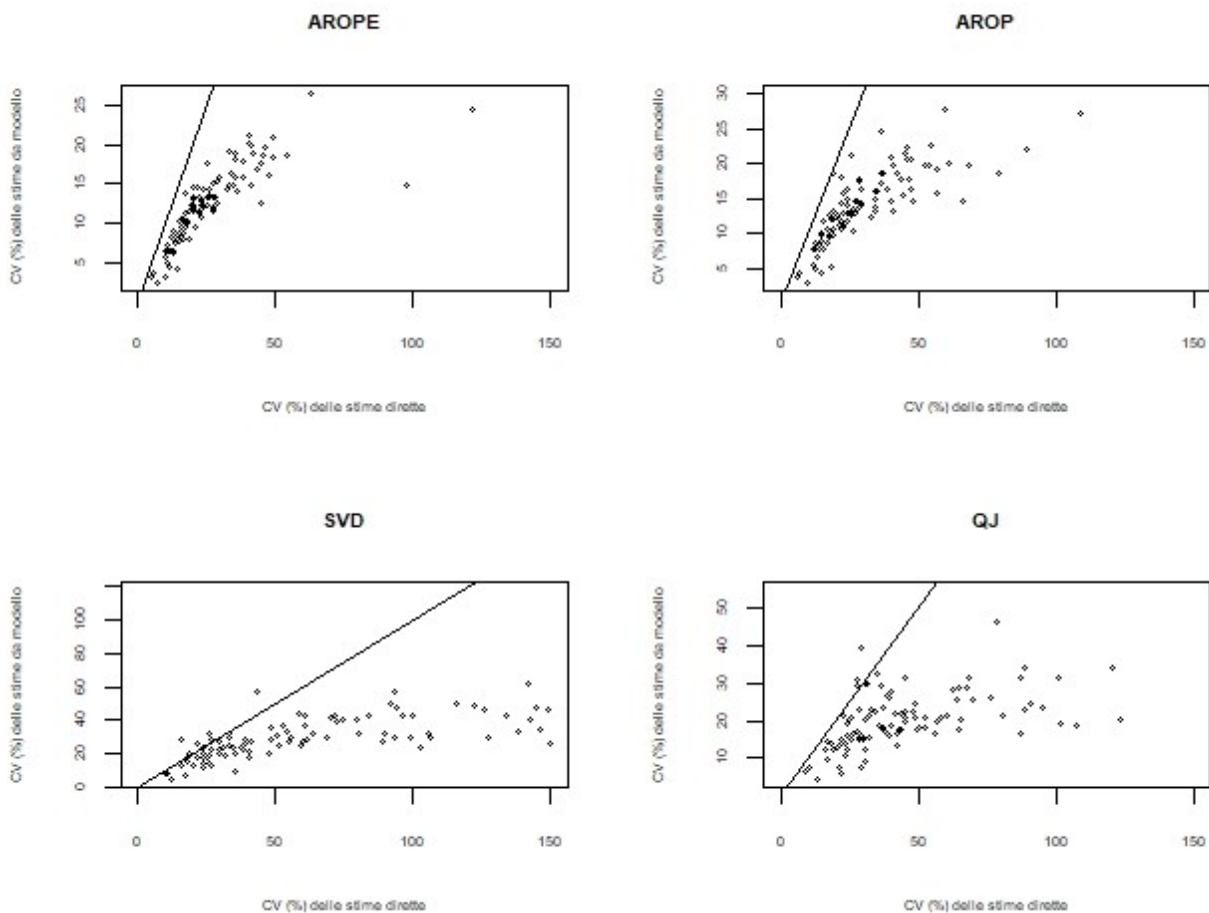
Nel Gr B si valuta la consistenza da disegno delle stime da modello. Come si può osservare, la differenza tra stime da disegno e stime da modello si riduce all'aumentare della numerosità campionaria e tale configurazione testimonia che il modello agisce modificando maggiormente le stime dirette nel caso in cui queste si basino su una numerosità campionaria piccola mentre, via via che la numerosità aumenta, il modello produce stime che si identificano sempre più alle stime dirette.

Gr. B – Differenza tra stime da modello e stime dirette vs numerosità campionaria



Il Gr. C, in cui si descrive graficamente la relazione tra coefficienti di variazione delle stime dirette e delle stime da modello, è di aiuto per valutare la capacità dei modelli utilizzati di ridurre la variabilità delle stime dirette. Come si può osservare, i CV delle stime da modello sono inferiori per la quasi totalità delle aree ai corrispondenti CV riferiti alle stime dirette, in alcuni casi decisamente inferiori. Questo risultato, unito a quello commentato in precedenza, conferma che i modelli di stime per piccole aree utilizzati hanno raggiunto l'obiettivo di ridurre in maniera cospicua la variabilità delle stime dirette.

Gr. C – Coefficiente di Variazione % delle stime dirette^a vs Coefficiente di Variazione % delle stime da modello



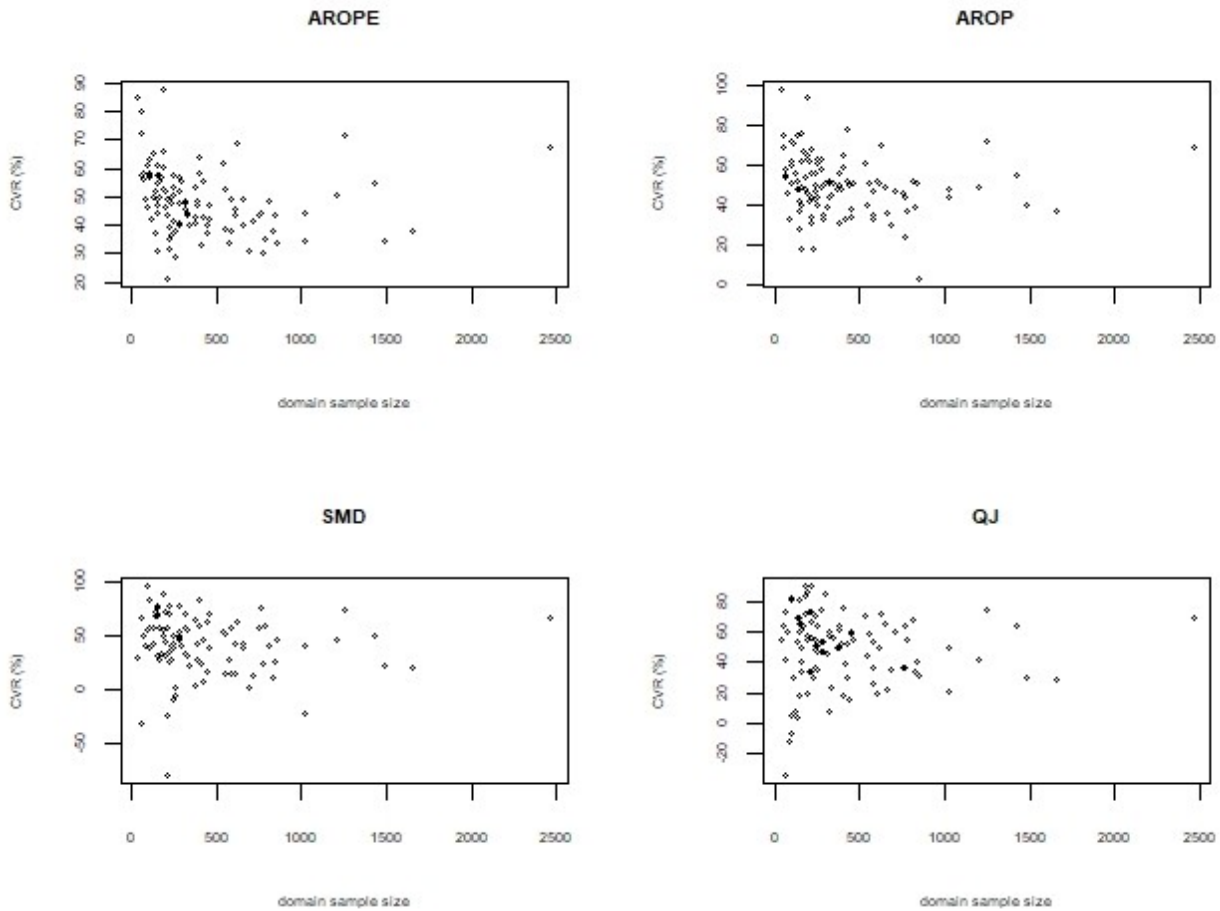
^a Il valore massimo dell'asse delle ascisse (CV% delle stime dirette) è stato fissato a 150 per consentire al grafico di mostrare l'andamento generale della relazione tra CV delle stime dirette e da modello (in 2 o 3 province, non appartenenti alla regione Emilia-Romagna, il CV delle stime dirette risulta più elevato del valore massimo qui menzionato).

Quale ulteriore indicatore di performance, consideriamo la riduzione del CV, definita per ogni area d come segue:

$$CVR_d = 100 \times \left(1 - CV_d^{\text{mod}} / CV_d^{\text{dir}}\right);$$

Nel Gr. D si riporta la riduzione % del CV vs la numerosità dell'area. Come ci si attende a seguito dell'impegno di un modello di stima per piccole aree, la riduzione è, in generale, più cospicua per le aree caratterizzate da numerosità campionaria più piccola.

Gr D - Riduzione % del CV vs numerosità dell'area



Nella Tab. A alcune statistiche di sintesi, riferite ancora a tutte le province italiane, relative al CV delle stime da modello. Le statistiche considerate mostrano un deciso livello di affidabilità per le stime riferite ad AROPE, con CV % leggermente più elevati per l'At Risk Of Poverty. Le stime di SVD e QJ soffrono, invece, di livelli di variabilità più elevati, come ci si attendeva sia perché riguardano un numero più contenuto di individui, sia perché la capacità predittiva delle covariate disponibili è meno spiccata per questi ultimi due parametri rispetto ad AROPE ed AROP.

Tab. A – Statistiche di sintesi dei CV % delle stime da modello

	Coefficiente di Variazione %			
	AROPE	AROP	SVD	QJ
Min.	2.37	3.04	4.04	4.20
1st Qu.	8.58	10.71	21.80	16.09
Median	12.17	13.50	29.24	20.35
Mean	12.29	14.14	33.22	20.92
3rd Qu.	14.86	17.66	41.69	24.60
Max.	26.57	30.15	117.90	54.77

In breve, i grafici e le statistiche di sintesi considerati testimoniano che il modello utilizzato per produrre stime per piccole aree produce stime decisamente più affidabili rispetto a quelle dirette, in generale prossime a quelle da disegno e comunque caratterizzate da consistenza da disegno.

Nelle successive sezioni presenteremo quindi le stime da modello dei quattro parametri considerati, per provincia e per i distretti sociosanitari della regione Emilia-Romagna, ed i rispettivi Coefficienti di Variazione.

Le stime per provincia

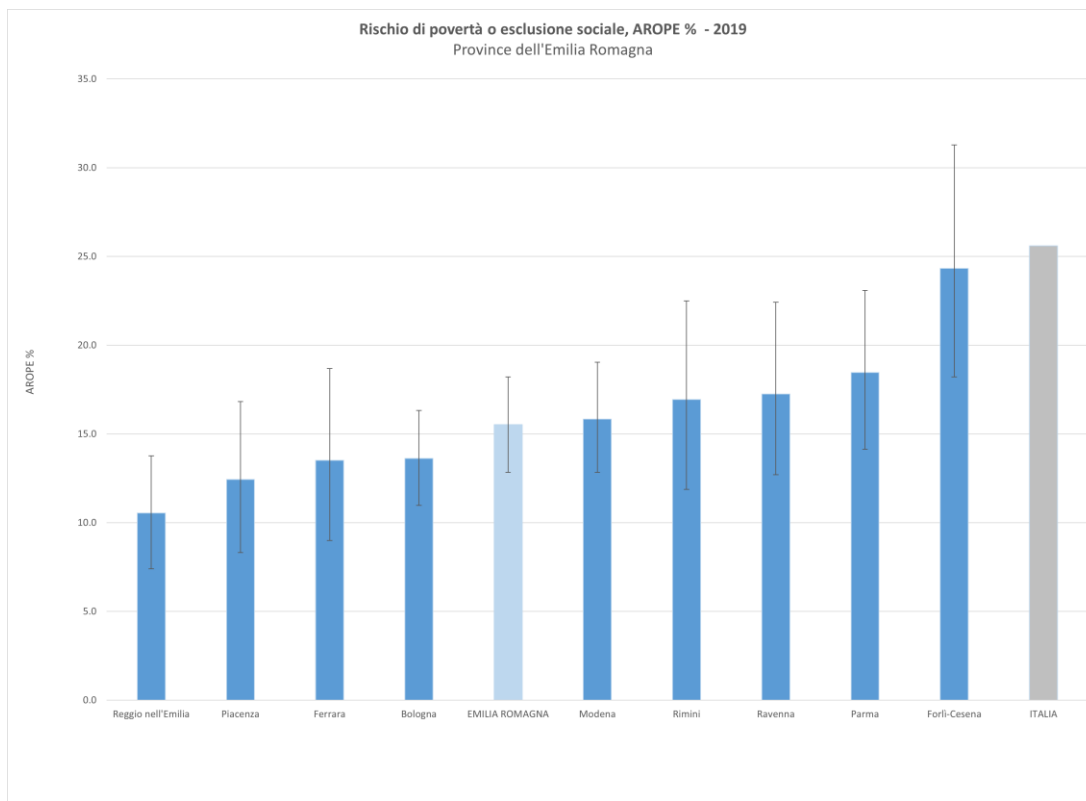
La Tabella 2.1 contiene le stime da modello dei quattro parametri di interesse, per le province dell'Emilia-Romagna, per la regione Emilia-Romagna e per l'Italia, ed i rispettivi coefficienti di variazione, riferiti al 2019. Nei grafici 1-4 riportiamo le stime dei quattro indicatori per provincia, ordinati per valore della stima, a cui aggiungiamo le stime riferite all'Emilia-Romagna nel suo insieme ed all'Italia. Su ognuna delle barre dei grafici è riportato un segmento che individua l'intervallo di confidenza per la stima. A seguire si riportano le mappe per i 4 parametri, riferite alle province italiane.

Come si diceva, accanto alle stime puntuali, si riporta il relativo CV. I CV sono decisamente accettabili per tutte le stime prodotte. La variabilità delle stime di SVD e QJ è più elevata rispetto a quella che è stata ottenuta per gli altri parametri; ciò è dovuto al più contenuto potere previsivo delle covariate utilizzate (paragrafo B.3). Nel commentare le stime ottenute di SVD e QJ si noti quindi che i relativi intervalli di confidenza sono più ampi rispetto a quelli riferiti ai restanti parametri.

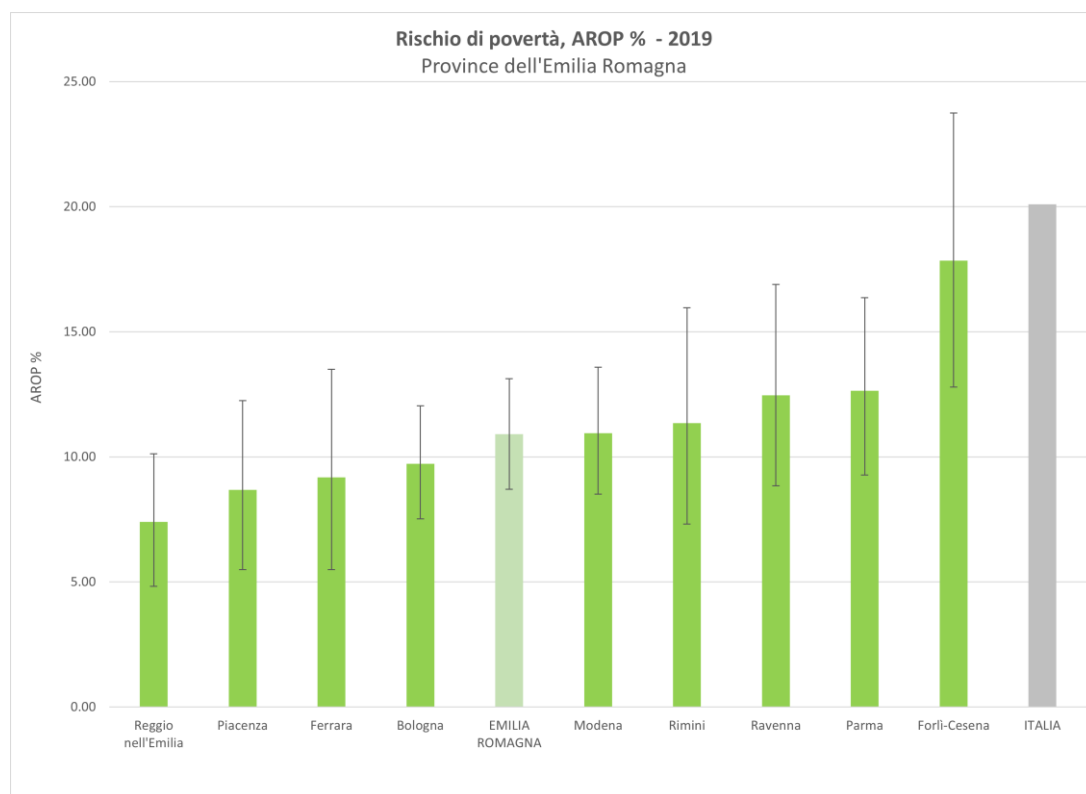
Tab. 2.1 - Stime da modello dei parametri obiettivo (%) e relativi CV%, per le province dell'Emilia-Romagna, anno 2019

codice prov	Provincia	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà - AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ
stime					
33	PC	12.44	8.68	2.55	4.12
34	PR	18.46	12.65	6.69	6.97
35	RE	10.54	7.40	1.74	2.86
36	MO	15.83	10.95	4.55	4.56
37	BO	13.62	9.73	3.49	5.06
38	FE	13.51	9.19	3.66	5.87
39	RA	17.25	12.47	3.22	5.54
40	FC	24.32	17.85	6.80	6.28
99	RN	16.94	11.36	3.01	3.36
	ER	15.53	10.91	3.96	4.94
	Italia	25.60	20.10	7.40	10.00
CV%					
33	PC	17.53	19.79	37.62	29.05
34	PR	12.22	14.24	24.04	19.57
35	RE	15.69	17.90	35.58	29.63
36	MO	9.95	11.87	20.57	17.51
37	BO	10.09	11.91	20.17	16.66
38	FE	18.54	22.03	41.09	31.69
39	RA	14.17	16.49	29.04	22.33
40	FC	13.78	15.91	32.12	22.18
99	RN	16.10	19.10	39.85	34.27
	ER	8.83	10.35	23.82	15.18

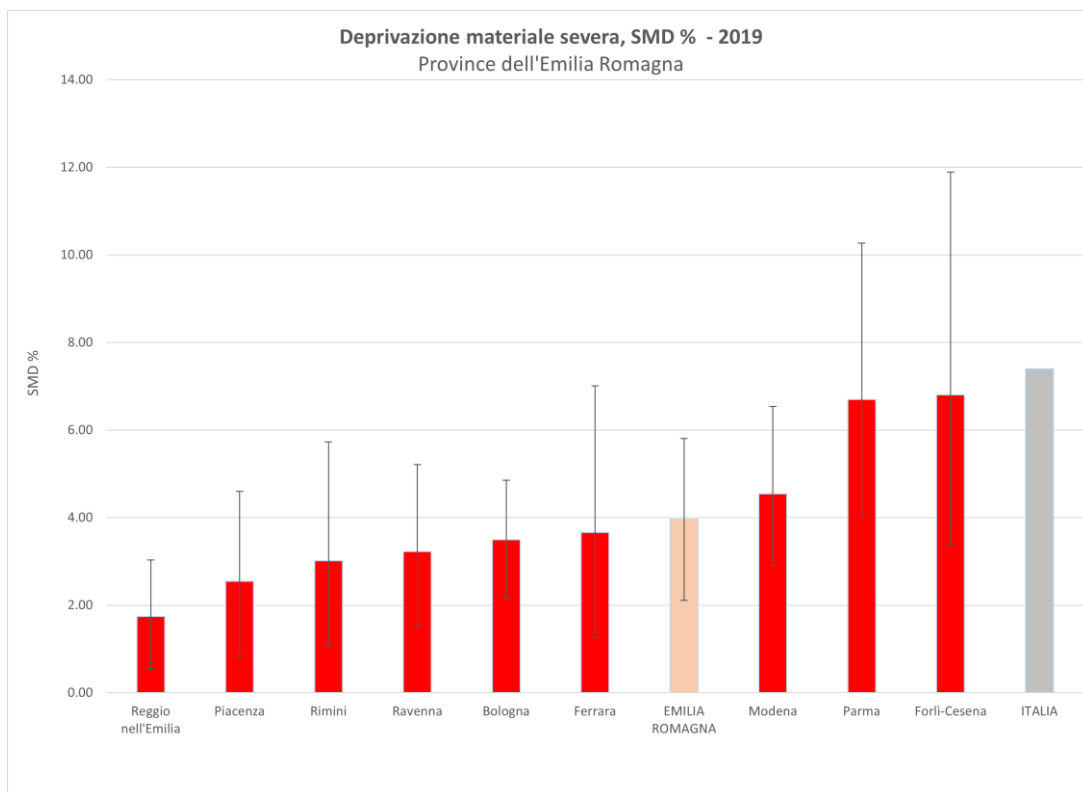
Gr. 1 — Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROPE) per le province dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



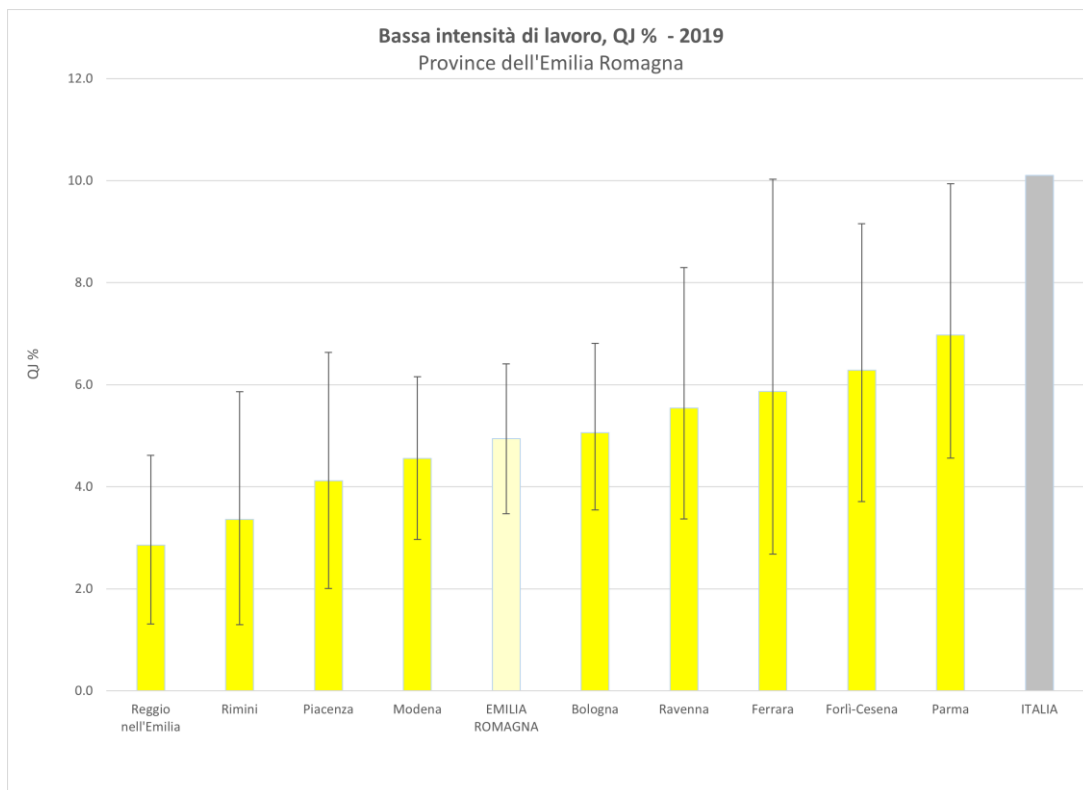
Gr. 2 - Stime da modello del Tasso di rischio povertà (AROP) per le province dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)

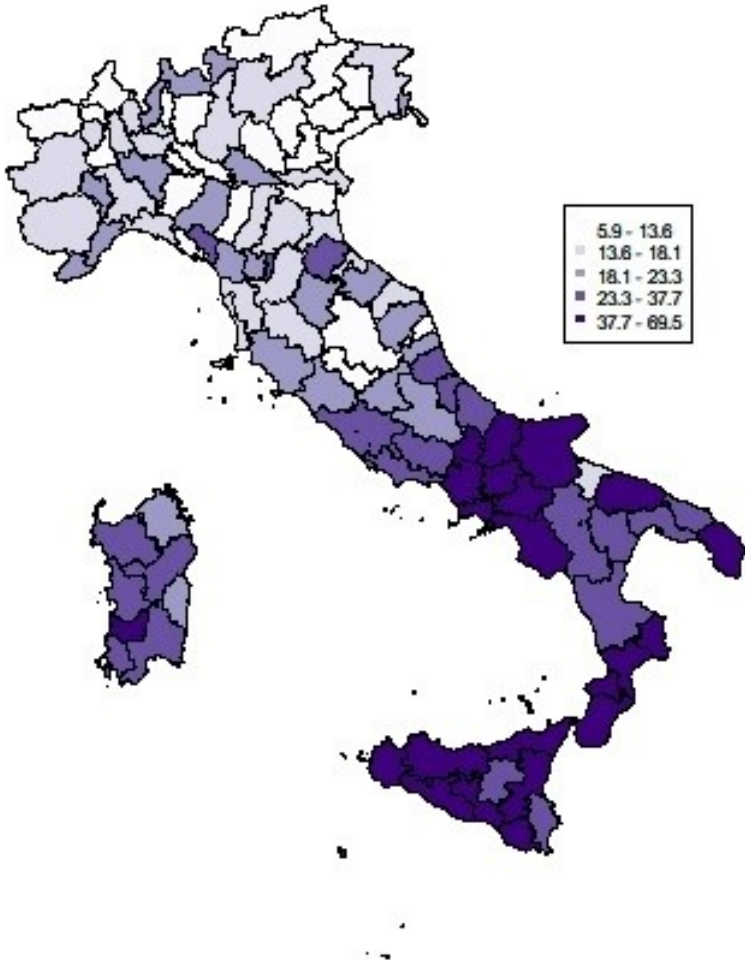


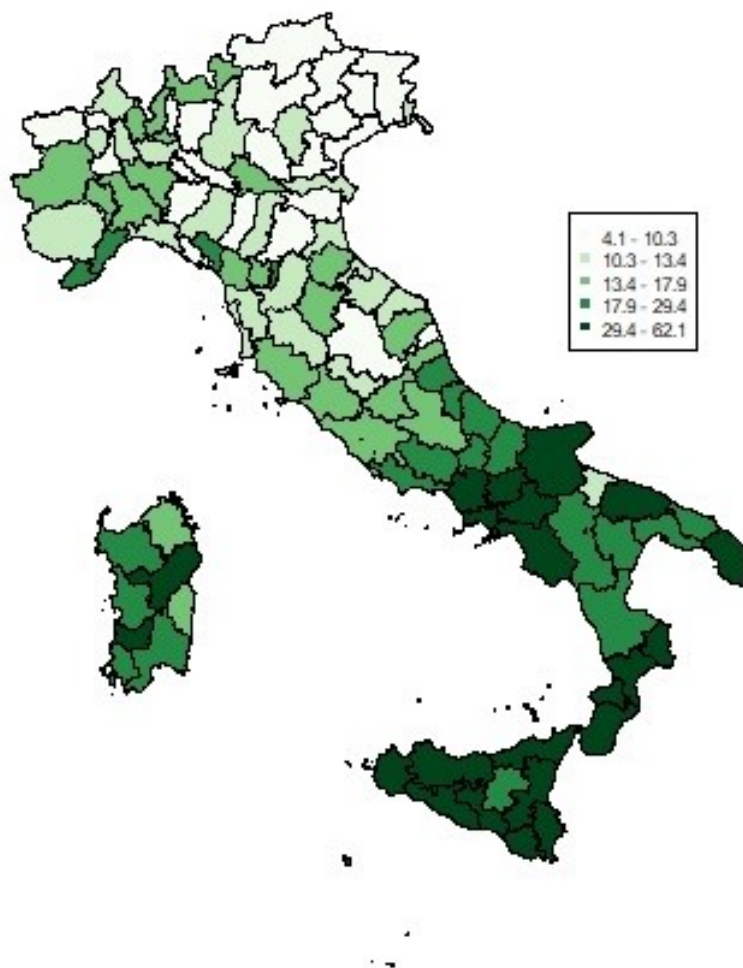
Gr. 3 - Stime da modello del Tasso di deprivazione materiale severa (SMD) per le province dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)

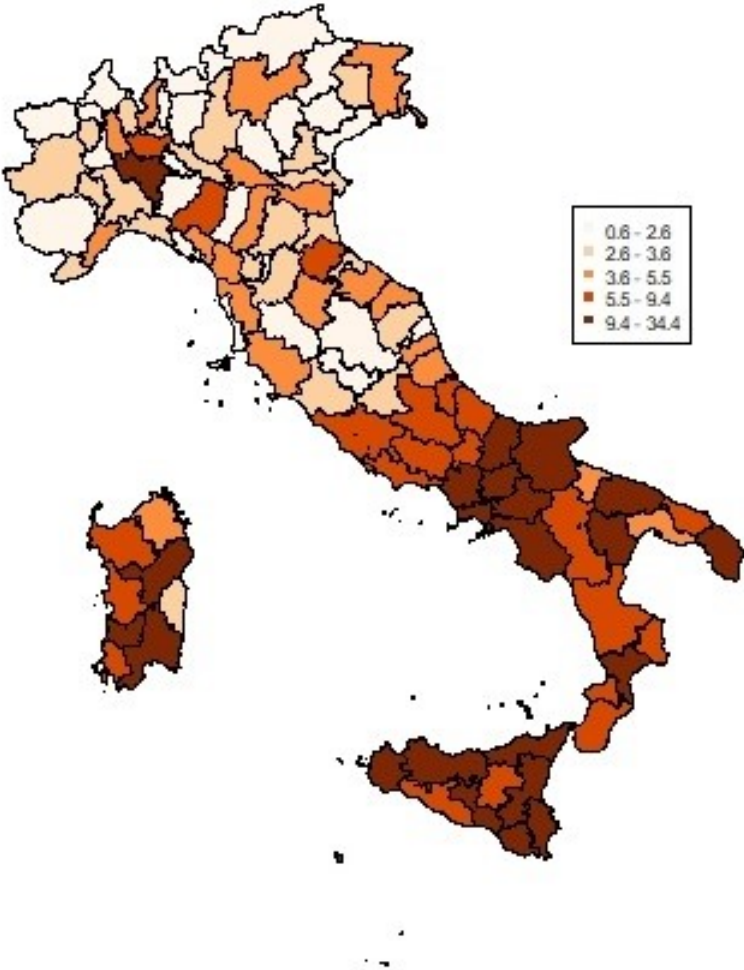


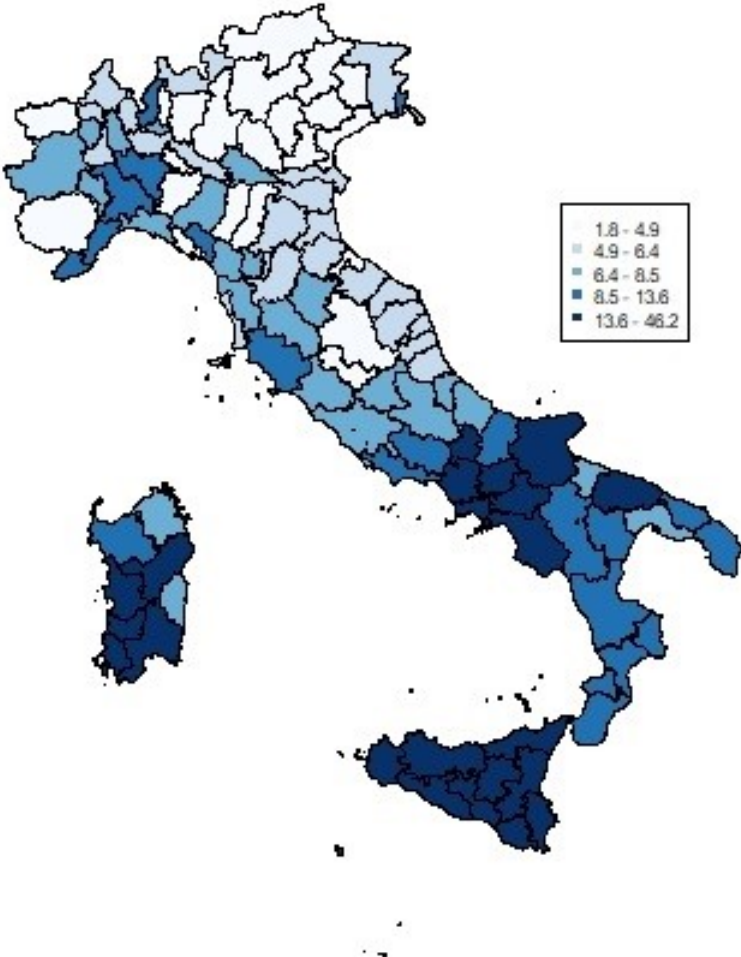
Gr. 4 - Stime da modello del Tasso di bassa intensità lavorativa (QJ) per le province dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)









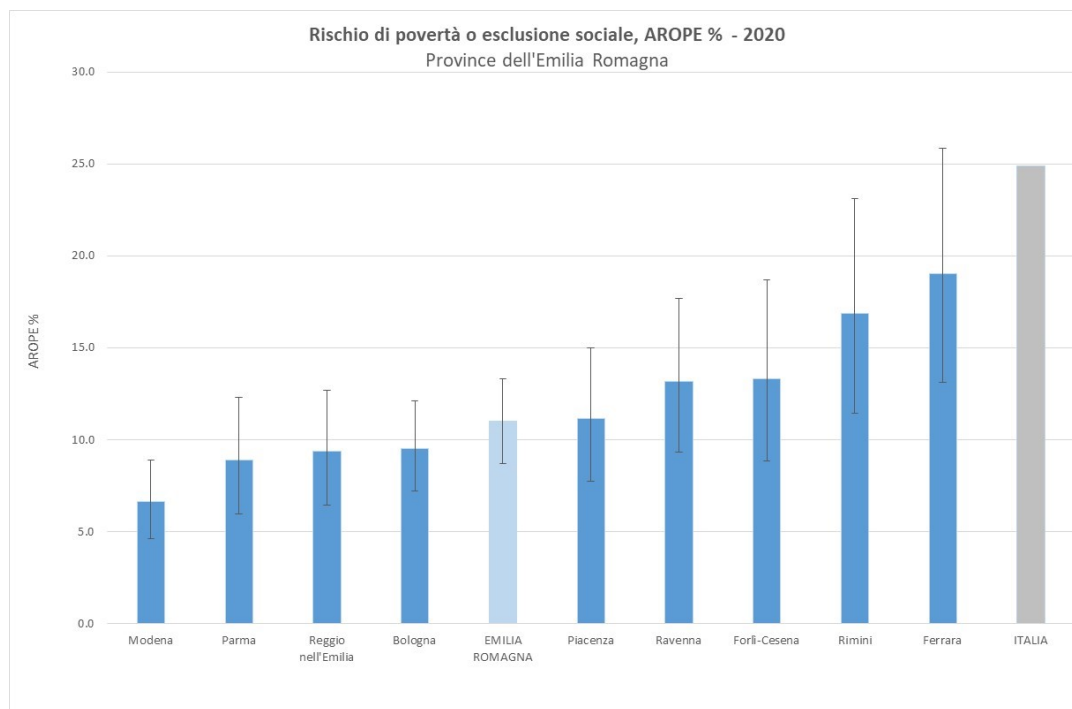


Nel seguito riportiamo le stime ottenute per il 2020 con riferimento alle province (Tab. 2.2, grafici 5-8 e relative mappe). Il set dei risultati relativi al 2020 viene presentato con la medesima struttura utilizzata per il 2019. Come già notato, le numerosità 2020 sono decisamente inferiori rispetto a quelle 2019. Come ci si attendeva, questa diminuzione si riflette, in generale, su un aumento della variabilità delle stime dirette e quindi delle stime da modello 2020 che risultano caratterizzate da una minore affidabilità rispetto a quelle 2019.

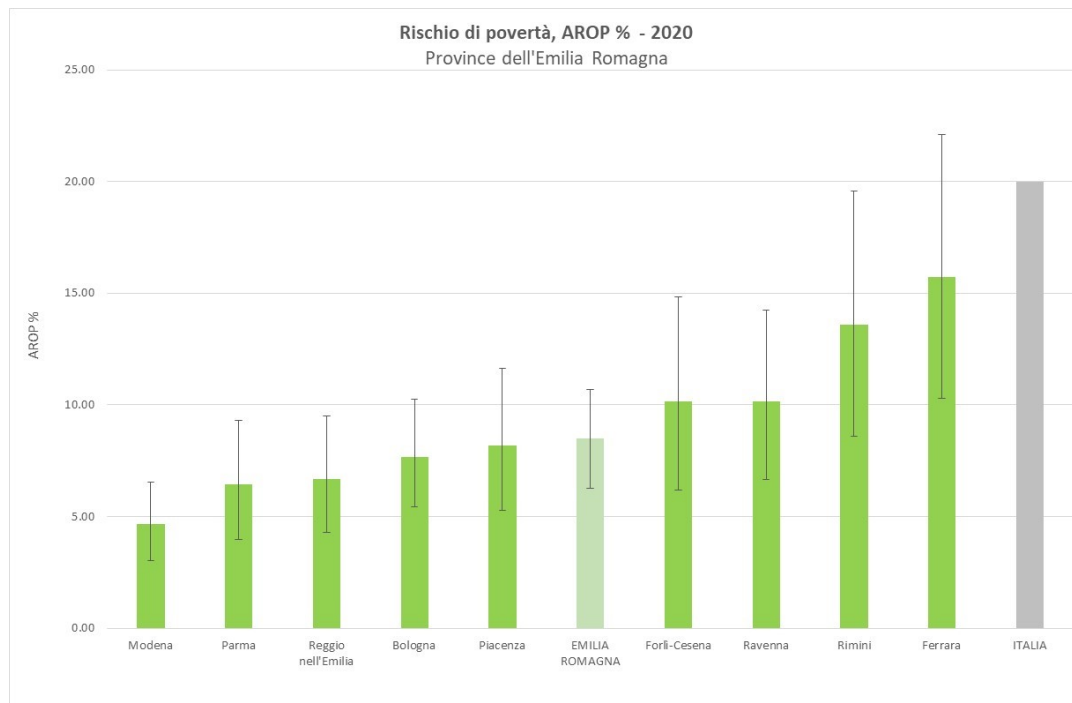
Tab. 2.2 - Stime da modello dei parametri obiettivo (%) e relativi CV%, per le province dell'Emilia-Romagna, anno 2020

codice prov	Provincia	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà – AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ
stime					
33	PC	11.2	8.2	1.7	3.8
34	PR	8.9	6.4	1.3	3.4
35	RE	9.4	6.7	1.2	3.8
36	MO	6.6	4.7	1.0	2.3
37	BO	9.5	7.6	0.9	4.4
38	FE	19.0	15.7	1.8	8.3
39	RA	13.2	10.2	1.5	5.4
40	FC	13.3	10.1	1.6	5.6
99	RN	16.9	13.6	2.2	6.9
	ER	11.0	8.5	1.3	4.5
	Italia	24.9	20.0	3.5	11.1
CV%					
33	PC	16.7	19.3	33.6	33.9
34	PR	18.1	20.9	34.6	34.8
35	RE	17.1	19.9	31.3	29.9
36	MO	16.2	19.1	29.2	29.0
37	BO	13.3	15.8	27.8	22.5
38	FE	17.1	19.5	45.0	30.2
39	RA	16.4	19.0	34.2	29.6
40	FC	18.6	21.6	40.9	34.7
99	RN	17.8	20.7	42.1	32.4
	ER	11.6	13.3	31.1	17.2

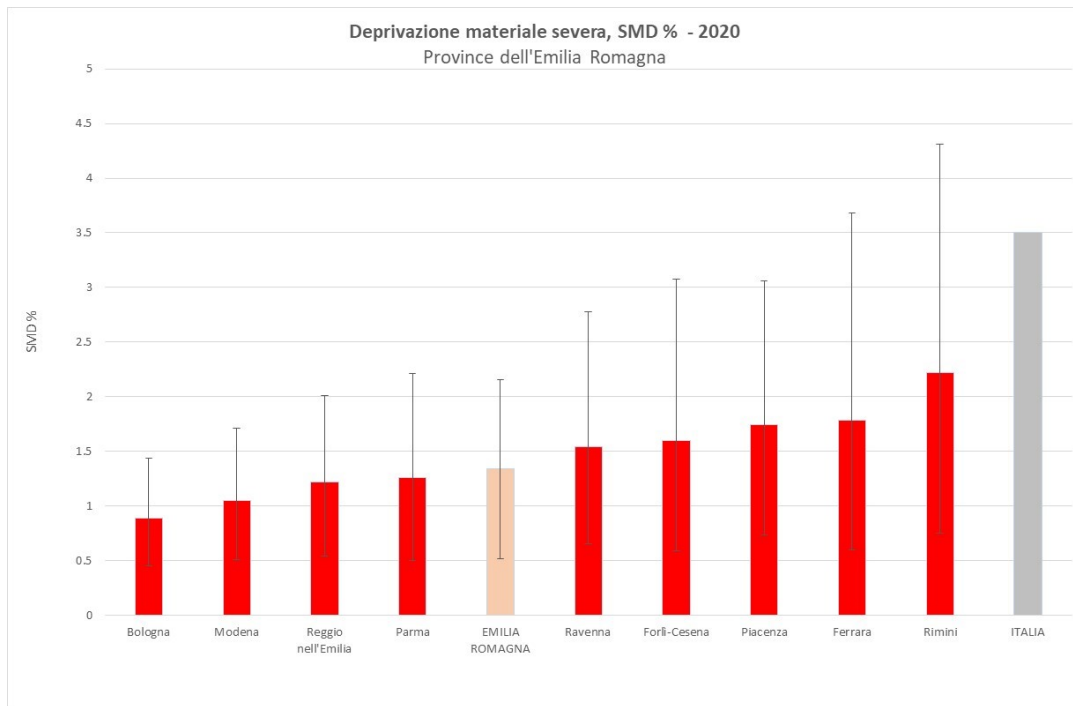
Gr. 5 — Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROPE) per le province dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



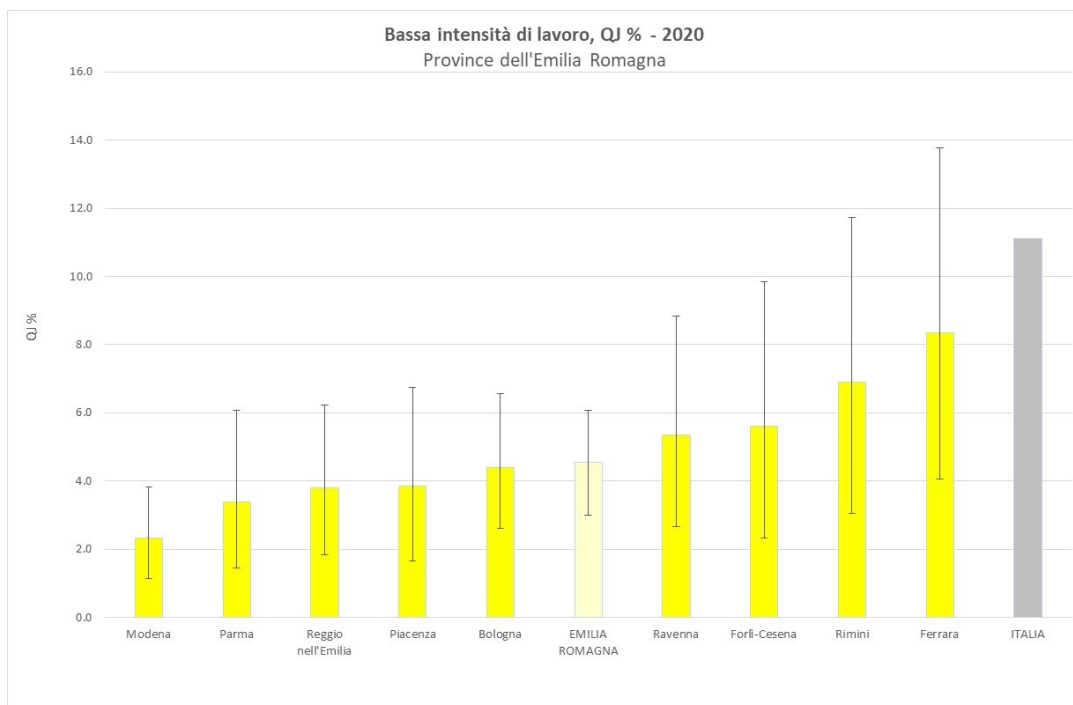
Gr. 6 - Stime da modello del Tasso di rischio povertà (AROP) per le province dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)

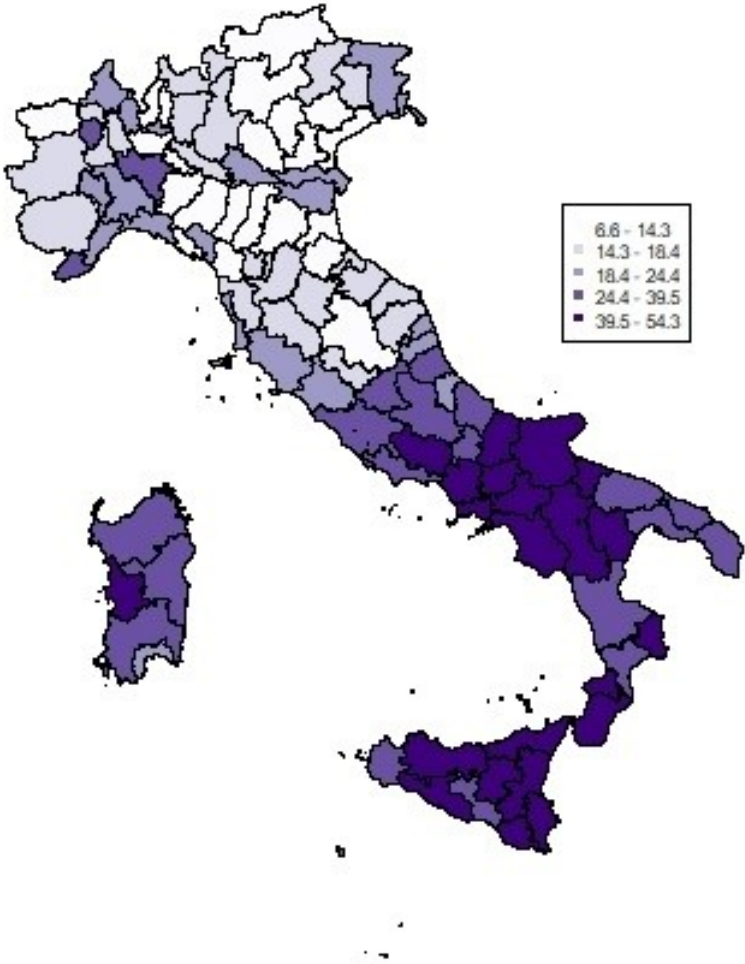


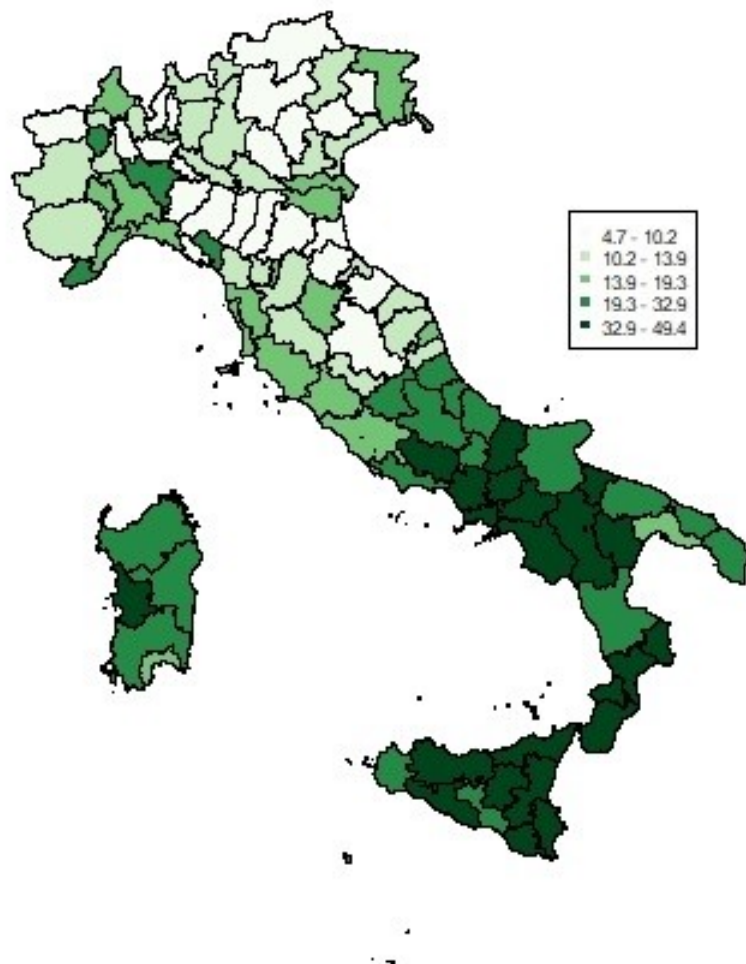
Gr. 7 - Stime da modello del Tasso di deprivazione materiale severa (SMD) per le province dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)

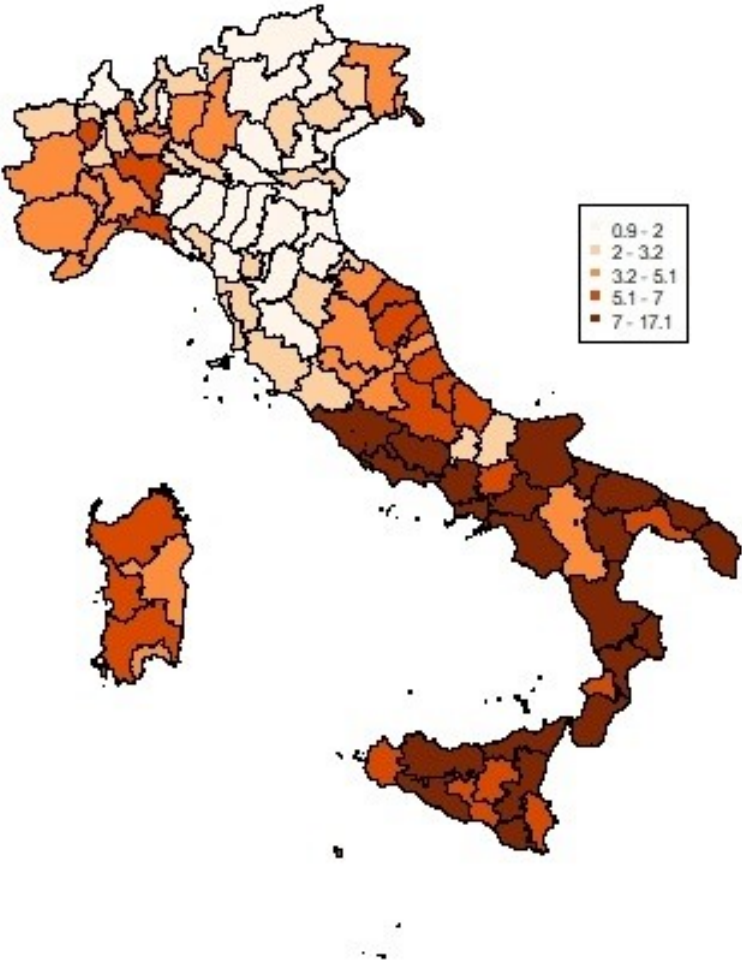


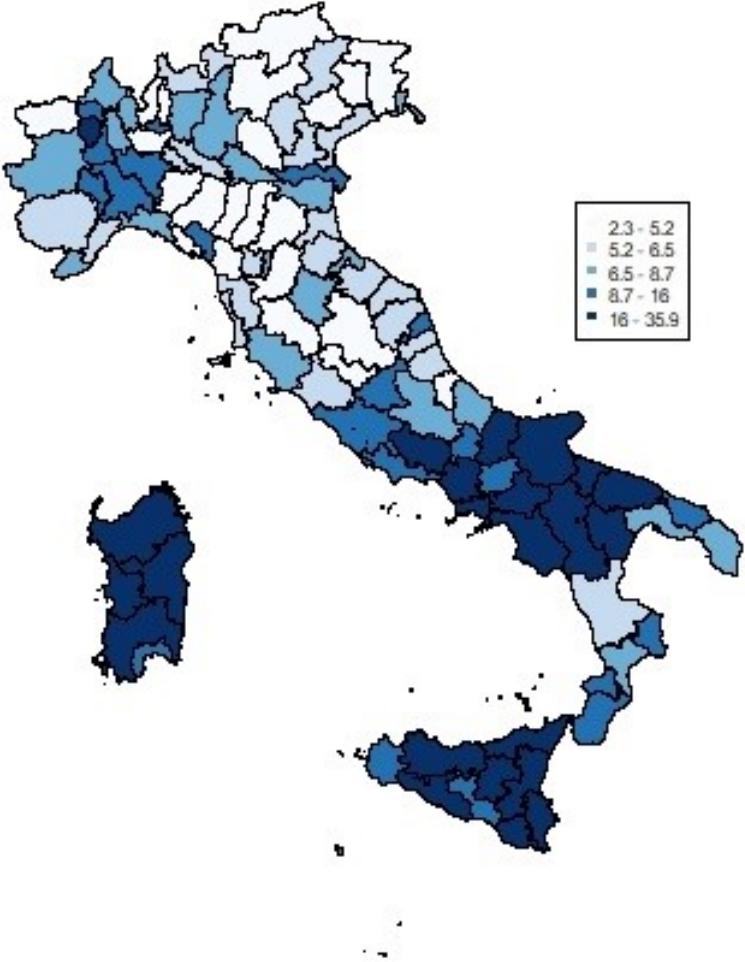
Gr. 8 - Stime da modello del Tasso di bassa intensità lavorativa (QJ) per le province dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)





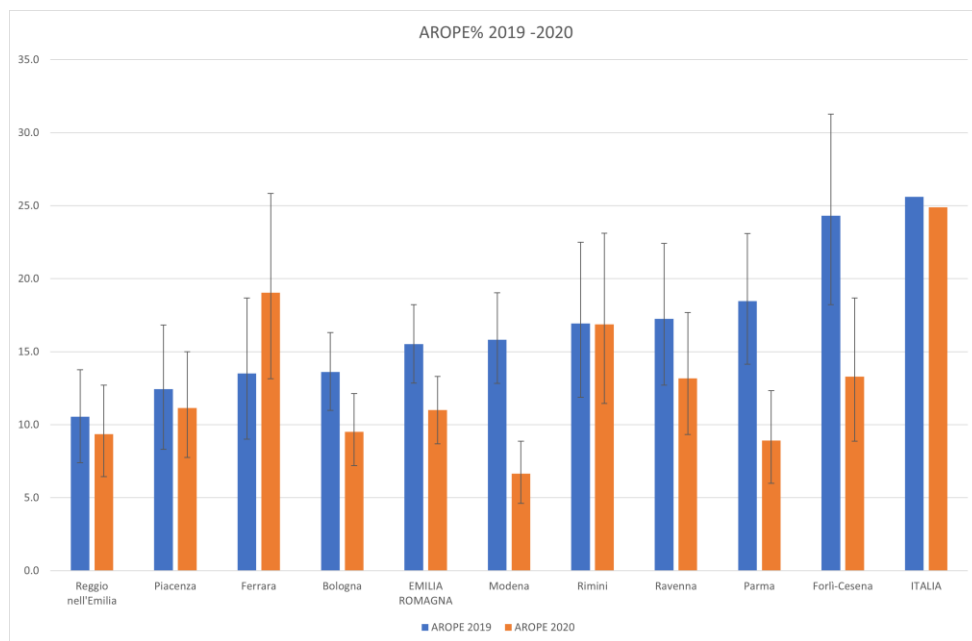




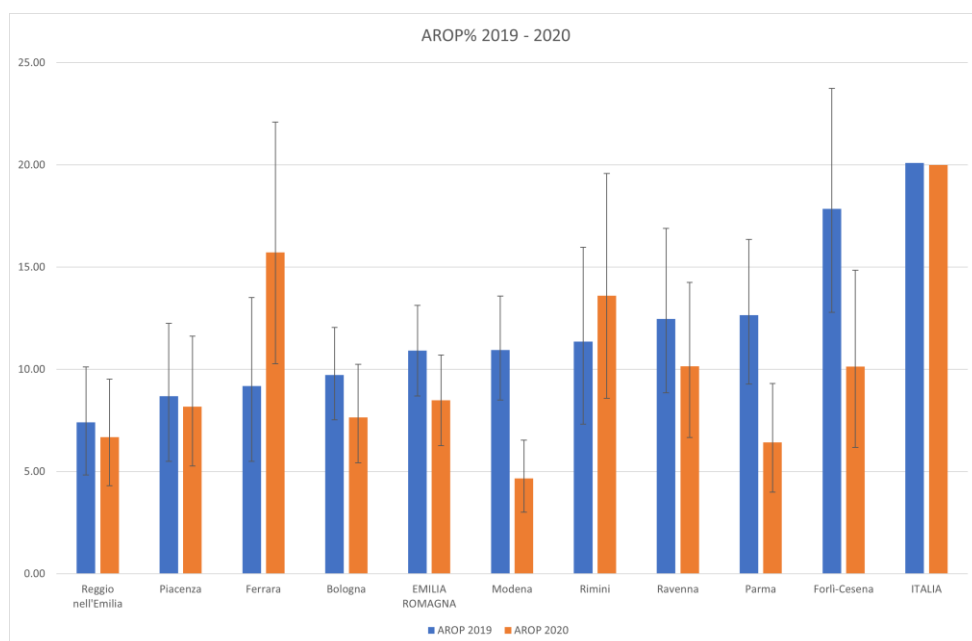


Nel seguito si confrontano le stime dal modello ottenute per le province dell'Emilia-Romagna per il 2019 e 2020 (Gr. 5.1-8.1). Tali confronti vanno effettuati con qualche cautela in quanto le stime dei due anni non sono ottenute sotto le stesse condizioni in relazione alle numerosità campionarie molto diverse.

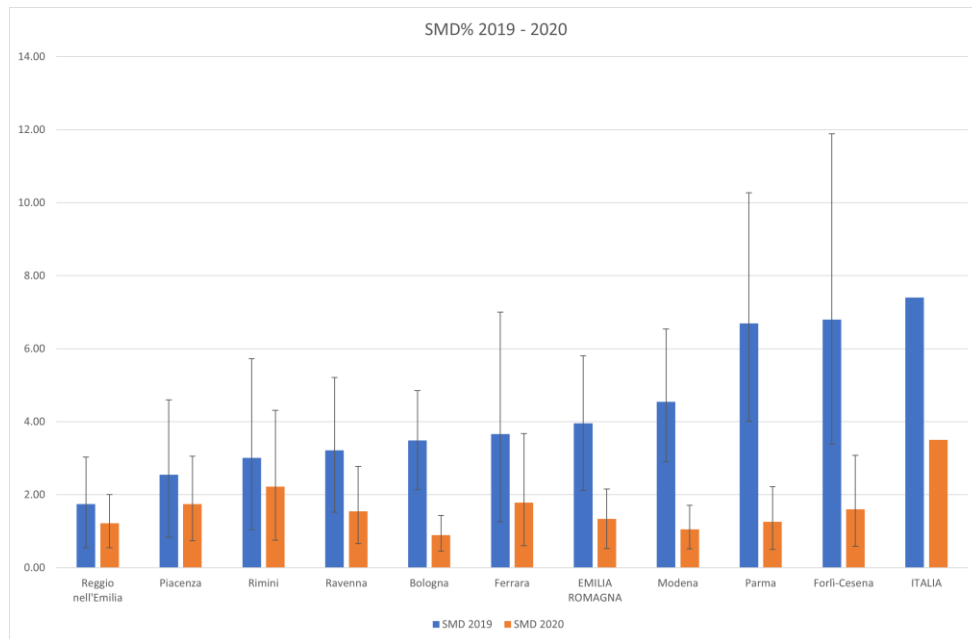
Gr. 5.1 -- Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROPE) per le province dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



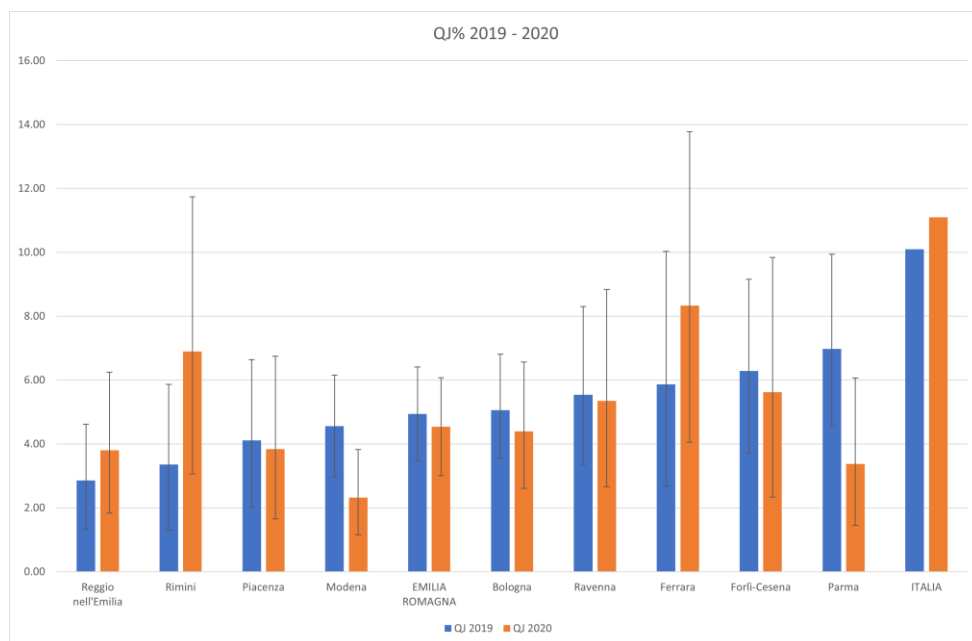
Gr. 6.1 -- Stime da modello del Tasso di rischio povertà (AROP) per le province dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Gr. 7.1 -- Stime da modello del Tasso di deprivazione materiale severa (SMD) per le province dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Gr. 8.1 -- Stime da modello del Tasso di bassa intensità lavorativa (QJ) per le province dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Le stime per i distretti sociosanitari

La Tabella 3.1 contiene le stime da modello dei quattro parametri di interesse (considerati in percentuale), relative al 2019, per i distretti sociosanitari dell'Emilia-Romagna, per la regione Emilia-Romagna e per l'Italia. La Tabella 4.1 riporta i rispettivi coefficienti di variazione percentuali.

I CV sono accettabili per i parametri AROPE e AROP. La variabilità delle stime riferita a SVM e QJ è tuttavia più elevata a causa del contenuto potere previsivo delle covariate disponibili (vedi paragrafo B.5).

Nei grafici 9-12 riportiamo inoltre le stime dei quattro indicatori per l'anno 2019 ordinati per valore della stima riferita alle province, a cui aggiungiamo le stime riferite all'Emilia-Romagna nel suo insieme ed all'Italia. Su ognuna delle barre dei grafici è riportato un segmento che individua l'intervallo di confidenza per la stima.

Si noti che, SMD e QJ assumono, per alcuni distretti, valori molto piccoli, a volte vicini allo 0. In questi casi, come testimoniato dalla letteratura (Ospina and Marmolejo-Ramos, 2019), il CV può non essere adatto a quantificare l'errore di stima. In tali casi, una più appropriata valutazione della variabilità degli stimatori la si può effettuare mediante gli intervalli di confidenza presenti nei grafici sopra menzionati.

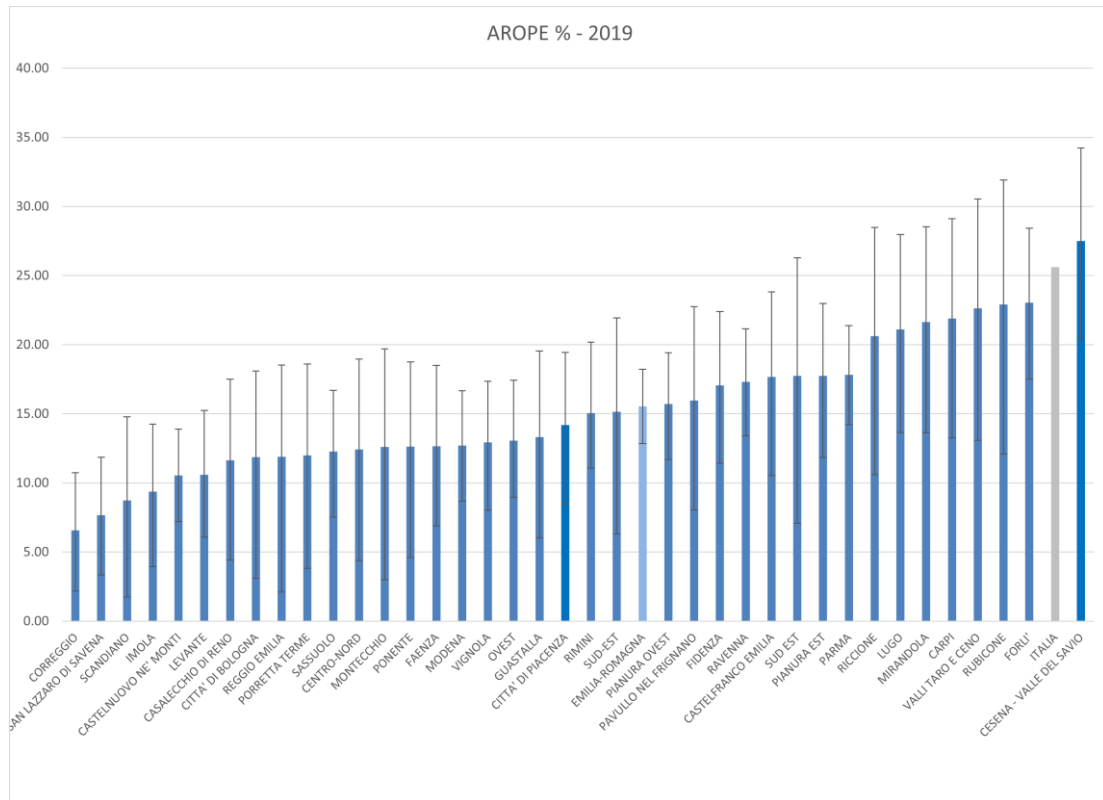
Tab. 3.1 - Stime da modello (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà – AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n ₂₀₁₉
Bologna	CASALECCHIO DI RENO	11.64	7.86	2.56	4.11	35
Bologna	CITTA' DI BOLOGNA	11.87	8.65	3.74	5.44	0
Bologna	IMOLA	9.37	6.23	1.81	2.62	59
Bologna	PIANURA EST	17.74	12.25	6.33	8.62	208
Bologna	PIANURA OVEST	15.72	11.46	3.55	4.89	292
Bologna	PORRETTA TERME	12.00	9.76	2.28	5.90	33
Bologna	SAN LAZZARO DI SAVENA	7.67	5.63	2.15	3.08	123
Ferrara	CENTRO-NORD	12.44	8.25	3.74	5.39	0
Ferrara	OVEST	13.07	8.72	3.82	6.08	128
Ferrara	SUD-EST	15.16	10.78	3.31	5.88	0
Forli	CESENA - VALLE DEL SAVIO	27.48	20.53	9.22	8.52	129
Forli	FORLI'	23.04	16.90	5.83	5.51	79
Forli	RUBICONE	22.91	16.35	5.68	5.00	0
Modena	CARPI	21.90	15.99	5.56	6.86	47
Modena	CASTELFRANCO EMILIA	17.66	11.91	5.88	6.01	116
Modena	MIRANDOLA	21.64	15.61	6.05	6.87	73
Modena	MODENA	12.70	7.99	4.83	3.01	197
Modena	PAVULLO NEL FRIGNANO	15.96	12.75	2.58	5.79	52
Modena	SASSUOLO	12.26	8.23	3.06	3.57	113
Modena	VIGNOLA	12.92	8.80	3.09	2.45	111
Parma	FIDENZA	17.06	12.05	4.09	6.48	108
Parma	PARMA	17.83	11.52	8.24	6.38	179
Parma	SUD EST	17.74	14.11	3.81	6.79	0
Parma	VALLI TARO E CENO	22.63	15.95	7.25	9.48	50
Piacenza	CITTA' DI PIACENZA	14.18	9.99	2.50	4.48	53
Piacenza	LEVANTE	10.58	7.35	1.87	3.42	135
Piacenza	PONENTE	12.63	8.74	3.55	4.58	0
Ravenna	FAENZA	12.67	8.98	2.19	3.27	101
Ravenna	LUGO	21.10	15.73	4.25	6.88	64
Ravenna	RAVENNA	17.31	12.34	3.14	5.86	114
Reggio	CASTELNUOVO NE' MONTI	10.54	7.21	1.65	2.45	174
Reggio	CORREGGIO	6.57	4.12	0.99	1.99	122
Reggio	GUASTALLA	13.31	9.13	2.08	3.21	60
Reggio	MONTECCHIO	12.60	9.17	2.59	4.00	0
Reggio	REGGIO EMILIA	11.89	8.69	2.39	3.92	0
Reggio	SCANDIANO	8.73	7.66	0.63	2.78	42
Rimini	RICCIONE	20.63	14.28	4.22	4.66	0
Rimini	RIMINI	15.06	9.86	2.40	2.70	180
	ER	15.53	10.91	3.96	4.94	3177
	Italia	25.60	20.10	7.40	10.00	43400

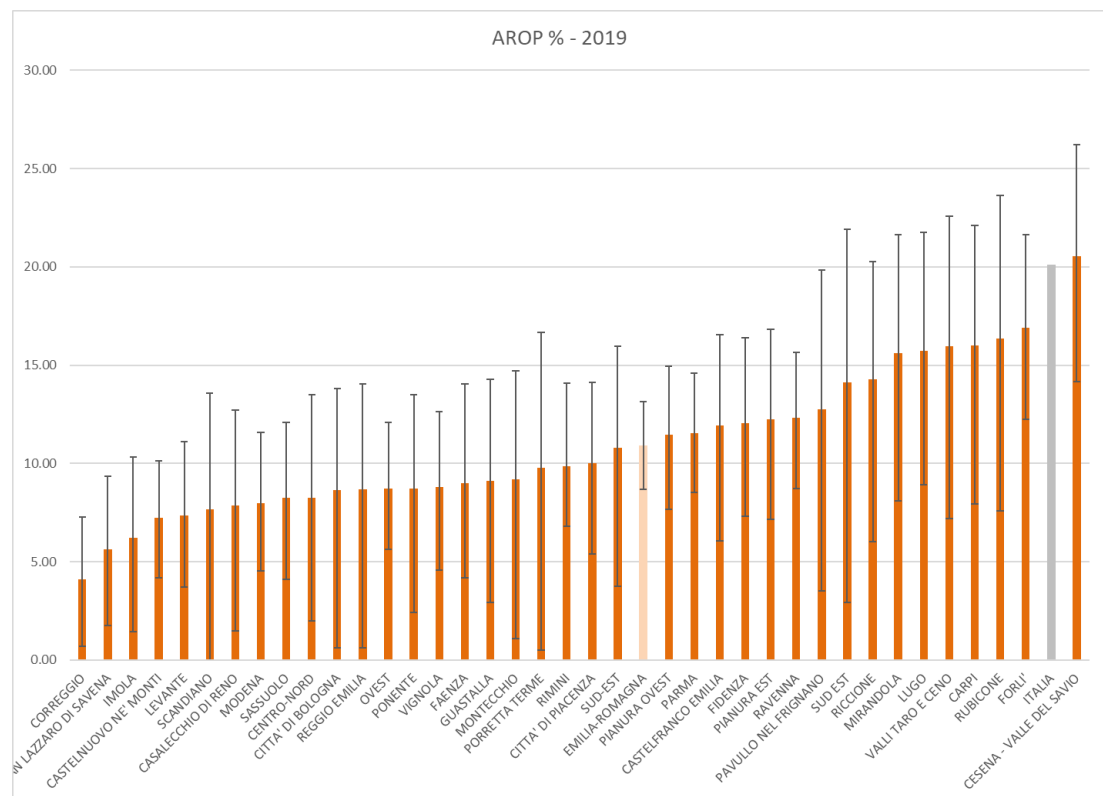
Tab. 4.1 - Coefficienti di variazione (CV %) delle stime da modello, per i distretti sociosanitari dell'Emilia-Romagna e in regione, anno 2019

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà – AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n ²⁰¹⁹
Bologna	CASALECCHIO DI RENO	29.11	36.66	50.12	49.57	35
Bologna	CITTA' DI BOLOGNA	32.11	37.67	47.68	43.96	0
Bologna	IMOLA	28.10	35.39	43.00	51.20	59
Bologna	PIANURA EST	16.17	19.71	27.08	23.04	208
Bologna	PIANURA OVEST	12.64	16.16	21.74	23.46	292
Bologna	PORRETTA TERME	31.46	42.38	65.86	58.73	33
Bologna	SAN LAZZARO DI SAVENA	27.69	33.84	42.38	42.76	123
Ferrara	CENTRO-NORD	29.22	35.32	47.62	37.25	0
Ferrara	OVEST	16.56	19.02	26.35	19.67	128
Ferrara	SUD-EST	26.52	29.52	45.41	33.65	0
Forlì	CESENA - VALLE DEL SAVIO	12.84	14.80	25.50	21.62	129
Forlì	FORLI'	11.92	13.85	28.94	25.10	79
Forlì	RUBICONE	21.66	24.42	44.36	44.01	0
Modena	CARPI	18.21	22.55	36.82	35.33	47
Modena	CASTELFRANCO EMILIA	18.94	22.64	40.35	33.46	116
Modena	MIRANDOLA	17.76	21.91	41.42	35.42	73
Modena	MODENA	16.02	22.34	28.60	39.74	197
Modena	PAVULLO NEL FRIGNANO	23.17	32.66	52.61	54.90	52
Modena	SASSUOLO	19.08	24.55	39.38	43.78	113
Modena	VIGNOLA	18.76	23.12	36.70	48.74	111
Parma	FIDENZA	16.37	19.13	41.56	27.92	108
Parma	PARMA	10.26	13.56	17.20	18.05	179
Parma	SUD EST	27.64	33.90	58.42	46.46	0
Parma	VALLI TARO E CENO	19.28	24.27	43.29	32.90	50
Piacenza	CITTA' DI PIACENZA	19.43	22.42	43.84	35.00	53
Piacenza	LEVANTE	21.93	25.37	46.16	36.09	135
Piacenza	PONENTE	28.85	32.37	40.67	39.60	0
Ravenna	FAENZA	23.21	27.54	55.36	44.79	101
Ravenna	LUGO	17.27	20.48	42.65	31.92	64
Ravenna	RAVENNA	11.44	14.11	29.77	19.76	114
Reggio	CASTELNUOVO NE' MONTI	16.05	20.73	32.71	36.69	174
Reggio	CORREGGIO	32.54	41.04	63.49	61.78	122
Reggio	GUASTALLA	25.71	32.35	64.18	59.46	60
Reggio	MONTECCHIO	34.41	37.50	47.41	51.52	0
Reggio	REGGIO EMILIA	35.65	39.06	51.64	53.56	0
Reggio	SCANDIANO	38.00	46.99	97.88	78.89	42
Rimini	RICCIONE	22.30	25.29	38.78	37.08	0
Rimini	RIMINI	15.61	18.70	34.92	32.71	180
	ER	8.83	10.35	23.82	15.18	3177

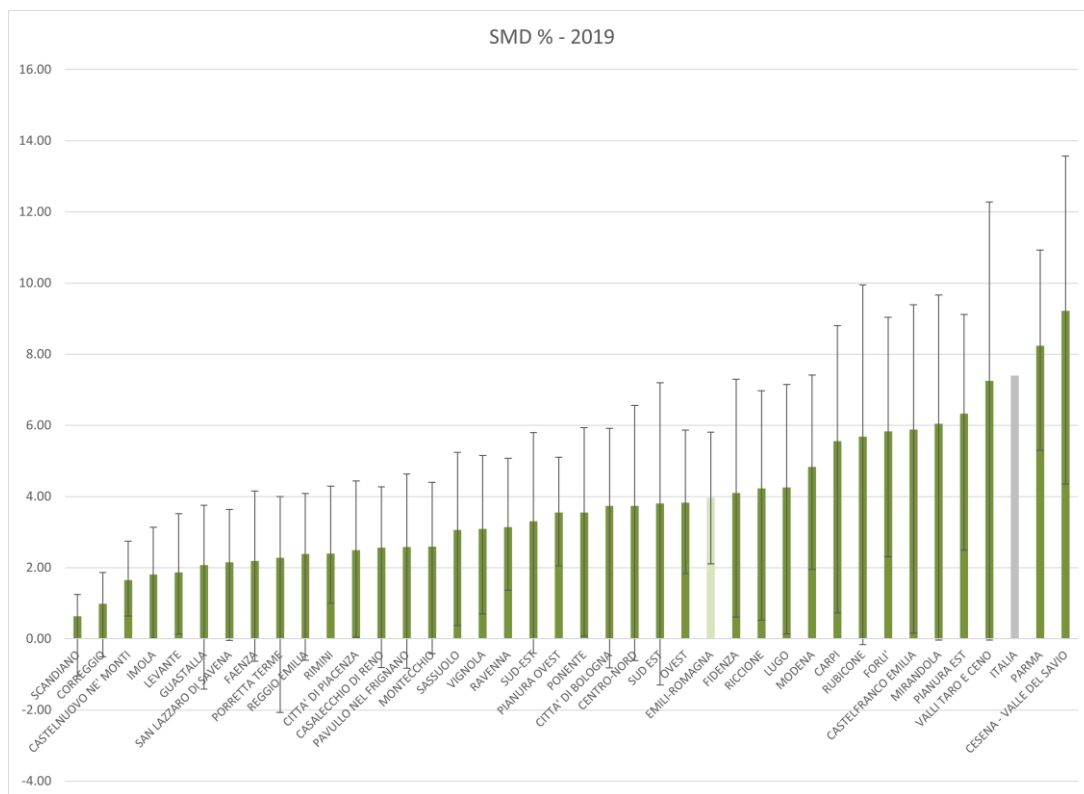
Gr. 9 — Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROPE) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



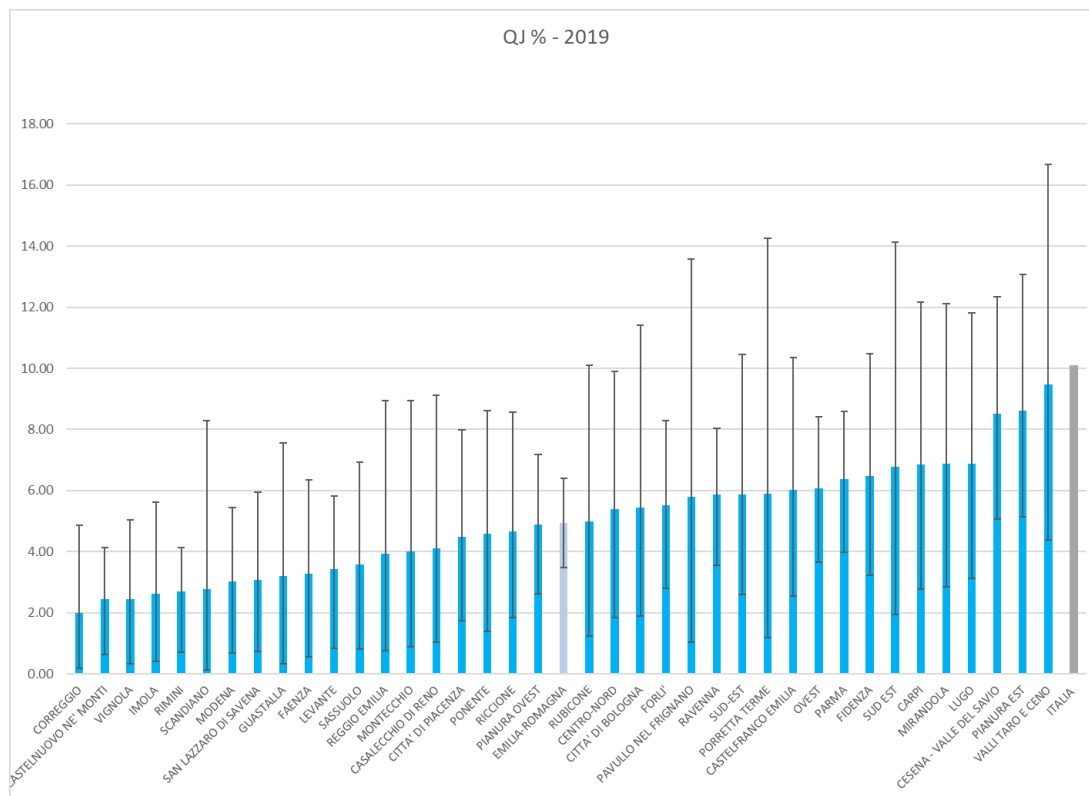
Gr. 10 - Stime da modello del Tasso di rischio povertà (AROP) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



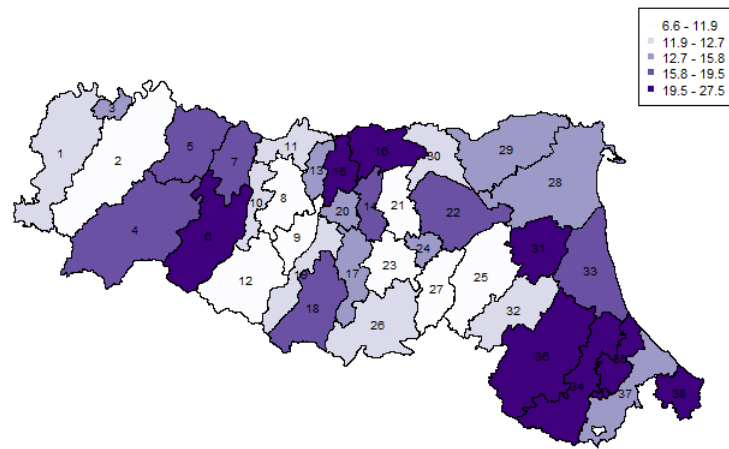
Gr. 11 - Stime da modello del Tasso di deprivazione materiale severa (SMD) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



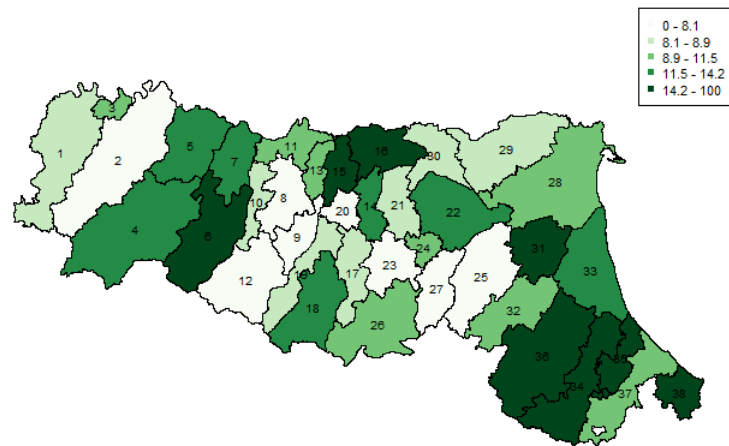
Gr. 12 - Stime da modello del Tasso di bassa intensità lavorativa per i distretti sociosanitari dell'Emilia-Romagna, anno 2019 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



AROE - 2019



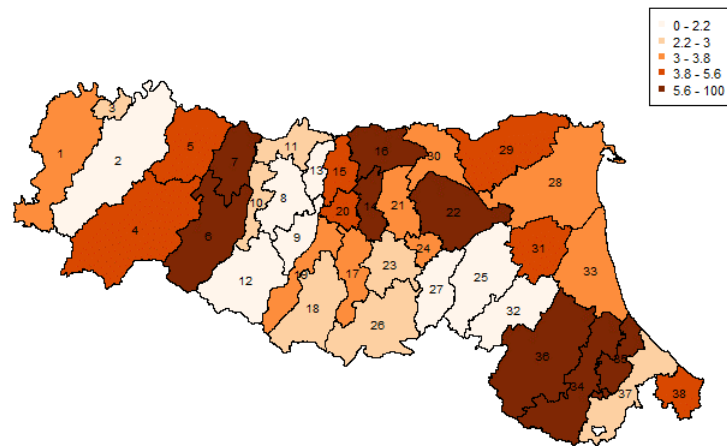
AROP - 2019



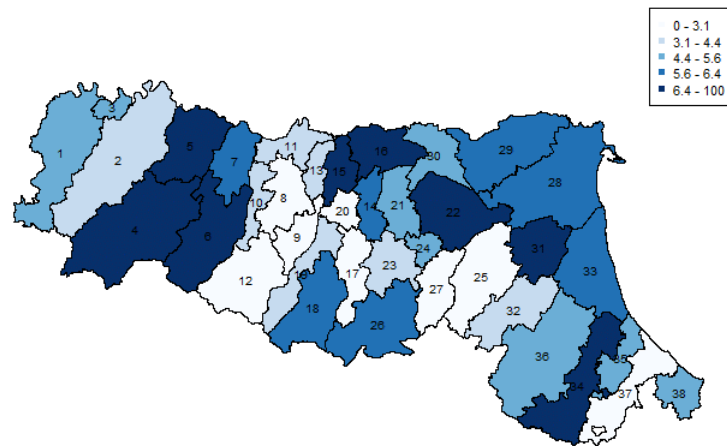
Legenda

1 Ponente, 2 Levante, 3 Citta Di Piacenza, 4, Valli Taro e Ceno, 5 Fidenza, 6 Sud Est, 7 Parma, 8 Reggio Emilia, 9 Scandiano, 10 Montecchio, 11 Guastalla, 12 Castelnuovo ne Monti, 13 Correggio, 14 Castelfranco Emilia, 15 Carpi, 16 Mirandola, 17 Vignola, 18 Pavullo Nel Frignano, 19 Sassuolo, 20 Modena, 21 Pianura Ovest, 22 Pianura Est, 23 Casalecchio di Reno, 24 Citta di Bologna, 25 Imola, 26 Porretta Terme, 27 San Lazzaro di Savena, 28 Sud-Est, 29 Centro-Nord, 30 Ovest, 31 Lugo, 32 Faenza, 33 Ravenna, 34 Cesena - Valle del Savio, 35 Rubicone, 36 Forlì, 37 Rimini, 38 Riccione

SMD - 2019



QJ - 2019



Legenda

1 Ponente, 2 Levante, 3 Citta Di Piacenza, 4, Valli Taro e Ceno, 5 Fidenza, 6 Sud Est, 7 Parma, 8 Reggio Emilia, 9 Scandiano, 10 Montecchio, 11 Guastalla, 12 Castelnuovo ne Monti, 13 Correggio, 14 Castelfranco Emilia, 15 Carpi, 16 Mirandola, 17 Vignola, 18 Pavullo Nel Frignano, 19 Sassuolo, 20 Modena, 21 Pianura Ovest, 22 Pianura Est, 23 Casalecchio di Reno, 24 Citta di Bologna, 25 Imola, 26 Porretta Terme, 27 San Lazzaro di Savena, 28 Sud-Est, 29 Centro-Nord, 30 Ovest, 31 Lugo, 32 Faenza, 33 Ravenna, 34 Cesena - Valle del Savio, 35 Rubicone, 36 Forlì, 37 Rimini, 38 Riccione

Nel seguito riportiamo le stime ottenute per il 2020 con riferimento alle province (Tab. 3.2, grafici 13-16 e relative mappe). Il set dei risultati relativi al 2020 viene presentato con la medesima struttura utilizzata per il 2019. Come già notato, le numerosità 2020 sono decisamente inferiori rispetto a quelle 2019. Come ci si attendeva, questa diminuzione si riflette, in generale, su un aumento della variabilità delle stime dirette e quindi delle stime da modello 2020 che risultano caratterizzate da una minore affidabilità rispetto a quelle 2019. Tale diminuzione, già notata per le province, è ancora più sentita per i distretti sociosanitari, che hanno una dimensione campionaria più piccola rispetto alle province stesse.

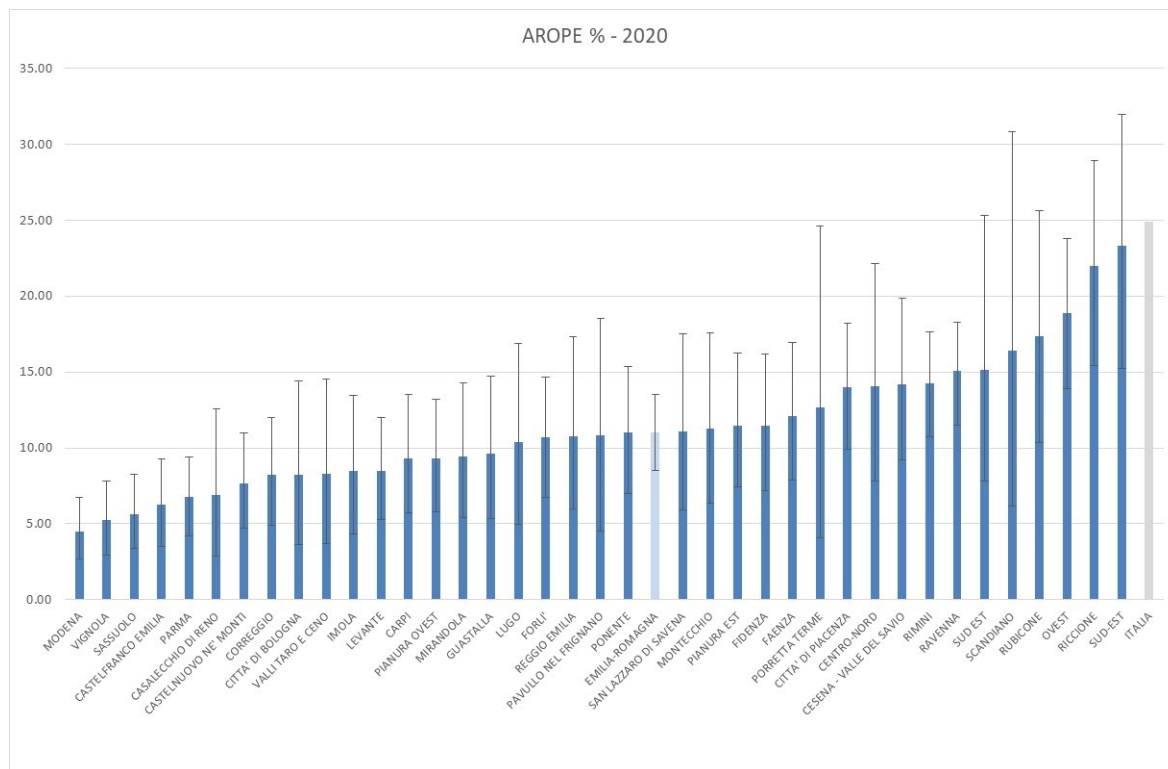
Tab. 3.2 - Stime da modello (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà - AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n ₂₀₁₉
Bologna	CASALECCHIO DI RENO	6.9	4.4	0.4	2.3	23
Bologna	CITTA' DI BOLOGNA	8.2	5.4	0.7	3.8	43
Bologna	IMOLA	8.5	5.6	0.6	3.6	45
Bologna	PIANURA EST	11.4	8.5	1.8	7.0	126
Bologna	PIANURA OVEST	9.3	8.9	0.8	3.5	188
Bologna	PORRETTA TERME	12.7	9.7	0.6	7.9	20
Bologna	SAN LAZZARO DI SAVENA	11.1	8.5	1.0	6.1	78
Ferrara	CENTRO-NORD	14.1	10.0	2.9	6.9	0
Ferrara	OVEST	18.9	15.9	1.6	7.2	64
Ferrara	SUD-EST	23.3	20.0	1.3	11.6	40
Forli	CESENA - VALLE DEL SAVIO	14.2	11.0	1.4	6.6	59
Forli	FORLI'	10.7	7.9	1.2	3.6	50
Forli	RUBICONE	17.4	13.6	2.8	8.4	0
Modena	CARPI	9.3	6.6	1.2	3.8	27
Modena	CASTELFRANCO EMILIA	6.2	4.0	0.9	2.4	64
Modena	MIRANDOLA	9.5	6.4	1.3	3.6	42
Modena	MODENA	4.5	3.3	1.2	1.3	131
Modena	PAVULLO NEL FRIGNANO	10.8	8.4	0.8	4.1	35
Modena	SASSUOLO	5.6	3.9	0.9	2.1	75
Modena	VIGNOLA	5.2	3.5	0.7	1.1	84
Parma	FIDENZA	11.5	8.0	1.4	5.1	63
Parma	PARMA	6.8	5.0	1.2	2.4	131
Parma	SUD EST	15.1	12.8	2.2	6.0	0
Parma	VALLI TARO E CENO	8.3	5.0	0.6	2.3	26
Piacenza	CITTA' DI PIACENZA	14.0	10.5	2.4	5.0	51
Piacenza	LEVANTE	8.5	6.2	1.1	2.3	76
Piacenza	PONENTE	11.0	7.8	1.8	4.3	69
Ravenna	FAENZA	12.1	9.0	1.9	4.9	116
Ravenna	LUGO	10.4	6.8	0.8	3.3	22
Ravenna	RAVENNA	15.1	12.4	1.8	6.6	47
Reggio	CASTELNUOVO NE' MONTI	7.7	5.7	0.9	2.5	84
Reggio	CORREGGIO	8.2	5.3	0.9	3.9	84
Reggio	GUASTALLA	9.6	6.1	1.3	3.7	57
Reggio	MONTECCHIO	11.3	8.0	2.1	4.9	0
Reggio	REGGIO EMILIA	10.8	7.5	2.1	5.2	0
Reggio	SCANDIANO	16.4	13.5	0.7	7.8	19
Rimini	RICCIONE	22.0	17.9	2.1	10.4	53
Rimini	RIMINI	14.3	11.4	2.3	5.1	71
	ER	11.0	8.48	1.34	4.5	2163
	Italia	24.9	20.0	3.5	11.5	29252

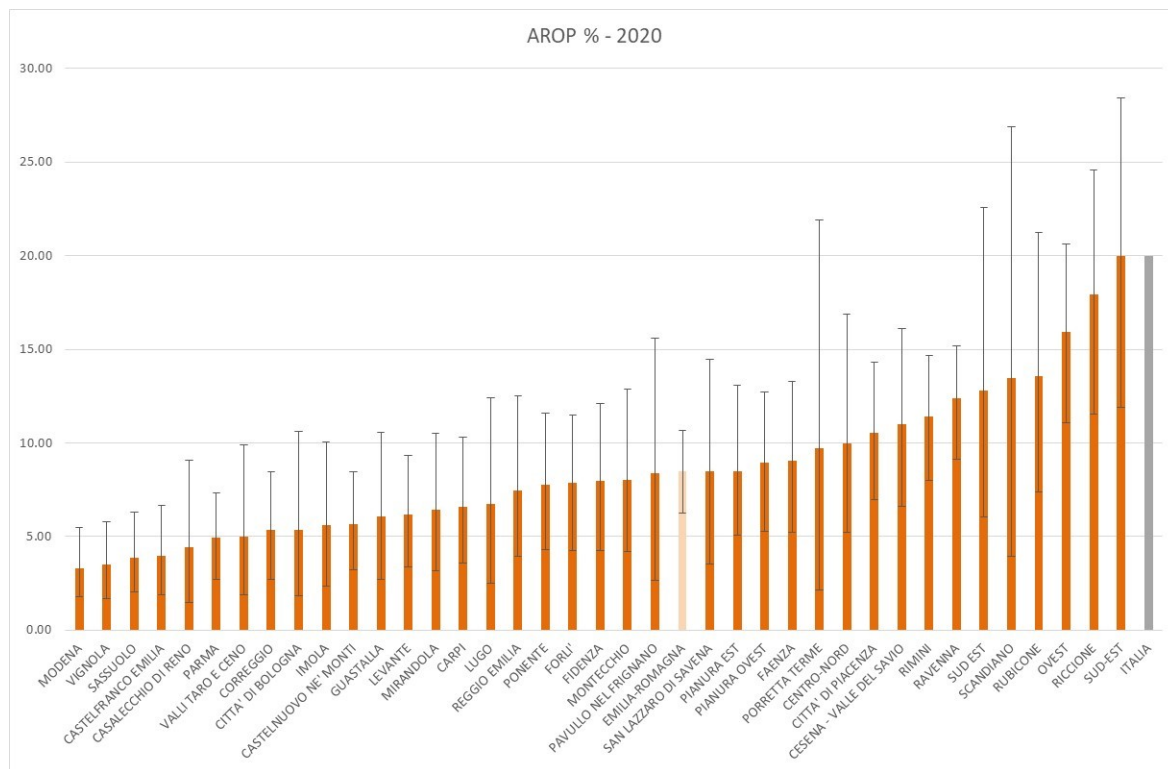
Tab. 4.2 - Coefficienti di variazione (CV %) delle stime da modello, per i distretti sociosanitari dell'Emilia-Romagna e in regione, anno 2020

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà – AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n ²⁰¹⁹
Bologna	CASALECCHIO DI RENO	36.4	45.0	75.1	68.3	23
Bologna	CITTA' DI BOLOGNA	33.7	43.8	76.4	56.6	43
Bologna	IMOLA	27.5	35.7	60.6	51.0	45
Bologna	PIANURA EST	19.8	24.0	35.3	29.7	126
Bologna	PIANURA OVEST	20.6	21.4	44.8	36.2	188
Bologna	PORRETTA TERME	42.8	53.4	125.9	62.8	20
Bologna	SAN LAZZARO DI SAVENA	26.9	33.1	67.6	42.6	78
Ferrara	CENTRO-NORD	26.4	29.8	31.5	44.0	0
Ferrara	OVEST	13.5	15.1	30.0	29.6	64
Ferrara	SUD-EST	18.1	20.8	56.6	31.9	40
Forlì	CESENA - VALLE DEL SAVIO	18.9	22.2	42.4	33.4	59
Forlì	FORLI'	18.6	23.2	38.2	40.5	50
Forlì	RUBICONE	22.7	26.0	31.4	35.8	0
Modena	CARPI	21.2	26.3	49.0	38.8	27
Modena	CASTELFRANCO EMILIA	23.9	30.9	54.2	42.7	64
Modena	MIRANDOLA	23.8	29.0	59.3	43.2	42
Modena	MODENA	23.3	28.2	42.6	46.3	131
Modena	PAVULLO NEL FRIGNANO	33.8	40.0	106.7	61.3	35
Modena	SASSUOLO	22.3	28.1	48.7	44.2	75
Modena	VIGNOLA	24.3	30.3	51.3	56.5	84
Parma	FIDENZA	20.2	24.7	42.3	36.3	63
Parma	PARMA	19.8	23.3	28.9	38.1	131
Parma	SUD EST	29.9	33.9	49.8	52.6	0
Parma	VALLI TARO E CENO	33.4	41.0	67.2	65.5	26
Piacenza	CITTA' DI PIACENZA	14.9	17.5	27.5	29.8	51
Piacenza	LEVANTE	20.3	24.3	43.3	48.7	76
Piacenza	PONENTE	19.2	23.7	39.1	38.1	69
Ravenna	FAENZA	19.0	22.6	38.2	37.9	116
Ravenna	LUGO	29.4	38.0	62.5	63.1	22
Ravenna	RAVENNA	11.6	12.7	23.4	20.9	47
Reggio	CASTELNUOVO NE' MONTI	20.8	23.9	41.0	39.9	84
Reggio	CORREGGIO	22.2	27.7	49.0	42.5	84
Reggio	GUASTALLA	25.3	32.7	54.8	48.4	57
Reggio	MONTECCHIO	25.5	28.1	32.1	40.1	0
Reggio	REGGIO EMILIA	26.9	29.5	33.7	41.3	0
Reggio	SCANDIANO	38.2	44.7	116.8	65.8	19
Rimini	RICCIONE	15.8	18.5	48.7	27.7	53
Rimini	RIMINI	12.4	14.8	22.9	28.6	71
	ER	11.6	13.3	31.1	17.2	2163

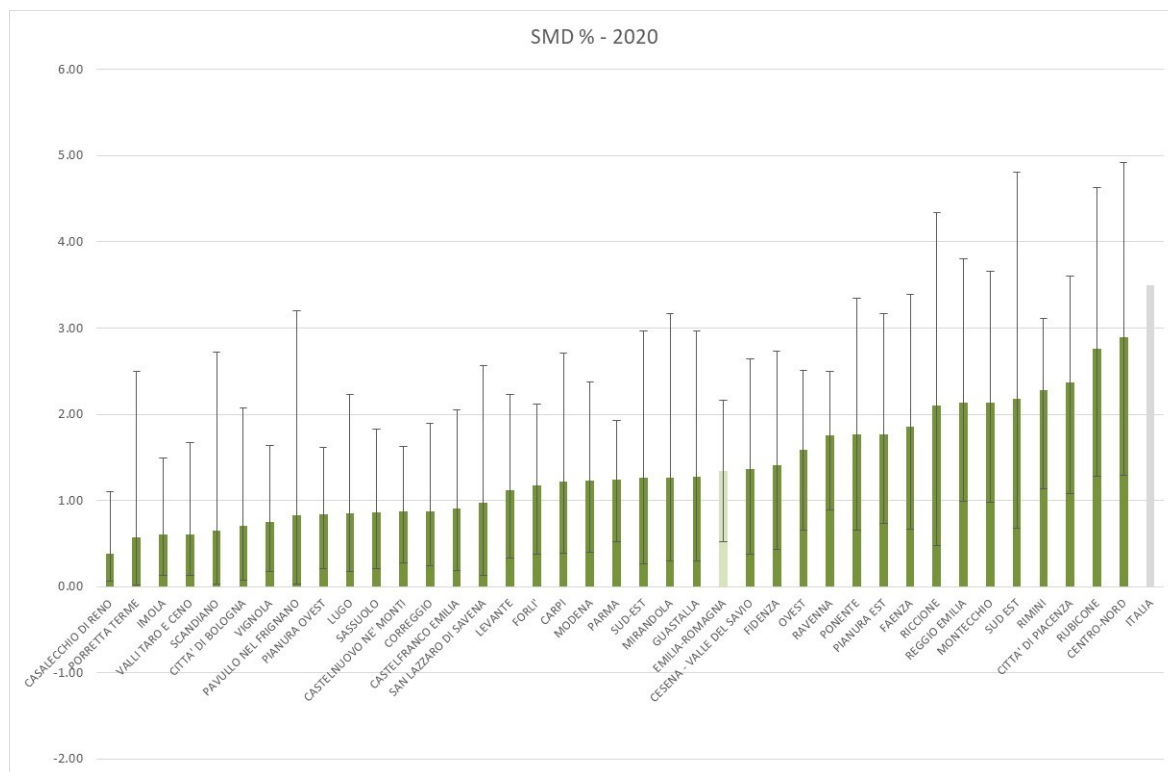
Gr. 13 — Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROEPE) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



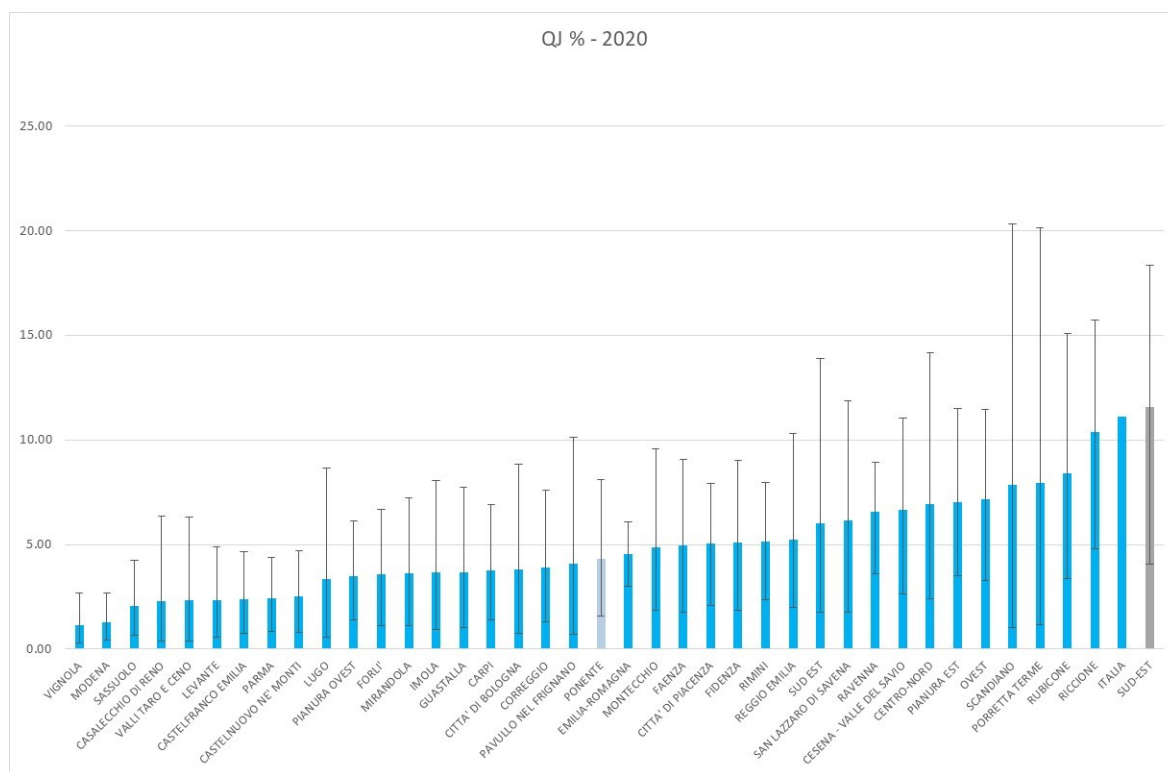
Gr. 14 - Stime da modello del Tasso di rischio povertà (AROP) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



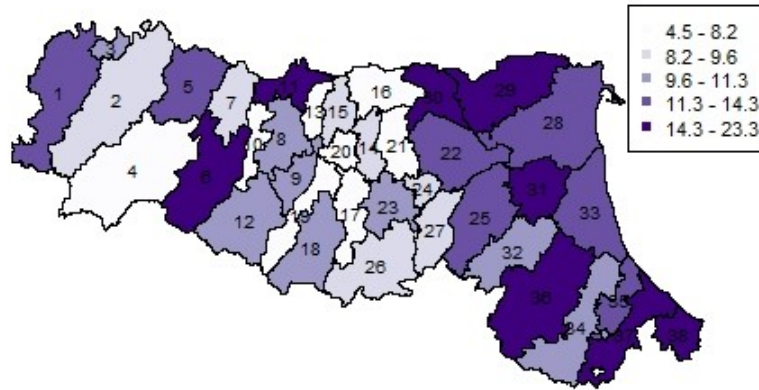
Gr. 15 - Stime da modello del Tasso di deprivazione materiale severa (SMD) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



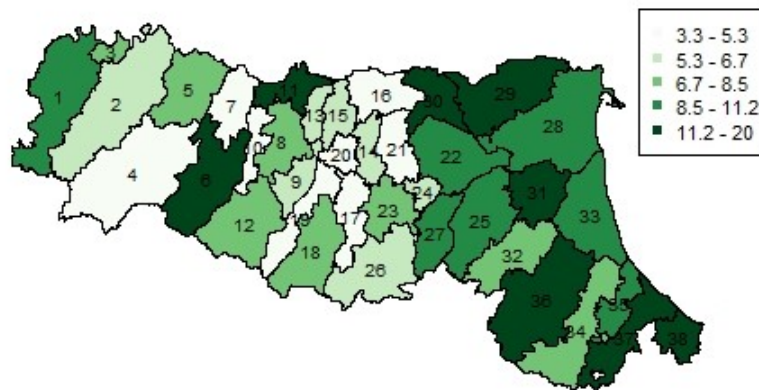
Gr. 16 - Stime da modello del Tasso di bassa intensità lavorativa per i distretti sociosanitari dell'Emilia-Romagna, anno 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



AROPE – 2020



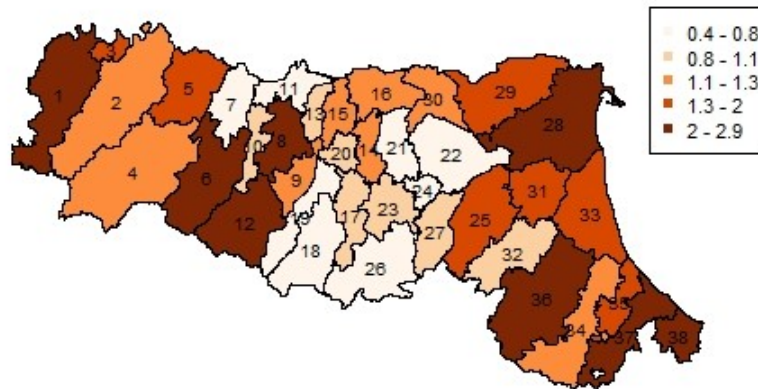
AROP - 2020



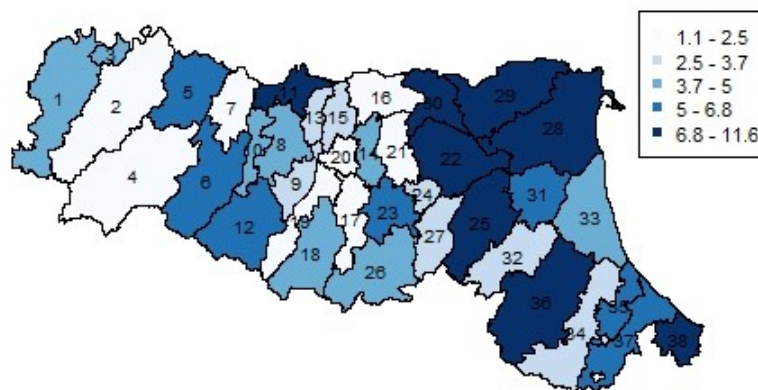
Legenda

1 Ponente, 2 Levante, 3 Citta Di Piacenza, 4, Valli Taro e Ceno, 5 Fidenza, 6 Sud Est, 7 Parma, 8 Reggio Emilia, 9 Scandiano, 10 Montecchio, 11 Guastalla, 12 Castelnuovo ne Monti, 13 Correggio, 14 Castelfranco Emilia, 15 Carpi, 16 Mirandola, 17 Vignola, 18 Pavullo Nel Frignano, 19 Sassuolo, 20 Modena, 21 Pianura Ovest, 22 Pianura Est, 23 Casalecchio di Reno, 24 Citta di Bologna, 25 Imola, 26 Porretta Terme, 27 San Lazzaro di Savena, 28 Sud-Est, 29 Centro-Nord, 30 Ovest, 31 Lugo, 32 Faenza, 33 Ravenna, 34 Cesena - Valle del Savio, 35 Rubicone, 36 Forlì, 37 Rimini, 38 Riccione

SMD – 2020



QJ – 2020

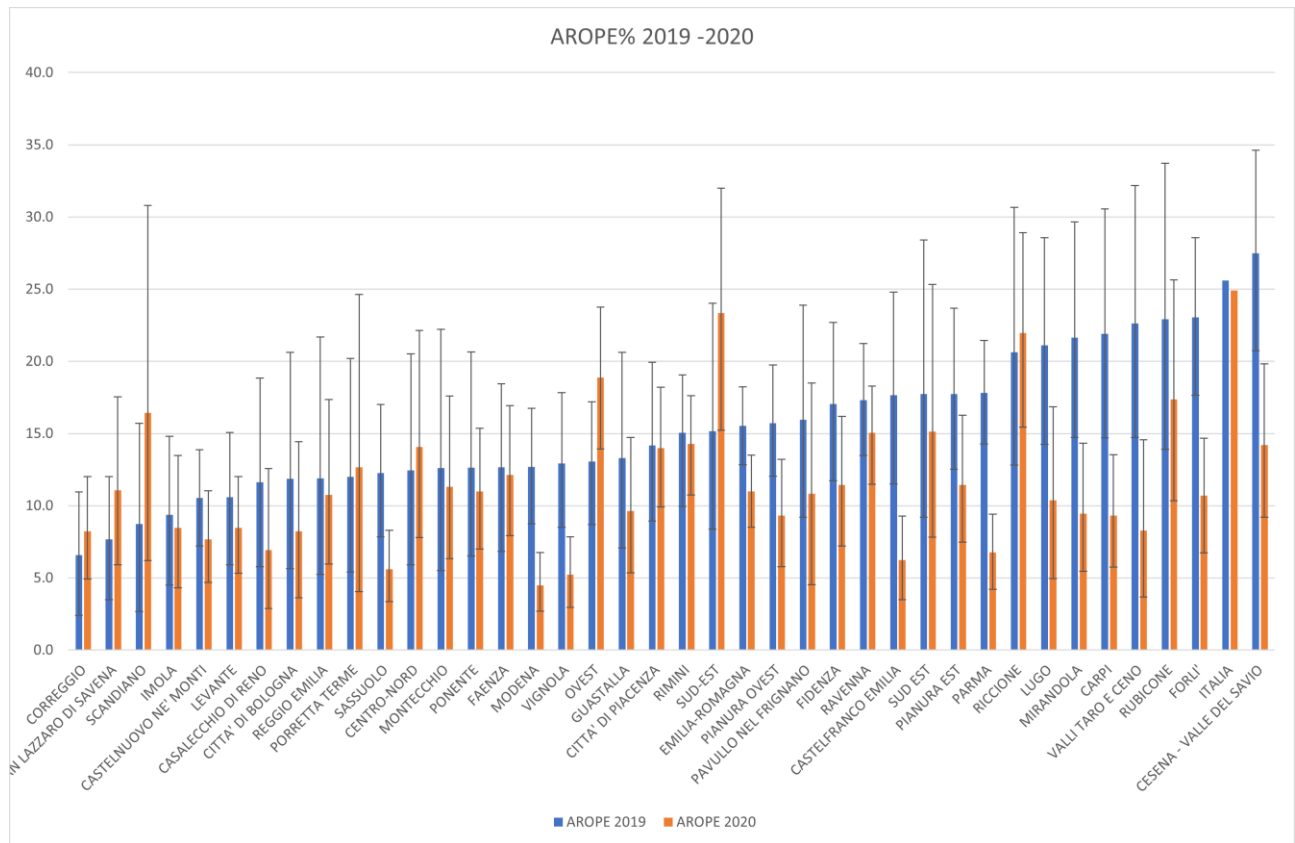


Legenda

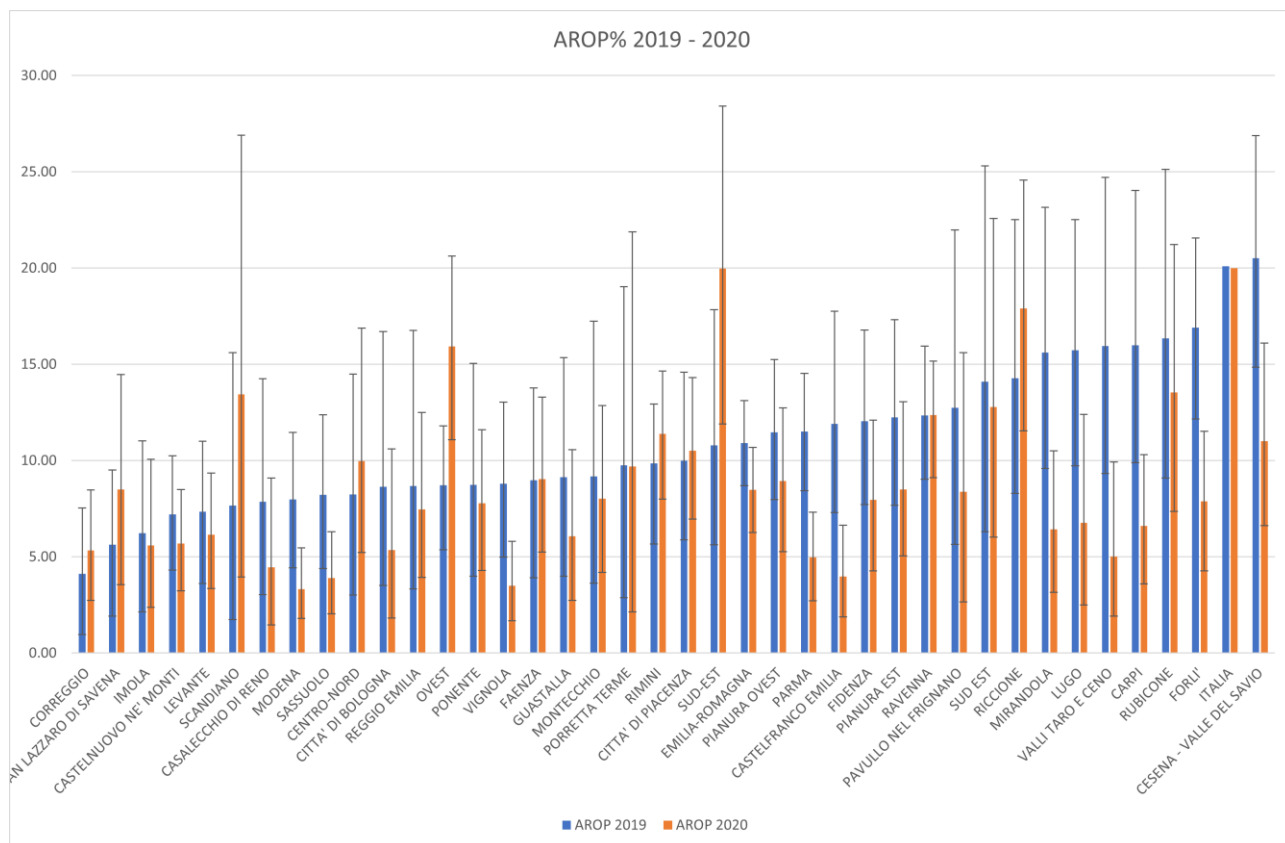
1 Ponente, 2 Levante, 3 Citta Di Piacenza, 4, Valli Taro e Ceno, 5 Fidenza, 6 Sud Est, 7 Parma, 8 Reggio Emilia, 9 Scandiano, 10 Montecchio, 11 Guastalla, 12 Castelnuovo ne Monti, 13 Correggio, 14 Castelfranco Emilia, 15 Carpi, 16 Mirandola, 17 Vignola, 18 Pavullo Nel Frignano, 19 Sassuolo, 20 Modena, 21 Pianura Ovest, 22 Pianura Est, 23 Casalecchio di Reno, 24 Citta di Bologna, 25 Imola, 26 Porretta Terme, 27 San Lazzaro di Savena, 28 Sud-Est, 29 Centro-Nord, 30 Ovest, 31 Lugo, 32 Faenza, 33 Ravenna, 34 Cesena - Valle del Savio, 35 Rubicone, 36 Forlì, 37 Rimini, 38 Riccione

Nel seguito si confrontano le stime dal modello ottenute per i distretti dell'Emilia-Romagna per il 2019 e 2020 (Gr. 13.1-16.1). Tali confronti vanno effettuati con qualche cautela in quanto le stime dei due anni non sono ottenute sotto le stesse condizioni in relazione alle numerosità campionarie molto diverse.

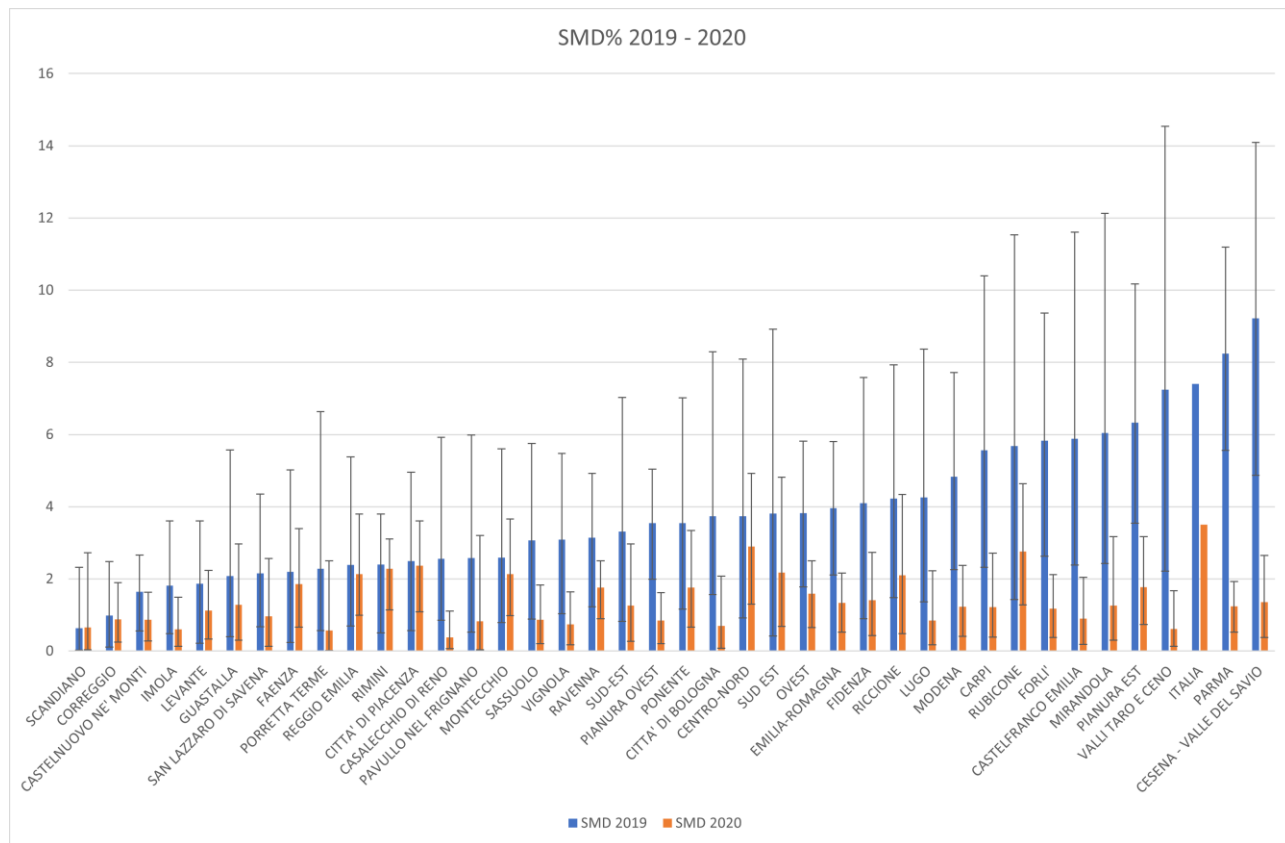
Gr. 13.1 -- Stime da modello del Tasso di rischio povertà ed esclusione sociale (AROPE) per i distretti dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



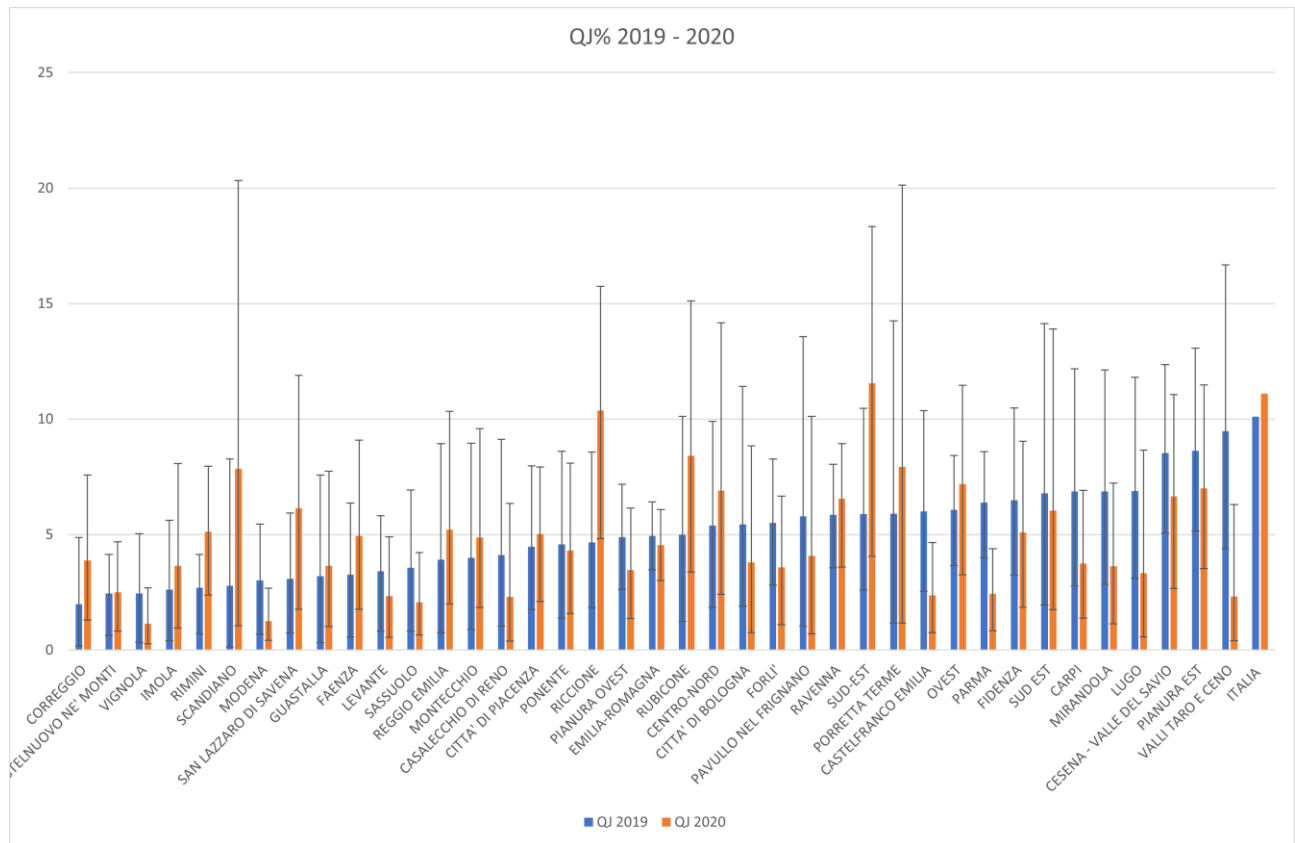
Gr. 14.1 -- Stime da modello del Tasso di rischio povertà (AROP) per i distretti dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Gr. 15.1 -- Stime da modello del Tasso di deprivazione materiale severa (SMD) per i distretti dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Gr. 16.1 -- Stime da modello del Tasso di bassa intensità lavorativa (QJ) per i distretti dell'Emilia-Romagna, anno 2019 e 2020 (i segmenti sulle barre rappresentano gli intervalli di confidenza)



Appendice A

L'indagine EU-SILC, la produzione delle stime dirette e della loro varianza

A1. Indagine EU-SILC. Caratteristiche generali

L'indagine EU-SILC (European Statistics on Income and Living Conditions) è stata istituita con il regolamento dell'Unione Europea n. 1177/2003. In Italia viene regolarmente condotta, con cadenza annuale, a partire dal 2004.

L'indagine sui redditi e le condizioni di vita è finalizzata alla produzione di stime annuali su queste tematiche in modo armonizzato per tutti i paesi dell'Unione. Più in dettaglio è pensata per produrre due tipologie di stime annuali: cross-section e longitudinali. Le stime cross-section forniscono la base per confronti coerenti tra le condizioni di vita, l'incidenza e la gravità di povertà ed esclusione sociale tra i paesi membri. Le stime longitudinali, rese possibili dalla struttura del piano di campionamento (gli individui campionati sono "seguiti" per quattro anni) sono finalizzate a rendere possibile l'analisi della dinamica micro-economica di questi indicatori.

La rilevazione del reddito monetario è uno degli obiettivi principali dell'indagine EU-SILC. Il reddito è il mezzo principale attraverso cui gli individui acquistano beni e servizi sul mercato ed è quindi naturalmente impiegabile come *proxy* delle risorse che gli individui e le famiglie hanno a disposizione per soddisfare i propri bisogni. L'indagine EU-SILC costituisce, in questo senso, la base informativa, attraverso cui costruire, in modo omogeneo per tutti i paesi e le regioni UE, indicatori di benessere e povertà basati sul reddito disponibile. Tuttavia, l'indagine non è limitata alla sola misurazione del reddito, ma include anche numerose domande sulle condizioni di salute, economiche, nonché di una molteplicità di variabili che misurano le condizioni materiali di vita delle famiglie e che permettono di ricostruire un quadro della diffusione del benessere e dell'incidenza della povertà che prescindono da una misura quantitativa di capacità come è il reddito. Larga parte della comunità scientifica vede infatti come riduttivi gli studi sulla povertà e l'esclusione sociale che si limitino a considerare il solo aspetto monetario di quest'ultima. Un'analisi multidimensionale della povertà richiede che vengano misurate le *capabilities* (possibilità che gli individui hanno di fare, essere o avere ciò che desiderano) e i *funzionamenti* (ossia i risultati effettivamente ottenuti). Il questionario EU-SILC è orientato alla misurazione dei mancati funzionamenti, quali indicatori di un effettivo stato di povertà materiale. La misurazione del reddito è complessa ed errori di calcolo, nonché dichiarazioni per difetto da parte dei rispondenti, sia involontarie, sia deliberate rappresentano una possibile fonte di distorsione difficile da quantificare. Per questa ragione in un'analisi più esaustiva della povertà è bene affiancare stime monetarie e stime non monetarie della povertà.

L'indagine EU-SILC viene condotta in ogni paese dall'Istituto Nazionale di Statistica, sotto il coordinamento di Eurostat. La parte italiana di EU-SILC, a cui faremo riferimento come IT-SILC, è quindi sotto la responsabilità dell'ISTAT. L'indagine viene condotta attraverso interviste personali

A1.1 Popolazione obiettivo

La popolazione obiettivo dell'indagine IT-SILC è costituita da tutti i componenti delle famiglie residenti in Italia, ancorché temporaneamente all'estero. Sono esclusi i membri di convivenze istituzionali e quelli di famiglie residenti all'estero, ancorché temporaneamente in Italia.

Per famiglia si intende la famiglia di fatto, in accordo con la definizione adottata da ISTAT in tutte le altre rilevazioni: un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti o aventi dimora abituale nello stesso comune.

Tutti i membri delle famiglie campionate di età superiore a 15 anni sono intervistati.

Una volta che un individuo entra a far parte del campione EU-SILC vi rimane per successivi 3 anni (per un totale di 4 rilevazioni), sia che rimanga all'interno della famiglia in cui si trovava al tempo della prima rilevazione, sia che vada a far parte di una nuova. Di conseguenza, il campione annuale IT-SILC è rinnovato ogni anno per un quarto.

A1.2 Periodo di rilevazione e periodo di riferimento

La rilevazione dei dati avviene dal mese di giugno al mese di novembre di ogni anno. La scelta di questo periodo è dettata dalla volontà di permettere alle famiglie di utilizzare nelle risposte le informazioni contenute nella loro dichiarazione annuale dei redditi.

Le domande relative al reddito fanno riferimento al reddito percepito nell'anno precedente a quello dell'intervista. Possiamo quindi distinguere anno di rilevazione e l'anno di riferimento del reddito. All'anno dell'intervista fanno riferimento le domande relative alle condizioni di vita della famiglia.

A1.3 Piano di campionamento

Il campione annuale EU-SILC per l'Italia è composto dall'unione di 4 campioni longitudinali. A regime, ogni anno viene rinnovato un quarto del campione; le unità selezionate parteciperanno a quattro successive ondate dell'indagine per poi uscire dal campione. Ovviamente per la prima ondata, nel 2004, l'intero campione è stato selezionato direttamente dalla popolazione. Nel 2005 è iniziato il rinnovamento ad un quarto per volta.

Complessivamente, il piano di campionamento è quindi a panel ruotato, secondo questo schema in cui ciascuna le lettere maiuscole indicano i campioni longitudinali:

	2008	2009	2010	2011	2012	2013	2014
I edizione	ABCD						
II edizione		BCDE					
III edizione			CDEF				
IV edizione				DEFG			
V edizione					EFGH		
VI edizione						FGHI	
VII edizione							GHIL

Per garantire stime sufficientemente precise per tutti i domini pianificati dell'indagine (di cui si dirà) e per ottemperare ai requisiti di numerosità minima stabiliti da Eurostat, ogni campione longitudinale di EU-SILC è composto di 8,000 famiglie.

La numerosità effettiva del campione trasversale per un dato anno t è quindi costituita:

- dalla frazione delle 8,000 famiglie nuove campionate che hanno aderito all'indagine
- alle famiglie provenienti dai campioni longitudinali entrati negli anni $t-1$, $t-2$, $t-3$, al netto ovviamente di quelle cadute per attrito.

Come in tutte le indagini con una componente panel, l'indagine EU-SILC prevede regole di inseguimento delle unità di rilevazione (follow-up rules) che da un'occasione di rilevazione alla successiva mutano di residenza o formano una nuova famiglia. Le regole generali di inseguimento sono state fissate a livello europeo e sono quindi simili per tutti i paesi. In linea di principio, tutte le unità intervistate (membri delle famiglie campionate di età maggiore o uguale a 15) vengono inseguite se formano un nuovo nucleo e/o si spostano all'interno del territorio nazionale nel periodo di vita del panel (le quattro rilevazioni successive). La nuova famiglia entra a far parte del campione e tutti i suoi componenti di età adeguata sono intervistati. Non sono inseguite le unità che si trasferiscono all'estero, se non in casi particolari, ovvero quando il trasferimento è solamente temporaneo, ossia di meno di sei mesi di durata prevista (Eurostat, 2006).

Il campione longitudinale delle famiglie “nuove” (l’intero campione trasversale per la prima edizione del 2004) è selezionato secondo un piano di campionamento stratificato a due stadi, sostanzialmente analogo a quello adottato da ISTAT per altre grandi indagini di argomento sociale ed economico.

Le unità di primo stadio sono costituite dai comuni, stratificati per regione e dimensione demografica. In particolare, all’interno di ciascuna regione i comuni sono divisi in autorappresentativi e non autorappresentativi in ragione della loro dimensione demografica. I comuni autorappresentativi vengono sempre inclusi nel campione mentre quelli non autorappresentativi vengono campionati in modo casuale stratificato con probabilità proporzionale alla dimensione demografica. La soglia che separa i comuni autorappresentativi dagli altri è fissata in modo diverso da regione a regione con l’obiettivo di garantire l’autoponderazione del campione a livello regionale. Complessivamente il numero degli strati è pari a 288 (per maggiori dettagli, si veda ISTAT, 2008)

Le unità di secondo stadio, le famiglie, sono campionate in modo sistematico dalle liste anagrafiche.

Le elaborazioni contenute in questo rapporto sono basate sui dati dell’edizione 2008-2013 dell’indagine. Per quest’ultima la numerosità effettiva del campione trasversale (ossia al netto di non risposta ed attrito) è di 18487 famiglie per un totale di 44622 individui.

Segnaliamo che la dimensione effettiva del campione EU-SILC per l’Italia si è andata progressivamente riducendo nel corso del tempo, per una molteplicità di cause che includono un attrito (ossia perdita di osservazioni tra occasioni successive nella componente panel) abbastanza severo. Nella wave 2008 – la più remota considerata in questo studio - il numero complessivo di famiglie rispondenti fu di 20928 per un totale di 52433 intervistati. Si è dunque registrata una contrazione del 11.6% in termini di famiglie e del 14.9% in termini di individui nell’arco del periodo considerato. Per la regione Emilia-Romagna la contrazione è stata più contenuta rispetto alla media nazionale: -7.9% per le famiglie, -8.8% per gli individui intervistati.

A1. 4 Pesì campionari

Nel file dei dati relativi ad EU-SILC rilasciato da ISTAT sono pubblicati i coefficienti di riporto all’universo (pesi campionari) sia per le componenti longitudinali che per il campione trasversale nel suo complesso. Poiché in questo lavoro ci interessiamo solamente di stime *cross-section* faremo sempre riferimento a questi ultimi.

Data una variabile obiettivo y per cui è stato osservato un campione di valori (y_1, \dots, y_n) i pesi campionari pubblicati in corrispondenza delle unità rispondenti (w_1, \dots, w_n)

permettono di ottenere una stima approssimativamente corretta e consistente del totale di popolazione applicando la seguente formula:

$$\hat{Y} = \sum_{i=1}^n w_i y_i$$

che per la media diventa

$$\hat{\bar{Y}} = N^{-1} \sum_{i=1}^n w_i y_i$$

dove $N = \sum_{i=1}^n w_i$.

Intuitivamente i pesi campionari sono dati dall'inverso delle probabilità di inclusione, in modo tale che se tutte le unità della popolazione avessero la stessa probabilità di essere incluse nel campione, allora i pesi sarebbero costanti e pari all'inverso del tasso di campionamento.

In pratica la loro costruzione è più complessa e il loro contenuto informativo è maggiore. In primo luogo, attraverso i pesi campionari si tenta di neutralizzare il più possibile l'effetto distorsivo della *non-risposta*. In secondo luogo, al fine di ottenere stime più efficienti si calibrano i pesi in modo tale che le stime del totale, per variabili il cui totale a livello di popolazione è noto, riproducano esattamente questo valore, secondo il metodo della "calibrazione".

I pesi campionari inclusi nei file IT-SILC rilasciati da ISTAT raggiungono questi obiettivi attraverso una tecnica di riponderazione in due passi. Nel primo passo utilizzando un criterio noto come "criterio delle celle di riponderazione" sono stati corretti, almeno parzialmente, gli effetti della mancata risposta totale. Il secondo passo costituisce un passo di calibrazione vero e proprio.

Nel caso di medie e totali, gli stimatori per calibrazione rappresentano la scelta *standard* di ISTAT che li applica in tutte le grandi indagini di argomento economico e sociale. L'illustrazione dei fondamenti teorici di questi stimatori eccede gli obiettivi di questo rapporto e pertanto si rimanda il lettore alla letteratura sul tema (in primis a Deville e Särndal, 1992 e a Särndal e Lundström, 2005).

Per quanto riguarda la correzione degli effetti della non risposta, alla base del metodo adottato vi sono alcune valutazioni di natura empirica sulle caratteristiche dei non rispondenti. È stato osservato infatti che "la difficoltà ad ottenere l'intervista è crescente al crescere dell'ampiezza demografica del comune di residenza e al diminuire del numero di componenti della famiglia perché spesso non si riesce proprio ad effettuare il contatto; le famiglie con persona di riferimento straniera hanno minori probabilità di partecipare all'indagine, in parte per l'alta mobilità degli stranieri sul territorio, in parte per le difficoltà o la diffidenza che possono avere persone di lingua e cultura differenti rispetto agli intervistatori. Si riscontrano inoltre notevoli differenze nella partecipazione all'indagine a seconda della "regione di residenza" (Ceccarelli e

Cutillo, 2006). Sulla base di queste considerazioni, la popolazione delle famiglie residenti è stata classificata sulla base di questi criteri:

1. dimensione demografica del comune (5 modalità);
2. cittadinanza della persona di riferimento (2 modalità);
3. regione di residenza (21 modalità essendo le province autonome di Trento e Bolzano considerate separatamente);
4. numero di componenti (4 modalità)

Per ciascuna delle celle è stato valutato il numero di rispondenti effettivo m_z^r e il numero di rispondenti teorico, m_z ($z=1,\dots,Z$), sulla base del campione estratto. La correzione della non risposta avviene, in linea di principio, moltiplicando il “peso base” w_i^{base} , definito come l’inverso della probabilità di inclusione (π_i), per il fattore $\eta_z = \frac{m_z}{m_z^r}$. In pratica quando la numerosità m_z^r in alcune celle è al di sotto di una certa soglia, si procede ad un collasso mento delle celle stesse per ottenere correzioni numericamente più stabili.

Questa procedura di riponderazione finalizzata alla correzione degli effetti della non-risposta si basa sull’ipotesi che le famiglie all’interno di ciascuna cella abbiano una probabilità di risposta uniforme, ovvero che i non-rispondenti formino, all’interno di ciascuna cella, un sotto-campione casuale. Benché la tecnica delle celle di riponderazione sia stata sviluppata indipendentemente dalla teoria degli stimatori per calibrazioni, può essere ricondotta a quest’ultima, nel caso in cui i totali a cui si calibra siano quelli campionari (si vedano a questo riguardo Särndal e Lundström (2005)).

Riassumendo la procedura di correzione della non risposta porta alla produzione di un

set di “pesi intermedi”: $w_i^{int} = \frac{1}{\pi_i} \frac{m_z}{m_z^r}$, $i \in s$

dove s indica l’intero campione trasversale.

Riguardo alla calibrazione a totali noti, ognuno dei quattro campioni longitudinali è vincolato alla:

1. popolazione residente per ripartizione territoriale, sesso e classe d’età al 31/12 dell’anno di riferimento del reddito.
2. il numero di famiglie residenti per regione riferite alla fine dell’anno di rilevazione

Inoltre, i pesi relativi all'intero campione trasversale, che sono poi quelli che ci interessano, sono vincolati a rispettare i seguenti totali:

1. popolazione residente per regione, sesso e classi d'età alla fine dell'anno di riferimento del reddito;
2. popolazione residente per regione, sesso e classi d'età alla fine dell'anno di rilevazione;
3. popolazione residente straniera maggiorenne per ripartizione geografica e sesso alla fine dell'anno di riferimento del reddito;
4. popolazione residente straniera maggiorenne per ripartizione geografica e provenienza (UE e non UE) alla fine dell'anno di riferimento del reddito

Le stime dirette ottenute applicando i pesi pubblicati da ISTAT insieme ai dati di EU-SILC per l'Italia presenta aspetti problematici. I domini rappresentati dai distretti non sono infatti "domini pianificati" e sommando i pesi relativi alle unità del campione incluse nel dominio non si ottiene la dimensione esatta della popolazione (noto attraverso le anagrafi); inoltre se si considerano sotto-campioni classificando per sesso, età o altre variabili, il loro peso relativi in termini di somma dei pesi può essere diverso da quello reale nella popolazione del distretto. La calibrazione è infatti effettuata a livello regionale, il più dettagliato per cui ISTAT pubblica ufficialmente le stime.

Al fine di ottenere un sistema di pesi che, all'interno di ciascun dominio riporti alla effettiva dimensione della popolazione, il processo di calibrazione dei pesi può essere modificato introducendo informazioni ausiliarie aggiuntive e facendo in modo che la somma dei pesi all'interno di ciascun dominio sommi alla dimensione della sua popolazione così come risulta dalle anagrafi comunali. Nella stessa direzione si può richiedere che i pesi dei sotto-campioni di distretto ottenuti classificando per sesso, classi d'età e nazionalità dei rispondenti sommino alla dimensione che queste sottopopolazioni hanno a livello di distretto. Questa idea, attraente in linea di principio, richiede in pratica che i campioni per cui si ricalibrano i pesi siano grandi. Una ricalibrazione a livello di piccoli domini può porre i pesi "sotto stress", aumentandone in modo anche cospicuo la variabilità. L'obiettivo di avere stimatori con una potenziale minore distorsione viene così ottenuto al prezzo di un forte incremento della variabilità degli stessi, secondo un "effetto disegno" ben noto in letteratura.

Una soluzione di compromesso, considerata in una versione preliminare di questo rapporto, può essere quella di utilizzare le province come livello di calibrazione, al posto dei distretti. Tecnicamente, ciò potrebbe portare ad un sistema di pesi che riporti ad una popolazione più "vicina" a quella del distretto, limitando tuttavia l'effetto di inflazione

della varianza dovuto alla calibrazione di sotto-campioni numericamente molto piccoli. Un'analisi accurata ha rilevato tuttavia che questa soluzione non è soddisfacente: le differenze di struttura nelle popolazioni sono in genere più forti tra i distretti della stessa provincia che non tra una provincia e l'altra, per cui si finisce con l'ottenere pesi simili per struttura a quelli ricalibrati a livello regionale ma più variabili e che comunque non riportano ai totali di distretto e non rappresentano accuratamente la struttura della loro popolazione.

A1. 5 Domini pianificati e domini di interesse

L'obiettivo principale di EU-SILC consiste nello stimare i principali indicatori relativi alla povertà e alla distribuzione del reddito per il paese nel suo complesso e per le sue Regioni (comprese le province autonome di Trento e Bolzano) con un livello di precisione sufficientemente elevato.

Le Regioni e le due province autonome citate rappresentano quindi i domini pianificati dell'indagine EU-SILC.

L'oggetto di studio di questo rapporto sono i distretti sociosanitari e le province della Regione Emilia-Romagna. I distretti sono ripartizioni geografiche di dimensioni ridotte, che non solo non rappresentano domini pianificati per l'indagine ma per cui la dimensione campionaria disponibile per la stima dei parametri è in genere molto ridotta e talvolta del tutto assente.

La numerosità campionaria (in termini di individui rispondenti) per ciascuno dei 38 distretti sociosanitari della regione Emilia-Romagna per l'anno di indagine 2019 va da un minimo di 0 (che è osservato in 9 distretti) ad un massimo di 292. La dimensione campionaria mediana è 68.5 (primo quartile 33.5, terzo quartile 122.8) mentre quella media è di 83.6. Si tratta di numeri nella maggior parte inadeguati a garantire stime di efficienza adeguata adottando i dati forniti dall'indagine come sola base informativa. Gli stessi numeri, calcolati per gli anni di indagine precedenti, non si discostano di molto e portano a conclusioni analoghe sulla possibilità di utilizzare EU-SILC come sola base informativa.

A2. Reddito disponibile equivalente

Uno degli indicatori che formano l'oggetto di questo studio, l'AROP, viene stimato sulla base del reddito disponibile equivalente.

In termini informali, "disponibile" significa effettivamente utilizzabile da parte delle famiglie per l'acquisto di beni e servizi. Calcolare il reddito equivalente significa calcolare un reddito pro-capite per i membri del nucleo familiare in cui però il numero dei componenti non consiste nel semplice conteggio ma viene corretto per le economie di scala che i membri delle famiglie realizzano convivendo e mettendo insieme le risorse. Il reddito equivalente si ottiene dividendo il reddito disponibile totale di una famiglia per un coefficiente (dimensione equivalente) che tiene in conto il numero dei componenti e composizione per età della famiglia. Il reddito disponibile totale della famiglia è la somma dei redditi netti percepiti dai suoi membri ed è la proxy che viene utilizzata per descrivere l'insieme delle risorse che le famiglie hanno a disposizione per raggiungere i propri obiettivi di benessere.

Nella tabella A1 riportiamo la definizione dettagliata dell'aggregato "reddito disponibile totale familiare", facendo esplicito riferimento a quali variabili nel database di EU-SILC sono state coinvolte nel calcolo. Tale definizione accoglie in modo sostanzialmente fedele le raccomandazioni del manuale di Camberra.

Tabella A1 *Definizione di reddito netto familiare*

<p>Addendi positivi</p>	<p>reddito lordo da lavoro dipendente di tutti i componenti della famiglia (sia monetario, sia quasi-monetario (PY010G), sia non monetario (PGY020G); contributi sociali a carico dei datori di lavoro (PY030G); guadagni o perdite lorde da lavoro autonomo compresi diritti e royalties (PG050G); valore dei beni prodotti ai fini dell'autoconsumo (PY070G); sussidi e assegni di disoccupazione (PY090G); pensioni di anzianità (PY100G), pensioni di reversibilità ai superstiti (PY110G), indennità di malattia (PY120G), invalidità (PY130G); borse di studio, premi ed altre sovvenzioni relative all'istruzione (PY140G); componenti di reddito lordo attribuibili alla famiglia nel suo complesso: fitto figurativo (HY030G), redditi derivanti dalla locazione di beni o terreni (HY040G), indennità inerenti alla famiglia e ai figli (HY050G), prestazioni sociali erogate a persone a rischio di emarginazione sociale e non classificate altrove (HY060G), indennità di alloggio erogate dagli enti pubblici (HY070G), trasferimenti attivi di denaro tra famiglie quali assegni ricevuti dal coniuge separato (HY080G); interessi, dividendi, utili da investimenti di capitale in imprese non costituite in società (HY090G); reddito percepito da membri della famiglia di età inferiore ai 16 anni (HY110G)</p>
<p>Addendi negativi</p>	<p>contributi sociali a carico dei datori di lavoro (PY030G); interessi pagati sui mutui ipotecari (HY100G); imposte correnti sul patrimonio (HY120G); trasferimenti passivi di denaro tra famiglie quali assegni ricevuti dal coniuge separato (HY130G); imposte sul reddito e contributi sociali a carico del lavoratore al netto di rimborsi/entrate per conguagli fiscali, le imposte sul reddito alla fonte (HY140G)</p>

Occorre notare come ISTAT attui una serie di misure volte a contenere gli effetti della non risposta, degli errori e dell'*under-reporting* sul calcolo del reddito. Nel caso di non-risposta totale da parte di uno dei componenti della famiglia si procede all'imputazione del dato mancante sulla base di modelli statistici (ISTAT, 2008). Per quanto riguarda errori ed *under-reporting*, quest'ultimo per varie ragioni particolarmente rilevante nel caso di percettori di reddito da lavoro autonomo, il reddito risultante dal questionario IT-SILC viene confrontato con quello risultante dalle dichiarazioni fiscali. Nell'ottica della minimizzazione dell'ammontare dell'*under-reporting* viene imputato il valore massimo tra quello che figura nel questionario e quello fiscalmente dichiarato. Chiaramente questo tipo di controlli è reso possibile dal livello di efficacia raggiunto dalle procedure di *linkage* tra questionari IT-SILC e dati di fonte fiscale.

Notiamo infine che, per effetto della definizione adottata è possibile osservare redditi famigliari netti negativi. Tra gli addendi positivi non sono inclusi poste quali “vincite alla lotteria”, “guadagni in conto capitale”, “aumenti delle disponibilità monetarie in conseguenza di denaro preso in prestito” classificate come variazioni istantanee della ricchezza e non come componenti di reddito. D’altra parte, tra gli addendi negativi sono inclusi trasferimenti ad altre famiglie (e tra essi gli assegni di mantenimento del coniuge separato) nonché le perdite per attività di lavoro autonomo.

A2.1 Scala di equivalenza

La dimensione equivalente della famiglia è stata calcolata utilizzando la scala di equivalenza OCSE modificata.

$$EQ_SIZE = 1 + 0.5(CF_{14+} - 1) + 0.3(CF_{13-})$$

dove CF_{14+} è il numero di componenti della famiglia con età ≥ 14 e CF_{13-} il numero di quelli di età inferiore.

Il reddito disponibile equivalente di ciascun individuo incluso nel campione è stato quindi calcolato come rapporto tra il reddito disponibile totale della famiglia di cui è membro diviso per il fattore di equivalenza. Si noti che, come conseguenza di questa definizione tutti i membri della stessa famiglia hanno lo stesso reddito disponibile equivalente. La distribuzione campionaria del reddito disponibile equivalente è caratterizzata da un’asimmetria molto forte.

È ragionevole ritenere che i redditi molto alti osservati non siano dovuti ad errori nella misurazione ma rappresentino dati corretti. Tuttavia, occorre tenere in conto che la loro influenza sulla stima di parametri quali medie e indici di concentrazione può essere molto grande. In particolare, lo sarà quando la stima viene calcolata a livello di domini di piccole dimensioni, in cui la presenza di outliers non è in alcun modo compensata dall’elevato numero di osservazioni. Inoltre, proprio per la loro rarità questi valori anomali si distribuiranno in modo casuale in alcune, ma non in tutte le aree, finendo con l’aver un effetto distorsivo su eventuali confronti tra le stime dello stesso parametro in domini diversi.

Riguardo alla produzione delle stime “dirette” si sono poste tre alternative di azione a riguardo di questi *outliers*:

1. non correzione. Si tratta dell’opzione meno “invasiva” e che mantiene inalterato il dataset fornito da ISTAT. Ciò, tuttavia, comporta il peggioramento delle proprietà statistiche degli stimatori e ne rende più difficoltosa l’interpretazione;
2. fissazione di una soglia arbitraria e: 2a) eliminazione delle osservazioni il cui valore del reddito disponibile equivalente eccede la soglia; 2b) per le osservazioni

per cui il valore del reddito disponibile equivalente eccede la soglia, esso viene imputato al valore della soglia (winsorizzazione);

3. fissazione di una soglia arbitraria e modellazione parametrica della coda della distribuzione sia per le aree che presentano outliers, sia per quelle che ne sono prive. Un simile modo di trattare il problema, con riferimento all'indagine EU-SILC è descritto ad esempio in Balog e Thornburn (2007).

Nel presente lavoro abbiamo scelto la via del trattamento degli outliers mediante una winsorizzazione della coda destra della distribuzione e una modellazione paretiana della coda sinistra della distribuzione. Sono stati imputati a zero 140 redditi equivalenti (individuali) pari a 0 e 143 redditi equivalenti che risultavano negativi.

La modellazione della coda sinistra della distribuzione consiste nell'individuazione di una soglia oltre la quale si assume che la distribuzione di Pareto fornisca un modello adatto per i dati e successivamente nella stima del parametro di forma che caratterizza questa distribuzione. Per una motivazione dettagliata di questa scelta si rimanda alla letteratura e in particolare a Van Kerm (2007) e Alfons e Templ (2013) che l'hanno proposta proprio nell'analisi dei dati EU-SILC. La soglia è stata individuata, a livello nazionale, utilizzando il metodo di Van Kerm (Van Kerm, 2007), mentre il parametro di forma che caratterizza la distribuzione è stato stimato con il metodo del "partial density component" (Vandewall et al., 2007). Per i calcoli è stato utilizzato il pacchetto Laeken di R (Alfons e Templ, 2013). Le stime del parametro di forma per la distribuzione Pareto e della soglia di reddito stimata con il metodo qui descritto sono rispettivamente pari a 2.68 ed a 50538 euro. Il numero di osservazioni oltre la soglia è pari a 773.

A3. Definizione degli indicatori di interesse

L'obiettivo di questo rapporto è stimare, per ciascuno dei distretti e per ogni provincia della regione Emilia-Romagna, i seguenti indicatori

1. tasso di rischio di povertà o esclusione sociale (anche detto AROPE, At Risk Of Poverty or Social Exclusion), dato dalla quota di individui che sono caratterizzati da almeno una delle condizioni individuate dagli indicatori 2, 3 e 4 seguenti.
2. tasso di Rischio di povertà (anche detto AROP, At Risk Of Poverty), dato dalla quota degli individui che dispongono di un reddito equivalente inferiore ad una soglia (di povertà) individuata nel 60% del reddito mediano nazionale equivalente
3. tasso di Deprivazione materiale grave (anche detto SMD, Severe Material Deprivation) dato dalla quota di individui che vivono in famiglie che sperimentano almeno quattro tra nove sintomi di disagio (vedi Introduzione)

4. tasso di Bassa intensità lavorativa (anche detto QJ, Quasi-Joblessness), dato dalla quota di individui adulti che vivono in famiglie ad intensità di lavoro molto bassa

Si noti che l'indicatore di cui al punto 4 non è calcolabile per famiglie composte solo da bambini, da studenti di età inferiore a 25 anni e/o da persone di età pari o superiore a 60 anni. Per queste famiglie, quindi, la condizione di rischio di povertà o esclusione sociale (indicatore di cui al punto 1) viene definita solo sulla base degli indicatori 2 e 3.

Il tasso di rischio di povertà (At-Risk-of Poverty, AROP)

Per semplicità indichiamo la variabile reddito equivalente con Y e con y_{jd} il valore che essa assume in corrispondenza dell'individuo j -esimo ($j=1,\dots,N_d$) incluso nel dominio d -esimo. Ricordiamo che $d=1,\dots,9$ e $d=1,\dots,38$ rispettivamente per le province e per i distretti sociosanitari.

Il parametro è definito nel modo seguente:

$$AROP_d = \frac{\sum_{j=1}^{N_d} \mathbf{1}(y_{jd} \leq PT)}{N_d}$$

in cui N_d è la numerosità di popolazione e PT è la soglia di povertà relativa che coincide con il 60% del reddito disponibile equivalente mediano nazionale. Notiamo che $\mathbf{1}(y_{jd} \leq PT)$ è una funzione indicatrice che assume valore 1 se $y_{jd} \leq PT$, 0 altrimenti.

Tasso di deprivazione materiale severa (Severe Materiale Deprivation, SMD)

La deprivazione materiale è definita come una situazione di involontaria incapacità di sostenere spese per determinati beni o servizi. Si considerano i seguenti segnali di deprivazione, rilevati tramite l'indagine EU-SILC:

1. essere in arretrato nel pagamento di almeno una tra le bollette di gas, luce, acqua, oppure essere in arretrato nel pagamento dell'affitto, mutuo o altro prestito;
2. non potersi permettere di riscaldare adeguatamente l'abitazione di residenza;
3. incapacità di affrontare spese impreviste (fissate nella misura di 800 euro) con risorse proprie;
4. incapacità di fare un pasto adeguato almeno una volta ogni due giorni;

5. impossibilità di permettersi una settimana di ferie lontano da casa una volta l'anno;
6. non potersi permettere un televisore a colori;
7. non potersi permettere il frigorifero;
8. non potersi permettere l'automobile;
9. non potersi permettere il telefono;

Il concetto di "pasto adeguato ogni due giorni" si traduce nella domanda: *“La sua famiglia, se volesse, potrebbe permettersi di mangiare carne, pollo o pesce (o equivalente vegetariano) almeno una volta ogni due giorni?”*

Il tasso di deprivazione materiale è definito come la frazione di persone che vivono in famiglie con almeno 4 segnali di deprivazione materiale. Questo indicatore prescinde dai problemi di misurazione implicati dal riferimento ad una misura quantitativa della capacità come il reddito e introduce il riferimento ad una situazione di involontaria incapacità di sostenere spese relative ad un determinato paniere di beni e servizi. Notiamo che il tasso di deprivazione materiale considera la famiglia come orizzonte fondamentale per la definizione dello stato di povertà: le risposte vengono fornite da un solo rispondente tra i componenti del nucleo intervistato.

L'indicatore di deprivazione severa viene definito attribuendo ad ogni individuo j il numero di segnali di deprivazioni conteggiate per la famiglia a cui l'individuo j appartiene. Si definisce quindi l'indicatore come segue:

$$SMD_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \mathbf{1}(r_{dj} \geq 4)$$

in cui r_{dj} è il numero di segnali di deprivazione riferito all'individuo j -simo nel dominio d -simo. Se $r_{dj} \geq 4$ la funzione indicatrice $\mathbf{1}(r_{dj} \geq 4)$ assume valore pari a 1 (e quindi attribuisce all'individuo j lo status di deprivato), 0 altrimenti.

Il tasso di bassa intensità lavorativa (Quasi Joblessness, QJ)

L'indicatore di bassa intensità lavorativa (anche detto QJ, Quasi-Joblessness) è dato dalla quota di individui di età compresa tra 0 e 59 anni, che vivono in famiglie ad intensità di lavoro molto bassa, cioè inferiore al 20%. L'intensità di lavoro è definita come rapporto fra il numero totale di mesi lavorati dai componenti della famiglia durante l'anno di riferimento dei redditi (quello precedente all'anno di rilevazione) e il numero totale di mesi teoricamente disponibili per attività lavorative.

I mesi lavorati sono misurati in termini di lavoro equivalente a tempo pieno. Per le persone che hanno dichiarato di aver lavorato a tempo parziale, una stima del numero

di mesi in termini di equivalente a tempo pieno è calcolata sulla base del numero di ore abitualmente lavorate al momento dell'intervista. Il potenziale totale della famiglia è dato dal numero totale di mesi che potrebbero, in teoria, essere stati lavorati dagli adulti nella stessa famiglia.

Sono considerate in età lavorativa le persone in età 18-59 con esclusione degli studenti di età compresa tra i 18 e i 24 anni e le persone che pur non avendo raggiunto i 59 auto-identifichino il loro stato come "pensionato" e percepiscano effettivamente un reddito da pensione.

Nei prossimi anni l'indice sarà rivisto per estendere l'età lavorativa fino ai 64 anni.

Più in dettaglio, l'indicatore di bassa intensità di lavoro viene definito attribuendo ad ogni individuo j l'intensità di lavoro conteggiata per la famiglia a cui l'individuo j appartiene. Al fine di individuare gli individui a bassa intensità di lavoro si assume una soglia pari a 0.2, e quindi l'individuo è a "bassa intensità di lavoro" se è caratterizzato da una percentuale di intensità di lavoro inferiore al 20%. Si definisce quindi l'indicatore come segue:

$$QJ_d = \frac{1}{N_{d(0-59)}} \sum_{j=1}^{N_{d(0-59)}} \mathbf{1}(z_{dj} \leq 0.2)$$

in cui z_{dj} è l'intensità di lavoro riferita all'individuo j -simo nel dominio d -simo; $N_{d(0-59)}$ indica il numero di individui nella fascia d'età considerata dall'indicatore. Se $z_{dj} \leq 0.2$ la funzione indicatrice $\mathbf{1}(z_{dj} \leq 0.2)$ assume valore pari a 1 (e quindi attribuisce all'individuo j lo status di bassa intensità di lavoro), 0 altrimenti. Come già anticipato, si noti che l'intensità di lavoro non è definita, e quindi non viene calcolata, per gli individui che vivono in famiglie composte solo da bambini, da studenti di età inferiore a 25 anni e/o da persone di età pari o superiore a 60 anni. In questo caso il valore della funzione indicatrice $\mathbf{1}(z_{dj} \leq 0.2)$ viene posto uguale a 0 nella definizione dell'indice di rischio di povertà ed esclusione sociale.

Il tasso di rischio di povertà ed esclusione sociale (At-Risk-of Poverty and Social Exclusion, AROPE)

L'indicatore di povertà o esclusione sociale è definito quindi nel modo seguente

$$AROPE_d = \frac{1}{N_d} \sum_{j=1}^{N_d} \mathbf{1}(y_{dj} \leq PT \cup z_{dj} \leq 0.2 \cup r_{dj} \geq 4)$$

L'individuo identificato come a rischio di povertà o esclusione sociale è, quindi, caratterizzato dall'essere identificato come povero oppure deprivato in maniera severa, oppure a bassa intensità di lavoro.

A4. Definizione degli stimatori per la produzione delle stime dirette

Passiamo ora ad illustrare gli stimatori “diretti” dei parametri ora introdotti. Per stimatore “diretto” si intende, nel gergo delle piccole aree, uno stimatore di un parametro descrittivo di un dominio di interesse che utilizza l'informazione sulla variabile obiettivo y relativa al solo dominio in questione. In pratica con stimatori diretti si indicano stimatori con proprietà ottimali rispetto al disegno campionario, ossia rispetto alla distribuzione di probabilità indotta dalla selezione casuale del campione. Gli stimatori che ora introdurremo sono tutti consistenti rispetto al disegno, nel senso che a questo termine viene attribuito nella letteratura sulla teoria dei campioni (si veda ad esempio Isaki e Fuller, 1982).

Assumiamo di denotare con w_{dj} il fattore di riporto all'universo (vedi anche par. 6) per l'individuo j -esimo nel dominio di interesse d -esimo. Riportiamo di seguito le formule degli stimatori “diretti”; con un abuso di notazione utilizziamo gli stessi simboli per denotare il parametro da stimare e lo stimatore diretto che utilizziamo per stimarlo.

- Tasso di povertà:
$$AROP_d = \frac{\sum_{j \in s_d} w_{dj} \mathbf{1}(y_{dj} \leq PT)}{\sum_{j \in s_d} w_{dj}}$$

dove s_d indica l'insieme delle unità campionate nel dominio d .

- Tasso di deprivazione materiale severa:

$$SMD_d = \frac{\sum_{j \in s_d} w_{dj} \mathbf{1}(r_{dj} \geq 4)}{\sum_{j \in s_d} w_{dj}}$$

- Tasso di bassa intensità lavorativa:

$$QJ_d = \frac{\sum_{j \in s_d(0-59)} w_{dj} \mathbf{1}(z_{dj} \leq 0.2)}{\sum_{j \in s_d(0-59)} w_{dj}}$$

dove $s_d(0-59)$ indica l'insieme degli individui campionato di età compresa tra 0 e 59 anni.

- Tasso di povertà o esclusione sociale

$$AROPE_d = \frac{\sum_{j \in s_d} w_{dj} \mathbf{1}(y_{dj} \leq PT \cup z_{dj} \leq QJ \cup r_{dj} \geq 4)}{\sum_{j \in s_d} w_{dj}}$$

A5. Stima della varianza delle stime dirette

Il disegno campionario di EU-SILC è complesso e non riconducibile a nessun disegno “notevole” a cui poter applicare stimatori delle varianze (siano essi stimatori diretti, per linearizzazione o basati su metodi di ri-campionamento) già proposti in letteratura.

Il campione si rinnova per un quarto ogni anno e la nuova componente è campionata con un disegno a due stadi con stratificazione delle unità di primo stadio. I comuni campione non sono gli stessi ogni anno e a quelli campionati, altri si aggiungono a causa delle operazioni di inseguimento delle unità panel. Questi aspetti, oltre alla complessità delle operazioni di calibrazione condotte da ISTAT rendono problematica la ricostruzione esatta delle informazioni relative al disegno e alla costruzione dei pesi effettuate da ISTAT.

Non è possibile calcolare le probabilità di inclusione del secondo ordine che sarebbero necessarie per gli stimatori “diretti” delle varianze. Le operazioni di calibrazione effettuate da ISTAT sul campione originario che, in molti algoritmi di ri-campionamento devono essere ri-effettuate per ciascun campione estratto nella simulazione non sono ricostruibili con esattezza. Inoltre, per effetto della natura panel del campione, non è possibile ricostruire – sulla base delle informazioni disponibili – quali unità di primo livello, ossia comuni, siano incluse nel campione per esser state effettivamente campionate e quali vi compaiano per esclusivo effetto delle regole di inseguimento.

Nella letteratura applicata che si è occupata della stima della varianza di stimatori calcolati utilizzando i dati EU-SILC, è emerso come l'applicazione di stimatori delle varianze che tenessero conto di alcuni elementi essenziali del disegno, ancorché non di tutti, porti a stime sufficientemente affidabili (Biewen and Jenkins, 2006; Goedemé, 2013). La nostra procedura di stima della varianza segue questo tipo di impostazione.

Una prima semplificazione del disegno consiste nel considerare l'intero campione EU-SILC come se fosse estratto simultaneamente, trascurando la natura di panel ruotato del campione. Gli elementi che consideriamo essenziali nel disegno e includiamo nella stima della varianza sono i seguenti:

- il raggrupparsi degli individui all'interno di *cluster* definiti dalle famiglie. È un aspetto molto importante poiché nonostante tutti gli indicatori siano definiti in

termini di individui, il reddito equivalente (o lo stato di deprivazione materiale) è lo stesso per tutti gli individui all'interno dello stesso *cluster*;

- la stratificazione delle unità;
- la variabilità dei pesi, di entità non trascurabile.

L'articolazione del campionamento della componente rotante del panel in due stadi non è quindi tenuta esplicitamente in conto. Questa scelta è motivata, dalla debole correlazione intra-cluster riscontrata nella maggior parte degli studi sul tema (Goedemé, 2013) e supportata dalla pratica diffusa nella letteratura, in cui spesso si devono fare i conti con la non disponibilità della localizzazione geografica delle osservazioni.

Una volta adottate queste semplificazioni, è stato utilizzato uno stimatore delle varianze per linearizzazione di tipo standard, così come esso è implementato nel pacchetto di R `survey` (Lumley, 2004, 2017) per tutti gli stimatori considerati, che sono tutte proporzioni. Più nel dettaglio si tratta di un algoritmo ricorsivo noto in letteratura e descritto in Bellhouse (1985) per i disegni multi-stage stratificati. Il pacchetto `survey` offre anche la possibilità di calcolare le varianze secondo una tecnica nota come Ultimate Cluster correntemente utilizzata anche da ISTAT. Nel caso del problema in esame le due tecniche portano esattamente agli stessi risultati.

In precedenti edizioni del rapporto veniva usato un algoritmo di tipo bootstrap basato su alcune semplificazioni del disegno in linea con quelle che abbiamo appena descritto. Gli algoritmi bootstrap permettono di ottenere le varianze per qualsiasi stimatore (anche non lineare) al prezzo di un notevole “sforzo” computazionale. Per due *wave* consecutive dell'indagine abbiamo confrontato le stime bootstrap con quelle ottenute utilizzando gli stimatori per linearizzazione. I risultati sono molto simili. Ciò conferma le aspettative teoriche: a parità di disegno non vi sono ragioni per ritenere che stimatori bootstrap e per linearizzazione si comportino in modo sostanzialmente diverso. Abbiamo quindi deciso di utilizzare gli stimatori per linearizzazione perché sono più “leggeri” dal punto di vista computazionale.

Infine, nella presentazione dei risultati, le stime regionali e riferite all'Italia sono prodotte con la metodologia utilizzata correntemente dall'Istat, mentre le stime delle loro varianze le otteniamo con la metodologia di stima della varianza qui illustrata.

Nel successivo paragrafo A6 riportiamo le stime dirette ed i rispettivi CV relativi al 2019 per le province dell'Emilia-Romagna (Tab. A2.1) e le stime dirette per i distretti (Tab. A3.1) ed i rispettivi CV% (Tab. A4.1) per i distretti. Nelle tabelle Tab. A2.2, A3.2 ed A4.2 le medesime stime riferite al 2020.

Introduciamo qui il tema del lisciamiento delle varianze riferite a tutti gli stimatori considerati, anticipando alcune considerazioni che saranno più chiare nel paragrafo B4 dell'Appendice B in cui descriviamo i modelli utilizzati nel contesto della metodologia di stima per piccole aree. I modelli che usiamo sono definiti a livello di piccola area (in altre parole nella stima di questi modelli la piccola area costituisce l'unità statistica) e richiedono, per essere identificati, la varianza dello stimatore diretto. La stima di tale varianza, tuttavia, soffre dello stesso problema di cui soffrono le stime dirette dei parametri oggetto di interesse, e cioè si basano su una numerosità campionaria scarsa. Per ovviare a questo problema è pratica usuale nel contesto della stima per piccole aree utilizzare metodi di "lisciamiento" delle stime dirette delle varianze che consente, appunto, di evidenziare, attraverso un modello di regressione, la relazione di fondo tra varianza e stima diretta del parametro oggetto di studio, attenuando il rumore generato, in alcune osservazioni (cioè in alcune piccole aree) dalla numerosità campionaria ridotta. Tali metodologie vanno sotto il nome di Generalized Variance Functions (Wolter, 2007). Nel seguito illustriamo la procedura seguita in questo contesto, adatta a tutti gli stimatori di interesse, in quanto tutti riferiti a proporzioni. Per semplificare la trattazione consideriamo un generico stimatore p_d (riferito al dominio d , con $d=1,\dots,D$) della proporzione θ_d e la sua varianza $V(p_d)$.

Se il piano di campionamento fosse casuale semplice con reintroduzione (srs) allora

$$V(p_d) = \frac{\theta_d(1-\theta_d)}{n_d^*}$$

dove n_d^* è il numero di individui campionati nel dominio d . Se consideriamo le varianze degli stimatori $\hat{V}_{lin}(p_d)$ ottenute con il metodo della linearizzazione, per un campione di disegno complesso, possiamo aspettarci, approssimativamente, che

$$\frac{\hat{p}_d(1-\hat{p}_d)}{\hat{V}_{lin}(p_d)} \propto n_d^*$$

Varie ragioni rendono questa relazione soltanto approssimativa: *i)* il campionamento è per cluster (le famiglie) e non per individui. Tutti i membri di una famiglia sono poveri o non poveri allo stesso modo, ma non tutte le famiglie hanno la stessa dimensione e la distribuzione delle famiglie per dimensione può variare da un dominio all'altro; *ii)* nel calcolo delle stime usiamo i pesi; *iii)* sia le \hat{p}_d , sia $\hat{V}_{lin}(p_d)$ sono caratterizzate da un errore campionario non trascurabile.

Ad esempio, con riferimento ai dati riferiti al 2013, nella figura A.1 n_d^* e

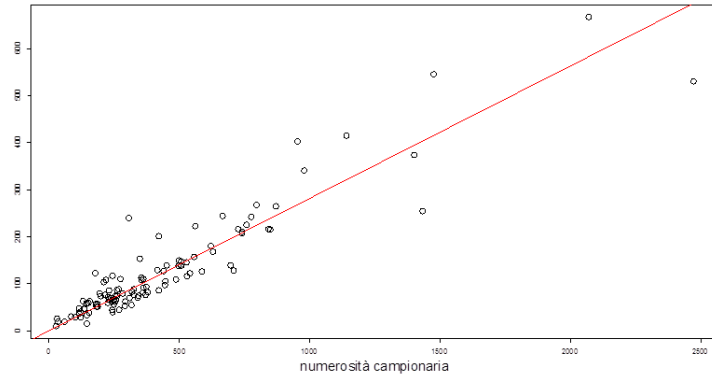
$\frac{1}{\hat{V}_{lin}(p_d)}[\hat{p}_d(1-\hat{p}_d)]$ sono rappresentati in un diagramma a dispersione in cui ciascun

punto corrisponde ad una provincia. Alla nuvola dei punti è sovrapposta una retta di equazione $y = \alpha x$ che descrive l'adattamento di una funzione lineare ai dati. Le stime del parametro α e la misura di adattamento R^2 del modello lineare ai dati sono rispettivamente pari a 0,25 ed a 0,85. L'adattamento della retta ai dati è soddisfacente.

Il coefficiente angolare, ampiamente inferiore a 1 indica che $\hat{V}_{lin}(p_d) > \frac{\hat{p}_d(1-\hat{p}_d)}{n_d^*}$ e ciò è

coerente con la struttura a cluster del campionamento. La differenza tra la stima ottenuta per il coefficiente angolare e 1 (che avremmo se il campionamento fosse casuale semplice) è attribuibile al fatto che il disegno è comunque meno efficiente di quello casuale semplice per via del campionamento a due stadi e che soprattutto, l'introduzione dei pesi nelle formule degli stimatori se da un lato li rende potenzialmente meno distorti, dall'altro ne inflaziona varianza.

Figura A.1 *Numerosità campionaria di dominio n_i^* e rapporto $\frac{1}{\hat{V}_{lin}(p_d)}[\hat{p}_d(1-\hat{p}_d)]$*



Nel paragrafo B4.1 dell'Appendice 2 riprenderemo il tema della specificazione e stima del modello di lisciamento della varianza nel contesto della specificazione del modello per piccole aree.

A6. STIME DIRETTE

Tab. A2.1- Stime dirette e rispettivi coefficienti di variazione (CV%) relativi alle province dell'Emilia-Romagna, anno 2019

codice prov	provincia	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà - AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n
<i>stima</i>						
33	PC	10.59	7.67	0.59	4.71	188
34	PR	20.39	12.55	7.65	6.33	337
35	RE	7.95	6.08	0.52	4.98	398
36	MO	15.81	10.59	5.68	3.74	709
37	BO	13.72	9.90	2.39	6.37	750
38	FE	11.00	5.77	4.01	6.25	128
39	RA	18.90	16.25	1.40	5.92	279
40	FC	33.56	23.12	11.75	1.61	208
99	RN	11.63	8.37	2.71	2.37	180
	ER	15.53	10.91	3.96	4.94	3177
	Italia	25.60	20.10	7.40	10.00	43400
<i>CV%</i>						
33	PC	31.85	41.04	73.23	47.40	
34	PR	29.81	29.19	76.50	38.75	
35	RE	23.18	27.56	81.64	32.41	
36	MO	19.10	25.19	34.42	31.53	
37	BO	17.73	21.65	59.00	32.96	
38	FE	38.88	49.25	82.21	56.84	
39	RA	24.55	27.61	58.20	47.18	
40	FC	19.03	26.23	33.21	74.72	
99	RN	32.30	40.95	58.15	60.38	
	ER	8.83	10.35	23.82	15.18	
	Italia	2.03	2.51	4.69	4.15	

Tab. A3.1 - Stime dirette dei tassi (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019

Denominazione distretto	Provincia	Rischio di povertà ed esclusione sociale - AROPE	Rischio di povertà - AROP	Deprivazione materiale severa - SMD	Bassa intensità lavorativa - QJ	n
CASALECCHIO DI RENO	Bologna	5.59	5.59	0.00	9.51	35
PORRETTA TERME	Bologna	6.89	1.47	0.00	7.96	33
SAN LAZZARO DI SAVENA	Bologna	4.06	3.50	0.00	0.92	123
IMOLA	Bologna	5.01	5.01	0.00	0.00	59
PIANURA EST	Bologna	20.72	11.89	6.67	11.70	208
PIANURA OVEST	Bologna					0
CITTA' DI BOLOGNA	Bologna	16.59	14.67	0.65	4.35	292
OVEST	Ferrara					0
CENTRO-NORD	Ferrara	11.00	5.77	4.01	6.25	128
SUD-EST	Ferrara					0
FORLI'	Forli	29.51	19.76	11.99	0.00	79
CESENA - VALLE DEL SAVIO	Forli	36.29	25.38	11.59	2.81	129
RUBICONE	Forli					0
CARPI	Modena	40.19	39.11	1.66	2.30	47
MIRANDOLA	Modena	27.02	22.33	8.84	5.62	73
MODENA	Modena	9.93	2.06	6.91	3.85	197
SASSUOLO	Modena	9.96	7.26	0.00	3.68	113
PAVULLO NEL FRIGNANO	Modena	12.03	10.60	0.00	8.44	52
VIGNOLA	Modena	9.34	7.60	1.44	0.40	111
CASTELFRANCO EMILIA	Modena	19.10	9.75	12.66	3.50	116
PARMA	Parma	21.28	8.59	11.56	1.55	179
FIDENZA	Parma	13.90	10.56	1.25	12.87	108
VALLI TARO E CENO	Parma					0
SUD EST	Parma	31.48	31.48	7.47	10.56	50
CITTA' DI PIACENZA	Piacenza	13.91	6.38	0.00	15.08	53
LEVANTE	Piacenza	9.26	8.18	0.83	1.76	135
PONENTE	Piacenza					0
RAVENNA	Ravenna	21.68	21.68	1.14	2.48	114
LUGO	Ravenna	27.18	21.37	3.81	13.43	64
FAENZA	Ravenna	9.98	6.61	0.00	4.58	101
GUASTALLA	Reggio					0
CORREGGIO	Reggio	9.72	6.79	2.93	6.03	60
SCANDIANO	Reggio	4.68	1.73	0.00	5.75	122
CASTELNUOVO NE' MONTI	Reggio	5.01	5.01	0.00	0.00	42
MONTECCHIO	Reggio					0
REGGIO EMILIA	Reggio	11.23	10.27	0.29	4.83	174
RIMINI	Rimini	11.63	8.37	2.71	2.37	180
RICCIONE	Rimini	5.59	5.59	0.00	9.51	0
	ER	15.53	10.91	3.96	4.94	3177
	Italia	25.60	20.10	7.40	10.00	43400

Tab. A4.1 - Stime dei coefficienti di variazione delle stime dirette dei tassi (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2019

Denominazione distretto	Provincia	Rischio di povertà - AROP	Deprivazione materiale severa - SMD	Bassa intensità lavorativa - QJ	Rischio di povertà ed esclusione sociale - AROPE
CASALECCHIO DI RENO	Bologna	98.25		98.67	98.25
PORRETTA TERME	Bologna	102.81		98.58	75.67
SAN LAZZARO DI SAVENA	Bologna	52.51		101.93	47.49
IMOLA	Bologna	81.87			81.87
PIANURA EST	Bologna	38.94	62.39	45.88	26.58
PIANURA OVEST	Bologna				
CITTA' DI BOLOGNA	Bologna	28.95	72.83	47.60	25.91
OVEST	Ferrara				
CENTRO-NORD	Ferrara	49.25	82.21	56.84	38.88
SUD-EST	Ferrara				
FORLI'	Forli	55.50	46.27		37.51
CESENA - VALLE DEL SAVIO	Forli	27.73	46.29	74.62	21.09
RUBICONE	Forli				
CARPI	Modena	40.19	102.57	105.72	38.65
MIRANDOLA	Modena	54.15	60.36	80.17	44.95
MODENA	Modena	60.18	52.37	54.25	38.93
SASSUOLO	Modena	56.11		84.87	46.18
PAVULLO NEL FRIGNANO	Modena	66.48		82.46	59.65
VIGNOLA	Modena	57.19	100.35	102.16	48.94
CASTELFRANCO EMILIA	Modena	60.76	59.27	69.24	41.68
PARMA	Parma	45.17	89.55	63.07	46.75
FIDENZA	Parma	43.45	100.66	45.39	34.54
VALLI TARO E CENO	Parma				
SUD EST	Parma	47.55	96.57	96.12	47.55
CITTA' DI PIACENZA	Piacenza	60.01		55.09	37.75
LEVANTE	Piacenza	50.48	73.78	74.65	45.36
PONENTE	Piacenza				
RAVENNA	Ravenna	39.47	100.78	73.15	39.47
LUGO	Ravenna	42.97	71.77	70.54	37.04
FAENZA	Ravenna	62.25		75.95	46.78
GUASTALLA	Reggio				
CORREGGIO	Reggio	61.72	100.46	64.55	53.01
SCANDIANO	Reggio	100.14		51.31	49.88
CASTELNUOVO NE' MONTI	Reggio	72.21			72.21
MONTECCHIO	Reggio				
REGGIO EMILIA	Reggio	34.23	100.98	53.34	32.03
RIMINI	Rimini	40.95	58.15	60.38	32.30
RICCIONE	Rimini				
	ER	10.35	23.82	15.18	8.83
	Italia	2.03	2.51	4.69	4.15

Tab. A2.2- Stime dirette e rispettivi coefficienti di variazione (CV%) relativi alle province dell'Emilia-Romagna, anno 2020

codice prov	Provincia	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà – AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n
stime						
33	PC	6.2	1.2	8.5	4.6	196
34	PR	5.9	0.8	6.7	4.8	220
35	RE	7.7	0.4	10.9	3.9	244
36	MO	3.4	2.3	6.8	2.2	458
37	BO	12.3	0.6	14.6	5.8	523
38	FE	19.0	1.5	20.1	8.5	104
39	RA	10.2	1.2	13.4	6.5	185
40	FC	12.2	1.8	13.4	5.3	109
99	RN	13.1	4.2	16.6	6.7	124
	ER	11.0	8.5	1.3	4.5	2163
	Italia	24.9	20.0	3.5	11.1	29252
CV%						
33	PC	40.1	52.9	61.4	52.8	
34	PR	36.8	40.2	100.5	56.1	
35	RE	29.0	36.1	100.8	49.9	
36	MO	27.2	33.0	59.4	42.7	
37	BO	22.4	25.9	99.9	37.8	
38	FE	43.2	45.9	101.7	56.1	
39	RA	31.7	40.5	86.4	45.5	
40	FC	37.4	40.4	101.3	55.1	
99	RN	34.1	39.7	70.5	49.4	
	ER	11.6	13.3	31.1	17.2	

Tab. A3.2 - Stime dirette dei tassi (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà - AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n
Bologna	CASALECCHIO DI RENO	3.4	3.4	0.0	0.0	23
Bologna	CITTA' DI BOLOGNA	16.2	16.2	0.0	3.1	43
Bologna	IMOLA	0.0	5.7	0.0	8.8	45
Bologna	PIANURA EST	13.9	21.1	2.5	10.5	126
Bologna	PIANURA OVEST	9.1	9.1	0.0	0.0	188
Bologna	PORRETTA TERME	0.0	4.5	0.0	6.7	20
Bologna	SAN LAZZARO DI SAVENA	13.1	13.1	0.0	7.8	78
Ferrara	CENTRO-NORD	19.8	21.2	1.9	6.1	0
Ferrara	OVEST					64
Ferrara	SUD-EST	15.7	15.7	0.0	15.8	40
Forli	CESENA - VALLE DEL SAVIO	16.6	18.8	0.0	7.8	59
Forli	FORLI'	6.8	6.8	3.9	2.4	50
Forli	RUBICONE					0
Modena	CARPI	8.3	10.3	2.2	4.1	27
Modena	CASTELFRANCO EMILIA	3.6	4.1	0.0	0.5	64
Modena	MIRANDOLA	6.7	22.3	13.2	2.7	42
Modena	MODENA	1.7	6.2	2.8	5.0	131
Modena	PAVULLO NEL FRIGNANO	2.5	4.9	2.4	0.0	35
Modena	SASSUOLO	3.5	6.3	0.0	3.8	75
Modena	VIGNOLA	3.1	3.1	0.0	0.0	84
Parma	FIDENZA	8.3	10.8	0.0	11.0	63
Parma	PARMA	3.5	3.5	1.3	2.0	131
Parma	SUD EST	13.6	13.6	0.0	0.0	0
Parma	VALLI TARO E CENO					26
Piacenza	CITTA' DI PIACENZA	13.5	16.1	1.3	9.4	51
Piacenza	LEVANTE	2.2	2.9	0.7	0.0	76
Piacenza	PONENTE	1.8	5.7	1.8	4.2	69
Ravenna	FAENZA	6.9	11.4	0.3	5.9	116
Ravenna	LUGO	0.0	2.2	0.0	4.8	22
Ravenna	RAVENNA	22.3	22.3	4.0	9.5	47
Reggio	CASTELNUOVO NE' MONTI	11.4	18.4	0.0	16.9	84
Reggio	CORREGGIO	4.2	6.0	1.8	1.8	84
Reggio	GUASTALLA					57
Reggio	MONTECCHIO					0
Reggio	REGGIO EMILIA	15.5	16.3	0.0	1.0	0
Reggio	SCANDIANO	2.7	8.3	0.0	6.6	19
Rimini	RICCIONE	31.2	32.0	0.0	10.1	53
Rimini	RIMINI	5.8	10.5	5.9	4.3	71
	ER	11.0	8.48	1.34	4.5	2163
	Italia	24.9	20.0	3.5	11.5	29252

Tab. A4.2 - Stime dei coefficienti di variazione delle stime dirette dei tassi (%) per i distretti sociosanitari dell'Emilia-Romagna, anno 2020

Provincia	Denominazione	Tasso di rischio di povertà ed esclusione sociale - AROPE	Tasso di rischio di povertà - AROP	Tasso di deprivazione materiale severa - SMD	Tasso di bassa intensità lavorativa - QJ	n
Bologna	CASALECCHIO DI RENO	103.1	103.1			23
Bologna	CITTA' DI BOLOGNA	38.1	38.1		62.3	43
Bologna	IMOLA	95.3			97.6	45
Bologna	PIANURA EST	36.0	51.2	99.2	57.5	126
Bologna	PIANURA OVEST	70.8	70.8			188
Bologna	PORRETTA TERME	102.6			103.0	20
Bologna	SAN LAZZARO DI SAVENA	56.0	56.0		96.9	78
Ferrara	CENTRO-NORD	49.5	53.3	102.1	79.0	0
Ferrara	OVEST					64
Ferrara	SUD-EST	60.4	60.4		71.1	40
Forli	CESENA - VALLE DEL SAVIO	42.9	47.8		64.8	59
Forli	FORLI'	73.4	73.4	101.6	105.7	50
Forli	RUBICONE					0
Modena	CARPI	57.0	65.1	103.1	108.0	27
Modena	CASTELFRANCO EMILIA	89.8	99.7		103.7	64
Modena	MIRANDOLA	58.4	76.2	92.5	104.5	42
Modena	MODENA	50.1	71.9	76.2	66.3	131
Modena	PAVULLO NEL FRIGNANO	75.4	103.4	103.5		35
Modena	SASSUOLO	45.9	54.5		76.2	75
Modena	VIGNOLA	99.7	99.7			84
Parma	FIDENZA	50.4	61.5		65.6	63
Parma	PARMA	60.6	60.6	100.6	101.1	131
Parma	SUD EST	92.1	92.1			0
Parma	VALLI TARO E CENO					26
Piacenza	CITTA' DI PIACENZA	51.7	61.5	101.9	64.9	51
Piacenza	LEVANTE	81.4	100.7	102.3		76
Piacenza	PONENTE	55.4	76.2	102.0	82.4	69
Ravenna	FAENZA	43.6	67.9	101.8	58.8	116
Ravenna	LUGO	100.5			115.2	22
Ravenna	RAVENNA	45.8	45.8	99.4	78.0	47
Reggio	CASTELNUOVO NE' MONTI	63.9	95.1		100.4	84
Reggio	CORREGGIO	55.4	63.1	102.4	78.6	84
Reggio	GUASTALLA					57
Reggio	MONTECCHIO					0
Reggio	REGGIO EMILIA	42.1	44.3		101.9	0
Reggio	SCANDIANO	54.8	100.0		66.9	19
Rimini	RICCIONE	40.7	41.9		52.0	53
Rimini	RIMINI	48.2	63.9	70.6	102.0	71
	ER	11.0	8.48	1.34	4.5	2163
	Italia	24.9	20.0	3.5	11.5	29252

Appendice B

Specificazione e selezione dei modelli per la stima per piccole aree

B1. Modelli per piccole aree. Introduzione

L'inadeguata rappresentatività dei campioni EU-SILC disponibili per la maggior parte dei distretti sociosanitari e delle province della regione Emilia-Romagna rende necessario l'utilizzo di modelli di stima per piccole aree in cui l'informazione fornita dal campione venga integrata con informazioni di natura amministrativa.

In questa breve introduzione, discutiamo alcune scelte di fondo da cui dipendono quelle più specifiche che saranno illustrate nei paragrafi successivi.

I modelli per piccole aree possono essere classificati in "area level" e "unit level". Nei modelli "area level" le stime dirette ottenute per ciascuno dei domini in questione (vedi appendice A) vengono messe in relazione con informazione ausiliaria riferita allo stesso livello di disaggregazione geografica o ad un livello superiore. Nei modelli "unit level" il valore delle variabili oggetto di studio rilevati per ciascuna unità statistica nel campione viene messa in relazione con valori di variabili anch'esse note a livello di ciascuna unità del campione.

A seconda dei problemi e delle informazioni disponibili possono essere preferibili gli uni o gli altri. Noi lavoreremo con modelli "area level" per le seguenti ragioni: i) l'insieme di informazioni disponibili a livello di individuo o famiglia e per cui i totali siano noti a livello di distretto o di provincia sono relativamente poche e non particolarmente esplicative; ii) i modelli "area level" permettono di ottenere stime consistenti rispetto al disegno. In altre parole, nei casi dei pochi domini per cui la dimensione del campione è adeguata, lo stimatore ottenuto sulla base del modello converge allo stimatore diretto; iii) i modelli "area level" sono, in termini di informazioni richieste, più parsimoniosi di quelli "unit level" quando le relazioni tra variabili obiettivo e variabili ausiliarie sono di tipo non lineare. Ai primi basta infatti che sia nota la media delle variabili ausiliarie a livello di dominio; ai secondi serve invece che il valore delle variabili ausiliarie sia noto per ciascun individuo / famiglia della popolazione.

Per quanto riguarda la produzione delle stime per i parametri di interesse per l'Emilia-Romagna sia a livello sia di provincia che a livello di distretto, è necessario fare una premessa. In generale, come è motivato nel paragrafo B2, si è scelto di includere nella specificazione e stima dei modelli stime dirette e variabili ausiliarie relative alle province delle rimanenti regioni italiane.

Questa seconda appendice è organizzata nel modo seguente. Nel paragrafo B2 viene discussa la scelta di includere nella specificazione e stima dei modelli stime dirette e variabili ausiliarie relative alle province delle rimanenti regioni italiane. Nel paragrafo B3 presenteremo l'informazione ausiliaria presa in considerazione nel processo di specificazione dei modelli e il processo di selezione delle variabili che ha portato alla scelta delle variabili effettivamente incluse. I paragrafi da B4 a B6 contengono una descrizione dei modelli che sono stati utilizzati per produrre le stime per piccole aree.

Nel paragrafo B7 vengono presentate una serie di misure diagnostiche tese a validare i modelli scelti.

B2. Inclusione delle provincie delle rimanenti regioni italiane nel processo di stima.

Sul piano logico, possiamo pensare ai modelli di tipo “area level” come a modelli in cui le osservazioni sono costituite dalle quantità descrittive dei domini oggetto di studio. Il numero dei domini definisce un sostanziale analogo della dimensione del campione nei problemi classici di statistica. Inoltre le quantità descrittive dei domini non sono note, ma soltanto stimate; in un certo senso possiamo pensare all’errore campionario associato alle stime dirette come ad un errore di misurazione che influenzi la nostra osservazione delle quantità descritte di popolazione.

Per queste ragioni, è semplice comprendere che nei modelli di tipo “area level” maggiore è il numero di aree su cui vengono stimati i parametri di interesse, maggiore sarà la precisione con cui questi ultimi sono stimati. In Emilia-Romagna vi sono 38 distretti sociosanitari, ma le stime dirette sono disponibili solamente per 29, i campioni relativi ai rimanenti 9 distretti essendo vuoti. Ciò implica che per stimare un modello “area level” sono disponibili solamente 29 “osservazioni” caratterizzate da un elevato errore di campionamento. Un discorso analogo vale per le province dell’Emilia-Romagna.

Abbiamo pertanto deciso di coinvolgere nel processo di specificazione e stima dei modelli anche le 101 provincie delle rimanenti regioni italiane.

Vi è più di una ragione che motiva tale decisione:

- quando l’obiettivo è la stima a livello di distretto, va osservato che in molti casi i distretti sociosanitari istituiti in altre Regioni non sono paragonabili a quelli della regione Emilia-Romagna per dimensioni e ruolo;
- le provincie rappresentano aree per le quali è in generale disponibile un’informazione campionaria maggiore di quella che è disponibile per i distretti; le stime dirette a livello provinciale risultano quindi più affidabili di quelle a livello di distretto sociosanitario;
- la relazione tra parametri di interesse e informazione ausiliaria utilizzata nei modelli è più evidente se si utilizzano aree su tutto il territorio nazionale di quanto lo sia concentrando l’attenzione solo su sotto-aree del territorio (ad esempio i distretti dell’Emilia-Romagna); più è forte tale relazione tanto più le stime ottenute a livello di singole aree saranno precise.

Gli stimatori, sia diretti che da modello, dei tassi sono additivi nel senso che la stima a livello provinciale si può ottenere come opportuna media aritmetica ponderata delle stime dei distretti che appartengono alla stessa provincia. Se vale la proprietà di additività si possono seguire due strade alternative per la produzione di stime da

modello a livello provinciale: 1) utilizzare un modello per le sole province italiane; 2) utilizzare un modello stimato sui dati dei distretti della regione Emilia-Romagna e delle province delle rimanenti regioni e poi ottenere le stime per provincia dell'Emilia-Romagna sfruttando la proprietà di additività. Si è preferito seguire questa seconda strada sia perché garantisce un livello di coerenza maggiore tra stime dei distretti e quelle delle province, sia perché gli stimatori da modello ottenuti sono risultati più precisi di quelli ottenuti utilizzando modelli specificati per le sole province.

B3. Analisi e selezione dell'informazione ausiliaria

In un modello di stima per piccole aree del tipo che consideriamo qui ("area level") l'informazione ausiliaria deve essere definita allo stesso livello di dettaglio geografico dei domini considerati oppure ad un livello di aggregazione maggiore. Le fonti che abbiamo preso in considerazione in questo studio sono:

1. Ministero dell'Economia e delle Finanze. Sono stati utilizzati i file annuali contenenti i principali dati fiscali di tutti i comuni italiani, resi disponibili dal Dipartimento delle Finanze. (http://www1.finanze.gov.it/finanze2/pagina_dichiarazioni/dichiarazioni.php)
2. I registri anagrafici dei comuni, accessibili attraverso i database ISTAT, al sito web: <http://demo.istat.it/>. Per gli anni precedenti al Censimento (2011) si è fatto riferimento alla ricostruzione intercensuaria della popolazione.
3. Database ISTAT sulla superficie dei comuni e delle province italiane (<http://www.istat.it/it/archivio/82599>).

Le variabili costruite e la loro descrizione è contenuta nella tabella B1. Le informazioni disponibili a livello comunale sono state aggregate per ottenere le covariate da inserire nei modelli a livello di distretto, per l'Emilia-Romagna, e a livello di provincia per le rimanenti regioni.

Tabella B1 *Descrizione delle informazioni ausiliarie considerate nella costruzione dei modelli.*

Etichetta	Descrizione	Fonte
f_dich	Numero dichiaranti/popolazione totale	Fiscale/anagrafe
rmedio_pop	Reddito totale/popolazione totale	fiscale/anagrafe
rmedio_dich	Reddito totale/Numero dichiaranti	fiscale
dens	Densità di popolazione	anagrafe
freq_giov	Popolazione 0_14/popolazione totale	anagrafe
dip_giov	Popolazione 0_14/popolazione 15_64	anagrafe
dip_sen	Popolazione 65 e oltre /popolazione 15_64	anagrafe
freq_stra	Popolazione straniera/popolazione totale	anagrafe
masc_stra	Numero stranieri maschi/numero stranieri femmine	anagrafe
medialog_f*	Media del log-reddito	fiscale
hr_f*	Frequenza di individui con reddito<10000	fiscale
logit_hr*	Logit della variabile precedente	fiscale
gini_f*	Indice di gini	fiscale
logit_gini*	Logit della variabile precedente	fiscale
rich_f*	Tasso di individui con reddito inferiore a due volte quello mediano	fiscale
logit_rich*	Logit della variabile precedente	fiscale
taddetti	Numero di addetti delle imprese su popolazione totale	ASIA
tlaureati	Numero di laureati su popolazione totale	ISTAT

*si tratta di un'approssimazione perché i dati fiscali sono disponibili in forma raggruppata con le seguenti classi:

da 0 a 10000

da 10000 a 15000

da 15000 a 26000

da 26000 a 55000

da 55000 a 75000

da 75000 a 120000

oltre 120000

B4. I modelli di stima per piccole aree

B4.1 Modello per i tassi

Indichiamo per semplicità con p_{dt} , lo stimatore da disegno del tasso t riferito all'area d , e con θ_{dt} il parametro oggetto di stima ossia il tasso t per la popolazione del dominio d , con $t = 1, \dots, 4$ ($t=1$ definisce *AROP*, $t=2$ definisce *SMD*, $t=3$ definisce *QJ*, $t=4$ definisce *AROPE*) e $d = 1, \dots, D$, dove D è il numero delle aree considerate. L'indice d identifica quindi, di volta in volta, le province ed i distretti. Il noto modello di Fay-Herriot (Rao, 2003), basato sull'ipotesi di distribuzione Normale dello stimatore, non è adatto per la modellazione di una proporzione $p_{dt} \in (0,1)$, un modello basato su questa assunzione potrebbe ad esempio condurre a stime al di fuori dell'intervallo $(0,1)$. Una assunzione adatta a modellare un tasso è quella che fa riferimento ad un modello di tipo Beta. La distribuzione Beta è, infatti, una distribuzione di probabilità continua, definita sull'intervallo unitario.

Poiché la distribuzione Beta non è definita per valori di p_{dt} esattamente uguali a zero o a uno, la distribuzione è stata modificata per tenere conto di casi di stime dirette esattamente uguali a zero, caso non raro nella stima di parametri di povertà. La nuova distribuzione sarà indicata come *Beta_infl*.

Se un tasso segue una distribuzione *Beta_infl*, allora:

$$f(p_{dt}) = (1 - \theta_{dt}^*)^{m_d} \mathbf{1}(p_{dt} = 0) \left\{ 1 - (1 - \theta_{dt}^*)^{m_d} \right\} dBeta(A_{dt}, B_{dt}) \mathbf{1}(p_{dt} > 0) \quad (\text{B.1})$$

dove m_d è il numero di famiglie campionate nel dominio d .

Inoltre,

$$A_{dt} = \theta_{dt}^* \phi_{dt}, \quad B_{dt} = (1 - \theta_{dt}^*) \phi_{dt}, \quad \theta_{dt}^* = E(p_{dt} | p_{dt} > 0, \theta_{dt}, \phi_{dt})$$

(una definizione di ϕ_{dt} è data alla fine del paragrafo).

Sotto questa specificazione il parametro di interesse θ_{dt} è definito come

$$\theta_{dt} = \theta_{dt}^* \left\{ 1 - (1 - \theta_{dt}^*)^{m_d} \right\} = E(p_{dt} | \theta_{dt}^*, \phi_{dt})$$

Si assume che

$$\text{logit}(\theta_{dt} | \alpha_t, \beta_t, v_d) = \alpha_t + \mathbf{x}_{dt}^T \beta_t + v_{dt} + \varphi_{dt}, \quad (\text{B.2})$$

Al fine di sfruttare la relazione spaziale tra le stime, si ipotizza che gli effetti casuali φ_{dt} siano correlati spazialmente ma indipendenti tra tassi (Mercer et al., 2015; Porter et al., 2014). Intuitivamente si ritiene che ciò che avviene in una certa area sia in relazione a ciò che avviene nelle aree vicine in misura maggiore di quanto non accada per le aree lontane. Si assume quindi che $\boldsymbol{\varphi}_t \sim ICAR(\sigma_{\varphi t})$. La distribuzione ICAR (Intrinsic Conditional AutoRegressive) è largamente utilizzata per il liscio spaziale e sfrutta la struttura di vicinato indotta dalla contiguità spaziale tra le aree.

Questa specificazione ammette la possibilità di “prendere a prestito” informazione anche dalle aree spazialmente “vicine” in modo da sfruttare la relazione eventualmente presente tra fenomeni in differenti dislocazioni geografiche, in cui ciò che si osserva in un’area è determinato, almeno in parte, da ciò che accade in una area vicina. Per esprimere formalmente la struttura della dipendenza geografica si utilizza una matrice, detta dei pesi o dei ritardi spaziali, i cui elementi riflettono l’intensità della connessione esistente tra ogni coppia di regioni. Le misure di questa intensità possono essere diverse. Nella forma più semplice ci si basa sul concetto di contiguità binaria secondo cui la struttura delle prossimità è espressa da valori 0-1. Se due unità spaziali hanno un confine in comune, di lunghezza maggiore di zero, saranno considerati contigui e saranno contrassegnati dal valore 1. Viceversa, se non sono contigui la loro accoppiata avrà valore 0. Più tecnicamente

$$\boldsymbol{\varphi}_t | \sigma_{\varphi t} \propto \exp \left\{ -\frac{1}{2\sigma_{\varphi t}^2} \boldsymbol{\varphi}_t^T \tilde{\mathbf{K}}_t^- \boldsymbol{\varphi}_t \right\} \quad (\text{B.3})$$

In cui $\tilde{\mathbf{K}}_s^-$ è l’inversa generalizzata della matrice $\mathbf{K} = \mathbf{D} - \mathbf{W}$; \mathbf{D} è una matrice diagonale che contiene il numero di connessioni per ogni area e \mathbf{W} è la matrice che riflette l’intensità della connessione esistente tra ogni coppia di regioni.

Gli effetti casuali v_{dt} sono a priori incorrelati tra le aree ma presentano una struttura di dipendenza tra i tassi. Essi sono definiti come segue:

$$v_t \sim NMV(\mathbf{0}, \boldsymbol{\Sigma}) , \quad (\text{B.4})$$

con

$$\boldsymbol{\Sigma} \sim \text{inverse.Wishart}(\lambda + 4 - 1, 2\lambda \text{diag}(\sigma_{v1}^2, \sigma_{v2}^2, \sigma_{v3}^2, \sigma_{v4}^2))$$

$$\text{e } \forall t \ \sigma_{v_t}^2 \sim \text{gamma}(.5, 1).$$

Si può dimostrare che sotto questa specificazione la distribuzione marginale degli effetti casuali v_{dt} è di tipo *variance-gamma* (si veda Bibby and Sørensen, 2003 per maggiori dettagli su questa distribuzione). L’uso di tale distribuzione presenta notevoli vantaggi rispetto alla tradizionale ipotesi di normalità degli effetti casuali (si veda Frühwirth-

Schnatter and Wagner, 2010, e Fabrizi and Trivisano, 2016): essa rappresenta un modello opportuno quando è di interesse tenere conto contemporaneamente della presenza di outliers e di effetti casuali piccoli in valore assoluto a causa, per esempio, di variabili ausiliarie con forte potere esplicativo. Inoltre, la distribuzione a priori della correlazione tra gli effetti casuali è di tipo uniforme se si sceglie $\lambda = 2$ e la distribuzione marginale della radice quadrata degli elementi sulla diagonale della matrice Σ è di tipo *half.t*, una distribuzione a priori largamente in uso per la deviazione standard di effetti casuali.

Nel seguito, per semplificare la notazione, omettiamo il pedice t ; la trattazione è quindi valida per ognuno dei quattro parametri da stimare.

Al fine di dare ulteriori dettagli sulla specificazione del modello, si consideri l'ipotesi di distribuzione Beta per ciascun stimatore da disegno dei tassi, ovvero:

$$p_d | \theta_d, \phi_d \sim \text{Beta}(\theta_d \phi_d, (1 - \theta_d) \phi_d). \quad (\text{B.5})$$

Tale ipotesi è di fatto coincidente con l'ipotesi *Betainfl* in tutti i casi in cui la stima da disegno è maggiore di zero.

$$\text{Tale assunzione implica } E(p_d | \theta_d, \phi_d) = \theta_d, \quad V(p_d | \theta_d, \phi_d) = \frac{\theta_d(1 - \theta_d)}{\phi_d + 1} \quad (\text{B.6})$$

in cui θ_d è il tasso per la popolazione del dominio d , cioè il parametro da stimare.

Riprendiamo qui la questione del lisciamento delle varianze, introdotta nel par A5 dell'Appendice A, ottenute in stima con il metodo della linearizzazione e sviluppiamo l'idea lì introdotta con riferimento alla stima di una proporzione, sotto l'ipotesi di distribuzione Beta dello stimatore stesso. Ricordiamo che tale operazione è di rilievo affinché, nella stima del modello, possano essere considerate in input stime stabili delle varianze degli stimatori di piccola area.

Si consideri che, al fine di assicurare l'identificabilità del modello, la varianza da disegno dello stimatore p_d , è assunta nota, come usuale nella stima per piccole aree. Nel caso in esame, come è evidente dalla (B.2), la varianza da disegno dello stimatore p_d è funzione del parametro obiettivo θ_d e del parametro ϕ_d che, quindi, viene assunto noto. Come già ricordato nel par. A5, in un disegno di tipo casuale semplice con reintroduzione (srs),

la varianza dello stimatore di una proporzione è pari a $V_{srs}(p_d) = \frac{\theta_d(1 - \theta_d)}{n_d^*}$.

Ora, data una certa ampiezza campionaria pari a n_d^* , un campione con disegno casuale semplice con reintroduzione è meno informativo di un disegno complesso del tipo adottato in EU-SILC, in cui vi è dipendenza tra le osservazioni. Consideriamo un

campione di numerosità inferiore ad n_d^* , diciamo $n_d^*\psi$, in cui ψ , tale che $0 < \psi \leq 1$, funge da fattore di riduzione della numerosità n_d^* , riduzione dovuta alla dipendenza tra le osservazioni. In altre parole, ψ non è altro che l'inverso del *design effect* (deff), definito nella teoria dei campioni come guadagno o perdita di precisione delle stime rispetto all'analogia stima ottenibile con un campione casuale semplice di uguale numerosità:

$$n_d^*\psi = \frac{n_d^*}{deff}.$$

Denotiamo ora con $\hat{V}_{lin}(p_d)$ la stima ottenuta per la varianza dello stimatore di proporzione, utilizzando la procedura descritto nel par A4 (Appendice 1) e con

$$V(p_d) = \frac{\theta_d(1-\theta_d)}{n_d^*\psi}$$

la varianza riferita al disegno complesso.

Si noti che
$$\frac{V_{srs}(p_d)}{V(p_d)} = \psi = \frac{1}{deff}$$

Considerando quindi le stime delle varianze $\hat{V}_{lin}(p_d)$ ottenute per il campione effettivo e lo stimatore di θ_d , possiamo specificare il seguente modello di regressione

$$\frac{p_d(1-p_d)}{\hat{V}(p_d)} = \kappa n_d^* + \varepsilon_d \tag{B.7}$$

e stimarlo sulla base delle stime ottenute per p_d , utilizzando il metodo dei minimi quadrati. Potremo allora porre $\hat{\kappa}_d = \hat{\kappa} n_d^*$ con $\hat{\kappa}$ che è la stima dei minimi quadrati del modello di regressione specificato in B.5, ed è interpretabile direttamente come misura di efficienza del piano di campionamento, nel senso che se $\hat{\kappa} < 1$ il disegno sarà meno efficiente di quello casuale semplice.

Sotto l'ipotesi che gli stimatori p_d abbiano una distribuzione di tipo Beta come specificato in B.1, il denominatore della varianza di p_d è pari a $\phi + 1$ che funge, quindi, da ampiezza campionaria sotto il disegno complesso, da cui $\phi_d + 1 = n_d^*\psi$.

B4.2 Specificazione delle distribuzioni a priori per gli iperparametri

L'analisi bayesiana di un modello prevede che a tutti i parametri incogniti sia assegnata una distribuzione a priori. Nei paragrafi precedenti, per alleggerire la presentazione, non abbiamo discusso la specificazione delle distribuzioni a priori relative agli iperparametri: coefficienti di regressione e componenti di varianza. Alla luce della complessità del modello abbiamo scelto a priori standard (dette anche “reference” in letteratura), approssimativamente non-informative.

In particolare, per ognuna delle componenti dei vettori di coefficienti di regressione sono state assunte distribuzioni normali diffuse, ossia caratterizzate da varianze molto elevate $N(0,1000)$. Benché approssimativamente uniformi, queste distribuzioni offrono il vantaggio, rispetto a distribuzioni uniformi improprie (uniformi su tutto l'asse reale) di portare ad un calcolo più agevole delle *full conditional*, un elemento importante nella costruzione degli algoritmi di tipo Markov Chain Monte Carlo che abbiamo utilizzato per lo studio di questi modelli. Per i parametri σ_{ϕ_t} , in linea con la letteratura, è stata scelta una distribuzione a priori di tipo *half.normal*(0,10)

B5. Analisi e selezione dell'informazione ausiliaria

Alcune delle 14 variabili ausiliarie descritte in Tabella B1 sono fortemente correlate tra loro, rappresentando informazione ridondante. È stato quindi necessario effettuare una loro selezione.

Il processo di selezione è avvenuto in due passi, utilizzando i dati relativi alle province.

Nel primo passo, dal gruppo iniziale di variabili ausiliarie (regressori) si è passati ad un gruppo più ristretto eliminando, tra quelle che sono fortemente correlate tra loro, le variabili ausiliarie che presentavano minore correlazione con le stime dirette dei parametri di interesse. Nella tabella B2 sono riportate le variabili selezionate e nella Tab B.4.2 le correlazioni tra le covariate e le stime dirette dei parametri di interesse:

Tabella B2 *Descrizione delle informazioni ausiliarie considerate nella costruzione de modelli dopo il primo passo di selezione.*

Etichetta	Descrizione	Fonte
f_dich	Numero dichiaranti/popolazione totale	Fiscale/anagrafe
dens	Densità di popolazione	anagrafe
dip_giov	Popolazione 0_14/popolazione 15_64	anagrafe
dip_sen	Popolazione 65 e oltre /popolazione 15_64	anagrafe
freq_stra	Popolazione straniera/popolazione totale	anagrafe
masc_stra	Numero stranieri maschi/numero stranieri femmine	anagrafe
logit_hr	Logit del head count ratio	fiscale
Logit_gini	Logit dell'indice di gini	fiscale
Logit_rich	Logit del tasso di individui con reddito superiore a 2 volte il reddito mediano nazionale	Fiscale
taddetti	Numero di addetti delle imprese su popolazione totale	ASIA
tlaureati	Numero di laureati su popolazione totale	ISTAT

Tabella B.4.2 *Coefficienti di correlazione tra le informazioni ausiliarie e le stime dirette*

		stime dirette			
		AROP	SMD	QJ	AROPE
Covariate	f_dich	-0.76	-0.61	-0.51	-0.81
	dens	0.14	0.17	-0.12	0.14
	dip_giov	0.00	0.06	-0.07	-0.02
	dip_sen	-0.45	-0.46	-0.27	-0.48
	freq_stra	-0.62	-0.45	-0.44	-0.63
	masc_stra	0.33	0.27	0.22	0.37
	logit_hr	0.73	0.54	0.46	0.78
	Logit_gini	0.54	0.50	0.33	0.60
	Logit_rich	-0.64	-0.42	-0.40	-0.68
	taddetti	-0.64	-0.46	-0.47	-0.67
	tlaureati	0.58	0.51	0.31	0.64

Nel secondo passo, un insieme di variabili ausiliarie ancora più ristretto è stato selezionato utilizzando, solo ai fini della selezione delle variabili, distribuzioni a priori per i parametri di regressione diverse da quelli ipotizzate in precedenza e stimando modelli simili a quelli descritti nelle sezioni B4.1-B4.4.

Sia β_i l' i -esima componente del vettore dei parametri di regressione; si assume per ogni i la seguente distribuzione a priori

$$\begin{aligned} \beta_i &\sim N(0, \sigma_i^2) \\ \sigma_i^2 &= 0.001(1 - \lambda_i) + c\lambda_i \\ \lambda_i &\square \text{Bernoulli}(.5) \end{aligned} \tag{B.8}$$

La media a posteriori del parametro λ_i è un indicatore dell'importanza della covariata i -esima in una scala tra 0 e 1. Il metodo di selezione delle covariate indotto dalla prior B.8 è noto in letteratura come *Stochastic search variable selection* ed è stato introdotto da George and McCulloch (1993). Si è preferito operare una scelta conservativa (nel senso di non escludere troppe variabili) ponendo $c=10$ e includendo nei modelli tutte le covariate con media a posteriori di λ_i maggiore di 0.2.

Di seguito i risultati per ognuno dei parametri di povertà di interesse.

Tabella B3 - *Media a posteriori del parametro λ_i , modello per AROP*

Covariata	Media a posteriori	Selezionata
f_dich	0.93	Si
dens	0.21	Si
dip_giov	0.11	No
dip_sen	0.24	Si
freq_stra	0.20	No
masc_stra	0.19	No
logit_hr	0.33	Si
Logit_gini	0.17	No
Logit_rich	0.17	No
taddetti	0.21	Si
tlaureati	0.11	No

Tabella B4 - *Media a posteriori del parametro λ_i , modello per SMD*

Covariata	Media a posteriori	Selezionata
f_dich	0.87	Si
dens	0.20	Si
dip_giov	0.28	Si
dip_sen	0.22	Si
freq_stra	0.20	Si
masc_stra	0.18	No
logit_hr	0.25	Si
Logit_gini	0.44	Si
Logit_rich	0.25	Si
taddetti	0.25	Si
tlaureati	0.47	Si

Tabella B5 - *Media a posteriori del parametro λ_i , modello per QJ*

Covariata	Media a posteriori	Selezionata
f_dich	0.86	Si
dens	0.13	No
dip_giov	0.41	Si
dip_sen	0.22	Si
freq_stra	0.24	Si
masc_stra	0.16	No
logit_hr	0.26	Si
Logit_gini	0.24	Si
Logit_rich	0.24	Si
taddetti	0.32	Si
tlaureati	0.19	No

Tabella B6 - *Media a posteriori del parametro λ_i , modello per AROPE*

Covariata	Media a posteriori	Selezionata
f_dich	0.87	Si
dens	0.26	Si
dip_giov	0.13	No
dip_sen	0.19	No
freq_stra	0.15	No
masc_stra	0.15	No
logit_hr	0.37	Si
Logit_gini	0.26	Si
Logit_rich	0.17	No
taddetti	0.18	No
tlaureati	0.11	No

B6. Simulazione delle distribuzioni a posteriori

Le distribuzioni a posteriori dei parametri inclusi nei modelli considerati sono state studiate utilizzando metodi Markov Chain Monte Carlo. In particolare, abbiamo utilizzato il software Rstan (The Stan Development, 2020). L'idea fondamentale di questi metodi è quella di generare campioni da catene di Markov la cui distribuzione stabile coincida con la distribuzione a posteriori del parametro oggetto di interesse.

Il campione generato dalla catena potrà quindi essere diviso in due parti: una prima parte (o "burn in") relativo alla transizione della catena verso la sua distribuzione stabile e la parte successiva alla convergenza, che è quello che può effettivamente essere utilizzato per descrivere la distribuzione a posteriori. Dal punto di vista tecnico occorre quindi approntare le diagnostiche necessarie a verificare l'avvenuta convergenza. Noi abbiamo scelto di generare catene lunghe di 8,000 valori per ciascuno dei parametri d'interesse e di eliminare le prime 4,000 come burn in. Che questo taglio fosse effettivamente sufficiente è stato accertato con gli strumenti dell'analisi visuale delle traiettorie delle catene, della generazione di catene parallele a partire da valori iniziali distanziati tra loro e calcolando la statistica di Gelman e Rubin.

B7. La selezione dei modelli

Alcune delle ipotesi alla base dei modelli presentati nelle sezioni B4.1-B4.4 sono il frutto di un'accurata procedura di selezione in cui alcune delle ipotesi centrali, sono state sottoposte alla verifica dei dati.

Sono state infatti considerate le seguenti specificazioni alternative dei modelli:

- Modello 1: effetti casuali non strutturati indipendenti tra i quattro parametri. Di fatto equivale a specificare 4 modelli indipendenti per i quattro parametri. Si può ottenere come caso particolare della specificazione illustrata nella sezione B.4.1 assumendo $\Sigma = \text{diag}(\sigma_{vd1}^2 / \tau_1, \sigma_{vd2}^2 / \tau_2, \sigma_{vd3}^2 / \tau_3, \sigma_{vd4}^2 / \tau_4)$ con $\tau_t \sim \text{gamma}(2,1)$ e $\sigma_{vdt}^2 \sim \text{gamma}(.5,1)$. Questo modello assume un tipo di shrinkage poco severo ed è detto local-global shrinkage.
- Modello 2: come il modello precedente, ma con un tipo di shrinkage di tipo global-global, ovvero uno shrinkage più severo. Si ottiene dal modello della sezione B.4.1 assumendo $\Sigma = \text{diag}(\sigma_{v1}^2 / \tau_1, \sigma_{v2}^2 / \tau_2, \sigma_{v3}^2 / \tau_3, \sigma_{v4}^2 / \tau_4)$ con $\tau_t \sim \text{gamma}(2,1)$ e $\sigma_{vt}^2 \sim \text{gamma}(.5,1)$.
- Il Modello multivariato della sezione B.4.1

La selezione tra modelli è stata basata sul criterio looic (Vehtari et al. 2017a, 2017b). Tra due modelli in competizione si sceglie quello con il looic minore. Ai fini della selezione sono stati considerati i modelli stimati sui dati delle 110 province italiane.

Di seguito sono riportati i risultati per ciascuno dei parametri di povertà di interesse:

Tabella B8 - Deviance Information criterion per i tre modelli considerati

	AROP	SMD	QJ	AROPE
Modello 1	-310.2	-380.6	-376.8	-291.9
Modello 2	-301.5	-381.0	-363.6	-280.2
Modello 3	-358.0	-398.0	-394.1	-372.2

Come si può vedere per ognuno dei quattro parametri di povertà di interesse, il looic risulta minore per il Modello 3, cioè il modello che sfrutta la relazione tra i 4 parametri da stimare.

B.8 Il benchmarking delle stime da modello

Una proprietà desiderabile per un insieme di stime per piccole aree è che siano coerenti per aggregazione. Ad esempio, le stime del numero di cittadini con una certa caratteristica (e.g. materialmente deprivati) ottenute per i distretti di una provincia devono, qualora sommate, coincidere con la stima ottenuta a livello provinciale. Allo stesso modo i totali stimati a livello provinciale devono sommare alla stima regionale.

Formalmente, se consideriamo un dominio (sottoinsieme della popolazione) d , suddiviso in G_d sotto-domini, la proprietà di aggregazione può essere scritta come

$$\sum_{g=1}^{G_d} \hat{Y}_{dg} = \hat{Y}_d$$

Dove \hat{Y}_{dg} è una stima del totale a livello del g -esimo sotto-dominio e \hat{Y}_d la relativa stima a livello di dominio d .

Questa proprietà implica che se N_{dg}, N_d rappresentano le dimensioni di sottodomini e dominio, anche le medie devono soddisfare una proprietà di coerenza per aggregazione:

$$\hat{\bar{Y}}_d = \frac{1}{N_d} \hat{Y}_d = \frac{1}{N_d} \sum_{g=1}^{G_d} \hat{Y}_{dg} = \sum_{g=1}^{G_d} \frac{N_{dg}}{N_d} \hat{\bar{Y}}_{dg}$$

dove $\hat{\bar{Y}}_{dg}$ è la stima della media a livello di sotto-dominio. Notiamo come in questo caso l'aggregazione non consiste in una somma semplice ma in una somma pesata, dove il peso di ciascun sotto-dominio consiste nel suo *share* della popolazione di dominio poiché $\sum_g N_{dg} = N_d$.

Le stime puntuali ottenute come medie delle distribuzioni a posteriori non godono di questa proprietà nel senso che:

- le stime da modello delle medie delle diverse variabili obiettivo a livello di distretto se aggregate non coincidono con la stima da modello ottenuta a livello provinciale;
- le stime da modello delle medie delle diverse variabili obiettivo a livello di provincia non coincidono con la stima da disegno a livello regionale.

Affinché le stime da modello siano coerenti per aggregazione devono essere modificate attraverso un'operazione detta di *benchmarking*. Ciò può essere realizzati in modi diversi e un certo numero di articoli nella letteratura sulle piccole aree si sono occupati di questo tema (si vedano ad esempio Datta et al. 2011, Steort et al. 2020).

Indichiamo con $\hat{\theta}_{dg}$ la stima da modello della media di popolazione nel g -esimo sotto-dominio di d ; sia $W_{dg} = \frac{N_{dg}}{N_d}$ la frazione di popolazione del sotto-dominio stesso. In generale $\sum_{g=1}^d W_{dg} \hat{\theta}_{dg} \neq \hat{\theta}_d$.

L'obiettivo è quello di cercare un nuovo set di stimatori $\tilde{\theta}_{dg}$ che soddisfino il vincolo

$$\sum_{g=1}^d W_{dg} \tilde{\theta}_{dg} = \hat{\theta}_d \quad (1)$$

La soluzione può essere ottenuta impostando un problema di ottimo vincolato. Nelle precedenti edizioni del rapporto la funzione obiettivo da minimizzare in $\tilde{\theta}_d = \{\tilde{\theta}_{dg}\}_{g=1}^d$ era stata scelta in forma quadratica

$$L(\tilde{\theta}_d, \hat{\theta}_d) = (\tilde{\theta}_d - \hat{\theta}_d)^2 \quad (2)$$

dove $\hat{\theta}_d$ è definito analogamente a $\tilde{\theta}_d$. Le stime *benchmarkate* si ottengono minimizzando (2) sotto il vincolo (1). Più nel dettaglio il vincolo non è stato applicato direttamente alla media a posteriori ma ad ogni singola estrazione $\theta_{dg}^{(r)}$, $r = 1, \dots, R$ dalla distribuzione ottenuta attraverso l'algoritmo MCMC. Procedere in questo modo permette di proiettare l'intera distribuzione a posteriori su un sotto-spazio che rispetta il vincolo (1). Si veda in questo senso Dunson and Neelan (2003). Ciò garantisce che la media a posteriori

$$\tilde{\theta}_{dg} = R^{-1} \sum_{r=1}^R \tilde{\theta}_{dg}^{(r)}$$

rispetti il vincolo (1). Inoltre, le deviazioni standard, così come gli intervalli di probabilità a posteriori ottenuti da queste distribuzioni, forniscono misure affidabili di incertezza da associare alle stime *benchmarkate*.

In questa nuova versione del rapporto proponiamo una modifica della procedura adottata in precedenza, definita nella proposta di De Nicolò et al. (2023). L'ottenimento delle stime *benchmarkate* come minimizzazione della funzione di perdita quadratica (2) si traduce in una trasformazione lineare di ciascuna estrazione dalla distribuzione non vincolata:

$$\tilde{\theta}_{dg}^{(r)} = \theta_{dg}^{(r)} + \left(\tilde{\theta}_d^{(r)} - \sum_g W_{dg} \theta_{dg}^{(r)} \right) W_{dg} \quad (3)$$

Lavorando con dei tassi per cui i valori ammissibili sono costituiti dall'intervallo (0,1) questa trasformazione non garantisce che i valori trasformati cadano al di fuori da (0,1). Come conseguenza pratica, poteva capitare che l'estremo inferiore di un intervallo di probabilità a posteriori fosse negativo. In linea teorica, non sarebbe stato possibile escludere neppure una media a posteriori (*benchmarkata*) negativa, anche se questo non è mai stato osservato in pratica.

La soluzione che proponiamo in questo rapporto consiste nel considerare una scelta diversa da (2). In particolare, consideriamo funzioni di perdita nella famiglia nota in letteratura come Bregman e a sostegno nell'intervallo (0,1). Le funzioni di perdita di Bregman generalizzano la perdita quadratica nel senso che, qualora non vincolate il loro minimo consiste nel valore medio. La funzione di perdita scelta garantisce che il suo minimo vincolato al rispetto di (1) sia compreso nell'intervallo (0,1). La funzione di perdita può essere scritta in questo modo:

$$(\tilde{\theta}_d, \hat{\theta}_d) = \sum_g W_{dg} \left\{ \tilde{\theta}_{dg} \log \left(\frac{\tilde{\theta}_{dg}}{\hat{\theta}_{dg}} \right) + (1 - \tilde{\theta}_{dg}) \log \left(\frac{1 - \tilde{\theta}_{dg}}{1 - \hat{\theta}_{dg}} \right) \right\}$$

La soluzione non esiste in forma chiusa come nel caso di (3), ma può essere facilmente ottenuta attraverso un algoritmo di ottimizzazione numerica. Come nella soluzione precedente, il vincolo non è stata applicato direttamente alla media a posteriori ma ad ogni singola estrazione $\theta_{dg}^{(r)}$, $r = 1, \dots, R$ dalla distribuzione ottenuta attraverso l'algoritmo MCMC. In questo modo si ottengono distribuzioni a posteriori le cui medie rispetto la coerenza per aggregazione che era richiesta, il cui sostegno è (0,1) e da cui possono facilmente essere ottenute misure descrittive dell'incertezza come deviazioni standard e intervalli di probabilità a posteriori.

Riferimenti bibliografici

Aaberge R., Melby I. (2005) The sensitivity of income inequality to choice of equivalence scales, *The Review of Income and Wealth*, **44**, 565-569.

Alfons A., Templ M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, **54**, 15.

Atkinson A.B. (1998) Poverty in Europe, *Blackwell Publishing*, New York.

Atkinson A.B., Rainwater L.B., Smeeding T.M. (1995) Income distribution in OECD countries, Paris: OECD.

Ayala L., Jurado A., Pérez-Mayo J. (2011) Income poverty and multidimensional deprivation: lessons form cross-regional analysis, *The Review of Income and Wealth*, **57**, 40-60.

Balog M., Thornburn D. (2007) Extreme values treatment for samples from skew income distributions, *Statistics in Transition*, **8**, 139-153.

Balduzzi R., Moirano F. (2011) La rete dei distretti sanitari in Italia (editoriale), *Quaderno di Monitor n. 8*, Agenas.

Bellentani M.D., Damiani G., Ronconi A., Catania S., Bugliari Armenio L. (2011) Report dell'indagine, in *La rete dei distretti sanitari in Italia*, *Quaderno di Monitor n. 8*, Agenas.

Banca d'Italia (2006), I bilanci delle famiglie italiane nell'anno 2004, *Supplementi al Bollettino Statistico*, Nuova serie, n.7.

Bellù L.G., Liberati P. (2005), Impacts of Policies on Poverty. The Definition of Poverty, *EASYPol*, FAO.

Bibby B.M., Sorensen M. (2003), Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes, in Rachev, S.T. (ed.) *Handbook of Heavy Tailed Distributions in Finance*, Elsevier Science B.V., 211-248.

Carbonaro G. (1985) Nota sulla scala di equivalenza, Presidenza del Consiglio dei Ministri, Commissione Indagine sulla Povertà "La Povertà in Italia", Studi di base.

Citro C.F., Michaels R.T. (1995), *Measuring poverty: a new approach*, National Academic Press, Washington D.C.

Curcetti C. (2011) Regione Emilia-Romagna, in *La rete dei distretti sanitari in Italia*, *Quaderno di Monitor n. 8*, Agenas.

Datta G.S., Ghosh M., Steorts, R., Maples J. (2011). Bayesian benchmarking with applications to small area estimation, *TEST*, **20**, 574-588.

De Nicolò S., Fabrizi E., Gardini A. (2023) Mapping poverty at multiple geographical scales, arXiv:2306.12674.

Deville J. C., Särndal C. H. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Dunson D.B. and Neelan B. (2003). Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models, *Biometrics*, **59**, 286-295.

Eurostat (2007) Comparative EU statistics on income and living conditions: issues and challenges. Proceedings of the EU-SILC conference, Helsinki 6-8 November 2006. Eurostat Methodologies and working papers

Fabrizi E., Ferrante M.R., Pacei S. (2008) Measuring sub-national income poverty using a Small-Area multivariate approach, *The Review of Income and Wealth*, **54**, 597-615.

Fabrizi E., Trivisano C. (2016) Small area estimation of the Gini concentration coefficient, *Computational Statistics & Data Analysis*; (99): 223-234.

Ferrante M.R., Pacei S. (2017) Small domain estimation of business statistics by using multivariate skew normal models. *J. R. Stat. Soc. Ser. A* 2017, 180, 1057–1088.

Fruhwirth-Schnatter S., Wagner H. (2010). Bayesian variable selection for random intercept modelling of Gaussian and non-Gaussian data. In J. Bernardo, M. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, M. West (eds.), *Bayesian Statistics*, 9, 165–200, Oxford Univ. Press.

Gelman A., Meng X., Stern H.S. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion), *Statistica Sinica*, **6**, 733-807.

George E. I., McCulloch, R. E. (1993). Variable selection via Gibbs sampling." *J. Am. Stat. Assoc.*, 85: 398-409.

Isaki C.T., Fuller W.A. (1982) Survey design under the regression superpopulation model, *Journal of the American Statistical Association*, **77**, 89-96.

ISTAT (2004) La povertà assoluta: informazioni sulla metodologia di stima, documento scaricabile da:

www.istat.it/salastampa/comunicati/non_calendario/20040503_00/poverta_mag04.pdf

ISTAT (2008) Reddito e condizioni di vita. Indagine sulle condizioni di vita. Anno 2005, scaricabile da <http://www.istat.it/>

ISTAT (2015), Il benessere equo e sostenibile delle province, in <http://www.besdelleprovince.it/>

ISTAT (2015), Il benessere equo e sostenibile in Italia, in [http://www.istat.it/it/files/2015/12/Rapporto BES 2015.pdf](http://www.istat.it/it/files/2015/12/Rapporto%20BES%202015.pdf)

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., & Clark, S. (2015). Space-time smoothing of complex survey data: small area estimation for child mortality. *The annals of applied statistics*, 9(4), 1889.

Ospina R., and Marmolejo-Ramos F. (2019), Performance of Some Estimators of Relative Variability, *Frontiers in Applied Mathematics and Statistics*, pp.- 1-20.

Porter, A. T., Holan, S. H., Wikle, C. K., & Cressie, N. (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10, 27-42.

Rao J.N.K. (2003), *Small Area Estimation*, Wiley series in Survey Mehtodology, John Wiley and Sons.

Rubin D. (1984) Bayesian justifiable and relevant frequency calculation for the applied statistician, *Annals of Statistics*, **12**, 1151-1172.

Särndal C. H., Lundström S. (2005) *Estimation in surveys with nonresponse*, *John Wiley and Sons*, New York.

Sen A. (1983) Poor, relatively speaking, *Oxford Economic Papers*, New series, **35**, 153-169.

Steorts R. C., Schmid T., Tzavidis, N. (2020). Smoothing and benchmarking for small area estimation. *International Statistical Review*, 88, 580-598.

Thomas A., O' Hara B., Ligges U., Sturz S. (2006) Making BUGS Open, *R News*, **6**, 12-17.

The Stan Development Team (2020) *Stan Modeling Language User's Guide and Reference Manual*. <https://mc-stan.org/>.

Tillé Y., Matei A. (2008) The Sampling package. Functions for drawing and calibrating samples, scaricabile da <http://cran.r-project.org/web/packages>.

Vandewalle B., Beirlant J., Christmann A., Hubert M. (2007). A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics and Data Analysis*, 51, 6252–6268.

Van Kerm, P. (2007). *Extreme incomes and the estimation of poverty and inequality indicators from EUSILC*. Working Paper Series 2007-01. Centre d'Etudes de

Populations, de Pauvrete et de Politiques Socio- Economiques International Network for Studies in Technology, Environment, Alternatives, Development.

Vehtari, A., Gelman, A., and Gabry, J. (2017a). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 27(5), 1413-1432. doi:10.1007/s11222-016-9696-4.

Vehtari, A., Gelman, A., and Gabry, J. (2017b). Pareto smoothed importance sampling. arXiv preprint: <https://arxiv.org/abs/1507.02646>

Wolter K. (2007), *Introduction to Variance Estimation*, Springer.