Capitolo 5. Povertà a livello locale

La stima per piccole aree rappresenta uno strumento molto utile quando si deve misurare la povertà e la disuguaglianza a livello regionale, ma i dati campionari sono disponibili solo a livello nazionale. In questo caso sono necessarie tecniche statistiche e metodologie economiche per utilizzare informazioni ausiliarie.

Il termine piccola area può essere riferito (Rao, 2003) sia ad aree geografiche di piccole dimensioni, sia a domini formati da sub-popolazioni definite sulla base di particolari caratteristiche demografiche o sociali.

In letteratura sono classificati come modelli per piccole aree quei modelli che utilizzano informazioni ausiliarie disponibili a livello di piccola area e a livello di singola unità campionaria (nucleo familiare o individuo).

Esiste una vasta gamma di tecniche di stima per piccole aree, e si tratta di un ambito di ricerca in continua espansione. L'adattabilità e l'efficienza di una tecnica rispetto ad un'altra, varia a seconda della specificità delle situazioni e della natura dei dati a disposizione.

I metodi di stima per piccole aree possono essere classificati secondo il tipo di inferenza in tre gruppi:

(i) metodi basati sul disegno (o campionari);

(ii) metodi assistiti da modello;

(iii) metodi basati sul modello (approccio predittivo).

Per i metodi del gruppo (i) il parametro di interesse viene stimato utilizzando i procedimenti campionari classici basati sulla distribuzione di probabilità indotta dal disegno di campionamento. Con questo metodo il parametro è pensato come una costante e gli stimatori sono corretti rispetto al disegno di campionamento applicato. La loro variabilità però, cresce al diminuire della numerosità del campione e può accadere che nessuna unità campionaria sia presente nella piccola area, impedendo così di ottenere una stima del parametro di interesse di piccola area.

Questa classe è composta solo da metodi diretti, e ne fanno parte gli stimatori di espansione, tra i quali il più utilizzato è quello di Horvitz e Thompson.

Per i metodi del gruppo (ii) l'inferenza è basata sul disegno e sul modello. L'obiettivo è quello di ottenere stimatori corretti indipendentemente dalla scelta del modello, sfruttando le informazioni derivanti dal disegno campionario.
Questa classe è formata dallo stimatore diretto di regressione e da molti altri indiretti, tra i quali gli stimatori sintetici e quelli combinati.
Per i metodi del gruppo (iii) l'aspetto rilevante è costituito dal fatto che il parametro oggetto di studio non è pensato come una costante, ma come una variabile casuale. Questo approccio prevede l'introduzione di un modello probabilistico di super-popolazione, relativo alla distribuzione del fenomeno tra le aree, da cui derivare il predittore ottimo corretto a livello di piccola area (Chiandotto, 1996). Appartengono a questa categoria i modelli di piccola area (*Small Area Models*).
Questi modelli prevedono la presenza di effetti casuali di area (*Area Level Random Effects Model*, Fay e Herriot, 1979), che vengono utilizzati quando l'informazione ausiliaria è disponibile solo a livello di area, e i *Nested Error Unit Level Regression Model* (Battese *et al*., 1988), utilizzati quando le covariate specifiche delle unità sono disponibili per ogni singolo individuo.

5.1. Modelli con effetti casuali di area

Come già anticipato, questi modelli possono essere utilizzati quando l'informazione ausiliaria esiste allo stesso livello di disaggregazione territoriale per il quale devono essere calcolati gli indici di povertà e disuguaglianza.
Questi modelli collegano i parametri di interesse alle variabili ausiliarie a livello di piccole aree, considerando gli effetti casuali indipendenti. Il modello base include gli effetti casuali specifici di ogni area. Il vettore di p variabili ausiliarie a livello di piccola area è:

$$\mathbf{x}_i = \left(x_{i,1}, x_{i,2}, ..., x_{i,p}\right) \tag{5.1}$$

I parametri di interesse θ_i (totali, medie, proporzioni, eccetera) possono essere così indicati:

$$\theta_i = x_i \beta + z_i v_i \tag{5.2}$$

dove $i=1,\ldots,m$, z_i sono costanti positive note, β è il parametro di regressione del vettore $px1$, m sono le piccole aree e v_i sono variabili casuali indipendenti e identicamente distribuite con media 0 e varianza σ_v^2. Inoltre si ipotizza che gli stimatori diretti $\hat{\theta}_i$ siano disponibili per le piccole aree, non distorti dal disegno, e che sia valido il seguente modello:

$$\hat{\theta}_i = \theta_i + e_i \qquad (5.3)$$

dove e_i sono gli errori campionari nell'area i, indipendenti, con media 0 e varianza ψ_i, questo significa che si tratta di stimatori corretti rispetto al disegno.

Combinando le equazioni (5.2) e (5.3) riportate sopra, si ottiene il seguente modello lineare ad effetti misti di Fay e Herriot (1979):

$$\hat{\theta}_i = x_i\beta + z_i v_i + e_i \qquad (5.4)$$

Esso considera gli effetti casuali di area v_i, gli errori di campionamento e_i ed assume la loro indipendenza.

Questo è un caso particolare del modello lineare misto con una struttura della covarianza diagonale, così come la maggior parte dei modelli di stima per piccole aree suggeriti in letteratura.

Gli effetti fissi di questi modelli configurano il valore medio della variabile di risposta y, mentre gli effetti casuali governano la struttura di varianza-covarianza.

Utilizzando i risultati generali del modello lineare ad effetti fissi e casuali si può determinare il predittore ottimo lineare e corretto, BLUP (*Best Linear Unbiased Predictor*) per il modello di stima per piccole aree a livello di area per θ_i:

$$\hat{\theta}_i = \gamma_i\hat{\theta}_i + (1-\gamma_i)x_i\hat{\beta} \qquad (5.5)$$

dal quale si deduce che esso è una media ponderata dello stimatore diretto $\hat{\theta}_i$ e dello stimatore sintetico di regressione $x_i\hat{\beta}$, dove $\hat{\beta}$ è lo stimatore BLUE (*Best Linear Unbiased Estimator*) di β.

Il coefficiente γ è noto come fattore di restringimento (*shrinkage factor*) e misura il rapporto tra la varianza del modello e quella totale. Nelle applicazioni pratiche il valore della componente di varianza non è noto, e quindi è necessario stimarlo. Si ottiene così uno stimatore a due stadi detto predittore empirico ottimo lineare e corretto EBLUP (*Empirical Best Linear Unbiased Predictor*).

L'applicazione di questa metodologia risulta utile quando si hanno a disposizione informazioni ausiliarie disaggregate a livello di area, ma non a livello di singolo individuo o nucleo familiare. Con questo metodo si ottengono risultati per aree relativamente piccole, in modo tale da avere un numero di unità sufficienti per stimare il modello lineare misto. Questo rappresenta una limitazione quando si vogliono calcolare le statistiche e i rispettivi errori standard ad un livello di aggregazione maggiore; infatti l'errore standard delle statistiche complesse non è additivo e pertanto non può essere calcolato come una media o somma pesata degli errori standard delle misure del più alto livello di disaggregazione.

5.2. Poverty mapping

Questa metodologia, facente parte delle metodologie di stima per piccole aree, combina le informazioni censuarie e quelle campionarie per produrre delle mappe disaggregate a livello territoriale. Queste mappe sono necessarie per descrivere la distribuzione spaziale della povertà e della disuguaglianza in un paese; non si tratta però esclusivamente di mappe, ma di database ad alta disaggregazione.

La procedura è più impegnativa rispetto al metodo EBLUP per quanto riguarda i dati che sono necessari (dati censuari a livello micro), benché non sia richiesto un abbinamento tra i dati censuari e campionari a livello di micro disaggregazioni.

L'idea di base è quella di stimare un modello di regressione lineare con le componenti della varianza a livello locale (*small area*), utilizzando le informazioni provenienti dai campioni più piccoli, le informazioni aggregate dei censimenti, e dove possibile, integrarle con altre fonti.

La variabile dipendente del modello di regressione è costituita dal reddito disponibile familiare o dal consumo. La stima della distribuzione di queste variabili può essere utilizzata per generare la distribuzione in ogni sottopopolazione censuaria, condizionata alle caratteristiche osservate nella sottopopolazione stessa.

Dalla stima della distribuzione di una variabile monetaria nei dati censuari, o in ogni sottopopolazione, può essere fatta una stima delle misure di povertà o di ineguaglianza.

Per valutare la precisione delle stime è necessario che gli errori standard di queste misure siano calcolati utilizzando le procedure appropriate che vedremo successivamente.

Di seguito un'analisi più approfondita dell'impiego e delle varie fasi di questa metodologia.

Profili dettagliati della povertà sono molto utili sia ai governi, per stabilire obiettivi geografici di politiche pubbliche e di decentralizzazione, sia per i ricercatori e per le istituzioni multilaterali ai fini del monitoraggio dell'impatto delle spese e degli investimenti pubblici in più settori (sanità, istruzione, trasporti, eccetera).

I paesi in via di sviluppo utilizzano queste mappe per decidere come ripartire le risorse tra le agenzie locali e le amministrazioni, come primo passo per raggiungere la popolazione più indigente.

Possono però insorgere alcuni problemi, poiché la maggior parte delle informazioni a disposizione si riferisce a disaggregazioni limitate; infatti le indagini campionarie dettagliate che includono misure ragionevoli di reddito o consumo, sono solitamente poco rappresentative o di dimensione troppo piccola a bassi livelli di disaggregazione per ottenere stime attendibili. Al contrario, la maggior parte dei dati censuari, o dati provenienti da altre fonti, nonostante la numerosità elevata, raccoglie poche informazioni e non sufficientemente dettagliate per calcolare indicatori monetari di povertà o disuguaglianza.

Poiché ottenere molte informazioni per grandi campioni è troppo costoso, si è cercato di risolvere questo problema combinando dati campionari e censuari, traendo vantaggio dalle informazioni dettagliate dei primi e dalla numerosità dei secondi (Elbers *et al.*, 2003).

Per utilizzare congiuntamente le informazioni provenienti da queste due fonti mediante tecniche econometriche, i dati devono avere alcuni requisiti come il riferimento allo stesso periodo temporale (o a periodi ragionevolmente simili), ed un certo numero di variabili in comune.

Possono inoltre essere combinati ai dati campionari e censuari anche dati provenienti da altre fonti; per meglio spiegare il fenomeno può essere utilizzato il GIS (*Geographic Information System*) che fornisce, in modo molto preciso, una serie di informazioni

ambientali e di caratteristiche geografiche del territorio, e può essere di ausilio per creare mappe digitali.

Il processo di *poverty mapping*, può essere scomposto in tre fasi (*stage*), che analizzeremo più in dettaglio successivamente:

- fase 0: durante la quale si stabilisce la comparabilità delle fonti dei dati, si identificano le variabili comuni e si cerca di comprendere la strategia di campionamento;
- fase 1: durante la quale si stima il modello di consumo basandosi sulle variabili comuni e sui dati campionari;
- fase 2: durante la quale si utilizzano i parametri stimati, si applicano ai dati del censimento, si simula il consumo, e si stimano le misure di povertà e di disuguaglianza.

5.2.1. Stage Zero

Lo scopo principale di questa prima fase di creazione di una mappa di povertà, è quello di identificare una serie di variabili paragonabili, presenti nel censimento e nelle indagini campionarie. Questa fase è molto importante, e nel caso in cui non venga svolta correttamente e si presentino problemi in quelle successive, dovrà essere ripetuta.

L'alto grado di confrontabilità delle variabili selezionate è il punto cruciale per l'accuratezza delle stime del benessere.

In questo stadio vengono selezionate una serie di potenziali variabili esplicative, provenienti dalle due fonti, un sottoinsieme delle quali verrà poi utilizzato e inserito nel modello di regressione per stimare le misure di benessere nel censimento.

L'obiettivo principale è quello di determinare se le variabili dell'indagine campionaria contengano effettivamente le stesse informazioni delle corrispondenti variabili censuarie. Questo poiché, anche se le domande dell'indagine e quelle del censimento sono formulate identicamente, ci possono essere sottili differenze nel modo in cui le stesse vengono poste o ordinate, generando una diversità nelle informazioni raccolte. Può anche accadere che a causa di variazioni dialettali nell'interpretazione, le variabili possano essere comparabili per alcune aree, mentre per altre no. Verificare la comparabilità delle variabili consiste nel vedere se queste sono statisticamente

distribuite in modo uguale, sia per i nuclei familiari della popolazione censuaria, che per i nuclei familiari dell'indagine campionaria. Questa procedura viene eseguita a livello nazionale e per ogni livello di disaggregazione per il quale si hanno i dati campionari rappresentativi della popolazione. La serie di variabili comuni viene inizialmente identificata con un confronto meticoloso dei questionari, studiando, se necessario, i manuali distribuiti agli intervistatori dell'indagine e del censimento.

I criteri qualitativi utilizzati per individuare le variabili candidate sono: (i) Le domande e le risposte sono formulate nello stesso modo? (ii) I criteri utilizzati nelle domande e nelle risposte sono uguali (per esempio, le domande relative all'occupazione sono state poste alle persone maggiori di una certa età, identica per entrambi i casi)? (iii) Le opzioni di risposta sono uguali? (iv) Le istruzioni degli intervistatori sono le stesse relativamente a domande identiche?

In alcuni casi quando le alternative di risposta sono diverse, si cerca, se possibile, di aggregare alcune categorie in una delle due fonti, in modo da ottenere sempre categorie confrontabili. Invece in altri casi si possono avere variabili comuni costruite combinando le informazioni provenienti da diverse domande.

Il passo successivo all'individuazione delle variabili, consiste nella verifica di ipotesi di uguaglianza nelle distribuzioni delle stesse tra i nuclei familiari censuari e quelli campionari; per verificare quest'ipotesi funzionale, per le variabili qualitative, si utilizza il test χ^2.

Il confronto statistico si basa sulle seguenti statistiche ottenute per ogni variabile sia all'interno del censimento che dell'indagine per ogni strato: la media, l'errore standard, e i valori del 1°, 5°, 10°, 25°, 50°, 75°, 90°, 95° e 99° percentile.

Si deve testare se la media di una variabile censuaria si trova all'interno dell'intervallo di confidenza al 95% costruito per la media della stessa variabile, ma ottenuta sui dati campionari. Nel caso in cui questa cada al di fuori dell'intervallo, si deve cercare di capire quale sia la fonte della discrepanza, andando a vedere i dati e verificando le definizioni della variabile mediante le istruzioni per la comparabilità. Qualora tra alcune variabili non sia possibile trovare delle affinità, è necessario escluderle dall'analisi.

Per quanto riguarda le variabili dicotomiche, si deve accertare che le loro medie non siano inferiori al 3% e maggiori del 97%, in modo tale che le variabili costruite contengano delle variazioni tra i nuclei familiari; solitamente quelle con una bassa

variabilità tra i nuclei familiari, generano osservazioni con un'elevata influenza nella prima fase della regressione.

Alla fine si deve fare un confronto trasversale tra gli strati per testare l'uniformità delle variabili; quelle che vengono incluse nella prima fase di regressione, sono solitamente confrontabili in tutti gli strati.

5.2.2. Stage One: IL MODELLO DI REGRESSIONE

In questa seconda fase si procede a modellizzare la spesa per consumi pro-capite, al livello geografico più basso per il quale l'indagine è rappresentativa.

Si deve creare un accurato modello empirico della trasformazione logaritmica di y_{ch}, la spesa pro-capite del nucleo familiare h nel cluster campionario c; in questa fase devono essere considerate le differenze geografiche dei livelli dei prezzi.

Di seguito un'approssimazione lineare della distribuzione condizionata di y_{ch}:

$$\ln y_{ch} = E\left[\ln y_{ch} \left| x_{ch}{}^{T} \right. \right] + u_{ch} = x_{ch}{}^{T}\beta + u_{ch} \qquad (5.6)$$

dove $x_{ch}{}^{T}$ è il vettore delle covariate e u_{ch} è il vettore dei disturbi $u \approx F(0,\Sigma)$.

Utilizzando un'approssimazione lineare del valore atteso condizionato, viene modellato il logaritmo della spesa pro-capite per consumi, riferita ad ogni nucleo familiare:

$$\ln y_{ch} = x_{ch}\beta + u_{ch} \qquad (5.7)$$

Questo modello è stimato attraverso i GLS (*Generalized Least Squares*) utilizzando i dati campionari.

Per ottenere le stime GLS si deve stimare Σ, la matrice di varianza-covarianza associata all'errore. Infatti precedenti esperienze con le analisi campionarie (Elbers *et al.*, 2002) hanno dimostrato che per far sì che il modello sia propriamente specificato, questo deve avere una struttura dell'errore complessa, per tener conto sia di un'eventuale correlazione all'interno dei cluster (autocorrelazione spaziale), che dell'eteroschedasticità e di una distribuzione non-normale. Il termine di disturbo può essere rappresentato così:

$$u_{ch} = \eta_c + \varepsilon_{ch} \qquad (5.8)$$

L'errore complessivo del nucleo familiare h all'interno del cluster c, può essere scomposto in due componenti: η_c è la componente relativa al cluster, e ε_{ch} è la

componente dell'errore relativa al nucleo familiare, indipendenti l'una dall'altra e incorrelate con la natura delle variabili esplicative.

η_c riflette la parte di errore dovuta ad alcune caratteristiche della locazione, comuni a tutti i nuclei familiari.

La componente relativa ai nuclei familiari, ε_{ch}, riflette alcune caratteristiche dei nuclei stessi che non sono correlate con l'effetto corrispondente alla locazione.

Dal momento in cui l'effetto relativo al cluster può ridurre notevolmente la precisione delle stime delle misure di benessere, è necessario introdurre alcune esplicative nella serie di covariate che spieghino la variazione della spesa per consumi dovuta alla locazione. Questo si ottiene introducendo nel modello le medie di tutte le covariate calcolate su tutti i nuclei familiari del censimento, nelle sue *Enumeration Area* (EA) (che solitamente corrispondono ai cluster nel campione).

Durante questa fase iniziale dello *stage one*, si deve verificare, qualora siano stati impiegati, se è necessario o no utilizzare i pesi.

Il test di Hausman considera, come ipotesi di base, che le regressioni siano omogenee tra gli strati, cosicché entrambi gli stimatori sia quello ottenuto con una procedura pesata, che quello ottenuto con una non pesata, siano non distorti (la differenza tra i due ha valore atteso pari a zero). Se l'eterogeneità e l'effetto dovuto al disegno campionario sono importanti, i due valori attesi differiranno.

Per applicare il test viene stimato un modello di regressione ausiliario, che contiene tutti gli usuali regressori più una serie di termini di interazione, ottenuti combinando ogni regressore con i pesi dei nuclei familiari (*expansion factor*). Viene poi verificata l'ipotesi che i parametri di questi regressori di interazione siano congiuntamente uguali a zero, mediante un test F.

Successivamente si procede con un test per verificare la normalità della distribuzione delle due componenti dei residui; questo è molto importante per la conoscenza della distribuzione nella quale si decide di rappresentare gli errori stimati per la fase finale della stima. Questa verifica è basata sui residui $\hat{\eta}_c$ e sui residui standardizzati $e_{ch}^{*} = \frac{e_{ch}}{\hat{\sigma}_{\varepsilon,ch}} - \left[\frac{1}{H} \sum_{ch} \frac{e_{ch}}{\hat{\sigma}_{\varepsilon,ch}^{2}} \right]$, dove H è il numero di nuclei familiari dell'indagine. Il secondo termine e^{*}_{ch} non deve essere inserito qualora le regressioni di questo *stage* non siano pesate.

Per stimare $\sum$ si procede come segue. Il modello (5.7) viene per primo stimato con gli OLS (*Ordinary Least Squares*) (pesato con i pesi dell'indagine campionaria, qualora sia necessario utilizzarli).

I residui di questa regressione vengono utilizzati come stima del disturbo complessivo ($\hat{u}_{ch}$).

Si scinde questo nelle due componenti:

$$\hat{u}_{ch} = \hat{\eta}_c + e_{ch} \tag{5.9}$$

Le componenti stimate relative ai clusters, $\hat{\eta}_c$, sono le medie all'interno dei cluster dei residui complessivi. Le componenti relative ai nuclei familiari, e_{ch}, sono i residui complessivi al netto della componente della locazione.

Si deve inoltre stimare $\hat{\sigma}_\eta^{\,2}$, ovvero la varianza di η_c e $\hat{V}\left(\sigma_\eta^{\,2}\right)$, cioè la varianza di $\sigma_\eta^{\,2}$.

Tenendo conto dell'eteroschedasticità nella componente relativa ai nuclei familiari, modelliamo $e_{ch}^{\,2}$ utilizzando una serie di variabili selezionate per spiegare nel modo migliore la sua variazione. Le componenti, z_{ch}, che spiegano al meglio le variazioni di $e_{ch}^{\,2}$, sono scelte al di fuori delle potenziali esplicative.

Si utilizza un modello logistico per la stima della varianza di ε_{ch} condizionata alle z_{ch}, ponendo come limiti della stima, zero ed un massimo, *A*, uguale a $1{,}05 * \max\left\{e_{ch}^{\,2}\right\}$:

$$\ln\left[\frac{e_{ch}^{\,2}}{A - e_{ch}^{\,2}}\right] = z^T{}_{ch}\hat{\alpha} + r_{ch} \tag{5.10}$$

Ponendo $\exp\left\{z^T{}_{ch}\hat{\alpha}\right\} = B$ ed utilizzando il metodo delta, la varianza di ε_{ch} viene stimata così:

$$\sigma_{\varepsilon,ch}^{\,2} = \left[\frac{AB}{1+B}\right] + \frac{1}{2}\operatorname{var}(r)\left[\frac{AB(1-B)}{(1+B)^3}\right] \tag{5.11}$$

La varianza di η_c viene invece stimata con metodi non parametrici, tenendo però sempre conto dell'eteroschedasticità di ε_{ch}. Queste stime degli errori vengono utilizzate per produrre due matrici quadrate di dimensione *H*. La prima è una matrice a blocchi, dove ogni blocco corrisponde a un cluster e gli elementi che si trovano all'interno di ogni blocco sono le $\hat{\sigma}_\eta^{\,2}$. La seconda è una matrice diagonale, i cui elementi sono

relativi ai nuclei familiari, e cioè sono $\hat{\sigma}_{\varepsilon,ch}{}^2$. La somma di queste due matrici dà $\hat{\Sigma}$, cioè la matrice di varianza-covarianza del residuo complessivo.

Nel caso in cui il campione non sia auto-pesato, le regressioni precedenti sono stimate utilizzando i pesi delle probabilità campionarie, che sono date dall'inverso delle probabilità di inclusione all'interno del campione.

Una volta calcolata $\hat{\Sigma}$, il modello originale può essere stimato con i GLS. Questo metodo di stima produce una serie di stime di $\hat{\beta}_{GLS}$, cioè i coefficienti di regressione relativi all'equazione (5.7). L'output dei GLS include anche le matrici di varianza-covarianza, associate alle stime, $\hat{V}\left(\hat{\beta}_{GLS}\right)$. Oltre a queste stime, nel secondo *stage* vengono utilizzate anche $\hat{\alpha}$, $V(\hat{\alpha})$, $\hat{\sigma}_{\eta}{}^2$ e $\hat{V}\left(\sigma_{\eta}{}^2\right)$.

Per ogni modello (uno per ogni strato) la significatività dell'effetto del cluster può essere testata con il test per gli effetti casuali di Breusch and Pagan (1980), che verifica l'ipotesi che $V(\eta_c) = 0$.

Si può inoltre calcolare qual è la parte di variabilità della componente relativa al cluster rispetto alla varianza del residuo totale, utilizzando questo rapporto: $Rho = \dfrac{\sigma_{\eta}{}^2}{\sigma_{u}{}^2}$.

5.2.3. Stage Two: SIMULAZIONE

Nel secondo *stage* dell'analisi, vengono applicati i parametri stimati nel primo *stage* ai dati censuari. Attraverso questi si ottengono le stime del log della spesa per consumi pro-capite per ogni nucleo familiare del censimento, e si simulano i corrispondenti disturbi.

Solitamente vengono effettuate 100 simulazioni, dove per ogni simulazione, *r,* si estrae dalla corrispondente distribuzione un vettore di parametri ottenuti nel primo *stage*. Vengono così estratti vettori di parametri α e β ($\tilde{\beta}^r$ e $\tilde{\alpha}^r$) dalle distribuzioni normali multivariate e le relative matrici di varianza-covarianza $\beta \approx N\left(\hat{\beta}_{GLS}, V\left(\hat{\beta}_{GLS}\right)\right)$ e $\alpha \approx N\left(\hat{\alpha}, V(\hat{\alpha})\right)$. Per quanto riguarda la simulazione di η_c, sono necessari due passi; questo perché anche la varianza di η_c è stimata con un errore. Il primo passo consiste

nell'ottenere la varianza simulata di η_c, cioè la varianza della componente di errore relativa ai clusters (σ_η^2). Elbers *et al.* (2003) hanno proposto di estrarla da una distribuzione Gamma con media pari a $\hat{\sigma}_\eta^2$ e varianza $\hat{V}\left(\sigma_n^2\right)$. Il secondo passo consiste nell'estrazione casuale di η_c, e se l'ipotesi di non-normalità viene respinta si assume che $\eta_c \approx N\left(0, \sigma_c^2\right)$.

Il processo per simulare ε_{ch} richiede l'utilizzo dei risultati della stima dell'equazione (5.10). Combinando il coefficiente α con i dati del censimento, per ogni nucleo familiare censito si stima la varianza definita nell'equazione (5.11). Infine si ottiene ε_{ch} che, sempre nell'ipotesi di normalità di distribuzione, viene estratto da una $N\left(0, \sigma_{ch}^2\right)$.

Nel caso in cui, invece, si sia accertata la non-normalità della distribuzione di entrambi, η_c e ε_{ch}, per generarli, si sceglie una distribuzione t-Student con un numero di gradi di libertà tali da consentire che la curtosi di queste componenti sia il più possibile vicino a quella dei residui ottenuti nel primo *stage* ($\hat{\eta}_c$ o e_{ch}). Si potrebbe altrimenti evitare di fare qualsiasi assunzione sulla forma specifica della distribuzione dei disturbi, estraendoli direttamente dai residui stimati: per ogni cluster i residui estratti sono $\tilde{\eta}_c$ e per ogni nucleo familiare $\tilde{\varepsilon}_{ch}$.

Giunti a questo punto, e avendo a disposizione tutte le componenti dell'equazione (5.7), si procede con il calcolo della spesa per consumi pro-capite per ogni nucleo familiare, $\hat{y}_{ch}^r$, basata sulla somma di $x'_{ch}\tilde{\beta}^r$ e il termine di disturbo (η_c^r e ε_{ch}^r), utilizzando il metodo *bootstrap*:

$$\ln \hat{y}_{ch}^r = \exp\left(x_{ch}'\tilde{\beta}^r + \tilde{\eta}_c^r + \tilde{\varepsilon}_{ch}^r\right) \tag{5.12}$$

Alla fine, il set completo di spese pro-capite simulate, $\hat{y}_{ch}^r$, è utilizzato per calcolare il valore atteso di ogni misura di povertà considerata.

Questa procedura viene ripetuta per 100 volte, estraendo nuovi $\tilde{\alpha}^r$, $\tilde{\beta}^r$, $\left(\tilde{\sigma}_\eta^2\right)^r$ e i termini di disturbo per ogni simulazione. La media delle 100 simulazioni fornisce la stima puntuale delle misure di benessere o di povertà desiderate, mentre la deviazione standard fornisce l'errore standard della stima puntuale.

Il metodo utilizzato per la stima di y_{ch} viene definito *bootstrap*, e consiste in una serie di procedure statistiche che utilizzano numeri casuali generati dai computer per simulare la distribuzione di uno stimatore.

Nel caso delle mappe di povertà vengono costruiti una serie di valori stimati:

$$y_{ij}^{\ b} = x_{ij}\beta^b + h_i^b + e_{ij}^{\ b} \qquad b=1,...B \tag{5.13}$$

in modo tale da tener conto della variabilità delle stime.

$\hat{\beta}$ è uno stimatore corretto di β, con varianza V_β. Possiamo così estrarre indipendentemente l'una dall'altra ogni β^b da una distribuzione Normale multivariata, con media $\hat{\beta}$ e matrice di varianza-covarianza V_β. Gli effetti causati dal cluster h_i^b sono presi da una distribuzione empirica di h_i. Per tener conto dell'eteroschedasticità nei residui relativi ai nuclei familiari, per prima cosa si estrae α^b da una distribuzione Normale multivariata, con media $\hat{\alpha}$ e matrice di varianza-covarianza V_α, si combina questa con z_{ij} per ottenere una varianza stimata e si utilizza il risultato per aggiustare l'effetto a livello di nucleo familiare.

$$e_{ij}^{\ b} = e_{ij}^{\ *b} \times \sigma_{e,ij}^{\ b} \tag{5.14}$$

dove $e_{ij}^{\ *b}$ rappresenta un'estrazione casuale dalla distribuzione empirica di $e_{ij}^{\ *}$.

Ogni serie completa di valori *bootstrap* $y_{ij}^{\ b}$, per un fissato valore di *b*, fornisce una serie di stime per piccole aree. Nel caso di stime delle misure di povertà si pone in forma esponenziale ogni *y* per ottenere le stime della spesa $E_{ij} = \exp\left(y_{ij}\right)$, e poi si applica l'equazione (7). La media e la deviazione standard di una particolare stima per piccole aree, attraverso tutti i valori di *b*, porta ad una stima puntuale e agli errori standard relativi a quell'area.

Analizzando la precisione degli stimatori, possiamo affermare che la differenza tra l'indice *W* (indice di povertà) e $\tilde{\mu}$, lo stimatore del valore atteso di *W*, a livello locale, può essere scritto come:

$$W - \tilde{\mu} = \left(W - \mu\right) + \left(\mu - \hat{\mu}\right) + \left(\hat{\mu} - \tilde{\mu}\right) \tag{5.15}$$

In questo caso l'errore di previsione ha tre componenti: la prima dovuta alla presenza di un termine di disturbo nel modello descritto nel primo *stage*, che fa sì che la spesa per consumi attesa di un nucleo familiare differisca da quella effettiva (errore

idiosincratico); la seconda causata dal fatto che i parametri del modello del primo *stage* sono stimati (errore da modello); la terza legata all'inesattezza del metodo utilizzato per stimare $\hat{\mu}$ (errori di stima). Quest'ultima componente viene considerata da Elbers *et al.* (2003) trascurabile.

La numerosità della popolazione in un cluster non influisce sull'errore dovuto al modello, mentre l'errore idiosincratico aumenta al diminuire del numero di nuclei familiari nella popolazione di riferimento.

Un elemento importante di questa seconda fase, che richiede una certa attenzione, è l'errore standard, il quale riflette l'incertezza della stima.

Una rudimentale regola consiste nel considerare due volte l'errore standard, su ogni lato della stima puntuale, per rappresentare gli estremi dell'intervallo all'interno del quale ci aspettiamo che cada il vero valore.

Quando due o più stime per piccole aree vengono comparate, per esempio quando si decide in merito alla priorità di un'area rispetto ad un'altra per fornire aiuti, l'errore standard informa su quanto ogni stima sia accurata e se le differenze osservate nelle stime siano indicative di reali divari tra le aree.

La dimensione dell'errore standard dipende da una serie di fattori. Quanto più bassa è la rappresentatività del modello ($y_{ij} = x_{ij}\beta + h_i + e_{ij}$), in termini di un piccolo R^2 o grandi σ_h^2 o σ_e^2, maggiori saranno la variabilità non spiegata dalle variabili di interesse e l'errore standard delle stime per piccole aree. La dimensione della popolazione, sia in termini di numero di nuclei familiari che di numero di cluster in ogni area, è un altro importante fattore, ovvero all'aumentare della numerosità, l'errore standard diminuirà. Anche la dimensione campionaria utilizzata per costruire il modello è molto rilevante. Il metodo *bootstrap* incorpora la variabilità delle stime dei coefficienti di regressione $\hat{\alpha}$ e $\hat{\beta}$ e, se il campione è piccolo, queste saranno molto incerte e gli errori standard elevati. Questo problema è influenzato anche dal numero di variabili esplicative incluse nelle informazioni ausiliarie, *x* e *z*: maggiore è il loro numero e maggiore è l'instabilità dei coefficienti di regressione. Possiamo sempre aumentare l'apparente potere esplicativo del modello (con il conseguente aumento dell'R^2) incrementando il numero delle variabili esplicative, ma in realtà aumenterebbe l'incertezza delle stime con una conseguente complessiva perdita di precisione. Ci può anche essere un'elevata

variabilità degli errori standard a causa della metodologia *bootstrap*, che utilizza un campione finito di stime per approssimare la distribuzione degli stimatori. Questo può essere diminuito, con una notevole dispersione di tempo, aumentando il numero di simulazioni *bootstrap B*. Infine, l'integrità delle stime e degli errori standard dipende dalla correttezza del modello utilizzato, sia quello riferito alla popolazione che quello riferito al campione. Questo è strettamente collegato ad una buona combinazione di dati censuari e campionari per fornire valide informazioni ausiliarie.

Si devono inoltre evitare le relazioni spurie o artefatte che statisticamente sembrano essere vere nel campione, ma che in realtà non appartengono alla popolazione. Queste sono dovute ad un'errata scelta delle variabili comuni o ad un'eccessiva numerosità. Tale situazione può portare a stime con un basso ma non veritiero errore standard. Per questa ragione è molto rilevante compiere un ultimo passo nella creazione di mappe di povertà, ovvero la verifica sul campo.

5.3 Metodo EB per la stima di misure di povertà tradizionali e Fuzzy per piccole aree[1]

Questo paragrafo si propone di confrontare risultati di analisi basate su differenti metodi di stima per piccole aree, compresa una nuova metodologia volta a ottenere i migliori stimatori empirici di parametri di dominio lineari e non lineari usando modelli di regressione a livello di unità. Tale metodo si propone inoltre di risolvere problemi computazionali dovuti a popolazioni numerose o a misure di povertà più complesse. Lo scopo è quello di stimare l'indice tradizionale di povertà (head count ratio HCR) e gli indicatori fuzzy monetario (FM) e non monetario (FS) come parametri non lineari.

Il metodo proposto è basato su una versione modificata dell' *Empirical Best (EB) prediction* proposto da Molina e Rao (2009) ed è applicato per la stima degli indicatori HCR, FM e FS per le province e i comuni della Toscana.

[1] Questo paragrafo, redatto dalla Dott.ssa Caterina Ferretti, costituisce un **approfondimento**.

5.3.1 Indicatori fuzzy per piccole aree

Sia $U = \{y_1,...,y_N\}$ una popolazione di dimensione *N*, E_i una variabile di benessere (per esempio il reddito equivalente) per l'individuo *i*, $F_{(M),i}$ la funzione di distribuzione di E_i e $L_{(M),i}$ il valore della curva di Lorenz relativa a E_i. L'indicatore fuzzy monetario per l'individuo *i* (FM_i) è definito, seguendo l'approccio IFR (*Integrated Fuzzy and Relative Approach*, Betti *et al*. 2006), come combinazione dell'indicatore $(1 - F_{(M),i})$ proposto da Cheli e Lemmi (1995) e dell'indicatore $(1 - L_{(M),i})$ proposto da Betti and Verma (1999). Formalmente:

$$FM_i = \left(1 - F_{(M),i}\right)^{\alpha-1}\left(1 - L_{(M),i}\right) = \left\{\frac{1}{N-1}\sum_{j=1}^{N} \mathrm{I}\{E_j > E_i\}\right\}^{\alpha-1} \left\{\frac{\sum_{j=1}^{N} E_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{N} E_j}\right\} \qquad (5.16)$$

dove $\mathrm{I}\{E_j > x\} = 1$ se $E_j > x$, 0 altrimenti, $(1 - F_{(M),i})$ è la proporzione di individui meno poveri rispetto alla persona considerata e $(1 - L_{(M),i})$ è la quota di reddito totale equivalente ricevuta da tutti gli individui meno poveri rispetto alla persona considerata.
Per l'intera popolazione l'indice di povertà definito sopra è dato da:

$$FM = \frac{1}{N}\sum_{i=1}^{N} FM_i \qquad (5.17)$$

Per ciascun dominio *d* (*d* = 1, …,*D*) si definisce l'indice fuzzy monetario come:

$$FM_d = \frac{1}{N_d}\sum_{i=1}^{N_d} FM_i \qquad (5.18)$$

Dato un campione casuale di dimensione $n < N$ estratto dalla popolazione, $s \subseteq U, s = \{E_1,...,E_n\}$, lo stimatore diretto di FM_i è dato da:

$$\hat{FM}_i^{DIR} = \left(1 - F_{(M),i}\right)^{\alpha-1}\left(1 - L_{(M),i}\right) = \left\{\frac{\sum_{j=1}^{n} w_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{n} w_j}\right\}^{\alpha-1} \left\{\frac{\sum_{j=1}^{n} w_j E_j \mathrm{I}\{E_j > E_i\}}{\sum_{j=1}^{n} w_j E_j}\right\} \qquad (5.19)$$

dove w_j è il peso campionario per l'individuo *j*. L'indice complessivo per l'intero campione è dato da:

$$\hat{FM}^{DIR} = \frac{\sum_{i=1}^{n} w_i \hat{FM}_i^{DIR}}{\sum_{i=1}^{n} w_i} \tag{5.20}$$

Analogamente, per un dominio *d* possiamo definire:

$$\hat{FM}_d^{DIR} = \frac{\sum_{i=1}^{n_d} w_i \hat{FM}_i^{DIR}}{\sum_{i=1}^{n_d} w_i} \tag{5.21}$$

Seguendo l'approccio IFR, data una popolazione $U = \{s_1,...,s_N\}$, può essere definito anche l'indice fuzzy supplementare:

$$FS_i = \left(1 - F_{(s),i}\right)^{\alpha-1}\left(1 - L_{(s),i}\right) = \left\{\frac{1}{N-1}\sum_{j=1}^{N} \mathrm{I}\{s_j > s_i\}\right\}^{\alpha-1}\left\{\frac{1}{N-1}\sum_{j=1}^{N} s_j \mathrm{I}\{s_j > s_i\}\right\} \tag{5.22}$$

dove $\mathrm{I}\{s_j > x\} = 1$ se $s_j > x$, 0 altrimenti. $(1 - F_{(S),i})$ è la proporzione di individui con un livello di privazione inferiore rispetto alla persona considerata, $F_{(S),i}$ è il valore della funzione di distribuzione dello score valutata per l'individuo *i*, $(1 - L_{(S),i})$ è la quota dello score di totale mancanza di privazione relativa a tutti gli individui meno deprivati che la persona considerata e $L_{(S),i}$ è il valore della curva di Lorenz dello score per l'individuo *i*.

Per l'intera popolazione l'indice supplementare definito sopra è dato da:

$$FS = \frac{1}{N}\sum_{i=1}^{N} FS_i \tag{5.23}$$

Per ciascun dominio *d* (*d* = 1, …,*D*) definiamo l'indice fuzzy non monetario come:

$$FS_d = \frac{1}{N_d}\sum_{i=1}^{N_d} FS_i \tag{5.24}$$

Dato un campione casuale di dimensione $n < N$ estratto dalla popolazione, $s \subseteq U, s = \{s_1,...,s_n\}$, lo stimatore diretto di FS_i è espresso da:

$$\hat{FS}_i^{DIR} = \left(1 - F_{(S),i}\right)^{\alpha-1}\left(1 - L_{(S),i}\right) = \left\{\frac{\sum_{j=1}^{n} w_j \mathrm{I}\{s_j > s_i\}}{\sum_{j=1}^{n} w_j}\right\}^{\alpha-1}\left\{\frac{\sum_{j=1}^{n} w_j s_j \mathrm{I}\{s_j > s_i\}}{\sum_{j=1}^{n} w_j s_j}\right\} \tag{5.25}$$

dove w_j è il peso campionario per l'individuo *j*. Per l'intero campione si ha:

$$\hat{FS}^{DIR} = \frac{\sum_{i=1}^{n} w_i \hat{FS}_i^{DIR}}{\sum_{i=1}^{n} w_i} \tag{5.26}$$

Analogamente, per un dominio *d* definiamo:

$$\hat{FS}_d^{DIR} = \frac{\sum_{i=1}^{n_d} w_i \hat{FS}_i^{DIR}}{\sum_{i=1}^{n_d} w_i} \tag{5.27}$$

In tutte le precedenti formule, il parametro α è stimato in modo tale che gli indicatori FM e FS siano uguali all'indice tradizionale di povertà calcolato utilizzando la linea di povertà ufficiale (60% della mediana del reddito equivalente).

5.3.2 Il miglior stimatore empirico

Consideriamo un vettore casuale $\mathbf{y}$ contenente i valori di una variabile casuale per le unità di una popolazione finita tale che $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')$ dove $\mathbf{y}_s$ è il sub-vettore degli elementi campionati e $\mathbf{y}_r$ il sub-vettore degli elementi non campionati. L'obiettivo è predire il valor di una funzione misurabile reale $\delta = h(\mathbf{y})$ del vettore casuale $\mathbf{y}$ usando i dati campionati $\mathbf{y}_s$. Il miglior stimatore (BP) di δ è la funzione di $\mathbf{y}_s$ che minimizza l'errore quadratico medio dello stimatore $\hat{\delta}$. Formalmente:

$$\hat{\delta}^B = \delta^0 = E_{y_r}(\delta \mid \mathbf{y}_s) \tag{5.28}$$

dove il valore atteso dipende dalla distribuzione condizionata di $\mathbf{y}_r$ e il risultato è una funzione dei dati campionati $\mathbf{y}_s$.

Generalmente, $\hat{\delta}^B$ dipende da un vettore di parametri non noti $\boldsymbol{\theta}$ che può essere sostituito con un opportuno stimatore, ottenendo così un BP empirico di δ.

E' interessante notare che, quando $\mathbf{y}$ segue una distribuzione Normale con vettore medio $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ per una matrice nota $\mathbf{X}$, matrice di covarianza positiva $\mathbf{V}$, e la quantità da predire, δ, è una funzione lineare di $\mathbf{y}$, allora il BP di δ è uguale al BLUP di δ.

5.3.3 Il miglior stimatore empirico per misure fuzzy

Consideriamo l'indicatore monetario fuzzy FM dato dalla formula (5.18). Il BP di FM_d è definito come:

$$\hat{FM}_d^B = E_{\mathbf{y}_r}(FM_d \mid \mathbf{y}_s) \tag{5.29}$$

Per ottenere il BP di FM_d, è necessario esprimere FM_d in termini di un vettore $\mathbf{y}_d$, per il quale la distribuzione condizionata del vettore non campionato $\mathbf{y}_{dr}$ dato i dati campionati $\mathbf{y}_{ds}$ sia nota. La distribuzione della variabile E_{di} è raramente Normale, tuttavia, più volte è possibile trovare una trasformazione di E_{di} la cui distribuzione è approssimativamente Normale. Supponiamo che esista una trasformazione univoca $Y_{di} = T(E_{di})$ della variabile E_{di}, che segue una distribuzione Normale, $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$. Sia $\mathbf{y}_d = (\mathbf{y}_{ds}', \mathbf{y}_{dr}')'$ il valore delle variabili trasformate Y_{di} rispettivamente per il campione e per le unità non campionate nel dominio *d*. Possiamo quindi definire FM_{di} come:

$$FM_{di} = \left\{\frac{1}{N-1}\sum_{j=1}^{N} \mathrm{I}\left\{T^{-1}(Y_j) > T^{-1}(Y_{di})\right\}\right\}^{\alpha-1} \left\{\frac{\sum_{j=1}^{N} T^{-1}(Y_j)\mathrm{I}\left\{T^{-1}(Y_j) > T^{-1}(Y_{di})\right\}}{\sum_{j=1}^{N} T^{-1}(Y_j)}\right\} =: h_\alpha(Y_{di}) \tag{5.30}$$

Allora, $FM_d = \frac{1}{N_d}\sum_{i=1}^{N_d} FM_{di}$ è una funzione non lineare di $\mathbf{y}$.

Attraverso la decomposizione di FM_d in termini di elementi campionati e non campionati abbiamo:

$$FM_d = \frac{1}{N_d}\left(\sum_{i\in s_d} FM_{di} + \sum_{i\in r_d} FM_{di}\right) \tag{5.31}$$

Allora, il BP di FM_d diventa:

$$\hat{FM}_d^B = \frac{1}{N_d}\left(\sum_{i\in s_d} FM_{di} + \sum_{i\in r_d} \hat{FM}_{di}^B\right) \tag{5.32}$$

dove $\hat{FM}_{di}^B$ è il BP di $FM_{di} = h_\alpha(Y_{di})$ dato da:

$$\hat{FM}_{di}^B = E_{\mathbf{y}_r}(h_\alpha(Y_{di}) \mid \mathbf{y}_s) = \int_R h_\alpha(y) f_{Y_{di}}(y \mid \mathbf{y}_s)dy, \quad i \in r_d \tag{5.33}$$

dove $f_{Y_{di}}(y \mid \mathbf{y}_s)$ è la distribuzione condizionata di Y_{di} dato $\mathbf{y}_s$. Data la complessità della funzione $h_\alpha(y)$, non è possibile ottenere un'esplicita espressione del valore atteso

(5.33). Comunque, poiché $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_r')$ è normalmente distribuito con vettore medio $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_r \end{pmatrix}$ e matrice di covarianza $\mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}$, la distribuzione di $\mathbf{y}_r \mid \mathbf{y}_s$ è data da:

$$\mathbf{y}_r \mid \mathbf{y}_s \sim N(\boldsymbol{\mu}_{r|s}, \mathbf{V}_{r|s}) \tag{5.34}$$

dove $\boldsymbol{\mu}_{r|s} = \boldsymbol{\mu}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s)$ e $\mathbf{V}_{r|s} = \mathbf{V}_r - \mathbf{V}_{rs}\mathbf{V}_s^{-1}\mathbf{V}_{rs}$.

E' possibile così approssimare il valore atteso (5.33) attraverso simulazioni Monte Carlo. Dalla (5.34) è generato un numero elevato L di vettori $\mathbf{y}_r$ e il vettore $\mathbf{y}_r^{(l)}$ ottenuto alla l-ma replicazione è unito al vettore $\mathbf{y}_s$ in modo tale da formare il vettore della popolazione $\mathbf{y}^{(l)} = (\mathbf{y}_s', (\mathbf{y}_r^{(l)})')'$. Usando gli elementi di $\mathbf{y}^{(l)}$ per l'area d, sono calcolati i parametri di interesse di piccola area $\delta_d^{(l)} = h(\mathbf{y}_d^{(l)})$. Un'approssimazione Monte Carlo del BP di Y_{di} è data da:

$$\hat{FM}_{di}^{B} \approx \frac{1}{L}\sum_{l=1}^{L} h_\alpha(Y_{di}^{(l)}), \quad i \in r_d \tag{5.35}$$

Generalmente, $f_{Y_{di}}(y \mid \mathbf{y}_s)$ dipende da un vettore non noto di parametri precedentemente stimati usando i metodi di massima verosimiglianza o massima verosimiglianza ristretta ottenendo così il miglior stimatore empirico (EBP).

5.3.4 Metodo EB per un modello ad effetti nidificati

Un possibile modello per gli elementi del vettore della popolazione $\mathbf{y}$ che può essere usato per calcolare l'EBP è il modello di regressione ad errori nidificati (Battese, Harter, Fuller, 1988) che definisce una relazione lineare, per tutte le aree, tra le variabili trasformate Y_{di} e i vettori $\mathbf{x}_{di}$ delle p variabili esplicative e include un effetto specifico di area u_d ed errori campionari e_{di}.Formalmente:

$$\begin{aligned} &Y_{di} = \mathbf{x}_{di}'\beta + u_d + e_{di}, \qquad j = 1,...,N_d, \quad d = 1,...D, \\ &u_d \sim iid\, N(0, \sigma_u^2), \quad e_{di} \sim iid\ N(0, \sigma_e^2) \end{aligned} \tag{5.36}$$

I vettori $\mathbf{y}_d = \underset{1 \le i \le N_d}{\text{col}}\,(Y_{di})$ sono indipendenti con $\mathbf{y}_d \sim N(\boldsymbol{\mu}_d, \mathbf{V}_d)$, dove $\boldsymbol{\mu}_d = \mathbf{X}_d\boldsymbol{\beta}$ e $\mathbf{V}_d = \sigma_u^2 \mathbf{1}_{N_d}\mathbf{1}_{N_d}' + \sigma_e^2 \mathbf{I}_N$. Allora, dalla formula (5.34) è possibile ricavare la distribuzione di $\mathbf{y}_{dr} \mid \mathbf{y}_{ds}$ e le rispettive media $\boldsymbol{\mu}_{dr|s}$ e varianza $\mathbf{V}_{dr|s}$. Per evitare problemi

computazionali dovuti alla complessità del processo, al posto del modello (5.36) possiamo utilizzare il seguente modello, notando che la matrice di covarianza condizionata $\mathbf{V}_{dr|s}$ corrisponde alla matrice di covarianza del vettore $\mathbf{y}_{dr}$ dato da:

$$\mathbf{y}_{dr} = \boldsymbol{\mu}_{dr|s} + v_d \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{dr}, \qquad v_d \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\varepsilon}_{dr} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_\varepsilon^2 \mathbf{I}_{N_d - n_d}) \qquad (5.37)$$

dove $\gamma_d = \sigma_u^2(\sigma_u^2 + \sigma_e^2 / n_d)^{-1}$ e n_d è la dimensione campionaria nel dominio *d*.

5.3.5 Il metodo proposto

Data la complessità delle misure fuzzy rispetto all'indice tradizionale di povertà e considerando quindi alcuni problemi computazionali il metodo EB descritto delle sezioni precedenti è stato modificato come segue: dal campione originale è estratto un campione della stessa dimensione di quest'ultimo e probabilità proporzionali ai pesi campionari. Quindi, il campione è rappresentativo dell'intera popolazione. Dopo ciò, i valori $\mathbf{y}_{di}$ sono generati utilizzando la formula (5.37):

$$\mathbf{y}_{di} = \boldsymbol{\mu}_{di|s} + v_d \mathbf{1}_{N_d - n_d} + \boldsymbol{\varepsilon}_{di}, \qquad v_d \sim N(0, \sigma_u^2(1-\gamma_d)), \quad \boldsymbol{\varepsilon}_{di} \sim N(\mathbf{0}_{N_d - n_d}, \sigma_\varepsilon^2 \mathbf{I}_{N_d - n_d}) \qquad (5.38)$$

sostituendo i parametri con la loro stima. Seguendo i passi descritti nella sezione 5.3.3, otteniamo un nuovo stimatore "diretto" che è rappresentativo dell'intera popolazione. Come sarà descritto nella prossima sezione, è stato condotto uno studio di simulazione basato sul modello per osservare la performance del metodo proposto per indicatori tradizionali e fuzzy per piccole aree. Dati i problemi computazionali già menzionati, per l'indice tradizionale (HCR) riportiamo i risultati della simulazione nella quale abbiamo confrontato gli stimatori diretti, gli stimatori EB originali e i nuovi stimatori EB, mentre per l'indice FM ci siamo limitati agli stimatori diretti e ai nuovi EB. Come possiamo notare dai risultati, il nuovo metodo conserva proprietà simili rispetto a quello tradizionale, ma permette di superare i problemi computazionali dovuti a popolazioni numerose o a misure di povertà più complesse come l'indicatore FM.

5.3.6 Bootstrap parametrico per la stima dell'errore quadratico medio

L' errore quadratico medio dello stimatore EB $\hat{FM}_d^{EB}$ rispetto al modello è dato da:

$$MSE(\hat{FM}_d^{EB}) = E(\hat{FM}_d^{EB} - FM_d) = V(\hat{FM}_d^{EB} - FM_d) + \left[E(\hat{FM}_d^{EB} - FM_d)\right]^2 \qquad (5.39)$$

Data la difficoltà di calcolare questa espressione analiticamente per misure di povertà, possiamo ottenere l'errore quadratico medio dello stimatore attraverso un bootstrap parametrico come descritto in Molina and Rao (2009). Questo metodo implica i seguenti passi:

1. Stimare il modello 3.9 con i dati campionari $(\mathbf{y}_s, \mathbf{X}_s)$ e ottenere gli stimatori $\hat{\beta}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ rispettivamente di β, σ_u^2 e σ_e^2, usando un metodo opportuno (per esempio il metodo REML).
2. Generare $u_d^* \sim \text{iid } N(0, \hat{\sigma}_u^2),\ d = 1,...,D$, e indipendentemente, generare $e_{di}^* \sim \text{iid } N(0, \hat{\sigma}_e^2),\ i = 1,..., N_d$.
3. Costruire il modello di super-popolazione bootstrap usando u_d^*, e_{di}^*, $\mathbf{x}_{di}$ e $\hat{\beta}$:

$$Y_{di}^* = \mathbf{x}_{di}'\hat{\beta} + u_d^* + e_{di}^* \tag{5.40}$$

4. Sotto il modello di super-popolazione bootstrap 3.13, generare un numero elevato B di popolazioni bootstrap indipendenti ed identicamente distribuite $Y_{di}^{*(b)}$ e calcolare i parametri della popolazione bootstrap $FM_d^{*(b)}$, $b = 1,\ldots,B$.
5. Da ciascuna popolazione bootstrap b generata al passo 4, prendere il campione con gli stessi indici del campione iniziale e calcolare i bootstrap EBP, $\hat{FM}_d^{EB*(b)}$ come descritto nella sezione 5.3.3, usando i dati campionari bootstrap $\mathbf{y}_s^*$ e i valori noti della popolazione $\mathbf{x}_{di}$.
6. Un'approssimazione Monte Carlo per lo stimatore dell'errore quadratico medio di $\hat{FM}_d^{EB}$ è dato da:

$$mse_*(\hat{FM}_d^{EB}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{FM}_d^{EB*(b)} - \hat{FM}_d^{*(b)})^2 \tag{5.41}$$

5.3.7. Esperimento di simulazione basato sul modello

Per studiare la performance dei nuovi stimatori EB proposti, abbiamo simulato popolazioni di dimensione $N = 20000$, composte da $D = 80$ aree con $N_d = 250$ elementi in ciascuna area $d = 1, .., D$. Le variabili risposta per le unità della popolazione Y_{dj} sono state generate utilizzando il modello (5.36) considerando come variabili ausiliarie due dummies $X_1 \in \{0,1\}$ and $X_2 \in \{0,1\}$ più un'intercetta. I valori delle due dummies per le

unità della popolazione sono state generate da due distribuzioni Bernoulli con probabilità di successo crescente con l'indice di area per X_1 e costante per X_2. Formalmente:

$$p_{1d} = 0.3 + 0.5d/80, \quad p_{2d} = 2, \qquad d = 1,...,D \tag{5.42}$$

Le variabili benessere rappresentano l'esponenziale delle variabili risposta del modello, quindi consideriamo una trasformazione logaritmica. Un insieme di indici campionari s_d è stato estratto indipendentemente in ciascuna area d con $n_d = 50$ usando un campionamento casuale semplice senza ripetizione. I valori delle variabili ausiliarie per le unità della popolazione e gli indici campionari sono stati tenuti fissi per tutte le simulazioni Monte Carlo. L'intercetta e i coefficienti di regressione associati alle due variabili ausiliarie per generare la popolazione sono $\beta = (3, 0.03, -0.04)'$. La varianza relativa agli effetti casuali di area è $\sigma_u^2 = (0.15)^2$ e la varianza degli errori $\sigma_e^2 = (0.5)^2$. La linea di povertà z è tenuta fissa a z = 12, che corrisponde al 60% della mediana della variabile benessere per una data popolazione generata. Le simulazioni Monte Carlo eseguite sono state I = 1000. Quindi, I vettori di popolazione $\mathbf{y}^{(i)}$ sono stati generati dal vero modello e per ciascuna popolazione i, abbiamo proceduto con i seguenti passi:

i. Per ciascuna popolazione sono stati calcolati l'incidenza di povertà $\left(HCR_d^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} I\left(E_{dj}^{(i)} < z\right),\ E_{dj}^{(i)} = \exp(Y_{dj}^{(i)}) \right)$ e l'indicatore monetario fuzzy $\left(FM_d^{(i)} = \frac{1}{N_d} \sum_{j=1}^{N_d} FM_{dj}^{(i)} \right)$.

ii. Per queste misure di povertà sono stati calcolati gli stimatori diretti usando la parte campionaria dell' i-mo vettore della popolazione $\mathbf{y}_s^{(i)}$.

iii. Il modello ad errori nidificati dato dalla formula (5.36) è stato stimato con i dati campionari $\left(\mathbf{y}_s^{(i)}, \mathbf{X}_s\right)$ e i parametri sono stati sostituiti dalle loro stime.

iv. L = 50 vettori non campionati $\mathbf{y}_r^{(il)}, l = 1,...,L$ sono stati generati dalla distribuzione condizionata (5.34) usando la (5.37) e il vettore della popolazione $\mathbf{y}^{(il)}$ è stato formato unendo i dati campionati $\mathbf{y}_s^{(i)}$ ai dati non campionati generati $\mathbf{y}_r^{(il)}$. A questo punto sono stati calcolati gli stimatori EBP attraverso approssimazioni Monte Carlo.

v. Dal campione originale, è stato estratto un campione della stessa dimensione del campione originale e probabilità proporzionale ai pesi campionari. $L = 50$ $\mathbf{y}_{di}$ valori sono stati generati dalla (5.38) e sono state calcolate approssimazioni Monte Carlo dei nuovi EBP per le misure di povertà.

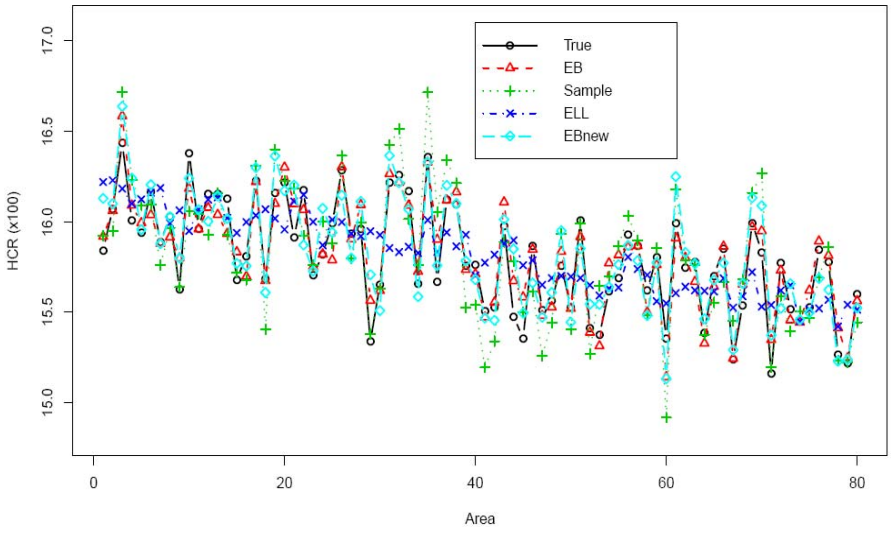
vi. Sono state calcolate le medie sulle popolazioni Monte Carlo per i veri valori delle misure di povertà, distorsioni ed errori quadratici medi sulle popolazioni Monte Carlo $i = 1, \ldots, I$ dei tre stimatori.

vii. Sono stati calcolati anche gli stimatori ELL (Elbers *et al.*, 2003) delle misure di povertà. Innanzitutto è stato stimato il modello (5.36) con i dati campionati $\mathbf{y}_s$ e sono state generate $A = 50$ popolazioni censuarie utilizzando un algoritmo bootstrap parametrico (per dettagli Molina and Rao, 2009). Per ciascuna popolazione sono state calcolate le misure di povertà e i risultati sono stati ottenuti come media sulle A popolazioni.

Come spiegato precedentemente, per problemi computazionali, per l'indice FM non sono stati eseguiti i passi iv e vii.

Le Figure 5.1, 5.2 e 5.3 mostrano rispettivamente i trends, le distorsioni e gli errori quadratici medi per gli stimatori relativi all'HCR.

Figura 5.1. Trend sulle popolazioni simulate dei veri valori e degli stimatori EB, ELL, diretto e nuovo EB per l' HCR per ciascuna area d



I veri valori e i quattro stimatori hanno gli stessi valori assoluti. Possiamo notare che le performance degli stimatori EB standard e dei nuovi proposti sono molto simili, quindi il nuovo metodo non ha perdita di efficienza. Inoltre, distorsioni in valore assoluto tra gli stimatori non sono significative (Figura 5.2), ma la Figura 5.3 mostra per gli stimatori standard EB e nuovo EB significativi miglioramenti nell'errore quadratico medio rispetto agli stimatori diretti e a quelli ELL.

Figura 5.2. Distorsione ($\times 100$) sulle popolazioni simulate degli stimatori EB, ELL, diretto e nuovo EB per l' HCR per ciascuna area d

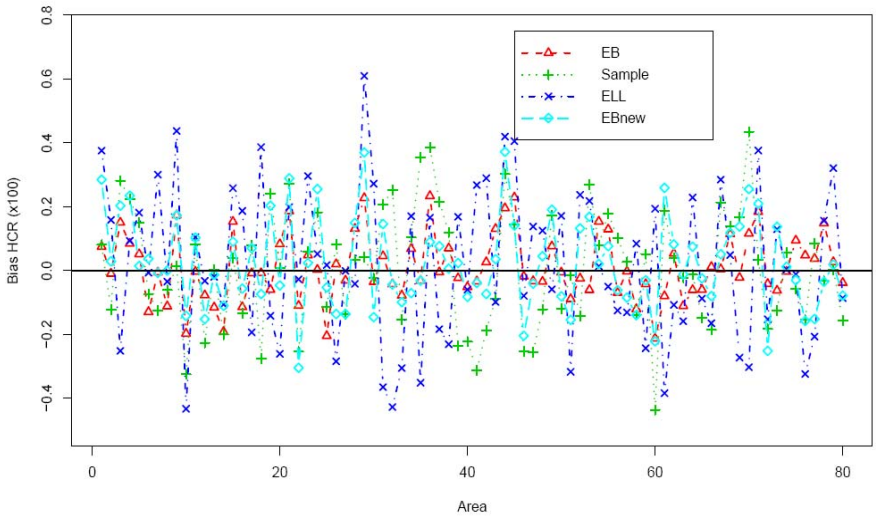
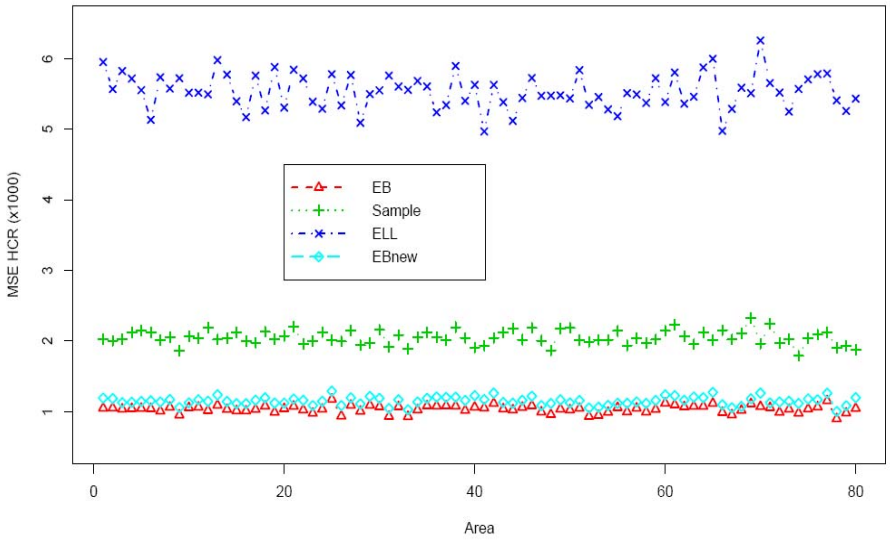


Figura 5.3. Errore quadratico medio ($\times 1000$) sulle popolazioni simulate degli stimatori EB, ELL, diretto e nuovo EB per l' HCR per ciascuna area d



Analogamente, le Figure 5.4, 5.5 e 5.6 mostrano rispettivamente i trends, le distorsioni e gli errori quadratici medi degli stimatori diretti e nuovi EB per l'indice FM.

Come per l' HCR, i veri valori e i due stimatori hanno stesso valore assoluto. Inoltre, le distorsioni non sono significative in valore assoluto tra gli stimatori (Figura 5.5), ma la Figura 5.6 mostra per i nuovi stimatori EB miglioramenti nell'errore quadratico medio rispetto agli stimatori diretti.

Figura 5.4. Trend sulle popolazioni simulate dei veri valori e degli stimatori nuovi EB e diretti per FM per ciascuna area d

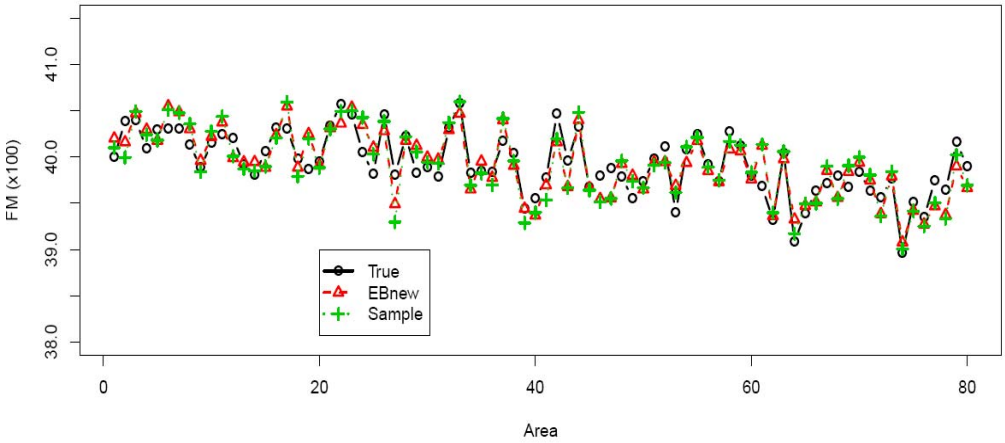


Figura 5.5. Distorsione (×100) sulle popolazioni simulate dei nuovi stimatori EB e degli stimatori diretti per FM per ciascuna area d

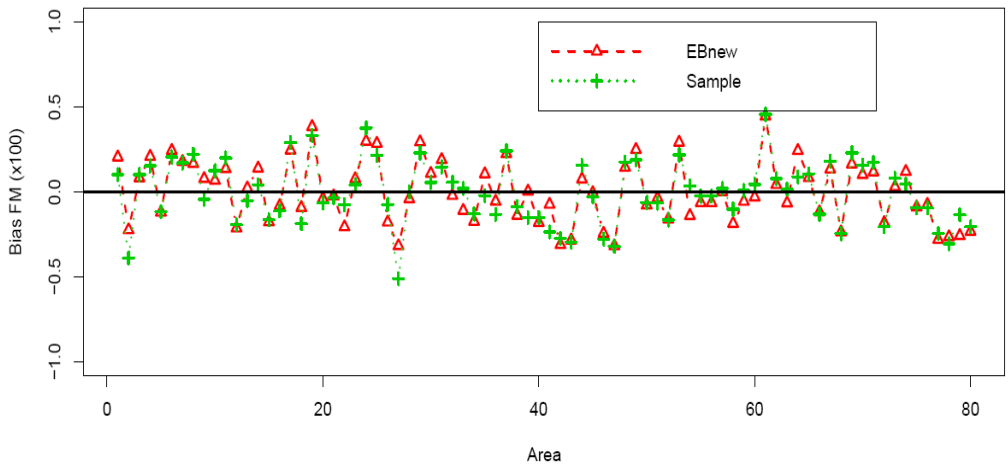
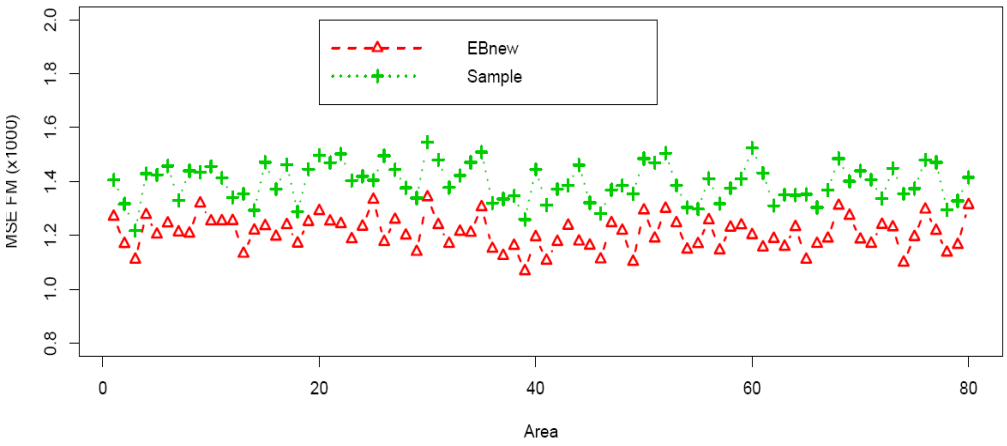


Figura 5.6. Errore quadratico medio ($\times 1000$) sulle popolazioni simulate dei nuovi stimatori EB e degli stimatori diretti per FM per ciascuna area d



5.3.8 Applicazione ai dati della Toscana

Il metodo EB modificato descritto nelle precedenti sezioni è stato applicato per stimare l'head count ratio (HCR), gli indicatori fuzzy monetario (FM) e supplementare nelle 10 province e nei 54 comuni campionati della Toscana. I dati utilizzati si riferiscono all'indagine EU-SILC 2004.

In Toscana, il campione regionale è basato su un disegno campionario a due stadi: in ciascuna provincia i comuni sono le unità campionarie di primo grado suddivise in strati in rapporto alla dimensione della popolazione. Da questi, attraverso un campionamento sistematico vengono selezionate le famiglie, ovvero le unità di secondo grado. Alcune province, in particolare le più piccole, possono contenere un numero limitato di comuni campionati e molti comuni non sono inclusi nel campione. Per esempio nel 2004 solo 54 comuni su 287 sono presenti nell'indagine. Quindi, dati gli elevati errori degli stimatori diretti a livello provinciale o l'impossibilità di calcolare questi a livello di comune, sono necessarie tecniche di stima per piccole aree.

Nella seguente analisi le piccole aree d'interesse sono le 10 province della Toscana con dimensione campionaria da 155 individui per la provincia di Grosseto a 1403 per la provincia di Firenze e i 54 comuni della Regione con dimensione campionaria da 23 individui per il comune di Sovicille a 408 per quello di Firenze. La dimensione campionaria complessiva è di 4426 individui.

La variabile benessere considerata è il reddito equivalente netto. Per ovviare al problema di valori negativi di tale variabile abbiamo proceduto seguendo suggerimenti

di Eurostat, ovvero tutti i valori inferiori al 15% della mediana del reddito familiare sono stati posti uguali tale soglia. Questa strategia non ha effetti sulla linea di povertà e quindi sugli stimatori diretti (Eurostat, 2006; Ciampalini *et al*., 2009; Neri *et al*., 2009).
Il reddito netto annuale equivalente è stato trasformato considerando il suo logaritmo al fine di ottenere una distribuzione approssimativamente Normale. Tale variabile trasformata è stata utilizzata come variabile risposta nel modello ad errori nidificati (5.36). Come variabili ausiliarie abbiamo considerato gli indicatori di 5 gruppi di variabili età, l'indicatore relativo alla nazionalità italiana, gli indicatori di 3 livelli della variabile educazione e 3 categorie della variabile occupazione. La linea di povertà per il calcolo dell'HCR è stata determinata come il 60% della mediana del reddito equivalente individuale ed è uguale a 9,372.24 Euro.
Sono stati calcolati gli stimatori diretti e i nuovi stimatori EB per l'HCR e gli indici FM e FS. Nella presente analisi abbiamo fissato il parametro alpha uguale a 2, quindi abbiamo eluso ogni legame numerico con l'approccio tradizionale. Questo perché il principale obiettivo dell'analisi è quello di sviluppare metodologie per la stima di misure fuzzy in piccolo aree piuttosto che un confronto con l'approccio tradizionale.
Nella tabella 5.1 sono riportati i valori degli stimatori diretti e dei nuovi stimatori EB per l'HCR e i corrispondenti coefficienti di variazione (CV) per ciascuna provincia della Toscana. Il valore medio sulle 10 province è 16.4%. Le province più povere sono concentrate principalmente a nord ovest della Regione. La provincia di Massa ha la più alta percentuale di individui poveri (22.4%) seguita da Lucca (18.2%) e Pisa (17.8%). Dall'altro lato, le province di Arezzo (13.0%) e Firenze (14.4%) sono le più ricche. Gli errori quadratici medi dei nuovi stimatori EB per l' HCR sono stati calcolati attraverso lo stimatore parametrico bootstrap (5.41) con B = 500. Il coefficiente di variazione è dato da $\text{cv}(\hat{HCR}_d^{newEB}) = \{\text{mse}(\hat{HCR}_d^{newEB})^{1/2} \big/ \hat{HCR}_d^{newEB}\}$. I risultati riportati nella Tabella 5.1 mostrano che i coefficienti di variazione dei nuovi stimatori EB sono inferiori a quelli degli stimatori diretti e la riduzione in termini di coefficiente di variazione tende ad essere maggiore per domini con dimensione campionaria più piccola. La sola eccezione è la provincia di Firenze che ha un'elevata dimensione campionaria, per la quale i coefficienti di variazione dei nuovi stimatori EB sono maggiori di quelli degli stimatori diretti. Dati problemi computazionali, non abbiamo potuto calcolare gli errori quadratici medi dei nuovi stimatori EB per le misure di povertà fuzzy.

Tabella 5.1. Dimensione della popolazione, dimensione campionaria, stimatori diretto e nuovo EB per l'HCR, CV degli stimatori diretto e nuovo EB (x100) per le province Toscane

Province	Dim. popolazione	Dim. campione	$\hat{HCR}_d^{DIR}$	$\hat{HCR}_d^{newEB}$	cv $\hat{HCR}_d^{DIR}$	cv $\hat{HCR}_d^{newEB}$
Arezzo	304121	416	0.087	0.130	19.09	12.42
Firenze	1119377	1403	0.133	0.144	8.32	10.89
Grosseto	149082	155	0.124	0.147	26.10	11.10
Livorno	290122	339	0.131	0.149	15.61	10.30
Siena	278495	338	0.110	0.156	19.31	10.80
Prato	319320	416	0.170	0.159	14.06	10.77
Pistoia	267076	344	0.174	0.169	13.75	9.87
Pisa	335777	399	0.168	0.178	15.54	8.88
Lucca	265293	315	0.215	0.182	13.30	8.88
Massa Carrara	251471	301	0.260	0.224	13.09	7.30
valore medio			**0.157**	**0.164**		

La Tabella 5.2 mostra rispettivamente gli stimatori diretto e il nuovo EB degli indici FM e FS e la combinazione dei due.

I nuovi stimatori EB dell'indice FM forniscono la stessa indicazione circa la povertà monetaria nelle piccole aree rispetto all'indice tradizionale e anche le differenze tra province sono simili seguendo i due approcci. I valori dell'indice FM sono maggiori di quelli dell'HCR poiché è presente una certa concentrazione in ciascuna provincia di individui con reddito equivalente appena sopra la linea di povertà. Le province di Arezzo (36.5%) e Firenze (38.0%) si confermano le più ricche, mentre la provincia di Massa ha la percentuale più elevata di individui poveri (47.5%) seguita da Lucca (42.6%) e Pisa (42.3%).

Successivamente, è stato calcolato per le province l'indice fuzzy supplementare complessivo. In questo caso la variabile benessere è il punteggio complessivo e come variabile risposta nel modello ad errori nidificati abbiamo considerato la trasformazione clog-log. Come variabili ausiliarie abbiamo utilizzato le stesse variabili scelte per il calcolo dell'HCR e degli indici FM.

In riferimento alla privazione non-monetaria, la posizione di alcune province è completamente opposta rispetto alla povertà monetaria (tabella 5.2). Per esempio, la provincia di Massa ha alti valori di FM e bassi valori di FS, mentre per le province di Firenze e Livorno vale l'opposto.

La dimensione non monetaria è stata combinata con quella monetaria al fine di ottenere le misure di privazione manifesta (MAN) e latente (LAT) che corrispondono rispettivamente all'intersezione e all'unione degli insiemi sfocati. Possiamo notare

alcune differenze tra le province (Tabella 5.2). La minima sovrapposizione si ha per Grosseto 27.6% mentre la più elevata per Pistoia 39.4%. In generale, il rapporto MAN/LAT è inferiore nelle aree con livelli di privazione più bassi e maggiore per livelli più alti. Elevati valori di tale rapporto implicano che coesistono differenti tipi di privazione e quindi questo significa che in aree dove i livelli di relativa privazione sono già elevati, privazione monetaria e non monetaria tendono con maggiore probabilità ad affliggere gli stessi individui della popolazione. Per contro, bassi valori implicano assenza di tale sovrapposizione a livello micro.

Tabella 5.2 Dimensione campionaria, stimatori diretto e nuovo EB degli indici FM e FS, privazione latente (LAT) e manifesta (MAN) ,rapporto MAN/LAT per le province toscane

Province	Dim. Camp.	$\hat{FM}_d^{\alpha DIR}$	$\hat{FM}_d^{\alpha newEB}$	$\hat{FS}_d^{\alpha DIR}$	$\hat{FS}_d^{\alpha newEB}$	LAT	MAN	MAN/LAT
Arezzo	416	0.354	0.365	0.262	0.297	0.494	0.167	0.338
Firenze	1403	0.376	0.380	0.371	0.368	0.542	0.206	0.380
Grosseto	155	0.390	0.383	0.158	0.203	0.460	0.127	0.276
Livorno	339	0.379	0.392	0.370	0.372	0.552	0.212	0.384
Siena	338	0.381	0.396	0.306	0.321	0.526	0.191	0.363
Prato	416	0.402	0.404	0.332	0.347	0.545	0.206	0.377
Pistoia	344	0.424	0.414	0.411	0.380	0.570	0.225	0.394
Pisa	399	0.433	0.423	0.353	0.348	0.558	0.212	0.381
Lucca	315	0.424	0.426	0.398	0.360	0.566	0.220	0.388
Massa Carrara	301	0.496	0.475	0.330	0.316	0.578	0.213	0.368
valore medio		**0.406**	**0.406**	**0.329**	**0.331**			

La Figura 5.7 mostra rispettivamente i cartogrammi dell' head count ratio, degli indicatori fuzzy monetario e supplementare e del rapporto manifesta/latente nelle province della Toscana costruiti usando i nuovi stimatori EB.

La stessa analisi è stata condotta per i 54 comuni campionati della Toscana. I valori degli stimatori diretti con i relativi coefficienti di variazione e i valori dei nuovi stimatori EB per l'HCR sono riportati nella Tabella 5.3. E' presente una certa distorsione probabilmente dovuta alla non perfetta approssimazione della variabile trasformata logaritmo del reddito alla distribuzione Normale. Comunque i coefficienti di variazione degli stimatori diretti per la maggior parte dei comuni sono molto elevati e quindi è necessario utilizzare metodi di stima per piccole aree.

La tabella 5.3 mostra differenze significative tra i comuni, dal valore più basso per Siena (4.6%) a quello più elevato per Massa (45.9%). La media sui comuni è 15.8%. In 13 comuni il tasso si povertà supera il 20% mentre per 10 è inferiore al 12%.

In accordo con i risultati delle province, i comuni più poveri sono concentrati principalmente nel nord ovest della Toscana e in particolare nella provincia di Massa. Il comune di Massa ha la più alta percentuale di poveri (45.9%) seguito da Certaldo (29.0%) e San Godenzo (27.6%). Dall'altro lato, il comune di Siena (4.6%), Montemurlo (6.1%) e Pescia (7.8%) sono i più ricchi.

Figura 5.7. Cartogrammi dell' head count ratio, degli indicatori fuzzy monetario e supplementare e del rapporto manifesta/latente nelle province della Toscana.

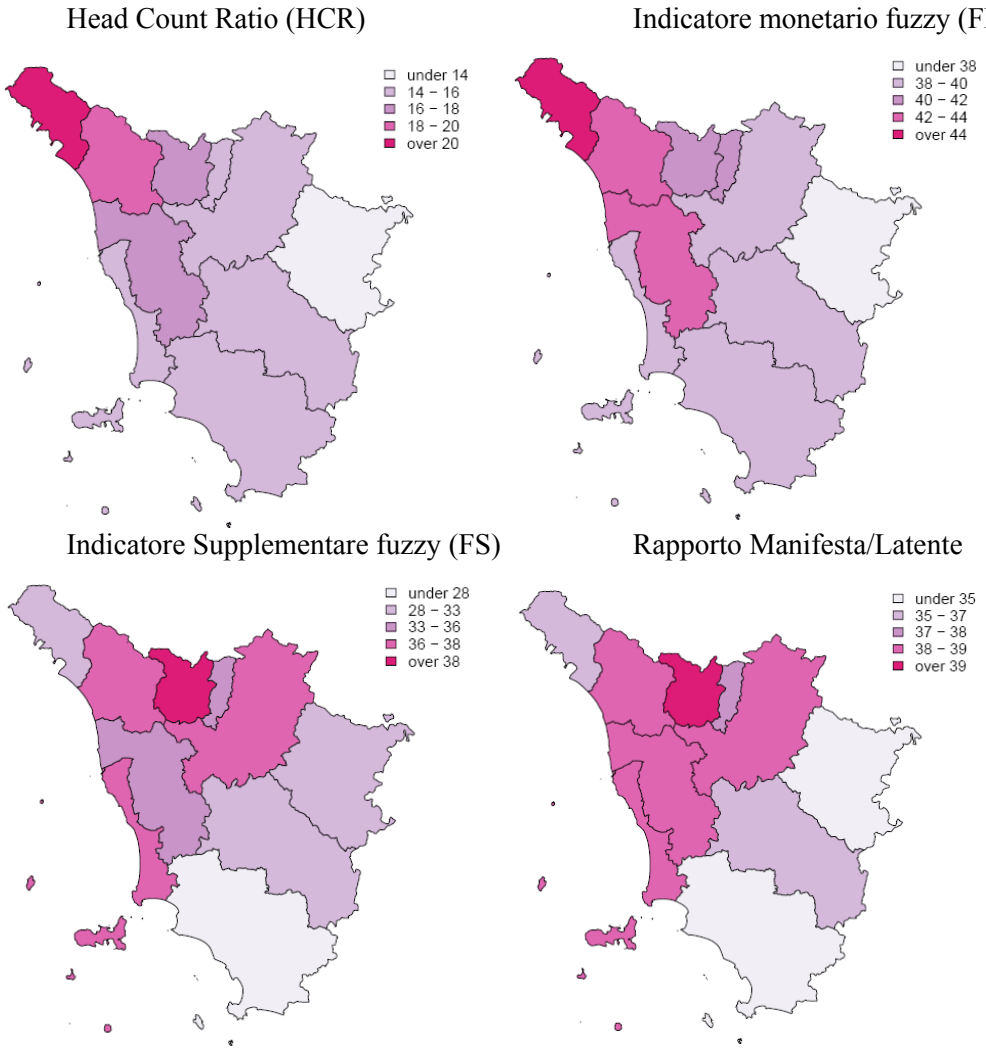


Tabella 5.3. Dimensione della popolazione, dimensione campionaria, stimatori diretti e nuovi EB per l'HCR, coefficienti di variazione degli stimatori diretti (x100) per i 54 comuni della Toscana campionati

Comuni	Dim. popolazione	Dim. campione	$\hat{HCR}_d^{DIR}$	$\hat{HCR}_d^{newEB}$	cv $\hat{HCR}_d^{DIR}$ (x 100)
Siena	66988	75	0.000	0.046	
Montemurlo	53258	75	0.000	0.061	

Pescia	43181	56	0.075	0.078	49.97
Pian di Sco	44789	71	0.000	0.083	
Firenze	362591	408	0.132	0.087	16.04
Bagno a Ripoli	76415	94	0.061	0.089	40.96
Bibbiena	48765	67	0.072	0.101	50.23
Grosseto	71211	89	0.066	0.105	42.50
Arezzo	111864	145	0.094	0.110	29.86
Reggello	48522	65	0.037	0.117	70.79
Vaiano	50055	69	0.076	0.121	41.39
Podenzana	41063	69	0.110	0.126	42.65
Livorno	148969	171	0.125	0.128	22.29
Santa Fiora	41672	32	0.096	0.129	58.62
Capannori	94622	120	0.122	0.129	30.86
Calcinaia	33621	43	0.000	0.129	
Campi Bisenzio	57764	74	0.038	0.130	58.74
Cecina	83711	100	0.165	0.134	26.23
Carmignano	35869	56	0.177	0.137	35.77
Lastra a Signa	50455	73	0.091	0.138	43.43
Pontassieve	49999	65	0.123	0.142	44.72
Pisa	122490	122	0.229	0.144	25.00
Castel Fiorentino	58114	74	0.146	0.149	38.13
San Giuliano Terme	76690	100	0.060	0.151	40.88
Licciana Nardi	51761	63	0.233	0.152	29.10
Pistoia	98096	116	0.161	0.154	24.57
Pelago	42455	62	0.117	0.158	39.71
Empoli	77761	109	0.112	0.162	28.48
Castiglion Fiorentino	46579	72	0.086	0.163	40.61
Fucecchio	60021	78	0.129	0.169	38.54
Scandicci	43404	45	0.199	0.172	33.92
Figline Valdarno	42683	65	0.154	0.174	34.75
Bucine	52124	61	0.163	0.175	37.29
Rapolano Terme	49621	61	0.079	0.175	42.85
Lucca	90046	103	0.231	0.188	22.52
Campiglia Marittima	57442	68	0.095	0.188	39.46
Asciano	43708	54	0.070	0.189	57.92
Incisa in val d'Arno	56247	75	0.175	0.190	29.55
Torrita di Siena	39173	58	0.100	0.191	49.86
Montepulciano	55948	67	0.255	0.191	31.02
Quarrata	48249	65	0.215	0.195	27.66
Pieve a Nievole	28590	54	0.401	0.202	28.18
Carrara	53770	66	0.167	0.207	34.03
Pomarance	58223	85	0.225	0.220	26.63
Prato	180138	216	0.246	0.233	16.02
Isola del Giglio	36199	34	0.273	0.241	37.90
Buggiano	48960	53	0.117	0.242	37.87
Lari	44753	49	0.235	0.247	32.35
Viareggio	80625	92	0.305	0.250	19.14
Fivizzano	53400	51	0.246	0.257	29.04
Sovicelle	23057	23	0.234	0.273	43.41
San Godenzo	43566	52	0.293	0.276	27.92
Certaldo	49380	64	0.279	0.290	24.15
Massa	51477	52	0.517	0.459	21.10

valore medio			**0.148**		**0.158**			

La tabella 5.4 mostra rispettivamente gli stimatori diretti e i nuovi stimatori EB per gli indici FM e FS e la combinazione dei due.

I comuni di Siena (22.1%) e Montemurlo (25.3%) si confermano i più ricchi in termini monetari mentre il comune di Massa ha la più alta percentuale di poveri (67.5%) seguito da Certaldo (53.8%) e San Godenzo (52.9%).

Relativamente alla privazione non monetaria, l'ordinamento dei comuni è abbastanza simile alla privazione monetaria agli estremi della distribuzione. Per contro, alcuni comuni hanno una posizione completamente opposta a quello della privazione monetaria. Per esempio, Fivizzano, Isola del Giglio e San Godenzo hanno alti valori di FM e bassi valori di FS, mentre per Firenze e Pian di Sco vale l'opposto.

La combinazione di povertà monetaria e non monetaria assume il valore minimo per Santa Fiora e Montemurlo (25.8%) e il massimo per Massa (49.3%).

Tabella 5.4. Dimensione campionaria, stimatori diretti e nuovi EB degli indici FM e FS, privazione latente (LAT) e manifesta (MAN), rapporto MAN/LAT per i 54 comuni della Toscana campionati

Comuni	Dim. Camp.	$\hat{FM}_d^{\alpha DIR}$	$\hat{FM}_d^{\alpha newEB}$	$\hat{FS}_d^{\alpha DIR}$	$\hat{FS}_d^{\alpha newEB}$	LAT	MAN	MAN/LAT
Siena	75	0.129	0.221	0.233	0.319	0.423	0.118	0.278
Montemurlo	75	0.200	0.253	0.162	0.210	0.368	0.095	0.258
Pescia	56	0.288	0.294	0.280	0.265	0.432	0.127	0.293
Pian di Sco	71	0.257	0.299	0.285	0.391	0.512	0.178	0.348
Firenze	408	0.295	0.304	0.443	0.438	0.549	0.192	0.350
Bagno a Ripoli	94	0.276	0.314	0.249	0.263	0.445	0.132	0.297
Bibbiena	67	0.278	0.328	0.286	0.304	0.478	0.154	0.323
Grosseto	89	0.334	0.331	0.191	0.198	0.417	0.112	0.269
Arezzo	145	0.351	0.344	0.201	0.227	0.444	0.127	0.286
Reggello	65	0.296	0.353	0.208	0.232	0.454	0.132	0.290
Vaiano	69	0.309	0.357	0.302	0.341	0.516	0.182	0.354
Podenzana	69	0.356	0.368	0.134	0.199	0.448	0.119	0.265
Campi Bisenzio	74	0.320	0.368	0.353	0.374	0.543	0.199	0.367
Capannori	120	0.346	0.369	0.325	0.286	0.494	0.161	0.327
Santa Fiora	32	0.385	0.372	0.088	0.185	0.444	0.114	0.258
Livorno	171	0.354	0.373	0.442	0.413	0.568	0.218	0.384
Cecina	100	0.397	0.377	0.333	0.345	0.532	0.190	0.358
Carmignano	56	0.330	0.378	0.533	0.522	0.639	0.261	0.409
Calcinaia	43	0.302	0.379	0.311	0.334	0.526	0.186	0.354
Pontassieve	65	0.409	0.385	0.271	0.289	0.506	0.169	0.333
Lastra a Signa	73	0.347	0.386	0.330	0.372	0.552	0.206	0.373
Pisa	122	0.426	0.388	0.425	0.361	0.547	0.202	0.369
San Giuliano Terme	100	0.388	0.398	0.288	0.287	0.510	0.175	0.342
Licciana Nardi	63	0.424	0.399	0.248	0.265	0.503	0.161	0.320

Castel Fiorentino	74	0.426	0.402	0.433	0.390	0.568	0.223	0.394
Pistoia	116	0.398	0.402	0.296	0.283	0.514	0.170	0.330
Pelago	62	0.409	0.403	0.470	0.458	0.608	0.253	0.416
Empoli	109	0.468	0.417	0.349	0.329	0.547	0.199	0.363
Castiglion Fiorentino	72	0.439	0.421	0.348	0.334	0.552	0.203	0.367
Scandicci	45	0.409	0.421	0.415	0.376	0.574	0.223	0.388
Fucecchio	78	0.450	0.422	0.302	0.346	0.559	0.209	0.374
Figline Valdarno	65	0.470	0.426	0.307	0.321	0.552	0.194	0.352
Rapolano Terme	61	0.386	0.430	0.165	0.206	0.500	0.136	0.273
Bucine	61	0.438	0.432	0.246	0.305	0.544	0.193	0.354
Asciano	54	0.458	0.435	0.465	0.414	0.603	0.246	0.408
Lucca	103	0.438	0.440	0.342	0.332	0.564	0.208	0.368
Quarrata	65	0.450	0.441	0.539	0.508	0.656	0.293	0.446
Campiglia Marittima	68	0.419	0.442	0.343	0.341	0.567	0.216	0.381
Incisa in val d'Arno	75	0.450	0.445	0.391	0.395	0.600	0.241	0.401
Montepulciano	67	0.486	0.447	0.308	0.319	0.562	0.204	0.363
Torrita di Siena	58	0.469	0.450	0.368	0.363	0.586	0.226	0.386
Pieve a Nievole	54	0.507	0.454	0.508	0.411	0.611	0.254	0.416
Carrara	66	0.490	0.468	0.486	0.432	0.629	0.271	0.430
Pomarance	85	0.494	0.479	0.345	0.355	0.601	0.233	0.387
Prato	216	0.501	0.493	0.357	0.348	0.606	0.235	0.388
Buggiano	53	0.521	0.495	0.519	0.514	0.684	0.325	0.476
Isola del Giglio	34	0.505	0.500	0.187	0.231	0.565	0.166	0.295
Lari	49	0.548	0.501	0.355	0.359	0.615	0.245	0.398
Fivizzano	51	0.480	0.505	0.217	0.266	0.584	0.188	0.321
Viareggio	92	0.502	0.510	0.506	0.494	0.685	0.320	0.467
Sovicelle	23	0.557	0.518	0.330	0.312	0.612	0.218	0.356
San Godenzo	52	0.589	0.529	0.251	0.276	0.602	0.203	0.337
Certaldo	64	0.557	0.538	0.356	0.372	0.644	0.265	0.411
Massa	52	0.702	0.675	0.562	0.461	0.761	0.375	0.493
valore medio		**0.397**	**0.399**	**0.343**	**0.345**			

BIBLIOGRAFIA

BALDI P., LEMMI A., SCICLONE N., (a cura di) (2005). Ricchezza e povertà. Condizioni di vita e politiche pubbliche in Toscana. Franco Angeli.

BALLINI F., BETTI G., NERI L. (2005). Towards a common methodology for a poverty mapping in Europe: old and new Member States, presented at Comparative Economic Analysis of Household Behaviour (CEAHB): Old and New EU Member. Department of Economics, Warsaw University, 30 September – 2 October.

BATTESE G.E., HARTER R.M., FULLER W.A. (1988). An error-component model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 401, pp. 28-36

BETTI G., BALLINI F. (2008). Variance Estimates of Poverty and Inequality in Albania. Eastern European Economics. Vol.46, n°6 November December 2008, pp.87-101.

BREUSCH T. S., PAGAN A. R. (1980). The Lagrange Multiplier Test and Its Applications to Model Specification. *Econometrics*. Review of Economic Studies, Blackwell Publishing, vol. 47(1), pp. 239-53, January.

CHELI B., LEMMI A. (1995). A 'Totally' Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic notes*, 24, n.1, pp. 115-134.

CHIANDOTTO B. (1996). L'informazione statistica a livello territoriale: significatività, problemi e limiti. Terza Conferenza Nazionale Statistica, 24-26 Novembre, Roma.

COX D.R., HINCKLEY D.V. (1974). Theoretical Statistics. Chapman and Hall, London.

ELBERS C., LANJOUW J.O., LANJOUW P. (2002). Micro-level Estimation of Welfare. Policy Research Working Paper N° 2911, World Bank, Washington DC.

ELBERS C., LANJOUW J.O., LANJOUW P. (2003). Micro-level Estimation of Poverty and Inequality. *Econometria,* 71 (1), pp. 355-364.

FAY R.E., HERRIOT R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *Journal of the American Statistical Association*, 74, pp. 269-277.

HAUSMAN J. A. (1978). Specification Tests in Econometrics. *Econometrica*, Vol. 46, No. 6., pp. 1251-1271.

RAO J.N.K. (2003). Small area estimation. John Wiley & Sons, New York.

TRIVELLATO U. (1998). Sul monitoraggio della povertà: progressi e questioni aperte. Atti della XXXIX Riunione Scientifica della Società Italiana di Statistica.