Progetto interdipartimentale
"Informazione statistica territoriale e settoriale per le politiche strutturali 2001-2008"

QCS OBIETTIVO 1 (2000-2006)
PON ASSISTENZA TECNICA E AZIONI DI SISTEMA
misura I.3

. . . . . .

Azione A

Occupati residenti e persone in cerca di occupazione per SLL 2001. Medie 2004 e 2005

Relazione metodologica[1]

. . . . . .

Roma, febbraio 2006

[1] A cura di Michele D'Alo', Stefano Falorsi, Fabrizio Solari, ISTAT - Servizio PSM

Indice

1. PREMESSA

Lo scopo del presente rapporto è quello di illustrare sinteticamente gli aspetti metodologici sottostanti alla produzione delle stime del numero di occupati e di persone in cerca di occupazione nei Sistemi Locali del Lavoro (SLL) per gli anni 2004 e 2005. Per la serie storica 1996-2002, l' ISTAT ha fornito le stime relative ai SLL definiti al censimento del 1991, sulla base dell'indagine trimestrale sulle Forze di Lavoro (FL), utilizzando uno stimatore per piccole aree di tipo *composto* classico secondo un approccio inferenziale basato sul disegno.

Rispetto alle precedenti, le stime qui presentate sono un naturale progresso metodologico e un adeguamento, tenendo conto della migliorata disponibilità di informazioni. Le principali differenze riguardano:

- Le stime per il periodo 1998-2002 erano riferite alla geografia dei Sll allora disponibile (784), definiti con riferimento ai risultati del Censimento del 1991;
- Le stime per il periodo 1998-2002 utilizzavano le informazioni provenienti dall'indagine trimestrale sulle forze di lavoro. A partire dal 2004 l'indagine sulle forze di lavoro è stata completamente modificata secondo quanto richiesto dal regolamento comunitario 577/98[2]. Tale regolamento prevede infatti lo svolgimento dell'indagine da effettuarsi durante tutte le 52 settimane di un anno, invece che con riferimento alla prima settimana di ogni trimestre.
- È stato utilizzato un nuovo modello statistico di stima che ha consentito un miglioramento nella precisione e qualità complessiva delle stime.
- Nel corso degli ultimi anni è migliorata la disponibilità di informazioni territoriali fini (a livello comunale) sulla popolazione per sesso e classi di età. La qualità e completezza di tali informazioni sono infatti determinanti nella stima degli occupati e delle persone in cerca di occupazione.

A partire dall'anno 2004, oltre alla definizione dei nuovi SLL (secondo le informazioni del censimento 2001), la strategia inferenziale è stata modificata sia perché i dati sono raccolti attraverso la nuova indagine FL di tipo continuo, sia perché sono stati studiati, verificati e implementati differenti metodi di stima per piccole aree, risultati più efficienti in base alle sperimentazioni effettuate nell'ambito del gruppo di lavoro per la produzione delle stime di interesse.

Gli aspetti legati alla stima per piccole area sono stati, inoltre, il tema del progetto europeo EURAREA che si proponeva come l'obiettivo il miglioramento delle metodologie esistenti e a cui l'ISTAT ha partecipato nel corso degli anni 2001-2005. Gli sviluppi metodologici ed i risultati delle sperimentazioni effettuate nell'ambito del progetto hanno permesso di applicare i metodi studiati in alcune delle più importanti indagini campionarie dell'ISTAT.

Nel contesto in esame la partecipazione al suddetto progetto ha permesso di verificare empiricamente la validità dei metodi proposti e la loro applicabilità ai contesti reali di indagine. Ciò ha permesso di sostituire lo stimatore composto utilizzato finora per la produzione delle stime con un stimatore basato su modello che sfrutta la correlazione spaziale tra le piccole aree di interesse. La scelta tra differenti stimatori è stata effettuata sulla base di studi di simulazione, basata sui dati del censimento della popolazione.

In questa nota saranno illustrati gli aspetti generali legati alla produzione di stime per piccole aree; gli stimatori considerati nella sperimentazione con riferimento al disegno di

[2] Istat, "La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione", Metodi e Norme n. 32, Roma, 2006

campionamento adottato nell'indagine; gli aspetti principali degli studi empirici condotti e l'analisi dei principali risultati.

2. ASPETTI GENERALI

Le indagini campionarie su larga scala hanno generalmente la finalità di stimare una vasta gamma di parametri (in genere totali e/o medie) non solo relativi all'intera popolazione oggetto di studio, ma anche a sottopopolazioni di quest'ultima riferite ad aree geografiche oppure a classificazioni di tipo socio-demografico o economico. Le stime dirette dei parametri relativi ad una data sottopopolazione sono basate unicamente sui dati osservati sulle unità campionarie ad essa appartenenti. Tuttavia, nella maggior parte delle indagini reali, la numerosità campionaria complessiva non è tale da garantire l'attendibilità delle stime dirette per tutte le sottopopolazioni di interesse. Si utilizza il termine *piccola area* per indicare ogni sottopopolazione per la quale non è possibile produrre stime dirette con una adeguata precisione campionaria.

Lo studio di metodologie statistiche per la produzione di stime per piccole aree sta assumendo notevole importanza sia a livello nazionale che internazionale; infatti, nel corso degli ultimi anni è cresciuta l'esigenza di adeguare le strutture e le procedure mediante le quali si attuano - ai vari livelli di governo (Ministeri, Regioni, Province, Camere di Commercio, Comprensori di Comuni, ecc.) - le scelte politiche e gli adempimenti amministrativi. Ciò ha determinato un ampliamento ed una specializzazione della domanda di statistiche riferite a piccole aree. A tali esigenze, i maggiori centri di informazione statistica a livello nazionale hanno dato in passato una risposta parziale attraverso tecniche di ampliamento del campione.

L'ISTAT conduce diverse indagini campionarie sulle famiglie e sulle imprese aventi la finalità di produrre stime con riferimento ad una vasta gamma di fenomeni di natura sociale, demografica ed economica. Tuttavia, poiché la dimensione campionaria delle suddette indagini è sufficiente a fornire stime di ragionevole precisione soltanto per domini territoriali pianificati (ad esempio regioni, ripartizioni geografiche) e per alcune importanti sottopopolazioni di interesse, non è sempre possibile ottenere stime che rispondano idoneamente a specifici obiettivi locali. In particolare, sino al 1990, per poter osservare - mediante l'indagine FL - la struttura e la dinamica dell'offerta di lavoro in realtà economiche e sociali sub-regionali, l'ISTAT ha fatto ricorso al sovradimensionamento del campione base.

Per risolvere i problemi finanziari, organizzativi e di qualità dei dati raccolti derivanti dall'adozione di tale tecnica l'ISTAT, così come i più importanti centri di informazione statistica nazionali ed internazionali, ha dedicato notevoli risorse allo studio delle metodologie alla base di stimatori di tipo indiretto, noti in letteratura come stimatori per piccole aree.

Gli stimatori per piccole aree più rilevanti dal punto di vista teorico e di maggiore diffusione applicativa possono essere classificati sia in funzione del contesto di rilevazione a cui si riferiscono, indagine occasionale o ripetuta nel tempo, sia in funzione dell'approccio inferenziale alla base della loro costruzione, approccio basato sul disegno o basato sul modello.

Rispetto ai metodi di stima diretti, gli stimatori per piccole aree permettono di migliorare il livello di precisione, utilizzando i valori della variabile d'interesse osservati sulle unità campionarie di un'area, detta *macroarea*, contenente la piccola area e/o relativi ad altre occasioni d'indagine oltre a quella corrente. Infatti, il riferimento alla macroarea ed eventualmente alle precedenti occasioni di indagine permette di incrementare la numerosità campionaria effettiva su cui calcolare le stime.

L'inferenza si basa generalmente su un modello che esprime il legame tra le osservazioni relative alle piccole aree appartenenti alla macroarea e/o riferite alle precedenti occasioni di indagine, sfruttando la conoscenza di informazioni ausiliarie, correlate alla variabile di interesse, desunte dal censimento oppure da archivi di tipo amministrativo. Tali metodi, pur introducendo una certa componente distorsiva legata alla validità del modello ipotizzato, consentono generalmente una riduzione della variabilità delle stime prodotte rispetto a quella ottenibile con i metodi di tipo diretto.

Poiché non si verifica mai una perfetta aderenza tra modello ipotizzato ed i rispettivi fenomeni rilevati, gli stimatori in questione sono sempre soggetti a distorsioni di difficile misurazione. Per la loro effettiva utilizzazione è necessario, pertanto, tener conto di alcuni problemi di natura teorica ed applicativa. Tali problemi riguardano la robustezza dei metodi, in particolare l'individuazione di opportuni criteri per la diagnosi della validità delle ipotesi alla base degli stimatori, la costruzione di stimatori che tengano conto di situazioni più complesse e più aderenti alla realtà; un ulteriore problema è quello dell'individuazione di tutte le fonti che possono fornire informazioni qualitativamente affidabili per la scelta delle variabili ausiliarie. Infine, per valutare le proprietà empiriche dei metodi di stima in esame nei reali contesti di indagine, è necessario effettuare verifiche su dati censuari o su dati provenienti da pseudopopolazioni.

3. LA STRATEGIA CAMPIONARIA DI RIFERIMENTO

La strategia di campionamento dell'indagine campionaria sulle FL si basa su un disegno di campionamento a due stadi con stratificazione delle unità primarie e uno stimatore appartenente alla classe degli stimatori di ponderazione vincolata (Deville e Särndal, 1992).

Per quanto concerne il disegno di campionamento, le unità primarie sono costituite dai comuni e quelle secondarie dalle famiglie. Nell'ambito di ciascuna provincia i comuni sono suddivisi in strati, sulla base della dimensione demografica degli stessi. Da ciascuno degli strati si seleziona, senza reimmissione e con probabilità proporzionale all'ampiezza demografica, un prefissato numero di comuni campione, all'interno dei quali si estrae in modo sistematico una predeterminata frazione di famiglie. Le variabili d'interesse sono rilevate su tutti i componenti delle famiglie selezionate. A partire dal primo trimestre del 2004 l'indagine si basa su un nuovo disegno rilevazione di tipo continuo, in modo che la rilevazione è effettuata in tutte le settimane dell'anno. La rotazione trimestrale del campione di famiglie è invece uguale a quella dell'indagine precedente (si veda ad esempio Falorsi e Falorsi, 1996) e la stima annua è ottenuta come media delle stime riferite ai singoli trimestri. Ulteriori dettagli relativi al nuovo disegno di campionamento sono illustrati in ISTAT (2006).

Per domini di stima costituiti dall'unione di H strati (ad es. le province, le regioni ecc.), se il parametro oggetto di inferenza è la media della variabile di interesse nella popolazione, è possibile definire la sua espressione formale con riferimento alla simbologia del campionamento a due stadi, come:

$$\theta = \frac{1}{P} \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{c=1}^{N_h} \sum_{j=1}^{M_{hc}} Y_{ahcj} ,$$

in cui, si è indicato con: h (h = 1, ..., H) l'indice di strato; c l'indice di comune; j l'indice di famiglia; P il numero di unità elementari nella popolazione; N_h il numero di comuni nello strato h; M_{hc} il numero di famiglie residenti nel comune c dello strato h; Y_{ahcj} è il totale della variabile di interesse y relativo ai P_{ahcj} componenti della famiglia j del comune c dello strato h inclusi nel generico post strato a (a = 1, ..., A) definito dalle combinazioni delle modalità del sesso con le classi di età. Per stimare il parametro θ si utilizza uno stimatore di ponderazione vincolata espresso da

$$\hat{\theta} = \frac{1}{\hat{P}} \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{j=1}^{m_{hc}} Y_{ahcj} w_{hcj}, \qquad (1)$$

essendo

$$\hat{P} = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{j=1}^{m_{hc}} w_{hcj},$$

dove $w_{hcj} = k_{hcj}\,\gamma_{hcj}$ è il peso finale assegnato alla famiglia j del comune c dello strato h, uguale al prodotto tra il peso base k_{hcj} e un fattore correttivo γ_{hcj}. Il peso base è espresso come

$$k_{hcj} = \frac{P_h}{n_h} \frac{M_{hc}}{P_{hc}} \frac{}{m_{hc}},$$

in cui P_h è il numero totale di persone residenti nello strato h, n_h è il numero di comuni campione selezionati nello strato h, m_{hc} è il numero di famiglie campione selezionate nel comune c dello strato h e P_{hc} è il numero totale di persone del comune c dello strato h.

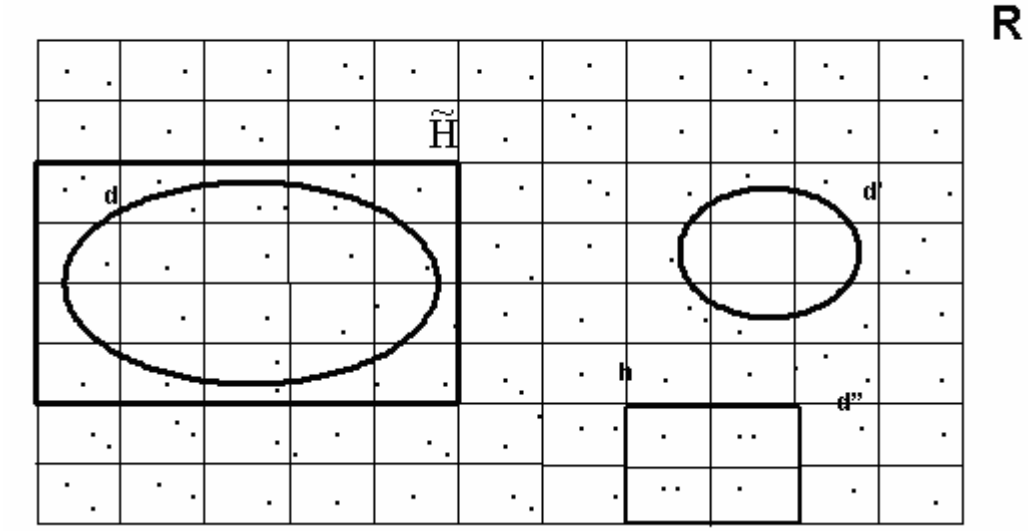
Il fattore correttivo è ottenuto come soluzione di un sistema di minimo vincolato alle numerosità N_a delle popolazioni negli A poststrati, per a = 1, ..., A. Pertanto, per il generico poststrato, vale la seguente uguaglianza

$$\hat{P}_a = \sum_{h=1}^{H} \sum_{c=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hcj} = P_a.$$

3.1 La stima di parametri riferiti ai SLL

La strategia campionaria descritta nel paragrafo precedente è mirata all'ottenimento di stime riferite alla popolazione P_R relativa ad una generica area geografica R individuata dall'unione di H strati completi. Il problema della determinazione di stime per piccole aree è schematizzato nella figura 1 in cui h indica il generico degli H strati costituenti una partizione di R ed in cui i comuni campione sono contraddistinti da punti. Nella figura sono rappresentate tre differenti piccole aree. La prima, indicata con d, esemplifica il caso di una piccola area *non pianificata* costituita dall'unione di H_d strati, alcuni dei quali non interi in quanto intersecati dalla linea perimetrale della piccola area; la seconda, denotata con d', costituisce un caso particolare della precedente in cui nessuna unità campionaria appartiene alla piccola area; la terza, indicata con d", costituisce un esempio di una piccola area *pianificata*, costituita dall'unione di strati elementari completi. In particolare, i SLL sono aree non pianificate del tipo d e d', un esempio di aree del tipo d" è dato dalle province.

Figura 1: *Alcuni importanti esempi di piccola area*



Con riferimento al disegno di campionamento descritto, tenendo conto della stratificazione territoriale e della relazione tra domini pianificati e non pianificati, è utile introdurre il generico parametro di interesse relativo alla piccola area d di dimensione P_d espresso da

$$\theta_d = \frac{1}{P_d} \sum_{a=1}^{A} \sum_{h=1}^{H_d} \sum_{c=1}^{N_{dh}} \sum_{j=1}^{M_{hc}} Y_{ahcj} \qquad (2)$$

e il corrispondente stimatore diretto di ponderazione vincolata

$$\hat{\theta}_d = \frac{1}{\hat{P}_d} \sum_{a=1}^{A} \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w_{hcj} Y_{ahcj} \qquad (3)$$

la cui espressione è analoga a quella dello stimatore $\hat{\theta}$, dato dalla (1), riferita alle sole unità campionarie appartenenti alla piccola area d, in cui H_d è il numero minimo di strati la cui unione contiene la piccola area d, N_{dh} e n_{dh} indicano rispettivamente il numero di comuni dello strato h appartenenti alla piccola area d e il corrispondente numero di comuni campione.

Lo stimatore diretto dato dalla (3) utilizza le sole unità appartenenti al sottocampione relativo alla piccola area di interesse, per tale motivo le stime dirette sono spesso caratterizzate da un eccessiva variabilità; in particolare ciò si verifica per domini di stima non pianificati, nei quali la dimensione campionaria è spesso molto piccola, se non addirittura nulla. In tal caso, per calcolare stime dirette affidabili sarebbe necessario definire una numerosità campionaria notevolmente più elevata sia a livello complessivo che per ciascun dominio pianificato. Alternativamente, così come usualmente accade, si può ricorre all'utilizzo di tecniche di stima per piccole aree.

4. METODOLOGIA ALLA BASE DEGLI STIMATORI PER PICCOLE AREE

Con riferimento al contesto d'indagine descritto, si descrive la struttura formale degli stimatori per piccole aree considerati nell'analisi sperimentale, scelti per la loro rilevanza dal punto di vista teorico e per la loro ormai diffusa applicazione. In particolare, rientra nell'approccio basato sul disegno, oltre allo stimatore diretto, lo stimatore di regressione generalizzata (Generalized Regresion). Rientrano, invece, nell'approccio predittivo lo stimatore sintetico di regressione e il predittore EBLUP (Empirical Best Linear Umbiased Predictor) basati su un modello lineare ad effetti misti a livello di unità elementare e gli analoghi stimatori basati su un modello lineare ad effetti misti a livello di area.

La differenza tra gli stimatori sintetici e i corrispondenti stimatori EBLUP è che nei primi si tiene conto della sola componente fissa del modello mentre nei secondi la stima degli effetti fissi si combina con la stima della componente casuale specifica di area, permettendo di definire stimatori che tengono conto di una fonte di variabilità specifica di area e non spiegata dagli effetti fissi del modello.

Nella specificazione del modello alla base di tali stimatori si assume che gli effetti casuali di area si distribuiscano normalmente, siano indipendenti ed omoschedastici; tuttavia, è spesso possibile che le manifestazioni di un fenomeno siano correlate a quelle relative alle aree limitrofe, più di quanto possano esserlo con quelle relative alle aree più distanti. Per tenere conto dell'informazione spaziale, si considera uno stimatore EBLUP basato su un modello definito a livello di unità elementare, in cui la componente aleatoria di area viene definita in modo da tener conto di una struttura di autocorrelazione spaziale tra le osservazioni, sulla base della distanza euclidea tra i centroidi delle aree d'interesse.

Al fine di rendere più agevole la successiva trattazione formale, è utile introdurre una notazione semplificata in cui si trascura la struttura del piano di campionamento utilizzato. In tal caso, al posto dell' indice j di unità, c di comune, h di strato ed a di post-strato è possibile considerare

l'indice i per indicare complessivamente gli indici a h c j. Il parametro di interesse relativo alla piccola area, dato dalla (2) può, pertanto, essere scritto nel modo seguente

$$\theta_d = \frac{1}{P_d} \sum_{i \in U_d} Y_i \,, \qquad (4)$$

in cui U_d è l'insieme delle unità appartenenti alla piccola area d, di dimensione N_d, Y_i è il valore della variabile di interesse osservato sull'i-esima unità della popolazione, per $i \in U_d$. In tal caso, lo stimatore diretto può essere espresso come

$$\hat{\theta}_d = \frac{1}{\hat{P}_d} \sum_{i \in s_d} w_i Y_i \,, \qquad (5)$$

dove s_d indica l'insieme delle unità campionarie appartenenti alla piccola area d, di dimensione n_d, e w_i è il peso finale assegnato all'i-esima unità del campione $(i \in s_d)$.

4.1 Stimatore di regressione generalizzata

Lo stimatore GREG è utilizzato nella maggior parte delle indagini su larga scala eseguite dagli Istituti Nazionali di Statistica, qualora siano noti i totali di popolazione per una o più variabili ausiliarie e sia possibile osservare tali informazioni sulle unità appartenenti al campione. Tale stimatore è approssimativamente corretto sotto il piano di campionamento ed usa in modo efficiente l'informazione ausiliaria, calibrando le stime rispetto alle covariate considerate nel modello. Nella stima di parametri riferiti alla popolazione obiettivo o a sottopolazioni della stessa, lo stimatore GREG da luogo a guadagni di efficienza rispetto allo stimatore espansione, in funzione del grado di correlazione esistente tra la variabile di interesse e le covariate considerate .

Lo stimatore GREG costituisce un caso particolare degli stimatori di calibrazione (Deville e Särndal 1992) e può essere esplicitato aggiungendo allo stimatore diretto un termine di aggiustamento. Tale termine è individuato dalle differenze calcolate tra le medie delle singole covariate note nella popolazione e le rispettive stime calcolate con le osservazioni campionarie moltiplicate per i corrispondenti coefficienti di regressione stimati; in formule si ha

$$\hat{\theta}_d^{GREG} = \frac{1}{\hat{P}_d} \sum_{i \in s_d} w_i y_i + \left(\overline{\mathbf{X}}_d - \frac{1}{\hat{P}_d} \sum_{i \in s_d} w_i \mathbf{x}_i \right)^T \hat{\boldsymbol{\beta}} \,. \qquad (6)$$

Nell'espressione precedente $\overline{\mathbf{X}}_d = \left(\overline{X}_{d,1},...,\overline{X}_{d,p}\right)^T$ indica il vettore delle medie delle p covariate nella popolazione, mentre $\hat{\boldsymbol{\beta}}$ è la stima dei coefficienti di regressione del modello lineare standard alla base della costruzione dello stimatore, espresso da

$$y_{di} = \mathbf{x}_{di}^T \boldsymbol{\beta} + \varepsilon_{di},$$

in cui

$$E(\varepsilon_{di}) = 0; \quad Var(\varepsilon_{di}) = \sigma_\varepsilon^2, \qquad \forall i = 1, \ldots, n_d \text{ e } d = 1, \ldots, D,$$

dove $\mathbf{x}_{di} = \left(x_{di,1},...,x_{di,p}\right)^T$ è il vettore delle osservazioni campionarie delle p covariate.

Stimato il coefficiente di regressione $\boldsymbol{\beta}$ mediante l'usuale metodo dei minimi quadrati ponderati

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s_d} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in s_d} w_i \mathbf{x}_i \, y_i ,$$

l'espressione dello stimatore (6) è equivalente alla seguente

$$\hat{\theta}_d^{GREG} = \frac{1}{\hat{P}_d} \sum_{i \in s_d} w_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) + \overline{\mathbf{X}}_{.d}^T \hat{\boldsymbol{\beta}} . \qquad (7)$$

4.2 Stimatore sintetico basato su un modello a livello di unità.

Tale stimatore sintetico è costruito sulla base di un modello lineare ad effetti misti con variabili ausiliarie specificate a livello di unità, effetti casuali di area ed errori accidentali distribuiti normalmente e tra loro indipendenti (Battese et. al. 1988), che può essere formalizzato nel seguente modo

$$y_{di} = x_{di}^T \boldsymbol{\beta} + u_d + e_{di},$$

$$\forall i = 1, \ldots, N_d \text{ e } d = 1, \ldots, D. \qquad (8)$$

$$u_d \sim \text{iid } N(0, \sigma_u^2), \quad e_{di} \sim \text{iid } N(0, \sigma_e^2),$$

Il modello (8) può essere scritto in forma più compatta mediante una rappresentazione matriciale. In tal caso, considerando la formulazione relativa alle sole unità campionarie, risulta

$$\mathbf{y}_s = \mathbf{x}_s \boldsymbol{\beta} + \mathbf{z}_s \mathbf{u} + \mathbf{e}_s ,$$

dove $\mathbf{y}_s$ è il vettore delle osservazioni campionarie, $\mathbf{x}_s$ è la matrice delle covariate osservate sulle unità campionarie, $\mathbf{e}_s$ è il vettore n-dimensionale degli errori accidentali, $\mathbf{z}_s$ è la matrice di incidenza delle unità campionarie in ogni area, $\mathbf{u}$ il vettore delle componenti casuali di area. Determinato lo stimatore dei minimi quadrati ponderati di $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}_s \hat{\mathbf{V}}^{-1}{}_s \mathbf{x}_s^T\right)^{-1} \mathbf{x}_s^T \hat{\mathbf{V}}_s \, \mathbf{y}_s \,, \tag{9}$$

in cui la stima $\hat{\mathbf{V}}_s$ della matrice di varianza di $\mathbf{y}_s$

$$\hat{\mathbf{V}}_s = \hat{\sigma}_e^2 \mathbf{I}_s + \hat{\sigma}_u^2 \mathbf{z}_s \mathbf{z}_s^T$$

è ottenuta stimando iterativamente mediante il metodo della massima verosimiglianza ristretta (REML) le componenti di varianza σ_e^2 e σ_u^2, lo stimatore sintetico (SI_A) basato su un modello lineare ad effetti misti definito a livello di unità è

$$\hat{\theta}_d^{SI\_A} = \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}}, \tag{10}$$

in cui $\bar{\mathbf{x}}_d = \left(\bar{x}_{d,1},\ldots,\bar{x}_{d,p}\right)^T$ indica il vettore delle medie campionarie relativo alle *p* variabili ausiliarie.

4.3 Stimatore sintetico basato su un modello a livello di area

Il modello alla base dello stimatore sintetico in questione è definito a livello aggregato. In tal caso si definisce una relazione tra le stime dirette del parametro di interesse e le medie delle variabili ausiliarie riferite alle piccole aree. Il modello può essere così espresso:

$$\hat{\theta}_d = \bar{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d + \bar{e}_d,$$

$$u_d \sim \text{iid } N\left(0, \sigma_u^2\right),\ \bar{e}_d \sim \text{iid } N\left(0, \sigma_e^2 / n_d\right), \tag{11}$$

dove n_d è la dimensione campionaria nell'area d e $\bar{\mathbf{X}}_d^T$ è il vettore delle medie delle p variabili ausiliarie nell'area d. In forma matriciale il modello (11) può essere riscritto nel modo seguente

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u} + \bar{\mathbf{e}},$$

$$\mathbf{u} \sim \text{MN}\left(\mathbf{0}, \sigma_u^2 \mathbf{I}\right), \quad \bar{\mathbf{e}} \sim \text{MN}\left(\mathbf{0}, \mathbf{D}\right),$$

con

$$\mathbf{D} = \begin{pmatrix} \sigma_e^2 / n_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_e^2 / n_D \end{pmatrix}.$$

Nei modelli di tipo aggregato, al fine di evitare problemi di identificabilità, le varianze campionarie si assumono generalmente note, tuttavia, disponendo delle informazioni a livello

individuale e nell'ipotesi di omoschedasticità della componente accidentale, la varianza σ_e^2 può essere stimata mediante la seguente espressione

$$\hat{\sigma}_{\mathrm{e}}^2 = \frac{1}{\mathrm{n} - \mathrm{n}^{(\mathrm{D})}} \sum_{\mathrm{i}} \sum_{\mathrm{d}} \left(\mathrm{y}_{\mathrm{di}} - \overline{\mathrm{y}}_{\mathrm{d}}\right)^2 , \qquad (12)$$

dove n è il numero di individui appartenenti al campione complessivo ed $\mathrm{n}^{(\mathrm{D})}$ è il numero di aree presenti nel campione.

Lo stimatore dei minimi quadrati ponderati del vettore dei coefficienti di regressione $\boldsymbol{\beta}$ è dato da

$$\hat{\boldsymbol{\beta}} = (\overline{\mathbf{X}}^T \hat{\mathbf{V}}^{-1} \overline{\mathbf{X}})^{-1} \overline{\mathbf{X}}^T \hat{\mathbf{V}}^{-1} \overline{\mathbf{y}} , \qquad (13)$$

dove $\overline{\mathbf{y}}$ è il vettore delle medie campionarie, $\overline{\mathbf{X}}$ è la matrice composta dalle righe $\overline{\mathbf{X}}_d^T$, $\hat{\mathbf{V}} = \hat{\sigma}_{\mathrm{u}}^2 \mathbf{I} + \hat{\mathbf{D}}$ è una matrice diagonale con elementi pari a $\hat{\sigma}_{\mathrm{u}}^2 + \hat{\sigma}_{\mathrm{e}}^2 / \mathrm{n}_{\mathrm{d}}$, in cui σ_{u}^2 e $\boldsymbol{\beta}$ sono stimati in modo iterativo.

L'espressione finale dello stimatore sintetico (SI_B) basato su un modello definito a livello di area è:

$$\hat{\theta}_{\mathrm{d}}^{SI\_B} = \overline{\mathbf{X}}_{\mathrm{d}}^{\mathrm{T}} \hat{\boldsymbol{\beta}} . \qquad (14)$$

4.4 Stimatore EBLUP basato su un modello a livello di unità

Lo stimatore in questione, indicato nel seguito come stimatore EB_A, così come lo stimatore SI_A, è basato sul modello lineare misto definito nella (8). Nell'ambito dell'approccio predittivo, il miglior predittore lineare corretto (BLUP) è ottenuto minimizzando gli errori quadratici medi all'interno della classe degli stimatori lineari non distorti. Lo stimatore BLUP dipende dalle componenti di varianza σ_{u}^2 e σ_{e}^2 che sono generalmente incognite; è necessario, quindi, calcolare una loro stima. Stimati i coefficienti di regressione tramite l'espressione (9), le componenti di varianza del modello possono essere stimate mediante differenti metodi, tra cui quello della massima verosimiglianza ristretta (REML) (Cressie, 1992) utilizzato nel presente lavoro. Il miglior predittore lineare empirico (EBLUP) è uno stimatore di tipo composto che, trascurando il fattore di correzione per popolazioni finite, è dato da

$$\hat{\theta}_{\mathrm{d}}^{\mathrm{EB\_A}} = \gamma_{\mathrm{d}} \left[\overline{\mathrm{y}}_{\mathrm{d}} + \left(\overline{\mathbf{X}}_{\mathrm{d}}^{\mathrm{T}} \hat{\boldsymbol{\beta}} - \overline{\mathbf{x}}_{\mathrm{d}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}\right)\right] + \left(1 - \gamma_{\mathrm{d}}\right) \overline{\mathbf{X}}_{\mathrm{d}}^{\mathrm{T}} \hat{\boldsymbol{\beta}}, \qquad (15)$$

dove

$$\gamma_{\mathrm{d}} = \frac{\hat{\sigma}_{\mathrm{u}}^2}{\hat{\sigma}_{\mathrm{u}}^2 + \hat{\sigma}_{\mathrm{e}}^2 / \mathrm{n}_{\mathrm{d}}} ,$$

è il peso associato alla componente campionaria; inoltre, $\overline{\mathrm{y}}_{\mathrm{d}}$ e $\overline{\mathbf{x}}_{\mathrm{d}}^{\mathrm{T}}$ sono rispettivamente il vettore delle medie campionarie della variabile di interesse y e delle covariate nell'area d, $\overline{\mathbf{X}}_{\mathrm{d}}$

è il vettore dei valori medi di popolazione delle covariate, $\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2, \hat{\sigma}_u^2$ sono le stime dei parametri del modello lineare definito a livello di unità, descritto nel paragrafo 3.2.

4.5 Stimatore EBLUP basato su un modello a livello di area

Tale stimatore, denotato nel resto del lavoro con EB_B, al pari di quello SI_B, si basa sul modello lineare normale ad effetti misti definito a livello di piccola area, dato dalla (11). Dopo aver stimato le componenti di varianza, mediante il metodo REML ed i coefficienti di regressione, con l'espressione (13), il miglior predittore lineare corretto ottenuto sulla base del modello (11) è pari alla combinazione ponderata dello stimatore diretto e dello stimatore SI_ B:

$$\hat{\theta}_d^{EB\_B} = \gamma_d \hat{\theta}_d + (1-\gamma_d)\overline{\mathbf{X}}_{.d}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \quad , \qquad (16)$$

dove

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$$

è il peso dello stimatore diretto, la varianza di campionamento σ_e^2 è stimata mediante la (12), mentre $\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}_u^2$ si ottengono iterativamente, come descritto nel paragrafo 3.3.

4.6 Stimatore EBLUP con struttura di autocorrelazione spaziale

Nel paragrafo 3.2 si è introdotto un modello lineare in cui si ipotizza che gli effetti di area siano tra loro indipendenti. Al fine di migliorare le proprietà dello stimatore, è possibile ipotizzare che le osservazioni rilevate nelle aree di interesse siano legate a quelle rilevate nelle altre aree in funzione della loro prossimità geografica, considerando effetti di area che abbiano una struttura di autocorrelazione spaziale. Il modello lineare ad effetti misti di riferimento è quello a livello di unità formalizzato nella (8); tuttavia, in questo caso gli effetti casuali di area u_d, anziché essere definiti indipendenti, sono tra loro correlati in funzione della distanza euclidea tra le aree.

L'espressione dello stimatore EBLUP basato su una struttura di autocorrelazione spaziale, indicato nel seguito come EB_SP, è ottenuta sempre sulla base del modello lineare ad effetti misti (8). Considerando il modello relativo alle sole unità campionarie si può scrivere

$$\mathbf{y}_s = \mathbf{x}_s\boldsymbol{\beta} + \mathbf{z}_s\mathbf{u} + \mathbf{e}_s ,$$

dove $\mathbf{e}_s$ ed $\mathbf{u}$ hanno una distribuzione multinormale con vettore delle medie nullo e matrice di varianza pari rispettivamente a $\sigma_e^2\mathbf{I}_n$ e $\sigma_u^2\mathbf{A}$, in cui, posto $\delta_{dd'} = 0$ se $d = d'$ e $\delta_{dd'} = 1$ altrimenti, la matrice $\mathbf{A}$ è data da

$$A = [a_{dd'}] = \left\{ \left[1 + \delta_{dd'} \exp\left(\frac{dist(d,d')}{\alpha} \right) \right]^{-1} \right\}, \qquad (17)$$

dove $dist(d,d')$ è la distanza tra le aree d e d' ed α è un parametro di scala incognito.

In tal caso, la matrice di varianza e covarianza di $\mathbf{y}_s$ è $\sigma_e^2 \mathbf{V}_s$, con $\mathbf{V}_s = \mathbf{I}_n + \varphi \mathbf{z}_s \mathbf{A} \mathbf{z}_s^T$, essendo $\varphi = \sigma_u^2 / \sigma_e^2$, e l'espressione dello stimatore BLUP nella generica area d è

$$\hat{\theta}_d = \frac{1}{P_d} \left(y_d + \left[(X_d - \mathbf{x}_d)\hat{\boldsymbol{\beta}} + (P_d - n_d) \sum_{d'=1}^{D} \left[y_{d'} - \mathbf{x}_{d'}' \hat{\boldsymbol{\beta}} \right] \tau_{d',d} \right] \right) , \qquad (18)$$

in cui:

y_d è il numero totale di persone nello stato occupazionale di interesse osservate nel sottocampione relativo all'area d;

$\mathbf{X}_d$ è il vettore dei totali di popolazione delle covariate nella piccola area d;

$\mathbf{x}_d$ è il vettore dei totali campionari delle covariate nella piccola area d;

$\hat{\boldsymbol{\beta}}$ è l'usuale stimatore dei minimi quadrati ponderati dei coefficienti di regressione;

$P_d - n_d$ è la differenza tra il numero unità elementari della popolazione appartenenti all'area d e il corrispondente numero unità osservate nel campione.

$\tau_{d',d}$ è il generico elemento della seguente matrice

$$\mathbf{T}^* = [\tau_{d',d}] = \left(\mathbf{z}_s' \mathbf{z}_s + \varphi^{-1} \mathbf{A}^{-1} \right)^{-1} = \left(\mathrm{diag}[n_d] + \varphi^{-1} \mathbf{A}^{-1} \right)^{-1}. \qquad (19)$$

Lo stimatore empirico con correlazione spaziale degli effetti di area (EB_SP), corrispondente alla (18), si ottiene stimando le componenti di varianza mediante il metodo della verosimiglianza ristretta, utilizzando un algoritmo di stima iterativo descritto in Saei e Chambers (2004).

4.7 Lo stimatore composto utilizzato finora per la stima nei SLL

Lo stimatore finora utilizzato dall'ISTAT per la produzione delle stime del numero di persone in cerca di occupazione è uno stimatore composto, dato da una combinazione lineare tra uno stimatore diretto di tipo rapporto ed uno sintetico definito in un ottica inferenziale basata sul disegno. Con riferimento alla simbologia introdotta nel paragrafo 2, lo stimatore ha la seguente espressione formale

$$_c\hat{\theta}_d = \alpha_d \, {}_r\hat{\theta}_d + (1 - \alpha_d) \, {}_s\hat{\theta}_d , \qquad (20)$$

dove α_d è una quantità costante indipendente dal campione estratto $(0 \le \alpha_d \le 1)$. Lo stimatore diretto ${}_r\hat{\theta}_d$ è lo stimatore rapporto post-stratificato, in cui si sono considerati $\tilde{A} = 4$ post-strati definiti dall'incrocio delle modalità del sesso con due classi di età (0–| 40 , 40 e oltre), espresso da

$$ {}_r\hat{\theta}_d = \sum_{a=1}^{\tilde{A}} \frac{\hat{Y}_{da}}{\hat{P}_{da}} P_{da} = \sum_{a=1}^{A} \hat{\bar{Y}}_{da} P_{da} \quad , \qquad (21) $$

in cui:

$$ \hat{Y}_{da} = \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w_{hcj} Y_{ahcj} \qquad (22) $$

e

$$ \hat{P}_{da} = \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w_{hcj} \qquad (23) $$

sono, rispettivamente, gli stimatori di ponderazione vincolata del totale della variabile di interesse e del numero, P_{da}, di unità della popolazione nel post-strato a e nell'area d.

Un'espressione alternativa dello stimatore (21) è data da

$$ {}_r\hat{\theta}_d = \sum_{a=1}^{A'} \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w_{hcj} \frac{P_{da}}{\hat{P}_{da}} Y_{ahcj} = \sum_{a=1}^{A'} \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w'_{hcj} Y_{ahcj} \quad , \qquad (24) $$

in cui il nuovo sistema di pesi finali $\{w'_{hcj}\}$ soddisfa la relazione

$$ P_{da} = \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} w'_{hcj} \quad . \qquad (25) $$

Lo stimatore sintetico ${}_s\hat{\theta}_d$, per il quale sono stati considerati gli stessi 28 post-strati utilizzati per la costruzione dello stimatore di ponderazione vincolata (1), è definito dall'espressione

$$ {}_s\hat{\theta}_d = \sum_{a=1}^{A} \frac{\hat{Y}_a}{\hat{P}_a} P_{da} = \sum_{a=1}^{A} \hat{\bar{Y}}_a P_{da} \, . \qquad (26) $$

Lo stimatore ${}_r\hat{\theta}_d$ è basato sulla costruzione di poststrati nell'ambito di ciascun SLL mentre nello stimatore ${}_s\hat{\theta}_d$ i poststrati sono definiti a livello di regione. Per la costruzione dello stimatore ${}_r\hat{\theta}_d$ in ciascun SLL è stato necessario definire un numero ridotto di post-strati ($A' = 4$), in modo da garantire la presenza di unità campione in ogni poststrato.

I pesi α_d sono stati determinati in modo da minimizzare l'errore quadratico medio dello stimatore composto e sono calcolati mediante la seguente espressione

$$\alpha_d = \frac{EQM({}_s\hat{\theta}_d)}{EQM({}_s\hat{\theta}_d) + Var({}_r\hat{\theta}_d)} = \frac{Var({}_s\hat{\theta}_d) + Bias^2({}_s\hat{\theta}_d)}{Var({}_s\hat{\theta}_d) + Bias^2({}_s\hat{\theta}_d) + Var({}_r\hat{\theta}_d)} \quad . \qquad (27)$$

Le quantità $Var({}_s\hat{\theta}_d)$, $Var({}_r\hat{\theta}_d)$ e $Bias^2({}_s\hat{\theta}_d)$ sono state calcolate utilizzando i dati del censimento 1991.

5. CONFRONTO TRA GLI STIMATORI ATTRAVERSO METODI DI SIMULAZIONE

Per affrontare il problema della scelta di uno stimatore per piccole aree per la stima dei parametri di interesse a partire dalle informazioni disponibili per una data indagine campionaria, può essere utile svolgere un'analisi comparativa tra gli stimatori considerati sulla base di uno studio simulativo condotto su una popolazione reale, artificiale o su una pseudopopolazione.

Nel primo caso è necessario disporre dei valori delle variabili di interesse e delle variabili ausiliarie per tutte le unità della popolazione ad una data il più possibile prossima a quella a cui si riferisce l'indagine; nel secondo caso si assume che le variabili di interesse e quelle ausiliarie abbiano specifiche forme distribuzionali, da cui è possibile generare i valori della variabile di interesse e delle covariate relativamente alla generica unità; nel terzo caso, la pseudopopolazione è ottenuta direttamente da un campione, non necessariamente relativo all'indagine di riferimento, su cui sono stati osservati i valori della variabile di interesse e delle variabili ausiliarie, replicando le informazioni relative a ciascuna unità campionaria un numero di volte pari al loro peso finale.

L'analisi delle proprietà empiriche degli stimatori considerati si basa sul metodo Monte Carlo, selezionando un determinato numero di campioni, R, dalla popolazione di interesse (reale, artificiale o pseudo) in conformità al disegno di campionamento adottato per l'indagine. Per ognuno di tali campioni si costruiscono le stime dei parametri di interesse utilizzando gli stimatori posti a confronto. Con riferimento a ciascuno stimatore, la distribuzione empirica determinata sulla base degli R campioni è utilizzata per la valutazione degli stimatori. In particolare si costruiscono degli indici sintetici di tali distribuzioni, utili a misurare la distorsione e la variabilità degli stimatori.

Per ciascuna piccola area, le proprietà degli stimatori esaminati, in termini di distorsione e variabilità, sono generalmente valutate sulla base dei valori assunti dalle seguenti statistiche

$$DR_d = \frac{1}{R}\left[\sum_{r=1}^{R} \frac{{}_r\hat{\theta}_d - \theta_d}{\theta_d}\right] \times 100, \qquad (28)$$

$$REQMR_d = \sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{{}_r\hat{\theta}_d - \theta_d}{\theta_d}\right]^2\right)} \times 100 \quad . \qquad (29)$$

I suddetti criteri di valutazione, espressi in termini percentuali, misurano rispettivamente la Distorsione Relativa e la Radice dell'Errore Quadratico Medio Relativo. Nelle espressioni si è

indicato con r (r = 1, ..., R) la generica replica, con θ_d e $_r\hat{\theta}_d$ il parametro di interesse e l'r-sima stima nell'area d.

Calcolando la media su tutte le aree d (d = 1, ..., D) dei valori assoluti delle statistiche DR_d e $REQMR_d$ si ottengono dei criteri di valutazione globali, la cui espressione è data da

$$\overline{DR} = \frac{1}{D}\sum_{d=1}^{D}\left| DR_d \right| , \qquad (30)$$

$$\overline{REQMR} = \frac{1}{D}\sum_{d=1}^{D} REQMR_d . \qquad (31)$$

Un secondo gruppo di criteri di valutazione globali si ottengono invece considerando il valore massimo dei valori assoluti della Distorsione Relativa ed il valore massimo della Radice dell'Errore Quadratico Medio Relativo riferiti alle piccole aree di interesse. Tali indicatori, che danno una misura delle proprietà degli stimatori di tipo *conservativo*, sono espressi come:

$$MDR = \max_d \left| DR_d \right| , \qquad (32)$$

$$MREQMR = \max_d \left| REQMR_d \right| . \qquad (33)$$

6. DESCRIZIONE DELLA SPERIMENTAZIONE

6.1 Premessa

Obiettivo dell'applicazione condotta è quello di confrontare gli stimatori per piccole aree descritti nel paragrafo 3 e di proporre una metodologia alternativa allo stimatore composto utilizzato dall'ISTAT per la stima delle persone occupate ed in cerca di occupazione nei SLL nel periodo 1996-2002.

A tal fine, è stata condotta una prima simulazione sui dati del Censimento Generale della Popolazione del 1991, con riferimento a quattro regioni italiane: Toscana, Lazio, Umbria e Marche. Dall'archivio sono stati estratti R = 500 campioni con il disegno di campionamento a due stadi (comuni-famiglie) utilizzato per l'indagine Forze Lavoro, con riferimento alle numerosità campionarie definite per tutte le province delle regioni considerate. Nello studio empirico come parametro di riferimento è stato considerato il tasso di persone in cerca di occupazione nei SLL, la cui stima a livello di piccola area presenta errori più elevati rispetto alla stima del tasso di occupazione. La scelta di considerare soltanto quattro regioni è dovuta alla complessità computazionale di un eventuale studio di simulazione condotto su tutti i SLL dell'intero territorio nazionale definiti sulla base del censimento 1991.

Uno studio ***assestante ha riguardato la scelta della macroarea di riferimento, effettuata sulla base dei risultati ottenuti applicando gli stimatori proposti ai dati di indagine del 2001, considerando modelli definiti a livello nazionale, ripartizionale e regionale.

Un secondo studio di simulazione è stato condotto sui nuovi SLL definiti in base ai dati del censimento 2001 sull'intero territorio nazionale sia con riferimento alle persone in cerca, sia per gli occupati, estraendo R = 500 campioni con il nuovo disegno di campionamento a due stadi utilizzato per l'indagine FL a partire dall'anno 2004. Quest'ultimo studio di simulazione è stato condotto per differenti ragioni. La prima era quella di valutare le proprietà empiriche degli stimatori esaminati e di confermare i risultati raggiunti con la precedente sperimentazione alla luce dei nuovi SLL 2001 e del nuovo disegno di campionamento dell'indagine continua utilizzato a partire dall'anno 2004. E' utile sottolineare che in questa sperimentazione non è stato preso in considerazione lo stimatore composto utilizzato fino al *** 200 poiché sarebbe stato molto gravoso ricalcolare i valori dei coefficienti α sui nuovi SLL sulla base dei dati censuari. Un'altra ragione è stata quella di effettuare una verifica empirica degli effetti della scelta di introdurre dei vincoli di additività delle stime prodotte a livello di SLL rispetto alle stime provinciali ottenute in base all'indagine FL.

6.2 Analisi dei risultati

Nel presente paragrafo sono riportati i risultati delle simulazioni effettuate. In particolare nella tabella 1 sono presentati gli indici di valutazione globale per ciascuno degli stimatori considerati nella simulazione condotta sui dati del censimento 1991. Si può notare che lo stimatore GREG è il migliore in termini di distorsione, tuttavia tale risultato è una conseguenza insita nella costruzione dello stimatore GREG, il quale considera la componente sintetica solo nel caso in cui nella piccola area non vi siano osservazioni campionarie. Inoltre, presentando un elevato errore quadratico medio relativo, esso è poco competitivo rispetto agli stimatori per piccole aree da modello.

Rispetto allo stimatore composto, finora utilizzato dall'ISTAT, è importante evidenziare che tutti gli stimatori basati su un modello a livello di unità portano notevoli miglioramenti nei risultati, sia in termini di distorsione che in termini di errore quadratico medio, nei valori medi e nei valori massimi. Appare, inoltre, evidente che lo stimatore EB_A è quello che fornisce i risultati migliori in termini generali, pur presentando un errore quadratico medio leggermente maggiore rispetto allo stimatore sintetico. Lo stimatore EB_SP basato su un modello con effetti di area strutturati spazialmente risulta il migliore in termini di distorsione.

Appare, quindi, opportuno sostituire lo stimatore composto con un nuovo stimatore, che in base ai risultati appena descritti può essere identificato con lo stimatore EB_SP (minore RD) o con quello EB_A (minore REQMR).

La scelta di utilizzare lo stimatore basato sulla correlazione spaziale è dipesa dalla maggiore robustezza di tale stimatore rispetto alle covariate utilizzate e alla macroarea di riferimento. La stessa simulazione condotta su un insieme di variabili ausiliarie differenti mostra, infatti, come lo stimatore EB_A sia molto più influenzato dal tipo di informazione ausiliaria disponibile, mentre le prestazioni dello stimatore EB_SP risultano pressoché invariate nelle due simulazioni. Al fine di distinguere i risultati ottenuti nelle due simulazioni sarà indicato con *caso A* lo studio condotto sull'insieme di variabili ausiliarie effettivamente disponibili, date da:

- Età (suddivisa in classi);
- Sesso;
- Tasso persone in cerca occupazione al precedente censimento;

sarà, invece, indicato con *caso B* lo studio condotto su variabili ausiliarie, per le quali i totali noti sono disponibili su file censuario:

- Età (considerate come variabile continua);
- Sesso;
- Livello d'istruzione;
- Tasso di persone in cerca occupazione al precedente censimento.

Quindi, nella tabella 1 sono riportati i risultati ottenuti nel caso A, mentre nella tabella 2 quelli ottenuti nel caso B; maggiori dettagli dell'analisi empirica effettuata sono riportati in D'Alo' *et al.*, 2004b.

Tabella 1: -*Caso A- Distorsione relativa, Radice dell'Errore quadratico medio relativo - valori medi e massimi.*

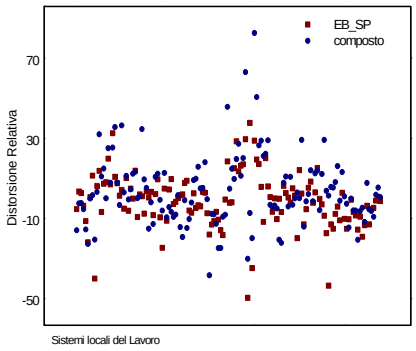
STIMATORE	$\overline{RD}$	$\overline{REQMR}$	**MRB**	**MREQMR**
GREG	4.53	29.23	46.45	59.40
SI_A	10.81	11.84	49.45	52.85
SI_B	10.85	27.64	50.58	61.66
EB_A	10.07	11.90	46.64	47.53
EB_B	10.81	27.63	50.44	61.52
EB _SP	9.60	12.33	49.52	49.89
COMPOSTO	12.41	14.66	82.52	82.83

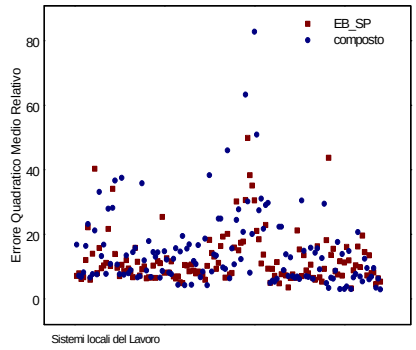
Tabella 2: -*Caso B- Distorsione relativa, Radice dell'Errore quadratico medio relativo - valori medi e massimi.*

STIMATORE	$\overline{RD}$	$\overline{REQMR}$	**MRB**	**MREQMR**
GREG	6.68	30.68	36.61	53.87
SI_A	12.09	13.35	58.52	58.71
SI_B	10.14	13.27	39.04	39.29
EB_A	11.23	13.37	52.22	53.05
EB_B	9.56	13.28	34.51	36.07
EB _SP	9.36	12.47	43.79	45.10

La maggiore robustezza dello stimatore spaziale rispetto agli altri stimatori considerati con riferimento alle variabili ausiliarie disponibili e rispetto alla macroarea di riferimento è evidente in un altro studio simulativo condotto in base all'indagine sui consumi delle famiglie per la stima sulle province italiane del numero di famiglie al di sotto della soglia di povertà (D'Alo' *et al.*, 2006).

Figura 1: *Confronto tra lo stimatore EB_SP e stimatore composto in termini di Distorsione relativa e Radice dell'Errore quadratico medio relativo*





La figura 1, riporta i valori degli indici (28) e (29) relativi allo stimatore EB_SP (18) e allo stimatore composto (20), al fine di confrontare in termini di distorsione e variabilità media nei SLL degli stimatori considerati nello studio di simulazione. Dall'esame del grafico si ha la conferma del miglioramento di prestazioni che l'uso dello stimatore EB_SP comporta rispetto allo stimatore composto finora utilizzato per la produzione delle stime.

Le due tabelle successive riportano i risultati delle simulazioni condotte sui dati del censimento 2001 nei 686 nuovi SLL dell'intero territorio nazionale. In particolare, nella tabella 3 gli indici di valutazione si riferiscono alla stima delle persone in cerca di occupazione, nella tabella 4 alla stima degli occupati. Anche in questo caso, l'analisi congiunta delle due tabelle permette di identificare nello stimatore EBLUP con correlazione spaziale (18) quello che dà luogo alle migliori prestazioni. Infatti, nella tabella 3, escluso lo stimatore GREG che presenta la distorsione più bassa ed un elevato errore quadratico medio, tra gli stimatori basati su modello, l'EB_B è quello con distorsione minore, ma con un alto errore quadratico medio, mentre lo stimatore SI_A è quello con errore quadratico medio più basso. Tuttavia, lo stimatore EB_SP è quello che dà luogo a prestazioni migliori se si considerano congiuntamente gli indici e questo

sia in termini di valore medio, sia di valori massimi. Anche, gli indicatori relativi alle stime degli occupati permettono di confermare la scelta dello stimatore EB_SP, sia in termini di distorsione sia di errore quadratico medio.

Tabella 3: *Distorsione relativa, Radice dell'Errore quadratico medio relativo - valori medi e massimi simulazione censimento 2001 persone in cerca.*

STIMATORE	$\overline{\text{RD}}$	$\overline{\text{REQMR}}$	**MRD**	**MREQMR**
GREG	6.32	36.59	83.27	107.63
SI_A	14.79	16.00	129.75	130.74
SI_B	11.81	30.99	77.31	178.54
EB_A	12.43	16.93	116.86	119.13
EB_B	10.98	31.03	77.30	178.50
EB _SP	11.37	16.50	102.18	105.47

Tabella 4: - *Distorsione relativa, Radice dell'Errore quadratico medio relativo - valori medi e massimi simulazione censimento 2001 occupati.*

STIMATORE	$\overline{\text{RD}}$	$\overline{\text{REQMR}}$	**MRB**	**MREQMR**
GREG	1.68	7.27	19.07	22.09
SI_A	3.98	4.14	33.26	33.27
SI_B	3.60	6.84	27.11	28.94
EB_A	3.04	4.13	21.13	21.45
EB_B	3.50	6.79	26.97	28.83
EB _SP	2.82	4.03	20.18	20.51

I risultati esposti nelle tabelle 3 e 4 si riferiscono a stimatori basati su modelli in cui le macroaree sono definite a livello di quattro ripartizioni geografiche (Nord-Est, Nord-Ovest, Centro e Sud). La scelta della macroarea di riferimento è stata fatta in base ai risultati ottenuti applicando gli stimatori proposti ai dati di indagine relativi ai quattro trimestri del 2001 e calcolando l'errore relativo percentuale rispetto al valore degli occupati e delle persone in cerca al censimento 2001, ovvero

$$\text{ER} = \frac{1}{\text{D}}\left[\sum_{\text{d}=1}^{\text{D}} \frac{{}_{\text{r}}\hat{\theta}_{\text{d}} - \text{C}_{\text{d}}}{\text{C}_{\text{d}}}\right] \times 100 \qquad (34)$$

L'indice sintetico *ER* è calcolato come media nei D = 686 SLL degli scostamenti relativi tra le stime calcolate con gli stimatori proposti $\hat{\theta}_{\text{d}}$ rispetto al dato del censimento C_{d}. I risultati per le stime delle persone in cerca e per gli occupati sono riportati rispettivamente nelle tabelle 5 e 6. Appare evidente, soprattutto per la stima delle persone in cerca di occupazione, che gli

stimatori basati su un modello a livello di unità sono migliori nel caso la sua specificazione sia a livello di ripartizione geografica. Da osservare che non è possibile stimare i parametri del modello a livello di area se la sua specificazione è regionale e che la stima dei parametri, in questo caso, è più efficiente se si considera come livello di specificazione del modello l'intero territorio nazionale. Naturalmente ciò è dovuto al ridotto numero di osservazioni da cui dipende la stima dei parametri del modello.

Tabella 5: *Errori relativi percentuali stime persone in cerca 2001 rispetto al dato del censimento.*

STIMATORE	**Macroarea**					
	Italia		***Ripartizione***		**Regione**	
	media	max	media	max	media	max
SI_A	17.7	113.7	16.6	80.4	21.0	97.3
SI_B	25.7	145.1	32.0	206.3	-	-
EB_A	26.6	188.1	22.6	202.8	25.5	224.9
EB_B	31.6	191.3	34.7	215.4	-	-
EB _SP	27.4	192.7	23.3	205.8	25.8	231.7

Tabella 6: *Errori relativi percentuali stime occupati 2001 rispetto al dato del censimento.*

STIMATORE	**Macroarea**					
	Italia		***Ripartizione***		**Regione**	
	media	max	media	max	media	max
SI_A	9.0	55.1	9.0	54.0	10.0	52.7
SI_B	8.9	55.6	9.3	45.8	-	-
EB_A	8.6	35.2	8.9	39.9	9.9	48.8
EB_B	9.2	36.4	9.5	47.5	-	-
EB _SP	8.8	41.0	9.0	42.3	9.1	51.5

In conclusione, le numerose analisi empiriche condotte hanno permesso di validare una nuova metodologia di stima degli occupati e delle persone in cerca di occupazione nei SLL alternativa allo stimatore composto usato in precedenza. In generale, lo stimatore che, tra quelli oggetto di sperimentazione, ha evidenziato i migliori risultati in base agli indicatori di valutazione (30), (31), (32) e (33) è l'EBLUP basato su un modello a livello di unità elementare, con struttura di autocorrelazione spaziale definita sulla base delle distanze euclidee tra i centroidi dei SLL e considerando come macroarea di specificazione del modello le quattro ripartizioni geografiche. Inoltre, da studi precedenti effettuati nell'ambito del gruppo di lavoro è stato verificato che l'uso delle matrici delle distanze stradali e quella dei tempi di percorrenza tra i centroidi dei SLL, in

alternativa alla matrice delle distanze euclidee utilizzata per definire la struttura di correlazione spaziale, non apportano significative differenze dei livelli di stima.

7. ADDITTIVITÀ DELLE STIME DI OCCUPATI E PERSONE IN CERCA DI OCCUPAZIONE PER SLL A LIVELLO DI DOMINI PIANIFICATI

7.1 Premessa

Al fine di facilitare la lettura e di rendere "coerenti" le stime relative allo stato occupazionale nei SLL in relazione alle stime prodotte con l'indagine FL nei domini di stima pianificati nel disegno dell'indagine (regioni e province), l'ISTAT ha ritenuto importante introdurre vincoli di additività delle stime a livello di SLL rispetto alle stime prodotte a livello provinciale.

Tale operazione di "aggiustamento" delle stime, da un lato rende le stime relative ai SLL coerenti a livello numerico con le stime medie annue calcolate a livello provinciale, dall'altro introduce problemi di coerenza concettuale legati alla diversa metodologia di stima utilizzata, oltre a quelle di verifica dell'impatto di tale procedura sulle stime finali. Il problema di natura concettuale è di difficile soluzione, infatti le stime a livello di dominio pianificato determinate con l'indagine FL sono calcolate utilizzando uno stimatore di calibrazione nell'ambito di un approccio inferenziale classico; mentre le stime per SLL sono determinate con un stimatore EBLUP specificando una struttura di autocorrelazione spaziale tra le osservazioni, in un ottica inferenziale predittiva. Per cui, sia le stime, sia l'errore quadratico medio ad esso associate sono difficilmente adattabili e confrontabili; infatti nel primo caso dipendono dal disegno di campionamento adottato, nel secondo dal modello di superpopolazione specificato ed ipotizzato.

Eluso il problema della coerenza formale dei due differenti approcci, resta quello relativo al metodo di "riproporzionamento" da utilizzare e la valutazione dell'impatto di tale procedura sulle stime, considerando, tra l'altro, che i SLL intersecano i domini pianificati nell'indagine FL.

Per tale motivo la procedura di adattamento delle stime di interesse si sviluppa secondo tre fasi: la prima di calcolo delle stime a livello di sottosistema provinciale; la seconda di determinazione dei correttori; la terza di calcolo delle stime finali per SLL che soddisfino l'additività delle stime a livello provinciale.

Anche in questo caso, la valutazione dell'impatto sulle stime è stata effettuata mediante simulazione, ripropozionando a livello provinciale le stime ottenute in base allo stimatore EB_SP nei 500 campioni estratti dai dati censuari del 2001. Di seguito si illustra la procedura utilizzata per la determinazione delle stime finali per SLL ($\hat{\theta}_{\mathrm{d}}^{\mathrm{EB\_SP\_PR}}$) che soddisfano il vincolo addittività, rispetto alla stima provinciale, delle stime per sottosistema provinciale.

Calcolo delle stime per sottosistemi provinciali

Indicati con l'indice dp, i sottosistemi provinciali, costituiti da 872 aree ottenute dall'intersezione dei 686 SLL con le 103 province, le stime per i sottosistemi provinciali possono essere ottenute in due modi:

1. in base alla proporzione di popolazione di 15 anni o più appartenente a ciascun sottosistema

$$\hat{\theta}_{dp}^{Pop} = \frac{\hat{\theta}_d}{P_d} P_{dp}, \quad (35)$$

in cui $\hat{\theta}_{dp}^{Pop}$ e la stima per sottosistema locale, $\hat{\theta}_d$ e la stima per SLL, P_{dp} e la popolazione per sottosistema provinciale e P_d è la popolazione provinciale;

2. calcolando per ciascun sottosistema uno stimatore sintetico sulla base della stima dei parametri di regressione del modello sottostante allo stimatore EB_SP

$$\hat{\theta}_{dp}^{beta} = \mathbf{X}_{dp}\beta_{EB\_SP}, \quad (36)$$

in cui $\hat{\theta}_{dp}^{beta}$ è la stima per sottosistema locale, $\mathbf{X}_{dp}$ è il vettore dei totali di popolazione delle covariate nel sottosistema provinciale dp, β_{EB_SP} è l'usuale stimatore dei minimi quadrati ponderati dei coefficienti di regressione calcolato per lo stimatore (18).

La differenza tra la (35) e la (36) è che la determinazione della stima per sottosistema nel primo caso dipende dalla popolazione complessiva di sottosistema, nel secondo caso dalle covariate utilizzate nel modello sottostante le stime per SLL, e quindi dalla composizione per sesso e classi d'età della popolazione.

Calcolo dei correttori di quadratura provinciale

I correttori per il riproporzionamento provinciale da applicare alle stime dei sottosistemi sono dati dal rapporto tra le stime provinciali prodotte dalla rilevazione FL e le corrispondenti stime provinciali ottenute sommando le stime relative ai sottosistemi che appartengono a ciascuna provincia. Per cui se quest'ultime sono calcolate con la (35) si ha:

$$q_p^{Pop} = \frac{\hat{Y}_p^{RFL}}{\hat{\theta}_p^{Pop}} = \frac{\hat{Y}_p^{RFL}}{\sum_{dp \in p} \hat{\theta}_{dp}^{Pop}}$$

se sono calcolate con la (36), risulta

$$q_p^{beta} = \frac{\hat{Y}_p^{RFL}}{\hat{\theta}_p^{beta}} = \frac{\hat{Y}_p^{RFL}}{\sum_{dp \in p} \hat{\theta}_{dp}^{beta}}.$$

Calcolo delle stime finali per sistemi locali del lavoro

Le stime finali per SLL sono ottenute applicando i correttori calcolati nel passo precedente alle stime dei sottosistemi provinciali, ossia

$$\hat{\theta}_{dp}^{q\_Pop} = \hat{\theta}_{dp}^{Pop} \cdot q_p^{Pop} \quad (37)$$

e

$$\hat{\theta}_{dp}^{q\_beta} = \hat{\theta}_{dp}^{beta} \cdot q_{p}^{beta} \qquad (38)$$

e sommando le suddette stime relative ai sottosistemi appartenenti a ciascun SLL si ha

$${}_{Pop}\hat{\theta}_{d}^{EB\_SP\_PR} = \sum_{dp \in d} \hat{\theta}_{dp}^{q\_Pop}, \qquad (39)$$

$${}_{beta}\hat{\theta}_{d}^{EB\_SP\_PR} = \sum_{dp \in d} \hat{\theta}_{dp}^{q\_beta} . \qquad (40)$$

Le stime nei sottosistemi provinciali, calcolate con la (37) o con la (38), rispettano i vincoli di additività con le stime prodotte dalla rilevazione FL a livello provinciale.

7.2 Risultati della sperimentazione

Nella tabella 7 sono riportati i risultati dell'analisi condotta al fine di scegliere il metodo di riproporzionamento e di verifica dell'impatto sulle stime prodotte, confrontando la distorsione relativa e l'errore quadratico medio nello spazio dei 500 campioni estratti per lo studio empirico, tra lo stimatore EB_SP, gli stimatori calibrati a livello provinciale.

L'esame della tabella mostra come il riproporzionamento delle stime a livello provinciale, porta ad un aumento dell'errore quadratico medio relativo, questo aumento è notevole, per la stima delle persone in cerca, nel caso in cui per calcolare la stima per sottosistema si utilizza l'espressione (35), più contenuta nel caso in cui si utilizza l'espressione (36). Tuttavia le stime ottenute con la (39) presentano un livello più basso della distorsione relativa. Tali risultati sono confermati dai risultati ottenuti per la stima degli occupati. Apprezzabili miglioramenti si hanno per quanto riguarda i corrispondenti valori massimi, cosa tra l'altro prevedibile, in quanto la principale conseguenza di un processo di riproporzionamento è quello di uno smussamento delle stime. Questo è anche il motivo per il quale, vincolare le stime SLL alle stime regionali, porta ad un miglioramento dei due indici di valutazione considerati, sia in termini di valori medi, sia in termini di massimi. Il miglior comportamento delle stime riproporzionate in base alle stime regionali può essere spiegato considerando: la maggiore precisione di quest'ultime stime rispetto a quelle provinciali ed al minor numero di SLL che tagliano le regioni e per le quali è necessario calcolare stime sintetiche.

Tabella 7: *Confronto della Distorsione relative, Radice dell'Errore quadratico medio relativo - valori medi e massimi - simulazione censimento 2001 - tra lo stimatore EB_SP e lo stimatore riproporzionato in base alla stima provinciale e regionale – persone in cerca e occupati.*

STIMATORE	$\overline{RD}$	$\overline{REQMR}$	**MRB**	**MREQMR**
PERSONE IN CERCA				
EB _SP	11.37	16.50	102.18	105.47
EB _SP_PR^{Pop}	9.07	20.16	95.74	99.49
EB _SP_REG^{Pop}	10.70	16.62	86.76	88.32
EB _SP_PR^{beta}	12.16	17.25	92.98	96.27
EB _SP_REG^{beta}	10.96	16.00	91.66	92.65
OCCUPATI				
EB _SP	2.82	4.03	20.18	20.51
EB _SP_PR	2.44	4.61	20.39	21.00
EB _SP_REG	2.73	4.07	19.85	20.20
EB _SP_PR^{beta}	3.06	4.26	26.48	26.70
EB _SP_REG^{beta}	2.75	3.99	20.10	20.48

In appendice A sono riportati i grafici relativi all'andamento dei valori medi della distorsione relativa e della radice dell'errore quadratico medio relativo, date rispettivamente dall'espressioni (28) e (29), per tutti gli stimatori considerati nella simulazione sui dati del censimento 2001, per i 686 SLL ordinati in base alla loro dimensione demografica. In particolare, le figure 2 e 3 si riferiscono alla stima delle persone in cerca di occupazione, le figure 4 e 5 alla stima delle persone occupate.

8. LA DETERMINAZIONE DEGLI ERRORI QUADRATICI MEDI DELLE STIME

In questo paragrafo si riporta l'espressione dell'errore quadratico medio associato alle stime prodotte con lo stimatore proposto:

$$EQM\left(\hat{\boldsymbol{\theta}}_{EB\_SP}\right) = G_1(\omega) + G_2(\omega) + G_3(\omega), \qquad (41)$$

in cui, indicato con $\omega = \left(\sigma^2, \varphi, \alpha\right)'$ il vettore dei parametri dai cui dipende l'espressione, risulta

$$G_1(\omega) = \sigma^2 \left(\mathbf{P}_d - \mathbf{n}_d\right)\mathbf{T}^* \left(\mathbf{P}_d - \mathbf{n}_d\right)',$$

$$G_2(\omega) = \sigma^2 \left[\left(\left(\mathbf{X}_d - \mathbf{x}_d\right) - \left(\mathbf{P}_d - \mathbf{n}_d\right)\mathbf{T}^* \mathbf{x}_d\right)\left(\mathbf{x}_s \hat{\mathbf{V}}_s \mathbf{x}_s^T\right)^{-1} \right.$$
$$\left. \left(\left(\mathbf{X}_d - \mathbf{x}_d\right) - \left(\mathbf{P}_d - \mathbf{n}_d\right)\mathbf{T}^* \mathbf{x}_d\right)' \right],$$

nelle quali

- $\left(\mathbf{P}_d - \mathbf{n}_d\right)$ è il vettore colonna delle differenze in ciascun area tra la numerosità della popolazione e quella campionaria;
- $\mathbf{T}^*$ é dato dalla (19)
- $\mathbf{X}_d$ è il vettore dei totali di popolazione delle covariate nella piccola area d;
- $\mathbf{x}_d$ è il vettore dei totali campionari delle covariate nella piccola area d;
- $\mathbf{x}_s$ è la matrice di dimensione $\left(n \times p\right)$ delle covariate osservate sulle unità campionarie;
- $\hat{\mathbf{V}}_s$ è la stima della matrice di varianza e covarianza di $\mathbf{y}_s$

$$G_3(\omega) = \sigma^2 \left[\mathrm{tr}\left(\mathbf{G}_{dd'}\mathbf{B}\right)\right],$$

in cui **B** è matrice di varianze e covarianze asintotica degli stimatori REML del vettore di componenti di varianza, mentre $\mathbf{G}_{dd'} = \mathrm{Cov}\left(\frac{\partial \hat{\theta}_d}{\partial \gamma}, \frac{\partial \hat{\theta}_{d'}}{\partial \gamma}\right)$.

Una stima della (41) è data da

$$\mathrm{eqm}\left(\hat{\boldsymbol{\theta}}_{EB\_SP}\right) = g_1(\hat{\omega}) + g_2(\hat{\omega}) + 2\,g_3(\hat{\omega}), \qquad (42)$$

ottenuta sostituendo al vettore dei paramenti incogniti ω la stima REML delle sue componenti, ossia $\hat{\omega} = \left(\hat{\sigma}^2, \hat{\varphi}, \hat{\alpha}\right)'$. Da tale matrice è possibile l'errore relativo delle stime delle persone in cerca e degli occupati nei SLL calcolate mediante lo stimatore proposto. Ulteriori dettagli sulle espressioni sono riportate in Saei, Chambers (2004).

L'espressione (42) è indicata per stimare l'EQM associato alle stime trimestrali; poiché le stime in questione sono stime annuali, ottenute come media delle quattro stime trimestrali, la stima dell'EQM deve tener conto dell'effetto rotazione dei campioni. Infatti, con riferimento al trimestre t (t = 1, ..., 4), la stima del numero di occupati e di persone in cerca di occupazione è basato su quattro gruppi di rotazione r (r = 1, ..., 4) ognuna delle quali fornisce una differente stima $\hat{\theta}_d^{(t,r)}$ del parametro incognito

$$\hat{\theta}_d^{(t,r)} = \frac{1}{\hat{P}_d} \sum_{a=1}^{A} \sum_{h=1}^{H_d} \sum_{c=1}^{n_{dh}} \sum_{j=1}^{m_{hc}} 4\, w_{hcj}\, Y_{ahcj} \ .$$

Di conseguenza, le stime trimestrali sono ricavate calcolando la media aritmetica delle stime basate sui gruppi di rotazione facenti parte del campione trimestrale. La stima annuale è, quindi, data dalla media aritmetica delle quattro stime trimestrali, in simboli

$$\hat{\theta}_d = \frac{1}{16} \sum_{t=1}^{4} \sum_{r=1}^{4} \hat{\theta}_d^{(t,r)} \ .$$

Sotto l'ipotesi di indipendenza tra i gruppi di rotazione, la varianza della stima $\hat{\theta}_d$ è data da

$$\mathrm{Var}\left(\hat{\theta}_d\right) = \frac{1}{16^2}\left(\sum_{t=1}^{4} \sum_{r=1}^{4} \mathrm{Var}\left(\hat{\theta}_d^{(t,r)}\right) + \sum_{t' \neq t} \mathrm{Cov}\left(\hat{\theta}_d^{(t',r)}, \hat{\theta}_d^{(t,r)}\right)\right). \qquad (43)$$

Per derivare la (43) in forma esplicita è necessario introdurre le ipotesi di omoschedasticità e costanza nel tempo della struttura di autocovarianza, ossia:

$$\mathrm{Var}\left(\hat{\theta}_d^{(t,r)}\right) = 4\,\mathrm{Var}\left(\hat{\theta}_d^{t}\right) = \mathrm{cost}, \ \ t = 1, \ \ldots, \ 4,$$

$$\mathrm{Cov}\left(\hat{\theta}_d^{(t',r)}, \hat{\theta}_d^{(t+k,r)}\right) = \rho_k\, \mathrm{Var}\left(\hat{\theta}_d^{(t,r)}\right),$$

dove ρ_k indica il coefficiente di correlazione relativo alle unità elementari osservate a k trimestri di distanza. Quindi, tenendo conto che in un anno si verificano sei sovrapposizioni ad un trimestre di distanza e una sovrapposizione a tre trimestri di distanza, la (43) si può riscrivere nel seguente modo:

$$Var\left(\hat{\theta}_d\right) = \frac{1}{4}\overline{Var}\left(\hat{\theta}_d^{t}\right)\left(\frac{3}{16}\rho_1 + \frac{1}{32}\rho_3 + 1\right),$$

dove $\overline{Var}\left(\hat{\theta}_d^{t}\right)$ indica la medie delle varianze trimestrali. La varianza annuale si ottiene, quindi, dalla media delle varianze trimestrali dividendo questa per 4 e moltiplicandola per $\left(3/16\,\rho_1 + 1/32\,\rho_3 + 1\right)$, che rappresenta il fattore correttivo della varianza della stima annuale, ottenuta partendo dalle varianze trimestrali, considerando che i campioni trimestrali sono parzialmente sovrapposti.

Per quanto concerne la determinazione dei coefficienti ρ_1 e ρ_3 si riportano nella seguente tabella i valori calcolati in un recente studio interno eseguito dell'ISTAT per un possibile ridisegno dell'indagine FL, valori che non si discostano in modo sostanziale da quelli riportati nello studio eseguito da Falorsi e Falorsi (1996).

Tabella 8: *Valori dei coefficienti di correlazione tra le unità campionarie appartenenti ai gruppi di rotazione.*

Parametro	ρ_1	ρ_3
Persone in cerca di occupazione	0.413	0.336
Occupati	0.906	0.877

Poiché la varianza campionaria in un approccio da disegno è equiparabile all'errore quadratico medio in un approccio da modello, una stima dell'errore quadratico medio che tiene conto della sovrapposizione dei campioni trimestrali può essere calcolata moltiplicando la medie delle stime trimestrali (42) per la quantità $(3/16\,\rho_1 + 1/32\,\rho_3 + 1)/4$.

Per tener conto dell'effetto del riproporzionamento provinciale nella stime dell'errore quadratico medio, è stata studiata la relazione lineare passante per l'origine nei SLL tra la radice quadrata dell'EQM associato alla stima riproporzionata e la radice dell'EQM associato alle stime EB_SP. Per ciascuno dei SLL i due valori dell'EQM sono stati calcolati come media degli scarti quadratici relativi alle 500 repliche, ossia

$$MSQ = \frac{1}{R}\left(\sum_{r=1}^{R}\left({}_r\hat{\theta}_d - \theta_d\right)^2\right)$$

Nella tabella successiva sono riportati il coefficiente di determinazione R^2 e di regressione $\hat{\beta}$ delle rette di regressione, per la stima dell'EQM associato alla stima delle persone in cerca e degli occupati.

Tabella 9: *Valori dei coefficienti di correlazione tra le unità campionarie appartenenti ai gruppi di rotazione.*

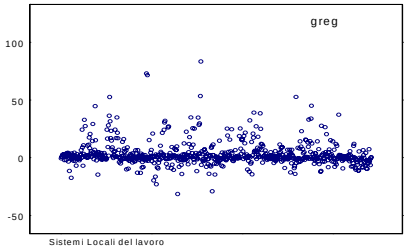
Parametro	R^2	$\hat{\beta}$
Persone in cerca di occupazione	0.9624	1.1110
Occupati	0.9388	1.0634

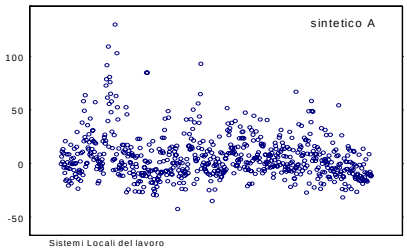
Di conseguenza, al fine di tener conto dell'effetto del riproporzionamento provinciale nella stime dell'errore quadratico medio è sufficiente moltiplicare le stime della radice dell'errore quadratico medio per i coefficienti di regressione riportati nella tabella precedente.

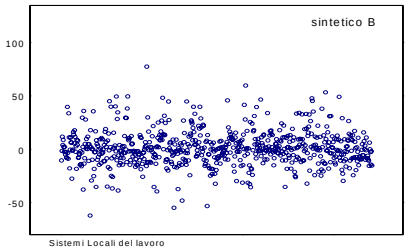
In appendice B sono riportate le stime e i corrispondenti errori per il biennio 2004-05 per le variabili persone in cerca di occupazione e occupati, mentre i due grafici nell'appendice C mostrano l'andamento territoriale del tasso di disoccupazione stimato per i SLL negli anni 2004 e 2005.

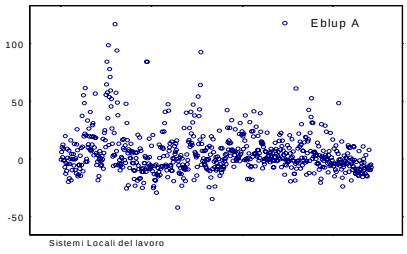
Appendice A

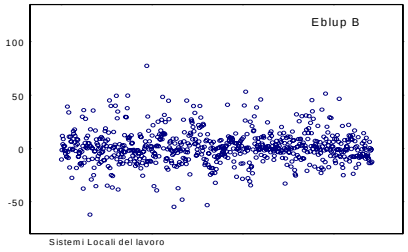
Figura A1: *Distorsione relativa stime persone in cerca di occupazione.*

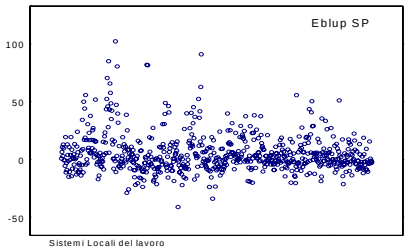


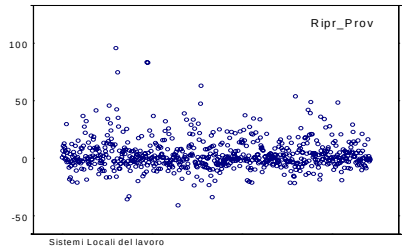


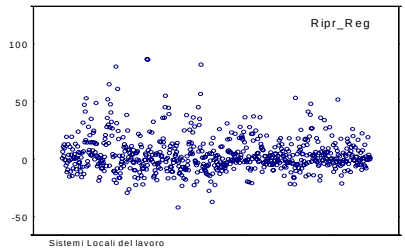


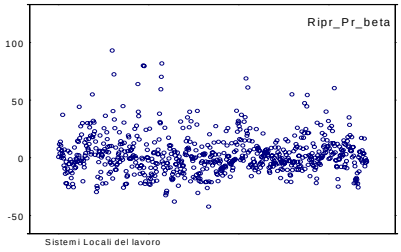



Eblup B
100
50
0
-50
Sistemi Locali del lavoro


Eblup SP
100
50
0
-50
Sistemi Locali del lavoro


Ripr_Prov
100
50
0
-50
Sistemi Locali del lavoro


Ripr_Reg
100
50
0
-50
Sistemi Locali del lavoro


Ripr_Pr_beta
100
50
0
-50
Sistemi Locali del lavoro

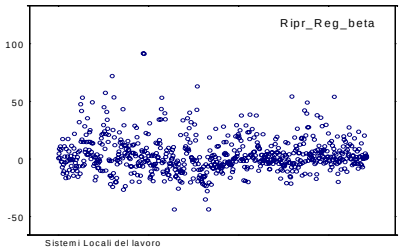
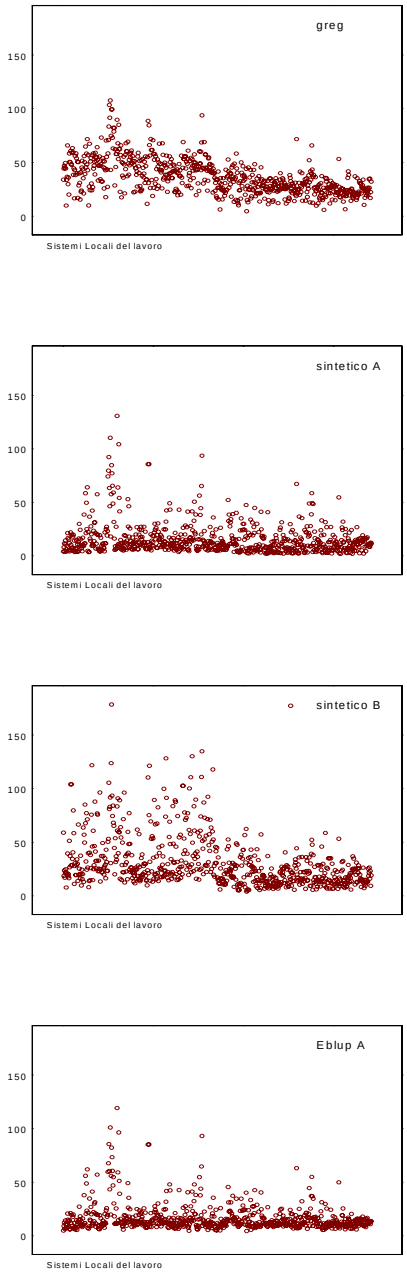
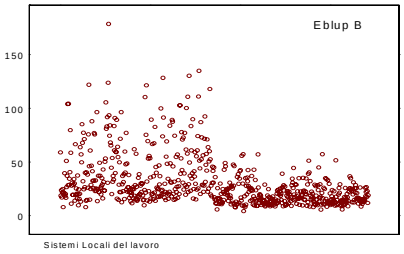
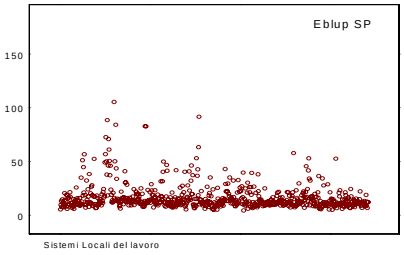
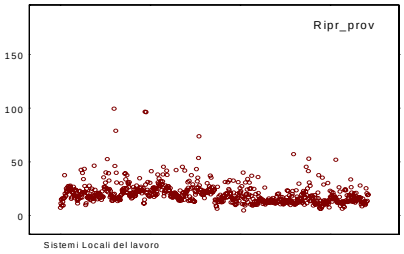

Ripr_Reg_beta
100
50
0
-50
Sistemi Locali del lavoro

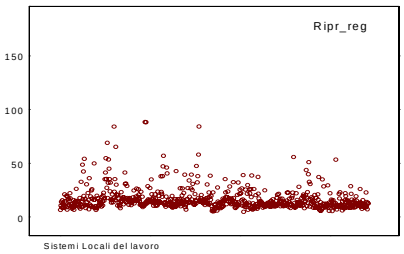
Figura A2: *Radice Errore quadratico medio relativo stime persone in cerca di occupazione.*

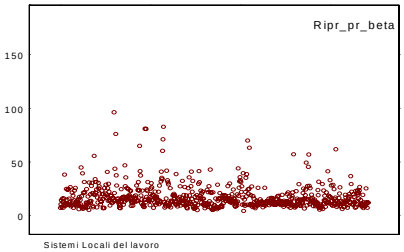



Eblup B
150
100
50
0
Sistemi Locali del lavoro


Eblup SP
150
100
50
0
Sistemi Locali del lavoro


Ripr_prov
150
100
50
0
Sistemi Locali del lavoro


Ripr_reg
150
100
50
0
Sistemi Locali del lavoro


Ripr_pr_beta
150
100
50
0
Sistemi Locali del lavoro

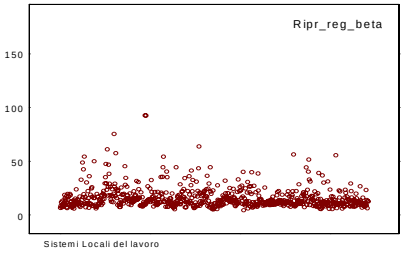
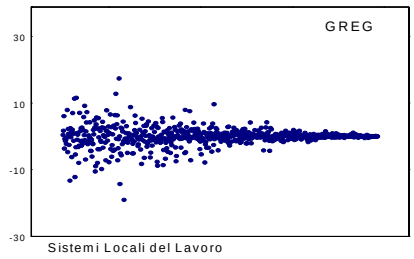
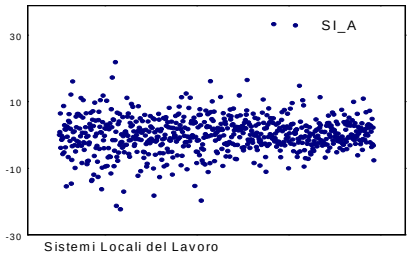
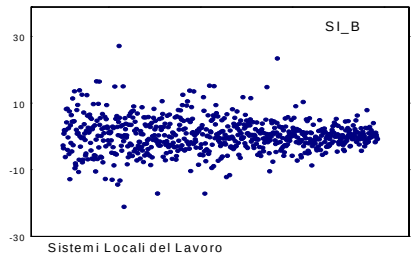
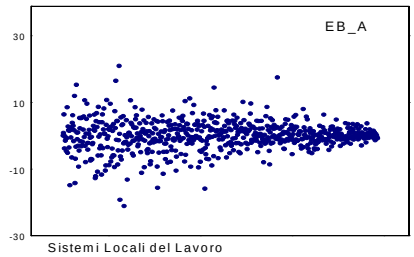

Ripr_reg_beta
150
100
50
0
Sistemi Locali del lavoro

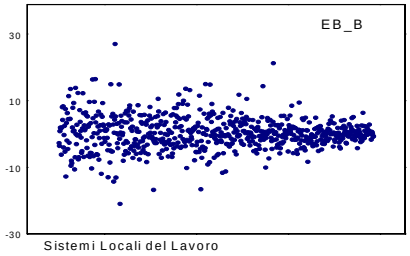
Figura A3: *Distorsione relativa stime occupati.*

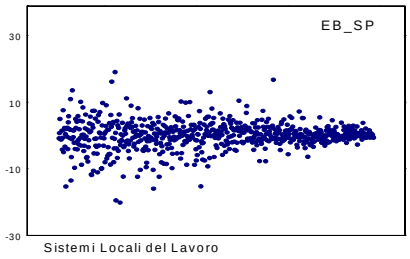


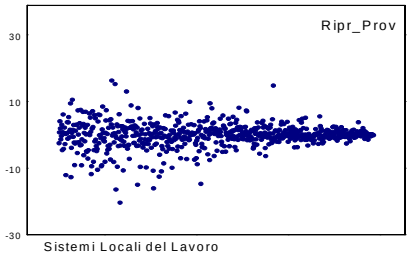


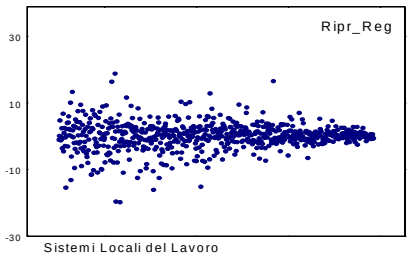


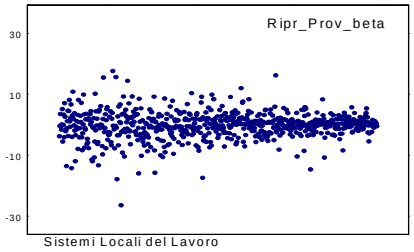



EB_B
30
10
-10
-30
Sistemi Locali del Lavoro


EB_SP
30
10
-10
-30
Sistemi Locali del Lavoro


Ripr_Prov
30
10
-10
-30
Sistemi Locali del Lavoro


Ripr_Reg
30
10
-10
-30
Sistemi Locali del Lavoro


Ripr_Prov_beta
30
10
-10
-30
Sistemi Locali del Lavoro

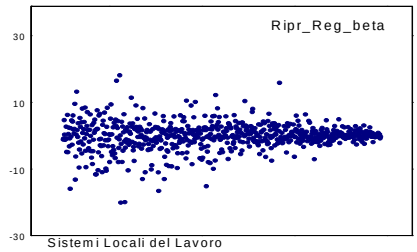
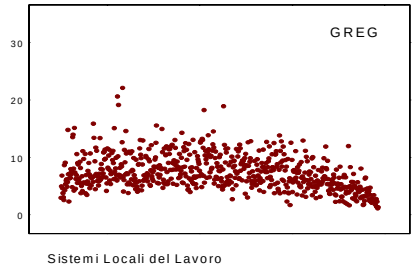
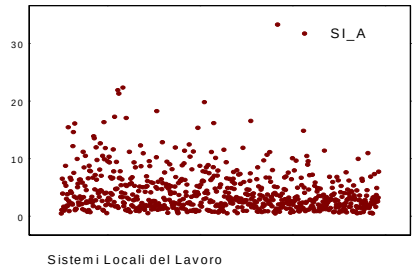
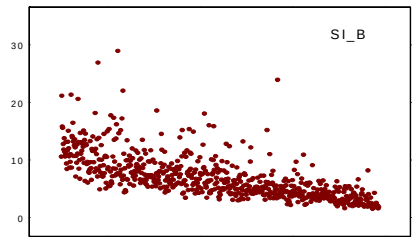
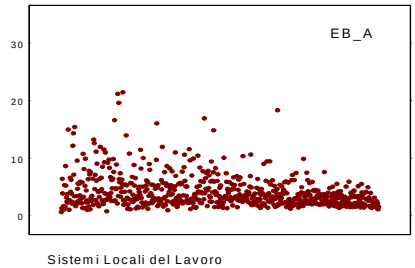

Ripr_Reg_beta
30
10
-10
-30
Sistemi Locali del Lavoro

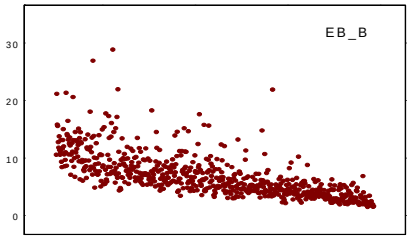
Figura A4: *Radice Errore quadratico medio relativo stime occupati.*



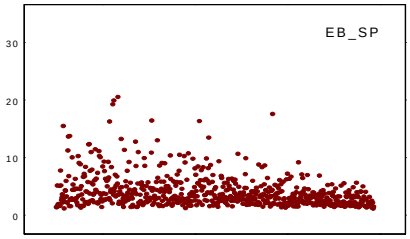




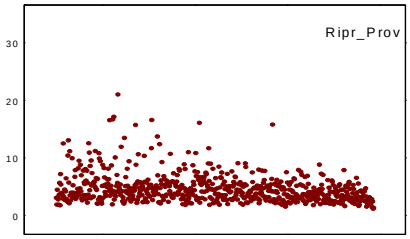




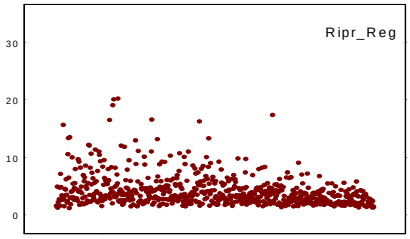
Sistemi Locali del Lavoro



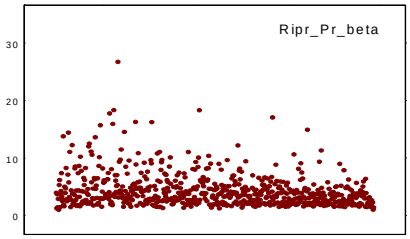
Sistemi Locali del Lavoro

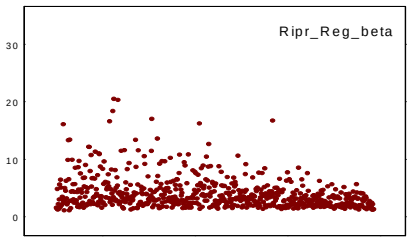


Sistemi Locali del Lavoro



Sistemi Locali del Lavoro


Ripr_Pr_beta
30
20
10
0
Sistemi Locali del Lavoro


Ripr_Reg_beta
30
20
10
0
Sistemi Locali del Lavoro

Appendice B

Figura C1: Distribuzione territoriale del tasso di disoccupazione per l'anno 2004.

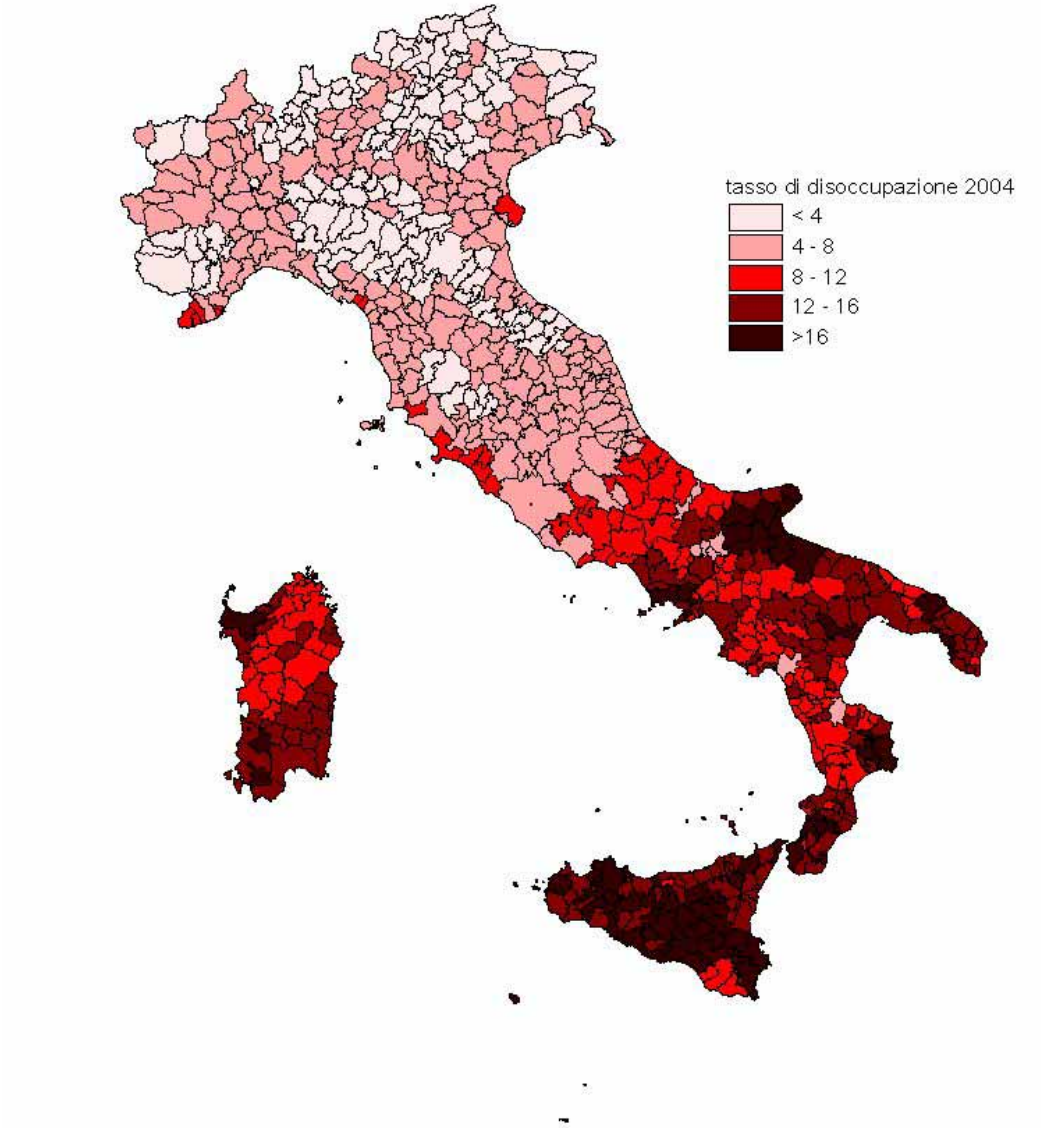
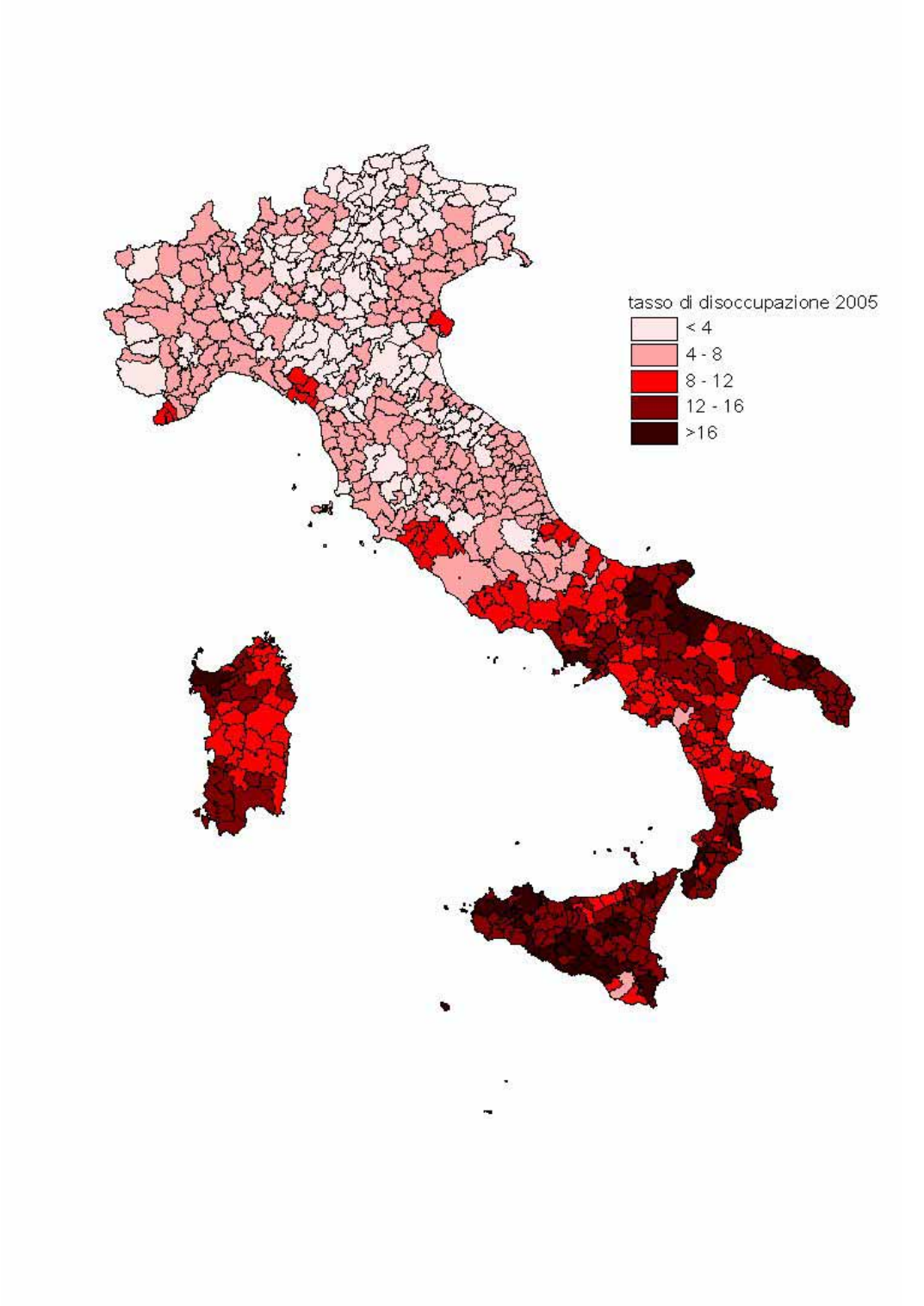


Figura C2: Distribuzione territoriale del tasso di disoccupazione per l'anno 2005.



Riferimenti Bibliografici

Battese, G.E., Harter, R.M., And Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crops Using Survey and Satellite Data. *Journal American Statistical Association*, 83, 28-36.

Cressie, N. (1992). REML Estimation In Empirical Bayes Smoothing of Census Undercount, *Survey Methodology*, 18, 75-94.

D'Alo', M., Falorsi, S., Solari, S., (2004a). Linear Mixed Model with Spatially Correlated Area Effects. SAS Program and Documentation. In *PROJECT REFERENCE VOLUME*. Vol. 3 EURAREA Consortium- https://www.statistics.gov.uk/ eurarea

D'Alo', M., Di Consiglio, L., Falorsi, S., Solari, S., (2004b). The Impact of the Auxiliary Information in the Estimation of Unemployment Rate at Sub-Regional Level: Futher Investigation on the Italian Results in the EURAREA Project. In *Proceeding of the European Conference on Quality and Methodology in Official Statistics* Mainz, Germany, 24-26 Maggio 2004.

D'Alo', M., Di Consiglio, L., Falorsi, S., Solari, S., (2006). Small Area Estimation of the Italian Poverty Rate. *Statistics In Transition*. 7, 771-784.

Deville J.C., Särndal C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

EURAREA Consortium (2004) PROJECT REFERENCE VOLUME, Vol. 1 - https://www.statistics.gov.uk/ eurarea.

Falorsi, P.D., Falorsi, (1996). Indagine sulle Forze di Lavoro: Descrizione della Strategia di Campionamento e Valutazione dell'errore Campionario dei Principali Indicatori Provinciali del Mercato del Lavoro. *Documenti ISTAT* N.1/1996.

Falorsi, P.D., Falorsi, S , Russo, A., (2004). Linear Mixed Model with Correlated Area Time Effects in Small Area Estimation. In *Atti Della XLII Riunione Scientifica SIS*, Bari 9-11 Giugno 2004

Fay, R.E., Herriot, R.A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association,* 74, 269-277.

Ghosh, M., Rao, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, 55-93.

ISTAT (2006). La Rilevazione sulle Forze di Lavoro: Contenuti, Metodologie, Organizzazione. , *Metodi e Norme*, n.32.

Prasad E, Rao J.N.K (1990). The Estimation of the Mean Squared Error of Small Area Estimation, *Journal of the American Statistical Association*, 85, 163-161.

Rao J.N.K, (2003) *Small Area Estimation*. John Wiley & Sons, Hoboken.

Saei, A. And Chambers,R (2004). In *PROJECT REFERENCE VOLUME. Vol. 2 EURAREA Consortium*- https://www.statistics.gov.uk/ eurarea

Searle S.R., Casella G., Mcculloch C.E. (1992). *Variance Components*. John Wiley & Sons, New York