**Università degli Studi di Pisa**
Dipartimento di Statistica e Matematica
Applicata all'Economia

**Report n. 240**

# Stimatore Combinato e Correlazione Spaziale nella Stima per Piccole Aree

**Alessandra Petrucci-Nicola Salvati-Monica Pratesi**

Pisa, Giugno 2003

- Stampato in Proprio –

Via Cosimo Ridolfi, 10 - 56124 PISA - Tel. Segr. Amm. 050 2216231 Segr. Stud. 050 2216317 Fax 050 2216375
Cod. Fisc. 80003670504 - P. IVA 00286820501 - Web http://statmat.ec.unipi.it/ - E-mail: dipstat@ec.unipi.it

# ERRATA CORRIGE

$$\theta_i = x_i^T \beta + \rho_s W_i u_j + u_i^*$$

$$\tilde{\theta}_i = \theta_i + e_i \quad con\ i = 1...m$$

# Stimatore Combinato e Correlazione Spaziale nella Stima per Piccole Aree*

Alessandra Petrucci e Nicola Salvati**

Departement of Statistics "G. Parenti", University of Florence

Monica Pratesi***

Departement of Mathematics and Statistics, University of Pisa

## INDICE



---

* Il lavoro è stato svolto in completa collaborazione dai tre autori, tuttavia il paragrafo 2 è stato curato da Monica Pratesi, il paragrafo 3 è stato curato da Nicola Salvati e i paragrafi 1, 4 e 5 sono stati curati da Alessandra Petrucci.

** *Address for correspondence*: A. Petrucci, N. Salvati, Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni 59, I–50134, Florence, Italy. E–mail:alex@ds.unifi.it e salvati@ds.unifi.it

*** *Address for correspondence*: M.Pratesi, Dipartimento di Matematica e Statistica Applicata all'Economia, Università di Pisa, Via Ridolfi 10, I–50126, Pisa, Italy. E–mail: m.pratesi@ec.unipi.it

**Abstract.** The demand of reliable statistics for small areas, when only reduced sizes of the samples are available, has promoted the development of statistical methods from both the theoretical and empirical point of view. In this perspective the model based methodologies allow for the construction of efficient estimators and their confidence intervals. These small area estimators have several fields of application: from the production of social data to the production of environmental data.

These models consider the random area effects as independent. In practice, it should be more reasonable to assume that the random effects between the neighboring areas (for instance the neighborhood could be define by a distance criterium) are correlated and the correlation decays to zero as distance increases.

The empirical analysis is carried out on some simulated examples obtained on a pseudo real population (Gosh and Rao, 1994). The preliminary results are promising and show an appreciable improvement of the statistical properties of the small area estimators.

## 1. IL PROBLEMA OGGETTO DI STUDIO

Nelle indagini su vasta scala i dati sono ottenuti, solitamente, sulla base di un disegno di campionamento complesso (es.: stratificazione per ripartizioni geografiche e ulteriore suddivisione sulla base di alcuni caratteri delle unità di osservazione) e le dimensioni campionarie sono determinate per garantire l'affidabilità delle stime ad un certo livello prestabilito di ripartizione geografica, dominio o dell'intero territorio nazionale. Tuttavia, sempre più spesso vengono richieste, da enti pubblici o privati, stime relative a territori e domini ridotti e non considerati come domini al momento del disegno dell'indagine (es.: regioni, province, comuni, individui in certe classi di età). Se si utilizzassero per questi domini gli stimatori basati direttamente sul disegno dell'indagine potremmo ottenere stime scarsamente affidabili.

Le soluzioni possibili per questo problema si collocano in due diversi filoni di ricerca. In estrema sintesi le alternative sono:

a) l'incremento della numerosità del campione fino ad ottenere il livello di precisione desiderato per il territorio o il dominio oggetto d'interesse; questo comporta, in fase di organizzazione dell'indagine, la definizione dell'area oggetto di studio per la quale si vogliono produrre le stime e in alcuni casi un elevato aumento dei costi di rilevazione e dei tempi di analisi dei dati;

b) la formulazione di stimatori che consentano di migliorare l'efficienza delle stime rispetto a quello che si otterrebbe sulla base del solo disegno di campionamento.

Questo contributo si colloca nel secondo filone di ricerca.

Lo scopo è valutare la possibilità di migliorare le stime delle caratteristiche d'interesse ($\theta_i$: valore totale, media, etc.) attraverso l'introduzione di un modello di analisi spaziale nello stimatore combinato (Pfeffermann, 2002). A differenza dei comuni modelli di stima per piccole aree, che ipotizzano effetti casuali non correlati tra zone, l'applicazione di un modello spaziale permette di considerare la plausibile presenza di correlazione tra valori del fenomeno in "aree vicine" secondo opportuni criteri di distanza.

Nel lavoro si farà riferimento a "piccole aree" intese come aree geografiche di piccole dimensioni in senso territoriale come province, comuni o sezioni di censimento. Lo studio non si occuperà invece dei "piccoli domini" ovvero gruppi di popolazioni di una data età, sesso o razza all'interno di una grande area geografica come ad esempio una nazione.

Da ora in poi per indicare l'insieme dei metodi di stima per piccole aree si utilizzerà l'acronimo inglese SAE (Small Area Estimation).

## 2. LA METODOLOGIA

Nello studio vengono utilizzati sia metodi di stima propri dell'ambito SAE, sia tecniche caratteristiche dell'impostazione nota come analisi statistica spaziale.

È noto che i metodi SAE sono distinti in metodi "basati su disegno di campionamento" e metodi "basati su modello".

I metodi basati su disegno (o campionari) forniscono direttamente l'espressione di stimatori per il parametro caratteristico di piccola area, senza assumere esplicitamente un modello di distribuzione del fenomeno fra le aree. Tra questi si ricordano:

1) i metodi demografici ("Demographic Methods"), noti anche come SAT (Symptomatic Accounting Techiniques) secondo la denominazione di Purcell e Kish (1980) . Queste tecniche integrano i dati dell'ultimo censimento disponibile con le informazioni dei registri amministrativi a livello locale. I metodi utilizzano variabili "sintomatiche", considerate cioè in stretta relazione con le variazioni di popolazione, e disponibili nelle banche dati delle amministrazioni locali (ad esempio, il numero di nascite e di morti in un certo anno, il numero delle abitazioni esistenti e di quelle di nuova costruzione, il numero di iscrizioni scolastiche). I principali metodi SAT sono la tecnica dei tassi di sopravvivenza ("Vital Rates", Bogue, 1950), il metodo composito ("Composite Method", Bogue and Duncan, 1959), il metodo delle componenti ("The Census Component Method", U.S. Bureau of the Census, 1966), "Administrative Records" (Starsinic, 1974), "Housing Unit" (Smith and Lewis, 1980) e i metodi di regressione campionari (Ericksen, 1974);

2) il metodo degli stimatori sintetici ("Synthetic Estimators", NCHS, 1968; Levy, 1971). In questo caso il parametro caratteristico di piccola area è stimato sulla base di informazioni tratte da aree che si possono definire "simili" rispetto a quella oggetto di studio. Il metodo è apprezzato per la sua semplicità e applicabilità a qualsiasi disegno campionario;

3) i metodi compositi (Shaible, 1979; Särndal, 1981; Fay e Herriot, 1979). Tali metodi combinano stime campionarie classiche e stime sintetiche.

I metodi basati sul modello ("Small Area Models", Holt et. al., 1979), assumono invece esplicitamente un modello di distribuzione del fenomeno tra le aree e prevedono anche la presenza di effetti casuali di area. I parametri possono essere stimati secondo un approccio classico o in modo bayesiano (Gosh e Rao, 1994).

Un ulteriore e più recente impostazione sintetizza gli stimatori ottenuti con i metodi campionari e quelli ricavati con i metodi basati su modello. In questo ambito il presente contributo propone una metodologia di stima che segue la logica dei metodi compositi e propone uno stimatore combinato prevedendo la specificazione di un modello spaziale. Tale modello appartiene alla categoria dei modelli ad effetti misti (Mixed Effects Models) e assume che gli effetti casuali tra "aree vicine" siano correlati.

### 2.1. Stimatore Combinato

Lo stimatore combinato $\hat{\theta}_i^C$, per la i-esima piccola area, viene definito come media ponderata di uno stimatore diretto $\hat{\theta}_i^D$, basato sul disegno, e di uno stimatore indiretto $\hat{\theta}_i^I$, ottenuto dalla specificazione del modello, e può essere scritto come:

$$\hat{\theta}_i^C = w_i\hat{\theta}_i^D + (1 - w_i)\hat{\theta}_i^I \quad con\ i = 1...m, \qquad [1]$$

dove $w_i$ rappresenta un peso opportunamete scelto ($0 < w_i < 1$) e $m$ è il numero delle piccole aree.

Lo stimatore diretto del valore d'interesse $\theta_i$ relativo all'area i-esima è ottenuto sulla base dei valori della variabile obiettivo osservati con riferimento alle sole unità del campione appartenenti alla piccola area $i$. Tale stimatore è corretto sulla base del disegno, ma può essere caratterizzato da una variabilità molto elevata.

Gli stimatori indiretti del carattere d'interesse $\theta_i$ tentano di ridurre la variabilità degli stimatori diretti utilizzando informazioni ausiliarie. Tali informazioni possono essere relative all'area più ampia che contiene la piccola area, oppure disponibili a livello di piccola area stessa. Il legame fra la variabile di studio e le informazioni ausiliarie è generalmente formalizzato attraverso un modello.

Lo stimatore combinato bilancia la potenziale distorsione di uno stimatore indiretto e l'instabilità di uno stimatore diretto.

Molti degli stimatori proposti in letteratura assumono la forma [1]. Anche la specificazione di modelli ad effetti misti per la distribuzione del fenomeno tra le aree, conduce a stimatori di tipo combinato.

Nell'applicazione degli stimatori combinati il peso $w_i$ riveste un ruolo importante. Esso può essere ottenuto minimizzando l'errore quadratico medio (Mean Square Error - MSE) dello stimatore combinato rispetto al peso stesso e assumendo $cov(\hat{\theta}_i^D, \hat{\theta}_i^I) = 0$:

$$w_i(opt) = \frac{MSE(\hat{\theta}_i^I)}{[MSE(\hat{\theta}_i^I) + V(\hat{\theta}_i^D)]}. \qquad [2]$$

Il peso $w_i(opt)$ può essere stimato sostituendo a $MSE(\hat{\theta}_i^I)$ e a $V(\hat{\theta}_i^D)$ le rispettive stime campionarie. I pesi che si ottengono possono risultare molto instabili (Gosh e Rao, 1994).

In questo studio si pone particolare attenzione al calcolo del peso $w_i$. Per la stima del parametro caratteristico di area si combinano le stime dirette e le stime ottenute da un modello ad effetti misti. L'elemento di novità è costituito dall'inserimento in questo modello di una componente che tenga conto della correlazione presente tra "aree vicine" secondo un opportuno criterio di distanza.

Un'ultima questione riguarda la presenza delle aree oggetto d'indagine nel campione estratto. Infatti si distingue il caso in cui almeno una unità della piccola area sia presente

nel campione, dal caso in cui nessuna unità della zona analizzata sia stata estratta nel campione. In quest'ultimo caso sono necessarie ulteriori precisazioni sulla metodologia di stima che rimandiamo a ricerche future.

### *2.2. Modelli a Effetti Misti*

Uno dei modelli più semplici è il "Nested Error Unit Level Regression Model" (Battese et al., 1988); si suppone di conoscere sia il valore di variabili ausiliarie $x_1...x_p$ per ogni unità del campione, sia il valore medio di area delle stesse variabili ausiliarie.

Il modello assume la forma

$$y_{ij} = x_{ij}^{'}\beta + u_i + \epsilon_{ij} \qquad [3]$$

$$i = 1...m;\ j = 1...n_i$$

con $x_{ij}$ valore della variabile ausiliaria per l'unità $j$ nell'area $i$, $u_i$ e $\epsilon_{ij}$ termini di errore mutuamente indipendenti con media 0 e varianza rispettivamente $\sigma_u^2$ e $\sigma_\epsilon^2$; $u_i$ rappresenta l'effetto casuale dell'area, mentre $\epsilon_{ij}$ si riferisce all'effetto casuale delle unità campionarie; $m$ è il numero delle piccole aree e $n_i$ rappresenta il numero delle unità campionarie dell'area i-esima.

Il modello

$$\bar{Y}_i = \bar{X}_i\beta + u_i + \bar{\epsilon}_i \quad con\ i = 1...m \qquad [4]$$

esprime la relazione funzionale tra valori medi di area e le variabili ausiliarie di area con $\bar{\epsilon}_i = \sum_{j=1}^{N_i} \frac{\epsilon_{ij}}{N_i} \cong 0$ per elevato numero delle unità dell'area i-esima ($N_i$).

Da quanto sopra deriva che i parametri obiettivo, cioè i valori del carattere oggetto di studio (valore totale, media), possono essere definiti come:

$$\theta_i = \bar{X}_i\beta + u_i \quad con\ i = 1...m \qquad [5]$$

e nel caso in cui si conoscano le varianze $\sigma_u^2$, $\sigma_\epsilon^2$, il migliore stimatore (predittore) lineare non distorto (Best Linear Unbiased Predictor - BLUP) di $\theta_i$ è:

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{X}_i - \bar{x}_i)^{'}\hat{\beta}_{GLS}] + (1-\gamma_i)\bar{X}_i^{'}\hat{\beta}_{GLS} \quad con\ i = 1...m, \qquad [6]$$

dove $\hat{\beta}_{GLS}$ è lo stimatore di $\beta$ ottenuto applicando i minimi quadrati generalizzati (Generalized Least Square - GLS) ai dati osservati e $\gamma_i = \frac{\sigma_u^2}{(\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_i})}$ rappresenta il peso $w_i$ di uno stimatore combinato. Il coefficiente $\gamma_i$ è noto come *"shrinkage factor" - s.f. -* (Ghosh e Rao, 1994). Esso è ricavato minimizzando il $MSE(\hat{\theta}_i)$ rispetto a $\gamma_i$. In questo modo lo s.f. $\gamma_i$ risulta essere quello migliore dato il modello prescelto ([6]).

Le varianze $\sigma_u^2$ e $\sigma_\epsilon^2$ raramente sono conosciute; vengono quindi stimate con procedure standard per le componenti di varianza (stima di massima verosimiglianza (MLE); stima MLE ristretta (Cressie, 1992)). I predittori ottenuti sono noti come Empirical BLUP di $\theta_i$.

Nel corso degli anni sono state mosse alcune critiche all'uso di questo modello di stima. In sostanza si ritiene che i predittori non colgano la variabilità tra le medie di area; in altre parole le variazioni tra i predittori sono più piccole delle variazioni tra le medie di area. In particolare Thomsen (Gosh e Rao, 1994) e Thomsen e Holmoy (1998) ritengono che i predittori lineari nella forma [6] tendono a sovrastimare il valore medio di area con limitati effetti casuali e sottostimarlo quando gli effetti casuali sono elevati.

Un altro modello, frequentemente applicato e discusso ampiamente in letteratura, è quello ad effetti casuali a livello di area ("Area Level Random Effects Model") utilizzato per la prima volta da Fay e Herriot (1979). Il modello è applicato quando si hanno a disposizione informazioni ausiliarie solo a livello di piccola area. Il modello è definito come:

$$\theta_i = x_i'\beta + u_i$$

$$\tilde{\theta}_i = \theta_i + e_i \quad con\ i = 1...m, \qquad [7]$$

dove $\tilde{\theta}_i$ è lo stimatore campionario diretto (ad esempio la media $\bar{y}_i$), $e_i$ rappresenta l'errore di campionamento con media nulla e varianza nota $Var_D(e_i) = \sigma^2_{Di}$ e $m$ numero delle piccole aree. Il modello presentato considera, quindi, sia gli effetti casuali di area $u_i$, sia gli errori di campionamento $e_i$ e assume la loro indipendenza.

Il predittore BLUP di $\theta_i$ è:

$$\hat{\theta}_i = \gamma_i\tilde{\theta}_i + (1-\gamma_i)x_i'\hat{\beta}_{GLS}$$

$$= x_i'\hat{\beta}_{GLS} + \gamma_i(\tilde{\theta}_i - x_i'\hat{\beta}_{GLS}) \quad con\ i = 1...m \qquad [8]$$

che è ancora uno stimatore composito con peso $\gamma_i = \frac{\sigma^2_u}{(\sigma^2_u+\sigma^2_{Di})}$. Anche in questo caso le varianze $\sigma^2_u$ e $\sigma^2_{Di}$ sono generalmente sconosciute e stimate da campione. Come nel caso precedente, il predittore che si ottiene è definito Empirical BLUP.

Tutti i modelli presentati precedentemente ipotizzano che gli effetti casuali di area siano indipendenti. In pratica sarebbe più ragionevole assumere che gli effetti casuali tra "aree vicine" siano correlati, con correlazione che tende a 0 all'aumentare della distanza. Tale ipotesi, comune nell'analisi spaziale (Cressie, 1993), non trova ancora applicazione nella stima per piccole aree.

In un recente contributo di Pfeffermann (2002) lo s.f. è ottenuto ipotizzando che le aree siano correlate, anche se in modo non spaziale. Pfeffermann (2002) ha considerato un modello del tipo [7], privo di covariate e con uguale numerosità campionaria ($n_i = n$) in ogni area.

Il modello ottenuto risulta:

$$\bar{y}_i = \mu + u_i + e_i = \theta_i + e_i;\ Var_D(e_i) = \frac{\sigma^2_e}{n} = \sigma^2;\ Var(u_i) = \sigma^2_u\ con\ i = 1...m$$

con $u_i$ e $e_i$ indipendenti entro e fra le aree, $\sigma^2$ e $\sigma^2_u$ conosciuti, $\mu$ sconosciuto ed $m$ numero delle piccole aree. Applicando questo modello, il predittore ottimale (BLUP) del valore

medio della piccola area i-esima è:

$$\hat{\theta}_i = \gamma \bar{y}_i + (1-\gamma)\bar{y} \quad con\ i = 1...m \qquad [9]$$

con $\bar{y} = \sum_{i=1}^{m} \frac{\bar{y}_i}{m}$ e $\gamma = \frac{\sigma_u^2}{(\sigma_u^2+\sigma^2)}$. Si può notare che lo s.f., sulla base delle semplificazioni applicate, è uguale per ogni piccola area $i$.

Se si suppone l'esistenza di correlazione semplice tra gli effetti di area, $Corr(u_i, u_k) = \rho$ con $\rho > 0$ per $i \neq k$, il valore ottimale del predittore BLUP del valore medio della piccola aree i-esima assume la stessa forma del predittore [9]:

$$\theta_i^* = \tilde{\gamma} \bar{y}_i + (1-\tilde{\gamma})\bar{y} \quad con\ i = 1...m \qquad [10]$$

tuttavia cambia lo s.f.:

$$\tilde{\gamma} = \frac{[\sigma_u^2(1-\rho)]}{[\sigma_u^2(1-\rho)+\sigma^2]}. \qquad [11]$$

La domanda a cui si cerca di rispondere è quanto si perde in termini di efficienza utilizzando $\hat{\theta}_i$ invece di $\theta_i^*$. È stato dimostrato da Pfeffermann (2002) che all'aumentare del numero delle aree ($m$) l'inserimento della correlazione semplice tra le aree migliora l'efficienza relativa dello stimatore $\theta_i^*$, infatti $\lim_{m\to\infty} \frac{MSE(\theta_i^*)}{MSE(\hat{\theta}_i)} = \frac{[\tilde{\gamma}]}{[1-\rho(1-\gamma)]}$, che tende a 0 al tendere di $\rho$ ad 1. Da questo discende che, nel caso in cui la correlazione $\rho$ tra gli effetti casuali di area sia elevata, la perdita di efficienza delle stime, qualora si usi $\hat{\theta}_i$ invece che $\theta_i^*$, è significativa. I passaggi algebrici per il calcolo del s.f. nel caso in cui non esista correlazione ($\rho = 0$) tra gli effetti casuali di area e nel caso in cui si assuma la presenza di correlazione ($Corr(u_i, u_k) = \rho$ con $\rho > 0$ per $i \neq k$) tra aree sono riportati in APPENDICE A.

## 3. INSERIMENTO DELLA CORRELAZIONE SPAZIALE NELLA STIMA PER PICCOLE AREE

Pfeffermann (2002) nel suo studio fa riferimento alla correlazione tra aree, ma non considera la posizione geografica delle aree stesse. Appare più realistico considerare la posizione di ciascuna area rispetto alle altre e assumere che gli effetti casuali tra "aree vicine" sulla base di un opportuno criterio di distanza siano correlati, con correlazione che tende ad annullarsi all'aumentare della distanza stessa.

In presenza di correlazione spaziale si può assumere che il vettore degli effetti di area $u = \rho_s W u + u^*$ con $\rho_s$ coefficiente di autoregressione spaziale, $u^* \sim N(0, \sigma^2_{u^*})$ disturbo e

$$u \sim (0, \sigma^2_{u^*}(I - \rho_s W)^{-1}(I - \rho_s W^T)^{-1}). \qquad [12]$$

Infatti:

$$u = \rho_s W u + u^*$$

$$u^* = u - \rho_s W u = u(I - \rho_s W)$$

$$\Rightarrow u = \frac{u^*}{(I - \rho_s W)}$$

$$\Rightarrow var(u) = \sigma^2_{u^*}(I - \rho_s W)^{-1}(I - \rho_s W^T)^{-1}.$$

Nei modelli di regressione lineare, la dipendenza spaziale può essere introdotta in due modi: attraverso un regressore nella forma di *lag* spaziale ($Wy$) (modello 1); oppure nella struttura dell'errore ($E[u_i, u_j] \neq 0$) (modello 2). Il primo modello viene utilizzato per valutare l'esistenza e la forza dell'interazione spaziale. Il secondo modello è appropriato quando si tende a correggere la potenziale distorsione determinata dalla correlazione spaziale tra i dati, indipendentemente dal fatto che il modello d'interesse sia spaziale oppure non spaziale.

Formalmente un modello spaziale autoregressivo (modello 1) per la variabile di studio $y$ è espresso nella forma:

$$y = \rho_s W y + X\beta + \epsilon \qquad [13]$$

dove $\rho_s$ è il coefficiente di autoregressione spaziale, $\epsilon$ rappresenta il vettore degli errori ($n \times 1$), $W$ è la matrice dei pesi ($n \times n$), $X$ è una matrice $n \times k$, dove $k$ è il numero delle variabili esplicative e $n$ rappresenta il numero delle osservazioni; infine $y$ è il vettore delle variabili risposta ($n \times 1$).

Nel caso di modelli con errori spazialmente autocorrelati (modello 2) la struttura spaziale può essere specificata in modi diversi determinando una diversa struttura della matrice varianze-covarianze che assume la forma:

$$E[u_i, u_j] = \Omega(\tau)$$

dove $\tau$ è il vettore dei parametri che possono riguardare un processo SAR (Spatial Autoregressive) o SMA (Spatial Moving Average). Un modello con struttura di errori SAR è del tipo:

$$y = X\beta + u \qquad [14]$$

con $u = \rho_s W u + u^*$. Da cui otteniamo per $y$:

$$y = X\beta + (I - \rho_s W)^{-1}u \Rightarrow y = \rho W y + X\beta - \rho_s W X\beta + u. \qquad [15]$$

Il modello per $y$ è con *lag* spaziale, con variabili esogene legate spazialmente ($WX$) e un insieme di $k$ di vincoli sui coefficienti.

Nell'ambito di questo lavoro si assume il modello 2:

$$\theta_i = x_i'\beta + \rho_s W_i u_j$$

$$\tilde{\theta}_i = \theta_i + e_i \quad con\ i = 1...m \qquad [16]$$

dove $\tilde{\theta}_i$ è lo stimatore campionario diretto (ad esempio la media $\bar{y}_i$), $e_i$ rappresenta l'errore di campionamento con media nulla e varianza conosciuta $Var_D(e_i) = \sigma^2_{Di}$ e $m$ è il numero delle piccole aree.

Il predittore BLUP di $\theta_i$ dato il modello 2 assume la forma di uno stimatore composito:

$$\breve{\theta}_i = \breve{\gamma}_i\tilde{\theta}_i + (1 - \breve{\gamma}_i)x_i'\hat{\beta}_{ML} \quad con\ i = 1...m \qquad [17]$$

con peso $\breve{\gamma}_i$. Assumendo che le le variabili ausiliarie $x_i$ siano note a livello di area, la stima dei parametri $\hat{\beta}_{ML}$ può essere ottenuta attraverso la massimizzazione della verosimiglianza (Ord, 1975):

$$\hat{\beta}_{ML} = [(X - \rho_s WX)^T(X - \rho_s WX)]^{-1}(X - \rho_s WX)^T(y - \rho_s Wy). \qquad [18]$$

Per ottenere lo s.f. ($\breve{\gamma}_i$), come è accaduto per calcolare il fattore di restringimento in caso di correlazione semplice, il modello adottato è quello nullo con l'ipotesi aggiuntiva di uguale numerosità campionaria in ogni area:

$$\bar{y}_i = \mu + \rho_s W_i u_j + u_i^* + e_i = \theta_i + e_i \quad con\ i = 1...m. \qquad [19]$$

Il predittore BLUP del parametro $\theta_i$ (valore medio della piccola area) assume la forma[1]:

$$\breve{\theta}_i = \breve{\gamma}_i\tilde{\theta}_i + (1 - \breve{\gamma}_i)\bar{y} \quad con\ i = 1...m. \qquad [20]$$

[1] Da ora in poi si indica la $var(u_i)$ con $\sigma^2_{u_i}(\rho_s, W)$ e la $cov(u_i, u_j)$ con $\sigma_{u_i,u_j}(\rho_s, W)$. Inoltre si considera $u^* \sim N(0,1)$ e che il coefficiente di autoregressione spaziale ($\rho_s$) sia conosciuto, altrimenti sarebbe necessario stimarlo iterativamente, cosa che sarà analizzata in un momento successivo.

Lo s.f. si ottiene calcolando il $MSE(\breve{\theta}_i)$ e minimizzando la funzione rispetto a $\breve{\gamma}$:

$$\breve{\gamma}_i = \frac{\sigma^2_{u_i}(\rho_s, W)[1-\frac{2}{m}] + \frac{1}{m^2}[\sum_{i=1}^{m}(\sigma^2_{u_i}(\rho_s, W)) + 2\sum_i \sum_{j>i} \sigma_{u_i,u_j}(\rho_s, W)] -}{\sigma^2_{u_i}(\rho_s, W)[1-\frac{2}{m}] + \frac{1}{m^2}[\sum_{i=1}^{m}(\sigma^2_{u_i}(\rho_s, W)) + 2\sum_i \sum_{j>i} \sigma_{u_i,u_j}(\rho_s, W)] -}$$

$$\frac{-\frac{2}{m}\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s, W)}{-\frac{2}{m}\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s, W) + \sigma^2(1-\frac{1}{m})} \qquad [21]$$

$$con\ i = 1...m.$$

Si osserva che in questo caso lo s.f. varia da area ad area in quanto dipende dal numero di vicini di ciascuna area, che si riflette sulla struttura della matrice varianze-covarianze. $\breve{\gamma}_i$ rappresenta un minimo della funzione e varia tra 0 e 1. Tutti i calcoli relativi alla determinazione dello s.f. sono riportati in APPENDICE B. L'efficienza relativa dello stimatore $\breve{\theta}_i$ rispetto allo stimatore $\theta_i^*$, si misura attraverso il rapporto tra $MSE(\breve{\theta}_i)$, considerando la correlazione spaziale, e $MSE(\theta_i^*)$ nel caso di correlazione semplice fra i valori di area:

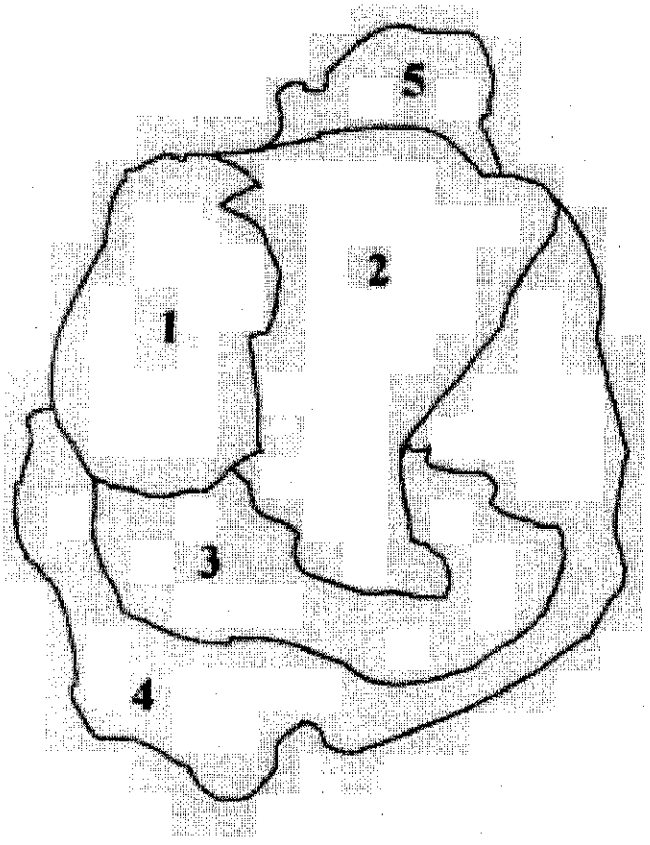
$$\frac{\sigma^2_{u_i}(\rho_s,W)(1-\tilde{\gamma})^2+\sigma^2(\tilde{\gamma}^2+\frac{(1-\tilde{\gamma}^2)}{m})+(1-\tilde{\gamma})^2\frac{1}{m^2}\{\sum_{i=1}^{m}\sigma^2_{u_i}(\rho_s,W)+2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s,W)\}-}{\tilde{\gamma}^2(\sigma^2_u+\sigma^2)+\frac{1}{m}\{(\sigma^2_u+\sigma^2)+(m-1)\rho\sigma^2_u\}(1-\tilde{\gamma}^2)+}$$

$$\frac{-2\frac{(1-\tilde{\gamma})}{m}[\sigma^2_{u_i}(\rho_s,W)+\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)]+2\frac{\tilde{\gamma}(1-\tilde{\gamma})}{m}[\sigma^2_{u_i}(\rho_s,W)+\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)]}{+\sigma^2_u(1-2\tilde{\gamma})-2(1-\tilde{\gamma})\frac{1}{m}[\sigma^2_u+(m-1)\rho\sigma^2_u]}.$$

Per analizzare come variano i $\tilde{\gamma}_i$ cambiando il coefficiente di autoregressione spaziale si è considerato un semplice esempio costituito da cinque regioni fittizie collegate tra loro secondo un certo schema di vicinanza (FIGURA 1).

Lo schema di vicinanza scelto è quello più semplice dove $w_{ij}$ è uguale a 1 se le aree sono contigue, altrimenti è 0.

FIGURA 1. Rappresentazione grafica delle 5 aree.



La matrice dei pesi è quindi:

$$W = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

La matrice di varianze-covarianze viene calcolata come soluzione del predetto prodotto matriciale:

$$\sigma^2_{u^*} \cdot \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} - \rho \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \right\}^{-1} \otimes$$

$$\otimes \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} - \rho \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \right\}^{-1}$$

dove $\rho_s$ esprime l'effetto di interazione spaziale e $\sigma^2_{u^*}$ rappresenta la variabilità tra regioni dovuta solo all'effetto area e non alla correlazione spaziale tra aree. Fissato $\rho_s = 0.25$ e $\sigma^2_{u^*} = 1$ risulta:

$$\begin{bmatrix} 4.96 & 4.64 & 4.32 & 4.32 & 1.38 \\ 4.64 & 5.62 & 4.64 & 4.64 & 1.87 \\ 4.32 & 4.64 & 4.96 & 4.32 & 1.38 \\ 4.32 & 4.64 & 4.32 & 4.96 & 1.38 \\ 1.38 & 1.87 & 1.38 & 1.38 & 1.58 \end{bmatrix}$$

mentre i $\tilde{\gamma}_i$ risultano:

$$\begin{bmatrix} \tilde{\gamma}_1 & 0.80 \\ \tilde{\gamma}_2 & 0.81 \\ \tilde{\gamma}_3 & 0.80 \\ \tilde{\gamma}_4 & 0.80 \\ \tilde{\gamma}_5 & 0.85 \end{bmatrix}$$

Fissando $\rho_s = 0.75$ la matrice varianze-covarianze diventa:

$$\begin{bmatrix} 0.47 & -0.07 & 0.15 & 0.15 & -0.34 \\ -0.07 & 0.57 & -0.07 & -0.07 & 0.62 \\ 0.15 & -0.07 & 0.47 & 0.15 & -0.34 \\ 0.15 & -0.07 & 0.15 & 0.47 & -0.34 \\ -0.34 & 0.62 & -0.34 & -0.34 & 1.16 \end{bmatrix}$$

e i $\tilde{\gamma}_i$ sono:

$$\begin{bmatrix} \tilde{\gamma}_1 & 0.35 \\ \tilde{\gamma}_2 & 0.26 \\ \tilde{\gamma}_3 & 0.35 \\ \tilde{\gamma}_4 & 0.35 \\ \tilde{\gamma}_5 & 0.61 \end{bmatrix}$$

Si nota come all'aumentare della correlazione spaziale i $\tilde{\gamma}_i$ assumono valori sempre più vicini allo 0 e quindi nell'espressione [21] la componente legata alle stime dirette risulta meno influente di quella da modello.

### *3.1. Le Simulazioni*

Allo scopo di mostrare empiricamente la distribuzione che assume lo s.f. $\tilde{\gamma}_i$ e di studiare l'efficienza dello stimatore $\breve{\theta}_i$ rispetto a $\theta_i^*$ sono state realizzate alcune simulazioni.

La procedura adottata per eseguire le simulazioni ha previsto in primo luogo l'estrazione di un valore $\rho_s$ da una funzione di densità di probabilità uniforme con limite inferiore e superiore pari al campo di variazione del coefficiente di autoregressione ($-1 < \rho_s < 1$). In questo modo sono state attribuite le stesse probabilità di essere estratto a ciascun valore compreso tra $-1$ e $1$[2].

Anche la varianza da disegno $\sigma^2$ è stata ottenuta campionando da una funzione di densità di probabilità uniforme tra 1 e 100. La FIGURA 2 mostra la distribuzione di frequenza di $\rho_s$ e di $\sigma^2$: sull'asse delle ascisse sono rappresentati rispettivamente i valori del coefficiente di autoregressione spaziale e della varianza da disegno mentre sull'asse delle ordinate si trova la frequenza assoluta dei valori. Il terzo grafico della FIGURA 2 rappresenta la distribuzione di frequenza congiunta di $\rho_s$ e $\sigma^2$. Ci sono alcune coppie di valori che assumono una frequenza maggiore delle altre, ma in definitiva la superficie che si ottiene è abbastanza piatta: questo significa che sono state campionate numerose coppie di valori fra tutte quelle possibili. La procedura di campionamento di $\rho_s$ e di $\sigma^2$ è stata utilizzata per tutte le simulazioni presentate.

Il criterio di definizione del numero $m$ di piccole aree è stato invece diverso. Alcune simulazioni sono state eseguite fissando $m = 5$ (FIGURE 3 e 5). In altri casi $m$ è generato casualmente utilizzando una funzione di densità di probabilità uniforme con limite inferiore 5 e limite superiore 100.

Per quanto riguarda la determinazione della struttura di vicinanza, si è proceduto nel seguente modo. È stato fissato il numero $m$ delle piccole aree ed è stato attribuito valore 1 ai pesi $w_{ij}$ nel caso in cui il valore estratto da una funzione di densità uniforme $[0, 1]$ fosse maggiore di 0.5, altrimenti è stato assegnato valore 0.

Una volta stabiliti i valori di $\rho_s$, $\sigma^2$, $m$ e fissata la struttura di vicinanza, è stato calcolato il valore di $\tilde{\gamma}_i$.

La FIGURA 3 mostra l'istogramma di frequenza di $\tilde{\gamma}_i$. Esso è simile a quella di una funzione esponenziale negativa con una elevata frequenza dei valori di $\tilde{\gamma}_i$ inferiori a 0.2. Questo significa che nell'espressione dello stimatore BLUP $\breve{\theta}_i$ di $\theta_i$ viene attribuito maggior

[2] Si osserva che questa assunzione comporta la conoscenza del coefficiente di autoregressione spaziale, che altrimenti poteva essere stimato in modo iterativo.

peso alle stime da modello rispetto a quelle dirette. Le simulazioni rappresentate in FIGURA 4 confermano l'andamento della distribuzione di $\check{\gamma}_i$. In questo caso il numero delle aree non è fissato a priori, ma varia fra 5 e 100. Rispetto ai risultati delle simulazioni precedenti si nota l'elevata frequenza dei valori di $\check{\gamma}_i$ intorno all'unità. Questo può essere spiegato dal fatto che lo s.f. nel caso di correlazione spaziale ([21]) è il rapporto fra due quantità uguali, ad eccezione della presenza di $\sigma^2$ al denominatore, quindi quando la varianza da disegno è molto piccola lo s.f. tende ad 1.

Le FIGURE 5 e 6 mettono in rilievo le differenze tra la distribuzione di frequenza dello s.f. con correlazione spaziale ([21]), la distribuzione di frequenza dello s.f. con correlazione semplice ([11]) e quella dello s.f. senza correlazione. In tutte e due le figure si evidenzia la peculiarità della distribuzione di $\check{\gamma}_i$ che assume valori prossimi allo 0 con maggiore frequenza rispetto a $\tilde{\gamma}_i$ e a $\gamma_i$.

Le simulazioni realizzate suggeriscono che i $\check{\gamma}_i$ hanno un campo di variazione compreso tra 0 e 1 e assumono una distribuzione quasi iperbolica, se non fosse per un incremento della frequenza dei valori nell'intorno di 1. Se si confrontano le distribuzioni di $\check{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$, si può notare come l'esistenza di correlazione spaziale determini una frequenza di valori vicini allo 0 nella distribuzione dei $\check{\gamma}_i$ molto più elevata rispetto alle distribuzioni di $\tilde{\gamma}_i$ e $\gamma_i$.

FIGURA 2. Rappresentazione della distribuzione di $\sigma^2$ e $\rho_s$, con $m = 5$ (5000 simulazioni).

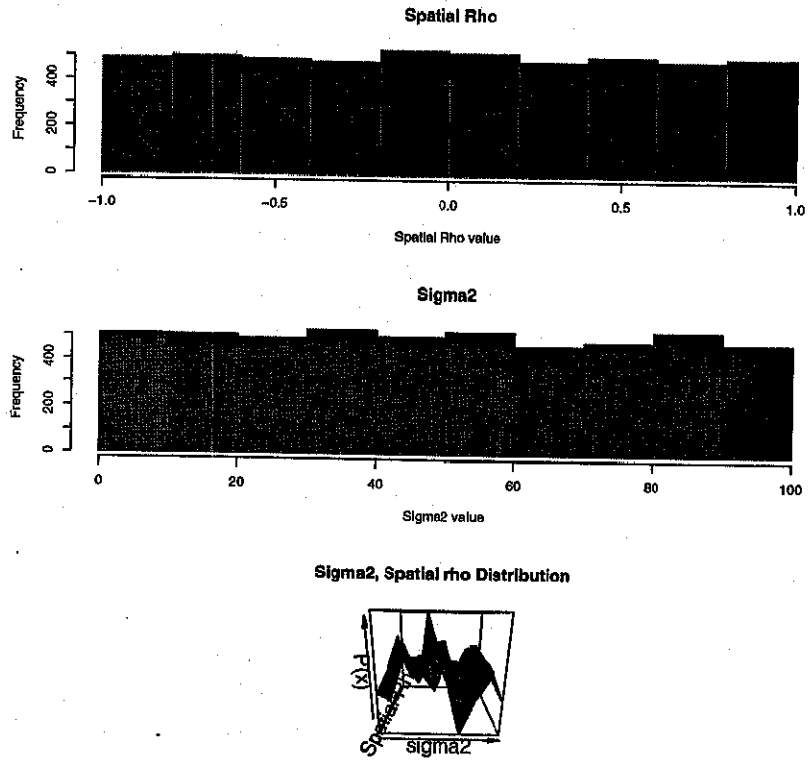



FIGURA 3. Rappresentazione della distribuzione di frequenza di $\breve{\gamma}_i$ con $m = 5$ (5000 simulazioni).

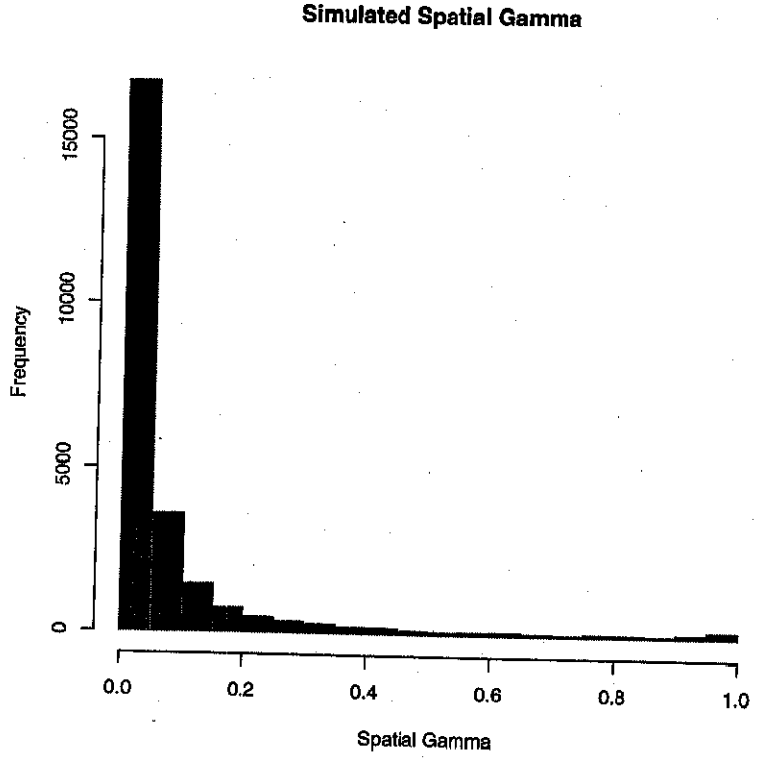


FIGURA 4. Rappresentazione della distribuzione di $\tilde{\gamma}_i$, e delle coppie di $\rho_s$ e $\sigma^2$ con $m$ che varia tra 5 e 100 (2000 simulazioni).

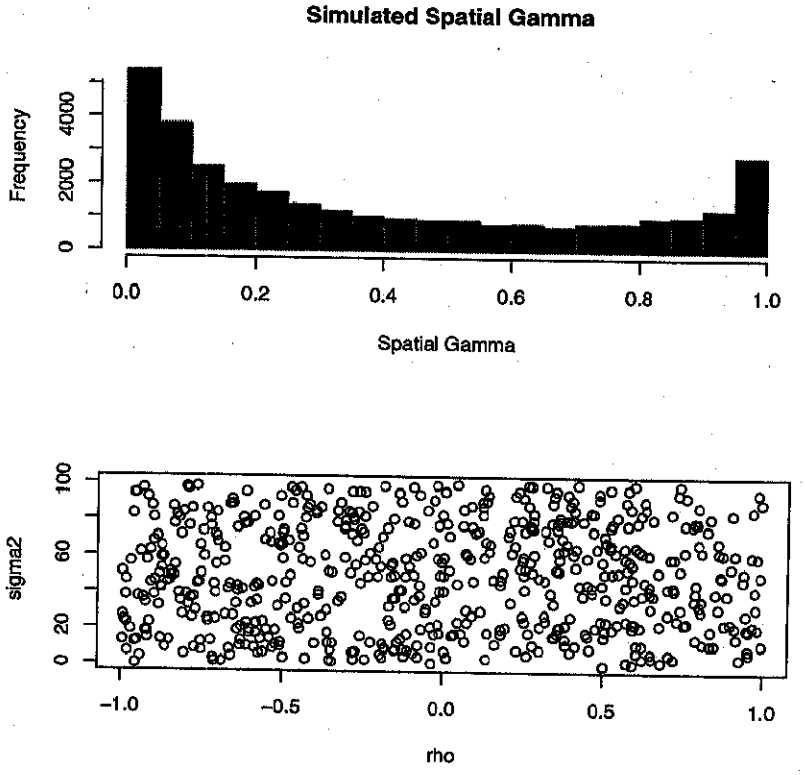



FIGURA 5. Rappresentazione della distribuzione di $\tilde{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$ con $m = 5$ (5000 simulazioni).

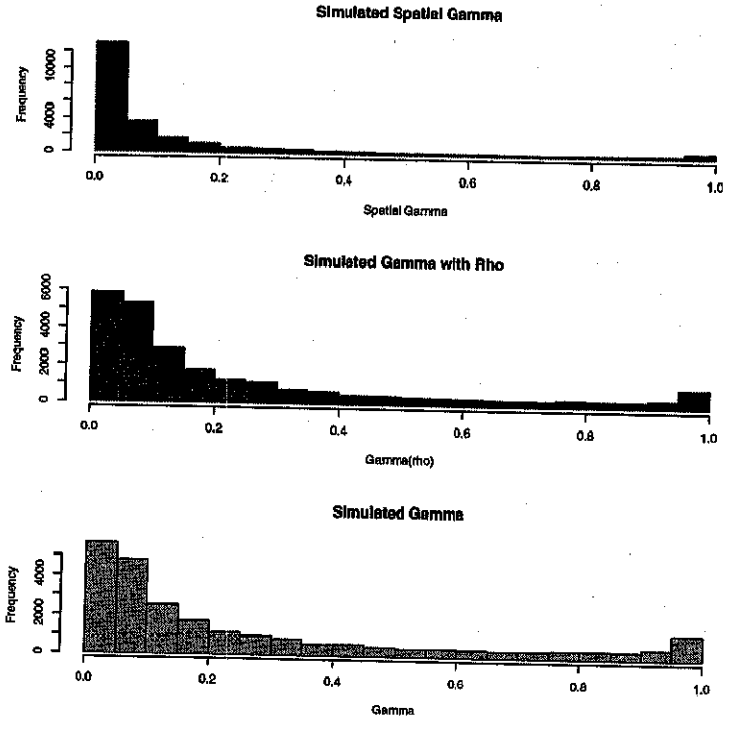
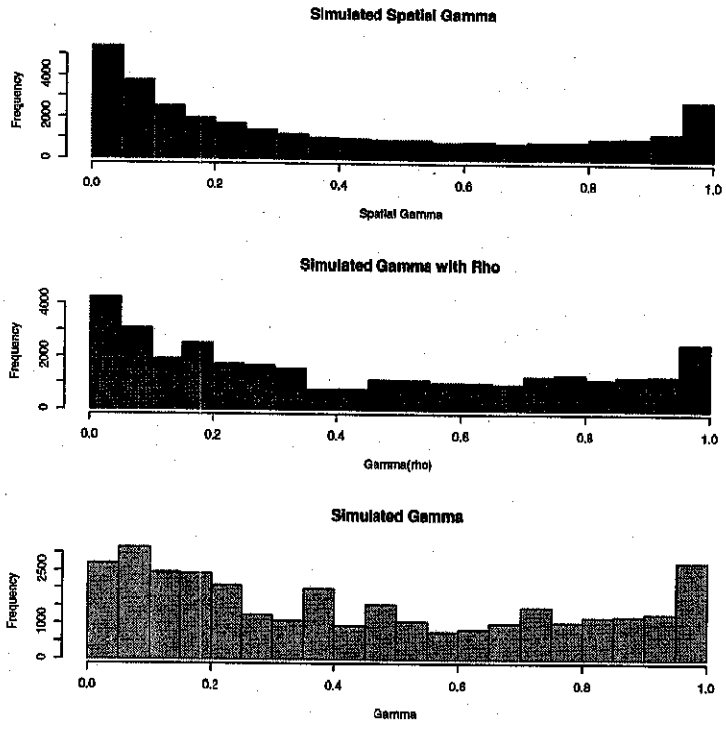


FIGURA 6. Rappresentazione della distribuzione di $\tilde{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$ con $m$ compreso tra 5 e 100 (2000 simulazioni).




Per verificare se l'utilizzo del modello con errori spazialmente correlati migliora l'efficienza dello stimatore BLUP $\breve{\theta}_i$ rispetto agli stimatori BLUP $\theta_i^*$ e $\hat{\theta}_i$, sono state eseguite ulteriori simulazioni formulando alcune ipotesi che riguardano $\rho$, $\rho_s$, $m$ e $\sigma^2$. La procedura di simulazione per definire la struttura di vicinanza è uguale a quella utilizzata per le simulazioni realizzate precedentemente. Per quanto riguarda gli altri parametri sono stati fissati e combinati tra loro. Tutti gli scenari e i risultati sono riportati nella TABELLA 1. La tabella deve essere letta considerando i valori assunti da tre variabili:

a) $m$, che rappresenta il numero delle piccole aree, assume tre valori: 5, 20, 50;

b) $\rho$ e $\rho_s$ esprimono rispettivamente il coefficiente di correlazione semplice e il coefficiente di autoregressione spaziale; per ciascun coefficiente si assumono valori significativi e si combinano in modo da ottenere 6 possibili risultati: $\rho = 0.5$ e $\rho_s = 0.5$, correlazione positiva ed uguale; $\rho = 0.25$ e $\rho_s = 0.75$, correlazione spaziale positiva e maggiore della correlazione semplice; $\rho = 0.75$ e $\rho_s = 0.25$, correlazione spaziale positiva e minore della correlazione semplice; le stesse combinazioni sono replicate per valori negativi dei coefficienti. Infine, per completare i risultati è stato considerato il caso di assenza di correlazione;

c) $\sigma^2$, che esprime la varianza da disegno ed assume tre valori: $\sigma^2 = 1$, $\sigma^2 = 10$ e $\sigma^2 = 100$.

TABELLA 1. Verifica empirica dell'efficienza dello stimatore BLUP $\breve{\theta}_i$.

| | $\rho$ | $\rho_s$ | $\sigma^2 = 1$ | $\sigma^2 = 10$ | $\sigma^2 = 100$ |
|---|---|---|---|---|---|
| $m = 5$ | 0.5 | 0.5 | 0.74 | 0.51 | 0.77 |
| | No corr. | 0.5 | 0.71 | 0.41 | 0.63 |
| | 0.25 | 0.75 | 0.73 | 0.37 | 0.43 |
| | No corr. | 0.75 | 0.72 | 0.36 | 0.38 |
| | 0.75 | 0.25 | 0.91 | 0.87 | 0.97 |
| | No corr. | 0.25 | 0.74 | 0.52 | 0.79 |
| | −0.5 | −0.5 | 0.93 | 0.83 | 0.94 |
| | No corr. | −0.5 | 1.00 | 0.96 | 0.98 |
| | −0.25 | −0.75 | 1.03 | 1.06 | 1.02 |
| | No corr. | −0.75 | 1.07 | 1.15 | 1.05 |
| | −0.75 | −0.25 | 0.82 | 0.64 | 0.87 |
| | No corr. | −0.25 | 0.89 | 0.80 | 0.94 |
| $m = 20$ | 0.5 | 0.5 | 0.56 | 0.17 | 0.18 |
| | No corr. | 0.5 | 0.55 | 0.16 | 0.12 |
| | 0.25 | 0.75 | 0.55 | 0.15 | 0.08 |
| | No corr. | 0.75 | 0.54 | 0.14 | 0.07 |
| | 0.75 | 0.25 | 0.60 | 0.29 | 0.47 |
| | No corr. | 0.25 | 0.56 | 0.18 | 0.20 |
| | −0.5 | −0.5 | 0.61 | 0.25 | 0.37 |
| | No corr. | −0.5 | 0.63 | 0.30 | 0.47 |
| | −0.25 | −0.75 | 0.66 | 0.33 | 0.50 |
| | No corr. | −0.75 | 0.67 | 0.37 | 0.56 |
| | −0.75 | −0.25 | 0.58 | 0.20 | 0.27 |
| | No corr. | −0.25 | 0.60 | 0.25 | 0.38 |
| $m = 50$ | 0.5 | 0.5 | 0.52 | 0.12 | 0.05 |
| | No corr. | 0.5 | 0.52 | 0.11 | 0.04 |
| | 0.25 | 0.75 | 0.52 | 0.11 | 0.03 |
| | No corr. | 0.75 | 0.52 | 0.11 | 0.03 |
| | 0.75 | 0.25 | 0.54 | 0.16 | 0.15 |
| | No corr. | 0.25 | 0.52 | 0.12 | 0.06 |
| | −0.5 | −0.5 | 0.54 | 0.15 | 0.11 |
| | No corr. | −0.5 | 0.55 | 0.16 | 0.15 |
| | −0.25 | −0.75 | 0.56 | 0.17 | 0.16 |
| | No corr. | −0.75 | 0.57 | 0.18 | 0.19 |
| | −0.75 | −0.25 | 0.53 | 0.13 | 0.08 |
| | No corr. | −0.25 | 0.53 | 0.14 | 0.11 |

Per ogni possibile combinazione dei tre parametri si è misurata l'efficienza relativa dello stimatore BLUP $\breve{\theta}_i$ rispetto allo stimatore BLUP $\theta_i^*$ calcolando il rapporto tra il $MSE(\breve{\theta}_i)$ e $MSE(\theta_i^*)$. Le stesse modalità di calcolo sono state seguite per misurare l'efficienza relativa dello stimatore BLUP $\breve{\theta}_i$ rispetto allo stimatore BLUP $\hat{\theta}_i$ $(\frac{MSE(\breve{\theta}_i)}{MSE(\hat{\theta}_i)})$.

La TABELLA 1 mostra come il rapporto tra gli $MSE$ sia sempre inferiore ad 1 tranne nel caso in cui si registra una correlazione spaziale negativa di $-0.75$ e una correlazione semplice di $-0.25$ o assenza di correlazione con il numero delle aree uguale a 5. Questo può essere spiegato considerando il significato della correlazione spaziale negativa che si presenta quando aree vicine assumono valori diversi; mentre nel caso di correlazione spaziale positiva si formano dei gruppi di aree con valori omogenei e quindi il modello migliora l'efficienza degli stimatori, nel caso di correlazione spaziale negativa il modello, con un numero limitato di aree, non può tradurre l'informazione in un miglioramento delle stime. Invece, all'aumentare del numero di aree $(m)$ i rapporti tra i $MSE$ assumono valori molto inferiori all'unità, evidenziando un miglioramento sostanziale delle stime di area.

In conclusione è possibile affermare che all'aumentare del numero delle aree oggetto di studio l'introduzione della posizione geografica delle aree stesse e della correlazione spaziale permette di ottenere un miglioramento significativo dell'efficienza dello stimatore BLUP di $\theta_i$.

FIGURA 7. Distribuzione di $\breve{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$ con $m = 5$, $\rho_s > 0$, $\rho = 0$ e $\sigma^2 = 1$ (5000 simulazioni).

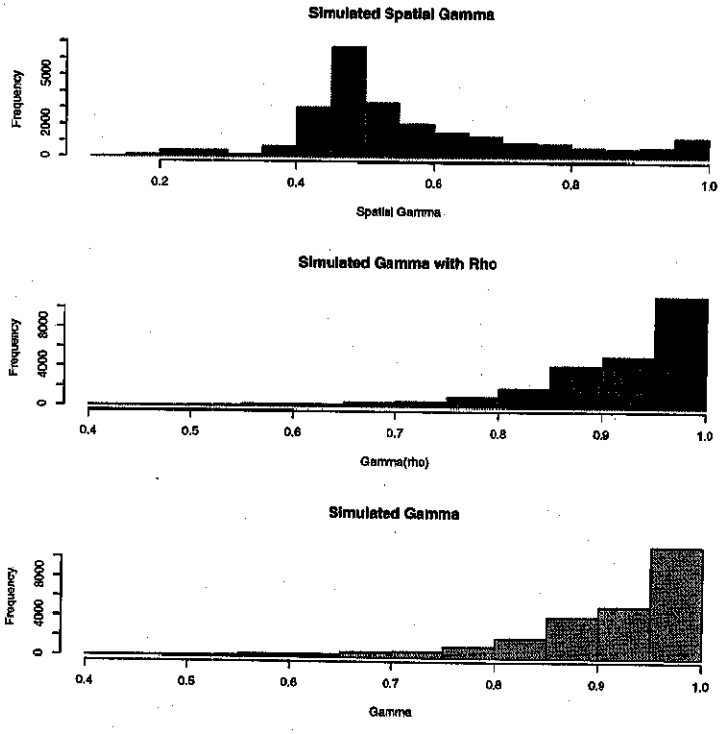
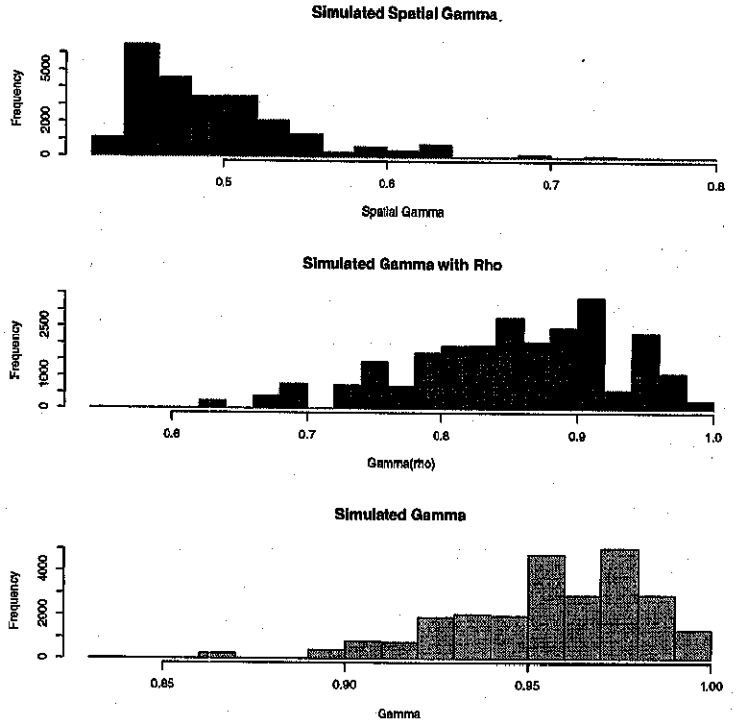


FIGURA 8. Distribuzione di $\tilde{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$ con $m = 5$, $\rho_s = 0.75$, $\rho = 0.5$ e $\sigma^2 = 100$ (5000 simulazioni).




Le distribuzioni di $\tilde{\gamma}_i$, $\tilde{\gamma}_i$ e $\gamma_i$ in due dei 108 possibili risultati delle simulazioni realizzate sono riportate in FIGURA 7 e 8. Nel caso in cui non ci sia correlazione semplice ($\rho = 0$) e con varianza piccola, uguale a 1, la distribuzione del fattore di restringimento si concentra attorno al valore unitario; si attribuisce, ovviamente, maggior peso alle stime dirette rispetto a quelle da modello che non apportano alcuna informazione (FIG. 7). La distribuzione di $\tilde{\gamma}_i$ assume quasi una forma campanulare centrata in 0.5: anche se la varianza da disegno è piccola, l'informazione apportata dal coefficiente di interazione spaziale permette di attribuire, nello stimatore BLUP di $\theta_i$, un peso elevato alle stime da modello.

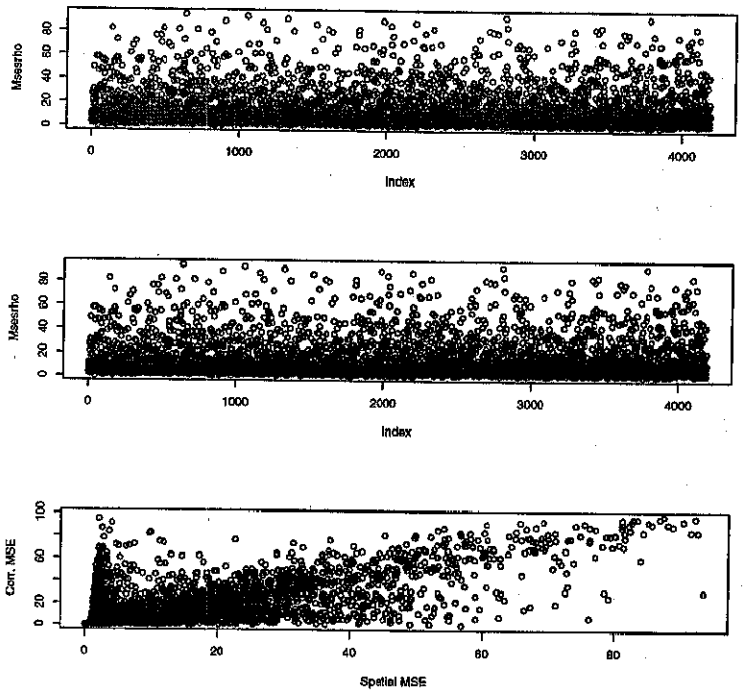
Nella FIGURA 8 la distribuzione di $\tilde{\gamma}_i$ si differenzia dalle altre due anche nel caso di elevata varianza da disegno ($\sigma^2 = 100$). La differenza è molto evidente rispetto alla distribuzione di $\gamma_i$ ($\rho = 0$), e merita comunque di essere sottolineata anche rispetto alla distribuzione di $\tilde{\gamma}_i$ ($\rho = 0.5$).

Un'ultima simulazione è stata eseguita per studiare l'andamento di $MSE(\breve{\theta}_i)$ e di $MSE(\theta_i^*)$. I valori del numero delle aree ($5 \leq m \leq 100$), del coefficiente di autoregressione spaziale ($-1 < \rho_s < 1$), del coefficiente di correlazione semplice ($-1 < \rho < 1$) e la varianza da disegno ($1 < \sigma^2 < 100$) sono stati campionati dalle rispettive distribuzioni uniformi definite sui campi di variazione. I risultati per $MSE(\breve{\theta}_i)$ e $MSE(\theta_i^*)$ sono riportati in FIGURA 9. In particolare il terzo grafico è una combinazione di due grafici precedenti ed esprime per ogni valore di $MSE(\breve{\theta}_i)$ il valore assunto da $MSE(\theta_i^*)$. Se si considera la retta bisettrice del primo e terzo quadrante, si può dire che tutti i punti che si trovano al disopra di questa retta esprimono una migliore efficienza dello stimatore BLUP $\breve{\theta}_i$; invece

tutti i punti che si trovano al disotto della retta testimoniano una migliore efficienza dello stimatore BLUP $\theta_i^*$. Si nota che i punti si trovano soprattutto al di sopra della bisettrice.

Tutte le procedure di simulazione sono state implementate con un programma ad hoc in ambiente *"R"*.

FIGURA 9. Rappresentazione (scatter) di $MSE(\breve{\theta}_i)$ e $MSE(\theta_i^*)$ con $5 \leq m \leq 100$, $-1 < \rho_s < 1$, $-1 < \rho < 1$ e $1 < \sigma^2 < 100$ (5000 simulazioni).



## 4. UNA VERIFICA EMPIRICA

Allo scopo di verificare le possibilità applicative della metodologia sviluppata, è stato realizzato un esperimento simulato e sono stati stimati i valori medi ($\hat{\bar{y}}_i$) di 16 piccole aree in cui è suddivisa un'ipotetica zona. Al fine di conoscere i veri valori medi per ogni piccola area ($\bar{Y}_i$) e quindi poter calcolare successivamente degli indici che permettono di verificare la bontà delle stime, sono stati utilizzati dati noti in letteratura (Gosh e Rao, 1994) sulla base di uno studio di popolazione di Särndal e Hidiroglou (1989). Sono state introdotte alcune semplificazioni che tuttavia non pregiudicano il risultato: per ogni piccola area si ipotizza un numero uguale di unità campionarie ($n_i = n \; \forall \; i$); si suppone di conoscere la varianza da disegno e si fissa pari a 1/10 di $\sigma_u^2$; infine ciascun valore ottenuto dalla stima diretta è stata attribuito casualmente ad ogni piccola area in cui è suddivisa l'ipotetica provincia (FIGURA 10).

FIGURA 10. Le 16 aree oggetto di studio.


In TABELLA 2 sono riportate le stime dirette $\tilde{\bar{y}}_i$ insieme al valore vero per ogni zona ($\bar{Y}_i$). È stato calcolato il valore dell'interazione spaziale, $\rho_s$, a livello di popolazione utilizzando i valori veri secondo lo schema di vicinanza che risulta dalla contiguità delle aree (FIGURA 10) e risulta pari a $-0.077$; mentre il coefficiente di correlazione $\rho$ con *lag* di 1 vale $-0.085$.

Si sono quindi calcolati i valori di $\gamma_i$, $\tilde{\gamma}_i$ e $\check{\gamma}_i$ (TABELLA 3). Si può notare come $\gamma_i$ e $\tilde{\gamma}_i$ sono uguali per ogni area e sono molto vicini all'unità, danno cioè molto peso, nello stimatore BLUP del parametro $\theta_i$, alla stima diretta, mentre $\check{\gamma}_i$ è diverso da area ad area,

essendo calcolato considerando la struttura di vicinanza, e attribuisce quasi lo stesso peso alla stima diretta e a quella da modello.

TABELLA 2. Stime dirette del valore medio $\tilde{\bar{y}}_i$ e valore vero $\bar{Y}_i$ per ogni area.

| Area | $\tilde{\bar{y}}_i$ | $\bar{Y}_i$ |
|---|---|---|
| 1 | 0 | 24.22 |
| 2 | 12.18 | 20.43 |
| 3 | 2.54 | 5.48 |
| 4 | 0 | 6.55 |
| 5 | 3.61 | 20.55 |
| 6 | 16.46 | 14.85 |
| 7 | 20.77 | 21.46 |
| 8 | 10.84 | 13.40 |
| 9 | 16.05 | 15.56 |
| 10 | −6.34 | 5.88 |
| 11 | 18.07 | 15.20 |
| 12 | 1.70 | 13.40 |
| 13 | 0 | 26.06 |
| 14 | 4.27 | 22.44 |
| 15 | 5.97 | 9.40 |
| 16 | 53.83 | 29.49 |

Utilizzando i valori di $\gamma_i$, $\tilde{\gamma}_i$ e $\check{\gamma}_i$, le medie di area sono state stimate con i predittori BLUP $\hat{\theta}_i$, $\theta_i^*$ e $\check{\theta}_i$.

La bontà delle stime è stata verificata attraverso l'uso di due indici (Gosh e Rao, 1994):

1) Average Square Error

$$ASE = \frac{1}{m}\sum_{i=1}^{m}(\hat{\bar{y}}_i - \bar{Y}_i)^2$$

2) Average Relative Error

$$ARE = \frac{1}{m}\sum_{i=1}^{m}\frac{(|\,\hat{\bar{y}}_i - \bar{Y}_i\,|)}{\bar{Y}_i}.$$

I risultati sono riportati in TABELLA 4 e mostrano come l'utilizzo del modello che tiene conto dell'informazione spaziale migliori le stime rispetto al modello con correlazione semplice e a quello senza correlazione. È importante sottolineare che in presenza di correlazione semplice negativa le stime sono peggiori rispetto al caso di assenza di correlazione.

TABELLA 3. $\gamma_i$, $\tilde{\gamma}_i$ e $\check{\gamma}_i$ per ogni area.

| Area | $\gamma_i$ | $\tilde{\gamma}_i$ | $\check{\gamma}_i$ |
|---|---|---|---|
| 1 | 0.90 | 0.91 | 0.53 |
| 2 | 0.90 | 0.91 | 0.50 |
| 3 | 0.90 | 0.91 | 0.50 |
| 4 | 0.90 | 0.91 | 0.52 |
| 5 | 0.90 | 0.91 | 0.50 |
| 6 | 0.90 | 0.91 | 0.52 |
| 7 | 0.90 | 0.91 | 0.53 |
| 8 | 0.90 | 0.91 | 0.52 |
| 9 | 0.90 | 0.91 | 0.53 |
| 10 | 0.90 | 0.91 | 0.53 |
| 11 | 0.90 | 0.91 | 0.51 |
| 12 | 0.90 | 0.91 | 0.50 |
| 13 | 0.90 | 0.91 | 0.53 |
| 14 | 0.90 | 0.91 | 0.51 |
| 15 | 0.90 | 0.91 | 0.51 |
| 16 | 0.90 | 0.91 | 0.51 |

TABELLA 4. Calcolo degli indici ASE e ARE.

| Modello | ASE | ARE |
|---|---|---|
| $\rho_s$ | 93.23 | 0.58 |
| $\rho$ | 160.00 | 0.59 |
| No corr. | 158.55 | 0.58 |

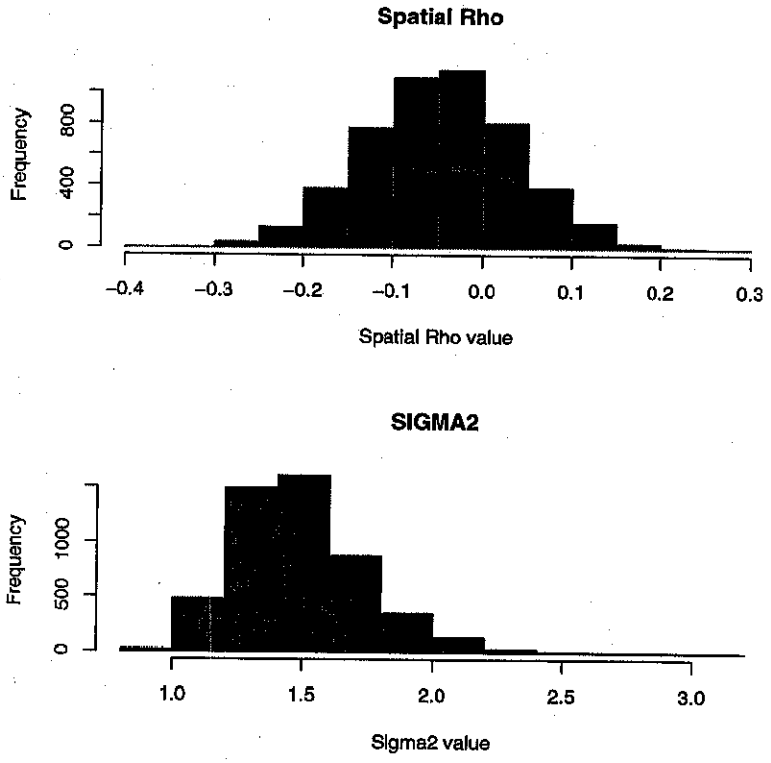
Allo scopo di verificare che i risultati ottenuti non derivano dalla particolare attribuzione dei valori alle aree, sono state eseguite delle simulazioni ipotizzando la scambiabilità delle aree e quindi immaginando diversi schemi di vicinanza. In altre parole, ogni area cambia la sua posizione rispetto alle altre aree in modo da ottenere strutture di vicinanza diverse, mentre mantiene il valore che gli era stato attribuito in modo casuale.

La correlazione semplice non varia nel caso si modifichi lo schema di vicinanza e quindi non cambiano né i valori di $\gamma_i$, né quelli di $\tilde{\gamma}_i$ e di conseguenza non variano le stime realizzate con lo stimatore BLUP. Si vuol verificare se i due indici $ARE$ e $ASE$, che indicano la bontà di adattamento delle stime, assumano distribuzioni di frequenza con valori inferiori a quelli assunti nel caso in cui venga applicato lo stimatore BLUP con correlazione semplice o in assenza di correlazione.

La FIGURA 11 mostra la distribuzione di frequenza del coefficiente di correlazione spaziale. I valori sono compresi tra $-0.3$ e $0.3$; questo significa che, comunque si modifichi lo schema di vicinanza, la correlazione spaziale non risulta elevata.

La distribuzione di frequenza della varianza da disegno ($\sigma^2$) assume una forma quasi campanulare centrata sul valore 1.5 che, nel calcolo dello stimatore BLUP di $\theta_i$, dovrebbe portare a valori dello s.f. vicini all'unità e cioè quasi tutto il peso delle stime sarebbe da imputare allo stimatore diretto. Invece, la distribuzione dello s.f., riportata in FIGURA 12, è caratterizzata da valori più frequenti intorno a 0.40. Quindi in presenza di correlazione spaziale, seppur debole, le stime da modello nello stimatore BLUP $\breve{\theta}_i$ hanno un peso mediamente maggiore di quello delle stime dirette.

FIGURA 11. Distribuzione di frequenza dei valori di $\rho_s$ e di $\sigma^2$ (5000 simulazioni).

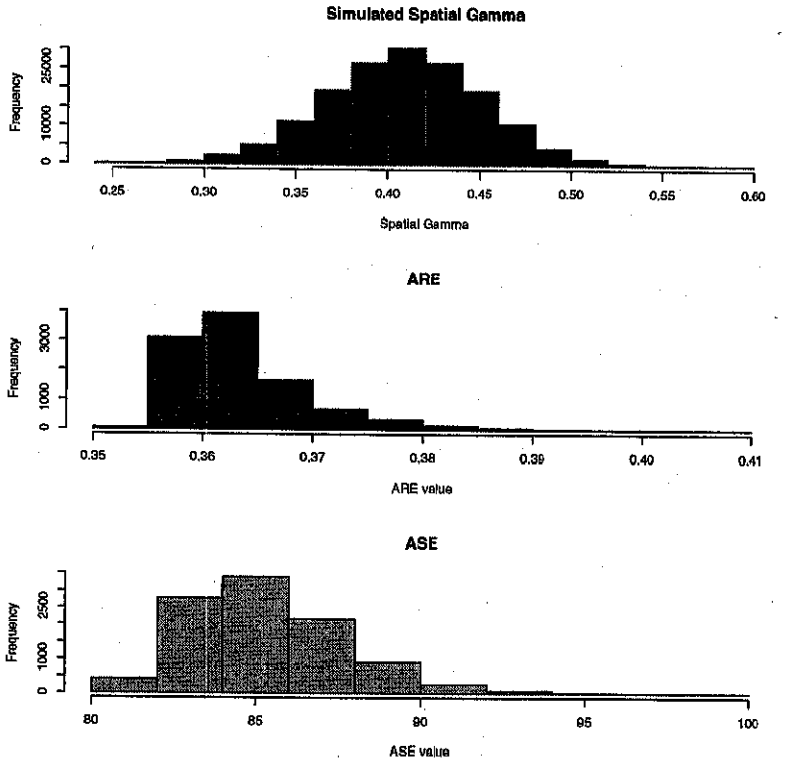



Per verificare l'impatto della diversa distribuzione dei $\breve{\gamma}_i$ sull'efficienza delle stime si sono calcolati i due indici *ARE* e *ASE*. La loro distribuzione di frequenza è rappresentata in FIGURA 12. Per entrambi gli indici il valore massimo del campo di variazione è inferiore al valore degli stessi indici calcolati nel caso di stime prodotte con lo stimatore BLUP $\hat{\theta}_i$ (assenza di correlazione) e $\theta_i^*$ (presenza di correlazione semplice) e sono riportati in TABELLA 4.

I grafici in FIGURA 12 evidenziano la similitudine della distribuzione di frequenza dei due indici: entrambe le distribuzioni sono caratterizzate da una marcata asimmetria a destra.

I risultati ottenuti in questo esperimento sembrano indicare che introducendo nello stimatore combinato un modello spaziale, secondo la metodologia presentata nei paragrafi precedenti, all'aumentare del numero di piccole aree, si ottiene un miglioramento della bontà delle stime, anche in presenza di debole interazione spaziale.

FIGURA 12. Distribuzione di frequenza di $\tilde{\gamma}_i$, dell'indice $ARE$ e $ASE$ (5000 simulazioni).



## 5. NOTE CONCLUSIVE

Questo studio introduce l'informazione spaziale nella metodologia di stima per piccole aree e mostra come questa consenta di migliorare le stime dei parametri $\theta_i$. In particolare, viene analizzato il ruolo dello *"shrinkage factor"* sia nel caso di correlazione semplice, sia in quello di interazione spaziale fra le aree. Utilizzando i $\breve{\gamma}_i$ ottenuti attraverso la minimizzazione del $MSE$, è stata misurata l'efficienza relativa dello stimatore $\breve{\theta}_i$ rispetto a $\theta_i^*$ e si è verificato come all'aumentare del numero delle piccole aree il vantaggio dell'introduzione dell'informazione spaziale sia sostanziale.

Senza dubbio esistono delle difficoltà e dei limiti nell'applicare la metodologia proposta. In primo luogo è necessario conoscere la posizione geografica (per esempio latitudine e longitudine) delle piccole aree che si vanno ad analizzare [3].

Un secondo problema riguarda la matrice dei pesi $W$ che non è univocamente determinata e può essere definita in numerosi modi diversi al variare dello schema di vicinanza tra le aree. Sarà quindi utile in studi successivi verificare i risultati ottenuti modificando $W$.

Si osserva inoltre che aumentando il dettaglio spaziale della piccola area possono cambiare i risultati della stima[4].

Infine nello studio sono state fatte alcune assunzioni quali la conoscenza dei valori della varianza da disegno ($\sigma^2$), della varianza tra aree ($\sigma_u^2$) e del coefficiente di autoregressione spaziale $\rho_s$. Si tratta di una situazione che si verifica raramente nella realtà e quindi in futuro dovrà essere affrontato il problema di stima di questi parametri.

[3] Non sempre è possibile ottenere queste informazioni. In alternativa si possono considerare le coordinate del centroide e assumere che il valore stimato sia riferito a quel punto.

[4] Si tratta del noto problema dell'unità areale modificabile (Modifiable Area Unit Problem - MAUP).

## APPENDICE A. DETERMINAZIONE DELLO SHRINKAGE FACTOR CON CORRELAZIONE SEMPLICE

Se si suppone l'esistenza di correlazione semplice tra gli effetti di area, $Corr(u_i, u_k) = \rho$ con $\rho > 0$ per $i \neq k$, il valore ottimale del predittore BLUP del valore medio della piccola aree $i-esima$ è:

$$\theta_i^* = \tilde{\gamma}\bar{y}_i + (1-\tilde{\gamma})\bar{y} \;\; con\; i = 1...m.$$

Per ottenere lo shrinkage factor si calcola il $MSE(\theta_i^*)$ e si minimizza rispetto a $\tilde{\gamma}$ (con $0 < \tilde{\gamma} < 1$).

$$MSE(\theta_i^*) = E\{[\tilde{\gamma}\bar{y}_i + (1-\tilde{\gamma})(\sum_{i=1}^{m}\frac{\bar{y}_i}{m}) - \mu - u_i]^2\} =$$

$$E\{[\tilde{\gamma}(\bar{y}_i - \mu) + (1-\tilde{\gamma})(\frac{\bar{y}_1 + \bar{y}_2 + ... + \bar{y}_m}{m} - \frac{m\mu}{m}) - u_i]^2\} =$$

$$\tilde{\gamma}^2 E\{(\bar{y}_i - \mu)^2\} + (1-\tilde{\gamma})^2\frac{1}{m^2}E\{[(\bar{y}_1 - \mu) + (\bar{y}_2 - \mu) + ... + (\bar{y}_m - \mu)]^2\} + E\{u_i^2\} -$$

$$-2\tilde{\gamma}E\{(\bar{y}_i - \mu)u_i\} - 2(1-\tilde{\gamma})E\{[\frac{\bar{y}_1 + \bar{y}_2 + ... + \bar{y}_m}{m} - \frac{m\mu}{m}]u_i\} +$$

$$+2E\{\tilde{\gamma}(\bar{y}_i - \mu)(1-\tilde{\gamma})(\sum_{i=1}^{m}\frac{(\bar{y}_i - \mu)}{m})\}.$$

Per calcolare il $MSE(\theta_i^*)$ si sviluppano singolarmente gli addendi della somma:

1) $\tilde{\gamma}^2 E\{(\bar{y}_i - \mu)^2\} = \tilde{\gamma}^2 var(\bar{y}_i) = \tilde{\gamma}^2(\sigma_u^2 + \sigma^2)$;

2) $(1-\tilde{\gamma})^2\frac{1}{m^2}E\{[(\bar{y}_1 - \mu) + (\bar{y}_2 - \mu) + ... + (\bar{y}_m - \mu)]^2\} =$

$$= (1-\tilde{\gamma})^2\frac{1}{m^2}\{m\; var(\bar{y}_i) + 2\sum_i\sum_{j>i} cov(\bar{y}_i, \bar{y}_j)\} =$$

Con $cov(\bar{y}_i, \bar{y}_j) = \rho\sigma_u^2$ e $2\sum_i\sum_{j>i} cov(\bar{y}_i, \bar{y}_j) = \frac{2m(m-1)}{2}\rho\sigma_u^2$ essendo $2\sum_i\sum_{j>i} cov(\bar{y}_i, \bar{y}_j)$ le combinazioni di $m$ elementi presi due alla volta.

$$= (1-\tilde{\gamma})^2\frac{1}{m^2}\{m(\sigma_u^2 + \sigma^2) + \frac{2m(m-1)}{2}\rho\sigma_u^2\} =$$

$$= (1-\tilde{\gamma})^2\frac{1}{m}\{(\sigma_u^2 + \sigma^2) + (m-1)\rho\sigma_u^2\};$$

3) $E\{u_i^2\} = \sigma_u^2$;

4) Con $\bar{y}_i - \mu = e_i + u_i$ e $e_i, u_i$ indipendenti si verifica che:

$$= -2\tilde{\gamma}E\{(\bar{y}_i - \mu)u_i\} = -2\tilde{\gamma}E\{(u_i + e_i)u_i\} =$$

$$= -2\tilde{\gamma}[E\{u_i^2\} + E\{u_i e_i\}] = -2\tilde{\gamma}\sigma_u^2$$

5) $-2(1-\tilde{\gamma})E\{[\frac{\bar{y}_1 + \bar{y}_2 + ... + \bar{y}_m}{m} - \frac{m\mu}{m}]u_i\} =$

$$= -2(1-\tilde{\gamma})\frac{1}{m}E\{[(u_1 + e_1) + ... + (u_m + e_m)]u_i\} =$$

$$= -2(1-\tilde{\gamma})\frac{1}{m}E\{u_1u_i + u_2u_i + ... + u_mu_i + e_1u_i + ... + e_mu_i\} =$$

$$= -2(1-\tilde{\gamma})\frac{1}{m}E\{u_1u_i + ... + u_iu_i + ... + u_mu_i\} =$$

$$= -2(1-\tilde{\gamma})\frac{1}{m}[E\{u_i^2\} + \sum_{j,j\neq i} E\{u_i,u_j\}] =$$

$$= -2(1-\tilde{\gamma})\frac{1}{m}[\sigma_u^2 + (m-1)\rho\sigma_u^2];$$

6) $2E\{\tilde{\gamma}(\bar{y}_i-\mu)(1-\tilde{\gamma})(\sum_{i=1}^{m}\frac{(\bar{y}_i-\mu)}{m})\} =$

$$= 2\frac{\tilde{\gamma}(1-\tilde{\gamma})}{m}E\{(\bar{y}_i-\mu)[(\bar{y}_1-\mu) + ... + (\bar{y}_i-\mu) + ... + (\bar{y}_m-\mu)]\} =$$

$$= 2\frac{\tilde{\gamma}(1-\tilde{\gamma})}{m}[E\{(\bar{y}_i-\mu)(\bar{y}_1-\mu) + ... + (\bar{y}_i-\mu)(\bar{y}_i-\mu) + ... + (\bar{y}_i-\mu)(\bar{y}_m-\mu)\}]$$

$$= 2\frac{\tilde{\gamma}(1-\tilde{\gamma})}{m}[var(\bar{y}_i) + (m-1)cov(\bar{y}_i,\bar{y}_j)] =$$

$$= 2\frac{\tilde{\gamma}(1-\tilde{\gamma})}{m}[(\sigma_u^2+\sigma^2) + (m-1)\rho\sigma_u^2].$$

Sommando tutti gli addendi e mettendo in evidenza alcuni fattori si ottiene:

$$\tilde{\gamma}^2(\sigma_u^2+\sigma^2) + \frac{1}{m}\{(\sigma_u^2+\sigma^2) + (m-1)\rho\sigma_u^2\}[(1-\tilde{\gamma})^2 + 2(\tilde{\gamma}(1-\tilde{\gamma}))] +$$

$$+\sigma_u^2(1-2\tilde{\gamma}) - 2(1-\tilde{\gamma})\frac{1}{m}[\sigma_u^2 + (m-1)\rho\sigma_u^2] =$$

$$= \tilde{\gamma}^2(\sigma_u^2+\sigma^2) + \frac{1}{m}\{(\sigma_u^2+\sigma^2) + (m-1)\rho\sigma_u^2\}(1-\tilde{\gamma}^2) +$$

$$\sigma_u^2(1-2\tilde{\gamma}) - 2(1-\tilde{\gamma})\frac{1}{m}[\sigma_u^2 + (m-1)\rho\sigma_u^2].$$

Derivando il $MSE(\theta_i^*)$ rispetto a $\tilde{\gamma}$ e minimizzando si ottiene lo s.f. in caso di correlazione semplice fra aree:

$$\tilde{\gamma} = \frac{[\sigma_u^2(1-\rho)]}{[\sigma_u^2(1-\rho)+\sigma^2]}.$$

Nel caso in cui non sia presente una correlazione semplice tra i dati ($\rho = 0$) lo s.f. è:

$$\gamma = \frac{\sigma_u^2}{[\sigma_u^2+\sigma^2]}.$$

## APPENDICE B. DETERMINAZIONE DELLO SHRINKAGE FACTOR NEL CASO DI CORRELAZIONE SPAZIALE TRA AREE

Se si suppone l'esistenza di correlazione spaziale tra gli effetti di area il valore ottimale del predittore BLUP del valore medio della piccola aree $i-esima$ è:

$$\breve{\theta}_i = \breve{\gamma}_i\tilde{\theta}_i + (1-\breve{\gamma}_i)\bar{y} \;\; con\; i = 1...m.$$

Per ottenere lo shrinkage factor si calcola il $MSE(\breve{\theta}_i)$ e si minimizza rispetto a $\breve{\gamma}$ (con $0 < \breve{\gamma} < 1$).

$$MSE(\breve{\theta}_i) = E\{[\breve{\gamma}\bar{y}_i + (1-\breve{\gamma})(\sum_{i=1}^{m}\frac{\bar{y}_i}{m}) - \mu - u_i]^2\} =$$

$$E\{[\breve{\gamma}(\bar{y}_i - \mu) + (1-\breve{\gamma})(\frac{\bar{y}_1+\bar{y}_2+...+\bar{y}_m}{m} - \frac{m\mu}{m}) - u_i]^2\} =$$

$$\breve{\gamma}^2 var(\bar{y}_i) + (1-\breve{\gamma})^2\frac{1}{m^2}E\{[(\bar{y}_1-\mu)+(\bar{y}_2-\mu)+...+(\bar{y}_m-\mu)]^2\} + E\{u_i^2\} -$$

$$-2\breve{\gamma}E\{(\bar{y}_i-\mu)u_i\} - 2(1-\breve{\gamma})E\{[\frac{\bar{y}_1+\bar{y}_2+...+\bar{y}_m}{m} - \frac{m\mu}{m}]u_i\} +$$

$$+2E\{\breve{\gamma}(\bar{y}_i-\mu)(1-\breve{\gamma})(\sum_{i=1}^{m}\frac{(\bar{y}_i-\mu)}{m})\}.$$

Anche in questa dimostrazione si considerano singolaremte gli addendi della somma:

1) $\breve{\gamma}^2 var(\bar{y}_i) = \breve{\gamma}^2(\sigma^2_{u_i}(\rho_s, W) + \sigma^2)$;

2) $(1-\breve{\gamma})^2\frac{1}{m^2}E\{[(\bar{y}_1-\mu)+(\bar{y}_2-\mu)+...+(\bar{y}_m-\mu)]^2\} =$

$$= (1-\breve{\gamma})^2\frac{1}{m^2}\{\sum_{i=1}^{m} var(\bar{y}_i) + 2\sum_i\sum_{j>i} cov(\bar{y}_i, \bar{y}_j)\} =$$

$$= (1-\breve{\gamma})^2\frac{1}{m^2}\{\sum_{i=1}^{m}(\sigma^2_{u_i}(\rho_s, W) + \sigma^2) + 2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s, W)\};$$

3) $E\{u_i^2\} = \sigma^2_{u_i}(\rho_s, W)$;

4) Con $\bar{y}_i - \mu = e_i + u_i$ e $e_i, u_i$ indipendenti si verifica che:

$$-2\breve{\gamma}E\{(\bar{y}_i-\mu)u_i\} = -2\breve{\gamma}E\{(u_i+e_i)u_i\} =$$

$$= -2\breve{\gamma}[E\{u_i^2\} + E\{u_ie_i\}] = -2\breve{\gamma}\sigma^2_{u_i}(\rho_s, W);$$

5) $-2(1-\breve{\gamma})E\{[\frac{\bar{y}_1+\bar{y}_2+...+\bar{y}_m}{m} - \frac{m\mu}{m}]u_i\} =$

$$= -2(1-\breve{\gamma})\frac{1}{m}E\{[(u_1+e_1)+...+(u_m+e_m)]u_i\} =$$

$$= -2(1-\breve{\gamma})\frac{1}{m}E\{u_1u_i + u_2u_i + ... + u_mu_i + e_1u_i + ... + e_mu_i\} =$$

$$= -2(1-\breve{\gamma})\frac{1}{m}E\{u_1u_i + ... + u_iu_i + ... + u_mu_i\} =$$

$$= -2(1-\check{\gamma})\frac{1}{m}[E\{u_i^2\} + \sum_{j,j\neq i} E\{u_i,u_j\}] =$$

$$= -2(1-\check{\gamma})\frac{1}{m}[\sigma^2_{u_i}(\rho_s,W) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)];$$

6) $2E\{\check{\gamma}(\bar{y}_i-\mu)(1-\check{\gamma})(\sum_{i=1}^m \frac{(\bar{y}_i-\mu)}{m})\} =$

$$= 2\frac{\check{\gamma}(1-\check{\gamma})}{m}E\{(\bar{y}_i-\mu)[(\bar{y}_1-\mu)+...+(\bar{y}_i-\mu)+...+(\bar{y}_m-\mu)]\} =$$

$$= 2\frac{\check{\gamma}(1-\check{\gamma})}{m}[E\{(\bar{y}_i-\mu)(\bar{y}_1-\mu)+...+(\bar{y}_i-\mu)(\bar{y}_i-\mu)+...+(\bar{y}_i-\mu)(\bar{y}_m-\mu)\}]$$

$$= 2\frac{\check{\gamma}(1-\check{\gamma})}{m}[var(\bar{y}_i) + \sum_{j,j\neq i} cov(\bar{y}_i,\bar{y}_j)] =$$

$$= 2\frac{\check{\gamma}(1-\check{\gamma})}{m}[(\sigma^2_{u_i}(\rho_s,W)+\sigma^2) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)].$$

Sommando tutti gli addendi si ottiene:

$$\check{\gamma}^2(\sigma^2_{u_i}(\rho_s,W)+\sigma^2) + (1-\check{\gamma})^2\frac{1}{m^2}\{\sum_{i=1}^m(\sigma^2_{u_i}(\rho_s,W)+\sigma^2) + 2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s,W)\}+$$

$$+\sigma^2_{u_i}(\rho_s,W) - 2\check{\gamma}\sigma^2_{u_i}(\rho_s,W) - 2(1-\check{\gamma})\frac{1}{m}[\sigma^2_{u_i}(\rho_s,W) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)]+$$

$$+2\frac{\check{\gamma}(1-\check{\gamma})}{m}[(\sigma^2_{u_i}(\rho_s,W)+\sigma^2) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)].$$

Il $MSE(\check{\theta}_i)$ risulta:

$$\sigma^2_{u_i}(\rho_s,W)(1-\check{\gamma})^2 + \sigma^2(\check{\gamma}^2 + \frac{(1-\check{\gamma}^2)}{m})+$$

$$(1-\check{\gamma})^2\frac{1}{m^2}\{\sum_{i=1}^m\sigma^2_{u_i}(\rho_s,W) + 2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s,W)\}-$$

$$-2\frac{(1-\check{\gamma})}{m}[\sigma^2_{u_i}(\rho_s,W) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)]+$$

$$+2\frac{\check{\gamma}(1-\check{\gamma})}{m}[\sigma^2_{u_i}(\rho_s,W) + \sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)].$$

Derivando il $MSE(\check{\theta}_i)$ rispetto a $\check{\gamma}$ e minimizzando si ottiene lo s.f. in caso di correlazione spaziale fra aree. Lo s.f., in questo caso, varia da area ad area in quanto dipende dal numero di vicini di ciascuna area, che si riflette sulla struttura della matrice varianze-covarianze.

$$\check{\gamma}_i = \frac{\sigma^2_{u_i}(\rho_s,W)[1-\frac{2}{m}] + \frac{1}{m^2}[\sum_{i=1}^m(\sigma^2_{u_i}(\rho_s,W)) + 2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s,W)]-}{\sigma^2_{u_i}(\rho_s,W)[1-\frac{2}{m}] + \frac{1}{m^2}[\sum_{i=1}^m(\sigma^2_{u_i}(\rho_s,W)) + 2\sum_i\sum_{j>i}\sigma_{u_i,u_j}(\rho_s,W)]-}$$

$$\frac{-\frac{2}{m}\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W)}{-\frac{2}{m}\sum_{j,j\neq i}\sigma_{u_i,u_j}(\rho_s,W) + \sigma^2(1-\frac{1}{m})}.$$

## Riferimenti bibliografici


Anselin L.. 1992. *"Spatial Econometrics: Method and Models"*.Kluwer Academic Publishers, Boston.

Cliff A.D., Ord J.K.. 1981. *"Spatial Processes. Models & Applications"*. Pion Limited, London.

Cressie N..1993. *"Statistics for Spatial Data"*. New York, Jhon Wiley & Sons.

Ericksen E.P. 1974. "A Regression method for Estimating Population Changes of Local Areas". *Journal of the American Statistical Association*, vol.69, n348, 867-875.

Estevao V.M., Särndal C.E.. 1999. "The use of Auxiliary Information in Design-Based Estimation for Domains". *Survey Methodology*, vol. 25, n2, 213-221.

Fay R.E., Herriot R.A. (1979). "Estimates of income for small places: an application of James-Stein procedures to census data". *J.Amer. Statist. Assoc., vol. 74 269-277.*

*Ghosh M., Rao J.N.K.. 1994. "Small Area Estimation: An Appraisal (with discussion)". Statistical Science*, 9, 65-93.

Haining R.. *"Spatial data analysis in the social and environmental sciences"*. Cambridge University press, Cambridge.

Noble A., Haslett S., Arnold G.. 2002. "Small Area Estimation via Generalized Linear Models". *Journal of Official Statistics*, vol. 18, n1, 45-60.

Openshaw S., Taylor P. 1981. "The Modifiable Unit Problem". *In Quantitative Geography, a British View*, edited by N. Wrigley. Pion. London, 127-144.

Ord K. (1975). "Estimation Methods for Models of Spatial Interaction". *Journal of the American Statistical Association*, Vol. 70, 349, 120-126.

Pfeffermann D.. 2002. "Small Area Estimation-New Developments and Directions". *International Statistical Review*, vol. 70, n1, 125-143.

Rao J.N.K.. 1999. "Some Recent Advances in Model-Based Small Area Estimation". *Survey Methodology*, vol. 25, n2, 175-186.

Rao J.N.K., Yu M.. 1994. "Small-area estimation by combining time-series and cross-sectional data". *The Canadian Journal of Statistics*, vol. 22, n4, 511-528.

Upton G.J.G., Fingleton B.. 1985. *"Spatial Data Analysis by Example"*. John Wiley & Sons.