UNIVERSITÀ DEGLI STUDI DI PISA

FACOLTÀ DI ECONOMIA
CORSO DI LAUREA IN
STATISTICA PER L'ECONOMIA E PER L'AZIENDA

TESI DI LAUREA

Metodi di stima per piccole aree applicate alle indagini correnti in agricoltura in Italia

CANDIDATO
Stefano Marchetti

RELATORE
Prof.ssa Monica Pratesi

CONTRORELATORE
Prof. Eugene M. Cleur

Anno Accademico 2004 - 2005

INDICE

ELENCO DELLE TABELLE

ELENCO DELLE FIGURE

RINGRAZIAMENTI

Voglio porgere i miei più sentiti ringraziamenti a tutti i professori del dipartimento di Statistica e Matematica Applicata all'Economia, che si sono dimostrati sempre molto disponibili e pronti a venire incontro alle esigenze degli studenti.
Un ringraziamento particolare va al Dott. Nicola Salvati e alla Prof.ssa Monica Pratesi che mi hanno seguito e supportato costantemente nella compilazione di questa tesi, nonché al Prof. Carlo Bianchi, al Prof. Marco Bottai e al Prof. Riccardo Cambini (in rigoroso ordine alfabetico!) che più hanno contribuito alla mia formazione culturale.
Colgo l'occasione per dedicare un pensiero al Prof. Gilberto Ghilardi che ha mi ha introdotto nell'affascinante mondo della statistica e mi ha aiutato nei momenti di maggiore difficoltà.

Infine ringrazio la mia ragazza, che mi ha sopportato durante gli anni di studio e mi ha spinto ad andare avanti facendomi raggiungere questo traguardo importante. Ringrazio tutti i miei amici che a loro modo mi hanno supportato ed aiutato.

A Serena …

INTRODUZIONE

Attualmente le indagini campionarie giocano un ruolo chiave per programmare lo svolgimento dell'attività di enti privati e pubblici. La rapida evoluzione dei mercati e del modo di vivere rende i dati censuari, rilevati una volta ogni dieci anni, utilizzabili solo marginalmente a tale scopo; inoltre essi si riferiscono solo ad alcuni aspetti, che possono risultare inadeguati rispetto alle esigenze conoscitive.

Molte indagini (soprattutto quelle su vasta scala) sono fatte in base ad un disegno campionario complesso, spesso con stratificazione geografica e rispetto ad alcune caratteristiche delle unità osservate, disegno che consente di ottenere stime affidabili soltanto per un livello prestabilito di ripartizione geografica o dominio.

La stima per piccole aree intende superare questo problema, consentendoci di ottenere stime affidabili per ripartizioni geografiche e domini più piccoli rispetto a quelli previsti dal disegno campionario originario. Grazie a questa procedura possiamo usare al meglio dati censuari e dati campionari per fornire velocemente stime affidabili per microaggregati, la cui analisi risulta oggi vincente sia per gli enti privati, che sono così in grado di ottenere informazioni precise e dettagliate per ogni mercato, sia per gli enti pubblici, che possono operare una pianificazione economica e territoriale più mirata e "personalizzata".

Un vantaggio ulteriore di questa metodologia deriva dalla possibilità di utilizzare fonti statistiche già esistenti, come i dati censuari, gli archivi amministrativi, i dati di diverse indagini campionarie, etc, combinandole insieme al fine di ottenere una stima affidabile per il carattere desiderato, nell'ambito dell'analisi svolta. Non risulta necessario, quindi, progettare indagini "ad hoc" per ottenere le informazioni desiderate, con il risultato di una notevole diminuzione di costi e tempi. La velocità nell'ottenere le stime e i costi sostenuti per ottenerle giocano un ruolo importante nel sistema competitivo di oggi, e questo sottolinea l'importanza di promuovere metodi di stima sofisticati e di avere persone in grado di applicarli. E' importante sottolineare che con il metodo di stima per piccole aree si ottengono stime affidabili su piccoli domini nonostante il disegno di campionamento prevedesse lo studio di domini più ampi (o addirittura diversi nei casi dove è possibile); siamo dunque svincolati dal disegno originario (in parte), fatta salva la possibilità di collegare i dati rilevati con i domini di interesse.

Nella tesi saranno trattati i modelli di stima per piccole aree che fanno uso di informazioni ausiliarie. Queste sono informazioni riguardanti il dominio di interesse (in genere area geografica) o le unità presenti in tale dominio; le informazioni possono

essere reperite da una qualunque fonte disponibile (sia ufficiale che non). La tesi si articola come segue.

Inizieremo con una breve introduzione ai tradizionali metodi di stima diretta. L'attenzione è sul campionamento stratificato e sulla stima diretta della media di area e della varianza (rispetto al disegno di campionamento in questione). Il metodo di stima tradizionale ci servirà come termine di paragone con il metodo di stima per piccole aree. Il metodo di stima per piccole aree verrà presentato in modo dettagliato dimostrando la correttezza e l'efficienza degli stimatori proposti; si farà uso del modello di Regressione multipla e del metodo di stima della Massima Verosimiglianza. Nello sviluppo della tesi la piccola area sarà intesa come porzione di territorio quale provincia, comune, frazione o altri tipi di aggregati.

Nella letteratura corrente nella stima per piccole aree lo stimatore più diffuso ipotizza effetti casuali non correlati tra zone (domini, aree geografiche); ciò significa ipotizzare che domini vicini (o confinanti) non si influenzano a vicenda. In questa tesi presenteremo anche un tipo di stimatore che rimuove questa ipotesi, spesso non aderente alla realtà, soprattutto quando i domini di interesse sono le aree geografiche.

Nella parte finale della tesi verrà presentata una simulazione di stima per piccole aree atta al confronto tra i diversi stimatori (diretto, per piccole aree senza ipotesi di correlazione spaziale) e un'applicazione pratica.

Utilizzando i dati dell'indagine sulla Struttura e Produzione delle aziende Agricole si è proceduto a stimare la produzione delle coltivazioni più diffuse in Toscana. L'indagine nasce per aggiornare i dati censuari sull'agricoltura, raccolti ogni dieci anni. Fatta dall'ISTAT unitamente alle Regioni, l'indagine ha come territorio d'interesse la regione; con la metodologia classica non si ottengono stime affidabili per domini più piccoli: stime affidabili per aree subregionali, quali i comuni o le province, sono invece ottenute con gli stimatori per piccola area proposti.

CAPITOLO 1

LA STIMA DIRETTA

1.1 PREMESSA

Per valutare l'efficienza dello stimatore per piccole aree abbiamo bisogno di un termine di paragone. Confronteremo i risultati con quelli ottenuti stimando i parametri di area in modo diretto da campione (stima diretta). Tale stima non usa modelli, è basata sul solo disegno di campionamento e sui dati del campione.

L'uso del campione casuale stratificato, infatti, permette di ottenere una stima diretta più efficiente (affidabile)[1] rispetto ad altri schemi di campionamento. Nel seguito faremo una veloce panoramica sulla stima di media e varianza di una popolazione nel campionamento casuale stratificato.

1.2 IL CAMPIONAMENTO CASUALE STRATIFICATO E LA STIMA DI MEDIA E VARIANZA CAMPIONARIE

In questo capitolo di carattere introduttivo daremo solo dei cenni sul metodo di campionamento stratificato e dei richiami per quanto riguarda la stima di un parametro.

Il campionamento è un processo tramite il quale si seleziona un insieme di individui, appartenenti ad una popolazione di riferimento, sui quali si osservano certi caratteri con lo scopo di fare inferenza.
Il processo di campionamento è diviso in cinque "steps" (passi):

a. Definizione della popolazione di riferimento (target).

b. Definizione di un "frame" (una popolazione da cui è possibile identificare ogni elemento e campionarlo; non sempre è possibile individuare le

[1] Si dice che uno stimatore corretto è più efficiente di un altro quando ha varianza attese minore. Infatti una stima per intervalli (il metodo più diffuso) del valore incognito X di una certa popolazione è $x \pm z_{\alpha/2} \cdot \sqrt{\mathrm{var}(x)}$, con x il valore calcolato dal campione (es.: la media campinaria), $z_{\alpha/2}$ una costante riferita alla distribuzione normale e var(x) la varianza del valore campionario. E' ovvio come diminuendo la varianza l'intervallo di stima si restringe.
NOTA: In teoria dei campioni spesso usano i termini stima e stimatore come sinonimi.

c. popolazioni di cui al punto a., ed infatti non sempre il frame corrisponde al target).

d. Definizione del metodo di campionamento (casuale, non casuale).

e. Estrazione del campione e raccolta dei dati.

f. Stima dei parametri.

g. Revisione del processo di campionamento.

Il campionamento casuale è un metodo di campionamento (punto c) mentre la stratificazione è un modo di trattare il frame di riferimento.

Questo metodo si basa sulla suddivisione della popolazione secondo uno o più caratteri (ad esempio sesso, regione di appartenenza, etc.), in questo modo otteniamo delle sottopopolazioni chiamate strati. Da ogni strato si estrae un campione casuale semplice[2]. Dai campioni estratti si rilevano (misurano) i caratteri (variabili) che ci interessano.

Questo metodo comporta una serie di vantaggi. Il più importante è la maggiore affidabilità delle stime rispetto al campionamento casuale semplice[3], inoltre si può analizzare la popolazione sia strato per strato sia nella sua interezza.

Per esempio se la popolazione di riferimento sono le aziende con sede in Italia, possiamo stratificare per zona (nord, centro e sud) di appartenenza. Otteniamo:

ZONA	*POPOLAZIONE*	*CAMPIONE*
Nord	N_1	n_1
Centro	N_2	n_2
Sud	N_3	n_3
Totale	$\boldsymbol{N}$	$\boldsymbol{n}$

[2] Il metodo del campione casuale semplice consiste nel numerare gli N elementi della popolazione di riferimento (nel caso del campione casuale stratificato gli elementi di uno strato) ed estrarne n (con $n \leq N$) senza reimmissione. Con questo metodo risulta che la probabilità di estrarre alla h-esima estrazione un certo elemento x_h dato che esso non è stato estratto nelle h-1 estrazioni precedenti è uguale a 1/N, qualsiasi sia h.

[3] La variabilità delle stime è minore qualora il criterio di stratificazione sia correlato con le variabili di studio e l'allocazione del campione negli strati sia proporzionale (cfr. Sardnal et al, 1992).

Dove N_i, $i=1…3$, è la numerosità della popolazione nello strato i (Nord, Centro e Sud) e n_i, $i=1…3$, è la numerosità campionaria (cioè il numero di elementi estratti) dello strato i. Ovviamente risulta $\sum_i N_i = N$ e $\sum_i n_i = n$, con N totale delle aziende in Italia ed n numero totale di aziende campionate. Sia N sia $N_1, N_2, ...,N_H$ si suppongono note. Possiamo immaginare di ottenere questo dato dalle camere di commercio, riferendoci all'esempio[4]. Supponiamo di rilevare diverse variabili tra cui il reddito dichiarato. Possiamo stimare il reddito medio per l'Italia e il reddito medio per zona.

In statistica uno stimatore è una funzione dei dati conosciuti usata per stimare un parametro sconosciuto di una popolazione di riferimento.

Si desidera che uno stimatore sia corretto, efficiente e consistente.

Nell'esempio il parametro sconosciuto è il reddito medio della popolazione, i dati conosciuti sono i redditi del campione e la funzione che usiamo è la media aritmetica ponderata:

$$\textit{Stima della media del reddito della popolazione} = \hat{\bar{X}} = \sum_{i=1}^{H} \hat{\bar{x}}_i \cdot W_i$$

Dove $\hat{\bar{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j$, con x_j osservazione j-esima ($j = 1…n_i$) del carattere x (reddito) nello strato n_i. $\hat{\bar{x}}_i$ è quindi la media aritmetica campionaria dello strato i-esimo. H è il numero di strati (3 nel nostro esempio). $W_i = \frac{N_i}{N}$, il peso che lo strato i-esimo (es.: Nord) ha all'interno della popolazione. La stima all'interno degli strati ($\hat{\bar{x}}$) ha, ovviamente, la stessa efficienza della stima ottenuta con campione casuale semplice, mentre la stima su tutta la popolazione è più efficiente rispetto a stime ottenute con il campione casuale semplice. Vediamo perché:

- Stima della media all'interno degli strati:

$$E[\hat{\bar{x}}_i] = E[\frac{1}{n_i} \sum_{j=1}^{n_i} x_j] = \frac{1}{n_i} \sum E[x_j] =$$

[4] Se $N_1, N_2, …, N_H$ sono incogniti vanno stimati. Una volta che sono stati stimati si procede come descritto nel testo ma è necessario aggiungere la variabilità di tale stima alla variabilità dello stimatore ottenuto con campione casuale stratificato.

$$= \frac{1}{n_i} \sum \frac{1}{N} x_j = \frac{1}{n_i} \sum \overline{X}_i = \frac{1}{n_i} \cdot n_i \overline{X}_i = \overline{X}_i$$

La media campionaria $\hat{\bar{x}}_i$ è uno stimatore corretto[5] della media della popolazione (il cui valore è sconosciuto) nello strato *i*.

- Stima della media della popolazione:

$$E[\hat{\bar{X}}] = E[\sum_{i=1}^{H} \hat{\bar{x}}_i \cdot w_i] = \sum E[\hat{\bar{x}}_i \cdot \frac{N_i}{N}] = \sum \frac{N_i}{N} E[\hat{\bar{x}}_i] =$$

$$\sum \frac{N_i}{N} \overline{X}_i = \frac{1}{N} \sum \overline{X}_i N_i = \frac{\sum \overline{X}_i N_i}{\sum N_i} = \overline{X}$$

La media pesata degli strati è una stima corretta della media della popolazione.

- Stima della varianza all'interno degli strati: la varianza del primo stimatore proposto è

$$V[\hat{\bar{x}}_i] = E[(\hat{\bar{x}}_i - E[\hat{\bar{x}}_i])^2] = \frac{S_{\bar{x}_i}^2}{n_i}$$

con $S_{\bar{x}_i}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{n_i} (X_j - \overline{X}_i)^2$. Il valore $S_{\bar{x}_i}^2$ è incognito poiché non conosciamo gli X_j,cioè gli elementi della popolazione. Per questo dobbiamo stimare la varianza sopraesposta; un metodo di stima proposto è:

$$\hat{V}[\hat{\bar{x}}_i] = \frac{\hat{S}_{\bar{x}_i}^2}{n_i}$$

con $\hat{S}_{\bar{x}_i}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_j - \bar{x}_i)^2$.

[5] Uno stimatore $\hat{\theta}$ di un certo parametro incognito θ si dice corretto quando $E[\hat{\theta}] = \theta$.

- Stima della varianza della popolazione:

$$\hat{V}\ [\hat{\bar{X}}] = E[(\hat{\bar{X}} - E[\hat{\bar{X}}])^2] = \sum_{i=1}^{H} W_i^2 \cdot \hat{V}(\hat{\bar{x}}_i).$$

- Nel caso del campione casuale semplice il metodo di stima della varianza è uguale al metodo di stima della varianza all'interno degli strati con la differenza che nel campione casuale semplice ho solo *n* ed $\hat{\bar{x}}$: la varianza dello stimatore assume la forma

$$V[\hat{\bar{x}}] = \frac{S_{\bar{x}}^2}{n}$$

con $S_{\bar{x}}^2 = \frac{1}{N-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$. Analogamente ai casi precedenti otteniamo una stima di (1.7) con:

$$\hat{V}\ [\hat{\bar{x}}] = \frac{\hat{S}_{\bar{x}}^2}{n}$$

dove $\hat{S}_{\bar{x}}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$.

Si dimostra che grazie alla ponderazione nella stima della varianza nel campione casuale stratificato essa risulta minore o uguale alla stima della varianza nel caso del campione casuale semplice:

$$\hat{V}_{\text{stratificato}}[\hat{\bar{X}}] \leq \hat{V}_{\text{semplice}}[\hat{\bar{x}}]$$

Questo comporta una riduzione dell'intervallo di confidenza che si traduce in una maggior efficienza delle stime. La stima ottenuta con il campionamento casuale stratificato è una stima efficiente[6].

[6] Si tratta di efficienza relativa: il rapporto $\hat{V}_{\text{stratificato}}/\hat{V}_{\text{casuale semplice}} < 1$.

E' importante sottolineare che la stima della varianza, e di conseguenza l'intervallo di confidenza che ne deriva, è strettamente legato alla dimensione campionaria *n*. Risulta ovvio, dalle formule appena presentate, che all'aumentare della dimensione campionaria (*n*) la stima della varianza diminuisce così come diminuisce l'intervallo di confidenza (quindi migliora la bontà della stima). Aumentare la dimensione campionaria porta, però, un aumento dei costi e dei tempi di rilevazione. Pertanto, ogni indagine campionarie avrà un numero di osservazioni necessarie ad ottenere stime accettabili rispetto all'entità a cui si riferisce (nell'esempio Nord, Centro e Sud).

Per completezza, ricordiamo che la stima diretta in questione è anche consistente, poiché converge in probabilità al vero valore assunto dalla popolazione[7].

Precisiamo che la procedura di stratificazione può essere fatta anche a posteriori, cioè dopo una rilevazione fatta con campionamento casuale semplice, si parla in questo caso di campionamento casuale post-stratificato. Ovviamente nella stratificazione a priori (classica) avremo una dimensione campionaria per strato che consente di ottenere un intervallo di stima accettabile, poiché siamo noi che decidiamo a priori la dimensione campionaria per ogni strato; per accettabile si intende un intervallo di confidenza i cui valori estremi non si allontanano in maniera considerevole dal valore medio, con una significatività, in genere, del 95%[8]; per esempio consideriamo accettabile una stima sul reddito medio con ampiezza 1000 su un valore di 10000 (10000€±1000) e non accettabile un'ampiezza 6000, sempre riferito a un valore medio di 10000 (10000€±6000).

Ottenere una stima accettabile non è garantito dal disegno di post-stratificazione; infatti è possibile che per alcuni strati non ci sia un numero di osservazioni sufficienti, o addirittura che non ci siano osservazioni. Ad esempio se post-stratifichiamo da un campione di *n* elementi riferito all'intero territorio nazionale suddividendolo per province, è possibile (anzi probabile) che per alcune province ci siano poche osservazioni (o nessuna). L'applicazione pratica che prenderemo in esempio rientra nel caso della post-stratificazione. Infatti gli strati considerati nella rilevazione SPA non

[7] Una stima $\hat{\theta}$ è consistente se $\lim_{n\to\infty} \Pr\{|\hat{\theta}-\theta|<\varepsilon\}=1$ per qualsiasi ε. La consistenza è una caratteristica importante per uno stimatore. Spesso in statistica non si possono ottenere stime corrette ma se le stime ottenute sono consistenti è possibile trarre comunque delle conclusioni.

[8] Significatività del 95% significa che c'è una probabilità uguale a 0.95 (del 95%) che il vero valore della popolazione sia compreso nell'intervallo di confidenza proposto. C'è dunque una probabilità di 0.05 (5%) che il vero valore non sia compreso nell'intervallo di confidenza. Tale probabilità (0.05) rappresenta l'errore di I specie (contrassegnato in letteratura dalla lettera α).

sono della stessa dimensione della piccola area a cui faremo riferimento. Esiste in questo caso la possibilità che in uno strato definito a posteriori non venga rilevata nessuna unità.

Possiamo concludere che la stima diretta ottenuta campionando gli strati è un perfetto punto di riferimento e di confronto per gli stimatori che proporremo nel seguito di questo lavoro.

CAPITOLO 2

LA STIMA PER PICCOLE AREE

2.1 DEFINIZIONE DI PICCOLA AREA

Vediamo innanzitutto l'evoluzione della definizione di piccola area negli ultimi anni.

Nel 1980 si ha una prima definizione di piccola area dovuta a Purcell e Kish che considerano piccole aree quelle formate da un numero di unità statistiche comprese tra 1/10 e 1/100 rispetto al totale della popolazione di riferimento.

Brackstone (1987) definisce piccola area una qualunque area per cui non si possono ottenere stime accurate (accettabili) con i metodi di stima classici dato un certo campione, ma è necessario utilizzare nuovi metodi di stima.

Secondo Rao (1994) le piccole aree sono aree geografiche di piccole dimensioni quali sezioni di censimento, frazioni, comuni, province, oppure sub-popolazioni identificate da certe caratteristiche come sesso, età, razza, etc. che appartengano ad una unità geografica più ampia, come ad esempio una nazione. Rao fonde il concetto spaziale di piccola area con la definizione di Brackstone, perciò si parla di piccola area, secondo Rao, se in quella zona geografica precedentemente descritta la numerosità campionaria specifica dell'area non consente di ottenere stime dirette di adeguata precisione (accettabili).

In effetti il termine piccola area non è inteso in senso assoluto ma in senso relativo poiché l'area è piccola rispetto alla dimensione campionaria. Per intenderci se un campione (siamo nell'assurdo) fosse estratto per l'intero pianeta potremmo considerare una nazione come piccola area. Sempre paradossalmente se estraiamo un qualsiasi campione da una popolazione con varianza tendente a zero non esisterebbe una piccola area (almeno secondo le definizioni di Rao e Brackstone) perché anche con una osservazione siamo in gradi di ottenere stime accettabili, anzi stime molto precise[9].

Per fare un esempio reale consideriamo l'US Survey (analisi campionaria di carattere generale fatta negli Stati Uniti). L'US Survey è un campione di numerosità

[9] Dato che l'intervallo di confidenza è $x \pm z_{\alpha/2} \cdot \sqrt{\text{var}(x)/n}$, se la varianza tende a 0 l'intervallo di confidenza degenere in un punto, in x.

10000 che copre tutta la popolazione degli Stati Uniti; le 10000 unità sono estratte nei 52 stati in base alla numerosità della popolazione residente. In questo modo la dimensione attesa del campione varia molto in base alla popolosità dello stato. Per esempio in California vengono rilevate 1207 unità, nello stato di New York 698, in Washington DC 22 e in Wyoming 18. Mentre per la California e lo stato di New York può funzionare uno stimatore post stratificato, per Washington DC e il Wyoming è necessaria una stima per piccole aree. In questo caso la "grande" area sono gli USA mentre piccole aree sono gli stati membri.
Questo esempio dimostra come il termine "piccola area" sia soprattutto legato al processo inferenziale e non ad una dimensione fisica.

Nella tesi la piccola area è intesa come unità geografica per la quale la numerosità campionaria, all'interno dell'area, non consente una stima affidabile con i metodi di stima tradizionali.

2.2 Panoramica sui Metodi di Stima per Piccole Aree

I metodi di stima per piccole aree si dividono in tre branche principali:

1. Metodi di stima per piccole aree basati su disegno.
2. Metodi di stima per piccole aree basati su modello.
3. Metodi di stima per piccole aree assistiti da modello.

2.2.1 Metodi di Stima per Piccole Aree Basati su Disegno

Sono metodi che si basano su sistemi di campionamento classico, come il campionamento stratificato presentato nel capitolo 1. Con tale metodo si riportano le osservazioni del campione alle piccole aree e si procede alla stima diretta per ognuna di esse. In alcuni casi, quando si sa a priori dell'indagine quali sono le piccole aree di interesse, si stabilisce di considerare le piccole aree come strati e di impostare un opportuno campionamento stratificato.
I metodi di stima diretti (classici) si basano sulla distribuzione di probabilità indotta dal campionamento e considerano il parametro da stimare, o sue funzioni, costante. In pratica, si presume che il parametro incognito da stimare sia compreso, con una certa probabilità (significatività), in un intervallo determinato dalla variabilità della popolazione, che viene stimata considerando la probabilità di inclusione degli elementi della popolazione nel campione.

L'obiettivo del disegno di campionamento è la minimizzazione dell'errore quadratico medio, grandezza legata alla varianza e alla distorsione dello stimatore, tenendo conto di un certo budget di spesa.

In statistica l'errore quadratico medio (d'ora in avanti MSE, acronimo di Mean Squared Error) di uno stimatore $\hat{\theta}$ di un parametro incognito non osservabile θ è definito come:

$$MSE = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + (Bias(\hat{\theta}))^2$$

Dove $V(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta})^2]$ è la varianza attesa dello stimatore $\hat{\theta}$ ed $E[\hat{\theta}] = \sum_{s \in S} p(s)x(s)$ (se esiste una funzione di densità per la variabile casuale x) è il valore atteso (media) dello stimatore $\hat{\theta}$. Bias($\hat{\theta}$) = ($E[\hat{\theta}] - \theta$) è la distorsione attesa dello stimatore $\hat{\theta}$, cioè di quanto si discosta in media $\hat{\theta}$ dal vero valore incognito θ.

Se lo stimatore è corretto ($E[\hat{\theta}] = \theta$) segue che il MSE coincide con la varianza attesa poiché $E[\hat{\theta}] = \theta$. Infatti $V(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta})^2] = E[(\hat{\theta} - \theta)^2]$ = MSE.

Come abbiamo visto nel capitolo 1 la varianza di uno stimatore, e quindi il MSE in caso di correttezza della stima, è inversamente proporzionale alla numerosità campionaria; quest'ultima però è direttamente proporzionale al costo di rilevazione. Si cerca nella fase di progettazione dell'indagine di ottimizzare il trade-off che esiste tra costi e bontà delle stime, ottimizzazione che possiamo ottenere con processi matematici oppure affidarci all'istinto e alle conoscenze pregresse dei ricercatori (in poche parole andare "a naso").

Nella stima per piccole aree il problema della numerosità campionaria si accentua poiché per ottenere stime accettabili per ogni piccola area è necessaria una certa numerosità campionaria cosa che può tradursi in costi ingenti di rilevazione, soprattutto se le piccole area sono numerose.

Un'attenta stratificazione può ridurre la numerosità campionaria (ma non in modo drastico); il metodo comunemente utilizzato consiste nel campionare in ogni strato in modo proporzionale rispetto alla dimensione della popolazione (negli strati).

Con questa strategia, però, si ripresenta il problema di partenza, per cui in aree poco popolate la numerosità campionaria può risultare insufficiente per avere stime accettabili. Una proposta per ovviare a questo problema è quella di diminuire la numerosità del campione nelle aree molto popolate ed aumentarla altrove.

Nella realtà (almeno italiana) la maggior parte delle indagini (che garantiscono un archivio di qualità del frame)[10] è svolta su vasta scala; questo implica che lo strato è generalmente a livello di regione o provincia. Per stime su aggregati più piccoli, come comuni o frazioni, non si dispone di una numerosità campionaria necessaria per ottenere risultati accettabili. In questi casi è consigliabile utilizzare metodi alternativi, che verranno presentati in seguito (vedi capitolo 2 paragrafo 2.3 e seguenti e capitolo 3)[11].

Nel caso che si ritenga comunque utile procedere alla stima diretta gli stimatori diretti più diffusi sono quello di Horvitz e Thompson, quello post-stratificato e lo stimatore rapporto; per ognuno forniremo la stima del valore medio, o del totale, e la stima del MSE o della varianza (poiché in questi casi coincidono), fatta eccezione per lo stimatore rapporto per cui presenteremo solo la stima del totale[12]:

- Stimatore (del totale) Horvitz e Thompson per area

$$\hat{Y}_{\pi,i} = \sum_{s_i} I_j \cdot \frac{y_j}{\pi_j}$$

Dove $\hat{Y}_{\pi,i}$ è la stima della somma della variabile Y nell'area i-esima, con $Y = \sum_{U_i} y_i$. U_i è la popolazione dell'area i, s_i è il campione dell'area i, I_j è una funzione identica che ha valore 1 per $j \in s_i$ e 0 altrimenti, y_j è il valore della variabile Y osservata sull'unità j mentre π_j è la probabilità che j appartenga al campione condizionata dalla probabilità di inclusione del campione s_i ($\pi_j = Pr(j \in s \mid P(s))$). Si dimostra che $\hat{Y}_{\pi,i}$ è una stima corretta di Y.

La varianza attesa dello stimatore (e data la sua correttezza il MSE) è

$$V(\hat{Y}_{\pi,i}) = \sum\sum_{U_i} (\pi_{jk} - \pi_j \pi_k) \cdot \frac{y_j}{\pi_j} \frac{y_k}{\pi_k} \qquad (2.1)$$

[10] In riferimento all'archivio di qualità vedere ISTAT "Metodi e Norme"

[11] I metodi di stima per piccole aree basati su disegno utilizzano solo le informazioni campionarie e sono sostanzialmente derivati dai metodi di stima classici, per questo la bontà delle stime rimane legata alla numerosità campionaria.

[12] Non è possibile risolvere con questi metodi il problema della scarsa numerosità delle osservazioni nelle piccole aree, ne è previsto l'utilizzo di informazioni provenienti da altre indagini per il miglioramento delle stime (se non marginalmente nello stimatore rapporto).

Dove π_{jk} è la probabilità che gli elementi *j* e *k* appartengano al campione *s*, cioè *Pr*(*j*, *k* $\in$ *s*) = *Pr*($I_j I_k$ = 1). Ovviamente non possiamo calcolare la (2.1), poiché non osserviamo tutti gli elementi della popolazione, così procediamo ad una sua stima:

$$\hat{V}(\hat{Y}_{\pi,i}) = \sum\sum_{s_i} \frac{(\pi_{jk} - \pi_j \pi_k)}{\pi_j^2} \cdot \frac{y_j}{\pi_j} \frac{y_k}{\pi_k}$$

Per ogni $\pi_{jk} > 0$ la stima della varianza è corretta.

Generalmente siamo interessati più alla media di un certo carattere della popolazione che non alla somma dello stesso. In questo caso è sufficiente conoscere la numerosità della popolazione nell'area *i*, cioè N_i, per ottenere la stima del valore medio del carattere *Y*:

$$\hat{\bar{Y}}_{\pi,i} = \frac{\hat{Y}_{\pi,i}}{N_i}$$

La stima della varianza (utilizzando le note proprietà):

$$\hat{V}(\hat{\bar{Y}}_{\pi,i}) = \frac{\hat{V}(\hat{Y}_{\pi,i})}{N_i^2}$$

- Stimatore post-stratificato del totale della variabile *Y*:

$$\hat{Y}_{post,i} = \frac{N_i \sum_{s_i} w_j y_j}{\sum_{s_i} w_j} \qquad (2.2)$$

Dove N_i è la numerosità della popolazione nell'area *i* (se è incognita va stimata), w_j è il peso campionario dell'unità *j* ottenuto con il rapporto $\frac{n_i}{N_i}$ (con n_i numerosità del campione nell'area *i*). Lo stimatore proposto è corretto.

La stima della varianza (considerando N_i noto):

$$\hat{V}\ [\hat{Y}_{post,i}] = (1-\frac{n_i}{N_i})N_i^2\frac{\hat{S}_y^2}{n_i} \qquad (2.3)$$

Dove $\hat{S}_y^2 = \frac{1}{n_i - 1}\sum_{s_i}(y_k - \bar{y}_i)^2$; la stima proposta è corretta e interpretabile anche come MSE, data la correttezza dello stimatore.

- Stimatore rapporto:

Si caratterizza per l'uso di una covariata:

$$\hat{Y}_{r,i} = X_i\hat{R}_i \quad (2.9)$$

Dove X_i è la variabile ausiliare osservata sull'area i, $\hat{R}_i = \frac{\hat{Y}_{e,i}}{\hat{X}_{e,i}}$ è la stima del rapporto $R_i = \frac{Y_i}{X_i}$. Y_i rappresenta il totale della variabile Y nell'area i. $\hat{Y}_{e,i} = \sum_{s_i}\frac{y_k}{\pi_k}$ e $\hat{X}_{e,i} = \sum_{s_i}\frac{x_k}{\pi_k}$ sono le stime rispettivamente della somma di Y e X nell'area i.

Risulta evidente che se la stima di R è corretta $\hat{Y}_{r,i} = Y_i$. $\hat{R}$ è una stima approssimativamente corretta (poiché risulta da una approssimazione lineare di Taylor).
La stima proposta è la più semplice tra gli stimatori rapporto; esistono stimatori di $\hat{Y}_{r,i}$ più complessi che tengono conto contemporaneamente degli strati e delle piccole aree.
Data la difficoltà di derivazione della stima della varianza, questa non verrà trattata, per essa si rimanda a R.L. Chambers, C.J. Skinner, (2003); inoltre nel proseguimento della tesi non verrà utilizzato lo stimatore rapporto (nemmeno lo stimatore Horvitz Thompson), per questo è stato presentato solo marginalmente.

Degli stimatori proposti useremo solo lo stimatore post-stratificato (2.2, 2.3) per confrontare i risultati con le stime ottenute con metodi presentati in seguito (vedi capitolo 2 paragrafo 2.3 e seguenti e capitolo 3). Gli altri sono stati presentati per

completezza, considerando che sono i metodi di stima tradizionali più usati nella categoria dei metodi di stima basati su disegno.

2.2.2 METODI DI STIMA PER PICCOLE AREE BASATI SU MODELLO

Questi metodi privilegiano lo studio del legame tra *y* ed eventuali variabili ausiliarie, trascurando il disegno campionario in favore di modelli probabilistici in cui si considerano gli effetti casuali a livello a di area e tra le variabili ausiliare incluse nel modello.

L'approccio prevede l'inserimento di un modello probabilistico di superpopolazione relativo alla relazione tra variabili d'interesse e le variabili ausiliarie da cui deriva il revisore ottimo a livello di piccola area (Chiandotto, 1996).

Nei metodi basati su modello siamo interessati a fare inferenza sul parametro da stimare, facendo delle assunzioni iniziali sulla popolazione che genera il valore di studio *y*. Generalmente nell'approccio basato su modello si assume di avere una superpopolazione infinita di variabili osservabili *y*, ognuna con una certa media e varianza.

Il disegno di campionamento è del tutto irrilevante ai fini inferenziali come discusso in Rubin (1976), Scott (1977a) e Sugden e Smith (1984).

E' di fondamentale importanza specificare che, data l'indipendenza della stima dal disegno campionario e dato l'utilizzo di un modello probabilistico, si può ottenere una stima anche per un dominio non campionato; ovviamente ciò è possibile sotto specifiche condizioni. La trattazione di questi casi non sarà approfondita in questo contesto (vedi Chambers, R. L., Skinner C. J. (2003), Saey, A., Chambers, R. L. (2003)).

Anche con questi metodi siamo interessati principalmente a due aspetti, ottenere la stima corretta, efficiente e consistente[13] di un parametro ed ottenere la stima corretta del suo MSE.

Adesso prendiamo in considerazione il problema che dobbiamo affrontare secondo il nuovo approccio. Abbiamo una serie di osservazioni, estratte da una certa popolazione, riferite ad un'area geografica. Tale area è suddivisa (da noi, o a livello geografico, o a livello politico) in sotto-aree, piccole aree, di dimensioni tangibilmente più piccole, senza che nessuna area del territorio ne resti esclusa. Si presume che in ogni

[13] Quando uno stimatore ha proprietà di correttezza, efficienza e consistenza è chiamato stimatore ottimo.

piccola area ci siano poche osservazioni (ma almeno 1), comunque in numero non sufficiente per fare stime "classiche" ed ottenere buoni risultati. Per ogni piccola area si suppone inoltre di avere delle informazioni ausiliarie (covariate). Tramite un modello statistico che si avvale delle osservazioni e delle informazioni ausiliarie si stima la media (o il totale) di un certo carattere della popolazione ed il suo MSE, con l'obiettivo di ottenere stime corrette ottime.

I metodi basati su modello per la stima per piccole aree si dividono in due categorie identificate dalle informazioni disponibili sull'area:

1. Modelli a livello di area: si stima il parametro utilizzando covariate specifiche della piccola area nella sua interezza (cioè ho un unico dato riassuntivo della piccola area per ogni covariata). Tale metodo è da usare quando non si conoscono le covariate per tutte le unità osservate.

2. Modelli a livello di unità: si stima il parametro utilizzando le covariate osservate su ogni unità della piccola area.

2.2.2.1 MODELLO A LIVELLO DI AREA

Prima di presentare il modello specifichiamo che si considerano m piccole aree denotate in pedice con l'indice i ($i = 1 \ldots p$), il parametro d'interesse è θ, i vettori (sempre colonna) e le matrici sono in grassetto e gli stimatori sono identificati dal segno ^ posto sopra il parametro; consideriamo p covariate, indicate con la lettera $\boldsymbol{x}$, con $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ il vettore delle covariate dell'area i.
Le osservazioni del campione le denotiamo con $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^T$ e indichiamo con $Y = [Y_1, Y_2, \ldots, Y_m]^T$ i valori riferiti alla popolazione.

Nel modello a livello di area (Area Level Random Effect Model) si assume un legame tra il parametro d'interesse nell'area i, θ_i, e la media della variabile da stimare nell'area $\overline{Y}_i$:

$$\theta_i = g(\overline{Y}_i) \tag{2.4}$$

Noi supponiamo un legame lineare, cioè la funzione g è linere, rispetto alle covariate $\boldsymbol{x}_i$:

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + z_i u_i \tag{2.5}$$

Dove $i = 1 \ldots m$, $\boldsymbol{\beta}$ è il vettore dei parametri di regressione, z_i è una costante positiva nota e u_i sono variabili casuali[14] indipendenti e identicamente distribuite con media 0 ($E[u_i] = 0$) e varianza σ_u^2 ($V[u_i] = \sigma_u^2$).

Per stimare il parametro θ_i ci si avvale del modello:

$$\hat{\theta}_i = \theta_i + e_i \tag{2.6}$$

Dove e_i è l'errore di campionamento nell'area i, con $E[e_i|\theta_i] = 0$ e $V[e_i|\theta_i] = \sigma_{ei}^2$.

Dalla (2.4) e (2.5) segue il modello lineare ad effetti misti proposto da Fay e Herriot (1979):

$$\hat{\theta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + z_i u_i + e_i$$

La stima (2.6) è composta da una componente sistematica $\boldsymbol{x}_i^T \boldsymbol{\beta}$, e da due componenti casuali $z_i u_i$ che rappresenta gli effetti dovuti all'area i (la piccola area trattata) ed e_i che rappresenta l'errore dovuto al campionamento. Si ipotizza che le u_i e le e_i siano non correlate (cioè $E[u_i e_i] = 0$, non esiste un effetto casuale tra area ed errore). Secondo questa ipotesi e le ipotesi fatte precedentemente (rispetto alla (2.4) e (2.5)) si dimostra che:

$$E[\hat{\theta}] = \boldsymbol{x}_i^T \boldsymbol{\beta} + z_i E[u_i] + E[e_i] = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

La stima è corretta[15] poiché assume i valori medi $\boldsymbol{x}_i^T \boldsymbol{\beta}$ ipotizzati[16]. Ci preoccuperemo in seguito della stima degli effetti casuali di area e dei $\boldsymbol{\beta}$.

Possiamo velocemente derivare la varianza dello stimatore tramite le ipotesi fatte per (2.4) e (2.5):

[14] Vedi A. M. Mood e F. A. Graybill, Introduzione alla statistica McGraw-Hill 1991, per un approfondimento sulle variabili casuali.

[15] In generale: lo stimatore è corretto, la stima (quale determinazione dello stimatore) non ha proprietà; vedi in proposito ai termini stima e stimatore la NOTA di pag. 3.

[16] Se $E[e_i|\theta_i] = 0$ non è un'assunzione valida la stima del modello può essere discutibile (Rao 2003).

$$V(\hat{\theta}) = z_i^2 \sigma_u^2 + \sigma_{ei}^2$$

dove σ_u^2 è la varianza dell'effetto casuale di area ,u_i, e σ_{ei}^2 è la varianza dovuta al campionamento. Per ora ci limitiamo a supporre noti questi valori. Procederemo ad una loro stima, e quindi alla stima della varianza, dopo aver introdotto il modello a livello di unità e i metodi di stima assistiti da modello.

2.2.2.2 Modello a Livello di Unità

I metodi di stima a livello di unità assumono che le covariate siano note per ogni unità presente nel campione. Le osservazioni campionarie sono indicate con y_{ij}, con $i = 1 \ldots m$ e $j = 1 \ldots n_i$, dove i indica l'area e j l'unità osservata all'interno dell'area; y_{ij} è l'unità j-esima osservata nell'area i appartenente ad un campione di n elementi suddiviso in m aree ($n = \sum_{i=1}^{m} n_i$). Le covariate, riferite ad ogni osservazione campionaria, sono indicate con $\boldsymbol{x}_{ij} = [x_{ij1}, x_{ij2}, \ldots, x_{ijp}]^T$, dove x_{ij} è il vettore delle p covariate relative all'elemento ij del campione (il j-esimo dell'area i). Seguendo la logica del modello a livello di area possiamo immaginare un legame lineare tra l'osservazione y_{ij}, le covariate (x_{ij}) e l'effetto casuale dipendente dall'area i:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij} \qquad (2.7)$$

dove $\boldsymbol{x}_{ij}^T \boldsymbol{\beta}$ è il valore atteso di y_{ij}, u_i sono gli effetti casuali dell'area i ed e_{ij} sono gli effetti casuali dovuti al campionamento. Le ipotesi sulla (2.7) sono che $E[u_i] = 0$, $V[u_i] = \sigma_{ui}^2$ e $E[e_{ij}] = 0$, $V[e_{ij}] = \sigma_{ei}^2$; si suppone altresì che $E[u_i e_i] = 0$ (gli effetti casuali di area e di campionamento non sono correlati tra loro). Considerate le ipotesi fatte risulta immediato che:

$$E[y_{ij}] = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + E[u_i] + E[e_{ij}] = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}$$

Mentre la varianza attesa di y_{ij} risulta:

$$V[y_{ij}] = \sigma_{ui}^2 + \sigma_{ei}^2$$

La stima dei parametri da cui dipende il modello presentato verrà presentata in seguito, nei paragrafi 2.6 e seguenti di questo capitolo e nel capitolo 3.

2.2.3 Metodi di Stima per Piccole Aree Assistiti da Modello

Questi metodi rappresentano un punto d'incontro tra i metodi basati su disegno e i metodi basati su modello. Il termine è dovuto a Särdnal (1993) ed è riferito ai metodi per i quali l'inferenza si basa sia su disegno che su modello. I metodi di stima assistiti da modello sono molteplici, ma tutti si basano sull'esistenza di uno stimatore corretto per il parametro di area, ottenuto sulla sola base delle informazioni campionarie e quindi indipendente da un qualsiasi modello. Su tale stima si integra un modello deciso dal ricercatore che lega la variabile oggetto di studio con le covariate osservate.

Presenteremo soltanto due metodi di stima per piccole aree assistiti da modello: lo stimatore sintetico e lo stimatore combinato.

2.2.3.1 Lo Stimatore Sintetico

Lo stimatore sintetico è uno stimatore indiretto che, utilizzando la stima diretta della variabile d'interesse ottenuta su una grande area, deriva stime per piccole aree, appartenenti alla grande area di riferimento, assumendo la similitudine per certe caratteristiche tra la grande area e le piccole aree (Gonzalez, 1973).

Vediamo il caso più semplice di stima della media di una generica variabile Y ($\overline{Y}$). Se non conosciamo informazioni ausiliare sulle piccole aree siamo "costretti" a ipotizzare che la media di piccola area sia uguale (molto simile) alla media della grande area:

$$\hat{\overline{Y}}_{\text{sin},i} = \hat{\overline{Y}} = \frac{\hat{Y}}{\hat{N}} = \frac{\sum_s w_j y_j}{\sum_s w_j}$$

dove s è l'insieme delle osservazioni presenti nella grande area, y_j sono le osservazioni e w_j sono i pesi campionari di ogni osservazione e "i" rappresenta la piccola area. Se utilizziamo un disegno di campionamento stratificato, si considera come grande area lo

strato, al cui interno ci sono le piccole aree; otteniamo così lo stimatore del totale (Purcell e Linacre, 1976):

$$\hat{Y}_{sin,i} = \sum_{s_h} N_{h,i} \hat{\bar{Y}}_h = \sum_h N_{h,i} \frac{\sum_{s_h} w_j y_j}{\sum_{s_h} w_j}$$

Dove *h* rappresenta lo strato (la somma sull'insieme h rappresenta la somma per tutti gli strati), *i* è la piccola area, s_h è il campione estratto dallo strato *h* ed $N_{h,i}$ è la dimensione della popolazione dello strato *h* che include anche la popolazione della piccola area *i*.
Tuttavia questi stimatori che si basano su un'ipotesi di omogeneità ("la piccola area è uguale alla grande area") risultano distorti se tale ipotesi non descrive bene la realtà.
Il MSE dello stimatore è, seguendo la definizione:

$$MSE(\hat{Y}_{sin,i}) = E[(\hat{Y}_{sin,i} - Y_i)^2] = E[((\hat{Y}_{sin,i} - \hat{Y}_i) + (\hat{Y}_i - Y_i))^2] =$$
$$= E[(\hat{Y}_{sin,i} - \hat{Y}_i)^2] - V(\hat{Y}_{sin,i} - \hat{Y}_i) + V(\hat{Y}_{sin,i})$$

dove $\hat{Y}_i$ è uno stimatore diretto e corretto di Y_i. La stima del MSE proposto risulta instabile (Gonzalez e Waksberg, 1973), perciò si stima il MSE rispetto allo stimatore sintetico della media, $\hat{\bar{Y}}_{sin,i}$, che però richiede la conoscenza della popolazione in ogni piccola area, indicata con N_i:

$$\widehat{MSE}(\hat{\bar{Y}}_{sin,i}) = \frac{\widehat{MSE}(\hat{Y}_{sin,i})}{N_i^2} = \frac{(\hat{Y}_{sin,i} - \hat{Y}_i)^2 - \hat{V}(\hat{Y}_{sin,i} - \hat{Y}_i) + \hat{V}(\hat{Y}_{sin,i})}{N_i^2}$$

Con questo sistema si riesce a stabilizzare la stima del MSE($\hat{Y}_{sin,i}$). Per un ulteriore approfondimento sull'argomento vedere (Gonzalez e Waksberg, 1973).

Nel caso in cui si conoscono informazioni ausiliarie sulle variabili osservate si utilizza lo stimatore sintetico di regressione, che per la stima del totale è:

$$\hat{Y}_{sin,reg,i} = \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}$$

Dove $\boldsymbol{X}_i$ rappresenta l'insieme delle covariate relative all'area *i*, con $\boldsymbol{X}_i$ matrice n_i per *p* dove il generico elemento x_{jk} indica la *j*-esima osservazione del carattere *k*-esimo (nell'area *i*). La stima di $\boldsymbol{\beta}$ si ottiene con:

$$\hat{\boldsymbol{\beta}} = \frac{\sum_s (w_j \boldsymbol{x}_j y_j)/c_j}{\sum_s (w_j \boldsymbol{x}_j \boldsymbol{x}_j^T)/c_j} \tag{2.8}$$

Lo stimatore $\hat{Y}_{sin,reg,i}$ proposto è distorto. Tale distorsione è uguale a $\boldsymbol{X}_i^T \boldsymbol{\beta} - Y_i$ dove $\boldsymbol{\beta}$ è calcolato con la (2.8) su tutta la popolazione. E' vero però che se β_i (calcolato con la (2.8) su tutta la popolazione dell'area *i*) è simile al coefficiente di regressione β ed è vero il legame tra la variabile di interesse e le covariate allora la distorsione è trascurabile. Secondo Rao questo è vero se $c_j = \boldsymbol{v}^T \boldsymbol{x}_j$ con $\boldsymbol{v}$ vettore di costanti.

2.2.3.2 LO STIMATORE COMBINATO

Questo stimatore, in perfetta sintonia con la filosofia dei metodi assistiti da modello, combina una stima diretta con una stima indiretta al fine di ridurre la variabilità complessiva. In pratica si utilizza la stima diretta, affetta da un'elevata variabilità, come base da correggere con l'aiuto di una stima indiretta, che si avvale di informazioni ausiliarie.

Il concetto appena espresso si traduce in una stimatore che utilizza la media aritmetica ponderata per bilanciare il peso dello stimatore diretto con il peso dello stimatore indiretto:

$$\hat{Y}_{c,i} = \gamma_i \hat{Y}_i^D + (1-\gamma_i)\hat{Y}_i^I$$

Dove $\hat{Y}_{c,i}$ è lo stimatore combinato del totale, $\hat{Y}_i^D$ è lo stimatore diretto del totale e $\hat{Y}_i^I$ è lo stimatore indiretto del totale, tutti riferiti alla piccola area *i*. γ_i è il peso assegnato allo stimatore diretto nell'area *i* ed ovviamente $(1-\gamma_i)$ è il peso dello stimatore indiretto per l'area *i*. La somma dei pesi è uguale a 1 ($\gamma_i + (1-\gamma_i) = 1$) e si considera $0 \leq \gamma_i \leq 1$.

Lo stimatore $\hat{Y}_{c,i}$ è strettamente legato al parametro γ_i, che è sconosciuto. Si capisce quindi che la scelta del peso riveste un ruolo chiave nella stima combinata. Il

valore del peso γ_i si ottiene minimizzando il MSE dello stimatore combinato. Innanzitutto definiamo il MSE dello stimatore combinato:

$$MSE(\hat{Y}_{c,i}) = E[(\hat{Y}_{c,i} - Y_i)^2] = E[(\gamma_i \hat{Y}_i^D + (1-\gamma_i)\hat{Y}_i^I - Y_i)^2] =$$
$$= E[(\gamma_i \hat{Y}_i^D + (1-\gamma_i)\hat{Y}_i^I - (\gamma_i \hat{Y}_i + (1-\gamma_i)\hat{Y}_i))^2] =$$
$$= \gamma_i^2 E[(\hat{Y}_i^D - Y_i)^2] + (1-\gamma_i)^2 E[(\hat{Y}_i^I - Y_i)^2] + 2\gamma_i(1-\gamma_i)E[(\hat{Y}_i^D - Y_i)(\hat{Y}_i^I - Y_i)] \quad (2.9)$$

Minimizzando la (2.9) rispetto a γ_i si ottiene γ_i^*, il peso che minimizza il MSE:

$$\gamma_i^* = \frac{MSE(\hat{Y}_i^I) - E[(\hat{Y}_i^I - Y_i)(\hat{Y}_i^D - Y)]}{MSE(\hat{Y}_i^D) + MSE(\hat{Y}_i^I) - 2E[(\hat{Y}_i^I - Y_i)(\hat{Y}_i^D - Y)]} \quad (2.10)$$

La (2.10) può essere approssimata come:

$$\gamma_i^* \simeq \frac{MSE(\hat{Y}_i^I)}{V(\hat{Y}_i^D) + MSE(\hat{Y}_i^I)} \quad (2.11)$$

L'approssimazione vale se $\hat{Y}_i^D$ è uno stimatore corretto (ipotesi che abbiamo fatto all'inizio) e se $COV(\hat{Y}_i^D, \hat{Y}_i^I) = 0$. Infatti se uno stimatore è corretto il *MSE* = *V* (l'errore quadratico medio coincide con la varianza, vedi paragrafo 2.2.1), nel nostro caso $MSE(\hat{Y}_i^D) = V(\hat{Y}_i^D)$, e se ipotizziamo che non ci sia una relazione tra la stima diretta e quella indiretta risulta $E[(\hat{Y}_i^I - Y_i)(\hat{Y}_i^D - Y)] = COV(\hat{Y}_i^D, \hat{Y}_i^I) = 0$. Ottenere un peso che minimizza il MSE è un vantaggio notevole della stima combinata. Infatti non solo si tiene conto dei metodi di stima basati su disegno e basati su modello ma si riesce a ponderarli in maniera efficiente rispetto ai rispettivi MSE.

Il peso ottenuto, γ_i^*, è compreso tra 0 e 1. Vale 1 se la varianza dello stimatore diretto è 0 (caso limite), in questo caso lo stimatore combinato coinciderà con la stima diretta:

$$\hat{Y}_{c,i} = \gamma_i^* \hat{Y}_i^D + (1-\gamma_i^*)\hat{Y}_i^I = 1 \cdot \hat{Y}_i^D + (1-1)\hat{Y}_i^I = \hat{Y}_i^D$$

Vale 0 se il MSE dello stimatore indiretto è 0, in questo caso la stima combinata coincide con la stima indiretta:

$$\hat{Y}_{c,i} = \gamma_i^* \hat{Y}_i^D + (1-\gamma_i^*)\hat{Y}_i^I = 0 \cdot \hat{Y}_i^D + (1-0)\hat{Y}_i^I = \hat{Y}_i^I$$

Con l'utilizzo del peso, così calcolato, siamo in grado di dare maggiore importanza allo stimatore con una minore variabilità, senza dover rinunciare (se non nei casi limite) alle informazioni apportate dall'altro stimatore. La stima combinata risulta essere un compromesso tra due stimatori differenti bilanciati da una scelta "intelligente" del peso; si dimostra che se $max(0, 2\gamma_i^* - 1) \leq \gamma_i \leq min(2\gamma_i^*, 1)$ allora lo stimatore combinato è migliore in termini di MSE degli stimatori che lo compongono.

Per calcolare γ_i^* abbiamo bisogno della stima della varianza di $\hat{Y}_i^D$ e della stima del MSE di $\hat{Y}_i^I$. Un metodo proposto è:

$$\hat{\gamma}_i^* = \frac{\widehat{MSE}(\hat{Y}_i^I)}{(\hat{Y}_i^I - \hat{Y}_i^D)^2} \tag{2.12}$$

Dove $\widehat{MSE}(\hat{Y}_i^I) = (\hat{Y}_i^I - \hat{Y}_i^D)^2 - \hat{V}(\hat{Y}_i^D)$, con $\hat{V}(\hat{Y}_i^D)$ stimato con i metodi proposti nel capitolo 1. Tuttavia la stima del MSE dello stimatore indiretto risulta instabile, per questo Purcell e Kish (1980) propongono di utilizzare un peso comune per tutte le aree, ottenuto minimizzando la somma dei MSEs, riferiti alla singola area, rispetto a γ_i:

$$min \sum_{i=1}^{m} MSE(\hat{Y}_{c,i}) \text{ rispetto a } \gamma_i \rightarrow \gamma^* = \frac{\sum_{i=1}^{m} MSE(\hat{Y}_i^I)}{\sum_{i=1}^{m} (V(\hat{Y}_i^D) + MSE(\hat{Y}_i^I))} \tag{2.13}$$

Dove *m* sono il numero di piccole aree e le altre notazioni mantengono lo stesso significato così come le ipotesi restano le stesse (la (2.13) segue intuitivamente dalla (2.11)). Per stimare i pesi ci si basa sulla (2.12), in questo modo si ottiene immediatamente:

$$\hat{\gamma}^{*} = \frac{\sum_{i=1}^{m} \widehat{MSE}(\hat{Y}_i^I)}{\sum_{i=1}^{m} (\hat{Y}_i^I - \hat{Y}_i^D)^2} = \frac{\sum_{i=1}^{m} ((\hat{Y}_i^I - \hat{Y}_i^D)^2 - \hat{V}(\hat{Y}_i^D))}{\sum_{i=1}^{m} (\hat{Y}_i^I - \hat{Y}_i^D)^2}$$

Con questo metodo otteniamo un peso stabile. Tuttavia se ne sconsiglia l'utilizzo se non c'è omogeneità tra le varianze della stima diretta. Se in una o più aree sono presenti caratteristiche peculiari lo stimatore diretto potrebbe apportare molta informazione nella stima combinata, ma utilizzando un peso comune, ottenuto considerando tutte le aree, il bilanciamento tra gli stimatori è omogeneo e non si considera il contributo maggiore dello stimatore diretto.

Esistono altri metodi per calcolare i pesi di uno stimatore combinato (per un approfondimento Gosh e Rao 1994. In ultima istanza se non si trovano procedure soddisfacenti si può scegliere soggettivamente il valore dei pesi.

Nonostante i problemi che possono sorgere per stimare γ_i lo stimatore combinato è molto usato in letteratura. Nei successivi capitoli verranno proposti modelli di stima basati su questo stimatore.

2.3 Best Linear Unbiased Predictor

L'acronimo BLUP significa Best Linear Unbiased Predictor, predittore ottimo lineare e corretto. Il BLUP deriva dall'evoluzione e dall'adattamento dei modelli lineari ad effetti misti alla stima per piccole aree. Storicamente i primi modelli inerenti alla stima di sotto-domini vedono la luce nel 1806 in un lavoro pionieristico di Legendre. Si parla di modelli ad effetti fissi che spiegano la variabilità tra aree in relazione ad una variabile di interesse legata a caratteristiche note. Le prime stime per piccole aree con modelli ad effetti fissi si devono a Levy e French (1977), per le stime indirette, Schaible e al. (1977), per le stime composte e Holt e al. (1979), Sarndal (1984) e Marker (1999), per la stima dei predittori.

Ricoprono un ruolo di maggiore importanza, per quanto riguarda il BLUP, i modelli lineari ad effetti misti; in particolare tali modelli hanno la capacità di individuare relazioni lineari tra effetti fissi ed effetti casuali (Saei e Chambers, 2003). Il BLUP, che è un metodo per modelli ad effetti misti, si deve ad Henderson che lo ha sviluppato in una serie di papers tra il 1948 e il 1975.

Il primo campo di applicazione del BLUP è stato in ambito biologico, per studiare trend genetici su popolazioni animali.

Il metodo BLUP sviluppato da Henderson prevede che la varianza degli effetti casuali, nel modello ad effetti misti, sia nota. Ovviamente non è sempre possibile avere tale informazione, quindi va stimata in base ai dati a disposizione. Dal 1977 sono stati proposti diversi metodi per stimare la varianza degli effetti casuali, alternativi a quelli già proposti da Henderson stesso. Harville (1977) rivisita i diversi metodi di stima, tra cui la massima verosimiglianza, che è il metodo utilizzato in questo contesto. Quando non si conosce la varianza degli effetti casuali e ne facciamo una stima non si parla più di BLUP ma si parla di EBLUP. L'EBLUP verrà trattato nel paragrafo successivo.

L'uso dei modelli ad effetti misti nell'ambito della stima per piccole aree con dati censuari o campionari si deve soprattutto a Fay e Herriot (1979), Gosh e Rao (1994), Rao (1999) e Pfeffermann (1999). Si utilizza il modello ed effetti misti perché si immagina che un vettore di valori relativa a una popolazione finita sia riferito ad una superpopolazione (popolazione infinita); in questo modo, la stima della media in una piccola area è funzione della previsione degli effetti casuali tra aree, che non sono osservabili, in un modello ad effetti misti in base alla distribuzione della variabile di studio nella superpopolazione (Saei e Chambers, 2003).

In questo paragrafo presenteremo i modelli lineari ad effetti fissi e misti ed il modello BLUP, cercando di darne una prospettiva generale.

2.3.1 Il Modello Lineare ad Effetti Fissi

In statistica un modello lineare ad effetti fissi è tipicamente espresso come:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e} \tag{2.14}$$

Dove $\boldsymbol{y}$ è il vettore della variabile di studio di dimensione N, $\boldsymbol{X}$ è la matrice delle covariate di dimensione Nxp, $\boldsymbol{\beta}$ è il vettore dei coefficienti di regressione (che sono inosservabili) di dimensione p ed $\boldsymbol{e}$ è un vettore, di dimensione N, i cui valori sono determinati da una variabile casuale distribuita normalmente con media 0, varianza σ_e^2 e $\mathrm{COV}(e_i e_j) = 0$. Quindi il vettore **e** ha media $E[\boldsymbol{e}] = \mathbf{0}$ e matrice di varianza-covarianza $E[\boldsymbol{e}\boldsymbol{e}^T] = \sigma_e^2 \boldsymbol{I}$, con $\boldsymbol{I}$ matrice identica NxN. N è la dimensione della popolazione. Questo modello può essere applicato anche ai dati campionari, in questo caso avremo la stessa relazione lineare della (2.14) dove $\boldsymbol{y}$ è il vettore della variabile studio osservata sul campione di dimensione n, $\boldsymbol{X}$ è la matrice nxp delle covariate osservate sul campione, $\boldsymbol{\beta}$

ha le stesse caratteristiche ed $\boldsymbol{e}$ è sempre un disturbo white noise[17]. $\boldsymbol{\beta}$ è un valore incognito non osservabile che può essere stimato con diversi metodi, il più diffuso è il metodi dei minimi quadrati[18]. Chiamiamo la stima $\hat{\boldsymbol{\beta}}$. Adesso possiamo utilizzare il modello per "predire" un valore non incluso nel campione rispetto alla variabile di studio *y*:

$$y_{k,i} = \boldsymbol{x}_{k,i}^T \hat{\boldsymbol{\beta}}$$

Dove $y_{k,i}$ è l'unità *k* nell'area *i* (che non è inclusa nel campione), $\boldsymbol{x}_{k,i}$ è il vettore delle covariate rispetto all'unità *k* nell'area *i*. Una predizione è possibile a patto di conoscere le covariate delle unità non osservate nel campione; questa non è una ipotesi sconsiderata poiché si possono trovare informazioni in altre indagini o in indagini svolte precedentemente. Con questo approccio è stato proposto un predittore per la stima della media della popolazione nella piccola area *i*:

$$\hat{\bar{y}}_i = \frac{1}{N_i}\left(\sum_{k=1}^{n_i} y_{k,i} + \sum_{k=n_i+1}^{N_i} \boldsymbol{X}_{k,i}^T \hat{\boldsymbol{\beta}}\right) \qquad (2.14\text{bis})$$

Con N_i popolazione totale dell'area *i*, n_i dimensione campionaria area *i*, $y_{k,i}$ come sopra e $\boldsymbol{X}_{k,i}$ è la matrice ottenuta unendo i vettori $\boldsymbol{x}_{k,i}$ per $k = n_i+1, \ldots, N_i$.

Una trattazione più approfondita sulla predizione nella stima per piccole aree con modelli ad effetti fissi si può trovare in Gosh e Rao (1994).

2.3.2 Il Modello Lineare ad Effetti Misti

Il modello lineare ad effetti fissi prevede che la variabilità della variabile di studio $\boldsymbol{y}$ sia interamente spiegata dalle covariate $\boldsymbol{X}$. Ciò che rimane è soltanto un "disturbo" ($\boldsymbol{e}$), almeno in teoria. In pratica analizzando i residui $\boldsymbol{res} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ si nota che spesso differiscono dal disturbo atteso $\boldsymbol{e}$. Questo significa che i residui contengono informazioni sulla variabilità della $\boldsymbol{y}$ e che il modello è stato mal specificato. In

[17] Si intende appunto un insieme di valori generati da una variabile casuale distribuita normalmente con media 0, varianza σ_e^2 e assenza di autocorrelazione (covarianza 0), $e \sim N(0, \sigma_e^2)$, mentre se ci esprimiamo a livello di vettore $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$.

[18] Vedi Econometryc Analysis per un approfondimento sulla regressione e sui metodi di stima tradizionali.

particolar modo nell'ambito della stima per piccole aree utilizzando il modelli ad effetti fissi non si considera la variabilità dovuta alla relazione tra le aree, che rimane appunto "imprigionata" nei residui.

Il modello lineare ad effetti misti supera i limiti del modello ad effetti fissi, poiché è in grado di scindere l'informazione sulla variabilità di $\boldsymbol{y}$ apportata dalle covariate $\boldsymbol{X}$ dall'informazione presente negli effetti casuali tra aree.

Il modello proposto è:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

Vediamo in dettaglio componente per componente al fine di fare chiarezza:

- Trattiamo il caso di un qualsiasi campione di dimensione n, con n_i osservazioni in ognuna delle m piccole aree. Le osservazioni rilevate su tutti gli individui riguardano il carattere oggetto di studio y e p covariate.

- $\boldsymbol{y}$ è il vettore delle osservazioni, ha dimensione n ed è una variabile casuale con una qualunque distribuzione.

- $\boldsymbol{X}$ è la matrice delle p covariate osservate sul campione, ha dimensione nxp e l'elemento x_{jk} indica il valore della variabile k per l'individuo i-esimo. $\boldsymbol{X}$ non è considerata una variabile casuale (quindi $E[\boldsymbol{Xy}] = \boldsymbol{X}E[\boldsymbol{y}]$).

- $\boldsymbol{\beta}$ è il vettore dei coefficienti di regressione, ha dimensione p ed è un valore incognito e non osservabile, per questo motivo deve essere stimato.

- $\boldsymbol{Z}$ è una matrice di costanti note, ha dimensione nxm[19]. $\boldsymbol{Z}$ serve per selezionare gli effetti di $\boldsymbol{u}$ in base all'area di appartenenza[20], è una matrice di selezione.

- $\boldsymbol{u}$ è un vettore di variabili casuali. Nella stima per aree rappresenta l'effetto casuale tra aree. Ha dimensione m e:
  - $E[\boldsymbol{u}] = \boldsymbol{0}$ (ogni u_i ha media 0)

[19] Può avere dimensione mxm nel caso di rilevazioni singole per ogni area.
[20] Questo concetto verrà approfondito in paragrafi seguenti.

- $E[\boldsymbol{u}\boldsymbol{u}^T] = \boldsymbol{G}$ matrice incognita di varianza-covarianza.
- Nella stima per piccole aree che consideriamo in questo capitolo si assume che $E[u_i u_j] = 0$ per ogni $i \neq j$. Quindi $\boldsymbol{G}$, in questo caso, è una matrice diagonale.

- $\boldsymbol{e}$ è un vettore di variabili casuali distribuite normalmente, ha dimensione *nxe*:
  - $E[\boldsymbol{e}] = \boldsymbol{0}$ (ogni e_k ha media 0)
  - $E[\boldsymbol{e}\boldsymbol{e}^T] = \sigma_e^2 \boldsymbol{R}$, con $\boldsymbol{R}$ matrice diagonale nota e σ_e^2 incognito (ogni e_k ha varianza σ_e^2 e la covarianza tra gli e_k ed e_j è 0 per ogni $k \neq j$).

- Si ipotizza anche che la variabile casuale $\boldsymbol{e}$ e la variabile casuale $\boldsymbol{u}$ siano indipendenti. Questo si traduce in $E[\boldsymbol{u}\boldsymbol{e}^T] = E[\boldsymbol{e}\boldsymbol{u}^T] = 0$.

Il valore atteso e la varianza di $\boldsymbol{y}$ sono (vedi dettagli in Appendice A, pag. 138):

- $E[\boldsymbol{y}] = \boldsymbol{X}\boldsymbol{\beta}$

- $V(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \sigma_e^2 \boldsymbol{R}$

  Dove $\boldsymbol{G}$ e $\sigma_e^2 \boldsymbol{R}$ sono le componenti di varianza e sono sconosciute.

La variabile casuale $\boldsymbol{y}$ secondo il modello lineare ad effetti misti risulta essere:

$$\boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \sigma_e^2 \boldsymbol{R})$$

2.3.3 LO STIMATORE BLUP

Utilizzando il modello lineare ad effetti misti possiamo costruire uno stimatore per una qualunque piccola area:

$$\tilde{\mu} = \boldsymbol{l}^T \tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T \tilde{\boldsymbol{u}} \qquad (2.15)$$

Dove $\boldsymbol{l}$ è il vettore delle covariate dell'area di interesse, $\boldsymbol{m}$ è un vettore di 1 in corrispondenza dell'area di interesse e 0 altrimenti[21], mentre $\tilde{\boldsymbol{\beta}}$ e $\tilde{\boldsymbol{u}}$ sono stime[22] di $\boldsymbol{\beta}$ e $\boldsymbol{u}$.

Il modello di stima specificato affinché appartenga alla categoria BLUP deve ovviamente essere ottimo, lineare e corretto. Per questo, è necessario ottenere delle stime di $\boldsymbol{\beta}$ e $\boldsymbol{u}$ tali che:

$$\hat{\mu} = \boldsymbol{a}^T \boldsymbol{x} + b \text{ Stimatore lineare, con } \boldsymbol{a} \text{ e } b \text{ noti}$$

$$E[\hat{\mu}] = \mu \text{ Stima corretta}$$

$$E[(\hat{\mu} - \mu)^2] \leq E[(\tilde{\mu} - \mu)^2] \text{ Stima ottima}$$

Dove $\tilde{\mu}$ è un generico stimatore corretto e $\hat{\mu}$ è lo stimatore ottimo tra gli stimatori corretti.

Deriviamo le stime di $\boldsymbol{\beta}$ e $\boldsymbol{u}$ affinché la (2.15) sia una stima BLUP. Innanzitutto risulta evidente che lo stimatore $\tilde{\mu}$ è una combinazione lineare dei parametri incogniti β e u e dei vettori noti $\boldsymbol{l}$ ed $\boldsymbol{m}$. Più rigorosamente:

$$\text{Posto } \boldsymbol{a} = \begin{bmatrix} \boldsymbol{l} \\ \boldsymbol{m} \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{u}} \end{bmatrix}, \ b = 0 \rightarrow \hat{\mu} = \boldsymbol{a}^T \boldsymbol{x} + b = \boldsymbol{l}^T \tilde{\boldsymbol{\beta}} + \boldsymbol{m}^T \tilde{\boldsymbol{u}}$$

Per quanto riguarda la correttezza vediamo come possiamo stimare $\boldsymbol{\beta}$ e $\boldsymbol{u}$.

Si considera il generico stimatore $\hat{\mu} = \boldsymbol{a}^T \boldsymbol{y} + b$ dove $\boldsymbol{y}$ è il modello lineare ad effetti misti, cioè $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$. Affinché lo stimatore lineare con modello ad effetti misti sia corretto rispetto alla stima $\mu = \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{u}$, cioè $E[\hat{\mu}] = E[\mu]$, è necessario e sufficiente che $\boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{l}^T$ e b = 0 (Rao, 2003). In termini più schematici si ha la proposizione i:

i. sia $\hat{\mu} = \boldsymbol{a}^T \boldsymbol{y} + b$ e $\mu = \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{u}$ allora $E[\hat{\mu}] = E[\mu] \Leftrightarrow \boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{l}^T$ e b = 0.

[21] La composizione di questo vettore cambia a seconda che si utilizzi un modello a livello di unità o un modello a livello di area; questo aspetto verrà approfondito in seguito.

[22] Notare che β e u sono stime generiche, senza nessuna garanzia di correttezza, efficienza e consistenza.

Dimostrazione di i.:
considerando le ipotesi e le derivazioni fatte sinora risulta:

- $E[\,\hat{\mu}\,] = E[\,\boldsymbol{a}^T\boldsymbol{y} + b\,] = E[\,\boldsymbol{a}^T(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}) + b\,] =$
  $E[\boldsymbol{a}^T\boldsymbol{X\beta} + \boldsymbol{a}^T\boldsymbol{Zu} + \boldsymbol{a}^T\boldsymbol{e} + b] = \boldsymbol{a}^T\boldsymbol{X\beta} + b$

- $E[\,\mu\,] = E[\,\boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{u}\,] = \boldsymbol{l}^T\boldsymbol{\beta}$

Posto $E[\,\hat{\mu}\,] = E[\,\mu\,]$ risulta:

$$\boldsymbol{a}^T\boldsymbol{X\beta} + b = \boldsymbol{l}^T\boldsymbol{\beta} \rightarrow \boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{l}^T \text{ e } b = 0 \tag{2.16}$$

Rispettando la (2.16) per ottenere lo stimatore ottimo si trovano i $\boldsymbol{\beta}$ e gli $\boldsymbol{u}$ che minimizzano il MSE di $\hat{\mu}$:

$$MSE(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] \tag{2.17}$$

Rispettando sempre la (2.17) risulta inoltre che:

$$V(\hat{\mu} - \mu) = E[((\hat{\mu} - \mu) - E[(\hat{\mu} - \mu)])^2] = E[((\hat{\mu} - \mu) - 0)^2] = E[(\hat{\mu} - \mu)^2] = MSE(\hat{\mu}) \tag{2.18}$$

Quindi secondo la (2.18), possiamo minimizzare $V(\hat{\mu} - \mu)$. Per le note proprietà della varianza risulta che $V(\hat{\mu} - \mu) = V(\hat{\mu}) + V(\mu) - 2COV(\hat{\mu}, \mu)$.
Calcoliamo il valore di ciascun componente della varianza (2.18) per poi minimizzare rispetto a $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ (vedi approfondimento in appendice A pag. 138)

1ª componente:

$$V(\hat{\mu}) = \boldsymbol{a}^T(\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T + \boldsymbol{V})\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{a}$$

2ª componente:

$$V(\mu) = \boldsymbol{m}^T\boldsymbol{Gm}$$

3ª componente:

$$COV(\hat{\mu}, \mu) = \boldsymbol{a}^T \boldsymbol{ZGm}$$

Riprendendo la (2.18) e le tre componenti di varianza possiamo scrivere:

$$V(\hat{\mu} - \mu) = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{ZGm}$$

Adesso utilizzando il moltiplicatore di Lagrange si minimizza rispetto al vettore $\boldsymbol{a}$ considerando il vincolo $\boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{l}^T$ (il vincolo di correttezza). Il moltiplicatore di lagrange garantisce l'ottimo di una funzione considerando n vincoli; più rigorosamente si considera (approfondimento in appendice A pag. 140):

$$h(x, \lambda) = f(x) + \sum_{k=1}^{n} g_k(x)$$

Dove i vincoli $g_k(x)$ devono essere scritti nella forma $g_k(x) = 0$. La nostra funzione lagrangiana sarà dunque:

$$f(x) \to f(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{ZGm}$$

$$g_1(x) \to g_1(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{l}^T = 0$$

$$\text{Quindi } h(x, \lambda) \to h(\boldsymbol{a}, \lambda) = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{ZGm} + (\boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{l}^T)\lambda$$

Derivando la funzione lagrangiana rispetto al vettore $\boldsymbol{a}$ e ponendola uguale a 0 si trova il valore di λ che sostituito in $\boldsymbol{a}$ ci fornisce le stime ottime di $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ (si ricorda $\hat{\mu} = \boldsymbol{a}^T \boldsymbol{y} + b$):

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}$ è la stima di $\boldsymbol{\beta}$ (2.19)

- $\hat{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ è la stima di $\boldsymbol{u}$ (2.20)

Quindi $\hat{\mu} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.

La stima di $\boldsymbol{\beta}$ che abbiamo ottenuto è la stessa stima che otteniamo utilizzando il metodo GLS (Generalized Least Square), che garantisce stime corrette anche in presenza di autocorrelazione degli errori ed eteroschedasticità[23]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

$$\text{Con } \hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

Riepilogando $\hat{\mu} = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{u}}$ è lo stimatore BLUP dove le stime di $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ sono calcolate secondo la (2.19) e (2.20). Tuttavia tale metodo di stima ha degli inconvenienti, infatti esso prevede l'inversione della matrice $\boldsymbol{V}$, un calcolo che può essere molto oneroso a livello computazionale (anche per gli odierni computers).

Un metodo alternativo si deve ad Henderson (1950) che assumendo la normalità di $\boldsymbol{u}$ ed $\boldsymbol{e}$ definisce i predittori ottimi lineari e corretti $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ in base alla funzione di densità congiunta di $\boldsymbol{y}$ ed $\boldsymbol{u}$.

Introduciamo alcune nozioni per il calcolo della funzione di densità congiunta. Consideriamo n variabili casuali indipendenti distribuite come una normale standard $(Z_1, \ldots, Z_n)$ con ogni $Z_i \sim N(0,1)$ si dimostra che il vettore $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^T$ è distribuito come una normale multivariata con $E[\boldsymbol{Z}] = \boldsymbol{0}$ e $V[\boldsymbol{Z}] = \boldsymbol{I}$, con funzione di densità $f_{(z)} = (2\pi)^{-n/2}\exp(-½(\boldsymbol{Z}^T\boldsymbol{Z}))$. Sia $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{AZ}$, con $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^T$ e $\boldsymbol{X} = [X_1, \ldots, X_n]^T$, si dimostra che $E[\boldsymbol{X}] = \boldsymbol{\mu}$ e $V[\boldsymbol{X}] = \boldsymbol{AA}^T = \boldsymbol{M}$ ed $\boldsymbol{X}$ ha una funzione di densità $f_{(x)} = 2\pi^{-\frac{n}{2}}|\boldsymbol{M}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}[(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{M}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]\}$. La definizioni di funzione di densità congiunta delle variabili aleatore X e Y è $f_{X,Y}(x,y) = f_{Y|X}(y\,|\,x)f_X(x) = f_{X|Y}(x\,|\,y)f_Y(y)$ dove $f_{Y|X}(y\,|\,x)$ è la densità di Y dato X (cioè considero la X come una variabile non casuale) e $f_X(x)$ è la densità di X. Nel nostro caso ci serve la distribuzione di $\boldsymbol{u}$ e la distribuzione di $\boldsymbol{y}$ dato $\boldsymbol{u}$:

1. $f_u(\boldsymbol{u}) = 2\pi^{-\frac{m}{2}}|\boldsymbol{G}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u})\}$ considerato che $\boldsymbol{u} \sim N_{mult}(\boldsymbol{0},\boldsymbol{G})$ (2.21)

[23] Vedere W. H. Greene Econometric analysis Prentice Hall 2003, per un approfondimento.

2. $f_{y|u}(y \mid u) = 2\pi^{-\frac{n}{2}} \left|\boldsymbol{R}\right|^{-\frac{1}{2}} \exp\{-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})]\}$ (2.22)

(considerato che $(\boldsymbol{y} \mid \boldsymbol{u}) \sim N_{mult}(\boldsymbol{X\beta} - \boldsymbol{Zu}, \boldsymbol{R})$)

La (2.21) e la (2.22) si ricavano facilmente considerando le assunzioni finora fatte e la generica funzione di densità di una normale multivariata, inoltre la (2.22) si basa su:

- $E[\boldsymbol{y} \mid \boldsymbol{u}] = E[(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}) \mid \boldsymbol{u}] = \boldsymbol{X\beta} + \boldsymbol{Zu} + E[\boldsymbol{e}] = \boldsymbol{X\beta} + \boldsymbol{Zu}$

- $V(\boldsymbol{y} \mid \boldsymbol{u}) = E[(\boldsymbol{y} - E[\boldsymbol{y} \mid \boldsymbol{u}])(\boldsymbol{y} - E[\boldsymbol{y} \mid \boldsymbol{u}])^T] = E[(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} - \boldsymbol{X\beta} - \boldsymbol{Zu})(\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{+e} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T] = E[\boldsymbol{ee}^T] = \boldsymbol{R}$ [24]

La funzione di densità congiunta che ricaviamo dalla (2.21) (2.22) è:

3. $f_{u,y}(\boldsymbol{u}, \boldsymbol{y}) = f_{y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_u(\boldsymbol{u}) =$

$= 2\pi^{-\frac{n}{2}-\frac{m}{2}} \left|\boldsymbol{R}\right|^{-\frac{1}{2}} \left|\boldsymbol{G}\right|^{-\frac{1}{2}} \exp\{-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})] - \frac{1}{2}[\boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}]\}$ (2.23)

Massimizzando la (2.23) si trovano gli stimatori ottimi di $\boldsymbol{\beta}$ ed $\boldsymbol{u}$, secondo il noto principio della massima verosimiglianza[25]. Questo equivale a massimizzare l'esponente della (2.23), cioè calcolare la log-verosimiglianza trascurando la parte costante, Henderson (1950):

$$\varphi(\boldsymbol{\beta}, \boldsymbol{u}) = -\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})] - \frac{1}{2}[\boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u}] \quad (2.24)$$

Derivando la (2.24) rispetto a $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ e ponendo uguali a zero le derivate si ottengono le "equazioni del modello ad effetti misti" (Henderson, 1950):

[24] Si ricorda che $E[\boldsymbol{ee}^T] = \sigma_e^2 \boldsymbol{R}$, con R nota e σ_e^2 incognito. Basta considerare σ_e^2 nella diagonale di $\boldsymbol{R}$ con $\boldsymbol{R}$ incognita, vedi nota 18.

[25] Vedere A. M. Mood e F. A. Graybill, Introduzione alla statistica, McGraw-Hill 1991, per un approfondimento sulla massima verosimiglianza.

$$\begin{cases} \dfrac{\partial \varphi(\boldsymbol{\beta},\boldsymbol{u})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{u} - \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} = 0 \\ \dfrac{\partial \varphi(\boldsymbol{\beta},\boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\boldsymbol{u} - \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} = 0 \end{cases} \tag{2.25}$$

La soluzione della (2.25) sono le stime ottime lineari e corrette di $\boldsymbol{\beta}$ e $\boldsymbol{u}$:

$$\begin{aligned} \boldsymbol{\beta}^* = &[\boldsymbol{X}^T(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{X}]^{-1}\boldsymbol{X}^T \cdot \\ &\cdot(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1})\boldsymbol{y} \end{aligned} \tag{2.26}$$

$$\boldsymbol{u}^* = [(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}](\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*) \tag{2.27}$$

Ponendo:

$$(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}) = \boldsymbol{V}^{-1}$$

$$(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} = \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}$$

le stime $\boldsymbol{\beta}^*$ e $\boldsymbol{u}^*$ corrispondono alle stime $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{u}}$. $\boldsymbol{\beta}^*$ e $\boldsymbol{u}^*$ sono definite "soluzioni del modello ad effetti misti" e sono più semplici da calcolare rispetto a $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{u}}$, poiché nella (2.26) e (2.27) non è richiesta l'inversione della matrice $\boldsymbol{V}$.
Per completezza, notiamo che $\varphi(\boldsymbol{\beta},\boldsymbol{u})$ non è una vera funzione di verosimiglianza poiché $\boldsymbol{u}$ non è osservabile (Robinson, 1991), essa viene definita funzione di verosimiglianza "penalizzata".

Il processo di stima BLUP sinora analizzato prevede che si conoscono $\boldsymbol{G}$ ed $\boldsymbol{R}$, quindi $\boldsymbol{V}$, che sono in realtà incogniti. Supponendo che $\boldsymbol{V}$ dipenda da un vettore di informazioni $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_h]^T$ noto, la stima BLUP è funzione delle osservazioni $\boldsymbol{y}$ e del vettore $\boldsymbol{\delta}$:

$$t(\boldsymbol{\delta},\boldsymbol{y}) = \hat{\mu} = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{u}} \tag{2.27bis}$$

Allo stimatore ottenuto dobbiamo associarvi una misura di variabilità, il MSE. Vediamo come ottenere il MSE(t($\boldsymbol{\delta,y}$)) (vedere approfondimento appendice A, pag. 141):

1. Si riscrive il modello in modo da scomporlo in due componenti:

$$t(\boldsymbol{\delta}, \boldsymbol{y}) = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{u}} = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

   Dove $\boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ è la prima componente che può essere interpretata come lo stimatore BLUP reso noto $\boldsymbol{\beta}$: $t^*(\boldsymbol{\delta,y,\beta})$ e $(\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ è la seconda parte, che possiamo interpretare come una componente aggiuntiva per conoscere $\boldsymbol{\beta}$. Riassumendo:

$$t(\boldsymbol{\delta,y}) = t^*(\boldsymbol{\delta,y,\beta}) + (\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad (2.28)$$

2. Si calcola il MSE($t(\boldsymbol{\delta,y})$) rispetto alla scomposizione fatta sopra:

$$MSE(t(\boldsymbol{\delta}, \boldsymbol{y})) = MSE[t(\boldsymbol{\delta}, \boldsymbol{y}, \beta)] + V[(\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta})$$

   Per semplificare la presentazione delle formule si considera:

   - $\boldsymbol{b}^T = \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}$
   - $\boldsymbol{d}^T = \boldsymbol{l}^T - \boldsymbol{b}^T\boldsymbol{X} = \boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X}$

   Con $\boldsymbol{b}^T$ e $\boldsymbol{d}^T$ costanti rispetto al valore atteso.

3. Si calcolano le due componenti:

$$\text{MSE}[t^*(\boldsymbol{\delta,y,\beta})] = g_1(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G})\boldsymbol{m}$$

$$V[(\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = g_2(\boldsymbol{\delta}) = \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$$

Abbiamo visto che

$$MSE(t(\boldsymbol{\delta}, \boldsymbol{y})) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G})\boldsymbol{m} + \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$$

Dove $g_1(\boldsymbol{\delta})$ è la componente dovuta agli effetti casuali di area e $g_2(\boldsymbol{\delta})$ è dovuta alla stima dei coefficienti di regressione $\boldsymbol{\beta}$.

Lo stimatore BLUP si può utilizzare solo se sono note le matrici varianza-covarianza $\boldsymbol{R}$ (nella sua componente diagonale σ_{ei}^2) e $\boldsymbol{G}$. Questo ne limita l'utilizzo solo a casi estremamente particolari. Per questo motivo è stata introdotta, come accennato all'inizio del paragrafo 2.3, la stima di tali componenti. Se si considerano le componenti di varianza-covarianza stimate si parla di EBLUP e non più di BLUP.

2.4 Empirical Best Linear Unbiased Predictor

EBLUP è l'acronimo di Empirical Best Linear Unbiased Predictor, predittore empirico ottimo, lineare e corretto.

Considerando di poter stimare le componenti di varianza σ_{ei}^2 e $\sigma_{u_i}^2 (i = 1, ..., m)$ ed ottenere una stima delle matrici $\boldsymbol{G}$ e $\boldsymbol{V}$, il predittore empirico ottimo, lineare e corretto assume la forma:

$$t(\hat{\boldsymbol{\delta}}, \boldsymbol{y}) = l^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{G}}\boldsymbol{Z}^T\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Dove $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$. Kackar e Harville (1981) dimostrano che $t(\hat{\boldsymbol{\delta}}, \boldsymbol{y})$ è uno stimatore corretto di μ se $\hat{\boldsymbol{\delta}}$ è uno stimatore invariante rispetto alle traslazioni[26]. Sempre Kackar e Harville (1981) dimostrano che lo stimatore di massima verosimiglianza è uno stimatore invariante rispetto alle traslazioni, e quindi è in grado di fornire stime corrette di μ[27].

Con questi presupposti non si considera più $\boldsymbol{\delta}$ noto e si cerca di ottenerne una stima. Nei sottoparagrafi successivi presenteremo un metodo di stima generale, mentre successivamente tale metodo sarà rielaborato per essere applicabile al caso pratico della stima per piccole aree.

[26] Questo concetto non viene trattato in modo esaustivo in questo contesto. Per un approfondimento vedere Kackar e Harville (1981) oppure Searle, McCulloch e Casella (1992) per quanto riguarda l'invarianza rispetto alle traslazioni dello stimatore proposto da Henderson.

[27] Per completezza esiste anche un altro stimatore corretto di μ, lo stimatore di massima verosimiglianza ristretta; tale stimatore non viene trattato in questa tesi. Per una panoramica sull'utilizzo della stima di massima verosimiglianza ristretta nell'ambito della stima per piccole aree vedere Salvati (2003).

Con il metodo di stime della massima verosimiglianza possiamo stimare le generiche componenti del vettore $\boldsymbol{\delta}$ e il vettore dei coefficienti di regressione $\boldsymbol{\beta}$.

2.4.1 La Stima di Massima Verosimiglianza delle Componenti di Varianza

L'approccio di stima di massima verosimiglianza richiede di specificare il tipo di distribuzione della variabile casuale del modello ad effetti misti, $\boldsymbol{y}$. Generalmente si assume che $\boldsymbol{y}$ sia distribuito normalmente. Fatta questa assunzione possiamo scrivere la funzione di verosimiglianza:

$$\text{si assume } \boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ}^T + \boldsymbol{R})$$

Date n osservazioni di $\boldsymbol{y}$ la funzione di verosimiglianza è:

$$l(\boldsymbol{\beta}, \boldsymbol{\delta}) = 2\pi^{-\frac{n}{2}} \left| \boldsymbol{ZGZ}^T + \boldsymbol{R} \right|^{-\frac{1}{2}} \exp\{-\frac{1}{2}[(\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{ZGZ}^T + \boldsymbol{R})^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})]\}$$

E la funzione di log-verosimiglianza risulta:

$$ll(\boldsymbol{\beta}, \boldsymbol{\delta}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|ZGZ^T + R\right| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{ZGZ} + \boldsymbol{R})^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$$

Da questa funzione bisogna stimare $\boldsymbol{\beta}$ e $\boldsymbol{\delta}$. Derivando rispetto al vettore $\boldsymbol{\beta}$ e rispetto ai singoli elementi del vettore $\boldsymbol{\delta}$, cioè $\delta_1, \ldots, \delta_h$, otteniamo due funzioni "score" che uguagliate a zero forniscono le stime desiderate (si ricorda che $\boldsymbol{ZGZ}^T + \boldsymbol{R} = \boldsymbol{V}$):

$$\frac{\partial ll(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$$

$$\frac{\partial ll(\boldsymbol{\beta}, \boldsymbol{\delta})}{\partial \delta_s} = -\frac{1}{2} tr(\boldsymbol{V}^{-1} \boldsymbol{V}_{(s)}) - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\beta})^T (-\boldsymbol{V}^{-1} \boldsymbol{V}_{(s)} \boldsymbol{V}^{-1}) (\boldsymbol{y} - \boldsymbol{X\beta})$$

Dove $\boldsymbol{V}_{(s)} = \frac{\partial \boldsymbol{V}}{\partial \delta_S}$ e $(-\boldsymbol{V}^{-1} \boldsymbol{V}_{(s)} \boldsymbol{V}^{-1}) = \frac{\partial \boldsymbol{V}^{-1}}{\partial \delta_s}$ (Rao, 2003).

La stima di $\boldsymbol{\beta}$ abbiamo già visto precedentemente come può essere ottenuta (GLS e massima verosimiglianza), adesso ci concentriamo sulla stima della matrice di varianza-

covarianza $\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$. Stimiamo il valore atteso della matrice $\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$, in questo modo sulla diagonale principale della matrice avremo i valori attesi rispettivamente della varianza di $\hat{\boldsymbol{\beta}}$ e delle componenti $\hat{\boldsymbol{\delta}}$ (cioè la stima delle componenti di varianza). Il valore atteso di $\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$ in letteratura prende il nome di matrice di informazione. Tramite il valore atteso delle derivate parziali seconde di $ll(\boldsymbol{\beta},\boldsymbol{\delta})$ si ottiene l'inversa della matrice di informazione (Rao, 2003):

$$\boldsymbol{\imath} = E[\,\partial^2 ll(\boldsymbol{\beta},\boldsymbol{\delta})\,]$$

dove per il generico elemento *s,k* la derivata parziale seconda è:

$$\boldsymbol{\imath}_{s,k}(\boldsymbol{\delta}) = \frac{1}{2} tr(\boldsymbol{V}^{-1}\boldsymbol{V}_{(s)}\boldsymbol{V}^{-1}\boldsymbol{V}_{(k)}) \quad (2.88)$$

Per ottenere la stima di massima verosimiglianza bisogna utilizzare l'algoritmo di *scoring*[28], un processo di tipo iterativo:

$$\boldsymbol{\delta}^{(r+1)} = \boldsymbol{\delta}^{(r)} + [\boldsymbol{\imath}(\boldsymbol{\delta}^{(r)})]^{-1} d[\hat{\boldsymbol{\beta}}(\boldsymbol{\delta}^{(r)}), \boldsymbol{\delta}^{(r)}]$$

Dove *r* indica il numero di iterazioni. L'algoritmo si ferma quando la differenza tra $\boldsymbol{\delta}^{(r+1)}$ e $\boldsymbol{\delta}^{(r)}$ è minore di un certo valore ε scelto a piacere (per esempio 0,01). Ovviamente la matrice di informazione è $\boldsymbol{\imath}^{-1}$. Quando l'algoritmo converge si ottengono le stime di massima verosimiglianza di $\boldsymbol{\beta}$ e $\boldsymbol{\delta}$, che indichiamo con $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\delta}}$. Otteniamo soprattutto la matrice di varianza-covarianza di $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\delta}}$ che ha una struttura diagonale a blocchi:

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}}) = \begin{bmatrix} (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} & 0 \\ 0 & \boldsymbol{\imath}^{-1}(\boldsymbol{\delta}) \end{bmatrix}$$

[28] Vedere K. Wolter, Introduction to variance estimation, Sprinter-Verlag, New York, per un approfondimento.

Dove $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$ rappresenta la matrice di varianza-covarianza relativa al vettore delle stime dei coefficienti di regressione ($\hat{\boldsymbol{\beta}}$) e $\boldsymbol{\iota}^{-1}(\boldsymbol{\delta})$, che è la matrice di informazione, rappresenta la matrice della stima delle componenti di varianza $\hat{\boldsymbol{\delta}}$. Entrambe le matrici sono diagonali. Per esempio considerando il vettore $\hat{\boldsymbol{\beta}}$ di dimensione *p* ed il vettore $\hat{\boldsymbol{\delta}}$ di dimensione *h* la matrice assume la seguente forma:

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}}) = \begin{bmatrix} \hat{\sigma}^2_{\hat{\beta}_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}^2_{\hat{\beta}_2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}^2_{\hat{\beta}_p} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\delta}_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\delta}_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{\delta}_h \end{bmatrix}$$

Utilizzando le componenti di questa matrice possiamo fare un test di verifica di ipotesi sui coefficienti di regressione. E' stato dimostrato in letteratura che $\hat{\beta}_i$ (*i = 1, ..., p*) si distribuisce come una *t di Student* con *n-p* gradi di libertà, dato un campione di *n* osservazioni. Il classico test *t* sottopone a verifica l'ipotesi nulla $\hat{\beta}_i = 0$, il punto critico t^* si ottiene da $(\hat{\beta}_i - 0)/(\sqrt{\hat{\sigma}^2_{\beta_i}/(n-p)})$. Per valori teorici della statistica *t* di Student minori di t^* si rifiuta l'ipotesi nulla in favore dell'ipotesi alternativa $\hat{\beta}_i \neq 0$ (il $\hat{\beta}_i$ è significativo).

2.4.2 La Stima del Mean Squared Error

Come per il BLUP anche per l'EBLUP vogliamo una misura della variabilità. Partendo dal MSE dello stimare BLUP e dalla stima delle componenti di varianza è possibile ottenere una stima del MSE dello stimatore EBLUP.

Il MSE del BLUP è in funzione del vettore delle componenti di varianza $\boldsymbol{\delta}$: $MSE(t(\boldsymbol{\delta},\boldsymbol{y})) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G})\boldsymbol{m} + \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$ (d'ora in

avanti per semplicità t($\boldsymbol{\delta}$,$\boldsymbol{y}$) = t($\boldsymbol{\delta}$)). In prima approssimazione basta sostituire alla componente incognita $\boldsymbol{\delta}$ la sua stima $\hat{\boldsymbol{\delta}}$:

$$MSE(t(\hat{\boldsymbol{\delta}})) = g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) = \boldsymbol{m}^T(\hat{\boldsymbol{G}} - \hat{\boldsymbol{G}}\boldsymbol{Z}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Z}\hat{\boldsymbol{G}})\boldsymbol{m} + \boldsymbol{d}^T(\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d} \qquad (2.29)$$

La (2.29) non è una stima corretta del MSE($t(\hat{\boldsymbol{\delta}})$) poiché si dimostra che $E[g_1(\hat{\boldsymbol{\delta}})] \neq g_1(\hat{\boldsymbol{\delta}})$ (Rao, 2003). In questo caso il MSE (che non è più uguale alla varianza) assume la seguente forma:

$$MSE(t(\hat{\boldsymbol{\delta}})) = V(t(\hat{\boldsymbol{\delta}})) + bias(t(\hat{\boldsymbol{\delta}}))^2 = MSE(t(\boldsymbol{\delta})) + E[t(\hat{\boldsymbol{\delta}}) - t(\boldsymbol{\delta})]^2 \qquad (2.30)$$

Il bias (la distorsione) non può essere calcolato in termini generici. Kackar e Harville hanno proposto di usare l'approssimazione di Taylor:

$$t(\hat{\boldsymbol{\delta}}) - t(\hat{\boldsymbol{\delta}}) \approx \boldsymbol{d}(\boldsymbol{\delta})^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$$

Con $\boldsymbol{d}(\boldsymbol{\delta}) = \partial t(\boldsymbol{\delta}) / \partial \boldsymbol{\delta}$ dove per alti valori di ($\hat{\boldsymbol{\delta}}$ - $\boldsymbol{\delta}$) il bias $\boldsymbol{d}(\boldsymbol{\delta})^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$ assume valori bassi. Supponendo la normalità delle osservazioni (e quindi di $\boldsymbol{e}$ ed $\boldsymbol{u}$):

$$\boldsymbol{d}(\hat{\boldsymbol{\delta}}) \approx \frac{\partial t^*(\boldsymbol{\delta}, \boldsymbol{\beta})}{\partial \boldsymbol{\delta}} = \left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}}\right)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{d}^*(\boldsymbol{\delta})$$

Dove $t^*(\boldsymbol{\delta}, \boldsymbol{\beta}) = t(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{y}) = \boldsymbol{l}^T\boldsymbol{\beta} + \boldsymbol{b}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ (vedi 2.28), $\boldsymbol{b}$ e $\boldsymbol{d}$ hanno lo stesso valore specificato nei paragrafi precedenti. Quindi una stima approssimativa della distorsione, supponendo la normalità, risulta essere:

$$E[t(\hat{\boldsymbol{\delta}}) - t(\boldsymbol{\delta})]^2 \approx E[\boldsymbol{d}(\boldsymbol{\delta})^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})]^2 \approx E[\boldsymbol{d}^*(\boldsymbol{\delta})^T(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})]^2 \qquad (2.31)$$

Sempre secondo Kackar e Harville la (2.31) si può approssimare con:

$$E[\boldsymbol{d}^*(\boldsymbol{\delta})^T(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})]^2 \approx tr[E[\boldsymbol{d}^*(\boldsymbol{\delta})\boldsymbol{d}^*(\boldsymbol{\delta})^T]\boldsymbol{V}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})] = tr[\left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}}\right)\boldsymbol{V}\left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}}\right)^T \boldsymbol{\iota}^{-1}(\boldsymbol{\delta})] = g_3(\boldsymbol{\delta}) \tag{2.32}$$

La (2.32) dipende dal vettore delle componenti di varianza $\boldsymbol{\delta}$ e rappresenta la distorsione del MSE dello stimatore EBLUP, per semplificare la scrittura considerando la (2.31) e (2.32) scriviamo:

$$bias(t(\hat{\boldsymbol{\delta}}))^2 = E[t(\hat{\boldsymbol{\delta}}) - t(\boldsymbol{\delta})]^2 \approx g_3(\boldsymbol{\delta}) \tag{2.33}$$

Dalla (2.28), (2.30) e (2.33) segue che:

$$MSE[t(\hat{\boldsymbol{\delta}})] \approx g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) + g_3(\boldsymbol{\delta})$$

Quindi la componente $g_3(\boldsymbol{\delta})$ è dovuta alla stima delle componenti di varianza.

Se sostituiamo ai termini $g_i(\boldsymbol{\delta})$ i corrispettivi $g_i(\hat{\boldsymbol{\delta}})$ utilizzando la stima delle componenti di varianza otteniamo uno stimatore per il $MSE[t(\hat{\boldsymbol{\delta}})]$:

$$\widehat{MSE}[t(\hat{\boldsymbol{\delta}})] \approx g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + g_3(\hat{\boldsymbol{\delta}})$$

Nonostante l'introduzione della componente $g_3(\hat{\boldsymbol{\delta}})$, la stima del MSE risulta ancora distorta poiché la stima del bias che abbiamo introdotto non influenza il MSE(t($\boldsymbol{\delta}$)), che non è corretto a causa della componente $g_1(\hat{\boldsymbol{\delta}})$ (come specificato a inizio paragrafo). Rao (2003) suggerisce una stima corretta di $g_1(\hat{\boldsymbol{\delta}})$ utilizzando l'espansione in serie di Taylor:

$$g_1(\hat{\boldsymbol{\delta}}) = g_1(\boldsymbol{\delta}) + (\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^T \nabla g_1(\boldsymbol{\delta}) + \frac{1}{2}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^T \nabla^2 g_1(\boldsymbol{\delta})(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}) + r \approx g_1(\boldsymbol{\delta}) + \Delta_1 + \Delta_2$$

Con $\nabla g_1(\boldsymbol{\delta}) = \left[\frac{\partial g_1(\boldsymbol{\delta})}{\partial \delta_1}, \cdots, \frac{\partial g_1(\boldsymbol{\delta})}{\partial \delta_h} \right]^T$ il vettore delle derivate parziali prime di $g_1(\boldsymbol{\delta})$ rispetto a δ, $\nabla^2 g_1(\hat{\boldsymbol{\delta}}) = \begin{bmatrix} \frac{\partial^2 g_1(\boldsymbol{\delta})}{\partial \delta_1^2} & \cdots & \frac{\partial^2 g_1(\boldsymbol{\delta})}{\partial \delta_1 \delta_h} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g_1(\boldsymbol{\delta})}{\partial \delta_h \delta_1} & \cdots & \frac{\partial^2 g_1(\boldsymbol{\delta})}{\partial \delta_h^2} \end{bmatrix}$ la matrice delle derivate parziali seconde rispetto a δ e r il resto del polinomio di Taylor (che per le sue proprietà può essere ignorato). Se supponiamo che $\hat{\boldsymbol{\delta}}$ è una stima corretta di $\boldsymbol{\delta}$, allora $E[\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}] = 0$ e il valore atteso della componente $g_1(\hat{\boldsymbol{\delta}})$ è secondo Rao (2003):

$$E[g_1(\hat{\boldsymbol{\delta}})] = g_1(\boldsymbol{\delta}) - \boldsymbol{b}_{\hat{\delta}}^T(\hat{\boldsymbol{\delta}}) \nabla g_1(\hat{\boldsymbol{\delta}}) + \frac{1}{2} tr[\nabla^2 g_1(\boldsymbol{\delta}) \boldsymbol{\iota}^{-1}(\boldsymbol{\delta})] \tag{2.34}$$

Dove $\boldsymbol{b}_{\hat{\delta}}^T(\boldsymbol{\delta}) = \frac{1}{2m} \iota^{-1}(\boldsymbol{\delta})_{col_{1 \le j \le m}} tr[\sum_i (X_i^T V_i^{-1} X_i)^{-1} \sum_i (X_i^T (-V_i^{-1} V_{i_{(j)}} V_i^{-1}) X_i)^{-1}]$. Prasad e Rao (1990) affermano che se la struttura della matrice $\boldsymbol{V}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}})$ è diagonale, come abbiamo ipotizzato, allora è possibile riscrivere la seguente uguaglianza:

$$\frac{1}{2} tr[\nabla^2 g_1(\boldsymbol{\delta}) \boldsymbol{\iota}^{-1}(\boldsymbol{\delta})] = -tr[\left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}} \right) \boldsymbol{V} \left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}} \right)^T \boldsymbol{\iota}^{-1}(\boldsymbol{\delta})] = -g_3(\boldsymbol{\delta}) \tag{2.35}$$

Grazie alla (2.34) e (2.35) possiamo calcolare il valore atteso di $g_1(\hat{\boldsymbol{\delta}})$:

$$E[g_1(\hat{\boldsymbol{\delta}})] = g_1(\boldsymbol{\delta}) + \boldsymbol{b}_{\hat{\delta}}^T(\hat{\boldsymbol{\delta}}) \nabla g_1(\hat{\boldsymbol{\delta}}) - g_3(\boldsymbol{\delta}) \tag{2.36}$$

Dalla (2.36), considerando che $E[g_2(\hat{\boldsymbol{\delta}})] = g_2(\boldsymbol{\delta})$ e $E[g_3(\hat{\boldsymbol{\delta}})] = g_3(\boldsymbol{\delta})$ (Rao, 2003), otteniamo uno stimatore approssimativamente corretto del MSE dello stimatore EBLUP:

$$\widehat{MSE}[t(\hat{\boldsymbol{\delta}})] \approx g_1(\hat{\boldsymbol{\delta}}) - \boldsymbol{b}_{\hat{\boldsymbol{\delta}}}^T(\hat{\boldsymbol{\delta}})\nabla g_1(\hat{\boldsymbol{\delta}}) + g_2(\hat{\boldsymbol{\delta}}) + 2g_3(\hat{\boldsymbol{\delta}})$$ [29]

Dove l'approssimazione dipende dal polinomio di Taylor. Generalmente le componenti $g_2(\hat{\boldsymbol{\delta}})$ e $g_3(\hat{\boldsymbol{\delta}})$ sono molto piccole rispetto alla componente $g_1(\hat{\boldsymbol{\delta}})$, che di fatto governa la variabilità della stima EBLUP.

2.4.3 EBLUP Nella Stima per Piccole Aree nel Modello a Livello di Area

In questo paragrafo si applicano le derivazioni sinora ottenute alla stima per piccole aree nel modello a livello di area, derivando EBLUP e stima del MSE. Si ricorda il modello a livello di area descritto nel paragrafo 2.2.2.1:

$$\hat{\theta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + z_i u_i + e_i$$

Dove $\boldsymbol{x_i}^T = [x_{i1}, \ldots, x_{ip}]^T$ è il vettore delle covariate dell'area i, $\boldsymbol{\beta}$ è il vettore dei coefficienti di regressione (inosservabile), z_i è una costante positiva nota, u_i è l'effetto casuale dell'area *i* che ipotizziamo essere indipendentemente e identicamente distribuito con media 0 e varianza σ_u^2 (u_i~i.i.d.(0, σ_u^2)) ed e_i è l'errore di campionamento che, secondo le ipotesi classiche, è indipendentemente distribuito con media 0 e varianza σ_{ei}^2 (e_i~ind(0, σ_{ei}^2)). Si ipotizza anche che u_i ed e_i sono indipendenti.
Posto $\boldsymbol{x_i}^T = \boldsymbol{l}^T$ e $m_i = z_i$ la stima di μ presentata nella (2.27bis) riferita alla piccola area *i* diventa:

$$\hat{\mu} = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{u}} \rightarrow \hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + z_i\hat{u}_i \qquad (2.37)$$

Secondo le ipotesi fatte, le matrici di varianza-covarianza del modello lineare ad effetti misti $\boldsymbol{G}$ ed $\boldsymbol{R}$ risultano diagonali e pari a:

[29] Dalla (2.36) segue che:

$$E[\widehat{MSE}[t(\hat{\boldsymbol{\delta}})]] = E[g_1(\hat{\boldsymbol{\delta}})] - E[\boldsymbol{b}_{\hat{\boldsymbol{\delta}}}^T(\hat{\boldsymbol{\delta}})\nabla g_1(\hat{\boldsymbol{\delta}})] + E[g_2(\hat{\boldsymbol{\delta}})] + 2E[g_3(\hat{\boldsymbol{\delta}})] =$$

$$= g_1(\boldsymbol{\delta}) + \boldsymbol{b}_{\hat{\boldsymbol{\delta}}}^T(\hat{\boldsymbol{\delta}})\nabla g_1(\hat{\boldsymbol{\delta}}) - g_3(\boldsymbol{\delta}) - \boldsymbol{b}_{\hat{\boldsymbol{\delta}}}^T(\hat{\boldsymbol{\delta}})\nabla g_1(\hat{\boldsymbol{\delta}}) + g_2(\boldsymbol{\delta}) + 2g_3(\boldsymbol{\delta}) = MSE[t(\hat{\boldsymbol{\delta}})]$$

$$\boldsymbol{G} = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \sigma_u^2 & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_u^2 \end{bmatrix}$$

e

$$\boldsymbol{R} = \begin{bmatrix} \sigma_{e_1}^2 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \sigma_{e_i}^2 & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_{e_m}^2 \end{bmatrix}$$

Da cui risulta $\boldsymbol{V}=\boldsymbol{ZGZ}^T+\boldsymbol{R}$:

$$\boldsymbol{V} = \begin{bmatrix} z_1^2\sigma_u^2 + \sigma_{e_1}^2 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \cdots & \cdots & \vdots \\ \vdots & \cdots & z_i^2\sigma_u^2 + \sigma_{e_i}^2 & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & z_m^2\sigma_u^2 + \sigma_{e_m}^2 \end{bmatrix}$$ [30]

Considerando che la stima di $\boldsymbol{u}$ è $\hat{\boldsymbol{u}} = \boldsymbol{GZ}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ possiamo dedurre la stima specifica di u_i e riscrivere la componente $\boldsymbol{m}^T\hat{\boldsymbol{u}}$ che, considerando la piccola area i, può essere scritta come $m_i\hat{u}_i = m_i\sigma_u^2 z_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$. Sostituendo $\boldsymbol{x_i}^T = \boldsymbol{l}^{\mathrm{T}}$ e $m_i = z_i$, come suggerito per la (2.106), e y_i con $\hat{\theta}_i$ otteniamo $z_i\hat{u}_i = z_i\sigma_u^2 z_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1}(\hat{\theta}_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) = \frac{z_i^2\sigma_u^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}(\hat{\theta}_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$. Posto $\frac{z_i^2\sigma_u^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2} = \gamma_i$, considerata la stima BLUP $t(\boldsymbol{\delta},\boldsymbol{y}) = \hat{\mu} = \boldsymbol{l}^T\hat{\boldsymbol{\beta}} + \boldsymbol{m}^T\hat{\boldsymbol{u}}$ e considerata la (2.37) possiamo scrivere la stima BLUP per la piccola area i:

[30] Considerato $\boldsymbol{Z}$ una matrice diagonale di dimensione m con il generico elementi $Z_{ii} = z_i$ dal prodotto $\boldsymbol{ZGZ}^T = \boldsymbol{V}$ risulta che il generico elemento $V_{ii} = Z_{i.}G_{.i}Z_{.i} = [0, \ldots, z_i, 0, \ldots,0]\mathrm{x}[0, \ldots, \sigma_u^2, 0, \ldots,0]^T\mathrm{x}[0, \ldots, z_i, 0, \ldots,0]^T = z_i\,\sigma_u^2 z_i = z_i^2\,\sigma_u^2$.

$$t_i(\sigma_u^2, \hat{\theta}_i) = \hat{\mu} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} + \boldsymbol{m}^T \hat{\boldsymbol{u}} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) \tag{2.38}$$

Rielaborando la (2.38) si verifica che la stima BLUP per piccole aree è uno stimatore combinato. Infatti:

$$t_i(\sigma_u^2, \hat{\theta}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$$

Dove γ_i è il peso dello stimatore combinato e $\hat{\boldsymbol{\beta}}$ è il vettore delle stime BLUE dei coefficienti di regressione. Utilizzando la stima GLS per ottenere $\hat{\boldsymbol{\beta}}$ considerando le informazioni riguardanti la piccola area *i* otteniamo che:

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \hat{\theta}_i}{z_i^2 \sigma_u^2 + \sigma_{e_i}^2}}{\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{z_i^2 \sigma_u^2 + \sigma_{e_i}^2}} \tag{2.39}$$

Infatti dato $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}$ se consideriamo invece dell'intera matrice $\boldsymbol{X}$ solo il vettore $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ e al posto del vettore $\boldsymbol{y}$ utilizziamo $\hat{\theta}$, in modo che l'osservazione $y_i = \hat{\theta}_i$, otteniamo $\hat{\boldsymbol{\beta}} = \sum_{i\_=1}^{m} (\boldsymbol{x}_i \boldsymbol{V}_{.i}^{-1} \boldsymbol{x}_i^T)^{-1} \boldsymbol{x}_i \boldsymbol{V}_{.i}^{-1} \hat{\theta}_i$ con $\boldsymbol{V}_{.i}$ uguale alla *i*-esima colonna della matrice $\boldsymbol{V}$. Con queste notazioni $\boldsymbol{X}^T\boldsymbol{X}$ può anche essere scritto come $\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T$ [31]. Svolgendo i prodotti si ottiene la (2.39).

Il peso della stima combinata tende a 1 quando l'errore campionario $\sigma_{e_i}^2$ tende a 0, in questo caso lo stimatore BLUP coincide con la stima diretta: $t_i(\sigma_u^2, \hat{\theta}_i) = \gamma_i \hat{\theta}_i + (1-\gamma_i)\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} = \gamma_i \hat{\theta}_i$. Grazie a questa proprietà si dice che lo stimatore BLUP è consistente rispetto al disegno.

[31] $X^TX =$

$$[x_{11},\ldots,x_{1m}]*\begin{bmatrix} x_{11} \\ \vdots \\ x_{m1} \end{bmatrix} + \ldots + [x_{i1},\ldots,x_{im}]*\begin{bmatrix} x_{1i} \\ \vdots \\ x_{mi} \end{bmatrix} + \ldots + [x_{p1},\ldots,x_{pm}]*\begin{bmatrix} x_{1p} \\ \vdots \\ x_{mp} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{m} x_{1k}^2 & \cdots & \sum_{k=1}^{m} x_{1k}x_{kp} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{m} x_{pk}x_{k1} & \cdots & \sum_{k=1}^{m} x_{pk}^2 \end{bmatrix} = \sum_{k=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

Il MSE[$t_i(\sigma_u^2, \hat{\theta}_i)$] lo otteniamo ponendo $\boldsymbol{\delta}$ = [σ_u^2] e utilizzando le derivazioni già ottenute. Da queste risulta:

$$MSE[t_i(\sigma_u^2, \hat{\theta}_i)] = g_{1_i}(\sigma_u^2) + g_{2_i}(\sigma_u^2)$$

Dalla definizione generica di $g_1(\boldsymbol{\delta})$ e $g_2(\boldsymbol{\delta})$ si passa a una definizione applicate alla stima per piccole aree con modello a livello di area, $g_{1_i}(\sigma_u^2)$ e $g_{2_i}(\sigma_u^2)$ (app. A pag. 143):

$$g_1(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{GZV}^{-1}\boldsymbol{Z}^T\boldsymbol{G})\boldsymbol{m} \rightarrow g_{1_i}(\sigma_u^2) = \frac{z_i^2\sigma_u^2\sigma_{e_i}^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2} = \gamma_i\sigma_{e_i}^2$$

$$g_2(\boldsymbol{\delta}) = \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d} \rightarrow = (1-\gamma_i)^2\boldsymbol{x}_i^T\left(\frac{\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}\right)^{-1}\boldsymbol{x}_i \qquad (2.41)$$

Considerando le sostituzioni $\boldsymbol{m}^T \rightarrow z_i, \boldsymbol{Z} \rightarrow z_i, \boldsymbol{G} \rightarrow \sigma_u^2, \boldsymbol{V} \rightarrow z_i^2\sigma_u^2 + \sigma_{e_i}^2$ che risultano dal processo di selezione (cioè non considero più un modello in cui stimo un vettore, ma un modello in cui si stima un solo elemento del vettore; utilizzo quindi solo un vettore di covriate e seleziono nelle matrici di varianza-covarianza solo i valori specifici dell'area d'interesse) e dalle ipotesi presentate a inizio paragrafo; si ricorda che $\boldsymbol{x_i}^T = \boldsymbol{l}^T$ e $y_i = \hat{\theta}_i$. Dalla matrice $\boldsymbol{X}$ ci interessa prendere, quindi, solo la riga delle covariate dell'area *i* (l'*i*-esima riga): $\boldsymbol{X} \rightarrow [x_{i1}, \ldots, x_{ip}] \rightarrow \boldsymbol{x}_i^T$.

Come per il caso generale g_1 è riferito alla variabilità dello stimatore mentre g_2 si riferisce alla variabilità dovuta alla stima di $\boldsymbol{\beta}$.

Lo stimatore BLUP presentato dipende dalla componente di varianza σ_u^2 che solo teoricamente può essere considerata nota. Nelle applicazioni reali è necessario stimare σ_u^2, in questo caso, come abbiamo precedentemente precisato, si parla di EBLUP:

$$t(\hat{\sigma}_u^2, \hat{\theta}_i) = \hat{\gamma}_i\hat{\theta}_i + (1-\hat{\gamma}_i)\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$$

Dove

$$\hat{\gamma}_i = \frac{z_i^2 \hat{\sigma}_u^2}{z_i^2 \hat{\sigma}_u^2 + \sigma_{e_i}^2}$$

e

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \hat{\theta}_i}{z_i^2 \hat{\sigma}_u^2 + \sigma_{e_i}^2}}{\sum_{i=1}^{m} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{z_i^2 \hat{\sigma}_u^2 + \sigma_{e_i}^2}}$$

Lo stimatore EBLUP conserva la proprietà di correttezza se u_i ed e_i sono distribuzioni simmetriche e se la stima $\hat{\sigma}_u^2$ è invariante rispetto alle traslazioni (Kackar e Harville 1981). Come precedentemente specificato (paragrafo 2.4) la stima di massima verosimiglianza rispetta le condizioni per la correttezza dello stimatore[32].

Utilizzando l'algoritmo di scoring otteniamo la stima $\hat{\sigma}_u^2$:

$$\sigma_u^{2^{(r+1)}} = \sigma_u^{2^{(r)}} + \left[\frac{1}{2} \sum_{i=1}^{m} \frac{z_i^4}{z_i^2 \sigma_u^{2^{(r)}} + \sigma_{e_i}^2} \right]^{-1} - \frac{1}{2} \sum_{i=1}^{m} \frac{z_i^2}{z_i^2 \sigma_u^{2^{(r)}} + \sigma_{e_i}^2} + \frac{1}{2} \sum_{i=1}^{m} z_i^2 \frac{(\hat{\theta}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(\hat{\sigma}_u^{2^{(r)}}))^2}{(z_i^2 \sigma_u^{2^{(r)}} + \sigma_{e_i}^2)^2} \tag{2.42}$$

Dove $\hat{\boldsymbol{\beta}}(\sigma_u^{2^{(r)}})$ è la stima di $\boldsymbol{\beta}$ calcolata utilizzando il valore $\sigma_u^{2^{(r)}}$ (che è la stima di σ_u^2 alla r-esma iterazione dell'algoritmo). La (2.42) si ottiene applicando l'algoritmo di scoring ai parametri presentati in questo paragrafo.

Il MSE[t($\hat{\sigma}_u^2, \hat{\theta}_i$)] in base alla definizione del MSE[t($\hat{\boldsymbol{\delta}}$)] è:

$$MSE[t(\hat{\sigma}_u^2, \hat{\theta}_i)] = g_{1_i}(\sigma_u^2) + g_{2_i}(\sigma_u^2) + g_{3_i}(\sigma_u^2)$$

Dove le prime due componenti sono già state introdotte mentre la terza componente è ricavata dalla definizione introdotta per il caso generico:

[32] Rispettano tali condizioni anche il metodo generalizzato dei momenti (noto anche come GMM) e la stima di massima verosimiglianza ristretta.

$$g_3(\boldsymbol{\delta}) = tr\left[\left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}}\right)\boldsymbol{V}\left(\frac{\partial \boldsymbol{b}^T}{\partial \boldsymbol{\delta}}\right)^T \boldsymbol{\imath}^{-1}(\boldsymbol{\delta})\right] \rightarrow g_{3_i}(\sigma_u^2) = \frac{\sigma_{e_i}^4 z_i^4}{(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^3} 2\left[\sum_{i=1}^m \frac{z_i^4}{(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^2}\right]^{-1} \tag{2.43}$$

Dove $2\left[\sum_{i=1}^m \frac{z_i^4}{(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^2}\right]^{-1}$ corrisponde al blocco riferito a σ_u^2 nella matrice di informazione, $\imath^{-1}(\sigma_u^2)$. Utilizzando le stime $\hat{\sigma}_u^2$ siamo in grado di fornire una stima del MSE per la stima EBLUP riferita alla piccola area *i*:

$$\widehat{MSE}[t(\hat{\sigma}_u^2, \hat{\theta}_i)] = g_{1_i}(\hat{\sigma}_u^2) - b_{\hat{\sigma}_u^2}^T(\hat{\sigma}_u^2)\nabla g_{1_i}(\hat{\sigma}_u^2) + g_{2_i}(\hat{\sigma}_u^2) + 2g_{3_i}(\hat{\sigma}_u^2)$$

Dove $g_{1_i}(\hat{\sigma}_u^2)$, $g_{2_i}(\hat{\sigma}_u^2)$ e $g_{3_i}(\hat{\sigma}_u^2)$ si ottengono sostituendo nella (2.40), (2.41) e (2.43) σ_u^2 con $\hat{\sigma}_u^2$ (Rao 2003)[33].

2.4.4 EBLUP Nella Stima per Piccole Aree nel Modello a Livello di Unità

Il modello a livello di unità considera tutte le osservazioni all'interno dell'area di riferimento. Per applicarlo è quindi necessario conoscere le covariate su tutte le unità campionate. La stima per l'osservazione *j*-esima della piccola area *i* può assumere la forma:

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij} \text{ (vedi paragrafo 2.2.2.2)}$$

Considerando invece il predittore empirico ottimo lineare e corretto della media dell'area *i*, il modello proposto può essere riscritto come:

$$\overline{Y}_i = \overline{\boldsymbol{X}}_i^T\boldsymbol{\beta} + u_i + \overline{e}_i$$

[33] La componente di varianza $\sigma_{e_i}^2$ si considera nota.

Dove $\bar{\boldsymbol{X}}_i$ è la matrice della media delle covariate per piccola area ed $\bar{e}_i$ la media degli errori per la piccola area i ($\bar{e}_i = \sum_{j=1}^{N_i} e_{ij}$, con N_i uguale alla numerosità totale nell'area i), che si presume tendere a 0 al crescere di N_i.

Procedendo in maniera parallela a quanto fatto nella stima con modello a livello di area si può ricavare la stima di massima verosimiglianza dei parametri richiesti. Per questo motivo in questo paragrafo non saranno fornite derivazioni formali.
Consideriamo il modello da utilizzare con le unità campionate che consente di ottenere il predittore ottimo lineare e corretto:

$$\boldsymbol{y}_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \sum_{j=}^{n_i} u_{ij} + e_i$$

Prasad e Rao (1990) applicando le nozioni generali (analogamente a quanto dimostrato nel paragrafo precedente) derivano il BLUP per la stima per piccole area a livello di unità:

$$t(\sigma_u^2, \sigma_{ei}^2) = \gamma_i [\bar{y}_i + (\bar{\boldsymbol{X}}_i + \bar{\boldsymbol{x}}_i)^T \hat{\boldsymbol{\beta}}] + (1-\gamma_i) \boldsymbol{X}_i^T \hat{\boldsymbol{\beta}} \text{ [34]} \qquad (2.44)$$

Dove σ_u^2 *e* σ_{ei}^2 si considerano noti, $\gamma_i = \dfrac{\sigma_u^2}{\sigma_u^2 + \dfrac{\sigma_{ei}^2}{w_{i.}}}$ è il peso della stima combinata, noto anche come fattore di restringimento, con $w_{i.} = \sum_{j=1}^{n_i} w_{ij}$ (generico peso per l'area i). $\bar{y}_i$ e $\bar{x}_i$ sono medie aritmetiche ponderate con pesi w_{ij}. La stima di $\boldsymbol{\beta}$ è una stima BLUE ottenuta con gli OLS applicati su dati trasformati: $\left(\sqrt{w_{ij}}(y_{ij} - \bar{y}_{iw}), \sqrt{w_{ij}}(x_{ij} - \bar{x}_{iw})\right)$ (Parsad e Rao, 1990); il metodo consiste semplicemente nel derivare una serie di coppie di valori (y^*, $\boldsymbol{x}^*$) e calcolare su di esse i coefficienti di regressione, con una stima OLS. La trasformazione, dovuta a Stukel (1991), garantisce di ottenere stime BLUE utilizzando gli OLS.

Applicando le derivazioni generali ottenute alla (2.44) possiamo scrivere il MSE[$t(\sigma_u^2, \sigma_{ei}^2)$]:

[34] Vedere anche Salvati (2003)

$$MSE[t(\sigma_u^2, \sigma_{ei}^2)] = g_{1_i}(\sigma_u^2, \sigma_{ei}^2) + g_{2_i}(\sigma_u^2, \sigma_{ei}^2)$$

Dove

$$g_{1_i}(\sigma_u^2, \sigma_{ei}^2) = \gamma_i \frac{\sigma_{ei}^2}{w_{i.}} \tag{2.45}$$

e

$$g_{2_i}(\sigma_u^2, \sigma_{ei}^2) = (\bar{\boldsymbol{x}}_i - \hat{\gamma}_i \boldsymbol{x}_{iw})^T \sum_i \frac{1}{\sigma_{ei}^2} (\sum_{j=1}^{n_i} w_{ij} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \hat{\gamma}_i \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T)(\bar{\boldsymbol{x}}_i - \hat{\gamma}_i \boldsymbol{x}_{iw}) \tag{2.46}$$

Non conoscendo nei casi pratici i valori σ_u^2 e σ_{ei}^2 bisogna stimarli. In questo caso si fa riferimento all'EBLUP.
Si usa un processo di stima articolato in due passi:

1. Si applica un modello di regressione alle coppie $\left(\sqrt{w_{ij}}(y_{ij} - \bar{y}_{iw}), \sqrt{w_{ij}}(x_{ij} - \bar{x}_{iw})\right)$ se l'area è campionata. Ottenuto il $\hat{\boldsymbol{\beta}}$ con gli OLS si calcolano i residui $\hat{\boldsymbol{e}}$. Con questi si fa una stima della varianza internaalle aree σ_e^2:

$$\hat{\sigma}_e^2 = \frac{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}}}{\sum_{i=1}^{m} (n_i - 1) - p}$$

2. Si applica un modello di regressione alle coppie $\left(\sqrt{w_{ij}} y_{ij}, \sqrt{w_{ij}} \boldsymbol{x}_{ij}\right)$ e si calcolano i residui $\hat{\boldsymbol{e}}_u$. Si calcola quindi un valore stimato di σ_u^2 come:

$$\tilde{\sigma}_u^2 = \sum_{i=1}^{m} w_{i.} \left[1 - w_{i.} \boldsymbol{x}_{iw}^T \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T\right)^{-1} \bar{\boldsymbol{x}}_{iw}\right] [\hat{\boldsymbol{e}}_u^T \hat{\boldsymbol{e}}_u - (n - p)]$$

La stima $\hat{\sigma}_u^2 = \max(0, \tilde{\sigma}_u^2)$; perché il metodo usato non garantisce la non negatività della varianza.

Utilizzando le stime appena ottenute possiamo scrivere la stima EBLUP per modelli a livello di unità:

$$t(\hat{\sigma}_u^2, \hat{\sigma}_e^2) = \hat{\gamma}_i[\overline{y}_i + (\overline{\boldsymbol{X}}_i + \overline{\boldsymbol{x}}_i)^T \hat{\boldsymbol{\beta}}] + (1-\hat{\gamma}_i)\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}$$

Seguendo le indicazioni dalla stima EBLUP generica, il MSE per la stima di $t(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ risulta:

$$MSE[t(\hat{\sigma}_u^2, \hat{\sigma}_e^2)] = g_{1_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{3_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$$

Dove $g_{1_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ e $g_{2_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ so ottengono sostituendo $\hat{\sigma}_u^2$ a σ_u^2 e $\hat{\sigma}_e^2$ a σ_e^2 nella (2.45) e (2.46). La componente $g_{3_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$, sempre ottenuta dalle derivazioni generali, è:

$$g_{3_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) = \frac{1}{w_{i.}^2}\left(\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{w_{i.}}\right)^{-3}[\hat{\sigma}_e^2 \imath^{-1}(\hat{\sigma}_u^2) + \hat{\sigma}_u^2 \imath^{-1}(\hat{\sigma}_e^2) - 2\hat{\sigma}_u^2 \hat{\sigma}_e^2 COV(\hat{\sigma}_u^2, \hat{\sigma}_e^2)]$$

Dove, si ricorda, $\imath^{-1}()$ è la matrice di informazione. Si ricorda altresì che le derivazioni valgono solo se gli errori e_{ij} e gli effetti di area u_i sono distribuiti normalmente, questo perché tutte le derivazioni sono ottenute con la stima di massima verosimiglianza (vedere paragrafo 2.7.1).

Rao (2003) suggerisce una stima per l'errore quadratico medio, sempre nel caso in cui $\hat{\boldsymbol{\delta}} = [\hat{\sigma}_u^2, \hat{\sigma}_e^2]^T$ è ottenuto con la massima verosimiglianza:

$$\widehat{MSE}[t(\hat{\sigma}_u^2, \hat{\sigma}_e^2)] = g_{1_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) - b^T_{(\hat{\sigma}_u^2, \hat{\sigma}_e^2)}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\nabla g_{1_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + g_{2_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) + 2g_{3_i}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$$

Dove il termine di distorsione (Rao, 2003)

$$\boldsymbol{b}^T_{(\hat{\sigma}_u^2, \hat{\sigma}_e^2)}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) =$$

$$= \frac{1}{2} \iota^{-1}(\sigma_u^2) tr\left[\left(\sum_i \frac{1}{\hat{\sigma}_e^2} \left(\sum_{j=1}^{n_i} w_{i.} \boldsymbol{x}_{iw} \boldsymbol{x}_{iw}^T - \hat{\gamma}_i \bar{\boldsymbol{x}}_{iw} \bar{\boldsymbol{x}}_{iw}^T \right) \right)^{-1} \left(\sum_i -(\hat{\sigma}_e^2 + w_{i.} \hat{\sigma}_u^2)^{-2} w_{i.}^2 \boldsymbol{x}_{iw} \boldsymbol{x}_{iw}^T \right) \right] +$$

$$+ \frac{1}{2} \iota^{-1}(\hat{\sigma}_e^2) tr\left[\left(\sum_i \frac{1}{\hat{\sigma}_e^2} \left(\sum_{j=1}^{n_i} w_{i.} \boldsymbol{x}_{iw} \boldsymbol{x}_{iw}^T - \hat{\gamma}_i \bar{\boldsymbol{x}}_{iw} \bar{\boldsymbol{x}}_{iw}^T \right) \right)^{-1} \cdot \right.$$

$$\left. \cdot \sum_i \left(-(\hat{\sigma}_e^2 \sum_j w_{i.} \boldsymbol{x}_{iw} \boldsymbol{x}_{iw}^T) + \frac{\hat{\sigma}_u^2 (2\hat{\sigma}_e^2 + w_{i.} \hat{\sigma}_u^2) w_{i.}^2 \boldsymbol{x}_{iw} \boldsymbol{x}_{iw}^T}{(\hat{\sigma}_e^2 + w_{i.} \hat{\sigma}_u^2)^2 \hat{\sigma}_e^4} \right) \right]$$

La matrice di informazione è una matrice diagonale di dimensione *2x2*:

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2) = \begin{bmatrix} \boldsymbol{V}(\hat{\boldsymbol{\beta}}) & 0 \\ 0 & \begin{bmatrix} \iota_{11}^{-1}(\hat{\sigma}_u^2) & 0 \\ 0 & \iota_{11}^{-1}(\hat{\sigma}_e^2) \end{bmatrix} \end{bmatrix}$$

Dove l'elemento ι_{11}^{-1} è la stima $\hat{\sigma}_u^2$ e l'elemento ι_{22}^{-1} è la stima $\hat{\sigma}_e^2$. $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$ è una matrice *pxp* (dove *p* è il numero delle covariate) diagonale con le stime della varianza dei coefficienti di regressione.

2.5 CONCLUSIONI

In questo capitolo abbiamo visto come ottenere una stima del predittore ottimo lineare e corretto (EBLUP) tramite i modelli lineari ad effetti misti. Date delle ipotesi di partenza è stato dimostrato come arrivare all'EBLUP e alla stima della sua variabilità tramite la stima del MSE. Successivamente, le formulazioni ottenute considerando il caso generale sono state applicate ai modelli di stima per piccole aree basati su modello, con particolare attenzione ai modelli a livello di area, mentre una trattazione meno esaustiva è stata proposta per i modelli a livello di unità.

Lo scopo di queste stime è ottenere un valore medio o il totale di un certo carattere per una piccola area di interesse. Considerando di voler stimare il valore medio di un carattere *y* riferito alla piccola area *i* appartenente a una popolazione di *N* individui sparsi su *m* piccole aree, dove N_i sono il numero di individui nell'area *i* e n_i sono le unità campionate nell'area *i*, sulle quali si osserva il carattere *y*:

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j=1}^{n_i} y_j + \sum_{j=n_i+1}^{N_i} \boldsymbol{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i \right) \tag{2.47}$$

La (2.47) la otteniamo introducendo nella (2.14bis) (vedere paragrafo 2.3.1), che utilizzava il modello lineare ad effetti fissi, il modello lineare ad effetti misti per predire il valore medio nelle aree non campionate. $\boldsymbol{x}_{ij}$ indica il vettore che otteniamo estraendo dalla matrice delle covariate $\boldsymbol{X}$ la riga j-esima tra le righe dell'area i-esima (dove la matrice $\boldsymbol{X}$ ha dimensione Nxp). In pratica si somma la variabile y per le unità campionate e per le unità non campionate si somma la stima di y, ottenuta con il modello lineare ad effetti misti; per ottenere il valore medio ovviamente bisogna dividere per la numerosità della popolazione nell'area i.
Se le informazioni sulle covariate sono note solo a livello di area la (2.47) diventa semplicemente:

$$\hat{\bar{Y}}_i = \left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}} + \hat{u}_i \right)$$

I modelli proposti presuppongono la conoscenza di un certo numero di covariate, quantomeno a livello di area. Per il modello a livello di unità è necessario avere informazioni ausiliarie su tutte le unità della popolazione.

Come abbiamo visto nei paragrafi precedenti, nei modelli di stima per piccole aree abbiamo ipotizzato che non ci siano effetti tra aree, cioè che non ci sia correlazione tra le aree. Questa è una restrizione molto forte soprattutto considerando le aree vicine tra loro. Infatti, intuitivamente, per aree vicine ci aspettiamo che esse siano tra loro correlate. Questa restrizione viene superata con lo stimatore spaziale che presenteremo nel prossimo capitolo.

CAPITOLO 3

LA STATISTICA SPAZIALE NELLA STIMA PER PICCOLE AREE

3.1 INTRODUZIONE

La statistica spaziale è quella parte della statistica che tratta le osservazioni tenendo conto esplicitamente della posizione in cui esse si manifestano nello spazio (Zani e Napoletano, 1992).

Nel caso particolare della stima per piccole aree la posizione delle osservazioni è usata per rimuovere l'ipotesi di indipendenza degli effetti casuali tra aree. Alla base di questo ragionamento c'è il fondamento della statistica spaziale dove, intuitivamente, quello che avviene in un sito dipende in modo rilevante da ciò che accade in siti che gli sono vicini, ma non è legato ad eventi che si verificano in luoghi lontani. Per sito si intende un zona delimitata appartenente ad uno spazio (che in genere è bi/tridimensionale). Quindi i caratteri che si rilevano in una piccola area sono legati in modo rilevante agli stessi caratteri rilevati nelle aree vicine.

Lo studio e l'analisi di numerosi fenomeni di interesse coinvolge, per diversi aspetti, la dimensione spaziale. La diffusione di un'epidemia, l'esposizione al rischio per determinate malattie, i flussi demografici, lo sviluppo urbanistico di una regione, la distribuzione delle precipitazioni, il costo dell'edilizia residenziale in una grossa città, la qualità dell'aria e dell'acqua, la fertilità del suolo, la dislocazione territoriale delle attività produttive sono solo alcuni, disparati, esempi di fenomeni indissolubilmente legati alla dimensione spaziale che negli ultimi anni ha conosciuto un notevole sviluppo.

In parallelo allo sviluppo della statistica spaziale sorge l'esigenza di disporre di un sistema informativo territoriale (in inglese GIS, acronimo di Geographical Information System). Ad oggi il sistema informativo territoriale (adottato nella maggior parte dei paesi) considera uno spazio a due sole dimensioni e il riferimento posizionale di un dato (una osservazione) sulla superficie è generalmente effettuato attraverso le coordinate geografiche. Tramite le coordinate geografiche si possono distinguere tre tipologie di rilevazione:

- PUNTI: si fa riferimento a un punto preciso del territorio, localizzato da una coordinata. Ad esempio le precipitazioni atmosferiche rilevate dalle stazioni metereologiche.

- LINEE: si fa riferimento ad un insieme di coordinate che giacciono su una retta o su più rette consecutive (una spezzata). Ad esempio le linee vengono utilizzate per analizzare gli itinerari turistici.

- AREE: si fa riferimento ad un insieme di coordinate adiacenti. L'area può essere regolare o irregolare. Possiamo dividere una superficie in quadrati oppure possiamo usare altri sistemi di divisione, ad esempio considerare confini geopolitici, come i confini regionali, provinciali, comunali, oppure altre classificazioni territoriali. L'analisi per aree si divide in due branche:

    i. si divide una superficie in aree (siano esse regolari o irregolari) e si associa a ciascuna area una variabile casuale le cui realizzazioni rappresentano le osservazioni spaziali (Griffith, 1988). In pratica si considera l'area come una entità e non si fa differenza per i vari punti che le appartengono.

    ii. si esamina la distribuzione di un insieme di punti su un area, la cui localizzazione è casuale (aleatoria) e l'interesse è rivolto all'analisi del "pattern" della distribuzione (Upton e Fingleton, 1985). In pratica si vuole sapere come sono distribuiti gli elementi su una superficie (generalmente si vuole sapere se sono concentrati in alcune aree o sono distribuiti casualmente).

La rappresentazione delle aree sul piano è generalmente demandata a software specifici, chiamati software GIS. Tramite questi è possibile georeferenziare un dato rilevato su una certa area appartenente a una superficie prestabilita. Per esempio possiamo individuare il punto che rappresenta idealmente Pisa all'interno di un'area irregolare, stabilita in base a criteri amministrativi, come ad esempio la Provincia di Pisa che a sua volta appartiene all'intera superficie considerata, nel caso di esempio, la Toscana.

Concludendo nella statistica spaziale l'obiettivo è quello di stabilire e quantificare la presenza di forme di dipendenza fra le osservazioni nello spazio (Bailey e Gartell, 1995).

3.2 Processo Stocastico Spaziale

I processi stocastici oggetti di studio sono legati allo spazio. Consideriamo lo spazio bidimensionale D dove un punto è individuabile con due valori, chiamati coordinata, contenuti nel vettore $\boldsymbol{s}$. Quindi $\boldsymbol{s} \in \Re^2$ e D viene considerato un sottoinsieme dello spazio Euclideo, con $D \subseteq \Re^2$. Si suppone che per ogni vettore $\boldsymbol{s}$, che localizza un punto, vi sia associato un vettore aleatorio $\boldsymbol{Y}(\boldsymbol{s})$. Con il variare di $\boldsymbol{s}$ sull'insieme D si genere un processo stocastico multivariato (multivariato perché si utilizza un vettore aleatorio anziché una variabile aleatoria):

$$\{\boldsymbol{Y}(\boldsymbol{s}) : \boldsymbol{s} \in D\}$$

Le osservazioni spaziali $\boldsymbol{y}(\boldsymbol{s})$ sono considerate una realizzazione del processo stocastico (Zani e Napoletano, 1992).

Esistono diversi approcci per modellare un processo stocastico spaziale. In questo contesto ne prendiamo in considerazione uno soltanto.

Se si considera un processo stocastico spaziale dove D è fisso ed è un sottoinsieme discreto di $\Re^2$, si individuano solo due processi stocastici spaziali: il SAR, acronimo di Simultaneously Autoregressive Model e il CAR, acronimo di Conditional Autoregressive Model. Il SAR è il modello che utilizzeremo per presentare la stima per piccole aree con correlazione tra aree.

Il SAR utilizza i principi dei processi stocastici temporali per applicarli ai processi legati allo spazio. Dato che in un processo stocastico temporale il valore della variabile aleatoria può essere influenzato sia dal valore attuale sia dai valori passati, oppure da nessuno dei due, intuitivamente in un processo stocastico spaziale il valore di una variabile che appartiene ad una cera area varia anche (o solo) in base ai valori che la stessa variabile assume nelle aree vicine.

Consideriamo A_i come l'area i-esima appartenente al territorio D. Si consideri inoltre che le aree in questione non abbiano parti in comune e che nessuna zona del territorio D sia scoperta, in breve:

$$A_i \in D$$
$$A_1 \cup A_2 \cup \ldots \cup A_m = D$$
$$A_i \cap A_j = \varnothing \; \forall i \neq j$$

Si definisce processo stocastico Gaussiano:

$$\{Y(A_i) : A_i \in (A_1, \ldots, A_m)\}$$

Un processo Gaussiano è un processo stocastico dove ogni combinazione lineare della variabile aleatoria (X_t) è distribuita normalmente.
Questo processo può essere rappresentato da un modello SAR:

$$Y(A_i) = \mu_i + \sum_{i=1}^{m} b_{ij}(Y(A_j) - \mu_j) + e_i$$

Dove e_i è distribuito normalmente con media 0 e varianza λ_i ($e_i \sim N(0, \lambda_i)$) e quindi il vettore $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{\Lambda})$ con $\boldsymbol{\Lambda}$ una matrice diagonale di dimensione m. b_{ij} sono costanti non necessariamente note, con $b_{ii} = 0$. Se m è un numero finito $\boldsymbol{B}$ è un amtrice quadrata di rango m dove l'elemento ij è uguale a b_{ij}. Senza darne una dimostrazione ci limitiamo a fornire il valore atteso del processo stocastico $Y(A_i)$:

$$E[Y(A_i)] = \mu_i$$

Considerando la distribuzione congiunta $\boldsymbol{Y} = [Y(A_1), \ldots, Y(A_m)]^T$ si dimostra che ha:

$$E[\boldsymbol{Y}] = \boldsymbol{\mu}$$

$$V[\boldsymbol{Y}] = (\boldsymbol{I}_m - \boldsymbol{B})^{-1} \boldsymbol{\Lambda} [(\boldsymbol{I}_m - \boldsymbol{B})^{-1}]^T$$

Dove $\boldsymbol{I}_m$ è una matrice identica di rango m e $\boldsymbol{\mu}$ è il vettore dei valori attesi μ_i ($\boldsymbol{\mu} = [\mu_1, \ldots, \mu_m]^T$).

Riassumendo, il processo stocastico spaziale Gaussiano rappresentato dal modello SAR risulta:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, (\boldsymbol{I}_m - \boldsymbol{B})^{-1} \boldsymbol{\Lambda} [(\boldsymbol{I}_m - \boldsymbol{B})^{-1}]^T) \tag{3.1}$$

3.3 I DATI SPAZIALI

Per utilizzare processi stocastici spaziali è necessario che i dati raccolti siano riconducibili alla coordinata, o almeno all'area, su cui sono stati rilevati. Dopodichè bisogna utilizzare un metodo di analisi che tenga conto della spazialità del dato. Per fare questo si utilizza la matrice di connessione. La matrice di connessione è una matrice in cui l'elemento *ij* indica la relazione spaziale tra l'unità *i* e l'unità *j*. La più semplice matrice di connessione che possiamo immaginare è fatta ponendo l'elemento *ij* uguale a 1 se le aree *i* e *j* sono confinanti e uguale a 0 altrimenti. Immaginiamo di avere una superficie quadrata divisa in 9 aree, anch'esse quadrate:

1	2	3
4	5	6
7	8	9

L'area 1 si considera confinante con l'area 4, l'area 5 e l'area 2. Con gli indici $i = j = 1, \ldots, 9$ si identificano le aree. La matrice di connessione, che ha dimensione *9x9*, viene rappresentata con $\boldsymbol{W}$:

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

Se osserviamo la matrice si notano due aspetti importanti (e ovvi):

a. Gli elementi sulla diagonale principale w_{ii} sono sempre uguali a 0; l'area non può confinare con se stessa.

b. La matrice $\boldsymbol{W}$ è simmetrica, quindi l'elemento $w_{ij} = w_{ji}$ per ogni i diverso da j (nel caso specifico l'uguaglianza è vera anche per $i = j$). E' scontato che l'area i confina con l'area j se e solo se l'area j confina con l'area i, questo genera la simmetria nella matrice di connessione.

Esistono altri metodi per creare la matrice di connessione. Si può immaginare una coordinata che identifica il centro di un'area, calcolata con il metodo del centroide o con altri metodi geometrici e non, tracciare un raggio di lunghezza k a partire dalla coordinata specificata e considerare "confinanti" tutte le aree che cadono nel cerchio immaginario generato da tale raggio, oppure considerare solo le aree il cui centro cade nel cerchio immaginario; in quest'ultimo caso sono confinati le aree il cui centro dista meno di una distanza k dal centroide. E' possibile che all'interno di un'area ci siano diverse unità rilevate, che abbiamo detto identificabili con una coordinata, da trattare indipendentemente l'una dall'altra. In questo caso si costruisce una matrice di connessione considerando elemento per elemento, ad esempio se abbiamo 500 unità rilevate avremo una matrice di connessione di *500x500*. I criteri di classificazione della contiguità sono gli stessi appena discussi.
Non sempre risulta banale dire se due aree sono confinanti. Nell'esempio delle 9 aree risulta evidente che l'area 1 confina con l'area 2 e con la 4, tuttavia il fatto che confini anche con l'area 5 è discutibile. Infatti per le aree 2 e 4 c'è un lato in comune con l'area 1, mentre tra le aree 1 e 5 c'è in comune soltanto un punto. Upton e Fingleton (1985) suggeriscono di utilizzare i lati per determinare la contiguità delle aree. Come abbiamo potuto constatare non è semplice stabilire un meccanismo per determinare il legame spaziale tra due o più aree; si possono immaginare le difficoltà per i casi in cui le aree sono irregolari (come quelle di comuni, province, sezioni di censimento, etc.).
Cliff e Ord (1981) propongono di utilizzare anche valori alternativi a 1 o 0. In questi casi l'elemento w_{ij} è espressione di una relazione tra l'unità i l'unità j che varia secondo certe caratteristiche, per questo viene considerato come "peso". Considerando la

distanza tra unità, o tra centri, e considerando il perimetro comune tra le aree Cliff e Ord suggeriscono uno schema generalizzato di pesi:

$$w_{ij} = \frac{1}{d_{ij}^{a}} \beta_{ij}^{b}$$

dove d_{ij} è la distanza tra le unità i e j, o tra i centri delle aree i e j, β_{ij} è la proporzione di perimetro tra le aree i e j (o delle aree che contengono le unità i e j, se si tratta di unità) e a e b sono parametri reali che servono a pesare le due informazioni prese in considerazione. La distanza è intesa come distanza euclidea; date due coordinate $\boldsymbol{c}_1 = (x_1, y_1)$ e $\boldsymbol{c}_2 = (x_2, y_2)$ la distanza euclidea si definisce come:

$$\|\boldsymbol{c}_1 - \boldsymbol{c}_2\| = \sqrt{(\boldsymbol{c}_1 - \boldsymbol{c}_2)^T (\boldsymbol{c}_1 - \boldsymbol{c}_2)} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

La matrice di connessione è un aspetto importante della statistica spaziali, infatti cambiandone la definizione i risultati ottenuti possono cambiare.

I dati spaziali hanno delle caratteristiche che li differenziano da altri tipi di osservazioni. Una caratteristica tipica dei dati spaziali è la *dipendenza*. Non si considera mai che i dati spaziali siano tra loro indipendenti. Questo deriva dalla prima legge della geografia di Tobler (1970), secondo la quale tutto è correlato con tutto, ma le cose più vicine sono più correlate delle cose lontane. Basti pensare ad esempio alla distribuzione del reddito nei centri abitati, spesso, individuata la zona più ricca ed allontanandoci da essa troviamo zone sempre meno ricche, fino ad arrivare alle periferie più lontane e povere. Quando c'è la possibilità che un fenomeno, legato a diversi luoghi di un territorio, sia influenzato dalla contiguità spaziale con luoghi in cui il fenomeno è osservato si parla di autocorrelazione spaziale (Upton e Fingleton (1985)), concetto approfondito in seguito.

Altro aspetto caratteristico dei dati spaziali è il legame multidirezionale. I dati sono legati e influenzati da tutte le unità che li circondano, al contrario in una serie storica l'influenza dei dati avviene longitudinalmente.

3.4 L'AUTOCORRELAZIONE SPAZIALE

Innanzitutto ricordiamo la definizione di autocorrelazione. In statistica l'autocorrelazione di una serie temporale discreta o di un processo stocastico X_t è semplicemente la correlazione che esiste tra un dato al tempo *t* o il valore di un processo stocastico al tempo *t* e un dato al tempo *t-k* o il valore del processo stocastico al tempo *t-k*, in formula:

$$\rho(k) = \frac{E[(X_i - \mu)(X_{i-k} - \mu)]}{\sigma_X^2}$$

Dove *k*, che spesso è chiamato "*lag*" o ritardo, rappresenta il numero di "passi" indietro nella serie storica o nel processo. Si suppone in questa formulazione che la serie sia omoschedastica. Si può calcolare il coefficiente di autocorrelazione per ogni ritardo, con $k = 1 \ldots n$ dove *n* indica la numerosità della serie o numero di passi del processo.

Nella statistica spaziale l'autocorrelazione assume un ruolo diverso, infatti non si muove nel "tempo" ma nello "spazio".

L'autocorrelazione spaziale è una proprietà che i dati spaziali possiedono ogni volta che c'è una sistematica variazione nei valori disposti su uno spazio (Cliff e Ord 1981). In pratica, quando il valore di un certo carattere, legato all'area in cui è stato rilevato, varia secondo una "legge" legata ai valori dello stesso osservati nelle aree vicine siamo in presenza di autocorrelazione; intuitivamente si può dire che se la correlazione che esiste tra due "oggetti" è dovuta alla loro posizione nello spazio allora i due oggetti sono autocorrelati spazialmente[35].

L'analisi dell'autocorrelazione spaziale si propone di verificare se ed in quale misura sussistono iterazioni spaziali in forza delle quali si realizzano, reciprocamente, influenze tra aree territoriali definite vicine (Salvati, 2004).

Se nelle coppie di luoghi contigui il fenomeno studiato assume determinazioni simili siamo in presenza di autocorrelazione spaziale positiva, si dice invece che c'è autocorrelazione spaziale negativa se nelle coppie di luoghi contigui il fenomeno presenta determinazioni divergenti (Badaloni e Vinci, 1988).

[35] Il termine autocorrelazione può essere sviante poiché si verifica tra entità diverse. Tuttavia le entità appartengono allo stesso spazio ed è per questo che si utilizza il suffisso "auto".

Per misurare l'autocorrelazione spaziale sono state fatte molte proposte, in proposito si veda Upton e Fingleton (1985). Haining (1980) propone tre processi stocastici spaziali per definire l'autocorrelazione spaziale:

1. *n* variabili continue $\{X_j\}$ sono assegnate ad *n* aree (diverse). Per ogni area la variabile è estratta (in totale si fanno *n* estrazioni) da $N(0, \sigma^2)$. Non c'è autocorrelazione spaziale se le *n* variabili sono totalmente indipendenti tra loro; questo è vero se e solo se $P(X_1<x_1, \ldots, X_n<x_n) = \prod_{i=1}^{n} P(X_i < x_i)$.

2. *n* variabili dicotomiche $\{X_j\}$ sono assegnate a *n* aree. Per ogni area la variabile è estratta da una Bernulli(0,5). L'assenza di autocorrelazione spaziale corrisponde alla totale indipendenza delle *n* variabili confrontate a coppie: questo è vero se e solo se $P(X_i = a, X_j = b) = P(X_i = a)$ x $(X_j = b)$ per ogni $i \neq j$.

3. I valori di una variabile Y_{ij} riferiti alla locazione di coordinate (i,j) è in qualche modo influenzata dai valori che la stessa variabile assume nelle aree vicine, Haining propone la seguente relazione:

$$Y_{i,j} = \rho(Y_{i-1,j} + Y_{i+1,j} + Y_{i,j-1} + Y_{i,j+1}) + e_{i,j}$$

Dove ρ indica una legge che lega i dati in oggetto ed $e_{ij} \sim N(0, \sigma^2)$ con $E[e_i e_j] = 0$.

I metodi 1 e 2 proposti da Hining richiedono la conoscenza della distribuzione di probabilità della variabile di studio *X*. Questo comporta dover fare assunzioni non vere o dover avere una conoscenza approfondita della variabile di studio.
Cliff e Ord (1981) riprendendo un lavoro proposto inizialmente da Knox(1964) e generalizzato da Mandel (1967) propongono un metodo alternativo per vedere se c'è autocorrelazione spaziale tra i dati: la "general cross product statistic". Essa è data dall'equazione:

$$r = \sum_i \sum_j w_{ij} y_{ij}$$

Dove w_{ij} è l'elementi *ij*-esimo della matrice di connessione (nella versione 0-1) che misura la prossimità "fisica" delle unità (o aree) *i* e *j*, mentre y_{ij} è una misura della prossimità tra le unità (o aree) *i* e *j* riferita ad altre caratteristiche: è una misura di prossimità non spaziale. Generalmente se la variabile osservata nell'area *i* è x_i e quella osservata nell'area *j* è x_j si definisce la matrice $\boldsymbol{Y}$ come:

$$y_{ij} = (x_i - x_j)^2$$

Quindi dato un carattere *x* rilevato sulle unità (o aree) e una matrice di connessione possiamo calcolare la statistica *r*.

Cliff e Ord (1981) hanno studiato la statistica *r* e hanno derivato media e varianza campionaria considerando che per un numero elevato di unità, o aree, tale statistica si può approssimare con una normale. Questo permette di fare un test di verifica d'ipotesi. L'ipotesi nulla proposta è assenza di autocorrelazione spaziale, contro l'ipotesi alternativa di presenza di autocorrelazione spaziale. I valori attesi della statistica *r* sono:

$$E[r] = \frac{S_0 T_0}{m(m-1)}$$

$$V[r] = \frac{S_1 T_1}{2m^{(2)}} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4m^{(3)}} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{m^{(4)}} - \left(\frac{S_0 T_0}{m^{(2)}}\right)^2$$

Dove:

$$S_0 = \sum_i \sum_j w_{ij} \quad con\ i \neq j$$

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \quad con\ i \neq j$$

$$S_2 = \sum_i (w_{i.} + w_{.i})^2 \quad con\ \ w_{i.} = \sum_j w_{ij};\ \ w_{.i} = \sum_j w_{ji}$$

$$m^{(2)} = m(m-1);\ \ m^{(3)} = m(m-1)(m-2);\ \ m^{(4)} = m(m-1)(m-2)(m-3)$$

con *m* che indica il numero di aree (o unità).

$$T_0 = \sum_i \sum_j y_{ij} \quad con\, i \neq j$$

$$T_1 = \frac{1}{2} \sum_i \sum_j (y_{ij} + y_{ji})^2 \quad con\, i \neq j$$

$$T_2 = \sum_i (y_{i.} + y_{.i})^2 \quad con \;\; y_{i.} = \sum_j y_{ij}; \;\; y_{.i} = \sum_j y_{ji}$$

Per fare il test di ipotesi si calcola il valore Z^* da confrontare con il valore Z teorico (dato un certo errore di I tipo); per ottenere il valore Z^* occorrono r, il suo valore atteso e la sua varianza attesa:

$$Z^* = \frac{|r - E[r]| - 1}{V[r]^{1/2}}$$

Dove "-1" è un fattore di correzione (si veda in proposito Upton e Fingleton, 1985).
Se il valore Z teorico (quel valore per il quale l'integrale sotteso alla curva della distribuzione normale standard da quel punto a più infinito è uguale ad α (l'errore di I tipo)) per un certo α, Z_α, è maggiore di Z^* allora non rifiutiamo l'ipotesi nulla; cioè dichiariamo che non c'è autocorrelazione spaziale. Viceversa, se $Z_\alpha < Z^*$, siamo in presenza di autocorrelazione spaziale.
La statistica r ha lo svantaggio di non fornire informazioni sul segno dell'autocorrelazione.
Moran già nel 1950 ha proposto un indice in grado di individuare l'autocorrelazione spaziale ed il suo segno. Indichiamo con x_i e x_j una qualsiasi variabile misurata rispettivamente nell'area i e nell'area j. L'indice di Moran, rappresentato con I_M, è definito come:

$$I_M = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Dove $\bar{x}$ è la media delle osservazioni della variabile di studio[36] e n sono il numero di aree (o di unità)[37]. La media attesa e la varianza attesa di I_M sono note:

[36] Gneralmente abbiamo indicato con x le covariate e con y la variabile interesse di studio, in questo caso con x indichiamo la variabile di studio e con y una sua funzione.

$$E[I_M] = -\frac{1}{m-1}$$

$$V[I_M] = \frac{m[(m^2-3m+3)S_1 - mS_2 + 3S_0^2] - \frac{a_4}{a_2^2}[m^{(2)}S_1 - 2mS_2 + 6S_0^2]}{(m-1)^{(3)}S_0^2} - \left(\frac{1}{(m-1)}\right)^2$$

Dove le lettere precedentemente introdotte non cambiano significato e $a_r = \frac{1}{m}\sum_i (x_i - \bar{x})^r$. La statistica $I_M \sim N(E[I_M],\ V[I_M])$ se $n \geq 20$ con assenza di autocorrelazione spaziale. Sotto ipotesi nulla di assenza di autocorrelazione si può fare un test statistico con lo stesso ragionamento descritto in precedenza. Il valore Z^* in questo caso è ottenuto con la formulazione classica:

$$Z^* = \frac{I_M - E[I_M]}{V[I_M]^{\frac{1}{2}}}$$

Se $Z_\alpha \leq Z^*$ rifiutiamo l'ipotesi nulla in favore dell'ipotesi alternativa di presenza di autocorrelazione spaziale e viceversa. L'indice di Moran ci da un'informazione sul segno dell'autocorrelazione. Infatti, una volta fatto il test e verificato che c'è autocorrelazione spaziale, se I_M ha valori positivi vuol dire che siamo in presenza di autocorrelazione spaziale positiva, viceversa se ha valori negativi. I_M non varia tra -1 e 1, ma si può ricondurre a questo campo di esistenza: bisogna dividere I_M per il suo limite massimo che è

$$|I_M| \leq \frac{m}{\sum_{i=1}^m \sum_{j=1, j\neq i}^m w_{ij}} \left[\frac{\sum_{i=1}^m (\sum_{j=1, j\neq i}^m w_{ij}(x_j - \bar{x}))^2}{\sum_{i=1}^m (x_i - \bar{x})^2} \right]^{\frac{1}{2}}$$

[37] Se si parla di unità, x_i è la variabile rilevata sull'unità i e $\bar{x}$ è la media di tutte le unità; se si parla si aree x_i è un valore rappresentativo dell'area i (può essere la media delle osservazione effettuate sull'area i), con $\bar{x}$ media delle medie di area.

In questo modo $I_M = 1$ indica la massima autocorrelazione spaziale positiva e $I_M = -1$ indica la massima negativa.

Hubert e al. (1981) dimostra che se poniamo $Y_{ij} = (x_i - \bar{x})(x_j - \bar{x})$ si può ottenere l'indice di Moran dividendo la statistica *r* per S_0 moltiplicato la varianza campionaria di *x*.

Un indice alternativo all'indice di Moran è l'indice di Geary, proposto nel 1954:

$$I_G = \frac{(m-1)\sum_i \sum_{j,j\neq i} w_{ij}(x_i - \bar{x})^2}{2S_0 \sum_i (x_i - \bar{x})^2}$$

I valori attesi di media e varianza sono:

$$E[I_G] = 1$$

$$V[I_G] = \frac{(m-1)S_1[m^2 - 3m + 3 - \frac{a_4}{a_2^2}(m-1)]}{S_0^2 m(m-2)(m-3)} + \frac{m^2 - 3 - \frac{a_4}{a_2^2}(m-1)^2}{m(m-2)(m-3)} -$$

$$- \frac{(m-1)S_2[m^2 + 3m - 6 - \frac{a_4}{a_2^2}(m^2 - m + 2)]}{4S_0^2 m(m-2)(m-3)}$$

Con lo stesso procedimento presentato per l'indice di Moran si verificano le ipotesi sull'autocorrelazione spaziale. Anche questo indice non è compreso tra -1 e 1 ma al crescere dell'indice oltre il valore 1 corrisponde una sempre maggiore autocorrelazione negativa mentre per valori che si avvicinano allo 0 corrisponde una crescente autocorrelazione positiva. 1, essendo il valore atteso, indica l'assenza di autocorrelazione spaziale.

Ci sono altri indici per valutare la presenza e l'intensità di autocorrelazione spaziale, per un approfondimento si rimanda a Upton e Fingleton (1985), comunque gli indici presentati sono tra i più usati.

3.5 MODELLI AUTOREGRESSIVI SPAZIALI

Quando si parla di modelli autoregerssivi ci si riferisce generalmente a processi stocastici di una serie storica. L'esempio più classico e rappresentativo di modello autoregressivo è il modello AR(1):

$$e_t = \rho e_{t-1} + \varepsilon_t$$

Dove $\varepsilon_t \sim N(0,\sigma_\varepsilon^2)$ e ρ è il coefficiente di autocorrelazione. In pratica un modello autoregressivo è un processo stocastico in cui il valore attuale del processo è determinato anche dal valore che aveva il processo stesso allo stadio precedente (nelle serie storiche nel periodo precedente). Esistono molti tipi di modelli autoregressivi, la loro classificazione varia in base ai criteri con cui bisogna manipolare la serie storica affinché sia stazionaria. Una serie storica si dice stazionaria se e solo se:

$$\begin{aligned} &E[X_t] = k \quad \forall t \\ &V[X_t] = k \quad \forall t \\ &COV[X_t, X_{t+k}] = k \quad \forall t, k \; con \; t \neq k \end{aligned}$$

Dove k è una costante. Il più semplice processo stocastico, noto come "white noise", è per definizione stazionario:

$$\begin{aligned} &X_t = \varepsilon_t \;\; dove \; \varepsilon_t \sim N(0,\sigma_\varepsilon^2) \; e \; COV[\varepsilon_t, \varepsilon_{t-d}] = 0 \; \forall t, d \; con \; t \neq d \\ &E[X_t] = 0 \\ &V[X_t] = \sigma_\varepsilon^2 \\ &COV[X_t, X_{t-d}] = 0 \end{aligned}$$

Altri processi invece, come l'AR(1), non sono stazionari per definizioni e bisogna operare trasformazioni sulla serie perché lo siano; nel caso AR(1) con ρ = 1, meglio noto in economia come "random walk":

$$\begin{aligned} &E[X_t] = X_0 \\ &V[X_t] = E[(X_t - X_0)^2] = t\sigma_\varepsilon^2 \end{aligned}$$

La varianza attesa non è costante per ogni *t*, quindi il modello non è stazionario. Per rendere la serie stazionaria bisogna trasformarla sottraendo da ogni termine il precedente, in questo modo:

$$X_t^{(trasformata)} = (X_t - X_{t-1}) = \varepsilon_t$$

che altro non è che un processo white noise (che abbiamo appena detto essere stazionario). Il processo si chiama AR(1) proprio perché risulta stazionario nelle differenze prime, da cui "(1)".

E' importante che un processo stocastico sia stazionario altrimenti non può essere modellato, come è stato dimostrato da Granger in una famosa simulazione Montecarlo.

Nel caso dei modelli autoregressivi spaziali il concetto non cambia molto. Anche in questo caso il modello può essere definito, non rigorosamente, come un processo stocastico spaziale dove un "punto" è definito in base al valore dei "punti" vicini. Più rigorosamente un processo stocastico spaziale si dice stazionario se le sue proprietà statistiche sono invarianti rispetto a traslazioni, cioè la sua distribuzione è inalterata quando l'origine dell'insieme degli indici viene traslato.

Nei modelli di regressione lineare, come quelli da noi considerati, la dipendenza spaziale può essere introdotta in due modi; vediamo la formulazione di un modello per ognuno dei due modi.

Nel primo modello la dipendenza spaziale è introdotta tramite un lag spaziale; il modello autoregressivo spaziale per la variabile di studio *y* è definito come:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

Dove $\boldsymbol{Wy}$ rappresenta il lag spaziale, con $\boldsymbol{W}$ matrice di connessione di dimensioni *mxm* e ρ coefficiente di autoregressione spaziale, $\boldsymbol{e}$ rappresenta il vettore degli errori di dimensione *m* ($\boldsymbol{e} \sim N(0,\sigma_e^2)$) e $\boldsymbol{X}$ di dimensione *mxp*, $\boldsymbol{\beta}$ di dimensione *p* non cambiano significato rispetto a quanto detto sinora.

Nel secondo modello la dipendenza spaziale si introduce direttamente nella struttura dell'errore, che per distinguerlo chiamiamo $\boldsymbol{v}$: $E[v_i v_j] \neq 0$ per ogni $i \neq j$. In questo modo si determina una matrice di varianza-covarianza piena:

$$V[\boldsymbol{v}] = \boldsymbol{\Omega}(\boldsymbol{\delta})$$

Dove $\boldsymbol{\delta}$ è il vettore dei parametri che riguardano un particolare processo stocastico spaziale. Utilizzeremo questo modello per ricavare un predittore ottimo lineare e corretto che tenga conto della dipendenza tra aree.

Nel paragrafo 3.2 abbiamo parlato di processi stocastici spaziali Gaussiani e di una loro rappresentazione tramite il modello SAR. Il modello autoregressivo spaziale che utilizzeremo si basa sul processo stocastico SAR:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{v} \tag{3.2}$$

Dove $\boldsymbol{v}$ è un processo stocastico autoregressivo multivariato di tipo SAR:

$$\boldsymbol{v} = \rho\boldsymbol{W}\boldsymbol{v} + \boldsymbol{e} \tag{3.2bis}$$

Con $\boldsymbol{e}$ vettore di errori indipendenti e normalmente distribuiti, $E[\boldsymbol{e}] = \boldsymbol{0}$ e $E[\boldsymbol{e}\boldsymbol{e}^T] = \sigma_e^2\boldsymbol{I}$. Dalla (3.2) e (3.2bis) si ricava una nuova formulazione del modello:

$$\boldsymbol{v} = \rho\boldsymbol{W}\boldsymbol{v} + \boldsymbol{e} \rightarrow \boldsymbol{v} - \rho\boldsymbol{W}\boldsymbol{v} = \boldsymbol{e} \rightarrow \boldsymbol{v} = (\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}\boldsymbol{e} \Rightarrow$$

$$\Rightarrow \boldsymbol{y} = \boldsymbol{X\beta} + (\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}\boldsymbol{e} \tag{3.3}$$

Dove $\boldsymbol{I}_m$ è una matrice identica di ordine m. Il processo stocastico autoregressivo $\boldsymbol{v}$ è stazionario, infatti:

$$E[v] = E[(\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}e] = \text{data l'indipendenza tra } \boldsymbol{e} \text{ e } (\boldsymbol{I}_m\text{-}\rho\boldsymbol{W})^{-1} =$$

$$= E[(\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}]E[\boldsymbol{e}] = 0$$

$$V[\boldsymbol{v}] = E[(\boldsymbol{v} - E[\boldsymbol{v}])(\boldsymbol{v} - E[\boldsymbol{v}])^T] = E[\boldsymbol{v}\boldsymbol{v}^T] = E[(\boldsymbol{I}_m - \rho\boldsymbol{W})\boldsymbol{e}\boldsymbol{e}^T(\boldsymbol{I}_m - \rho\boldsymbol{W})^T] =$$

$$= (\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}E[\boldsymbol{e}\boldsymbol{e}^T]((\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1})^T = (\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}\sigma_e^2(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)^{-1} =$$

$$= \sigma_e^2(\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)^{-1} = \sigma_e^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}$$

Anche le covarianze sono considerate costanti rispetto alle traslazioni. In letteratura non c'è accorda sulla stazionarietà del processo stocastico proposto. In questo lavoro si "abbraccia" la teoria di stazionarietà.
Considerando la (3.3) vediamo la distribuzione di $\boldsymbol{y}$:

$$E[\boldsymbol{y}] = E[\boldsymbol{X\beta} + \boldsymbol{v}] = \boldsymbol{X\beta} + E[\boldsymbol{v}] = \boldsymbol{X\beta} \tag{3.4}$$

$$V[\boldsymbol{y}] = E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T] = E[(\boldsymbol{y} - \boldsymbol{X\beta})(\boldsymbol{y} - \boldsymbol{X\beta})^T] =$$
$$= E[(\boldsymbol{X\beta} + \boldsymbol{v} - \boldsymbol{X\beta})(\boldsymbol{X\beta} + \boldsymbol{v} - \boldsymbol{X\beta})^T] = E[\boldsymbol{vv}^T] = \sigma_e^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1} \tag{3.5}$$

Dalla (3.4) e (3.5) concluiamo che il modello $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{v}$ è distribuito:

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma_e^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1})$$

Riprendendo la (3.1), $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, (\boldsymbol{I}_m - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}[(\boldsymbol{I}_m - \boldsymbol{B})^{-1}]^T)$, che ci da la distribuzione del processo stocastico SAR, si può verificare che posto $\boldsymbol{\mu} = \boldsymbol{X\beta}$, $\boldsymbol{B} = \rho\boldsymbol{W}$ e $\boldsymbol{\Lambda}=\sigma_e^2\boldsymbol{I_m}$ il modello (3.2) è un caso particolare del processo stocastico autoregressivo spaziale Gaussiano di tipo SAR.
Assumendo che le osservazioni y_i siano una realizzazione della variabile casuale $\boldsymbol{y}$, di cui conosciamo la distribuzione di probabilità, possiamo stimare i parametri incogniti $\boldsymbol{\beta}$, σ_e^2 e ρ con il metodo della massima verosimiglianza:

$$l(\boldsymbol{\beta}, \sigma_e^2, \rho) =$$
$$= 2\pi^{-\frac{m}{2}} \left|\sigma_e^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}\right|^{-\frac{1}{2}} \cdot$$
$$\cdot \exp\left\{-\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta})^T\left[[\sigma_e^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}\right]^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right]\right\}$$

Passando ai logaritmi e ponendo $\boldsymbol{A} = (\boldsymbol{I}_m\text{-}\rho\boldsymbol{W})$ per semplificare la notazione si ottiene:

$$ll(\boldsymbol{\beta}, \sigma_e^2, \rho) = -\frac{m}{2}\ln 2\pi - \frac{1}{2}\ln\left|\sigma_e^2 (\boldsymbol{A}^T \boldsymbol{A})^{-1}\right| - \frac{1}{2\sigma_e^2}\left[(\boldsymbol{y} - \boldsymbol{X\beta})^T [\boldsymbol{A}^T \boldsymbol{A}](\boldsymbol{y} - \boldsymbol{X\beta})\right] \quad (3.6)$$

Sfruttando una derivazione di Ord si può calcolare il determinante di $\boldsymbol{A}$ come:

$$\boldsymbol{A} = \boldsymbol{I}_m - \rho \boldsymbol{W} = \prod_{i=1}^{m}(1 - \rho\lambda_i)$$

Dove λ_i son gli autovalori della matrice di connessione $\boldsymbol{W}$. Derivando la funzione di log-verosimiglianza rispetto a σ_e^2 e ponendola uguale a 0 se ottiene la stima:

$$\hat{\sigma}_e^2 = \frac{1}{m}(\boldsymbol{y}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{y} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{y} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}) \quad (3.7)$$

Sostituendo la (3.7) nella funzione (3.6) si ottiene la funzione di log-verosimiglianza

$$ll(\boldsymbol{\beta}, \hat{\sigma}_e^2, \rho) = k - \frac{m}{2}\ln[\hat{\sigma}_e^2 (|\boldsymbol{A}|)^{-\frac{2}{n}}]$$

Dove k è una costante che non influenza il processo di stima. Sfruttando la proprietà (3.40) scriviamo:

$$ll(\boldsymbol{\beta}, \hat{\sigma}_e^2, \rho) = k - \frac{m}{2}\ln\left[\hat{\sigma}_e^2 \left(\prod_{i=1}^{m}(1 - \rho\lambda_i)\right)^{-\frac{2}{n}}\right] \quad (3.8)$$

Con le derivazioni ottenute si può iniziare un processo iterativo per stimare i parametri di interesse $\boldsymbol{\beta}$, σ_e^2 e ρ:

1. Considerando un modello di regressione $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ (tradizionale) si stima il vettore $\boldsymbol{\beta}$ con gli OLS, si ottiene $\hat{\boldsymbol{\beta}}$

2. Si attribuisce un valore iniziale a ρ secondo la formula:

$$\rho^{(0)} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}$$

3. Si calcola una stima di σ_e^2 tramite la (3.7) utilizzando i valori stimati in 1. e 2.

4. Si minimizza la (3.8) per stimare il coefficiente di autoregressione spaziale utilizzando l'algoritmo di ottimizzazione di Newton-Rapson:

$$\rho^{(r+1)} = \rho^{(r)} - \frac{f'(\rho^{(r)})}{f''(\rho^{(r)})}$$

$$\text{Con } f(\rho^{(r)}) = -\frac{2}{m}\sum_{i=1}^{m} \ln(1 - \rho\lambda_i) + \ln\left[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{A}^T \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right]$$

5. Si ricalcala una stima di σ_e^2 utilizzando la stima di ρ ottenuta in 4. nella (3.7)

6. Si stima nuovamente $\boldsymbol{\beta}$ con i GLS utilizzando le stime ottenute in 5. e 4.:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\rho}, \hat{\sigma}_e^2) = (\boldsymbol{X}^T (\hat{\sigma}_e^2 \boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\hat{\sigma}_e^2 \boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{y}$$

7. Si ritorna al passo 2. e si continua finché la differenza tra le stime ottenute in due iterazioni successive non sono minore di un valore prefissato a piacere, in termini specifici finché l'algoritmo non converge.

3.6 Spatial Empirical Best Linear Unbiased Predictor

Come accennato ad inizio capitolo siamo in gradi di rimuovere l'ipotesi di indipendenza degli effetti di area nel modello lineare ad effetti misti. La rimozione di tale ipotesi segna un passo importante per la precisione nella stima per piccole aree poiché sono molti i casi in cui tra i dati c'è autocorrelazione spaziale, in genere positiva.

Nel modello BLUP $\boldsymbol{u}$ rappresentava una variabile casuale multivariata distribuita normalmente con media 0 e varianza σ_u^2 e matrice di varianza covarianza diagonale (poiché supponevamo l'indipendenza tra gli effetti casuali di area). Consideriamo il modello autoregressivo spaziale $\boldsymbol{v}$ ed effettuiamo un cambio di notazione dell'errore $\boldsymbol{e}$: $\boldsymbol{e}$

= $\boldsymbol{u}$; questo per non confondere nel modello di stima Spatial EBLUP l'errore del modello autoregressivo con l'errore indotto dal campionamento, che continueremo a chiamare $\boldsymbol{e}$. Con la nuova notazione abbiamo:

$$\boldsymbol{v} = \rho \boldsymbol{W}\boldsymbol{v} + \boldsymbol{u} \rightarrow \boldsymbol{v} = (\boldsymbol{I}_m - \rho \boldsymbol{W})^{-1}\boldsymbol{u}\text{, dove } u_i \sim N(0, \sigma_u^2) \tag{3.9}$$

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R}) \tag{3.44}$$

Introducendo il modello autoregressivo spaziale nel modello BLUP otteniamo un modello di stima SAR per piccole aree:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}$$

Utilizzando la (3.9) si ottiene:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}(\boldsymbol{I}_m - \rho \boldsymbol{W})^{-1}\boldsymbol{u} + \boldsymbol{e}$$

Dove:

- $\boldsymbol{y}$ è il vettore, di dimensione m, della variabile di interesse per le piccole aree; quindi y_i è il valore della variabile di studio nell'area i.
- $\boldsymbol{X}$ è la matrice, di dimensioni mxp, delle variabili ausiliarie (covariate); la riga i-esima della matrice è il vettore $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{ip}]^T$, il vettore delle covariate dell'area i.
- $\boldsymbol{\beta}$ è il vettore, di dimensione p, dei coefficienti di regressione.
- $\boldsymbol{Z}$ è una matrice di costanti note di dimensione mxm; $\boldsymbol{Z}$ è la matrice di selezione, cioè un'identica di ordine m. In questo modo per ogni area si selezione dal vettore $[\boldsymbol{I_m}\text{-}\rho\boldsymbol{W}]^{-1}\boldsymbol{u}$ l'effetto di area i-esimo, che è calcolato considerando gli effetti delle aree vicine in base all'autocorrelazione spaziale ρ.
- $[\boldsymbol{I_m}\text{-}\rho\boldsymbol{W}]^{-1}\boldsymbol{u} = \boldsymbol{v}$ è il vettore degli effetti casuali di area, di dimensione m. $\boldsymbol{I}_m$ è una matrice identica di ordine m, ρ è il coefficiente di autocorrelazione spaziale (costante), $\boldsymbol{W}$ è la matrice di connessione (di qualsiasi genere), di dimensioni

mxm ed $\boldsymbol{u}$ è il vettore, di dimensione m, degli errori nel modello autoregressivo SAR, con $u_i \sim N(0,\sigma_u^2)$.

- $\boldsymbol{e}$ è il vettore, di dimensione m, degli errori dovuti al campionamento. $\boldsymbol{e} \sim N(\boldsymbol{0},\boldsymbol{R})$ dove $\boldsymbol{R}$ è una matrice diagonale, con l'elemento $R_{ii} = \sigma_{e_i}^2$. La matrice $\boldsymbol{R}$ si considera nota.
- m è il numero delle piccole aree.

Nelle specifiche date si ipotizza un modello a livello di area; è possibile specificare un modello a livello di unità con pochi cambi di notazione: $\boldsymbol{y}$ ha dimensione nxp e y_{ij} indica l'osservazione j-esima nell'area i, $\boldsymbol{X}$ è una matrice nxp, e su ogni riga ci sono le covariate per ogni unità, $\boldsymbol{Z}$ è una matrice nxm, strutturata come una matrice diagonale a blocchi, dove ogni blocco ha dimensione pari alla numerosità delle unità appartenenti ad una piccola area[38] ed $\boldsymbol{e}$ è un vettore di dimensione n; n sono il numero totale di unità osservate in m piccole aree, per il resto rimane tutto invariato.
Usando le definizioni proposte la variabile di interesse $\boldsymbol{y}$ ha (cfr. appendice A pag. 144):

$$E[\boldsymbol{y}] = \boldsymbol{X\beta}$$

$$V[\boldsymbol{y}] = = \boldsymbol{Z}\left[\sigma_u^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}\right]\boldsymbol{Z}^T + \boldsymbol{R} = \boldsymbol{ZGZ}^T + \boldsymbol{R}$$

Dove per una notazione più agevole $\sigma_u^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1} = \boldsymbol{G}$. Riassumendo:

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZGZ}^T + \boldsymbol{R}) \qquad (3.10)$$

La (3.10) in notazione è identica alla già nota distribuzione di $\boldsymbol{y}$ nel modello lineare ad effetti misti. La differenza è nella componente di varianza $\boldsymbol{G}$, come abbiamo appena

[38] Ad esempio se ho 3 aree con 2 osservazioni nella prima, 4 osservazioni nella seconda e 3 osservazioni nella terza la matrice Z, identica a blocchi, è: $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$, di dimensione (2+4+3)x3.

dimostrato, che considera gli effetti casuali tra aree con un modello autoregressivo SAR.

Lo stimatore ottimo lineare e corretto (BLUP) per piccole aree, $\hat{\mu} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, può essere riscritto considerando $\boldsymbol{l}^T = \boldsymbol{x}_i$, $\boldsymbol{m}^T = \boldsymbol{b}_i^T$, $\boldsymbol{y} = \boldsymbol{\theta}$. Si considera inoltre la stima di $\boldsymbol{\beta}$ e si considerano note (per ora) σ_u^2 e ρ:

$$\hat{\mu} = t(\sigma_u^2, \rho) = \boldsymbol{x}_i \hat{\boldsymbol{\beta}} - \boldsymbol{b}_i^T \boldsymbol{G}\boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.11}$$

Dove $\boldsymbol{x}_i$ è il vettore delle p covariate nell'area i, $\boldsymbol{b}_i^T$ è un vettore di zeri con 1 alla i-esima posizione (è un vettore di selezione) e la stima di $\boldsymbol{\beta}$ si ottiene con i GLS: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \hat{\boldsymbol{\theta}}$. $\hat{\boldsymbol{\theta}}$ è il vettore delle osservazioni della variabile di interesse.

Utilizzando tutte le derivazioni ottenute per il BLUP si può calcolare il MSE, tenendo sempre in considerazione la diversa struttura della matrice $\boldsymbol{G}$:

$$MSE[t(\sigma_u^2, \rho)] = g_{1_i}(\sigma_u^2, \rho) + g_{2_i}(\sigma_u^2, \rho)$$

Sempre considerando le sostituzioni fatte le componenti risultano:

$$g_{1_i}(\sigma_u^2, \rho) = \boldsymbol{b}_i^T (\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \boldsymbol{Z}\boldsymbol{G}) \boldsymbol{b}_i$$

$$g_{2_i}(\sigma_u^2, \rho) = (\boldsymbol{x}_i - \boldsymbol{b}_i^T \boldsymbol{G}\boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \boldsymbol{X})(\boldsymbol{X}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \boldsymbol{X})^{-1} \cdot$$
$$\cdot (\boldsymbol{x}_i - \boldsymbol{b}_i^T \boldsymbol{G}\boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} \boldsymbol{X})^T$$

Dove considerando $\boldsymbol{V} = (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)$ e $\boldsymbol{d}^T = \boldsymbol{l}^T - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1}\boldsymbol{X} = \boldsymbol{x}_i - \boldsymbol{b}_i^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1}\boldsymbol{X}$ (fatte le sostituzioni sopra menzionate) i risultati presentati corrispondono a quelli ottenuti per il MSE[$t(\boldsymbol{\delta}, \boldsymbol{y})$] (il MSE dello stimatore BLUP).

Nella realtà i valori di ρ e σ_u^2 non sono noti, è necessario quindi farne una stima. Conoscendo la distribuzione di $\boldsymbol{\theta}$ (= $\boldsymbol{y}$) possiamo usare il metodo della massima verosimiglianza e calcolare l'EBLUP Spaziale; spaziale grazie alla struttura della matrice di varianza-covarianza $\boldsymbol{G}$. Lo stimatore Spatial EBLUP, che dipende dalle stime

di ρ e σ_u^2, nonché dalla stima $\hat{\boldsymbol{\beta}}$ (che otteniamo con i GLS una volta noti ρ e σ_u^2), è utilizzando la (3.11)

$$t(\hat{\sigma}_u^2, \hat{\rho}) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{b}_i^T\hat{\boldsymbol{G}}\boldsymbol{Z}^T(\boldsymbol{R} + \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^T)^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.12}$$

Dove $\hat{\boldsymbol{G}} = \hat{\sigma}_u^2[(\boldsymbol{I}_m - \hat{\rho}\boldsymbol{W})(\boldsymbol{I}_m - \hat{\rho}\boldsymbol{W}^T)]^{-1}$, e d'ora in avanti $\boldsymbol{V} = (\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)$.
Considerando la (3.10), e la distribuzione normale multivariata, la funzione di verosimiglianza è

$$l(\beta, \sigma_u^2, \rho) = 2\pi^{-\frac{m}{2}}|V|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{V})^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right]\right\}$$

e la funzione di log-verosimiglianza

$$ll(\beta, \sigma_u^2, \rho) = -\frac{m}{2}\ln 2\pi - \frac{1}{2}|V| - \frac{1}{2}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{V})^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\right]$$

Derivando rispetto ai parametri ρ e σ_u^2 ottengo:

$$\frac{\partial ll(\boldsymbol{\beta}, \sigma_u^2, \rho)}{\partial \sigma_u^2} = -\frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1})(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$\frac{\partial ll(\boldsymbol{\beta}, \sigma_u^2, \rho)}{\partial \rho} = -\frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\sigma_u^2[-\boldsymbol{C}^{-1}(2\rho\boldsymbol{W}\boldsymbol{W}^T - 2\boldsymbol{W})\boldsymbol{C}^{-1}]\boldsymbol{Z}^T) -$$

$$-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\sigma_u^2[-\boldsymbol{C}^{-1}(2\rho\boldsymbol{W}\boldsymbol{W}^T - 2\boldsymbol{W})\boldsymbol{C}^{-1}]\boldsymbol{Z}^T\boldsymbol{V}^{-1})(\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Dove per semplificare la notazione $\boldsymbol{C} = [(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]$. Attraverso le derivate parziale seconde della funzione di "meno log-verosimiglianza" ($-ll(\boldsymbol{\beta}, \sigma_u^2, \rho)$) si ottiene l'inversa della matrice di informazione (si procede esattamente come per l'EBLUP):

$$\iota(\sigma_u^2,\rho)=\begin{bmatrix}\frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T) & \frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T)\\ \frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T) & \frac{1}{2}tr(\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T)\end{bmatrix}$$

Con $\boldsymbol{A}=\sigma_u^2[-\boldsymbol{C}^{-1}(2\rho\boldsymbol{W}\boldsymbol{W}^T)\boldsymbol{C}^{-1}]$. Utilizzando l'algoritmo di *scoring* si ottengono le stime di massima verosimiglianza:

$$\sigma_u^{2(r+1)}=\sigma_u^{2(r)}+\left[\iota(\sigma_u^{2(r)},\rho^{(r)})\right]^{-1}d\left[\hat{\beta}(\sigma_u^{2(r)},\rho^{(r)}),\sigma_u^{2(r)},\rho^{(r)}\right]$$

$$\rho^{(r+1)}=\rho^{(r)}+\left[\iota(\sigma_u^{2(r)},\rho^{(r)})\right]^{-1}d\left[\hat{\beta}(\sigma_u^{2(r)},\rho^{(r)}),\sigma_u^{2(r)},\rho^{(r)}\right]$$

La struttura della matrice di varianza-covarianza è di tipo diagonale a blocchi, identica nella simbologia a quella proposta per l'EBLUP:

$$\bar{\boldsymbol{V}}(\hat{\boldsymbol{\beta}},\hat{\sigma}_u^2,\hat{\rho})=\begin{bmatrix}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} & 0\\ 0 & \iota^{-1}(\sigma_u^2,\rho)\end{bmatrix}$$

Dove $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$ è una matrice *pxp* diagonale a con le varianze dei coefficienti di regressione; $\iota^{-1}(\sigma_u^2,\rho)$ è la matrice di informazione dove sulla diagonale principale ci sono le varianze rispettivamente di σ_u^2 e ρ e sulla diagonale secondaria la covarianza tra σ_u^2 e ρ ($COV[\sigma_u^2,\rho] = COV[\rho,\sigma_u^2]$).

3.7 Il Mean Squared Error dello Stimatore Spatial EBLUP

Per misurare la variabilità dello stimatore Spatial EBLUP si utilizza il MSE. Continuando con l'estendere i risultati ottenuti per l'EBLUP il MSE per lo stimatore spaziale è espresso come:

$$MSE[t(\hat{\sigma}_u^2,\hat{\rho})]\simeq g_{1_i}(\sigma_u^2,\rho)+g_{2_i}(\sigma_u^2,\rho)+g_{3_i}(\sigma_u^2,\rho)$$

Dove la terza componente (Salvati, 2004):

$$g_{3_i}(\hat{\sigma}_u^2,\hat{\rho}) = tr\left\{\begin{bmatrix} \boldsymbol{b}_i^T(\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1})) \\ \boldsymbol{b}_i^T(\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1} + \hat{\sigma}_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1})) \end{bmatrix}\cdot\boldsymbol{V}\cdot\right.$$

$$\left.\cdot\begin{bmatrix} \boldsymbol{b}_i^T(\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1})) \\ \boldsymbol{b}_i^T(\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1} + \hat{\sigma}_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1})) \end{bmatrix}^T \bar{\boldsymbol{V}}(\hat{\boldsymbol{\beta}},\hat{\sigma}_u^2,\hat{\rho})\right\}$$

Una stima del $MSE[t(\hat{\sigma}_u^2,\hat{\rho})]$ può essere ottenuta sostituendo in ognuna delle tre componenti i valori sconosciuti σ_u^2 e ρ con la loro stima ottenuta con la massima verosimiglianza. Purtroppo, in questo modo il valore atteso del MSE non è corretto a causa della distorsione nella prima componente (come nell'EBLUP). Conoscendo l'entità della distorsione si ricava una stima corretta del MSE:

$$\widehat{MSE}[t(\hat{\sigma}_u^2,\hat{\rho})] = g_{1_i}(\hat{\sigma}_u^2,\hat{\rho}) - \boldsymbol{b}^T_{\hat{\sigma}_u^2,\hat{\rho}}(\hat{\sigma}_u^2,\hat{\rho})\nabla g_{1_i}(\hat{\sigma}_u^2,\hat{\rho}) + g_{2_i}(\hat{\sigma}_u^2,\hat{\rho}) + 2g_{3_i}(\hat{\sigma}_u^2,\hat{\rho})$$

Dove le stime $\hat{\sigma}_u^2,\hat{\rho}$ sono ottenute con il metodo della massima verosimiglianza e

$$\nabla g_{1_i}(\hat{\sigma}_u^2,\hat{\rho}) = \begin{bmatrix} \boldsymbol{b}_i^T(\boldsymbol{C}^{-1} - [\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\sigma_u^2\boldsymbol{C}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1})\boldsymbol{Z}\sigma_u^2\boldsymbol{C}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}])\boldsymbol{b}_i \\ \boldsymbol{b}_i^T(\boldsymbol{A} - [\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\sigma_u^2\boldsymbol{C}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1})\boldsymbol{Z}\sigma_u^2\boldsymbol{C}^{-1} + \sigma_u^2\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}])\boldsymbol{b}_i \end{bmatrix}$$

$$\boldsymbol{b}_{\hat{\sigma}_u^2,\hat{\rho}}(\hat{\sigma}_u^2,\hat{\rho}) = \frac{1}{2m}\left\{\iota^{-1}(\hat{\sigma}_u^2,\hat{\rho})\begin{bmatrix} tr[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1})\boldsymbol{X}] \\ tr[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(-\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T\boldsymbol{V}^{-1})\boldsymbol{X}] \end{bmatrix}\right\}$$

Per una trattazione più approfondita vedere Salvati (2004).

Abbiamo proposto un modello di stima libero dall'ipotesi dell'indipendenza tra aree. Ciò è stato possibile introducendo un modello autoregressivo spaziale basato su un processo stocastico Gaussiano di tipo SAR. Esistono altri processi stocastici Gaussiani dai quali è possibile ottenere uno stimatore spaziale, come ad esempio il processo Gaussiano di tipo CAR (Conditional Autoregressive Model). In questa tesi trattiamo solamente il metodo di stima basato sul processo Gaussiano SAR.

3.8 CONCLUSIONI

Nelle applicazioni reali conoscendo l'ubicazione delle piccole aree è possibile determinare la matrice di vicinanza $\boldsymbol{W}$. Con l'indice di Moran e/o di Geary si verifica se per una data variabile di interesse tra le aree c'è autocorrelazione spaziale. Se c'è autocorrelazione spaziale si utilizza lo stimatore Spatial EBLUP.

Come per il modello di stima EBLUP, per stimare il totale, o la media, di una piccola area per una data variabile bisogna conoscere tutte le unità della piccola aree e le relative covariate; generalmente tale richiesta è soddisfatta da indagini censuarie e da archivi di vario tipo (erariali, camere di commercio, etc.).

Utilizzando le osservazioni campionarie di una certa area i e il predittore ottimo lineare e corretto spaziale applicato alle unità non campionate possiamo stimare il totale, o la media, con la formula (2.47) già presentata nel capitolo precedente:

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j=1}^{n_i} y_j + \sum_{j=n_i+1}^{N_i} \boldsymbol{x}_j \hat{\boldsymbol{\beta}} + \boldsymbol{b}_i^T \hat{\boldsymbol{G}} \boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^T)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right) \tag{3.13}$$

dove le notazioni mantengono lo stesso significato e $\boldsymbol{Z}$ matrice di selezione per il modello a livello di unità (vedi nota 36 per un esempio).
Se le informazioni sono note solo a livello di area la (3.13) diventa semplicemente uguale a:

$$\hat{\bar{Y}}_i = \left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}} + \boldsymbol{b}_i^T \hat{\boldsymbol{G}} \boldsymbol{Z}^T (\boldsymbol{R} + \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}^T)^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \right)$$

Dove $\boldsymbol{x}_i$ è il vettore delle covariate dell'area i.

CAPITOLO 4

STIMATORE DIRETTO, EBLUP E SPATIAL EBLUP: UN CASO DI SIMULAZIONE

4.1 STUDIO DI SIMULAZIONE

Lo scopo della simulazione è quello di confrontare la funzionalità e l'efficienza degli stimatori presentati. Usando come termine di paragone lo stimatore diretto post-stratificato si esplora il comportamento degli stimatori EBLUP e Spatial EBLUP. Sulla base delle teorie presentate ci aspettiamo che le stime EBLUP e Spatial EBLUP siano migliori, in termini di efficienza, dello stimatore post-stratificato; è inoltre ragionevole pensare che in presenza di una popolazione correlata spazialmente le stime Spatial EBLUP abbiano performance migliori delle stime EBLUP.

Uno studio di simulazione nel campionamento da popolazioni finite si suddivide in 3 steps principali:

1. Generazione della popolazione con determinate caratteristiche, sia sistematiche che aleatorie.

2. Campionamento dalla popolazione e calcolo delle stime oggetto di studio per un numero finito di volte.

3. Confronto delle stime ottenute nei diversi campioni con indici opportuni.

Spesso negli studi di simulazione si generano più popolazioni con diverse caratteristiche con lo scopo di verificare il comportamento degli stimatori in "situazioni" differenti.

Nel nostro studio di simulazione abbiamo generato otto popolazioni diverse utilizzando le specifiche del modello lineare ad effetti misti e il modello autoregressivo spaziale SAR. Abbiamo creato una variabile $\boldsymbol{y} = [y_1, \ldots, y_N]^T$ per le unità della popolazione con il seguente modello:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{v} + \boldsymbol{h}$$

Dove $\boldsymbol{y}$ è il vettore della variabile di studio che riguarda tutte le unità della popolazione, di grandezza N e divisa in 42 piccole aree; $\boldsymbol{X}$ è la matrice dell'unica variabile ausiliaria utilizzata, di dimensione $Nx1$, generata casualmente utilizzando, per ogni elemento, una variabile casuale uniforme con parametri $1 \leq a \leq 100$ e $101 \leq b \leq 10000$[39], $\boldsymbol{\beta} = [0.2]$. Il vettore $\boldsymbol{v}$, di dimensione 42, rappresenta il modello autoregressivo spaziale SAR; $\boldsymbol{v}$ è stato generato da una variabile casuale multivariata distribuita normalmente con media 0 e matrice di varianza-covarianza $\sigma_u^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}$, con la matrice di connessione $\boldsymbol{W}$ riferita a 42 piccole aree, realmente esistenti, dove w_{ij} è uguale a 1 se le aree i e j sono contigue altrimenti $w_{ij} = 0$; $\sigma_u^2 = 100$ e ρ è stato posto uguale a 0.9, 0.75, 0.50, 0.25, 0, -0.25, -0.50, -0.75 ottenedo perciò otto diversi modelli autoregressivi spaziali e di conseguenza otto popolazioni diverse. $\boldsymbol{h}$ è un vettore di disturbi, di dimensione $Nx1$, generato moltiplicando elemento per elemento il vettore degli errori $\boldsymbol{e}$ e la radice quadrata della variabile ausiliaria: $h_k = e_k\sqrt{x_{k1}}$, dove gli elementi di $\boldsymbol{e}$ sono generati da una variabile casuale distribuita normalmente con media 0 e varianza $\sigma_e^2 = 1.34$, $\boldsymbol{e} \sim N(0,1.34)$. I parametri sono stati scelti seguendo l'esempio di Rao e Choudry (1995).

N è stato calcolato sommando la numerosità totale nelle 42 piccole aree. La popolazione di ogni piccola area è stata generata utilizzando una variabile casuale uniformemente distribuita con parametri a e b uguali rispettivamente a 100 e 350.

Nelle otto popolazioni generate la variabile di interesse per l'unità j-esima appartenente alla piccola area i è data da:

$$y_{ij} = 0.2x_{ix} + v_i + e_{ij}\sqrt{x_{ij}}$$

Da ogni popolazione ottenuta è stato estratto un campione di 800 unità garantendo un minimo di 2 unità per ognuna delle 42 piccole aree. Utilizzando il campione estratto è stata calcolata la stima diretta post-stratificata, la stima EBLUP e la stima Spatial EBLUP. Si è calcolata inoltre la stima del MSE per ogni stimatore; tale stima è stata ottenuta dalla tre componenti, g_1, g_2 e g_3, per gli stimatori EBLUP e Spatial EBLUP, ed

[39] Si ricorda che una variabile casuale uniforme continua definita sull'intervallo (a,b) ha funzione di densità $f(x) = \begin{cases} \frac{1}{b-a}, & \text{se } a < x < b \\ 0, & \text{altrimenti} \end{cases}$, con media attesa $\frac{a+b}{2}$ e varianza attesa $\frac{(a+b)^2}{12}$.

è stata calcolata una stima della distorsione della prima componente per rispettarne la correttezza.

Tale procedimento è stato ripetuto 1000 volte per ogni popolazione (con un tempo di elaborazione medio di circa 8290 secondi per popolazione).

Considerando una popolazione per volta, per confrontare i risultati ottenuti si utilizzano quattro indici, per ogni stimatore, riassuntivi dei 1000 campionamenti fatti:

- ARB, distorsione assoluta media relativa (Average Relative Bias)

$$ARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{Y}_{it}}{Y_i}-1\right)\right|$$

- EFF, efficienza media relativa (average relative EFFiciency)

$$EFF = \left(\frac{\overline{MSE}[\hat{Y}(pst)]}{\overline{MSE}[\hat{Y}]}\right)^{\frac{1}{2}}$$

- ARE, errore assoluto medio relativo (Average Relative Error)

$$ARE = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\left(\left|\frac{\hat{Y}_{it}}{Y_i}\right|-1\right)$$

- RRMSE, radice del mean squared error (errore quadratico medio) medio relativo (average Relative Root MSE)

$$RRMSE = \frac{1}{m}\sum_{i=1}^{m}\frac{\left(\overline{MSE}[\hat{Y}_i]\right)^{\frac{1}{2}}}{Y_i}$$

Dove $\overline{MSE}$, il MSE medio, è

$$\overline{MSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{Y}_{it} - Y_i\right)^2$$

m = 42 è il numero delle piccole aree, T = 1000 indica il numero di estrazioni campionarie fatte, $\hat{Y}_{it}$ è la stima per la piccola area i ottenuta dall'estrazione campionaria t-esima e Y_i è il vero valore della variabile nell'area i (che ovviamente è noto). Nella denominazione degli indici il termine relativo si riferisce al fatto che le stime sono confrontate con il vero valore della popolazione, confronto che è possibile effettuare, in genere, solo con una simulazione.

Per ottenere le stima EBLUP e Spatial EBLUP è necessaria la stima dei parametri σ_e^2, σ_u^2, ρ e $\boldsymbol{\beta}$. Nei metodi di stima presentati nei capitoli precedenti la varianza dovuta al campionamento σ_e^2 è stata considerata nota. Per simulare al meglio un caso reale non l'abbiamo considerata tale; come proxy per il valore da assegnare a σ_e^2 abbiamo usato la varianza campionaria della stima diretta:

$$\hat{\sigma}_{e_i}^2 = \left(1 - \frac{n_i}{N_i}\right)\frac{\sigma_{Y_i}^2}{N_i}$$

Dove $\hat{\sigma}_{e_i}^2$ è calcolata per l'area i, n_i è la numerosità campionaria dell'area i, N_i è la numerosità della popolazione nell'area i e $\sigma_{Y_i}^2$ è la varianza della popolazione nell'area i; quest'ultimo valore, che generalmente è incognito, lo consideriamo noto per non introdurre un ulteriore fattore di variabilità.

Le componenti $\boldsymbol{\beta}$, σ_u^2 e ρ si stimano con il metodo di massima verosimiglianza proposto nei capitoli precedenti.

Per la procedura di simulazione e per il calcolo degli stimatori è stato usato un programma scritto ad hoc in ambiente R[40].

4.1.1 I RISULTATI OTTENUTI

Si confrontano gli stimatori EBLUP e Spatial EBLUP con lo stimatore diretto. La stima diretta post-stratificata è stata così calcolata:

[40] R è un linguaggio di programmazione interpretato open source. E' possibile scaricarlo dal sito ufficiale http://www.r-ptoject.org.

$$\hat{\bar{Y}}(pst)_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

e per una stima del MSE:

$$\widehat{MSE}[\hat{Y}(pst)_i] = \sigma^2_{\hat{Y}(pst)_i} = \left(1 - \frac{n_i}{N_i}\right)\frac{\sigma^2_{y_i}}{n_i}$$

Dove $\sigma^2_{y_i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{\bar{Y}}(pst)_i\right)^2$ è la varianza campionaria.

Per ogni stimatore abbiamo calcolato l'intervallo di confidenza e controllato se il valore vero della popolazione vi è contenuto. Considerando uno stimatore generico $\hat{\theta}$ l'intervallo di confidenza con una significatività del 95% è:

$$\hat{\theta} \pm 1.96\sqrt{\widehat{MSE}(\hat{\theta})} \tag{4.1}$$

I risultati ottenuti per le otto popolazioni generati sono riportati nelle tabelle 1, 2 e 3. In tabella 1 sono riportati gli indici ARB, EFF, ARE e RRMSE per ognuno dei tre stimatori, mentre in tabella 2 troviamo la media delle stime del MSE per i tre stimatori, ottenuta con la media aritmetica semplice dei MSE stimati in ogni estrazione campionaria.

Osservando i risultati si nota come l'EBLUP e lo Spatial EBLUP ottengano performance migliori rispetto alla stima diretta in termini di EFF, ARE e RRMSE, indipendentemente dal coefficiente di auotocorrelazione spaziale della popolazione. In termini di efficienza (EFF) lo Spatial EBLUP risulta migliore dell'EBLUP per le popolazioni con coefficiente di autocorrelazione spaziale (ρ) diverso da zero. Questo è d'altronde il risultato che ci aspettavamo, lo Spatial EBLUP ottiene le performance migliori poiché considera il legame che c'è tra le aree vicine. L'indice EFF, molto intuitivo, considera come base di confronto l'efficienza dello stimatore post-stratificato, il metodo di stima classico. In termini percentuali l'EBLUP ha un indice di efficienza medio per tutte le popolazioni di 132.80% (con un errore standard di 13.61%) mentre lo Spatial EBLUP di 139.20% (errore standard 10.31%). Quindi, in media, lo stimatore EBLUP è del 30% più efficiente della stima post-stratificata mentre lo stimatore Spatial

EBLUP lo è del 40%. Questo significa che il MSE, che determina l'intervallo di confidenza, di EBLUP e Spatial EBLUP è sempre minore del MSE della stima post-stratificata; lo possiamo verificare anche dai risultati in tabella 4.2.

Qualunque sia ρ la stima dei MSE di EBLUP e Spatial EBLUP è minore di quella dello stimare post-stratificato. Dalla tabella 4.2 possiamo osservare che per valori assoluti crescenti di ρ la stima Spatial EBLUP ha un MSE stimato minore rispetto a quello dell'EBLUP. Per ρ uguale a 0 le stime del MSE di EBLUP e Spatial EBLUP sono praticamente identiche. Infatti se ρ è uguale a 0 il modello teorico dello Spatial EBLUP si riduce a quello dell'EBLUP, ci aspetteremmo, quindi, dei risultati identici; bisogna considerare invece che nello Spatial EBLUP la stima di ρ, quindi di un parametro in più rispetto all'EBLUP, comporta un aumento della variabilità nella stima della distorsione della prima componente g_1, che riscontriamo nella componente g_3 (che è appunto la stima della distorsione (bias) dovuta alla stima delle componenti di varianza nel MSE). Questo comportamento si verifica sempre, come è possibile vedere nella colonna "Media G3 stim." (g_3 stimato) della tabella 4.2. D'altronde è proprio grazie all'autocorrelazione spaziale che la componente g_1, che governa di fatto il MSE degli stimatori, della stima Spatial EBLUP è sempre più piccola rispetto a quella della stima EBLUP. Questo è il risultato che ci aspettavamo, infatti g_1 è la componente riferita agli effetti casuali di area che, grazie al modello autoregressivo spaziale SAR, sono stimati meglio dallo Spatial EBLUP. Non è presente in tabella 4.2 il termine di stima della distorsione della componente g_1 dovuto all'utilizzo della stima di massima verosimiglianza (presentato nei capitoli 2 e 3, rispettivamente per EBLUP e Spatial EBLUP); è stato ritenuto inopportuno introdurlo poiché non apporta informazioni ulteriori.

L'unico indice che risulta migliore per la stima post stratificata è l'ARB (Average Relative Bias), che misura la distorsione della media calcolata con uno stimatore rispetto al vero valore medio della popolazione. In media abbiamo un ARB di 0.44% (errore standard 0.17%) significativamente più piccolo rispetto ad un ARB di 6.59% (1.26%) per l'EBLUP e di 6.50% (0.84%) per lo Spatial EBLUP. La distorsione che otteniamo per EBLUP e Spatial EBLUP è figlia dell'introduzione del modello lineare ad effetti misti. Questo non ci deve preoccupare poiché non siamo interessati ad una stima puntuale, che ha probabilità infinitesimali, ma ad una stima per intervalli, che dipende dal MSE. Come abbiamo già visto il MSE è più piccolo per gli stimatori presentati nella tesi e come possiamo vedere dalla tabella 4.3 la copertura per una stima

con significatività 95% è pressoché perfetta. Quindi gli stimatori EBLUP e Spatial EBLUP sono più distorti rispetto allo stimatore post-stratificato ma il risultato finale di un'analisi è molto più preciso (l'intervallo di confidenza è più piccolo). La copertura è calcolata guardando per ogni estrazione campionaria e per ogni area la percentuale di volte che il valore vero cade nell'intervallo di confidenza, calcolato secondo la (4.1) con una significatività del 95%. Significatività 95%, si ricorda, vuol dire che estraendo infiniti campioni da una popolazione, nel 95% dei casi la media campionaria ha un valore compreso nell'intervallo di confidenza. Nel nostro caso, come si vede dalla tabella 3, con 1000 campioni per popolazione ci avviciniamo molto al risultato limite.
Focalizzando i risultati messi in luce dalla simulazione risulta che:

- Lo stimatore Spatial EBLUP è più efficiente dello stimatore EBLUP soprattutto all'aumentare del valore assoluto del coefficiente di autocorrelazione spaziale (aspetto sottolineato dall'indice "EFF", tabella 4.1).

- La stima del MSE dello Spatial EBLUP è sempre minore rispetto a quella dell'EBLUP; ciò implica che le stime dello Spatial EBLUP sono le più precise (aspetto sottolineato dall'indice "Media MSE Stim.", tabella 4.2)

- Gli stimatori EBLUP e Spatial EBLUP sono sempre preferibili rispetto allo stimatore post-stratificato. Lo stimatore EBLUP è preferibile allo Spatial EBLUP in caso di autocorrelazione spaziale prossima a 0 (aspetti denotati dall'insieme dei risultati ottenuti).

Insieme ai risultati che abbiamo commentato presentiamo la tabella 4.4 che contiene i dati relativi alla struttura in media delle popolazioni generate nelle 42 aree e degli 8000 campioni estratti, 1000 per popolazione.

TABELLA 4.1. Confronto degli stimatori per piccola area, indici: ARB, EFF, ARE e RRMSE.

ρ	INDICE	$\hat{Y}$ (pst)	EBLUP	SEBLUP
0,9	ARB	0,89%	9,57%	8,19%
0,9	EFF	100,00%	108,27%	122,59%
0,9	ARE	27,94%	27,31%	24,68%
0,9	RRMSE	18,28%	17,39%	13,94%
0,75	ARB	0,41%	5,44%	6,85%
0,75	EFF	100,00%	119,01%	145,85%
0,75	ARE	16,09%	15,20%	14,23%
0,75	RRMSE	20,35%	19,10%	17,82%
0,5	ARB	0,31%	6,00%	5,58%
0,5	EFF	100,00%	145,77%	152,05%
0,5	ARE	11,88%	9,39%	9,00%
0,5	RRMSE	14,98%	11,30%	10,89%
0,25	ARB	0,42%	6,12%	5,98%
0,25	EFF	100,00%	132,60%	135,71%
0,25	ARE	11,96%	10,33%	10,24%
0,25	RRMSE	15,09%	12,59%	12,50%
0	ARB	0,34%	5,60%	5,68%
0	EFF	100,00%	154,79%	153,45%
0	ARE	11,62%	8,53%	8,70%
0	RRMSE	14,64%	10,37%	10,60%
-0,25	ARB	0,34%	6,90%	6,68%
-0,25	EFF	100,00%	134,57%	136,77%
-0,25	ARE	13,15%	11,41%	11,35%
-0,25	RRMSE	16,59%	13,87%	13,85%
-0,5	ARB	0,39%	5,93%	5,90%
-0,5	EFF	100,00%	129,62%	127,20%
-0,5	ARE	11,46%	10,01%	10,15%
-0,5	RRMSE	14,51%	12,08%	12,32%
-0,75	ARB	0,43%	7,16%	7,16%
-0,75	EFF	100,00%	137,80%	140,00%
-0,75	ARE	14,19%	12,87%	12,73%
-0,75	RRMSE	17,95%	15,90%	15,72%

TABELLA 4.2. Comparativa degli stimatori per piccola area, indici: media della Stima del MSE e valore medio della stima delle sue tre componeti[41].

P	STIMATORE	$\widehat{MSE}$ medio	$\hat{g}_1$ medio	$\hat{g}_2$ medio	$\hat{g}_3$ medio
0,9	$\hat{Y}$ (pst)	75,86	0,00	0,00	0,00
0,9	EBLUP	64,71	59,59	0,77	0,55
0,9	SEBLUP	50,48	41,46	1,87	0,99
0,75	$\hat{Y}$ (pst)	76,11	0,00	0,00	0,00
0,75	EBLUP	53,74	51,94	1,12	0,62
0,75	SEBLUP	35,78	36,79	1,86	1,01
0,5	$\hat{Y}$ (pst)	102,58	0,00	0,00	0,00
0,5	EBLUP	48,27	36,95	2,52	0,95
0,5	SEBLUP	44,37	34,12	2,28	2,19
0,25	$\hat{Y}$ (pst)	90,81	0,00	0,00	0,00
0,25	EBLUP	51,65	43,73	1,78	0,97
0,25	SEBLUP	49,31	42,16	1,52	2,21
0	$\hat{Y}$ (pst)	86,48	0,00	0,00	0,00
0	EBLUP	36,09	33,96	2,10	1,18
0	SEBLUP	36,73	32,36	2,45	3,09
-0,25	$\hat{Y}$ (pst)	87,16	0,00	0,00	0,00
-0,25	EBLUP	48,13	37,07	1,94	0,86
-0,25	SEBLUP	46,59	36,50	2,22	2,60
-0,5	$\hat{Y}$ (pst)	62,18	0,00	0,00	0,00
-0,5	EBLUP	37,01	29,63	1,47	0,70
-0,5	SEBLUP	38,43	28,41	1,67	1,86
-0,75	$\hat{Y}$ (pst)	91,97	0,00	0,00	0,00
-0,75	EBLUP	48,43	47,60	1,78	0,95
-0,75	SEBLUP	46,92	44,21	2,52	2,67

[41] La somma delle tre componenti non è uguale al MSE perché non è stata introdotta nella tabelle l'ulteriore termine di distorsione della componente g_1.

TABELLA 4.3. Percentuale di copertura ottenuta considerando un intervallo di confidenza del 95%.

ρ	STIMATORE	Copertura 95%
0,9	$\hat{Y}$ (pst)	95%
0,9	EBLUP	95%
0,9	SEBLUP	95%
0,75	$\hat{Y}$ (pst)	95%
0,75	EBLUP	95%
0,75	SEBLUP	95%
0,5	$\hat{Y}$ (pst)	94%
0,5	EBLUP	95%
0,5	SEBLUP	95%
0,25	$\hat{Y}$ (pst)	95%
0,25	EBLUP	95%
0,25	SEBLUP	95%
0	$\hat{Y}$ (pst)	93%
0	EBLUP	94%
0	SEBLUP	94%
-0,25	$\hat{Y}$ (pst)	93%
-0,25	EBLUP	93%
-0,25	SEBLUP	94%
-0,5	$\hat{Y}$ (pst)	95%
-0,5	EBLUP	95%
-0,5	SEBLUP	95%
-0,75	$\hat{Y}$ (pst)	95%
-0,75	EBLUP	95%
-0,75	SEBLUP	95%

TABELLA 4.4. Popolosità media delle piccole aree per le 8 popolazioni generate, dimensione campionaria media estratta da ogni piccola area e sua percentuale rispetto alla popolazione di area.

N. Area	*Numerosità Media Popolazione*	*Numerosità Media Campionaria*	*%*
1	216	18	8,47%
2	239	20	8,49%
3	203	17	8,36%
4	217	18	8,31%
5	225	19	8,49%
6	236	20	8,34%
7	234	19	8,33%
8	236	20	8,37%
9	232	19	8,33%
10	234	20	8,46%
11	250	21	8,32%
12	222	19	8,35%
13	223	19	8,46%
14	213	18	8,44%
15	222	19	8,39%
16	225	19	8,37%
17	238	20	8,38%
18	213	18	8,46%
19	222	19	8,38%
20	224	19	8,40%
21	234	20	8,42%
22	229	19	8,37%
23	228	19	8,49%
24	243	21	8,43%
25	223	19	8,41%
26	227	19	8,36%
27	232	19	8,36%
28	241	20	8,45%
29	245	20	8,35%
30	236	20	8,39%
31	231	20	8,47%
32	213	18	8,36%
33	245	20	8,31%
34	232	19	8,25%
35	212	18	8,31%
36	229	19	8,40%
37	212	18	8,34%
38	228	19	8,35%
39	240	20	8,41%
40	199	17	8,42%
41	211	18	8,38%
42	226	19	8,35%
Totale	***9538***	***800***	***8,39%***

4.2 CONCLUSIONI

Lo studio di simulazione è risultato molto utile: esso conferma i risultati attesi in base alle teorie elaborate. E' uno studio di simulazione di carattere generale dove non è stato approfondito un aspetto specifico ma si è valutato l'insieme delle prestazioni degli stimatori. L'uso di diverse matrici di connessione, una numero maggiore o minore di piccole aree e l'impiego di più covariate possono cambiare i risultati ottenuti. Tuttavia alcuni di questi aspetti non sono stati ancora approfonditi nella letteratura specifica della stima per piccole aree.

Si sottolinea che uno studio di simulazione è solo il primo passo per verificare le performance di uno stimatore che devono essere riscontrate anche nell'utilizzo pratico, cioè su dati reali.

CAPITOLO 5

INDAGINE STRUTTURA E PRODUZIONE DELLE AZIENDE AGRICOLE (SPA)

5.1 INTRODUZIONE

In questo capitolo presenteremo le principali fonti statistiche in ambito agricolo in Italia. Cercheremo, tramite le fonti principali, di delineare la situazione attuale dell'agricoltura italiana, ponendo l'attenzione soprattutto sull'aspetto quantitativo.

Verrà posta una particolare attenzione sull'indagine SPA (Struttura e Produzione delle aziende Agricole), svolta dall'ISTAT, di cui presenteremo i dati raccolti.

Sui dati dell'indagine SPA verrà proposta una stima per piccole aree, sia di tipo EBLUP sia Spatial EBLUP, dove le piccole aree sono rappresentate dalle Superfici Economiche Locali (SEL) della regione Toscana.

5.2 LE PRINCIPALI FONTI STATISTICHE IN AGRICOLTURA

La fonte più completa a disposizione in Italia sono i Censimenti. L'ISTAT (Istituto Nazionale di Statistica) ha sempre raccolto dati a livello censuario solo per ciò che riguarda la popolazione, con il noto Censimento, di cadenza decennale, "Popolazione e Abitazioni". Nel primo dopoguerra si è resa necessaria una raccolta di dati su uno spettro più ampio di argomenti come l'industria e l'agricoltura. Oggi l'ISTAT raccoglie dati a livello censuario nel settore industriale e terziario con il censimento "Industria e Servizi" (giunto sino all'ottava edizione) e nel settore primario, l'agricoltura, con il "Censimento sull'Agricoltura" (giunto alla quinta edizione). Aggiunge a questi il censimento "Istituzioni Private e Non-Profit", fatto nel 2001.

I Censimenti richiedono però grandi investimenti dal punto di vista economico e temporale nonché lunghi tempi di attesa per la disponibilità dei dati. Per questo motivo l'ISTAT ed altri enti pubblici effettuano indagini campionarie con cadenze più frequenti rispetto al censimento.

Presentiamo una panoramica sulle principali indagini in campo agricolo e rimandiamo all'appendice B per un approfondimento sul tema.

Le indagini principali in ambito agricolo sono l'indagine sulla struttura e produzione delle aziende agricole (vedi paragrafo 5.2.1) e l'indagine RICA-REA: La RICA-REA

nasce nel 2002-2003 dall'unione di due indagini campionarie, RICA e REA: la REA (Risultato Economico delle Aziende Agricole) è condotta dall'ISTAT in collaborazione con l'INEA, gli Assessorati all'Agricoltura e gli Uffici di Statistica delle Regioni e Province Autonome con l'obiettivo di ottenere informazioni microeconomiche (a livello di ogni singola azienda agricola) sui risultati economici delle aziende nell'anno solare di riferimento; la RICA, acronimo di Rete d'Informazione Contabile Agricola, è uno strumento informativo finalizzato alla conoscenza della condizione economica delle aziende agricole europee.

L'INEA (Istituto Nazionale Economia Agraria) svolge attività di ricerca, di rilevazione, analisi e previsione nel campo strutturale e socio economico del settore agro-industriale, forestale e della pesca producendo informazioni importanti sui vari aspetti del settore agricolo.

Altre fonti che forniscono dati sull'agricoltura sono l'Eurostat, il Registro delle Imprese, gli Enti Regionali, l'ENARPRI (European Network of Agricultural and Rural Policy Research Institutes) e la Banca d'Italia (limitatamente alle informazioni di carattere economico).

5.2.1 L'INDAGINE STRUTTURA E PRODUZIONE DELLE AZIENDE AGRICOLE

Nella nuova strategia di rilevazione, il "Sistema delle Statistiche Agricole", l'indagine sulla Struttura e Produzione delle Aziende Agricole (SPA) gioca un ruolo centrale. Come abbiamo accennato, il suo ruolo è quello di indagare approfonditamente sulle attività svolte dalle aziende agricole. L'indagine SPA inizia nel 1995, con cadenza annuale; l'ultima indagine SPA disponibile è quella del 2003, svolta nel periodo 15 ottobre 15 dicembre dall'ISTAT, dalle Regioni e dalle Province autonome. La rilevazione viene svolta anche in attuazione della Regolamento CE (Comunità Europea) relativo alla organizzazione delle indagini comunitarie sulla struttura delle aziende agricole.

In base alla definizione ISTAT l'azienda agricola è "una unità tecnico economica costituita da terreni, anche in appezzamenti non contigui, ed eventualmente da impianti ed attrezzature varie, in cui si attua la produzione agraria, forestale e zootecnica ad opera di un conduttore, cioè persona fisica, società od ente che ne sopporta il rischio sia da solo (conduttore coltivatore e conduttore con salariati e/o compartecipanti), sia in associazione ad un mezzadro o colono parziario".

Il piano di campionamento adottato per l'indagine SPA 2003 è del tipo ad uno stadio stratificato con inclusione certa delle aziende di maggior dimensione. La dimensione complessiva del campione è di 55.030 aziende, selezionate tra le circa 2.590.000 aziende agricole italiane[42]. Vengono considerate, in questa indagine, quelle aziende agricole che:

1. hanno una superficie agricola utilizzata (SAU) uguale o superiore ad un ettaro (SAU $\geq$ 1 ettaro).

2. hanno una superficie agricola utilizzata inferiore ad un ettaro, qualora esse producano in una determinata misura per la vendita, oppure qualora la loro unità di produzione oltrepassi determinati limiti fisici.

La stratificazione delle aziende avviene in due fasi.
Nella prima fase sono state individuate le aziende autorappresentative (quelle aziende che hanno probabilità di inclusione nel campione uguale a 1) sulla base della loro dimensione economica e/o della loro superficie agricola utilizzata e/o del numero di capi animali espressi in termini di UBA (UBA è acronimo di Unità di Bovino Adulto, questa variabile è ottenuta come combinazione lineare del numero di capi animali presenti in ciascuna azienda; i coefficienti utilizzati per ottenere il numero di UBA in ciascuna azienda sono i seguenti: 0.8 per Bovini e bufalini, 0.14 per ovini e caprini, 0.6 equini, 0.27 per suini, 0.014 per avicoli, 0.028 per conigli, 0.2 per struzzi). Complessivamente sono state individuate 6.972 unità autorappresentative.
Nella seconda fase le aziende sono state suddivise in 407 strati utilizzando criteri geografici, dimensionali (espressi in termini di superficie agricola utilizzata (SAU) e/o numero di capi animali (UBA) e/o reddito lordo standard (RLS)) e tipologici. La numerosità campionaria per regione è riportata nella seguente tabella:

[42] In realtà l'universo delle aziende agricole è diviso in due sotto-universi: le aziende agricole che rispondono alla definizione adottata dalla CEE e le restanti (cioè aziende agricole considerate tali in Italia ma non rispondenti alla definizione CEE). Poiché le aziende che non appartengono all'universo CEE rappresentano meno dell'1% del totale delle aziende è stato deciso di trascurare questa distinzione.

TABELLA 5.1. Dimensione del campione per regione nell'indagine SPA 2003 (Fonte ISTAT).

Regioni e Province autonome	***Dimensione del campione***
Piemonte	3.700
Valle D'Aosta	430
Lombardia	5.300
Bolzano-Bozen	700
Trento	700
Veneto	4.000
Friuli -V. G.	1.300
Liguria	1.350
Emilia -Romagna	3.250
Toscana	3.000
Umbria	1.500
Marche	1.350
Lazio	3.850
Abruzzo	1.350
Molise	1.250
Campania	3.300
Puglia	6.700
Basilicata	1.100
Calabria	3.000
Sicilia	5.150
Sardegna	2.750
TOTALE	***55030***

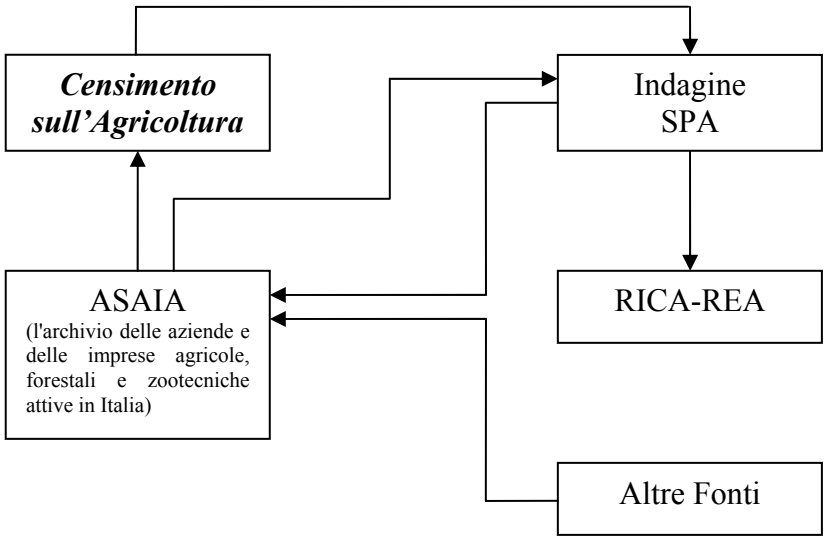
Per quanto riguarda la Toscana la stratificazione è avvenuta in base alle province e alla classe di UDE (Unità di Dimensione Europea, pari a 1200€ per ciascuna unità di Reddito Lordo Standard): classe 1 = UDE $\in$ [1-2], classe 2 = UDE $\in$ (2-30] e classe 3 = UDE $\in$ (30 e oltre). In questo modo si sono creati 25 strati.

L'indagine viene svolta tramite questionario con intervista diretta da. Con il questionario (intervista del conduttore aziendale) si rilevano dati sulle diverse coltivazioni e sugli allevamenti, nonché informazioni strutturali sulla forma organizzativa, sulla manodopera impiegata, sui rapporti dell'azienda con il mercato, sulle pratiche ambientali e sulle eventuali attività extragricole condotte in azienda (agriturismo, trasformazione dei prodotti agricoli, ecc.). Quest'ultime ricoprono un ruolo di interesse crescente, infatti è sempre maggiore il numero di aziende agricole che opera in attività "collaterali", tali aziende sono denominate "aziende agricole multifunzionali".

L'indagine è utile anche per soddisfare alcune esigenze di Contabilità Nazionale (conti economici nazionali e conto satellite dell'agricoltura) e per raccogliere informazioni aggiornate per la programmazione e la sorveglianza delle diverse attività connesse alle politiche agricole regionali.

5.2.2 IL RUOLO DELL'INDAGINE SPA NEL SISTEMA DELLE RILEVAZIONI IN AGRICOLTURA

Negli anni seguenti al 1990 c'è stato un riassetto delle statistiche in campo agricolo[43] che ha portato ad avere nel 2003 tre punti fermi: il Censimento sull'Agricoltura, l'indagine sulla Struttura e la Produzione delle Aziende Agricole e l'indagine RICA-REA. Si è instaurato un meccanismo che può essere sintetizzato con il seguente schema:



Con l'incrocio di diverse fonti statistiche è stato creato l'ASAIA (vedi appendice B pag. 144), un elenco di tutte le aziende agricole presenti sul territorio nazionale. Nel 2000 tutte le aziende presenti nel database ASAIA sono state censite. Da questa popolazione di aziende è stato estratto il campione per l'indagine sulla Struttura e la Produzione delle Aziende Agricole (SPA) con il fine di integrare i dati censuari e tenere aggiornato l'elenco ASAIA. Dal campione dell'indagine SPA si estrae un sottocampione per l'indagine RICA-REA. L'indagine SPA serve per approfondire gli aspetti produttivi, l'indagine RICA-REA approfondisce gli aspetti economico-gestionali.

[43] Si specifica che non solo nelle statistiche agricole ci sono state delle riforme, tutto l'assetto delle statistiche istituzionali è stato riorganizzato per una maggiore efficienza e precisione.

L'elenco ASAIA è aggiornato anche sulla base delle altre fonti statiche presenti sul territorio.

5.3 La Situazione Attuale dell'Agricoltura in Italia

5.3.1 Il Settore Agricolo Secondo i Dati Censuari

Una prima analisi emerge confrontando i dati del Censimento sull'Agricoltura del 1990 con quello del 2000. Alla data di riferimento del Censimento (22 ottobre 2000) sono state rilevate in Italia 2.593.090 aziende agricole (considerando anche quelle zootecniche e forestali), con superficie totale pari a 19,6 milioni di ettari, di cui 13,2 milioni di superficie agricola utilizzata (SAU). Rispetto al Censimento del 1990, il numero delle aziende è nel complesso diminuito di 430 mila unità (-14,2%), a fronte di una riduzione più contenuta della superficie totale per 3,1 milioni di ettari (-13,6%), di cui 1,8 milioni di SAU (-12,2%).

Da un punto di vista strutturale la distribuzione delle aziende e delle relative superfici per classi di estensione mostra come nel settore agricolo risulti ancora dominante la presenza di micro-aziende o di aziende nelle quali la SAU ricopre una parte esigua della superficie totale aziendale. Infatti, tenuto conto che le aziende senza SAU sono pari all'1,6% del numero complessivamente censito, sono 1.163.793 (pari a circa il 45% del totale) le aziende che hanno meno di un ettaro di SAU, con un grado di copertura pari appena al 4,8% della superficie totale agricola italiana e al 3,9% della SAU complessivamente rilevata. In numero contenuto sono, invece, le aziende con almeno 20 ettari che tuttavia, pur rappresentando solo il 4,6% del totale, coprono il 55,3% della superficie totale e il 54,8% della SAU. Emerge da questi dati una concentrazione molto accentuata della distribuzione della superficie agricola.

Le attività più diffuse tra le aziende agricole sono le coltivazioni legnose agrarie, che sono presenti nel 71,7% del totale, dedite prevalentemente alla olivicoltura (1,2 milioni di aziende), alla viticoltura (790 mila aziende), ma anche alla frutticoltura e agrumicoltura (circa 650 mila aziende). La relativa superficie investita rappresenta il 18,6% della SAU e il 12,5% della superficie totale, anche in questo caso con prevalenza delle superfici di olivo (8,2% della SAU) e di vite (5,4% della SAU). Tra il 1990 e il 2000 ha preso forma un fenomeno di trasformazione delle coltivazioni di vite. Infatti è aumentata la produzione di uva destinata a produrre vini DOC e DOCG mentre è

drasticamente diminuita la coltivazione di vite destinata alla produzione di altri vini (-36% delle aziende e -34.2% delle supefici).
Particolarmente diffusa è la coltivazione dei seminativi, che sono presenti nel 59,9% delle aziende e coprono il 55,6% della SAU e il 37,4% della superficie totale delle aziende.
Abbastanza diffusa tra le aziende agricole è la presenza di boschi, che sono presenti nel 23.3% di esse. Tale superficie boschiva copre il 23.2% della superficie agricola totale, valore che segna una netta diminuzione rispetto al 1990 (-17.5%). Tuttavia tale valore non deve allarmare poiché il dato non è omogeneo (alcune grosse aziende agricole statali sono state nel corso degli anno trasformate in parchi).

Scendendo a livello regionale sono state rilevate in Toscana 139.872 aziende agricole (comprese quelle zootecniche e forestali), la cui superficie totale assomma a 1.627.461 ettari, di cui 857.699 di superficie agricola utilizzata (SAU). Rispetto al Censimento del 1990, il numero delle aziende è diminuito del 6,6% (pari a 9.869 unità), a fronte di una riduzione dell'8,4% della superficie totale (pari a 149.102 ettari) e del 7,5% della superficie agricola utilizzata (pari a 69.870 ettari).
La distribuzione delle aziende per classi di superficie agricola utilizzata (SAU) mostra come il settore agricolo sia tuttora caratterizzato dalla massiccia presenza di micro-aziende. Sono infatti 63.544 (pari al 45,4% del totale) le aziende con meno di 1 ettaro di SAU, le quali coprono soltanto il 4,7% della superficie totale e il 3,1% della SAU complessivamente rilevate nella regione. Se si considerano tutte le aziende con meno di 5 ettari, la quota sale al 76,9% del totale regionale, cui corrispondono quote del 16,3% della superficie totale e di appena il 14,7% della SAU. Le aziende con oltre 20 ettari di SAU sono 8.589 e, pur rappresentando solo il 6,1% del totale, coprono il 61,3% della superficie totale e il 63,8% della SAU. Anche in Toscana, come in Italia, c'è un'alta concentrazione nella distribuzione del territorio.
Considerando il titolo di possesso dei terreni, continuano ad essere ampiamente prevalenti le aziende che hanno solo terreni di proprietà (88,1%). Il loro numero è però diminuito, fra il 1990 e il 2000, del 9,7%, contrariamente a quanto è avvenuto per le aziende con terreni solo in affitto (il cui numero è rimasto pressoché invariato) e per quelle con terreni parte in proprietà e parte in affitto, che sono aumentate del 50,8%, con analogo incremento percentuale della superficie agricola utilizzata.

Spostando l'attenzione sulle coltivazioni, la coltura più importante, in termini di superficie investita, è quella dei seminativi, praticata dal 67,9% delle aziende. I seminativi coprono il 63,0% della SAU e il 33,2% della superficie totale delle aziende. Rispetto al 1990, tuttavia, il numero delle aziende coltivatrici è diminuito del 16,7% (dunque ben più di quello delle aziende in complesso, diminuite del 6,6%), mentre la superficie dei seminativi si è ridotta in misura molto minore (-5,2%), in questo modo il suo valore medio è aumentato da 5,00 a 5,69 ettari per azienda coltivatrice. Gioca un ruolo importante anche la coltivazione delle legnose agrarie, praticate dal 75,7% delle aziende, dedite prevalentemente alla coltura dell'olivo, della vite e dei fruttiferi. La relativa superficie investita rappresenta il 21,4% della SAU e l'11,3% della superficie totale delle aziende. Rispetto al 1990 il numero delle aziende che praticano questo tipo di coltivazioni è diminuito sensibilmente (-8,1%), mentre assai più contenuta è stata la diminuzione della relativa superficie investita (-4,8%), il cui valore medio è aumentato da 1,67 a 1,73 ettari per azienda coltivatrice. In particolare, sono aumentati notevolmente sia il numero delle aziende olivicole (+11,9%), sia l'estensione della superficie investita a olivo (+9,1%), che copre l'11,3% della SAU e il 6,0% della superficie totale delle aziende. Non si registrano, invece, variazioni significative nel valore medio della superficie investita a olivo per azienda coltivatrice, che rimane piuttosto basso (1,23 ettari). Per quanto riguarda la coltivazione di vite, la superficie investita copre il 6,8% della SAU e il 3,6% della superficie totale delle aziende ed è diminuita del 17,3% rispetto al 1990. Tale diminuzione, però, non riguarda le produzioni di vite di alta qualità che sono in netta espansione: la vite per la produzione di vini DOC e DOCG, infatti, segna un incremento del 49,8% in termini di aziende coltivatrici e del 21,6% in termini di superficie investita, mentre diminuisce del 44,0% la superficie investita nella produzione di altri vini, con una riduzione del 32,9% delle aziende coltivatrici. Questo è determinato sia da un aumento delle viti considerate DOC e DOCG, sia da un aumento della competitività determinata dalla produzione di alta qualità e dall'aumento di produzioni inimitabili. Tra i due censimenti, la superficie investita a fruttiferi è quasi triplicata, con un aumento del 42,8% delle aziende coltivatrici.

Le colture boschive conservano un peso molto rilevante sulla superficie totale delle aziende (40,2%), nonostante abbiano subito, rispetto al 1990, una sensibile contrazione della superficie investita (-7,5%). In realtà, l'entità della riduzione delle superfici boschive è amplificata dall'uscita dal campo di osservazione del Censimento di alcune

grandi aziende forestali pubbliche, convertite nel corso degli anni Novanta in aree protette e in quanto tali non più rilevate come aziende silvicole. L'incidenza delle colture boschive è particolarmente alta (89,6%) nelle aziende senza SAU, che sono, in questa regione, prevalentemente forestali. Incidenze superiori alla media regionale si osservano inoltre nelle aziende più grandi (oltre 100 ettari di SAU) e in quelle più piccole (fino a 1 ettaro di SAU), dove le colture boschive coprono rispettivamente il 44,4 e il 49,8% della superficie totale delle aziende.

5.3.2 Il Settore Agricolo nel 2003

Dopo aver visto a grandi linee l'andamento nel settore agricolo rispetto al decennio intercensuario passiamo ad una panoramica sull'annata agricola 2003.

Un fattore determinante nell'agricoltura è il clima. Il settore nel 2003 è stato negativamente influenzato da questo fattore, dal punto di vista meteorologico, infatti, il 2003 è stato tra gli anni peggiori dell'ultimo decennio. Nel 2003 l'agricoltura italiana ha fatto registrare un pesante calo sia della produzione (-4,7%) sia del valore aggiunto (-5,7%), in termini reali, confermando purtroppo un trend negativo iniziato dal 2000. Per compensare l'effetto negativo della congiuntura climatica non sono stati sufficienti, né il favorevole andamento dei prezzi registrato per il 2003, né la riduzione, oramai strutturale, dei costi di produzione (-10% nell'ultimo decennio). Alla luce di ciò, l'immagine che se ne trae è quella di un settore che, sebbene evidenzi al suo interno interessanti dinamiche evolutive di segno positivo, nel suo insieme non appare, tuttavia, in grado di fare fronte a fatti contingenti negativi.

A livello di singoli settori produttivi, le flessioni più rilevanti hanno interessato, tra gli altri, i comparti delle colture industriali e dei seminativi.

Il settore è fortemente sostenuto dall'intervento pubblico con una spesa di 18 miliardi di euro, pari al 58,5% del valore aggiunto del settore. Questo intervento pesa per il 32.6% sull'AGEA (AGenzia per le Erogazioni in Agricoltura), per il 24.8% sulle Regioni, per il 3.5% sul Ministero delle Politiche Agricole e Forestali ed il restante 39.1% è suddiviso tra altri enti ed agenzie pubbliche che sostengono il settore. A sua volta l'intervento pubblico è stato fortemente condizionato dai finanziamenti europei. Questo comporta una forte dipendenza dell'economia primaria dagli interventi UE che stanno subendo in questi ultimi anni un totale riassetto, dovuto in parte all'ingresso di nuovi paesi nell'unione.

Nel 2003 la produzione agricola, inclusa la silvicoltura e la pesca, è lievemente aumentata in valore rispetto al 2002 (0,9%), in termini monetari. Il risultato è la sintesi di una diminuzione delle quantità prodotte del 4,4% e di un incremento dei prezzi del 5,5%. Riportiamo nella tabella 5.2 la produzione delle principali coltivazioni del comparto vegetale:

TABELLA 5.2. Principali produzioni vegetali (Fonte INEA).

		Quantità		***Valore***	
		x1000 t.	***Var % 02-03***	***Mln. Euro***	***Var % 02-03***
SEMINATIVI (Coltivazioni principali)	*Frumento tenero*	2.517	-23,2	671	-17,1
	Frumento duro	3.727	-12,7	1.109	-9,3
	Mais	8.985	-14,9	1.841	-9,5
	Riso	1.360	-0,8	438	-8,6
	Barbabietola da zucchero	7.137	-43,9	338	-20,8
	Tabacco	124	-1,5	370	3,1
	Soia	425	-25	176	-10,8
	Girasole	242	-30,8	83	-31,1
	Patate	1.604	-7,3	555	-10,5
	Pomodori	6.634	15,4	1.206	24
COLTIVAZIONI LEGNOSE AGRPARIE (Coltivazioni principali)	*Uva da tavola*	1.176	3,2	562	7,5
	Uva da vino	3.537	1,2	994	4,5
	Vino (x1000 hl)	18.937	-1,7	1.993	2,9
	Olive vendute	294	-10,2	157	-8,5
	Olio	484	-7,4	1.946	-4,9
	Mele	1.947	-11,5	722	-9,2
	Pere	822	-11	411	-8,5
	Pesche e nettarine	1.357	-14,7	636	-1,6
	Arance	1.962	13,8	667	19,9
	Limoni	549	12,8	277	24,6
	Mandarini e clementine	589	2,5	266	7,1
	Actinidia	365	-3,9	272	1,9

Dove i valori sono espressi in termini di prezzi base (base 2000).

I risultati sul complesso della produzione agricola sono sintetizzati, per comparto, nella tabella 5.3, espressa in termini di prezzi base (base 2000):

TABELLA 5.3. Produzione per comparti, anno 2003 (Fonte INEA).

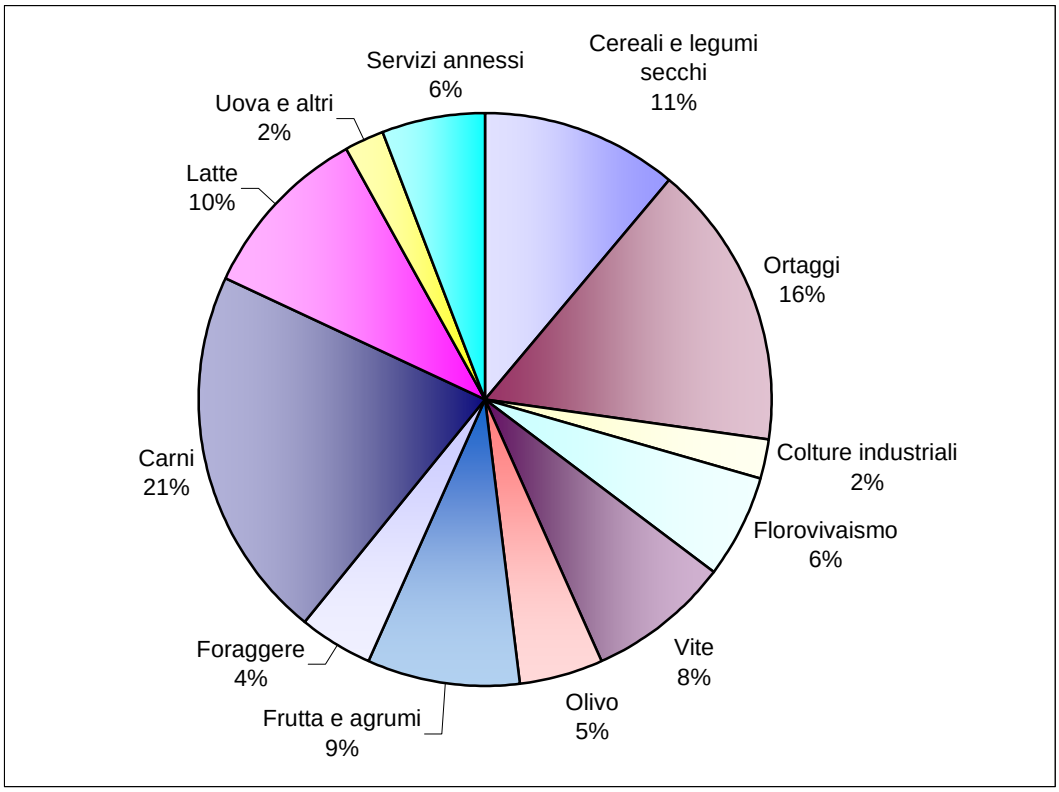
	Italia		***Variazione % 02-03***	
	Mln. Euro	***%***	***Quantità***	***Prezzi***
Erbacee	14.739	31,7	-8	8,7
Arboree	10.507	22,6	-5,5	5,4
Foraggere	1.811	3,9	-16,5	6,5
Zootecnia	14.765	31,8	-0,2	3,5
Servizi Annessi	2.642	5,7	1,3	2,3
Silvicoltura	399	0,8	-5,2	2,1
Pesca	1.621	3,5	5	3,6
Totale	***46.484***	***100***	***-4,4***	***5,5***

La produzione dei principali settori, escludendo pesca e silvicoltura, è sintetizzata, invece, dalla tabella 5.4, espressa sempre in termini di prezzi base, e dalla figura 5.1:

TABELLA 5.4. Produzione agricola per i principali settori (Fonte INEA).

Settori	***Mio. Euro***
Cereali e legumi secchi	4.964
Ortaggi	7.153
Colture industriali	989
Florovivaismo	2.557
Vite	3.564
Olivo	2.130
Frutta e agrumi	3.888
Foraggere	1.811
Carni	9.354
Latte	4.415
Uova e altri	997
Servizi annessi	2.642
TOTALE	***44.464***

FIGURA 5.1. Rappresentazione grafica della produzione agricola per i principali settori agricoli (Fonte INEA).



Per concludere, per i paesi aderenti nel 2002 all'Unione Europea (UE 15) il volume della produzione agricola è diminuito del 3,3% rispetto al 2002. La diminuzione ha interessato soprattutto il comparto delle colture vegetali (-6,3%) ed in particolare i cereali (-10,6%), la barbabietola da zucchero (-8,4%), le patate (-8,8%), la vite (-10,2%) e l'olivo (-23,8%). Per i nuovi paesi membri (UE 10) si è registrata una flessione media di circa il 10% della produzione agricola complessiva (esclusa Polonia e Malta).

TABELLA 5.5. Produzione agricole nei paesi UE (Fonte INEA).

	Produzione	
	Mln. Euro	***%***
Belgio	7.056	2,2
Danimarca	8.348	2,6
Germania	41.454	13,2
Grecia	12.189	3,9
Spagna	37.335	11,9
Francia	64.813	20,6
Irlanda	5.746	1,8
Italia	43.639	13,9
Lussemburgo	256	0,1
Olanda	20.114	6,4
Austria	5.704	1,8
Portogallo	6.258	2,0
Finlandia	4.288	1,4
Svezia	4.710	1,5
Regno Unito	24.465	7,8
UE 15	***286.375***	***91,1***
Polonia	13.241	4,2
Ungheria	6.077	1,9
Nuovi paesi UE	28.013	8,9
UE 25	**314.388**	**100**

Tutti i dati presentati nel paragrafo sono forniti dall'INEA.

5.4 ANALISI PER PICCOLE AREE SUI DATI DELL'INDAGINE STRUTTURA E PRODUZIONE DELLE AZIENDE AGRICOLE (SPA) 2003

Questo paragrafo si articola in due parti: analisi panoramica dei dati d'interesse dell'indagine e stima per piccole aree con stima post-stratificata, EBLUP e Spatial EBLUP.

5.4.1 ANALISI PANORAMICA DEI DATI CAMPIONARI

L'indagine sulla Struttura e Produzione delle Aziende Agricole del 2003 in Toscana prevedeva di rilevare 3000 aziende, rispondenti a certe caratteristiche (vedi paragrafo 5.2.1), tramite questionario con intervista diretta del conduttore. E' stato

possibile intervistare l'83,43% delle 3000 previste dal disegno campionario, pari a 2504 aziende[44]. Considerando le province il campione è così distribuito:

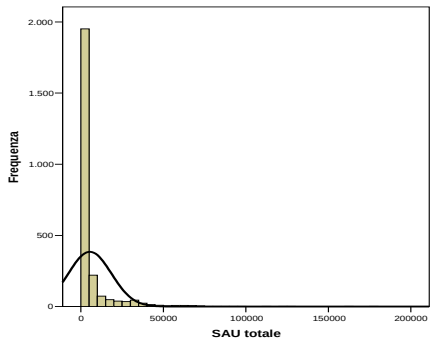
TABELLA 5.6. Distribuzione del campione dell'indagine SPA 2003 per Province.

Provincia	***Totale***	***%***
Arezzo	278	11,10%
Firenze	284	11,34%
Grosseto	473	18,89%
Livorno	147	5,87%
Lucca	279	11,14%
Massa-Carrara	103	4,11%
Pisa	204	8,15%
Pistoia	375	14,98%
Prato	27	1,08%
Siena	334	13,34%
Totale complessivo	***2504***	***100,00%***

La superficie agricola utilizzata, espressa in ettari, è ottenuta sommando le superfici agricole destinate a Seminativi, Coltivazioni Legnose Agrarie, Orti Familiari e Prati Permanenti e Pascoli.

La SAU media per azienda è 5459,96 ettari (errore standard 259,22) per un totale di 13586572 ettari di terreno agricolo utilizzato in Toscana. La dimensione della SAU per azienda spazia da un minimo di 2 ettari ad un massimo di 178383. L'indice di concentrazione di Gini è 0,796, quindi ci sono poche aziende con SAU di grandi dimensioni e molte aziende con SAU piccole; ciò rispecchia la tipica struttura della dimensione aziendale italiana che ha, non solo nel settore agricolo, una moltitudine di medie e piccole aziende a fronte di un numero esiguo di grandi aziende.

FIGURA 5.2. Distribuzione di frequenza per SAU in Toscana.



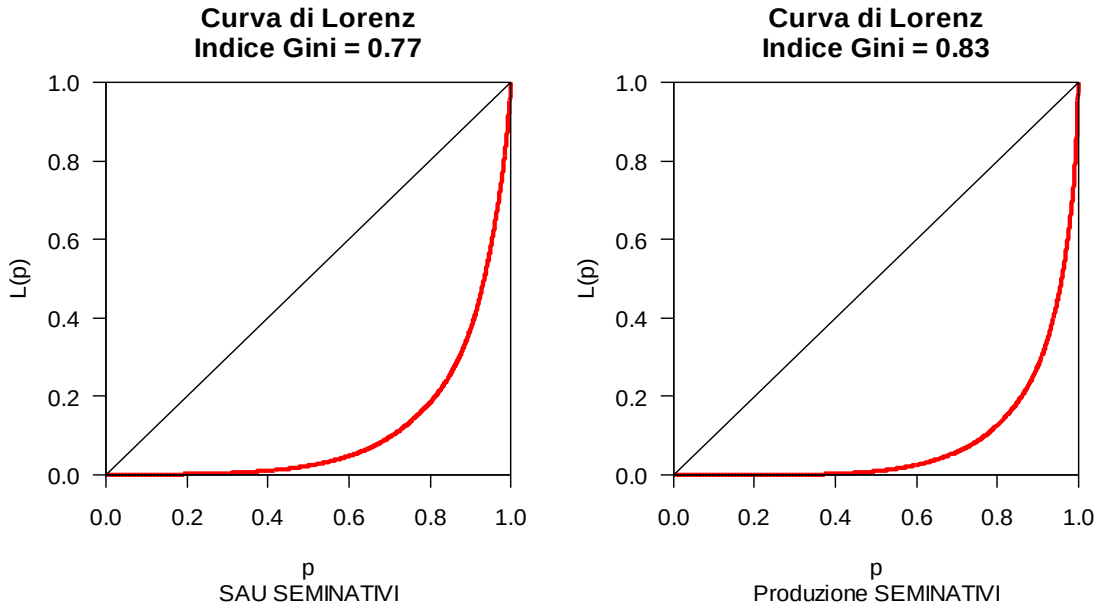
[44] Per gli aspetti riguardanti il trattamento della mancata risposta si veda Guarnera U. Luzi O. (2005)

Analizziamo la dimensione produttiva delle aziende agricole per i settori dei Seminativi e delle Coltivazioni Legnose Agrarie.

Il settore dei Seminativi è formato dalle coltivazioni di Cereali per la Produzione di Granella, Colture Proteiche per la Produzione di Granella, Patata, Barbabietola da Zucchero, Piante Sarchiate da Foraggio, Piante Industriali, Ortive, Fiori e Piante Ornamentali, Piantine, Foraggere avvicendate e sementi. La produzione di Fiori e Piante Ornamentali e di Piantine non è rilevata. Nella SAU dei seminativi non vengono inclusi i terreni a riposo.

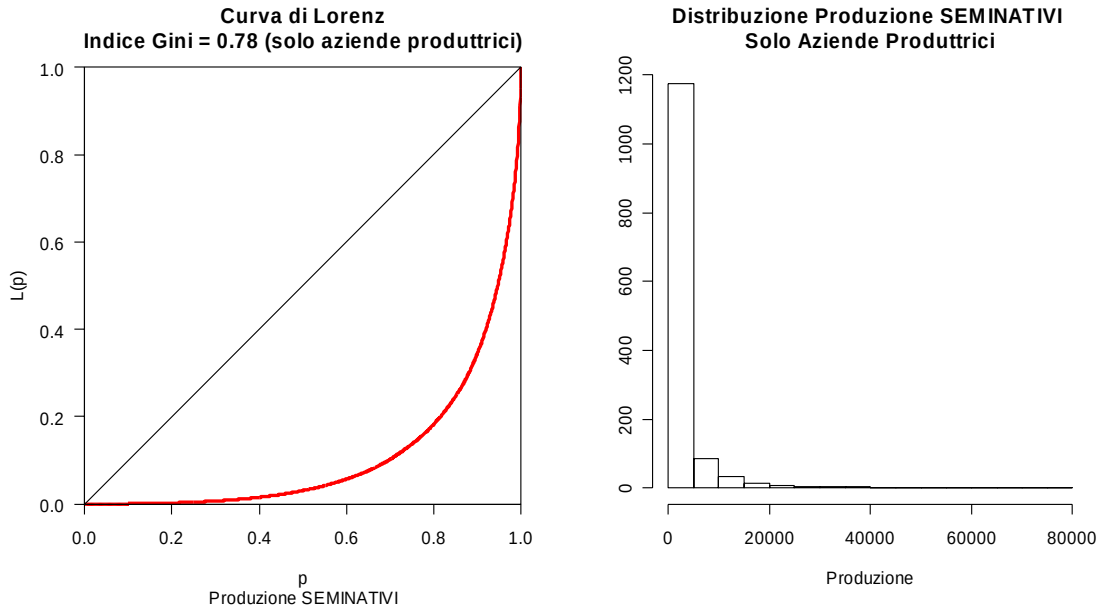
Delle 2504 aziende osservate nel campione 1765 hanno terreni adibiti alla coltivazione di seminativi ma sono 1333 le aziende che hanno effettivamente prodotto. La SAU totale destinata ai seminativi è 9023124 ettari (il 66,4% della SAU totale presente in Toscana) con una media di 5123,86 (errore standard 253,72) ettari per azienda, con SAU che varia da 1 a 112650 ettari. La produzione complessiva di seminativi è stata di 3001811 quintali con una media di 1704,61 (errore standard 123,80) quintali per azienda, con una produzione che varia tra 1 e 76504 quintali. L'indice di concentrazione di Gini per la SAU è 0,77 mentre per la produzione è 0,83. Anche utilizzando la curva di Lorenz (Figura 4) si osserva come sia la SAU sia, in misura maggiore, la produzione nel settore dei seminativi è concentrata in poche aziende di grandi dimensioni (parlando in termine di estensione territoriale e di produzione, non in termini di redditività).

FIGURA 5.3. Curva di Lorenz per SAU e Produzione di Seminativi.



Se consideriamo solo le aziende che sono riuscite a produrre (1333 su 1765) l'indice di concentrazione della produzione è 0,72, segnalando una concentrazione ovviamente meno accentuata[45], anche se sempre importante; anche considerando la distribuzione di frequenza si osserva un'asimmetria nella produzione di seminativi.

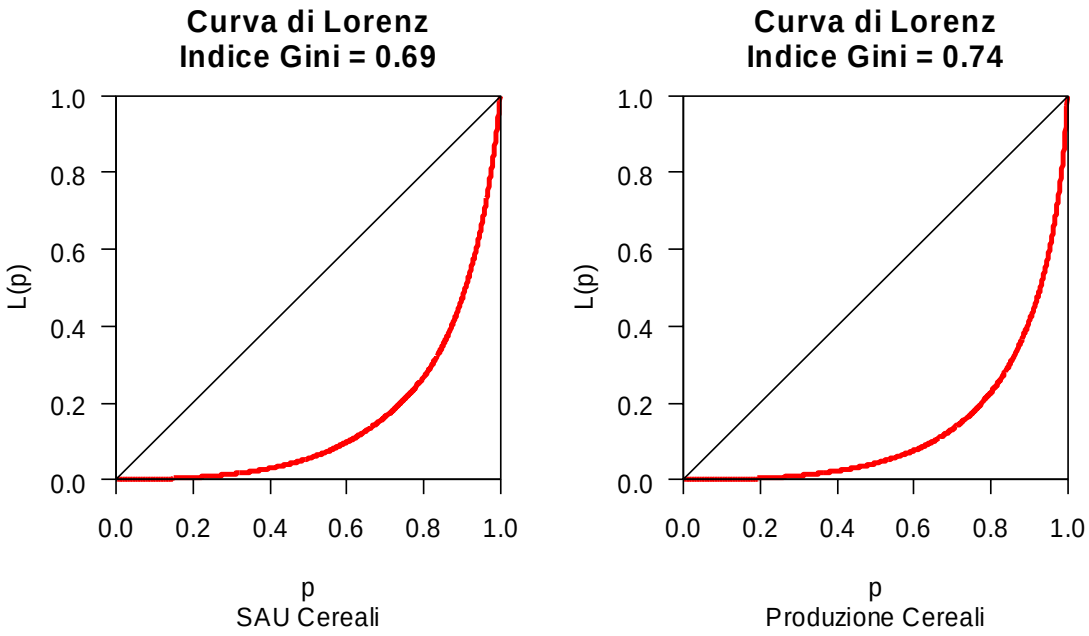
FIGURA 5.4. Curva di Lorenz e Distribuzione di Frequenza per Produzione di Seminativi per le aziende che hanno avuto un raccolto.



Le coltivazioni più importanti nel settore dei seminativi sono i Cereali, con 1192 aziende con SAU dedicata e 1097 aziende hanno che ottenuto un raccolto. La SAU dedicata alla coltivazione di cereali è pari a 4724726 ettari per una produzione totale di 1228789 quintali. La SAU media per azienda è 4066,03 (errore standard 202,29) ettari per una produzione media di 1057,48 (errore standard 65,90) quintali di cereali. La SAU varia tra 2 e 61457 ettari mentre la produzione varia tra 1 e 31000 quintali. Gli indici di Gini per la SAU e la produzione, in linea con quelli del settore, sono rispettivamente 0,69 e 0,74.

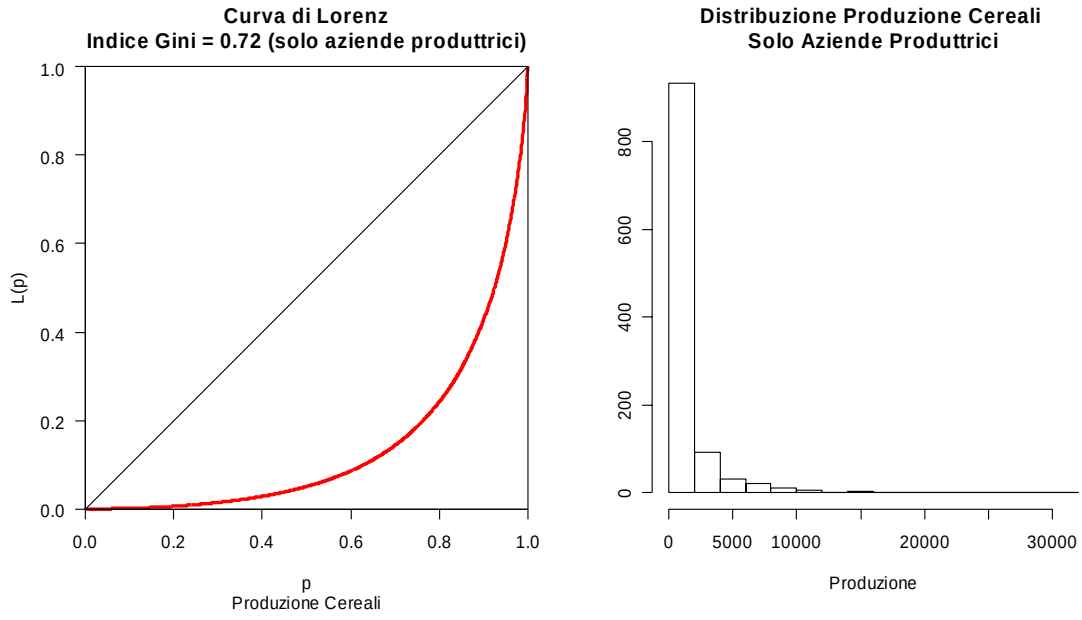
[45] Se si escludono le aziende con produzione 0 dall'indice di concentrazione di Gini allora l'area compresa tra la curva di Lorenz e la retta di equidistribuzione diminuirà, poiché manca il tratto di curva che sta sull'ascissa (dovuto alle aziende con produzione 0). L'indice di Gini può essere interpretato come il rapporto tra l'area suddetta e ½ (= area sottesa alla retta di equidistribuzione). Se il numeratore diminuisce allora diminuisce il valore dell'indice di concentrazione.

FIGURA 5.5. Curva di Lorenz per SAU e Produzione di Cereali.



Se consideriamo solo le 1097 imprese che sono riuscite ad ottenere un raccolto, l'indice di concentrazione della produzione passa da 0,74 a 0,72 (resta invariato). Per la Toscana il 10% delle imprese più grandi produce il 51% (circa) dei cereali e se consideriamo il 20% delle imprese la quota sulla produzione sale al 70%.

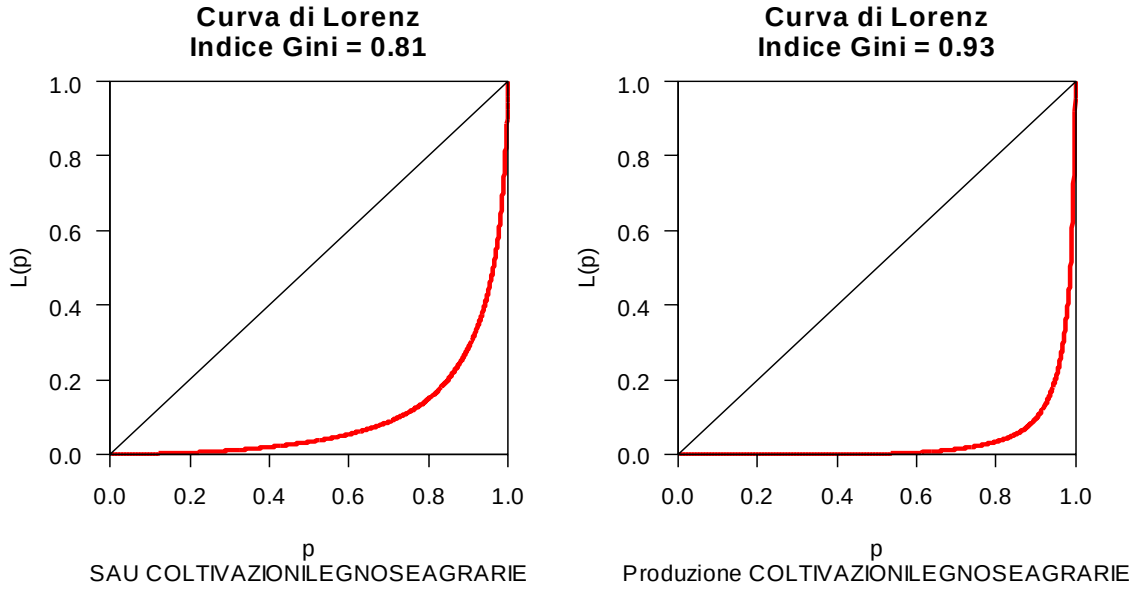
FIGURA 5.6. Curva di Lorenz e Distribuzione di Frequenza per Produzione di Cereali per le aziende che hanno avuto un raccolto.



Il settore delle Coltivazioni Legnose Agrarie è formato dalle coltivazioni di Vite, Olivo, Agrumi, Frutta Fresca (di origine temperata e sub-tropicale), Frutta in guscio, Vivai, Coltivazioni Legnose Agrarie in Serra, Altre Coltivazioni Legnose Agrarie. I vivai (fruttiferi, piante ornamentali, etc.) non si considerano a livello di produzione, viene rilevata solo la dimensione della SAU.

Delle 2504 aziende osservate 1969 hanno terreni adibiti alle coltivazioni legnose agrarie, ma sono 1112 le aziende che hanno effettivamente prodotto nell'annata 2003. La SAU totale destinata alle coltivazioni legnose agrarie è 2270641 ettari (il 16,7% della SAU totale presente in Toscana) con una media di 1153,20 (errore standard 85,19) ettari per azienda, con SAU che varia da 1 a 95841 ettari. La produzione complessiva di coltivazione legnose agrarie è stata di 343457 quintali con una media di 174,43 (errore standard 19,63) quintali per azienda, con una produzione che varia tra 1 e 16120 quintali. L'indice di concentrazione di Gini per la SAU è 0,81 mentre per la produzione è 0,93. Anche utilizzando la curva di Lorenz (Figura 18) si osserva come sia la SAU nel settore delle coltivazioni legnose agrarie sia, in misura maggiore, la produzione è concentrata in poche aziende di grandi dimensioni in modo ancora più accentuato rispetto al settore dei seminativi.

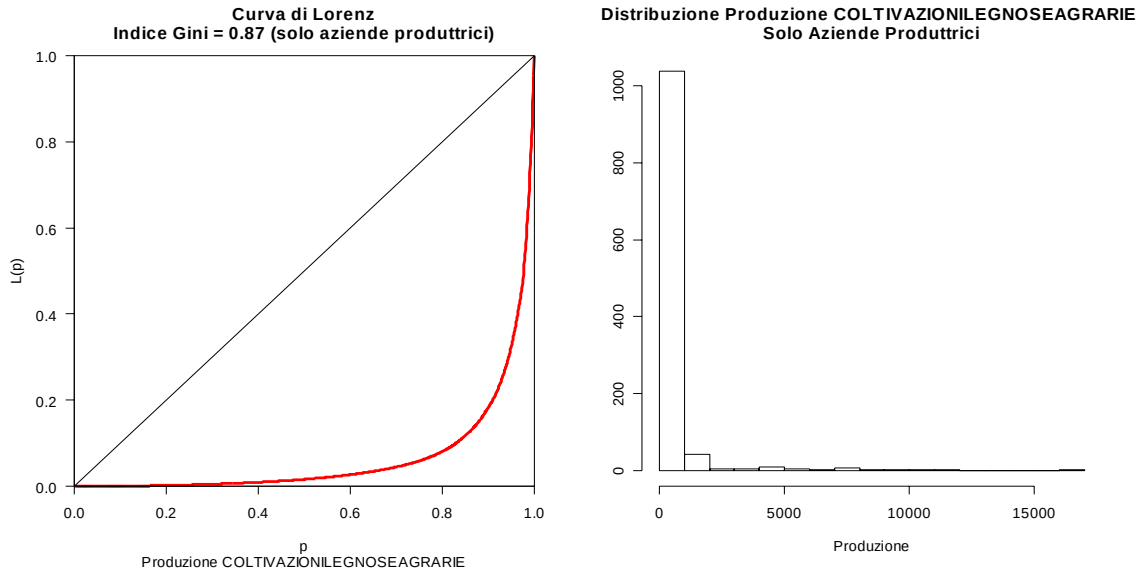
FIGURA 5.7. Curva di Lorenz per SAU e Produzione di Coltivazioni Legnose Agrarie.



Se consideriamo solo le aziende che sono riuscite a produrre (1112 su 1969, cioè il 56,5%) la concentrazione nella produzione è sempre di altissima intensità con un indice

di Gini uguale a 0,87; anche considerando la distribuzione di frequenza per la produzione si confermano i risultati esposti.

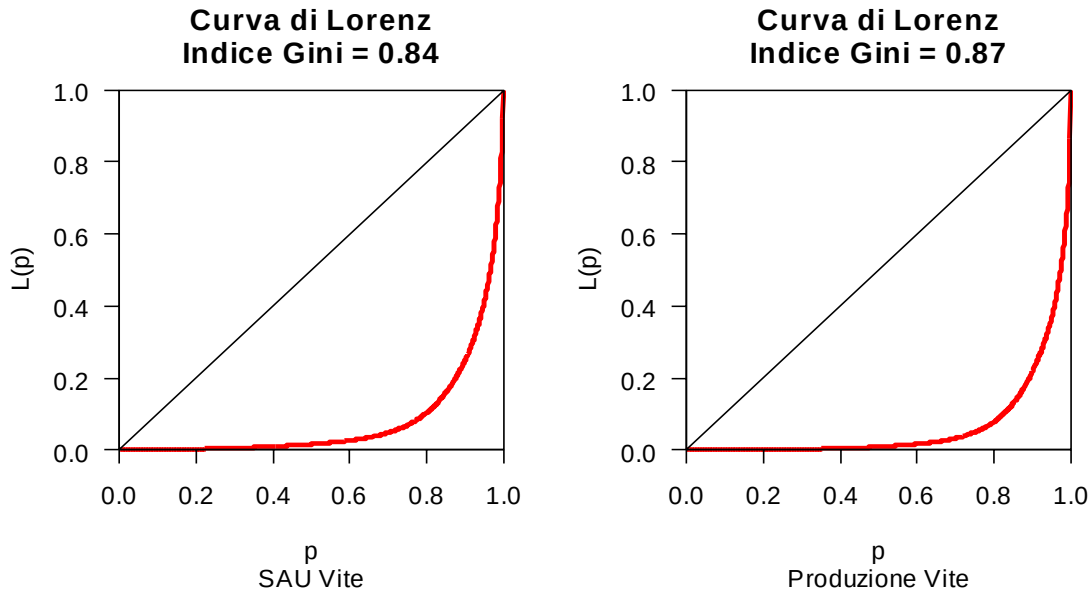
FIGURA 5.8. Curva di Lorenz e Distribuzione di Frequenza per le Coltivazioni Legnose Agrarie per le aziende che hanno avuto un raccolto.



Tra le coltivazioni legnose agrarie in Toscana le più importanti da un punto di vista sia di SAU, sia di produzione, nonché di fama internazionale in campo enogastronomico, sono la Vite e l'Ulivo, che danno origine a prodotti legati al territorio di altissima qualità.

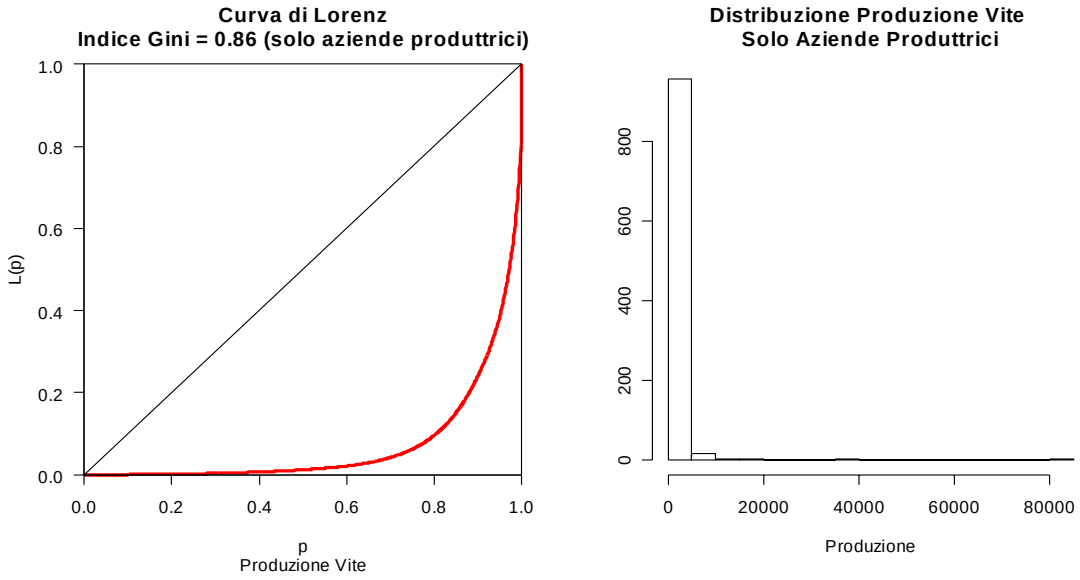
Le viti comprendono la produzione di vini DOC e DOCG, di altri vini e di uva da tavola. La SAU totale dedicata alle viti da 1093 aziende del campione è 1075152 ettari, pari al 47,3% della SAU destinata alle coltivazioni legnose agrarie, con una media di 983,67 (errore standard 109,55) ettari per azienda e la dimensione della SAU per azienda varia tra 1 e 83676 ettari: la più grande SAU, in Toscana, dedicata ad una sola coltivazione. La produzione totale ammonta a 611879 quintali con una media di 559,82 (errore standard 91,07) quintali per azienda; a livello di singola azienda (delle 977 che hanno prodotto) si passa da una produzione minima di 1 quintale ad una massima di 83100. L'indice di concentrazione per la SAU e la produzione di ortive è rispettivamente di 0,84 e 0,87.

FIGURA 5.9. Curva di Lorenz per SAU e Produzione di Vite.



Considerando solo le aziende che hanno effettivamente avuto un raccolto, l'indice di concentrazione è 0,86. Il 10% delle maggiori aziende produce il 76% circa dei derivati della vite e se consideriamo il 20% delle aziende la quota sulla produzione sale al 90% circa.

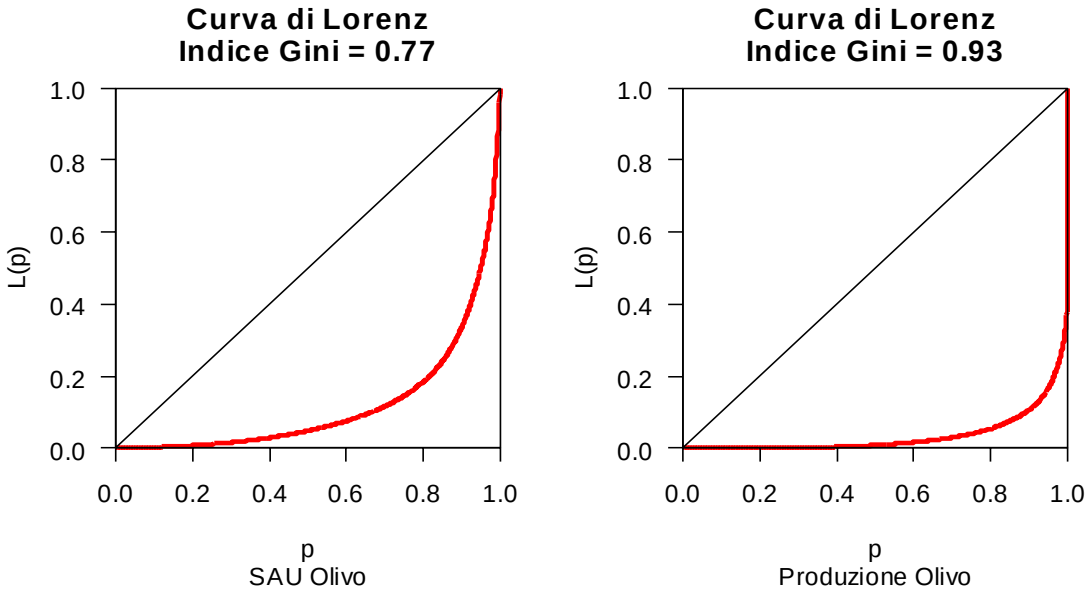
FIGURA 5.10. Curva di Lorenz e Distribuzione di Frequenza per le Viti per le aziende che hanno avuto un raccolto.



Segue alla coltivazione di vite, per diffusione sul territorio, la coltivazione di olivo (per la produzione sia di olio, sia di olive da tavola) con 1451 aziende con SAU dedicata e 1119 aziende che anno raccolto nell'annata agricola 2003. La SAU dedicata alla

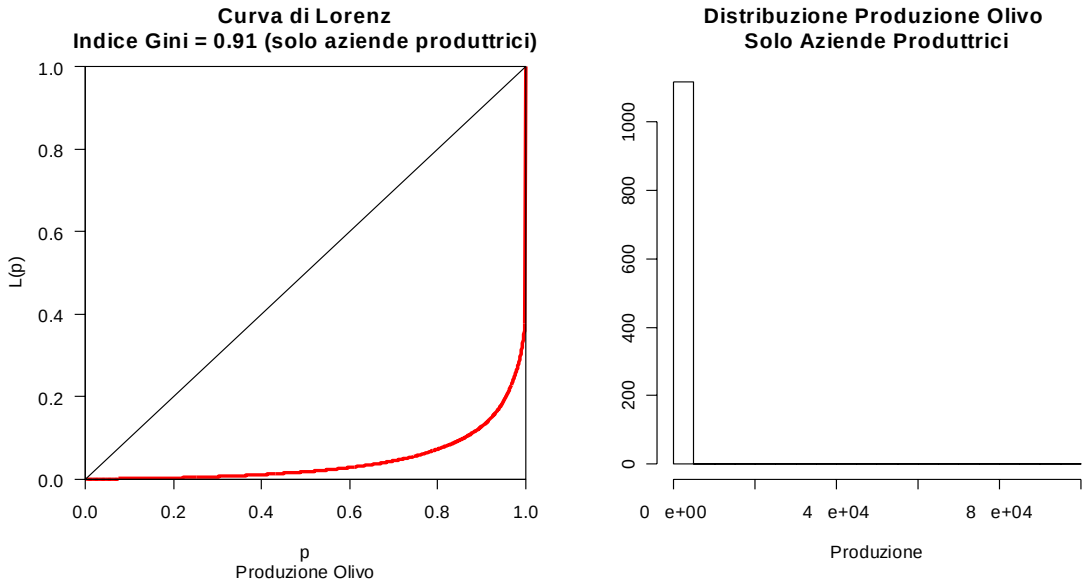
coltivazione di olivo è pari a 822206 ettari per una produzione totale di 156946 quintali. La SAU media per azienda è 566,65 (errore standard 38,60) ettari per una produzione media di 108,16 (errore standard 67,38) quintali di olive. La SAU varia tra un minimo di 1 e un massimo di 16000 ettari, mentre la produzione varia tra 1 e 97700 quintali. Gli indici di Gini per la SAU e la produzione, in linea con quelli del settore, sono rispettivamente 0,77 e 0,93.

FIGURA 5.11. Curva di Lorenz per SAU e Produzione di Olivo.



Considerando solo le aziende che hanno ottenuto un raccolto di olive l'indice di concentrazione di Gini è 0,81. Il 10% delle aziende di maggiore produzione ha raccolto circa l'87% delle olive prodotte in Toscana. In questo settore c'è un azienda outlier, infatti da sola produce il 62,3% della produzione totale di olive nella regione. Questa azienda ha sede nel comune di Montalcino (SI) e produce 97700 quintali di olive contro i 1500 quintali della seconda aziende produttrice della Toscana; questo in base alle osservazioni campionarie.

FIGURA 5.12. Curva di Lorenz e Distribuzione di Frequenza per le Olive per le aziende che hanno avuto un raccolto.



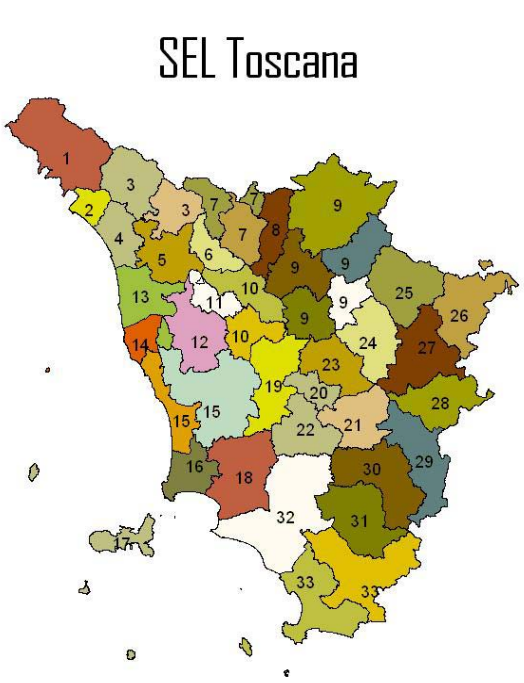
L'indagine sulla Struttura e Produzione delle Aziende Agricole rileva anche le produzioni derivate da agricoltura biologica, la produzione di latte, dati sul personale, informazioni sulle superficie irrigate e sui capi di bestiame allevati. Vengono rilevate, inoltre, informazioni su alcuni aspetti giuridico-economici e informazioni relative ad un settore oggetto di approfondimenti che cambia ogni anno. Nella SPA 2003 sono state rilevate informazioni sulle attività multifunzionali dell'azienda agricola (come attività di agriturismo, artigianato, lavorazione di prodotti derivati dall'agricoltura, produzione di energia rinnovabile, etc.).
In questo contesto ci siamo limitati all'analisi delle principali produzioni in campo agricolo in Toscana (seminativi, coltivazioni legnose agrarie e alcune coltivazioni specifiche di questi settori), escludendo le colture biologiche e il settore zootecnico.

Sintetizzando i dati analizzati, scopriamo un settore caratterizzato dalla presenza di un numero ristretto di aziende di grandi dimensioni che producono la maggior parte delle coltivazioni affiancate da una moltitudine di piccole e piccolissime aziende. L'annata 2003 è stata caratterizzata da fenomeni di siccità e alluvionali, evento che può aver avvantaggiato le grandi aziende che hanno avuto i mezzi per fronteggiare il clima sfavorevole.

5.4.2 STIMA PER SUPERFICIE ECONOMIA LOCALE DELLE PRINCIPALI PRODUZIONI AGRICOLE IN TOSCANA

L'analisi a livello regionale può essere insufficiente per conoscere la struttura della produzione agricola rispetto al territorio. Si è interessati in particolar modo alla produzione nel settore dei seminativi e delle coltivazioni legnose agrarie. In particolare siamo interessati alla stima della produzione media per azienda per superficie economica locale (SEL) (di un settore o di particolari coltivazioni).
Le regioni italiane sono divise in SEL; La Toscana è divisa in 33 SEL:

FIGURA 5.13. Suddivisione della Toscana in SEL[46].



1	LUNIGIANA
2	AREA DI MASSA E CARRARA
3	VALLE DEL SERCHIO
4	VERSILIA
5	AREA LUCCHESE
6	VAL DI NIEVOLE
7	AREA PISTOIESE
8	AREA PRATESE
9	AREA FIORENTINA
10	CIRCONDARIO DI EMPOLI
11	VALDARNO INFERIORE
12	VAL D'ERA
13	AREA PISANA
14	AREA LIVORNESE
15	VAL DI CECINA
16	VAL DI CORNIA
17	ARCIPELAGO
18	COLLINE METALLIFERE
19	ALTA VAL D'ELSA
20	AREA URBANA SENESE
21	CRETE SENESI - VAL D'ARBIA
22	VAL DI MERSE
23	CHIANTI SENESE
24	VALDARNO SUPERIORE SUD
25	CASENTINO
26	ALTA VALTIBERINA
27	AREA ARETINA
28	VAL DI CHIANA ARETINA
29	VAL DI CHIANA SENESE
30	AMIATA - VAL D'ORCIA
31	AMIATA GROSSETANO
32	AREA GROSSETANA
33	ALBEGNA - FIORA

Le SEL 3, 7, 9, 10, 15 e 33 sono suddivise in più parti (3.1, 3.2; 7.1, 7.2; 9.1, 9.2, 9.3, 9.4, 9.5; 10.1, 10.2; 15.1, 15.2; 33.1, 33.2), vista l'importanza di questi territori a livello agricolo, in modo da avere in tutto 42 SEL. Riportiamo nella tabella seguente la numerosità campionaria dell'indagine SPA per ogni SEL:

[46] Le SEL ulteriormente suddivise hanno la denominazione in comune. Per questo sulla cartina i numeri delle SEL suddivise sono ripetuti più volte.

TABELLA 5.7. Numerosità campionaria per SEL (indagine SPA 2003).

SEL	*Num. Camp.*	*SEL*	*Num. Camp.*
1	82	15.1	59
2	21	15.2	62
3.1	31	16	63
3.2	13	17	6
4	114	18	64
5	121	19	55
6	128	20	21
7.1	9	21	59
7.2	238	22	23
8	27	23	34
9.1	40	24	62
9.2	25	25	20
9.3	45	26	23
9.4	42	27	82
9.5	21	28	91
10.1	60	29	92
10.2	43	30	58
11	20	31	70
12	74	32	138
13	48	33.1	90
14	19	33.2	111

Come si può osservare dalla tabella, la numerosità campionaria è piuttosto bassa in alcune SEL, si va da 6 a 138 aziende. Infatti anche in quelle aree dove la numerosità supera le 100 osservazioni non si ottengono stime soddisfacenti utilizzando lo stimatore post-stratificato, come si vedrà in dettaglio nelle prossime pagine.

Dopo aver introdotto, nei capitoli 2 e 3, la metodologia adottata per le stime analizziamo la produzione media per azienda di seminativi e delle coltivazioni legnose agrarie per SEL, analizzando anche le coltivazioni principali di questi settori.

La stima dei dati a disposizione suggerisce l'uso di un modello lineare ad effetti misti per la previsione della produzione media per azienda per ogni SEL. I dati sono, infatti, presumibilmente affetti da correlazione spaziale ed i risultati del nostro studio di simulazione suggeriscono di usare la metodologia di stima proposta (par. 2.4.3 e 3.6).

Al fine di scegliere opportuni repressori per la parte fissa del modello sono state fatte prove con i dati censuari: in questo modo le covariate del modello non sono stime e quindi non aggiungono un ulteriore fattore di variabilità alle stime in questione.
Inizialmente è stato applicato un modello regressione lineare ai dati sulla produzione con le variabili ausiliarie SAU totale e Forma giuridica ("dummy variable"). E' risultata

significativa solo la SAU totale. In base a questo risultato si è deciso di applicare il modello di regressione lineare univariato alle altre variabili riguardanti la dimensione aziendale: SAU specifica per una coltivazione, UDE (Unità di Dimensione Economica dell'azienda), OTE (Orientamento Tecnico Economico) e UBA (Unità di Bovino Adulto). Queste variabili sono legate tra loro, quindi non sono state usate insieme in un modello di regressione lineare multiplo per evitare problemi di multicollinearità. La SAU specifica per coltivazione ha fornito i risultati migliori in termini di significatività dei coefficienti e di indice di determinazione, il modello lineare più utile è dunque:

$$y_{i_j} = \alpha_i + \beta_i x_{1_{ij}} + \varepsilon_{i_j}$$

Dove y_{i_j} è la produzione dell'azienda j nella SEL i, α_i è la costante per la SEL i, β_i è il coefficiente di regressione della SEL i, $x_{1_{ij}}$ è la SAU dedicata alla produzione dall'azienda j-esima nella SEL i ed ε_{ij} è un disturbo white noise. Per il modello di regressione abbiamo calcolato le stime OLS di α e β, la loro significatività, l'indice di determinazione (R^2) e un test di normalità.

Dalle applicazioni fatte per la scelta dei regressori è evidente che i residui non sono risultati distribuiti normalmente (in base al test di Shapiro), indice, questo, come si sa, di un modello mal specificato. Ci aspettavamo questo risultato che concorda con quanto detto nel capitolo 2: il modello lineare ad effetti fissi non è sufficiente per dati spaziali e bisogna ricorrere al modello lineare ad effetti misti. Nell'economia di questo studio la regressione lineare è stata usata solo per individuare potenziali regressori.

Per stimare la produzione media per azienda delle coltivazioni d'interesse per SEL, si sono applicati gli stimatori: post-stratificato, EBLUP e Spatial EBLUP.

Lo stimatore post-stratificato è calcolato con la (5.1) e la sua varianza con la (5.2):

$$\overline{y}_i = \frac{\sum_{j=1}^{n_i} w_{i_j} y_{i_j}}{\sum_{j=1}^{n_i} w_{i_j}} \qquad (5.1)$$

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} w_{i_j} (y_{i_j} - \bar{y}_i)^2}{\sum_{j=1}^{n_i} w_{i_j}} \cdot \left(1 - \frac{n_i}{N_i}\right) \tag{5.2}$$

Dove $\bar{y}_i$ è la produzione media per la SEL *i*-esima, w_{i_j} è il peso dell'azienda *j*-esima nella SEL *i* (tale valore è fornito dall'ISTAT ed è calcolato in base al disegno campionario), $\hat{\sigma}_i^2$ è la varianza della produzione nella SEL *i*. N_i e n_i sono rispettivamente il numero di aziende presenti nella SEL *i* (dato ottenuto dal censimento sull'agricoltura del 2000) e il numero delle aziende campionate appartenenti alla SEL *i*. Con la (5.1) la (5.2) abbiamo calcolato una stima con intervallo di confidenza del 95%:

$$\bar{y}_i \pm 1{,}96 \cdot \sqrt{\frac{\hat{\sigma}_i^2}{n_i}}$$

Le stime EBLUP e Spatial EBLUP sono calcolate utilizzando il modello a livello di area; si rimanda alle formule (2.37) per lo stimatore EBLUP e (3.12) per lo stimatore Spatial EBLUP.

Per stimare $\hat{\sigma}_u^2$ e $\hat{\rho}$ si utilizzano i metodi di massima verosimiglianza presentati, mentre per la componente di varianza σ_e^2, che in realtà non è nota, utilizziamo come sua proxy la varianza della stima post stratificata, presentata nella (5.2). Ciò comporta un aumento del MSE: in pratica è stato dimostrato che si sovrastima il MSE, con una conseguente perdita di efficienza, ma non di correttezza. $\boldsymbol{I}_m$ è una matrice identica di dimensione 42 (numero di SEL). La matrice di connessione $\boldsymbol{W}$ (di dimensione *42x42*) è stata standardizzata[47], in questo modo il coefficiente di autocorrelazione spaziale ρ è compreso tra $-1/\min(\lambda_W)$ e $1/\max(\lambda_W)$, dove λ_W è l'autovalore della matrice $\boldsymbol{W}$ (standardizzata). Essendo $\boldsymbol{W}$ una matrice decomponibile (poiché ha sicuramente un elemento uguale a zero sulla diagonale secondaria) non negativa, per il teorema di Perron-Frobenius (II proprietà) il raggio spettrale della matrice è compreso tra il minimo e il massimo delle somme per riga, che nel caso della matrice $\boldsymbol{W}$ coincidono e

[47] Per standardizzare una matrice si divide ogni elemento per la somma degli elementi appartenenti alla riga dello stesso: $w_{ij}^s = w_{ij} / \sum_{j=1}^{n} w_{ij}$.

sono pari a 1. Questo implica che il coefficiente di autocorrelazione spaziale assume al massimo valore 1[48].

L'intervallo di confidenza al 95% è calcolato utilizzando la stima del MSE presentata nei capitoli 2 e 3, rispettivamente per EBLUP e Spatial EBLUP. I risultati della stima della produzione media per azienda in ogni SEL sono presentati dettagliatamente nell'appendice D, sia per gli stimatori EBLUP e Spatial EBLUP sia per lo stimatore post-stratificato. Nei punti seguenti presenteremo la stima dei parametri, di EBLUP e Spatial EBLUP, che ci hanno permesso di calcolare la produzione media per azienda per SEL nelle coltivazioni d'interesse, georeferenziando quest'utlima sulla carta geografica della Toscana (utilizzando un sistema GIS). Tramite la rappresentazione geografica è facile individuare le SEL in cui le aziende sono orientate a certi tipi di coltivazioni e se esistono dei "distretti", cioè dei gruppi di SEL contigue caratterizzate da produzione medie per azienda elevate. Si può osservare, inoltre, le differenze dei tre stimatori usati sul territorio oggetto d'interesse.

SEMINATIVI

Per vedere se c'è autocorrelazione spaziale tra le SEL per la produzione di seminativi utilizziamo l'indice di Moran:

TABELLA 5.8. Indice di Moran per la produzione di Seminativi.

Statistica I_M	***Err. Std.***	***p-val.***
0,405	4,33	0,000

In base ai valori ottenuti si rifiuta l'ipotesi nulla in favore dell'ipotesi alternativa di presenza di autocorrelazione spaziale, che risulta positiva.

Presentiamo, riassunti in una tabella, la stima dei parametri di EBLUP e Spatial EBLUP e i relativi errori standard:

[48] Notare che la standardizzazione della matrice di connessione non comporta variazioni nei pesi tra le aree confinanti poiché per ogni riga gli elementi diversi da zero hanno uguale valore.

TABELLA 5.9. Stima dei parametri di EBLUP e Spatial EBLUP per la produzione di Seminativi per SEL.

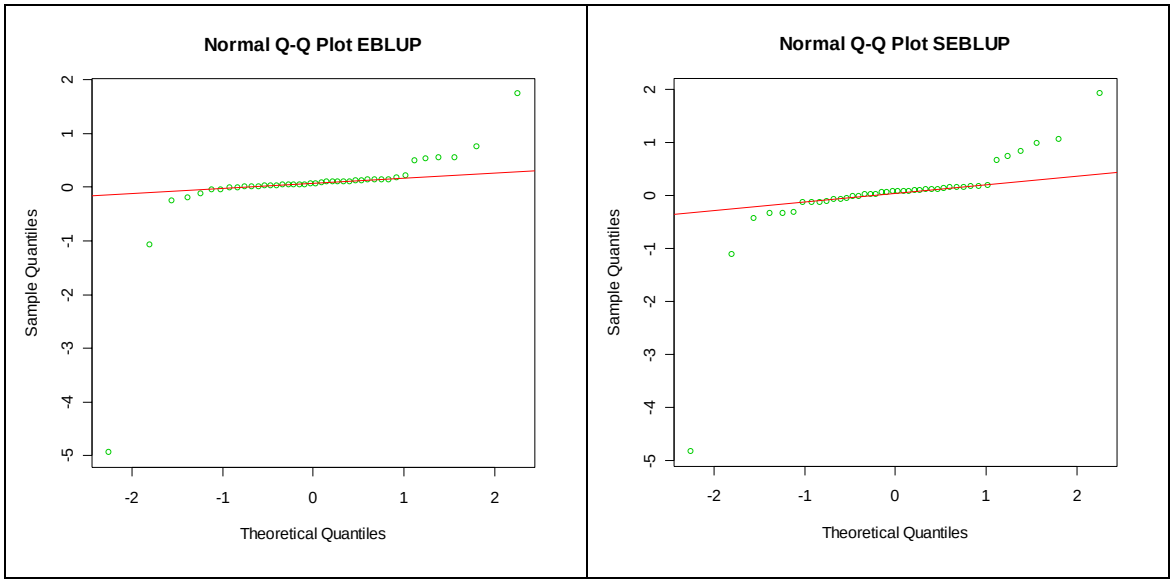
	EBLUP			***Spatial EBLUP***		
	Valore	***Err. Std.***	***p-val.***	***Valore***	***Err. Std.***	***p-val.***
Costante	-1,34	2,39	0,578	-1,46	2,51	0,563
β_1	0,25	0,026	0,000	0,253	0,026	0,000
σ_u^2 ***stimato***	11,99	13,87	0,392	11,683	0,074	0,000
P stimato	-	-	-	0,236	0,730	0,749
Test Shapiro[49]	0,45	-	0,000	0,57	-	0,000

La costante non è statisticamente significativa, utilizzando un intervallo di confidenza sia del 95% sia del 99%, per entrambi gli stimatori; al contrario β_1 è significativo al 95% e al 99% per entrambi. Dal coefficiente di regressione vediamo che 1 ettaro di SAU produce 0,25 quintali, in media, di seminativi, indipendentemente dall'area in cui si trova; l'effetto della produzione dovuto all'area, infatti, è funzione di σ_u^2 (si ricorda $\hat{u}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \cdot \left(\bar{y}_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}} \right)$). La stima di σ_u^2 non è significativamente diversa da 0 per la stima EBLUP, mentre è significativa, al 95% e al 99%, per lo Spatial EBLUP. La stima di ρ non è significativa, quindi nel modello proposto non si riscontra un effetto sulla produzione di seminativi tra le SEL, che risultava invece dall'indice di Moran. Con il test Shapiro si verifica la normalità degli errori e si rifiuta l'ipotesi nulla di distribuzione normale degli effetti di area, sia con una significatività del 95% sia con una del 99%. Tuttavia, è stato dimostrato che anche in presenza di errori (effetti di area e campionari) non normali le stime, sia EBLUP sia Spatial EBLUP, sono robuste; si veda in proposito Rao, 2003.

La distribuzione degli effetti di area si può rappresentare graficamente con il grafico "quantile-quantile":

[49] Il test di normalità serve per verificare se un campione di n elementi è stato estratto da una popolazione distribuita normalmente. Il test Wilks-Shapiro (in generale test Shapiro) si basa sulla statistica seguente: $W = \sum_{i=1}^{n} \left(a_i \tilde{x}_i \right)^2 / \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$, dove $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ è il vettore delle osservazioni campionarie, $\tilde{\boldsymbol{x}}$ è il vettore $\boldsymbol{x}$ ordinato in modo crescente e $\boldsymbol{a}$ è un vettore di costanti ottenuto dalle medie, varianze e covarianze di una normal order statistic di un campione di ampiezza n: $\boldsymbol{a} = \boldsymbol{m}^T \boldsymbol{V}^{-1} \left(\boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m} \right)^{-\frac{1}{2}}$, dove $\boldsymbol{m} = [m_1, \ldots, m_n]^T$ è il vettore dei valori attesi di una standard normal order statistic e $\boldsymbol{V}$ è la matrice di varianza-covarianza generata da $\boldsymbol{m}$.

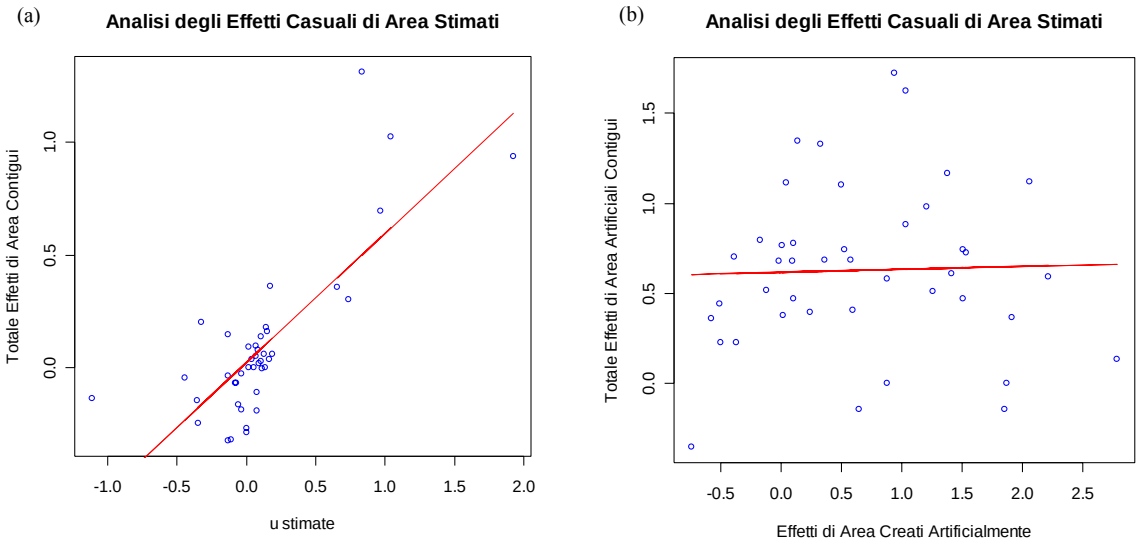
FIGURA 5.14. Rappresentazione degli effetti di area[50] per i Seminativi.



Gli effetti di area distribuiti normalmente giacciono su una retta inclinata positivamente di quarantacinque gradi, in questo caso invece si nota una tendenza negli effetti di area (una "S" rovesciata), indice del fatto che essi, per entrambi gli stimatori, non sono distribuiti normalmente.

Per verificare la tendenza degli effetti casuali tra aree graficamente si utilizzano due rappresentazioni grafiche:

FIGURA 5.15. La stima degli effetti casuali (u) contro una media degli effetti casuali stimati per le aree contigue (a) e una media per aree scelte casualmente (b). Seminativi.

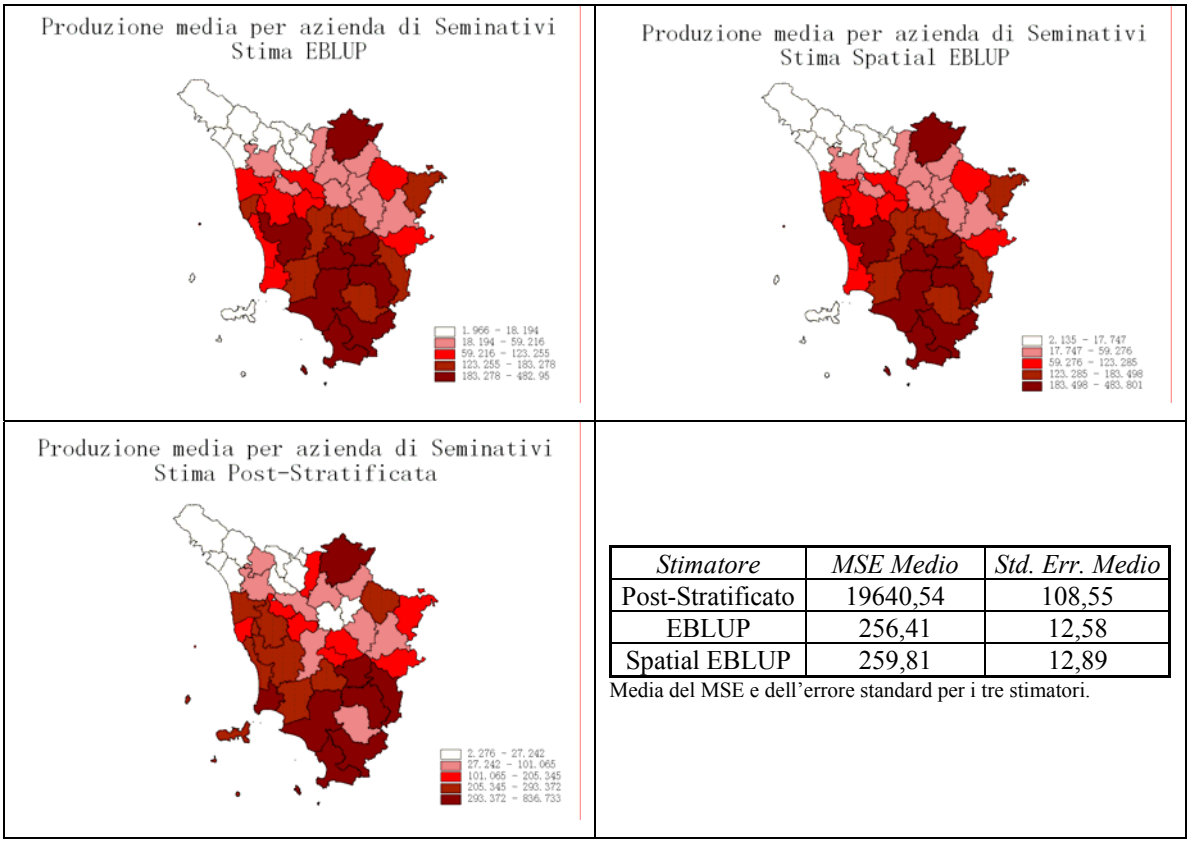


[50] SEBLUP è acronimo di Spatial EBLUP.

Un metodo per valutare il modello applicato è basato sull'utilizzo di due rappresentazioni grafiche: gli effetti casuali di area contro il "Totale degli effetti di area contigui" (Figura 5.15 (a)), che è una media degli effetti di area stimati tra SEL contigue; è interessante mettere il grafico ottenuto a confronto con uno simile dove sulle ascisse utilizziamo una media degli effetti di area scelti casualmente (Figura 5.15 (b)). Se non c'è autocorrelazione spaziale negli effetti di area, allora ci si aspetta che il coefficiente angolare della retta ottenuta interpolando gli effetti di area con i minimi quadrati sia significativamente non diverso da zero. Questo è il caso della Figura 5.16 (b): il raffronto tra i grafici (a) e (b) ci induce a credere che tra i dati ci sia autocorrelazione spaziale. Dalla dispersione dei punti intorno alla retta nel grafico in Figura 5.15 (a) si intuisce che c'è molta variabilità negli effetti di area stimati; per questo motivo nel modello non si ottiene una stima significativa di ρ.

Georeferenziamo le stime di produzione media per azienda in ogni SEL. In questo modo è possibile vedere sia quali sono le aree interessate nella produzione di seminativi, sia le differenze nei risultati tra i tre modelli di stima: classica, EBLUP e Spatial EBLUP.

FIGURA 5.16. Produzione media di Seminativi per azienda per SEL, anno 2003.



Stimatore	*MSE Medio*	*Std. Err. Medio*
Post-Stratificato	19640,54	108,55
EBLUP	256,41	12,58
Spatial EBLUP	259,81	12,89

Media del MSE e dell'errore standard per i tre stimatori.

La rappresentazione delle stime EBLUP e Spatial EBLUP è molto simile, infatti l'unica divergenza, anche se minima, è nel primo quantile della produzione media per SEL, che varia da 1,966 a 18,194 quintali per la stima EBLUP contro un intervallo da 2,135 a 17,747 quintali per la stima Spatial EBLUP. Questo è stato un risultato atteso poiché abbiamo riscontrato un coefficiente di autocorrelazione spaziale prossimo a zero e non significativo. Le stime ottenute con lo stimatore post-stratificato risultano nettamente differenti nei valori medi di produzione che sono molto superiori. D'altronde lo stimatore post-stratificato ha in media uno standard error circa nove volte superiore a quello di EBLUP e Spatial EBLUP: si intuisce quindi come le stime effettuate con i modelli proposti siano nettamente più efficienti.

COLTIVAZIONI LEGNOSE AGRARIE

Per vedere se c'è autocorrelazione spaziale tra le SEL per la produzione di coltivazioni legnose agrarie utilizziamo l'indice di Moran:

TABELLA 5.10. Indice di Moran per la produzione di Coltivazioni Legnose Agrarie.

Statistica I_M	***Dev. Std.***	***p-val.***
0,019	0,44	0,661

In base ai valori ottenuti non si rifiuta l'ipotesi nulla di assenza di autocorrelazione spaziale.

Presentiamo, riassunti in una tabella, la stima dei parametri di EBLUP e Spatial EBLUP e i relativi errori standard:

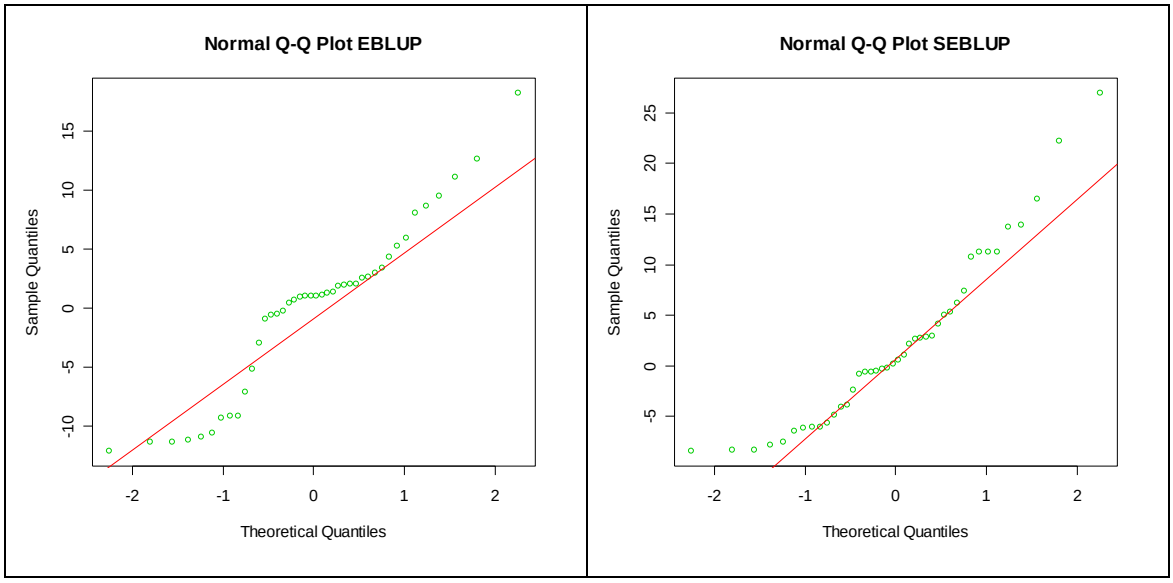
TABELLA 5.11. Stima dei parametri di EBLUP e Spatial EBLUP per la produzione delle Coltivazioni Legnose Agrarie per SEL.

	EBLUP			***Spatial EBLUP***		
	Valore	***Err. Std.***	***p-val.***	***Valore***	***Err. Std.***	***p-val.***
Costante	11,22	2,91	0,000	8,66	3,46	0,016
β_1	0,0015	0,017	0,929	-0,005	0,011	0,660
σ_u^2 stimato	91,13	29,79	0,004	19,913	0,135	0,000
ρ stimato	-	-	-	0,885	15,712	0,955
Test Shapiro	0,94	-	0,02	0,92	-	0,007

La costante è statisticamente significativa, utilizzando un intervallo di confidenza del 95% per entrambi gli stimatori, mentre al 99% è significativa solo per lo stimatore EBLUP; β_1 non è significativo ne al 95% ne al 99% per entrambi. Dal coefficiente di regressione vediamo che non c'è un legame lineare tra la produzione di coltivazioni legnose agrarie e SAU in un modello lineare ad effetti misti. La stima di σ_u^2 è significativamente diversa da 0 sia per la stima EBLUP sia per lo Spatial EBLUP. La stima di ρ non è significativa, quindi nel modello proposto non si riscontra un effetto sulla produzione di coltivazioni legnose agrarie tra le SEL, come risulta dall'indice di Moran. Con il test Shapiro si verifica la normalità degli errori e si rifiuta l'ipotesi nulla di distribuzione normale degli effetti di area con una significatività del 95%, mentre non si rifiuta al 99% nel caso della stima EBLUP.

La distribuzione degli effetti di area si può rappresentare graficamente con il grafico "quantile-quantile":

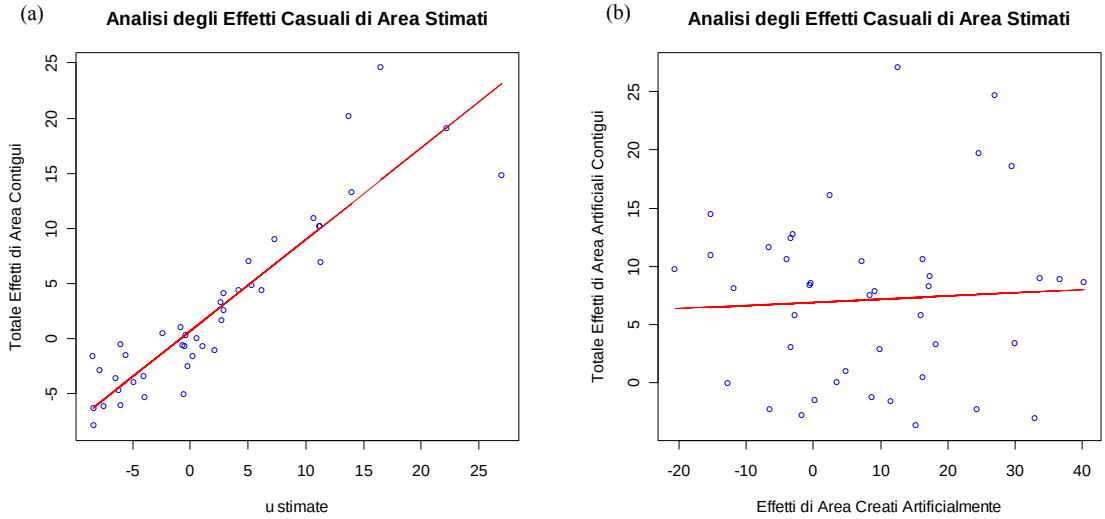
FIGURA 5.17. Rappresentazione degli effetti di area per le Coltivazioni Legnose Agrarie.



Gli effetti casuali di area seguono un andamento oscillatorio, indice di asimmetria nella loro distribuzione.

Verifichiamo la tendenza degli effetti casuali tra aree utilizzando le rappresentazioni grafiche proposte per i seminativi:

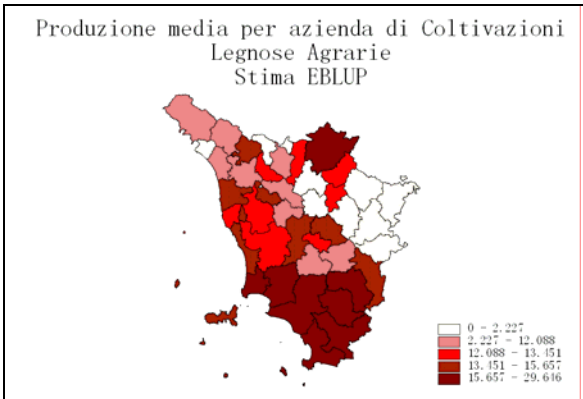
FIGURA 5.18. La stima degli effetti casuali (u) contro una media degli effetti casuali stimati per le aree contigue (a) e una media per aree scelte casualmente (b). Coltivazioni Legnose Agrarie

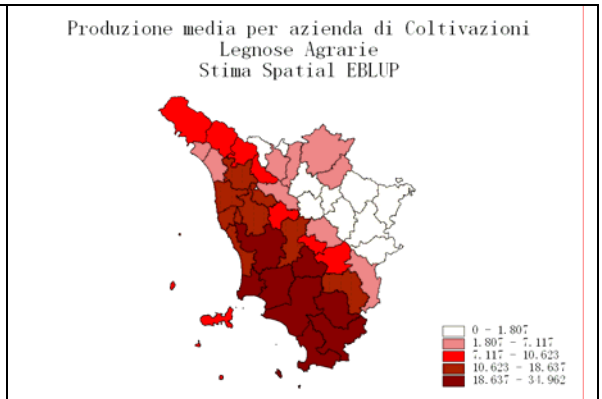


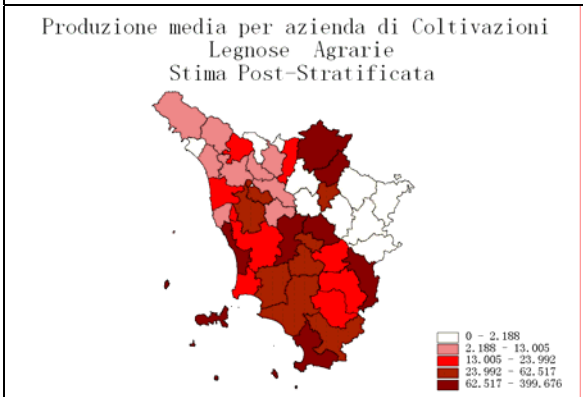
Il coefficiente angolare della retta ottenuta interpolando gli effetti di area con i minimi quadrati è non diversa da zero in figura (b) (dove le aree sono scelte casualmente) mentre è sicuramente positiva in figura (a). Questo ci induce a pensare che ci si autocorrelazione spaziale positiva tra i dati. Dalla dispersione dei punti intorno alla retta nel grafico in Figura 5.18 (a) si intuisce che c'è molta variabilità negli effetti di area stimati; per questo motivo nel modello non si ottiene una stima significativa di ρ.

Georeferenziamo le stime di produzione media per azienda in ogni SEL. In questo modo è possibile vedere sia quali sono le aree interessate nella produzione di coltivazioni legnose agrarie, sia le differenze nei risultati tra i tre modelli di stima: classica, EBLUP e Spatial EBLUP.

FIGURA 5.19. Produzione media di Coltivazioni Legnose Agrarie per azienda per SEL, anno 2003.







Stimatore	*MSE Medio*	*Std. Err. Medio*
Post-Stratificato	1467,49	20,82
EBLUP	47,11	5,72
Spatial EBLUP	17,81	3,69

Media del MSE e dell'errore standard per i tre stimatori.

Per le coltivazioni legnose agrarie abbiamo ottenuto, nella stima Spatial EBLUP, un ρ positivo uguale a circa 0,9 (non significativo) da cui deriva la differenza nei valori medi di produzione tra EBLUP e Spatial EBLUP. Considerando lo standard error medio dello Spatial EBLUP si potrebbe concludere che questo stimatore produca i risultati migliori, bisogna considerare, però, che il ρ non è significativo e che secondo l'indice di Moran non c'è autocorrelazione spaziale nella produzione considerata; per questo la stima EBLUP risulta più vicina alle ipotesi del modello, considerando anche la normalità degli effetti di casuali di area con un test con significatività al 99%.

Dopo aver presentato le stime sulla produzione del settore dei seminativi e delle coltivazioni legnose agrarie scendiamo più in dettaglio stimando la produzione dei cereali per il settore dei seminativi e la produzione di olive e vite per il settore delle coltivazioni legnose agrarie.

CEREALI

Per vedere se c'è autocorrelazione spaziale tra le SEL per la produzione di cereali utilizziamo l'indice di Moran:

TABELLA 5.12. Indice di Moran per la produzione di Cereali.

Statistica I_M	***Dev. Std.***	***p-val.***
0,487	5,16	0,000

In base ai valori ottenuti si rifiuta l'ipotesi nulla in favore dell'ipotesi alternativa di presenza di autocorrelazione spaziale, che in base alla statistica di Moran è positiva.

Presentiamo, riassunti in una tabella, la stima dei parametri di EBLUP e Spatial EBLUP e i relativi errori standard:

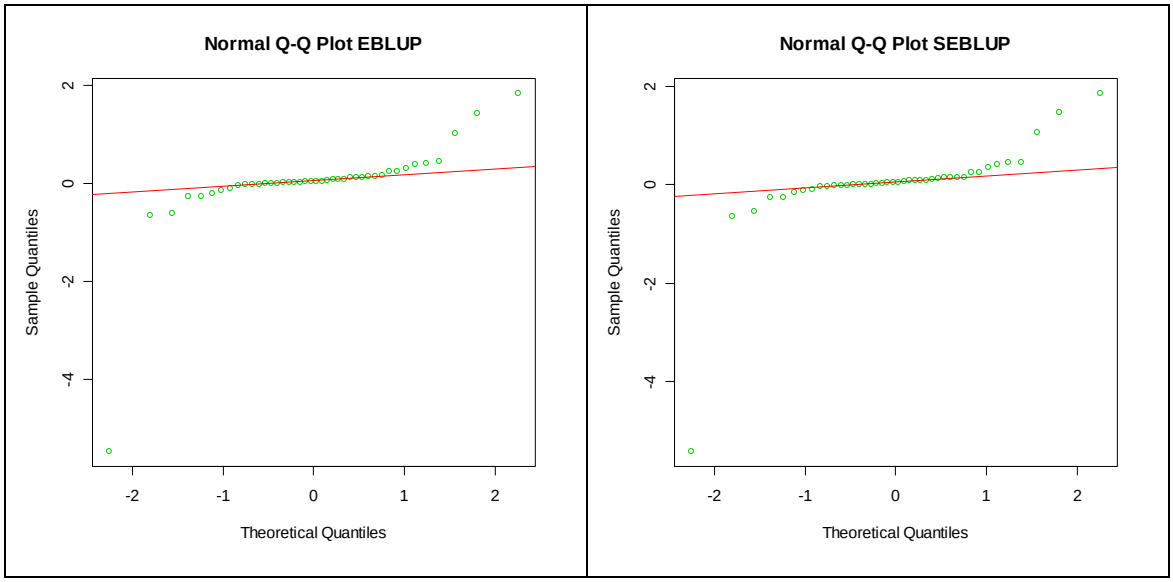
TABELLA 5.13. Stima dei parametri di EBLUP e Spatial EBLUP per la produzione di Cereali per SEL.

	EBLUP			***Spatial EBLUP***		
	Valore	***Err. Std.***	***p-val.***	***Valore***	***Err. Std.***	***p-val.***
Costante	-0,76	1,23	0,540	-0,80	1,24	0,520
β_1	0,26	0,03	0,000	0,26	0,027	0,000
σ_u^2 stimato	7,05	4,73	0,144	7,003	0,21	0,000
ρ stimato	-	-	-	0,046	1,02	0,964
Test Shapiro	0,49	-	0,000	0,49	-	0,000

La costante non è statisticamente significativa, utilizzando un intervallo di confidenza sia del 95% sia del 99% per entrambi gli stimatori; al contrario β_1 è significativo al 95% e al 99% sia per EBLUP sia per Spatial EBLUP. Dal coefficiente di regressione deduciamo che un ettaro di SAU per i cereali produce, in media ed escludendo gli effetti dovuti alla SEL di appartenenza, circa 0,26 quintali di cereali; questo valore è peraltro quasi uguale al β_1 stimato per il settore dei seminativi. La stima di σ_u^2 non è significativamente diversa da 0 per la stima EBLUP, mentre per lo Spatial EBLUP è significativa sia al 95%, sia al 99%. La stima di ρ non è significativa, quindi nel modello proposto non si riscontra un effetto sulla produzione di cereali tra le SEL, in contrasto con i risultati ottenuti dall'indice di Moran. Si rifiuta l'ipotesi nulla di distribuzione normale degli effetti di area con una significatività sia del 95%, sia del 99%, per entrambe le stime.

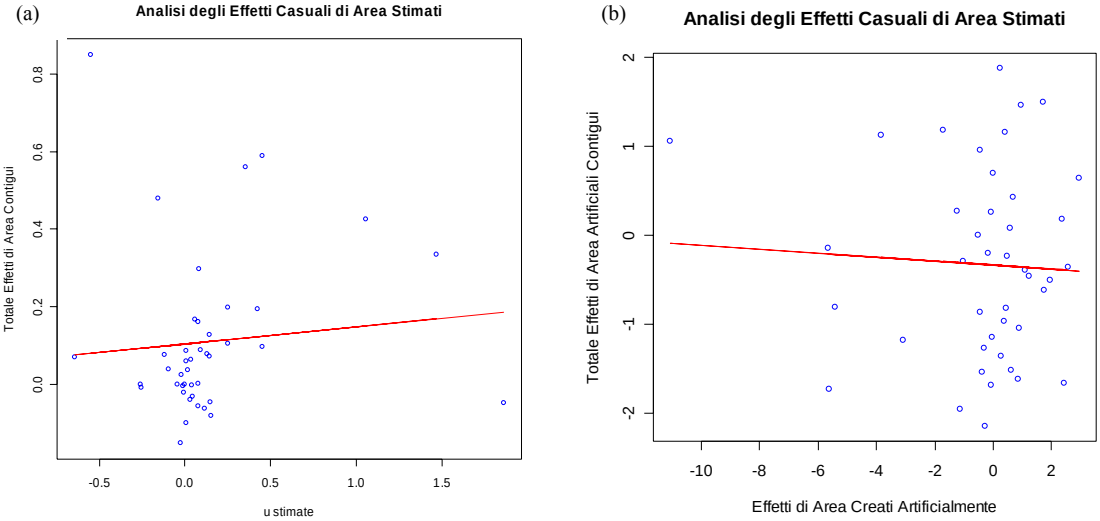
Osserviamo la distribuzione degli effetti di area con il grafico "quantile-quantile":

FIGURA 5.20. Rappresentazione degli effetti di area per i Cereali.



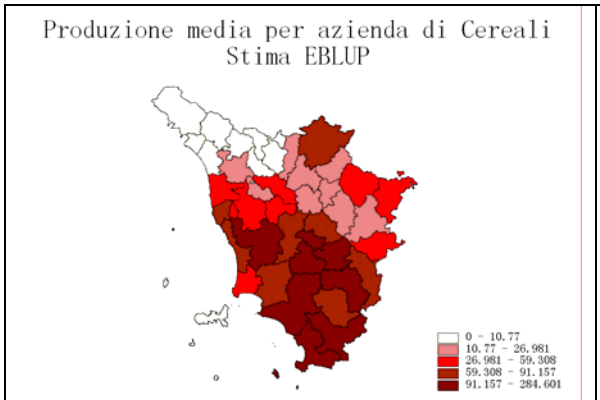
Gli effetti casuali di area, che in media sono molto prossimi a zero, seguono un andamento non lineare. I due modelli mostrano effetti casuali di area stimati quasi identici, ciò è dovuto ad un ρ circa uguale a 0 nella stima Spatial EBLUP.
Verifichiamo la tendenza degli effetti casuali tra aree utilizzando le rappresentazioni grafiche ormai note:

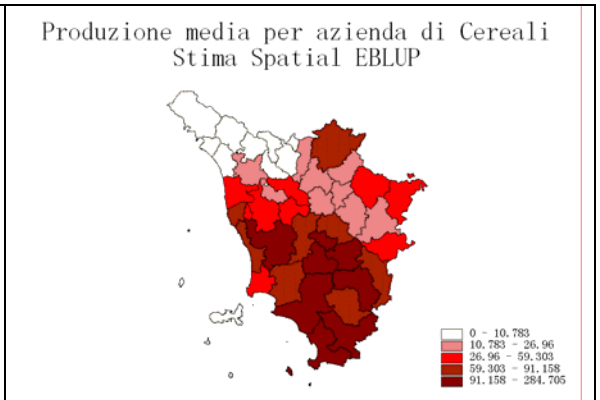
FIGURA 5.21. La stima degli effetti casuali (u) contro una media degli effetti casuali stimati per le aree contigue (a) e una media per aree scelte casualmente (b). Cereali.

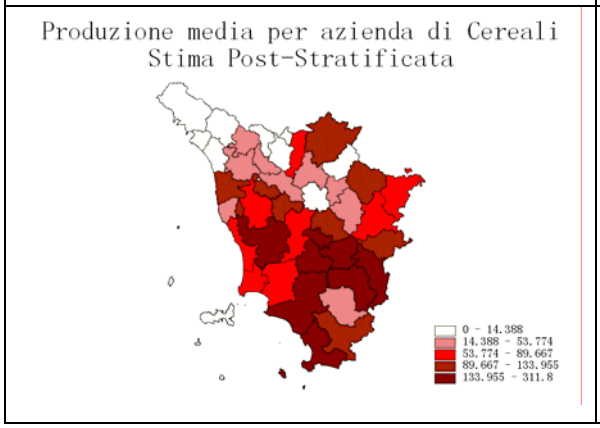


Il coefficiente angolare della retta ottenuta interpolando gli effetti di area con i minimi quadrati è non diversa da zero sia in figura (b) sia in figura (a). Questo ci induce a pensare che non ci si autocorrelazione spaziale tra i dati, in sintonia con la stima di ρ. Presentiamo la rappresentazione geografica delle stime di produzione media di cereali per azienda in ogni SEL, in modo da cogliere le differenze nei risultati tra i tre modelli di stima: classica, EBLUP e Spatial EBLUP.

FIGURA 5.22. Produzione media di Cereali per azienda per SEL, anno 2003.







Stimatore	*MSE Medio*	*Std. Err. Medio*
Post-Stratificato	2890,47	42,70
EBLUP	70,68	6,45
Spatial EBLUP	71,33	6,55

Media del MSE e dell'errore standard per i tre stimatori.

Per i cereali abbiamo ottenuto, nella stima Spatial EBLUP, un ρ circa uguale 0, per questo con i metodi EBLUP e Spatial EBLUP si ottengono gli stessi risultati. Il MSE medio dell'EBLUP è minore, anche se di poco, di quello dello stimatore spaziale: risultato che conferma gli studi di simulazione fatti nel capitolo 4. Infatti la componente g_3 del MSE dello Spatial EBLUP, che rappresenta la variabilità di ρ e σ_u^2, è sempre maggiore di quella dell'EBLUP, che rappresenta solo la variabilità di σ_u^2; quando il ρ non apporta informazioni sulla variabilità degli effetti casuali di area la componente g_1, che li rappresenta, non cambia tra i due stimatori ed il risultato finale è rappresentato come in questo caso da una maggiore efficienza della stima EBLUP. Bisogna sottolineare, comunque, come la perdita di efficienza della stima Spatial EBLUP è

minima, in presenza di ρ uguali a 0, contro un guadagno notevole in termini di efficienza che si ottiene nei casi in cui ρ è diverso da 0.
Con la stima post-stratificata si ottengono valori medi maggiori rispetto a quelli ottenuti con EBLUP e Spatial EBLUP. Con il MSE medio ottenuto non siamo in grado di offrire stime fruibili, ci limitiamo ad usarlo come paragone per vedere la maggior efficienza degli stimatori presentati.

VITE

Per vedere se c'è autocorrelazione spaziale tra le SEL per la produzione di vite si utilizza l'indice di Moran:

TABELLA 5.14. Indice di Moran per la produzione di Vite.

Statistica I_M	***Dev. Std.***	***p-val.***
0,001	0,27	0,790

In base ai valori ottenuti non si rifiuta l'ipotesi nulla di assenza di autocorrelazione spaziale.

Presentiamo, riassunti in una tabella, la stima dei parametri di EBLUP e Spatial EBLUP e i relativi errori standard:

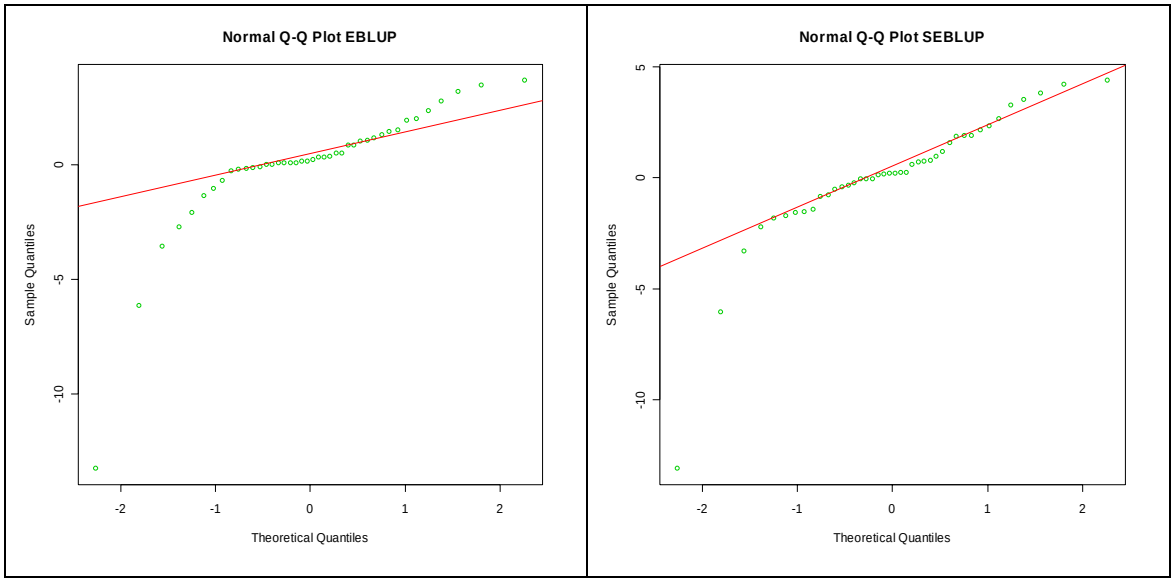
TABELLA 5.15. Stima dei parametri di EBLUP e Spatial EBLUP per la produzione di Vite per SEL.

	EBLUP			***Spatial EBLUP***		
	Valore	***Err. Std.***	***p-val.***	***Valore***	***Err. Std.***	***p-val.***
Costante	-3,72	1,86	0,052	-4,43	2,13	0,044
B_1	0,64	0,09	0,000	0,67	0,10	0,000
σ_u^2 stimato	20,49	8,57	0,022	20,23	0,12	0,000
P stimato	-	-	-	0,31	2,55	0,904
Test Shapiro	0,71	-	0,000	0,96	-	0,169

La costante è statisticamente significativa, utilizzando un intervallo di confidenza del 95% per lo Spatial EBLUP, mentre al 99% non è significative per nessuno dei due; β_1 è significativo al 95% e al 99% per entrambi gli stimatori. Dal coefficiente di regressione vediamo che la produzione media per azienda di un ettaro di SAU è circa 0,6 quintali (non varia rispetto agli stimatori) di uva. La stima di σ_u^2 è significativamente diversa da 0 sia per la stima EBLUP sia per lo Spatial EBLUP, con intervallo di confidenza al

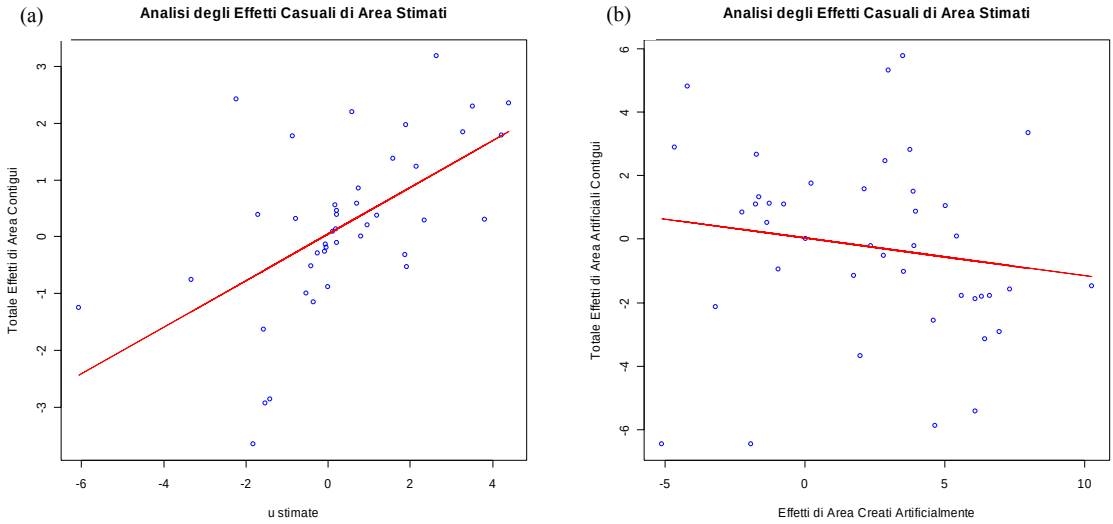
95%; ampliando tale intervallo al 99% risulta significativa solo per lo Spatial EBLUP. La stima di ρ non è significativa, quindi nel modello proposto non si riscontra un effetto sulla produzione di vite tra le SEL, come risulta dall'indice di Moran. Gli effetti di area sono distribuiti normalmente nel modello Spatial EBLUP, sia con intervallo di confidenza del 95%, sia del 99%, accade il contrario per la stima EBLUP.
Rappresentiamo gli effetti di area con il grafico "quantile-quantile":

FIGURA 5.23. Rappresentazione degli effetti di area per le Viti.



Gli effetti casuali di area seguono un andamento oscillatorio, indice di asimmetria nella loro distribuzione; tale tendenza è molto meno accentuata nel grafico "Normal Q-Q Plot SEBLUP", relativo alla stima Spatial EBLUP: in questo caso gli effetti si considerano distribuiti normalmente.
Verifichiamo la tendenza degli effetti casuali tra aree utilizzando le rappresentazioni grafiche ormai note:

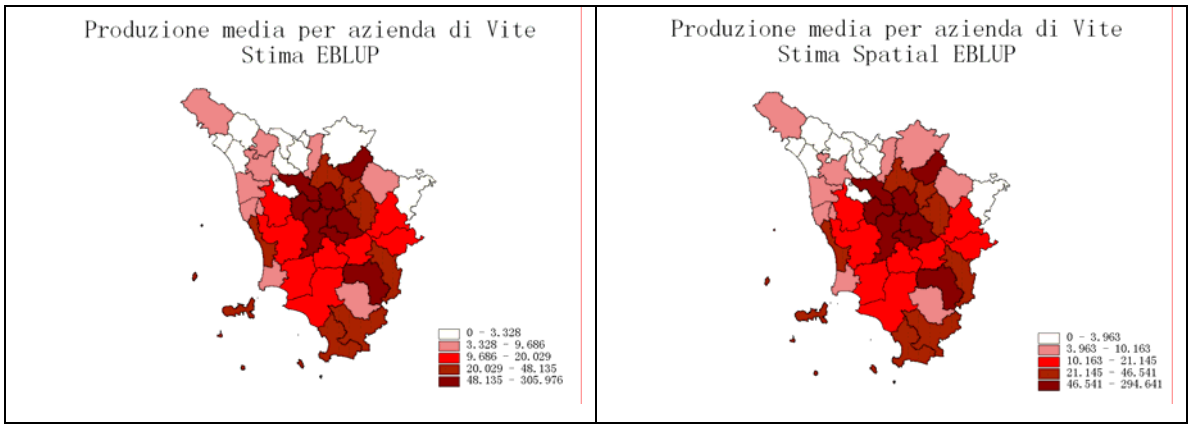
FIGURA 5.24. La stima degli effetti casuali (u) contro una media degli effetti casuali stimati per le aree contigue (a) e una media per aree scelte casualmente (b). Viti.

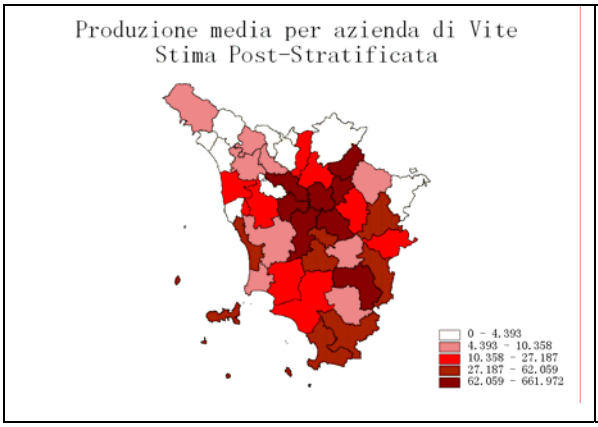


Il coefficiente angolare della retta ottenuta interpolando gli effetti di area con i minimi quadrati è all'incirca zero in figura (b) (dove le aree sono scelte casualmente) mentre è sicuramente positiva in figura (a). Questo ci indurrebbe a pensare che ci si autocorrelazione spaziale positiva tra i dati, anche se non molto accentuata. Dalla dispersione dei punti intorno alla retta nel grafico in Figura 5.24 (a) si intuisce che c'è molta variabilità negli effetti di area stimati; per questo motivo nel modello non si ottiene una stima significativa di ρ.

Utilizzando la rappresentazione su carta geografica utilizzata precedentemente raffiguriamo la produzione media di vite per azienda in ogni SEL.

FIGURA 5.25. Produzione media di Vite per azienda per SEL, anno 2003.





Stimatore	MSE Medio	Std. Err. Medio
Post-Stratificato	21304,25	45,63
EBLUP	87,63	6,00
Spatial EBLUP	94,72	6,24

Media del MSE e dell'errore standard per i tre stimatori.

Per la vite riscontriamo delle differenza nei valori medi di produzione tra EBLUP e Spatial EBLUP: con quest'ultimo si ottengono valori medi per SEL distribuiti sul territorio in modo graduato ("smooth"). Considerando lo standard error medio dell'EBLUP si nota come questo stimatore produca i risultati migliori; infatti ρ non è significativo ed in questi casi lo stimatore spaziale non produce le stime più efficienti. Per la stima della produzione media di vite per azienda per SEL, lo stimatore EBLUP fornisce i risultati migliori in termini di efficienza però gli effetti di area non sono distribuiti normalmente, da cui si deduce che il modello è mal posto. Al contrario per le stime ottenute con lo Spatial EBLUP gli effetti di area risultano normalmente distribuiti, per questo motivo, dato che in termini di errore standard medio non ci sono differenze sostanziali, è meglio fare affidamento su quest'ultime.

OLIVE

Verifichiamo la presenza di autocorrelazione spaziale tra le SEL per la produzione di olive con l'indice di Moran:

TABELLA 5.16. Indice di Moran per la produzione di Olive.

Statistica I_M	***Dev. Std.***	***p-val.***
0,05	0,72	0,471

In base ai valori ottenuti non si rifiuta l'ipotesi nulla di assenza di autocorrelazione spaziale.

Presentiamo, riassunti in una tabella, la stima dei parametri di EBLUP e Spatial EBLUP e i relativi errori standard:

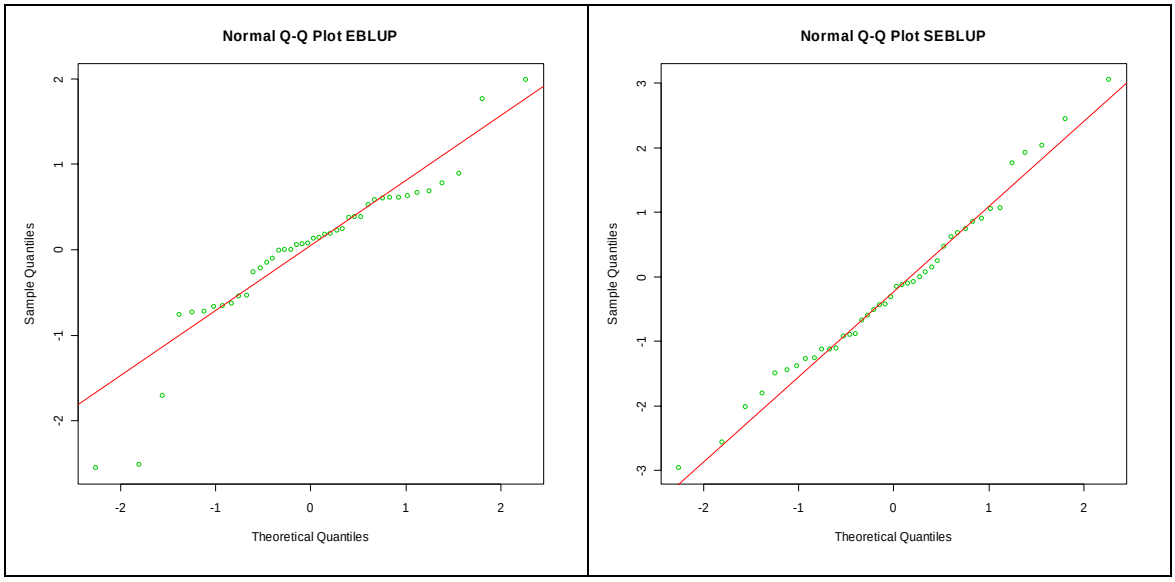
TABELLA 5.17. Stima dei parametri di EBLUP e Spatial EBLUP per la produzione di Olive per SEL.

	EBLUP			***Spatial EBLUP***		
	Valore	***Err. Std.***	***p-val.***	***Valore***	***Err. Std.***	***p-val.***
Costante	0,24	0,53	0,649	0,06	0,79	0,943
β_1	0,069	0,01	0,000	0,075	0,01	0,000
σ_u^2 ***stimato***	1,90	0,78	0,020	1,64	1,51	0,282
ρ ***stimato***	-	-	-	0,71	6,63	0,916
Test Shapiro	0,91	-	0,002	0,98	-	0,861

La costante non è statisticamente significativa, utilizzando un intervallo di confidenza sia del 95% sia del 99% per nessuno dei due stimatori; β_1, al contrario, è significativo al 95% e al 99% per EBLUP e Spatial EBLUP. Dal coefficiente di regressione vediamo che la produzione media per azienda di un ettaro di SAU è circa 0,7 quintali (non varia molto rispetto ai due stimatori) di olive. La stima di σ_u^2 è significativamente diversa da 0 per lo stimatore EBLUP solo con un intervallo di confidenza del 95%, mentre nel caso dello Spatial EBLUP σ_u^2 stimato non è significativo (ne al 95% ne al 99%). Anche la stima di ρ non è significativa: nonostante il valore di ρ stimato sia molto alto (0,71) l'elevata variabilità rende instabile tale stima. Gli effetti di area sono distribuiti normalmente nel modello Spatial EBLUP, sia con intervallo di confidenza del 95%, sia del 99%, contrariamente alla stima EBLUP.

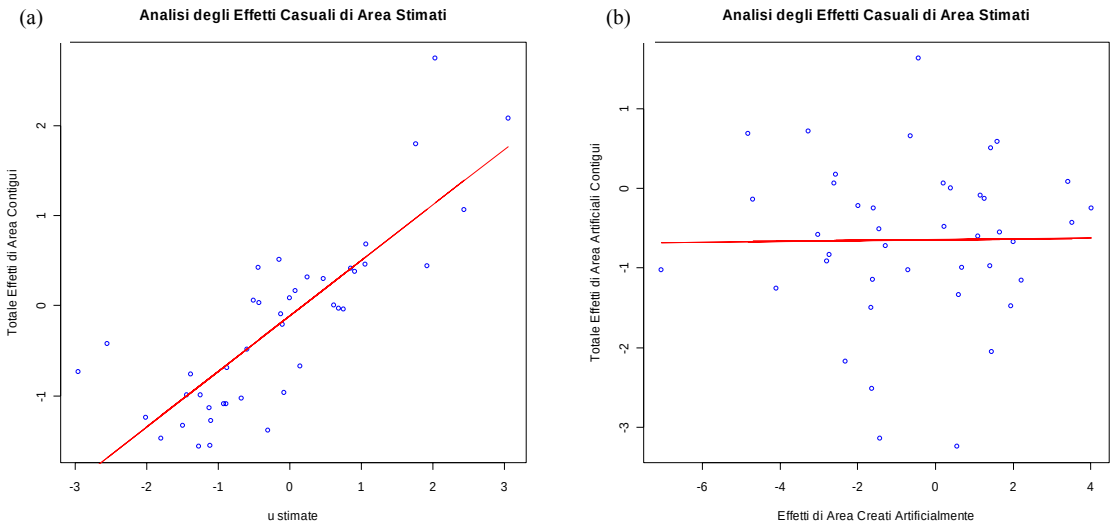
Rappresentiamo gli effetti di area con il grafico "quantile-quantile":

FIGURA 5.26. Rappresentazione degli effetti di area per le Olive.



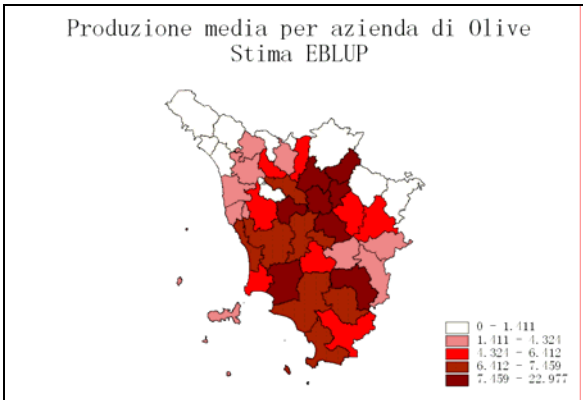
Gli effetti casuali di area seguono un andamento oscillatorio, indice di asimmetria nella loro distribuzione nel caso della stima EBLUP. Tale tendenza è molto meno accentuata nel caso della stima Spatial EBLUP: in questo caso gli effetti casuali di area si considerano distribuiti normalmente (come risulta, in modo rigoroso, dal test Shapiro).
Verifichiamo la tendenza degli effetti casuali tra aree utilizzando le rappresentazioni grafiche ormai note:

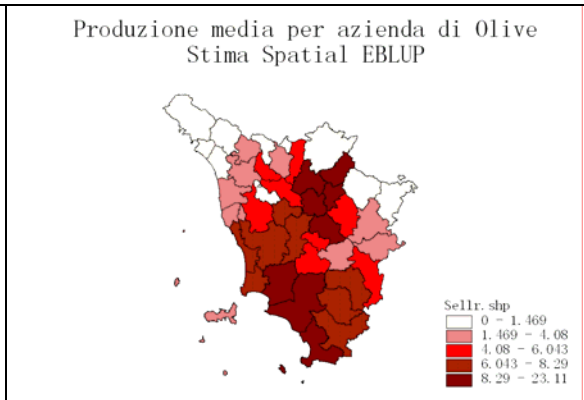
FIGURA 5.27. La stima degli effetti casuali (u) contro una media degli effetti casuali stimati per le aree contigue (a) e una media per aree scelte casualmente (b). Olive.

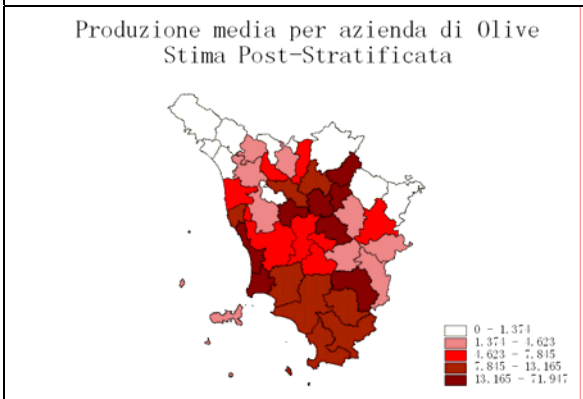


Il coefficiente angolare della retta ottenuta interpolando gli effetti di area con i minimi quadrati è all'incirca zero in figura (b) (dove le aree sono scelte casualmente) mentre è sicuramente positiva in figura (a). Questo ci indurrebbe a pensare che ci si autocorrelazione spaziale positiva tra i dati. Dalla dispersione dei punti intorno alla retta nel grafico in Figura 5.27 (a) si vede che c'è molta variabilità negli effetti di area stimati; infatti, nel modello non si ottiene una stima significativa di ρ. In definitiva c'è una tendenza tra gli effetti di area per SEL contigue, ma tale tendenza è molto variabile per cui non si può utilizzare in modo proficuo.
Utilizzando la rappresentazione su carta geografica utilizzata precedentemente raffiguriamo la produzione media di olive per azienda in ogni SEL.

FIGURA 5.28. Produzione media di Olive per azienda per SEL, anno 2003.







Stimatore	MSE Medio	Std. Err. Medio
Post-Stratificato	2649,28	12,21
EBLUP	1,85	1,19
Spatial EBLUP	2,12	1,26

Media del MSE e dell'errore standard per i tre stimatori.

Per la produzione di olive riscontriamo delle differenze minime nei valori medi di EBLUP e Spatial EBLUP. Considerando lo standard error medio dell'EBLUP si nota come questo stimatore produca i risultati migliori, anche se la differenza rispetto allo standard error medio dello stimatore spaziale è minima. Bisogna considerare che gli effetti di area per la stima Spatial EBLUP sono distribuiti normalmente, al contrario dell'EBLUP, quindi nonostante le stime di alcuni parametri non siano significative il modello non è mal posto. Conviene per questo motivo utilizzare le stime ottenuto con lo stimatore Spatial EBLUP.

5.4.3 NOTE CONCLUSIVE

Gli stimatori EBLUP e Spatial EBLUP sono stati applicati alla stima per SEL nell'indagine SPA. L'applicazione a dati reali ha confermato i risultati della simulazione presentata nel capitolo 4 questi stimatori sono risultati nettamente più efficienti rispetto allo stimatore post-stratificato, ed in caso di autocorrelazione spaziale dei dati lo Spatial EBLUP ha ottenuto le performance migliori considerando gli indici utilizzati (ARB, ARE, EFF, RRMSE). Il comportamento degli stimatori non è diverso utilizzando i dati relativi all'indagine SPA: la stima post-stratificata è nettamente

peggiore, in termini di efficienza, rispetto alle stime ottenute con i modelli EBLUP e Spatial EBLUP; questo è vero in particolar modo in alcune SEL. Tra gli stimatori EBLUP e Spatial EBLUP, come ci aspettavamo, in presenza di autocorrelazione spaziale lo Spatial EBLUP è migliore in termini di efficienza (cfr. appendice D).

Nella stima della produzione media per azienda per SEL delle coltivazioni più diffuse in Toscana abbiamo riscontrato un'elevata variabilità degli effetti di area, ciò non ha permesso di ottenere, nella maggior parte dei casi, stime significative di ρ e σ_u^2. Probabilmente si sono riscontrate delle difficoltà ad adattare il modello lineare ad effetti misti ai dati poiché l'annata agricola 2003 è stata caratterizzata dal maltempo (siccità e alluvioni), per questo le produzioni sono risultate molto variabili. Inoltre il maltempo non ha colpito uniformemente il territorio, creando così un ulteriore fattore di variabilità. Dal momento che questi problemi stanno diventando sempre più frequenti, una soluzione potrebbe essere quella di riuscire ad inserire nel modello di stima un fattore di variabilità legato all'andamento climatico.

Considerando i risultati della simulazione e i risultati ottenuti sulla stima della produzione media per azienda per SEL in toscana, risulta evidente come i metodi di stima per piccole aree, EBLUP e Spatial EBLUP, siano indispensabili per ottenere risultati fruibili relativi a domini ristretti, fondamentali per una conoscenza approfondita delle tematiche legate al territorio. Tale conoscenza è sicuramente un fattore di successo nella scelta delle politiche di intervento operate dalle pubbliche amministrazioni, sia centrali sia locali. Si può ben immaginare come queste metodologie possono essere esportate nel "mondo aziendale", dove informazioni dettagliate, ottenute ad un costo contenuto, possono risultare un fattore di vantaggio determinante nella competizione che si sta creando nel mercato globale.

Gli stimatori presentati in questa tesi sono soggetti, come tutti i modelli, ad ipotesi restrittive che non sempre trovano adeguata conferma nei casi reali. Nel nostro caso abbiamo ipotizzato di conoscere senza errore la varianza campionaria a livello di piccola area, ipotesi che spesso non è rispettata nell'applicazione a dati reali. Ciò sottolinea l'importanza di continuare la ricerca in questo campo, in modo da ottenere stimatori sempre più efficienti e applicabili ai dati disponibili nella maggior parte delle indagini.

APPENDICE A

APPROFONDIMENTI SUGLI STIMATORI EBLUP E SPATIAL EBLUP

Si riportano, in questa appendice, le derivazioni utilizzate per ottenere gli stimatori EBLUP e Spatial EBLUP.

- Valore atteso e varianza di $\boldsymbol{y}$ nel modello lineare ad effetti misti (pag. 29):

$$E[\boldsymbol{y}] = E[\boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}] = E[\boldsymbol{X\beta}] + E[\boldsymbol{Zu}] + E[\boldsymbol{e}] = \boldsymbol{X\beta} + \boldsymbol{Z}E[\boldsymbol{u}] + E[\boldsymbol{e}] = \boldsymbol{X\beta}$$

$$V(\boldsymbol{y}) = E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T] = E[(X\beta + Zu + e - X\beta)(X\beta + Zu + e - X\beta)^T] =$$
$$= E[(\boldsymbol{Zu} + \boldsymbol{e})(\boldsymbol{u}^T\boldsymbol{Z}^T + \boldsymbol{e}^T)] = E[\boldsymbol{Zuu}^T\boldsymbol{Z}^T + \boldsymbol{Zue}^T + \boldsymbol{eu}^T\boldsymbol{Z}^T + \boldsymbol{ee}^T] =$$
$$\boldsymbol{Z}E[\boldsymbol{uu}^T]\boldsymbol{Z}^T + \boldsymbol{Z}E[\boldsymbol{ue}^T] + E[\boldsymbol{eu}^T]\boldsymbol{Z} + E[\boldsymbol{ee}^T] = \boldsymbol{ZGZ}^T + \sigma_e^2\boldsymbol{R}$$

- Componenti di varianza dello stimatore BLUP (pag. 31):

Il MSE dello stimatore di μ può essere scritto come:

$$V(\hat{\mu} - \mu) = E[((\hat{\mu} - \mu) - E[(\hat{\mu} - \mu)])^2] = E[((\hat{\mu} - \mu) - 0)^2] = E[(\hat{\mu} - \mu)^2] = MSE(\hat{\mu})$$

Quindi dobbiamo minimizzare $V(\hat{\mu} - \mu)$ per ottenere uno stimatore ottimo (tra gli stimatori corretti). Per le note proprietà della varianza risulta che:

$$V(\hat{\mu} - \mu) = V(\hat{\mu}) + V(\mu) - 2COV(\hat{\mu}, \mu) \qquad \text{(A.1)}$$

Analizziamo ogni componente singolarmente:

1ª componente:

$$V(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}])(\hat{\mu} - E[\hat{\mu}])^T] = E[(a^T\boldsymbol{y} + b - \boldsymbol{a}^T\boldsymbol{X\beta} - b)(a^T\boldsymbol{y} + b - \boldsymbol{a}^T\boldsymbol{X\beta} - b)^T] =$$

$$= E[(\boldsymbol{a}^T \boldsymbol{y} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}^T \boldsymbol{a} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a})] = E[\boldsymbol{a}^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{y}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{y}^T \boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a}] \tag{A.2}$$

Consideriamo separatamente per facilità di calcolo E[$\boldsymbol{a}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{a}$] (utilizzando la nota proprietà la somma dei valori attesi è ugale al valore atteso della somma):

$$E[\boldsymbol{a}^T \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{a}] = E[\boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e})(\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{u}^T \boldsymbol{Z}^T + \boldsymbol{e}^T)\boldsymbol{a}] =$$
$$= \boldsymbol{a}^T E[\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{u}^T \boldsymbol{Z}^T + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{e}^T + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{u}^T \boldsymbol{Z}^T + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{e}^T +$$
$$+ \boldsymbol{e}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{e}\boldsymbol{u}^T \boldsymbol{Z}^T + \boldsymbol{e}\boldsymbol{e}^T]\boldsymbol{a} = \boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + 0 + 0 + 0 + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + 0 + 0 + 0 + \boldsymbol{R})\boldsymbol{a} = {}^{51}$$
$$= \boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})\boldsymbol{a}$$

reinserendo nella (A.2) questa derivazione otteniamo:

$$\text{(A.2)} = \boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})\boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} =$$
$$= \boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{V})\boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a}$$

2ª componente:

$$V(\mu) = E[(\mu - E[\mu])(\mu - E[\mu])^T] = E[(\boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{u} - \boldsymbol{l}^T \boldsymbol{\beta})(\boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{u} - \boldsymbol{l}^T \boldsymbol{\beta})^T] =$$
$$E[(\boldsymbol{m}^T \boldsymbol{u})(\boldsymbol{m}^T \boldsymbol{u})^T] = E[\boldsymbol{m}^T \boldsymbol{u}\boldsymbol{u}^T \boldsymbol{m}] = \boldsymbol{m}^T E[\boldsymbol{u}\boldsymbol{u}^T]\boldsymbol{m} = \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m}$$

3ª componente:

$$COV(\hat{\mu}, \mu) = E[(\hat{\mu} - E[\hat{\mu}])(\mu - E[\mu])^T] = E[(\boldsymbol{a}^T \boldsymbol{y} + b - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta} - b)(\boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{u} - \boldsymbol{l}^T \boldsymbol{\beta})^T] =$$
$$= E[(\boldsymbol{a}^T \boldsymbol{y} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{u}^T \boldsymbol{m})] = E[\boldsymbol{a}^T \boldsymbol{y}\boldsymbol{u}^T \boldsymbol{m} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{u}^T \boldsymbol{m}] = \boldsymbol{a}^T E[\boldsymbol{y}\boldsymbol{u}^T]\boldsymbol{m} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta} E[\boldsymbol{u}^T]\boldsymbol{m} =$$
$$= \boldsymbol{a}^T E[(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e})\boldsymbol{u}^T]\boldsymbol{m} - 0 = \boldsymbol{a}^T E[\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{u}^T + \boldsymbol{Z}\boldsymbol{u}\boldsymbol{u}^T + \boldsymbol{e}\boldsymbol{u}^T]\boldsymbol{m} =$$
$$= \boldsymbol{a}^T (0 + \boldsymbol{Z}\boldsymbol{G} + 0)\boldsymbol{m} = \boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$$

Riprendendo la (A.1) e le tre componenti di varianza possiamo scrivere:

$$V(\hat{\mu} - \mu) = \boldsymbol{a}^T (\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T + \boldsymbol{V})\boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} =$$

[51] Per semplificare la scrittura si considera $\sigma_{ei}^2 \boldsymbol{R} = \boldsymbol{R}$, quindi si assume $\boldsymbol{R}$ incognita a causa dei valori sulla diagonale $\sigma_{ei}^2 \boldsymbol{R}_{ii}$.

$$= \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} - \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$$

- Minimizzazione con il moltiplicatore di Lagrange della varianza dello stimatore BLUP (pag. 32)

$$h(x,\lambda) = f(x) + \sum_{k=1}^{n} g_k(x)$$

Dove i vincoli $g_k(x)$ devono essere scritti nella forma $g_k(x) = 0$. La nostra funzione lagrangiana sarà dunque:

$$f(x) \to f(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$$

$$g_1(x) \to g_1(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{l}^T = 0$$

Quindi $h(x,\lambda) \to h(\boldsymbol{a},\lambda) = \boldsymbol{a}^T \boldsymbol{V}\boldsymbol{a} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{m} - 2\boldsymbol{a}^T \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} + (\boldsymbol{a}^T \boldsymbol{X} - \boldsymbol{l}^T)\lambda$

Derivando la funzione lagrangiana rispetto al vettore $\boldsymbol{a}$ e ponendola uguale a 0 si trova il valore di λ che sostituito in $\boldsymbol{a}$ ci fornisce le stime ottime di $\boldsymbol{\beta}$ ed $\boldsymbol{u}$ (si ricorda $\hat{\mu} = \boldsymbol{a}^T \boldsymbol{y} + b$):

1. $\dfrac{\partial h(\boldsymbol{a},\lambda)}{\partial \boldsymbol{a}} = 2\boldsymbol{V}\boldsymbol{a} - 2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} + 2\boldsymbol{X}\lambda$ [52]

2. $\dfrac{\partial h(\boldsymbol{a},\lambda)}{\partial \boldsymbol{a}} = 0 \Rightarrow 2\boldsymbol{V}\boldsymbol{a} - 2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} + 2\boldsymbol{X}\lambda = 0 \to \boldsymbol{V}\boldsymbol{a} + \boldsymbol{X}\lambda = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$

3. dalla 2. ricavo $\boldsymbol{a}$: $\boldsymbol{V}\boldsymbol{a} = -\boldsymbol{X}\lambda + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} \to \boldsymbol{a} = -\boldsymbol{V}^{-1}\boldsymbol{X}\lambda + \boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$

4. Sostituendo la 3. nel vincolo $\boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{l}^T$ ricavo il valore di λ:

   $a^T X = l^T \to (-\boldsymbol{V}^{-1}\boldsymbol{X}\lambda + \boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{m})^T \boldsymbol{X} = \boldsymbol{l}^T \to (-\lambda^T \boldsymbol{X}^T \boldsymbol{V}^{-1^T} + \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{Z}^T \boldsymbol{V}^{-1^T})\boldsymbol{X} = \boldsymbol{l}^T \to$

   $\to -\boldsymbol{\lambda}^T \boldsymbol{X}^T \boldsymbol{V}^{-1^T} \boldsymbol{X} + \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{Z}^T \boldsymbol{V}^{-1^T} \boldsymbol{X} - \boldsymbol{l}^T = 0 \to \lambda^T =$

[52] In realtà per ottenere $2\boldsymbol{X}\lambda$ utilizzo un moltiplicatore di lagrange invece che uguale a λ uguale a 2λ, poiché questo non altera i risultati e facilita i calcoli poiché possiamo dividere la derivata per 2.

$$= \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{Z}^T \boldsymbol{V}^{-1^T} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1^T} \boldsymbol{X})^{-1} - \boldsymbol{l}^T (\boldsymbol{X}^T \boldsymbol{V}^{-1^T} \boldsymbol{X})^{-1} \rightarrow$$

$$\rightarrow \lambda = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} - (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{l}$$

5. Sostituendo la 4. nella 2. si ottiene:

$$\boldsymbol{a} = -\boldsymbol{V}^{-1} \boldsymbol{X} [-(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{l} + (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}] + \boldsymbol{V}^{-1} \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} =$$

$$= \boldsymbol{V}^{-1} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{l} - \boldsymbol{V}^{-1} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m} + \boldsymbol{V}^{-1} \boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}$$

6. Infine risulta:

$$\boldsymbol{a}^T = \boldsymbol{l}^T (\boldsymbol{X}^T \boldsymbol{V}^{-1^T} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1^T} - \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{Z}^T \boldsymbol{V}^{-1^T} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1^T} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1^T} + \boldsymbol{m}^T \boldsymbol{G}^T \boldsymbol{Z}^T \boldsymbol{V}^{-1^T}$$

La 6. può essere riscritta considerando che le matrici $\boldsymbol{G}$ e $\boldsymbol{V}$, essendo matrici di varianza-covarianza, sono simmetriche, quindi $\boldsymbol{G}^T = \boldsymbol{G}$ e $\boldsymbol{V}^T = \boldsymbol{V}$:

$$\boldsymbol{a}^T = \boldsymbol{l}^T (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1}$$

7. Sotto i vincoli di correttezza $\boldsymbol{a}^T \boldsymbol{X} = \boldsymbol{l}^T$ e b = 0 possiamo scrivere:

$$\hat{\mu} = \boldsymbol{a}^T \boldsymbol{y} + b = \boldsymbol{l}^T (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{y} =$$

$$= \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X} \hat{\boldsymbol{\beta}} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{y} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Quindi $\hat{\mu} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$

- MSE dello stimatore BLUP, MSE[t($\boldsymbol{\delta}$,$\boldsymbol{y}$)] (pag. 36):

1. Si riscrive il modello $t(\boldsymbol{\delta},\boldsymbol{y})$ in modo da scomporlo in due componenti:

$$t(\boldsymbol{\delta}, \boldsymbol{y}) = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} + \boldsymbol{m}^T \hat{\boldsymbol{u}} = \boldsymbol{l}^T \hat{\boldsymbol{\beta}} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) =$$

$$= \boldsymbol{l}^T \hat{\boldsymbol{\beta}} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \boldsymbol{l}^T \boldsymbol{\beta} - \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta} =$$

$$= \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{l}^T \hat{\boldsymbol{\beta}} - \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X}\hat{\boldsymbol{\beta}} =$$

$$= \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{l}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) =$$

$$= \boldsymbol{l}^T \boldsymbol{\beta} + \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{l}^T - \boldsymbol{m}^T \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{V}^{-1} \boldsymbol{X}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

Dove $\boldsymbol{l}^T\boldsymbol{\beta}+\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})$ è la prima componente che può essere interpretata come lo stimatore BLUP reso noto $\boldsymbol{\beta}$: $t^*(\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{\beta})$ e $(\boldsymbol{l}^T-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ è la seconda parte, che possiamo interpretare come una componente aggiuntiva per conoscere $\boldsymbol{\beta}$.

Riassumendo $t(\boldsymbol{\delta},\boldsymbol{y})=t^*(\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{\beta})+(\boldsymbol{l}^T-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$

2. Si calcola il MSE($t(\boldsymbol{\delta},\boldsymbol{y})$) rispetto alla scomposizione fatta sopra:

$$MSE(t(\boldsymbol{\delta},\boldsymbol{y}))=MSE[t(\boldsymbol{\delta},\boldsymbol{y},\beta)]+V[(\boldsymbol{l}^T-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]=g_1(\boldsymbol{\delta})+g_2(\boldsymbol{\delta})$$

Per semplificare la presentazione delle formule si considera:

- $\boldsymbol{b}^T=\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}$
- $\boldsymbol{d}^T=\boldsymbol{l}^T-\boldsymbol{b}^T\boldsymbol{X}=\boldsymbol{l}^T-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{X}$

Con $\boldsymbol{b}^T$ e $\boldsymbol{d}^T$ costanti rispetto al valore atteso.

3. Si calcola la prima componente:

$$\text{MSE}[t^*(\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{\beta})]=g_1(\boldsymbol{\delta})=E[(t(\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{\beta})-\boldsymbol{\mu})(t(\boldsymbol{\delta},\boldsymbol{y},\boldsymbol{\beta})-\boldsymbol{\mu})^T]=$$

$$=E[(\boldsymbol{l}^T\boldsymbol{\beta}+\boldsymbol{b}^T(\boldsymbol{y}-\boldsymbol{X\beta})-\boldsymbol{l}^T\boldsymbol{\beta}-\boldsymbol{m}^T\boldsymbol{u})(\boldsymbol{l}^T\boldsymbol{\beta}+\boldsymbol{b}^T(\boldsymbol{y}-\boldsymbol{X\beta})-\boldsymbol{l}^T\boldsymbol{\beta}-\boldsymbol{m}^T\boldsymbol{u})^T]=$$

$$=E[(\boldsymbol{b}^T(\boldsymbol{y}-\boldsymbol{X\beta})-\boldsymbol{m}^T\boldsymbol{u})((\boldsymbol{y}^T-\boldsymbol{\beta}^T\boldsymbol{X}^T)\boldsymbol{b}-\boldsymbol{u}^T\boldsymbol{m})]=$$

$$=E[\boldsymbol{b}^T(\boldsymbol{y}-\boldsymbol{X\beta})(\boldsymbol{y}^T-\boldsymbol{\beta}^T\boldsymbol{X}^T)\boldsymbol{b}-\boldsymbol{b}^T(\boldsymbol{y}-\boldsymbol{X\beta})\boldsymbol{u}^T\boldsymbol{m}-\boldsymbol{m}^T\boldsymbol{u}(\boldsymbol{y}^T-\boldsymbol{\beta}^T\boldsymbol{X}^T)\boldsymbol{b}+\boldsymbol{m}^T\boldsymbol{uu}^T\boldsymbol{m}]=$$

$$=E[\boldsymbol{b}^T\boldsymbol{yy}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{y\beta}^T\boldsymbol{X}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{X\beta y}^T\boldsymbol{b}+\boldsymbol{b}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{yu}^T\boldsymbol{m}+$$

$$\boldsymbol{b}^T\boldsymbol{X\beta u}^T\boldsymbol{m}-\boldsymbol{m}^T\boldsymbol{uy}^T\boldsymbol{b}+\boldsymbol{mu\beta}^T\boldsymbol{X}^T\boldsymbol{b}+\boldsymbol{m}^T\boldsymbol{uu}^T\boldsymbol{m}]=$$

$$=\boldsymbol{b}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{b}+\boldsymbol{b}^T\boldsymbol{ZGm}-\boldsymbol{b}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{b}+\boldsymbol{b}^T\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T\boldsymbol{b}-\boldsymbol{b}^T\boldsymbol{ZGm}-$$

$$-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{b}+\boldsymbol{m}^T\boldsymbol{Gm}=\boldsymbol{m}^T\boldsymbol{Gm}-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{b}=\boldsymbol{m}^T\boldsymbol{Gm}-\boldsymbol{m}^T\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{ZGm}=$$

$$=\boldsymbol{m}^T(\boldsymbol{G}-\boldsymbol{GZ}^T\boldsymbol{V}^{-1}\boldsymbol{ZG})\boldsymbol{m}$$

Si ricorda che:

- $E[\boldsymbol{yy}^T]=\boldsymbol{X\beta\beta}^T\boldsymbol{X}^T+\boldsymbol{V}$ vedi (2.48) e (2.48bis)
- $E[\boldsymbol{yu}^T]=E[(\boldsymbol{X\beta}+\boldsymbol{Zu}+\boldsymbol{e})\boldsymbol{u}^T]=\boldsymbol{X\beta}E[\boldsymbol{u}^T]+\boldsymbol{Z}E[\boldsymbol{uu}^T]+E[\boldsymbol{eu}^T]=\boldsymbol{ZG}$
- $\boldsymbol{G}=\boldsymbol{G}^T,\boldsymbol{V}^{-1}=\boldsymbol{V}^{-1^T},\boldsymbol{b}=\boldsymbol{V}^{-1^T}\boldsymbol{ZG}^T\boldsymbol{m}=\boldsymbol{V}^{-1}\boldsymbol{ZGm}$

4. Si calcola la seconda componente:

$$V[(\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = g_2(\boldsymbol{\delta}) = V((\boldsymbol{l}^T - \boldsymbol{b}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) = V(\boldsymbol{d}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) =$$

$$= E[(\boldsymbol{d}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))(\boldsymbol{d}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^T] = E[\boldsymbol{d}^T\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T\boldsymbol{d} - \boldsymbol{d}^T\hat{\boldsymbol{\beta}}\boldsymbol{\beta}^T\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{\beta}\hat{\boldsymbol{\beta}}^T\boldsymbol{d} + \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d}] =$$

$$= \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d} + \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d} + \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d} = \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$$

Considerando:

$$E[\boldsymbol{d}^T\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T\boldsymbol{d}] = \boldsymbol{d}^T E[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}]\boldsymbol{d} =$$

$$= \boldsymbol{d}^T[(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} +$$

$$+(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}]\boldsymbol{d} = \boldsymbol{d}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{d} + \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$$

Abbiamo dimostrato che:

$$MSE(t(\boldsymbol{\delta}, \boldsymbol{y})) = g_1(\boldsymbol{\delta}) + g_2(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G})\boldsymbol{m} + \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d}$$

- Passaggio dalla formula generica della stima delle componenti di varianza alla formulazione per la stima a livello di area (pag. 47):

$$g_1(\boldsymbol{\delta}) = \boldsymbol{m}^T(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}\boldsymbol{V}^{-1}\boldsymbol{Z}^T\boldsymbol{G})\boldsymbol{m} = \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{m} - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}\boldsymbol{V}^{-1}\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{m} \rightarrow$$

$$\rightarrow g_{1_i}(\sigma_u^2) = z_i\sigma_u^2 z_i - z_i\sigma_u^2 z_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1} z_i\sigma_u^2 z_i = z_i^2\sigma_u^2 - \frac{z_i^2\sigma_u^2 \cdot z_i^2\sigma_u^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2} = \frac{z_i^2\sigma_u^2\sigma_{e_i}^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2} = \gamma_i\sigma_{e_i}^2$$

Considerando le sostituzioni $\boldsymbol{m}^T \rightarrow z_i, \boldsymbol{Z} \rightarrow z_i, \boldsymbol{G} \rightarrow \sigma_u^2, \boldsymbol{V} \rightarrow z_i^2\sigma_u^2 + \sigma_{e_i}^2$.

$$g_2(\boldsymbol{\delta}) = \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{d} = (\boldsymbol{l}^T - \boldsymbol{m}^T\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{l} - \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{m}) \rightarrow$$

$$\rightarrow$$

$$g_{2_i}(\sigma_u^2) = (\boldsymbol{x}_i^T - z_i\sigma_u^2 z_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1}\boldsymbol{x}_i^T)(\boldsymbol{x}_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1}\boldsymbol{x}_i^T)^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_i z_i\sigma_u^2 z_i(z_i^2\sigma_u^2 + \sigma_{e_i}^2)^{-1}) =$$

$$= (\boldsymbol{x}_i^T - \frac{z_i^2\sigma_u^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}\boldsymbol{x}_i^T)\left(\frac{\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}\right)^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_i\frac{z_i^2\sigma_u^2}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}) =$$

$$= (\boldsymbol{x}_i^T - \gamma_i\boldsymbol{x}_i^T)\left(\frac{\boldsymbol{x}_i\boldsymbol{x}_i^T}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}\right)^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_i\gamma_i) = (1-\gamma_i)\boldsymbol{x}_i^T\left(\frac{\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T}{z_i^2\sigma_u^2 + \sigma_{e_i}^2}\right)^{-1}(1-\gamma_i)\boldsymbol{x}_i =$$

$$= (1-\gamma_i)^2 \boldsymbol{x}_i^T \left(\frac{\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T}{z_i^2 \sigma_u^2 + \sigma_{e_i}^2} \right)^{-1} \boldsymbol{x}_i$$

- Valore atteso e varianza di $\boldsymbol{y}$ nel modello Spatial EBLUP (pag. 75):

$$E[\boldsymbol{y}] = E[\boldsymbol{X\beta} + \boldsymbol{Z}(\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1}\boldsymbol{u} + \boldsymbol{e}] = \boldsymbol{X\beta} + \boldsymbol{Z}(\boldsymbol{I}_m - \rho\boldsymbol{W})^{-1} E[\boldsymbol{u}] + E[\boldsymbol{e}] = \boldsymbol{X\beta}$$

$$V[\boldsymbol{y}] = E[(\boldsymbol{y} - E[\boldsymbol{y}])(\boldsymbol{y} - E[\boldsymbol{y}])^T] =$$

$$= E[(\boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e} - \boldsymbol{X\beta})(\boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e} - \boldsymbol{X\beta})^T] = E[(\boldsymbol{Zv} + \boldsymbol{e})(\boldsymbol{v}^T\boldsymbol{Z}^T + \boldsymbol{e}^T)] =$$

$$= E[\boldsymbol{Zvv}^T\boldsymbol{Z}^T + \boldsymbol{Zve}^T + \boldsymbol{ev}^T\boldsymbol{Z}^T + \boldsymbol{ee}^T] = \boldsymbol{Z}E[\boldsymbol{vv}^T]\boldsymbol{Z}^T + \boldsymbol{Z}E[\boldsymbol{ve}^T] + E[\boldsymbol{ev}^T]\boldsymbol{Z}^T + E[\boldsymbol{ee}^T] =$$

$$= \boldsymbol{Z}\left[\sigma_u^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1}\right]\boldsymbol{Z}^T + \boldsymbol{R} = \boldsymbol{ZGZ}^T + \boldsymbol{R}$$

Dove $\sigma_u^2[(\boldsymbol{I}_m - \rho\boldsymbol{W})(\boldsymbol{I}_m - \rho\boldsymbol{W}^T)]^{-1} = \boldsymbol{G}$.

APPENDICE B

APPROFONDIMENTO SULLE PRINCIPALI FONTI STATISTICHE DEL SETTORE AGRICOLO.

B.1 IL CENSIMENTO

La fonte di riferimento per il settore agricolo è il censimento sull'agricoltura, che ha cadenza decennale. L'ultima edizione è dell'anno 2000. Nell'ambito agricolo il censimento gioca un ruolo d'importanza strategica rispetto agli altri settori. In Italia, ad oggi, non esiste un archivio anagrafico su cui sono registrate tutte le imprese agricole presenti sul territorio. Per questo motivo anche per conoscere soltanto il numero delle aziende operanti è necessario un censimento. Il motivo per cui non esiste un archivio completo per le aziende agricole è legato al legislatore e non verrà discusso in questa sede.

L'ISTAT fino al censimento del 1990 non effettuava indagini mirate ad aggiornare l'elenco delle aziende agricole tra un censimento e l'altro. Dopo il 1990 visto un rinnovato vigore in ambito agricolo l'ISTAT intraprende indagini durante il decennio con lo scopo di tenere aggiornato l'elenco delle aziende agricole. Per il censimento del 2000 l'ISTAT aveva effettuato questo tipo di indagini tra il 1993 e il 1997 ed inoltre ha utilizzato i dati derivanti da quattro archivi: Sistema informativo agricolo nazionale (SIAN), anagrafe tributaria del Ministero delle Finanze, registro delle imprese agricole tenuto presso le Camere di commercio, archivio residente presso l'AIMA.

L'integrazione degli archivi statistici e amministrativi ha consentito di creare un archivio provvisorio di circa tre milioni di soggetti, il quale è stato accuratamente verificato dai comuni italiani.

La fase preliminare ha reso possibile verificare il numero e la dispersione territoriale delle aziende agricole, forestali e zootecniche, per mettere a fuoco il campo d'azione degli oltre 20.000 rilevatori destinati a visitare "porta a porta" tutti i conduttori d'azienda. Inoltre si vogliono creare in futuro le condizioni per mettere a punto un vero e proprio sistema informativo territoriale integrato (GIS); in altre parole, l'insieme delle informazioni ottenute tramite i censimenti generali dell'agricoltura integrato con i censimenti della popolazione e delle abitazioni, dell'industria e dei servizi verrà riferita

ai singoli segmenti del territorio (le sezioni di censimento), tracciandone così un identikit dettagliato. Con il riferimento alle sezioni di censimento sarà possibile passare ad aggregati di dimensioni maggiori come comuni, province, regioni, etc.

Una nuova concezione del censimento agricolo è stata resa necessaria dalla normativa EU, che ha riformato l'agricoltura a livello comunitario ed anche da una trasformazione che era già in atto nel settore. Settore che è in continua trasformazione sia perché il prodotto agricolo italiano sta conquistando paesi europei e non (basti pensare al vino, ai formaggi, alla cucina italiana), sia perché l'arrivo di nuovi paesi nell'unione europea ha portato cambiamenti nella politica comunitaria. Il censimento del 2000 è inserito in un quadro più ampio costituito dal progetto di costruzione del "Sistema delle statistiche agricole", al quale l'ISTAT sta lavorando dal 1997 coinvolgendo le regioni e il ministero delle politiche agricole. Le infrastrutture di base del Sistema sono: l'archivio ASAIA (l'archivio delle aziende e delle imprese agricole, forestali e zootecniche attive in Italia), la carta di copertura del suolo e l'area-frame, altri archivi di natura amministrativa che possono rappresentare importanti supporti per l'integrazione delle informazioni necessarie alla conduzione delle rilevazioni. La strategia generale prevede d'impiegare rilevazioni specifiche per le tematiche di carattere settoriale nonché di assegnare alla rilevazione sui risultati economici e, soprattutto, a quella sulla struttura delle aziende agricole (SPA) il compito di approfondire insieme alle caratteristiche principali del settore anche temi diversi, (ambiente, aspetti legati allo sviluppo rurale, alle caratteristiche socio-demografiche dei conduttori e alle attività emergenti svolte dalle aziende). Merita una particolare attenzione l'ASAIA, una delle più importanti novità del censimento agricolo 2000. Avvalendosi dell'utilizzo sinergico di una serie di banche dati, l'ISTAT ha messo a punto lo schedario delle aziende agricole esistenti oggi il quale, una volta verificata l'esattezza dei dati, potrà essere costantemente aggiornato. Questo rende possibile analisi ulteriori di tipo campionario senza dovere fare sforzi per identificare la popolazione di riferimento. L'idea di avere un elenco di tutte le unità statistiche di interesse (le aziende agricole in questo caso) e di conoscerne i dati principali è anche alla base della stima per piccole aree; è sufficiente estrarre un campione, rilevare la variabile di interesse e tramite l'elenco ed i suoi dati aggiuntivi, le covariate, fare una stima per piccole aree.

B.2 INDAGINE RICA-REA

Questa indagine nasce nel 2002-2003 dall'unione di due indagini campionarie, RICA e REA.

B.2.1 INDAGINE REA

La REA, acronimo di Risultati Economici delle Aziende Agricole, è un'indagine effettuata su un campione casuale di aziende agricole attraverso intervista diretta. L'indagine, con cadenza annuale, è condotta dall'ISTAT in collaborazione con l'INEA, gli Assessorati all'Agricoltura e gli Uffici di Statistica delle Regioni e Province Autonome con l'obiettivo di ottenere informazioni microeconomiche (a livello di ogni singola azienda agricola) sui risultati economici delle aziende nell'anno solare di riferimento. Queste informazioni sono tratte dal Conto Economico, dallo Stato Patrimoniale e dalla Contabilità Analitica, nel caso che esista una rilevazione contabile sistematica nell'azienda; altrimenti vengono ricostruite o stimate dal conduttore d'azienda, utilizzando tutti gli strumenti a sua disposizione al momento dell'intervista, con l'ausilio del rilevatore.

Le informazioni microeconomiche ottenute hanno un duplice scopo: 1. essere integrate nel nuovo "Sistema delle Statistiche Agricole" per lo studio e la programmazione economica (politiche agricole comunitarie, nazionali e regionali), 2. soddisfare l'obbligo comunitario di tenuta della Contabilità Nazionale seguendo il nuovo sistema contabile (SEC 95).

L'indagine utilizza un campione di aziende agricole scelte tra quelle che hanno partecipato all'indagine sulla *Struttura e Produzioni delle aziende Agricole* (SPA). Nell'indagine svolta nel 2002 sono state campionate 16774 aziende suddivise tra regioni, province autonome e 5 province per la regione Puglia:

TABELLA B.1. Dimensione campionaria per regione nell'indagine REA 2002 (Fonte ISTAT).

Regione	***Casi particolari***	***Campione***
Piemonte		1.204
Valle D'Aosta		83
Lombardia		218
Trentino Alto Adige	*Trento*	174
	Bolzano	118
Veneto		1.282
Friuli Venezia Giulia		339
Liguria		162
Emilia Romagna		1.491
Toscana		1.599
Umbria		733
Marche		839
Lazio		414
Abruzzo		*n.p.*
Molise		326
Campania		658
Puglia	*Brindisi*	211
	Bari	597
	Lecce	276
	Taranto	445
	Foggia	291
Basilicata		932
Calabria		899
Sicilia		2.468
Sardegna		1.015
TOTALE		***16.774***

L'indagine REA rileva solo l'aspetto economico ma essendo un sottodominio dell'indagine SPA si possono ottenere, per le aziende campionate, tutte le informazioni di carattere ausiliario raccolte nella SPA per uno studio approfondito sulla redditività.

B.2.2 INDAGINE RICA

L'indagine RICA, acronimo di Rete d'Informazione Contabile Agricola, è uno strumento informativo finalizzato alla conoscenza della condizione economica delle aziende agricole europee. E' stata istituita nel 1965 con norme comunitarie che ne stabiliscono i principi e l'organizzazione.

L'indagine campionaria annuale, svolta con un'impostazione analoga in tutti i Paese Membri dell'Unione Europea (UE), fornisce le informazioni che confluiscono nella base dati europea che costituisce il fulcro dell'intero sistema di monitoraggio sull'agricoltura europea. La RICA è l'unica fonte armonizzata di dati microeconomici: i principi su cui si basa la raccolta dei dati sono infatti i medesimi in tutti i paesi e sono indicati in appositi regolamenti. Le aziende agricole che partecipano alla RICA vengono selezionate sulla base di un piano di campionamento redatto in ciascun Paese Membro. Il campo di osservazione dell'indagine non coincide con l'universo delle aziende agricole ma include solo quelle la cui dimensione in termini economici è tale da poterle definire commerciali. La metodologia comune adottata permette di rappresentare i risultati secondo tre dimensioni principali: la regione geografica, la dimensione economica e l'orientamento tecnico economico. Le informazioni RICA vengono utilizzate per studi e ricerche di carattere microeconomico e trovano ampia utilizzazione nella gestione delle politiche agricole per fini di programmazione e di valutazione.

La Rica, funzionando a livello europeo, ha una struttura abbastanza complessa. Un'unità tecnica all'interno della Commissione Europea è responsabile della RICA Europea (EU-RICA). L'Unità "Analisi della situazione delle aziende agricole" facente capo alla DG AGRI coordina il flusso dei dati dai singoli Paesi Membri, gestisce la banca dati europea e cura la diffusione dei dati a livello europeo. A livello comunitario, un comitato di gestione (Comitato comunitario RICA), costituito dai rappresentanti delle Reti contabili di tutti i paesi membri e presieduto da un dirigente della Direzione Generale dell'Agricoltura, discute in merito alla metodologia ed all'organizzazione della Rete, promuove studi e ricerche inerenti e propone gli atti normativi che regolano i sistemi RICA europeo e nazionali. In ciascun Paese Membro, la responsabilità e la gestione della RICA nazionale sono affidate ad un Organo di collegamento, che in Italia si identifica con l'Istituto Nazionale di Economia Agraria (INEA). Anche la RICA nazionale è guidata da un comitato di gestione, detto Comitato Nazionale RICA. In Italia, tale Comitato è composto da rappresentanti INEA, dai rappresentanti delle Regioni/Provincie-autonome, da funzionari dell'ISTAT, da esponenti delle Organizzazioni Professionali agricole ed è presieduto da un dirigente del Ministero delle Politiche Agricole e Forestali.

L'universo di riferimento coincide con quello dell'indagine SPA, quindi è stato deciso di utilizzare una metodologia di campionamento comune.

A partire dal 2003 (con la rilevazione 2002-2003) le indagini RICA e REA sono state unite in un'unica indagine. L'obiettivo è quello di coinvolgere le imprese agricole per ottenere una stretta collaborazione. Partecipando all'indagine RICA-REA le imprese ottengono il bilancio della propria azienda accompagnato da un'analisi della conduzione aziendale e le informazioni necessarie per poter aderire agli intervinti di politica agricola come i contributi europei e i finanziamenti pubblici. Inoltre è attivo un sistema di assistenza tecnica alla gestione, attraverso cui individuare eventuali miglioramenti nelle scelte tecniche ed economiche. L'accesso ad informazioni corrette raccolte nelle singole aziende agricole permette l'impostazione di misure di politica agraria efficaci, eque e modulate sugli aspetti specifici locali e quindi rispondenti alle effettive necessità del mondo imprenditoriale agricolo.
Il campione RICA-REA è un sottocampione dell'indagine SPA. Anche la riorganizzazione delle indagine RICA e REA rientra nel progetto per il nuovo Sistema delle Statistiche Agricole.

B.3 Registro delle Imprese

Le aziende agricole sono soggette ad una normativa speciale per l'iscrizione al registro delle imprese. Esse sono registrate in sezioni speciali del registro delle imprese. L'iscrizione a tale registro riguarda gli imprenditori agricoli definiti dall'articolo 2135 del codice civile e le società semplici. Con questo meccanismo non si ha la certezza di includere tutte le aziende agricole, un insieme diverso da quello definito dal legislatore.
Il registro delle imprese è tenuto in modalità informatica dalla società InfoCamere. Questo archivio è una fonte statistica di principale importanza soprattutto per la creazione di un elenco esaustivo delle imprese operanti in diversi settori.

Nel corso degli anni, InfoCamere ha messo a disposizione dell'analisi statistica le sue competenze e l'esperienza sul trattamento dei grandi archivi informatici delle Camere di Commercio creando due elaborati:

1. Movimprese: L'analisi trimestrale delle variazioni dell'anagrafe delle imprese italiane: natalità e mortalità delle imprese per territorio, forma giuridica e settore di attività economica.

2. MUD: La sintesi statistica delle "Dichiarazioni Ambientali" rese dalle imprese alle Camere di Commercio attraverso il MUD, il Modello di Dichiarazione ambientale.

Dalla banca dati di InfoCamere è possibile ottenere informazioni con dettaglio sino alla singola impresa. Le informazioni disponibile sono di tipo amministrativo: natura giuridica, data di costituzione, capitale sociale, codice fiscale, attività svolta, cariche amministrative, organi sociali, etc.
Il registro delle imprese agricole è utilizzato tra le fonti per costituire l'elenco delle imprese da censire nel "Censimento sull'Agricoltura".

B.4 Ente Ragione Toscana

Le Regioni unitamente all'ISTAT e le amministrazioni provinciali e comunali si occupano di raccogliere dati riguardanti il territorio e la sua economia.
Sul fronte agricolo la Regione Toscana raccoglie ed elabora dati del censimento e di altre indagini:

- Dati congiunturali sulle coltivazioni per Province relativi alle superfici ed alle produzioni delle coltivazioni erbacee ed arboree esistenti in Toscana. Tali dati sono rilevati attraverso accertamenti estimativi effettuati dagli Assessorati all'Agricoltura delle Province, nel corso di un programma di rilevazioni a carattere congiunturale che l'ISTAT svolge annualmente.
- Indicatori statistici sintetici della struttura delle aziende agricole per Province, SEL, Comunità Montane (5° Censimento Generale dell'Agricoltura).
- Indicatori statistici sintetici della produzione delle aziende agricole per Province, SEL, Comunità Montane (5° Censimento Generale dell'Agricoltura).
- Indicatori statistici sintetici della struttura delle aziende agricole per Province, Regioni Agrarie, Comunità Montane (Censimento dell'agricoltura, dati non aggiornati).
- Indicatori statistici sintetici della produzione delle aziende agricole per Province, Regioni Agrarie, Comunità Montane (Censimento dell'agricoltura, dati non aggiornati).

I dati raccolti ed elaborati dalla Regione Toscana non sono una fonte alternativa a quelle già descritte, ma sono una parte integrante delle fonti ufficiali italiane quali il censimento e le indagini di principale importanza svolte dall'ISTAT.

B.5 EUROSTAT

A livello europeo l'Eurostat in campo agricolo collezione una grossa quantità di dati. Suddivide il settore primario in:

1. Agricoltura
2. Pesca
3. Foreste

In campo agricolo le statistiche sono organizzate in otto settori:

1. Indicatori agricoli principali
2. Economia per l'agricoltura e le foreste
3. Struttura delle aziende agricole
4. Alimentazioni animali
5. Prezzi e indici dei prezzi in agricoltura
6. Prodotti agricoli
7. Campionamento europeo sui frutteti
8. Viticoltura

Per ogni settore sono fornite statistiche a diversi livelli di aggregato, senza mai scendere su dimensioni inferiori alla nazione. Le indagini sono fatte in collaborazione con i paesi membri come abbiamo visto nel caso RICA.

L'ente statistico europeo pubblica ogni trimestre un bollettino sull'agricoltura. Questa pubblicazione contiene dati sui raccolti, sugli allevamenti e sui prezzi commentati da esperti del settore. Le ultime edizioni contengono anche una relazione sul patrimonio forestale europeo. Il database di riferimento utilizzato è il FAME/NewCronos, attivo sin dal 1995.

B.6 ISTITUTO NAZIONALE DI ECONOMIA AGRARIA

Una fonte importante, soprattutto per elaborazioni e rapporti, è l'INEA, Istituto Nazionale di Economia Agraria. L'INEA svolge attività di ricerca, di rilevazione, analisi e previsione nel campo strutturale e socio economico del settore agro-industriale, forestale e della pesca. Negli ultimi anni l'attività dell'Istituto si è ampliata nelle attività di supporto alla Pubblica Amministrazione per l'attuazione delle politiche agricole, in primo luogo quelle che discendono dall'Unione Europea. L'INEA è coinvolto dai servizi della Commissione Europea, dal Ministero per le Politiche Agricole e da numerose regioni in attività di assistenza tecnica, monitoraggio e valutazione delle politiche strutturali e di mercato (Organizzazioni comuni di mercato). Filoni recenti di attività vedono l'INEA impegnato anche in materia di sviluppo rurale e su temi riguardanti la valorizzazione delle risorse ambientali e la gestione delle risorse idriche, aspetti che guideranno in un prossimo futuro le politiche agricole nazionali, comunitarie e mondiali.

L'INEA fornisce una banca dati, veicolata da Internet, creata da un gruppo di regioni italiane (Piemonte, Lombardia, Veneto, Friuli Venezia Giulia, Emilia Romagna, Toscana, Campania, Puglia, Basilicata, Sicilia) in cui sono archiviate le ricerche in materia agroambientale, finanziate dalle regioni negli ultimi 4 anni (2000/2003). L'iniziativa ha la finalità generale di dotare le Regioni, con particolare riguardo ai decisori pubblici, di uno strumento informativo sugli aspetti salienti della ricerca agricola finanziata e promossa dalle Regioni stesse (istituzioni coinvolte, obiettivi, principali contenuti, risorse attivate), in modo da avviare un processo di coordinamento volto alla migliore allocazione delle risorse disponibili.

La banca dati ha altri obiettivi operativi, oltre a quello sopra menzionato: la diffusione generalizzata di informazioni sulle iniziative di ricerca promosse dalle regioni, la disponibilità di dati utili alla elaborazione di statistiche e analisi, la costruzione del primo tassello di un sistema di comunicazione interregionale e la verifica dell'evoluzione della ricerca agricola regionale in termini di finanziamenti, obiettivi, contenuti.

L'INEA pubblica ogni anno un fascicolo intitolato "L'agricoltura Italiana Conta". Nel 2004 è stata pubblicata la sedicesima edizione. Al suo interno sono analizzati i principali temi di interesse per il settore primario: il ruolo dell'agricoltura nel sistema economico nazionale; i rapporti con l'industria alimentare e il settore distributivo, il mercato, le istituzioni e le politiche agricole. Con questa pubblicazione l'INEA vuole

fornire uno strumento che, alla facile ed immediata consultazione, associ la qualità e la completezza dei dati offerti. Il rapporto sull'agricoltura redatto dall'INEA si può considerare come punto di riferimento per l'analisi dei dati agricoli italiani. Dal 2003 pubblica anche il "Rapporto sullo Stato dell'Agricoltura Italiana", questo volume ha l'obiettivo di proporsi come base di riferimento per sviluppare riflessioni e proposte riguardo agli assetti presenti e futuri della politica agraria.

B.7 ENARPRI

L'ENARPRI, European Network of Agricultural and Rural Policy Research Institutes (istituto di ricerca per la rete europea delle politiche agricole e rurali), mira a creare una rete di rapporti tra i paesi membri al fine di scambiarsi informazioni e politiche di ricerca. L'ENARPRI è una istituzione recente, nata nel 2003, a cui fanno capo istituzioni statistiche di 13 paesi membri dell'unione: CEPS per il Belgio, FAL per la Germania, FOI per la Danimarca, IEEP per il Regno Unito, INEA per l'Italia, INRA per la Francia, IRWIN-PAN per la Polonia, LEI per l'Olanda, MTT per la Finlandia, TEAGASC per l'Irlanda, AUA per la Grecia, UPM-ETSIA per la Spagna e VUZE per la Repubblica Ceca.

Il tema centrale della rete di rapporti è lo studio sull'impatto degli accordi commerciali bilaterali e multilaterali tra regioni eruopee che sono avvenuti o sono in fase di negoziato in europa e tra europa e resto del mondo. Non meno importante è lo studio di politiche per un'agricoltura sostenibile.

L'ENARPRI ha l'ulteriore obiettivo di unire le forze dei paesi partecipanti per sviluppare nuove metodologie di controllo delle politiche agricole, applicabili anche paese per paese, e di una loro valutazione sia a livello comunitario sia a livello nazionale.

B.8 BANCA D'ITALIA

Pur non essendo una vera fonte statistica per l'agricoltura la Banca d'Italia rilascia informazioni sull'andamento economico del settore nella "Nota sull'Andamento dell'Economia in Toscana", redatto ovviamente per ogni regione. La pubblicazione ha carattere sintetico ma è utile per tracciare a grandi linee l'andamento del settore agricolo da un punto di vista produttivo.

B.9 ALTRE FONTI

Le fonti che abbiamo presentato sono le più importanti a livello nazionale e per l'insieme del settore agricolo. Lo scopo di questo paragrafo non è di fornire una lista completa ed esauriente di tutte le fonti statistiche, istituzionali e non, presenti nel settore agricolo ma quello di presentare le fonti statistiche determinante delle politiche agricole. Altre basi di dati per il settore agricolo sono tenute da organizzazione di settore e da sindacati, quale col diretti, nonché dalle organizzazioni riferite a specifiche branche dell'agricoltura, come ad esempio le coltivazioni biologiche.

APPENDICE C

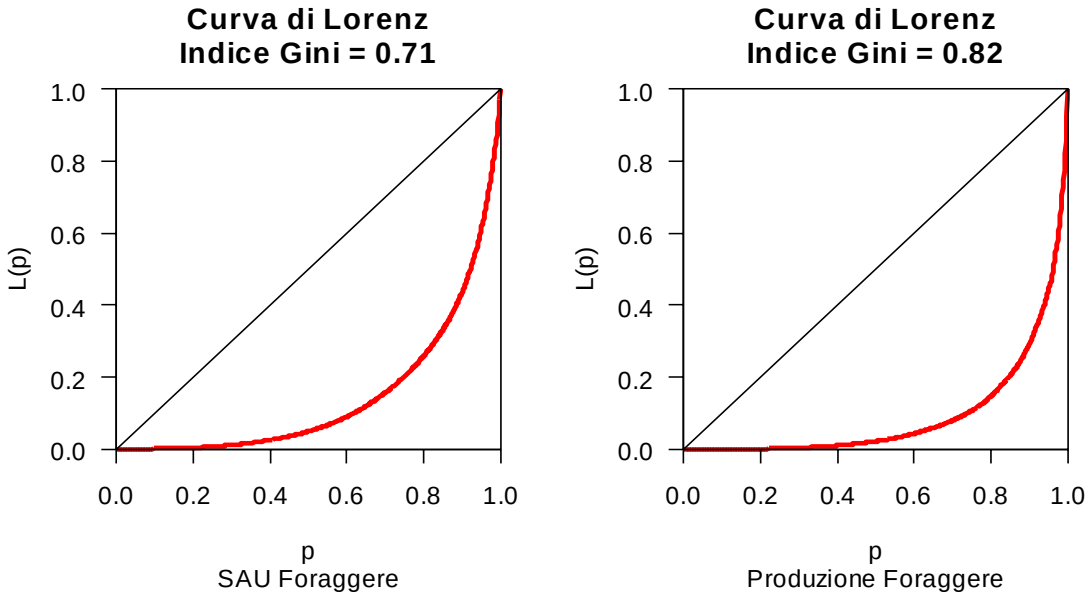
ANALISI DESCRITTIVA SUI DATI DELL'INDAGINE SPA NON TRATTATI NELLA STIMA PER PICCOLE AREE.

SEMINATIVI

FORAGGERE AVVICENDATE

Per le foraggere avvicendate (erba medica, granoturco in erba, etc.) ci sono 644 aziende con SAU dedicata e 569 aziende produttrici. La SAU dedicata alla coltivazione di foraggere è pari a 1979992 ettari per una produzione totale di 965281 quintali. La SAU media per azienda è 3074,52 (errore standard 232,49) ettari per una produzione media di 1498,88 (errore standard 181,52) quintali di cereali. La SAU varia tra un minimo di 2 e un massimo di 53084 ettari, mentre la produzione varia tra 1 e 70150 quintali. Gli indici di Gini per la SAU e la produzione, in linea con quelli del settore, sono rispettivamente 0,71 e 0,82.

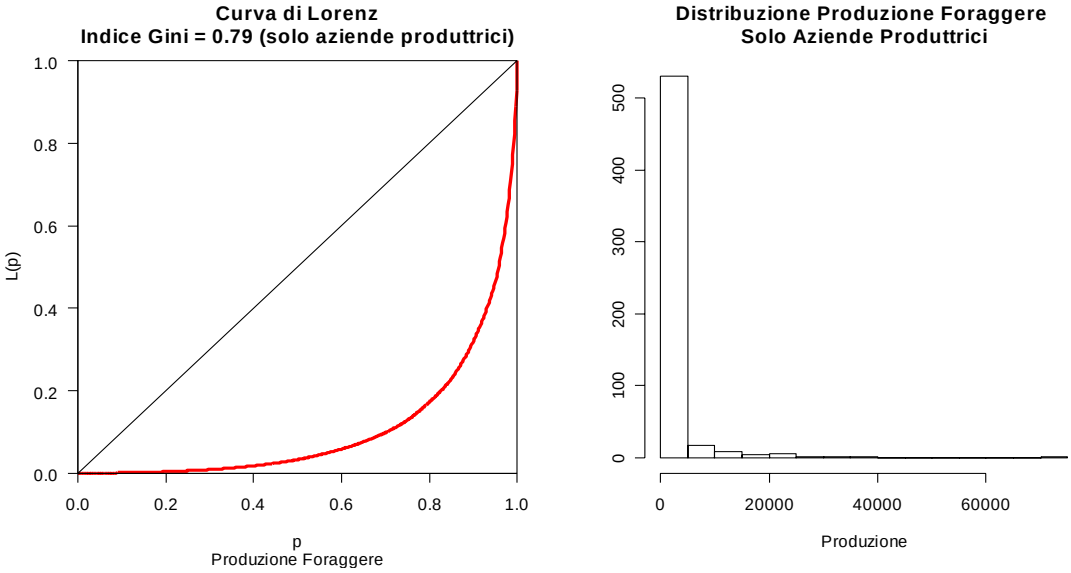
FIGURA C.1. Curva di Lorenz per SAU e Produzione di Foraggere.



Se consideriamo solo le aziende che sono riuscite a produrre foraggere (569) l'indice di concentrazione è 0,79. In questo comparto il 10% delle aziende produce circa il 69% del

totale, mentre il 20% delle aziende riesce a produrre l'83% del totale delle foraggere nella regione Toscana.

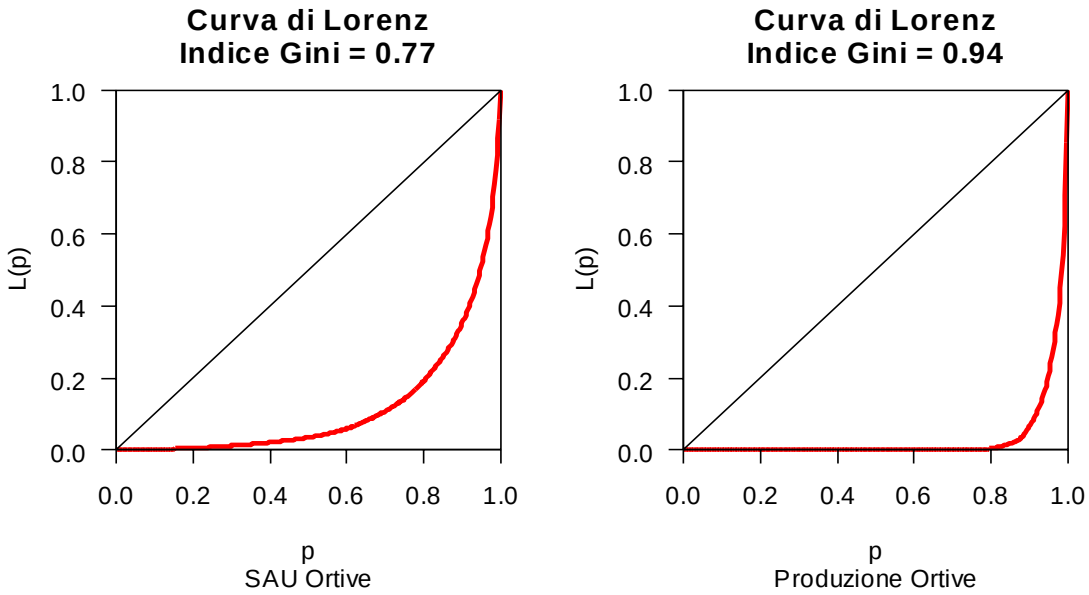
FIGURA C.2. Curva di Lorenz e Distribuzione di Frequenza per Produzione di Foraggere per le aziende che hanno avuto un raccolto.



ORTIVE

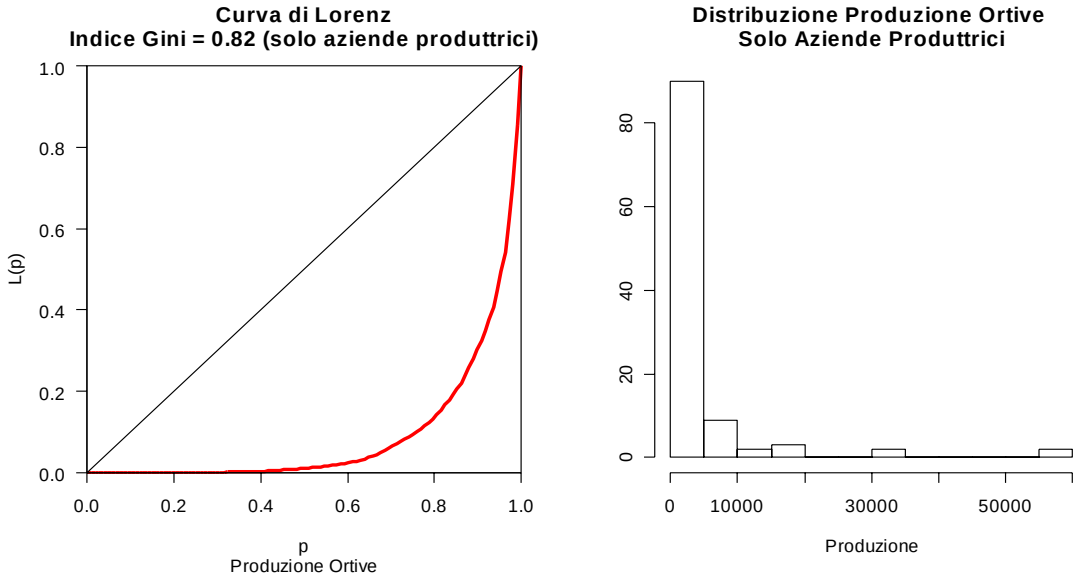
Le ortive, sia in piena aria sia in serra, sono state prodotte da 108 aziende su 320 dotate di SAU dedicata. Il fatto che solo il 33,7% delle aziende, che hanno un terreno destinato alle ortive, hanno effettivamente avuto una produzione è sintomo di una coltivazione che ha risentito in maniera particolare della condizione climatica sfavorevole verificatasi nel 2003 (come specificato nel rapporto INEA). La SAU totale dedicata alle ortive è 128116 ettari con una media di 400,36 (errore standard 56,36) ettari per azienda e la dimensione della SAU per azienda varia tra 2 e 10364 ettari: un valore molto più piccolo rispetto alle coltivazioni di cereali e foraggere. La produzione totale ammonta a 389263 quintali con una media di 1216,45 (errore standard 311,80) quintali per azienda; a livello di singola azienda (delle 108 che hanno prodotto) si passa da una produzione minima di 2 quintali ad una massima di 57000. L'indice di concentrazione per la SAU e la produzione di ortive è rispettivamente di 0,77 e 0,94; l'indice di concentrazione della produzione è prossimo a 1 ma bisogna considerare che molte aziende non hanno prodotto.

FIGURA C.3. Curva di Lorenz per SAU e Produzione di Ortive.



Se consideriamo soltanto le 108 che hanno prodotto ortive l'indice di concentrazione è 0,82, un valore sempre molto alto. Infatti in questo comparto il 10% delle aziende maggiori produce il 72% circa del totale mentre il 20% produce il 88% circa del totale. In questo comparto si registra dunque una concentrazione più alta rispetto agli altri già presi in esame.

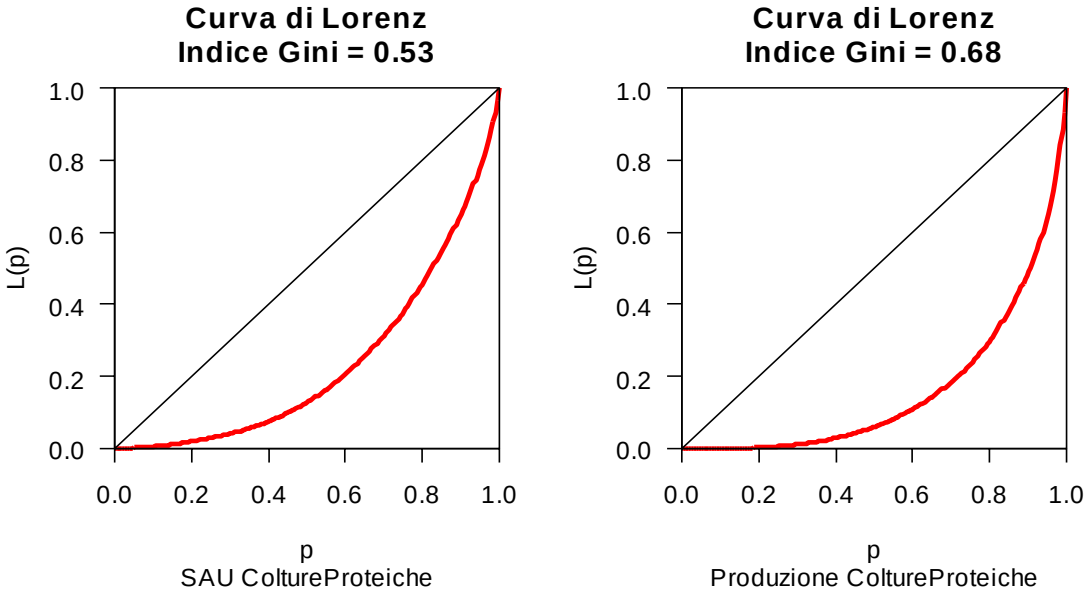
FIGURA C.4. Curva di Lorenz e Distribuzione di Frequenza per Produzione di Ortive per le aziende che hanno avuto un raccolto.



COLTURE PROTEICHE

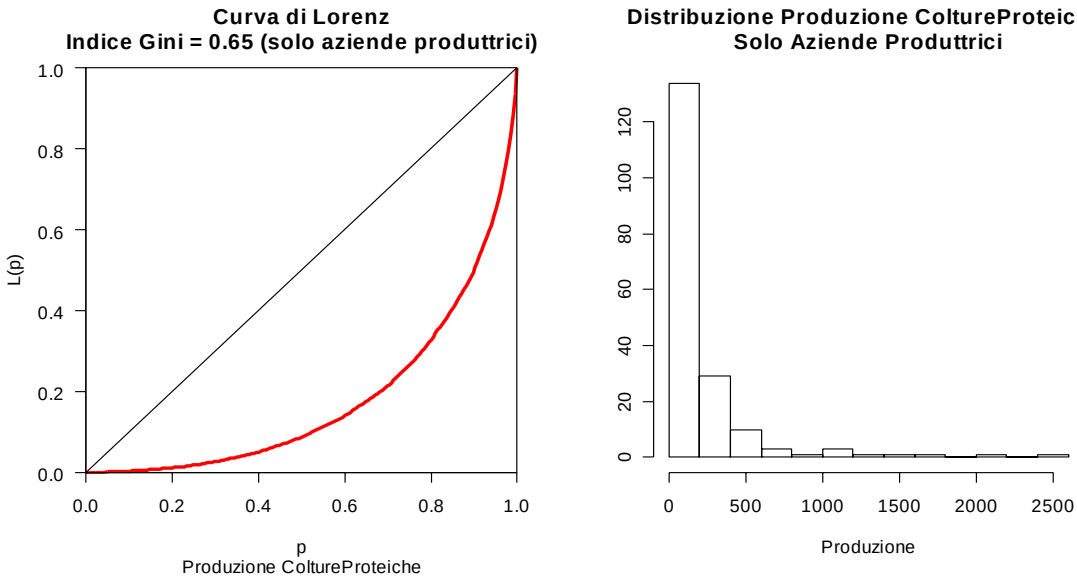
Il comparto delle Colture Proteiche per la produzione di granella comprende le coltivazioni di pisello, pisello secco, fagiolo secco, fava, lupino dolce, lenticchia, cece, vecce e altri legumi secchi. La SAU totale in Toscana per queste coltivazioni è pari a 412660 ettari per una produzione totale di 36988 quintali di colture proteiche. Le aziende dotate di SAU per la coltivazione di colture proteiche sono 214 con una SAU media di 1928,32 (errore standard 148,38); la SAU delle aziende spazia tra 2 e 14265 ettari. La produzione media per azienda è 172,84 quintali (errore standard 20,98) con un minimo di produzione per singola azienda pari a 1 quintale ed un massimo di 2500 quintali. L'indice di concentrazione è 0,53 per la SAU e 0,68 per la produzione, notevolmente più bassi rispetto agli indici del settore dei Seminativi.

FIGURA C.5. Curva di Lorenz per SAU e Produzione di Colture Proteiche.



Se consideriamo soltanto le 185 aziende che hanno ottenuto una produzione l'indice di concentrazione è 0,65. Il 10% delle aziende produce circa la metà del totale mentre l'80% circa di tutta la produzione è raccolta dal 31% delle aziende.

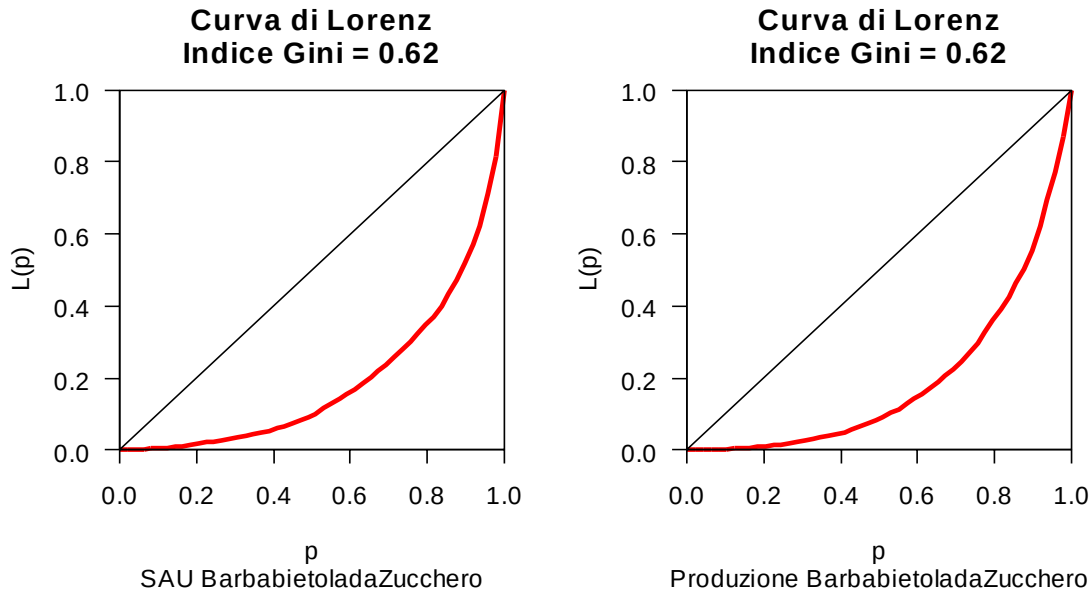
FIGURA C.6. Curva di Lorenz e Distribuzione di Frequenza per Produzione di Colture Proteiche per la Produzione di Granella per le aziende che hanno avuto un raccolto.



BARBABIETOLA DA ZUCCHERO

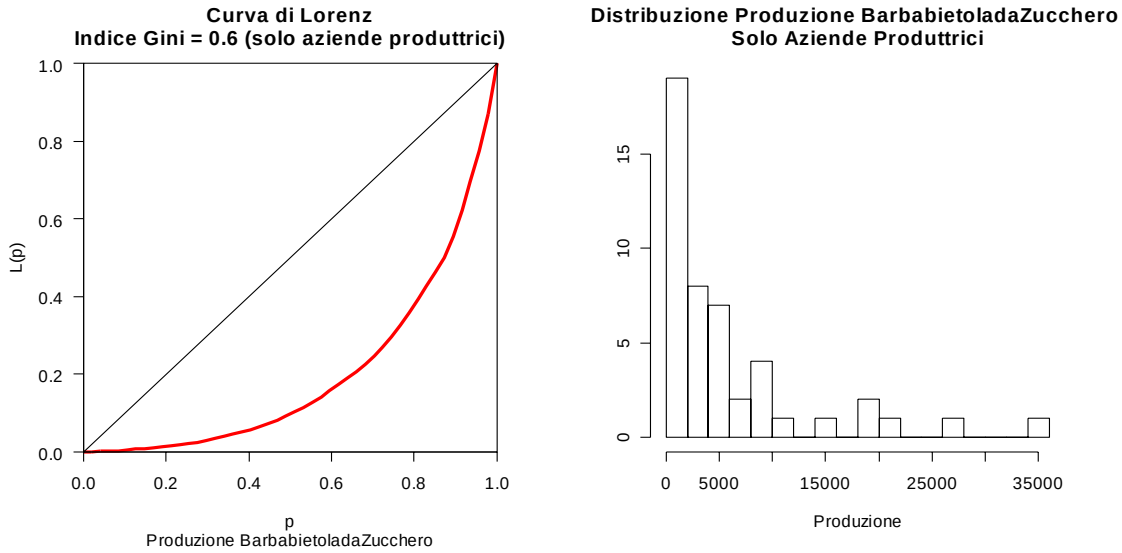
Prodotta da poche aziende ma in grande quantità la barbabietola da zucchero SAU totale pari a 113626 ettari. Le 49 aziende con terreno adibito per questa coltivazione hanno SAU media 2318,90 (errore standard 520,75) ettari, con un'ampiezza compresa tra 50 e 21185 ettari. La produzione di barbabietola da zucchero è stata di 271914 quintali con media 5549,27 (errore standard 1062,23) quintali. La produzione minima è stata di 50 quintali mentre la massima di 35000. L'indice di concentrazione è 0,62 sia per la SAU sia per la produzione.

FIGURA C.7. Curva di Lorenz per SAU e Produzione di Barbabietola da Zucchero.



Considerando solo le aziende che hanno prodotto barbabietola da zucchero, 47, l'indice di concentrazione è 0,60. In questo comparto la concentrazione è meno elevata rispetto al settore dei seminativi, ma comunque significativa, infatti il 36% delle aziende detiene l'80% circa della produzione.

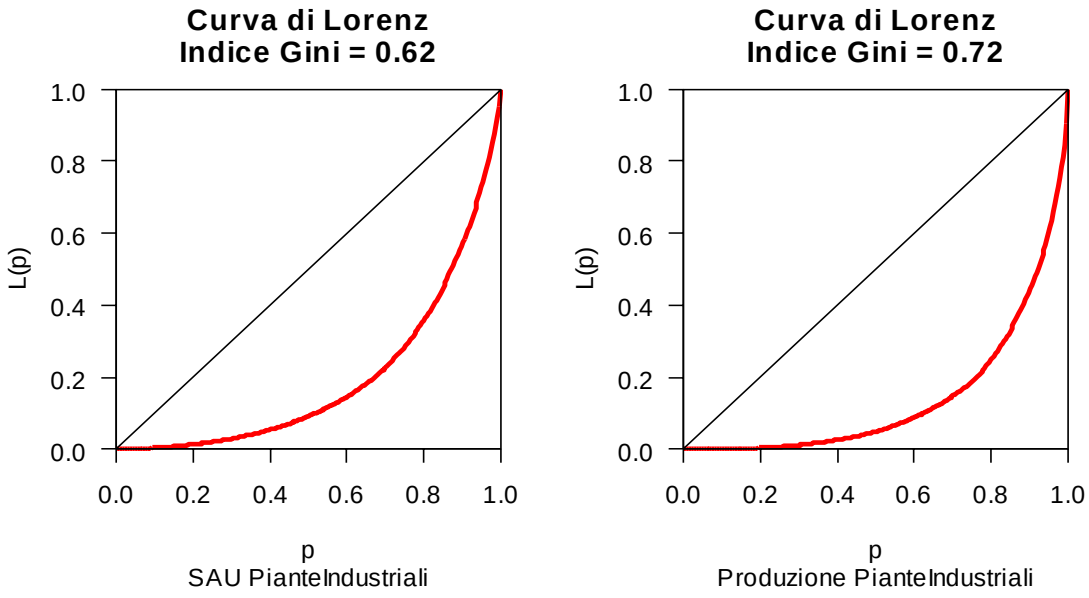
FIGURA C.8. Curva di Lorenz e Distribuzione di Frequenza per Barbabietola da Zucchero per le aziende che hanno avuto un raccolto.



PIANTE INDUSTRIALI

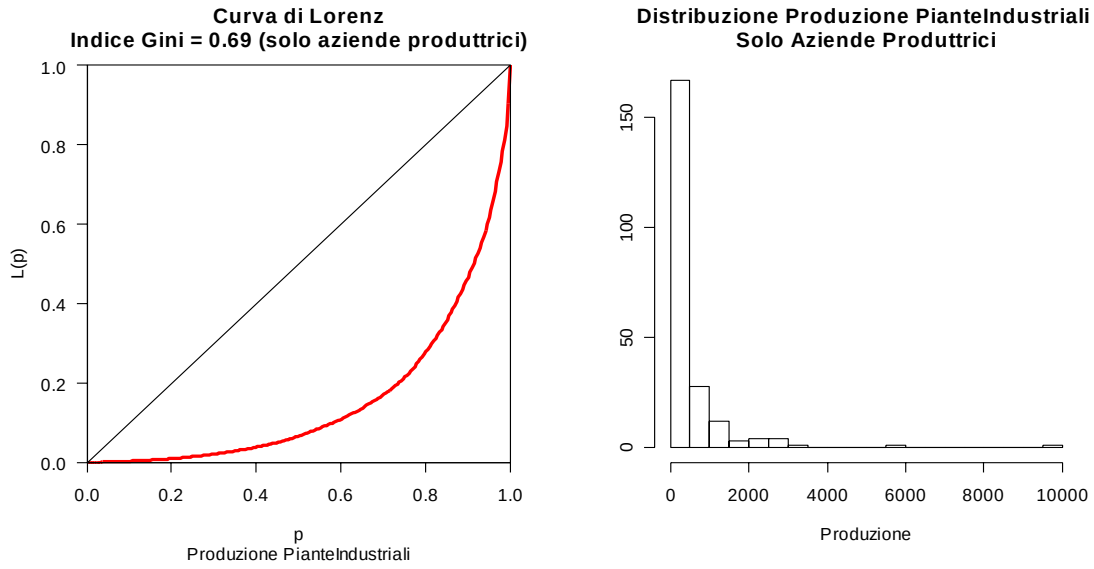
La SAU totale in Toscana per la coltivazione di Piante Industriali è pari a 519889 ettari per una produzione totale di 101932 quintali. Le piante industriali comprendono tabacco, luppolo, piante tessili, piante da semi oleosi (colza, girasole, soia, etc.), piante aromatiche, spezie, piante medicinali, piante da condimento e altre piante. Le aziende dotate di SAU per la coltivazione di piante industriali sono 246 con una SAU media di 2113,37 (errore standard 190,26); la SAU delle aziende spazia tra 1 e 25031 ettari. La produzione media per azienda è di 414,36 quintali (errore standard 57,08) con un minimo di produzione per singola azienda pari a 4 quintali ed un massimo di 9600 quintali. L'indice di concentrazione è 0,62 per la SAU e 0,72 per la produzione.

FIGURA C.9. Curva di Lorenz per SAU e Produzione di Piante Industriali.



Le aziende che hanno ottenuto un raccolto positivo di piante industriali sono 221. Per queste aziende l'indice di concentrazione della produzione è 0,69. Il 31% delle aziende produce l'80% circa delle piante industriali, quindi anche il comparto delle piante industriali presente una struttura formata da poche aziende (circa 65) a produzione elevata e molte aziende con produzioni ridotte.

FIGURA C.10. Curva di Lorenz e Distribuzione di Frequenza per Piante Industriali per le aziende che hanno avuto un raccolto.



ALTRE COLTIVAZIONI DI SEMINATIVI

La patata, le piante sarchiate da foraggio e le sementi sono coltivazioni di importanza marginale in Toscana. Riportiamo in una tabella le informazioni principali su queste coltivazioni:

TABELLA C.1. Coltivazioni di minore importanza a livello produttivo in Toscana nel 2003, dati campionari indagine SPA.

Coltivazione		***Numero aziende***	***Totale***	***Media (err.std.)***	***Minimo***	***Massimo***	***Concen-trazione***
Patata	SAU (ettari)	50	1709	34,18 (10,14)	1	400	0,72
	Produzione (quintali)	45	2387	47,74 (24,85)	1	1200	0,87
Piante Sarchiate da foraggio	SAU (ettari)	2	780	390 (190)	200	580	0,24
	Produzione (quintali)	1	40	20	40	40	-
Sementi	SAU (ettari)	18	46082	2560,11 (623,30)	172	9000	0,49
	Produzione (quintali)	17	2109	117,17 (48,91)	0	880	0,68

Le sementi sono quelle non comprese nelle coltivazione precedentemente presentate. Come si nota dalla tabella 13 la coltivazione di sementi ha un certo rilievo rispetto alla

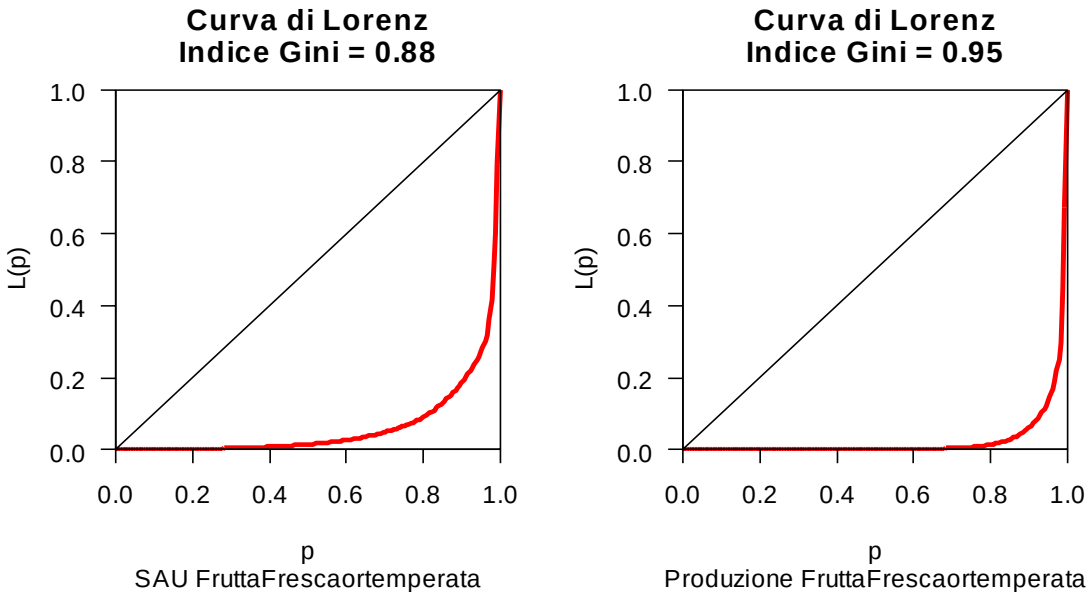
SAU ma la scarsa produzione complessiva e il ristretto numero di aziende ne limitano l'importanza. Sia la patata che le piante sarchiate da foraggio hanno SAU molto piccole e le produzioni sono insignificanti.

COLTIVAZIONI LEGNOSE AGRARIE

FRUTTA FRESCA DI ORIGINE TEMPERATA

La produzione di frutta fresca di origine temperata ammonta a 55136 quintali, una quantità minima se paragonata alla produzione di vite e olivo. La SAU dedicata a questa coltivazione è 80492 ettari con una media di 444,71 (errore standard 139,98) ettari ad azienda, mentre la produzione media è di 304,62 (errore standard 129,90) quintali. 181 aziende hanno SAU dedicata alla frutta fresca che varia tra 1 e 16500 ettari con un indice di concentrazione uguale a 0,88, mentre la produzione per ogni azienda, tra le 105 che hanno prodotto frutta nell'anno 2003, varia tra 1 e 18000 quintali con una concentrazione di 0,95.

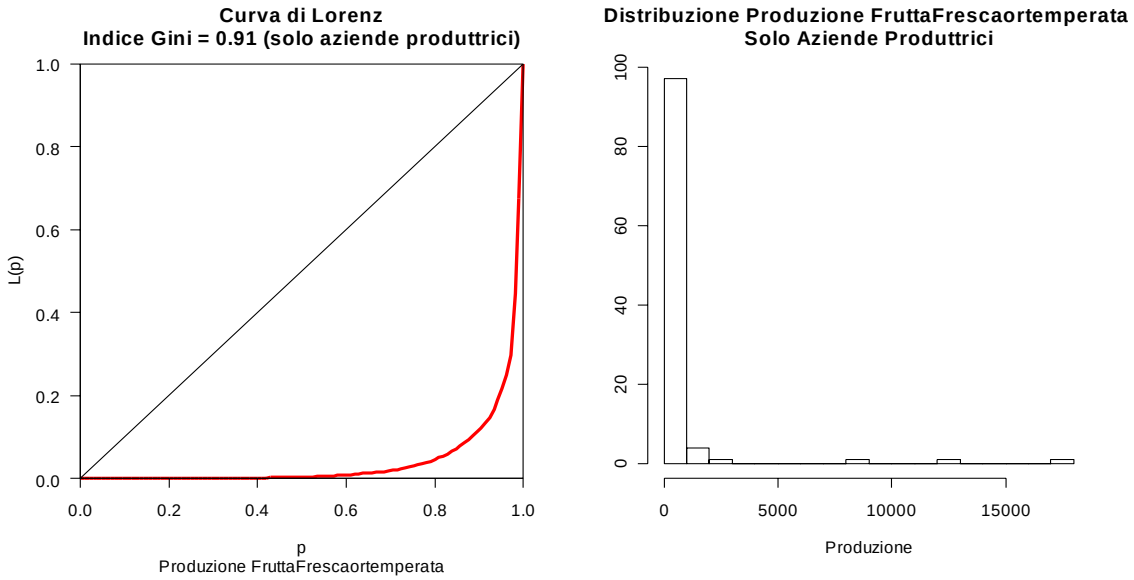
FIGURA C.11. Curva di Lorenz per SAU e Produzione di Frutta fresca di origine temperata.



Se consideriamo solo le 105 aziende che hanno prodotto frutta di origine temperata l'indice di concentrazione è uguale a 0,91. Le 10 aziende con produzione maggiore hanno raccolto l'88,7% della frutta fresca prodotta in Toscana. Anche in questo

comparto si contano due otliers, infatti le prime due aziende produttrici, da sole, hanno prodotto il 55,3% del totale di frutta fresca di origine temperata.

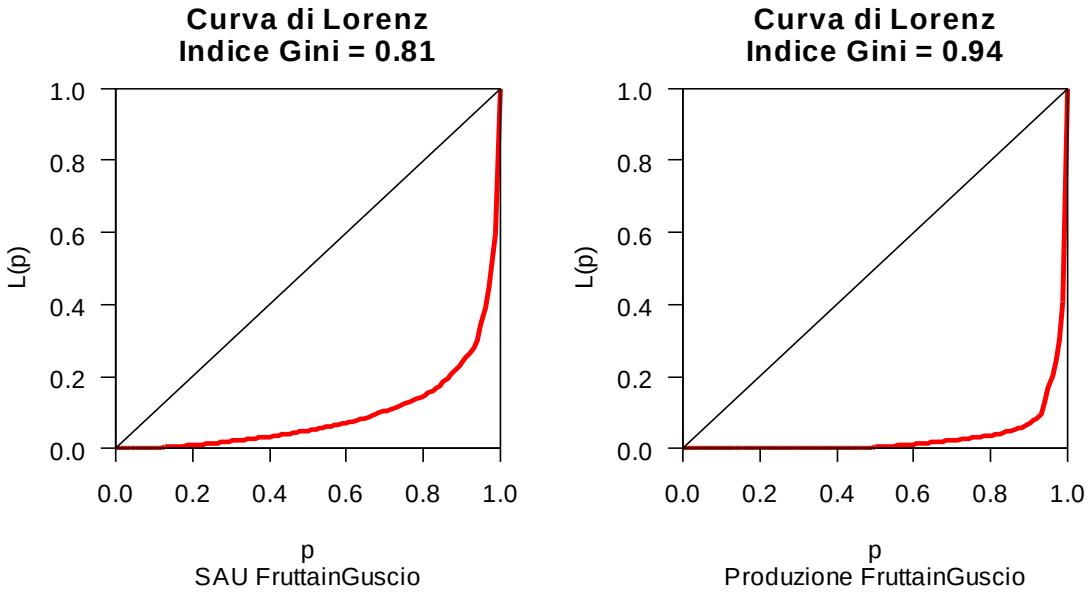
FIGURA C.12. Curva di Lorenz e Distribuzione di Frequenza per la Frutta Fresca di Origine Temperata per le aziende che hanno avuto un raccolto.



FRUTTA FRESCA IN GUSCIO

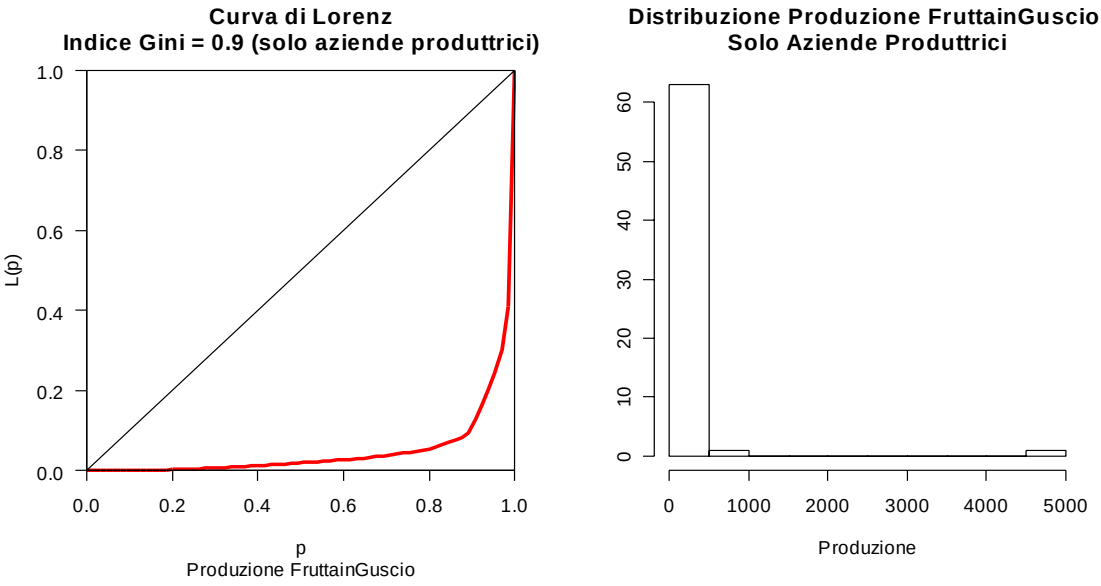
Oltre alla frutta fresca gioca un ruolo non marginale, anche se non di primo piano, la frutta in guscio. La SAU totale dedicata a questa coltivazione è 86816 ettari con una produzione media di 851,14 (errore standard 356,24) ettari per 102 aziende. La produzione nell'anno agricolo 2003 è stata di 8459 quintali, con una media di 8459 (errore standard 82,93) quintali; le aziende che hanno avuto una produzione sono state 65, cioè il 60% delle aziende con terreno dedicato. Questo significa che il 40% circa dei terreni dedicati alla frutta in guscio non ha prodotto, probabilmente per cause legata al clima sfavorevole. Gli indici di Gini per la SAU e la produzione sono rispettivamente 0,81 e 0,94.

FIGURA C.13. Curva di Lorenz per SAU e Produzione di Frutta in guscio.



Considerando solo le 65 aziende che sono riuscite a produrre l'indice di concentrazione è 0,90. Anche in questo comparto troviamo un outlier: un'azienda di Poppi (AR) che produce 5000 quintali di frutta in guscio equivalenti al 60% circa della produzione totale; la SAU di questa azienda è di 6600 ettari contro i 35000 ettari di SAU dell'azienda che vanta la maggior estensione ma che nell'anno 2003 non ha prodotto.

FIGURA C.14. Curva di Lorenz e Distribuzione di Frequenza per la Frutta in Guscio per le aziende che hanno avuto un raccolto.



ALTRE COLTIVAZIONI LEGNOSE AGRARIE

Le coltivazioni di frutta fresca di origine sub-tropicale, di agrumi, di legnose agrarie in serra e le altre coltivazioni legnose rivestono un ruolo di scarsa rilevanza. Eccetto le altre coltivazioni legnose i comparti citati hanno una produzione prossima a zero e SAU totali di dimensioni insignificanti. Le alte coltivazioni legnose hanno SAU totale pari a 14824 ettari con una media per 18 aziende di 823,56 (errore standard 248,81). La produzione totale ottenuta da solo 2 aziende è pari a 17 quintali; l'indice di concentrazione è 0,62 per la SAU e 0,93 per la produzione.

APPENDICE D

TABELLA D.1. Stima post-stratificata della produzione media per azienda di seminavi per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	12,08	6,62	0,00	25,06	54,82%
2	2,28	2,61	0,00	7,40	114,76%
3.1	13,52	11,54	0,00	36,14	85,35%
3.2	27,95	22,65	0,00	72,34	81,04%
4	19,77	13,01	0,00	45,27	65,80%
5	63,57	31,17	2,48	124,66	49,03%
6	18,74	13,00	0,00	44,22	69,36%
7.1	4,27	3,64	0,00	11,39	85,24%
7.2	2,58	1,19	0,25	4,90	46,09%
8	152,01	164,54	0,00	474,51	108,24%
9.1	566,18	208,45	157,61	974,75	36,82%
9.2	54,72	77,49	0,00	206,59	141,60%
9.3	50,54	41,67	0,00	132,21	82,44%
9.4	27,24	28,19	0,00	82,50	103,48%
9.5	26,92	29,79	0,00	85,32	110,66%
10.1	73,06	35,10	4,26	141,86	48,05%
10.2	178,34	80,77	20,02	336,65	45,29%
11	167,16	168,88	0,00	498,17	101,03%
12	227,43	127,95	0,00	478,20	56,26%
13	238,12	280,82	0,00	788,52	117,93%
14	136,87	91,02	0,00	315,28	66,50%
15.1	208,03	106,16	0,00	416,10	51,03%
15.2	217,72	90,05	41,22	394,22	41,36%
16	355,50	143,52	74,20	636,80	40,37%
17	254,17	226,23	0,00	697,57	89,01%
18	226,57	186,57	0,00	592,25	82,35%
19	68,71	35,60	0,00	138,49	51,82%
20	173,54	165,70	0,00	498,32	95,49%
21	555,57	353,23	0,00	1247,90	63,58%
22	293,37	255,05	0,00	793,27	86,94%
23	130,27	174,32	0,00	471,94	133,81%
24	35,67	44,98	0,00	123,83	126,11%
25	272,97	130,15	17,88	528,05	47,68%
26	184,49	67,74	51,72	317,26	36,72%
27	92,88	110,13	0,00	308,74	118,57%
28	205,34	111,86	0,00	424,58	54,47%
29	399,21	239,33	0,00	868,30	59,95%
30	443,82	135,24	178,75	708,90	30,47%
31	101,06	53,82	0,00	206,55	53,25%
32	495,47	150,53	200,43	790,51	30,38%
33.1	836,73	273,22	301,22	1372,24	32,65%
33.2	327,22	65,40	199,05	455,40	19,99%

TABELLA D.2. Stima EBLUP della produzione media per azienda di seminavi per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	5,70	4,09	0,00	13,73	71,83%
2	1,97	2,78	0,00	7,42	141,52%
3.1	8,15	4,16	0,00	16,30	51,00%
3.2	5,11	4,14	0,00	13,23	81,04%
4	9,24	4,15	1,10	17,38	44,95%
5	19,22	4,25	10,89	27,55	22,11%
6	18,19	4,26	9,84	26,55	23,42%
7.1	4,04	3,45	0,00	10,79	85,33%
7.2	3,16	1,28	0,65	5,66	40,47%
8	39,02	5,20	28,84	49,21	13,32%
9.1	218,64	22,32	174,90	262,38	10,21%
9.2	42,65	5,44	31,99	53,30	12,75%
9.3	30,86	4,72	21,62	40,11	15,28%
9.4	44,66	5,57	33,73	55,58	12,48%
9.5	36,09	5,02	26,26	45,92	13,90%
10.1	60,12	6,74	46,91	73,32	11,21%
10.2	123,25	12,64	98,47	148,04	10,26%
11	59,22	6,70	46,08	72,35	11,32%
12	103,40	10,72	82,39	124,41	10,37%
13	109,92	11,36	87,65	132,19	10,34%
14	145,69	14,89	116,50	174,88	10,22%
15.1	116,61	12,00	93,09	140,13	10,29%
15.2	256,59	26,21	205,23	307,95	10,21%
16	120,56	12,39	96,28	144,84	10,28%
17	16,54	4,19	8,34	24,75	25,30%
18	136,03	13,94	108,71	163,35	10,25%
19	183,28	18,66	146,70	219,85	10,18%
20	178,59	18,23	142,86	214,32	10,21%
21	482,95	49,71	385,51	580,39	10,29%
22	219,26	22,39	175,37	263,14	10,21%
23	147,99	15,14	118,32	177,66	10,23%
24	44,33	5,55	33,45	55,22	12,52%
25	108,21	11,18	86,29	130,12	10,33%
26	166,95	17,01	133,61	200,28	10,19%
27	51,97	6,12	39,97	63,96	11,77%
28	88,46	9,29	70,24	106,68	10,51%
29	148,03	15,14	118,36	177,70	10,23%
30	308,23	31,54	246,40	370,05	10,23%
31	147,68	15,07	118,13	177,22	10,21%
32	233,93	23,88	187,13	280,73	10,21%
33.1	292,27	29,91	233,65	350,89	10,23%
33.2	265,20	27,03	212,21	318,19	10,19%

TABELLA D.3. Stima Spatial EBLUP della produzione media per azienda di seminavi per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	5,77	6,90	0,00	19,30	119,57%
2	2,14	3,10	0,00	8,20	144,97%
3.1	8,61	5,83	0,00	20,04	67,69%
3.2	5,12	5,40	0,00	15,70	105,33%
4	9,36	5,45	0,00	20,03	58,18%
5	19,33	4,64	10,23	28,43	24,03%
6	17,75	5,84	6,31	29,19	32,89%
7.1	3,69	4,30	0,00	12,11	116,53%
7.2	3,16	1,34	0,54	5,77	42,35%
8	38,49	6,57	25,61	51,37	17,08%
9.1	218,93	22,36	175,12	262,75	10,21%
9.2	42,63	5,53	31,79	53,47	12,98%
9.3	30,75	4,88	21,19	40,31	15,86%
9.4	44,54	5,63	33,51	55,57	12,63%
9.5	36,04	5,10	26,05	46,04	14,15%
10.1	59,64	7,12	45,69	73,59	11,94%
10.2	123,29	12,65	98,50	148,07	10,26%
11	59,28	6,75	46,04	72,51	11,39%
12	103,55	10,74	82,50	124,60	10,37%
13	110,17	11,43	87,78	132,57	10,37%
14	145,90	14,90	116,69	175,11	10,22%
15.1	116,73	12,01	93,19	140,28	10,29%
15.2	256,97	26,22	205,58	308,35	10,20%
16	120,68	12,40	96,38	144,99	10,27%
17	16,46	4,20	8,23	24,68	25,50%
18	136,12	13,93	108,81	163,43	10,24%
19	183,50	18,67	146,91	220,08	10,17%
20	178,72	18,23	142,99	214,45	10,20%
21	483,80	49,76	386,27	581,33	10,29%
22	219,51	22,39	175,62	263,40	10,20%
23	148,04	15,14	118,37	177,71	10,23%
24	44,28	5,61	33,28	55,28	12,67%
25	108,30	11,22	86,32	130,29	10,36%
26	167,18	17,02	133,82	200,55	10,18%
27	51,98	6,18	39,86	64,11	11,90%
28	88,53	9,31	70,28	106,79	10,52%
29	148,24	15,15	118,55	177,93	10,22%
30	308,72	31,56	246,86	370,58	10,22%
31	147,91	15,08	118,36	177,46	10,19%
32	234,30	23,89	187,49	281,12	10,19%
33.1	292,79	29,92	234,14	351,43	10,22%
33.2	265,63	27,05	212,60	318,65	10,18%

TABELLA D.4. Stima post-stratificata della produzione media per azienda di cereali per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	4,18	2,43	0,00	8,94	58,06%
2	0,10	0,14	0,00	0,37	135,07%
3.1	2,22	1,79	0,00	5,72	80,46%
3.2	22,05	11,87	0,00	45,32	53,86%
4	14,39	12,72	0,00	39,32	88,40%
5	53,77	27,17	0,51	107,04	50,53%
6	18,00	12,96	0,00	43,41	72,01%
7.1	1,25	0,62	0,04	2,46	49,27%
7.2	2,32	1,00	0,36	4,28	43,05%
8	81,57	88,59	0,00	255,20	108,61%
9.1	97,74	55,61	0,00	206,73	56,89%
9.2	12,19	11,38	0,00	34,50	93,34%
9.3	28,51	24,15	0,00	75,84	84,72%
9.4	8,76	13,90	0,00	36,00	158,63%
9.5	21,14	22,93	0,00	66,08	108,44%
10.1	31,39	20,03	0,00	70,64	63,80%
10.2	133,95	66,16	4,27	263,64	49,39%
11	114,76	95,36	0,00	301,66	83,09%
12	89,67	55,02	0,00	197,50	61,36%
13	106,56	90,66	0,00	284,25	85,08%
14	40,83	43,66	0,00	126,40	106,93%
15.1	75,27	37,14	2,49	148,06	49,33%
15.2	137,64	61,99	16,13	259,14	45,04%
16	53,87	23,98	6,86	100,87	44,52%
17	0,00	0,00	0,00	0,00	-
18	61,92	52,30	0,00	164,43	84,46%
19	54,48	29,10	0,00	111,52	53,41%
20	146,57	106,70	0,00	355,71	72,80%
21	276,51	112,85	55,34	497,69	40,81%
22	198,11	95,14	11,63	384,59	48,02%
23	89,72	84,10	0,00	254,55	93,73%
24	15,13	32,83	0,00	79,48	216,98%
25	109,25	59,38	0,00	225,64	54,36%
26	85,52	27,76	31,12	139,92	32,45%
27	58,75	32,57	0,00	122,58	55,43%
28	104,22	56,30	0,00	214,56	54,02%
29	144,39	60,74	25,35	263,43	42,06%
30	311,80	100,51	114,79	508,81	32,24%
31	47,47	18,01	12,16	82,78	37,95%
32	155,49	43,42	70,38	240,60	27,93%
33.1	157,50	66,14	27,85	287,14	42,00%
33.2	122,20	34,35	54,87	189,54	28,11%

TABELLA D.5. Stima EBLUP della produzione media per azienda di cereali per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	2,64	2,05	0,00	6,66	77,90%
2	0,10	0,14	0,00	0,37	136,55%
3.1	2,50	1,66	0,00	5,75	66,39%
3.2	1,50	2,88	0,00	7,14	192,02%
4	4,13	2,87	0,00	9,75	69,45%
5	11,95	2,99	6,08	17,82	25,05%
6	10,77	2,94	5,00	16,53	27,31%
7.1	1,18	0,62	0,00	2,39	52,52%
7.2	2,35	0,99	0,41	4,29	42,16%
8	21,71	3,42	15,01	28,41	15,74%
9.1	91,16	9,31	72,91	109,41	10,21%
9.2	14,56	3,05	8,59	20,54	20,92%
9.3	16,03	3,13	9,89	22,17	19,54%
9.4	15,97	3,12	9,85	22,08	19,55%
9.5	21,48	3,39	14,84	28,12	15,78%
10.1	31,43	3,99	23,60	39,25	12,70%
10.2	57,79	6,18	45,67	69,91	10,70%
11	26,98	3,73	19,67	34,29	13,82%
12	52,20	5,69	41,05	63,35	10,90%
13	56,35	6,06	44,46	68,23	10,76%
14	68,02	7,12	54,06	81,98	10,47%
15.1	65,96	6,91	52,41	79,50	10,48%
15.2	161,30	16,23	129,50	193,10	10,06%
16	50,46	5,50	39,68	61,24	10,90%
17	0,00	0,00	0,00	0,00	-
18	66,51	6,98	52,82	80,19	10,50%
19	85,00	8,70	67,95	102,05	10,23%
20	93,17	9,52	74,52	111,83	10,22%
21	284,60	28,63	228,48	340,73	10,06%
22	101,01	10,27	80,88	121,14	10,17%
23	73,14	7,61	58,23	88,05	10,40%
24	18,47	3,25	12,10	24,84	17,60%
25	40,79	4,73	31,53	50,06	11,59%
26	59,31	6,27	47,02	71,60	10,57%
27	22,74	3,45	15,97	29,51	15,19%
28	43,18	4,92	33,54	52,83	11,39%
29	80,80	8,31	64,51	97,10	10,29%
30	184,98	18,60	148,52	221,43	10,05%
31	77,56	7,94	62,00	93,12	10,24%
32	111,21	11,23	89,21	133,22	10,10%
33.1	137,67	13,87	110,48	164,86	10,08%
33.2	123,68	12,44	99,30	148,07	10,06%

TABELLA D.6. Stima Spatial EBLUP della produzione media per azienda di cereali per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	2,61	3,00	0,00	8,49	114,84%
2	0,10	0,14	0,00	0,37	136,68%
3.1	2,52	1,77	0,00	5,98	70,18%
3.2	1,49	3,69	0,00	8,71	248,18%
4	4,09	3,67	0,00	11,29	89,81%
5	11,93	3,12	5,82	18,05	26,15%
6	10,78	3,55	3,83	17,73	32,88%
7.1	1,17	0,63	0,00	2,40	53,23%
7.2	2,35	1,02	0,35	4,34	43,39%
8	21,71	3,96	13,94	29,47	18,25%
9.1	91,16	9,33	72,87	109,44	10,23%
9.2	14,54	3,06	8,54	20,53	21,06%
9.3	15,99	3,16	9,79	22,19	19,77%
9.4	15,93	3,13	9,80	22,06	19,63%
9.5	21,44	3,42	14,74	28,15	15,95%
10.1	31,41	4,07	23,42	39,39	12,97%
10.2	57,77	6,19	45,65	69,90	10,71%
11	26,96	3,74	19,63	34,29	13,86%
12	52,19	5,69	41,04	63,34	10,90%
13	56,34	6,07	44,44	68,24	10,78%
14	68,01	7,12	54,05	81,97	10,47%
15.1	65,94	6,91	52,39	79,49	10,49%
15.2	161,33	16,23	129,53	193,14	10,06%
16	50,44	5,50	39,66	61,22	10,90%
17	0,00	0,00	0,00	0,00	-
18	66,49	6,98	52,81	80,18	10,50%
19	85,00	8,70	67,95	102,05	10,23%
20	93,17	9,52	74,52	111,83	10,21%
21	284,71	28,64	228,56	340,85	10,06%
22	101,02	10,27	80,88	121,15	10,17%
23	73,13	7,61	58,21	88,04	10,41%
24	18,44	3,26	12,05	24,83	17,69%
25	40,78	4,74	31,48	50,07	11,63%
26	59,30	6,28	47,00	71,60	10,58%
27	22,72	3,46	15,94	29,50	15,22%
28	43,17	4,92	33,52	52,82	11,40%
29	80,80	8,31	64,51	97,09	10,29%
30	185,01	18,60	148,56	221,46	10,05%
31	77,56	7,94	62,00	93,12	10,23%
32	111,21	11,23	89,21	133,21	10,09%
33.1	137,70	13,87	110,50	164,89	10,08%
33.2	123,68	12,44	99,29	148,07	10,06%

TABELLA D.7. Stima post-stratificata della produzione media per azienda di coltivazioni legnose agrarie per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	8,29	1,12	6,10	10,48	13,47%
2	2,19	0,56	1,09	3,29	25,69%
3.1	10,39	2,59	5,31	15,46	24,91%
3.2	14,51	4,52	5,65	23,37	31,16%
4	6,04	1,01	4,06	8,02	16,75%
5	10,69	3,37	4,08	17,31	31,56%
6	13,01	5,54	2,15	23,86	42,59%
7.1	0,07	0,08	0,00	0,23	122,76%
7.2	4,11	1,44	1,29	6,94	35,00%
8	19,97	25,17	0,00	69,31	126,03%
9.1	70,37	19,77	31,62	109,13	28,10%
9.2	183,53	123,28	0,00	425,15	67,17%
9.3	0,95	0,73	0,00	2,37	76,52%
9.4	0,00	0,00	0,00	0,00	-
9.5	25,71	35,14	0,00	94,58	136,67%
10.1	11,20	8,12	0,00	27,11	72,49%
10.2	10,22	13,33	0,00	36,34	130,45%
11	28,13	18,84	0,00	65,06	66,97%
12	27,13	24,55	0,00	75,24	90,50%
13	22,98	12,39	0,00	47,26	53,93%
14	12,68	4,29	4,28	21,09	33,81%
15.1	74,75	52,13	0,00	176,93	69,74%
15.2	14,49	7,55	0,00	29,29	52,09%
16	23,99	5,47	13,27	34,71	22,79%
17	99,83	55,02	0,00	207,67	55,11%
18	27,82	9,03	10,13	45,51	32,44%
19	70,21	51,13	0,00	170,42	72,82%
20	44,03	51,85	0,00	145,66	117,75%
21	13,39	13,28	0,00	39,43	99,18%
22	29,52	60,43	0,00	147,97	204,69%
23	399,68	158,52	88,98	710,37	39,66%
24	1,61	2,25	0,00	6,02	140,07%
25	0,00	0,00	0,00	0,00	-
26	0,00	0,00	0,00	0,00	-
27	1,62	2,46	0,00	6,44	151,89%
28	0,39	0,39	0,00	1,16	100,39%
29	66,19	39,91	0,00	144,41	60,29%
30	20,90	7,24	6,71	35,09	34,63%
31	18,39	5,39	7,83	28,94	29,29%
32	34,68	8,76	17,51	51,84	25,26%
33.1	75,72	25,14	26,45	124,99	33,20%
33.2	62,52	12,80	37,42	87,61	20,48%

TABELLA D.8. Stima EBLUP della produzione media per azienda di coltivazioni legnose agrarie per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	8,33	1,11	6,15	10,51	13,34%
2	2,22	0,56	1,12	3,32	25,29%
3.1	10,45	2,52	5,51	15,39	24,10%
3.2	13,92	4,17	5,74	22,10	29,99%
4	6,10	1,01	4,12	8,07	16,53%
5	10,76	3,23	4,44	17,09	29,97%
6	12,58	4,92	2,94	22,22	39,11%
7.1	0,07	0,08	0,00	0,23	121,22%
7.2	4,28	1,43	1,48	7,08	33,40%
8	12,49	9,21	0,00	30,53	73,72%
9.1	22,66	8,96	5,09	40,22	39,55%
9.2	12,80	10,54	0,00	33,45	82,30%
9.3	1,01	0,72	0,00	2,43	71,74%
9.4	0,00	0,00	0,00	0,00	-
9.5	12,52	9,57	0,00	31,27	76,43%
10.1	11,34	6,43	0,00	23,94	56,68%
10.2	11,13	8,18	0,00	27,16	73,45%
11	14,75	8,85	0,00	32,11	60,01%
12	13,45	9,18	0,00	31,45	68,27%
13	15,66	7,88	0,20	31,11	50,36%
14	12,46	3,99	4,64	20,27	32,02%
15.1	13,51	9,64	0,00	32,41	71,34%
15.2	13,31	6,12	1,32	25,31	45,98%
16	20,88	4,86	11,34	30,41	23,30%
17	13,91	9,68	0,00	32,89	69,59%
18	20,11	6,81	6,76	33,46	33,86%
19	13,63	9,97	0,00	33,17	73,15%
20	12,57	9,68	0,00	31,54	76,97%
21	12,09	8,05	0,00	27,87	66,59%
22	11,86	9,66	0,00	30,80	81,46%
23	13,74	14,12	0,00	41,42	102,82%
24	2,12	2,21	0,00	6,45	103,79%
25	0,00	0,00	0,00	0,00	-
26	0,00	0,00	0,00	0,00	-
27	2,23	2,40	0,00	6,93	107,76%
28	0,41	0,39	0,00	1,17	95,82%
29	14,41	9,54	0,00	33,11	66,20%
30	17,50	6,00	5,74	29,26	34,28%
31	16,71	4,81	7,29	26,13	28,77%
32	24,05	6,69	10,94	37,16	27,81%
33.1	19,56	9,22	1,50	37,63	47,11%
33.2	29,65	7,95	14,06	45,23	26,82%

TABELLA D.9. Stima Spatial EBLUP della produzione media per azienda di coltivazioni legnose agrarie per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	8,12	1,09	5,98	10,27	13,47%
2	2,29	0,56	1,19	3,38	24,44%
3.1	9,36	2,33	4,79	13,93	24,92%
3.2	10,53	3,31	4,04	17,02	31,44%
4	6,12	1,00	4,17	8,07	16,27%
5	11,12	2,97	5,30	16,93	26,71%
6	8,55	3,61	1,47	15,63	42,27%
7.1	0,07	0,08	0,00	0,24	117,87%
7.2	4,13	1,40	1,40	6,87	33,79%
8	3,18	4,55	0,00	12,09	142,98%
9.1	7,03	4,84	0,00	16,51	68,83%
9.2	2,91	5,37	0,00	13,43	184,77%
9.3	1,01	0,72	0,00	2,43	71,66%
9.4	0,00	0,00	0,00	0,00	-
9.5	1,57	4,44	0,00	10,28	282,28%
10.1	7,12	4,61	0,00	16,16	64,82%
10.2	10,02	4,99	0,24	19,80	49,79%
11	11,23	5,16	1,12	21,34	45,93%
12	13,46	5,23	3,20	23,71	38,89%
13	12,52	4,54	3,63	21,42	36,23%
14	13,36	3,56	6,39	20,34	26,63%
15.1	15,21	5,00	5,42	25,01	32,84%
15.2	19,26	5,12	9,23	29,30	26,58%
16	18,83	4,05	10,89	26,77	21,52%
17	8,90	5,56	0,00	19,80	62,50%
18	21,73	4,82	12,29	31,17	22,17%
19	13,42	5,88	1,89	24,94	43,82%
20	10,62	5,71	0,00	21,81	53,73%
21	7,67	5,81	0,00	19,06	75,80%
22	19,20	5,96	7,52	30,89	31,04%
23	4,32	7,18	0,00	18,39	166,32%
24	1,81	2,11	0,00	5,95	117,06%
25	0,00	0,00	0,00	0,00	-
26	0,00	0,00	0,00	0,00	-
27	0,62	2,24	0,00	5,01	359,18%
28	0,44	0,39	0,00	1,20	88,98%
29	7,06	4,62	0,00	16,11	65,44%
30	18,64	4,76	9,31	27,96	25,53%
31	21,68	4,03	13,78	29,57	18,58%
32	34,96	5,97	23,27	46,66	17,07%
33.1	24,37	5,58	13,44	35,31	22,89%
33.2	30,37	5,68	19,24	41,51	18,71%

TABELLA D.10. Stima post-stratificata della produzione media per azienda di vite per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	7,01	1,00	5,05	8,97	14,27%
2	0,71	0,44	0,00	1,58	61,67%
3.1	2,13	0,96	0,25	4,00	44,98%
3.2	7,63	3,50	0,78	14,49	45,83%
4	1,59	0,82	0,00	3,19	51,37%
5	6,43	3,00	0,55	12,32	46,68%
6	7,61	5,18	0,00	17,76	68,02%
7.1	0,00	0,00	0,00	0,00	-
7.2	3,88	3,36	0,00	10,46	86,58%
8	14,91	22,46	0,00	58,93	150,69%
9.1	1,57	2,85	0,00	7,15	181,69%
9.2	122,34	110,20	0,00	338,34	90,08%
9.3	27,06	19,06	0,00	64,42	70,43%
9.4	165,20	76,11	16,02	314,38	46,07%
9.5	62,11	41,55	0,00	143,55	66,91%
10.1	182,96	67,38	50,89	315,04	36,83%
10.2	203,88	70,86	65,00	342,76	34,75%
11	0,00	0,00	0,00	0,00	-
12	22,36	13,92	0,00	49,63	62,24%
13	17,26	11,22	0,00	39,24	64,98%
14	4,39	1,83	0,80	7,98	41,71%
15.1	32,37	48,53	0,00	127,48	149,92%
15.2	8,38	6,26	0,00	20,65	74,74%
16	5,97	1,92	2,20	9,74	32,22%
17	29,33	8,84	12,00	46,67	30,15%
18	13,12	5,66	2,04	24,21	43,11%
19	79,54	51,42	0,00	180,33	64,65%
20	36,20	50,78	0,00	135,73	140,28%
21	10,36	13,00	0,00	35,84	125,49%
22	28,06	56,64	0,00	139,07	201,87%
23	385,26	145,24	100,60	669,92	37,70%
24	27,19	17,45	0,00	61,38	64,17%
25	6,51	3,03	0,58	12,44	46,48%
26	1,32	1,03	0,00	3,35	78,45%
27	28,69	14,06	1,14	56,24	48,99%
28	17,84	12,07	0,00	41,49	67,63%
29	56,84	38,82	0,00	132,93	68,30%
30	661,97	910,68	0,00	2446,91	137,57%
31	5,20	3,78	0,00	12,62	72,75%
32	18,08	6,56	5,21	30,94	36,30%
33.1	62,06	19,54	23,75	100,37	31,49%
33.2	52,40	12,05	28,77	76,03	23,01%

TABELLA D.11. Stima EBLUP della produzione media per azienda di vite per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	6,94	0,99	5,01	8,87	14,21%
2	0,74	0,44	0,00	1,60	59,37%
3.1	2,08	0,94	0,23	3,93	45,39%
3.2	5,97	2,92	0,25	11,69	48,87%
4	1,49	0,81	0,00	3,07	54,51%
5	6,55	2,62	1,42	11,68	39,99%
6	4,51	3,63	0,00	11,63	80,44%
7.1	0,00	0,00	0,00	0,00	-
7.2	3,24	2,84	0,00	8,80	87,83%
8	9,56	4,63	0,48	18,63	48,44%
9.1	2,39	2,52	0,00	7,33	105,36%
9.2	66,57	9,92	47,13	86,02	14,90%
9.3	21,07	5,00	11,28	30,87	23,72%
9.4	165,59	23,57	119,39	211,79	14,23%
9.5	31,39	5,87	19,89	42,88	18,69%
10.1	74,07	10,82	52,87	95,27	14,60%
10.2	84,62	12,22	60,67	108,56	14,44%
11	0,00	0,00	0,00	0,00	-
12	19,19	4,80	9,79	28,59	24,99%
13	5,27	4,42	0,00	13,93	83,83%
14	4,98	1,74	1,56	8,39	35,02%
15.1	23,81	5,30	13,42	34,21	22,27%
15.2	9,70	3,91	2,03	17,36	40,32%
16	7,08	1,83	3,50	10,66	25,80%
17	26,05	4,78	16,68	35,41	18,34%
18	12,90	3,79	5,47	20,34	29,40%
19	103,41	14,81	74,38	132,44	14,32%
20	50,64	7,94	35,08	66,21	15,68%
21	18,82	4,81	9,39	28,24	25,55%
22	15,76	4,86	6,22	25,29	30,87%
23	297,10	42,52	213,76	380,44	14,31%
24	29,44	5,56	18,54	40,35	18,90%
25	5,61	2,64	0,44	10,77	47,02%
26	1,39	1,02	0,00	3,39	73,32%
27	18,54	4,74	9,25	27,83	25,57%
28	11,67	4,48	2,89	20,45	38,38%
29	46,83	7,44	32,24	61,42	15,89%
30	82,25	12,01	58,72	105,79	14,60%
31	5,27	3,06	0,00	11,27	58,13%
32	15,34	4,02	7,46	23,23	26,22%
33.1	35,03	5,94	23,39	46,68	16,96%
33.2	27,71	5,00	17,92	37,51	18,04%

TABELLA D.12. Stima Spatial EBLUP della produzione media per azienda di vite per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	6,90	1,00	4,93	8,86	14,52%
2	0,75	0,44	0,00	1,61	58,78%
3.1	2,10	0,95	0,24	3,96	45,19%
3.2	6,23	3,04	0,27	12,20	48,85%
4	1,46	0,81	0,00	3,06	55,55%
5	6,40	2,66	1,18	11,61	41,60%
6	4,92	3,78	0,00	12,33	76,98%
7.1	0,00	0,00	0,00	0,00	-
7.2	3,44	2,93	0,00	9,18	85,03%
8	9,82	4,96	0,10	19,54	50,51%
9.1	2,32	2,55	0,00	7,33	109,75%
9.2	68,39	10,44	47,92	88,86	15,27%
9.3	21,08	5,25	10,78	31,38	24,92%
9.4	170,63	24,45	122,70	218,55	14,33%
9.5	32,21	6,15	20,16	44,26	19,09%
10.1	74,99	11,18	53,08	96,91	14,91%
10.2	85,01	12,49	60,53	109,48	14,69%
11	0,00	0,00	0,00	0,00	-
12	17,38	5,01	7,57	27,20	28,80%
13	4,76	4,71	0,00	14,00	98,94%
14	4,92	1,76	1,46	8,37	35,83%
15.1	22,14	5,73	10,90	33,38	25,90%
15.2	8,59	4,07	0,60	16,57	47,45%
16	7,01	1,85	3,38	10,64	26,42%
17	26,26	4,82	16,82	35,70	18,35%
18	12,26	3,89	4,63	19,88	31,73%
19	106,14	15,26	76,23	136,05	14,38%
20	51,66	8,13	35,71	67,60	15,75%
21	19,16	5,04	9,29	29,03	26,29%
22	15,75	5,06	5,83	25,67	32,13%
23	306,51	44,25	219,78	393,24	14,44%
24	30,20	5,78	18,88	41,53	19,13%
25	5,47	2,68	0,22	10,73	49,00%
26	1,40	1,03	0,00	3,42	73,24%
27	18,69	5,03	8,82	28,56	26,94%
28	11,55	4,71	2,32	20,78	40,75%
29	47,91	7,70	32,81	63,00	16,08%
30	84,68	12,42	60,33	109,03	14,67%
31	5,55	3,18	0,00	11,79	57,35%
32	16,09	4,28	7,71	24,48	26,58%
33.1	36,79	6,36	24,33	49,26	17,29%
33.2	28,71	5,50	17,92	39,49	19,17%

TABELLA D.13. Stima post-stratificata della produzione media per azienda di olive per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	1,02	0,30	0,43	1,62	29,68%
2	0,99	0,35	0,31	1,68	35,12%
3.1	1,36	0,85	0,00	3,03	62,53%
3.2	4,01	1,17	1,70	6,31	29,32%
4	1,37	0,38	0,63	2,11	27,45%
5	3,72	0,82	2,12	5,33	22,00%
6	5,07	1,25	2,63	7,52	24,57%
7.1	0,07	0,08	0,00	0,23	122,76%
7.2	2,97	0,55	1,89	4,06	18,60%
8	5,06	3,55	0,00	12,02	70,10%
9.1	0,32	0,52	0,00	1,35	162,33%
9.2	44,03	28,47	0,00	99,83	64,66%
9.3	7,87	3,84	0,34	15,39	48,80%
9.4	56,29	16,65	23,66	88,91	29,58%
9.5	16,25	8,59	0,00	33,09	52,90%
10.1	8,39	7,41	0,00	22,91	88,39%
10.2	26,57	9,58	7,79	45,34	36,06%
11	0,00	0,00	0,00	0,00	-
12	4,62	1,24	2,20	7,05	26,75%
13	4,71	3,56	0,00	11,69	75,54%
14	8,29	3,75	0,94	15,65	45,25%
15.1	33,97	9,78	14,80	53,15	28,79%
15.2	6,33	1,64	3,12	9,54	25,88%
16	17,02	4,70	7,80	26,23	27,62%
17	4,00	2,48	0,00	8,86	62,02%
18	8,06	3,47	1,26	14,86	43,06%
19	5,39	2,37	0,73	10,04	44,07%
20	7,85	5,27	0,00	18,18	67,20%
21	2,29	0,89	0,56	4,03	38,61%
22	4,78	6,67	0,00	17,86	139,56%
23	46,97	19,78	8,19	85,75	42,12%
24	3,55	2,13	0,00	7,73	59,99%
25	0,98	0,66	0,00	2,27	67,56%
26	0,62	0,23	0,17	1,07	37,23%
27	5,34	1,60	2,20	8,48	30,00%
28	4,38	1,52	1,40	7,36	34,72%
29	3,85	1,00	1,88	5,81	26,12%
30	71,95	330,49	0,00	719,70	459,35%
31	10,04	3,29	3,60	16,49	32,73%
32	13,16	4,16	5,00	21,33	31,63%
33.1	13,16	5,42	2,54	23,78	41,19%
33.2	9,15	1,73	5,75	12,54	18,93%

TABELLA D.14. Stima EBLUP della produzione media per azienda di olive per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	1,05	0,30	0,46	1,63	28,54%
2	1,00	0,34	0,33	1,67	34,25%
3.1	1,12	0,76	0,00	2,61	67,85%
3.2	2,73	0,95	0,87	4,59	34,79%
4	1,39	0,37	0,67	2,12	26,47%
5	3,45	0,73	2,01	4,89	21,23%
6	4,51	0,98	2,60	6,43	21,65%
7.1	0,07	0,08	0,00	0,24	118,42%
7.2	3,09	0,53	2,06	4,12	17,02%
8	6,03	1,41	3,27	8,78	23,34%
9.1	0,57	0,50	0,00	1,55	88,06%
9.2	15,02	2,27	10,56	19,47	15,13%
9.3	13,76	2,04	9,77	17,75	14,79%
9.4	23,58	3,32	17,06	30,09	14,10%
9.5	11,12	1,82	7,55	14,70	16,41%
10.1	6,70	1,50	3,76	9,63	22,39%
10.2	8,30	1,58	5,20	11,41	19,08%
11	0,00	0,00	0,00	0,00	-
12	4,47	0,98	2,56	6,38	21,80%
13	3,07	1,35	0,44	5,71	43,78%
14	3,79	1,35	1,14	6,44	35,68%
15.1	7,67	1,54	4,66	10,68	20,03%
15.2	6,63	1,17	4,35	8,92	17,58%
16	6,68	1,43	3,88	9,49	21,40%
17	1,84	1,29	0,00	4,36	69,87%
18	8,12	1,49	5,19	11,04	18,42%
19	7,26	1,37	4,57	9,94	18,89%
20	6,77	1,48	3,87	9,68	21,84%
21	3,33	0,79	1,79	4,88	23,64%
22	4,70	1,42	1,92	7,49	30,23%
23	18,73	2,71	13,42	24,04	14,47%
24	5,14	1,26	2,67	7,60	24,49%
25	0,94	0,62	0,00	2,14	65,86%
26	0,63	0,23	0,19	1,08	36,05%
27	4,56	1,11	2,39	6,74	24,31%
28	3,92	1,08	1,80	6,04	27,62%
29	4,23	0,86	2,55	5,91	20,28%
30	7,68	1,58	4,57	10,78	20,64%
31	6,67	1,39	3,95	9,39	20,83%
32	7,64	1,47	4,76	10,52	19,23%
33.1	7,23	1,49	4,31	10,14	20,61%
33.2	6,01	1,15	3,77	8,26	19,05%

TABELLA D.15. Stima Spatial EBLUP della produzione media per azienda di olive per SEL.

SEL	*Media*	*Errore Standard*	*Limite inferiore*	*Limite superiore*	*Coefficiente di variazione*
1	1,07	0,30	0,48	1,65	28,07%
2	0,98	0,34	0,31	1,65	34,80%
3.1	1,18	0,75	0,00	2,64	63,30%
3.2	2,76	0,92	0,96	4,57	33,39%
4	1,43	0,37	0,71	2,16	25,72%
5	3,44	0,73	2,00	4,88	21,30%
6	4,62	0,96	2,73	6,51	20,88%
7.1	0,07	0,08	0,00	0,24	116,69%
7.2	3,06	0,53	2,02	4,09	17,29%
8	4,98	1,39	2,25	7,71	27,93%
9.1	0,46	0,51	0,00	1,45	110,21%
9.2	15,06	2,59	9,97	20,14	17,23%
9.3	13,51	2,32	8,96	18,06	17,17%
9.4	23,98	3,68	16,76	31,20	15,36%
9.5	11,02	2,01	7,09	14,95	18,20%
10.1	5,54	1,51	2,58	8,50	27,27%
10.2	7,48	1,54	4,47	10,50	20,57%
11	0,00	0,00	0,00	0,00	-
12	4,33	0,98	2,40	6,26	22,73%
13	3,32	1,32	0,73	5,91	39,83%
14	4,15	1,45	1,31	6,99	34,93%
15.1	8,60	1,65	5,37	11,83	19,16%
15.2	7,28	1,29	4,76	9,81	17,67%
16	7,12	1,53	4,13	10,11	21,46%
17	1,68	1,31	0,00	4,25	77,78%
18	8,84	1,55	5,80	11,87	17,52%
19	6,95	1,44	4,12	9,78	20,76%
20	5,93	1,49	3,02	8,85	25,08%
21	3,16	0,80	1,59	4,74	25,37%
22	4,29	1,54	1,27	7,31	35,95%
23	18,66	2,90	12,98	24,35	15,54%
24	4,82	1,29	2,30	7,35	26,72%
25	0,76	0,62	0,00	1,98	81,79%
26	0,65	0,23	0,20	1,10	35,36%
27	4,03	1,09	1,89	6,17	27,06%
28	3,31	1,06	1,24	5,39	31,94%
29	4,03	0,86	2,34	5,72	21,41%
30	7,98	1,63	4,78	11,17	20,43%
31	8,13	1,51	5,18	11,08	18,52%
32	9,85	1,82	6,28	13,43	18,51%
33.1	9,24	1,58	6,13	12,34	17,15%
33.2	7,21	1,33	4,60	9,82	18,47%

BIBLIOGRAFIA

BAILEY, T.C., GATREL, A.C. (1995): *Interactive Spatial Data Analysis*. Longman, London.

BRACKSTONE, G.J. (1987): *Small Area Data: Policy Issues and Technical Challenges, in Plateck R., Rao J. N. K., Särdnal C. E. and Singh M. P. (Eds), Small Area Statistics*. Wiley, New York, 3-20.

CHAMBERS, R.L., SKINNER, C.J. (2003): *Analysis of Survey Data*. John Wiley & Sons, Ltd.

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

CLIFF, A.D., ORD, J.K. (1981): *Spatial Processes. Model and Applications.* Pion Limited, London.

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, 68, 626-632

GHOSH, M., RAO, J.N.K (1994): *Small Area Estimation: An Apprisal (with discussion).* Statistical Science, 9, 1, 55-93.

GONZALEZ, M.E. (1973): *Use and Evaluationof Synthetic Estimates.* Proceeding of the Social Statistics Section, American Statistical Association, 33-36.

GONZALEZ, M.E., WAKESBERG, J. (1973): *Estimation of the Error of Synthetic Estimates.* Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

GREENE, W.H. (2003): *Econometric Analysis.* Prentice Hall.

GRIFFITH, D.A. (1988): *Advanced Spatial Statistics.* Kluwer, Dordrecht.

GUARNERA, U., LUZI, O. (2005): *Valutazione del trattamento degli errori di misura e di risposta nell'indagine SPA.* Convegno AGRI@STAT, "Verso un nuovo sistema di statistiche agricole", Firenze, 30 – 31 maggio 2005.

KACKAR, R.N., HARVILLE, D.A. (1981): *Unbiasdness of Two-stage Estimation and Prediction Procedures for Mixed Models.* Communication in Statistics, Series A, 10, 1249-1261.

KACKAR, R.N., HARVILLE, D.A. (1984): *Approximation for standard errors of estimators for fixed and random effects in mixed models.* Journal of the American Statistical Association, 79, 853-862.

MOOD, M.A., GRAYBILL, A.F., DUANE, C.B. (1991): *Introduzione alla Statisitca.* McGraw-Hill.

PRASAD, N., RAO, J.N.K. (1990): *The Estimation of the Mean Squared Error of Small-Area Estimators.* Journal of the American Statistical Association, 85, 409, 163-71.

PRATESI, M., SALVATI, N. (2004): *Small Area Estimation: the EBLUP estimator with autoregressive random area effects.* Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa.

PRATESI, M., SALVATI, N. (2004): *Spatial EBLUP in agricoltural surveys: an application based on Italian Census data.* Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa.

PURCELL, N.J., LINACRE, A. (1976): *Technique for the Estimation of Small Area Characteristics.* 3rd Australian Statistical Conference, Melbourne.

PURCELL, N.J., KISH, L. (1980): *Postcensal estimates for local areas (or domains).* International Statistical Review, 48, 3-18

RAO, J.N.K. (2003): *Small Area Estimation.* Wiley, London.

ROBINSON, J.K. (1991): *That BLUP is a Good Thing: The Estimation of Random Effects.* Statistical Science, 6, 1, 15-51.

RUBIN, D.B. (1976): *Inference and missing data*. Biometrika, 63, 581-92.

SAEI, A., CHAMBERS, R.L. (2003): *Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects.* Southampton Statistical Sciences Research Institute.

SAEI. A., CHAMBERS. R.L. (2003): *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models.* Southampton Statistical Sciences Research Institute.

SAEI, A., CHAMBERS, R.L. (2003): *Empirical Best Linear Unbiased Prediction for Out of Samples Areas*. Southampton Statistical Sciences Research Institute.

SAEI, A., CHAMBERS, R.L (2003): *Out of Samples Estimation for Small Area Level Data*. Southampton Statistical Sciences Research Institute.

SALVATI, N. (2004): *La correlazione spaziale nella stima per piccole aree: metodi proposti e casi di studio.* Tesi di Dottorato in Statistica Applicata, Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze.

SÄRDNAL, C.E., SWENSSON, B., WRETMAN, J.H. (1992): *Model Assisted Survey Sampling.* Springer-Verlag, New York.

SCOTT, A.J. (1977a): *Some comments on the problem of randomisation in survey sampling.* Sankhya, C, 39, 1-9.

SEARLE, S.R., McCULLOCH, C.E., CASELLA, J. (1992): *Variance Components.* Wiley, New York.

SUGDEN, R.A., SMITH, T.M.F. (1984): *Ignorable and informative designs in survey sampling inference.* Biometrika, 71, 495-506.

TONNELLATO, S.F. (2003): *Introduzione alla statistica spaziale.* Dipartimento di Statistica, Università Ca'Foscari Venezia.

UPTON, G., FINGLETON, B. (1985): *Spatial Data Analysis by Example. Point pattern and quantitative data. Volume 1.* John Wiley & Sons, Ltd.

WOLTER, K. (1985): *Introduction to variance estimation.* Springer-Verlag, New York.

ZAINI, S., NAPOLITANO, P. (1992): *Problemi di rilevazione e di presentazione dei dati spaziali. In Avanzamenti metodologici e statistiche ufficiali.* ISTAT.