Daniela Cocchi* Maria Rosaria Ferrante**
Silvia Pacei*

La stima di proporzioni per piccole aree nell'indagine sulla struttura delle aziende agricole

Serie Ricerche 2000, n.2

*Dip. Sc. Statistiche "P. Fortunati", Università di Bologna
**Dip. Economia Politica, Università della Calabria

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Università degli studi di Bologna

Il lavoro è frutto della collaborazione tra i tre autori. Per quanto riguarda la stesura, Daniela Cocchi ha curato i capitoli 1 e 5, Maria Rosaria Ferrante ha curato i paragrafi 2.1 e 2.2 del capitolo 2 ed il capitolo 3, Silvia Pacei ha curato il paragrafo 2.3 del capitolo 2 ed il capitolo 4.

Finito di stampare nel mese di Marzo 2000
presso le Officine Grafiche Tecnoprint
Via del Legatore 3, Bologna

INDICE

1. Introduzione

Il presente lavoro è frutto di una ricerca condotta nell'ambito dell'attività di una Convenzione fra la Regione Emilia Romagna ed il Dipartimento di Scienze Statistiche sulla "Rappresentatività delle indagini strutturali dell'UE in Emilia Romagna". Gli obiettivi posti dalla Regione erano innanzitutto di vagliare la possibilità di ottenere stime soddisfacenti dalle informazioni rilevate nell'ambito dell'indagine ISTAT sulla "Struttura delle aziende agricole", per provincia e per zona altimetrica (livelli territoriali ridotti rispetto a quello per cui le stime correnti risultano affidabili) mediante le procedure di stima correnti. Nel caso in cui le stime così ottenute non fossero risultate affidabili, si richiedeva di mettere a punto una procedura di stima cosiddetta per "piccole aree" che consentisse di ottenere stime con un livello di variabilità accettabile per provincia e zona altimetrica. Poiché la grandezza ritenuta più interessante dalla Regione ai fini della caratterizzazione di provincia e zona altimetrica in termini di struttura agricola è stata la quota (e quindi il numero) delle aziende agricole in classi di "superficie agricola utilizzata" (SAU), una *proxy* della dimensione dell'azienda, ci si è concentrati sulla stima di tale parametro.

Come già in parte richiamato, la rilevazione ISTAT fornisce stime affidabili a livello regionale. L'impiego di uno stimatore "diretto" per provincia e per altimetria potrebbe dunque condurre a risultati particolarmente instabili, a causa dello scarso numero di unità selezionate per tali "piccole aree" (da ora in poi PA). Tuttavia, le stime riferite alle province ed alle zone altimetriche risulterebbero di indubbia utilità in un'ottica di programmazione e pianificazione delle politiche economiche agrarie, in particolare per gli organi preposti a gestire ed indirizzare tali politiche.

Tra i metodi suggeriti in letteratura per risolvere questo problema appaiono particolarmente interessanti quegli stimatori che connettono le aree in questione attraverso modelli statistici e sfruttano informazioni supplementari di tipo censuario o amministrativo. Infatti, poiché la teoria del campionamento basata esclusivamente sul disegno campionario non offre soluzioni soddisfacenti al problema, sembra opportuno abbandonare o, almeno, integrare l'ottica dell'inferenza basata

sul disegno con la teoria inferenziale basata sul modello. La stima riferita alla piccola area risulta, in tale ambito, una previsione effettuata sulla base di un modello di superpopolazione che descrive il fenomeno d'interesse.

Tuttavia, i metodi basati sul modello suggeriti in letteratura per la stima relativa a PA sono prevalentemente concepiti per variabili quantitative, ossia per la stima di medie o totali mediante l'adattamento di modelli di regressione lineare, mentre decisamente limitato è il numero di contributi in tema di stima di proporzioni (Chand e Alexander, 1995). Le tecniche inferenziali classiche, che si basano sull'ipotesi di normalità distributiva della variabile di interesse, impiegate, ad esempio, da Fay ed Herriot (1979) per stimare il reddito entro PA, non sono direttamente applicabili al problema della stima di proporzioni per variabili discrete. In particolare, negli scarsi ma interessanti contributi che affrontano il problema della stima di proporzioni per PA, ci si concentra principalmente sull'impiego dello stimatore composto ottenuto come media ponderata dello stimatore diretto e di uno stimatore sintetico (MacGibbon e Tomberlin, 1989; Thomsen e Holmøy, 1996), e sulla stima di modelli di regressione logistica eventualmente con effetti casuali specifici per le PA (Chand e Alexander, 1995; Farrell *et al.,* 1997). Ovviamente, non è detto che queste tecniche garantiscano un incremento di efficienza ed il miglioramento delle stime deve essere valutato caso per caso.

In vista della carenza in letteratura di una trattazione organica ed esaustiva delle tecniche di stima di proporzioni per PA, nel presente lavoro saranno dapprima presentati i principali metodi utili a risolvere i problemi sopra richiamati e le relative caratteristiche e proprietà degli stimatori, valutate in termini comparativi. Oltre all'usuale stimatore diretto, basato sul disegno e notoriamente inefficiente, si prenderanno in considerazione altri stimatori per PA specifici per proporzioni o frequenze. Lo stimatore sintetico, ad esempio, si deriva semplicemente sulla base di un'ipotesi di uguaglianza tra le stime ottenute per un livello di aggregazione maggiore rispetto a quello della PA, e quelle relative alle PA stesse. Tale metodo consente di ridurre la varianza dello stimatore diretto, ma può produrre stimatori distorti.

Lo stimatore composto, invece, basandosi su una media ponderata dei due stimatori menzionati (sintetico e diretto), si propone di individuare un punto di bilanciamento tra distorsione e variabilità ad essi associate, così da minimizzare l'errore quadratico medio delle stime finali (Ghosh e Rao, 1994). Le scelte riguardanti il tipo di stimatori (sia sintetico che diretto) e la tecnica di costruzione dei pesi di ponderazione definiscono vari tipi di stimatore composto (basato o meno su variabili ausiliarie, ed in questo ultimo caso su un modello di regressione logistica, ad effetti casuali, ecc.). Fra le soluzioni alternative suggerite in letteratura si è scelto in questa sede di impiegare, per la stima della quota di aziende agricole per classi di SAU, uno stimatore composto in cui: i) la parte sintetica è costituita dalla stima ottenuta ad un livello territoriale più ampio rispetto a quello provinciale o di zona altimetrica; ii) la parte diretta viene costruita come uno stimatore di espansione; iii) i pesi sono funzione della differenza fra le due stime: all'aumentare della differenza fra la stima diretta e la stima sintetica cresce il peso attribuito alla componente diretta e diminuisce quello attribuito alla parte sintetica.

Le motivazioni alla base di tali scelte sono state le seguenti. In primo luogo si è ritenuto opportuno utilizzare uno stimatore che non facesse ricorso a variabili ausiliarie. L'impiego di uno stimatore che consentisse un aumento di efficienza mediante tali variabili presupponeva infatti la disponibilità, seppur in forma aggregata, dell'informazione relativa alla popolazione delle PA e quindi il ricorso ai dati censuari, che tuttavia risultavano ormai troppo datati (si riferiscono al 1990) per essere ritenuti affidabili anche per l'anno in cui sono state rilevate le informazioni campionarie utilizzate in questo lavoro (1995). In secondo luogo ci è sembrato rilevante selezionare uno stimatore che non presentasse difficoltà di calcolo troppo ampie e fosse applicabile in modo corrente. In altre parole, abbiamo ritenuto importante indicare una tecnica di stima in grado di fornire stime abbastanza soddisfacenti di altre proporzioni di interesse, oltre a quella qui trattata, e la cui procedura di calcolo fosse facilmente applicabile per la stima di qualsiasi proporzione. Anche per quest'ultimo motivo si sono esclusi gli stimatori basati su modelli di regressione che prevedono l'inclusione di variabili esplicative differenti a seconda dell'obiettivo della stima.

2. Gli stimatori per piccole aree

L'esigenza di disporre di risultati non solo per la popolazione nel suo complesso, ma anche per particolari sottogruppi di essa, detti appunto PA, è avvertita ormai sempre più spesso nell'ambito dell'analisi di dati provenienti da indagini campionarie. Tuttavia, per vincoli di costo o poiché tale esigenza viene manifestata solo dopo che il campione è stato progettato, non è solitamente possibile garantire numerosità sufficientemente elevate per i singoli domini. In particolare il numero di osservazioni estratte da ciascuna delle PA è generalmente troppo basso per consentire di ottenere stime caratterizzate da una sufficiente precisione (Sarndal et al., 1992).

E' necessario dunque adottare particolari tecniche di stima che consentano di ottenere stimatori che godono di proprietà accettabili. Uno degli schemi classificatori per gli stimatori per PA proposto più di frequente è quello secondo cui vengono raggruppati nelle due grandi categorie degli stimatori diretti, cioè basati strettamente sul disegno, e indiretti, cioè basati su un modello (Ghosh e Rao, 1994). Nella prima classe troviamo gli stimatori di espansione, di poststratificazione, rapporto e di regressione, mentre gli stimatori indiretti includono gli stimatori sintetici, composti, predittori lineari corretti ottimi, bayesiani empirici e bayesiani gerarchici (Rao e Choudhry, 1995). Nel presente lavoro illustreremo, tuttavia, solo i metodi adatti alla stima di proporzioni (o, equivalentemente, di numerosità) e quindi relativi a variabili categoriche che, come già richiamato, sono stati spesso trascurati rispetto al caso di variabili quantitative. Per una trattazione esaustiva e per eventuali approfondimenti sul problema delle PA e sugli altri tipi di stimatori si vedano Ghosh e Rao (1994), Rao e Choudhry (1995), Sarndal et al. (1992), Singh et al. (1994) e, per il caso italiano, Cocchi (1993), Fabbris et *al.* (1988), Falorsi et al. (1994), Gori e Marchetti (1990), Russo (1995).

Ad una rassegna ragionata della letteratura più recente sull'argomento seguirà la stima del numero di aziende agricole distinte per classi di superficie agricola utilizzata (una ***proxy*** della classe dimensionale) condotta mediante lo stimatore che più proficuamente può essere impiegato in tale ambito. Si noti che, prima di passare all'impiego di uno stimatore per PA, è necessario valutare la precisione ottenibile impiegando le numerosità campionarie di cui si dispone. Potrebbe verificarsi, ad esempio, che in alcuni domini il campione, pur se non disegnato appositamente, presenti una numerosità sufficiente ad ottenere stime abbastanza affidabili con l'impiego delle tecniche di stima classiche.

2.1 *Stimatori* diretti

Lo stimatore diretto del totale relativo all'area i-sima (i=1,. ..,K) è ottenuto sulla base dei valori della variabile obiettivo osservati con riferimento alle sole unità del campione appartenenti alla piccola area i. La forma più semplice di stimatore diretto è lo stimatore di espansione di Horvitz-Thompson (*exp*), che nel caso di stima del numero di unità statistiche che portano una certa caratteristica C, ${}_{C}N_i$, è dato da:

$$ {}_{C}\hat{N}_{i,exp} = \sum_{j \in s_i} w_j d_j \qquad (1) $$

in cui s_i è la parte del campione complessivo che appartiene alla piccola area i, d_j è una variabile dicotomica posta pari ad 1 se l'unità j possiede la caratteristica oggetto di studio e pari a 0 se non la possiede, e w_j è il peso campionario dato dal reciproco della probabilità di inclusione. Tale stimatore è corretto sulla base del disegno ma, come già accennato, può essere caratterizzato da una variabilità molto elevata poiché l'ampiezza di s_i potrebbe essere insufficiente per raggiungere una precisione accettabile dello stimatore. Stimatori diretti più efficienti di quello exp sono gli stimatore di postratificazione, rapporto e regressione. Tali stimatori sfruttano l'informazione relativa a variabili ausiliarie correlate alla variabile obiettivo e disponibile in popolazione per la PA.

Ad esempio, se la numerosità della popolazione N, riferita alla piccola area è nota è possibile costruire lo stimatore di

poststratificazione (*post*) che risulta più stabile dello stimatore di espansione:

$$ {}_C\hat{N}_{i,post} = N_i \frac{\sum_{j \in s_i} w_j d_j}{\sum_{j \in s_i} w_j} = N_i \frac{{}_C\hat{N}_{i,exp}}{\hat{N}_i} \qquad (2) $$

Lo stimatore rapporto è simile allo stimatore di poststratificazione ma impiega una variabile ausiliaria al posto della numerosità di popolazione della piccola area. Se indichiamo con X_i il totale noto di una variabile X, ausiliaria per la variabile oggetto di stima, lo stimatore rapporto *(rapp)* sarà:

$$ {}_C\hat{N}_{i,rapp} = X_i \frac{\sum_{j \in s_i} w_j d_j}{\hat{X}_{i,esp}} = X_i \frac{{}_C\hat{N}_{i,exp}}{\hat{X}_{i,esp}} = X_i \hat{R}_i \qquad (3) $$

Gli stimatori illustrati sono riferiti ad un disegno di campionamento di tipo casuale semplice e vanno quindi adattati a casi di disegno campionario complesso.

Generalmente gli stimatori diretti sono corretti o approssimativamente corretti rispetto al disegno, ma soffrono di una elevata variabilità, tanto maggiore quanto più la dimensione campionaria riferita alla piccola area è ridotta.

2.2 Stimatori sintetici basati implicitamemte su un modello

Gli stimatori sintetici (Purcell e Kish, *1979)* tentano di ridurre la variabilità associata allo stimatore diretto mediante una qualche assunzione di similitudine tra la piccola area e l'area di dimensioni più ampie. In altre parole, l'obiettivo viene perseguito utilizzando l'informazione relativa all'area più ampia che contiene la piccola area e che, quindi, è riferita ad un numero di osservazioni più elevato. Lo

stimatore diretto relativo al campione complessivo (che d'ora in poi chiameremo area), che si assume uguale per ognuno dei sottocampioni (PA) e che gode della proprietà di correttezza, viene "aggiustato" con un fattore "di scala" relativo alla piccola area.

Uno degli stimatori sintetici più semplici è il cosiddetto *stimatore sintetico della media* (*symm*) costruito assumendo che la proporzione relativa alla i-sima piccola area, P_i, sia uguale alla proporzione P relativa al campione complessivo *s*. Tale assunzione conduce al seguente stimatore:

$$ {}_C\hat{N}_{i,symm} = N_i \sum_{j \in s} w_j d_j \Big/ \sum_{j \in s} w_j = N_i \hat{p} \qquad (4) $$

in cui p è lo stimatore della proporzione nell'area di dimensione più grande, mentre N_i è la numerosità della popolazione della piccola area i e costituisce il fattore di aggiustamento.

Seguendo la medesima procedura è possibile costruire stimatori SYN di stratificazione, di poststratificazione o del rapporto, che impiegano variabili ausiliarie nel procedimento di stima. In questi casi la relazione di uguaglianza tra area e piccola area riguarda rispettivamente le medie di strato, le medie di poststrato od il rapporto tra le variabili.

Lo stimatore sintetico di poststratificazione *(syn-pos)* impiega, ad esempio, la stima della proporzione riferita ai poststrati:

$$ {}_C\hat{N}_{i,syn\text{-}post} = \sum_p N_{p,i} \frac{\sum_{j \in s} w_j d_j}{\sum_{j \in s} w_j} = \sum_p N_{p,i} \frac{\sum_{j \in s_p} w_{jp} d_{jp}}{\sum_{j \in s_p} w_{jp}} = \sum_p N_{p,i} \hat{p}_p \qquad (5) $$

mentre lo stimatore sintetico del rapporto *(syn-rapp)* utilizza la stima del rapporto tra la variabile obiettivo ed una variabile ausiliaria, con riferimento all'area di dimensioni più ampie:

$$ {}_C\hat{N}_{i,syn-rapp} = X_i \frac{\sum_{j \in s} w_j d_j}{\hat{X}_{exp}} = X_i \frac{{}_C\hat{N}_{exp}}{\hat{X}_{exp}} = X_i \hat{R} \qquad (6) $$

2.2.1 Lo stimatore sintetico di regressione logistica

Il modello alla base degli stimatori sintetici può essere o meno esplicito e più o meno complesso. Mentre nel precedente paragrafo sono stati illustrati alcuni stimatori basati su un modello implicito (la similitudine tra area e piccola area), nel seguito verrà descritto lo stimatore sintetico di proporzioni basato su un modello di regressione. In tale caso il modello alla base della stima, che risulterà un modello di regressione logistica, è il seguente:

$$ g_i = x_i^T \beta \qquad (7) $$

dove g_i è il parametro di interesse riferito alla piccola area, $x_i = (x_{i1} \ldots x_{is})^T$ è il vettore delle s variabili esplicative, β è il vettore dei coefficienti di regressione.

Si assume che per la stima diretta, $\mathbf{g}_i$, sia $\mathbf{g}_i = \mathbf{g}_i + e_i$ (i=1,...,m) dove gli e_i sono errori campionari con $E(e_i)=0$ e $V(e_i|g_i)=\delta_i^2$. Poiché l'obiettivo è la stima di una proporzione, si definisce $\hat{g}_i$ come una trasformata della stima campionaria della proporzione $\hat{p}_{i,dir}$ (Chand e Alexander, *1995)* al fine di giungere ad un modello di tipo lineare. Scegliendo, ad esempio, la trasformata logit si perviene ad un modello di regressione logistica aggregato in cui le varianze δ_i^2 sono ottenute mediante lo sviluppo in serie di Taylor della funzione $\hat{g}_i$ e richiedono, dunque, il calcolo della varianza dello stima diretta con riferimento al disegno campionario.

Il parametro del modello di regressione viene stimato mediante il metodo GLS:

$$ \tilde{\tilde{\beta}} = \left[\sum_i x_i x_i^T / \delta_i^2 \right]^{-1} \left[\sum_i x_i \hat{g}_i / \delta_i^2 \right] \qquad (8) $$

Lo stimatore sintetico risulterà quindi una previsione desunta dalla stima del modello, $\tilde{\tilde{g}}_i = x_i^T \tilde{\tilde{\beta}}$, da cui si potrà poi trarre lo stimatore della proporzione $p_{.....}$ e quindi quello della numerosità, mediante la trasformazione inversa.

L'errore quadratico medio della stima sintetica è abbastanza complesso da stimare. Comunque, sotto l'ipotesi che la covarianza tra lo stimatore diretto e quello sintetico sia nulla, uno stiinatore asintoticamente corretto dell'errore quadratico medio di questo ultimo è dato da (Ghosh e Rao, *1994):*

$$ mse(\hat{p}_{i,syn}) = (\hat{p}_{i,syn} - \hat{p}_{i,dir})^2 - v(\hat{p}_{i,dir}) \qquad (9) $$

in cui $\hat{p}_{i,dir}$ è un qualsiasi stimatore diretto e $v(\hat{p}_{i,dir})$ è la stima corretta della varianza di tale stimatore. L'ipotesi di indipendenza fra i due stimatori si considera generalmente plausibile poiché lo stimatore diretto è caratterizzato da una maggiore variabilità di quello sintetico (Ghosh e Rao, 1994).

Se il modello assunto è correttamente specificato, gli stimatori basati sul modello godono di proprietà desiderabili. La loro varianza infatti risulta funzione della varianza di stimatori basati sul campione complessivo, ed è decisamente ridotta rispetto a quelli di tipo diretto. Tuttavia, se l'ipotesi di uguaglianza tra area e piccola area con riferimento agli stimatori coinvolti non è verificata, gli stimatori sintetici sono seriamente distorti.

2.3. Gli stinzatori che tentano di ridurre la distorsione dello stimatore sintetico

Come abbiamo visto, gli stimatori sintetici possono produrre stime distorte quando le ipotesi su cui si basano, non verificabili mediante le informazioni campionarie, non sono soddisfatte. Al fine di risolvere tali

problemi vi sono delle tecniche che tentano di trovare un bilanciamento tra distorsione e variabilità associate agli stimatori menzionati. I due stunatori presi in considerazione in questo lavoro perseguono tale obiettivo avvalendosi di due strategie alternative: i) lo stimatore composto che è costituito da una media ponderata dello stimatore diretto e di quello sintetico effettuata con persi opportunamente scelti; ii) lo stimatore basato sul modello ad effetti casuali che introduce nel modello variabili latenti riferite alla piccola area (Ghosh e Rao, 1994).

*2.3.*I . Lo *stimatore composto*

Un metodo naturale per bilanciare la possibile distorsione associata allo stimatore sintetico e l'inefficienza tipica dello stimatore diretto consiste nel calcolare una media ponderata dei due stimatori (Ghosh e Rao, 1994).

Nel caso di una proporzione si avrà:

$$\hat{p}_{i,comp} = w_i \hat{p}_{i,dir} + (1 - w_i)\hat{p}_{i,syn} \qquad (10)$$

dove $\hat{p}_{i,dir}$ denota la stima diretta, $\hat{p}_{i,syn}$ quella sintetica e w_i è un peso scelto in modo appropriato e tale da essere compreso fra 0 e 1. La stima sintetica può essere desunta da uno stimatore basato sul modello che sfrutta informazioni ausiliarie, oppure può essere una stima corretta per un'area più grande che contiene la piccola area (ad esempio regione-provincia). Il problema principale nell'impiego dello stimatore composto è quello di scegliere i pesi w_i in modo appropriato.

Un insieme ottimo di pesi può essere ottenuto minimizzando l'errore quadratico medio di $\hat{p}_{i,comp}$ rispetto a w_i, assumendo $cov(\hat{p}_{i,dir}, \hat{p}_{i,syn}) = 0$:

$$w_i^* = MSE(\hat{p}_{i,syn}) / [MSE(\hat{p}_{i,syn}) + V(\hat{p}_{i,dir})] \qquad (11)$$

Tale peso può essere stimato sostituendo le stime di $MSE(\hat{p}_{i,syn})$ e di $V(\hat{p}_{i,dir})$, ma il peso che ne risulta potrebbe essere molto instabile. In questi casi conviene ricorrere ad altri criteri per il calcolo dei pesi, basati sulla distanza fra le due stime considerate (Thomsen e Holmøy, 1996):

$$w_i = (\hat{p}_{i,dir} - \hat{p}_{i,syn})^2 / [\hat{p}_{i,dir}(1 - \hat{p}_{i,dir})/n_i + (\hat{p}_{i,dir} - \hat{p}_{i,syn})^2] \qquad (12)$$

Da uno studio condotto sulla base di simulazioni (Thomsen e Holmøy, 1996) l'impiego dello stimatore composto con pesi così calcolati consente un sostanziale guadagno di efficienza rispetto all'uso dello stimatore diretto. Tuttavia, tale miglioramento nelle stime si riduce quando la proporzione da stimare nella piccola area è molto bassa, ossia si avvicina a zero.

Il passaggio alla stima delle numerosità entro le classi può avvenire moltiplicando ciascuna proporzione per il totale N, delle unità nella popolazione della i-esima area, ottenuto come somma dei pesi relativi alle osservazioni campionarie di tale area.

Considerando fissi i pesi w_i, l'errore quadratico medio dello stimatore composto ottenuto secondo la (10) è dato da:

$$MSE(\hat{p}_{i,comp}) = w_i^2 V(\hat{p}_{i,dir}) + (1 - w_i)^2 MSE(\hat{p}_{i,syn}) \qquad (13)$$

e può essere a sua volta stimato sostituendo in questa espressione una stima corretta della varianza della stima diretta $v(\hat{p}_{i,dir})$ ed una stima approssimativamente corretta dell'errore quadratico medio di $\hat{p}_{i,syn}$. Quest'ultima può essere ottenuta, come già visto nel precedente paragrafo, tramite la (9).

2.3.2. Lo *stimatore desunto* dal *modello* ad *effetti* casuali

Alla base dello stimatore desunto dal modello ad effetti casuali vi è ancora

un modello di regressione che esprime il parametro di interesse g_i come una combinazione lineare di effetti fissi ed una variabile latente, l'effetto casuale:

$$g_i = x_i^T \beta + t_i, \qquad i = 1,\ldots,m \tag{14}$$

in cui tutti i simboli assumono lo stesso significato riportato nel par. 2.2 ed i t_i sono gli effetti casuali, ossia variabili casuali IID con $E(t_i) = 0$ e $V(t_i) = \tau^2$. Tali effetti specifici non osservabili rappresentano le caratteristiche delle PA non incluse nella regressione e vengono introdotti, come vedremo meglio in seguito, allo scopo di giungere ad un appropriato winproniesso fra la stima diretta, corretta e inefficiente, e la stima desunta dal modello che mette in relazione le informazioni sulle aree (MacGibbon e Tomberlin, 1989). Come nel modello precedentemente illustrato, $g_i = g_i + e_i$ $(i=1,\ldots,m)$ dove gli e_i sono errori campionari e $\hat{g}_i$ è anwra una trasformata logit della stima campionaria della proporzione p, (Chand e Alexander, 1995).

Sotto il modello definito dalla (14) lo stimatore BLUP (*best linear unbiased predictor*) della trasformata g_i della proporzione si ottiene come media ponderata della stima diretta g_i e della stima sintetica ottenuta dal modello $x_i^T \hat{\beta}$:

$$\tilde{g}_i = \gamma_i \hat{g}_i + (1-\gamma_i) x_i^T \hat{\beta} \qquad i = 1,\ldots,m \tag{15}$$

dove β è lo stimatore BLUP di β ottenuto impiegando il metodo del minimo χ^2, che equivale all'applicazione dei minimi quadrati generalizzati:

$$\hat{\beta} = \left[\sum_i x_i x_i^T / \left(\tau^2 + \delta_i^2\right)\right]^{-1} \left[\sum_i x_i \hat{g}_i / \left(\tau^2 + \delta_i^2\right)\right], \tag{16}$$

ed il peso γ_i rappresenta la quota di incertezza dovuta al modello rispetto alla varianza totale:

$$\gamma_i = \tau^2 / \left(\tau^2 + \delta_i^2\right) \tag{17}$$

Lo stimatore BLUP è valido per qualsiasi disegno campionario, poiché in esso si modellano solo i g_i e non le singole unità osservate: la complessità del disegno viene incorporata mediante gli stimatori campionari della proporzione e della varianza. Inoltre, lo stimatore $\tilde{g}_i$ è consistente rispetto al disegno, perché $\gamma_i \to 0$ se le varianze $\delta_i^2 \to 0$.

Sostituendo τ^2 con una stima asintoticamente consistente si ottiene uno stimatore in due stadi, $\tilde{g}_i$, noto con il nome di stimatore BLUP empirico o EBLUP. Uno stimatore semplice e con-etto per τ^2, per il quale inoltre non occorrono ipotesi distributive, è quello ottenuto con il metodo dei momenti (Ghosh e Rao, 1994):

$$\hat{\tau}^2 = (m-s)^{-1}\left[\sum_i \left(g_i - x_i^T b\right)^2 - \sum_i \delta_i^2 + \sum_i \left\{\delta_i^2 x_i^T \left(\sum_i x_i x_i^T\right)^{-1} x_i\right\}\right] \tag{18}$$

in cui b è lo stimatore dei minimi quadrati ordinari di β. Per le proprietà di tale stimatore si veda (Harville, 1991). Se il modello esplicitato risulta adeguato, i residui standardizzati $r_i = \left(g_i - x_i^T \beta\right) / \sqrt{\left(\hat{\tau}^2 + \delta_i^2\right)}$ risultano approssimativamente distribuiti come $N(0,1)$.

Uno stimatore approssimativamente corretto dell'errore quadratico medio di $\hat{\tilde{g}}_i$ è costituito da tre parti, in cui la prima parte è dovuta alla quota di incertezza dovuta al modello rispetto alla varianza totale, la seconda è dovuta alla stima dei parametri β e la terza è dovuta alla stima della varianza degli effetti casuali (Ghosh e Rao, 1994).

Mediante la trasformazione logit inversa applicata a $\tilde{g}_i$ si ottiene la stima della proporzione $\tilde{p}_i = e^{\tilde{g}_i} / \left(1 + e^{\tilde{g}_i}\right)$. L'errore quadratico medio di tale stimatore è ottenuto nuovamente mediante lo sviluppo in serie di Taylor della funzione $\tilde{p}_i$.

Come è già stato evidenziato, lo stimatore (15) può essere visto come uno stimatore composto ottenuto da una media ponderata di uno stimatore diretto e di uno sintetico basato su un modello di regressione ad effetti casuali, in cui i pesi incorporano la variabilità dell'effetto casuale stesso.

3. La stima della precisione della soluzione *standard*

Come specificato nell'introduzione, è stata in primo luogo valutata la precisione delle stime ottenute con la procedura usuale di impiego dello stimatore diretto a livello di piccola area utilizzando il coefficiente di variazione, che fornisce una valutazione relativa della precisione delle stime. Le stime sono qui calcolate con riferimento alla provincia o alla zona altimetrica, ed al disegno campionario impiegato dall'indagine.

Nel caso dell'indagine ISTAT sulla struttura delle aziende agricole il disegno campionario è stratificato e gli strati intersecano i domini (province e zone altimetriche). Ciò rende più complessa la stima della varianza della stima diretta, che può essere ottenuta come:

$$v(\hat{p}_{i,dir}) = \frac{1}{\hat{N}_i^2} \sum_h N_h^2 \left(\frac{1-f_h}{n_h} \right) \left[\frac{n_{hi} p_{hi}(1-p_{hi}) + n_{hi}\left(1-\frac{n_{hi}}{n_h}\right)(p_{hi}-\hat{p}_{i,dir})^2}{n_h - 1} \right] \tag{19}$$

dove con h si denotano gli strati, con p_{hi} le proporzioni nelle celle hi, con f_h la frazione di campionamento nello strato h. Si ricorda che $\hat{N}_i = \sum_h N_h (n_{hi}/n_h)$, in quanto le numerosità di popolazione nelle celle definite dallo strato e dalla piccola area, N_{hi}, non sono immediatamente disponibili, ed occorre stimarle sotto l'ipotesi $N_{hi}/N_h = n_{hi}/n_h$, così come suggerito da Särndal *et* ***al.*** (1992).

La stima del coefficiente di variazione (CV) è quindi data da:

$$cv(\hat{p}_{i,dir}) = \frac{\sqrt{v(\hat{p}_{i,dir})}}{\hat{p}_{i,dir}} \cdot 100 \tag{20}$$

Si noti che nella espressione (19) compaiono varianze riferite alle celle formate dall'incrocio dello strato con la piccola area. Il numero

delle celle in cui la numerosità n_{hi} risulta pari ad 1, benché lo strato non sia auto-rapresentativo ossia composto da quell'unica unità, è abbastanza elevato. Pertanto, per calcolare le varianze nell'incrocio fra strato e piccola area $(p_{hi}(1-p_{hi}))$ è stato necessario effettuare una operazione di collassamento di tali celle in modo che contengano almeno due unità.

I coefficienti di variazione per provincia e per zona altimetrica sono stati calcolati con riferimento alla variabile "quota di aziende per classi di SAU", per l'importanza che la misura dell'utilizzazione dei terreni da parte delle aziende assume negli studi sulle aziende agricole. La SAU viene espressa in ettari ed è considerata suddivisa in sette classi.

Nella Tabella 1 sono riportate le stime dei CV riferiti al numero (od alla quota) delle aziende per classi di SAU. Dal loro esame emerge la grande influenza delle numerosità campionarie sui risultati. In particolare, come del resto c'era da attendersi, il CV risulta accettabile per quelle classi in cui il numero di aziende campionate è abbastanza elevato. Ad esempio, per le classi di SAU di 5-10 e 10-50 ettari i CV per provincia risultano accettabili. Se, invece, si considerano le prime classi di SAU, caratterizzate da numerosità molto basse, i CV risultano decisamente elevati.

Tabella 1: Coefficienti di variazione per provincia del numero di aziende per classi di Superficie Agricola Utilizzata (SAU) e relativa numerosità campionaria

	Classi di SAU (in ettari)							
	0		(0,1]		(1,5]		(5,10]	
PROVINCIA	CV	n	CV	n	CV	n	CV	n
Piacenza	42,1	2	35,1	13	14,2	42	11,9	57
Parma	09,6	39	41,6	10	11,7	75	10,5	98
R. Emilia	13,0	27	39,1	11	10,2	110	11,8	118
Modena	15,1	7	40,2	11	7,9	117	9,6	128
Bologna	42,1	3	38,5	14	8,4	105	8,3	116
Ferrara	66,0	3	40,0	11	16,1	34	9,2	74
Ravenna	191,5	1	18,1	16	17,8	82	8,5	96
Forlì	59,1	3	21,3	18	10,5	102	13,4	72
Rimini	--	0	13.4	17	30.9	49	31,2	22

	Classi di SAU (in ettari)					
	(10,50]		(50,100]		oltre 100	
PROVINCIA	CV	n	CV	n	CV	n
Piacenza	09,6	259	11,7	89	12,8	63
Parma	07,4	338	23,4	62	29,9	39
R. Emilia	08,6	320	25,2	47	13,6	62
Modena	8,9	322	12,4	86	20,3	55
Bologna	8,7	229	14,6	60	11,2	90
Ferrara	8,3	225	14,4	54	16,3	76
Ravenna	17,7	143	26,7	19	18,0	49
Forlì	19,7	141	13,2	52	21,3	47
Rimini	116,7	28	79,8	11	117,0	4

Per la classificazione secondo la zona altimetrica (Tabella 2) si ottengono risultati simili. In particolare, per i casi in cui all'incrocio tra zona altimetrica e classe di SAU vi è un numero rilevante di aziende, i CV risultano soddisfacenti (è il caso, ad esempio, della pianura per le classi di SAU 1-5, 5-10 e 10-50).

Tabella 2: Coefficienti di variazione per zona altinietrica del numero di aziende per classi di Superficie Agricola Utilizzata (SAU) e relativa numerosità canipionaria

	Classi di SAU (in ettari)							
	0		(0,1]		(1,5]		(5,10]	
ZONA ALTIM.	CV	n	CV	n	CV	n	CV	n
Montagna	10,1	32	33,4	12	6,6	106	10,2	96
Collina	15,1	24	13,4	35	10,4	158	6,2	199
Pianura	12,5	58	11,8	74	5,6	452	4,8	486

	Classi di SAU (in ettari)					
	(10,50]		(50,100]		oltre 100	
ZONA ALTIM.	CV	n	CV	n	CV	n
Montagna	9,5	204	21,4	55	15,1	42
Collina	9,4	535	11,5	132	15,1	107
Pianura	4,8	1.266	6,8	293	9,4	336

Pertanto, si può asserire che dall'attuale campione o, più in generale, mantenendo le attuali numerosità e disegno campionari, solo in alcuni casi è possibile ottenere stime caratterizzate da una precisione accettabile. In sostanza cioè, i risultati ottenuti non ci consentono di fornire regole di comportamento generali utili a stabilire se una data stima può essere ottenuta con una certa affidabilità per ognuna delle province o per ogni zona altimetrica. Infatti, tale conclusione può essere tratta solo per alcune PA. Per questo motivo si è ricorsi ad uno stimatore specifico per PA, lo stimatore composto, che consente in molti casi di ottenere un guadagno di efficienza.

4. La stima delle proporzioni e delle numerosità mediante lo stimatore composto

L'obiettivo di valutare la distribuzione delle aziende agricole delle provincie e delle zone altimetriche dell'Emilia Roinagna rispetto alla SAU suddivisa in classi è stato perseguito, come si è detto, utilizzando il campione regionale del 1995, in cui la stratificazione è stata effettuata in base alle variabili "classe di SAU", "classe di capi bovini", "classe di capi suini" e "classe di capi ovini e caprini".

Si è proceduto alla stima delle proporzioni, e quindi delle nuinerosità, delle aziende agricole in ciascuna classe di SAU della provincia di volta in volta considerata impiegando lo stimatore composto riportato nella (10). Tale stimatore consente, come si è visto, di giungere ad un compromesso fra l'inefficienza dello stimatore diretto e la distorsione dello stimatore sintetico, senza comportare la complessità di calcolo che caratterizza invece gli stimatori per proporzioni basati su modelli di regressione logistica. Inoltre, ha permesso in un'altra esperienza di ottenere soluzioni soddisfacenti quanto i metodi esplicitamente basati su un modello (Ferrante e Pacei, 1998).

Nell'impiego di uno stimatore composto occorre effettuare due scelte: quella della componente sintetica e quella dei pesi da attribuire alle due componenti. Come parte sintetica può essere proposta la quantità oggetto di studio relativa alla popolazione e nota, ad esempio, dal Censimento, oppure una stima più attendibile di tale quantità proveniente da altre fonti. Nel nostro caso, non vi sono altre fonti in grado di fornire stime più attendibili delle proporzioni di interesse. A sua volta il Censimento, che potrebbe essere visto come la fonte più ovvia di informazione per la correzione degli stimatori, rappresenta il termine di paragone con cui si intendono confrontare i risultati finali per valutare eventuali cambiamenti intervenuti negli ultimi anni nel "panorama" agricolo della regione. Si è quindi tentato di produrre stime che non dipendessero fortemente dai suoi risultati e si è deciso di considerare come componente sintetica la stima dell'analoga proporzione nella regione ottenuta dalla stessa indagine ISTAT sulle aziende agricole, ossia la stima tramite gli stessi dati della quantità di

interesse riferita ad un area più ampia che contiene la piccola area, come talvolta si effettua nell'ambito della stima per domini (Thomsen e Holmøy, 1996). Tale stima della proporzione nella regione può ritenersi senz'altro soddisfacente data l'ampiezza del campione regionale (n=4.702). Tuttavia, va sottolineato che si tratta di una scelta subottimale, dovuta alla mancanza di informazioni ausiliarie sulla popolazione delle PA che siano legate alla proporzione di interesse e che siano state raccolte recentemente.

Per quanto riguarda la scelta dei pesi w_i, quelli calcolati in modo da minimizzare l'errore quadratico medio di $\hat{p}_{i,comp}$ conducevano a risultati abbastanza instabili, perché alcune delle stime ottenute con la (14) assumevano valori negativi. Pertanto, si è deciso di utilizzare per il loro calcolo l'espressione (12), in cui i pesi dipendono dalla differenza fra stima diretta e stima sintetica: al suo aumentare cresce il peso attribuito alla componente diretta e diminuisce quello attribuito alla parte sintetica.

La metodologia proposta è stata poi impiegata per ottenere delle stime delle analoghe proporzioni per zona altimetrica anziché per provincia. Il passaggio alle stime delle numerosità è stato effettuato moltiplicando le stime composte delle proporzioni per le stime delle numerosità della popolazione nelle province (o nelle zone altimetriche) desunte dai pesi campionari.

L'errore quadratico medio dello stimatore composto viene stimato mediante l'espressione (13). Da tali stime si perviene poi a quelle dei coefficienti di variazione.

Le Tabelle 3 e 4 riportate di seguito contengono i risultati ottenuti sulla base delle proposte effettuate, ossia le stime delle proporzioni, delle numerosità e dei coefficienti di variazione, e le numerosità campionarie con cui sono state ottenute. Tali risultati mettono in evidenza come le distribuzioni della variabile osservata si differenzino per provincia e per altimetria. Tuttavia, nell'effettuare i confronti, occorre tenere presente che, soprattutto per quanto riguarda l'analisi per provincia, in corrispondenza di alcune modalità della variabile considerata il numero delle osservazioni disponibili era ancora troppo basso per consentire di ottenere stime soddisfacenti delle proporzioni e

delle numerosità anche con lo stimatore composto che è stato impiegato. Ci si riferisce in particolare alle prime due classi della variabile SAU ("Senza superficie agricola utilizzata" e "Meno di un ettaro"), ed alla provincia di Rimini. Infatti, in questi casi le stime dei coefficienti di variazione dei $\hat{p}_{i,comp}$ sono più elevate che negli altri e si può notare un'elevata discrepanza fra questi risultati e quelli del Censimento 1990.

La validità, in linea di principio, della proposta di stimatori composti specifici per piccoli domini di studio può essere apprezzata notando che, le stime ottenute per i coefficienti di variazione dello stimatore composto impiegato sono sempre inferiori a quelle desunte per lo stimatore diretto. Si ha pertanto un guadagno di efficienza con lo stimatore composto che, come si può notare, non è sempre della stessa entità, bensì varia in funzione sia della numerosità campionaria che del valore della proporzione. Infatti, lo stimatore composto appare meno efficiente soprattutto per le stime particolarmente basse delle proporzioni, ossia lontane dal valore medio della distribuzione, come accade per le proporzioni di aziende che appartengono alle classi di SAU "0 ettari" o "100 ettari e oltre" di tutte le province e di tutte le zone altimetriche. Questo limite dello stimatore composto è stato riscontrato anche in altri lavori (Thomsen e Holmøy, 1996; Ferrante e Pacei, 1998).

Le stime più attendibili sono quelle ottenute per la classe di SAU "10-50 ettari" (i coefficienti di variazione si aggirano sui 3-4 punti percentuali) in corrispondenza della quale è stato rilevato un numero di aziende sempre nettamente superiore a quello osservato per le altre classi, tranne che nella provincia di Rimini.

Analizzando le stime delle distribuzioni delle aziende agricole per classi di SAU desunte per le varie province e zone altimetriche, si può osservare che la percentuale di aziende senza superficie agricola utilizzata è sempre molto bassa e raggiunge il massimo per Piacenza con il 4%. Per quanto riguarda le classi di SAU più elevate ("10-50 ettari", "50-100 ettari" e "100 ettari e oltre"), si osserva che la proporzione di aziende in esse si riduce passando dalle province occidentali (Piacenza, Parma, Reggio Emilia, Modena, Bologna e Ferrara) a quelle orientali della regione (Ravenna, Forlì e Rimini).

Poiché per la provincia di Rimini non è stata selezionata neppure una azienda con SAU pari a zero, le prime due classi di SAU sono state riunite in un'unica classe. La particolarià appena accennata, unitamente al fatto che la variabile "provincia" non entra nel disegno, ha dato luogo anche al risultato seguente. Confrontando le stime ottenute per questa provincia con quelle desunte per le altre, la quota di aziende con SAU compresa fra 0 ed 1 ettaro appare particolarmente alta a Rimini, mentre quella delle aziende con SAU compresa fra 5 e 10 ettari risulta particolarmente bassa in questa provincia.

La maggior parte delle aziende si colloca nella classe di SAU "da 1 a 5 ettari" in tutte le province (dal 31% al 47%) tranne in quella di Rimini. Per tale provincia, come è già stato detto, i risultati sono pero meno attendibili. In particolare le percentuali di aziende nella classe di SAU "da 1 a 5 ettari" sono elevate nelle provincie di Forlì, Modena e Bologna, dove corrispondono a circa 7.500-8.500 unità.

Proseguendo con le distribuzioni per classi di SAU all'interno di ciascuna zona altimetrica, si riscontrano differenze rilevanti fra le proporzioni solamente per le classi di SAU "da 0 ad 1 ettaro" e "da 1 a 5 ettari". Infatti, in montagna la proporzione di aziende che appartiene alla classe di SAU "da 0 ad 1 ettaro" risulta più bassa che in collina ed in pianura, mentre per la quota di aziende con SAU compresa fra 1 e 5 ettari si verifica il contrario: risulta molto più elevata in montagna che in collina ed in pianura.

Le stime ottenute per la collina e la pianura risultano in generale accettabili tranne che in corrispondenza della classe "Senza superficie agricola utilizzata" in cui i coefficienti di variazione superano il 10%. Invece, per quanto riguarda la montagna, solo le numerosità campionarie raggiunte per le classi intermedie ("da 1 a 5 ettari", "da 5 a 10 ettari" e "da 10 a 50 ettari") consentono di ottenere stime caratterizzate da una bassa variabilità, ossia stime i cui coefficienti di variazione sono inferiori al 10%.

Tabella 3. Distribuzione delle aziende agricole delle provincie dell'Emilia Romagna per classi di SAU.

Classi di SAU in ettari	$P_{i,comp}$	$N_{i,comp}$	$cv_{i,comp}$	n_i
PIACENZA				
0	0,040	572	36,27	2
(0-1)	0,146	2080	23,85	13
[1-5)	0,359	5087	9,00	42
[5-10)	0,167	2371	8,41	57
[10-50)	0,260	3693	3,61	259
[50-100)	0,025	366	7,84	89
100 e oltre	0,007	99	10,99	63
PARMA				
0	0,006	110	7,75	39
(0-1)	0,116	1992	23,58	10
[1-5)	0,403	6883	8,81	75
[5-10)	0,177	3030	8,04	98
[10-50)	0,273	4661	3,23	338
[50-100)	0,014	248	10,44	62
100 e oltre	0,004	76	22,21	39
REGGIO EMILIA				
0	0,006	68	7,36	27
(0-1)	0,097	1036	23,56	11
[1-5)	0,403	4286	7,80	110
[5-10)	0,201	2146	7,97	118
[10-50)	0,250	2664	3,12	320
[50-100)	0,015	168	10,29	47
100 e oltre	0,006	70	19,45	62
MODENA				
0	0,005	87	17,35	7
(0-1)	0,087	1455	23,46	11
[1-5)	0,473	7882	5,59	117
[5-10)	0,205	3420	4,94	128
[10-50)	0,201	3351	1,94	322
[50-100)	0,018	314	8,86	86
100 e oltre	0,005	96	15,98	55
BOLOGNA				
0	0,001	24	33,28	3
(0-1)	0,093	1799	23,77	14
[1-5)	0,444	8507	6,82	105
[5-10)	0,216	4146	6,31	116
[10-50)	0,203	3890	3,35	229
[50-100)	0,016	314	9,72	60
100 e oltre	0,007	147	9,67	90

Tabella 3 (continua): Distribuzione delle aziende agricole delle provincie dell'Emilia Romagna per classi di SAU.

Classi di SAU in ettari	$p_{i,comp}$	$N_{i,comp}$	$cv_{i,comp}$	n_i
FERRARA				
0	0,001	12	55,61	3
(0-1)	0,133	1832	24,11	11
[1-5)	0,324	4472	9,57	34
[5-10)	0,193	2664	7,98	74
[10-50)	0,311	4286	3,73	225
[50-100)	0,018	258	9,42	54
100 e oltre	0,008	110	11,90	76
RAVENNA				
0	0,0001	1	119,00	1
(0-1)	0,305	4789	14,05	16
[1-5)	0,316	4962	8,91	82
[5-10)	0,203	3193	5,71	96
[10-50)	0,157	2465	4,22	143
[50-100)	0,011	181	9,43	19
100 e oltre	0,006	95	9,39	49
FORLI'				
0	0,001	19	37,27	3
(0-1)	0,237	3820	13,78	18
[1-5)	0,462	7450	5,34	102
[5-10)	0,143	2315	6,97	72
[10-50)	0,128	2071	4,33	141
[50-100)	0,016	264	8,66	52
100 e oltre	0,005	91	11,40	47
RIMINI				
[0-1)	0,641	8208	10,20	17
[1-5)	0,276	3542	26,99	49
[5-10)	0,053	683	28,21	22
[10-50)	0,037	480	98,37	28
[50-100)	0,001	25	59,80	11
100 e oltre	0,001	11	95,06	4

* Per questa provincia sono state riunite in una sola classe le prima due modalità della SAU poiché non vi erano aziende-campione senza superficie agricola utilizzata

Tabella 4. Distribuzione delle aziende agricole delle zone altimetriche dell'Emilia Romagna per classi di SAU.

Classi di SAU in ettari	$p_{i,comp}$	$N_{i,comp}$	$cv_{i,comp}$	n_i
MONTAGNA				
0	0,006	121	8,74	32
(0-1)	0,095	1843	20,58	12
[1-5)	0,523	10137	4,27	106
15-10)	0,182	3538	9,49	96
[10-50)	0,171	3324	6,92	204
[50-100)	0,014	276	15,56	55
100 e oltre	0,005	115	10,80	42
COLLINA				
0	0,002	94	14,02	24
(0-1)	0,275	9935	9,30	35
[1-5)	0,319	11500	7,33	158
15-10)	0 184	6655	6,34	199
[10-50)	0,198	7143	7,44	535
[50-100)	0,016	589	6,51	132
100 e oltre	0,005	191	9,10	107
PIANURA				
0	0,007	634	10,79	58
(0-1)	0,192	15528	7,78	74
[1-5)	0,391	31523	5,77	452
15-10)	0,172	13897	5,86	486
[10-50)	0,213	17206	2,93	1 266
[SO-100)	0,016	1333	5,99	293
100 e oltre	0,006	524	6,65	336

Il confronto per provincia fra i risultati ottenuti applicando lo stimatore composto ai dati dell'indagine campionaria ISTAT del 1995 con quelli desunti dall'ultimo Censimento dell'agricoltura (1990), è stato realizzato considerando insieme le province di Forlì e di Rimini, poiché quest'ultima non era ancora stata istituita all'epoca del Censimento. Il numero complessivo delle aziende agricole risulta diminuito nel tempo di circa 14.000 unità. Tale riduzione sembra essersi verificata in tutte le zone altimetriche (soprattutto in "montagna") ed in tutte le province ad esclusione di quelle di Ravenna e di Forlì-Rimini, in cui il numero delle aziende agricole stimate dai dati campionari supera quello risultante dal Censimento.

Nelle Tabelle 5 e 6 sono riportati i risultati ottenuti in questo lavoro con lo stimatore per PA utilizzato ed i corrispondenti risultati del

Censimento, rispettivamente per provincia e per zona altimetrica.

Dall'analisi della Tabella 5 emerge che, stando ai risultati ottenuti con lo stimatore composto ($\hat{p}_{i,comp}$), è diminuita rispetto al Censimento (P_i) la quota di aziende con SAU compresa nella classe "fino ad un ettaro" nelle province situate a ovest della regione, come Reggio-Emilia (-12 punti percentuali), Modena (-9 punti percentuali) e Bologna (-6 punti percentuali). In queste stesse province è aumentata, invece, la quota di aziende di dimensione più elevata, in particolare con SAU compresa fra 10 e 50 ettari, con un incremento massimo di 9 punti percentuali nella provincia di Reggio-Emilia. Per quanto riguarda, invece, le province di Ravenna e di Forlì-Rimini, si verifica il contrario: sono aumentate in proporzione le aziende appartenenti alla classe di SAU "fino ad un ettaro" di 12-15 punti percentuali, mentre sono diminuite quelle appartenenti alla classe di SAU appena superiore, "da 1 a 5 ettari", di circa 8 punti percentuali in entrambe le province. Queste differenti variazioni rispetto al Censimento, emerse per le aziende agricole delle province occidentali e delle province orientali della regione, possono essere spiegate anche dal diverso orientamento tecnico-economico che in esse prevale. Infatti, com'è noto, nelle province occidentali sono molto più diffuse le aziende con allevamenti, mentre in quelle orientali prevalgono le aziende produttrici di frutta.

Infine, risultano lievemente diminuite, tranne che a Piacenza, le aziende senza SAU ma, come si è visto, le stime composte ottenute per questa classe sono quelle meno attendibili. Molto vicine fra loro sono invece le percentuali di aziende nelle due classi di SAU più elevate, "da 50 a 100 ettari" ed "oltre 100 ettari", e nella classe di SAU "da 5 a 10 ettari", alla quale corrispondono sempre le stime più attendibili.

Osservando la Tabella 6, invece, si nota che le distribuzioni percentuali delle aziende per classi di SAU risultanti dal Censimento e da questo lavoro siano molto simili per quanto riguarda la pianura. In collina sembrerebbe aumentata rispetto al Censimento la quota di aziende che appartengono classe di SAU [0-1) e diminuita quella delle aziende appartenenti alla classe [1-5). Il contrario sembra essersi verificato per la montagna.

Tabella 5. Distribuzione percentuale delle aziende agricole delle provincie dell'Emilia Romagna per classi di SAU ottenuta con lo stimatore composto ($\hat{p}_{i,comp}$) e desunta dal Censimento dell'agricoltura del 1990 (P_i).

Classi di SAU in ettari	$P_{i,comp} \cdot 100$	$P_i \cdot 100$
PIACENZA		
0	4,0	1,6
(0-1)	14,6	18,5
[1-5)	35,9	36,0
[5-10)	16.7	18.3
[10-50)	26,0	22,9
[50-100)	2,5	2,0
100 e oltre	0,7	0,7
PARMA		
0	0.6	2,4
(0-1)	11,6	11,6
[1-5)	40,3	37,9
[5-10)	17,7	21,1
[10-50)	27,3	25.3
[SO-100)	1,4	1.3
100 e oltre	0,4	0,4
REGGIO EMILIA		
0	0,6	2,3
(0-1)	9,7	21,4
[1-5)	40,3	41,2
[5-10)	20,1	17,0
[10-50)	25,0	16,7
[50-100)	1,5	1,0
100 e oltre	0,6	0,4
MODENA		
0	0,5	4,0
(0-1)	8,7	15,9
[1-5)	47,3	41,9
[5-10)	20,5	18,8
[10-50)	20,1	17,8
[50-100)	1,8	1,2
100 e oltre	0.5	0,4
BOLOGNA		
0	0,1	1,9
(0-1)	9,3	14,6
[1-5)	44,4	41,2
[5-10)	21,6	20,8
[10-50)	20,3	19,6
[50-100)	1,6	1,3
100 e oltre	0,7	0,6

Tabella 5 (continua). Distribuzione percentuale delle aziende agricole delle provincie dell'Emilia Romagna per classi di SAU ottenuta con lo stimatore coniposto ($\hat{p}_{i,comp}$) e desunta dal Censimento dell'agricoltura del 1990 (P_i)

Classi di SAU in ettari	$p_{i,comp} \cdot 100$	$P_i \cdot 100$
FERRARA		
0	0,1	0,8
(0-1)	13,3	17,2
[1-5)	32,4	27,1
[5-10)	19,3	22,4
[10-50)	31,1	29,8
[50-100)	1,8	1,8
100 e oltre	0,8	0,9
RAVENNA		
0	0,01	1,0
(0-1)	30,5	15,6
[1-5)	31,6	39,5
[5-10)	20,3	23,8
[10-50)	15,7	18,6
[50-100)	1,1	0,9
100 e oltre	0,6	0,6
FORLI' e RIMINI		
0	0,07	1,0
(0-1)	41,5	29,4
[1-5)	37,9	45,3
[5-10)	10,3	13,0
[10-50)	8,8	10,0
[50-100)	1,0	0,9
100 e oltre	0,4	0,4

Tabella 6. Distribuzione percentuale delle aziende agricole delle zone altimetriche dell'Emilia Romagna per classi di SAU ottenuta con lo stinzatore composto ($\hat{p}_{i,comp}$) e desunta dal Censimento dell'agricoltura del 1990 (P_i).

Classi di SAU in ettari	$p_{i,comp} \cdot 100$	$P_i \cdot 100$
MONTAGNA		
[0-1)	10.1	15.3
[1-5)	52.3	46,2
[5-10)	18.2	20.3
[10-50)	17,1	16,9
[50-100)	1,4	0.9
100 e oltre	0,5	0,4
COLLINA		
[0-1)	27,7	20,3
[1-5)	31,9	39,5
[5-10)	18,4	18,6
[10-50)	19,8	19,8
[50-100)	1,6	1,4
100 e oltre	0,5	0,5
PIANURA		
[0-1)	19.9	19.3
[1-5)	39,1	39,0
[5-10)	17,2	19,4
[10-50)	21,3	20,4
[50-100)	1,6	1,4
100 e oltre	0,6	0,6

5. Conclusioni

In questo lavoro si è affrontato il problema della stima di proporzioni e di numerosità all'intemo di PA, sulla base delle informazioni rilevate in Emilia Romagna nell'ambito dell'indagine ISTAT sulle aziende agricole condotta nel 1995. I domini di interesse sono stati le province e le zone altimetriche ed, in particolare, ci si è concentrati sulla stima delle proporzioni e delle numerosità delle aziende agricole per classi di SAU in tali domini. La SAU si ritiene, infatti, insieme alla OTE (orientamento tecnico economico) ed alla UDE (unità di dimensione economica), una delle caratteristiche più rilevanti delle aziende agricole poiché ne rappresenta in qualche modo l'ampiezza. Per questo motivo l'interesse non è stato rivolto alla stima della SAU totale o media, bensì alla stima della distribuzione delle aziende per classi di SAU.

In una prima fase della ricerca si è proceduto alla valutazione della precisione delle stime tradizionali ottenibili dal campione realizzato per le PA considerate. Poiché da tale analisi è emerso che l'impiego dello stimatore diretto di espansione conduce a stime delle proporzioni di interesse particolarmente inefficienti per la maggior parte dei domini, si è deciso di ricorrere ad una tecnica di stima specifica per PA. Fra i metodi proposti in letteratura è stato scelto uno stimatore composto, che costituisce una media ponderata dello stimatore diretto e di uno stimatore sintetico, in cui i pesi sono funzioni della differenza fra le due stime.

La scelta di questo stimatore è giustificata sia dal fatto che ha già consentito di ottenere risultati soddisfacenti quanto i metodi basati esplicitamente sul modello in altri lavori, sia sulla sua semplicità di calcolo. L'obiettivo in questa fase del lavoro è stato, infatti, quello di individuare un metodo facilmente applicabile alla stima di una qualsiasi proporzione.

I risultati ottenuti mettono in evidenza il considerevole guadagno di efficienza che lo stimatore composto impiegato permette di ottenere rispetto allo stimatore diretto, soprattutto in corrispondenza di alcune PA e di alcune proporzioni.

A questo proposito vanno effettuate due osservazioni. Innanzi tutto, il guadagno di efficienza non si verifica in uguale misura in tutte le aree,

bensì dipende dalla numerosità campionaria del dominio e dalla grandezza della stima. Lo stimatore composto appare infatti meno efficiente per la stima delle proporzioni più basse, ossia più vicine allo zero. Inoltre, va osservato che in corrispondenza di alcune modalità della variabile considerata il numero delle osservazioni è troppo basso per consentire di ottenere stime soddisfacenti delle proporzioni, anche con stimatori specifici per PA. Si tratta, in particolare, delle prime due classi della variabile SAU ("senza superficie agricola utilizzata" e "meno di un ettaro") e della provincia di Rimini, per la quale il numero di unità selezionate è molto più basso che per le altre province. Una soluzione a tale problema è da ricercarsi nella possibilità di prevedere un sovracampionamento nelle aree in cui l'impiego di stimatori specifici per PA non permette di raggiungere una variabilità accettabile, oppure nella possibilità di utilizzare variabili ausiliarie più informative ed affidabili.

Le stime ottenute con lo stimatore composto permettono di effettuare qualche confronto fra le province e fra le zone altimetriche, oltre che di valutare eventuali cambiamenti intervenuti nel periodo intercorso fra il Censimento e l'anno in cui ha avuto luogo la rilevazione dell'ISTAT (1990-1995).

Dai confronti fra province emerge che le proporzioni di aziende appartenenti alle classi di SAU più elevate si riducono passando dalle province dell'Emilia (ovest) a quelle della Romagna (est). Per quanto riguarda l'altimetria si osservano alcune differenze rilevanti solo fra la montagna e le altre due zone altimetriche, collina e pianura, nelle quali le distribuzioni delle aziende per classi di SAU appaiono fra loro molto simili.

Infine, dal confronto fra i risultati del Censimento del 1990 e quelli ottenuti in questo lavoro emergono alcuni cambiamenti, che si differenziano per provincia e per zona altimetrica. Nelle province occidentali fino a Ferrara, infatti, risulta una diminuzione della quota di aziende di dimensione più ridotta (fino ad un ettaro di SAU) ed un aumento delle quote di aziende di dimensione un po' più elevata (da 10 a 50 ettari). Nelle province di Ravenna e Forlì, invece, sembrano essersi verificate tendenze opposte. Per quanto riguarda la zona altimetrica, poi, in collina appare aumentata rispetto al Censimento la quota di aziende

che appartengono classe di SAU [O-1) e diminuita quella delle aziende appartenenti alla classe [1-5), mentre il contrario sembra essersi verificato per la montagna.

Stando ai risultati ottenuti si può concludere che i metodi di stima per PA meritano di essere impiegati ed ulteriormente sviluppati poiché risolvono una buona parte delle difficoltà legate alle basse numerosità per certi sottogruppi della popolazione. Resta l'interrogativo dell'adeguatezza delle scelte effettuate nella costruzione dello stimatore composto. In particolare, la scelta della stima della proporzione a livello regionale come componente sintetica, causata dalla mancanza di dati sulla popolazione, non può considerarsi una soluzione ottimale e potrebbe aver inficiato la qualità dei risultati.

Riferimenti bibliografici

N. Chand, C. H. Alexander (1995), *Indirect Estimation of Rates and Proportions for Small Areas with Continuos Measurement*, Sect. on Survey Res. Meth., American Stat. Ass., 549-554.

D. Cocchi (1993), *Linear Bayes Small Area Estimation*, International Statistical Conference on Small Area Statistics and Survey Design, Varsavia, 30 sep. – 3 oct. 1992, Central Statistical Office.

L. Fabbris, A. Russo, I. Sanetti (1988), *Storia e prinie proposte in tema di sovvacanipionaniento a livello regionale, provinciale e sub-provinciale per indagini sulle Forze di Lavoro*, Università degli Studi di Padova, Dipartimento di Scienze Statistiche, Rapporto di ricerca n. 4.

P. D. Falorsi, S. Falorsi, A. Russo (1994), *Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey*, Survey Methodology, n. 20, pp. 171-176.

P. J. Farrel, B. MacGibbon, T. J. Tomberlin (1997), *Empirical Bayes Small-Area Estimation Using Logistic Regression Models and Summary Statistics*, Journal of Business and Economic Statistics, American Stat. Ass., 15, 1, 101-108.

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

M. R. Ferrante, S. Pacei (1998), *Stima di Proporzioni per PA nell'Indagine ISTAT sulle Aziende Agricole*, Atti della XXXIX Riunione Scientifica della Società Italiana di Statistica, Sorrento, 14-17 Aprile.

M. Ghosh, J. N. K. Rao (1994), *Small Area Estimation: An Appraisal*, Statistics Science, vol. 9, n. 1, pp. 55-93.

E. Gori, G. M. Marchetti (1990), *Modelli lineari e componenti di varianza nella stima per PA*, Atti della XXXV Riunione Scientifica della SIS, Padova, 18-21 Aprile 1990.

D. A. Harville (1991), *Commento a "That BLUP is a Good Thing: The Estimation of Random Effects" di G. K. Robinson*, Statistical Science, n. 6, pp. 35-39.

ISTAT (1991), *4° Censimento generale dell'agricoltura- 1990, Caratteristiche tipologiche delle aziende agricole*, fascicolo Emilia Romagna.

B. MacGibbon, T. J. Tomberlin (1989), *Small* Area Estimates of Proportions Via *Empirical* Bayes Techniques, Survey Methodology, vol. 15, n. 2, pp. 237-252.

N. J. Purcell, L. Kish (1979), Estimation for *Small Domain*, Biometrics, vol. 35, pp. 365-384.

J. N. K. Rao, G. H. Choudhry (1995), *Small* Area Estimation: *Overview* and *Empirical* Study, Business Survey Methods, John Wiley & Sons.

A. Russo (1995), Stimatori per PA: problemi aperti, Atti del Convegno "100 anni di indagini campionarie", Roma, 31 maggio-1 giugno 1995, CISU.

C. E. Sarndal, B. Swensson, J. Wretmann (1992), *Model Assisted Survey Sampling*, Springer & Verlag: New York.

A. C. Singh, H. J. Mantel, B. W. Thomas (1994), *Time* Series *EBLUPs for Small* Areas *Using Survey* Data, Survey Methodology, Statistics Canada, vol. 20, n. 1, pp. 33-43.

I. Thomsen, A. M. K. Holmøy (1996), Combining *Data from Surveys* and Administrative Record System. The Norwegian Experience, 5th Independent Conference, Reykjavik, Iceland, 2-5 luglio.