istat
working
papers

N.21
2015

Disegni di campionamento alternativi per la selezione di gruppi casuali di Comuni e sezioni nell'ipotesi base dell'indagine di copertura del Censimento permanente

Claudia De Vitiis, Luigi Marcone e Francesca Inglese

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.21
2015

Disegni di campionamento alternativi per la selezione di gruppi casuali di Comuni e sezioni nell'ipotesi base dell'indagine di copertura del Censimento permanente

Claudia De Vitiis, Luigi Marcone e Francesca Inglese

Istat

Istat Working Papers

Disegni di campionamento alternativi per la sezione di gruppi casuali di Comuni e sezioni nell'ipotesi base dell'indagine di copertura del Censimento permanente

N. 21/2015

ISBN 978-88-458-1879-0

Disegni di campionamento alternativi per la selezione di gruppi casuali di Comuni e sezioni nell'ipotesi base dell'indagine di copertura del Censimento permanente[1]

Claudia De Vitiis, Luigi Marcone e Francesca Inglese

Sommario

Il presente lavoro illustra lo studio di disegni di campionamento alternativi per la progettazione dell'indagine di copertura del Censimento permanente (C-sample) che, nell'ipotesi base, prevedeva la selezione di cinque gruppi casuali (Comuni o sezioni di censimento) non sovrapposti da attribuire a ciascun anno del ciclo completo della rilevazione. L'obiettivo principale del lavoro si sostanzia dunque nell'individuazione di una partizione costituita da cinque sottoinsiemi ugualmente rappresentativi dell'universo di riferimento. Per la selezione dei gruppi casuali sono stati presi in considerazione due disegni di campionamento, il disegno di campionamento bilanciato e il disegno di campionamento stratificato. Mentre nel primo la selezione dei gruppi casuali è bilanciata rispetto a totali noti di variabili ausiliarie correlate alle variabili di interesse dell'indagine, nel secondo la selezione è effettuata all'interno degli strati tramite campionamento sistematico.

Parole chiave: selezione bilanciata, stratificazione, campionamento sistematico.

Abstract

The present work is aimed at the study of alternative sampling designs for the Permanent Census post-enumeration survey which envisaged the selection of five non-overlapping random groups (municipalities or census sections) to be assigned to each year of the full survey cycle. The main objective of the work is to identify five equally representative subsets of the reference population. For the selection of random groups balanced sampling design and stratified sampling design were taken into account. While by the first sampling design the random group selection is balanced with respect to total known auxiliary variables related to the survey target variable , by the second design type the selection is carried out within strata through systematic sampling.

Keywords: balanced sampling, stratification, systematic sampling.

[1] Le opinioni espresse impegnano esclusivamente gli autori e non implicano alcuna responsabilità da parte dell'Istat.

Dedicato a Luigi Marcone

Indice

1. Introduzione

Nella progettazione del Censimento permanente basato su un sistema di indagini sul campo, era prevista la raccolta sistematica delle liste anagrafiche comunali riferite al 1 gennaio di ciascun anno e la realizzazione di due indagini, denominate C-sample e D-sample, caratterizzate da un ciclo di rilevazione quinquennale, finalizzate rispettivamente alla correzione degli errori di copertura delle LAC e al potenziamento del potere informativo in esse contenuto.

Il presente lavoro illustra lo studio di disegni di campionamento alternativi nell'ipotesi base formulata per la progettazione dell'indagine di copertura del Censimento permanente (C-sample) dove, per il ciclo completo della rilevazione (cinque anni), era prevista la selezione di cinque gruppi casuali non sovrapposti (*random groups*) di unità statistiche appartenenti a universi di riferimento diversamente definiti a seconda della dimensione del Comune. In tale ipotesi, per i Comuni con popolazione inferiore a 50.000 abitanti (Comuni Non Autorappresentativi - NAR) l'universo di riferimento è costituito dai Comuni stessi, mentre, per i Comuni con popolazione superiore a 50.000 abitanti (Comuni Autorappresentativi - AR), l'universo di riferimento è costituito dalle sezioni di censimento[1].

L'obiettivo del lavoro consiste nella costruzione di una partizione costituita da cinque sottoinsiemi ugualmente rappresentativi dell'universo di riferimento. Per la selezione dei cinque gruppi casuali (Comuni o sezioni di censimento) sono stati studiati diversi disegni di campionamento: il disegno di campionamento bilanciato e il disegno di campionamento stratificato. Nel primo disegno di campionamento la selezione dei gruppi casuali è bilanciata rispetto a totali noti di variabili ausiliarie correlate alle variabili di interesse dell'indagine C-sample, nel disegno di campionamento stratificato la selezione è effettuata all'interno degli strati con la tecnica del campionamento sistematico.

La metodologia a cui si è fatto riferimento per la selezione simultanea di un insieme di gruppi casuali bilanciati non sovrapposti (Tillé and Favre, 2004) è costituita dal metodo *cube* (Deville e Tillé, 2004). Tale metodo è stato preferito ad altri noti in letteratura per la selezione di campioni bilanciati (Tillé, 2006, 2011) in quanto fornisce una soluzione generale al problema. Si tratta inoltre di un metodo già utilizzato per il nuovo Censimento della popolazione in Francia (INSEE; Durr e Grosbras, 2010; Durr e Clanché 2010).

L'obiettivo specifico dello studio di disegni di campionamento bilanciati è stato di valutarne la fattibilità in un contesto territoriale, quello delle regioni italiane, fortemente disomogeneo rispetto alla numerosità e alla dimensione demografica dei Comuni. A tal fine sono state ipotizzate strategie di bilanciamento alternative, prendendo in considerazione per la selezione bilanciata di Comuni sia diverse popolazioni di riferimento che differenti insiemi di variabili ausiliarie per la definizione dei vincoli di bilanciamento. Tali strategie sono state valutate effettuando selezioni ripetute dei cinque gruppi casuali, ovvero attraverso una simulazione. Alcune sperimentazioni sono state svolte, attraverso dei test, per la selezione bilanciata di sezioni in alcuni grandi Comuni aventi una popolazione superiore ai 50.000 abitanti (Comuni definiti Auto Rappresentativi, AR) della regione Emilia Romagna. In questo caso per il disegno di campionamento si è tenuto conto di vincoli di bilanciamento analoghi a quelli utilizzati per i Comuni sotto i 50.000 abitanti (Comuni definiti Non Auto Rappresentativi, NAR) e le sezioni di censimento sono state stratificate sulla base della loro dimensione in termini di popolazione. Infine, per la selezione di Comuni, si è tenuto conto di un universo di riferimento comprendente anche agglomerati di Comuni di piccole dimensioni (sotto i 5.000 abitanti) definiti con diversi metodi di aggregazione.

Per quanto riguarda il disegno di campionamento stratificato sono state ipotizzate due diverse stratificazioni dell'universo dei Comuni NAR e una sola stratificazione per le sezioni di censimento dei Comuni AR.

Il confronto tra i disegni di campionamento è stato effettuato con riferimento alla regione Emilia

[1] Progetto per la pianificazione metodologica del censimento permanente della popolazione e delle abitazioni (2013).

Romagna e ha avuto l'obiettivo di fornire una prima valutazione dell'efficienza dei disegni di campionamento adottati. Tale valutazione è stata fatta sui campioni di sezioni di censimento selezionati dai cinque gruppi casuali, ottenuti con diverse strategie, confrontando la popolazione legale al 2011 con le stime del conteggio di popolazione.

È utile precisare che le sperimentazioni che hanno riguardato la selezione bilanciata dei gruppi casuali è stata studiata soltanto in alcune regioni e in alcuni Comuni della regione Emilia Romagna, mentre la strategia di selezione dei gruppi casuali basata sui disegni di campionamento stratificati è stata applicata all'intero territorio nazionale.

Il lavoro presenta la seguente struttura: nella sezione 2 si descrive brevemente il disegno dell'indagine di copertura C-sample nell'ipotesi base; nella sezione 3 si illustrano i due disegni di campionamento studiati evidenziandone caratteristiche e proprietà teoriche; la sezione 4 riporta le strategie di selezione dei gruppi casuali adottate con riferimento ai due disegni di campionamento, il disegno di campionamento bilanciato e quello stratificato; la sezione 5 è dedicata alle sperimentazioni effettuate per la selezione bilanciata di Comuni e sezioni di censimento; la sezione 6 presenta i risultati delle diverse strategie di selezione con riferimento alla regione Emilia Romagna; infine nella sezione 7 si riportano alcune riflessioni conclusive.

2. L'indagine di copertura C-sample nell'ipotesi base

L'obiettivo principale dell'indagine di copertura del Censimento permanente (C-sample) è costituito dalla stima del numero di unità per "sottocopertura" e "sovracopertura".

L'indagine, nell'ipotesi base formulata per la progettazione del Censimento permanente, prevedeva un ciclo completo della rilevazione di cinque anni e, quindi, la selezione di cinque gruppi casuali (*random groups*) di unità statistiche appartenenti a universi di riferimento diversamente definiti a seconda della dimensione del Comune.

In tale ottica, il disegno di indagine pone l'esigenza di costruire insiemi reciprocamente esclusivi da cui estrarre le unità campione per ciascun anno del ciclo completo dell'indagine C-sample. In particolare, la selezione coinvolge cinque campioni non sovrapposti di Comuni con popolazione inferiore a 50.000 abitanti (Comuni Non Autorappresentativi - NAR) e cinque campioni non sovrapposti di sezioni di censimento (o porzioni di territorio) per i Comuni con popolazione superiore ai 50.000 abitanti (Comuni Autorappresentativi - AR).

L'assenza di sovrapposizione deve essere garantita per: (i) i Comuni NAR, in modo che ciascuno di essi sia coinvolto a campione in un solo anno del ciclo; (ii) le sezioni di censimento dei Comuni AR, in modo che ciascuna di esse rientri al più in un solo anno del ciclo completo d'indagine. Nel primo caso, nell'arco di cinque anni, tutti i Comuni sono sottoposti a rilevazione, mentre nel secondo caso la rilevazione è condotta tutti gli anni ma solo su una porzione del territorio (Fortini M., 2013).

Il disegno di campionamento annuale dell'indagine C-sample è a due stadi per i Comuni NAR (Comuni-sezioni) e ad uno stadio per i Comuni AR (sezioni). Si tratta di un campione areale in cui, per ogni sezione campione, si ripercorre l'itinerario di sezione e si rilevano tutti i civici e tutte le famiglie incluse nel territorio della sezione stessa. La dimensione campionaria totale in termini di numero di famiglie attese da intervistare ogni anno è di 650.000.

Le numerosità campionarie attese di famiglie sono assegnate sulla base dell'ampiezza demografica dei Comuni, quindi assegnate in maniera esogena secondo il seguente schema:

1) fino a 1.000 abitanti, 60 famiglie;
2) da 1 a 2.000 abitanti, 120 famiglie;
3) da 2 a 3.000 abitanti, 180 famiglie;
4) da 3 a 4.000 abitanti, 240 famiglie;
5) da 4 a 5.000 abitanti, 300 famiglie;
6) da 5 a 50.000 abitanti, 625 famiglie;
7) da 50 a 150.000 abitanti, 1.250 famiglie;
8) oltre 150.000 abitanti, 4.170 famiglie.

Per gli aspetti relativi al calcolo delle numerosità campionarie necessarie per fascia di ampiezza

demografica dei Comuni coinvolti anno per anno nell'indagine si rimanda al contributo "Il censimento continuo" (Fortini, M. 2013).

3. Disegni di campionamento alternativi per la selezione dei gruppi casuali

3.1 Il disegno di campionamento bilanciato: il metodo cube

Al fine di fornire una definizione di disegno di campionamento bilanciato ed evidenziare le proprietà di detto disegno è necessario introdurre la seguente notazione simbolica. Sia *U* la popolazione di riferimento di dimensione *N* costituita dalle unità $i \in (1,.....,N)$ e sia il totale della variabile di interesse *y*

$$Y = \sum_{i \in U} y_i \,, \qquad (1)$$

il parametro della popolazione che si vuole stimare. Si suppone, inoltre, che il vettore dei valori $\mathbf{x}_i = (x_{i1} \ldots x_{ij} \ldots x_{ip})'$ delle *p* variabili ausiliarie sia noto per tutte le unità della popolazione

$$\mathbf{X} = \sum_{i \in U} \mathbf{x}_i \,, \qquad (2)$$

e che la sua stima sia

$$\hat{\mathbf{X}}_{HT} = \sum_{i \in U} \frac{\mathbf{x}_i S_i}{\pi_i} .(3)$$

Un disegno di campionamento è detto bilanciato rispetto alle variabili ausiliarie $x_1,...,x_p$, se e soltanto se soddisfa l'equazione di bilanciamento data da

$$\hat{\mathbf{X}}_{HT} = \mathbf{X}, \qquad (4)$$

ovvero, se e soltanto se è soddisfatta l'eguaglianza tra la stima del totale di un insieme di variabili ausiliarie e il totale noto nella popolazione delle stesse variabili. Tale equazione può essere scritta anche nella forma

$$\sum_{i \in U} \frac{\mathbf{x}_{ij} S_i}{\pi_i} = \sum_{i \in U} \mathbf{x}_{ij} \,, (5)$$

dove S_i è una variabile indicatrice che assume valore uno se l'unità *i* è stata selezionata nel campione *s* e valore zero in caso contrario.

Un disegno di campionamento bilanciato è caratterizzato dalle seguenti proprietà:

a. la stima del totale di un insieme di variabili ausiliarie ottenute utilizzando lo stimatore di Horvitz-Thompson è uguale al totale noto nella popolazione delle variabili ausiliarie stesse;
b. la varianza della stima del totale di una variabile di interesse dipende dalla relazione lineare che sussiste tra la variabile di interesse e le variabili ausiliarie;
c. è un metodo di selezione che preserva l'insieme delle probabilità di inclusione del campione.

Una soluzione generale al problema della selezione di campioni bilanciati con probabilità di inclusione sia costanti che variabili e un numero non ristretto di variabili di bilanciamento è stata fornita da Deville e Tillé (2004). I due autori hanno proposto il metodo cube che è un'estensione del metodo noto come *splitting method* (Deville e Tillé, 1998). Il metodo è basato su una trasformazione del vettore delle probabilità di inclusione fino ad ottenere un campione tale da soddisfare esattamente le probabilità di inclusione e le equazioni di bilanciamento. Il nome del metodo deriva dalla rappresentazione geometrica del disegno di campionamento, infatti un campione può essere rappresentato da un vettore di variabili indicatrici ovvero

$$\mathbf{s} = \left(I\left[1 \in s\right]..I\left[i \in s\right]..I\left[N \in s\right]\right)', \qquad (6)$$

dove $I\left[1 \in s\right]$ prende valore 1 se $i \in s$ e valore 0 altrimenti. In tale ottica, un campione può essere visto come il vertice di un cubo ad *N* dimensioni e il problema della selezione di un campione bilanciato deve essere riformulato, infatti un disegno di campionamento bilanciato si traduce nella scelta di un vertice dell' *N*-cubo (Deville e Tillé, 2004). Per una descrizione più approfondita del metodo si rinvia agli articoli degli autori riportati in bibliografia.

3.1.1 Selezione contemporanea di gruppi bilanciati

Il metodo cube può essere utilizzato per la selezione bilanciata di un insieme di campioni non sovrapposti. In questo caso la condizione necessaria alla selezione di campioni bilanciati è che il complemento del primo campione (e dei successivi campioni selezionati) bilanciato su determinate variabili ausiliarie sia anch'esso bilanciato sulle stesse variabili ausiliarie.

Quando la selezione dell'insieme dei campioni bilanciati è effettuata attribuendo alle unità probabilità di inclusione uguali allora anche il complemento del primo, del secondo e dei successivi campioni, è bilanciato. Il problema della selezione di più campioni contemporaneamente è stato affrontato da Tillé e Favre (2004) i quali mettono in luce la complessità della selezione di un insieme di campioni bilanciati con probabilità di inclusione variabili.

Per la selezione bilanciata di un insieme di campioni non sovrapposti con probabilità di inclusione fissate, è necessario che ad ogni selezione siano ridefinite le probabilità di inclusione e le variabili ausiliarie da assegnare alle unità non selezionate appartenenti al complemento del primo e dei successivi campioni selezionati.

Quanto enunciato può essere formalizzato definendo con il simbolo $\bar{S}$ il complemento del campione *s* nella popolazione *U*. Le probabilità di inclusione del complemento del campione *s* sono date da $\bar{\pi}_i = P\left(i \in \bar{s}\right) = 1 - \bar{\pi}_i$. Il campione $\bar{S}$ è detto bilanciato sulle variabili **x** se

$$\sum_{i \in \bar{s}} \frac{\mathbf{x}_i}{1-\pi_i} = \sum_{i \in U} \mathbf{x}_i \,. \qquad (7)$$

Se si vogliono selezionare più campioni bilanciati sulle variabili **x** con probabilità π_i, si seleziona

il primo campione bilanciato s_1, poi si seleziona il secondo campione s_2 dal complemento di s_1 con probabilità di inclusione $\frac{\pi_i}{1-\pi_i}$ bilanciato sulle variabili $(z_i)=\frac{x_i}{1-\pi_i}$ e così via.

La procedura descritta è implementata nella macro *fast cube* (Chauvet e Tillé 2007). L'algoritmo fornisce ad ogni passo di selezione sia il campione selezionato che il suo complemento bilanciato sulle stesse variabili ausiliarie. La macro va dunque ripetuta tante volte quanti sono i campioni bilanciati che si vogliono selezionare. Essa ha l'importante caratteristica di avere dei tempi di esecuzione veloci oltre a permettere una selezione sequenziale di campioni bilanciati. Per il suo funzionamento è necessario definire: (a) i parametri relativi al database che costituisce il frame di riferimento per la selezione dei campioni bilanciati; (b) i parametri relativi al campionamento e all'output. In quest'ultimo caso per la definizione dei parametri sono previste diverse opzioni. La macro fast cube contiene rispetto alla nota macro cube un ulteriore parametro, indicato come *compeq*, attraverso il quale, selezionando l'opzione 1, si richiede che anche il complemento del campione selezionato sia esso stesso bilanciato sulle stesse variabili. La macro restituisce come output anche il complemento del campione selezionato alle cui unità sono assegnate le probabilità di inclusione e le variabili ausiliarie trasformate.

3.2 Il disegno di campionamento stratificato

Il disegno di campionamento stratificato prevede la partizione della popolazione di riferimento in strati, più precisamente in sottopopolazioni che presentano un'elevata omogeneità rispetto alla variabile di indagine e l'estrazione di un campione casuale semplice in ogni strato.

La stratificazione di una popolazione oggetto di analisi equivale alla creazione di una partizione finita delle unità che compongono la stessa. Utilizzando la terminologia propria della matematica, ciascuno dei sottoinsiemi si chiama parte, classe o blocco della partizione, che nell'accezione statistica viene etichettato come strato. Gli obiettivi per la scelta di questa operazione possono essere diversi:

- ✓ impiegare aggregazioni di unità elementari utili per la ricerca: aree o insiemi di unità con valenza amministrativa (Regioni, Province, Sistemi Locali del Lavoro, …), gruppi di unità in possesso di caratteristiche rare, gruppi estremi e/o devianti;
- ✓ separare popolazioni fisicamente isolate o con caratteristiche speciali la cui inclusione nel piano della ricerca può essere decisa caso per caso, anche per i problemi connessi alle modalità di rilevazione delle informazioni (indagini con tecnica di somministrazione del questionario mista tra telefonia fissa e posta elettronica in cui l'individuazione dei gruppi è dettata dall'informazione di contatto disponibile);
- ✓ garantire il controllo in un procedimento di selezione casuale;
- ✓ individuare sottopopolazioni omogenee rispetto alle variabili oggetto di analisi in modo da aver stime più efficienti di quanto non lo sarebbero con un campione casuale semplice.

La progettazione di un campione stratificato consente in generale di ottenere un guadagno in efficienza. L'utilizzo di informazioni disponibili sulla popolazione per la costruzione di strati comporta un aumento della precisione delle stime tanto più consistente quanto più le variabili di stratificazione sono correlate con le variabili di indagine.

All'interno degli strati dunque si effettua la selezione delle unità campionarie. Definito il criterio di stratificazione della popolazione di riferimento, la selezione casuale di un certo numero di campioni (*random group*) può essere fatta utilizzando come tecnica di estrazione il campionamento sistematico avendo prima ordinato casualmente le unità all'interno degli strati. Nel caso di un passo di campionamento $k=N/n$ non intero si considera

$$1 \le r \le k(10)^d \qquad \text{con} \qquad d = 1,2,3,... \quad (8)$$

in modo da poter selezionare esattamente le *n* unità campione tutte incluse con uguale probabilità come l'intero più vicino a

$$\frac{r+10^d/2}{10^d};\ \frac{r+k+10^d/2}{10^d}\ ;\ \frac{r+2k+10^d/2}{10^d}\ ;...;\frac{r+(n-1)k+10^d/2}{10^d}\ . \quad (9)$$

Tale tecnica di estrazione garantisce, in ogni strato, la formazione di campioni non sovrapposti. Se l'obiettivo è di ottenere gruppi casuali di dimensione fissa e ugualmente rappresentativi dell'universo di riferimento rispetto a qualche caratteristica, allora occorre scegliere la variabile ausiliaria idonea a tale rappresentazione, ordinare all'interno di ogni strato le unità in base a tale variabile, costruire dei sottoinsiemi costituiti da tante unità quanti sono i gruppi casuali desiderati e procedere alla selezione sistematica delle unità in ogni sottoinsieme definito nello strato. In questo modo, alle unità statistiche è assegnata una probabilità di inclusione costante pari all'inverso del numero di gruppi casuali che si vogliono ottenere.

4. Strategie di selezione dei gruppi casuali

4.1 La selezione nel disegno di campionamento bilanciato

Un disegno di campionamento bilanciato consente di definire dei gruppi di unità statistiche, costituenti la popolazione di riferimento, secondo regole precise che assicurano la stessa dimensione e la stessa struttura di popolazione stabilita dall'insieme di variabili ausiliarie con le quali si definiscono i vincoli di bilanciamento.

Le variabili ausiliarie devono essere correlate con le variabili di interesse e stabili nel tempo. Ad esempio, le variabili di tipo demografico come sesso e classi di età assicurano l'omogeneità dei gruppi casuali per le strutture di popolazione, mentre i totali di popolazione per sub-aree territoriali assicurano che le stesse siano rappresentate in ogni gruppo. Ciascun gruppo costituisce un campione rappresentativo dell'insieme delle unità appartenenti all'area territoriale di riferimento (livello di bilanciamento).

L'efficienza di un disegno di campionamento bilanciato dipende dal numero di unità da ripartire entro una determinata area territoriale (deve essere sufficientemente elevato) e dalla omogeneità delle stesse sia in termini di dimensione demografica che per altre caratteristiche eventualmente prese in considerazione nella definizione dei vincoli di bilanciamento.

Nel nostro contesto di studio, la definizione di una strategia di selezione bilanciata di gruppi casuali (Comuni o sezioni di censimento) richiede che siano operate scelte circa la popolazione di riferimento, il livello di bilanciamento e i vincoli di bilanciamento.

In conformità agli obiettivi della C-sample, la soglia fissata a 50.000 abitanti ha costituito la discriminante per l'individuazione di due diversi insiemi di unità statistiche su cui effettuare la selezione dei cinque gruppi casuali. Tali insiemi definiscono due universi di riferimento, il primo costituito dai Comuni di piccole e medie dimensioni e il secondo costituito dalle sezioni di censimento dei Comuni di grandi dimensioni. In tale ottica, devono essere studiate opportune strategie per la selezione bilanciata dei Comuni, da una parte, e per la selezione bilanciata delle sezioni di censimento, dall'altra.

Per quanto riguarda la selezione bilanciata di Comuni sono state individuate diverse popolazioni di riferimento: Comuni con una popolazione inferiore a 5.000, Comuni con una popolazione compresa tra 5.000 e 50.000 e il complesso dei Comuni con una popolazione inferiore a 50.000. Una popolazione di riferimento ulteriore è stata definita tenendo conto dell'insieme di unità costituite dai Comuni di piccole dimensioni (sotto 5.000 abitanti) aggregati in agglomerati secondo metodi diversi e dai rimanenti Comuni aventi una popolazione compresa tra 5.000 e 50.000 abitanti.

Il livello di bilanciamento è costituito dalla regione per la selezione bilanciata dei Comuni e dal

Comune per la selezione bilanciata delle sezioni di censimento nei Comuni di grandi dimensioni (sopra i 50.000 abitanti).

I vincoli di bilanciamento sono stati definiti utilizzando diversi criteri: per la selezione bilanciata di Comuni si è tenuto conto di due diverse strutture di popolazione definite dalle variabili sesso, età e cittadinanza, oppure di variabili ausiliarie sostanzialmente riconducibili a totali di popolazione di determinate aree geografiche; per la selezione bilanciata delle sezioni di censimento si è fatto riferimento a vincoli analoghi a quelli considerati per i Comuni, insieme a diverse ipotesi di stratificazione delle sezioni per tipo di località o per classi di popolazione.

4.2 La selezione nel disegno di campionamento stratificato

La selezione dei cinque gruppi casuali di Comuni e di sezioni di censimento secondo il disegno di campionamento stratificato è stata effettuata considerando due diverse ipotesi di stratificazione per i Comuni NAR (sotto i 50.000 abitanti) ed una sola stratificazione per le sezioni di censimento basata sulla tipologia di località.

Nella prima ipotesi (campione stratificato A), la stratificazione dei Comuni NAR, che costituiscono le unità di primo stadio, è costituita dalle 110 province, mentre nella seconda ipotesi (campione stratificato B) gli strati sono definiti dall'incrocio provincia (110) e due classi dimensionali di Comuni (Comuni fino a 20.000 abitanti e Comuni con una popolazione tra 20.000 e 49.999 abitanti).

Per quanto riguarda i Comuni con una popolazione superiore ai 50.000 abitanti (Comuni AR), per i quali è prevista la selezione di cinque gruppi casuali di sezioni, è stata adottata un'unica stratificazione costituita da due soli strati che identificano le sezioni di censimento sulla base della tipologia: i due strati sono costituiti dalle sezioni *località di centro* e *altro tipo di località* all'interno di ciascun Comune.

La scelta degli strati è stata guidata sia dalla necessità di mantenere un certo controllo sulla procedura casuale di selezione dei Comuni, sia dalla riconosciuta omogeneità delle grandezze già definite "*sovracopertura*" e "*sottocopertura*" delle anagrafi in confronto al dato desunto dalla popolazione legale censuaria, nelle diverse fasce di ampiezza demografica definite per ogni provincia.

La selezione dei cinque gruppi casuali (Comuni o sezioni di censimento) è in ogni caso effettuata all'interno di ciascun strato e la tecnica di estrazione è quella del campionamento sistematico.

In ciascuno strato di Comuni NAR definito secondo le due ipotesi è stata operata una ripartizione degli stessi in un numero di sottogruppi al più pari a cinque sulla base di un ordinamento crescente per numero di abitanti desunto dalla popolazione legale. L'etichetta da 1 a 5, associata casualmente a ciascun Comune nei sottogruppi, ha consentito di definire i *random group* sulle cui unità procedere alla allocazione del numero di sezioni campione ripartite proporzionalmente per tipologia di località (centro abitato, altra località) e necessarie a raggiungere la quota predefinita del numero atteso di famiglie.

Le sezioni di censimento di ciascuno dei 142 comuni AR sono state associate a un'etichetta casuale tra 1 e 5 in un ordinamento casuale per tipologia di località. Per garantire la non sovrapposizione dei campioni di sezioni, la selezione dei gruppi casuali è stata effettuata per ciascun Comune all'interno degli strati definiti dalla tipologia di località, prescindendo in questo caso dall'ordinamento delle singole sezioni per ampiezza demografica.

E' utile sottolineare che, in alcuni casi, la costruzione di sottoinsiemi di Comuni costituiti da cinque unità all'interno di ogni strato non è stata sempre possibile. Là dove necessario sono state operate scelte ad hoc.

La costruzione di sottoinsiemi di unità è stata ulteriormente complicata quando è stata considerata una soglia di popolazione dei Comuni delimitata da 5.000 abitanti. Le province di Monza Brianza e Brindisi non hanno, infatti, Comuni fino a 5.000 abitanti; per 12 province si è creato un solo gruppo ricorrendo, quando necessario, a donazioni di unità dalla fascia di ampiezza demografica immediatamente superiore a 5.000 abitanti per raggiungere almeno le 5 unità; per 9 province i Comuni fino a 5.000 abitanti sono stati ripartiti in due classi, per 13 province in 3 classi, per 6 in 4 gruppi e per le restanti 68 province i Comuni sono stati collocati nei rispettivi quintili.

Nella Tavola 1 è riportato il numero di Comuni NAR per ciascun anno del ciclo della rilevazio-

ne secondo le due ipotesi di stratificazione e il numero di sezioni per i Comuni AR. Dalla tabella è possibile notare che la selezione dei cinque gruppi di Comuni effettuata a partire dalla definizione di due differenti stratificazioni della popolazione porta a risultati abbastanza simili, tuttavia se si guarda al tipo di stratificazione A emerge che il quarto gruppo casuale ha un numero di Comuni (1.600) che è il più elevato rispetto a tutti gli altri gruppi selezionati, mentre il terzo gruppo casuale ottenuto a partire dalla stratificazione di tipo B ha il numero più piccolo di Comuni (1.573).

Tavola 1 – Numero di Comuni NAR e numero di Sezioni (Comuni AR) per gruppo casuale

Gruppi casuali	Tipo di stratificazione		Numero di sezioni (Comuni AR)
	A	B	
	Numero di comuni NAR		
1	1.587	1.585	26.104
2	1.585	1.599	26.098
3	1.590	1.573	26.101
4	1.600	1.595	26.127
5	1.588	1.598	26.098
Totale	7.950	7.950	130.528

5. Sperimentazioni per la selezione bilanciata dei gruppi casuali

5.1 La selezione di Comuni

La selezione bilanciata dei Comuni ha l'obiettivo di ottenere cinque campioni ognuno dei quali rappresentativo dell'intera popolazione relativamente ai criteri individuati, che siano tra loro non sovrapposti e il più possibile equivalenti. Per queste ragioni la selezione è stata effettuata con probabilità uguali.

La sperimentazione per la selezione casuale di cinque gruppi di Comuni è stata condotta in regioni molto diverse tra loro sia per il numero di Comuni coinvolti che per la dimensione demografica degli stessi e secondo strategie di bilanciamento alternative. Per la valutazione della qualità del bilanciamento delle diverse strategie è stata effettuata una simulazione che ha coinvolto 500 selezioni di cinque gruppi di Comuni con probabilità uguali $\pi_k = 1/5$.

Tale valutazione è stata basata sul coefficiente di variazione di tutte le stime calcolate sui 2.500 campioni selezionati. Il valore assunto dall'indice di variabilità delle stime ottenute dalla selezione ripetuta di cinque campioni bilanciati fornisce, in sostanza, una misura di sintesi della qualità del bilanciamento; valori elevati del coefficiente di variazione sono indicativi del fatto che l'errore di bilanciamento può assumere valori piccoli per alcuni campioni e valori elevati per altri. Calcolato sull'insieme delle stime, esso permette, inoltre, di confrontare strategie di bilanciamento differenti.

5.1.1 Vincoli sulla struttura demografica della popolazione

Le popolazioni di riferimento definite nelle regioni Piemonte, Emilia Romagna e Calabria, sono costituite dai Comuni sotto la soglia di 50.000 abitanti distinti in piccoli e medi.

La selezione è stata effettuata ipotizzando due diverse strutture di popolazione per la definizione dei vincoli di bilanciamento. Nella prima sono state considerate 10 classi di età distinte per maschi e femmine, nella seconda sono state considerate cinque classi di età (senza distinzione di sesso), le variabili *numero di maschi*, *numero di femmine* e *numero di stranieri*.

Le tabelle che seguono presentano il prospetto in cui sono descritte le variabili ausiliarie utilizzate per la definizione dei vincoli di bilanciamento (Tavola 2) e un'analisi delle popolazioni di riferimento nelle tre regioni considerate che descrive la numerosità, la dimensione media dei Comuni in termini di popolazione e la variabilità della dimensione demografica dei Comuni (Tavola 3).

Tavola 2 – Variabili ausiliarie per la definizione dei vincoli di bilanciamento

Primo criterio di bilanciamento		Secondo criterio di bilanciamento	
Variabili ausiliarie	Descrizione	Variabili ausiliarie	Descrizione
X_1	N. maschi, età <20	X_1	N. individui, età <20
X_2	N. maschi, età 20-39	X_2	N. individui, età 20-39
X_3	N. maschi, età 40-59	X_3	N. individui, età 40-59
X_4	N. maschi, età 60-74	X_4	N. individui, età 60-74
X_5	N. maschi, età >75	X_5	N. individui, età >75
X_6	N. femmine, età <20	X_6	N. maschi
X_7	N. femmine, età 20-39	X_7	N. femmine
X_8	N. femmine, età 40-59	X_8	N. stranieri
X_9	N. femmine, età 60-74	-	-
X_{10}	N. femmine, età >75	-	-

Tavola 3 – Numero di Comuni, media e coefficiente di variazione della dimensione dei Comuni per Regione e soglie di popolazione

Regione	Totale comuni	Comuni < 5.000			Comuni 5.000- 50.000		
		N. comuni	Dimensione media	CV %	N. comuni	Dimensione media	CV %
Piemonte	1.206	1.073	1.228	91,2	127	14.158	75,7
Emilia Romagna	348	158	2.652	49,5	177	11.882	59,4
Calabria	409	327	2.008	59,1	77	10.824	66,7

Fonte: Popolazione legale 2011

Dalla Tavola 3 si nota che le regioni sono molto diverse tra di loro soprattutto per il numero di Comuni ma anche per l'omogeneità degli stessi riguardo alla loro dimensione. E' evidente che la regione Emilia Romagna ha un numero di Comuni sotto la soglia di 5.000 abitanti molto contenuto ed anche più omogenei, infatti il coefficiente di variazione della dimensione demografica è pari al 49,5%. La regione Calabria ha invece un numero di Comuni di medie dimensioni alquanto contenuto rispetto alle altre regioni considerate.

Le successive tabelle riportano i risultati della simulazione per la selezione dei Comuni con una popolazione inferiore ai 5.000 abitanti e per la selezione dei Comuni con una popolazione compresa tra 5.000 e 50.000 abitanti nelle tre regioni effettuata secondo i due criteri di bilanciamento descritti. Per ogni variabile ausiliaria utilizzata come vincolo di bilanciamento è riportato il coefficiente di variazione percentuale delle stime (totali delle variabili ausiliarie) ottenute dalle selezioni ripetute.

Tavola 4 – Coefficiente di variazione delle stime – Regione Piemonte

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni < 5.000	Comuni 5.000- 50.000		Comuni < 5.000	Comuni 5.000- 50.000
X_1	0,76	5,04	X_1	0,76	5,14
X_2	0,73	5,05	X_2	0,73	5,15
X_3	0,69	5,14	X_3	0,69	5,31
X_4	0,69	5,54	X_4	0,69	5,66
X_5	0,67	5,5	X_5	0,68	5,75
X_6	0,76	5,07	X_6	0,69	5,25
X_7	0,74	5,1	X_7	0,7	5,35
X_8	0,71	5,33	X_8	1,13	6,17
X_9	0,72	5,67	-	-	-
X_{10}	0,71	5,73	-	-	-

Tavola 5 – Coefficiente di variazione delle stime– Regione Emilia Romagna

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni < 5.000	Comuni 5.000- 50.000		Comuni < 5.000	Comuni 5.000-50.000
X_1	2,39	3,15	X_1	2,41	3,29
X_2	2,38	3,11	X_2	2,4	3,28
X_3	2,34	3,23	X_3	2,35	3,39
X_4	2,35	3,47	X_4	2,39	3,63
X_5	2,45	3,66	X_5	2,41	3,7
X_6	2,45	3,19	X_6	2,29	3,32
X_7	2,46	3,18	X_7	2,33	3,42
X_8	2,4	3,35	X_8	3,16	3,71
X_9	2,49	3,68	-	-	-
X_{10}	2,53	3,68	-	-	-

Tavola 6 – Coefficiente di variazione delle stime - Regione Calabria

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni < 5.000	Comuni 5.000- 50.000		Comuni < 5.000	Comuni 5.000-50.000
X_1	1,65	9,15	X_1	1,68	9,15
X_2	1,53	9,48	X_2	1,53	9,58
X_3	1,41	9,1	X_3	1,38	9,3
X_4	1,39	9,02	X_4	1,32	8,98
X_5	1,39	8,15	X_5	1,3	7,82
X_6	1,69	9,09	X_6	1,43	9,1
X_7	1,52	9,49	X_7	1,4	9,07
X_8	1,42	9,46	X_8	5,05	13,62
X_9	1,37	9,13	-	-	-
X_{10}	1,37	7,68	-	-	-

Segue, infine, una tabella che presenta una sintesi dei risultati della selezione bilanciata dei Comuni (Tavola 7). Tale sintesi si concretizza per ogni regione, popolazione di riferimento e criterio di bilanciamento adottato, nella individuazione del valore più elevato del coefficiente di variazione percentuale tra quelli calcolati per singola stima. Tale valore è in qualche modo indicativo della situazione peggiore a cui la selezione bilanciata può portare.

Tavola 7 – Valore massimo del Coefficiente di Variazione % delle stime per Regione e criterio di bilanciamento

Regione	Valore massimo del CV % delle stime			
	Primo criterio di bilanciamento		Secondo criterio di bilanciamento	
	Comuni < 5.000	Comuni 5.000- 50.000	Comuni < 5.000	Comuni 5.000- 50.000
Piemonte	0,76	5,73	1,13	6,17
Emilia Romagna	2,53	3,68	3,16	3,71
Calabria	1,69	9,49	5,05	13,62

Dalla Tavola 7 emerge che un disegno di campionamento bilanciato di Comuni produce generalmente risultati migliori nella regione Emilia Romagna, tranne quando il bilanciamento è effettuato per i Comuni con meno di 5.000, dove i valori più bassi del coefficiente di variazione delle stime si registrano per le altre due regioni prese in esame. Tali risultati sono certamente condizionati dal maggior numero di unità coinvolte nel bilanciamento, soprattutto nella regione Piemonte che

ha un numero di comuni piccoli molto elevato (1.073).

Le migliori performance del bilanciamento nella regione Emilia Romagna sono determinate da una più equilibrata distribuzione dei Comuni nelle due soglie di popolazione considerate (sono anche Comuni più omogenei per dimensione demografica) e da una maggiore conformità della struttura di popolazione con quelle ipotizzate.

I risultati peggiori si evidenziano nella regione Calabria dove il coefficiente di variazione delle stime raggiunge un valore massimo del 9,49% quando la selezione dei Comuni è effettuata sull'insieme dei Comuni individuati all'interno della soglia 5.000-50.000 e adottando il primo criterio di bilanciamento (Tavola 6), ed è del 13,62% quando la selezione dei Comuni è effettuata all'interno della stessa soglia ma con il secondo criterio di bilanciamento (Tavola 6).

Nel primo caso è possibile che la struttura di popolazione ipotizzata non è particolarmente adatta alla struttura della popolazione della regione, infatti il valore elevato del coefficiente di variazione è riconducibile alla variabile *età* definita dalla classe 20-39. Per quanto riguarda il secondo criterio di bilanciamento il valore elevato del coefficiente di variazione è riconducibile alla variabile *numero di stranieri*. La Calabria è infatti una regione che non presenta un elevato numero di stranieri.

5.1.2 Vincoli sul totale di popolazione

La seconda sperimentazione ha interessato la selezione bilanciata dei Comuni delle regioni Piemonte ed Emilia Romagna appartenenti a classi di popolazione diversamente definite: Comuni sotto la soglia di 5.000, Comuni con una popolazione compresa tra 5.000 e 50.000, Comuni sotto la soglia di 50.000. In questo caso, per ogni universo di riferimento, il vincolo di bilanciamento è costituito dal totale di popolazione.

Nelle tabelle seguenti sono riportati i risultati della simulazione della selezione bilanciata dei cinque gruppi casuali di Comuni nelle regioni Piemonte (Tavola 8) ed Emilia Romagna (Tavola 9). Il coefficiente di variazione percentuale è calcolato sulle stime del totale della popolazione dei tre universi di Comuni.

All'indice di variabilità delle stime sono affiancati anche alcune statistiche, come la numerosità e la dimensione demografica media dei Comuni, e il coefficiente di variazione della popolazione dei Comuni, data la stretta relazione delle stesse con la qualità del bilanciamento.

Tavola 8 - Coefficiente di variazione % delle stime, Numero e dimensione media dei Comuni, coefficiente di variazione (popolazione) per universo di riferimento – Regione Piemonte

Popolazione di riferimento	CV % stime	N. comuni	Dimensione media	CV % popolazione
Comuni sotto 5mila	0,62	1.073	1.228	91,21
Comuni tra 5-50mila	5,04	127	14.158	75,75
Comuni sotto 50mila	2,69	1.200	2.596	207,58

Fonte: Popolazione legale 2011

Tavola 9 - Coefficiente di variazione % delle stime, Numero e dimensione media dei Comuni, coefficiente di variazione (popolazione) per universo di riferimento – Regione Emilia Romagna

Popolazione di riferimento	CV % stime	N. comuni	Dimensione media	CV % popolazione
Comuni sotto 5mila	2,18	158	2.652	49,55
Comuni tra 5-50mila	3,2	177	11.882	59,42
Comuni sotto 50mila	2,32	335	7.529	92,38

Fonte: Popolazione legale 2011

Un importante risultato che si evidenzia dalle due tabelle è la non buona performance del disegno di campionamento bilanciato per i Comuni di dimensione compresa tra 5.000-50.000 abitanti in entrambe le regioni (Piemonte ed Emilia Romagna).

5.1.3 La selezione di agglomerati di Comuni

Nelle sperimentazioni si è tenuto conto anche della costituzione di agglomerati di Comuni di piccole dimensioni (sotto i 5.000 abitanti) definiti secondo diverse ipotesi riportate in modo dettagliato nel rapporto tecnico prodotto nella prima fase di svolgimento del Progetto per la pianificazione metodologica del censimento permanente [2].

Nelle regioni Piemonte, Emilia Romagna e Calabria, gli agglomerati di Comuni sono stati definiti sulla base di due differenti approcci che tengono conto in ogni caso delle Unioni di Comuni già esistenti.

Il primo metodo (1) assume come vincolanti alcuni parametri territoriali (la contiguità tra Comuni dello stesso agglomerato è obbligatoria, l'appartenenza alla stessa Unione o Comunità montana è obbligatoria) e, ove necessario, adotta criteri di preferenza (come l'appartenenza allo stesso versante geografico e la similitudine demografica); la dimensione minima dell'agglomerato è fissata a 5.000 abitanti, mentre nessun limite massimo è fissato agglomerando insieme tutti i Comuni appartenenti alla stessa Unione.

Nel secondo metodo (2) nessun vincolo, eccetto la contiguità, è considerato stringente; tra i possibili Comuni limitrofi sono preferiti quelli appartenenti alla stessa Unione, allo stesso Sistema Locale del Lavoro, alla stessa regione SGI, con altimetria e demografia più simile, appartenenza alla stessa provincia.

Le tabelle che seguono riportano sia un'analisi degli agglomerati di Comuni costruiti sulla base dei due metodi descritti che i risultati della simulazione per la selezione dell'insieme delle unità identificate dagli agglomerati di Comuni e dai Comuni con popolazione compresa tra 5.000 e 50.000 abitanti nelle tre regioni considerate effettuata secondo gli stessi criteri di bilanciamento utilizzati in precedenza.

La Tavola 10 evidenzia che le tre regioni sono molto diverse tra di loro per il numero di agglomerati ma anche per la loro dimensione. Infatti la dimensione degli agglomerati definiti secondo il metodo 1 arriva ad un massimo di 118.332 abitanti nella regione Piemonte.

Tavola 10 - Numerosità e struttura degli agglomerati di Comuni piccoli per Regione

Regione	Metodo	N. totale comuni <5.000	N. di agglomerati	N. comuni in agglomerati	N. max di comuni in agglomerati	Popolazione	
						Minimo	Massimo
Piemonte	1	1.200	197	1.092	25	5.026	118.332
	2		190	1.077	24	5.012	27.306
Emilia Romagna	1	335	59	175	7	3.100	72.536
	2		59	172	7	4.996	62.237
Calabria	1	404	92	330	8	5.054	20.722
	2		96	331	8	5.029	20.046

[2] Rapporto tecnico non pubblicato redatto per il "Progetto per la pianificazione metodologica del censimento permanente della popolazione e delle abitazioni (2013) " dal sottogruppo incaricato del WP2: Talice S.*et al*, Parte I-Aggregazione di comuni piccoli; Inglese F. *et al*, Parte II- Disegno di campionamento bilanciato

Tavola 11 – Coefficiente di variazione delle stime – Regione Piemonte

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni e agglomerati di comuni			Comuni e agglomerati di comuni	
	Metodo 1	Metodo 2		Metodo 1	Metodo 2
X_1	3,02	2,93	X_1	3,03	2,92
X_2	3	2,9	X_2	3,03	2,93
X_3	2,95	2,86	X_3	3,04	2,98
X_4	3,15	3,05	X_4	3,23	3,15
X_5	3,03	2,96	X_5	3,10	3,08
X_6	3,04	2,94	X_6	3,00	2,91
X_7	3,05	2,97	X_7	3,10	3,02
X_8	3,14	3,05	X_8	3,62	3,52
X_9	3,34	3,24	-	-	-
X_{10}	3,15	3,07	-	-	-

Tavola 12 – Coefficiente di variazione delle stime – Regione Emilia Romagna

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni e agglomerati di comuni			Comuni e agglomerati di comuni	
	Metodo 1	Metodo 2		Metodo 1	Metodo 2
X_1	2,86	2,77	X_1	2,73	2,65
X_2	2,79	2,69	X_2	2,68	2,62
X_3	2,84	2,73	X_3	2,77	2,7
X_4	2,94	2,85	X_4	2,95	2,86
X_5	3,02	2,92	X_5	2,91	2,83
X_6	2,9	2,79	X_6	2,7	2,63
X_7	2,84	2,75	X_7	2,79	2,71
X_8	2,96	2,85	X_8	3,04	2,94
X_9	3,17	3,06	-	-	-
X_{10}	3,04	2,93	-	-	-

Tavola 13 – Coefficiente di variazione delle stime – Regione Calabria

Primo criterio di bilanciamento			Secondo criterio di bilanciamento		
Variabili ausiliarie	CV % delle stime		Variabili ausiliarie	CV % delle stime	
	Comuni e agglomerati di comuni			Comuni e agglomerati di comuni	
	Metodo 1	Metodo 2		Metodo 1	Metodo 2
X_1	5,3	5,27	X_1	5,15	5,03
X_2	5,42	5,38	X_2	5,33	5,21
X_3	5,03	5	X_3	5,04	4,93
X_4	4,83	4,78	X_4	4,76	4,64
X_5	3,88	3,85	X_5	3,75	3,58
X_6	5,26	5,23	X_6	4,92	4,8
X_7	5,48	5,46	X_7	4,91	4,79
X_8	5,28	5,24	X_8	8,32	8,25
X_9	4,89	4,84	-	-	-
X_{10}	3,66	3,6	-	-	-

Tavola 14 – Valore massimo del Coefficiente di Variazione % delle stime per Regione e criterio di bilanciamento

Regione	Valore massimo del CV % delle stime - Comuni e agglomerati di comuni			
	Primo criterio di bilanciamento		Secondo criterio di bilanciamento	
	Metodo 1	Metodo 2	Metodo 1	Metodo 2
Piemonte	3,34	3,24	3,62	3,52
Emilia Romagna	3,17	3,06	3,04	2,94
Calabria	5,48	5,46	8,32	8,25

È interessante notare come, quando la popolazione di riferimento della selezione bilanciata è costituita da aggregazioni di Comuni piccoli e Comuni con popolazione compresa tra 5.000 e 50.000 abitanti, i valori del coefficiente di variazione delle stime si attestano intorno al 3% tranne che nella regione Calabria dove, eventualmente, le strutture di popolazione adottate sono meno pertinenti con quelle della regione. La selezione bilanciata presenta in questo caso performance migliori di quella effettuata sull'universo dei Comuni compresi tra 5.000 e 50.000 abitanti.

Le sperimentazioni aventi come universo di riferimento aggregazioni di piccoli Comuni e Comuni con popolazione compresa tra 5.000 e 50.000 abitanti ottenuti con metodi diversi sono state effettuate considerando un unico vincolo di bilanciamento costituito dal totale della popolazione dei Comuni costituenti l'universo di riferimento.

Per la regione Piemonte gli agglomerati di piccoli Comuni sono stati definiti (metodo 3) tenendo conto, come vincoli, della delimitazione geografica della provincia e di una dimensione minima (5.000 abitanti) e massima (10.000 abitanti) della popolazione, mentre per la regione Emilia Romagna sono stati definiti sulla base di tre criteri diversi in cui si è tenuto conto di una dimensione minima di 2.500 abitanti (metodo a), di 3.500 abitanti (metodo b) oppure di una dimensione massima di 5.000 abitanti (metodo c). In ogni caso il vincolo principale rimane sempre la contiguità territoriale.

Tavola 15 - Coefficiente di variazione % delle stime, Numero e dimensione media dei Comuni, coefficiente di variazione (popolazione) per Regione e Comuni e agglomerati di comuni

Regione	Popolazione di riferimento	CV % stime	N. comuni	Dimensione media	CV % popolazione
Piemonte	Comuni e agglomerati (min 5.000)	2,77	319	9.766	78,95
Emilia Romagna	Comuni e agglomerati (min 2.500)	2,41	315	8.007	86,3
	Comuni e agglomerati (min 3.500)	2,48	286	8.819	77,17
	Comuni e agglomerati (max 5.000)	2,53	264	9.953	70,26

Fonte: Popolazione legale 2011

Se si analizzano i risultati della Tavola 15 si nota che, in generale, il coefficiente di variazione delle stime non si discosta di molto da quello relativo alla selezione dell'insieme dei Comuni sotto la soglia di 50.000 abitanti (Tabelle 8 e 9). In Piemonte questo risultato si ottiene anche se il numero di unità complessive su cui è effettuata la selezione bilanciata dei cinque gruppi casuali diminuisce di molto (319 agglomerati rispetto a 1200 comuni sotto i 5.000 abitanti) perché compensato da una notevole riduzione della variabilità della dimensione demografica delle unità coinvolte nella selezione (78,95%).

Per quanto riguarda invece l'Emilia Romagna la selezione bilanciata dei cinque gruppi di Comuni e agglomerati porta ad un peggioramento quando le unità coinvolte diminuiscono; in questo caso la lieve diminuzione della variabilità della dimensione demografica delle unità coinvolte nella selezione non riesce a compensare la diminuzione delle numerosità complessive dei Comuni coinvolti nella selezione bilanciata.

5.2 I test per la selezione bilanciata di sezioni

La selezione bilanciata delle sezioni di censimento ha l'obiettivo di suddividere, per ogni Co-

mune con popolazione superiore a 50.000 abitanti, le sezioni in cinque gruppi casuali ognuno dei quali rappresentativo dell'intera popolazione, che siano tra loro non sovrapposti e il più possibile equivalenti. Per queste ragioni la selezione dei gruppi casuali è stata effettuata assegnando alle sezioni una uguale probabilità di selezione.

Le tecniche di selezione utilizzate per la selezione dei campioni di sezioni ricalcano quanto già definito e descritto per la selezione dei Comuni con popolazione inferiore a 50.000 abitanti. Infatti gli schemi di campionamento che sono stati presi in considerazione seguono l'approccio della selezione bilanciata che consente di ottenere cinque campioni, o gruppi casuali, ciascuno dei quali rappresentativo della popolazione rispetto a predefiniti criteri, i criteri di bilanciamento.

Sono stati effettuati due test di selezione bilanciata su alcuni Comuni della regione Emilia Romagna, individuati a scopo esemplificativo per valutare la performance della selezione bilanciata delle sezioni su Comuni di differente dimensione. I Comuni sono Bologna, Ferrara e Faenza, ossia quello di maggiore dimensione demografica, uno di dimensione intermedia (132.545) e il più piccolo (57.748).

Nella tabella seguente (Tavola 16) sono riportate, con riferimento all'insieme di tutti i Comuni della regione Emilia Romagna con popolazione maggiore di 50.000 e per i tre comuni considerati, le distribuzioni del numero di sezioni e della popolazione per tipologia di sezione e per fascia dimensionale.

Tavola 16 - Distribuzione delle sezioni nei Comuni di Bologna, Ferrara e Faenza per tipo di località e fascia dimensionale di popolazione - Regione Emilia Romagna

Tipologia di sezione	Tutti i comuni >50.000		Bologna		Ferrara		Faenza	
	Numero	Popolazione	Numero	Popolazione	Numero	Popolazione	Numero	Popolazione
Centro abitato	15.568	1.677.487	1.917	366.439	1.247	119.492	418	45.633
Altra tipologia	2.421	142.505	156	4.898	428	13.053	168	12.115
<50	7.616	176.734	526	10.945	826	18.072	242	6.448
50-249	8.672	977.519	952	128.908	770	85.791	300	33.334
>=250	1.701	665.739	595	231.484	79	28.682	44	17.966

Fonte: Popolazione legale 2011

Dalla Tavola 16 emerge chiaramente che la stratificazione per tipologia non può essere funzionale alla definizione di uno schema campionario, dal momento che circa il 90% della popolazione ricade nella tipologia di centro abitato.

Per quanto riguarda invece le fasce di popolazione, si vede che la distribuzione è molto diversa tra i tre Comuni considerati e che pertanto sarebbe necessario definire una stratificazione, più fine, differente comune per comune. Quest'ultima è la ragione per la quale il test di selezione bilanciata dei gruppi casuali è stata effettuata, a scopo di esempio, solamente per il Comune di Bologna, definendo per questo una stratificazione ad hoc.

Il primo test ha preso in considerazione le stesse variabili di bilanciamento utilizzate per le sperimentazioni sui Comuni con popolazione al di sotto dei 50.000 abitanti, ossia le variabili, riportate nella Tavola 2, definite secondo il primo criterio di bilanciamento, mentre per il secondo si è presa in considerazione solamente la variabile popolazione totale.

Relativamente a questo secondo test, è stato fatto anche un tentativo di selezione bilanciata rispetto alla popolazione suddivisa per strati di sezioni; tuttavia, per le ragioni che saranno spiegate nel seguito, quest'ultimo test non è stato possibile effettuarlo in tutti e tre i comuni considerati.

I risultati della selezione bilanciata delle sezioni sono valutati tramite il calcolo dell'errore di bilanciamento basato sullo scarto relativo tra i totali noti delle variabili ausiliarie utilizzate come vincoli di bilanciamento e i totali stimati, tramite lo stimatore di Horvitz-Thompson, delle stesse variabili ausiliarie,

$$R = \frac{\left|\mathbf{X} - \hat{\mathbf{X}}_{HT}\right|}{\mathbf{X}} * 100 . \quad (10)$$

Nella tabella seguente (Tavola 17) sono riportate le principali statistiche relative all'errore di bilanciamento riscontrato nel primo test condotto sui tre Comuni. Come si vede e come era prevedibile, per il Comune di Bologna, che ha un elevato numero di sezioni, il bilanciamento funziona molto meglio, anzi, si può affermare che l'errore di bilanciamento è funzione decrescente della popolazione e del numero di sezioni del Comune.

Tavola 17 - Minimo e massimo dello scarto relativo tra totale noto della popolazione e stima del totale (HT) della popolazione per Comune – Regione Emilia Romagna

Comune	Popolazione	N. sezioni	Errore di bilanciamento %	
			Min	Max
Bologna	371.337	2.071	0,00	0,85
Ferrara	132.545	1.673	0,07	3,46
Faenza	57.748	586	0,08	9,30

Fonte: Popolazione legale 2011

Per quanto riguarda invece il secondo test, che segue lo schema della seconda sperimentazione sui Comuni al di sotto dei 50.000 abitanti, è stata testata la selezione di cinque gruppi bilanciati di sezioni con probabilità uguali e pari a 1/5 con un solo vincolo relativo alla popolazione totale del Comune.

Come già detto, solo per il Comune di Bologna è stata anche effettuata una selezione delle sezioni con vincolo sulla popolazione di cinque strati (<50, 50-150, 150-300, 300-400, >400). Infatti, si è verificato che non è possibile utilizzare soglie fisse per la definizione di tali strati, dal momento che in Comuni con differente dimensione e struttura la distribuzione delle sezioni per popolazione è del tutto dissimile. Tale circostanza rende non praticabile la selezione dei cinque gruppi casuali che rispettino la popolazione di detti strati prefissati.

I risultati della selezione bilanciata test con vincolo sulla popolazione comunale (ipotesi di bilanciamento senza stratificazione) e con vincolo sulla popolazione di cinque strati (ipotesi di bilanciamento con stratificazione) sono riportati nella Tavola 18, dove ciò che risulta evidente è che al decrescere della dimensione della popolazione e del numero di sezioni dei comuni aumenta l'errore di bilanciamento.

Tavola 18 - Minimo e massimo dello scarto relativo tra totali noti della popolazione e stime del totale (HT) della popolazione per le variabili di bilanciamento nelle due ipotesi (senza e con stratificazione) per Comune – Regione Emilia Romagna

Comune	Errore di bilanciamento %			
	Senza stratificazione		Con stratificazione	
	Min	Max	Min	Max
Bologna	0,007	0,277	0,009	2,19
Ferrara	0,019	0,283	-	-
Faenza	0,464	1,38	-	-

6. Confronto tra disegni di campionamento alternativi – un'applicazione alla regione Emilia-Romagna

6.1 Premessa

Al fine di condurre appropriate analisi di confronto tra i diversi schemi di campionamento, nella

selezione bilanciata dei gruppi casuali è stata utilizzata una strategia di bilanciamento che fosse il più simile possibile alle scelte adottate per la selezione effettuata secondo i disegni di campionamento stratificati.

In sostanza per la selezione dei Comuni con un numero di abitanti inferiore a 50.000 (Comuni NAR) è stato posto come vincolo di bilanciamento la dimensione della popolazione provinciale, mentre per la selezione dei gruppi casuali di sezioni di censimento nei Comuni con un numero di abitanti superiore a 50.000 (Comuni AR) i vincoli di bilanciamento sono costituiti dal totale della popolazione delle sezioni distinte in due tipologie di località.

Le variabili utilizzate qui come vincoli di bilanciamento sono le stesse utilizzate per la definizione degli strati nei due disegni di campionamento alternativi utilizzati. Un ulteriore aspetto da sottolineare è che per i campioni selezionati tramite disegno di campionamento bilanciato è stata comunque effettuata una prima valutazione dell'errore di bilanciamento.

Una volta selezionati i cinque gruppi casuali di Comuni NAR e i cinque gruppi casuali di sezioni di censimento per i Comuni AR secondo le diverse strategie, è stato ottenuto per ogni gruppo casuale un campione di sezioni di censimento. Il numero di sezioni di censimento da selezionare è stato definito sulla base della quota di famiglie attribuita annualmente nella rilevazione secondo lo schema descritto alla sezione 2.

L'universo di riferimento per procedere all'estrazione delle unità di secondo stadio, per i Comuni NAR, è formato dalle sezioni che contano almeno un individuo censito nella popolazione legale del 2011 di cui si dispone del dato relativo al numero di famiglie censite. In ogni Comune campione, in cui le sezioni di censimento sono state distinte in base a due tipologie di località (centro e altro) è stato calcolato il numero di famiglie medio per sezione di censimento. Questo ha consentito di allocare il numero di sezioni campione attese necessarie a raggiungere il target esogeno di famiglie associato a ciascuna località di ogni Comune campione[3]. Anche per l'estrazione delle sezioni campione è stata utilizzata la tecnica della selezione sistematica già illustrata.

6.2 La strategia di selezione bilanciata

Per la selezione bilanciata dei Comuni della regione Emilia Romagna è stata considerata come popolazione di riferimento soltanto quella costituita da tutti i Comuni con una popolazione inferiore ai 50.000 abitanti (Comuni NAR), mentre come vincoli di bilanciamento si è fatto riferimento al totale della popolazione delle province.

Nella selezione bilanciata delle sezioni di censimento, nei Comuni della regione Emilia Romagna con una popolazione superiore ai 50.000 abitanti (Comuni AR), per la definizione dei vincoli di bilanciamento si è tenuto conto del totale della popolazione delle sezioni distinte in due tipologie di località (centro abitato e altre località).

I risultati della selezione bilanciata dei Comuni e delle sezioni di censimento sono valutati tramite l'errore di bilanciamento (8). Nelle Tavole 19 e 20, che fanno riferimento rispettivamente ai risultati della selezione dei cinque gruppi casuali di Comuni NAR e dei cinque gruppi casuali di sezioni di censimento nei 13 Comuni AR, si riportano i valori minimo e massimo dell'errore di bilanciamento per ogni variabile ausiliaria utilizzata come vincolo.

I risultati della selezione bilanciata dei Comuni mettono in evidenza che la strategia adottata non fornisce risultati esaltanti in quanto l'errore di bilanciamento è molto elevato (le differenze tra totali noti e totali stimati sulle stesse variabili devono essere prossimi allo zero).

3 Per ciascuna località: Numero di sezioni campione =target di famiglie / (numero di famiglie /numero di sezioni).

Tavola 19 - Minimo e massimo dello scarto relativo tra totali noti e stime del totale (HT) della popolazione per le variabili di bilanciamento (N. di individui nelle province)

Provincia	Errore di bilanciamento %	
	Min	Max
Piacenza	4,53	22,64
Parma	2,65	28,5
Reggio nell'Emilia	0,41	13,47
Modena	5,34	17,93
Bologna	0,87	5,69
Ferrara	3,04	57,44
Ravenna	16,18	51,71
Forlì-Cesena	0,18	52,78
Rimini	1,15	28,26

Come si evince dalla tabella 19 soltanto il vincolo costituito dal totale della popolazione della provincia di Bologna conduce ad un errore di bilanciamento più contenuto. La spiegazione di tale risultato è sicuramente legato al fatto che la provincia di Bologna ha un numero di Comuni più elevato e più omogenei per dimensione demografica. I risultati peggiori si evidenziano quando si utilizzano come variabili di bilanciamento i totali di popolazione delle province di Ferrara, Ravenna e Forlì-Cesena che hanno un numero di Comuni inferiore a 30 unità.

Tavola 20 - Minimo e massimo dello scarto relativo tra totali noti e stime del totale (HT) della popolazione per le variabili di bilanciamento (N. di individui nelle sezioni di censimento per tipo di località)

Tipo di località	Errore di bilanciamento %	
	Min	Max
Centro abitato	0,003	2,263
Altro	0,120	22,629

I risultati della selezione bilanciata delle sezioni di censimento (Tavola 20) mettono in evidenza che la strategia adottata fornisce risultati buoni sulla prima variabile di bilanciamento, ovvero il numero di individui delle sezioni di censimento definite come località centro abitato, ma non per la seconda variabile di bilanciamento (numero di individui delle sezioni definite come altre località). In quest'ultimo caso il valore massimo dell'errore di bilanciamento è alquanto elevato (22,629 %).

Per quanto riguarda la prima variabile di bilanciamento, il valore maggiore dell'errore di bilanciamento si ha per la selezione delle sezioni del Comune Reggio nell'Emilia, mentre per la seconda variabile di bilanciamento, il valore più elevato dell'errore di bilanciamento si ha per la selezione delle sezioni del Comune di Carpi. Anche in questo caso la spiegazione sta nel fatto che le unità definite come altre località sono generalmente poche e soprattutto in alcuni Comuni.

6.3 Analisi dei risultati

Nelle tabelle che seguono si riportano alcune analisi descrittive dei campioni di sezioni estratti secondo le diverse impostazioni. Le Tavole 21 e 22 sono riferite alla selezione bilanciata dei Comuni NAR (335) e alla selezione bilanciata delle sezioni nei Comuni AR (13). Le successive tabelle fanno riferimento ai cinque gruppi casuali di Comuni e sezioni di censimento ottenuti con i disegni di campionamento stratificati.

Tavola 21 – Distribuzione del numero di sezioni e famiglie per gruppo casuale e tipologia di Comune - campionamento bilanciato

Gruppo casuale	Comuni NAR					Comuni AR			
	N. Comuni NAR	N. sezioni		N. famiglie		N. sezioni		N. famiglie	
		Universo	Campione	Universo	Campione	Universo	Campione	Universo	Campione
1	67	3.683	461	216.441	22.705	3.600	691	168.731	31.040
2	67	3.687	495	223.871	28.664	3.594	688	168.921	31.506
3	66	3.365	496	208.039	28.782	3.593	686	168.299	31.423
4	69	3.684	569	217.730	28.627	3.600	686	168.513	31.130
5	66	3.125	487	210.037	30.080	3.602	688	168.213	31.081
Totale	*335*	*17.544*	*2.508*	*1.076.118*	*138.858*	*17.989*	*3.439*	*842.677*	*156.180*

Tavola 22 – Distribuzione del numero di sezioni e famiglie per gruppo casuale nell'universo e nel campione e Tasso di campionamento (valori percentuali) - campionamento bilanciato

Gruppo casuale	Universo		Campione		Tasso di campionamento (%)	
	N. sezioni	N. famiglie	N. sezioni	N. famiglie	Sezioni	Famiglie
1	7.283	385.172	1.152	53.745	15,8	14,0
2	7.281	392.792	1.183	60.170	16,2	15,3
3	6.958	376.338	1.182	60.205	17,0	16,0
4	7.284	386.243	1.255	59.757	17,2	15,5
5	6.727	378.250	1.175	61.161	17,5	16,2
Totale	*35.533*	*1.918.795*	*5.947*	*295.038*	*16,7*	*15,4*

Tavola 23 – Distribuzione del numero di sezioni e famiglie per gruppo casuale e tipologia di Comune - campionamento stratificato A

Gruppo casuale	Comuni Nar					Comuni AR			
	N. Comuni NAR	N. sezioni		N. famiglie		N. sezioni		N. famiglie	
		Universo	Campione	Universo	Campione	Universo	Campione	Universo	Campione
1	66	3.306	479	207.869	28.254	3.599	697	165.272	30.729
2	67	3.298	489	210.812	27.464	3.600	671	171.937	30.189
3	68	3.589	502	220.539	28.732	3.597	709	165.080	30.979
4	69	3.832	542	229.093	29.282	3.599	683	169.078	30.685
5	65	3.519	499	207.805	26.985	3.594	665	171.310	30.779
Totale		17.544	2.511	1.076.118	140.717	17.989	3.425	842.677	153.361

Tavola 24 – Distribuzione del numero di sezioni e famiglie per gruppo casuale nell'universo e nel campione e Tasso di campionamento (valori percentuali) - campionamento stratificato A

Gruppo casuale	Universo		Campione		Tasso di campionamento (%)	
	N. sezioni	N. famiglie	N. sezioni	N. famiglie	Sezioni	Famiglie
1	6.905	373.141	1.176	58.983	17,0	15,8
2	6.898	382.749	1.160	57.653	16,8	15,1
3	7.186	385.619	1.211	59.711	16,9	15,5
4	7.431	398.171	1.225	59.967	16,5	15,1
5	7.113	379.115	1.164	57.764	16,4	15,2
Totale	*35.533*	*1.918.795*	*5.936*	*294.078*	*16,7*	*15,3*

(a) * 38.578 sono tutte le sezioni di censimento, dalle quali sono state eliminate quelle con 0 individui censiti.

Tavola 25 – Distribuzione del numero di sezioni e famiglie per gruppo casuale e tipologia di Comune - campionamento stratificato B

Gruppo casuale	Comuni Nar					Comuni AR			
	N. Comuni NAR	N. sezioni		N. famiglie		N. sezioni		N. famiglie	
		Universo	Campione	Universo	Campione	Universo	Campione	Universo	Campione
1	66	3.641	471	223.560	26.558	3.600	683	170.404	30.615
2	66	3.476	488	207.750	26.869	3.599	693	166.838	31.240
3	65	3.479	505	201.982	27.308	3.596	720	162.506	31.644
4	67	3.313	496	203.850	27.333	3.598	671	174.270	31.837
5	71	3.635	531	238.976	30.125	3.596	677	168.659	31.051
Totale		*17.544*	*2.491*	*1.076.118*	*138.193*	*17.989*	*3.444*	*842.677*	*156.387*

Fonte: Popolazione legale 2011

Tavola 26 – Distribuzione del numero di sezioni e famiglie per gruppo casuale nell'universo e nel campione e Tasso di campionamento (valori percentuali) - campionamento stratificato B

Gruppo casuale	Universo		Campione		Tasso di campionamento (%)	
	N. sezioni	N. famiglie	N. sezioni	N. famiglie	Sezioni	Famiglie
1	7.241	393.964	1.154	57.173	15,9	14,5
2	7.075	374.588	1.181	58.109	16,7	15,5
3	7.075	364.488	1.225	58.952	17,3	16,2
4	6.911	378.120	1.167	59.170	16,9	15,6
5	7.231	407.643	1.208	61.176	16,7	15,0
Totale	*35.533*	*1.918.803*	*5.935*	*294.580*	*16,7*	*15,4*

Fonte: Popolazione legale 2011
(a) * 38.578 sono tutte le sezioni di censimento, dalle quali sono state eliminate quelle con 0 individui censiti.

I tre disegni considerati includono nel campione circa 5.950 sezioni (16,5% circa del totale) per meno di 300.000 famiglie attese (15% circa del totale) nel ciclo completo di rilevazione.

La valutazione sull'efficienza dei disegni di campionamento è effettuata sulla base del confronto tra le stime del conteggio di popolazione (Tavola 28) ottenute a livello di singolo Comune - correggendo il numero di individui presenti in anagrafe (Lac) per la stima di individui recuperati (sottocopertura) e individui irreperibili al censimento (sovracopertura) -.

Nella Tavola 27 è riportata la popolazione legale nelle nove province della regione Emilia Romagna, variabile utilizzata per la valutazione dell'errore delle stime. Dalla Tavola 28, osservando i valori assunti dell'errore relativo percentuale – calcolato come differenza tra il valore vero della popolaone legale 2011 e il valore stimato della popolazione - si evidenzia che i campioni di sezioni selezionati dai cinque gruppi casuali ottenuti con i disegni di campionamento stratificati risultano produrre delle stime più efficienti, in particolare quando si adotta una stratificazione più fine. Ciò suggerisce che in generale, il campionamento bilanciato non è adatto quando si vuole effettuare la selezione in suddivisioni molto fini del territorio, in particolare nelle province.

Tavola 27 – Popolazione legale 2011 nelle province della Regione Emilia Romagna

Provincia	Numero legale di individui
Piacenza	284.616
Parma	427.435
Reggio nell'Emilia	517.317
Modena	685.778
Bologna	976.243
Ferrara	353.482
Ravenna	384.761
Forlì Cesena	390.738
Rimini	321.769
Totale	4.342.139

Tavola 28 – Stima del conteggio di popolazione nelle tre strategie di campionamento per provincia

		Stima del conteggio degli individui			Errore relativo %		
Gruppo casuale	Provincia	Bilanciato	Stratificato A	Stratificato B	Bilanciato	Stratificato A	Stratificato B
1	Piacenza	249.959	257.395	261.001	-12,2	-9,6	-8,3
	Parma	500.184	473.050	426.693	17,0	10,7	-0,2
	Reggio nell'Emilia	481.386	474.194	544.486	-6,9	-8,3	5,3
	Modena	622.074	620.393	689.422	-9,3	-9,5	0,5
	Bologna	927.172	953.193	1.013.844	-5,0	-2,4	3,9
	Ferrara	375.434	373.230	340.624	6,2	5,6	-3,6
	Ravenna	364.064	375.344	335.926	-5,4	-2,4	-12,7
	Forlì Cesena	340.369	358.382	375.175	-12,9	-8,3	-4,0
	Rimini	292.178	297.197	299.838	-9,2	-7,6	-6,8
	Totale	4.152.819	4.182.380	4.287.009	-4,4	-3,7	-1,3
2	Piacenza	264.231	269.704	271.246	-7,2	-5,2	-4,7
	Parma	381.757	385.476	431.730	-10,7	-9,8	1,0
	Reggio nell'Emilia	517.492	509.115	524.017	0,0	-1,6	1,3
	Modena	629.838	629.613	649.446	-8,2	-8,2	-5,3
	Bologna	972.343	959.223	934.136	-0,4	-1,7	-4,3
	Ferrara	441.388	439.660	384.401	24,9	24,4	8,7
	Ravenna	432.612	448.875	340.363	12,4	16,7	-11,5
	Forlì Cesena	383.765	371.875	368.827	-1,8	-4,8	-5,6
	Rimini	311.004	317.748	294.861	-3,3	-1,2	-8,4
	Totale	4.334.430	4.331.289	4.199.026	-0,2	-0,2	-3,3
3	Piacenza	271.159	277.167	293.836	-4,7	-2,6	3,2
	Parma	393.083	398.918	448.445	-8,0	-6,7	4,9
	Reggio nell'Emilia	562.262	550.196	499.875	8,7	6,4	-3,4
	Modena	681.575	685.458	660.556	-0,6	0,0	-3,7
	Bologna	1.020.405	999.354	952.357	4,5	2,4	-2,4
	Ferrara	294.538	295.285	317.263	-16,7	-16,5	-10,2
	Ravenna	475.924	420.460	381.818	23,7	9,3	-0,8
	Forlì Cesena	412.420	397.387	421.736	5,5	1,7	7,9
	Rimini	330.864	338.688	310.892	2,8	5,3	-3,4
	Totale	4.442.229	4.362.914	4.286.777	2,3	0,5	-1,3
4	Piacenza	301.403	290.356	285.668	5,9	2,0	0,4
	Parma	412.251	416.112	388.279	-3,6	-2,6	-9,2
	Reggio nell'Emilia	566.124	560.722	490.316	9,4	8,4	-5,2
	Modena	721.966	719.083	705.684	5,3	4,9	2,9
	Bologna	1.061.385	1.037.634	961.047	8,7	6,3	-1,6
	Ferrara	308.142	309.365	337.242	-12,8	-12,5	-4,6
	Ravenna	323.843	331.333	462.934	-15,8	-13,9	20,3
	Forlì Cesena	445.335	432.784	371.527	14,0	10,8	-4,9
	Rimini	413.555	376.639	331.400	28,5	17,1	3,0
	Totale	4.554.003	4.474.028	4.334.097	4,9	3,0	-0,2
5	Piacenza	319.057	305.880	283.950	12,1	7,5	-0,2
	Parma	445.289	453.515	412.982	4,2	6,1	-3,4
	Reggio nell'Emilia	458.712	489.239	514.749	-11,3	-5,4	-0,5
	Modena	771.836	774.343	721.808	12,5	12,9	5,3
	Bologna	899.349	925.675	988.146	-7,9	-5,2	1,2
	Ferrara	347.395	349.870	370.879	-1,7	-1,0	4,9
	Ravenna	329.262	335.893	387.681	-14,4	-12,7	0,8
	Forlì Cesena	343.583	359.183	363.078	-12,1	-8,1	-7,1
	Rimini	263.261	267.599	350.958	-18,2	-16,8	9,1
	Totale	4.177.745	4.261.197	4.394.231	-3,8	-1,9	1,2

7. Conclusioni

Le sperimentazioni condotte per la selezione bilanciata di cinque gruppi casuali costituiti dai Comuni mettono in evidenza risultati interessanti che ci permettono di delimitare in qualche modo le situazioni in cui l'utilizzo di un disegno di campionamento bilanciato non produce buoni risultati. È evidente che se si prende in considerazione la selezione dei Comuni di dimensione compresa tra 5.000-50.000 abitanti si hanno sempre pessime performance a prescindere dai vincoli di bilanciamento utilizzati.

E' chiaro che i vincoli di bilanciamento che definiscono strutture di popolazione (come quelle demografiche) devono essere studiate ad hoc e riflettere le caratteristiche proprie di una regione. In questo caso potrebbero essere utili le analisi delle piramidi dell'età per regione (relativamente ai Comuni sotto la soglia di 50.000) al fine di individuare regioni con strutture di popolazione simili.

Quando si introducono vincoli sul totale della popolazione delle province la qualità del bilanciamento peggiora considerevolmente tanto da far ritenere che un disegno di campionamento bilanciato non sia in tal caso da considerare.

L'altro aspetto importante che emerge dalle sperimentazioni è il risultato della selezione bilanciata di Comuni e agglomerati di Comuni di piccole dimensioni. E' vero che le performance del disegno di campionamento non migliorano (come già detto a causa della diminuzione delle unità di riferimento determinata dall'aggregazione di Comuni), ma se si riesce a costruire agglomerati in modo da abbattere notevolmente la variabilità delle dimensioni delle unità costituenti tali insiemi, allora si possono ottenere risultati migliori.

Un discorso a parte va fatto sulla selezione bilanciata delle sezioni di censimento per i Comuni di grandi dimensioni, dove i primi, anche se pochi, risultati delle sperimentazioni mettono in evidenza soprattutto l'inefficienza della selezione vincolata al totale di popolazione delle sezioni per tipo di località (dato il numero ridotto di sezioni di censimento non classificate come centro abitato). Inoltre, se si considera una suddivisione della popolazione per strati di sezioni è necessario che, anche in questo caso, sia definita una stratificazione ottimale basata sulle distribuzioni delle sezioni dei singoli comuni.

D'altra parte la minore efficienza del disegno di campionamento bilanciato risulta anche dall'analisi effettuata sui campioni di sezioni selezionati nella regione Emilia Romagna con disegni alternativi, anche se si tratta di un'analisi molto parziale che sarebbe opportuno replicare in più regioni e con più campioni.

La spiegazione di questi risultati è comunque individuabile nelle criticità di una selezione contemporanea di campioni bilanciati sottoposta a vincoli troppo stringenti, come quelli posti dalla rappresentatività della popolazione a livello di provincia, e a un numero non sufficiente di unità negli universi di riferimento. Quest'ultimo aspetto non ha costituito un limite all'applicazione del metodo alla progettazione del Censimento della popolazione in Francia, Paese che presenta una suddivisione amministrativa del territorio molto diversa dalla nostra. In questo caso, infatti, la selezione bilanciata di Comuni con meno di 10.000 abitanti è stata realizzata su un numero molto elevato di Comuni (30.000) mentre per i Comuni con più di 10.000 abitanti la selezione bilanciata è stata effettuata sugli indirizzi.

Riferimenti bibliografici

Chauvet G. e Tillé Y. 2007. *Application of Fast Sas Macros for Balancing Samples to the Selection of Addresses.* http://www.bentley.edu/cdgigs/vol1-2/chauvet.pdf

Deville J.C. e Tillé Y. 2004. Efficient balanced sampling: The cube method. *Biometrika*, 91, 4: 893-912.

Deville J.C.e Tillé Y. 1998. Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Durr J. M. e Grosbras J. M. 2010. *La renovation du recensement francais: principes et méthode.*

Durr J. M. e Clanché F. 2010. *The French Rolling Census: a decade of experience.*

Fortini M. 2012. Presentazione *Il censimento continuo della popolazione e delle abitazioni*.

Fortini M. 2013. Presentazione *Il censimento permanente: informazioni di contesto e obiettivi del gruppo di lavoro*.

INSEE - *Le recensement général de la population*. -http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/doc/presentation-recensement.htm

Tillé Y. 2011. Ten years of balanced sampling with the cube method: An appraisal. *Survey methodology*, December 2011: 215-226.

Tillé Y. (2006). Sampling Algorithms – Springer Series in Statistics. New York, USA 2006: Springer Science Business Media.

Tillé Y. e Favre A.C. 2004. Coordination, Combination and Extension of Balanced Samples. *Biometrika.* Vol. 91, N. 4, p. 913-927.