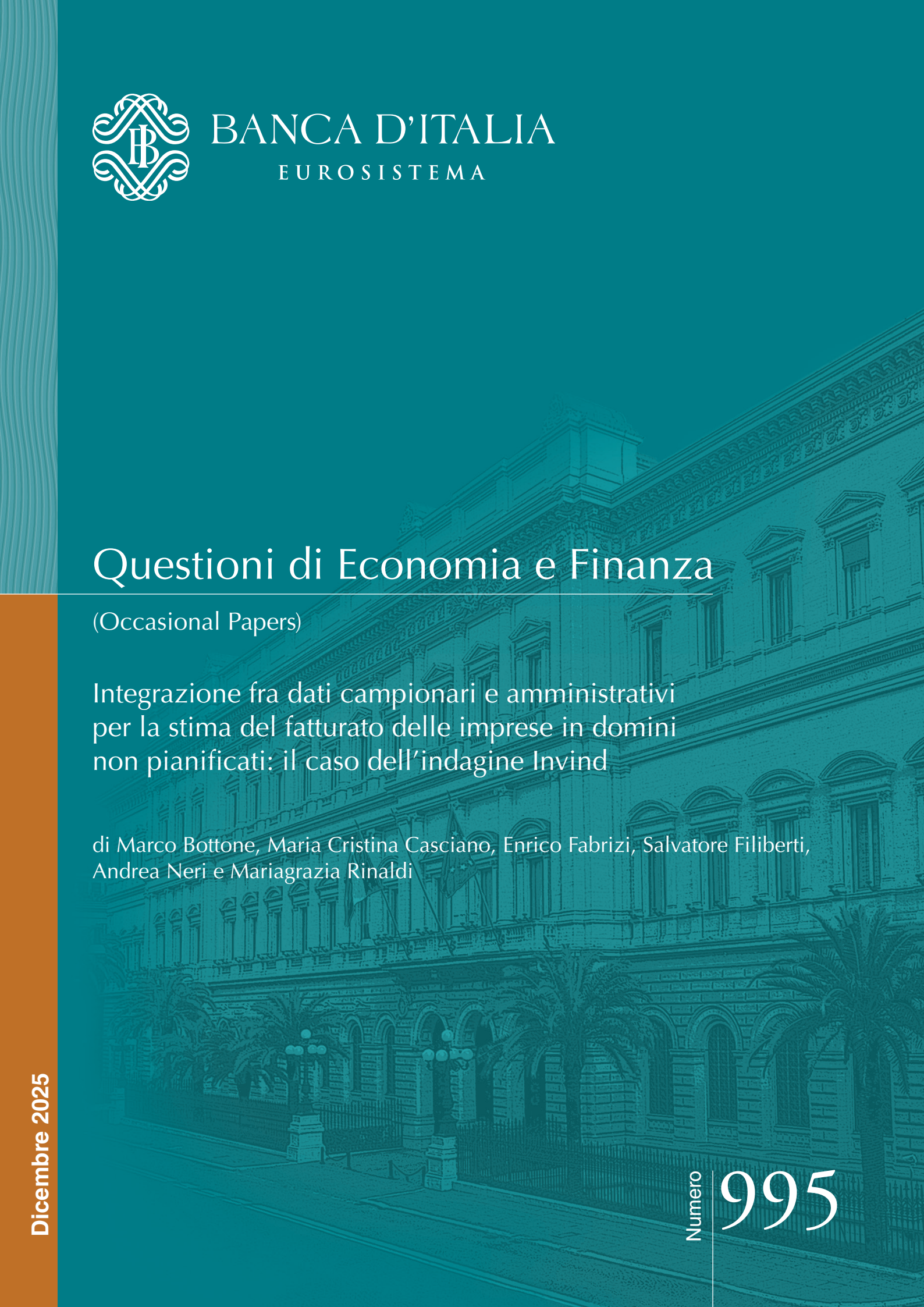
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

Integrazione fra dati campionari e amministrativi per la stima del fatturato delle imprese in domini non pianificati: il caso dell'indagine Invind

di Marco Bottone, Maria Cristina Casciano, Enrico Fabrizi, Salvatore Filiberti, Andrea Neri e Mariagrazia Rinaldi

Dicembre 2025

Numero 995

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Integrazione fra dati campionari e amministrativi per la stima del fatturato delle imprese in domini non pianificati: il caso dell'indagine Invind

di Marco Bottone, Maria Cristina Casciano, Enrico Fabrizi, Salvatore Filiberti, Andrea Neri e Mariagrazia Rinaldi

Numero 995 – Dicembre 2025

La serie Questioni di economia e finanza *ha la finalità di presentare studi e documentazione su aspetti rilevanti per i compiti istituzionali della Banca d'Italia e dell'Eurosistema. Le* Questioni di economia e finanza *si affiancano ai* Temi di discussione *volti a fornire contributi originali per la ricerca economica.*

La serie comprende lavori realizzati all'interno della Banca, talvolta in collaborazione con l'Eurosistema o con altre Istituzioni. I lavori pubblicati riflettono esclusivamente le opinioni degli autori, senza impegnare la responsabilità delle Istituzioni di appartenenza.

La serie è disponibile online sul sito www.bancaditalia.it .

ISSN 1972-6643 (online)

Grafica a cura della Divisione Editoria e stampa della Banca d'Italia

INTEGRAZIONE FRA DATI CAMPIONARI E AMMINISTRATIVI PER LA STIMA DEL FATTURATO DELLE IMPRESE IN DOMINI NON PIANIFICATI: IL CASO DELL'INDAGINE INVIND

di Marco Bottone[*], Maria Cristina Casciano[†], Enrico Fabrizi[‡], Salvatore Filiberti[†], Andrea Neri[*] e Mariagrazia Rinaldi[†]

Sommario

Il lavoro propone una metodologia che integra l'indagine Invind della Banca d'Italia con l'archivio Frame-SBS dell'Istat per la stima della variazione del fatturato delle imprese in domini non pianificati, ossia a un livello di dettaglio superiore a quello progettato in fase di disegno (ad esempio, per settori economici su scala regionale). L'integrazione fra i due archivi consente una significativa riduzione dell'errore quadratico medio (MSE) degli stimatori, grazie soprattutto alla forte autocorrelazione temporale della variabile analizzata. Oltre alla fase di stima, l'archivio Frame-SBS può essere impiegato nella progettazione del campione. Le analisi condotte mostrano che per ottenere stime affidabili nei domini a oggi non pianificati, sarebbe necessario incrementare la numerosità del campione nei servizi e in specifiche regioni, in particolare la Lombardia e il Lazio.

Classificazione JEL: C13; C81; C83.

Parole chiave: integrazione dati campionari e amministrativi, Stima su domini non pianificati, analisi economica regionale.

DOI: 10.32057/0.QEF.2025.995

* Banca d'Italia - Dipartimento di Economia e Statistica.

† Istat.

‡Università Cattolica del Sacro Cuore.

Le opinioni espresse sono quelle degli autori e non riflettono necessariamente quelle degli Istituti di cui fanno parte. Si ringraziano tutti i partecipanti del seminario sull'uso di Invind per l'analisi economica regionale per gli utili commenti e suggerimenti offerti.

1 Introduzione

Le indagini sulle imprese costituiscono una fonte informativa essenziale per l'analisi e la progettazione delle politiche economiche. Attraverso la raccolta diretta di dati su attività produttiva, investimenti, aspettative e performance aziendali, esse permettono di monitorare in modo tempestivo l'evoluzione del ciclo economico, rilevare gli effetti di shock esogeni e valutare l'impatto degli interventi pubblici.

Per garantire tempestività e contenere i costi di rilevazione, queste indagini si basano spesso su campioni di dimensioni limitate. Tale scelta, pur efficace a livello nazionale o per aggregati ampi, comporta la difficoltà a produrre stime statisticamente affidabili per domini più dettagliati, come singole regioni, settori specifici o classi dimensionali ristrette di imprese. Ne deriva una crescente esigenza di valorizzare al massimo i dati campionari disponibili attraverso tecniche che consentano di estendere la portata informativa delle indagini, fornendo stime accurate anche per *domini non pianificati*, ossia aggregati per i quali il disegno campionario originario non garantisce rappresentatività.

Il presente lavoro si propone di valutare i benefici derivanti dall'integrazione tra dati campionari e fonti amministrative per la produzione di stime affidabili su domini non pianificati. L'analisi si fonda sull'Indagine sulle imprese industriali e dei servizi (Invind), realizzata annualmente dalla Banca d'Italia. L'indagine è progettata per fornire stime affidabili a livello di macro-area, settore economico (ATECO a due cifre) e classe dimensionale, ossia nei cosiddetti *domini pianificati*. Tuttavia, essa non è strutturata per garantire la rappresentatività statistica a livello regionale o per altri sottogruppi più granulari di imprese.

L'applicazione empirica si concentra sulla stima della variazione del fatturato delle imprese su scala regionale, una variabile di grande rilevanza per cogliere tempestivamente l'andamento dell'attività economica a livello locale. La scelta di questo indicatore riflette la finalità originaria dell'indagine Invind, pensata per fornire informazioni rapide in un contesto in cui i conti nazionali erano disponibili con ritardi significativi.

Per migliorare l'accuratezza delle stime a livello regionale, si utilizza l'archivio Frame-SBS dell'Istat, che raccoglie dati amministrativi dettagliati sulla struttura produttiva delle imprese italiane. In particolare, il lavoro confronta due strategie complementari:

1. Un approccio ex ante, che sfrutta le informazioni contenute in Frame-SBS per costruire un disegno campionario stratificato più efficiente, mirato a garantire una migliore copertura dei domini regionali;

2. Un approccio ex post, che integra i dati amministrativi con quelli campionari in fase di stima, tramite modelli statistici in grado di "trasferire" informazione da domini ben rappresentati a quelli scarsamente coperti.

Entrambe le strategie fanno leva sull'informazione ausiliaria fornita dal sistema Frame-SBS con l'obiettivo comune di aumentare la precisione e la robustezza delle stime subnazionali, rispondendo così alla domanda crescente di indicatori territoriali affidabili.

2 Letteratura

Il tema della stima su domini non pianificati, ossia ambiti territoriali o settoriali per cui l'indagine campionaria non è progettata per garantire rappresentatività statistica, è ampiamente trattato nella letteratura. In tali casi, le dimensioni del campione disponibile possono risultare troppo contenute, o addirittura nulle, rendendo le stime dirette instabili o impossibili. Due grandi famiglie di approcci sono stati sviluppati per affrontare questa sfida: gli approcci diretti e quelli indiretti.

Gli approcci diretti si basano esclusivamente sui dati campionari raccolti all'interno del dominio di interesse. In pratica, si limita l'analisi ai soli dati disponibili per quella regione, settore o gruppo sociale. Questo metodo è concettualmente semplice, non richiede assunzioni aggiuntive e riflette esattamente le osservazioni empiriche. Tuttavia, la sua principale debolezza è la scarsa affidabilità delle stime in presenza di pochi dati: al diminuire della numerosità campionaria, aumenta infatti l'incertezza statistica, rendendo le stime molto variabili e poco utilizzabili per fini decisionali.

Gli approcci indiretti, invece, utilizzano anche informazioni esterne al dominio specifico per migliorare la qualità delle stime. L'idea è quella di "prestare forza informativa" da altri domini o da fonti ausiliarie (come archivi amministrativi o dati di registro) attraverso modelli statistici che stabiliscono relazioni tra le variabili. Per esempio, se una certa variabile (come il fatturato medio delle imprese) è ben conosciuta a livello nazionale o in altre regioni simili, è possibile sfruttare questa conoscenza per produrre stime più affidabili anche in regioni con campionamento scarso o nullo.

Questo approccio, pur più complesso da implementare, consente di ottenere stime più precise e robuste, e rappresenta la base della Small Area Estimation (SAE),una branca della statistica sempre più centrale nello sviluppo di indicatori granulari e tempestivi.

Un contributo fondamentale in questo ambito è rappresentato dal lavoro di Fay e Herriot (1979), che hanno introdotto il modello di base per la stima del reddito su piccole aree. Successivamente, Petrucci, Pratesi e Salvati (2005) hanno ampliato questo approccio includendo effetti spaziali, dimostrando che la correlazione tra aree geograficamente contigue può migliorare significativamente la precisione degli stimatori.

Diversi studi applicati hanno ulteriormente approfondito l'efficacia di questi modelli in contesti particolari. Chandra e Chambers (2011) e Karlberg (2014) si sono concentrati su variabili caratterizzate da distribuzioni fortemente asimmetriche o con numerosi valori nulli, proponendo

soluzioni innovative per affrontare queste sfide. D'Alò, Falorsi e Solari (2017) hanno esteso i modelli SAE a contesti spazio-temporali, dimostrando l'efficacia di tali approcci nella gestione di grandi dataset e nel miglioramento della precisione delle stime. Un ulteriore contributo rilevante è fornito da Luzi, Monducci, Righi e Vacca (2018), che hanno applicato il modello Fay-Herriot per stimare variabili economiche rilevanti, come l'ammortamento, utilizzando l'archivio Frame-SBS dell'Istat. I risultati di tale studio hanno evidenziato come l'integrazione tra dati amministrativi e campionari consenta di ridurre significativamente l'errore quadratico medio (MSE), migliorando la qualità delle stime per piccole aree.

Anche l'indagine Invind è stata utilizzata per sviluppare stimatori a livello locale. Cesari e Signorini (1991) hanno analizzato stimatori indiretti per variabili come investimenti, fatturato e occupazione, mostrando miglioramenti nella precisione delle stime, pur evidenziando errori standard elevati per regioni piccole o variabili particolarmente volatili. D'Alessio e Faiella (2001) hanno invece applicato stimatori indiretti per valutare la variazione degli investimenti a livello regionale, riscontrando un aumento di precisione rispetto agli stimatori diretti ma sottolineando la necessità di ampliare il campione per migliorare ulteriormente l'affidabilità delle stime.

Il nostro lavoro si inserisce nell'ambito degli stimatori indiretti e si allinea con lo studio di Luzi et al. (2018), condividendo l'approccio di integrazione tra dati campionari e l'archivio Frame-SBS. Tuttavia, il presente lavoro si distingue per due aspetti chiave che rappresentano il suo valore aggiunto. In primo luogo, vengono valutate le prestazioni di diversi stimatori, sfruttando in maniera più approfondita le potenzialità offerte dall'archivio Frame-SBS, mentre Luzi et al. si concentrano sull'applicazione di un singolo modello. In secondo luogo, il nostro studio affronta sia il miglioramento della precisione delle stime ex-post sia l'ottimizzazione del disegno campionario ex-ante, un aspetto non trattato nello studio precedente.

Il lavoro è strutturato come segue. Il paragrafo 3 descrive le principali caratteristiche delle basi dati utilizzate; il paragrafo 4 discute le difficoltà metodologiche legate all'utilizzo dell'indagine Invind per stime regionali; i paragrafi 5 e 6 illustrano gli stimatori disegnati per piccole aree; il paragrafo 7 analizza l'uso del Frame-SBS nella fase di disegno; infine, il paragrafo 8 sintetizza i principali risultati emersi.

3 Dati

3.1 L'indagine Invind

L'indagine sugli investimenti delle imprese industriali e dei servizi (Invind), è condotta dal 1973 dalla Banca d'Italia. La rilevazione è svolta nel periodo febbraio-maggio dalla Rete territoriale dell'Istituto presso un campione, in larga parte longitudinale, di circa 4.500 imprese dell'industria e dei servizi con almeno 20 addetti (almeno 10 nelle costruzioni). Oltre alle informazioni anagrafiche, alle imprese vengono posti circa 300 quesiti, di natura prevalentemente

quantitativa; di questi, circa l'85 per cento è ricorrente (spesa per investimenti, occupazione, retribuzioni, fatturato, prezzi, risultato d'esercizio, finanziamento e crediti commerciali), il resto muta di anno in anno per soddisfare richieste conoscitive del momento. L'orizzonte temporale di riferimento dei quesiti è sia consuntivo, sia previsivo.

L'indagine è basata su un campione stratificato. Gli strati sono costituiti dalle combinazioni di settore di attività economica (classificazione ATECO), classe dimensionale (in termini di addetti) e sede amministrativa dell'impresa (Piemonte e Valle d'Aosta sono inseriti in un unico strato). Relativamente alle imprese con almeno 5.000 addetti (meno di 100 nella popolazione) l'indagine è censuaria. L'unità di rilevazione è la sede amministrativa, di conseguenza nel caso di imprese con stabilimenti in regioni diverse, le grandezze economiche rilevate (come ad esempio fatturato, addetti e investimenti) vengono attribuite all'unica regione in cui è presente tale sede.

L'universo di riferimento dell'indagine, secondo l'Archivio Statistico sulle Imprese Attive (Asia), è costituito da circa 60.000 unità. [1]

3.2 L'archivio Frame-SBS

Frame-SBS è un sistema integrato di dati amministrativi e statistici, realizzato annualmente dall'Istat per la stima dei risultati economici delle imprese, integrando tutte le informazioni economiche disponibili relative a ognuna delle imprese attive operanti nel territorio nazionale. Tali informazioni sono elaborate attraverso opportune metodologie statistiche e riguardano i risultati economici (fatturato, costo di acquisto di beni e servizi, costo del personale, valore aggiunto) e le caratteristiche delle imprese (attività economica, localizzazione territoriale, numero di addetti e dipendenti). I risultati economici a livello di singola impresa sono ricostruiti attraverso un complesso processo di stima applicato ai dati individuali provenienti da più fonti amministrative, come Camere di commercio, Agenzia delle Entrate e Inps. Per le imprese con meno di 100 addetti (circa 4,3 milioni di unità) il processo combina le fonti amministrative con i risultati della Rilevazione (campionaria) sulle piccole e medie imprese e sull'esercizio di arti e professioni (PMI). Le informazioni relative alle imprese con 100 addetti ed oltre (circa 10.500 unità) derivano invece dalla Rilevazione (censuaria) sul sistema dei conti delle imprese (SCI). Nel Frame-SBS sono presenti, inoltre, informazioni che permettono la classificazione delle unità economiche secondo molteplici criteri, come ad esempio l'appartenenza a gruppi di imprese e/o lo svolgimento di attività di commercio con l'estero. Per le finalità del presente lavoro, dall'archivio Frame-SBS sono state selezionate le imprese che risultavano attive al 2015 con almeno 20 addetti e che operano nei settori di attività oggetto di rilevazione nell'indagine Invind della

[1] Non disponendo dei microdati Asia, la lista di imprese da rilevare è costruita sulla base di archivi di fonte INPS, Infocamere e altre liste reperite autonomamente dalle Filiali allo scopo di ridurre al minimo il rischio di copertura incompleta della popolazione.

Banca d'Italia. Tali imprese rappresentano la popolazione di riferimento per il 2015 e sono pari a 62.259 unità.

Grazie ad un accordo di collaborazione siglato fra Istat e Banca d'Italia è stato possibile, nel pieno rispetto della riservatezza delle informazioni fornite dalle imprese, agganciare il campione di Invind con le informazioni dell'archivio Frame-SBS. Questo aggancio ha permesso di valutare l'esistenza di possibili incongruenze fra le due fonti su due variabili chiave che sono utilizzate nelle analisi seguenti: la dimensione aziendale e il settore di attività (codice ATECO). Infatti, sebbene la dimensione aziendale sia definita in entrambe le fonti sulla base del numero di addetti medi in un determinato anno, in Frame-SBS tale valore è basato sull'archivio Asia, mentre nell'indagine Invind esso riflette la risposta fornita dall'impresa. Allo stesso modo, mentre in Frame-SBS la classificazione settoriale è definita direttamente dall'Istat attraverso una pluralità di fonti informative, nell'indagine Invind il codice ATECO è quello desumibile dalla lista di campionamento e poi riconfermato dall'impresa.

L'aggancio è stato fatto per l'indagine sul 2015. Su un campione di 4.395 intervistate, 295 (il 7 per cento circa) risulterebbero essere fuori dalla popolazione di riferimento secondo le informazioni di Frame-SBS. In particolare, circa la metà apparterrebbe all'industria alimentare (non oggetto di rilevazione nell'indagine) e la restante parte avrebbe un numero di addetti inferiore a 20 secondo Frame-SBS.

Fra le imprese agganciate le differenze in termini di settore e classe dimensionale sono nel complesso limitate (tavole 1, 2). Il settore di attività raggruppato in 6 codici ATECO (classificazione rilevante anche ai fini del calcolo dei pesi campionari) risulta lo stesso nelle due fonti in circa il 94 per cento dei casi. Un simile risultato si osserva per la classe di addetti (6 modalità, 93,5 per cento). Inoltre, si osserva che gli errori si concentrano su classi contigue a quelle corrette.

Tabella 1: Numero di imprese per classe di occupati medi nel 2015: confronto fra Frame-SBS e Invind
(valori percentuali)

	Classe di addetti Frame-SBS						
Classe di addetti Invind	20-49	50-99	100-199	200-499	500-999	1.000 e +	Totale
20-49	31,1	2,1	0,0	0,0	0,0	0,0	33,2
50-99	1,1	20,9	0,9	0,1	0,0	0,0	22,9
100-199	0,0	0,6	16,0	0,4	0,0	0,0	17,0
200-499	0,0	0,0	0,6	14,7	0,2	0,1	15,5
500-999	0,0	0,0	0,0	0,2	5,4	0,1	5,6
1.000 e +	0,0	0,0	0,0	0,1	0,2	5,5	5,7
Totale imprese	1.321	966	716	628	237	232	4.100

Tabella 2: Numero di imprese per settore di attività nel 2015: confronto fra Frame-SBS e Invind *(valori percentuali)*

	Settore ATECO Frame-SBS						
Settore ATECO Invind	Manif.	Alt. ind.	Comme.	Albe.	Trasp.	Altri servizi	Totale
Manifattura	65,3	0,4	1,2	0,0	0,2	0,2	67,3
Altre industrie	0,1	4,1	0,0	0,0	0,0	0,1	4,3
Commercio	0,5	0,0	11,8	0,0	0,1	0,0	12,4
Alberghi	0,0	0,0	0,0	2,2	0,0	0,0	2,2
Trasporti	0,1	0,1	0,1	0,0	8,5	0,2	8,9
Altri servizi	0,1	0,1	0,1	0,0	0,3	4,3	4,9
Totale imprese	2.705	191	543	90	373	198	4.100

4 Le principali sfide metodologiche

La costruzione di stimatori per la variazione del fatturato su domini non pianificati tramite l'archivio integrato Invind/Frame-SBS, richiede di affrontare due principali sfide metodologiche. La prima riguarda la corretta attribuzione delle variabili rilevate alle singole regioni. Nell'indagine Invind i dati raccolti sul fatturato (e le altre grandezze economiche) sono infatti integralmente attribuiti alla regione in cui l'impresa ha la sede legale. Tuttavia, il decentramento delle attività produttive su più regioni, in assenza di informazioni dettagliate sulla localizzazione e sull'attività delle singole sedi, può portare a un'erronea assegnazione delle variabili al dominio di interesse.

La seconda sfida metodologica è relativa allo sviluppo di un sistema di pesi regionali nell'indagine Invind. Il sistema di ponderazione attualmente utilizzato, infatti, applica vincoli a livello di macroarea geografica e condurrebbe a stime distorte su un livello di disaggregazione territoriale più fine. Tale problematica è direttamente collegata alla limitata numerosità campionaria, che in molte regioni riduce la precisione degli stimatori, rendendoli meno affidabili per l'analisi economica.

Poiché la collaborazione scientifica con l'Istat prevede la fornitura del Frame-SBS 2015, le analisi empiriche nel presente lavoro si basano sul campione Invind rilevato nel 2016, contenente i dati di consuntivo del 2015 e le attese per il 2016. Sebbene siano oggi a disposizione dati più recenti dell'indagine, la scelta è giustificata dalla natura metodologica del presente lavoro. L'obiettivo principale, infatti, non è quello di analizzare le più recenti tendenze economiche, ma quello di sviluppare e testare stimatori su domini non pianificati che sfruttano l'integrazione delle diverse fonti di dati nel momento in cui le stesse sono disponibili.

Come fase preliminare all'analisi, i due archivi sono stati quanto più possibile armonizzati, per permetterne un utilizzo congiunto. In particolare, dall'archivio Frame-SBS sono state escluse le imprese che non rientrano nella popolazione di riferimento di Invind, ossia quelle con meno di 20 addetti o appartenenti a settori non inclusi nell'indagine. Parallelamente, dai dati Invind sono state rimosse 295 imprese che, secondo il Frame-SBS, non appartengono alla

popolazione di riferimento. Infine, con l'obiettivo di eliminare le discrepanze tra le due fonti e assicurare una maggiore coerenza nei risultati, sono stati sostituiti i dati relativi al numero di addetti e al settore di attività economica rilevati nell'indagine Invind con quelli presenti per la stessa impresa in Frame-SBS, creando quindi una classificazione settoriale e dimensionale identica tra le due fonti.

4.1 Struttura regionale dell'attività produttiva

La distribuzione territoriale dell'attività produttiva rappresenta un potenziale elemento di distorsione nelle analisi regionali, dal momento che tutte le informazioni raccolte nell'indagine vengono attribuite alla sede legale dell'impresa e non si dispone di dati sulla dislocazione e sul grado di attività dei singoli stabilimenti produttivi sparsi nelle diverse regioni. La tavola 3, relativamente sia al numero di imprese sia al numero di addetti, riporta due statistiche che aiutano a comprendere la composizione delle unità locali per regione. In particolare, nelle prime due colonne, fatto cento il totale di unità locali presenti in una certa regione, si mostra la quota di quelle afferenti ad imprese con sede principale in un'altra regione. Nelle ultime due colonne invece, fatto cento il numero di unità locali afferenti ad imprese con sede principale in una certa regione, si mostra la quota di quelle residenti nella stessa regione.

I due fenomeni descrivono da un lato la capacità di ogni regione di attrarre unità produttive dal resto d'Italia e dall'altro la propensione delle imprese con sede principale in quella regione a spostare parte della produzione altrove. Per entrambi i fenomeni, emerge una chiara distinzione fra le regioni del nord e quelle del centro sud con un'evidente eccezione rappresentata dal Lazio. In particolare, si osserva che regioni come l'Abruzzo, il Molise, la Basilicata, la Calabria e la Sardegna si caratterizzano per la capacità di attrarre molte unità locali dalle altre regioni d'Italia e per un tessuto produttivo con una bassa propensione alla delocalizzazione in altre regioni. Al contrario, regioni come il Piemonte, la Lombardia, il Veneto e l'Emilia Romagna accolgono relativamente meno imprese con sede principale fuori confine e spostano, al contempo, una consistente fetta della produzione.

In generale, nella stima dei totali a livello regionale di una qualsiasi variabile Y interessata dal fenomeno della distribuzione territoriale della produzione tra più regioni (ad esempio fatturato, occupazione o investimenti), la distorsione sarà direttamente proporzionale alla quota della variabile Y impiegata in altre regioni e inversamente proporzionale alla quota impiegata nella stessa regione della sede principale. Per correggere tale distorsione, sarebbero necessarie informazioni ausiliarie sulla distribuzione territoriale del fenomeno di interesse, o di una sua proxy attendibile. Nel presente lavoro tali informazioni non sono disponibili: qualsiasi tentativo di ricostruzione implicherebbe ipotesi forti e non verificabili sulla ripartizione delle variabili tra regioni, con il rischio di introdurre ulteriore incertezza nella stima.

Tabella 3: Composizione unità locali per regione

	% Unità locali con sede principale in altre regioni		% Unità locali nella stessa regione della sede principale	
	Imprese	Addetti	Imprese	Addetti
Piemonte	43	28	67	77
Valle d'Aosta	58	37	84	89
Lombardia	23	13	64	72
Bolzano	27	15	74	81
Trento	33	22	81	88
Veneto	31	20	69	83
Friuli-V. G.	50	29	74	81
Liguria	56	33	72	88
E.-Romagna	34	21	67	81
Toscana	40	29	76	89
Umbria	42	26	66	82
Marche	39	26	79	91
Lazio	35	24	34	54
Abruzzo	54	35	73	86
Molise	65	61	78	88
Campania	33	32	75	85
Puglia	38	38	86	88
Basilicata	62	41	83	96
Calabria	56	41	93	96
Sicilia	40	35	93	96
Sardegna	48	38	96	94

Fonte: Asia unità Locali 2014. La percentuale di unità locali con sede principale in altre regioni è calcolata come somma delle unità locali presenti nella regione X ma aventi sede principale in altre regioni fratto il totale unità locali presenti nella regione X. La percentuale di unità locali nella stessa regione della sede principale è calcolata come somma delle unità locali presenti nella regione X ed aventi sede principale nella stessa regione, fratto il totale unità locali che fanno riferimento a imprese con sede principale nella regione X.

Per queste ragioni, l'analisi si concentra esclusivamente sugli aspetti metodologici legati alla costruzione di sistemi di ponderazione per migliorare la rappresentatività regionale dell'indagine. La questione della regionalizzazione dell'attività produttiva, pur rilevante, esula dall'ambito di questo lavoro ed è affrontabile solo con strumenti e fonti informative specifici.

4.2 Sistema di ponderazione

La seconda difficoltà metodologica è legata alla creazione di un adeguato sistema di ponderazione. Essendo nata con l'obiettivo di fornire stime principalmente a livello nazionale e per macro area, il sistema di ponderazione dell'indagine Invind non include attualmente vincoli a livello di regione, ma soltanto per le quattro principali aree geografiche (Nord-Ovest, Nord-Est,

Centro, Sud e Isole).

In mancanza di un sistema di ponderazione contenente vincoli a livello regionale, anche una semplice stima del numero di imprese per regione può risultare distorta. La tavola 4 mostra infatti come nella quasi totalità delle regioni, il numero di imprese stimato sulla base dell'indagine sia significativamente diverso da quello di Frame-SBS. La stessa tavola mostra inoltre come tale distorsione risulti solo debolmente associata al settore di attività delle imprese. Ad esempio, in alcune regioni (Lombardia, Trentino A.A., Liguria, Puglia, Molise) il campione tende a sovra-rappresentare le imprese industriali mentre in altre (Toscana, Umbria, Marche, Abruzzo) sono le imprese dei servizi ad avere un peso superiore a quello teorico secondo Frame-SBS.

Tabella 4: Distribuzione numero di imprese per regione e settore stimate su Frame-SBS (F) e Invind (I) per l'anno 2015

Regione	*Settore*													
	N Totale		% Manif.		% Altre ind.		% Commercio		% Alberghi		% Trasporti		% Altri Ser.	
	F	I	F	I	F	I	F	I	F	I	F	I	F	I
Piemonte e V. D'A.	4.804	7.539	52,5	52,7	2,6	1,4	14,6	15,6	4,2	4,5	13,2	13,8	13	12
Lombardia	16.161	8.657	48,4	51,5	2	1,4	17	20,1	4,7	5,8	12,2	11	15,7	10,1
Trentino A.A	1.572	5.918	33,7	38,3	3,9	4,5	23,4	13,1	17,4	3,4	12,7	14,9	9	25,8
Veneto	7.917	2.114	59,5	57,6	1,6	8,9	15,1	16,4	6,7	13,8	9,2	2,2	7,9	1,2
Friuli V. G.	1.414	7.180	56,7	54,6	2,6	1,7	14,3	19,4	4,7	8,5	10,7	8,2	11,2	7,7
Liguria	1.149	3.306	30,6	54,8	4,5	0,1	20,9	9	7,5	1,1	22,9	19,8	13,6	15,3
Emilia Romagna	6.519	4.822	53	52,9	1,8	0,7	18	18,7	5,7	6,4	11,5	11,3	10	10
Toscana	4.097	4.746	50,1	25,8	2,8	2,6	16,8	16,4	8,6	13,2	10,7	18,9	11	23,1
Umbria	882	1.817	49,8	42,1	3,3	0,8	18,4	23,7	6,5	10,5	10,7	16,1	11,5	6,8
Marche	1.999	3.535	66,3	55,4	2,4	2,1	14,9	16,3	3,2	4,6	6,8	11,2	6,4	10,5
Lazio	4.983	1.863	15,7	35,2	3,1	7,1	18,7	15,7	10,5	0,9	26,7	22,2	25,4	18,9
Abruzzo	997	1.180	48,7	32,2	5,1	3,7	17,8	23	5,9	0,9	10,8	10,6	11,7	29,7
Molise	156	333	41	49,1	5,1	2,6	13,5	2,7	5,8	18,6	19,9	26,9	14,7	0
Campania	3.718	1.989	34,7	31,4	3,5	0,9	20,3	25,4	10,8	13,2	18,4	19,3	12,4	9,8
Puglia	2.244	2.212	38,5	55,2	4,1	5,6	22,4	18,3	8,4	1,3	14	11,1	12,6	8,5
Basilicata	276	815	31,9	35,2	10,9	9,2	15,2	13,8	6,9	0	18,5	16,9	16,7	24,9
Calabria	669	1.177	20,3	19,1	7,3	8,4	29,3	20,6	10,9	18,7	17,6	24,5	14,5	8,7
Sicilia	1.901	1.997	23,8	19,8	6,1	2,9	27,5	31,1	10,5	14,6	16,9	16,7	15,2	15
Sardegna	801	1.058	18,5	21,7	7,4	10,6	25,2	23,7	14,7	18,1	15,5	13,9	18,7	12,1
Totale	62.259	62.259	45,5	45,5	2,8	2,8	17,9	17,9	7	7	13,6	13,6	13,3	13,3

I valori relativi all'Indagine Invind sono calcolati adoperando il sistema di ponderazione standard utilizzato per le stime ufficiali dell'indagine.

Chiaramente, la mancanza di vincoli adeguati nel sistema di ponderazione non produce una distorsione solo relativamente alla stima della numerosità di imprese per regione, ma si estende anche alla stima di variabili economiche molto rilevanti per l'indagine, come ad esempio il fatturato (tavola 5).

Tabella 5: Differenze fra valore del fatturato riportato in Invind e Frame-SBS (2015)

Regione	Percentili della differenza (migliaia di euro)					
	P25	P50	P75	P90	P95	media
Piemonte e Valle D'Aosta	- 6	11	195	687	1.915	314
Lombardia	- 16	0	103	1.481	3.128	-226
Trentino A.A	- 0	1	82	642	1.587	1.398
Veneto	- 2	0	143	695	1.580	450
Friuli V. G.	- 1	1	94	322	535	452
Liguria	- 54	0	332	1.364	4.096	-6.179
Emilia Romagna	- 1	2	137	578	1.214	543
Toscana	- 9	0	114	515	1.692	2.006
Umbria	- 178	0	231	1.378	2.366	204
Marche	- 23	6	91	524	701	48
Lazio	- 1	43	237	1.553	3.740	-4.334
Abruzzo	- 270	0	113	921	2.360	51
Molise	- 1.175	6	160	286	999	-174
Campania	- 113	0	194	1.504	6.435	302
Puglia	- 33	6	205	1.753	2.672	185
Basilicata	- 74	6	92	519	1.873	170
Calabria	- 67	10	304	1.504	2.272	532
Sicilia	- 61	1	191	1.198	2.396	1.212
Sardegna	- 22	15	447	2.266	6.095	486
Totale	-16	1	160	802	2.272	-289

Per ciascuna impresa inclusa nel campione Invind è stata calcolata la differenza fra il valore dichiarato nell'indagine e quello risultante in Frame-SBS.

L'uso di Invind a livello regionale richiederebbe quindi quantomeno una modifica dell'attuale sistema di ponderazione, che permetta l'inclusione di vincoli a livello regionale in aggiunta agli altri attualmente esistenti o, in alternativa, la creazione di un sistema di pesi ad hoc per le analisi regionali basato sui c.d. *stimatori di regressione generalizzata* o *stimatori per piccole aree*, che sfruttano la disponibilità di informazioni ausiliarie note per ciascuna unità del campione (come ad esempio quelle contenute in Frame-SBS) per migliorare l'accuratezza e la precisione degli stimatori nei domini di interesse. Un confronto tra una vasta scelta di stimatori potenzialmente utilizzabili per lo sfruttamento dell'indagine Invind a livello regionale è offerto nei paragrafi 5 e 6.

4.3 Numerosità campionaria

Gli stimatori "classici" attualmente utilizzati nell'indagine (basati cioè sulle sole informazioni campionarie della regione di interesse), oltre a produrre stime potenzialmente distorte, non consentono di ottenere in genere risultati sufficientemente precisi data l'esigua numerosità campionaria attualmente rilevata in molte regioni. Chiaramente, lo specifico livello di precisione di

ogni stima sarà funzione, oltre che della numerosità del campione, anche della variabilità del fenomeno di interesse e del dominio di studio. La figura 1 mostra a titolo di esempio i coefficienti di variazione delle stime regionali di tre statistiche generalmente pubblicate nel Rapporto annuale della serie Economie regionali: la variazione puntuale del fatturato; la variazione puntuale degli investimenti; la quota di aziende in utile. Usando come dominio di studio la regione, si ottiene la precisione maggiore dalla stima della variazione del fatturato, dove i coefficienti di variazione risultano prevalentemente compresi fra il 2 e 3 per cento. Molto meno affidabili sono invece gli stimatori della variazione degli investimenti, (con coefficienti di variazione tra il 10 e il 20 per cento) e della quota di aziende in utile (tra il 5 e il 10 per cento). Anche in questo caso l'utilizzo di stimatori per piccole aree, accennati nel sotto-paragrafo precedente, permette di migliorare la precisione delle stime. Tuttavia, come mostreremo, tali stimatori producono risultati fortemente correlati alle variabili ausiliarie utilizzate (ad esempio il valore della variabile in t-1 per la stima della variabile in t). Di conseguenza possono essere poco adatti a fenomeni che mostrano una spiccata variabilità temporale (come ad esempio la stima della variazione degli investimenti) o a seguito di significativi shock tra un anno e l'altro. Per questo motivo, dopo aver confrontato l'accuratezza e la precisione dei diversi stimatori utilizzabili per l'analisi regionale, il paragrafo 7 si soffermerà sul possibile uso di Frame-SBS in fase di disegno dell'indagine, al fine di migliorare ulteriormente l'accuratezza e la precisione dei risultati.

5 La ricerca del miglior stimatore per la variazione del fatturato su domini non pianificati: una simulazione basata su Frame-SBS

L'analisi che segue valuta diversi stimatori adottando due approcci complementari. Il primo si basa su una simulazione condotta utilizzando l'archivio Frame-SBS che consente un'analisi controllata delle proprietà statistiche in condizioni ideali. In particolare, permette di: (i) valutare ex-ante la qualità degli stimatori al netto di errori campionari, quali la non risposta o gli errori di misura che possono affliggere i dati d'indagine; (ii) esaminare le performance degli stimatori su un elevato numero di campioni indipendenti, potenzialmente estraibili dalla popolazione di riferimento, superando così il vincolo dell'unico campione disponibile nell'indagine.

Il secondo approccio utilizza direttamente i dati dell'indagine Invind, consentendo di testare ex-post l'efficacia degli stimatori in un contesto reale. Ciò permette di verificarne la robustezza in presenza di alcune criticità empiriche tipiche delle indagini campionarie, come la corretta attribuzione territoriale delle variabili economiche e le limitazioni sul numero di indagini a disposizione.

Figura 1: Coefficienti di variazione di alcuni stimatori regionali (valori percentuali)

Fonte: Invind 2015.

La simulazione si basa su 2.000 campioni casuali indipendenti, ciascuno composto da 4.100 imprese, distribuite in modo da riflettere la struttura per strato del campione Invind 2015. Su ognuno dei 2000 campioni sono testati nove stimatori. I primi cinque usano la tecnica di post-stratificazione attualmente adottata da Invind. Gli altri quattro sono stimatori per piccole aree (cfr il paragrafo A in appendice per una descrizione approfondita degli stimatori per piccole aree).

La variabile analizzata è la variazione del fatturato medio. I domini di studio sono regione, settore (industria/servizi) e classe dimensionale (fino a 100 addetti/oltre 100 addetti). Le imprese con più di 5.000 addetti sono state considerate auto-rappresentative e, di conseguenza, sono state incluse in ognuno dei 2000 campioni.

Nello specifico, gli stimatori post-stratificati analizzati includono:

- stimatore T1, ovvero quello attualmente utilizzato nell'indagine Invind e non prevede vincoli a livello regionale. È costruito con vincoli basati sulla distribuzione delle imprese per classe di addetti, settore di attività economica e area geografica. In particolare, i vincoli sono dati dall'interazione fra classe di addetti (5 categorie) e settore (11 categorie) e

dall'interazione fra area geografica (4 categorie), settore (6 categorie) e classe dimensionale (2 categorie).

- stimatore T2, costruito con vincoli basati sulla distribuzione della numerosità delle imprese per regione, classe di addetti, settore di attività e della numerosità degli addetti totali per regione. In particolare i vincoli sono il numero di addetti per regione (19 categorie) e l'interazione fra numero di imprese per regione e classe dimensionale (2 categorie) e fra numero di imprese per regione e settore (2 categorie).

- stimatore T3, costruito con vincoli basati sulla distribuzione della numerosità delle imprese per regione, classe di addetti, settore di attività e forma giuridica. In particolare i vincoli sono dati dall'interazione fra numero di imprese per regione e classe di addetti (2 categorie), fra numero di imprese per regione e settore (2 categorie) e fra numero di imprese per regione e forma giuridica (2 tipologie).

- stimatore T4, costruito con vincoli basati sulla distribuzione della numerosità delle imprese per regione, classe di addetti, settore di attività, forma giuridica e della numerosità degli addetti per regione. In pratica si tratta degli stessi vincoli usati per lo stimatore precedente ai quali si aggiunge il totale addetti per regione.

- stimatore T5, che ha una struttura simile allo stimatore T4 ma, per la stima in una data regione, oltre alle imprese rilevate in quella regione utilizza anche una parte di imprese rilevate nelle regioni confinanti. La quota di imprese recuperata dalle regioni confinanti sarà tale da raggiungere, nella regione di interesse, una frazione sondata pari al 30% in ogni strato[2]. I vincoli usati sono il totale fatturato nel 2014 per regione, il numero totale di addetti per regione e l'interazione fra il settore (2 categorie) e la regione. Nel campione aggiuntivo le imprese sono selezionate casualmente fra quelle con uguale classe dimensionale e settore di attività.

Gli stimatori per piccole aree implementati sono invece quelli descritti nell'appendice A e utilizzano come informazione ausiliaria il fatturato nel 2014 (ovvero quello che si avrebbe a disposizione al termine dell'indagine sul 2015):

- T7, stimatore "pseudo EBLUP"

- T8, stimatore "GREG"

- T9, stimatore "pseudo M-Quantile"

[2]Ad esempio, ipotizzando che nel Lazio ci siano due strati con frazione sondata pari al 20 e al 26% rispettivamente, lo stimatore T5 campionerà casualmente dalla Campania, l'Abruzzo e la Toscana una quota di imprese pari al 10 e il 4% di quelle del Lazio, aggiungendole al campione rilevato in quella regione.

- T10, stimatore "pseudo M-Quantile BC"

La tavola 6 mostra la distorsione (bias) e l'errore quadratico medio (MSE) complessivi (ovvero calcolati come media su tutti i 2000 campioni e tutti i domini di studio) e quelli medi per dimensione e settore [3]. Per permettere confronti tra i bias dei diversi domini, il valore ottenuto è stato riscalato per il fatturato medio nel dominio e moltiplicato per 100.

Tabella 6: Bias e MSE medi.

	Stimatori post stratificati					Stimatori piccole aree			
	t1	t2	t3	t4	t5	t7	t8	t9	t10
Bias medio									
20-99 addetti	14,65	-3,14	-1,86	-4,46	-12,71	-0,88	5,46	6,11	0,66
100-oltre	-6,78	2,02	0,48	2,52	6,61	0,16	0,44	0,39	1,70
Industria	-13,68	7,78	1,93	9,45	24,57	-1,50	2,18	1,99	7,52
Servizi	1,16	-0,57	-0,67	-0,59	-1,70	2,35	1,83	2,03	-0,32
Totale	-7,11	4,08	0,77	5,00	12,93	0,21	2,03	2,02	4,05
MSE medio									
20-99 addetti	20,40	19,31	19,60	19,35	11,17	7,40	7,14	5,86	1,51
100-oltre	27,31	23,94	19,65	24,55	20,72	9,68	7,47	4,53	3,90
Industria	41,51	36,84	32,86	37,48	29,26	11,99	9,83	5,87	4,73
Servizi	6,20	6,41	6,40	6,42	2,63	5,09	4,78	4,52	0,69
Totale	23,85	21,63	19,63	21,95	15,95	8,54	7,31	5,20	2,71

Le colonne T2-T5 mostrano che l'introduzione di un sistema di pesi con vincoli sulle distribuzioni marginali a livello regionale, rispetto allo stimatore attuale (T1) attenua in modo consistente la distorsione delle stime, indipendentemente dalla dimensione e dal settore delle imprese. Lo stimatore T5 costituisce un'eccezione: l'aggiunta di imprese esterne alla regione migliora la varianza, ma aumenta la distorsione a causa della mancata corrispondenza tra le imprese residenti e quelle confinanti.

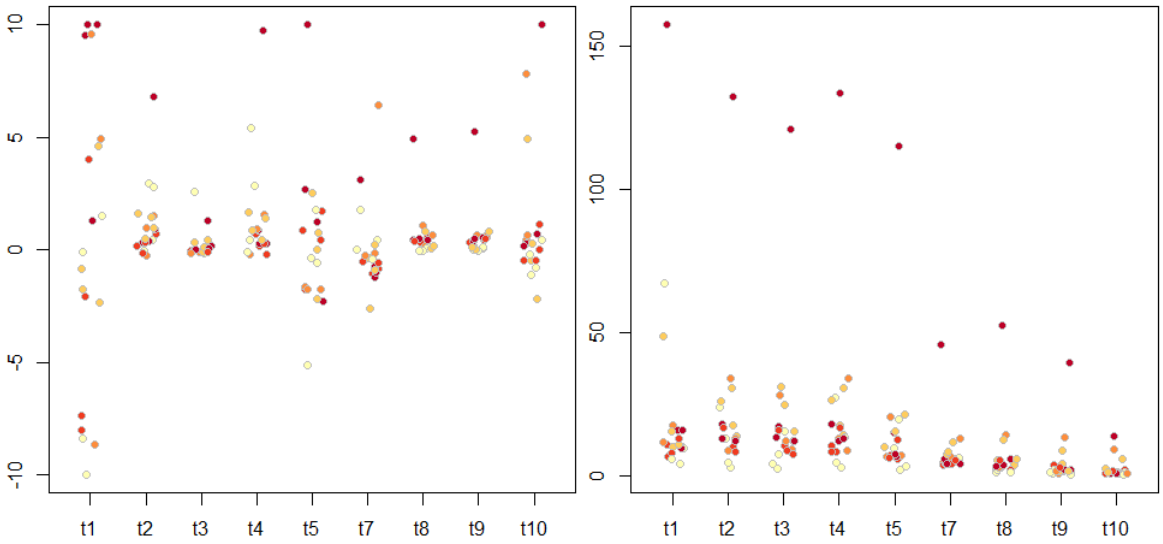
Gli stimatori per piccole aree (T7-T10) si dimostrano più efficienti su tutti i domini, riducendo l'errore quadratico medio (MSE) dal 65% all'87% rispetto ad altri metodi. Lo stimatore pseudo e-BLUP (T7) presenta una distorsione particolarmente bassa; tuttavia, questo non garantisce che sia sempre la scelta ottimale, poiché i risultati possono variare in base al dominio di studio analizzato.

A tal fine, le figure 2 e 3 mostrano, per ogni stimatore, rispettivamente le distribuzioni del bias e del MSE per regione e per l'incrocio di regione con il settore di attività economica.

La figura 2 (pannello a sinistra) confermerebbe in prima analisi che gli stimatori qui proposti (T2-T10), alternativi a quello attualmente utilizzato per le elaborazioni regionali (T1) producono stime che nella quasi totalità dei domini hanno una distorsione più contenuta. Tale risultato

[3] Il bias ed il MSE su ogni singolo campione sono calcolati rispettivamente come $Bias = 1/n \sum_{i=1}^{N}(x_i - x^*)$ e $MSE = 1/n \sum_{i=1}^{n}(x_i - x^*)^2$ dove x^* rappresenta il valore di fonte amministrativa.

Figura 2: Bias (sx) e MSE (dx) delle stime per regione

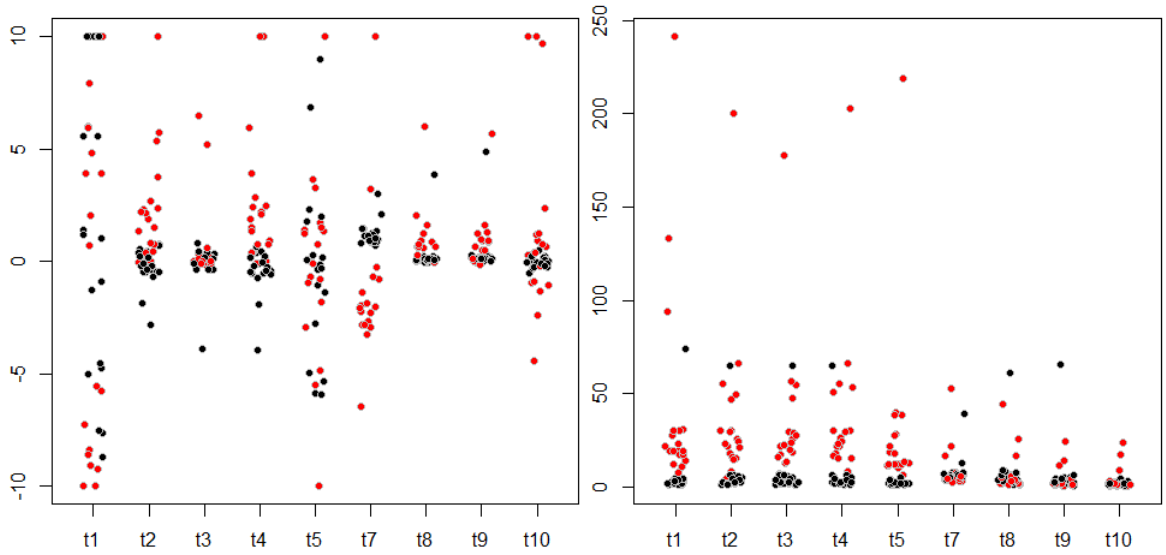


La figura mostra, per ognuno degli stimatori T1-T10, il bias (calcolato come $1/n \sum_{i=1}^{N}(x_i - x^*)$; figura di sinistra) e il Mean Squered Error (calcolato come $1/n \sum_{i=1}^{n}(x_i - x^*)^2$) delle stime regionali. Ogni puntino rappresenta quindi una regione e l'intensità del colore è proporzionata alla dimensione della regione, in termini di quota di imprese sul totale nazionale.

non sembrerebbe peraltro essere direttamente collegato alla dimensione delle regioni in termini di quota di imprese sul totale nazionale, identificata in figura dall'intensità del colore dei punti. Osservando l'intera distribuzione delle stime prodotte per le diverse regioni si osserva inoltre che solo per gli stimatori T3, T8 e T9 la bassa distorsione complessiva riassunta nella tavola 6 riflette una nuvola di punti effettivamente concentrata attorno allo zero. Diversamente, per lo stimatore T7, la minore distorsione complessiva appare legata principalmente alla contemporanea presenza di stime distorte positivamente e negativamente che si bilanciano. Inoltre, il pannello di destra della stessa figura mostra come gli stimatori per piccole aree (T7-T10) siano in grado di aumentare su tutte le regioni la precisione delle stime. Ciò è particolarmente evidente per lo stimatore T10, l'unico per il quale non si osserva nessun outlier nella distribuzione dei MSE.

Sebbene la figura 2 permetta di avere una visione di sintesi dell'incremento di correttezza e precisione delle stime nelle regioni tramite l'utilizzo degli stimatori proposti, la stessa non permette di catturare la presenza di eterogeneità della qualità delle stime nei domini di interesse, in particolare tra i principali settori e le classi dimensionali di impresa. A tal fine la figura 3 suggerisce che la minor distorsione degli stimatori T2-T10 è associata principalmente ai domini legati alle imprese dei servizi; la stessa analisi, replicata per le imprese di diversa dimensione, suggerisce che alle imprese con meno di 100 addetti è associato un bias che, in termini assoluti, è più contenuto ma, in rapporto al fatturato medio della classe, è superiore a quello che emerge per le imprese più grandi. Tali risultati sono anche confermati da una semplice analisi di regressione

Figura 3: Bias (sx) e MSE (dx) delle stime per regione e settore: in rosso domini con imprese dell'industria, in nero i servizi.



La figura mostra, per ognuno degli stimatori T1-T10, il bias (calcolato come $1/n \sum_{i=1}^{N}(x_i - x^*)$; figura di sinistra) e il Mean Squered Error (calcolato come $1/n \sum_{i=1}^{n}(x_i - x^*)^2$) delle stime regionali. Ogni puntino rappresenta quindi una regione e l'intensità del colore è proporzionata alla dimensione della regione, in termini di quota di imprese sul totale nazionale.

(di cui non riportiamo i risultati per brevità) del bias su una serie di dummy esplicative delle caratteristiche dei domini.

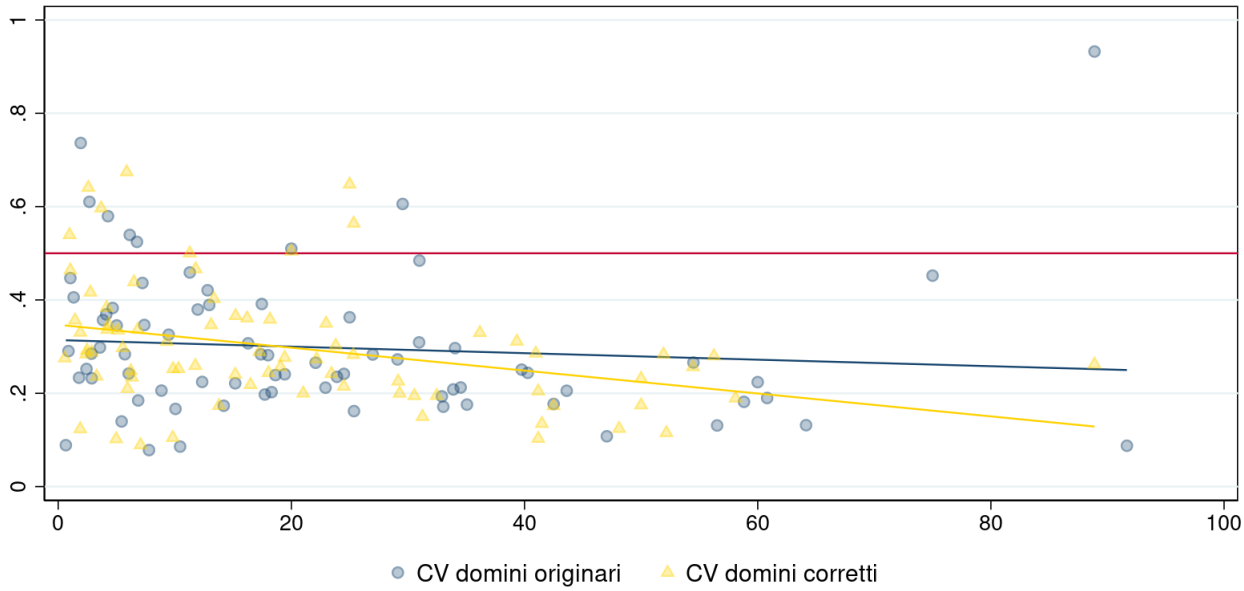
6 La stima del fatturato: un'applicazione all'indagine Invind

La simulazione del precedente paragrafo, basata sull'archivio Frame-SBS, ha permesso di studiare alcune proprietà asintotiche degli stimatori proposti in assenza dei problemi legati alla non risposta e agli errori di misurazione tipici delle indagini campionarie.

In questo paragrafo, tali stimatori vengono applicati all'indagine Invind 2015 per analizzare il fatturato su due tipi di domini: quelli basati sui dati raccolti nell'indagine (definiti domini originali) e quelli costruiti usando le informazioni del settore e classe dimensionale ottenute da Frame-SBS (definiti domini corretti). Vengono valutati il coefficiente di variazione (CV) e la distorsione degli stimatori T1-T10, assumendo che i valori medi del fatturato di fonte amministrativa siano quelli veri della popolazione. La figura 4 illustra la relazione tra i CV e la frazione sondata per entrambi i tipi di domini.

In linea con le aspettative, si osserva che la variabilità dei risultati ottenuti dai diversi stimatori è in parte spiegata dalla numerosità campionaria. L'inclinazione negativa della retta di regressione conferma che, all'aumentare della frazione sondata, diminuisce la discordanza tra gli stimatori.

Figura 4: Coefficiente di variazione su frazione sondata per i domini rilevati nell'indagine e quelli allineati a Frame-SBS.



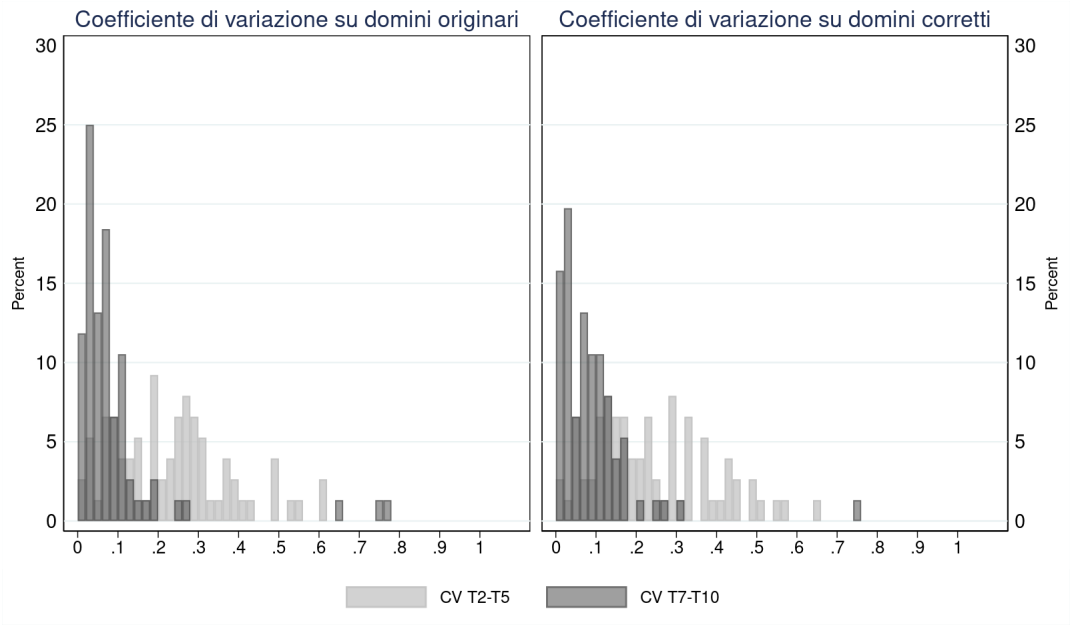
Tuttavia, non emerge una correlazione significativa tra l'ampiezza dei coefficienti di variazione (CV) e il settore o la dimensione delle imprese nei domini. Questo risultato potrebbe sembrare insolito, poiché ci si aspetterebbe una maggiore concordanza delle stime nei domini con imprese di grandi dimensioni. Nel presente studio, però, la soglia di addetti che separa le imprese grandi da quelle piccole è fissata a 100, mantenendo così un'elevata eterogeneità all'interno dei domini.

La figura 5 mostra la distribuzione dei CV del fatturato medio nei vari domini. I risultati mostrano come gli stimatori per piccole aree (T7-T10) abbiano una distribuzione dei CV con massa più spostata verso lo zero rispetto agli altri e questo a prescindere dal fatto che i domini siano o meno definiti sulla base di Frame-SBS.

La figura 6 mostra invece le differenze fra le stime di Frame-SBS e quella ottenute sulla base dei vari stimatori. Anche in questo caso gli stimatori per piccole aree (punti rossi) tendono a produrre differenze vicino allo zero e quindi a garantire una minore distorsione rispetto agli altri.

La maggiore efficienza e minore distorsione di questi stimatori dipendono dalla correlazione tra la variabile di studio e quella ausiliaria. Per il fatturato, c'è una forte correlazione a distanza di due anni, ma per variabili come gli investimenti, con una correlazione più debole, i risultati potrebbero cambiare. La figura 6 mostra che in alcuni domini, soprattutto di piccole dimensioni, ci sono differenze significative tra le stime basate su informazioni ausiliarie e i valori corretti di

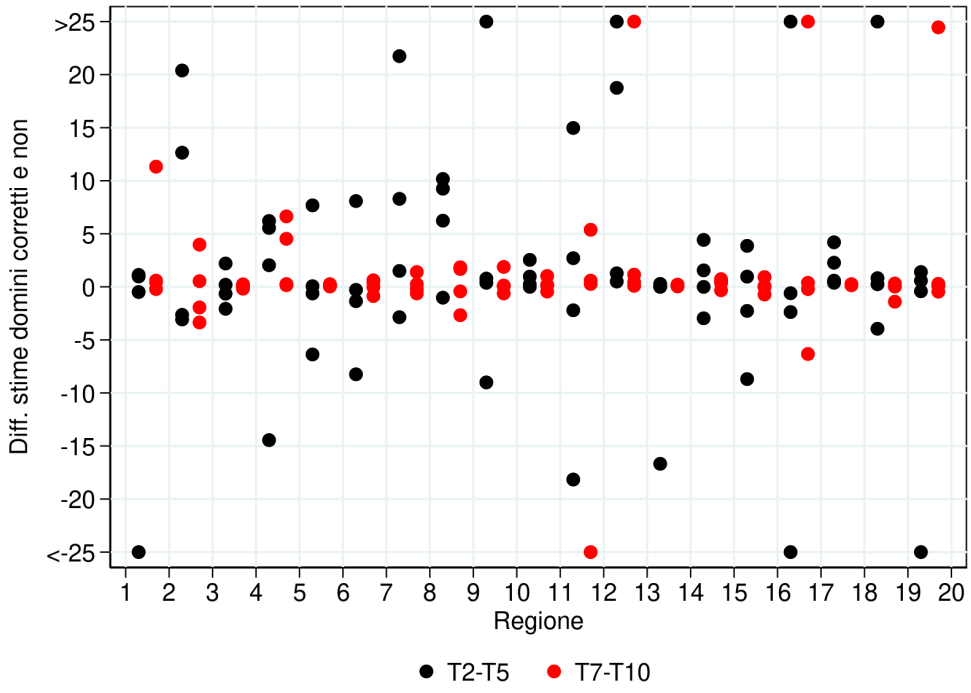
Figura 5: Confronto tra la distribuzione dei coefficienti di variazione medi per gli stimatori T2-T5 e T7-T9



Frame-SBS. Questo accade per due motivi: primo, il fatturato di Frame-SBS può variare molto tra gli anni, rendendo meno precisi gli stimatori basati su dati passati; secondo, alcune imprese possono appartenere a domini diversi a seconda che si utilizzino le informazioni campionarie o quelle amministrative. Gli stimatori per piccole aree, che assumono coerenza tra classificazioni nell'indagine e nella popolazione, possono risultare meno accurati. In tal caso, gli stimatori post-stratificati (T2-T5) potrebbero essere preferibili.

I risultati quindi suggeriscono che, per l'analisi regionale, sia opportuno valutare la robustezza dei risultati utilizzando più stimatori per poi scegliere il più adatto in base al fenomeno di interesse e alle caratteristiche del dominio di studio.

Figura 6: Differenza tra le stime nei domini originariamente rilevati e quelli corretti sulla base dei dati Frame-SBS. Per ogni regione (sull'ascisse) sono mostrati 4 punti ad indicare i valori assunti per le imprese dell'industria e dei servizi, e per quelle di piccole e grandi dimensione.



7 L'uso delle informazioni ausiliarie in fase di disegno

Abbiamo fin ora analizzato il modo in cui la disponibilità dell'archivio amministrativo Frame-SBS, possa migliorare la correttezza e la precisione delle stime campionarie nell'indagine Invind, tramite l'utilizzo di stimatori per piccole aree. Nei successivi paragrafi ci soffermeremo invece sul possibile utilizzo dello stesso archivio nella fase di disegno dell'indagine Invind. Per una discussione più generale sui disegni di campionamento nelle indagini sulle imprese si rimanda all'appendice B.

7.1 Definizione di un disegno campionario ottimale per l'indagine Invind

L'obiettivo di questo paragrafo è quello di utilizzare l'archivio Frame-SBS per individuare un nuovo disegno campionario per Invind che permetta di pianificare il livello di precisione atteso delle stime per diverse variabili di interesse ai diversi livelli di dettaglio richiesti e che consenta allo stesso tempo il controllo della dimensione campionaria.

I domini considerati sono di tre tipi: la regione in cui l'impresa ha la propria sede amministrativa, l'incrocio fra regione e due macrosettori di attività economica (Industria e Servizi) e l'incrocio fra regione e dimensione dell'impresa (piccole/grandi imprese, dove per grandi si intendono quelle con almeno 50 addetti).

La determinazione di un'allocazione campionaria multidominio e multiobiettivo per Invind è stata effettuata intervenendo su tre fattori: la scelta della soglia di censimento, la definizione degli strati, il tipo di disegno stratificato (standard o incompleto). Chiaramente, la partizione base che consente di ottenere i tre tipi di dominio come pianificabili è quella data dal prodotto delle tre variabili (ovvero il prodotto di 19 regioni, per 2 settori, per 2 classi dimensionali). Tuttavia, se la conoscenza di una o più variabili ausiliarie Y correlate con quelle target permette la suddivisione della popolazione in gruppi più omogenei rispetto a quella finora adottata, è possibile introdurre una stratificazione alternativa in grado di apportare una significativa riduzione della dimensione campionaria a parità di precisione attesa delle stime.

Una prima suddivisione si può effettuare scegliendo opportunamente una soglia di auto-rappresentatività delle unità, o soglia di censimento. Se determinata in base ad un criterio di ottimalità, tale soglia y^* è in grado di individuare una subpopolazione affetta da minore variabilità rispetto all'universo considerato nel suo complesso. La ricerca della y^* ottimale avviene normalmente attraverso un processo iterativo di non immediata implementazione. Un metodo sub-ottimale ma di facile applicazione è quello denominato *approx1* proposto da Hidiroglou (1996): consente di determinare la soglia y^* tale che l'errore campionario che si commette nella stima del totale Y, in termini di coefficiente di variazione, sia inferiore ad un ϵ fissato e la dimensione campionaria n sia minima. Considerando come variabile ausiliaria il numero di

addetti dell'impresa, l'applicazione del criterio di Hidiroglou distintamente per regione ha dato luogo alle soglie di censimento in tavola 7.

Tabella 7: Soglie di censimento ottime per regione

Soglia CENS								
Regione		n. addetti medi nell'anno	Regione		n. addetti medi nell'anno	Regione		n. addetti medi nell'anno
1	Piemonte - V.d'A.	1095,00	7	Emilia Rom.	1108,00	13	Molise	87,85
2	Lombardia	2101,00	8	Toscana	642,50	14	Campania	604,00
3	Trentino A.A.	420,50	9	Umbria	331,00	15	Puglia	466,00
4	Veneto	1042,50	10	Marche	423,50	16	Basilicata	199,00
5	Friuli V.G.	507,00	11	Lazio	1815,50	17	Calabria	246,00
6	Liguria	507,50	12	Abruzzo	388,00	18	Sicilia	454,00
						19	Sardegna	324,50

Come primo passo nella ricerca di un disegno più efficiente per Invind è stato verificato l'effetto sulla numerosità campionaria complessiva dell'introduzione, tra i caratteri di stratificazione, della variabile CENS, corrispondente alla soglia di censimento ottima. Le caratteristiche di questo primo disegno sperimentato sono riassunte in tavola 8.

Per ciascuna regione, tutte le imprese con *addetti*≥CENS sono state collocate in strati completamente osservati (*censiti*) alle cui unità è stata assegnata probabilità di inclusione pari ad 1. Per le imprese appartenenti agli strati *campionati* la probabilità di inclusione sarà invece pari al rapporto $\frac{n_h^*}{N_h}$, dove n_h^* rappresenta la dimensione campionaria ottima di strato ottenuta come risultato della procedura di allocazione multidominio.

Tabella 8: Disegno1- stratificazione per macro settori di attività economica, classi addetto, regioni.

n. Strati: 114	Strato: incrocio fra regione, censimento, settore e classi di addetti	
Variabili di Strato o Dominio	N. modalità	Descrizione
Censimento (Cens)	2	Soglia di censimento (in n. addetti) variabile per regione.
Regione (Reg)	19	Regioni, con Piemonte + V.d'A.=01, Trento + Bolzano=04.
Classe di addetti (Cladd)	3	Cladd: classi di addetto (20-49; 50-CENS, oltre CENS).
Settore (Ind)	2	M0=Industria: divisioni Nace da 06 a 39. M1=Commercio e Servizi: divisioni da 45 a 82.
Domini Pianificati:		DOM1= Reg; DOM2=Reg*Ind; DOM3=Reg*Cladd

L'allocazione negli strati è stata determinata valutando se, a parità di errore atteso sulle stime delle variabili, l'eventuale adozione di un disegno a stratificazione incompleta (S.INC) porti a una riduzione della numerosità campionaria complessiva rispetto al disegno a stratificazione completa (S.COMP)[4]. L'errore atteso è stato espresso in entrambi i casi in termini di coefficiente di variazione delle stime dei totali; per la sua stima, è necessaria la specificazione a

[4]La numerosità ottimale ottenuta dall'adozione del disegno a stratificazione completa e incompleta è stata calcolata usando rispettivamente i packages Bethel e MultiWay del software R.

priori di alcuni parametri di input a livello di strato, ossia: la dimensione N_h della popolazione e, relativamente alle variabili ausiliarie prescelte, la media e la varianza di popolazione.

Il calcolo dei parametri di input dipende dalla disponibilità di informazioni sulle variabili: quando queste sono note a livello di popolazione (ad esempio se presenti su Registro), è possibile calcolare la varianza di popolazione della variabile Y_p con la formula:

$$ {}_pS_h^2 = \frac{\sum_i^{N_h} ({}_py_{ih} - \overline{{}_pY_h})^2}{N_h - 1} \tag{1} $$

Nel caso non si disponga di informazioni complete da Registro si usa imputare una stima degli stessi parametri, ottenuta con dati rilevati in precedenti occasioni di indagine; posto w_h il peso finale ottenuto dalla calibrazione dei dati campionari dell'indagine al tempo t-1, la stima campionaria ${}_ps_h^2$ della varianza di strato di Y_p è espressa da:

$$ {}_ps_h^2 = \frac{\sum_i^{n_h} w_i \left({}_py_{ih} - \left(\frac{\sum w_{ip} y_i}{\sum w_i} \right) \right)^2}{\sum_i^{n_h} w_i - 1} \tag{2} $$

In questo lavoro, per il calcolo dei parametri di input relativi a Invind è stato possibile utilizzare le informazioni di Frame-SBS.

Nella progettazione del campione si è tenuto conto della necessità di includere tutte le unità degli strati censiti e di non superare, per ragioni logistiche e di costo, la soglia di 7.000 unità cui inviare il questionario. Il dimensionamento del campione è stato, quindi, ottenuto imponendo alle stime per dominio dei totali delle variabili guida *Addetti* e *Fatturato* un set di vincoli sull'errore massimo atteso fino al raggiungimento della numerosità complessiva desiderata.

Il risultato dell'applicazione dei due disegni di allocazione sotto tali condizioni è riportato nella tavola 9. Si osserva come una dimensione vicina a quella desiderata possa essere raggiunta solo imponendo un errore massimo per le stime non inferiore al 20%. Si può inoltre notare che non si ottiene un guadagno apprezzabile con l'adozione di un disegno a stratificazione incompleta (S.INC): ciò è dovuto al numero relativamente basso degli strati, i quali sono probabilmente anche caratterizzati da una scarsa omogeneità interna rispetto alle variabili ausiliarie utilizzate per guidare l'allocazione.

Tabella 9: Disegno1- numerosità campionaria ottima con disegni a stratificazione completa (S.COMP) e incompleta (S.INC)

Algoritmo di allocazione		CV pianificati su DOM1-3
S.COMP	S.INC	Variabili: Fatturato, Addetti
28.914	28.624	0,05
15.527	15.102	0,10
10.448	9.944	0,15
7.910	7.361	0,20

Per testare questa ipotesi si è pertanto sperimentato un secondo disegno con stratificazione più fine, ottenuta disaggregando il carattere IND che definisce l'attività economica dell'impresa nelle 63 divisioni di attività (Nace2- 2 cifre). Le caratteristiche del Disegno2 sono riportate in tavola 10.

Tabella 10: Disegno 2: stratificazione per divisioni di attività economica, classe di addetti regioni.

n. Strati: 2.135	Strato: incrocio regione, censimento, settore e classi di addetti	
Variabili di Strato o Dominio	N. modalità	Descrizione
Censimento (Cens)	2	Soglia di censimento (in n. addetti) variabile per regione .
Regione (Reg)	19	Regioni, con Piemonte + V.d'A.=01, Trento + Bolzano=04.
Settore (S2AT)	63	Divisioni di attività economica (da 06 a 82).
Classe di addetti (Cladd)	3	Cladd: classi di addetto (20-49; 50-CENS, oltre CENS).
Settore (Ind)	2	Ind: M0=Industria: divisioni Nace da 06 a 39. M1=Commercio e Servizi: divisioni da 45 a 82.
Domini Pianificati:		DOM1= Reg DOM2=Reg*Ind DOM3=Reg*Cladd

Il passaggio alla divisione di attività porta alla partizione dell'universo da 114 a 2135, con un netto miglioramento della loro omogeneità interna rispetto alle variabili ausiliarie: assumendo il coefficiente di variazione di popolazione come misura di dispersione delle suddette variabili nello strato, i percentili delle rispettive distribuzioni mostrano una riduzione nei valori di oltre il 50% passando alla stratificazione introdotta col Disegno2 (tabella 11).

Tabella 11: Distribuzione dei coefficienti di variazione delle variabili ausiliarie negli strati per disegno

	Percentili delle distribuzioni						
variabile	p_25	p_50	p_75	p_90	p_95	max	mean
	Disegno1 (114 strati)						
FATTURATO	0,993	1,451	2,023	3,031	4,903	25,217	2,077
ADDETTI	0,276	0,574	0,764	1,057	1,260	2,212	0,607
	Disegno2 (2135 strati)						
FATTURATO	0,339	0,729	1,090	1,508	1,812	12,095	0,781
ADDETTI	0,151	0,277	0,524	0,871	1,064	1,864	0,366

Da un punto di vista dell'allocazione campionaria, l'aumento di omogeneità negli strati si traduce in una netta riduzione della dimensione campionaria ottimale: come si evince dalla tavola 12, imponendo un vincolo di errore pianificato al 10% nel Disegno 2 si raggiunge la dimensione desiderata di circa 7.000 unità, numerosità che nel Disegno1 comporta un errore anche superiore al 20%.

Nel Disegno2 inoltre l'algoritmo di stratificazione incompleta con errore al 10% porta ad una ulteriore riduzione della dimensione campionaria a 4.917 unità: ciò dipende, come si è anticipato nel paragrafo precedente, dal fatto che tale algoritmo non è soggetto al vincolo di dover campionare almeno due unità per strato. Questo vincolo è invece previsto per l'algoritmo

Tabella 12: Disegno2- numerosità campionaria ottima con disegni a stratificazione completa (S.COMP) e incompleta (S.INC)

Algoritmo di allocazione		CV pianificati su DOM1-3
S.COMP	S.INC	Variabili: Fatturato, Addetti
12.344	11.526	0,05
8.214	6.593	0,08
6.909	4.917	0,10

a stratificazione completa e implica che in presenza di un numero elevato di strati con al più due unità, questi vengano censiti con una conseguente allocazione di un campione superiore di quello strettamente necessario per non eccedere l'errore massimo pianificato.

La possibilità di non censire gli strati di piccolissima dimensione fa sì che l'allocazione complessiva sia più efficiente e non determini sovra-campionamento e conseguente riduzione non pianificata degli errori attesi in tali strati.

Con l'algoritmo a stratificazione incompleta è pertanto possibile allocare negli strati la numerosità campionaria *esatta* per realizzare un errore atteso pari al CV max pianificato.

Tuttavia se da un lato questa caratteristica è ottimale in termini di response burden e minimizzazione dei costi di rilevazione, dall'altro non prevedere una minima percentuale di sovracampionamento in fase di allocazione può rivelarsi rischioso, in termini di errore campionario effettivamente realizzato e di distorsione delle stime finali, quando il fenomeno di interesse è caratterizzato da un'elevata mancata risposta totale.

La tavola 13 mostra la numerosità ottimale per ciascun dominio di studio formato dall'incrocio di regione e settore di attività (industria/servizi) confrontandola con quella prevista dall'attuale allocazione.

Dai risultati emerge che per molte regioni sarebbe necessario ridurre la numerosità delle imprese manifatturire ed aumentare quella dei servizi. Questo risultato è coerente con il fatto che la variabilità osservata in questo settore è maggiore rispetto a quella del manifatturiero. Ciò nonostante, in alcune grandi regioni quali la Lombardia, il Lazio e l'Emilia Romagna, sarebbe necessario aumentare la dimensione campionaria in entrambi i domini. Queste considerazioni non tengono naturalmente in considerazione i vincoli operativi che limitano inevitabilmente la possibilità di raggiungere tali obiettivi.

Tabella 13: Confronto tra la numerosità obiettivo calcolata con un disegno a stratificazione incompleta e quella di Invind 2015

Regione		Industria		Servizi	
		campione S.INC	campione Invind15	campione S.INC	campione Invind15
1	Piemonte - V.d'A.	229	272	125	21
2	Lombardia	417	359	241	119
3	Trentino A.A.	116	87	74	23
4	Veneto	224	251	185	121
5	Friuli V.G.	76	112	60	48
6	Liguria	87	124	56	60
7	Emilia Romagna	240	182	170	84
8	Toscana	265	215	153	116
9	Umbria	64	132	66	38
10	Marche	78	258	89	51
11	Lazio	476	146	65	75
12	Abruzzo	79	113	69	38
13	Molise	23	41	19	8
14	Campania	152	174	97	87
15	Puglia	177	339	140	64
16	Basilicata	36	79	35	29
17	Calabria	64	66	61	46
18	Sicilia	186	113	66	91
19	Sardegna	101	179	57	34

8 Conclusioni

Questo studio ha evidenziato come l'integrazione tra l'indagine Invind e l'archivio Frame-SBS possa migliorare significativamente le stime su domini non pianificati, in particolare a livello regionale o nell'incrocio tra regione e settore o classe di addetti. Anche la semplice introduzione di vincoli specifici a livello regionale nel sistema di ponderazione ha consentito di ottenere risultati più accurati rispetto al sistema di ponderazione usato per produrre stime a livello nazionale (o di macro area geografica). Inoltre, le analisi basate su simulazioni e su dati d'indagine hanno mostrato che gli stimatori per piccole aree (SAE) si distinguono per una maggiore stabilità e minore errore quadratico medio, pur comportando un possibile incremento del bias in alcune circostanze.

I risultati hanno confermato che l'efficienza degli stimatori è strettamente legata alla correlazione tra la variabile di studio e quella ausiliaria. La forte correlazione temporale del fatturato ha reso questi stimatori particolarmente adatti per l'analisi di questa variabile. Al contrario, per altre grandezze, come gli investimenti, che presentano una correlazione più debole, i risultati potrebbero essere meno soddisfacenti.

Un aspetto cruciale emerso è l'importanza di valutare la robustezza degli stimatori in fun-

zione delle caratteristiche dei domini e delle variabili analizzate. Per garantire stime affidabili, è fondamentale adottare un approccio flessibile che integri diverse metodologie, selezionando di volta in volta quella più appropriata al contesto.

Infine, l'archivio Frame-SBS si è dimostrato un supporto essenziale non solo per migliorare la precisione delle stime ex-post, ma anche per fornire indicazioni utili sulla dimensione ottimale del campione necessario per ottenere stime affidabili a livello locale. Questi risultati rappresentano un significativo passo avanti verso un'analisi più dettagliata e precisa delle economie regionali, offrendo una base solida per applicazioni future e sviluppi metodologici ulteriori.

Sebbene la questione della corretta attribuzione territoriale dell'attività produttiva non sia affrontata in questo lavoro, si riconosce che essa rappresenta un tema cruciale e metodologicamente distinto, che richiederebbe fonti informative specifiche o modelli di allocazione.

Infine, è importante sottolineare come l'approccio ex ante e quello ex post proposti nel lavoro non sono soluzioni alternative, ma strumenti complementari che possono essere impiegati in modo sinergico. Un disegno campionario ottimizzato con le informazioni di Frame-SBS migliora la copertura nei domini di interesse già in fase di rilevazione, rafforzando l'efficacia degli stimatori indiretti utilizzati in una fase successiva. Al tempo stesso, quando vincoli operativi limitano l'espansione del campione, l'approccio ex post consente di integrare efficacemente le informazioni disponibili, garantendo stime più accurate anche in contesti scarsamente rappresentati.

Bibliografia

Bethel J., (1989), Sample Allocation in Multivariate Surveys, *Survey Methodology*, 15, pp. 47-57.

Cesari, R., Signorini, L. F., (1991). Stime regionali con 'pochi dati': analisi e simulazione di stimatori alternativi per investimenti, occupazione e fatturato nelle imprese manifatturiere, Banca d'Italia, Temi di Discussione, n. 152.

Chandra, H., and R. Chambers. (2011). "Small Area Estimation for Skewed Data in Presence of Zeros." *The Bulletin of Calcutta Statistical Association*, **63**, 249–252.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R., Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chambers, R., Chandra, H., Salvati, N., Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society*, Ser. B, 76, 47-69.

Cicchitelli, G., Herzel, A., Montanari, G. E., (1992), *Il campionamento statistico*, Il mulino, Bologna.

Deville, J.C., Sarndal C.E. (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 141, 376-382.

D'Alessio G., D'Aurizio L., Faiella I (2007). Metodi di stima degli investimenti industriali a livello sub-nazionale nell'indagine Invind Mimeo, Banca d'Italia.

D'Alò, M., S. Falorsi, and F. Solari. (2017). "Space-time Unit-level EBLUP for Large Data Sets." *Journal of Official Statistics*, **33**(1), 61–77.

Fabrizi, E., Salvati, N., Pratesi, M. & Tzavidis, N. (2014). Outlier robust model-assisted small area estimation. *Biometrical Journal*, 56, 157-175.

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

74(366), 269–277. doi: 10.1080/01621459.1979.10482505.

Fuller W.A. (2009) *Sampling statistics*, New York: John Wiley ans Sons.

Hidiroglou M.A. (1996) " The Costruction of Self-Representing Stratum of Large Units in Survey Design". *The American Statistician*, 40, 27-31.

Karlberg, F. (2014). "Small Area Estimation for Skewed Data in the Presence of Zeros." *Statistics in Transition new series and Survey Methodology Joint Issue: Small Area Estimation 2014*, **16**(4), 541–562.

Koenker, R. (2005). *Quantile Regression*, Economic Society Monographs. New York: Cambridge University press.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Luzi, O., R. Monducci, A. Righi, and F. Vacca. (2018). "A Study of Small Area Estimation for Italian Structural Business Statistics." *Journal of Official Statistics*, **34**(2), 493–526.

Petrucci, A., M. Pratesi, and N. Salvati. (2005). "Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects." *Statistics in Transition*, **7**(3), 609–623.

Sarndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling.* SpringerVerlag.

Tzavidis, N., Marchetti, S., Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

You Y, Rao, J.N.K. (2002) A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights, *The Canadian Journal of Statistics,* 30, 431-439.

Appendice

A Una rassegna degli stimatori in presenza di informazioni ausiliarie

In questo paragrafo vengono descritti alcuni stimatori che potrebbero essere impiegati per far fronte ai problemi di bassa numerosità campionaria e di mancanza di un sistema di ponderazione con vincoli regionali. Una prima categoria di stimatori sono quelli di regressione generalizzata che sfruttano la disponibilità di informazioni ausiliarie, ossia di variabili note, per ciascuna unità del campione e il cui totale o la cui media sono noti a livello di dominio da altra fonte, tipicamente un archivio amministrativo o un censimento. Una seconda categoria che può essere utilizzata sono gli stimatori per domini di studio in cui l'integrazione tra l'informazione ausiliaria viene integrata direttamente al livello del dominio di interesse.

A.1 Gli stimatori di regressione generalizzata

Si supponga che la popolazione obiettivo P, la cui dimensione indichiamo con N sia suddivisa in D domini di studio $(P_1, \ldots, P_d, \ldots, P_D)$ di dimensione $(N_1, \ldots, N_d, \ldots, N_D)$, tali che $\sum_{d=1}^{D} N_d = N$. Si supponga inoltre che il campione estratto dalla popolazione (di dimensione complessiva n) sia caratterizzato da probabilità di inclusione π_{di} con $i = 1, \ldots, n_d$ e $d = 1, \ldots, D$, e quindi da pesi $w_{di} = \pi_{di}^{-1}$. Si indichi con n_d la dimensione del campione specifica di ciascun dominio. Se i domini rappresentano strati della popolazione, ovvero se il dominio è pianificato n_d è noto a priori, altrimenti sarà un numero casuale che dipenderà dallo specifico campione estratto. Si supponga di voler stimare la media di una variabile obiettivo y per ciascuno dei domini di studio, un insieme di parametri che denotiamo con $\bar{Y}_d$, $d = 1, \ldots, D$. Gli stimatori più semplici per $\bar{Y}_d$ sono gli stimatori "diretti", basati sull'applicazione delle formule standard che si utilizzano per gli stimatori dei parametri di popolazione alle osservazioni y_d specifiche di ciascun dominio.

- Stimatore di Horwitz-Thompson

$$\hat{\bar{Y}}_d^{HT} = \frac{1}{N_d} \sum_{i=1}^{n_d} \frac{y_{di}}{\pi_{di}} = \frac{1}{N_d} \sum_{i=1}^{n_d} w_{di} y_{di} \tag{3}$$

Si tratta di uno stimatore non-distorto, la cui efficienza è però strettamente legata al fatto che il dominio sia pianificato ossia al rispetto della relazione $\hat{N}_d = N_d$, dove $\sum_{i=1}^{n_d} w_{di} = \hat{N}_d$. Per domini non pianificati questa la differenza tra $\hat{N}_d$ e N_d può essere anche considerevole e causare l'instabilità dello stimatore (si veda Cicchitelli, Herzel, Montanari, 1988, pagg. 258-259).

- Stimatore per rapporto o di Hàjek

$$\hat{\bar{Y}}_d^{Ha} = \frac{1}{\hat{N}_d} \sum_{i=1}^{n_d} \frac{y_{di}}{\pi_{di}} = \frac{\sum_{i=1}^{n_d} w_{di} y_{di}}{\sum_{i=1}^{n_d} w_{di}} \tag{4}$$

Tecnicamente si tratta di uno stimatore per rapporto in cui la variabile ausiliaria è data dall'indicatrice di appartenenza al dominio. Come stimatore per rapporto, non è corretto, ma asintoticamente corretto (solitamente poco distorto anche in piccoli campioni): la distorsione è infatti di ordine $O(n_{di}^{-1})$. Alternativamente, lo stimatore di Hàjek può essere visto come uno stimatore di calibrazione in cui si utilizzino le indicatrici di appartenenza al dominio come variabili ausiliarie.

Entrambe questi stimatori sono caratterizzati da variabilità elevata quando le numerosità n_d sono piccole e pertanto spesso si può non essere in grado di garantire precisioni adeguate per $\hat{\bar{Y}}_d^{HT}$ e $\hat{\bar{Y}}_d^{Ha}$ per la maggior parte dei domini di studio. Una soluzione al problema è rappresentata dall'impiego di stimatori "indiretti" che utilizzano osservazioni y_{di} di tutto il campione per ottenere stime a livello di ciascun singolo dominio. In pratica si tratta spesso di stimatori che fanno ricorso ad informazioni ausiliarie, ossia a variabili $x_1, \ldots, x_p$ i cui valori sono noti per ciascuna unità campionata, ma di cui allo stesso tempo è nota la media a livello di dominio $\bar{X}_{d1}, \ldots, \bar{X}_{dp}$ da una fonte esterna.

Nella teoria della stima per popolazioni finite, una categoria molto ampia di stimatori che utilizzano informazioni ausiliarie è data dagli stimatori per calibrazione. L'idea di base è quella di sostituire i pesi base w_{di} utilizzati nello stimatore di Horwitz-Thompson con un sistema di pesi "calibrato" $\tilde{w}_{di}$ in modo tale da rispettare i vincoli $\sum_{d=1}^{D} \sum_{i=1}^{n_d} \tilde{w}_{di} \boldsymbol{x_{di}} = \boldsymbol{X}$ dove $\boldsymbol{x_{di}}$ rappresenta il vettore delle osservazioni ausiliarie relative a una singola unità campionaria e $\boldsymbol{X}$ è il vettore dei totali (noti) a livello dell'intera popolazione per le variabili ausiliarie. I pesi vengono ottenuti, nel rispetto del vincolo di calibrazione, minimizzando un'opportuna funzione distanza dai pesi originari w_{di}. Se si adotta una funzione di distanza quadratica nel calcolo dei pesi calibrati, gli stimatori per calibrazione assumono la forma di stimatori per regressione generalizzata; questa forma caratterizza a livello asintotico qualunque stimatore per calibrazione, indipendentemente dalla funzione di distanza scelta, sotto le condizioni descritte in Deville e Sarndal (1992). Lo stimatore per regressione generalizzata applicato ai domini di studio (d-GREG) assume la seguente forma:

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d^{HT} + (\bar{\boldsymbol{X}}_d - \hat{\bar{\boldsymbol{X}}}_d^{HT})^T \hat{\boldsymbol{B}} \tag{5}$$

dove $\bar{\boldsymbol{X}}_d = \bar{X_{d1}}, \ldots, \bar{X_{dp}}$ è il vettore delle medie di dominio note per le variabili ausiliarie e $\hat{\bar{\boldsymbol{X}}}_d^{HT}$ è il vettore dei corrispondenti stimatori di Horwitz-Thompson. Riguardo al coefficiente $\hat{\boldsymbol{B}}$ esso è definito come

$$\hat{\boldsymbol{B}} = \left(\sum_{d=1}^{D} \sum_{i=1}^{n_d} \frac{w_{di} \boldsymbol{x_{di}} \boldsymbol{x_{di}}^T}{c_{di}} \right)^{-1} \sum_{d=1}^{D} \sum_{i=1}^{n_d} \frac{w_{di} \boldsymbol{x_{di}} y_{di}}{c_{di}} \tag{6}$$

dove c_{di} è un insieme di costanti opportunamente scelte dal ricercatore, ad esempio per tener conto dell'eterogeneità dei residui associati alla relazione lineare tra la variabile obiettivo e quelle ausiliarie. E' evidente come lo stimatore $\hat{\boldsymbol{B}}$ presenti la struttura di uno stimatore di minimi quadrati. Lo stimatore per regressione generalizzato permette di raggiungere dei guadagni di efficienza notevoli rispetto allo stimatore $\hat{\bar{Y}}_d^{HT}$ ogni qual volta il vettore delle variabili $x_1, \ldots, x_p$ presenti un forte legame lineare con la variabile obiettivo. Nonostante l'obiettivo sia stimare la media a livello di dominio, $\hat{\boldsymbol{B}}$ è stimato sui dati di tutto il campione, assumendo che la relazione che lega variabile obiettivo e variabile ausiliaria sia la stessa per tutti i domini della popolazione. Lo stimatore d-GREG non è però una semplice previsione della media di y basata sulla relazione stimata con le variabili ausiliarie. Infatti può essere presentato come

$$\hat{\bar{Y}}_d^{GREG} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{B}} + \left(\hat{\bar{Y}}_d^{HT} - \hat{\boldsymbol{B}} \hat{\bar{\boldsymbol{X}}}_d^{HT} \right) \tag{7}$$

In modo da mettere in evidenza due componenti. La prima è uno ?stimatore sintetico? ossia la previsione della media di y basata sulla relazione stimata con le variabili ausiliarie che sfrutta il fatto che $\bar{\boldsymbol{X}}_d^T$ sia noto:

$$\hat{\bar{Y}}_d^{SYN} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{B}} \tag{8}$$

Lo stimatore $\hat{\bar{Y}}_d^{SYN}$ può essere utilizzato direttamente quando la relazione lineare tra le $x_1, \ldots, x_p$ e y è forte e si ha ragione di ritenere che nessuno dei domini devi in modo marcato da questa relazione, ovvero nessun dominio sia caratterizzato da un'eterogeneità non spiegata dalle variabili ausiliarie. Se queste condizioni sono soddisfatte si tratta di stimatori potenzialmente molto efficienti. Tuttavia, essendo "ciechi" rispetto a qualsiasi specificità propria del dominio di studio, possono essere, se le condizioni citate non sono soddisfatte, molto distorti. La seconda componente di $\hat{\bar{Y}}_d^{GREG}$ ha appunto lo scopo di correggere questa potenziale distorsione:

$$\left(\hat{\bar{Y}}_d^{HT} - \hat{\boldsymbol{B}} \hat{\bar{\boldsymbol{X}}}_d^{HT} \right) = \frac{1}{N_d} \sum_{i=1}^{n_d} w_{di} \left(y_{di} - \hat{\boldsymbol{B}} \boldsymbol{x_{di}} \right) \tag{9}$$

Essa può essere infatti letta come una correzione dello stimatore sintetico operata attraverso una stima della media dei residui di regressione specifica di ciascuna area. Lo stimatore $\hat{\bar{Y}}_d^{GREG}$ può rappresentare, qualora si disponga di un vettore di variabili ausiliarie con una buona capacità predittiva, una buona soluzione per la stima a livello di dominio. Tuttavia soffre di alcune limitazioni:

- non è robusto rispetto alla presenza di outliers nel campione;
- la sua efficienza è spesso inadeguata quando i domini di studio sono particolarmente piccoli.

Riguardo alla robustezza possiamo notare che outliers presenti "ovunque" nel campione possono influenzare la stima $\hat{\boldsymbol{B}}$ - si tratta infatti di uno stimatore per minimi quadrati; outliers specifici del dominio di studio possono invece avere un indebito impatto sulla seconda componente dello stimatore. Questa seconda componente ha a che fare con la relativa "inefficienza" di $\hat{\bar{Y}}_d^{GREG}$: la stima della media dei residui di area, se basata su un numero esiguo di osservazioni può essere instabile e implica una relativa inefficienza degli stimatori. Più in generale, quando la dimensione dei campioni specifici nei domini è nella maggior parte dei casi piccola si ricorre a stimatori "per piccole aree" che estendono la logica di sfruttamento delle informazioni ausiliarie un passo oltre, cercando di stimare in modo più efficiente la seconda componente dello stimatore $\hat{\bar{Y}}_d^{GREG}$ o, più in generale attraverso una caratterizzazione efficiente dell'eterogeneità del dominio rispetto alla relazione tra le y e le variabili ausiliarie.

A.2 Stimatori per piccole aree

Di seguito presenteremo alcuni stimatori per piccole aree, limitandoci a stimatori che godano di buone proprietà rispetto al disegno e in particolare della coerenza rispetto al disegno (design consistency; si veda Fuller, 2009, p. 41 per una definizione). Alla luce del contesto dell'applicazione in questo lavoro una particolare attenzione sarà dedicata al problema della robustezza rispetto alla presenza di osservazioni outlier. Il primo stimatore che prendiamo in considerazione è lo pseudo-EBLUP proposto da Rao e You (2002). Alla base della derivazione di questo stimatore, viene esplicitamente assunto un modello lineare misto a livello di singola unità campionata (impresa)

$$y_{id} = \boldsymbol{x}_{id}^t \boldsymbol{\beta} + \nu_d + e_{di} \tag{10}$$

$\nu_d : N(0, \sigma_\nu^2)$, $e_{di} : N(0, \sigma_e^2)$. ν_d è un effetto casuale utile per modellare la correlazione dei residui nella stessa area. La necessità di generalizzare il modello di regressione lineare semplice, implicitamente alla base del d-GREG che abbiamo discusso in precedenza, introducendo un effetto casuale è giustificata da Skinner (1991) alla luce del fatto che, nelle applicazioni i domini sono spesso caratterizzati da un'eterogeneità latente che le variabili ausiliarie non sono in grado di spiegare. Sulla base del modello lineare misto si può ottenere il miglior predittore lineare (BLUP) della media $\hat{\bar{Y}}_d$

$$\hat{\bar{Y}}_d^{p-BLUP} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_w + \gamma_d \left(\hat{\bar{Y}}_d^{HT} - \hat{\bar{\boldsymbol{X}}}_d^{HT} \hat{\boldsymbol{\beta}}_w \right) \tag{11}$$

$$= \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_w + \gamma_d \frac{1}{N_d} \sum_{i=1}^{n_d} w_{di} \left(y_{di} - \boldsymbol{x_{di}}^T \hat{\boldsymbol{\beta}}_w \right) \tag{12}$$

dove $\gamma_d = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_e^2 \sum_{i=1}^{n_d} w_{id}}$ e $\hat{\boldsymbol{\beta}}_w = \hat{\boldsymbol{\beta}}_w(\sigma_\nu^2, \sigma_e^2)$ è uno stimatore dei minimi quadrati generalizzati che tiene in considerazione il fatto che le osservazioni non siano assunte indipendenti e identicamente distribuite. La struttura è simile a quella di d-GREG evidenziata in A.1; tuttavia la componente di correzione basata sui residui specifici di dominio riceve un peso $\gamma_d \in [0, 1]$ che è funzione crescente di σ_ν^2, un parametro che descrive l'eterogeneità dei domini non spiegata dalle variabili ausiliarie. Le componenti della varianza che compaiono in γ_d e implicitamente in $\hat{\boldsymbol{\beta}}_w$ devono essere stimate. You e Rao (2002) propongono di utilizzare il metodo della massima verosimiglianza o della massima verosimglianza ristretta. Lo stimatore empirico corrispondente a $\hat{\bar{Y}}_d^{p-BLUP}$ sarà ottenuto sostituendo ai valori incogniti σ_ν^2, σ_e^2 e loro stime $\hat{\sigma}_\nu^2, \hat{\sigma}_e^2$ (il coefficiente $\hat{\gamma}_d$ è definito come sopra, ma operando questa sostituzione):

$$\hat{\bar{Y}}_d^{p-EBLUP} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_w + \hat{\gamma}_d \frac{1}{N_d} \sum_{i=1}^{n_d} w_{di} \left(y_{di} - \boldsymbol{x_{di}}^T \hat{\boldsymbol{\beta}}_w \right) \tag{13}$$

Si tratta di uno stimatore consistente rispetto al disegno; tuttavia la sua somiglianza con A.1 ci permette di ripetere la stessa osservazione in merito alla robustezza. Una generalizzazione del d-GREG, che, seppur basata su modello, non fa ricorso ad ipotesi distributive sulle componenti stocastiche del modello e si mostra robusta rispetto alla presenza di outliers è basata sull'utilizzo della regressione quantilica, che può essere descritta da questa famiglia di relazioni, che possiamo scrivere in funzione di ciascun quantile $q \in (0, 1)$.

$$y_{id} = \boldsymbol{x}_{id}^t \boldsymbol{\beta}(q) + \nu_d + e_{di}(q) \tag{14}$$

dove i residui godono della proprietà $q(e_{di}(q)) = 0$. Se $q = 0.5$ otteniamo la regressione mediana a cui è naturalmente associato lo stimatore di $\boldsymbol{\beta}(0.5)$ che minimizza gli scarti in valore assoluto, una tradizionale alternativa robusta allo stimatore dei minimi quadrati. La regressione quantilica generalizza la regressione mediana assumendo che esista un vettore di coefficienti di regressione specifico per ciascun quantile della distribuzione. Chambers e Tzavidis (2006) per primi hanno proposto l'applicazione della regressione quantilica alla stima per piccole aree. Questi autori osservano come ciascuna singola osservazione y_{di} giaccia sul piano di regressione quantilica caratteristico di un certo quantile q_{di}, e come si possa quindi sempre definire un $q_{di} \in (0, 1)$ tale per cui $y_{id} = \boldsymbol{x}_{id}^t \boldsymbol{\beta}(q)$. Se, condizionatamente alle variabili ausiliarie rimane dell'eterogeneità non spiegata a livello di dominio, i quantili specifici di osservazioni

appartenenti allo stesso dominio saranno simili tra loro: la specificità di ciascun dominio potrà essere quindi caratterizzata da un coefficiente di regressione specifico di dominio $\boldsymbol{\beta}(\bar{q}_d)$ dove $\bar{q}_d = n_d^{-1} \sum_{i=1}^{n_d} q_{di}$. Una descrizione, ancorché sommaria degli stimatori impiegati nella stima dei parametri della regressione quantilica è oltre gli obiettivi di questo documento; per questo si rimanda alla vasta letteratura sull'argomento (si vedano ad esempio Koenker e Bassett, 1978 oppure Koenker, 2005). Riguardo a Chambers e Tzavidis (2006) notiamo come nella loro proposta il concetto di quantile sia sostituito da quello più generale di M-quantile, in cui il ruolo della funzione di perdita assoluta viene sostituito da una qualsiasi funzione di perdita che rispetti determinate caratteristiche e in particolare l'unicità del quantile per cui $y_{id} = \boldsymbol{x}_{id}^t \boldsymbol{\beta}(q_{di})$. Il q-esimo m-quantile di una variabile casuale y e indicato con θ_q è definito come la quantità che minimizza

$$\int \rho_q(y - \theta_q) dF_y \tag{15}$$

dove F_y è la funzione di ripartizione di y e $\rho_q(u) =\mid q - I(u < 0) \mid \rho(u)$, è una funzione di perdita convessa e differenziabile. Notiamo come $\rho_q(u) = |u|$ permetta di ri-ottenere la definizione ordinaria di quantile. La funzione di influenza $\psi(u) = d\rho(u)/du$ svolge un ruolo importante nella regressione m-quantilica. Gli stimatori per piccole aree proposti in questo contesto da Chambers e Tzavidis (2006) e Tzavidis et al. (2010) sono stati applicati al contesto della stima per popolazioni finite secondo un'impostazione model-assisted da Fabrizi et al. (2014) con l'obiettivo di ottenere stimatori consistenti rispetto al disegno. In particolare per la stima della media a livello di dominio consideriamo il seguente stimatore:

$$\hat{\bar{Y}}_d^{WMQ-bc} = \bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_{w,\bar{d}} + \frac{1}{N_d} \sum_{i=1}^{n_d} w_{di} \left(y_{di} - \boldsymbol{x_{di}}^T \hat{\boldsymbol{\beta}}_{w,\bar{d}} \right) \tag{16}$$

Anche in questo caso la struttura è molto simile a A.1; le differenze principali consistono nel ricorso ad un vettore di coefficienti di regressione specifico per il quantile medio di dominio (in grado di catturare l'eterogeneità non spiegata a livello di dominio) e nell'utilizzo di uno stimatore, $\hat{\boldsymbol{\beta}}_{w,\bar{d}}$ robusto rispetto agli outliers che incorpora anche i pesi campionari nel processo di stima. Fabrizi et al. (2014) dimostrano come $\hat{\bar{Y}}_d^{WMQ-bc}$ sia uno stimatore consistente rispetto al disegno di $\bar{Y}$. Anche se non messo in evidenza ai fini di semplificare la notazione lo stimatore $\hat{\boldsymbol{\beta}}_{w,\bar{d}}$ dipende dalla scelta di una funzione di perdita $\rho(u)$ o equivalentemente dalla funzione di influenza $\psi(u)$.
Anche in $\hat{\bar{Y}}_d^{WMQ-bc}$ possiamo mettere in evidenza due componenti:

- $\bar{\boldsymbol{X}}_d^T \hat{\boldsymbol{\beta}}_{w,\bar{d}}$ è lo stimatore originariamente proposto da Chambers e Tzavidis (2006) modificato per incorporare i pesi campionari nella stima di $\boldsymbol{\beta}(\bar{q}_d)$ e che per questo possiamo indicare con $\hat{\bar{Y}}_d^{WMQ}$. Si tratta di uno stimatore robust projective nel senso che all'espres-

sione da Chambers (1986). La stima del coefficiente $\boldsymbol{\beta}(\bar{q}_d)$ è infatti robusta rispetto alla presenza di outliers. Questo stimatore, in presenza di una forte relazione lineare tra y e variabili ausiliarie può avere una varianza campionaria molto piccola, ma la sua distorsione può essere considerevole ($\boldsymbol{\beta}(\bar{q}_d)$ è più flessibile rispetto all'adozione di un unico coefficiente per l'intera popolazione, ma la struttura rimane quella di un semplice stimatore di proiezione);

- $\frac{1}{N_d}\sum_{i=1}^{n_d} w_{di}\left(y_{di} - \boldsymbol{x_{di}}^T\right)\hat{\boldsymbol{\beta}}_{w,\bar{d}}$ rappresenta una componente di correzione della distorsione non dissimile da quella impiegata in A.1. Questa componente non è robust predictive (Chambers, 1986): essa è infatti sensibile alla presenza di outliers nel campione specifico di dominio e può risultare abbastanza instabile.

Fabrizi et al. (2014) notano come $\hat{\bar{Y}}_d^{WMQ-bc}$ sia in generale più efficiente di $\hat{\bar{Y}}_d^{GREG}$, come questo caratterizzato da una distorsione contenuta (a differenza di $\bar{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}}_{w,\bar{d}}$) ma, in presenza di residui outliers caratterizzato da una varianza elevata. Per ottenere uno stimatore che sia ad un tempo robust projective e robust predictive, Chambers et al. (2014) propongono di basare la correzione della distorsione non già sui residui grezzi, ma su una loro trasformazione che limiti l'influenza di quelli più estremi. La loro proposta è avanzata in un contesto model based in cui i pesi campionari non vengono considerati. Una estensione del loro stimatore che incorpori anche i pesi campionari e sia quindi design-consistent è data dallo stimatore seguente:

$$\hat{\bar{Y}}_d^{WMQ-\phi bc} = \bar{\boldsymbol{X}}_d^T\hat{\boldsymbol{\beta}}_{w,\bar{d}} + \frac{1}{N_d}\sum_{i=1}^{n_d} w_{di}\phi\left(y_{di} - \boldsymbol{x_{di}}^T\hat{\boldsymbol{\beta}}_{w,\bar{d}}\right) \tag{17}$$

dove $\phi(u)$ è una funzione di influenza limitata, monotona non decrescente definita su tutto l'asse reale e tale che $\phi(0) = 0$. Questa funzione di perdita può essere diversa dalla ψ utilizzata nello stimatore $\hat{\boldsymbol{\beta}}_{w,\bar{d}}$; in pratica è consigliabile sceglierla in modo tale che $|\rho(u)| \leq |\phi(u)|$, ossia tale da essere meno "aggressiva" nella riduzione dell'impatto degli outliers.

B I disegni di campionamento nelle indagini sulle Imprese

Le rilevazioni sulle imprese hanno lo scopo di fornire dati su una pluralità di fenomeni di interesse, con il duplice impiego sia nell'analisi microeconomica che in quella aggregata, che fornisce stime relative all'intera popolazione e su specifici domini di studio. [5]

Una delle caratteristiche principali delle indagini sulle imprese è la forte asimmetria destra della distribuzione delle variabili di interesse. Questo fenomeno è particolarmente rilevante nel nostro Paese dove la struttura produttiva è caratterizzata da un numero molto elevato di piccole e medie imprese e da un numero piuttosto ridotto di grandi imprese.

[5] Per dominio di studio si intende una sottopopolazione individuata da una partizione (detta tipo di dominio) della popolazione oggetto di indagine.

Per questo motivo nell'ambito delle principali rilevazioni sulle imprese la popolazione delle unità viene normalmente suddivisa in due sottoinsiemi:

- le unità di maggiore dimensione (in termini di addetti o di fatturato) che entrano con certezza nel campione;
- le rimanenti unità di piccola dimensione che vengono osservate secondo un preciso schema di campionamento.

La scelta della popolazione di imprese da inserire con certezza nei campioni, la cosiddetta popolazione *censita*, può essere effettuata sulla base della rilevanza delle osservazioni (ovvero al contributo alla formazione dell'aggregato nazionale), oppure secondo criteri dettati dalle necessità organizzative o da propositi di riduzione della pressione statistica sulle imprese di minori dimensioni. In letteratura sono presenti alcune tecniche che permettono di definire la soglia di censimento secondo diversi criteri di ottimalità (Hidiroglou M.A., 1996).

Le rimanenti unità, che costituiscono la popolazione *campionata*, sono di solito osservate tramite un disegno di campionamento ad uno stadio stratificato, con selezione delle unità con probabilità uguali e senza reimmissione nell'ambito di ciascuno strato.

La stratificazione di base delle unità nella popolazione viene ottenuta incrociando le modalità delle variabili strutturali che identificano i tipi di dominio; in tal modo si ottiene la partizione meno fine della popolazione che consente di ottenere ciascun dominio come unione di strati elementari completi.
Per aumentare l'efficienza delle stime campionarie, gli strati di base possono essere ulteriormente suddivisi in modo da avere una partizione della popolazione in gruppi ancora più omogenei rispetto ad una o più variabili di interesse. Tuttavia, esiste un trade off tra la granularità della stratificazione e la dimensione campionaria ideale: all'aumentare del numero di strati si ottiene da un lato una maggiore precisione nelle stime dall'altro un minor numero di unità all'interno di ogni strato, che comportano una più elevata complessità organizzativa e un aumento dei costi per l'osservazione delle unità selezionate. Fissato un tipo di dominio, le unità di ciascuno strato appartengano ad uno solo dei suoi domini. In tal modo i domini di stima costituiscono *domini pianificati* (Cicchitelli et al. 1992).

Il fatto che tutti i domini di interesse siano di tipo pianificato presenta alcuni considerevoli vantaggi dal punto di vista della progettazione della rilevazione. In particolare, ciò permette di allocare il campione negli strati predefinendo i livelli di precisione attesa delle stime su tutti i domini di interesse.

B.1 Problema dell'allocazione nei disegni a stratificazione classica e a stratificazione incompleta

L'allocazione del campione negli strati viene realizzata secondo un criterio di tipo multivariato e multidominio. Tale criterio consiste in una generalizzazione del metodo di Neyman che consente di minimizzare la dimensione campionaria per determinati errori di campionamento attesi delle variabili d'interesse, relativamente a ciascun dominio di stima.

Per descrivere dal punto di vista formale quanto appena illustrato, denotiamo con U_h la popolazione di dimensione N_h del generico strato $h(h = 1, \ldots, H)$. Indichiamo inoltre con s^* il campione (selezionato senza reimmissione delle unità e con probabilità costanti per strato) di dimensione n^* e con s_h^* il campione di dimensione n_h^* dello strato h.

In un disegno stratificato "classico", individuare l'allocazione campionaria n_h^* equivale a definire il vettore di probabilità di inclusione π_h^* ottimale, dove: $pr[u_k \in s_h] = \pi_h^* = \frac{n_h^*}{N_h}$ e la probabilità di inclusione semplice dell'unità k-esima.

Si supponga di essere in assenza di mancate risposte e di voler produrre stime efficienti del totale ${}_pY_d$ relativo alla variabile di interesse Y_p sul dominio di stima d con $(p = 1, \ldots, P; d = 1, \ldots, D)$. Utilizzando lo stimatore di Horvitz-Thompson, l'espressione della stima del totale ${}_pY_d$ è data da:

$$ {}_p\tilde{Y}_d = \sum_{h=1}^{H} \sum_{k=1}^{n_h^*} \frac{N_h}{n_h^*} y_{pd,k} \tag{18} $$

L'efficienza della stima può essere formulata in termini di varianza attesa dello stimatore sotto il disegno prescelto, ossia, considerando le ipotesi precedenti, da:

$$ V({}_p\tilde{Y}_d) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h^*)}{n_h^*(N_h^* - 1)} \sum_{k=1}^{N_h} (y_{pd,k} - \frac{Y_{pd}}{N_h})^2 \tag{19} $$

Dal punto di vista formale, l'allocazione campionaria viene quindi determinata mediante la soluzione del seguente sistema di programmazione lineare:

$$ \sum_{k \in U_h} \pi_h c_h = \min \tag{20} $$

$$ V({}_p\tilde{Y}_d) \leq {}_p\tilde{Y}_d^*, \qquad p = 1, \ldots, P; d = 1, \ldots, D; h = 1, \ldots, H \tag{21} $$

$$ 0 < \pi_h \leq 1 \tag{22} $$

dove si è indicato con c_h il costo di rilevazione di una unità campionaria nello strato U_h; ${}_pV_d^*$ il valore massimo ammesso della varianza della stima $V({}_p\tilde{Y}_d)$.

In un disegno a stratificazione *completa* il vincolo sulla numerosità campionaria ottima, $n_h^* = N_h \pi_h^*$, è che sia intera positiva *per ogni strato*; si impone $n_h^* \geq 2$ per assicurare una stima consistente della varianza di strato, con l'ovvia eccezione in cui $n_h^* = \pi_h^* = 1$ quando $N_h = 1$.

La soluzione del sistema di ottimo vincolato, rappresentata da un vettore di numerosità campionarie di strato $[n_h]$ con $h = 1, \ldots, H$, può essere ottenuta mediante l'algoritmo illustrato in Bethel (1989). Tale algoritmo individua la soluzione ottima del problema di allocazione multivariata in maniera iterativa, partendo da una soluzione iniziale che coincide con la soluzione ottima nel caso univariato per la prima variabile sul primo dominio. In ciascuno dei passi successivi, la numerosità campionaria viene aumentata minimizzando la funzione obiettivo fino al soddisfacimento di tutti i vincoli. Bethel dimostra che tale algoritmo converge.

Un aspetto critico del disegno a stratificazione *completa o standard* è rappresentato dal vincolo di dimensione n_h^* intera positiva del campione di strato, che implica che gli strati di popolazione di bassissima numerosità ($N_h \leq 2$) vengano di fatto censiti; quando gli strati esigui sono numerosi, il doverli censire fa sì che in molti domini risulti allocato un campione totale più numeroso di quello strettamente necessario per soddisfare il vincolo di errore campionario massimo fissato sul dominio di interesse.

Per evitare che l'allocazione risulti inutilmente sovradimensionata rispetto alla precisione attesa richiesta per le stime, è utile in molti casi adottare un disegno a stratificazione *incompleta*, la cui caratteristica essenziale è la possibilità di allocare negli strati la numerosità campionaria esatta per realizzare un errore atteso pari all'errore massimo pianificato.

Il problema di ottimizzazione si formalizza pertanto come in precedenza, ma con i vincoli espressi come segue:

$$\sum_{h \in H_{jd}} \pi_h = 1 \quad quando \quad N_{jd} = 1 \tag{23}$$

$$n_{jd}^* = \sum_{h \in H_{jd}} N_h \pi_h^* \geq 1 \tag{24}$$

con $n_h^* = N_h$ e π_h^* anche molto piccolo o prossimo a 0 per qualche $h \in H_{jd}$ (alcuni strati possono risultare non campionati), ma n_{jd}^* intero positivo a livello di dominio j_d.

La soluzione del problema consiste quindi nella determinazione di un vettore di probabilità di inclusione ottime $[\pi_k^*]$ con $k \in U_k$, il quale fornisce una dimensione ottima attesa sul dominio di interesse: la numerosità campionaria da allocare nel dominio è infatti realizzata come media delle probabilità di inclusione delle unità negli strati che compongono il dominio j_d.

La ricerca della soluzione ottima nel caso di un disegno a stratificazione incompleta può essere realizzata con il software 'MultiWay Sample Allocation' (Istat, 2016).