Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Dottorato di Ricerca in Statistica Applicata
XX ciclo SECS-S/01

Il movimento dei clienti nelle strutture turistico ricettive della provincia di Firenze: stime anticipate da un campione autoselezionato

Graziano Scaffai

Tutor: **Prof. Andrea Giommi**

Co-tutor: **Prof. Monica Pratesi, Dott. Stefano Falorsi**

Coordinatore: **Prof. Guido Ferrari**

Ringraziamenti

Il presente lavoro è stato possibile grazie alla disponibilità dell'APT di Firenze nel rendere accessibili gli archivi relativi alla consistenza delle strutture ricettive e al movimento dei clienti, nonché alla collaborazione degli addetti della U. O. Strutture Ricettive e Statistica che hanno fornito il necessario supporto per l'utilizzo degli archivi medesimi.

Ringrazio la Regione Toscana per i dati messi a disposizione, in particolare la dott.ssa Francesca Dodero per le informazioni fornite, indispensabili per un corretto uso degli archivi.

Un grazie infine a mia moglie Luana e mia figlia Claudia che hanno pazientemente sopportato per tre anni le mie assenze, non di rado anche quand'ero presente.

INDICE

Capitolo 1

Capitolo 2

Capitolo 3

Capitolo 4

Capitolo 5

Capitolo 6

Capitolo 7

Capitolo 8 – Conclusioni

Riferimenti bibliografici

Capitolo 1

1.1. Obiettivo della tesi

Nella attuale tendenza a una localizzazione della produzione di beni e servizi sempre più indipendente dai confini nazionali, il settore economico del turismo può costituire un elemento strategico per lo sviluppo dei paesi avanzati, in particolare per l'Italia e le sue regioni grazie alla presenza di risorse non rinnovabili quali quelle costituite dal patrimonio artistico e culturale o difficilmente rinnovabili quali quelle ambientali integrate da un sistema diffuso di strutture di accoglienza.

La regione Toscana, la Provincia e la città di Firenze in particolare si collocano nelle prime posizioni quanto a risorse e strutture ricettive rendendole già oggi destinazioni rilevanti dei flussi turistici nazionali e internazionali.

Nel Prospetto 1, tratto dal "Movimento Turistico 2004" pubblicato dalla Regione Toscana è calcolato l'indice di turisticità per le tipologie di risorsa turistica prevalente della regione. Tale indice è dato dal numero medio di turisti presenti giornalmente su 1.000 residenti. Il luogo comune di città prese d'assalto dai turismi è infondato, concentrandosi questi in piccole porzioni delle città. Sono piuttosto i comuni della costa e quelli termali dove il numero dei turisti rispetto ai residenti ha un rilievo maggiore. Per la provincia di Firenze tale indice nel 2004 è stato di 26,8 turisti per 1.000 abitanti.

Prospetto 1 - Indice di turisticità delle tipologie di risorsa turistica della Toscana, anno 2004

Risorsa prevalente	Numero di comuni	Popolazione media 2004	Superficie (Kmq)	Densità popolazione	Densità turisti	Indice di turisticità
Balneare	33	676.178	2.986,6	226,4	13,3	58,8
Arte/Affari	63	1.762.683	6.307,0	279,5	5,5	19,6
Termale	18	141.924	1.717,6	82,6	5,5	66,0
Campagna/Collina	85	394.315	5.893,8	66,9	1,1	15,9
Altro Interesse	33	435.373	2.057,8	211,6	2,4	11,4
Montagna	55	171.698	4.027,3	42,6	0,6	14,7
Totale	287	3.582.170	22.990,2	155,8	4,2	27,2

Fonte: Regione Toscana, Settore Statistica.

Queste cifre, pur riferite al movimento dei clienti nelle strutture ricettive come in seguito specificato, indicano che c'è ancora spazio per i turisti purché si riesca a intercettare in tempo l'evoluzione della domanda. Ne è un esempio il successo delle strutture agrituristiche negli ultimi anni.

Un contributo alla gestione e sviluppo del settore turistico è fornito da una conoscenza sia strutturale sia congiunturale del fenomeno che permetta interventi tempestivi ed efficaci.

La principale informazione sul turismo deriva dalla rilevazione mensile Istat del movimento dei clienti nelle strutture turistico-ricettive: questa avviene con una indagine censuaria giornaliera presso tutte le tipologie ricettive del territorio nazionale. Nel successivo paragrafo esamineremo in dettaglio lo svolgimento di questa rilevazione nella provincia di Firenze.

I risultati, sia a livello locale (province, regioni) sia, a maggior ragione, a livello nazionale sono disponibili dopo molti mesi da quello cui si riferiscono, mentre agli utenti (amministratori ed esercenti in primo luogo) interessa disporre di informazione tempestiva sull'andamento congiunturale del fenomeno: quante sono state nel mese t le presenze (le notti vendute) e gli arrivi (le richieste di alloggio) e in particolare se queste quantità sono aumentate o diminuite e di quanto rispetto al corrispondente mese dell'anno precedente.

Il fenomeno in questi anni è soggetto a una evoluzione lenta, con variazioni mensili fisilogiche piuttosto modeste (entro il ±10%) salvo eventi eccezionali; pertanto qualunque informazione tempestiva avrà un valore se è in grado di cogliere variazioni di questa entità.

In questa tesi ci poniamo l'obiettivo di ottenere per la provincia di Firenze stime molto tempestive del totale delle presenze e degli arrivi utilizzando un sottoinsieme di strutture ricettive in grado di trasmettere via web alla Provincia i dati sul proprio movimento entro pochi giorni dalla fine del mese. Vogliamo poi valutare se e quanto tali stime sono in grado di soddisfare l'esigenza informativa sopra espressa; ciò è possibile, almeno empiricamente a posteriori, grazie all'esistenza dell'indagine censuaria.

Questo obiettivo si inserisce nel filone di ricerca, prevalente nella statistica pubblica, che va sotto il nome di stime anticipate, intendendo con ciò la produzione e diffusione di dati a un tempo precedente quello previsto dalla conclusione dell'indagine sia essa censuaria o campionaria. In pratica si concede qualcosa sul versante dell'accuratezza per ottenere maggior tempestività. Del resto la qualità non è un concetto assoluto ma è relativa alle esigenze dell'utente il quale, a fronte di una maggior tempestività, può essere disposto ad accettare una minor accuratezza.

1.2. L'indagine sul movimento dei clienti nelle strutture turistico-ricettive

La rilevazione del movimento dei clienti nelle strutture turistico-ricettive, rileva il movimento che si verifica nelle strutture che svolgono l'attività economica di offerta di alloggio. Non rientrano in questa rilevazione i movimenti che si verificano senza l'acquisto di almeno un pernottamento quali quelli legati all'escursionismo giornaliero o quelli delle vacanze in seconde case o in case di amici e parenti.

Il turismo è un fenomeno complesso dai confini non netti riferendosi ai comportamenti delle persone che non possono essere assegnati, anche in brevi segmenti del tempo, totalmente a una categoria (studente, consumatore, lavoratore, viaggiatore, turista, ecc.). Così un parente può alloggiare in un albergo di Firenze per assistere un malato ricoverato in ospedale, un impiegato può trascorrere un periodo di ferie verniciando le persiane della casa al mare. Il primo non mancherà di fare una visita alla città, il secondo qualche bagno.

La rilevazione che consideriamo si riferisce appunto ai clienti delle strutture turistico-ricettive. Tuttavia, sebbene non esistano rilevazioni sistematiche sulle motivazioni per cui una persona acquista alloggio per alcuni giorni presso una struttura ricettiva, normalmente questo avviene per motivo di svago, riposo, visita a nuovi luoghi, ecc. Insomma per quell'insieme di motivi che rendono la persona "etichettabile prevalentemente" come turista.

Si può quindi concludere che la rilevazione del movimento dei clienti nelle strutture ricettive misura quella consistente quota del turismo che si realizza utilizzando l'offerta imprenditoriale di alloggio.

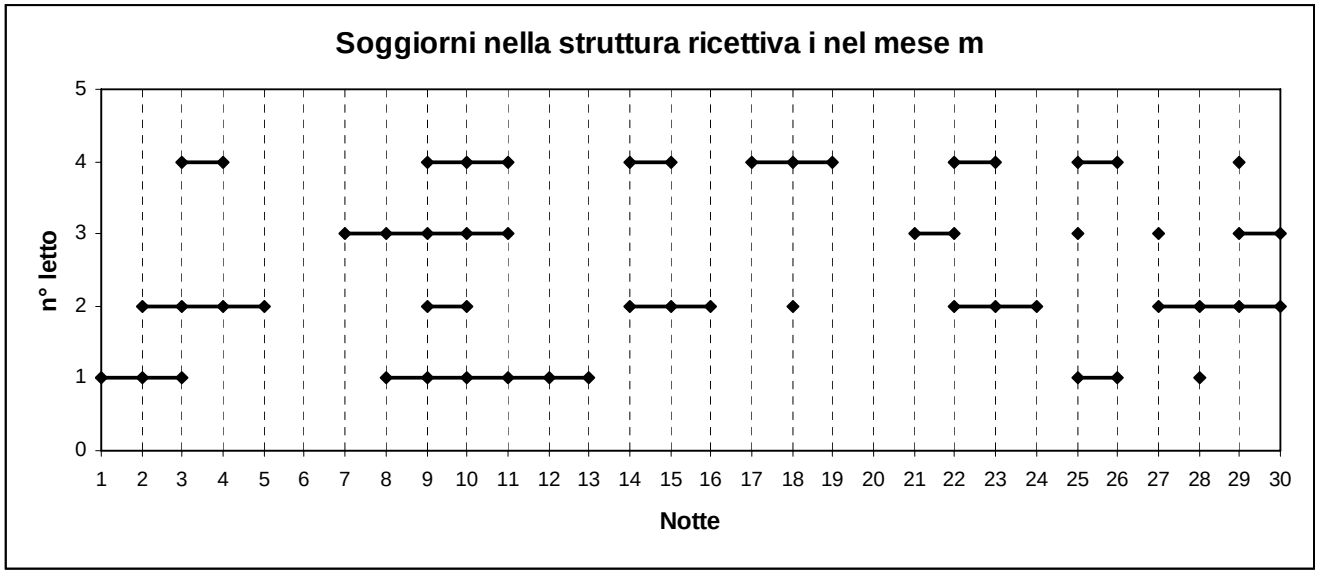
1.3. Definizione del modello concettuale statistico

Formalmente la popolazione statistica è costituita dai "soggiorni" che si realizzano nelle strutture turistico-ricettive presenti in un dato territorio (nel caso di nostro interesse la provincia di Firenze) in un fissato intervallo di tempo: di norma il mese, anche se interessano periodi più lunghi come i trimestri, l'intero anno, o più brevi come quelli settimanali a cavallo di date particolari dell'anno: la Pasqua, il Ferragosto, il Natale, di particolare rilevanza per il fenomeno.

Per queste unità statistiche si rilevano i seguenti caratteri: notte (data) di inizio, notte di fine, provenienza della persona che acquista il soggiorno (nazione per i clienti stranieri, regione per quelli italiani), alcune informazioni anagrafiche del cliente. La struttura turistico ricettiva presso la quale il soggiorno si realizza è l'unità di rilevazione .

Fissato un intervallo di tempo cui vogliamo riferire il movimento turistico e una struttura ricettiva, il manifestarsi dei soggiorni può essere schematizzato con il Grafico 1.1 seguente in cui per semplicità consideriamo un'ipotetica struttura che disponga di soli quattro posti letto.

Grafico 1.1

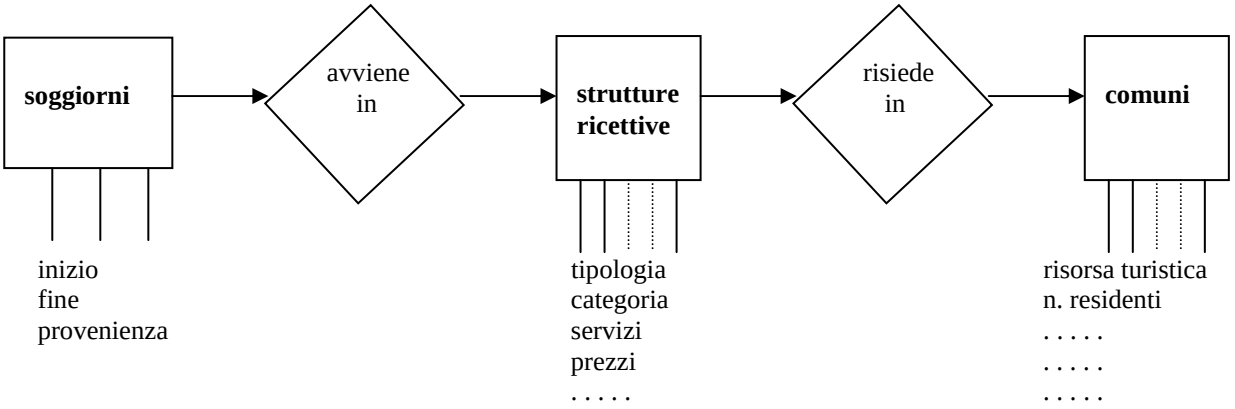


Caratterizzando i soggiorni con la loro notte di inizio e di fine, questi due estremi possono presentarsi in mesi diversi, per cui non è possibile in generale assegnare a tutti i soggiorni il mese.

Nell'esempio del grafico, il primo soggiorno nel posto letto 1 potrebbe essere iniziato sia nella prima notte del mese sia in una notte del mese precedente o addirittura di altro precedente; analogamente gli ultimi due soggiorni dei posti letto n° 2 e 3 potrebbero terminare con la 30-esima notte del mese o proseguire. La durata dei soggiorni può quindi risultare censurata a destra come a sinistra. Da qui deriva la impossibilità di ottenere una esatta distribuzione dei soggiorni per numero di notti e la relativa intensità in un dato periodo temporale.

Sull'unità statistica soggiorno si rilevano solo tre caratteri: notte di inizio, notte di fine, provenienza. Le statistiche su questa popolazione sarebbero pertanto assai povere. Se però consideriamo la relazione di questa popolazione con quella delle strutture ricettive e di questa seconda con la popolazione dei comuni, ecco che è possibile associare ai soggiorni numerosissimi caratteri. La figura seguente, schematizza le tre popolazioni ("entità" nel gergo informatico) e le relazioni fra di esse.

Figura 1.1



Alcune informazioni anagrafiche dei clienti vengono raccolte dalla struttura ricettiva a fini di pubblica sicurezza, ma non sono oggetto degli adempimenti previsti dalla rilevazione statistica.

Nella realizzazione pratica dell'indagine si aggregano le informazioni sui soggiorni a livello della unità statistica struttura ricettiva, che è l'unità di rilevazione, e per questa si ottengono, per ciascuna provenienza, i due caratteri: il numero di soggiorni che in essa hanno avuto inizio in un determinato mese e il numero di soggiorni in essere nelle notti del mese: queste due grandezze prendono in nome rispettivamente di *arrivi* e *presenze*.

Dalla conoscenza della struttura ricettiva in cui si manifestano i soggiorni e quindi il numero di arrivi e presenze in un periodo e dalla conoscenza del comune di residenza della struttura ricettiva si ottengono le distribuzioni degli arrivi e delle presenze per le tipologie di struttura ricettiva e le varie classificazioni del territorio: da quelle amministrative a quelle funzionali.

Notiamo infine che i soggiorni e quindi gli arrivi che si manifestano in una struttura ricettiva, e a maggior ragione in un aggregato territoriale o tipologico di strutture ricettive in un dato periodo non coincidono con il numero di turisti che hanno acquistato soggiorno in quell'aggregato: questo perché una persona può dar luogo a più soggiorni, ad esempio un francese può passare due notti in un albergo di Firenze e tre in uno di Viareggio dando così luogo a due arrivi. Pertanto gli arrivi registrati in un territorio in un dato periodo sono di norma maggiori dei turisti che hanno visitato (acquistando alloggio) quel territorio in quel periodo.

In uno studio condotto con una indagine ad hoc nel distretto turistico di Cefalù in Sicilia a fronte di 13.630 arrivi registrati negli alberghi del distretto per il mese di settembre 2005, i turisti sono risultati 11.753 (Parroco e Vaccina, 2006)

1.4. Il flusso dei dati

Operativamente questa rilevazione è molto complessa. Giornalmente tutte le strutture ricettive, registrano il numero dei clienti arrivati (gli arrivi), cioè quante persone hanno chiesto alloggio, il numero dei clienti partiti, il conseguente saldo giornaliero (le presenze nella notte incipiente): questo distintamente per le nazionalità dei clienti stranieri e per le regioni di provenienza dei clienti italiani. Tali registrazioni vengono effettuate prevalentemente su due tipologie di moduli cartacei: uno con le registrazioni dei movimenti giornalieri individuali (C59), l'altro con le registrazioni dei movimenti giornalieri aggregati (Tavola di spoglio A). Questi documenti, nella regione toscana vengono trasmessi[1] mensilmente alle Province, le quali verificano

[1] Vi sono nella pratica modalità differenziate di trasmissione dei dati mensili dalle strutture ricettive alle province. Nella provincia di Firenze, oltre la trasmissione per via telematica sono presenti le modalità manuale, per posta e fax. Nel

la completezza della rilevazione rispetto alle strutture presenti e attive ed effettuano alcuni controlli di qualità. Il materiale cartaceo viene poi inviato a società di registrazione che lo trasferiscono su archivi informatizzati. Successivamente tali archivi sono inviati alla Regione, la quale raccolti i dati da tutte le province lo invia all'Istat titolare della rilevazione e cura proprie pubblicazioni.

Si tratta in sostanza di un censimento giornaliero con riepiloghi mensili effettuato prevalentemente su questionari cartacei; è quindi comprensibile che i dati sugli arrivi e le presenze mensili con le articolazioni della tipologia ricettiva e della zona geografica siano disponibili solo molti mesi dopo quello cui si riferiscono. Attualmente (aprile 2007) ad esempio si stanno fornendo i dati regionali dell'anno 2005. La provincia di Firenze dispone al momento dei dati sul movimento dell'ottobre 2006.

Le Province, oltre ai dati sul movimento raccolgono, con la comunicazione annuale dei prezzi, una ampia serie di dati sulle caratteristiche strutturali degli esercizi ricettivi (prezzi relativi alle diverse modalità di offerta di alloggio, servizi presenti nelle camere e nell'intera struttura, lingue parlate, ecc.). Tale insieme di dati costituisce un ricco patrimonio informativo che oltre a soddisfare gli adempimenti amministrativi e quegli statistici verso la Regione e Istat, potrebbe essere sfruttato anche per fini informativi propri. Di particolare interesse sarebbe una analisi comparata prezzi–servizi fra diversi ambiti territoriali sia interni a una regione sia fra regioni diverse.

1. 5. Come ottenere dati tempestivi

Una catena regionale di supermercati conosce la mattina la distribuzione dei prodotti venduti il giorno precedente grazie a un sistema informativo che acquisisce con la registrazione automatica alle casse i codici dei prodotti venduti. La direzione regionale o provinciale può così valutare tempestivamente l'effetto di una promozione sulle vendite complessive, su quelle dei prodotti sostitutivi e complementari; decidere sui rifornimenti, ecc.

Un assessorato regionale o provinciale, una associazione di categoria deve invece aspettare mesi se non addirittura un intero anno per conoscere quale è stato l'andamento del turismo nell'estate nella regione o nella provincia. Una statistica siffatta serve più alla storia del turismo che alle decisioni. Si può obiettare che il sistema delle strutture ricettive presenti sul territorio non potrà mai aver una catena di comando come una struttura imprenditoriale; tuttavia non è utopistica, utilizzando le tecnologie disponibili, l'aspettativa di ottenere informazioni statistiche entro pochi giorni dalla fine del mese. Questo obiettivo può essere raggiunto con due strumenti:

- acquisire i dati in formato digitale al momento in cui gli eventi si manifestano,
- ottenere la collaborazione degli esercenti alla qualità e tempestività.

Il primo strumento è essenzialmente tecnico. Il computer e il collegamento internet sono ampiamente diffusi presso le strutture ricettive e il loro uso può essere incentivato offrendo ad esempio sevizi di rete e maggior visibilità.

---

comune di Firenze operano anche alcune ditte che raccolgono la modulistica presso gli esercizi loro clienti. Nei comuni della provincia la raccolta passa attraverso i vigili; in alcuni casi un comune può centralizzare la raccolta e spedire per posta o corriere i pacchi dei modelli statistici alla Provincia. Vi sono infine situazioni in cui i dati sul movimento sono prodotti tramite procedure informatizzate, quindi stampati su moduli cartacei che riproducono i modelli C59 o la tavola di Spoglio A e in tale forma inviati alla Provincia. In questi casi si perdono gran parte dei vantaggi dell'informatizzazione. Questo comportamento risulta, guardando il processo complessivo di raccolta dei dati, inefficiente e meriterebbe un approfondimento per capirne le ragioni.

Il secondo strumento richiede che la statistica si ponga fra gli obiettivi anche quello di fornire una informazione utile agli esercenti delle strutture ricettive, non solo ai decisori politici. Questo obiettivo pone una questione non banale per la Statistica e rinvia al problema oggettivo dell'asimmetria del procedimento statistico cui accenneremo nel successivo capitolo.

Va ricordato che solo l'indagine censuaria può fornire informazione dettagliata sul movimento turistico di una specifica nazionalità entro un territorio piccolo. Sono aumentate le presenze di francesi nei comuni dell'Amiata nell'estate 2007 rispetto all'anno precedente? Se interessa mantenere l'indagine sul movimento dei clienti nelle strutture ricettive censuaria, questa deve avvenire in modo informatizzato, come sottoprodotto dell'adempimento per la pubblica sicurezza oppure, se questo venisse a mancare, come sottoprodotto dell'attività gestionale della struttura ricettiva.

Sarebbe molto utile un esame dei costi di questa indagine, condotta con le attuali modalità, per evidenziare la sproporzione di questi con la qualità dell'informazione fornita.

In attesa che si realizzino le condizioni per una indagine censuaria completamente informatizzata, dati tempestivi sul movimento turistico vengono attualmente ottenuti attraverso indagini campionarie realizzate sia dalle amministrazioni pubbliche che dalle associazioni di categoria, specialmente in alcuni periodi dell'anno: il periodo pasquale, quello estivo. Tali indagini forniscono per loro natura stime di aggregati piuttosto vasti sia rispetto al territorio, sia rispetto alle tipologie ricettive e alle nazionalità. Inoltre richiedono agli esercizi del campione un adempimento ulteriore, da una semplice valutazione soggettiva dell'andamento alla replicazione della comunicazione degli arrivi e presenze verificatosi nel periodo di interesse, con ciò aumentando il disturbo del rispondente. Inoltre necessitano di una struttura organizzativa che realizzi l'indagine parallela a quella censuaria, con tempi di realizzazione e costi non trascurabili. Va infine tenuto presente che la popolazione degli esercizi ricettivi, così come accade per le imprese in generale, si presenta molto asimmetrica, numerose piccole e poche grandi unità, con una forte variabilità elementare, sì che le dimensioni dei campioni devono essere rilevanti per ottenere stime con una precisione sufficiente per le esigenze informative.

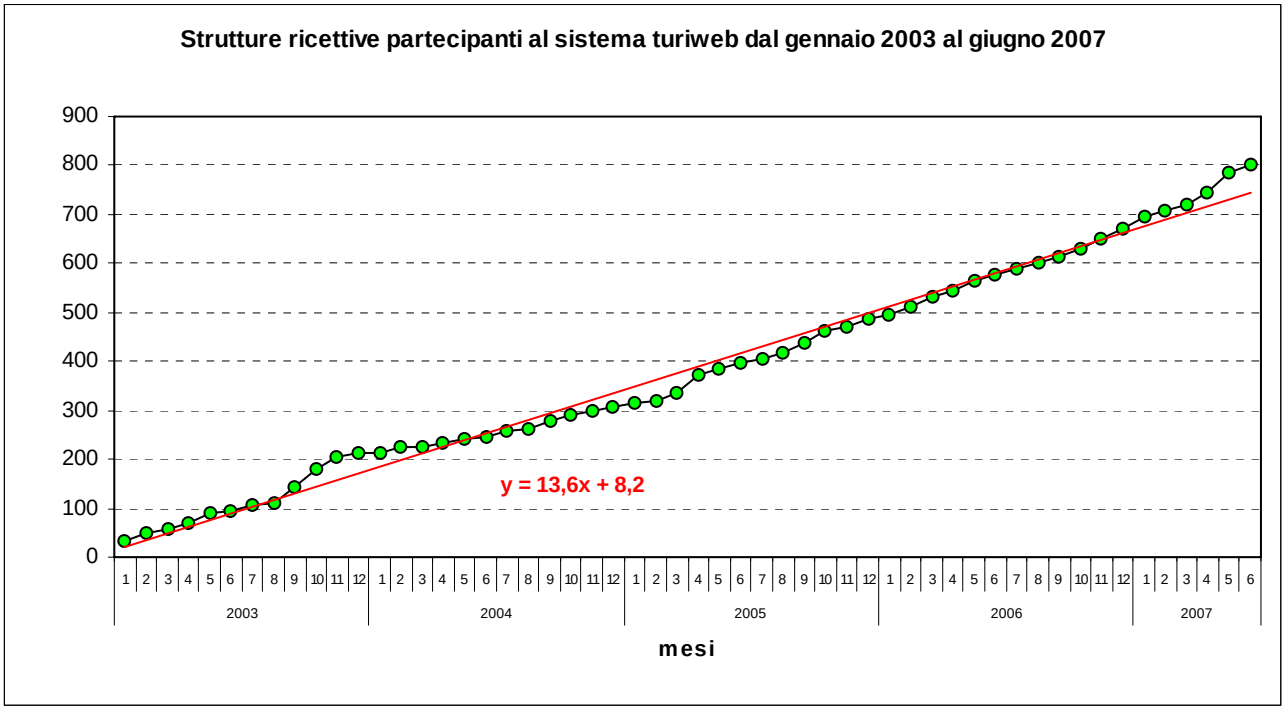
1.6. Il sistema turiweb della Provincia di Firenze

Alcune province, fra cui in primo luogo quella di Firenze, hanno promosso un sistema automatico di registrazione dei movimenti e trasmissione alla provincia via WEB dei dati. Le prime attivazioni nella provincia di Firenze si sono verificate nel mese di novembre 2002. Nei mesi successivi si sono aggiunte man mano altre strutture. La partecipazione al sistema è del tutto volontaria.

L'evoluzione del numero di esercizi presenti nel sistema turiweb è evidenziata nel grafico seguente (Grafico 1.2). Si nota una evoluzione pressoché lineare, salvo alcuni momenti in cui si registrano tassi di crescita superiori a quello del trend complessivo, dovuti ad interventi di promozione da parte della Provincia, come è avvenuto nei primi mesi del 2007.

Fenomeni di questo tipo hanno, di solito, un andamento sigmoidale: tassi di crescita crescenti all'inizio, tassi costanti nella fase intermedia e tassi decrescenti in quella finale. Il fenomeno di interesse sembra lontano da questa fase finale dove risulta sempre più difficile convincere i più resistenti a partecipare. Questo fa ben sperare sulla evoluzione del sistema turiweb nei prossimi mesi. L'informazione più recente (giugno 2007) indica in circa 800 il numero di esercizi del sistema turiweb.

Grafico 1.2



Tuttavia, assumendo un tasso di crescita di 14 esercizi al mese, occorreranno 120 mesi (10 anni) perché il sistema copra tutta la popolazione degli esercizi ricettivi della provincia di Firenze costituita da circa 2400 unità. Nel caso invece si mantenesse il tasso di crescita di 20 esercizi al mese verificatosi nei primi mesi del 2007 occorrerebbero "solo" 73 mesi (poco più di 6 anni).

Va anche considerato che molte strutture ricettive sono costituite da piccole imprese spesso a gestione familiare dove il digital divide può essere presente come lo è notoriamente nelle famiglie. Se quindi si vuol raggiungere l'obiettivo di coinvolgere tutte o quasi le strutture ricettive nel sistema turiweb in tempi più brevi, occorre assumere iniziative che incentivino l'ingresso del sistema.

Si pone allora il problema, in attesa della completa diffusione del sistema turiweb, di come utilizzare al meglio le informazioni che pochi giorni dopo la fine del mese sono disponibili presso la provincia di Firenze provenienti dal sottoinsieme delle strutture che volontariamente partecipano al sistema.

1.7. Opportunità ed elementi critici

L'aspettativa degli utenti è quella di ottenere, stime del movimento per l'intera provincia e possibilmente per alcuni ambiti territoriali e di tipologia dell'offerta, in particolare stime della variazione del movimento (arrivi e presenze) in un periodo dell'anno (un mese, un trimestre) rispetto al corrispondente periodo dell'anno precedente.

Lo statistico non può esimersi dalla sfida di soddisfare questa esigenza, adottando gli strumenti che meglio permettano di evitare o limitare la possibile distorsione delle stime derivante dalla autoselezione del campione. Valutiamo pertanto le opportunità offerte dalla presenza del sistema turiweb e gli elementi critici cui si va incontro usando il campione autoselezionato.

Opportunità

- Non si deve impiantare una rilevazione campionaria parallela, i dati sono disponibili molto tempestivamente e già informatizzati, i costi dell'utilizzo dei volontari sono limitati a quelli della implementazione dei metodi di stima e della loro esecuzione mensile.
- Siamo in presenza di una indagine censuaria, dovrebbe quindi essere possibile valutare, sulla base dei dati con questa ottenuti, quanto i metodi di stima che utilizzano il sottoinsieme turiweb forniscono valori prossimi ai parametri di interesse.
- Vi sono notevoli informazioni ausiliarie sulle singole unità: dalla loro localizzazione, dimensione, all'attrezzatura in esse presente, all'entità del movimento verificatosi in periodi precedenti che possono essere utilizzate sia per limitare la possibile distorsione, sia per aumentare l'efficienza delle stime. L'indagine censuaria stessa può fornire un aiuto nella realizzazione delle stime anticipate basate sulle unità del sistema turiweb.

Elementi critici

- Il fenomeno presenta in questi anni una evoluzione lenta: le variazioni da un anno all'altro o di un mese rispetto al corrispondente dell'anno precedente sono, in assenza di eventi eccezionali, normalmente inferiori al 10%. L'incertezza delle stime potrebbe essere dello stesso ordine di grandezza di questa variazione. In tal caso potrebbe esser perfino impossibile affermare se il fenomeno è cresciuto o meno in un dato periodo dell'anno rispetto al corrispondente periodo dell'anno precedente; e questa conclusione è difficilmente accettabile dall'utente non statistico.
- L'informazione ausiliaria è costituita da due fonti:

  1- La consistenza delle strutture, cioè la lista delle strutture ricettive presenti nel territorio con le informazioni sulla tipologia, dimensione (posti letto) e per gli alberghi sulla categoria (stelle).

  2- L'attrezzatura delle strutture ricettive dove viene riportata l'informazione sulle combinazioni di offerta di alloggio, sui servizi offerti, sui prezzi massimi praticati per tali combinazioni.

  Tali fonti non risultano esattamente aggiornate e presentano carenze che sono fonte di un ulteriore errore nelle stime da aggiungere a quello derivante dall'uso di un campione di volontari.
- La precisione delle stime campionarie, a parità di altre condizioni, dipende essenzialmente dalla numerosità del campione. Ora la popolazione delle strutture ricettive della provincia di Firenze è composta da circa 2400 unità, il sottoinsieme dei volontari da poco più di 700 unità (aprile 2007). Anche una buona strategia campionaria con un campione non autoselezionato di questa dimensione potrebbe fornire stime con incertezza eccessiva.
- Infine, si pone la questione della validità dell'inferenza all'intera popolazione delle strutture ricettive a partire da un sottoinsieme autoselezionato: la disponibilità locale della rete, la disponibilità e l'atteggiamento dei gestori verso lo strumento, possono influire sulla decisione a partecipare o meno al sistema turiweb, rendendo il processo di autoselezione non ignorabile.

1.8. Le strutture del sistema TURIWEB della provincia di Firenze

Abbiamo effettuato una analisi del sottoinsieme delle strutture ricettive del sistema turiweb con un duplice lo scopo.

- Fornire alla Amministrazione provinciale una conoscenza di questo segmento dell'offerta, evidenziando come si caratterizza rispetto alla localizzazione e alle tipologie ricettive, utile per il monitoraggio e la promozione del sistema stesso. Questo anche come ringraziamento per la disponibilità dell'Amministrazione nel mettere a disposizione i dati.
- Esaminare la distribuzione nel territorio e per tipologia ricettiva delle strutture ricettive del sistema turiweb al fine di valutare la "rappresentatività" di tale sottoinsieme. Se le strutture ricettive turiweb fossero concentrate in un'unica categoria ed area geografica, sarebbe arduo accingersi ad usare tale sottoinsieme per il nostro obiettivo.

I dati su cui viene effettuata l'analisi sono quelli dell'ottobre 2006. E' stato scelto questo mese in quanto l'ultimo per il quale erano disponibili, al momento della stesura di questa parte del testo, sia le informazioni del sottoinsieme turiweb, sia quelle per l'intera popolazione delle strutture ricettive.

Nel mese di ottobre 2006, nella provincia di Firenze sono risultate appartenenti al sistema turiweb 635 strutture ricettive sulle 2401, pari al 26,4%. Nella tavola 1 si esamina l'appartenenza al sistema turiweb per comune e per le aggregazioni di comuni secondo le 5 risorse turistiche prevalenti[2] presenti nella provincia.

Ci saremmo aspettati che la frequenza di appartenenza al sistema turiweb fosse più alta nei comuni della risorsa Arte e Affari e nel comune di Firenze in particolare. Nel capoluogo invece appartengono al sistema il 24,4% degli esercizi, contro un 30,2% riscontrato negli altri comuni della risorsa Arte e Affari. Una percentuale superiore alla media provinciale si registra anche per il gruppo di comuni della risorsa Altro Interesse; è questo un altro segnale della "vicinanza" dei comuni classificati in questa voce residua a quelli della risorsa Arte e Affari. Frequenze più basse della media provinciale risultano nel gruppo dei comuni della risorsa Campagna-Collina (23,9%). Le due risorse Montagna e Termale contengono rispettivamente due e un comune, per cui le frequenze di appartenenza al sistema turiweb sono poco significative.

Vi sono infine alcuni comuni dove la frequenza di strutture turiweb raggiunge o supera il 40%: Empoli (44%), Montelupo Fiorentino (50%), Dicomano (40%), Vaglia (46%), Bagno a Ripoli (58%), Impruneta (40%), Lastra a Signa (46%), Scandicci (47%). Metà di questi 8 comuni appartengono alla risorsa Altro Interesse, la quale risulta, almeno nella propensione ad utilizzare le nuove tecnologie per gli adempimenti statistici, più dinamica di realtà turistiche tradizionalmente più riconosciute e forse per questo più inclini a "vivere sugli allori".

[2] La classificazione dei comuni secondo la risorsa turistica prevalente, definita dalle province agli inizi degli anni 90' presenta qualche incoerenza: soprattutto per la mancanza di un criterio di assegnazione univoco e oggettivo a livello regionale. Per la provincia di Firenze in particolare la voce residua <<Altro Interesse>> comprende ben 18 comuni con 540 strutture. In alcuni casi si tratta di comuni della cintura di Firenze; in questi casi sembrerebbe più logica la loro appartenenza alla classe <<Arte e Affari>>. Una analisi della distribuzione congiunta degli arrivi per provenienza e risorsa turistica prevalente relativa all'anno 1993 (Scaffai, G., 1995) evidenziava analiticamente la "vicinanza" della distribuzione per nazionalità nella risorsa Arte e Affari con quella nella risorsa Altro Interesse, mentre ben distinte da queste risultavano le tre risorse "ambientali": Balneare, Campagna-Collina, Montagna, molto simili fra loro. Come per l'Arte e Affari, anche l'Altro Interesse infatti era caratterizzato da un ampio ventaglio di nazionalità di provenienza, mentre le tre risorse ambientali erano caratterizzate da arrivi provenienti dall'interno e da poche altre nazioni straniere.

La Provincia di Firenze ha adottato una propria classificazione del territorio basata su un criterio geografico-funzionale individuando 7 aree. Riportiamo per queste aree le percentuali di esercizi appartenenti al sistema turiweb:

Aree	Percentuale di strutture turiweb
1- il comune di Firenze	24,4%
2- Area Fiorentina (i comuni della cintura fiorentina)	36,7%
3- Mugello(i comuni della valle del Mugello e del versante adriatico)	20,0%
4- Montagna Fiorentina (i comuni della bassa valle della Sieve)	22,4%
5- Valdarno (i comuni del Valdarno)	23,8%
6- Chianti (i comuni del Chianti fiorentino)	30,6%
7- Empolese Valdelsa (i comuni della Valdelsa e della zona empolese)	25,5%

Tav. 1 - Strutture ricettive appartenenti e non al sistema TURIWEB per comune; provincia di Firenze, ottobre 2006

Comune	Valori assoluti			Valori percentuali		
	WEB	Non WEB	Totale	WEB	Non WEB	Totale
	Risorsa turistica ARTE e AFFARI					
Firenze	**216**	**671**	**887**	**24,4**	**75,6**	**100,0**
Certaldo	16	48	64	25,0	75,0	100,0
Empoli	11	14	25	44,0	56,0	100,0
Fiesole	10	25	35	28,6	71,4	100,0
Greve in Chianti	34	118	152	22,4	77,6	100,0
Montelupo Fiorentino	6	6	12	50,0	50,0	100,0
San Casciano in Val di Pesa	37	57	94	39,4	60,6	100,0
Tavarnelle Val di Pesa	23	43	66	34,8	65,2	100,0
Vinci	7	22	29	24,1	75,9	100,0
Totale altri ARTE e AFFARI	**144**	**333**	**477**	**30,2**	**69,8**	**100,0**
Totale ARTE e AFFARI	**360**	**1004**	**1364**	**26,4**	**73,6**	**100,0**
	Risorsa turistica MONTAGNA					
Firenzuola	4	22	26	15,4	84,6	100,0
Reggello	12	48	60	20,0	80,0	100,0
Totale MONTAGNA	**16**	**70**	**86**	**18,6**	**81,4**	**100,0**
	Risorsa turistica TERMALE					
Gambassi Terme	12	23	35	34,3	65,7	100,0
Totale TERME	**12**	**23**	**35**	**34,3**	**65,7**	**100,0**
	Risorca turistica CAMPAGNA - COLLINA					
Barberino di Mugello	9	22	31	29,0	71,0	100,0
Barberino Val d'Elsa	23	47	70	32,9	67,1	100,0
Borgo San Lorenzo	7	25	32	21,9	78,1	100,0
Capraia e Limite	4	7	11	36,4	63,6	100,0
Cerreto Guidi	2	13	15	13,3	86,7	100,0
Dicomano	10	15	25	40,0	60,0	100,0
Londa	2	8	10	20,0	80,0	100,0
Marradi	3	18	21	14,3	85,7	100,0
Montespertoli	18	45	63	28,6	71,4	100,0
Palazzuolo sul Senio	1	11	12	8,3	91,7	100,0
Pelago	2	16	18	11,1	88,9	100,0
San Godenzo	1	17	18	5,6	94,4	100,0
Vaglia	6	7	13	46,2	53,8	100,0
Vicchio	2	35	37	5,4	94,6	100,0
Totale CAMPAGNA COLLINA	**90**	**286**	**376**	**23,9**	**76,1**	**100,0**
	Risorsa turistica ALTRO INTERESSE					
Bagno a Ripoli	14	10	24	58,3	41,7	100,0
Calenzano	3	15	18	16,7	83,3	100,0
Campi Bisenzio	4	18	22	18,2	81,8	100,0
Castelfiorentino	2	25	27	7,4	92,6	100,0
Figline Valdarno	6	28	34	17,6	82,4	100,0
Fucecchio	2	8	10	20,0	80,0	100,0
Impruneta	20	30	50	40,0	60,0	100,0
Incisa Valdarno	4	15	19	21,1	78,9	100,0
Lastra a Signa	22	26	48	45,8	54,2	100,0
Montaione	14	63	77	18,2	81,8	100,0
Pontassieve	14	34	48	29,2	70,8	100,0
Rignano sull'Arno	10	21	31	32,3	67,7	100,0
Rufina	2	11	13	15,4	84,6	100,0
San Piero a Sieve	0	8	8	0,0	100,0	100,0
Scandicci	19	21	40	47,5	52,5	100,0
Scarperia	9	16	25	36,0	64,0	100,0
Sesto Fiorentino	5	19	24	20,8	79,2	100,0
Signa	7	15	22	31,8	68,2	100,0
Totale ALTRO INTERESSE	**157**	**383**	**540**	**29,1**	**70,9**	**100,0**
TOTALE GENERALE	**635**	**1766**	**2401**	**26,4**	**73,6**	**100,0**

Nella provincia è presente una offerta di alloggio molto differenziata quanto a tipologie: sono presenti ben 13 tipologie ricettive, sebbene cinque siano prevalenti rappresentando il 93% delle strutture. La tavola 2 riporta la frequenza di appartenenza al sistema turiweb per tipologia di struttura ricettiva.

Vi è una notevole differenza nella propensione all'uso della tecnologia. Si passa dal 56% delle residenze d'epoca, all'11% dei campeggi. Gli alberghi, che costituiscono la tipologia prevalente nella provincia con 564 unità, solo per il 22,2% partecipano al sistema, una frequenza inferiore a quella di strutture ricettive "meno imprenditoriali" quali gli affittacamere e gli alloggi privati.

Sebbene le tipologie di offerta ricettiva tendano ad evolversi verso forme sempre più varie e specifiche dell'ambiente in cui si trovano (si pensi agli alloggi agrituristici), gli esercizi alberghieri restano una tipologia fondamentale del sistema dell'accoglienza, specialmente per i turisti stranieri che visitano le città d'arte. La partecipazione degli alberghi è quindi fondamentale per lo sviluppo del sistema di trasmissione telematica del movimento mensile.

Le strutture più specializzate, che magari operano in particolari nicchie di mercato, risultano le più propense all'utilizzo delle nuove tecnologie. Dato che l'utilizzo delle nuove tecnologie informatiche é molto correlato all'età e al titolo di studio, un approfondimento di questo tema si potrebbe ottenere se si disponesse di informazioni sull'età e il titolo di studio dei gestori.

Tav. 2 - Strutture ricettive appartenenti al sitema TURIWEB per tipologia ricettiva; provincia di Firenze, ottobre 2006.

Tipologia ricettiva	Valori assoluti			Valori percentuali		
	WEB	Non WEB	Totale	WEB	Non WEB	Totale
Residenze d'epoca	14	11	25	56,0	44,0	100,0
Residence	18	24	42	42,9	57,1	100,0
Rifugi alpini	2	4	6	33,3	66,7	100,0
Affittacamere (professionale)	175	384	559	31,3	68,7	100,0
Ostelli	4	10	14	28,6	71,4	100,0
Residenza turistico alberghiere	2	5	7	28,6	71,4	100,0
Alloggi privati	83	229	312	26,6	73,4	100,0
Case per vacanze	71	196	267	26,6	73,4	100,0
Alloggi agrituristici	131	406	537	24,4	75,6	100,0
Alberghi	125	439	564	22,2	77,8	100,0
Case per ferie	8	42	50	16,0	84,0	100,0
Campeggi	2	15	17	11,8	88,2	100,0
Villaggi turistici	0	1	1	0,0	100,0	100,0
Totale	**635**	**1766**	**2401**	**26,4**	**73,6**	**100,0**

Con la tavola 3 si esamina l'appartenenza al sistema turiweb degli alberghi secondo la categoria (numero di stelle). La categoria più "virtuosa" è rappresentata dagli alberghi a 4 stelle che appartengono al sistema per il 37%, mentre quelli a 1 e 2 stelle solo per il 14% e 10% rispettivamente.

Dei 15 alberghi di massima categoria (le 5 stelle) solo 3 (pari al 20%) appartengono al sistema. Questi dati mettono in evidenza come gli alberghi di bassa categoria (1 e 2 stelle) insieme ai campeggi risultano le strutture meno propense all'uso dell'automazione per gli adempimenti statistici, ma anche gli alberghi a 3 stelle, che costituiscono la categoria alberghiera più diffusa, non brillano nell'utilizzo dell'automazione per questi adempimenti (25,7%), a fronte ad esempio del 31,3% degli affittacamere.

Tav. 3 - Alberghi appartenenti al sitema TURIWEB per stelle; provincia di Firenze, ottobre 2006.

Stelle	Valori assoluti			Valori percentuali		
	WEB	Non WEB	Totale	WEB	Non WEB	Totale
1	14	85	99	14,1	85,9	100,0
2	13	113	126	10,3	89,7	100,0
3	56	162	218	25,7	74,3	100,0
4	39	67	106	36,8	63,2	100,0
5	3	12	15	20,0	80,0	100,0
Totale	**125**	**439**	**564**	**22,2**	**77,8**	**100,0**

Non sappiamo se la mancata partecipazione al sistema turiweb sia un segnale del digital divide o indichi solo una diffidenza verso l'uso di internet per gli adempimenti statistici. Ma per quale motivo questa diffidenza dovrebbe esser più presente negli alberghi di medio bassa categoria piuttosto che negli affittacamere? Gli esercizi alberghieri, quelli di medio bassa categoria in particolare e i campeggi risultano, come abbiamo già accennato, meno dinamici di altre tipologie.

Viene da chiedersi: "Sono forse le strutture tradizionali in crisi?" Chi gestisce un albergo medio piccolo o un campeggio ha poca fiducia nello sviluppo del turismo a differenza dei gestori di strutture "più giovani" e quindi è meno propenso ad investire nelle nuove tecnologie? I clienti tendono a preferire "strutture alternative"? Oppure, in alcune realtà non si sente il bisogno di evolversi investendo verso le nuove tecnologie perché si usufruisce di un vantaggio competitivo di posizione, "vivendo un po' sugli allori". Per rispondere a queste domande occorrerebbe esaminare l'evoluzione della consistenza e del movimento delle diverse tipologie ricettive e nelle diverse aree del territorio, il che esula dai nostri attuali obiettivi. Qui, sulla base dei dati relativi alla partecipazione al sistema turiweb emerge un ritardo di queste strutture ricettive che abbiamo definito più tradizionali: gli alberghi specialmente di medio bassa categoria, i campeggi.

La tavola 4 riporta infine il movimento dei clienti della provincia nell'ottobre 2006, distinto secondo l'appartenenza o meno delle strutture ricettive al sistema turiweb. Le strutture del sistema pari al 26,4% hanno assorbito il 29,7 degli arrivi e il 28,6 delle presenze, una quota leggermente superiore alla loro consistenza. Tenendo conto che alcune strutture extralberghiere di notevole capacità ricettiva non sono presenti nel sistema, questo conferma quanto già evidenziato sulla maggior "dinamicità" delle strutture del sistema turiweb.

Tav. 4 - Movimento dei clienti nelle strutture ricettive del sistema TURIWEB e non; provincia di Firenze, ottobre 2006

Provenienza	Valori assoluti			Valori percentuali		
	WEB	Non WEB	Totale	WEB	Non WEB	Totale
			ARRIVI			
Italiani	29.001	69.376	98.377	29,5	70,5	100,0
Stranieri	78.402	184.572	262.974	29,8	70,2	100,0
Totale	107.403	253.948	361.351	29,7	70,3	100,0
			PRESENZE			
Italiani	66.227	182.589	248.816	26,6	73,4	100,0
Stranieri	216.930	525.352	742.282	29,2	70,8	100,0
Totale	283.157	707.941	991.098	28,6	71,4	100,0

1.9. Considerazioni di sintesi sulle strutture del sistema turiweb

L'esame esplorativo delle strutture ricettive del sistema turiweb conduce alle seguenti conclusioni.

- Il sistema turiweb è cresciuto dal gennaio 2003 al giugno 2007 a un tasso costante ma modesto: entrano mediamente 14 esercizi al mese. In occasione di campagne promozionali, tale tasso si è alzato, come ad esempio nei primi mesi del 2007. A un tasso di 14 esercizi al mese occorreranno 120 mesi per coprire l'intera popolazione delle strutture ricettive.

- Con riferimento all'ottobre 2006, il sistema turiweb risulta diffuso su tutto il territorio provinciale e in tutte le tipologie ricettive. La diffusione sul territorio e fra le tipologie ricettive è però assai differenziata. Le strutture ricettive "più tradizionali", in particolare alberghi di categoria medio-bassa e campeggi; presentano frequenze di partecipazione molto inferiori alla media provinciale. Viceversa le tipologie ricettive "più giovani" risultano maggiormente presenti nel sistema turiweb. Una eventuale promozione e/o incentivazione del sistema dovrebbe rivolgersi a queste tipologie ricettive al momento in ritardo, nonché in alcune località ove la partecipazione è molto limitata.

- Il comune di Firenze, nonostante la rilevanza turistica di livello nazionale non risulta particolarmente virtuoso con una frequenza di esercizi presenti nel sistema inferiore a quella del gruppo degli altri comuni classificati con risorsa prevalente Arte e Affari o facenti parte della sua cintura.

- Assumendo, con tutte le cautele del caso, la partecipazione al sistema come un indicatore della propensione ad investire nelle tecnologie informatiche si evidenzia un ritardo delle strutture che abbiamo definito "più tradizionali" (alberghi di medio bassa categoria, campeggi). Questo ritardo potrebbe esser dovuto a scarsa fiducia nello sviluppo o a scarso bisogno di evoluzione, specialmente per situazioni con un'elevata "rendita di posizione".

Riguardo infine all'obiettivo di utilizzare il sottoinsieme delle strutture turiweb per la stima del movimento emergono due aspetti:

- Le strutture turiweb sono presenti, sebbene in proporzioni differenziate, in tutte le aree territoriali, in tutte le tipologie ricettive, salvo due eccezioni. Con riferimento all'ottobre 2006 vi è un solo comune (San Piero a Sieve) e una sola tipologia di struttura, del resto costituita da una sola unità, (villaggi turistici) privi di rappresentanti turiweb.
- Le strutture turiweb si presentano però più dinamiche, in particolare coprendo il 26,4% della popolazione, hanno fatto registrare nel mese di ottobre 2006 il 29,7% degli arrivi e il 28,6% delle presenze. Nel caso degli alberghi, il livello qualitativo (rappresentato dalle stelle) differenzia la propensione a partecipare al sistema turiweb.

Capitolo 2

2.1. Il modello concettuale statistico

A fronte di un problema conoscitivo più o meno formalmente espresso da un committente[1] lo statistico costruisce uno "modello concettuale statistico" (MCS) definendo:

1. una popolazione obiettivo
2. i caratteri (le variabili) da misurare sulle unità che costituiscono la popolazione obiettivo
3. le modalità con cui tali caratteri si manifestano

La **popolazione obiettivo** è un insieme definito in modo che data una qualunque unità o evento si possa dire se appartiene o non appartiene all'insieme. Le unità della popolazione obiettivo prendono il nome di unità statistiche. Occorre che la definizione sia precisa e preveda scelte univoche per situazioni talvolta incerte. Si pensi alla definizione di famiglia, di impresa, abitazione, struttura turistico-ricettiva, ecc. Spesso le norme giuridiche forniscono una buon punto di riferimento, ma casi dubbi si trovano anche inaspettatamente.

I **caratteri** sono le "qualità caratteristiche" con cui l'unità statistica appartenente alla popolazione obiettivo si può presentare. I caratteri di cui l'unità statistica è portatrice sono pressoché illimitati in quanto una completa rappresentazione dell'unità richiederebbe la sua esatta ricostruzione. Viene in mente il paradosso dei cartografi di Borges (1984) la cui mappa dell'Impero uguagliava in grandezza l'Impero stesso, coincidendo puntualmente con esso e pertanto completamente inutile. La scelta dei caratteri da prendere in considerazione è un'operazione complessa e frutto di un trade-off fra le esigenze informative da un lato e le risorse e l'operatività dall'altro. Siamo di fronte a una inevitabile schematizzazione che la scelta di un numero limitato di caratteri effettua nella rappresentazione dell'unità statistica anche per un obiettivo informativo ben formulato e circoscritto.

I caratteri possono assumere in ciascuna unità statistica un insieme di **modalità** che il ricercatore definisce, sempre avendo presente gli obiettivi informativi. Per un dato carattere si possono considerare differenti modalità (modalità di risposta). Ai fini delle elaborazioni possibili ha rilevanza la distinzione in caratteri qualitativi sconnessi, qualitativi ordinati e quantitativi. Specifiche modalità di alcuni caratteri sono comuni a tutte le unità che costituiscono la popolazione obiettivo, e in effetti servono per la sua definizione e circoscrizione; in particolare una modalità del tempo e una dello spazio.

Non esiste una procedura per una formulazione univoca del MCS a partire dal problema conoscitivo, se non altro per la formulazione non completamente strutturata del problema e per la presenza di vincoli e opportunità nella implementazione concreta del processo di rilevazione dei dati. La definizione del MCS è frutto inoltre di ipotesi di lavoro sul funzionamento del fenomeno oggetto di studio. Modelli concettuali statistici diversi adottati per uno stesso obiettivo informativo conducono a risultati diversi. E' questo il motivo di accesi dibattiti fra coloro che, pur appoggiandosi a dati statistici, sostengono tesi diverse.

[1] Intendiamo qui con committente il destinatario dei risultati dell'indagine statistica: può trattarsi di una persona fisica o giuridica, come nel caso di una società che commissione una indagine di mercato, di un governo o dell'intera collettività nazionale interessata a monitorare l'evoluzione del tasso di disoccupazione.

E' allora fondamentale che i risultati delle rilevazioni statistiche siano accompagnati dalle informazioni che dichiarano gli obiettivi informativi, il MCS scelto e il processo di rilevazione ed elaborazione realizzato (i cosiddetti metadati). Questo può non bastare se l'utilizzatore dei dati non dispone di una elementare cultura statistica che gli permetta di riconoscere il MCS.

Nel caso della popolazione degli esercizi turistici ricettivi della provincia di Firenze, abbiamo visto che si rilevano su queste unità statistiche i caratteri quantitativi *arrivi* e *presenze* distintamente per le nazionalità dei clienti stranieri e per le regioni dei clienti italiani, in un dato periodo, di norma il mese. Eppure molto spesso si attribuisce al numero di arrivi, se non addirittura a quello delle presenze, il significato di numero di turisti che hanno visitato un territorio.

2.2. Popolazioni statistiche e modelli di popolazione

Il MCS può far riferimento a una popolazione obiettivo reale o ipotetica. Nel primo caso siamo di fronte a un insieme di unità o eventi che può avere una numerosità più o meno grande, ma é per sua natura finito. La Statistica, attraverso la rilevazione dei caratteri sulle unità e le sintesi quantitative fornisce l'informazione richiesta su questa popolazione. L'esempio di maggior rilievo è fornito dai censimenti delle popolazioni residenti e presenti negli stati nazionali.

Ogni popolazione, in quanto circoscritta dalle comuni modalità di alcuni caratteri possedute da tutte le sue unità può sempre esser considerata come una sottopopolazione di una popolazione più ampia allorché si riducono alcuni vincoli definitori. Inoltre vi sono esigenze informative che richiedono la specificazione di popolazioni potenziali quando si fa riferimento a processi o si voglia ricercare l'esistenza di relazioni fra i diversi caratteri valide oltre le limitate osservazioni.

Si voglia, ad esempio conoscere la qualità del servizio di trasporto effettuato dalla linea urbana 26 a Firenze. A tal fine si potrebbe considerare la popolazione obiettivo costituita dai "soggiorni" dei passeggeri su un mezzo della linea. Per ogni "soggiorno" si considerano i caratteri: fermata e istante di inizio, fermata e istante di fine, codice del mezzo, codice del conduttore. Altri caratteri potrebbero esser presi in considerazione relativi al giudizio dei passeggeri. Ma i primi potrebbero esser sufficienti una volta stabilito che la qualità del servizio é quantificata da indicatori che misurano il rispetto dell'orario, la frequenza, i tempi di percorrenza, l'affollamento a bordo. Per completare la definizione della popolazione occorre definire un periodo temporale: definiamo la settimana dal 23 al 29 gennaio 2006. Effettuata la rilevazione ed sintetizzati i dati si ottengono gli indicatori desiderati: ad esempio potrebbe risultare che vi è un eccessivo affollamento nell'intervallo fra la fermata 7 e la 8, che il ritardo è accumulato soprattutto nell'intervallo fra la fermata 3 e la 4 e così via. Una statistico pignolo potrebbe osservare che non è esatto concludere che vi è un eccessivo affollamento nella tratta fra la fermata 7 e la 8 , ma che la conclusione corretta è "vi è stato un eccessivo affollamento in tale tratta nella settimana dal 23 al 29 gennaio. In realtà è modesto l'interesse sul livello di qualità del servizio in quella specifica settimana, quello che effettivamente interessa è conoscere la qualità del processo produttivo del servizio, in relazione ad eventuali fattori connessi alla sua organizzazione, ai mezzi impiegati, alle condizioni del traffico ecc.

Inferenze a un collettivo più ampio di quello sul quale si effettua l'osservazione vengono effettuate quasi di norma nel nostro comportamento quotidiano: dalla decisione di scolare la pasta dopo aver assaggiato uno spaghetto, all'utilizzo dei dati censuari per definire interventi di pubblico interesse due mesi o un anno dopo la realizzazione del censimento.

In realtà il procedimento inferenziale è lo strumento di cui disponiamo per aumentare le conoscenze, individuando leggi, almeno operativamente valide al di là delle osservazioni necessariamente limitate: sulla base di queste leggi non solo volano gli aerei, stanno in piedi i ponti, si curano alcune malattie, ma si possono svolgere con sufficiente sicurezza le azioni più scontate come attraversare la strada: osserviamo per un breve tempo l'auto che sta sopraggiungendo,

stimiamo la sua velocità (il b del modello lineare), prevediamo la posizione dell'auto nell'istante in cui ci troveremo sulla sua traiettoria in base alla conoscenza che abbiamo della nostra velocità di attraversamento; se valutiamo questa posizione sufficientemente distante, decidiamo di attraversare.

Più le conclusioni inferenziali sono generali, più hanno valore; d'altro lato più ci si allontana dall'osservazione più tali conclusioni diventano incerte (e quindi di minor valore). Se la strada è molto larga e quindi più in là nel tempo deve essere la previsione sulla posizione dell'auto e nostra, più incerta è l'informazione per la nostra decisione. In questo caso preferiremmo usare il sottopassaggio. Il trade-off fra l'utilità dell'inferenza come procedimento di astrazione dalla limitata osservazione a categorie più generali e la conseguente incertezza era ben presente in Pascal (1982):

Fra tutti i frutti si distinguono le uve; e fra queste, le moscate, e Condrieu, e Desargues e quest'altra varietà. E' questo forse non è tutto. Una vite ha mai prodotto due grappoli uguali e in grappolo, due acini uguali? Io non ho mai giudicato della medesima cosa nello stesso esatto modo. Né posso giudicare della mia opera mentre la compongo: bisogna che faccia come i pittori, e me ne allontani, ma non troppo. Di quanto? Indovinatelo.

Tradizionalmente si considerano due tipologie di popolazioni verso le quali l'inferenza si rivolge. Popolazioni finite e modelli di popolazioni. Nel primo caso l'inferenza è richiesta soprattutto per motivi di tempo e risorse; invece di osservare tutta la popolazione ci si limita all'osservazione di un suo sottoinsieme. Queste popolazioni potrebbero essere, almeno in linea teorica, osservate anche completamente e pertanto il procedimento inferenziale è in qualche modo dettato da motivi di convenienza, non presenta una intrinseca necessità. In realtà vi sono situazioni in cui l'osservazione completa non é affatto conveniente come nel caso in cui questa conduce alla distruzione dell'unità, ma non per questo impossibile. Nel secondo caso siamo interessati a una popolazione ipotetica della quale possiamo osservare solo alcune manifestazioni. Dobbiamo ipotizzare una popolazione perché vogliamo ricavare regole operative che vogliamo applicare anche a unità non osservabili perché ancora non manifestate. O perché vogliamo dire qualcosa sugli acini dell'uva moscata che valga oltre quello specifico grappolo di quella pianta. Vogliamo ottenere un protocollo valido per curare il paziente che arriverà domani o per far decollare il prossimo aereo a partire dalla conoscenza di parametri che caratterizzano il processo e che sono sufficienti, almeno operativamente, per prevedere il comportamento del sistema. La formalizzazione di questa popolazione passa attraverso la specificazione di un modello probabilistico e l'inferenza ha qui un carattere di necessità essendo l'intera popolazione oggettivamente inaccessibile in quanto frutto esclusivamente di una costruzione intellettuale, ad esempio di una teoria fisica, economica, ecc.. E' però anche evidente che tale costruzione non potrà mai corrispondere alla realtà. Esiste forse un acino perfettamente sferico?

Lo statistico applicato non si può permettere come Pascal di solo sollevare la questione, ma deve fornire una risposta al quesito e senza consultare la sfera di cristallo del mago.

Sia per l'inferenza verso i parametri descrittivi della popolazione obiettivo, sia verso i parametri del modello, la Statistica ha l'obiettivo di mettere a punto metodologie che permettano di effettuare queste inferenze in modo scientifico, ovvero documentando il procedimento e quantificando probabilisticamente l'incertezza inevitabilmente presente nelle conclusioni.

Fra la fase in cui lo statistico definisce il MCS e si accinge ad attivare il procedimento inferenziale, vi è tutto un insieme di operazioni relative alla effettiva individuazione e osservazione delle unità, alla registrazione delle modalità che i caratteri assumono su tali unità, alla costruzione di archivi informatizzati su cui poter efficacemente operare. Queste operazioni, che costituiscono l'attività di rilevazione sul campo, non sono di esclusiva competenza dello statistico, ma lo coinvolgono in quanto i vincoli delle risorse, le scelte organizzative, le modalità operative interagiscono con le scelte strettamente statistiche e hanno un effetto sulla qualità finale dei dati.

Anche qualora non si voglia attivare il procedimento inferenziale, ma limitarsi semplicemente, si fa per dire, a una statistica descrittiva della popolazione definita con il MCS, è

ben noto che agiscono fattori di disturbo, così come in qualunque processo produttivo, che possono rendere il prodotto finale più o meno difforme da quello progettato. Liste molto estese elencano e classificano le fonti di possibili errori e i rimedi sia di tipo preventivo che correttivo per limitarne gli effetti (Lessler e Kalsbeek, 1992). Già a livello di definizione del MCS possono sorgere errori: questo non potrebbe essere adeguato all'esigenza informativa, troppo costoso, o addirittura infattibile. Si riprenda ad esempio l'ipotesi sopra descritta della rilevazione della qualità del servizio di trasporto della linea 26. Come operativamente acquisire l'informazione sul movimento dei passeggeri che usano questa linea? Con un marcapersone digitale alle porte? Impossibile. Con rilevatori? Quanti? Forse troppi. Dobbiamo riempire l'autobus di rilevatori?[2].

Assumendo l'esistenza di valori veri, almeno per i parametri quantitativi con i quali vogliamo caratterizzare la popolazione obiettivo, siano essi descrittivi o analitici, gli errori che si verificano nell'indagine sono le discrepanze fra i risultati ottenuti e la realtà. Questi errori vengono generalmente classificati in due principali categorie la seconda delle quali articolata in sottocategorie:

A - Errori campionari
B - Errori non campionari
    B1 - Errori di lista
    B2 - Errori di mancata risposta
    B3 - Errori di misura

Gli errori campionari sono dovuti al fatto di osservare solo una parte della popolazione invece della popolazione intera. Gli errori non campionari riguardano: l'imprecisione dello strumento di identificazione e accesso alle unità della popolazione obiettivo (errori di lista); quelli derivanti dalla impossibilità di ottenere l'informazione da alcune unità della popolazione obiettivo o del campione (errori di mancata risposta); quelli dovuti alla acquisizione di misure errate in particolare connessi agli strumenti e ai metodi di misura (errori di misura).

Notiamo, per inciso, che gli errori non campionari sono presenti sia nell'indagine censuaria, sia in quella campionaria; non è da escludere che nella prima siano più elevati a causa della maggior complessità del processo di contatto delle unità statistiche e raccolta delle informazioni.

Ovviamente sia i singoli errori, sia l'errore complessivo relativo a una quantità oggetto dell'indagine restano ignoti; se fossero noti potrebbero venir corretti. Quello che ci aspettiamo dalla Statistica è una quantificazione probabilistica di tali errori; dopo di che ci piacerebbe poter esprimere l'errore totale in funzione degli strumenti scelti per la realizzazione dell'indagine. Dato poi il sistema dei vincoli su tali strumenti, ricavare l'allocazione delle risorse che minimizza l'errore totale.

Ci sono troppe scelte interrelate e così tante variabili di cui si dovrebbe tener conto che una formulazione analitica del problema di ottimo è pressoché inconcepibile (Särndal, 1992). Già insormontabile risulta la quantificazione della gran parte degli strumenti usati per l'indagine e l'effetto che hanno sugli errori: si pensi a come si possa quantificare il comportamento dei rilevatori, e ammesso di averlo fatto a come agisce su ciascuna tipologia di errore. Il concetto di disegno complessivo dell'indagine (total survey design) serve soprattutto come modello concettuale per avere una visione complessiva del processo ed evitare di accanirsi su una scelta operativa magari costosa mirante alla riduzione di una fonte di errore senza accorgersi che da altre parti si aprono falle incontrollate.

Gli errori campionari sono quelli che hanno ricevuto maggior attenzione dagli studi metodologici e per i quali esiste attualmente una teoria consolidata ed organica. Per gli altri tipi di errori, anch'essi oggetto di studio, ma oggettivamente di maggior difficoltà di trattamento, non si dispone di una teoria altrettanto consolidata.

[2] L'osservazione stessa produce una alterazione del sistema osservato. Così immergendo un termometro in un bicchiere d'acqua non misuriamo la temperatura dell'acqua nel bicchiere, ma quella del sistema acqua-termometro.

2.3. Il problema dell'autoselezione posto dalle nuove tecnologie

La continua crescita delle richieste informative e la domanda di risultati tempestivi, a fronte della necessità di limitare i costi, ha reso in molti casi impraticabile l'osservazione di tutte le unità che costituiscono la popolazione obiettivo; per questo motivo negli ultimi decenni si sono fortemente diffuse e perfezionate le indagini campionarie, dalle grandi indagini degli istituti di statistica pubblica (si pensi alle indagini sulle forze di lavoro) ai sondaggi telefonici realizzati in una serata per fornire una tabella di percentuali al programma televisivo del giorno dopo.

La diffusione della tecnologia internet sta fornendo nuove opportunità per ottenere informazioni statistiche tempestive o addirittura in tempo reale a costi ridotti, superando in parte la necessità di limitare il numero delle unità osservate e quindi del campionamento.

Questa nuova tecnologia ha fatto nascere nuove modalità di indagine. Couper (2000) fornisce una classificazione delle tipologie di indagine svolte tramite internet. Si va dai sondaggi di intrattenimento senza alcuna pretesa scientifica, alle indagini svolte secondo gli standard tradizionali in cui lo strumento è solo un mezzo di comunicazione. Fra questi due estremi si collocano una decina di tipologie. L'utilizzo di internet per le indagini si diffonderà sempre più grazie alle potenzialità comunicative: oltre alle domande tradizionali, lo strumento permette di sottoporre agevolmente uno stimolo visivo, sonoro, un filmato, ecc. e registrare la risposta (la reazione).

Queste nuove modalità di raccolta dei dati avvengono sempre meno nel rispetto dei vincoli sulla selezione del campione di cui parleremo più avanti. Rispetto alla popolazione obiettivo cui si intende riferire i risultati, queste modalità conducono molto spesso all'osservazione di un campione autoselezionato, soprattutto a causa del digital divide. Sorge così una sfida alla Statistica perché, indipendentemente da questa, più o meno implicitamente, si da rilevanza ai risultati, se non altro attraverso la loro diffusione. Poco importa se una breve nota segnala che i risultati non hanno valore scientifico, il fatto stesso di essere pubblicati su un quotidiano dà loro una inevitabile autorevolezza. Il lettore non ha la possibilità di valutare la qualità dei dati ed è costretto a farsi lui un'inferenza di buon senso: se il sondaggio ha raccolto 5.000 opinioni fra i lettori o navigatori di un giornale politicamente orientato verso un polo, il lettore, già di buona cultura visto che legge un giornale, concluderà che i risultati forniscono una informazione di una qualche validità almeno per la popolazione dei lettori di quel giornale e in qualche modo della popolazione con quell'orientamento politico.

Come vedremo fra breve, alla base dell'inferenza statistica, sia essa realizzata con l'impostazione basata sul disegno o con quella basata sul modello, vi sono dei precisi criteri di selezione del campione. All'allontanarsi dal rispetto di questi criteri l'inferenza si fa sempre più critica. Situazioni in cui questi criteri non sono rispettati sono presenti in grado diverso in tutte le indagini, in particolare la difficoltà crescente nel tempo di ottenere le informazioni da tutte le unità della popolazione (nell'indagine censuaria), da quelle del campione selezionato dal ricercatore secondo determinati criteri (nell'indagine campionaria).

Come abbiamo accennato nel precedente capitolo, un modo per aumentare la tempestività, è quello di realizzare stime anticipate. Sostanzialmente questo obiettivo si persegue in due modi o con un misto dei due: selezionare un sottocampione dal quale raccogliere rapidamente i dati e quindi ottenere le stime anticipate; utilizzare il sottoinsieme dei rispondenti entro la scadenza temporale necessaria per la produzione delle stime anticipate. In questo secondo caso il ricercatore ha un limitato o nullo controllo del meccanismo con cui si autoselezionano i rispondenti tempestivi.

Ora, se definissimo valide le inferenze laddove sono rispettati i criteri di selezione e di osservazione del campione, i risultati di tutte le indagini sarebbero da rigettare. L'esperienza, la possibilità in alcuni casi di controlli di coerenza, la presenza di accurata documentazione del processo, ci permette di valutare almeno qualitativamente la qualità dei risultati. Tuttavia senza una

misura quantitativa della qualità dei risultati rimane aperta la questione di cosa supera e cosa no il controllo di qualità.

Un atteggiamento, stimolato dallo sviluppo di indagini statistiche sempre più lontane dal rispetto dei criteri di selezione ed osservazione di un campione di cui diremo fra poco, è quello di "prendere il toro per le corna" e individuare strumenti che permettano, eventualmente a determinate condizioni, l'inferenza a una popolazione finita sulla base di sottoinsiemi della popolazione obiettivo costituitisi per la volontà di alcuni appartenenti a tale popolazione. In altre parole: se e come sia possibile un'inferenza alla popolazione obiettivo a partire da un suo sottoinsieme di volontari autoselezionatisi.

Il problema dell'autoselezione non è però nuovo per la statistica; la tecnologia internet non ha fatto altro che renderlo più macroscopico. Nel campo dell'inferenza su una popolazione finita si presenta tutte le volte che il campione osservato differisce da quello selezionato a causa della non risposta. Il processo di risposta è in effetti un processo di autoselezione. Finché il tasso di risposta rimane alto, può anche essere trascurato, ma in molti casi questo tasso è bassissimo. Quando in una indagine telefonica si contattano 4.000 famiglie, selezionate anche nel massimo rispetto di un qualche criterio e si ottengono meno di 1.000 risposte dai quei componenti delle famiglie che hanno risposto al telefono, siamo in pratica di fronte a un campione di volontari. Indagini con bassissimi tassi di risposta sono pressoché la norma come si ricava visitando il sito dell'Autorità per le Garanzie nelle Comunicazioni all'indirizzo **http://www.AGCOM.it/sondaggi/sondaggi.htm**. Tali sondaggi sono realizzati da importanti istituti di ricerca fornitori dei principali quotidiani, partiti, sindacati, associazioni di consumatori, e costituiscono una attività economica rilevante in termini di addetti e fatturato. Probabilmente il loro fatturato è maggiore del budget dell'Istat. La tabella 2.1 che segue, seguente riporta alcuni elementi sintetici della documentazione prevista dalle normative vigenti per la diffusione dei risultati dei sondaggi.

Tabella 2.1

Alcuni elementi sintetici estratti delle 12 più recenti documentazioni sui sondaggi presenti nel sito dell'Autorità per le garanzie nelle Comunicazioni come previsto dalle delibere sulla materia n.153/02CSP e n.237/03/CSP; alla data del 5 luglio 2007 Data di raccolta

Argomento	Società realizzatrice	Committente	Data di realizzazione	Metodo	Rispondenti	Contatti	Rispondenti su Contatti	Tasso di risposta
I comportamenti e la morale	Demos & Pi. E Demetra	Il Gazzettino	29/11/2006	CATI	1.031	8.565	0,120	0,163
PACS	ISPO	WW.SONDAGGITALIA	26/01/2007	CATI	301	2.514	0,120	0,149
Rischi dei minori (internet e cellulari)	DOXA Spa	Save the Chikdren	28/01/2007	CATI	861	4.001	0,215	-
TFR e fondi pensione	Atlante-Il canocchiale.it	mensile Atlante	22/01/2007	WEB	193	-	-	-
Atteggiamenti nei confronti della pensione	Gfk Eurisko Srl	Gfk CR	27/07/2006	CATI	622	16.884	0,037	0,046
Uso di prodotti cosmetici	Gfk Eurisko Srl	Lacote	18/12/2006	CATI	2.017	12.899	0,156	0,018
Outing (i personaggi famosi)	IPSOS Public Affairs	Fascino PGT	29/01/2007	CATI	1.000	5.871	0,170	-
Notorietà e immagine dei personaggi famosi	TNS Infratest Spa	TNS Infratest	11/12/2006	AUTO	2.000	4.670	0,428	-
Clima politico e apettative di consumo	Dinamiche Srl	Dinamiche Srl	11/01/2007	CATI	1.000	2.251	0,444	-
Dialogo fra maggioranza e opposizione	SWG Srl	SWG Srl	25/01/2006	CATI	700	4.800	0,146	-
La cicogna è un lusso?	ISPO	Speciale TG1-RAI1	17/10/2006	CATI	800	4.906	0,163	-
Indagini per il Rapporto Italia 2007	CIERRE Ricerche Srl	EURISPES	05/12/2006	CATI	1.070	8.867	0,121	-

Nota: La data di realizzazione riporta solo l'ultimo giorno di svolgimento del sondaggio che in realtà è durato da 1 a 5 giorni, salvo due casi di 10 e 22 giorni

Per molti sondaggi non è possibile ricavare il tasso di risposta come rapporto fra il numero delle unità rispondenti e il numero delle unità contattate ed eligibili; ci dobbiamo accontentare del rapporto fra rispondenti e contatti che in effetti è inferiore, anche se non di molto, al tasso di risposta.

Già nel 1990 (Fabbris 1990) in occasione del Seminario di studio su "Rilevazioni per campione delle opinioni degli italiani" era emerso il distacco fra produttori di sondaggi che devono farli per rispondere a un'incalzante domanda e statistici teorici che sostengono debbano farsi nel rispetto delle metodologie. Non sembra che a tutt'oggi questo gap si sia ridotto, e forse le difficoltà a realizzare il sondaggio secondo paradigmi scientifici si sono aggravate. Non c'è dubbio su chi

abbia la meglio: tocca allo statistico, sporcandosi anche un po' le mani, fornire qualche soluzione, come invitava Finney (1974):

Se lo statistico si rifiuta, altri tenteranno l'inferenza, Non dovremmo accettar di applicare la nostra competenza ai dati che esistono, considerandone i limiti e le potenzialità, e in aggiunta cercare di migliorare i dati futuri?

Al che risponde Smith (1993):

Benché condivida in pieno la frase sui dati futuri, …, penso che vi siano situazioni nelle quali dovremmo accettare che i limiti della nostra disciplina ci possono richiedere di rifiutare l'analisi di dati di provenienza ignota, indipendentemente dal fatto che altri lo possano fare o no. La conoscenza dei nostri limiti dovrebbe esser parte della nostra responsabilità professionale.

Resta però la questione di dove si fissa la soglia rifiuto; questione non risolvibile senza una quantificazione della qualità dei dati.

Nell'obiettivo che ci siamo proposti vi sono comunque alcuni punti fermi: la popolazione obiettivo, i caratteri da rilevare, le modalità restano nella completa disponibilità del ricercatore[3].

2.4. Inferenza su popolazioni finite, le due impostazioni

La popolazione statistica finita costituita da un numero N di unità oggetto di campionamento statistico può essere intesa in due modi:

1. Una popolazione statistica realmente esistente, in un dato spazio e tempo la quale é completamente nota rilevando su tutte le unità i caratteri (le variabili) di interesse.
2. La realizzazione di una variabile casuale (superpopolazione) sulla cui legge probabilistica si formula una ipotesi e che è nota una volta che lo sono i suoi parametri.

Come abbiamo detto, nella realtà, specialmente al momento di utilizzare le conclusioni statistiche, è spesso presente un atteggiamento inferenziale a una popolazione più ampia: così dalla proporzione di pezzi difettosi prodotti oggi dalla linea di montaggio L si passa quasi inavvertitamente, ma anche "piuttosto naturalmente" a un giudizio sulla capacità della linea produttiva in un periodo più ampio; così come Pagnoncelli, presentando a Ballarò le tabelle delle stime delle percentuali dei giudizi degli italiani sulla crisi del governo Prodi non ha detto "la stima della percentuale di italiani che *ieri sera hanno ritenuto* si dovesse andare ad elezioni anticipate è dell'xx%" ma "la percentuale degli italiani che *ritengono* si debba andare …", con ciò estendendo il risultato puntuale quantomeno ad un intervallo temporale di alcuni giorni, ipotizzando implicitamente una costanza del giudizio nell'intervallo temporale. Questa "quasi naturale estensione" delle conclusioni sulla popolazione obiettivo finita a una popolazione più ampia rende in qualche modo "debole" il concetto stesso di popolazione finita. D'altro canto l'inferenza a una popolazione realmente esistente assegna una forza particolare ai risultati che difficilmente possono esser messi in dubbio criticando le ipotesi non verificabili che debbono assumersi per formalizzare un modello probabilistico. Almeno concettualmente è sempre possibile la verifica delle conclusioni

[3] Sempre pù spesso, per motivi soprattutto di costo, si cerca di ottenere informazione statistica a partire da fonti amministrative. In questi casi il modello concettuale statistico è definito per obiettivi amministrativi da persona diversa dal ricercatore statistico che li vuole usare con propri obiettivi.

inferenziali sulla popolazione obiettivo finita osservandone tutte le unità[4]. Nel caso delle stime anticipate che ci proponiamo di realizzare, siamo in effetti nella situazione in cui l'urna pochi o molti mesi più tardi verrà davvero rovesciata.

Connesse alle due concezioni della popolazione finita vi sono due teorie dell'inferenza campionaria:

1. Teoria basata sul disegno ed eventualmente assistita dal modello
2. Teoria basata sul modello

Effettuiamo un breve esame di queste due impostazioni con riferimento soprattutto ai vincoli che pongono sul meccanismo di selezione del campione da osservare per ottenere un'inferenza valida.

2.5. Teoria basata sul disegno ed eventualmente assistita dal modello

In questa teoria la popolazione finita è una popolazione statistica concreta la cui conoscenza si ottiene attraverso l'osservazione di tutte le unità; le sintesi statistiche elementari (totale, media, varianza, …) si calcolano come funzioni di tutte le N unità. Volendola caratterizzare completamente con un insieme di parametri occorre prenderne N; per questo motivo i valori $y_1, \ldots, y_N$ assumono il ruolo di parametro della popolazione.

L'obiettivo è quello di fare inferenza sulle sintesi statistiche elementari $f(y_1, \ldots, y_N)$ a partire da una strategia campionaria [p(s), t(s)], dove p(s) indica una distribuzione di probabilità sullo spazio S dei possibili campioni e t(s) uno stimatore di f(·).

Normalmente la funzione f è la combinazione lineare che fornisce il totale:

$$T_y = (1, \ldots, 1)' \, (y_1, \ldots, y_N) \qquad (2.1)$$

Altri parametri (media, varianza, rapporti) vengono espressi come funzioni lineari o meno di totali.

La distribuzione di probabilità p(s) viene definita dal ricercatore e non ipotizzata, in quanto realizzata dal meccanismo fisico scelto per la selezione del campione. E' inoltre l'unica si cui si basa l'inferenza.

Le proprietà statistiche dello stimatore t(s) do T_y sono valutate con riferimento alla distribuzione di probabilità p(s), ad esempio il valore atteso è dato da:

$$E_p\left[t(s)\right] = \sum_S t(s) \cdot p(s) \qquad (2.2)$$

dove la sommatoria si riferisce all'insieme S di tutti i possibili campioni.

Dalla conoscenza della distribuzione p(s) si ottiene la distribuzione della variabile casuale indicatore $\mathbf{I}(s) = [I_1(s), \ldots, I_i(s), \ldots, I_N(s)]$, dove:

$$I_i(s) = \begin{cases} 1 & \text{se } i \in s \\ 0 & \text{se } i \notin s \end{cases} \qquad (2.3)$$

[4] Questa possibilità di rovesciare l'urna presente quando la popolazione è reale produce, nello statistico che affronta l'inferenza verso una popolazione finita, la preoccupazione di esser soggetto a un controllo sebbene ipotetico. Molti scrupoli che assillano questo statistico non sono presenti, a mio parere, per lo statistico che modella i dati. Qui nessuno potrà mai rovesciare l'urna e scoprire il vero modello e quindi quanto le conclusioni inferenziali sono vicine o lontane dal vero.

tenendo conto che Prob(S = s) = Prob[**I**(s) = **i**(s)]

Si definiscono poi le probabilità di inclusione del primo e del secondo ordine:

$$\pi_i = \Pr ob[I_i(s) = 1] = \sum_{s:\, i \in s} p(s) \qquad \pi_{ij} = \Pr ob[I_i(s) = 1, I_j(s) = 1] = \sum_{s:\, i,j \in s} p(s) \tag{2.4}$$

con le quali si ricava: $E(I_i) = \pi_i$, $Var(I_i) = \pi_i(1-\pi_i)$, $Cov(I_i, I_j) = \pi_{ij} - \pi_i \pi_j$.

Si definisce lo stimatore corretto del totale $T_y = \sum_U y_i$ (Horvitz e Thompson, 1952) come combinazione lineare degli indicatori (y_i sono considerate costanti ignote):
:

$$\hat{T}_\pi = \sum_S \frac{y_i}{\pi_i} = \sum_U I_i \frac{y_i}{\pi_i} \tag{2.5}$$

la cui varianza è data da:

$$V_p(\hat{T}_\pi) = \sum\sum_U (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j} \tag{2.6}$$

ed è stimata correttamente da:

$$\hat{V}_p(\hat{T}_\pi) = \sum\sum_S \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \frac{y_i y_j}{\pi_i \pi_j} \tag{2.7}$$

A partire da questo stimatore si definiscono poi gli stimatori di funzioni lineari e non di totali. Dalla correttezza di $\hat{T}_\pi$ e dalla regolarità delle funzioni discende la correttezza degli stimatori di funzioni lineari dei totali e la consistenza degli stimatori di funzioni non lineari.

Infine sulla base del teorema del limite centrale, sebbene introducendo una successione crescente di popolazioni e di campioni, si perviene alla costruzione dell'intervallo di confidenza:

$$CI_\alpha = \hat{T}_\pi \pm z_{\alpha/2} \sqrt{\hat{V}(\hat{T}_\pi)} \tag{2.8}$$

Tale intervallo risulta necessariamente avere un livello di confidenza approssimato per la popolazione finita e il cui significato è quello di frequenza relativa nell'insieme dei possibili campioni ottenibili con il dato disegno dell'Indicatore di copertura Ic di T:

$$Ic = \begin{cases} 1 & \text{se } CI_\alpha \text{ contiene T} \\ 0 & \text{altrimenti} \end{cases} \tag{2.9}$$

In questa impostazione è fondamentale che il campione sia selezionato e osservato nel rispetto del disegno scelto altrimenti le proprietà statistiche dello stimatore non sarebbero

conosciute; una espressione come la (2.2) non corrisponderebbe alla realtà. Per la correttezza di $\hat{T}_\pi$ è necessario che: $\Pr ob(I_i = 1)$ sia proprio la π_i che compare nella (2.4). Se $\Pr ob(I_i = 1)$ fosse ignota, non sapremmo quanto vale $E(\hat{T}_\pi) - T$, cioè se lo stimatore (2.5) è o meno corretto.

Va rilevato che la selezione nel completo rispetto del disegno scelto, non è sufficiente, se non accompagnata dall'osservazione delle unità selezionate. Se il campione è osservato nel rispetto del disegno, la validità dell'inferenza è garantita dalla correttezza dello stimatore HT del totale e dalla correttezza dello stimatore della sua varianza.

A fronte di un problema concreto, questa impostazione mette a disposizione la strategia campionaria [p(s), t(s)]. Si tratta di uno strumento molto flessibile su cui agire per aumentare l'efficienza sulla base dell'informazione disponibile: questa può intervenire sia nella scelta di p(s), sia nella scelta di t(s). La presenza di informazioni ausiliarie, correlate con la/e variabile/i oggetto di studio permette di individuare stimatori ottenuti dalla specificazione di un modello descrittivo[5] di regressione che formalizza la relazione fra le variabili ausiliarie e quella/e di studio.

Lo stimatore GREG (stimatore di regressione generalizzata) è costruito nel modo seguente.

Si assume un modello descrittivo della popolazione che mette in relazione[6] la variabile oggetto di studio y e le variabili ausiliarie **X**.

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{Xb} \\ V(\mathbf{y}) &= \mathbf{V} = \mathrm{diag}(\sigma_1, \ldots, \sigma_N) \end{aligned} \qquad (2.10)$$

Si stima **b** dalle osservazioni campionarie come combinazione (non lineare) di stime di totali

$$\hat{\mathbf{b}} = (\mathbf{X}_s' \mathbf{V}_{ss}^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{V}_{ss}^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s \qquad (2.11)$$

dove: $\mathbf{V}_{ss} = \mathrm{diag}(\sigma_i)$ $\boldsymbol{\Pi}_s = \mathrm{diag}(\pi_i)$; l'indice s indica che le matrici sono quelle relative alle unità del campione.

Si ottiene lo stimatore del totale:

$$\hat{T}_{GREG} = \sum_U \mathbf{X}\hat{\mathbf{b}} + \sum_s \frac{y_i - \mathbf{X}\hat{\mathbf{b}}}{\pi_i} = \sum_U \hat{y}_i + \sum_s \frac{\hat{\varepsilon}_i}{\pi_i} \qquad (2.12)$$

Se poi la struttura di **V** è tale che $\mathbf{V1}_N \in \langle \mathbf{X} \rangle$ (il vettore delle varianze di **y** è combinazione lineare dei regressori), allora la (2.12) si semplifica nel modo seguente:

$$\hat{T}_{GREG} = \sum_U X\hat{b} = \sum_U \hat{y}_i \qquad (2.13)$$

[5] Così come si considera parametro descrittivo della popolazione finita il totale T_y, analogamente si possono considerare parametri descrittivi la media, i rapporti e i coefficienti di regressione, che formalmente altro non sono che funzioni di totali.

[6] Una qualche ambiguità sorge quando diciamo "modello descrittivo che mette in relazione **y** con **X**". Il concetto di relazione statistica conduce alla considerazione di un termine di errore sul quale formulare una ipotesi distributiva.

Il modello ha in questa impostazione un ruolo solo strumentale per la definizione dell'algoritmo dello stimatore. L'aderenza del modello alla realtà ha effetto sull'efficienza, ma non sulla consistenza dello stimatore, in quanto le proprietà statistiche restano valutate rispetto al disegno p(s).

Lo stimatore GREG può essere espresso in forme algebricamente alternative che evidenziano vari suoi significati (Särndal e altri, 1992). Consideriamo la quantità:

$$Q = \hat{T}_\pi + (T - \hat{T}_\pi) \tag{2.14}$$

dove si "corregge" lo stimatore HT con il suo errore di stima. Qualunque sia il campione S la quantità Q è priva di errore: Q = T . Ovviamente non possiamo calcolare Q non conoscendo T, che è l'obiettivo dell'inferenza. Supponiamo di disporre di un set di variabili ausiliarie **X** linearmente correlate alla **y**: potremmo allora stimare l'errore di stima con:

$$(T - \hat{T}_\pi\hat{)} = (\mathbf{T}_X - \hat{\mathbf{T}}_{X,\pi})'\hat{\mathbf{b}} \tag{2.15}$$

dove $\mathbf{T}_X = \sum_U \mathbf{x}_i$, $\hat{\mathbf{T}}_{X,\pi} = \sum_S \frac{\mathbf{x}_i}{\pi_i}$.

Se la relazione lineare **y** = **Xb** valesse esattamente si stimerebbe l'errore di stima senza errore e quindi:

$$\hat{T}_{GREG} = \hat{T}_\pi + (T - \hat{T}_\pi\hat{)} = \hat{T}_\pi + (\mathbf{T}_X - \hat{\mathbf{T}}_{X,\pi})'\hat{\mathbf{b}} \tag{2.16}$$

sarebbe uno stimatore senza errore. In pratica questo non si verifica, ma ci aspettiamo che tanto più è forte questa relazione, tanto più precisamente venga stimato l'errore di stima e con ciò tanto più efficiente sia $\hat{T}_{GREG}$ rispetto a $\hat{T}_\pi$.

Resta fermo che le proprietà statistiche degli stimatori $\hat{T}_\pi$ e $\hat{T}_{GREG}$ sono definite e valutate rispetto al disegno p(s). Per lo stimatore GREG la validità dell'inferenza è assicurata dalla consistenza dello stimatore del totale e dalla consistenza dello stimatore della sua varianza.

Un modo alternativo di esprimere le stimatore GREG è il seguente.

$$\hat{T}_{GREG} = \hat{T}_\pi + (\mathbf{T}_X - \hat{\mathbf{T}}_{X,\pi})'\hat{\mathbf{b}} = \sum_S \frac{y_i}{\pi_i} + (\mathbf{T}_X - \hat{\mathbf{T}}_{X,\pi})'(\mathbf{X'_S}\,\mathbf{V^{-1}}\mathbf{X_S})^{-1}\sum_S \frac{\mathbf{x_i} y_i}{\sigma_i^2 \pi_i} = \sum_S g_{i,S}\frac{y_i}{\pi_i} \tag{2.17}$$

dove: $g_{i,S} = 1 + (\mathbf{T}_X - \hat{\mathbf{T}}_{X,\pi})'(\mathbf{X'_S}\,\mathbf{V^{-1}}\mathbf{X_S})^{-1}\frac{\mathbf{x}_i}{\sigma_i^2}$ (2.18)

Lo stimatore GREG non è lineare, e per ricavare la sua varianza si considera la varianza della sua approssimazione lineare nell'intorno dei valori attesi dei totali. Si ottiene (Särndal e altri, 1992):

$$AV_p(\hat{T}_{GREG}) = \sum\sum_U (\pi_{ij} - \pi_i\pi_j)\frac{E_i E_j}{\pi_i \pi_j} \tag{2.19}$$

dove $E_i = y_i - \mathbf{Xb}$ è il residuo della regressione sull'intera popolazione finita.
Uno stimatore della varianza (2.19) è fornito da:

$$\hat{V}_p(\hat{T}_{GREG}) = \sum\sum_S \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}} \frac{g_{iS}e_i g_{jS}e_j}{\pi_i\pi_j} \qquad (2.20)$$

dove $e_i = y_i - \mathbf{X}\hat{\mathbf{b}}$ sono i residui della regressione sulle unità del campione e $g_{i,s}$ sono dati dalla (2.18)

Nella espressione (2.17) lo stimatore GREG può essere interpretato come uno stimatore in cui si "aggiustano" i pesi base $d_i = 1/\pi_i$ con pesi correttivi $g_{i,s}$. I pesi finali $(1/\pi_i) \cdot g_{i,s}$ godono della proprietà:

$$\sum_S g_{i,S} \frac{\mathbf{x_i}}{\pi_i} = \mathbf{T_X} \qquad (2.21)$$

La proprietà (2.21) può essere soddisfatta anche con pesi correttivi diversi da quelli che derivano dalla regressione. Sviluppando questa idea, sempre nell'ambito dell'impostazione basata sul disegno è stato definito (Deville e Särndal, 1992) un procedimento di stima che va sotto il nome di calibrazione.

Lo stimatore calibrato si definisce come:

$$\hat{T}_W = \sum_S w_i y \qquad (2.22)$$

Si ricavano i pesi w_i imponendo il vincolo:

$$\sum_S w_i \mathbf{x}_i = \mathbf{T_X} \qquad (2.23)$$

e minimizzando una funzione distanza G(**w**, **d**) fra il vettore $\mathbf{w} = (w_1, \ldots, w_s)$ e il vettore $\mathbf{d} = (1/\pi_1, \ldots, 1/\pi_s)$ dei pesi base. Siccome con i pesi **w**, come con quelli presenti nello stimatore GREG, si ottengono stime senza errore dei totali noti $\mathbf{T_X}$, c'è da aspettarsi che lo stimatore calibrato abbia una varianza inferiore, di molto inferiore se le **X** spiegano bene la y, rispetto allo stimatore HT. Lo stimatore GREG risulta poi un caso particolare di stimatore calibrato. Gli autori dimostrano poi che lo stimatore calibrato è asintoticamente equivalente allo stimatore GREG.

Recentemente (Särndal e Lundström, 2005), il metodo di stima basato sulla calibrazione è stato proposto, nell'ambito del campionamento in due fasi[7], per la stima in presenza di non risposta col duplice obiettivo: ridurre la varianza dello stimatore e l'eventuale bias dovuto alla non risposta, utilizzando l'informazione ausiliaria disponibile.

2.6. Teoria basata sul modello

In questa teoria la popolazione finita $y_1, \ldots, y_N$ è considerata la realizzazione di una variabile casuale $\mathbf{Y_1, \ldots, Y_N}$ la cosiddetta superpopolazione. L'obiettivo è ancora quello di fare inferenza su una funzione $f(y_1, \ldots, y_N)$ sulla base dell'osservazione di un campione $y_1, \ldots, y_n$.

[7] Per una descrizione del campionamento in due fasi si veda Särndal e altri (1992)

"prevedendo" le realizzazioni y_{n+1}, …, y_N non osservate. A tal fine è necessario specificare un modello probabilistico che permetta di formalizzare la relazione fra le y osservate e quelle non osservate, come normalmente si fa in ambito econometrico. La struttura stocastica di quello che osserviamo è specificata, a meno di parametri da stimare, da un modello, e non dipende dal meccanismo di selezione del campione.

Si considerano normalmente stimatori di combinazioni lineari del vettore $\mathbf{y} = (y_1, \ldots, y_N)'$ $\theta = \mathbf{g'y}$ quale ad esempio il totale T quando $\mathbf{g} = (1, \ldots, 1)'$. Per ogni campione s di dimensione n possiamo riordinare il vettore **y** in modo che i primi n elementi siano quelli del campione, sì che la combinazione lineare oggetto di stima sia espressa come:

$$\theta = \mathbf{g'y} = (\mathbf{g_s'}, \mathbf{g_r'})(\mathbf{y_s'}, \mathbf{y_r'})' = \mathbf{g_s'y_s} + \mathbf{g_r'y_r} \tag{2.24}$$

Nel caso del totale:

$$T_y = \mathbf{1'y} = \mathbf{1_s'y_s} + \mathbf{1_r'y_r} \tag{2.25}$$

realizzazione della variabile casuale $\mathbf{1_s'Y_s} + \mathbf{1_r'Y_r}$. Allora il problema di stimare $\mathbf{1'y}$ si riduce a quello di predire il valore $\mathbf{1_r'y_r}$ della variabile non osservata $\mathbf{1_r'Y_r}$.

Viene specificato un modello probabilistico per la popolazione:

$$\begin{aligned} E_M(\mathbf{Y}) &= \mathbf{Xb} \\ V_M(\mathbf{Y}) &= \mathbf{V} \end{aligned} \tag{2.26}$$

dove **X** é la matrice (N, p) delle p variabili ausiliarie note per tutte le N unità, **b** il vettore (p, 1) dei parametri, **V** la matrice (N, N) di varianza-covarianza che possiamo esprimere a blocchi come:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{rs} \\ \mathbf{V}_{sr} & \mathbf{V}_{rr} \end{bmatrix} \tag{2.27}$$

dove gli indici s e r indicano rispettivamente il campione e il suo complemento in U.

Si dimostra (Valliant e altri, 2000) che il predittore lineare corretto e ottimale (BLUP) di T_y è dato da:

$$\hat{T}_{ott} = \mathbf{1}_s' \mathbf{Y}_s + \mathbf{1}_r' \left[\mathbf{X}_r \hat{\mathbf{b}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \hat{\mathbf{b}}) \right] \tag{2.28}$$

dove:

$$\hat{\mathbf{b}} = (\mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{Y}_s \tag{2.29}$$

La varianza d'errore di $\hat{T}_{ott}$ è data da:

$$\begin{aligned} &\mathrm{Var}_M(\hat{T}_{ott} - T) = \mathbf{1}_r' (\mathbf{V}_{rr} - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr}) \mathbf{1}_r + \mathbf{1}_r' (\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s) \mathbf{A}^{-1} (\mathbf{X}_r - \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \mathbf{X}_s) \mathbf{1}_r \\ &\text{con} \quad \mathbf{A} = (\mathbf{X}_s' \mathbf{V}_{ss}^{-1} \mathbf{X}_s) \end{aligned} \tag{2.30}$$

Per la stima della varianza, si adottano metodi di stima robusti, come lo stimatore sandwich (Valliant e altri, 2000).

Nel caso in cui $\mathbf{V_{rs}} = \mathbf{0}$, cioé quando le $\mathbf{Y_s}$ e $\mathbf{Y_r}$ sono incorrelate, si ottiene:

$$\hat{T}_{ott} = \mathbf{1}_s' \mathbf{Y}_s + \mathbf{1}_r' \mathbf{X}_r \hat{\mathbf{b}} = \sum_s Y_i + \sum_r \hat{Y}_i = \sum_U \hat{Y}_i + \sum_s \left(Y_i - \hat{Y}_i\right) = \sum_U \hat{Y}_i + \sum_s \hat{\varepsilon}_i \qquad (2.31)$$

Se poi la struttura di $\mathbf{V}$ è tale $\mathbf{V1}_N \in \langle \mathbf{X} \rangle$ (il vettore delle varianze di $\mathbf{Y}$ è combinazione lineare dei regressori),

$$\hat{T}_{ott} = \mathbf{1}_N' \mathbf{X}\hat{\mathbf{b}} = \sum_U \hat{Y}_i \qquad (2.32)$$

Va notata la somiglianza fra la (2.12) e la (2.31) fra la (2.13) e la (2.32); tuttavia il vettore **b** non è stimato nello stesso modo nelle due impostazioni; vi sono però situazioni in cui i due stimatori sono algebricamente pressoché uguali e addirittura coincidono.

Le proprietà statistiche dello stimatore dipendono dal modello che specifica la distribuzione di probabilità del vettore $\mathbf{Y}$. Se questo è corretto non ha alcuna importanza ai fini di una inferenza valida quale campione viene selezionato purché il meccanismo di selezione non dipenda da Y (selezione ignorabile). Criteri di scelta del campione possono esser dettati da motivi di efficienza.

Nella realtà però difficilmente un modello, che è sempre una semplificazione, può descrivere esattamente la $\mathbf{Y}$ come in (2.26). Se il modello è malspecificato lo stimatore è distorto.

Per rendere lo stimatore robusto almeno rispetto a una classe di modelli occorre disporre di un campione bilanciato. Consideriamo la classe di modelli polinomiale con un solo regressore:

$$Y_i = \sum_{j=0}^{J} \delta_j b_j x_i^j + \varepsilon_i v_i^{1/2} \qquad (2.33)$$

dove gli errori sono $\varepsilon_i \sim (0, \sigma^2)$ e incorrelati, b_j sono i parametri incogniti da stimare, δ_j indicatori della presenza della j-esima potenza, v_i una funzione nota di i.

Supponiamo di aver adottato come modello di lavoro uno più parsimonioso di (2.33), ad esempio quello $Yi = b + \varepsilon_i$ da cui deriva lo stimatore per espansione $\hat{T}_0 = N\overline{Y}_s$. Tale stimatore risulterà distorto sotto il modello (2.33). Se però il campione soddisfa la condizione di bilanciamento:

$$\frac{1}{n}\sum_s x_k^i = \frac{1}{N}\sum_U x_k^i \qquad i = 1, 2, \ldots, J \qquad (2.34)$$

lo stimatore per espansione risulta corretto. Ovvero lo stimatore per espansione risulta corretto per la classe più ampia dei modelli polinomiali di grado J. E siccome con i polinomi si approssima qualunque funzione abbastanza regolare, la correttezza è garantita almeno approssimativamente qualunque sia il vero modello se ben approssimato dal polinomio di grado J.

Si estende poi il bilanciamento generalizzando la (2.34) al bilanciamento ponderato e a più covariate. Vanno fatte due osservazioni:

- Può essere difficile in pratica ottenere un perfetto bilanciamento soprattutto in presenza di numerosi regressori, in tal caso la sensibilità del valore atteso degli stimatori all'allontanamento dal bilanciamento è più elevata per quelli basati su modelli ridotti rispetto a quello polinomiale.
- Quanto poi alla varianza degli stimatori, anche nelle condizioni di bilanciamento, non é la stessa e tende ad essere più grande per quelli più ridotti.

L'impostazione basata sul modello richiede due condizioni:

- la disponibilità di adeguata informazione ausiliaria correlata con la Y in modo da poter formulare un modello il più possibile aderente alla realtà,
- **la selezione di un campione bilanciato per garantirsi la correttezza dello stimatore in presenza di malspecificazione del modello**.

2.7. Il dibattito fra le due teorie

Il campione bilanciato riporta sulla scena il concetto di campione rappresentativo che ha accompagnato la nascita delle indagini campionarie a cavallo dell'inizio del secolo scorso fornendogli una base teorica. Storica è l'applicazione fatta da Gini e Galvani (1929) di quello che allora veniva definito "metodo rappresentativo" consistente nel selezionare (a quell'epoca manualmente!) un sottoinsieme di unità della popolazione obiettivo in modo che i momenti della distribuzione congiunta di un set di variabili note fossero uguali nel campione e nella popolazione.

Gli autori si accorsero presto che una rappresentatività assoluta era impossibile.

Comunque, indipendentemente da qualsiasi considerazione di ordine teorico, l'esperienza da noi fatta dimostra che un campione può essere dichiarato rappresentativo per quanto riguarda alcuni aspetti di determinati caratteri, ma, almeno generalmente, non può venir dichiarato tale per tutti i caratteri e per tutti i loro aspetti. (pag. 22)

Gli autori definiscono allora una rappresentatività parziale (relativa):

. . . sostituendo alla consueta definizione di rappresentatività una definizione particolaristica che possa riferirsi a ad uno o ad alcuni determinati aspetti di uno o alcuni determinati caratteri ….. Con ciò non si vuole escludere l'astratta possibilità che un campione possa per tutti i caratteri esaminabili e per tutti i loro aspetti, essere perfettamente rappresentativo della totalità: si pensi al caso banale di un insieme A+A' costituito da due insiemi A e A' perfettamente uguali. L'insieme parziale A costituirà un campione rappresentativo in senso assoluto dell'insieme totale A+A'. (pag. 23)

Il tentativo era destinato a fallire perché, anche considerando un solo carattere, la popolazione finita di N unità è completamente descritta dai momenti di ordine da 1 a N (è quindi un "oggetto" a N dimensioni), mentre un campione di dimensione $n < N$ al più può fornire n momenti coincidenti con quelli della popolazione; a maggior ragione, considerando k caratteri e i momenti misti. Proprio matematicamente non può esistere un campione con la stessa distribuzione di frequenza relativa congiunta di una qualsiasi popolazione, salvo il caso di popolazioni le cui frequenze congiunte ammettono tutte un comune divisore d, cioè esiste una sottopopolazione di dimensione $n = N/d$ simile. In tal caso tutti i momenti semplici e misti di tutti gli ordini sono uguali. Ed è appunto il caso, ma l'unico, dell'esempio riportato dagli autori con $d = 2$.

Nel 1934, con un famoso articolo, Neyman contestava il metodo rappresentativo e poneva le basi della impostazione basata sul disegno definendo rappresentativo non tanto il campione quanto il binomio campione-stimatore che permettesse una valida conclusione inferenziale.

Se siamo interessati a un carattere collettivo X di una popolazione P e usiamo metodi di campionamento e di stima che ci permettano di assegnare ad ogni campione S un intervallo di confidenza $C = [X_1(S) , X_2(S)]$ tale che la frequenza dell'errore nell'affermazione $X \in C$ non supera il limite $1-\alpha$ in precedenza stabilito, indipendentemente dalle caratteristiche ignote della

popolazione, io sostengo un tale metodo di campionamento rappresentativo e il metodo di stima consistente.

Al concetto di campione rappresentativo Neyman contrappose quello di campione stratificato mostrando inoltre come l'allocazione ottimale non fosse quella proporzionale, ma quella che da allora prende il suo nome.

L'impostazione di Neyman ha dominato incontrastata fino a metà degli anni 50, quando l'articolo di Godambe (1955) ha evidenziato come nella classe degli stimatori lineari non poteva esistere uno stimatore di minima varianza per ogni popolazione. Successivamente sono emerse le difficoltà ad applicare all'impostazione basata sul disegno i concetti di verosimiglianza, di sufficienza e di condizionamento. L'applicazione dei principi dell'inferenza basata sulla verosimiglianza fallisce nel caso di popolazione finita soprattutto a causa dello spazio parametrico che è N-dimensionale, mentre i dati del campione sono solo n < N. Da qui è sorta l'impostazione alternativa basata sul modello nella quale si ipotizza un legame fra gli n valori osservati e gli N-n non osservati attraverso un modello probabilistico (Royall, 1970).

La fondamentale differenza fra le due teorie consiste nella distribuzione di probabilità alla base dell'inferenza. Nella teoria basata sul disegno (ed eventualmente assistita dal modello) la distribuzione di probabilità è quella costruita da chi progetta l'indagine al momento in cui stabilisce il meccanismo probabilistico di selezione del campione, e quindi delle unità dalla popolazione. Questa distribuzione è ovviamente nota e permette di ricavare le proprietà statistiche degli stimatori[8].

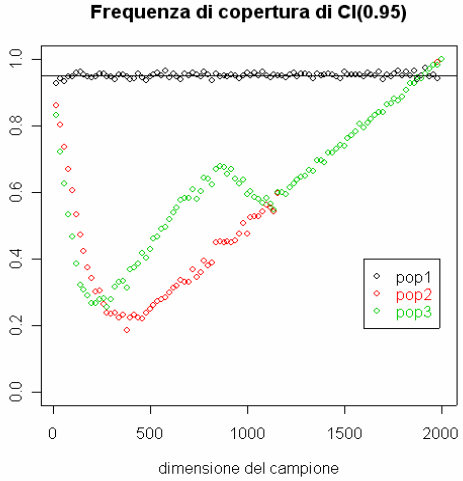
Nella teoria basata sul modello si assume una distribuzione di probabilità secondo lo schema classico dell'inferenza. I valori osservati e quelli del complemento non osservati sono collegati dal modello probabilistico assunto. Si stimano i parametri del modello sulle unità osservate e si predice

---

[8] Dovremmo aggiungere: "con alcune ipotesi, sebbene non molto restrittive, sulla popolazione". Se la popolazione finita è molto asimmetrica, o presenta alcuni valori anomali, resta valida la correttezza dello stimatore della varianza, ma l'approssimazione alla normale della distribuzione campionaria può risultare del tutto inadeguata per la costruzione di un intervallo di confidenza approssimato. Abbiamo considerato le tre popolazioni

```
pop1:       y1 =  0.001,  0.002, . . . , 1.997, 1.998,  1.999,   2
pop2:       y2 =  0.001,  0.002, . . . , 1.997, 1.998,  1,999, 200
pop3:       y3 =  0.001,  0.002, . . . , 1.997, 1.998,  50,    200
```

nella seconda popolazione abbiamo sostituito il valore 2 con 200; nella terza abbiamo sostituito il valore 1,999 e 2 con 50 e 200 rispettivamente. Attraverso un ciclo di simulazioni Monte Carlo abbiamo ricavato la frequenza di copertura dell'intervallo di confidenza per la stima del totale con campioni casuali semplici di dimensioni n= 20 , 40, 60, . . . , 1980, 2000. Il grafico seguente riporta il confronto fra il valore nominale 95% e i valori empirici della frequenza di copertura.



Nel caso della popolazione 2, il campione casuale semplice di dimensione non modesta (n=500), fornirebbe conclusioni inferenziali del tutto inaccettabili. Il fatto che lo stimatore del totale e quello della sua varianza sono corretti rispetto al disegno é una magra consolazione.

il valore delle Y per le unità non osservate, quindi il valore di una funzione delle Y dell'intera popolazione finita in una logica tipicamente econometrica.

Nella prima impostazione le proprietà degli stimatori (valore atteso, varianza, …) si riferiscono alla distribuzione di probabilità nell'insieme dei possibili campioni che si possono o si sarebbero potuti estrarre con quel criterio di selezione (si parla per questo di inferenza procedurale).

La seconda adotta gli strumenti dell'inferenza classica: assunto un modello parametrico, si stimano i parametri e le funzioni di questi che interessano; le proprietà statistiche di tali funzioni discendono dalle ipotesi esplicitate nella formulazione del modello probabilistico generatore delle osservazioni.

Vi sono stati accesi dibattiti, soprattutto in passato, sui meriti e sulla validità delle due impostazioni. L'impostazione basata sul disegno viene criticata su due versanti; quello teorico connesso ad alcuni principi e strumenti fondamentali dell'inferenza classica: il principio di verosimiglianza, la sufficienza, la individuazione di stimatori UMVUE; quello pratico: non è autosufficiente per l'inferenza analitica e per affrontare le situazioni degli errori non campionari.

L'impostazione basata sul modello viene criticata per la quasi insormontabile difficoltà a formulare un modello probabilistico che tenga conto delle numerose variabili dell'indagine, delle strutture identificabili della popolazione obiettivo, come i sottogruppi, per il rischio di distorsioni anche notevoli derivanti dalla malspecificazione del modello (Hansen, Madow, Tepping, 1983).

Attualmente in un clima più "ecumenico" e orientato alle soluzioni pratiche (Smith, 1994; Barnett, 1999) gli strumenti vengono valutati in riferimento all'informazione disponibile, alla conoscenza che si ha dei fenomeni oggetto di indagine, alle possibilità di rispondere a esigenze concrete, in particolare alle difficoltà crescenti che si incontrano nella realizzazione pratica delle indagini statistiche, lasciando in secondo piano (o in secondo tempo) gli argomenti strettamente teorici.

Quanto poi all'uso di modelli, già Cochran (1977), maestro dell'approccio basato sul disegno, utilizza il modello per introdurre lo stimatore di regressione. Più recentemente (Särndal e altri, 1992) si fa riferimento a modelli nella fase di progettazione e per la individuazione degli stimatori. Sul versante dell'approccio basato sul modello, si riconosce alla randomizzazione quantomeno una garanzia di imparzialità contro il rischio di fonti anche inconsce di distorsione; si riconosce inoltre che la selezione casuale fornisce almeno in media il bilanciamento, necessario in questo approccio per garantire la robustezza.

In entrambe le impostazioni, alla base della misura della precisione delle stime vi è il ricorso a varianti particolari del teorema del limite centrale per la costruzione di intervalli di confidenza approssimati.

Nella concreta applicazione occorre distinguere fra indagini ufficiali e private, indagini con grandi campioni o con piccoli campioni (Deville, 1991). Nelle indagini pubbliche con grandi campioni rivolte ad una utenza generale si preferisce l'impostazione basata sul disegno perché svincolata da ipotesi che potrebbero pregiudicare la correttezza degli stimatori e che garantisce l'efficienza grazie alla dimensione campionaria; si cerca inoltre di limitare gli errori di non risposta con due strumenti: in fase di progettazione sensilibizzando i rispondenti con l'autorevolezza degli istituti nazionali di statistica; in fase di stima sfruttando l'informazione ausiliaria con stimatori di regressione, o più in generale con stimatori calibrati. Nelle indagini pubbliche l'elemento "imparzialità" nella selezione delle unità che costituiscono il campione ha un ruolo rilevante. Nei sondaggi privati, di solito basati su piccoli campioni, da svolgersi in tempi brevissimi, prevale l'impostazione basata sul modello, nella versione del campionamento per quote, che può essere oggetto di accordo contrattuale con lo specifico cliente; in tali indagini la distorsione può essere tollerabile come quota non eccessiva dell'errore quadratico medio. Deville (1991) indica una dimensione fra 2.500 e 3.000 come soglia pratica per la scelta fra campionamento per quote e campionamento probabilistico.

2.8. Il problema dell'autoselezione

Molti autori (Little, 1982; Rubin,1976; Smith, 1983) hanno affrontato il problema del meccanismo di selezione sulla validità delle inferenze sia nell'impostazione basata sul disegno sia in quella basata sul modello.

Riprendendo la formalizzazione del problema da Smith (1983), il meccanismo di selezione di un campione S da una popolazione di N unità identificabili viene descritto da una variabile casuale vettoriale $\mathbf{I}_S = [I_1, \ldots, I_k, \ldots, I_N]$ dove:

$$I_k = \begin{cases} 1 & k \in S \\ 0 & k \notin S \end{cases} \tag{2.35}$$

Il vettore $\mathbf{I}_S$ determina quali unità sono selezionate nel campione S. Il meccanismo di selezione è completamente descritto allorché si specifica la distribuzione di probabilità di $\mathbf{I}_S$. In generale questa distribuzione può dipendere da informazione nota a priori **Z**, dalla variabile oggetto di studio **Y**, dal parametro ignoto ψ:

$$f(\mathbf{I}_S \mid \mathbf{Z}, \mathbf{Y}, \psi) \tag{2.36}$$

Nella impostazione basata sul disegno la distribuzione di $\mathbf{I}_S$ dipende solo da **Z**, cioè dalle informazioni note che il progettista della strategia campionaria decide di utilizzare per selezionare il campione.

$$f(\mathbf{I}_S \mid \mathbf{Z}) \tag{2.37}$$

Essendo **Z** nota, $f(\mathbf{I}_S \mid \mathbf{Z})$ risulta nota ed'é proprio il cosiddetto disegno p(s) che sta alla base di questa impostazione inferenziale.

Nella impostazione basata sul modello si richiede la specificazione in generale della distribuzione:

$$f(\mathbf{Y} \mid \mathbf{Z}, \theta) \tag{2.38}$$

con la quale si modella la relazione fra la variabile **Y** oggetto di studio e le variabili **Z** note. θ é un parametro da stimare.

La distribuzione congiunta della **Y** e della $\mathbf{I}_S$ è data da:

$$f(\mathbf{Y}, \mathbf{I}_S \mid \mathbf{Z}, \theta, \psi) = f(\mathbf{Y} \mid \mathbf{Z}, \theta)\, f(\mathbf{I}_S \mid \mathbf{Y}, \mathbf{Z}, \psi) \tag{2.39}$$

che fornisce la legge probabilistica congiunta del meccanismo di selezione del campione $\mathbf{I}_S$ e della realizzazione di **Y**.

Indicando con $(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}})$ la partizione di **Y** indotta da S, la probabilità marginale di osservare $\mathbf{I}_S$ e del realizzarsi di $\mathbf{Y}_S$, cioè la probabilità di osservare il campione S e del realizzarsi di dati valori per le unità che entrano a far parte del campione è data da:

$$f(\mathbf{Y}_S, \mathbf{I}_S \mid \mathbf{Z}, \theta, \Psi) = \int f(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}} \mid \mathbf{Z}, \theta) \cdot f(\mathbf{I}_S \mid \mathbf{Y}_S, \mathbf{Y}_{\bar{S}}, \mathbf{Z}, \Psi) d\mathbf{Y}_{\bar{S}} \tag{2.40}$$

Ignorare il meccanismo di selezione equivale a lavorare, per la stima del parametro θ del modello (2.38), con la distribuzione:

$$f(\mathbf{Y}_S \mid \mathbf{Z}, \theta) = \int f(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}} \mid \mathbf{Z}, \theta) \cdot d\mathbf{Y}_{\bar{S}} \tag{2.41}$$

Questo è possibile quando il meccanismo di selezione non dipende da **Y**. In tal caso infatti:

$$\begin{aligned} f(\mathbf{Y}_S, \mathbf{I}_S \mid \mathbf{Z}, \theta, \Psi) &= \int f(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}} \mid \mathbf{Z}, \theta) \cdot f(\mathbf{I}_S \mid \mathbf{Y}_S, \mathbf{Y}_{\bar{S}}, \mathbf{Z}, \Psi) d\mathbf{Y}_{\bar{S}} = \\ &= \int f(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}} \mid \mathbf{Z}, \theta) \cdot f(\mathbf{I}_S \mid , \mathbf{Z}, \Psi) d\mathbf{Y}_{\bar{S}} = \\ &= f(\mathbf{I}_S \mid , \mathbf{Z}, \Psi) \cdot \int f(\mathbf{Y}_S, \mathbf{Y}_{\bar{S}} \mid \mathbf{Z}, \theta) d\mathbf{Y}_{\bar{S}} = \\ &= f(\mathbf{I}_S \mid , \mathbf{Z}, \Psi) \cdot f(\mathbf{Y}_S \mid \mathbf{Z}, \theta) \end{aligned} \tag{2.42}$$

Allora, condizionatamente a **Z**, il meccanismo di selezione $\mathbf{I}_S$ e la variabile osservata $\mathbf{Y}_S$ sono indipendenti.

Negli schemi di campionamento probabilistici cosiddetti non informativi, come quelli espressi dalla (2.37) il meccanismo di selezione non dipende da **Y**. Ma anche schemi non probabilistici permettono di ignorare il meccanismo di selezione e stimare il parametro θ solo sulla base della distribuzione $f(\mathbf{Y}_S \mid \mathbf{Z}, \theta)$. Si tratta di tutti i meccanismi di selezione funzione solo della **Z**: ad esempio, per Z univariata, selezionare le unità con gli n valori Z più grandi o le unità che soddisfano il criteri del bilanciamento sulla variabile Z.

2.9. Strumenti per la stima a partire da un sottoinsieme austoselezionato

Come abbiamo detto nel Cap. 1 intendiamo ottenere stime tempestive sul movimento mensile (arrivi e presenze) dei clienti nelle strutture ricettive della provincia di Firenze, sulla base di un campione di strutture che volontariamente partecipano al sistema turiweb, e quindi sono in grado di fornire pochi giorni dopo la fine del mese i dati sul loro movimento.

Nell'approccio basato sul disegno si considera il disegno in due fasi. Nella prima fase si seleziona un campione s_a secondo il disegno $p_a(s_a)$. Dato s_a si seleziona un campione s di seconda fase secondo il disegno $p(s|s_a)$. In corrispondenza delle due fasi, si ricavano le probabilità di inclusione π_{ai} e $\pi_{i|sa}$, e con queste lo stimatore corretto:

$$\hat{t}_{\pi^*} = \sum_s \frac{y_i}{\pi_{ai}\pi_{i|sa}} \tag{2.43}$$

Si estende poi il metodo di stima assistito dal modello di regressione al campionamento in due fasi utilizzando l'informazione disponibile sia a livello di intera popolazione, sia quella raccolta a livello di campione di prima fase (Sarndal e altri, 1992). Il campionamento in due fasi viene usato, nell'approccio basato sul disegno, per trattare la non risposta. Il campione di prima fase s è quello selezionato per essere osservato, quello di seconda fase r è quello dei rispondenti. Poiché non si conosce, di norma, il meccanismo di risposta, la probabilità

$$\pi_{i|s} = \Pr(i \in r \mid s) \tag{2.44}$$

non è nota, e viene stimata, sulla base di un modello. Un modello semplice è quello dei gruppi di risposta omogenei (RHG: Response Homogeneity Group): il campione osservato è partizionato in H_s gruppi s_h (h = 1, 2, . . ., H) e si assume che in ciascuno di tali gruppi si abbia:

$$\pi_{i|s} = \Pr(i \in r \mid s) = \theta_{hs} > 0 \quad \forall i \in s_h$$
$$\pi_{ij|s} = \Pr(i, j \in r \mid s) = \Pr(i \in r \mid s)\Pr(j \in r \mid s) \quad \forall i \neq j \in s \tag{2.26}$$

Con questo modello si assume che all'interno dei gruppi il meccanismo di risposta sia ignorabile, e che quindi condizionatamente al campione selezionato e al gruppo vi sia indipendenza fra il meccanismo di risposta e la variabile oggetto di studio.

In pratica si affronta il problema della non risposta "aggiustando" la probabilità di inclusione delle unità nel campione s con la probabilità di rispondere dato che si è stati selezionati. Questo metodo integra il disegno con un modello, e si può considerare un approccio basato prevalentemente sul disegno se il tasso di non risposta è modesto.

Nel nostro caso il campione di prima fase coincide con l'intera popolazione, quello di seconda fase è selezionato con un meccanismo formalmente corrispondente a quello di risposta.

Ora, però in questa situazione tutta l'incertezza è connessa alla seconda fase, ed non ha più senso continuare a parlare di approccio basato sul disegno. Vedremo a proposito degli esercizi alberghieri alcuni metodi si stima che utilizzano le probabilità di appartenere al sistema turiweb (propensity scores).

Un approccio alla non risposta è quello proposto da Särndal e Lundström (2005), basato sulla calibrazione che usi totali noti di variabili ausiliarie che possiedano le due caratteristiche:

1. avere una notevole capacità esplicativa del meccanismo di risposta,
2. avere una notevole capacità esplicativa delle variabili di studio.

Specificate le variabili ausiliarie, i relativi totali noti, per un particolare algoritmo di calcolo dei pesi correttivi, si ritrovano gli stimatori GREG (stimatori di regressione generalizzata) inizialmente sviluppati con l'obiettivo di ridurre la varianza sfruttando variabili ausiliarie linearmente correlate alla variabile di studio. In presenza di autoselezione, la speranza è che gli stimatori calibrati, e gli stimatore GREG in particolare, siano in grado di ridurre l'MSE.

Se valesse una relazione lineare esatta fra la variabile di studio y e le ausiliarie **X**, gli stimatori calibrati, e GREG in particolare, fornirebbero stime senza errore.

… *Questo risultato suggerisce che quando esiste una forte relazione lineare fra y e **X** lo stimatore calibrato dovrebbe essere prossimo al valore obiettivo Y. Sia l'errore di campionamento, sia l'errore di nonrisposta dovrebbero essere sostanzialmente eliminati*. (Särndal e Lundström, 2005, Estimation in Surveys with nonresponse, pag.61).

In altre parole, se uno stimatore è molto efficiente e almeno consistente, come lo sono gli stimatore GREG, sarà "poco sensibile" al campione selezionato: fornirà stime che differiscono poco da un campione all'altro. E quindi "poco affetto" dalla possibile distorsione da nonrisposta e quindi da autoselezione.

Nell'approccio basato sul modello il problema dell'autoselezione non si porrebbe se si potesse disporre di un modello esattamente specificato per la variabile di studio e fosse soddisfatta la condizione di Smith: processo di selezione indipendente dalla variabile di studio condizionatamente a un set di varibili esplicative. In tal caso si avrebbe assenza di distorsione delle stime anche con un campione autoselezionato. L'autoselezione potrebbe agire solo sulla varianza degli stimatori. Nella realtà queste due condizioni non si verificano mai esattamente per cui è necessario imporre vincoli alla selezione del campione (es. bilanciamento). Nel nostro caso, non possiamo agire sulla selezione del campione; ma solo sul modello, e in ogni caso nella sua specificazione dobbiamo fare i conti con le covariate disponibili negli archivi disponibili.

Va infine rilevato che a parità di variabili ausiliarie disponibili, specificato un modello per la y, sia nell'approccio basato sul disegno (e assistito dal modello), sia in quello basato sul modello, in molti casi si ottengono stimatori che formalmente sono identici.

Nei prossimi capitoli sperimenteremo l'applicazione di alcuni stimatori che utilizzano le variabili ausiliarie disponibili e, disponendo dell'indagine completa, avremo modo di valutare le loro performances.

Capitolo 3

Nel presente capitolo verranno sperimentati alcuni metodi di stima del totale degli arrivi e delle presenze dei clienti nelle strutture ricettive della provincia di Firenze. La disponibilità dei dati della rilevazione censuaria per i mesi del periodo gennaio 2005 – dicembre 2006 e la conoscenza di quali strutture in tale periodo appartengono al sistema turiweb, permette di ricavare gli errori di stima che si sarebbero ottenuti con i diversi metodi e in tal modo valutare empiricamente la loro bontà.

3.1. I dati disponibili

Per il periodo gennaio 2005 - dicembre 2006 sono stati recuperati gli archivi mensili della rilevazione censuaria. Tali archivi contengono le informazioni su:

- il movimento dei clienti
- la consistenza
- lo stato di attività

Il movimento mensile riporta per ogni struttura il numero di arrivi e presenze per ciascuna nazionalità di provenienza per i clienti stranieri e per ciascuna regione per i clienti italiani.

La consistenza mensile riporta per ogni struttura i posti letto, le camere, i bagni, le giornate letto disponibili (posti letto x numero di giorni di apertura nel mese), .

L'aggiornamento della consistenza avviene contestualmente alla rilevazione del movimento. Solo dopo la fine del mese, a seguito delle verifiche sulla completezza delle comunicazioni, si conosce l'effettiva evoluzione della consistenza in termini di numero di esercizi, loro dimensione e stato di attività del mese (se aperto per almeno alcuni giorni del mese, se chiuso). Un aspetto critico può derivare dalle mancate comunicazioni cui possono essere associate tre situazioni: esercizio non più esistente, esercizio chiuso (quindi con movimento nullo), esercizio aperto che non ha comunicato il movimento. Di norma la risoluzione delle mancate comunicazioni, per quanto possibile, per gli esercizi attivi (sia aperti che chiusi) e quindi il consolidamento dell'archivio della consistenza richiede del tempo; è questo uno dei motivi del ritardo nella raccolta dei dati. Nelle nostre elaborazioni sui dati pregressi abbiamo assunto la consistenza delle strutture attive nel mese in questione come nota alla fine del mese. Nella implementazione pratica c'è da attendersi che questa informazione, necessaria per poter applicare i metodi di stima basati sulle comunicazioni tempestive delle strutture del sistema turiweb, non sia disponibile così tempestivamente.

Anche ammessa nota, prima o poi, la consistenza delle strutture attive, si possono verificare mancate comunicazioni del movimento, quale che sia il termine in cui si decide di "chiudere il mese" sia fra gli esercizi non appartenente che appartenenti al sistema turiweb.

La popolazione obiettivo e i relativi totali noti che abbiamo utilizzato per confrontare le stime sono appunto al netto delle mancate risposte così come consolidati dalla Provincia. In sostanza ci poniamo l'obiettivo di stimare non il "vero" movimento dei clienti, ma di anticipare il movimento che risulterà dall'indagine censuaria al netto delle mancate risposte e di altre fonti di errore quale l'errata risposta.

Nel mese di ottobre tutte le strutture ricettive sono tenute a comunicare alla Provincia l'attrezzatura ricettiva di cui dispongono e i prezzi massimi praticati per l'anno successivo. Per ciascuna tipologia di struttura ricettiva è previsto un modello rilevativo in relazione alle specifiche modalità di offerta di alloggio di tale tipologia. Tale comunicazione deve esser fatta anche dai nuovi

esercizi al momento di inizio dell'attività. Gli archivi che ne derivano forniscono una ricca informazione sui servizi offerti, sia a livello centralizzato, sia per singola unità di alloggio.

Abbiamo acquisito una selezione di variabili per gli esercizi alberghieri (alberghi in senso stretto e residenze turistico alberghiere) aggiornata al dicembre 2006 per la sperimentazione dei metodi di stima basati sui propensity scores.

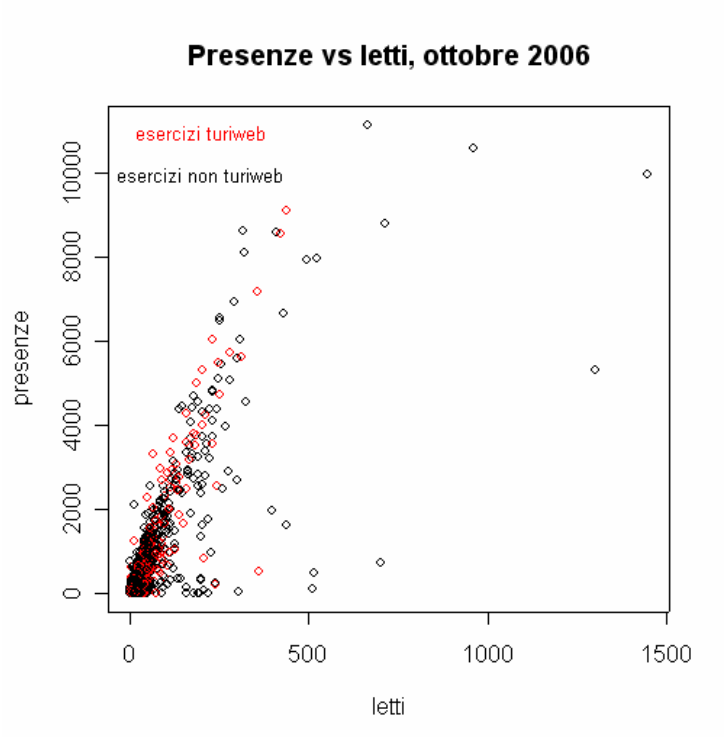
3.2. Stimatori delle presenze e degli arrivi in presenza di autoselezione

Come abbiamo visto il campione degli esercizi del sistema turiweb è distribuito su tutto il territorio provinciale e in tutte le tipologie ricettive; risulta però che gli arrivi e le presenze che si verificano in tale campione (29,7% e 28,6% rispettivamente) sono superiori alla quota che esso rappresenta dell'intera popolazione (26,4%). Di per sé questa differenza, del resto non eclatante, dice poco. Essa si può verificare anche selezionando un campione secondo un definito disegno; se così non fosse il semplice stimatore per espansione fornirebbe stime prive di errore. Dobbiamo però "ragionevolmente supporre" che il meccanismo di autoselezione sia potenzialmente distorsivo, cioè non sia ignorabile e richieda un qualche aggiustamento.

Una variabile rilevante, certamente assai correlata con le presenze e gli arrivi, e quindi con capacità esplicativa delle due variabili di studio, è la dimensione della struttura, espressa dai posti letto. Si può ragionevolmente supporre, inoltre, che tale variabile sia assai costante nel tempo.

Il grafico 3.1 che segue mostra per l'intera popolazione degli esercizi ricettivi della provincia di Firenze nel mese di ottobre 2006, la relazione fra letti e presenze con la distinzione dell'appartenenza o meno al sistema turiweb,

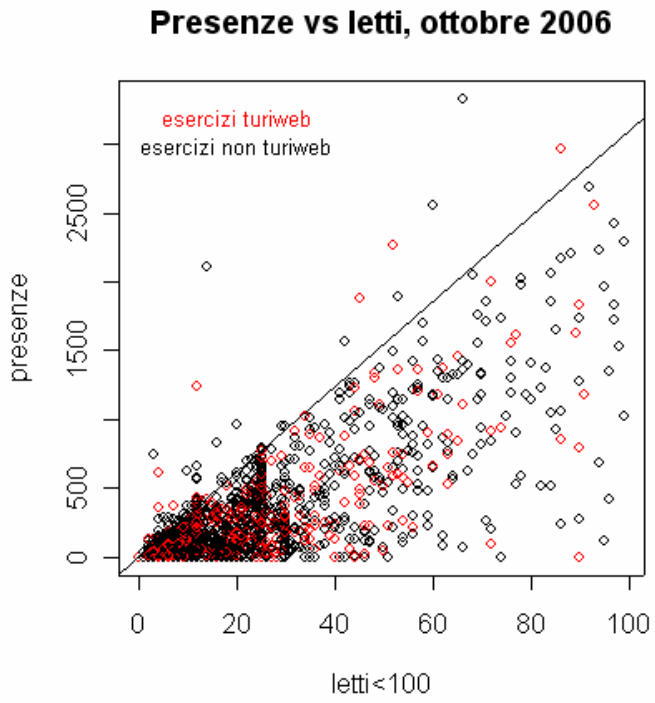
Grafico 3.1



La popolazione è molto asimmetrica, come accade di solito per le popolazioni di imprese: molte unità di piccole e poche di grandi dimensioni: 5, 122, 180 esercizi risultano disporre di 1, 2, 3 posti letto rispettivamente. Nel grafico successivo effettuiamo uno "zoom" nel rettangolo: 0 < letti<100, 0 < presenze < 3000. Disegniamo inoltre la retta per l'origine: Presenze = 31·Letti. Questa retta

rappresenta il limite superiore di presenze possibili per un esercizio in base al numero di letti di cui dispone

Grafico 3.2



I punti sopra la retta segnalano situazioni in cui le presenze registrate nel mese sono superiori a quelle che la dimensione permetterebbe. In alcuni casi questo è giustificato dalla possibilità di aggiungere occasionalmente un certo numero di letti, in qualche caso siamo in presenza di evidenti errori o delle presenze o dei posti letto. Vi sono poi unità con presenze nulle a fronte di dimensione anche rilevante: questo è dovuto al fatto che alcuni esercizi non hanno avuto effettivamente movimento oppure sono stati chiusi durante tutto il mese. Abbiamo considerato anche gli esercizi chiusi nella nostra analisi perché questa è un'informazione nota quando si conclude l'indagine censuaria.

I grafici suggeriscono un orientamento, per la stima del totale delle presenze e degli arrivi verso modelli per la popolazione lineari omogenei con eteroschedasticità, quali sono quelli che conducono allo stimatore rapporto.

Nelle nostre prime sperimentazioni, sulla base dell'esplorazione della relazione fra presenze/arrivi e posti letto, abbiamo considerato tre modelli semplici per la y (presenze/arrivi) da cui derivare gli stimatori:

mod 1 $\begin{cases} E_\xi(y_k) = \beta_g \\ V_\xi(y_k) = \sigma_g^2 \end{cases}$ y_k indipendenti (3.1)

mod 2 $\begin{cases} E_\xi(y_k) = \beta_g x_k \\ V_\xi(y_k) = \sigma_g^2 x_k \end{cases}$ y_k indipendenti (3.2)

mod 3 $\begin{cases} E_\xi(y_k) = \beta x_k \\ V_\xi(y_k) = \sigma^2 x_k \end{cases}$ y_k indipendenti (3.3)

mod 4 $\begin{cases} E_\xi(y_k) = \alpha_g + \beta_g x_k \\ V_\xi(y_k) = \sigma_g^2 \end{cases}$ y_k indipendenti (3.4)

dove: g = 1, . . ., G indica il gruppo (o post-strato) definito dalla combinazione di alcune variabili categoriche di cui diremo più avanti; x è la variabile ausiliaria quantitativa posti letto.

Dai modelli di cui sopra, derivano, nell'approccio basato sul modello, univoche espressioni algebriche per gli stimatori: stimatore post-stratificato, stimatore rapporto separato, stimatore rapporto, stimatore di regressione separata. Nell'approccio basato sul disegno, invece, la forma degli stimatori dipende anche dal disegno. Coincide con i tre sopra indicati se il disegno è SI, altrimenti no. Ad esempio dal modello (3.3) se il disegno è STSI si ottiene il cosiddetto rapporto combinato.

Ora, nel nostro caso non disponiamo di un disegno noto col quale specificare lo stimatore a partire dalla individuazione di un modello descrittivo della popolazione. In "qualche modo" dobbiamo usare (o provare) gli stimatori ben definiti all'interno di precise condizioni in una situazione dove queste condizioni non sono soddisfatte.

In sostanza cerchiamo stimatori che utilizzino in tutto o in parte le informazioni ausiliarie: la dimensione dell'esercizio ricettivo, la sua tipologia, la sua localizzazione turistica, il suo livello di qualità. Nel modello (3.1) non si considera la dimensione, che invece è presente nel (3.2) e (3.3). Tale modello, e lo stimatore che ne deriva, ci serve soprattutto come riferimento.

Per quanto riguarda le variabili qualitative di raggruppamento (post-stratificazione), abbiamo considerato due criteri di classificazione delle unità della popolazione in G gruppi.

- Il primo criterio considera le due variabili: tipologia di risorsa turistica, tipologia di struttura ricettiva. I G = 4 gruppi sono ottenuti incrociando le due modalità territoriali aggregate di risorsa turistica (Arte, Altra risorsa) con le due modalità aggregate di tipologia ricettiva (Alberghi, Altre strutture).
- Per il secondo criterio abbiamo prima definito un indice di qualità dato dal rapporto bagni/camere; quindi assegnato il livello di qualità Basso alle unità con indice inferiore a 1, Medio se pari a 1, Alto se maggiore di 1. Abbiamo poi incrociato le tre modalità del livello di qualità con le due modalità di risorsa turistica (Arte, Altra risorsa). In questo caso G = 6.

La prima post-stratificazione è piuttosto convenzionale, mirante soprattutto a un eventuale sviluppo dei metodi di stima per domini, la seconda cerca di inserire una variabile (la qualità) che potrebbe avere una maggior capacità esplicativa soprattutto del meccanismo di selezione. In entrambi i casi abbiamo cercato di limitare il numero dei gruppi per evitare che, specialmente nei primi mesi di cui disponiamo dei dati, alcuni gruppi fossero costituiti da un numero troppo piccolo di unità: questo potrebbe rendere molto instabili gli stimatori post-stratificato, rapporto separato e regressione separata.

Abbiamo considerato quattro stimatori.

- stimatore post-stratificato
- stimatore rapporto combinato
- stimatore rapporto separato

- stimatore di regressione separata

La variabile ausiliaria quantitativa per gli stimatori rapporto e regressione è la dimensione della struttura ricettiva (posti letto, spesso indicata semplicemente con letti).

Indichiamo con U l'intera popolazione, con g il generico gruppo, con U_g la popolazione del gruppo g, con V_g i volontari del gruppo g, con N_g la dimensione di U_g, con n_g la dimensione di V_g, con y_k il generico valore della variabile di studio, con x_k il generico valore della variabile ausiliaria posti letto. I primi tre stimatori del totale (arrivi/ presenze) assumono rispettivamente la forma

$$\hat{t}_{ps} = \sum_{g=1}^{G} \frac{N_g}{n_g} \sum_{Vg} y_k \qquad \hat{t}_{rc} = \left(\sum_{U} x_k \right) \frac{\sum_{g=1}^{G} \frac{N_g}{n_g} \sum_{Vg} y_k}{\sum_{g=1}^{G} \frac{N_g}{n_g} \sum_{Vg} x_k} \qquad \hat{t}_{rs} = \sum_{g=1}^{G} \left(\sum_{Ug} x_k \frac{\sum_{Vg} y_k}{\sum_{Vg} x_k} \right) \tag{3.5}$$

Särndal e Lundström (2005), sia sulla base di considerazioni teoriche, sia con una simulazione Monte Carlo, evidenziano il miglior comportamento, in presenza di nonrisposta, dello stimatore di regressione rispetto allo stimatore rapporto. Abbiamo quindi considerato anche lo stimatore di regressione separata del totale:

$$\hat{t}_{reg,sep} = \sum_{g=1}^{G} N_g \left[\frac{1}{n_g} \sum_{Vg} y_k + \left(\frac{1}{N_g} \sum_{Ug} x_k - \frac{1}{n_g} \sum_{Vg} x_k \right) \hat{B}_g \right] \tag{3.6}$$

dove $\hat{B}_g$ è il coefficiente di regressione calcolato in ciascun gruppo (post-strato) sulle osservazione delle unità rispondenti (gli esercizi del sistema turiweb nel nostro caso).

Lo stimatore (3.6), in presenza di non risposta, dovrebbe funzionare meglio degli stimatori rapporto, anche se il modello descrittivo della popolazione è quello lineare omogeneo eteroschedastico.

Per gli stimatori post-stratificato e rapporto combinato abbiamo usato solo la prima post-stratificazione, per il rapporto separato e regressione separata abbiamo invece effettuato le stime utilizzando entrambi i criteri di post-stratificazione. Non confidiamo molto nella bontà dei primi due stimatori. Abbiamo effettuato le stime con questi soprattutto per avere termini di paragone.

3.3. Misura dell'errore di stima

Disponendo dei dati censuari possiamo confrontato le stime ottenute con i diversi stimatori per i 24 mesi. Definiamo per nostra comodità l'errore relativo come il rapporto fra il totale stimato e il totale noto:

$$\varepsilon_R = \frac{\hat{t}}{t} \tag{3.7}$$

In effetti un errore relativo è convenzionalmente definito come:

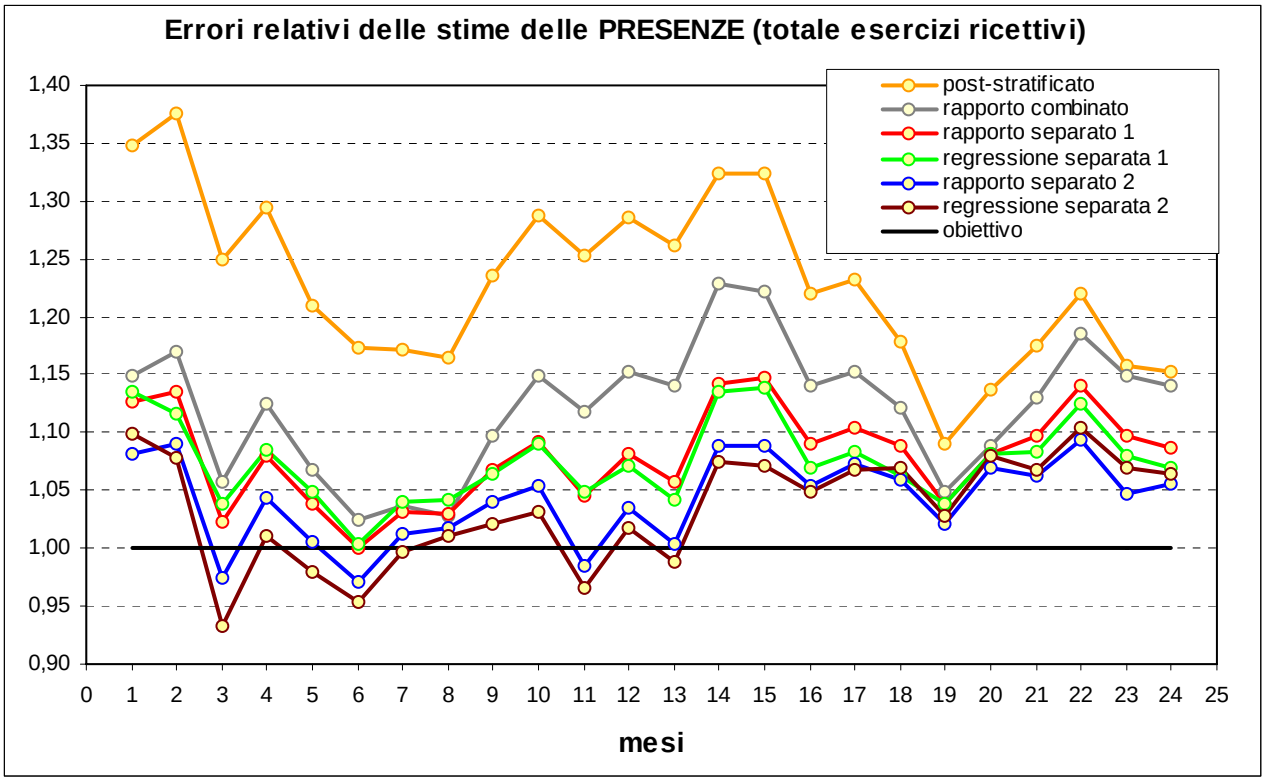
$$\nu_R = \frac{\hat{t} - t}{t} = \varepsilon_R - 1 \tag{3.8}$$

La differenza è nella costante 1. La scelta di lavorare con un errore relativo (sia esso espresso dalla (3.7) o dalla (3.8) è connessa soprattutto al tipo d'informazione che le stime dei totali devono fornire. L'interesse dell'utente non è tanto quello di conoscere il totale del movimento in un dato mese o periodo, quanto la variazione relativa del movimento rispetto al corrispondente mese o periodo dell'anno precedente. Espressa questa variazione come rapporto, la si può immediatamente confrontare con l'errore di stima relativo da noi adottato. Le stime avranno un buon valore informativo se il loro errore relativo sarà inferiore alle variazioni fisiologiche del fenomeno. Con i dati a nostra disposizione potremo confrontare le stime delle variazioni mensili con le variazioni mensili che si sono verificate nei mesi del 2006 rispetto al 2005.

3.4. Stime delle presenze e degli arrivi

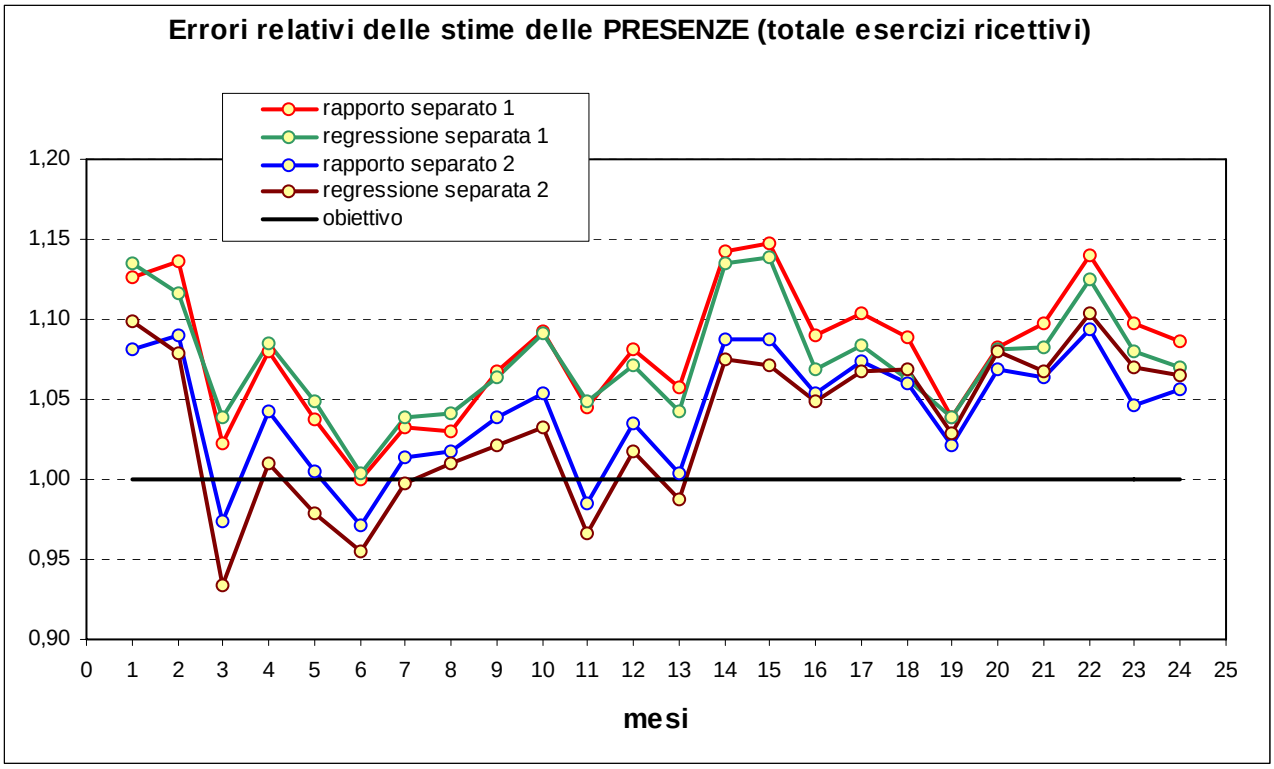
Il grafico seguente (Grafico 3.3) riporta gli errori relativi di stima, come definiti nella (3.7), per il totale delle presenze per l'intera popolazione, per i 24 mesi dal gennaio 2005 al dicembre 2006.

Grafico 3.3



Gli stimatori pos-stratificato e rapporto combinato utilizzano, come abbiamo detto, solo la prima post-stratificazione: i 4 gruppi ottenuti dall'incrocio delle due modalità di risorsa (Arte, Altra risorsa) con le due modalità di tipologia di esercizio (Alberghi, Altre strutture). Con rapporto separato 1 e regressione separata 1 abbiamo indicato le stime ottenute con questa prima post-stratificazione. Con rapporto separato 2 e regressione separata 2 le stime ottenute con la seconda post-stratificazione: i 6 gruppi ottenuti dall'incrocio dei tre livelli di qualità con le due modalità di risorsa. Il Grafico3.4 che segue cerca di evidenziare con una maggior risoluzione il comportamento dei quattro stimatori che forniscono risultati migliori.

Grafico 3.4



Il comportamento dei sei stimatori può essere quantitativamente confrontato usando alcune sintesi delle distribuzioni delle stime nei 24 mesi, riportate nella tabella seguente.

Tabella 3.1

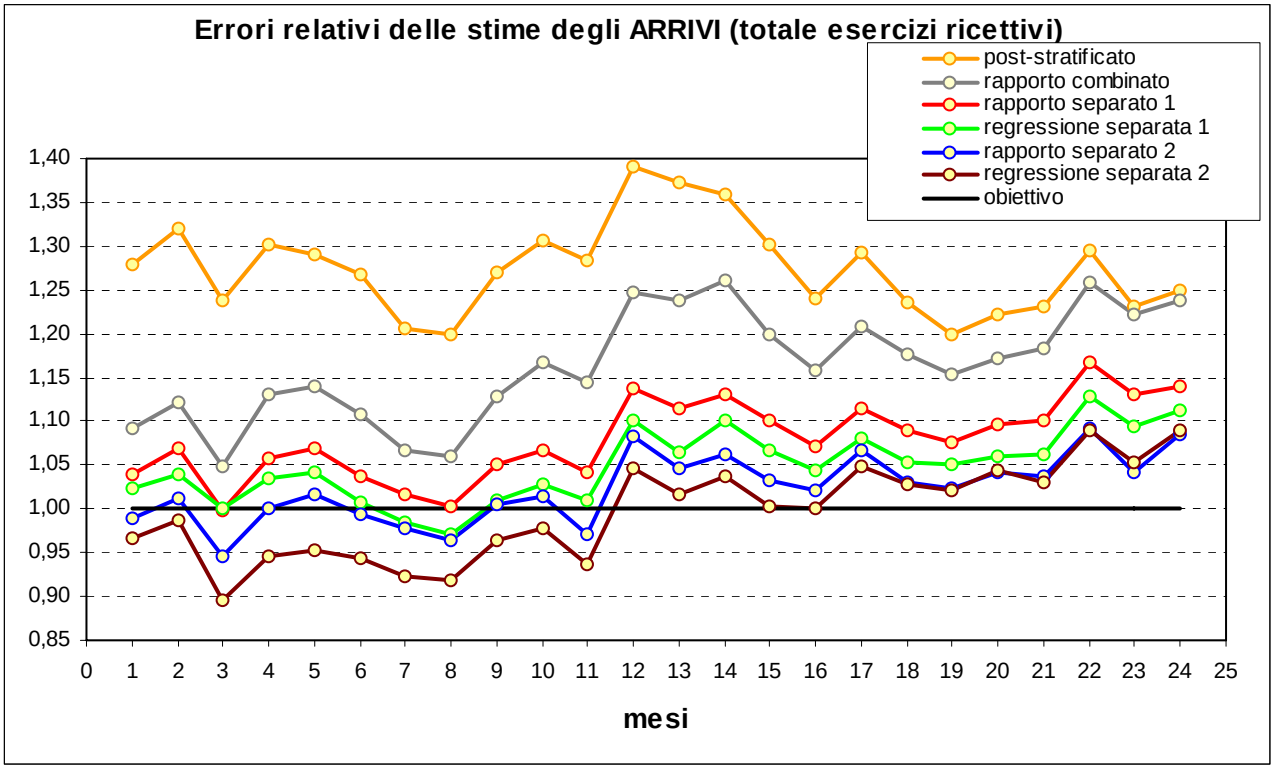
SINTESI DELLA DISTRIBUZIONE DEGLI ERRORI RELATIVI DELLE STIME (TOTALE PRESENZE)

	STIMATORE					
SINTESI	post stratificato	rapporto combinato	rapporto separato 1	regressione separata 1	rapporto separato 2	regressione separata 2
minimo	1,0910	1,0250	1,0000	1,0039	0,9712	0,9337
massimo	1,3760	1,2280	1,1480	1,1391	1,0942	1,1033
range	0,2850	0,2030	0,1480	0,1351	0,1230	0,1696
media	1,2300	1,1217	1,0801	1,0747	1,0428	1,0348
bias	0,2300	0,1217	0,0801	0,0747	0,0428	0,0348
varianza	0,0051	0,0030	0,0016	0,0012	0,0013	0,0021
mse	0,0580	0,0178	0,0080	0,0068	0,0031	0,0033

Gli stimatori tendono a sovrastimare il totale delle presenze, però questa sovrastima (il bias) decresce passando dal primo all'ultimo dei sei, come pure l'mse, pressoché uguale negli ultimi due. Gli stimatori che utilizzano nella post-stratificazione la qualità delle strutture presentano una maggior capacità di riduzione del bias, pressochè assente nel primo anno (2005), però sembrano più sensibili alla evoluzione del sistema turiweb. La varianza dei 24 errori relativi di stima è minima per la regressione separata 1 e il rapporto separato 2.

Una comportamento analogo, ma con differenze più costanti nel tempo, si verifica per il totale degli arrivi. Va notato che presenze e arrivi sono due variabili molto correlate. Il grafico seguente riporta gli errori relativi di stima degli arrivi per l'intera popolazione, per i mesi dal gennaio 2005 al dicembre 2006, ottenuti con i sei stimatori.

Grafico 3.5



Nella tabella seguente sono riportate le sintesi delle distribuzioni degli errori relativi di stima del totale degli arrivi per i 24 mesi.

Tabella 3.2

SINTESI DELLA DISTRIBUZIONE DEGLI ERRORI RELATIVI DELLE STIME (TOTALE ARRIVI)

	STIMATORE					
SINTESI	post stratificato	rapporto combinato	rapporto separato 1	regressione separata 1	rapporto separato 2	regressione separata 2
minimo	1,1990	1,0490	0,9980	0,9718	0,9462	0,8953
massimo	1,3910	1,2610	1,1670	1,1275	1,0910	1,0906
range	0,1920	0,2120	0,1690	0,1557	0,1448	0,1953
media	1,2743	1,1635	1,0800	1,0490	1,0230	0,9965
bias	0,2743	0,1635	0,0800	0,0490	0,0230	-0,0035
varianza	0,0026	0,0038	0,0020	0,0016	0,0014	0,0028
mse	0,0779	0,0305	0,0083	0,0040	0,0019	0,0029

L'andamento nei 24 mesi degli errori di stima è simile a quello visto per le presenze, con una maggior evidenza della gerarchia degli stimatori nella riduzione del bias. La regressione separata 2 presenta un bias nei 24 mesi pressoché nullo.
I quattro stimatori: rapporto e regressione separata con le due post-stratificazioni considerate sono buoni candidati per la stima delle presenze e degli arrivi mensili. Hanno tutti e quattro una varianza molto piccola.

Disponendo dell'indagine censuaria, vorremo applicare metodi di riduzione, se non proprio di eliminazione del bias, applicando la correzione di Rao di cui diremo fra breve. In vista della applicazione di questo strumento, la bontà degli stimatori va misurata prevalentemente con la varianza degli errori di stima osservati nei 24 mesi.

3.5. Il motivo della distorsione positiva

Gli stimatori rapporto e regressione separati con post-stratificazione 2, che tiene conto della qualità degli esercizi ricettivi, presentano per il primo anno un bias pressoché nullo e addirittura negativo nel caso degli arrivi. In tutti gli altri casi gli stimatori tendono a sovrastimare il totale delle presenze e degli arrivi. Il motivo di questa sovrastima, risiede nel fatto che, condizionatamente alle variabili esplicative, non vi è indipendenza fra il meccanismo di selezione e la variabile di studio: a parità di post-strato e posti letto gli esercizi del sistema turiweb hanno una media delle presenze (e degli arrivi) più elevata. Questa differenza potrebbe essere più accentuata nei mesi di bassa stagione quando il vantaggio competitivo ha maggior effetto. Nei periodi di alta stagione gli esercizi tendono a riempirsi comunque.

I due grafici seguenti evidenziano questo fatto nel mese di ottobre 2006: nei due grafici sono riportate le rette passanti per l'origine con pendenza data dal rapporto fra la media della variabile di studio e la media dei posti letto distintamente per gli esercizi appartenenti al sistema turiweb (web) e non appartenenti (no web), nei due post-strati: Alberghi-località d'Arte, Alberghi-Altre località.

Ora nello stimatore rapporto separato con post-stratificazione del tipo 1:

$$\hat{t}_{rs} = \sum_{g=1}^{G} \left(\sum_{Ug} x_k \frac{\sum_{Vg} y_k}{\sum_{Vg} x_k} \right) \qquad (3.9)$$

si stima la pendenza della retta per l'origine proprio col rapporto fra media delle presenze e media dei posti letto osservate sugli esercizi del sistema turiweb. Per questi esercizi la pendenza è più elevata, indicando una loro maggior efficienza nell'utilizzo dei posti letto disponibili; da qui la sovrastima dello stimatore rapporto separato.

Grafico 3.6

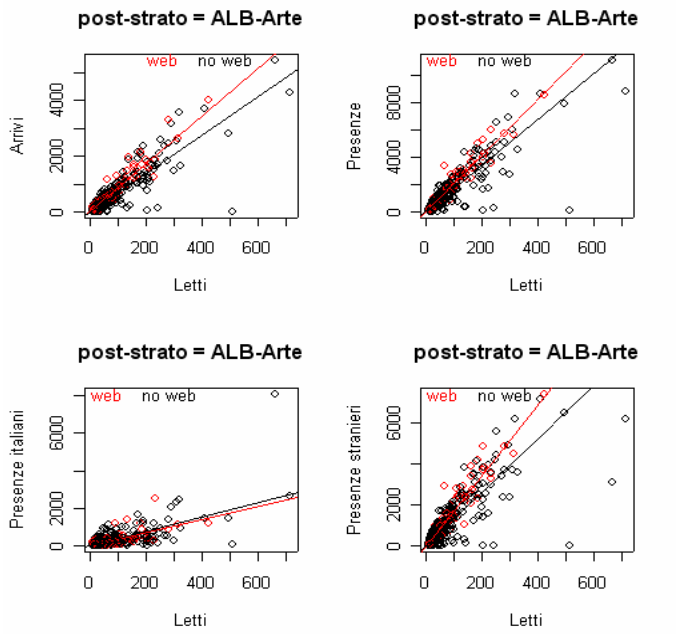
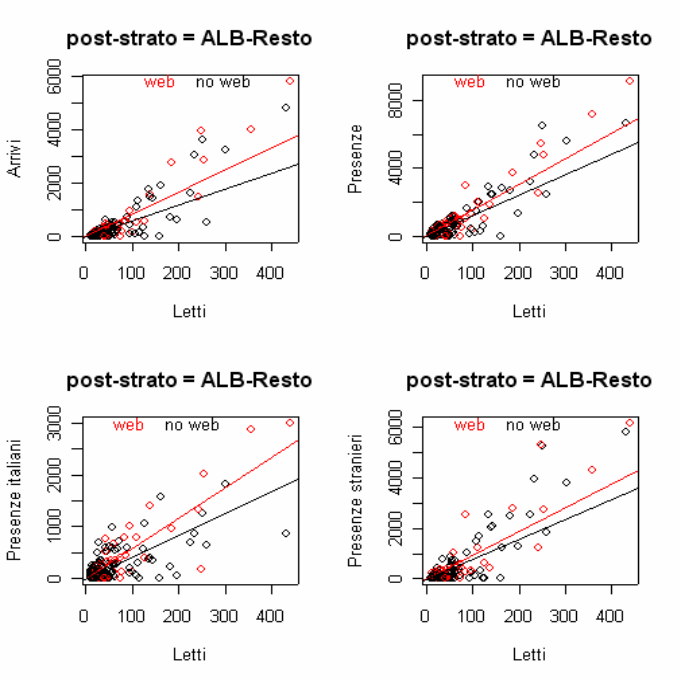


Grafico 3.7



Nel caso invece degli stimatori rapporto e regressione separati con post-stratificazione del tipo 2, la variabile qualità dell'esercizio ha una maggior capacità esplicativa del meccanismo di autoselezione e il bias viene ridotto e in alcuni casi pressoché eliminato. Si conferma infine l'affermazione di Särndal e Lundström (2005) che lo stimatore di regressione offre una buona possibilità di riduzione del bias in presenza di autoselezione, anche se la struttura dei dati consiglierebbe lo stimatore rapporto.

3.6. L'aggiustamento di Rao

Nelle indagini ripetute, soprattutto nel caso in cui i dati si ottengono con molto ritardo, vi è un forte interesse a fornire stime anticipate: queste vengono ottenute, di norma, sulla base del sottoinsieme di rispondenti entro una data stabilita precedente a quella programmata per la conclusione dell'indagine. Il meccanismo di selezione del sottoinsieme dei rispondenti tempestivi può essere stabilito da chi progetta l'indagine oppure no.

Nel primo caso si seleziona tale sottoinsieme secondo un disegno prestabilito; la tempestività è ottenuta con strumenti organizzativi che curano la raccolta dei dati dalle unità del campione preliminare. In questa situazione, sia nel caso di indagine periodica completa, sia campionaria, le stime anticipate sono ottenute con i metodi standard consolidati in assenza di specifiche fonti di errore connesse al procedimento anticipatorio. Tuttavia nella produzione di stime anticipate potrebbero non potersi attivare tutti gli accorgimenti per la riduzione della distorsione messi in atto nella produzione delle stime definite.

Quando invece il meccanismo di risposta entro un dato tempo è affidato alla libera iniziativa delle unità statistiche, non può essere completamente noto ed è possibile che tale meccanismo non

sia indipendente dalla variabile di studio e quindi introduca una distorsione, anche rilevante, nelle stime.

Nel nostro caso i rispondenti tempestivi sono coloro che partecipano al sistema turiweb e che quindi usano per la trasmissione dei dati uno specifico strumento. L'accesso allo strumento e il suo utilizzo possono dar luogo a una selezione non ignorabile. Abbiamo visto, esaminando gli errori di stima, che le strutture ricettive del sistema turiweb presentano una distribuzione delle presenze, condizionata alla localizzazione, tipologia e dimensione della struttura, diversa rispetto alle strutture non appartenenti al sistema; da qui la presenza di un bias positivo nelle stime.

Rao e altri (1989) hanno proposto due metodi per ridurre il possibile bias delle stime anticipate derivante da varie fonti in un'ottica che non entra nelle specifiche cause della distorsione ma cerca di ridurla sulla base dell'errore di stima che può essere osservato in virtù della presenza di una indagine ripetuta nel tempo. Uno è l'approccio di campionamento, l'altro quello di serie storica.

Descriveremo l'approccio di campionamento essendo quello da noi utilizzato, dopo una verifica svolta presso l'Istat in cui è stata evidenziata la leggera superiorità di questo approccio rispetto a quello di serie storica che utilizza il filtro di Kalman.

3.6.1. L'approccio di campionamento

Illustriamo l'approccio di campionamento con riferimento alla stima del totale delle presenze. Definiamo stima preliminare (o grezza) quella ottenuta sulla base delle sole osservazioni sul sottoinsieme dei rispondenti tempestivi (gli esercizi del sistema turiweb nel nostro caso) applicando uno degli stimatori visti nei paragrafi precedenti. Chiamiamo stima anticipata quella ottenuta con l'applicazione dell'aggiustamento dove entra in gioco l'informazione ottenuta dalle stime finali nelle precedenti occasioni di indagine. Indichiamo pertanto:

$\hat{P}_t^P$ la stima preliminare (grezza) al tempo t,

$\hat{P}_{t-k}^P$ la stima preliminare (grezza) al tempo t-k,

P_{t-k} la stima definitiva al tempo t-k o, nel nostro caso, il totale ottenuto dall'indagine completa al tempo t-k.

k = 1, 2, …, t-1 è l'indice dei tempi, nel nostro caso i mesi.

Si definisce uno stimatore anticipato del totale lo stimatore composto, media ponderata dello stimatore preliminare al tempo t e dello stimatore definitivo (o totale noto) al tempo t-k, incrementato della variazione ottenuta con gli stimatori preliminari $\Delta_0 = \hat{P}_t^P - \hat{P}_{t-k}^P$

$$\hat{P}_{t,\alpha} = \alpha\hat{P}_t^P + (1-\alpha)\left[P_{t-k} + (\hat{P}_t^P - \hat{P}_{t-k}^P)\right] = \alpha\hat{P}_t^P + (1-\alpha)\left[P_{t-k} + \Delta_0\right] = \hat{P}_t^P + (1-\alpha)(P_{t-k} - \hat{P}_{t-k}^P) \quad (3.10)$$

Per ricavare il valore di α si deriva il bias di $\hat{P}_{t,\alpha}$ sotto due "ragionevoli" assunzioni.

Assunzione 1: $$E(\hat{P}_t^P - \hat{P}_{t-k}^P) = E(P_t - P_{t-k}) \quad (3.11)$$

Assunzione 2: $$\left|B(\hat{P}_j^P)\right| > \left|B(P_j)\right|, \quad j = t, t-k \quad (3.12)$$

Si assume inoltre, per semplicità, che $B(P_t) = B(P_{t-k}) = 0$, da cui: $B(\hat{P}_t^P) = B(\hat{P}_{t-k}^P) = \delta$.

Sotto queste assunzioni si ottiene:

$$\left|B(\hat{P}_{t,\alpha})\right| = \left|B\left[\hat{P}_t^P + (1-\alpha)(P_{t-k} - \hat{P}_{t-k}^P)\right]\right| = \alpha\left|\delta\right|. \quad (3.13)$$

E quindi il bias è annullato per $\alpha = 0$. Si ottiene allora lo stimatore aggiustato.

$$\hat{P}_{t,0} = \hat{P}_t^P + (P_{t-k} - \hat{P}_{t-k}^p) = P_{t-k} + (\hat{P}_t^P - \hat{P}_{t-k}^P) \qquad (3.14)$$

In pratica si ottiene la stima anticipata al tempo t a partire dalla stima preliminare per il tempo t "corretta" con l'errore di stima che si è commesso al tempo t-k; oppure la stima anticipata si può vedere come la stima finale ottenuta al tempo t-k a cui si aggiunge l'incremento fra t-k e t ottenuto con le stime preliminari.

3.6.2. Applicazione dell'aggiustamento di Rao alle stime preliminari del totale delle presenze

Nel nostro caso, l'indagine periodica mensile è completa; le stime anticipate hanno l'obiettivo di anticipare il risultato che si otterrà al momento della conclusione dell'indagine censuaria. Non è detto che tale risultato fornisca il valore esatto del totale delle presenze (arrivi) nella popolazione degli esercizi attivi del mese t. Anche l'indagine censuaria può essere affetta da errori, tipicamente di sottocopertura e misura. Tuttavia, i risultati dell'indagine censuaria sono assunti come "ufficiali" dagli utenti e la bontà delle stime anticipate è misurata dallo scarto commesso rispetto a questi risultati, ovvero dall'errore di stima che possiamo osservare fra la stima anticipata e il risultato dell'indagine completa k mesi più tardi. Come abbiamo visto nel paragrafo 3.4, la stima preliminare è affetta da un errore complessivo dato da una quota dovuta alla distorsione e da una dovuta a fattori accidentali. La quota dovuta alla distorsione può essere ridotta se non proprio eliminata con l'aggiustamento, quella dovuta a fattori accidentali rimane.

Assumiamo che P_t, P_{t-k} siano parametri deterministici (i totali della popolazione ottenuti con l'indagine censuaria) per cui le assunzioni 1 e 2 assumono la forma:

Assunzione 1b $\quad E(\hat{P}_t^P - \hat{P}_{t-k}^P) = P_t - P_{t-k} \qquad (3.15)$

Assunzione 2b $\quad \left|B(\hat{P}_j^P)\right| > 0, \quad j = t,\ t-k \qquad (3.16)$

Di conseguenza, vale che: $B(P_t) = B(P_{t-k}) = 0$, e quindi a maggior ragione $B(\hat{P}_t^P) = B(\hat{P}_{t-k}^P) = \delta$. Il valore ottimale di α è ancora 0.

Abbiamo considerato e quindi intendiamo applicare, oltre all'aggiustamento additivo in (3.14) anche un aggiustamento "moltiplicativo":

$$\hat{P}_{t,0}^* = \hat{P}_t^P \cdot (P_{t-k} / \hat{P}_{t-k}^p) = P_{t-k} \cdot (\hat{P}_t^P / \hat{P}_{t-k}^P) \qquad (3.17)$$

che equivale formulare l'assunzione 1b nella forma: $E(\hat{P}_t^P / P_t) = E(\hat{P}_{t-k}^P / P_{t-k})$.

Le stime preliminari utilizzano l'informazione ausiliaria disponibile al tempo t fornita dalla tipologia di esercizio, tipologia di risorsa turistica, qualità (nel caso di post-stratificazione 2), dimensione dell'esercizio, con l'obiettivo di ridurre l'errore di stima derivante dall'autoselezione, secondo la logica presente nei due approcci: model assisted o model dependent. In particolare si cerca di ridurre, se non eliminare la eventuale quota distorsiva di tale errore. Da un "qualche" punto di vista l'aggiustamento di Rao fa entrare in gioco un'ulteriore informazione ausiliaria: le presenze

verificatesi nei mesi precedenti. Tale variabile può avere capacità esplicativa sia della variabile di studio, sia del meccanismo di appartenenza al sistema turiweb.

3.6.3. Analogia fra lo stimatore aggiustato nella versione additiva e moltiplicativa e gli stimatori differenza e rapporto

Le espressioni (3.14) e (3.17) sono formalmente analoghe agli stimatori del totale differenza e rapporto che si ottengono nell'approccio basato sul disegno e assistito dal modello

$$\begin{aligned} \hat{t}_{y,dif} &= \hat{t}_{y,\pi} + (t_x - \hat{t}_{x,\pi}) \\ \hat{t}_{y,rap} &= \hat{t}_{y,\pi} \frac{t_x}{\hat{t}_{x,\pi}} \end{aligned} \qquad (3.18)$$

ove si sostituisca agli stimatori disegno-corretti HT gli stimatori preliminari al tempo t e ai totali noti, i totali noti delle presenze al tempo t-k. Le proprietà statistiche di questi stimatori sono ricavate rispetto al disegno p(s): la distribuzione di probabilità sullo spazio campionario.

Nell'approccio basato sul modello gli stimatori differenza e rapporto del totale assumono la forma:

$$\begin{aligned} \hat{T}_{dif} &= \sum_S y_k + \left(\sum_U x_k - \sum_S x_k \right) \\ \hat{T}_{rap} &= \sum_S y_k \cdot \frac{\sum_U x_k}{\sum_S x_k} \end{aligned} \qquad (3.19)$$

In questo approccio le proprietà statistiche degli stimatori derivano dalle ipotesi che specificano il modello per i dati sottostante. Se tale modello specifica esattamente la relazione fra x e y i due stimatori sono corretti, anzi BLU.

Formalmente la differenza fra gli stimatori model assisted e model dependent è nella presenza o meno dei pesi (inverso delle probabilità di inclusione).

3.6.4. Aggiustamento di Rao e stima della variazione mensile delle presenze rispetto al corrispondente mese dell'anno precedente

L'interesse per le stime del totale mensile è limitato. Quello che più interessa è la stima della variazione relativa mensile rispetto al mese corrispondente dell'anno precedente.

Come abbiamo fatto per l'errore relativo di stima, definiamo (per nostra comodità) la variazione relativa mensile come rapporto fra il totale del mese t e quello del mese t-12 (in effetti si tratta del numero indice):

$$I_{t,\, t-12} = \frac{P_t}{P_{t-12}}. \qquad (3.20)$$

La stimiamo con:

$$\hat{I}_{t,\,t-12,\,k} = \frac{\hat{P}_{t,\,k}}{P_{t-12}}, \qquad (3.21)$$

dove: $\hat{P}_{t,\,k} = \hat{P}_t^{P} + (P_{t-k} - \hat{P}_{t-k}^{P})$ se applichiamo la correzione additiva alle stime preliminari

$\hat{P}_{t,\,k} = \hat{P}_t^{P} \dfrac{P_{t-k}}{\hat{P}_{t-k}^{P}}$ se applichiamo la correzione moltiplicativa.

Come abbiamo visto le due correzioni additiva e moltiplicativa fanno intervenire l'ulteriore informazione ausiliaria fornita dalle presenze nel mese t-k in un modo analogo a quello che si ha con lo stimatore differenza e rapporto.

La correzione di Rao nella versione (3.14) e (3.17) è giustificata dalla assunzione di un bias additivo/moltiplicativo mediamente costante. Se fosse esattamente così la variazione mensile potrebbe essere allora stimata utilizzando direttamente le stime preliminari.

La stima della variazione relativa mensile quando k = 12, correggendo le stime preliminari in modo additivo col bias osservato 12 mesi prima risulta:

$$\hat{I}_{t,\,t-12,\,12} = \frac{\hat{P}_{t,\,12}}{P_{t-12}} = \frac{\hat{P}_t^{P} + (P_{t-12} - \hat{P}_{t-12}^{P})}{P_{t-12}} = 1 + \frac{\hat{P}_t^{P} - \hat{P}_{t-12}^{P}}{P_{t-12}} \qquad (3.22)$$

Quindi applicare la correzione additiva con lag 12 implica ricavare la variazione assoluta fra i mesi corrispondenti dei due anni utilizzando le stime preliminari senza alcuna correzione. Con questa variazione assoluta si ricava poi la variazione relativa.

Nel caso di correzione moltiplicativa, si ha:

$$\hat{I}_{t,\,t-12,\,12} = \frac{\hat{P}_{t,\,12}}{P_{t-12}} = P_{t-12}\,\frac{\hat{P}_t^{P}}{\hat{P}_{t-12}^{P}}\,\frac{1}{P_{t-12}} = \frac{\hat{P}_t^{P}}{\hat{P}_{t-12}^{P}} \qquad (3.23)$$

Allora aggiustare le stime con la correzione moltiplicativa con lag 12 implica ricavare la variazione relativa mensile direttamente dalle stime preliminari senza alcun aggiustamento.

Si pone allora la questione, se sia meglio o no evitare la correzione di Rao e ricavare direttamente la variazione relativa mensile utilizzando direttamente le stime preliminari basate solo sull'osservazione dei volontari.

Nel capitolo successivo applicheremo la correzione di Rao alle stime grezze ottenute con i quattro stimatori:

rapporto separato	con post-stratificazione 1
regressione separata	con post-stratificazione 1
rapporto separato	con post-stratificazione 2
regressione separata	con post-stratificazione 2

Avremo così modo di verificare, per la nostra situazione, sia l'efficacia della correzione di Rao, sia la sua utilità nella stima della variazione relativa mensile.

Capitolo 4

4.1. Applicazione della correzione di Rao

I questo capitolo applicheremo la correzione di Rao alle stime preliminari del totale delle presenze dell'intera popolazione degli esercizi ricettivi ottenute con i quattro stimatori:

1. rapporto separato con post-stratificazione 1
2. regressione separata con post-stratificazione 1
3. rapporto separato con post-stratificazione 2
4. regressione separata con post-stratificazione 2

Si tratta degli stimatori che presentano la varianza (della distribuzione delle stime preliminari nei 24 mesi) più piccola dei sei considerati nel cap. 3. Per ciascuno stimatore applicheremo la correzione di Rao con k = 1, 2, . . ., 12, nella versione additiva e moltiplicativa, cercando cioè di correggere il bias attuale con quello osservato 1, 2, . . ., 12 mesi prima grazie alla presenza dell'indagine completa. La bontà della stima anticipata del totale (la stima preliminare aggiustata con la correzione di Rao) verrà misurata dalla precisione con cui essa fornisce la stima della variazione relativa mensile.

Definita la variazione relativa mensile come rapporto fra i totali:

$$I_i = \frac{P_i}{P_{i-12}} \quad i = 13,...,24 \tag{4.1}$$

la stimeremo con:

$$\hat{I}_{i,k} = \frac{\hat{P}_{i,k}}{P_i} \quad i = 13,...,24; \quad k = 1,...,12 \tag{4.2}$$

dove i indica il mese del 2006, $\hat{P}_{i,k}$ la stima anticipata del mese i corretta con il bias osservato k mesi mesi prima. Nel seguito, e nei grafici in particolare, chiamiamo questo valore di k "lag k".

Oltre a riportare in tabelle le sintesi delle distribuzione delle stime anticipate nei 24 mesi, per un apprezzamento visivo della loro bontà riporteremo un grafico con gli errori relativi di stima del totale mensile al lag 0 (stima preliminare senza correzione di Rao) e al lag 1. In una serie di grafici successivi effettuiamo il confronto fra le variazioni relative mensili effettive e quelle stimate con le stime anticipate ai lag 3, 5, 7, 12. Ci interessano soprattutto i lag 5, 7 perché corrispondono al numero di mesi di ritardo con cui l'indagine completa è in grado di fornire i risultati.

Nei grafici che riportano il confronto fra le variazioni relative mensili effettive e quelle stimate anticipatamente, è presente una linea di riferimento al valore 1,00: tale linea indica la situazione di variazione nulla:

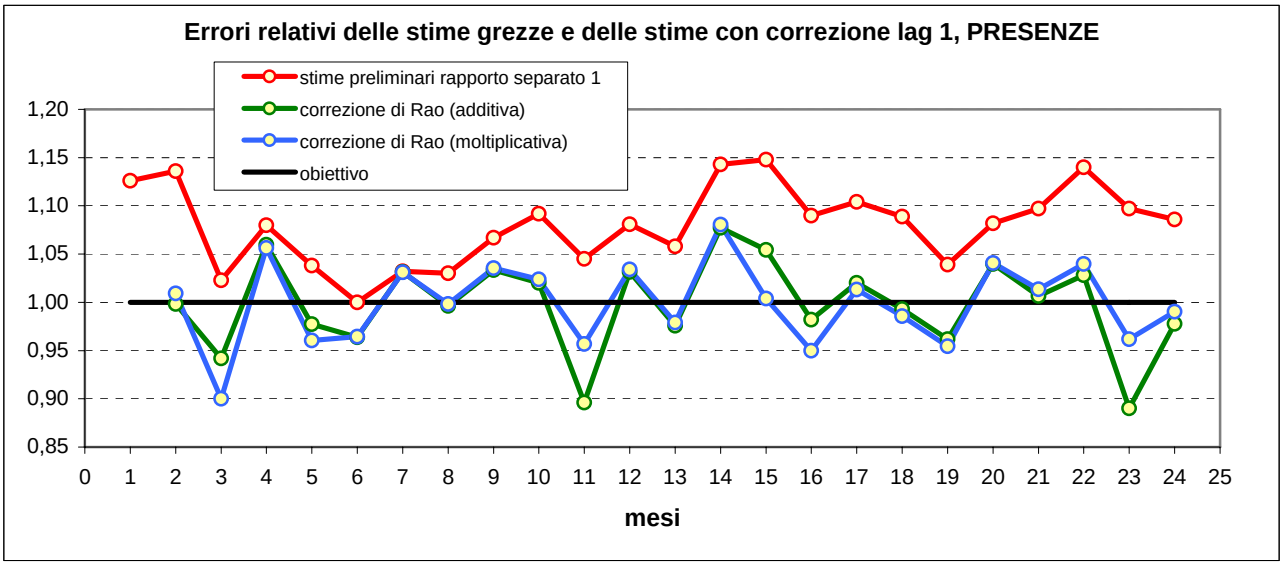
$$I_i = 1 <=> P_i = P_{i-12}$$

Tanto più la variazione effettiva è vicina a tale linea, tanto più è difficile con la stima anticipata concludere se I_i è stato maggiore o minore di 1.

4.2. Applicazione della correzione di Rao alle stime con rapporto separato 1

Il grafico seguente riporta gli errori relativi delle stime preliminari del totale mensile ottenute con il rapporto separato 1 e gli errori relativi delle stime anticipate del totale mensile ottenute con la correzione di Rao al lag 1 additiva e moltiplicativa.

Grafico 4.1



Nella successiva tabella 4.1 sono riportate le sintesi statistiche delle distribuzioni degli errori relativi delle stime anticipate: dal lag 0 (stime preliminari) fino al lag 12.

Tabella 4.1

PRESENZE TOTALI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

CORREZIONE ADDITIVA

LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	1,0000	0,8901	0,8543	0,8921	0,9134	0,9404	0,9107	0,8967	0,9331	0,9026	0,9495	0,9528	0,9515
masimo	1,1480	1,0773	1,1048	1,0869	1,1106	1,0629	1,0797	1,1000	1,1420	1,1470	1,0947	1,1125	1,1259
range	0,1480	0,1872	0,2505	0,1948	0,1972	0,1225	0,1690	0,2033	0,2089	0,2443	0,1452	0,1598	0,1744
media	1,0801	0,9981	0,9954	1,0033	1,0059	1,0119	1,0151	1,0193	1,0268	1,0287	1,0329	1,0416	1,0414
bias	0,0801	-0,0019	-0,0046	0,0033	0,0059	0,0119	0,0151	0,0193	0,0268	0,0287	0,0329	0,0416	0,0414
varianza	0,0016	0,0022	0,0032	0,0030	0,0026	0,0012	0,0022	0,0025	0,0029	0,0040	0,0023	0,0022	0,0018
mse	0,0080	0,0022	0,0032	0,0030	0,0026	0,0013	0,0024	0,0029	0,0036	0,0048	0,0033	0,0039	0,0035

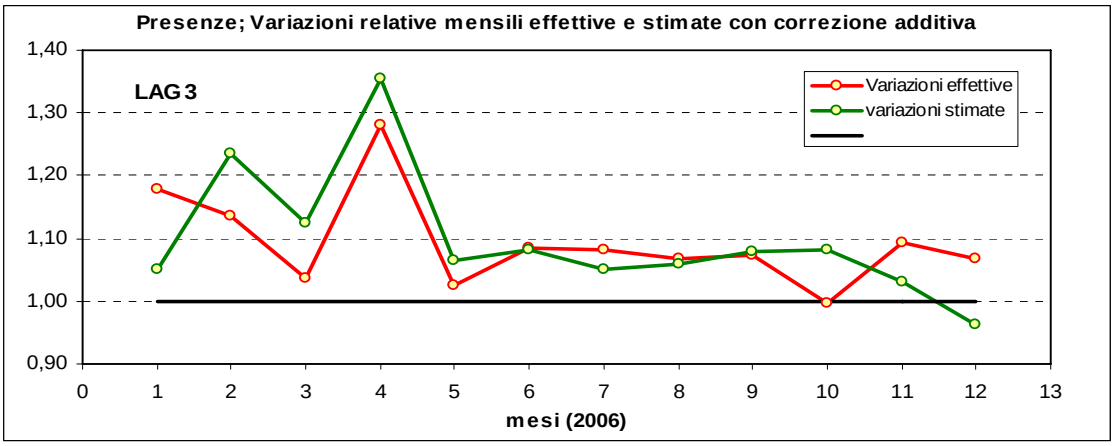
CORREZIONE MOLTIPLICATIVA

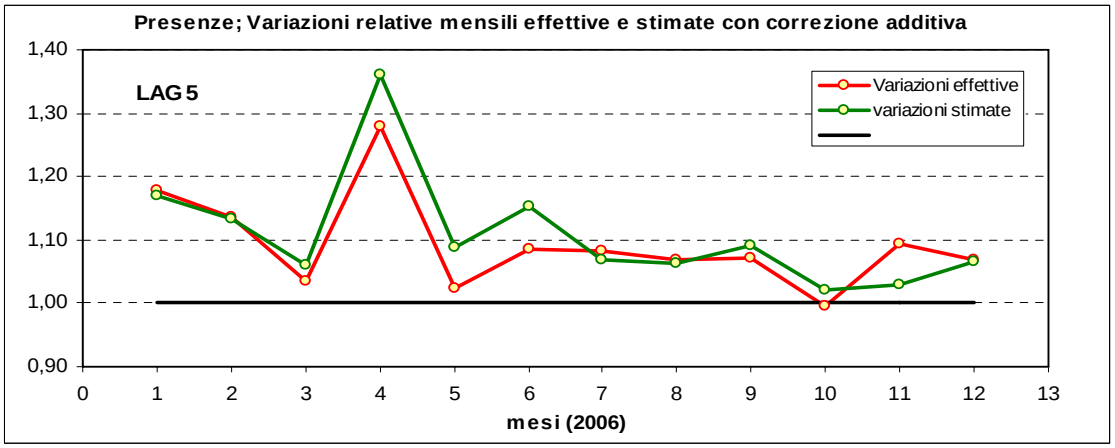
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	1,0000	0,9002	0,9085	0,9134	0,8808	0,8889	0,9068	0,9152	0,9477	0,9199	0,9284	0,9315	0,9400
massimo	1,1480	1,0805	1,0848	1,0972	1,0983	1,0717	1,1098	1,1142	1,1426	1,1471	1,1062	1,1179	1,1224
range	0,1480	0,1802	0,1762	0,1838	0,2175	0,1828	0,2030	0,1991	0,1949	0,2272	0,1778	0,1865	0,1824
media	1,0801	0,9993	0,9979	1,0030	1,0043	1,0065	1,0089	1,0125	1,0174	1,0207	1,0252	1,0340	1,0346
bias	0,0801	-0,0007	-0,0021	0,0030	0,0043	0,0065	0,0089	0,0125	0,0174	0,0207	0,0252	0,0340	0,0346
varianza	0,0016	0,0017	0,0022	0,0024	0,0032	0,0028	0,0029	0,0029	0,0026	0,0040	0,0029	0,0027	0,0020
mse	0,0080	0,0017	0,0022	0,0024	0,0032	0,0029	0,0030	0,0031	0,0029	0,0044	0,0035	0,0039	0,0032

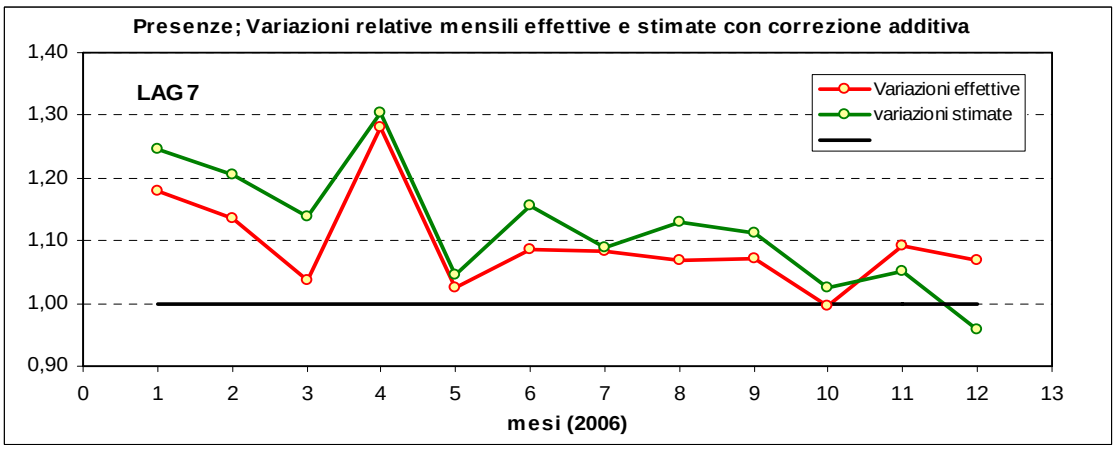
Si nota come il bias delle stime preliminari, dell'ordine dell'8% con questo stimatore, venga fortemente ridotto, sebbene tale riduzione si attenui all'aumentare del lag: pressoché annullato ai primi lag e dimezzato agli ultimi. La varianza sembra poco dipendente dal lag. In termini di mse, non si nota una maggior efficienza della correzione moltiplicativa rispetto all'additiva.

Nel grafico seguente si confrontano le variazioni relative effettive con quelle stimate dopo l'applicazione della correzione di Rao additiva e moltiplicativa ai lag 3, 5, 7, 12.

Grafico 4.2 (Rapporto separato 1 - Correzione additiva)







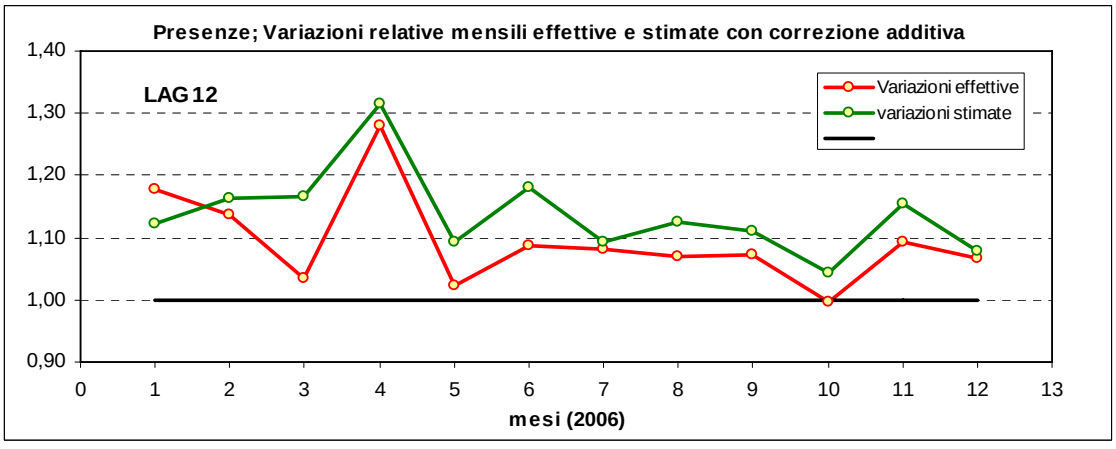
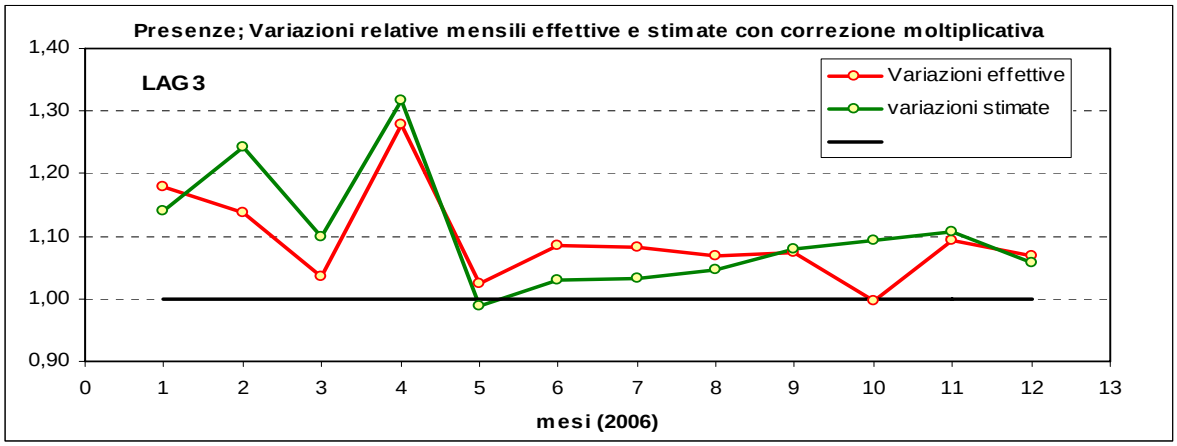
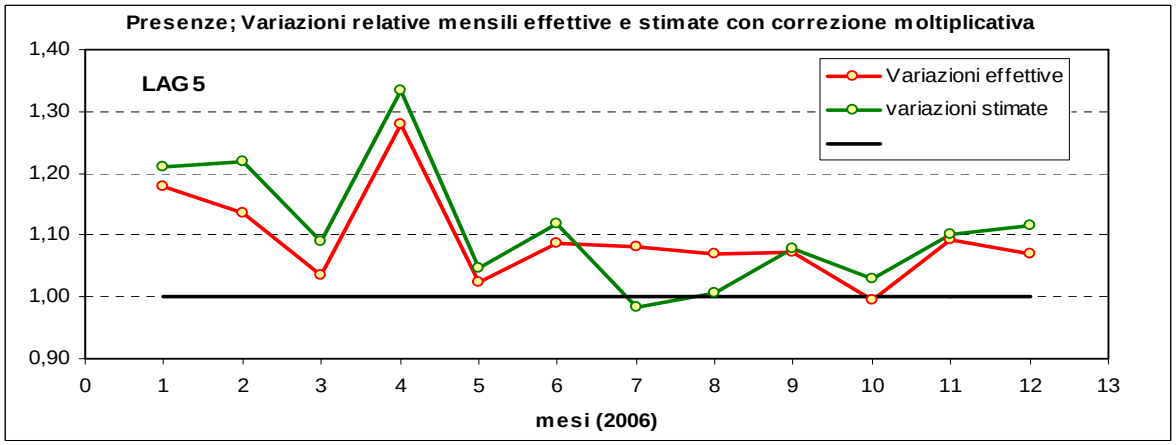
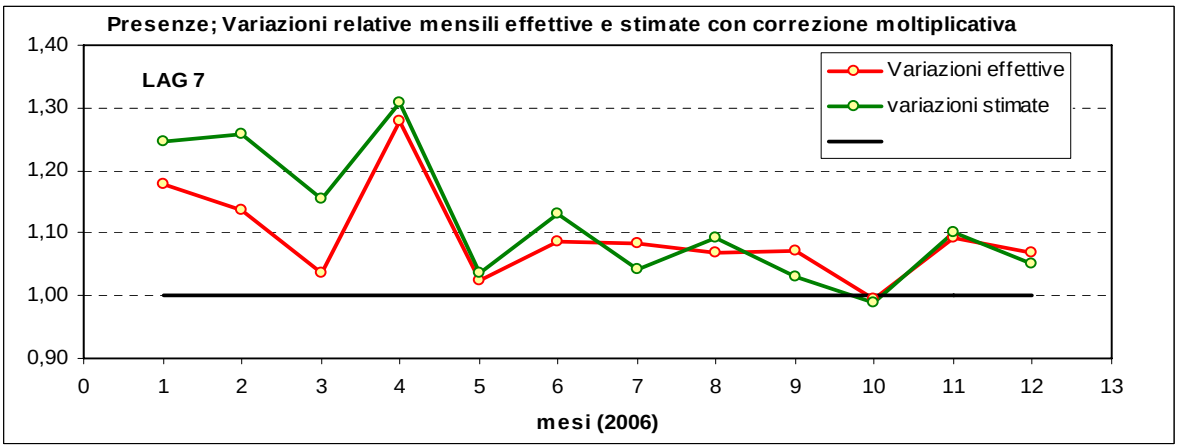
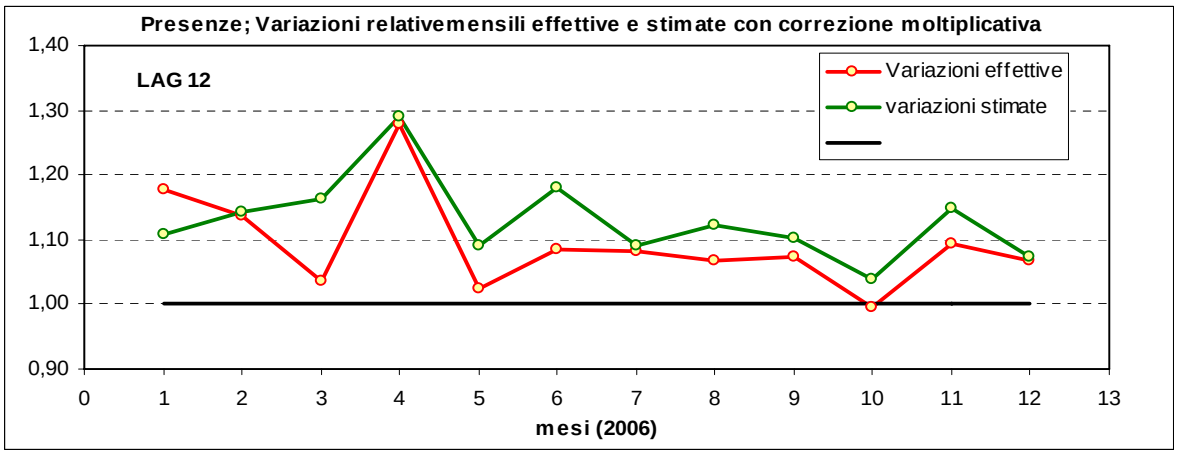


Grafico 4.3 (Rapporto separato 1 - Correzione moltiplicativa)









Le variazioni relative mensili stimate con entrambi i metodi di correzione, seguono abbastanza bene quelle effettive e solo in pochi casi si sarebbe ottenuta una conclusione errata circa il segno della variazione. Va però notato che le variazioni mensili effettive nell'anno 2006 rispetto al 2005 sono state sistematicamente positive e piuttosto elevate, mediamente del +8,4%.

Come abbiamo visto, al crescere del lag, sia nella correzione additiva sia in quella moltiplicativa, la capacità di ridurre il bias positivo delle stime decresce. Questo si ripercuote nelle stime delle variazioni relative mensili: aumentando il lag si tende a sovrastimare le variazioni, questo è più evidente al lag 12.

Applicare la correzione di Rao moltiplicativa al lag 12, per ottenere la stima anticipata del totale mensile e quindi calcolare la variazione relativa mensile, equivale a ricavare tale variazione direttamente dal rapporto fra le stime preliminari dei totali mensili. La tabella 4.1 e i grafici 4.2 e 4.3 evidenziano che invece conviene effettuare la correzione di Rao al lag più piccolo possibile, ottenere la stima anticipata del totale e con questa la stima anticipata della variazione relativa mensile.

Nella tabella 4.2 che segue si fornisce una misura della differenza fra variazioni relative mensili e variazioni stimate. L'ideale sarebbe avere una distribuzione delle differenze con media zero e varianza nulla.

Tabella 4.2

Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi 2006	Variazione effettiva	Differenze fra variazioni stimate ed effettive ai lag				
		lag 1	lag 3	lag 5	lag 7	lag 12
				Correzione additiva		
1	1,178	-0,028	-0,127	-0,008	0,067	-0,057
2	1,136	0,088	0,098	-0,004	0,069	0,027
3	1,036	0,056	0,090	0,026	0,104	0,130
4	1,280	-0,023	0,075	0,080	0,026	0,035
5	1,024	0,021	0,042	0,064	0,021	0,069
6	1,086	-0,008	-0,004	0,067	0,071	0,096
7	1,082	-0,041	-0,031	-0,013	0,006	0,011
8	1,069	0,042	-0,008	-0,005	0,061	0,057
9	1,073	0,007	0,006	0,018	0,039	0,037
10	0,995	0,028	0,086	0,026	0,029	0,048
11	1,093	-0,120	-0,063	-0,065	-0,043	0,061
12	1,068	-0,024	-0,106	-0,003	-0,110	0,011
Media	**1,0933**	**-0,0002**	**0,0047**	**0,0152**	**0,0282**	**0,0438**
Varianza		**0,0027**	**0,0054**	**0,0015**	**0,0031**	**0,0020**
				Correzione moltiplicativa		
1	1,178	-0,0247	-0,0366	0,0320	0,0677	-0,0707
2	1,136	0,0914	0,1068	0,0815	0,1227	0,0073
3	1,036	0,0041	0,0642	0,0529	0,1183	0,1267
4	1,280	-0,0644	0,0387	0,0551	0,0280	0,0117
5	1,024	0,0135	-0,0347	0,0224	0,0116	0,0660
6	1,086	-0,0154	-0,0558	0,0315	0,0455	0,0958
7	1,082	-0,0491	-0,0502	-0,0982	-0,0414	0,0080
8	1,069	0,0437	-0,0218	-0,0613	0,0241	0,0538
9	1,073	0,0147	0,0079	0,0067	-0,0436	0,0303
10	0,995	0,0397	0,0968	0,0325	-0,0063	0,0442
11	1,093	-0,0418	0,0151	0,0081	0,0069	0,0543
12	1,068	-0,0104	-0,0103	0,0481	-0,0176	0,0054
Media	**1,0933**	**0,0001**	**0,0100**	**0,0176**	**0,0263**	**0,0361**
Varianza		**0,0018**	**0,0029**	**0,0023**	**0,0027**	**0,0023**

Si evidenzia numericamente quello che già emergeva dai grafici: al crescere del lag di correzione si tende a sovrastimare le variazioni relative mensili. Ad esempio con la correzione moltiplicativa si sovrastima la variazione relativa mensile mediamente di circa l'1,0% al lag 3, di circa 2,6% al lag 7 . Non si evidenzia una performance migliore fra la correzione additiva e quella moltiplicativa.

Un'altra valutazione della bontà delle stime ottenute può ottenersi considerando un intervallo di confidenza di riferimento. Consideriamo una strategia basata su un disegno probabilistico stratificato con una allocazione uguale a quella data dalla distribuzione dei volontari nei quattro post-strati che abbiamo usato con lo stimatore rapporto separato. Per tale strategia, sulla base dei dati di popolazione noti, calcoliamo l'intervallo di confidenza (approssimato) al 95% per la stima del totale delle presenze. I due grafici che seguono confrontano gli errori relativi delle stime anticipate, ottenute con la correzione di Rao ai lag 1, 3, 5, 7 sia additiva che moltiplicativa con tale intervallo di confidenza.

Grafico 4.4

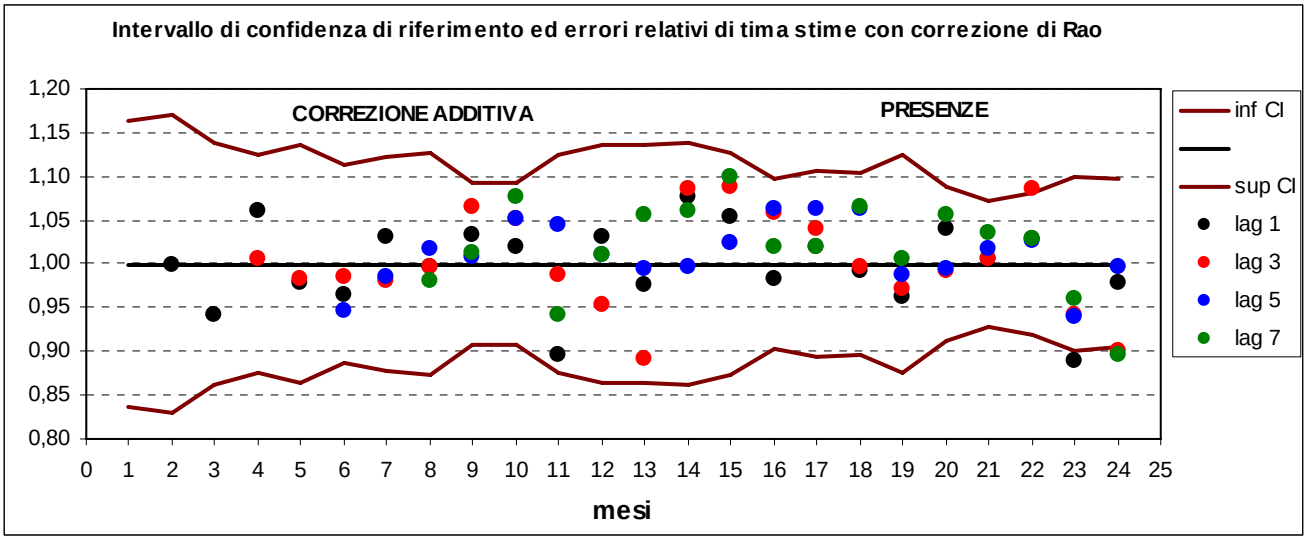
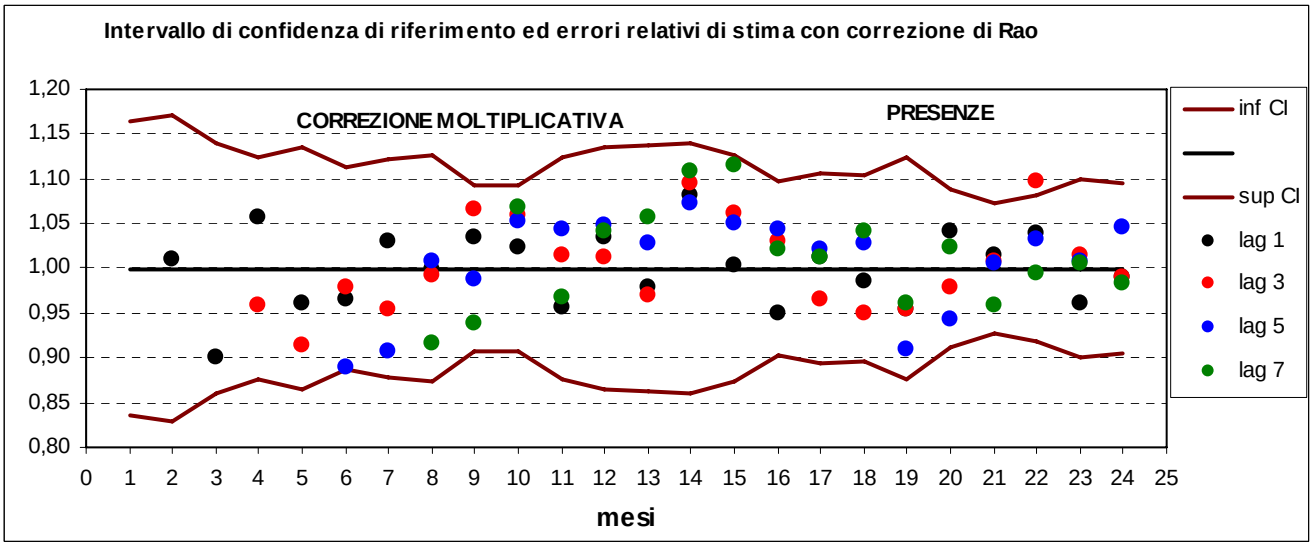


Grafico 4.5



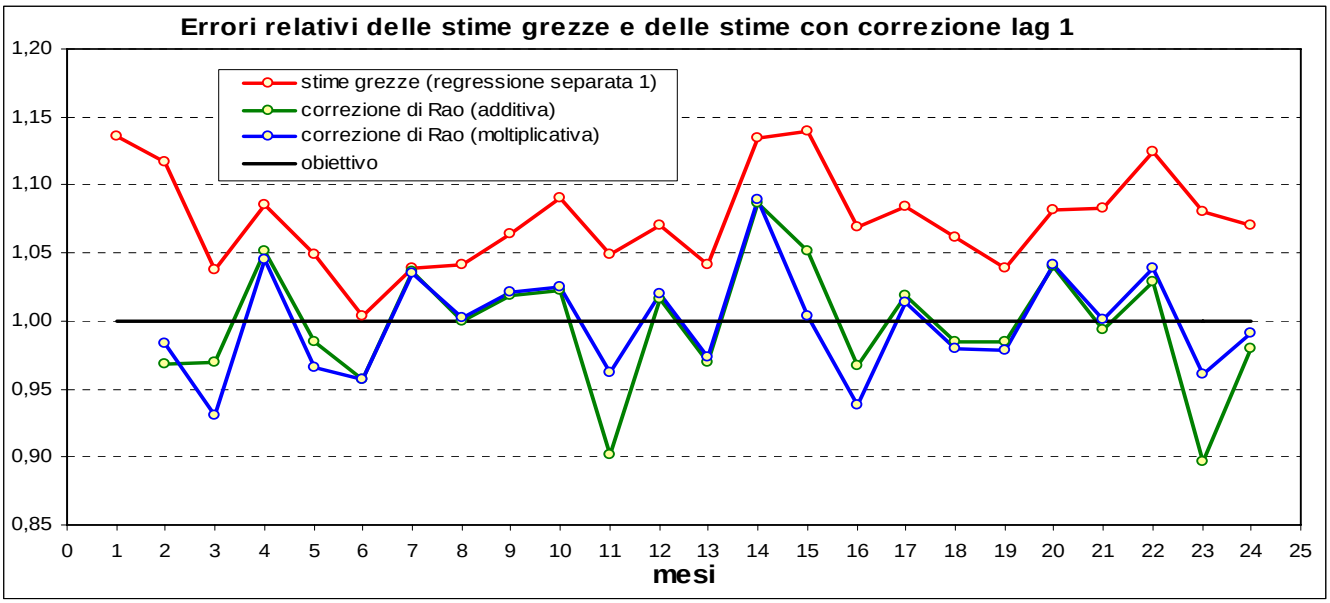
L’intervallo di confidenza di riferimento tende a restringersi nel tempo per effetto dell’aumento degli esercizi partecipanti al sistema turiweb. Si presenta più ampio nei mesi di bassa stagione perché in tali mesi la relazione lineare omogenea presenze-posti letto, alla base dello stimatore rapporto, è più debole e la variabilità elementare residua più alta.

Nel caso di correzione additiva, 4 degli 80 errori relativi di stima risultano fuori dell’intervallo (esattamente la proporzione attesa!), mentre nel caso di correzione moltiplicativa solo uno risulta esterno. La correzione moltiplicativa sembra fornire una maggior garanzia contro valori anomali.

4.3. Applicazione della correzione di Rao alle stime con regressione separata 1

Il grafico 4.6 seguente riporta gli errori relativi delle stime grezze ottenute con lo stimatore di regressione sparata 1 e gli errori relativi delle stime anticipate ottenute con la correzione di Rao al lag 1 additiva e moltiplicativa.

Grafico 4.6



Nella successiva tabella 4.3 sono riportate le sintesi statistiche delle distribuzioni degli errori relativi delle stime anticipate: dal lag 0 (stime grezze) fino al lag 12.

Tabella 4.3

PRESENZE TOTALI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

CORREZIONE ADDITIVA

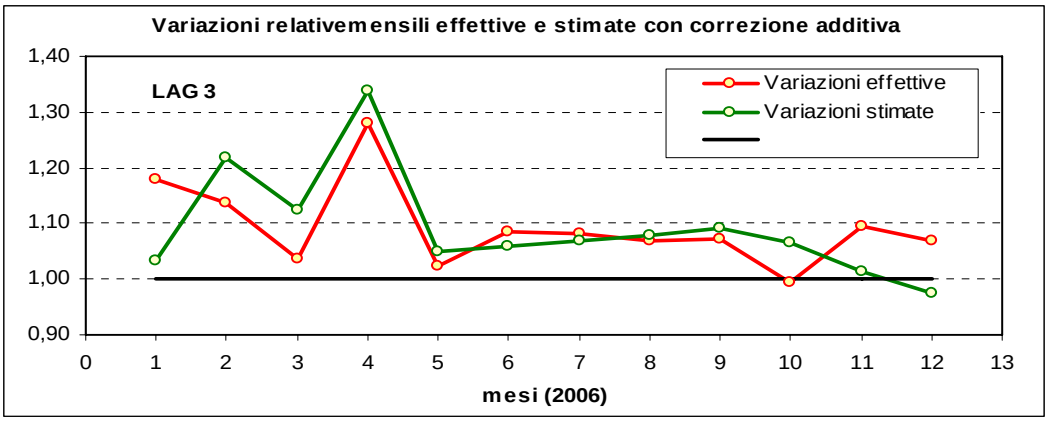
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	1,0039	0,8967	0,8641	0,8776	0,8982	0,9459	0,9446	0,9178	0,9262	0,8971	0,9462	0,9519	0,9270
massimo	1,1391	1,0870	1,1080	1,0858	1,0984	1,0473	1,0631	1,0731	1,1259	1,1333	1,0845	1,0941	1,1022
range	0,1351	0,1903	0,2439	0,2081	0,2002	0,1015	0,1185	0,1553	0,1997	0,2362	0,1384	0,1422	0,1752
media	1,0747	0,9968	0,9945	0,9997	1,0005	1,0063	1,0092	1,0105	1,0151	1,0147	1,0182	1,0257	1,0249
bias	0,0747	-0,0032	-0,0055	-0,0003	0,0005	0,0063	0,0092	0,0105	0,0151	0,0147	0,0182	0,0257	0,0249
varianza	0,0012	0,0020	0,0028	0,0028	0,0025	0,0009	0,0014	0,0017	0,0027	0,0036	0,0017	0,0017	0,0016
mse	0,0068	0,0020	0,0029	0,0028	0,0025	0,0009	0,0015	0,0018	0,0029	0,0038	0,0020	0,0024	0,0022

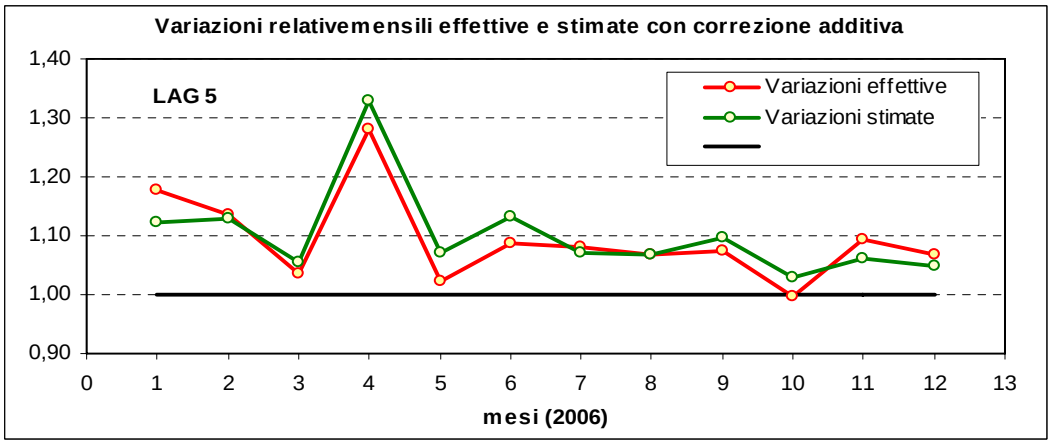
CORREZIONE MOLTIPLICATIVA

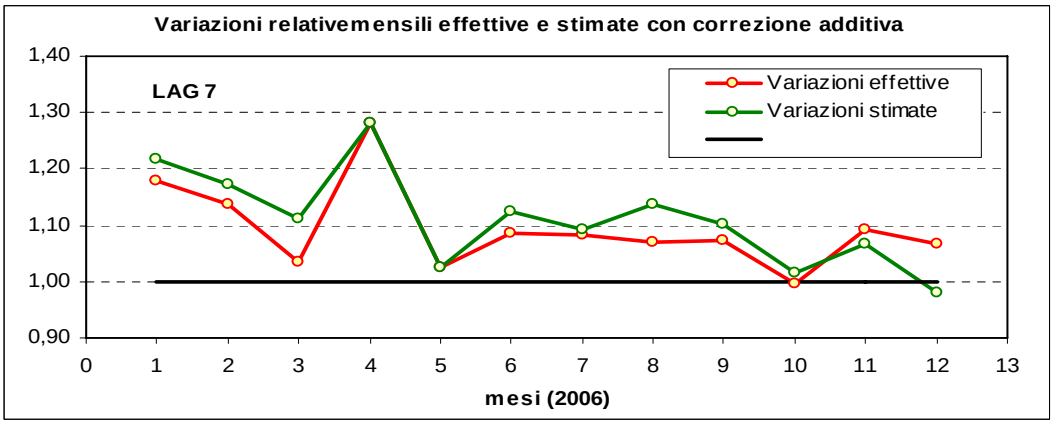
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	1,0039	0,9300	0,9143	0,9322	0,8993	0,8841	0,9153	0,9172	0,9372	0,9393	0,9240	0,9334	0,9176
massimo	1,1391	1,0890	1,0932	1,0828	1,0868	1,0662	1,0895	1,0936	1,1303	1,1346	1,0865	1,0930	1,0972
range	0,1351	0,1590	0,1789	0,1506	0,1875	0,1821	0,1741	0,1764	0,1931	0,1953	0,1625	0,1596	0,1795
media	1,0747	0,9981	0,9968	1,0004	1,0006	1,0020	1,0039	1,0053	1,0081	1,0094	1,0125	1,0195	1,0185
bias	0,0747	-0,0019	-0,0032	0,0004	0,0006	0,0020	0,0039	0,0053	0,0081	0,0094	0,0125	0,0195	0,0185
varianza	0,0012	0,0014	0,0020	0,0020	0,0027	0,0023	0,0021	0,0021	0,0023	0,0033	0,0021	0,0021	0,0017
mse	0,0068	0,0014	0,0020	0,0020	0,0027	0,0023	0,0021	0,0022	0,0023	0,0034	0,0023	0,0025	0,0021

Riguardo al bias, la varianza e di conseguenza l'mse la procedura di correzione di Rao, sia additiva che moltiplicativa associata allo stimatore di regressione separata conduce a risultati migliori, anche se non di molto, rispetto a quella che parte dalle stime preliminari (grezze) dello stimatore rapporto separato. Vediamo nei due grafici seguenti 4.7 e 4.8 il confronto fra variazioni relative mensili effettive e stimate.

Grafico 4.7 (Regressione separata 1 - Correzione additiva)







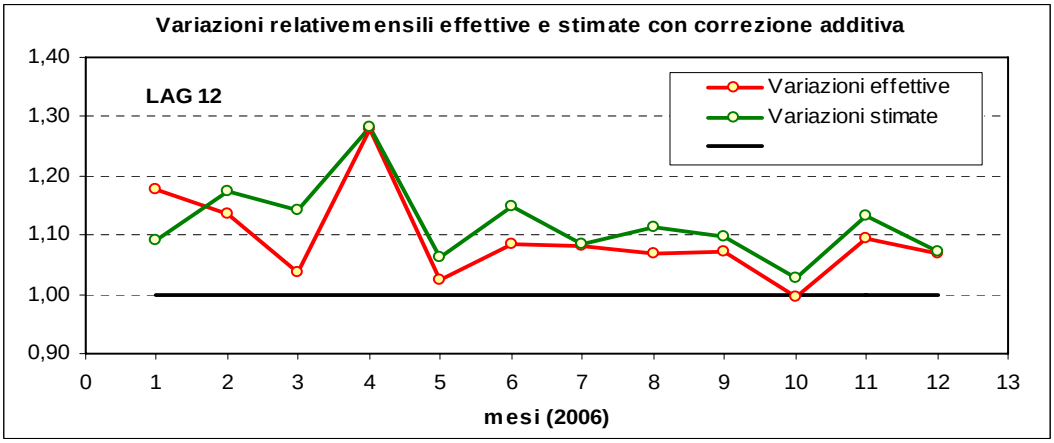
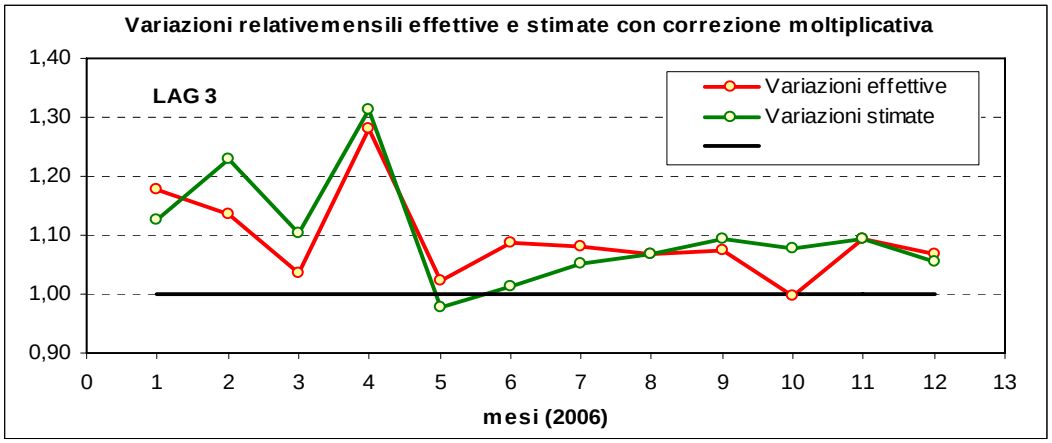
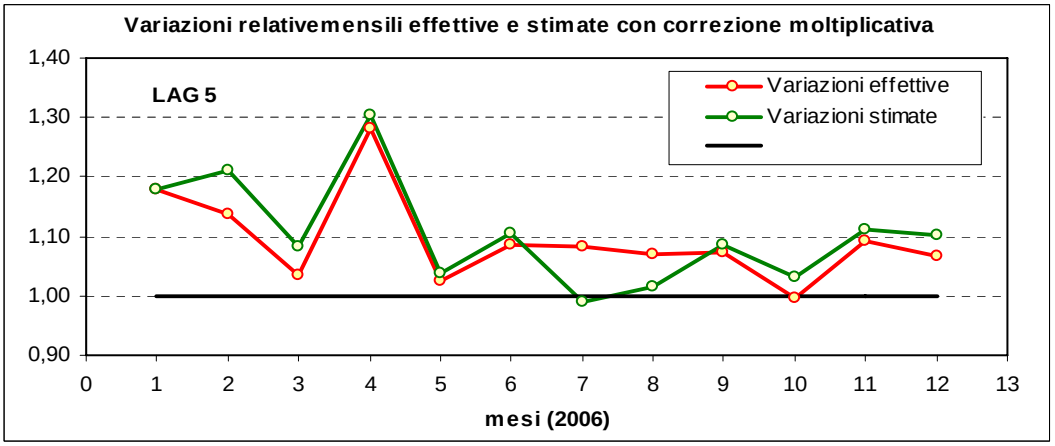
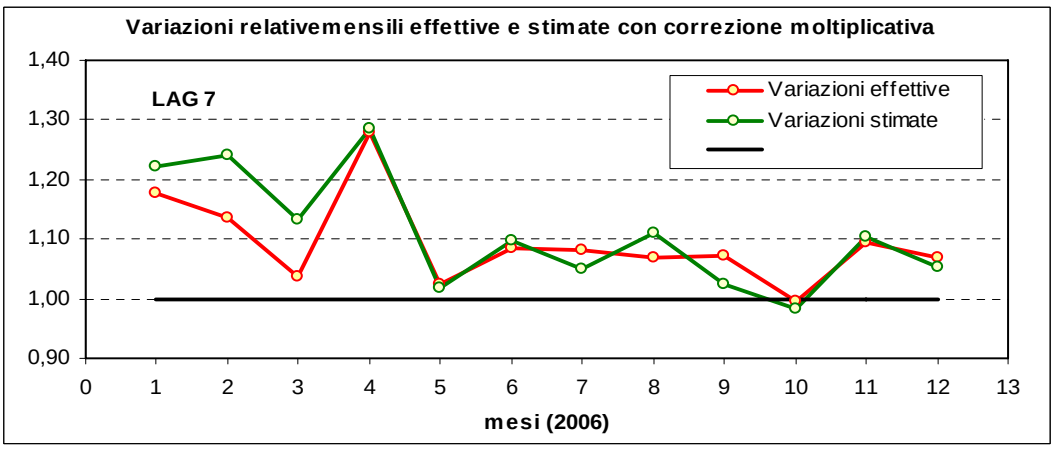
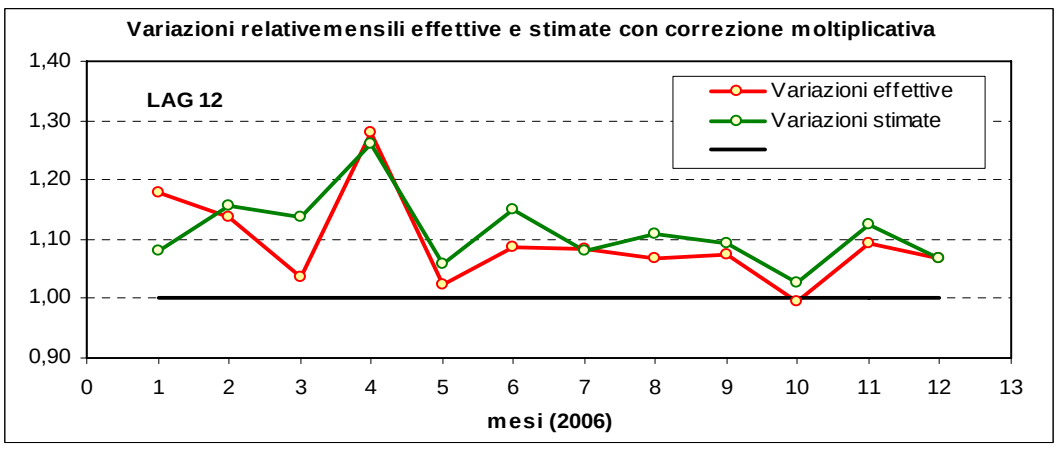


Grafico 4.8 (Regressione separata 1 - Correzione moltiplicativa)









Nella tabella 4.4 che segue si fornisce una misura della differenza fra variazioni relative mensili effettive e variazioni stimate a diversi mesi di anticipazione. L'ideale sarebbe avere, come abbiamo detto a proposito delle stime preliminari con rapporto separato, una distribuzione delle differenze con media zero e varianza nulla. Più la distribuzione è concentrata intorno allo 0 più il procedimento è buono.

Tabella 4.4

Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi 2006	Variazione effettiva	Differenze fra variazioni stimate ed effettive ai lag				
		lag 1	lag 3	lag 5	lag 7	lag 12
Correzione additiva						
1	1,178	-0,035	-0,144	-0,056	0,040	-0,086
2	1,136	0,099	0,082	-0,008	0,037	0,037
3	1,036	0,053	0,089	0,018	0,076	0,106
4	1,280	-0,042	0,059	0,050	0,002	0,003
5	1,024	0,020	0,024	0,048	0,001	0,037
6	1,086	-0,016	-0,028	0,046	0,039	0,063
7	1,082	-0,017	-0,014	-0,010	0,010	0,002
8	1,069	0,043	0,010	-0,001	0,067	0,045
9	1,073	-0,007	0,020	0,023	0,028	0,025
10	0,995	0,028	0,071	0,032	0,020	0,033
11	1,093	-0,113	-0,080	-0,032	-0,026	0,039
12	1,068	-0,021	-0,095	-0,019	-0,088	0,004
Media	**1,0933**	**-0,0008**	**-0,0004**	**0,0078**	**0,0172**	**0,0257**
Varianza		**0,0027**	**0,0051**	**0,0011**	**0,0018**	**0,0019**
Correzione moltiplicativa						
1	1,178	-0,0317	-0,0530	0,0005	0,0447	-0,0971
2	1,136	0,1011	0,0926	0,0753	0,1043	0,0187
3	1,036	0,0040	0,0661	0,0455	0,0970	0,1006
4	1,280	-0,0789	0,0330	0,0239	0,0056	-0,0196
5	1,024	0,0143	-0,0459	0,0125	-0,0068	0,0346
6	1,086	-0,0220	-0,0736	0,0207	0,0131	0,0627
7	1,082	-0,0237	-0,0307	-0,0917	-0,0325	-0,0008
8	1,069	0,0441	-0,0024	-0,0541	0,0405	0,0409
9	1,073	0,0015	0,0213	0,0141	-0,0489	0,0189
10	0,995	0,0383	0,0825	0,0375	-0,0126	0,0306
11	1,093	-0,0431	-0,0012	0,0189	0,0117	0,0323
12	1,068	-0,0102	-0,0129	0,0322	-0,0137	-0,0008
Media	**1,0933**	**-0,0005**	**0,0063**	**0,0113**	**0,0168**	**0,0184**
Varianza		**0,0020**	**0,0027**	**0,0018**	**0,0021**	**0,0021**

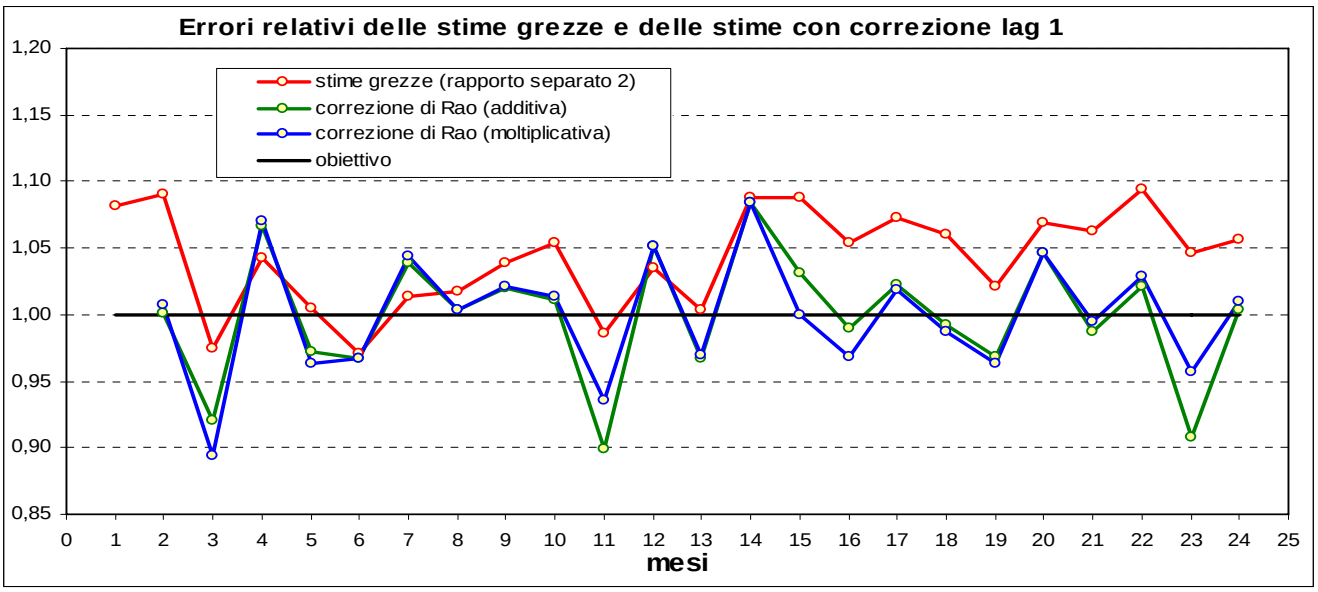
Per quanto riguarda la capacità di ridurre il bias, partendo dalle stime preliminari (grezze) ottenute con la regressione separata 1, si ottengono, dopo la correzione di Rao ai diversi lag, stime anticipate un po' migliori di quelle ottenute con il rapporto separato.

4.4. Applicazione della correzione di Rao alle stime con rapporto separato 2

In questo e nel successivo paragrafo applichiamo la correzione di Rao alle stime grezze ottenute con la post-stratificazione che usa una variabile di classificazione che sintetizza la qualità della struttura ricettiva e che ha una certa capacità esplicativa del meccanismo di autoselezione.

Il grafico 4.9 riporta, come in precedenza, gli errori di stima delle stime grezze e quelli delle stime anticipate ottenute con la correzione additiva e moltiplicativa al lag 1.

Grafico 4.9



Nella tabella le sintesi statistiche delle distribuzioni degli errori relativi delle stime anticipate: dal lag 0 (stime grezze) fino al lag 12.

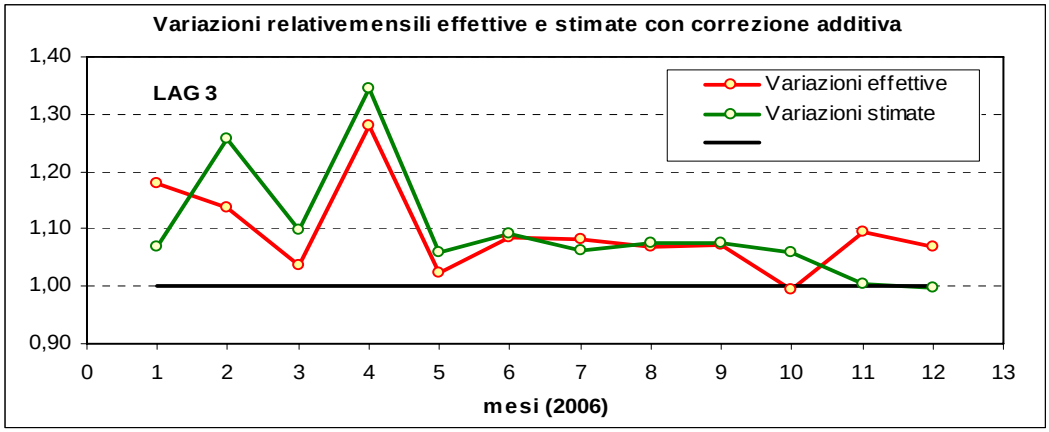
Tabella 4.5

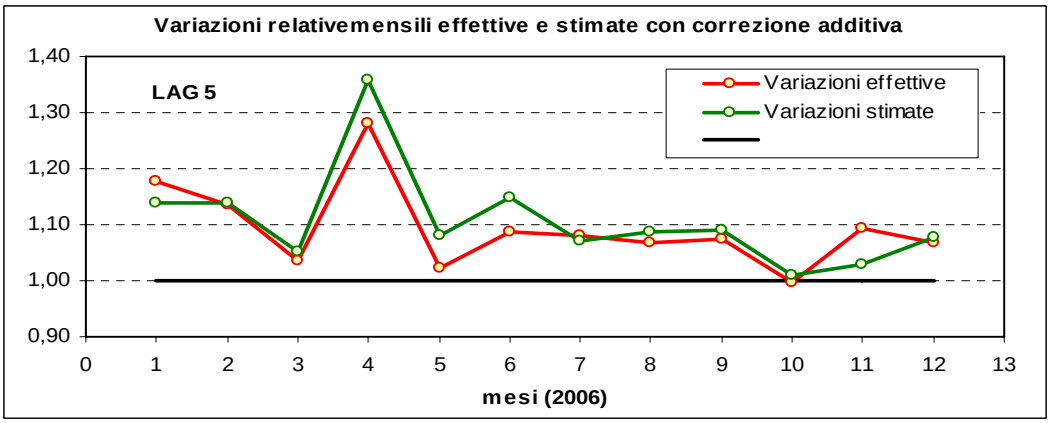
PRESENZE TOTALI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

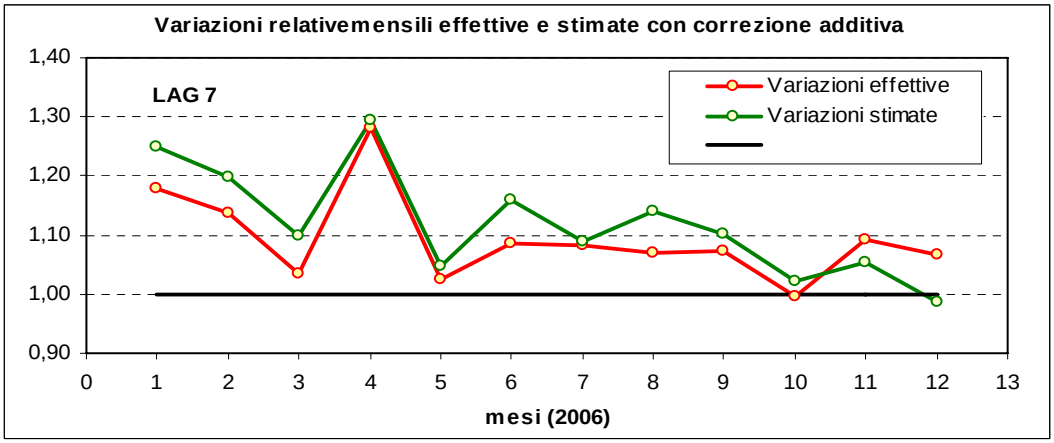
CORREZIONE ADDITIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9712	0,8987	0,9007	0,9070	0,9112	0,9362	0,9278	0,9232	0,9489	0,9229	0,9234	0,9341	0,9346
massimo	1,0942	1,0836	1,0852	1,1064	1,1000	1,0624	1,0916	1,0724	1,1529	1,1303	1,0845	1,1264	1,1130
range	0,1230	0,1849	0,1844	0,1994	0,1889	0,1262	0,1638	0,1493	0,2039	0,2074	0,1611	0,1924	0,1785
media	1,0428	0,9988	0,9952	1,0019	1,0031	1,0103	1,0142	1,0184	1,0258	1,0266	1,0288	1,0383	1,0369
bias	0,0428	-0,0012	-0,0048	0,0019	0,0031	0,0103	0,0142	0,0184	0,0258	0,0266	0,0288	0,0383	0,0369
varianza	0,0013	0,0021	0,0023	0,0026	0,0024	0,0012	0,0019	0,0020	0,0022	0,0034	0,0021	0,0026	0,0019
mse	0,0031	0,0021	0,0023	0,0026	0,0024	0,0013	0,0021	0,0023	0,0028	0,0041	0,0029	0,0041	0,0033
CORREZIONE MOLTIPLICATIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9712	0,8936	0,9006	0,9217	0,8909	0,8979	0,9333	0,9406	0,9609	0,9040	0,9111	0,9209	0,9281
massimo	1,0942	1,0838	1,0838	1,1041	1,1041	1,0691	1,0891	1,0817	1,1203	1,1203	1,0849	1,1170	1,1170
range	0,1230	0,1902	0,1832	0,1824	0,2132	0,1713	0,1558	0,1410	0,1595	0,2163	0,1738	0,1961	0,1889
media	1,0428	0,9999	0,9980	1,0034	1,0049	1,0080	1,0112	1,0146	1,0186	1,0217	1,0250	1,0348	1,0346
bias	0,0428	-0,0001	-0,0020	0,0034	0,0049	0,0080	0,0112	0,0146	0,0186	0,0217	0,0250	0,0348	0,0346
varianza	0,0013	0,0019	0,0019	0,0020	0,0026	0,0020	0,0019	0,0021	0,0017	0,0036	0,0026	0,0030	0,0022
mse	0,0031	0,0019	0,0019	0,0020	0,0027	0,0021	0,0020	0,0023	0,0020	0,0041	0,0032	0,0042	0,0034

Il comportamento delle stime anticipate ai diversi lag di correzione non è molto diverso da i due casi esaminati nei paragrafi 4.2 e 4.3, come risulta anche dai grafici 4.10 e 4.11 dove si confrontano le variazioni relative mensili effettive con quelle stimate.

Grafico 4.10 (Rapporto separato 2 - Correzione additiva)







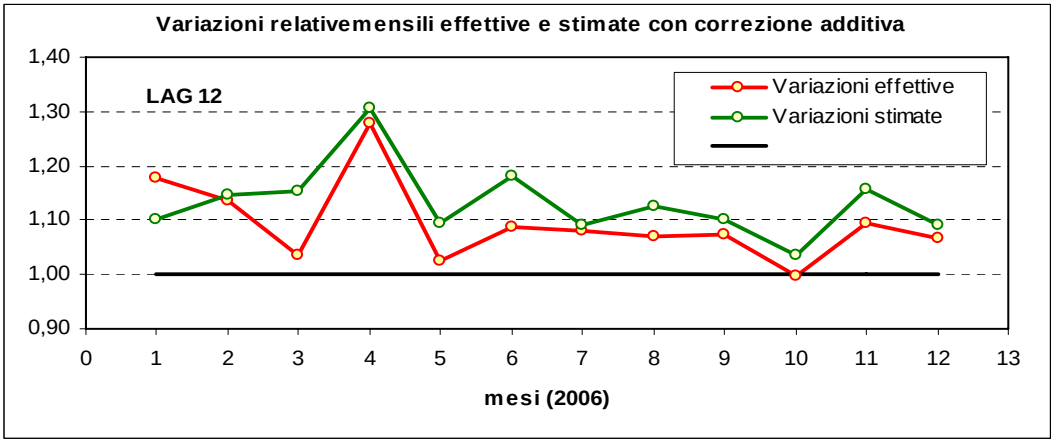
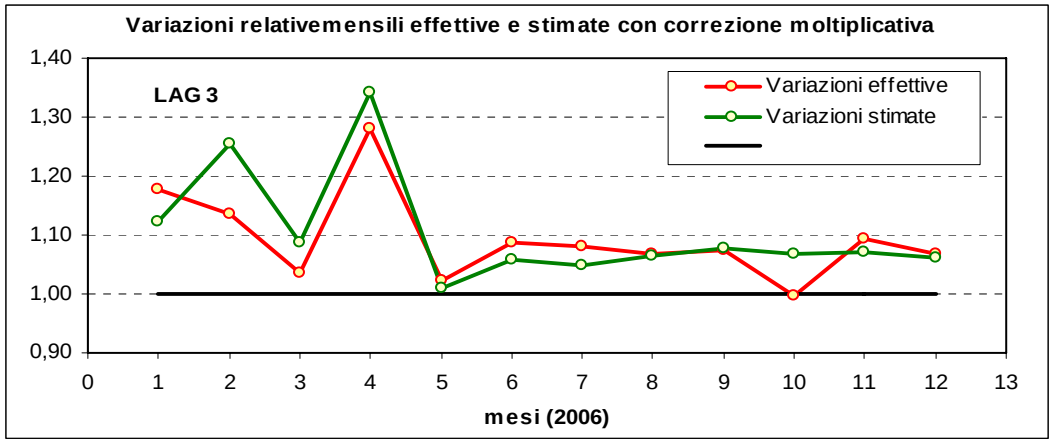
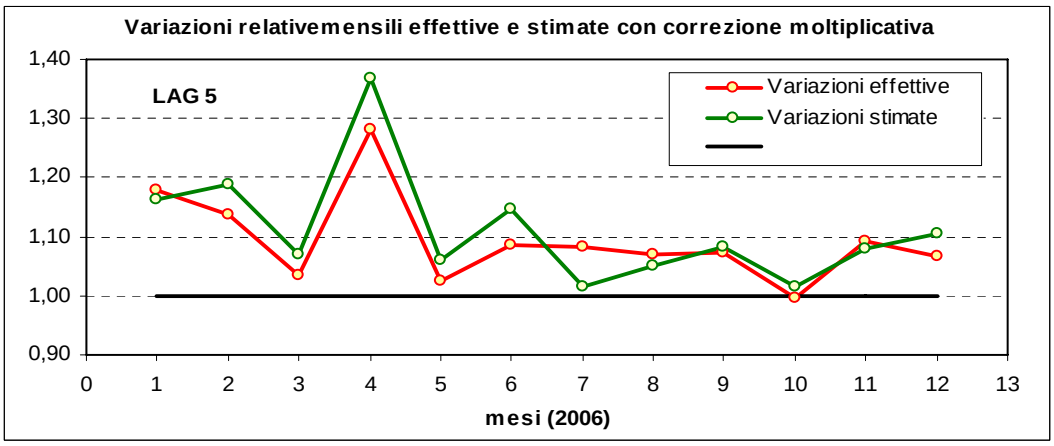
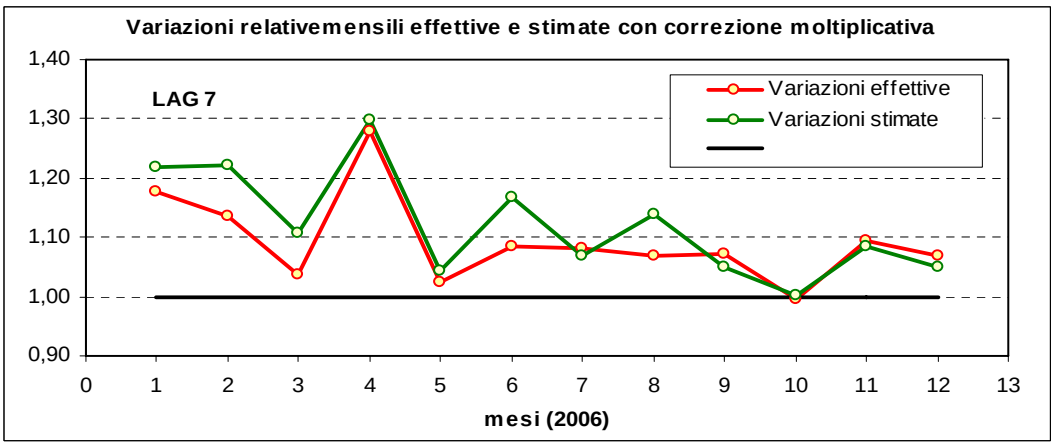
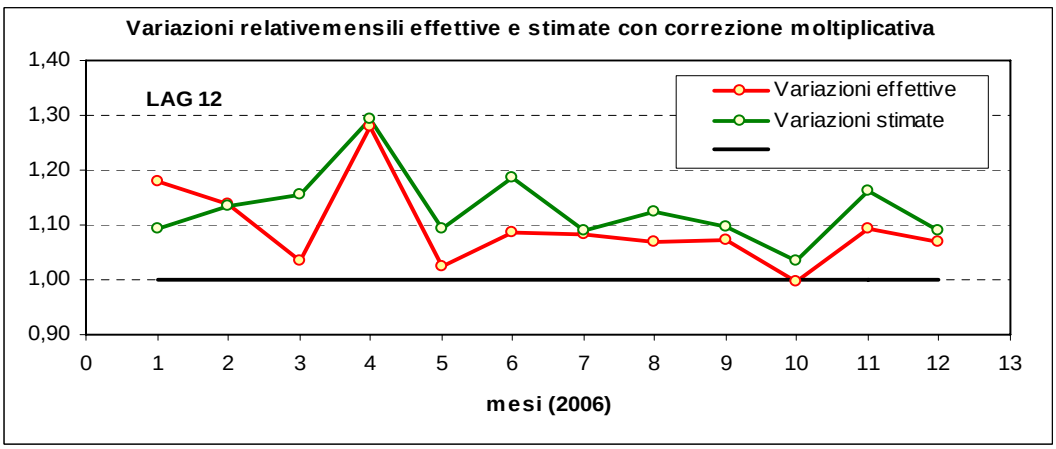


Grafico 4.11 (Rapporto separato 2 - Correzioni moltiplicative)









Nella tabella 4.6 che segue riportiamo una misura della differenza fra variazioni relative mensili e variazioni stimate in corrispondenza dei lag 1, 3, 5, 7, 12 di correzione.

Tabella 4.6

Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi 2006	Variazione effettiva	Differenze fra variazioni stimate ed effettive ai lag				
		lag 1	lag 3	lag 5	lag 7	lag 12
Correzione additiva						
1	1,178	-0,038	-0,110	-0,040	0,072	-0,077
2	1,136	0,095	0,121	0,001	0,061	0,010
3	1,036	0,032	0,063	0,017	0,062	0,117
4	1,280	-0,014	0,066	0,080	0,016	0,026
5	1,024	0,024	0,035	0,056	0,025	0,070
6	1,086	-0,008	0,005	0,063	0,073	0,094
7	1,082	-0,034	-0,021	-0,011	0,007	0,010
8	1,069	0,049	0,006	0,018	0,072	0,056
9	1,073	-0,013	0,002	0,017	0,028	0,029
10	0,995	0,021	0,065	0,014	0,028	0,040
11	1,093	-0,101	-0,091	-0,065	-0,038	0,065
12	1,068	0,004	-0,069	0,009	-0,082	0,025
Media	**1,0933**	**0,0014**	**0,0060**	**0,0133**	**0,0270**	**0,0387**
Varianza		**0,0022**	**0,0045**	**0,0015**	**0,0021**	**0,0022**
Correzione moltiplicativa						
1	1,178	-0,0360	-0,0556	-0,0157	0,0397	-0,0847
2	1,136	0,0952	0,1182	0,0532	0,0839	-0,0022
3	1,036	0,0000	0,0525	0,0338	0,0719	0,1211
4	1,280	-0,0405	0,0634	0,0885	0,0176	0,0132
5	1,024	0,0192	-0,0138	0,0374	0,0191	0,0699
6	1,086	-0,0135	-0,0280	0,0607	0,0821	0,0993
7	1,082	-0,0395	-0,0331	-0,0663	-0,0149	0,0086
8	1,069	0,0497	-0,0045	-0,0189	0,0691	0,0539
9	1,073	-0,0055	0,0034	0,0099	-0,0243	0,0248
10	0,995	0,0289	0,0710	0,0194	0,0056	0,0383
11	1,093	-0,0476	-0,0228	-0,0139	-0,0074	0,0677
12	1,068	0,0100	-0,0071	0,0365	-0,0170	0,0214
Media	**1,0933**	**0,0017**	**0,0120**	**0,0187**	**0,0271**	**0,0359**
Varianza		**0,0016**	**0,0025**	**0,0016**	**0,0015**	**0,0026**

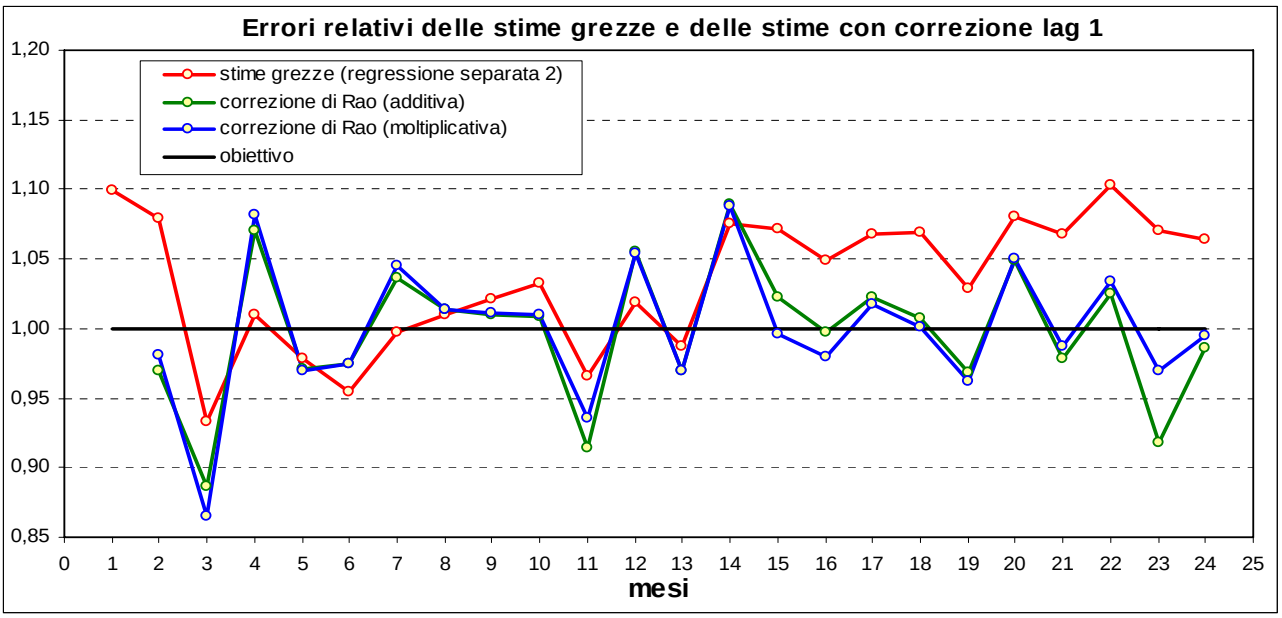
La bontà delle stime delle variazioni relative mensili in questo caso risulta leggermente inferiore a quella ottenuta a partire dalle stime grezze con regressione separata 1.

Dopo aver ricavato, nel paragrafo successivo, i confronti fra variazioni effettive e stimate per la regressione separata 2, effettueremo un confronto sintetico dei quattro procedimenti di stima anticipata delle variazioni relative mensili.

4.5. Applicazione della correzione di Rao alle stime con regressione separata 2

Il grafico 4.12 seguente riporta, come in precedenza, gli errori di stima delle stime grezze ottenute con lo stimatore di regressione separata e post-stratificazione 2, nonché gli errori di stima dopo la correzione di Rao al lag 1 additiva e moltiplicativa.

Grafico 4.12



Nella tabella 4.7 riportiamo le sintesi statistiche delle distribuzioni degli errori relativi delle 24 stime anticipate: dal lag 0 (stime grezze) fino al lag 12.

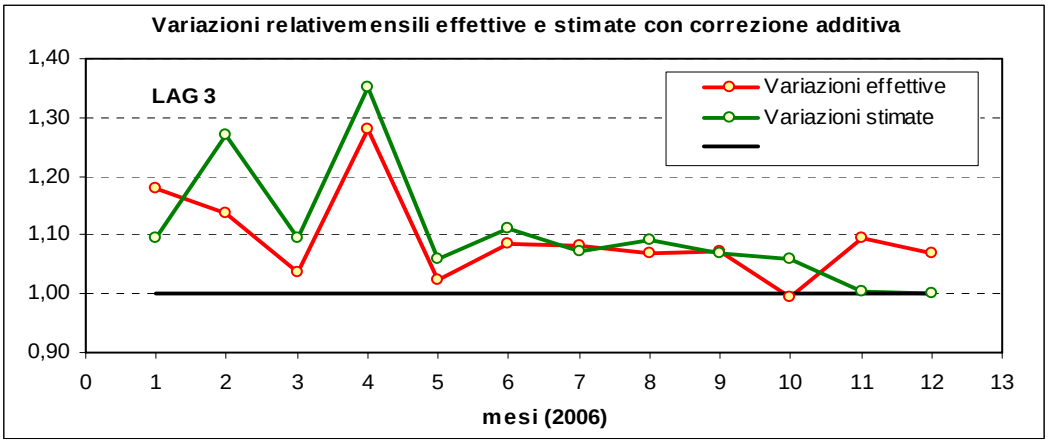
Tabella 4.7

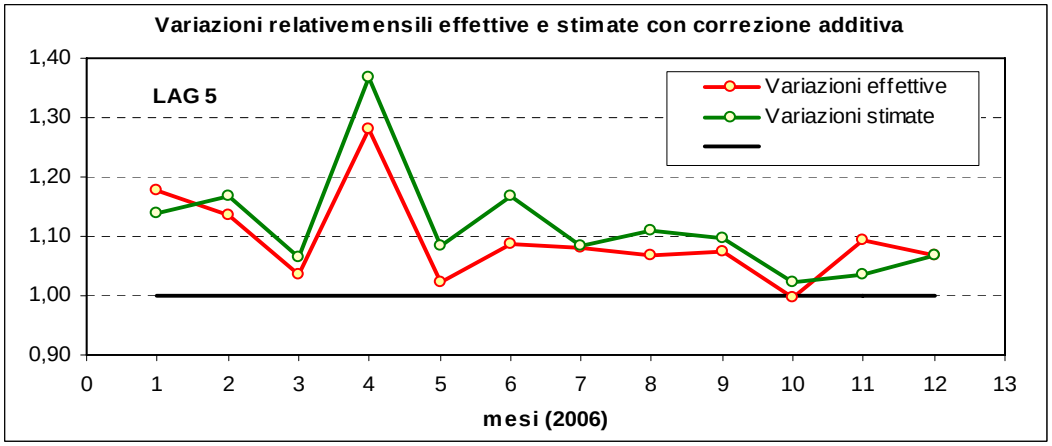
PRESENZE TOTALI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

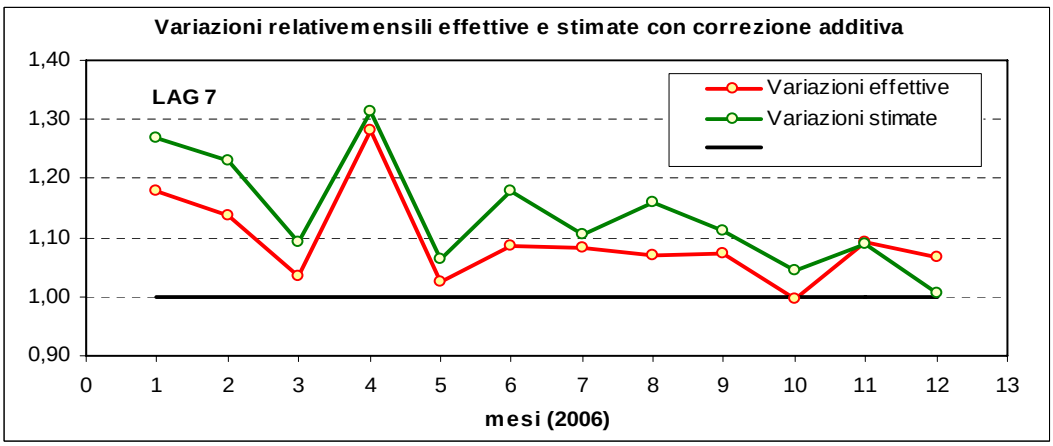
CORREZIONE ADDITIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9337	0,8869	0,8693	0,9189	0,8949	0,9120	0,9274	0,9410	0,9785	0,9112	0,8905	0,9266	0,9035
massimo	1,1033	1,0889	1,0802	1,1180	1,0992	1,0751	1,1066	1,0870	1,1773	1,1378	1,1010	1,1731	1,1351
range	0,1696	0,2019	0,2110	0,1991	0,2043	0,1631	0,1793	0,1460	0,1988	0,2266	0,2105	0,2466	0,2316
media	1,0348	0,9976	0,9955	1,0044	1,0078	1,0177	1,0243	1,0322	1,0412	1,0435	1,0456	1,0560	1,0544
bias	0,0348	-0,0024	-0,0045	0,0044	0,0078	0,0177	0,0243	0,0322	0,0412	0,0435	0,0456	0,0560	0,0544
varianza	0,0021	0,0023	0,0027	0,0027	0,0027	0,0018	0,0021	0,0021	0,0022	0,0039	0,0034	0,0043	0,0033
mese	0,0033	0,0023	0,0027	0,0028	0,0028	0,0021	0,0027	0,0031	0,0039	0,0058	0,0055	0,0074	0,0063
CORREZIONE MOLTIPLICATIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9337	0,8655	0,8495	0,9075	0,8847	0,8684	0,9073	0,9192	0,9296	0,8953	0,8788	0,9155	0,8986
massimo	1,1033	1,0883	1,0845	1,1127	1,1089	1,0860	1,1056	1,1070	1,1261	1,1222	1,1048	1,1511	1,1471
range	0,1696	0,2228	0,2350	0,2052	0,2242	0,2177	0,1983	0,1878	0,1965	0,2269	0,2260	0,2357	0,2486
media	1,0348	0,9998	0,9995	1,0073	1,0107	1,0163	1,0211	1,0266	1,0315	1,0365	1,0416	1,0535	1,0549
bias	0,0348	-0,0002	-0,0005	0,0073	0,0107	0,0163	0,0211	0,0266	0,0315	0,0365	0,0416	0,0535	0,0549
varianza	0,0021	0,0023	0,0026	0,0024	0,0032	0,0029	0,0024	0,0027	0,0021	0,0044	0,0040	0,0046	0,0038
mese	0,0033	0,0023	0,0026	0,0025	0,0033	0,0031	0,0029	0,0034	0,0031	0,0057	0,0057	0,0075	0,0069

Con questo stimatore la distribuzione delle 24 stime preliminari presenta un bias più contenuto dei tre casi precedenti, però una varianza più alta; nell'applicare la correzione di Rao, il bias viene ridotto ai lag bassi, ma poi tende a risalire; al lag 12 è superiore al quello delle stime preliminari. Vediamo il comportamento delle stime anticipate delle variazioni relative mensili.

Grafico 4.13 (Regressione separata 2 - Correzione additiva)







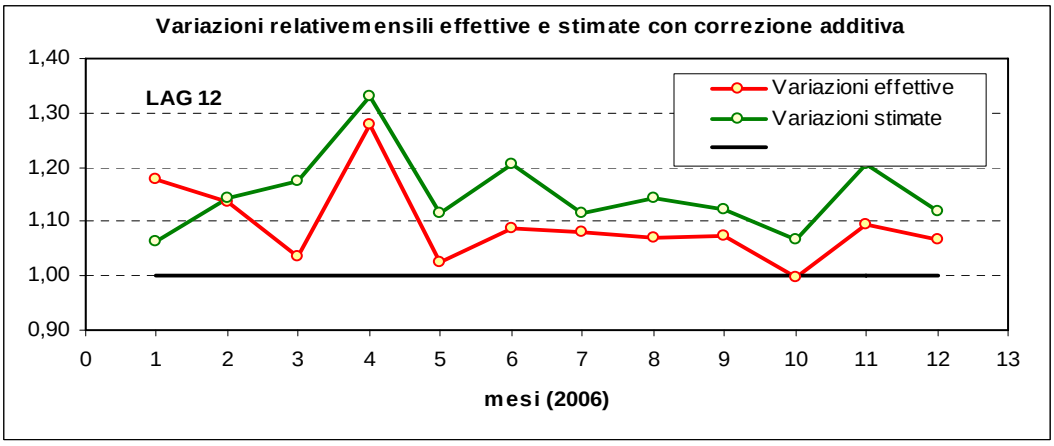
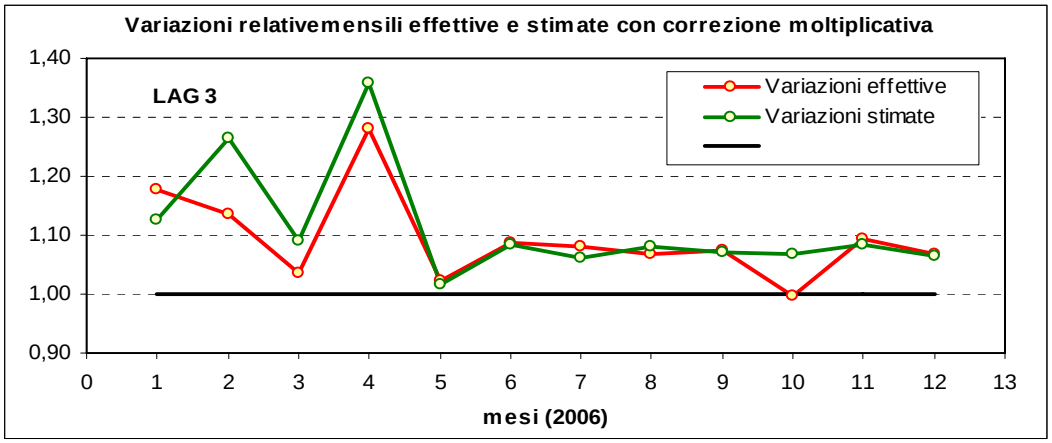
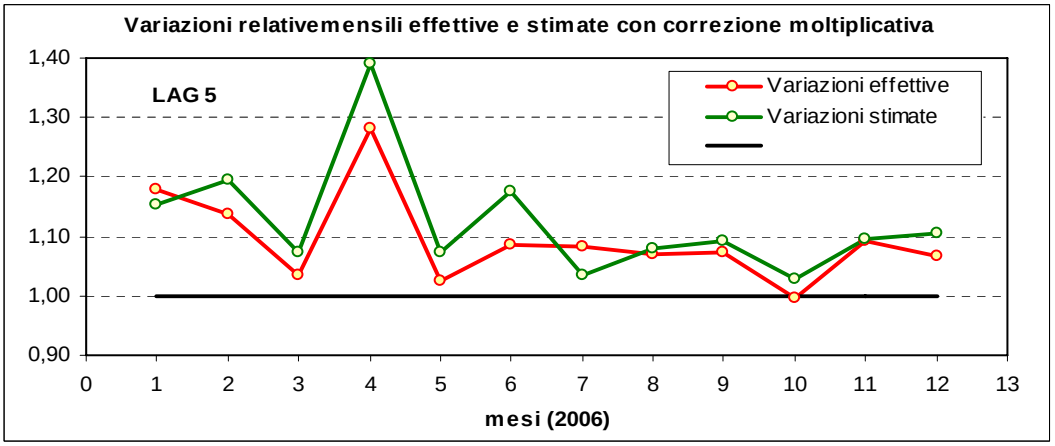
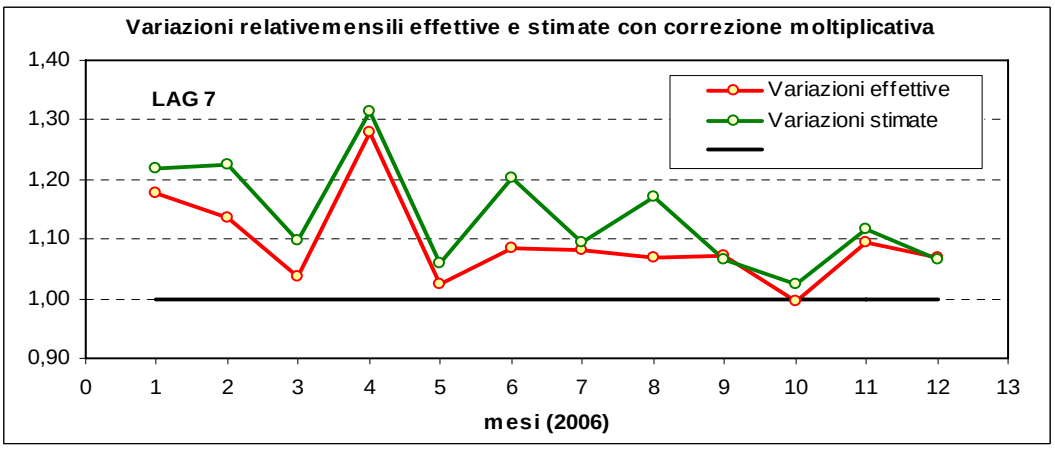
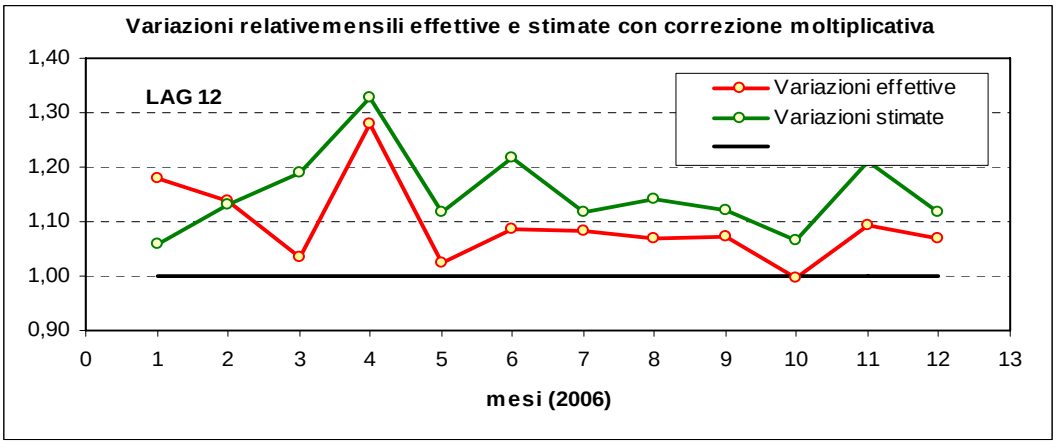


Grafico 4.14 (Regressione separata 2 - Correzione moltiplicativa)









Nella tabella 4.8 che segue riportiamo una misura della differenza fra variazioni relative mensili e variazioni stimate in corrispondenza dei lag 1, 3, 5, 7, 12 di correzione.

Tabella 4.8

Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi 2006	Variazione effettiva	Differenze fra variazioni stimate ed effettive ai lag				
		lag 1	lag 3	lag 5	lag 7	lag 12
Correzione additiva						
1	1,178	-0,036	-0,083	-0,041	0,092	-0,114
2	1,136	0,101	0,134	0,030	0,093	0,006
3	1,036	0,023	0,060	0,029	0,057	0,140
4	1,280	-0,004	0,071	0,089	0,033	0,052
5	1,024	0,023	0,035	0,060	0,039	0,091
6	1,086	0,008	0,027	0,082	0,094	0,121
7	1,082	-0,034	-0,009	0,002	0,023	0,034
8	1,069	0,053	0,023	0,041	0,092	0,076
9	1,073	-0,023	-0,005	0,025	0,038	0,050
10	0,995	0,025	0,064	0,028	0,050	0,071
11	1,093	-0,090	-0,089	-0,057	-0,005	0,111
12	1,068	-0,015	-0,069	-0,001	-0,063	0,051
Media	**1,0933**	**0,0025**	**0,0133**	**0,0240**	**0,0453**	**0,0573**
Varianza		**0,0022**	**0,0042**	**0,0018**	**0,0020**	**0,0040**
Correzione moltiplicativa						
1	1,178	-0,0352	-0,0510	-0,0265	0,0410	-0,1195
2	1,136	0,1003	0,1281	0,0590	0,0884	-0,0042
3	1,036	-0,0036	0,0539	0,0389	0,0623	0,1524
4	1,280	-0,0264	0,0795	0,1101	0,0341	0,0490
5	1,024	0,0185	-0,0065	0,0502	0,0354	0,0930
6	1,086	0,0014	-0,0018	0,0898	0,1162	0,1307
7	1,082	-0,0411	-0,0209	-0,0464	0,0113	0,0342
8	1,069	0,0536	0,0124	0,0092	0,1003	0,0740
9	1,073	-0,0130	-0,0021	0,0186	-0,0076	0,0477
10	0,995	0,0337	0,0721	0,0329	0,0299	0,0685
11	1,093	-0,0330	-0,0104	0,0007	0,0219	0,1178
12	1,068	-0,0056	-0,0028	0,0370	-0,0036	0,0486
Media	**1,0933**	**0,0041**	**0,0209**	**0,0311**	**0,0441**	**0,0577**
Varianza		**0,0016**	**0,0024**	**0,0018**	**0,0015**	**0,0046**

4.6. Confronto fra gli otto procedimenti di stima

Abbiamo considerato quattro stimatori: rapporto separato con post-stratificazione 1 e 2, regressione separata con post-stratificazione 1 e 2. Si veda il paragrafo 3.2 per la definizione delle due post stratificazioni. In tutti i quattro casi la variabile quantitativa (x) è data dai posti letto.

Per ciascuno dei quattro stimatori abbiamo applicato le due versioni di correzione del bias (additiva e moltiplicativa). Conviene riepilogare con una tabella e alcuni grafici la media e la varianza delle differenze fra le variazioni relative mensili effettive e quelle ottenute con le stime anticipate corrette ai lag 0, (nessuna correzione) e ai lag 1 – 12. Sulla base delle sintesi soprattutto grafiche, instauriamo un confronto che ci permetta la valutazione della procedura migliore, almeno fra le considerate.

Tabella 4.9

Medie e varianze delle 12 differenze fra variazioni relative mensili effettive e stimate (gen - dic 2006 / gen-dic 2005), senza correzione di Rao (lag 0) e con correzione al lag 1-12

Stime grezze	Tipo di correzione	Sintesi	Differenze fra variazioni stimate ed effettive ai lag												
			lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Rapporto separato post-stratificazione 1	additiva	Media	0,1061	-0,0002	0,0023	0,0047	0,0057	0,0152	0,0143	0,0282	0,0361	0,0344	0,0484	0,0501	0,0436
		Varianza	0,0011	0,0027	0,0043	0,0054	0,0038	0,0015	0,0029	0,0031	0,0042	0,0050	0,0020	0,0023	0,0020
	moltiplicativa	Media	0,1061	0,0001	0,0050	0,0099	0,0119	0,0175	0,0173	0,0263	0,0323	0,0324	0,0435	0,0427	0,0359
		Varianza	0,0011	0,0018	0,0025	0,0029	0,0027	0,0023	0,0025	0,0027	0,0032	0,0043	0,0024	0,0029	0,0023
Regressione separata post-stratificazione 1	additiva	Media	0,0910	-0,0008	-0,0003	-0,0004	-0,0009	0,0078	0,0064	0,0172	0,0213	0,0167	0,0293	0,0329	0,0257
		Varianza	0,0011	0,0027	0,0038	0,0051	0,0038	0,0011	0,0019	0,0018	0,0038	0,0049	0,0016	0,0016	0,0019
	moltiplicativa	Media	0,0910	-0,0005	0,0026	0,0063	0,0072	0,0113	0,0105	0,0168	0,0201	0,0180	0,0268	0,0270	0,0184
		Varianza	0,0011	0,0020	0,0026	0,0027	0,0026	0,0018	0,0020	0,0021	0,0029	0,0039	0,0018	0,0021	0,0021
Rapporto separato post-stratificazione 2	additiva	Media	0,0645	0,0014	0,0046	0,0060	0,0049	0,0133	0,0120	0,0270	0,0367	0,0343	0,0468	0,0472	0,0387
		Varianza	0,0007	0,0022	0,0026	0,0045	0,0031	0,0015	0,0022	0,0021	0,0031	0,0037	0,0013	0,0029	0,0022
	moltiplicativa	Media	0,0645	0,0017	0,0074	0,0120	0,0130	0,0187	0,0180	0,0271	0,0335	0,0338	0,0449	0,0438	0,0359
		Varianza	0,0007	0,0016	0,0016	0,0025	0,0021	0,0016	0,0014	0,0015	0,0019	0,0033	0,0015	0,0033	0,0026
Regressione separata post-stratificazione 2	additiva	Media	0,0658	0,0025	0,0086	0,0133	0,0144	0,0240	0,0261	0,0453	0,0578	0,0583	0,0722	0,0710	0,0573
		Varianza	0,0008	0,0022	0,0025	0,0042	0,0031	0,0018	0,0021	0,0020	0,0028	0,0032	0,0011	0,0041	0,0040
	moltiplicativa	Media	0,0658	0,0041	0,0135	0,0209	0,0237	0,0311	0,0325	0,0441	0,0523	0,0555	0,0695	0,0686	0,0577
		Varianza	0,0008	0,0016	0,0014	0,0024	0,0020	0,0018	0,0013	0,0015	0,0014	0,0029	0,0014	0,0043	0,0046

I valori della tabella sono riportati più efficacemente nei successivi quattro grafici.

Grafico 4.15

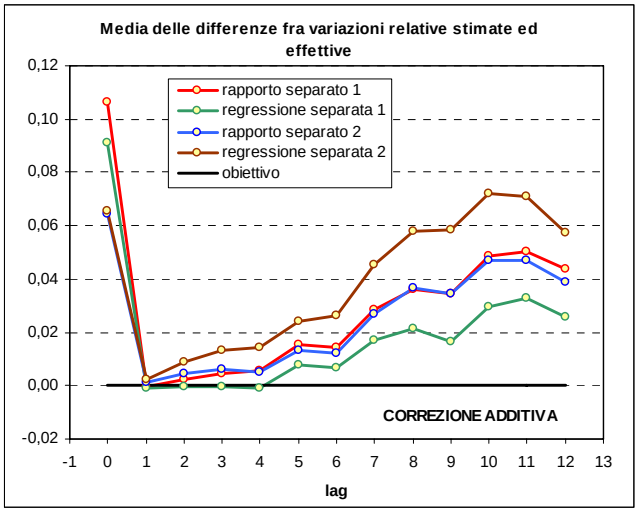


Grafico 4.16

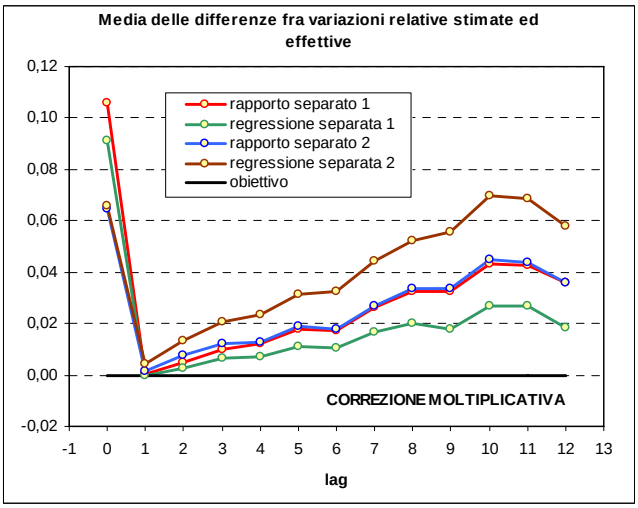


Grafico 4.17

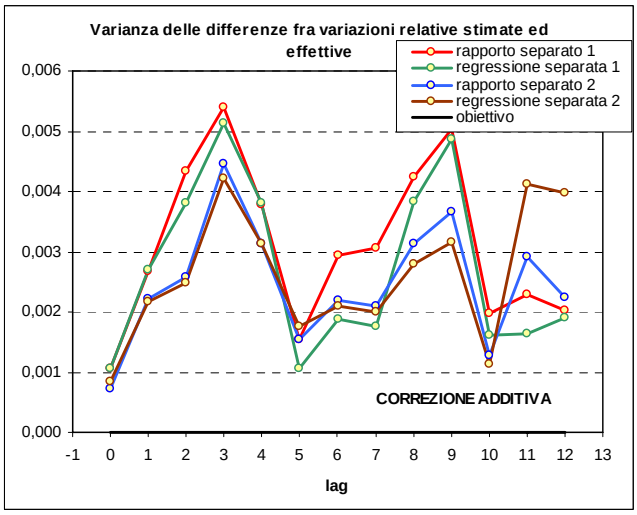
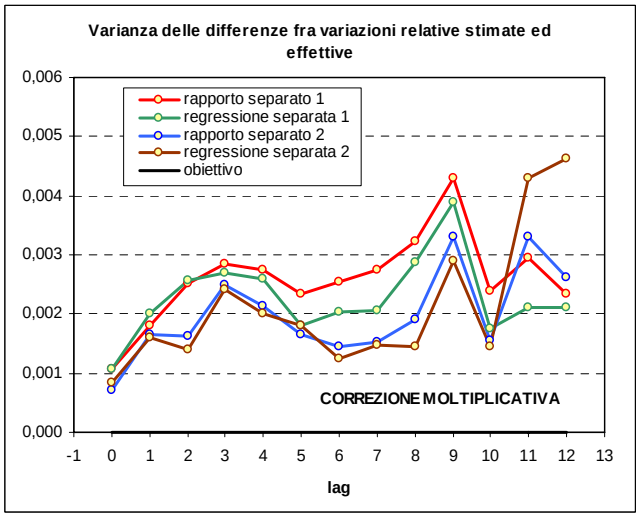


Grafico 4.18



Al lag 1 vi è pochissima differenza fra gli otto metodi di stima anticipata delle variazioni relative mensili: tutti i procedimenti di stima conducono a stime delle variazioni relative mensili sostanzialmente senza bias a prezzo di un leggero incremento della varianza. Le differenze si manifestano al crescer del lag.

Al crescere del lag la capacità di ridurre il bias diminuisce; fra i quattro, lo stimatore di regressione separata 1 fornisce stime anticipate delle variazioni relative mensili mediamente più vicine alle effettive a tutti i lag sia con la correzione additiva sia con quella moltiplicativa. La regressione separata 1 funziona meglio del rapporto separato 1 che è stato lo stimatore sul quale inizialmente avevamo puntato in considerazione della relazione fra presenze e posti letto.

Gli stimatori rapporto e regressione separati con post-stratificazione 2 forniscono stime preliminari migliori (sono quindi preferibili al lag 0); ma quando si applica la correzione di Rao, la regressione separata 2 conduce a sovrastimare le variazioni relative mensili più di quanto facciano gli altri stimatori.

All'aumentare del lag in effetti tutti gli stimatori tendono a sovrastimare sempre più le variazioni relative mensili. La capacità di eliminazione del bias della correzione di Rao non funziona appieno perché l'ipotesi di bias costante nel tempo non è soddisfatta. Questo fatto è più evidente nella regressione separata 2.

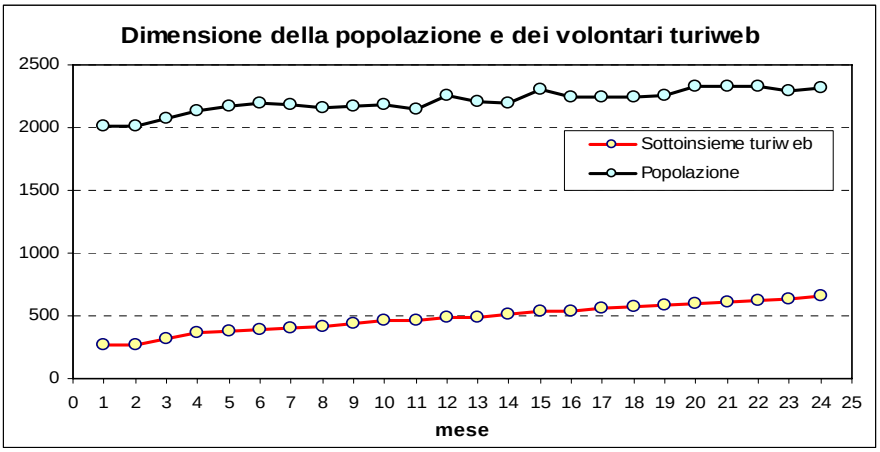
Quanto alla varianza (delle 12 differenze fra variazioni relative mensili stimate ed effettive), il metodo di correzione moltiplicativo appare un po' più stabile di quello additivo. Non si evidenzia una netta gerarchia degli stimatori come nel caso della media.

Dovendo scegliere fra le procedure di stima della variazione relativa mensile per una concreta applicazione, lo stimatore di regressione separata con post-stratificazione 1 (quella convenzionale) e correzione moltiplicativa del bias é un buon candidato.

4.7. Considerazione sulle evidenze empiriche

Tutte le considerazioni fin qui svolte si basano su evidenze empiriche relative a un periodo di 24 mesi (dal gennaio 2005 al dicembre 2006). Vi sono molti fattori che non sono entrati nell'analisi. Il primo fattore é l'evoluzione della popolazione e del sottoinsieme delle strutture turiweb; come evidenzia il grafico 4.19 seguente, la popolazione e il sottoinsieme turiweb sono cresciuti nei due anni considerati.

Grafico 4.19



Non solo non conosciamo il meccanismo di autoselezione in un dato mese, ma neppure come evolve nel tempo.

I campioni turiweb nei 24 mesi sono fortemente correlati, ci aspetteremmo che gli errori di stima seguissero un andamento "liscio"; invece in alcuni mesi si osservano bruschi cambiamenti dell'errore relativo delle stime grezze rispetto al mese precedente. Probabilmente vi sono fattori "esogeni" che modificano la distribuzione della variabile di studio in modo non indipendente dal meccanismo di selezione, fattori ad esempio climatici, economici, sia nel luogo di origine che in quello di destinazione dei flussi turistici.

Infine la nostra osservazione è riferita a due soli anni; il calcolo delle variazioni relative mensili a un solo anno. Questi intervalli temporali potrebbero essere modesti per evidenziare eventuali trend e stagionalità.

Infine, abbiamo assunto come "corretti" i dati dell'indagine censuaria. Ma non possiamo escludere che in qualche mese la rilevazione degli esercizi non turiweb sia affetta da specifici errori.

Dobbiamo infine riconoscere che ci mancano strumenti per una trattazione più approfondita degli errori di stima che, grazie all'indagine censuaria, possiamo osservare, quali quelli specifici di analisi delle serie storiche.

Nel prossimo capitolo effettueremo un approfondimento per il dominio alberghi che costituisce una sottopopolazione di particolar interesse e per la quale siamo in grado di utilizzare informazione ausiliaria più ricca di quella fin qui usata.

Capitolo 5

5.1. Approfondimento per la sottopopolazione degli esercizi alberghieri

Nel precedente capitolo abbiamo esaminato procedimenti di stima delle presenze mensili e delle variazioni relative rispetto al corrispondente mese dell'anno precedente per l'intera popolazione delle strutture turistico-ricettive della provincia di Firenze. Come mostra la tavola 2 del cap.1 le strutture ricettive sono articolate in ben tredici tipologie: dagli affittacamere, ai rifugi alpini, ai campeggi, agli alberghi, ecc. Inoltre queste strutture, anche all'interno della medesima tipologia, differiscono notevolmente per dimensione.

In questo capitolo intendiamo concentrare l'attenzione sulla sottopopolazione degli esercizi alberghieri per tre motivi.

1. Gli esercizi alberghieri costituiscono una quota rilevante della popolazione: 564 unità su un totale di 2400, pari al 35%, con 695.000 presenze su 1.155.000 pari al 60% (ottobre 2006). Gli utenti delle statistiche sul turismo sono pertanto interessati a disporre di informazione tempestiva su questo sottoinsieme.
2. Dal punto di vista metodologico intendiamo valutare se i procedimenti di stima visti per l'intera popolazione funzionano anche in una sottopopolazione: da un lato aumentano le difficoltà poiché si riduce la numerosità campionaria (84 unità a gennaio 2005, 134 unità a dicembre 2006), dall'altro si può sfruttare la disponibilità di maggior informazione ausiliaria per questo dominio. Per gli esercizi alberghieri disponiamo dell'informazione sulla categoria (le stelle, da 1 a 5). Si tratta di una informazione che può avere capacità esplicativa sia del meccanismo di autoselezione sia delle variabili di studio.
3. Infine, abbiamo recuperato per gli esercizi alberghieri alcune informazioni presenti negli archivi dell'attrezzatura e prezzi. Questo con l'obiettivo di sperimentare un metodo alternativo a quello fin qui visto: stimare la probabilità di appartenere al sistema turiweb utilizzando un set di informazioni che possono spiegare la propensione dell'unità a partecipare al sistema turiweb. Stimata questa probabilità, ricondursi a stimatori formalmente analoghi a quelli basati sul disegno in cui sono note le probabilità di inclusione.

5.2. Esplorazione delle relazioni fra letti e presenze, fra qualità e presenze

Sulla base delle informazioni disponibili nell'archivio della consistenza, costruiamo un indice di qualità dell'esercizio alberghiero che useremo nel seguito, e che chiameremo brevemente qualità, dato dal rapporto bagni/camere. Si tratta dello stesso indice utilizzato per l'intera popolazione nel capitolo precedente. La qualità, così definita, potrebbe avere capacità esplicativa sia delle presenze, sia del meccanismo di selezione.

Prima di applicare i procedimenti di stima visti per l'intera popolazione al dominio alberghi, esaminiamo la relazione fra letti e presenze e fra qualità e presenze per ciascuno dei 4 gruppi costituiti dalle stelle: 1, 2, 3, 4-5. Abbiamo aggregato le 4 e 5 stelle perché in questa ultima classe vi è un numero molto limitato di alberghi (15 unità di cui 3 appartenenti al sistema turiweb).

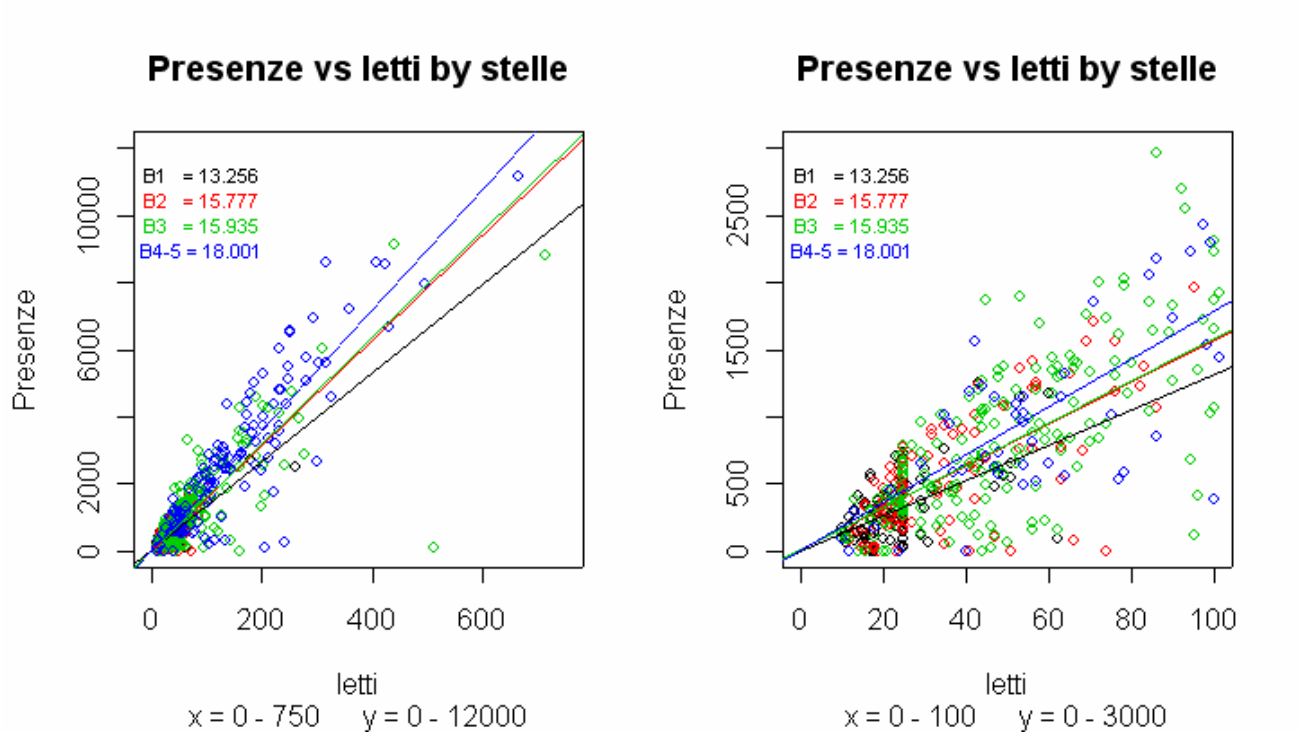
Abbiamo costruito due scatter per ciascuna situazione, in cui il secondo effettua uno zoom nella regione dove i punti si addensano maggiormente. I valori B_i riportati sugli scatter sono i rapporti fra il totale delle presenze e il totale dei letti calcolati per l'intera popolazione o per il sottoinsieme

turiweb nel caso di modello senza intercetta eteroschedastico, mentre sono i coefficienti slope delle rette di regressione nel caso di modello con intercetta omoschedastico.

L'esame di cui sopra fa riferimento al mese di ottobre 2006.

Grafico 5.1 (Scatter con rette di regressione, modello senza intercetta eteroschedastico)

Intera popolazione degli alberghi



Sottopopolazione degli alberghi del sistema turiweb

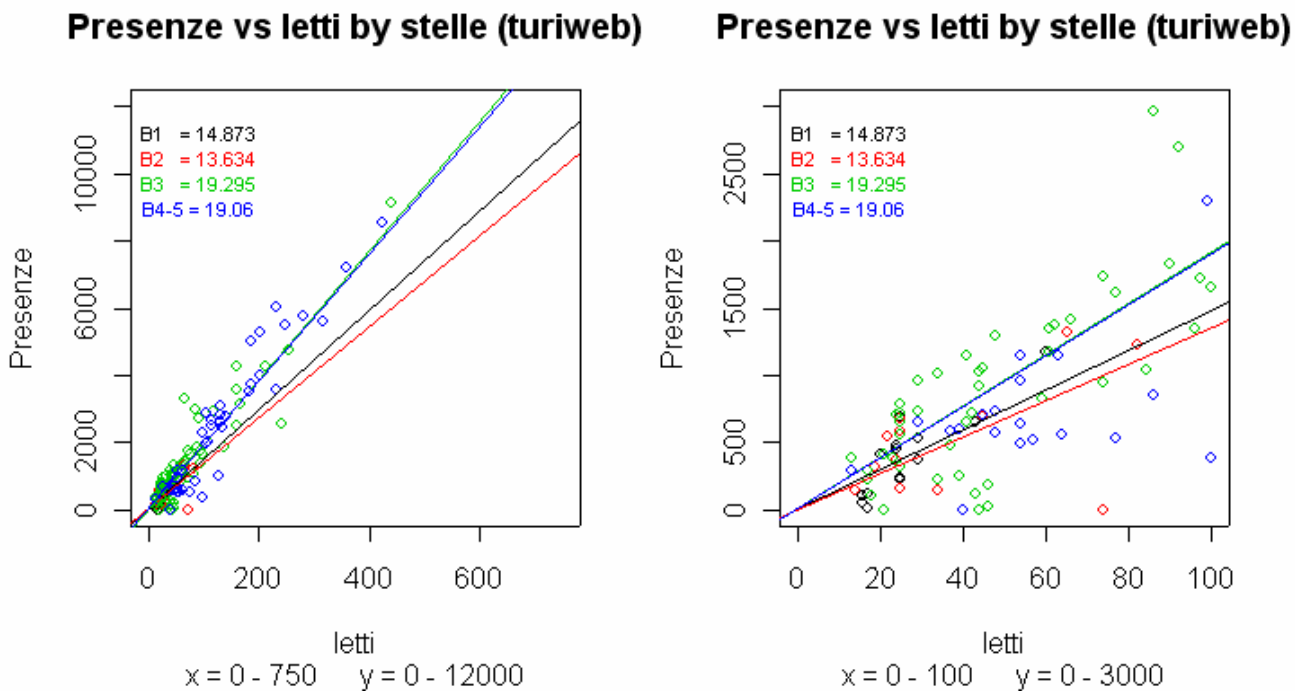
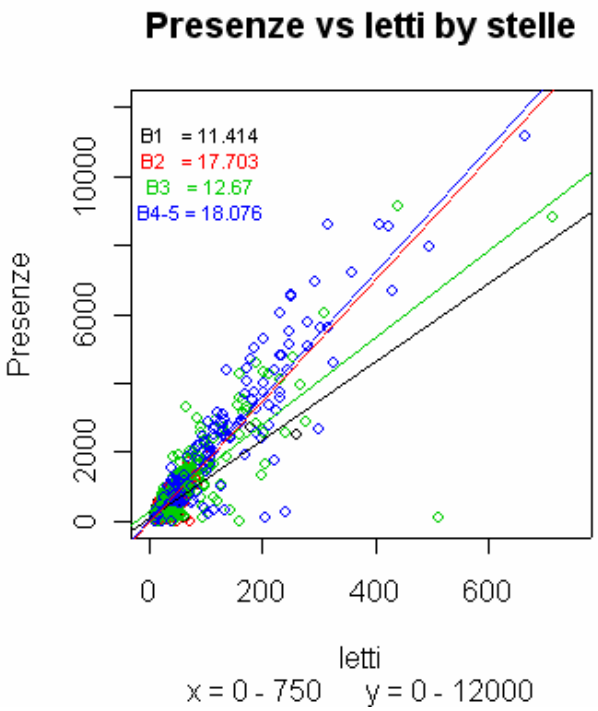
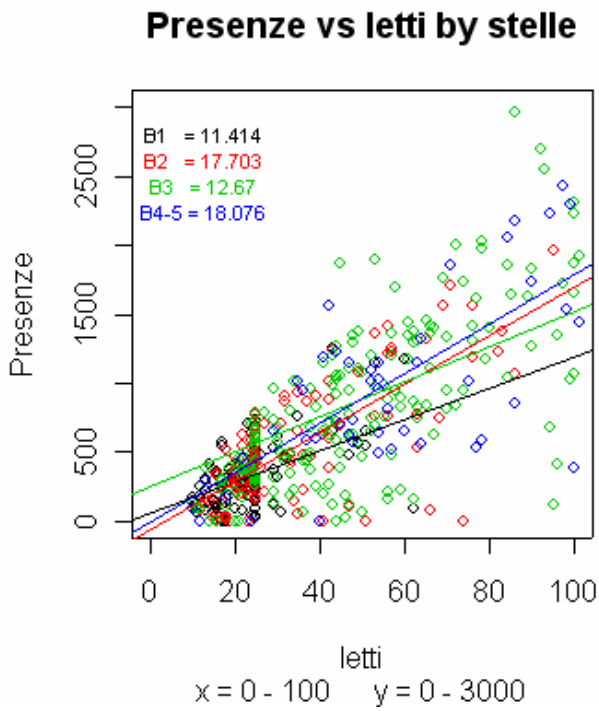


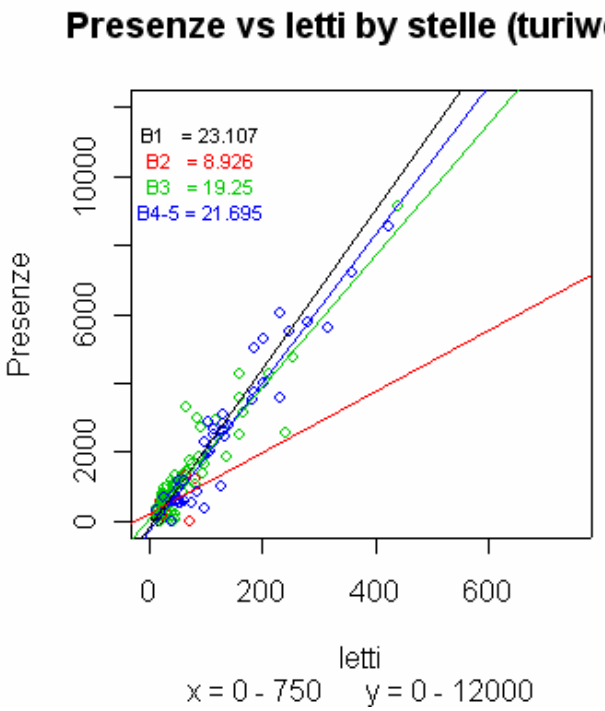
Grafico 5.2 (Scatter con rette di regressione: modello con intercetta omoschedastico)

Intera popolazione degli alberghi





Sottopopolazione degli alberghi del sistema turiweb



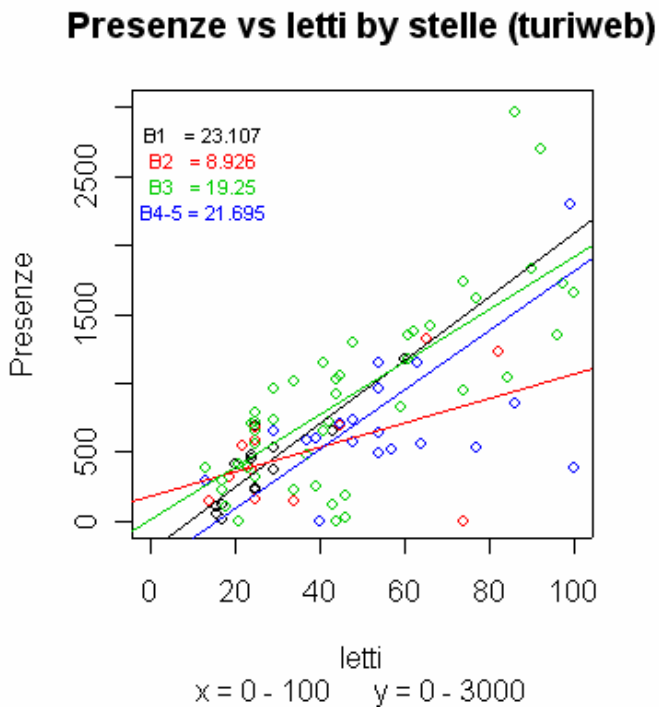
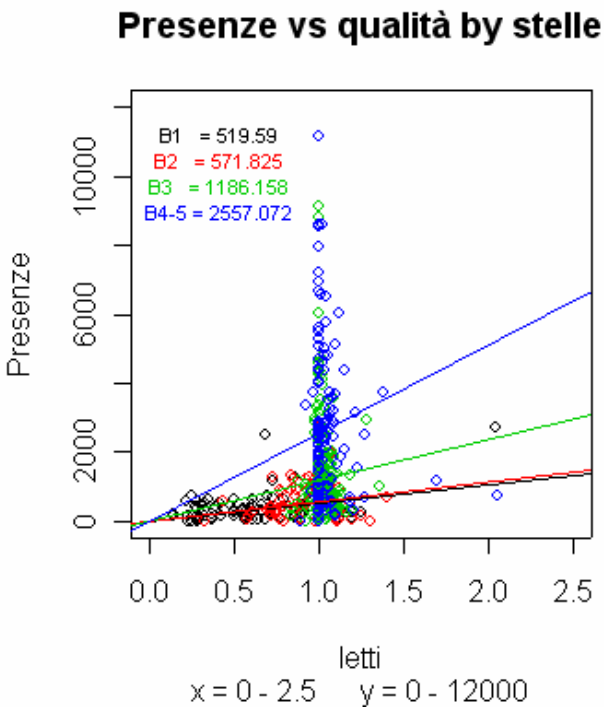
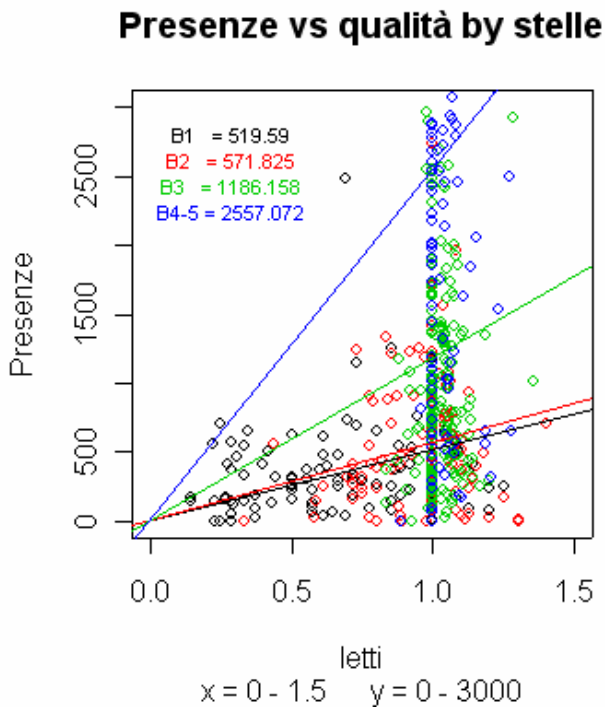


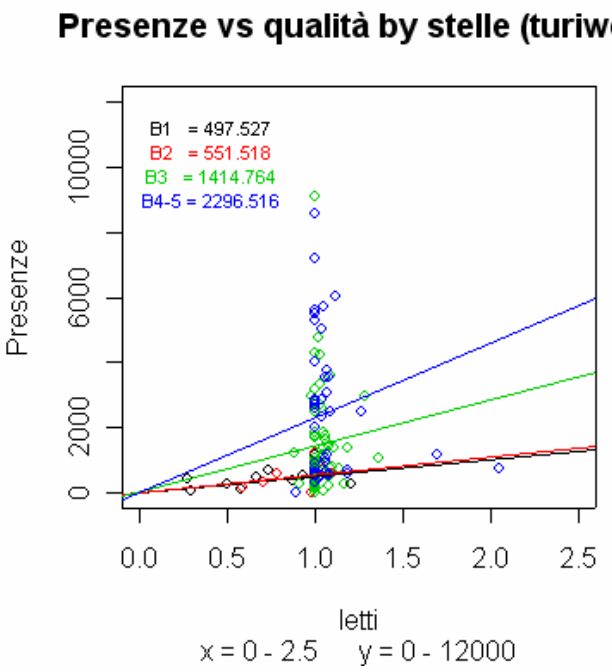
Grafico 5.3 (Scatter con rette di regressione, modello senza intercetta ed eteroschedasticità)

Intera popolazione





Sottopopolazione degli alberghi del sistema turiweb



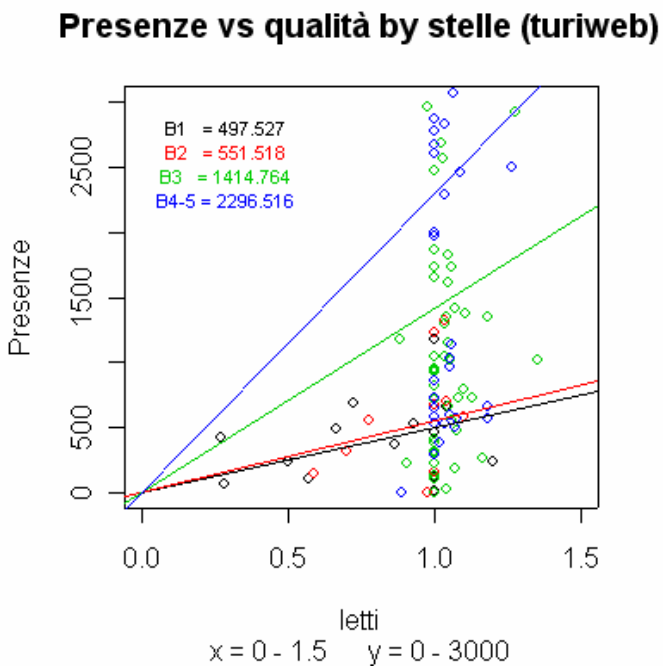
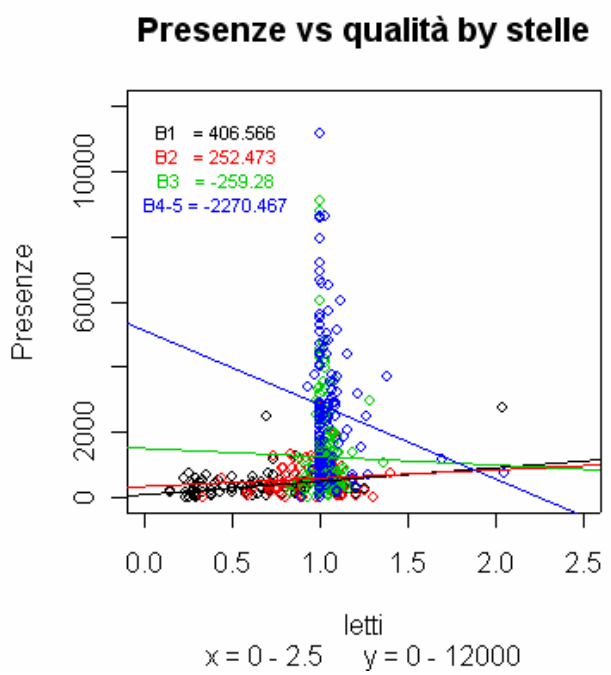
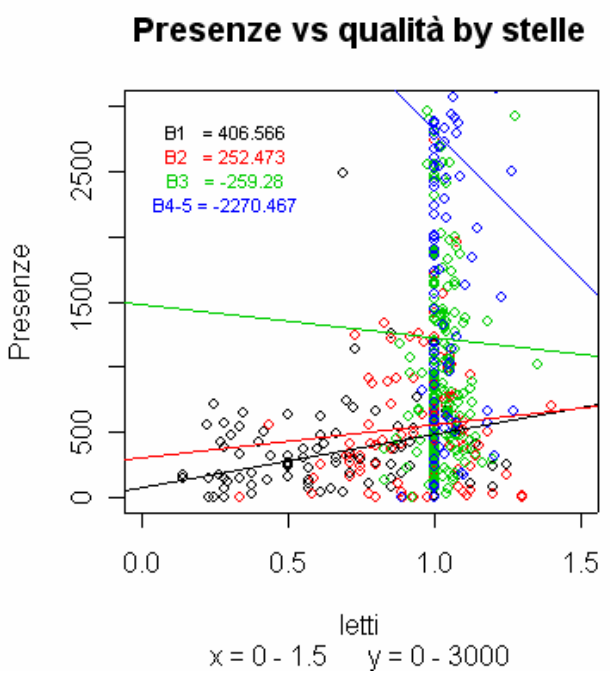


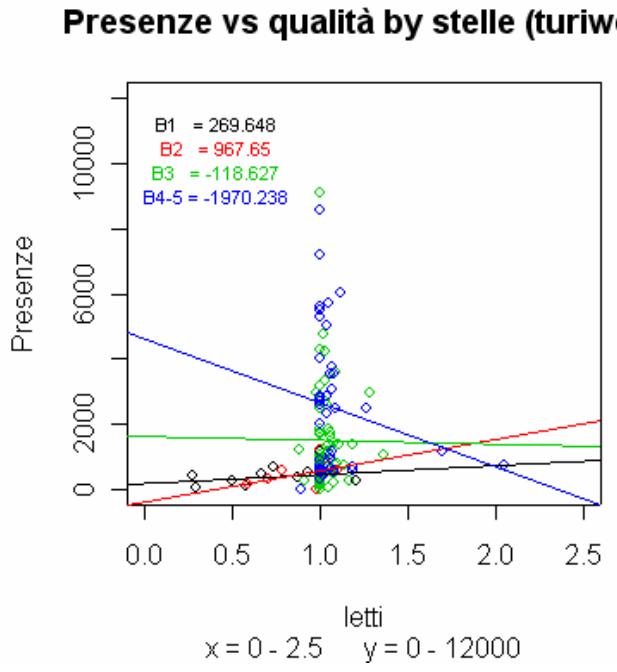
Grafico 5.4 (Scatter con rette di regressione: modello con intercetta omoschedastico)

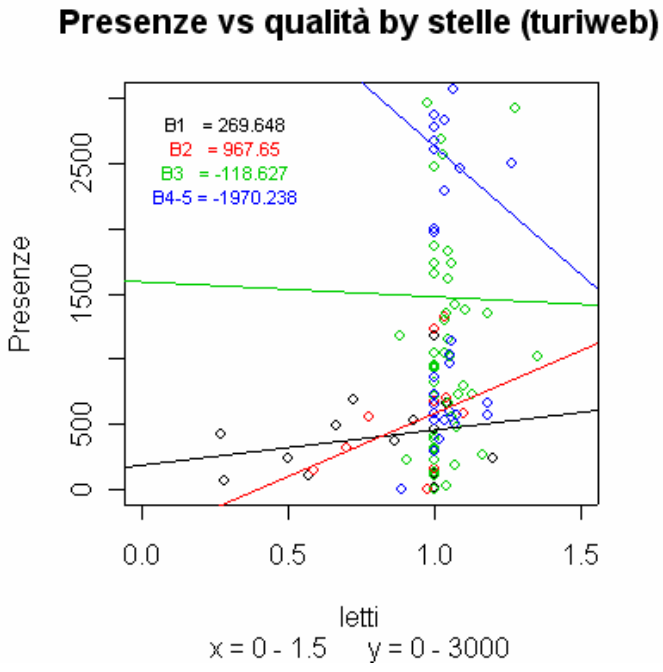
Intera popolazione





Sottopopolazione degli alberghi del sistema turiweb





I grafici presentati, evidenziano alcuni aspetti: la dimensione dell'albergo ha una notevole capacità esplicativa delle presenze, mentre la qualità poca o nulla, la maggioranza delle unità è concentrata intorno al valore di qualità 1 (numero di bagni pari al numero di camere). La relazione fra letti e presenze conferma quanto visto per l'intera popolazione degli esercizi ricettivi: gli alberghi del sistema turiweb realizzano mediamente più presenze a parità di posti letto di quanto avvenga per

l'intera popolazione degli alberghi. Questa maggior efficienza è però differenziata per stelle: nelle 2 stelle si osserva, almeno per il mese di ottobre 2006, una efficienza minore del gruppo turiweb.

La tabella seguente quantifica la correlazione fra presenze e letti, fra presenze e qualità per l'intera popolazione degli alberghi

Tabella 5.1

Correlazione fra presenze e letti, presenze e qualità; alberghi ottobre 2006

	1 stella	2 stelle	3 stelle	4 stelle	5 stelle	totale
cor(presenze, letti)	0,83	0,77	0,79	0,89	0,95	0,88
cor(presenze, qualità)	0,32	0,08	-0,01	-0,14	-0,09	0,22

La qualità potrebbe avere invece una capacità esplicativa del meccanismo di autoselezione. Per avere un'idea del suo effetto su tale meccanismo abbiamo considerato due modelli logistici:

M_1: $\text{logit}[\text{prob}(k \in \text{turiweb})|g] = a_g$ (5.1)

M_2: $\text{logit}[\text{prob}(k \in \text{turiweb})|(g, x_k)] = a_g + b \cdot x_k$ (5.2)

dove k indica il generico albergo, g = 1, 2, 3, 4 indica i quattro gruppi 1, 2, 3, 4-5 stelle, x è la variabile quantitativa qualità (rapporto bagni/camere).

Adattati i due modelli ai dati per i 24 mesi, riportiamo nella tabella 5.2 seguente la significatività dei coefficienti (1 se significativo, 0 altrimenti). Si nota che una volta presente la classificazione per stelle, la qualità risulta statisticamente significativa solo negli ultimi mesi. In sostanza è la variabile di classificazione stelle che ha la maggior capacità esplicativa del meccanismo di autoselezione; inoltre è il passaggio dalle 1-2 stelle alle 3 e 4-5 che segna un salto rilevante nella propensione a partecipare al sistema turiweb.

Tabella 5.2

Significatività dei coefficienti nei due modelli logistici (alfa = 0,1)
(1: coefficiente statisticamente significativo, 0: non significativo)

	Modello logistico 1				Modello logistico 2				
mesi	intercetta	stelle2	stelle3	stelle4-5	intercetta	qualità	stelle2	stelle3	stelle4-5
1	1	0	1	1	1	0	0	1	1
2	1	0	1	1	1	0	0	1	1
3	1	0	1	1	1	0	0	1	1
4	1	0	1	1	1	0	0	1	1
5	1	0	1	1	1	0	0	1	1
6	1	0	1	1	1	0	0	1	1
7	1	0	1	1	1	0	0	1	1
8	1	0	1	1	1	0	0	1	1
9	1	0	1	1	1	0	0	1	1
10	1	0	1	1	1	0	0	1	1
11	1	0	1	1	1	0	0	1	1
12	1	0	1	1	1	0	0	1	1
13	1	0	1	1	1	0	0	1	1
14	1	0	1	1	1	0	0	1	1
15	1	0	1	1	1	0	0	1	1
16	1	0	1	1	1	0	0	1	1
17	1	0	1	1	1	0	0	1	1
18	1	0	1	1	1	0	0	0	1
19	1	0	1	1	1	0	0	1	1
20	1	0	1	1	1	0	0	1	1
21	1	0	1	1	1	0	0	0	1
22	1	0	1	1	1	1	0	0	1
23	1	0	1	1	1	1	0	0	1
24	1	0	1	1	1	1	0	0	1

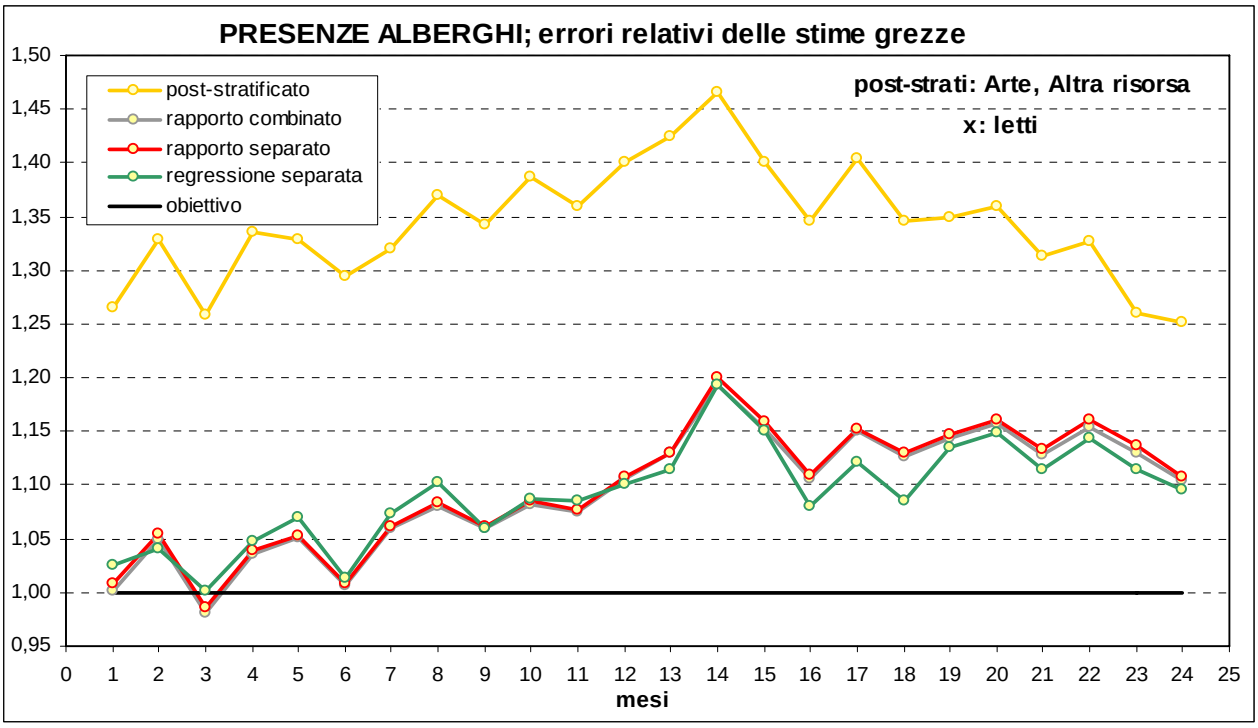
Va osservato che, ai fini dell'utilizzo di variabili ausiliarie nella formalizzazione degli stimatori, non si richiede che tali variabili abbiano una capacità esplicativa statisticamente significativa della y e del meccanismo di selezione: se ne hanno poca produrranno un limitato beneficio nel limitare la varianza e il bias.

L'esplorazione condotta indica che la variabile stelle ha una rilevante capacità esplicativa sia della presenze, sia del meccanismo di risposta, la dimensione ha rilevanza sulle presenze, la qualità, aggiunge poca informazione esplicativa, una volta che sono presenti come variabile esplicativa le stelle.

5.3. Stima delle presenze nel dominio alberghi

Effettuiamo una prima sperimentazione sul dominio alberghi applicando i quattro stimatori utilizzati per l'intera popolazione; in pratica ignoriamo le stelle e usiamo i due post-strati in cui si articola questo dominio (Località d'Arte, Altre località). Il grafico 5.5 seguente riporta gli errori relativi delle stime preliminari (grezze) del totale delle presenze mensili.

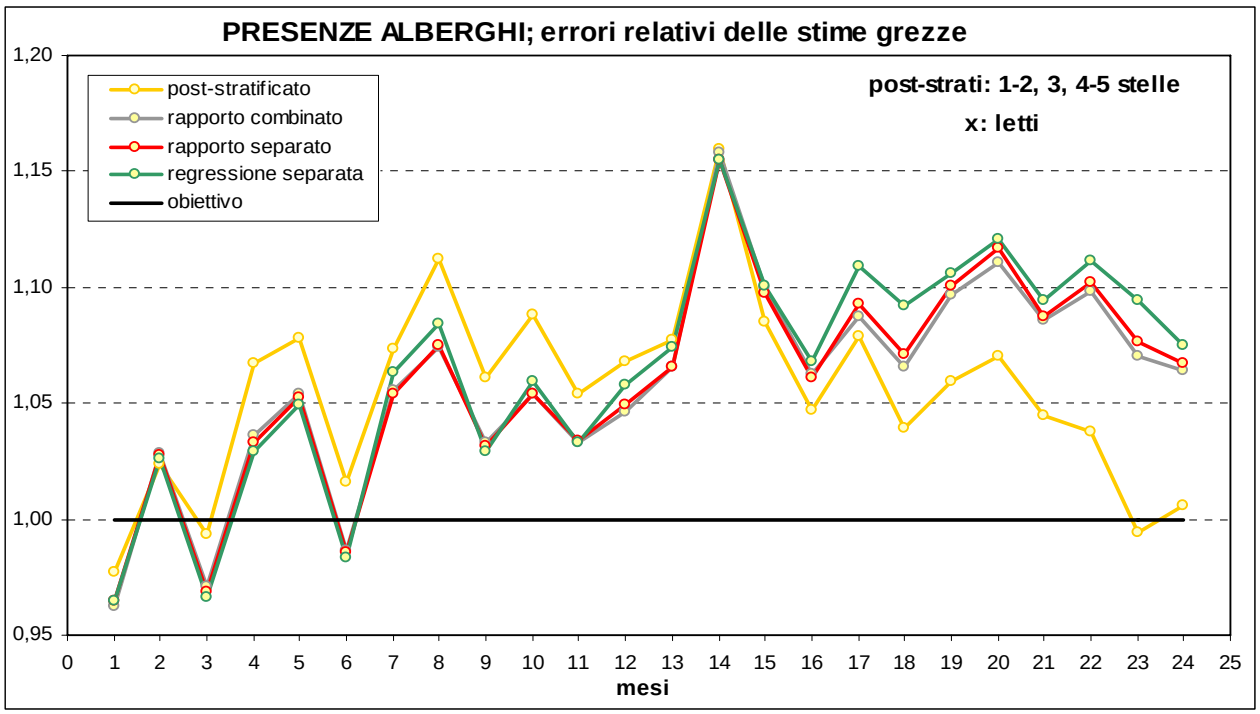
Grafico 5.5



E' evidente la riduzione del bias che si ottiene utilizzando la variabile ausiliaria letti che entra in gioco negli stimatori rapporto combinato, rapporto separato e regressione separata. Si nota infine come le stime ottenute con i tre ultimi metodi differiscono di poco.

Il grafico 5.6 successivo mostra gli errori relativi di stima dei quattro stimatori, allorché si utilizzi, invece della post-stratificazione per risorsa turistica (Località d'Arte, Altre località), la post-stratificazione per categoria (1, 2 stelle, 3 stelle, 4-5 stelle). Una volta presente la variabile ausiliaria categoria dell'albergo, la dimensione ha un effetto limitato nella riduzione del bias; addirittura nei mesi del secondo anno lo stimatore post-stratificato, che ignora la dimensione, si comporta meglio degli altri tre che invece la utilizzano.

Grafico 5.6



Abbiamo sperimentato anche due stimatori di regressione separata del totale delle presenze mensili. Indicando con P il totale delle presenze, formalmente essi sono espressi da:

$$\hat{P}_{reg1} = \sum_{g=1}^{4} N_g \left[\overline{P}_{Vg} + \left(\overline{x}_{1Ug} - \overline{x}_{1Vg} \right) \hat{B}_g \right]$$
$$\hat{P}_{reg2} = \sum_{g=1}^{4} N_g \left[\overline{P}_{Vg} + \left(\overline{x}_{1Ug} - \overline{x}_{1Vg} \right) \hat{B}_{1g} + \left(\overline{x}_{2Ug} - \overline{x}_{2Vg} \right) \hat{B}_{2g} \right] \qquad (5.3)$$

dove:

g = 1, 2, 3 ,4 sono gli indici dei quattro post-strati definiti dalle stelle;

x_1 è la variabile ausiliaria letti, x_2 la variabile ausiliaria qualità = bagni/camere;

Ug è il g-esimo post-strato;

Vg = Ug ∩V è l'insieme degli alberghi del gruppo Ug appartenenti al sistema turiweb;

Ng è la numerosità di Ug;

$\overline{P}_{Vg}$ $\overline{x}_{1Ug}$ $\overline{x}_{1Vg}$ $\overline{x}_{2Ug}$ $\overline{x}_{2Vg}$ sono ripettivamente: la media in Vg delle presenze, la media in Ug e Vg dei letti (pedice 1), la media in Ug e Vg dell'indice di qualità (pedice 2).

$\hat{B}_g$ è la stima dei minimi quadrati per g = 1, 2, 3, 4 del parametro slope, ottenuta sulle osservazioni del sottoinsieme Vg, del modello

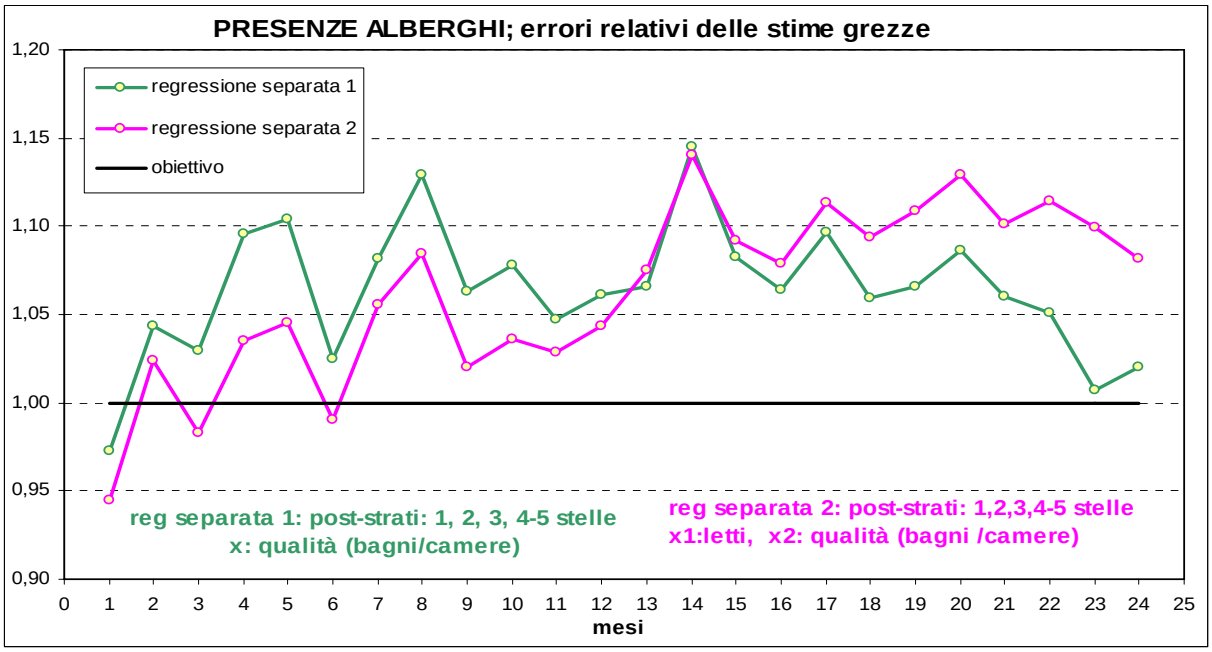
$$E(P_k) = A_g + B_g x_k \quad \forall k \in U_g$$
$$V(P_k) = \sigma_g^2 \quad \forall k \in U_g \qquad (5.4)$$

$\hat{B}_{1g}, \hat{B}_{2g}$ sono le stime dei minimi quadrati per g = 1, 2, 3, 4 dei due parametri slope, ottenute sulle osservazioni del sottoinsieme Vg, del modello

$$E(P_k) = A_g + B_{1g}x_{1k} + B_{21g}x_{2k} \quad \forall k \in U_g$$
$$V(P_k) = \sigma_g^2 \quad \forall k \in U_g \qquad (5.5)$$

Il grafico seguente riporta gli errori relativi delle stime preliminari per i 24 mesi ottenute con i due stimatori di regressione separata sopra indicati.

Grafico 5.7



Riportiamo nella tabella 5.3 alcune sintesi statistiche delle distribuzioni dei 24 errori relativi di stima ottenuti con i 10 metodi applicati

Il primo gruppo di errori di stima, sotto l'intestazione post-stratificazione: ARTE, ALTRA RISORSA, è ottenuto con gli stessi stimatori usati per l'intera popolazione degli esercizi ricettivi; il secondo gruppo sotto l'intestazione post-stratificazione: 1, 2, 3, 4-5 STELLE adotta la post-stratificazione per stelle, aggregando le 4 e 5 stelle poiché queste ultime sono in numero molto limitato. Infine sotto le due colonne intestate con regressione separata 1 e regressione separata 2 si sono applicati i due stimatori di regressione separata del totale delle presenze (5.3).

Tabella 5.3

Sintesi statistiche dei 24 errori relativi di stima ottenuti con i 10 metodi indicati

	post-stratificazione: ARTE, ALTRA RISORSA				post-stratificazione: 1, 2, 3, 4-5 STELLE					
Sintesi	post-stratificato	rapporto combinato	rapporto separato	regressione separata	post-stratificato	rapporto combinato	rapporto separato	regressione separata	regressione separata 1	regressione separata 2
minimo	1,2524	0,9814	0,9853	1,0017	0,9771	0,9622	0,9644	0,9650	0,9729	0,9449
massimo	1,4653	1,1929	1,1993	1,1940	1,1598	1,1577	1,1549	1,1547	1,1447	1,1401
range	0,2129	0,2115	0,2140	0,1923	0,1827	0,1955	0,1905	0,1897	**0,1718**	0,1952
media	1,3431	1,0941	1,0978	1,0918	1,0547	1,0584	1,0593	1,0644	1,0639	1,0633
bias	0,3431	0,0941	0,0978	0,0918	**0,0547**	0,0584	0,0593	0,0644	0,0639	0,0633
varianza	0,0028	0,0030	0,0030	**0,0021**	**0,0016**	**0,0018**	**0,0019**	0,0022	**0,0014**	**0,0023**
mse	0,1206	0,0118	0,0126	0,0105	**0,0046**	**0,0053**	**0,0054**	0,0064	**0,0055**	0,0063

In neretto si evidenziano i valori più piccoli, in particolare per la varianza.

La post-stratificazione per stelle è più efficiente di quella per risorsa turistica; si nota inoltre come l'ulteriore informazione ausiliaria fornita dai letti e/o dalla qualità sembra non avere un effetto migliorativo sul bias e sulla varianza rispetto a quanto ottenuto con la semplice post-stratificazione per stelle.

Il bias è positivo per tutti i 10 metodi di stima, sebbene per quelli che usano la post-stratificazione per stelle più ridotto. Intendendo applicare la correzione di Rao per aggiustare le stime grezze, i candidati per questo successivo trattamento sono quegli stimatori che forniscono errori relativi di stima con varianza più piccola. Sulla base dei risultati della tabella precedente applichiamo la correzione di Rao ai quattro stimatori:

- regressione separata con post-stratificazione: ARTE, ALTRA RISORSA, X=LETTI
- post-stratificato con post-stratificazione: 1, 2, 3, 4-5 STELLE
- rapporto separato con post-stratificazione: 1, 2, 3, 4-5 STELLE, X = LETTI
- regressione separata 1 con post-stratificazione: 1, 2, 3, 4-5 STELLE, X = QUALITA'

5.3.1. Osservazione sulla scelta degli stimatori

Va notato che la scelta degli stimatori, cui far seguire il procedimento di aggiustamento di Rao, è dettata prevalentemente dalle evidenze empiriche: dagli errori relativi di stima che possiamo osservare nei 24 mesi, dal gennaio 2005 al dicembre 2006, in particolare dalla varianza della distribuzione nei 24 mesi di tali errori di stima.

Non possiamo affermare che "buone" stime preliminari osservate nei 24 mesi siano tali anche nei mesi futuri: la popolazione evolve, seppur lentamente, il sottoinsieme delle strutture del sistema turiweb aumenta in dimensione. Infine, osservando i grafici delle stime grezze, si nota una divergenza nel comportamento degli stimatori considerati nei due anni: alcuni presentano un errore relativo di stima crescente nel primo anno poi stabile nel secondo; altri un errore relativo crescente nel primo anno e decrescente nel secondo. Da cosa questi andamenti dipendano è difficile dire.

In una applicazione concreta delle tecniche che stiamo provando, vale la pena di monitorare più metodi: la gerarchia che in qualche modo otteniamo con le attuali informazioni, potrebbe cambiare nel tempo: alcuni stimatori che ci saremmo aspettati migliori, in quanto usano maggior informazione ausiliaria, non risultano tali con queste specifiche 24 popolazioni e 24 campioni autoselezionati; ad esempio lo stimatore post-stratificato con post-stratificazione per stelle fornisce errori relativi di stima migliori del rapporto separato e della regressione separata con uguale post-stratificazione. Ora, questi ultimi due sono più efficienti del primo in presenza di correlazione rilevante[1] fra letti e presenze quando l'efficienza è riferita alla varianza rispetto al disegno; ciò non significa che per uno specifico campione, sia esso ottenuto secondo un disegno noto, sia esso frutto di un meccanismo ignoto di autoselezione, l'errore di stima sia più piccolo con lo stimatore rapporto separato che con lo stimatore post-stratificato. La maggior efficienza è una proprietà che vale nell'insieme dei possibili campioni selezionati secondo un disegno noto.

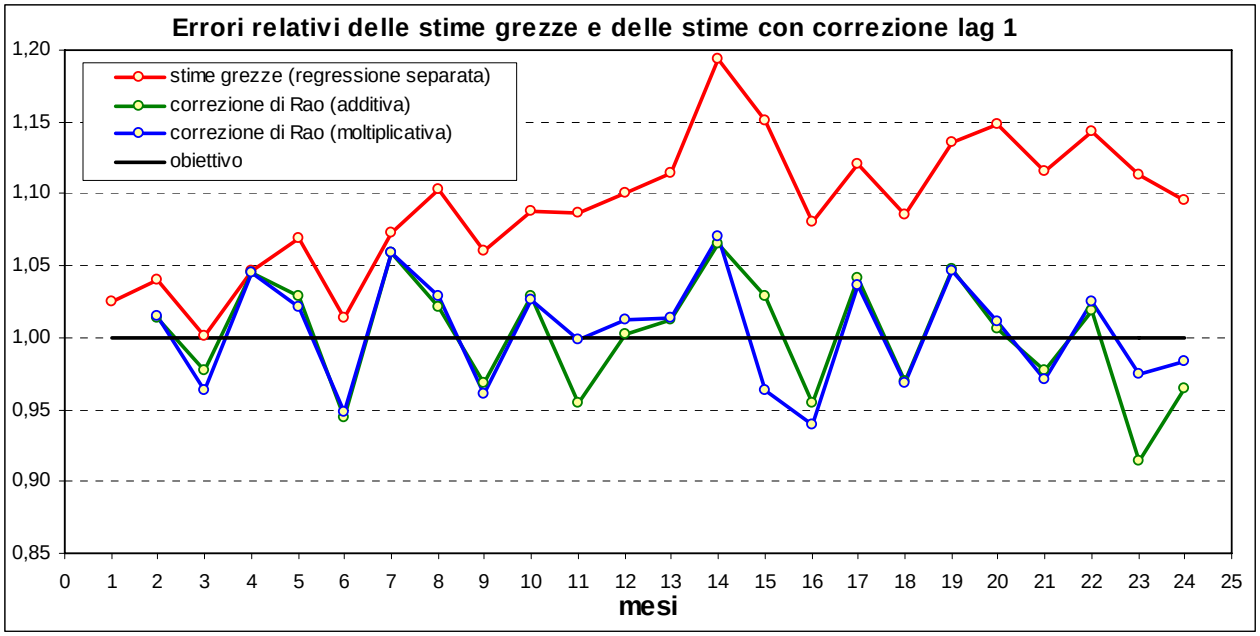
Nel nostro caso, essendo i 24 campioni di volontari molto correlati (il campione del mese m+1 è ottenuto dal campione del mese m con l'aggiunta di alcune unità), il miglior comportamento delle stime con il post-stratificato deriva dal quel "quasi unico" campione che si è manifestato. Non conoscendo il meccanismo di selezione, né disponendo di in modello per l'osservazione, la generalizzazione delle conclusioni a un popolazione più ampia quale quella che comprende i mesi futuri, risulta ardua.

[1] Abbiamo verificato per tutti i 24 mesi che la condizione per la maggior efficienza dello stimatore rapporto : $\text{cor}(\text{letti}, \text{presenze}) > \frac{\text{cv}(\text{letti})}{2 \cdot \text{cv}(\text{presenze})}$ vale in tutti i post-strati definiti dalla categoria (stelle).

5.4. Applicazione della correzione di Rao alle stime con regressione separata e post-stratificazione: ARTE, ALTRA RISORSA

Il grafico 5.8 seguente riporta gli errori relativi delle stime grezze ottenute con la regressione separata e post- stratificazione: Arte, Altra risorsa, nonché gli errori relativi delle stime anticipate ottenute con la correzione di Rao al lag 1.

Grafico 5.8



Nella tabella 5.4 seguente sono riportate le sintesi statistiche delle distribuzioni degli errori relativi delle stime anticipate: dal lag 0 (stime grezze) al lag 12.

Tabella 5.4

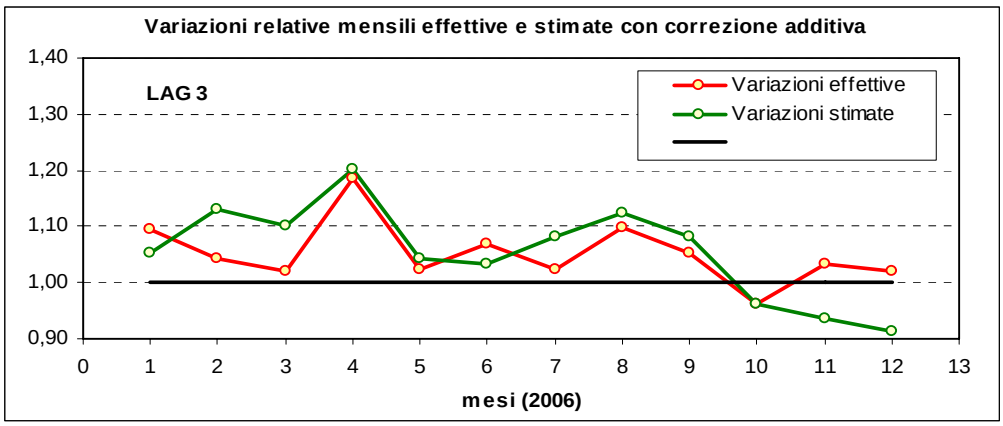
PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

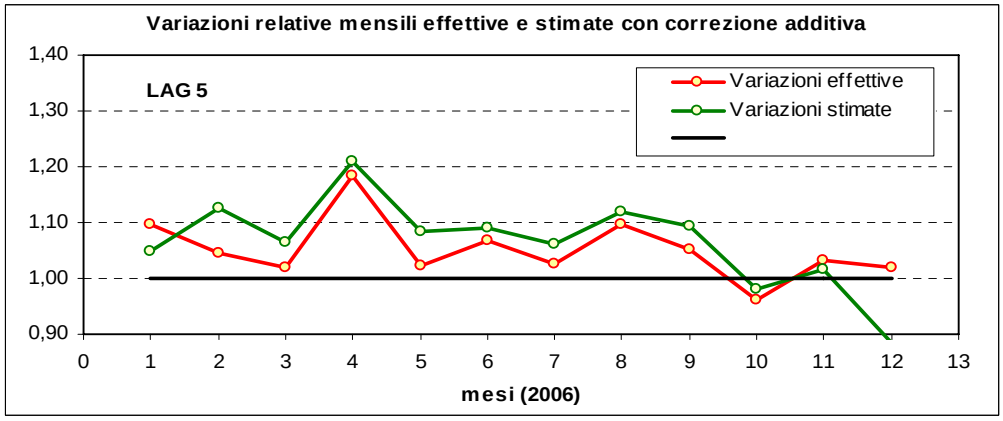
CORREZIONE ADDITIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	1,0017	0,9143	0,8660	0,8966	0,8577	0,8679	0,9399	0,8955	0,9346	0,8894	0,9297	0,9856	0,9975
massimo	1,1940	1,0653	1,0881	1,0818	1,0805	1,1016	1,0807	1,0920	1,1683	1,1346	1,1182	1,1914	1,1556
range	0,1923	0,1510	0,2221	0,1852	0,2228	0,2338	0,1408	0,1965	0,2337	0,2452	0,1886	0,2058	0,1580
media	1,0918	1,0020	1,0037	1,0096	1,0126	1,0206	1,0299	1,0342	1,0417	1,0445	1,0549	1,0653	1,0683
bias	0,0918	0,0020	0,0037	0,0096	0,0126	0,0206	0,0299	0,0342	0,0417	0,0445	0,0549	0,0653	0,0683
varianza	0,0021	0,0016	0,0024	0,0023	0,0031	0,0025	0,0019	0,0022	0,0029	0,0031	0,0021	0,0027	0,0019
mse	0,0105	0,0016	0,0024	0,0024	0,0033	0,0029	0,0028	0,0034	0,0047	0,0051	0,0051	0,0069	0,0066
CORREZIONE MOLTIPLICATIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	1,0017	0,9392	0,9051	0,9385	0,9089	0,9508	0,9617	0,9339	0,9571	0,9327	0,9177	0,9828	0,9961
massimo	1,1940	1,0709	1,0884	1,0994	1,0977	1,1262	1,0853	1,1129	1,1780	1,1352	1,1405	1,1920	1,1487
range	0,1923	0,1317	0,1833	0,1609	0,1888	0,1755	0,1236	0,1790	0,2209	0,2025	0,2228	0,2092	0,1526
media	1,0918	1,0036	1,0070	1,0135	1,0177	1,0222	1,0297	1,0324	1,0357	1,0395	1,0464	1,0576	1,0632
bias	0,0918	0,0036	0,0070	0,0135	0,0177	0,0222	0,0297	0,0324	0,0357	0,0395	0,0464	0,0576	0,0632
varianza	0,0021	0,0013	0,0017	0,0014	0,0022	0,0018	0,0014	0,0019	0,0026	0,0028	0,0028	0,0029	0,0019
mse	0,0105	0,0013	0,0018	0,0016	0,0025	0,0023	0,0023	0,0029	0,0039	0,0043	0,0049	0,0062	0,0059

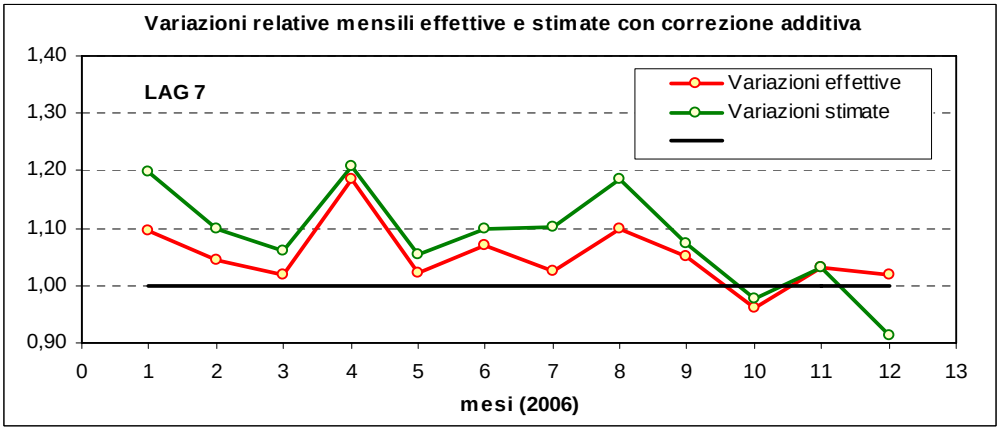
I grafici 5.9 e 5.10 successivi confrontano le variazioni relative mensili effettive con quelle stimate, dopo l'applicazione della correzione di Rao additiva e moltiplicativa ai lag 3, 5, 7, 12.

La linea nera indica la situazione di nessuna variazione relativa mensile: $P_1/P_{1\text{-}12} = 1$.

Grafico 5.9 (Regressione separata - Correzione additiva)







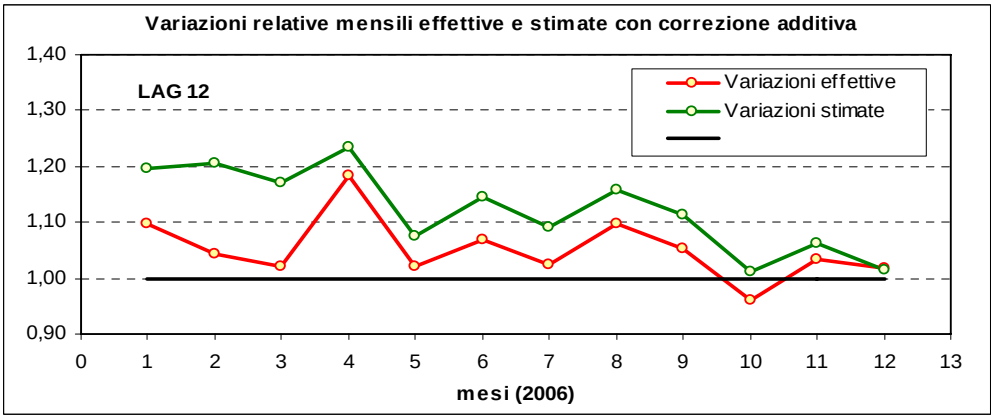
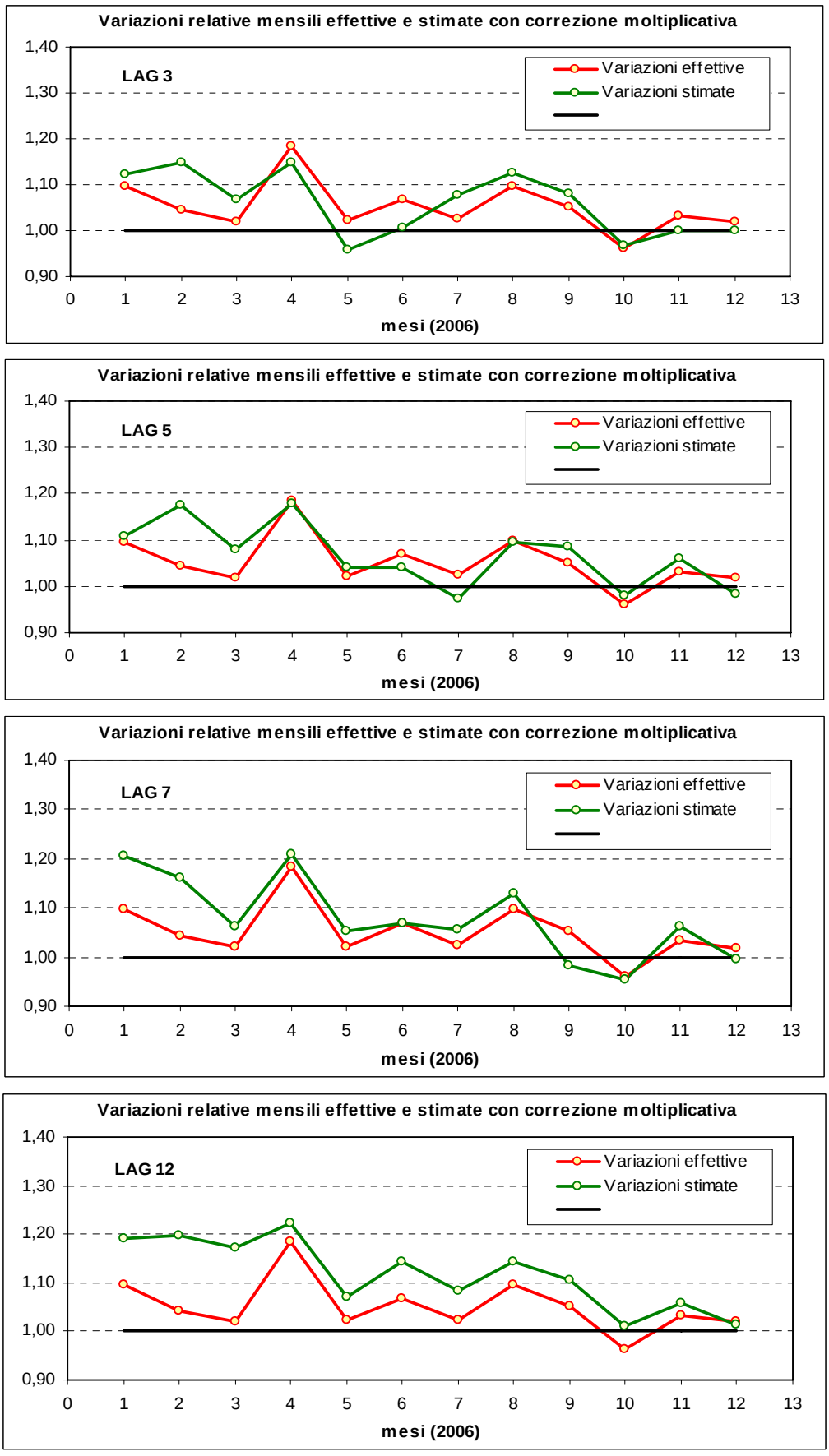


Grafico 5.10 (Regressione separata - Correzione moltiplicativa)



Nella tabella 5.5 seguente riportiamo una misura della bontà delle stime anticipate delle variazione relative mensili. Abbiamo qui calcolato la media e la varianza delle 12 differenze fra variazioni relative stimate ed effettive. Vedremo alla fine del capitolo un riepilogo dei quattro procedimenti di stima che ne evidenzia la bontà, congiuntamente alla correzione di Rao, nel fornire stime anticipate delle variazioni relative mensili.

Tabella 5.5

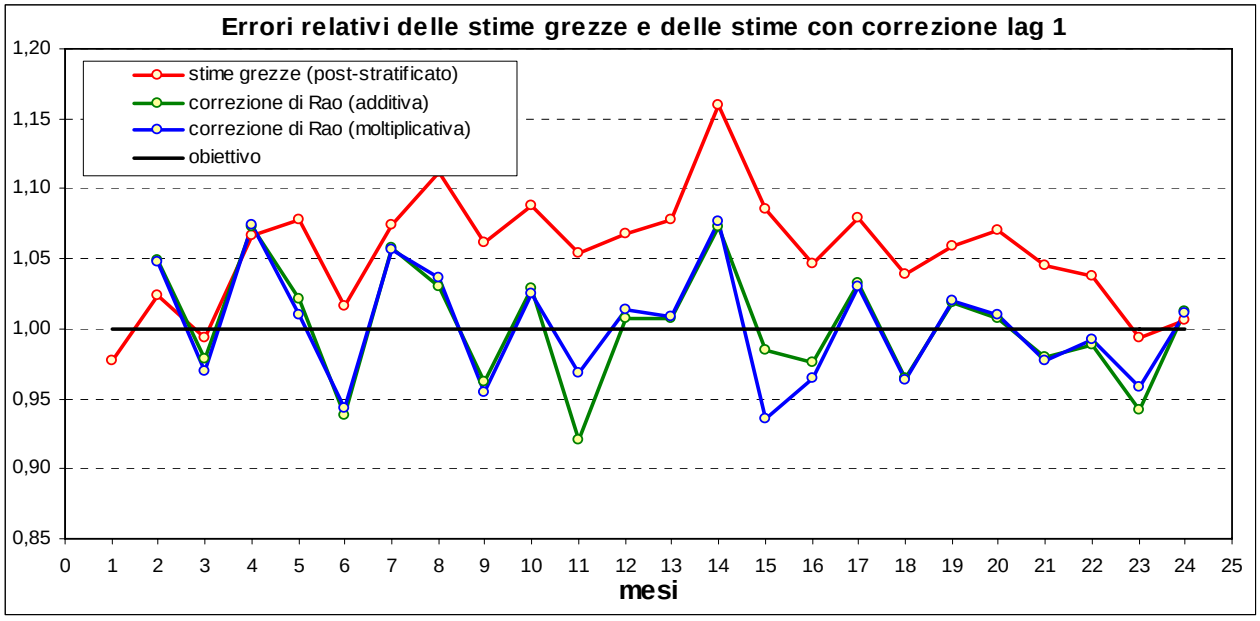
Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi	Variazione	Differenze fra variazioni stimate ed effettive ai lag												
2006	effettiva	lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Correzione additiva														
Media	**1,0517**	**0,1306**	**-0,0005**	**0,0009**	**0,0033**	**0,0074**	**0,0133**	**0,0192**	**0,0317**	**0,0375**	**0,0382**	**0,0561**	**0,0669**	**0,0717**
Varianza		**0,0009**	**0,0020**	**0,0033**	**0,0036**	**0,0039**	**0,0030**	**0,0021**	**0,0026**	**0,0040**	**0,0037**	**0,0027**	**0,0031**	**0,0020**
Correzione moltiplicativa														
Media	**1,0517**	**0,1306**	**-0,0005**	**0,0022**	**0,0069**	**0,0117**	**0,0149**	**0,0191**	**0,0269**	**0,0310**	**0,0341**	**0,0474**	**0,0588**	**0,0663**
Varianza		**0,0009**	**0,0016**	**0,0025**	**0,0023**	**0,0030**	**0,0021**	**0,0017**	**0,0024**	**0,0035**	**0,0033**	**0,0036**	**0,0033**	**0,0021**

5.5. Applicazione della correzione di Rao alle stime con stimatore post-stratificato, post-stratificazione: 1, 2, 3, 4-5 STELLE

Il grafico seguente riporta gli errori relativi delle stime preliminari (grezze) ottenute con lo stimatore post-stratificato con post- stratificazione: Arte, Altra risorsa, nonché gli errori relativi delle stime anticipate dei totali mensili ottenute con la correzione di Rao al lag 1.

Grafico 5.11



La tabella seguente riporta, al solito, le sintesi statistiche delle distribuzioni degli errori relativi delle 24 stime anticipate dei totali mensili: dal lag 0 (stime grezze) al lag 12.

Tabella 5.6

PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

CORREZIONE ADDITIVA

LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,9771	0,9211	0,9174	0,8961	0,8929	0,9053	0,8806	0,8751	0,8929	0,8752	0,8688	0,9315	0,9387
massimo	1,1598	1,0734	1,0943	1,0898	1,0904	1,1183	1,0988	1,1260	1,1291	1,1005	1,0864	1,1700	1,1368
range	0,1827	0,1524	0,1769	0,1937	0,1975	0,2129	0,2182	0,2509	0,2361	0,2253	0,2176	0,2385	0,1981
media	1,0547	1,0023	0,9992	0,9986	0,9952	0,9971	0,9990	0,9982	1,0002	0,9981	1,0010	1,0101	1,0093
bias	0,0547	0,0023	-0,0008	-0,0014	-0,0048	-0,0029	-0,0010	-0,0018	0,0002	-0,0019	0,0010	0,0101	0,0093
varianza	0,0016	0,0017	0,0022	0,0028	0,0031	0,0027	0,0030	0,0035	0,0038	0,0036	0,0032	0,0044	0,0040
mse	0,0046	0,0017	0,0022	0,0028	0,0031	0,0027	0,0030	0,0035	0,0038	0,0036	0,0032	0,0045	0,0041

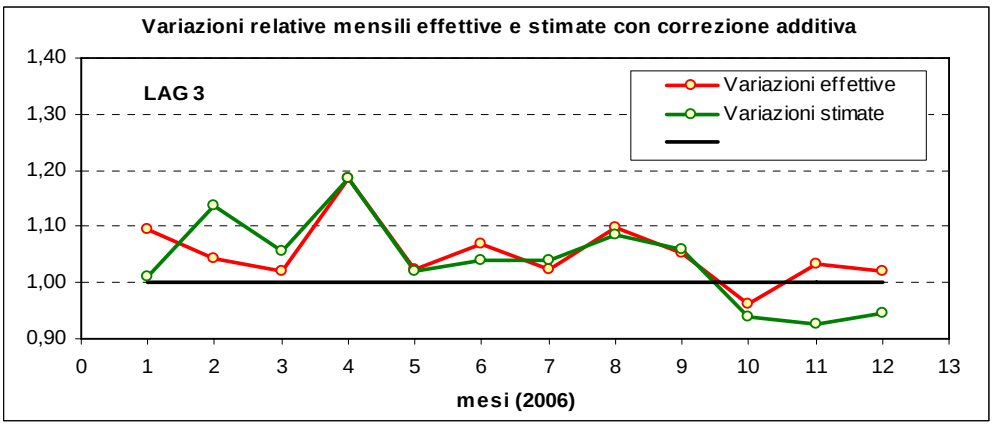
CORREZIONE MOLTIPLICATIVA

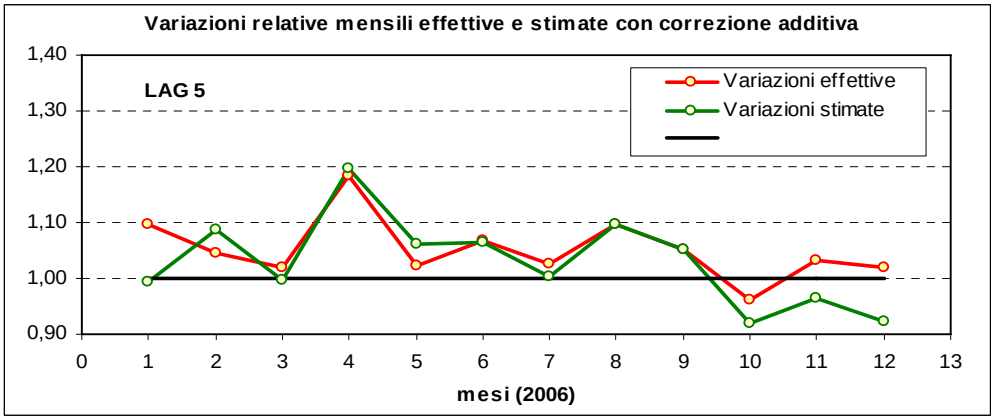
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,9771	0,9354	0,9027	0,9287	0,8961	0,9135	0,9215	0,9011	0,8945	0,8570	0,8670	0,9304	0,9414
massimo	1,1598	1,0766	1,0946	1,1007	1,1034	1,1196	1,0988	1,1384	1,1413	1,1140	1,0869	1,1674	1,1326
range	0,1827	0,1412	0,1919	0,1720	0,2073	0,2061	0,1773	0,2373	0,2468	0,2571	0,2198	0,2369	0,1912
media	1,0547	1,0021	1,0010	1,0030	1,0024	1,0019	1,0043	1,0031	1,0014	1,0006	1,0002	1,0076	1,0086
bias	0,0547	0,0021	0,0010	0,0030	0,0024	0,0019	0,0043	0,0031	0,0014	0,0006	0,0002	0,0076	0,0086
varianza	0,0016	0,0016	0,0020	0,0020	0,0026	0,0023	0,0025	0,0035	0,0038	0,0040	0,0038	0,0045	0,0039
mse	0,0046	0,0016	0,0020	0,0020	0,0026	0,0023	0,0025	0,0035	0,0038	0,0040	0,0038	0,0045	0,0040

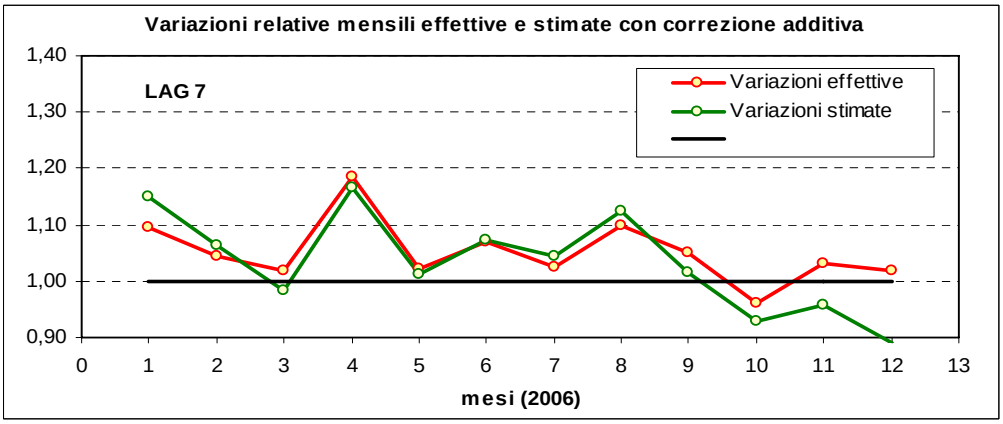
I grafici seguenti 5.12 e 5.13 confrontano le variazioni relative mensili effettive con quelle stimate, dopo l'applicazione della correzione di Rao additiva e moltiplicativa ai lag 3, 5, 7, 12.

La linea nera indica la situazione di nessuna variazione relativa mensile: $P_1/P_{1-12} = 1$.

Grafico 5.12 (Post-stratificato - Correzione additiva)







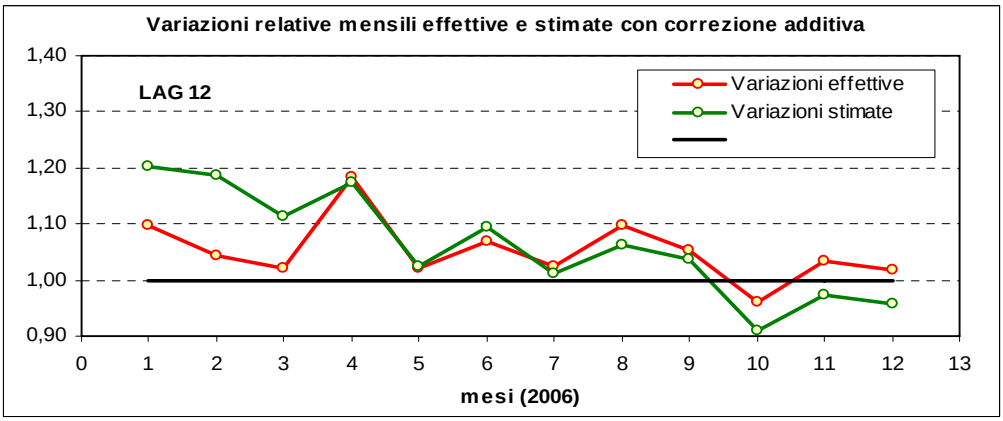
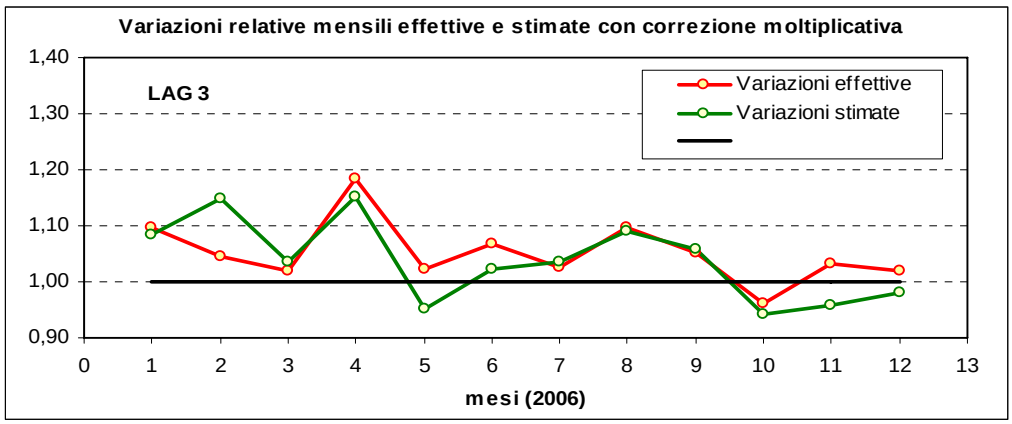
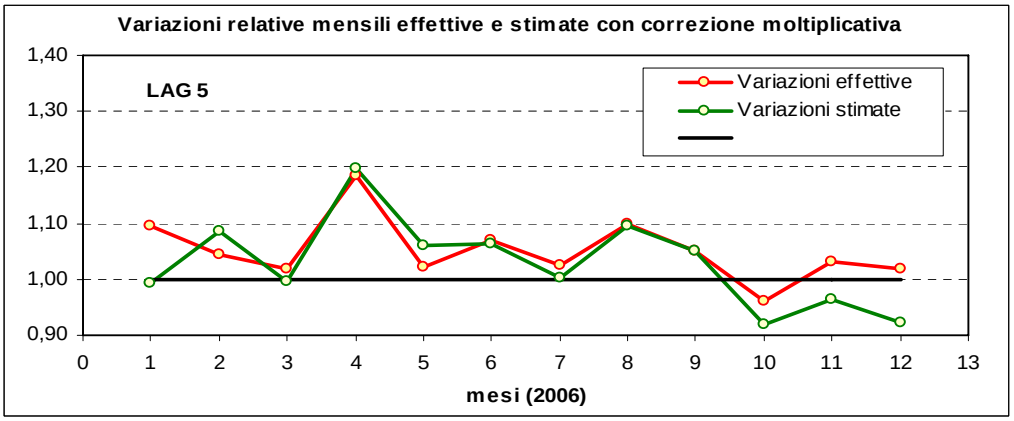
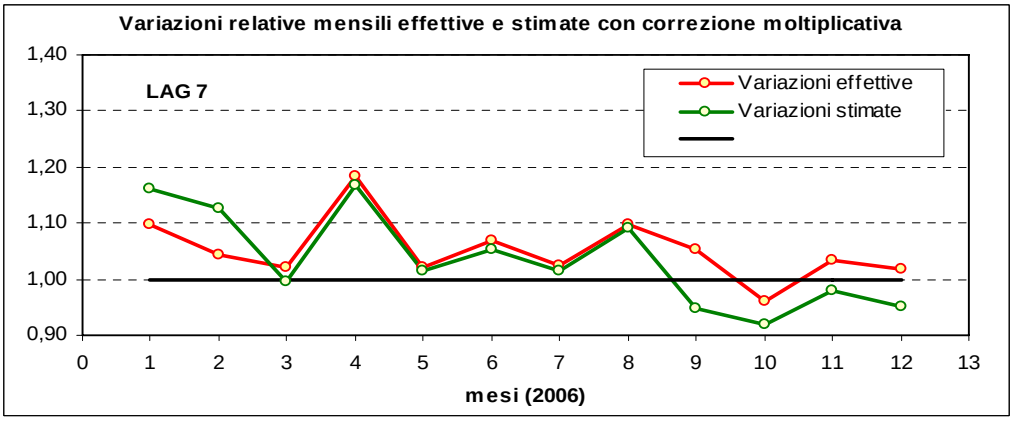
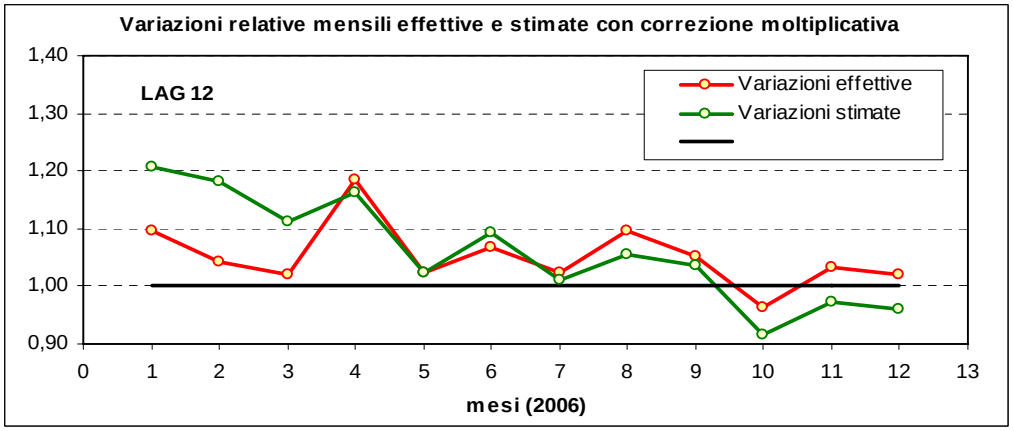


Grafico 5.13 (Post-stratificato - Correzione moltiplicativa)









Nella tabella 5.7 seguente riportiamo una misura della bontà delle stime anticipate delle variazione relative mensili. Abbiamo qui calcolato la media e la varianza delle 12 differenze fra variazioni relative stimate ed effettive.

Tabella 5.7

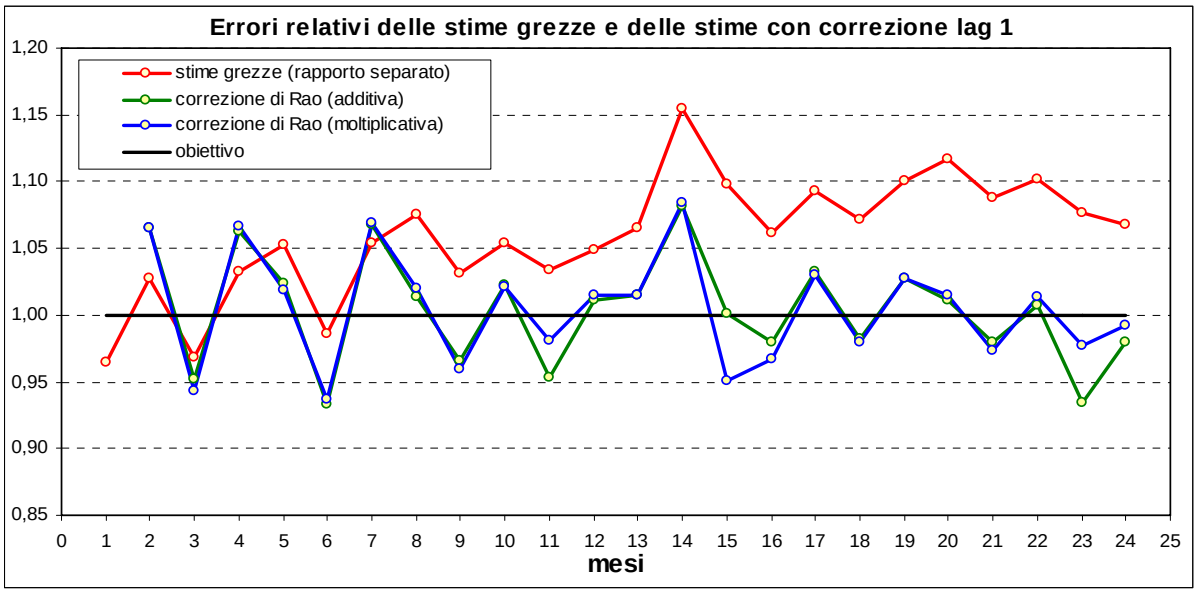
Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi 2006	Variazione effettiva	Differenze fra variazioni stimate ed effettive ai lag												
		lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Correzione additiva														
Media	**1,0517**	**0,0614**	**-0,0013**	**-0,0060**	**-0,0152**	**-0,0192**	**-0,0221**	**-0,0253**	**-0,0175**	**-0,0162**	**-0,0209**	**-0,0084**	**0,0036**	**0,0103**
arianza		**0,0018**	**0,0012**	**0,0018**	**0,0028**	**0,0020**	**0,0021**	**0,0018**	**0,0022**	**0,0039**	**0,0028**	**0,0033**	**0,0046**	**0,0043**
Correzione moltiplicativa														
Media	**1,0517**	**0,0614**	**-0,0048**	**-0,0099**	**-0,0136**	**-0,0148**	**-0,0179**	**-0,0202**	**-0,0165**	**-0,0163**	**-0,0179**	**-0,0093**	**0,0006**	**0,0094**
arianza		**0,0018**	**0,0015**	**0,0022**	**0,0021**	**0,0022**	**0,0019**	**0,0015**	**0,0025**	**0,0039**	**0,0033**	**0,0040**	**0,0045**	**0,0043**

5.6. Applicazione della correzione di Rao alle stime con stimatore rapporto separato, post-stratificazione: 1, 2, 3, 4-5 STELLE

Il grafico 5.14 seguente riporta gli errori relativi delle stime grezze ottenute con il rapporto separato e post- stratificazione: Arte, Altra risorsa, nonché gli errori relativi delle stime anticipate ottenute con la correzione di Rao al lag 1.

Grafico 5.14



La tabella seguente riporta, al solito, le sintesi statistiche delle distribuzioni dei 24 errori relativi delle stime anticipate dei totali mensili: dal lag 0 (stime grezze) al lag 12.

Tabella 5.8

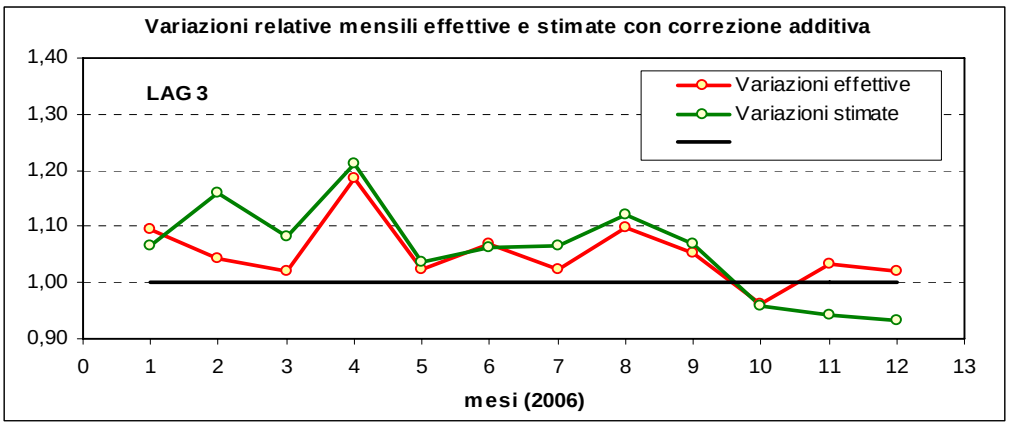
PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

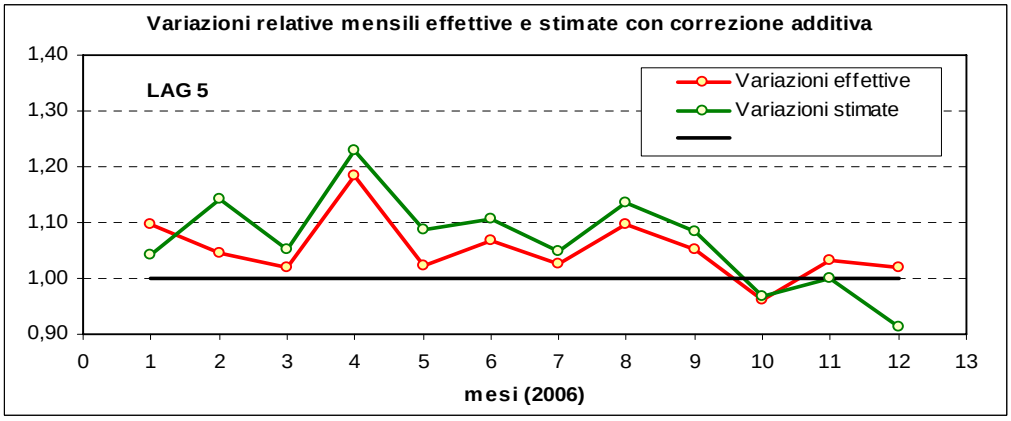
CORREZIONE ADDITIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9644	0,9333	0,9037	0,9136	0,8799	0,8975	0,9423	0,9129	0,9604	0,9338	0,9349	1,0049	1,0190
massimo	1,1549	1,0818	1,0982	1,1108	1,0794	1,1036	1,0732	1,0962	1,1822	1,1147	1,1090	1,2039	1,1283
range	0,1905	0,1485	0,1945	0,1972	0,1995	0,2061	0,1309	0,1833	0,2218	0,1808	0,1741	0,1990	0,1093
media	1,0593	1,0045	1,0047	1,0098	1,0113	1,0180	1,0256	1,0298	1,0372	1,0400	1,0487	1,0616	1,0648
bias	0,0593	0,0045	0,0047	0,0098	0,0113	0,0180	0,0256	0,0298	0,0372	0,0400	0,0487	0,0616	0,0648
varianza	0,0019	0,0017	0,0021	0,0021	0,0029	0,0023	0,0019	0,0022	0,0028	0,0022	0,0018	0,0026	0,0012
mse	0,0054	0,0017	0,0021	0,0022	0,0030	0,0026	0,0025	0,0031	0,0042	0,0038	0,0041	0,0064	0,0054
CORREZIONE MOLTIPLICATIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9644	0,9366	0,9190	0,9465	0,9274	0,9532	0,9671	0,9417	0,9540	0,9321	0,9243	1,0021	1,0172
massimo	1,1549	1,0841	1,1005	1,1171	1,0959	1,1193	1,0928	1,1146	1,1718	1,1136	1,1180	1,1921	1,1329
range	0,1905	0,1475	0,1816	0,1707	0,1686	0,1661	0,1257	0,1729	0,2178	0,1815	0,1937	0,1900	0,1157
media	1,0593	1,0052	1,0076	1,0139	1,0177	1,0214	1,0287	1,0316	1,0349	1,0391	1,0448	1,0573	1,0630
bias	0,0593	0,0052	0,0076	0,0139	0,0177	0,0214	0,0287	0,0316	0,0349	0,0391	0,0448	0,0573	0,0630
varianza	0,0019	0,0017	0,0018	0,0014	0,0023	0,0016	0,0013	0,0019	0,0023	0,0023	0,0021	0,0025	0,0014
mse	0,0054	0,0017	0,0018	0,0016	0,0026	0,0020	0,0021	0,0029	0,0036	0,0038	0,0041	0,0058	0,0053

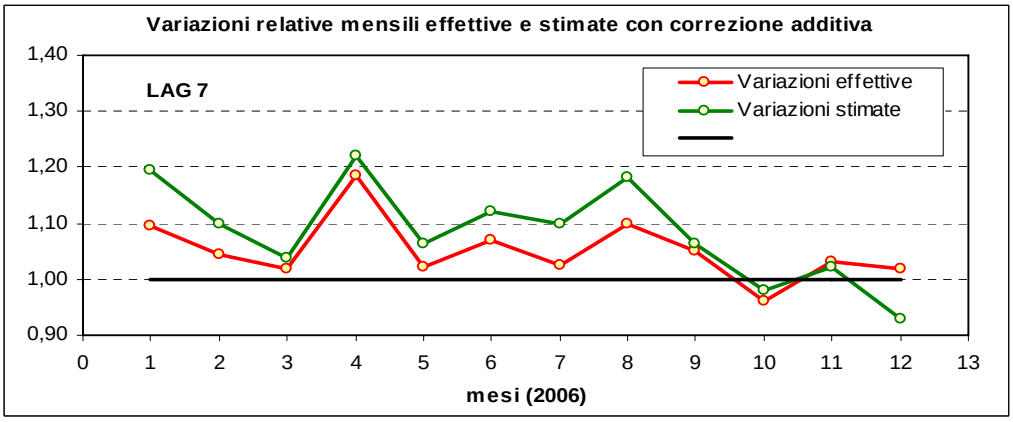
I grafici seguenti 5.15 e 5.16 confrontano le variazioni relative mensili effettive con quelle stimate, dopo l'applicazione della correzione di Rao additiva e moltiplicativa ai lag 3, 5, 7, 12.

La linea nera indica la situazione di nessuna variazione relativa mensile: $P_1/P_{1-12} = 1$.

Grafico 5.15 (Rapporto separato - Correzione additiva)







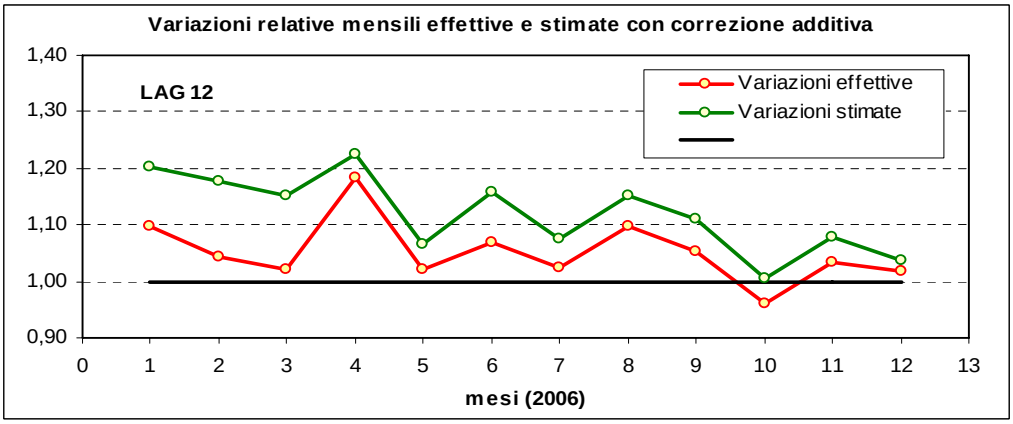
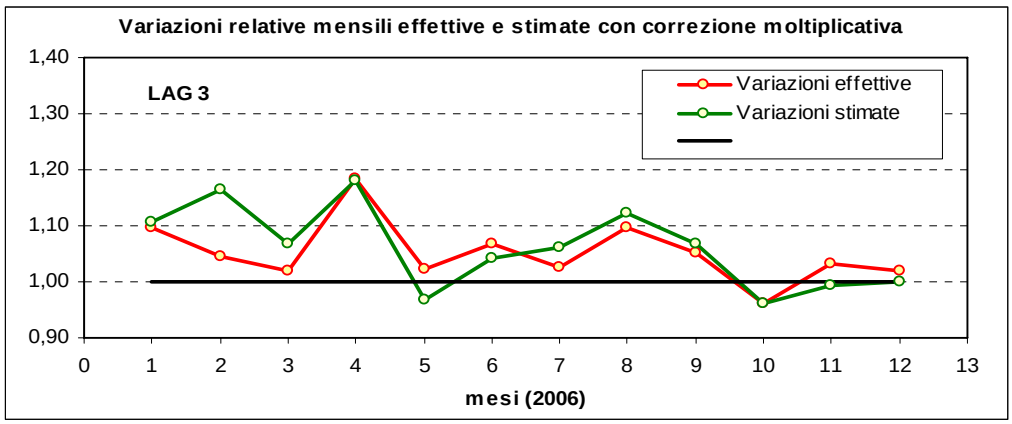
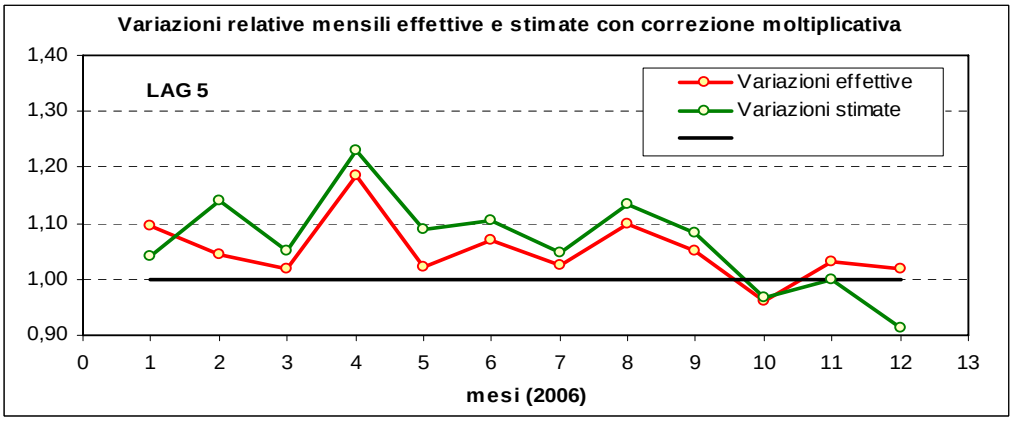
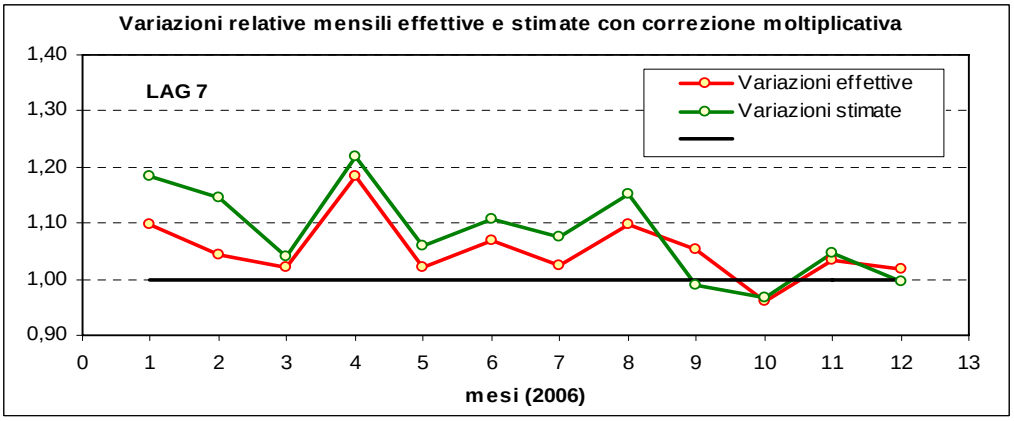
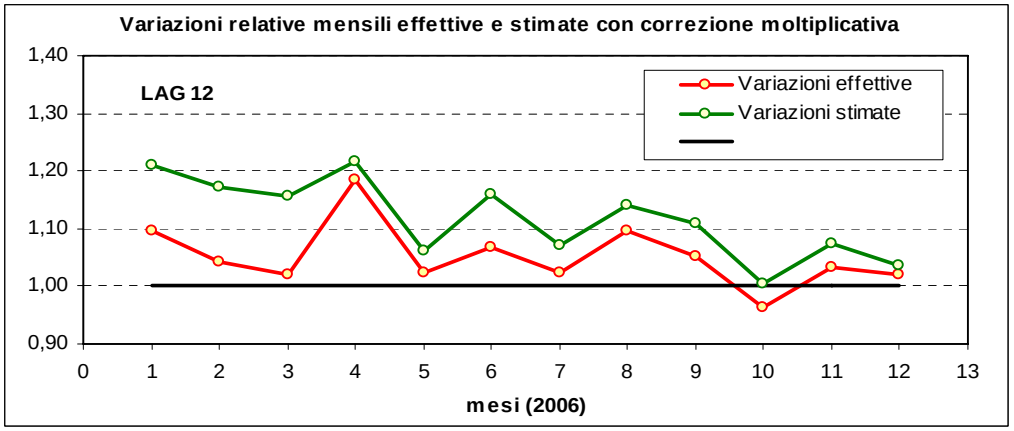


Grafico 5.16 (Rapporto separato - Correzione moltiplicativa)









Nella tabella seguente riportiamo una misura della bontà delle stime anticiapate delle variazione relative mensili. Abbiamo qui calcolato la media e la varianza delle 12 differenze fra variazioni stimate ed effettive.

Tabella 5.9

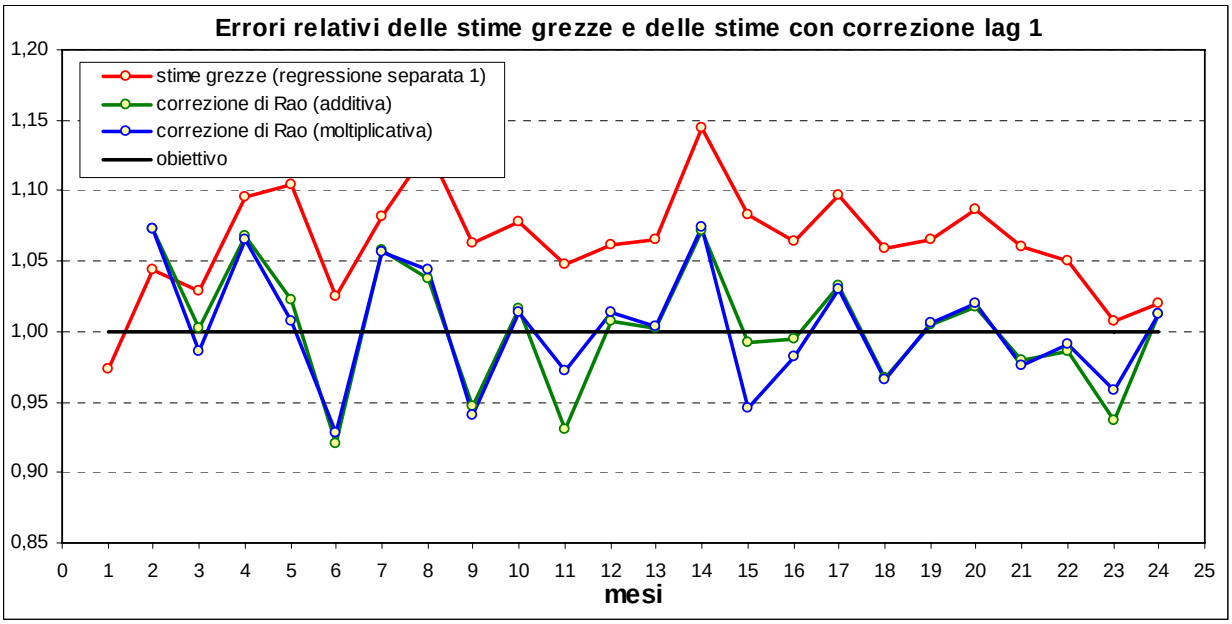
Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi	Variazione	Differenze fra variazioni stimate ed effettive ai lag											
2006	effettiva	lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Correzione additiva														
Media	**1,0517**	**0,0955**	**0,0024**	**0,0050**	**0,0071**	**0,0110**	**0,0154**	**0,0183**	**0,0322**	**0,0374**	**0,0368**	**0,0528**	**0,0624**	**0,0680**
Varianza		**0,0007**	**0,0013**	**0,0025**	**0,0031**	**0,0030**	**0,0028**	**0,0019**	**0,0023**	**0,0039**	**0,0025**	**0,0023**	**0,0030**	**0,0013**
Correzione moltiplicativa														
Media	**1,0517**	**0,0955**	**0,0017**	**0,0054**	**0,0099**	**0,0147**	**0,0180**	**0,0208**	**0,0298**	**0,0333**	**0,0354**	**0,0477**	**0,0570**	**0,0661**
Varianza		**0,0007**	**0,0013**	**0,0021**	**0,0020**	**0,0022**	**0,0017**	**0,0011**	**0,0018**	**0,0031**	**0,0025**	**0,0027**	**0,0029**	**0,0015**

5.7. Applicazione della correzione di Rao alle stime con stimatore regressione separata 1, post-stratificazione: 1, 2, 3, 4-5 STELLE

Il grafico 5.17 seguente riporta gli errori relativi delle stime grezze ottenute con la regressione separata 1 (dove l'ausiliaria x è data dall'indici di qualità: bagni/camere) e post- stratificazione: 1, 2, 3, 4-5 STELLE, nonché gli errori relativi delle stime anticipate ottenute con la correzione di Rao al lag 1.

Grafico 5.17



La tabella seguente riporta, al solito le sintesi statistiche delle distribuzioni degli errori relativi delle stime anticipate: dal lag 0 (stime grezze) al lag 12.

Tabella 5.10

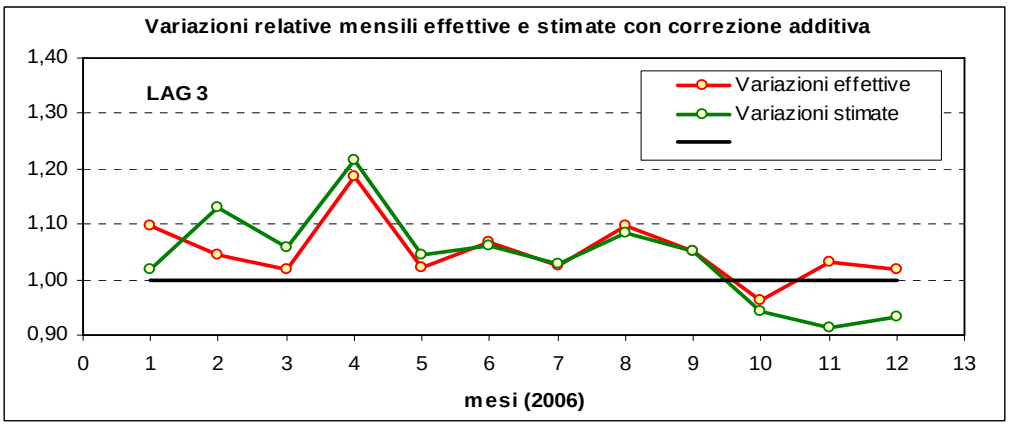
PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

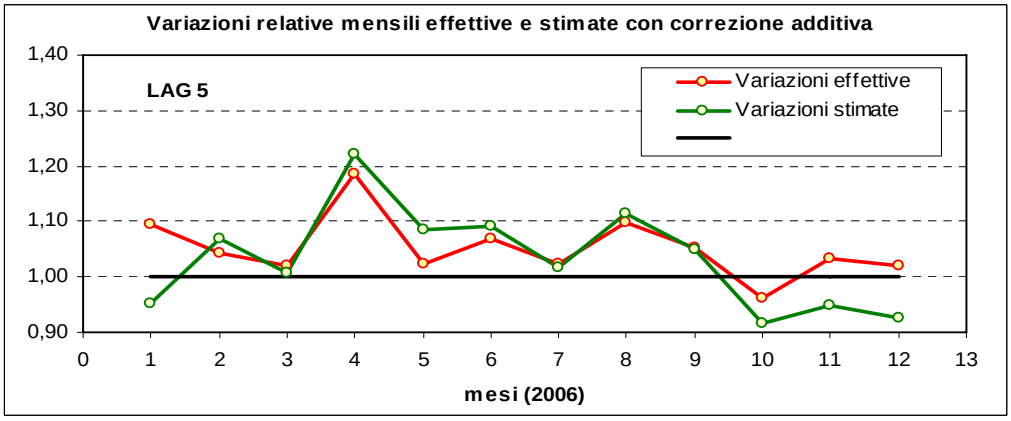
CORREZIONE ADDITIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9729	0,9202	0,9168	0,8772	0,8684	0,8679	0,8683	0,8602	0,8890	0,8994	0,8960	0,9531	0,9594
massimo	1,1447	1,0725	1,1024	1,1130	1,1189	1,1028	1,1052	1,1459	1,0982	1,0920	1,0684	1,0988	1,1030
range	0,1718	0,1523	0,1857	0,2358	0,2505	0,2349	0,2369	0,2857	0,2092	0,1925	0,1724	0,1457	0,1435
media	1,0639	1,0035	0,9998	0,9975	0,9931	0,9935	0,9938	0,9931	0,9946	0,9936	0,9980	1,0079	1,0088
bias	0,0639	0,0035	-0,0002	-0,0025	-0,0069	-0,0065	-0,0062	-0,0069	-0,0054	-0,0064	-0,0020	0,0079	0,0088
varianza	0,0014	0,0018	0,0026	0,0035	0,0039	0,0033	0,0041	0,0046	0,0038	0,0029	0,0019	0,0024	0,0023
mse	0,0055	0,0018	0,0026	0,0035	0,0039	0,0033	0,0042	0,0046	0,0038	0,0029	0,0019	0,0025	0,0024
CORREZIONE MOLTIPLICATIVA													
LAG	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
minimo	0,9729	0,9277	0,9297	0,9266	0,9251	0,9310	0,9186	0,9260	0,9177	0,8797	0,8909	0,9433	0,9607
massimo	1,1447	1,0741	1,1026	1,1264	1,1351	1,0976	1,1123	1,1612	1,1173	1,1079	1,0763	1,1122	1,0969
range	0,1718	0,1463	0,1730	0,1998	0,2101	0,1666	0,1937	0,2351	0,1996	0,2282	0,1854	0,1688	0,1362
media	1,0639	1,0029	1,0016	1,0026	1,0012	1,0001	1,0023	1,0015	0,9995	0,9994	0,9994	1,0064	1,0069
bias	0,0639	0,0029	0,0016	0,0026	0,0012	0,0001	0,0023	0,0015	-0,0005	-0,0006	-0,0006	0,0064	0,0069
varianza	0,0014	0,0017	0,0021	0,0022	0,0026	0,0020	0,0022	0,0032	0,0028	0,0026	0,0022	0,0028	0,0023
mse	0,0055	0,0017	0,0021	0,0022	0,0026	0,0020	0,0022	0,0032	0,0028	0,0026	0,0022	0,0028	0,0024

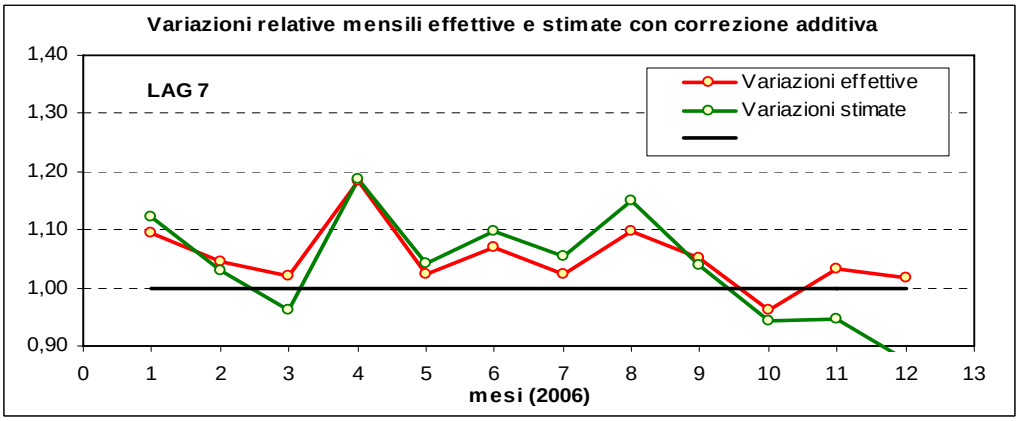
I grafici 5.18 e 5.19 seguenti confrontano le variazioni relative mensili effettive con quelle stimate, dopo l'applicazione della correzione di Rao additiva e moltiplicativa ai lag 3, 5, 7, 12.

La linea nera indica, al solito, la situazione di nessuna variazione relativa mensile: $P_1/P_{1-12} = 1$.

Grafico 5.18 (Regressione separata1 - Correzione additiva)







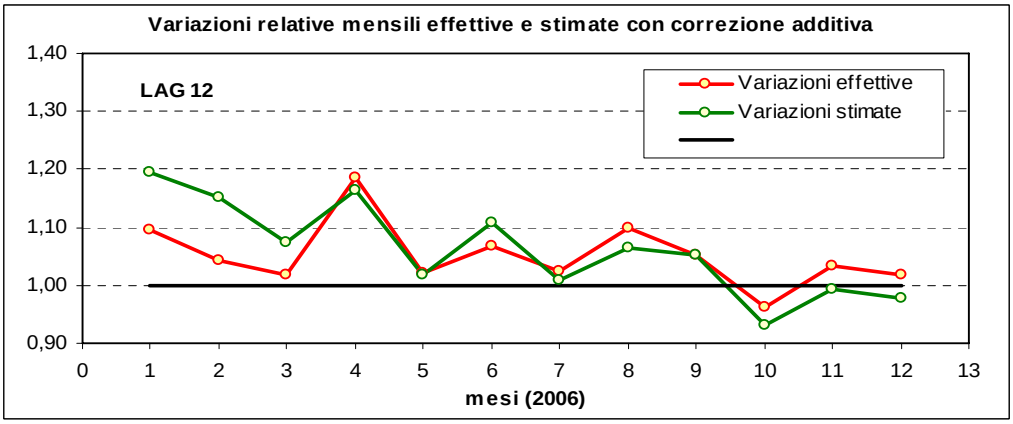
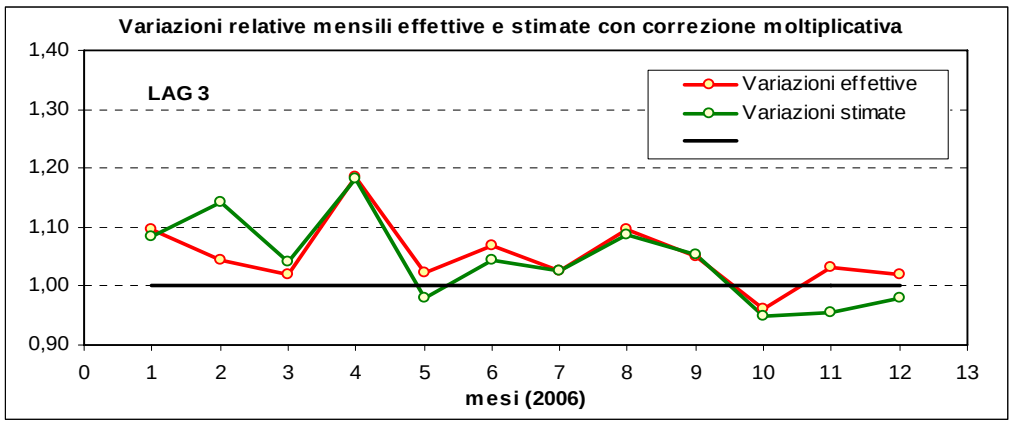
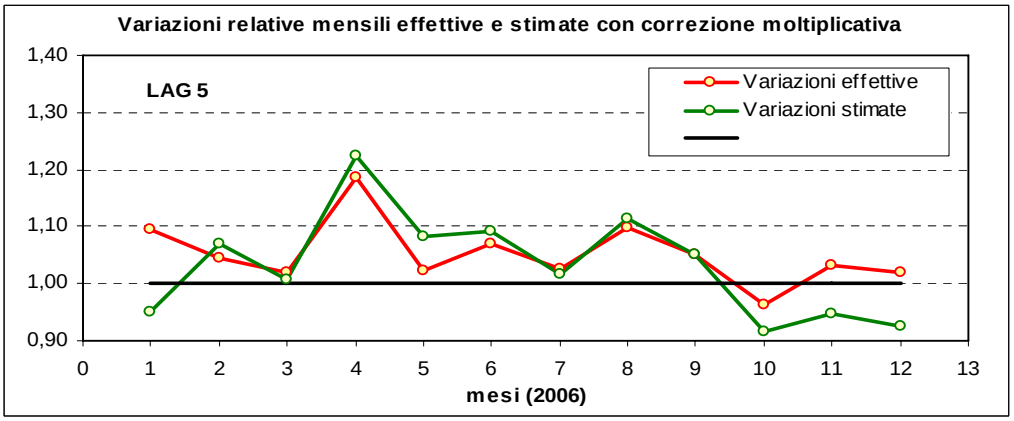
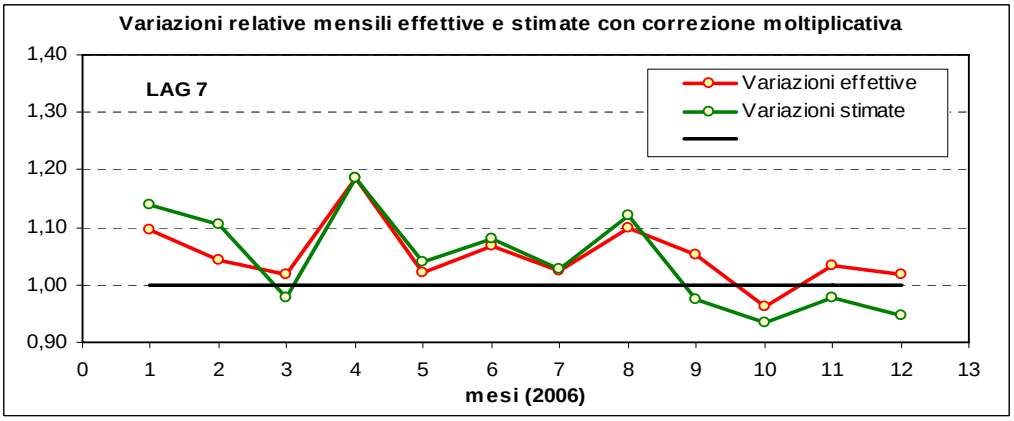
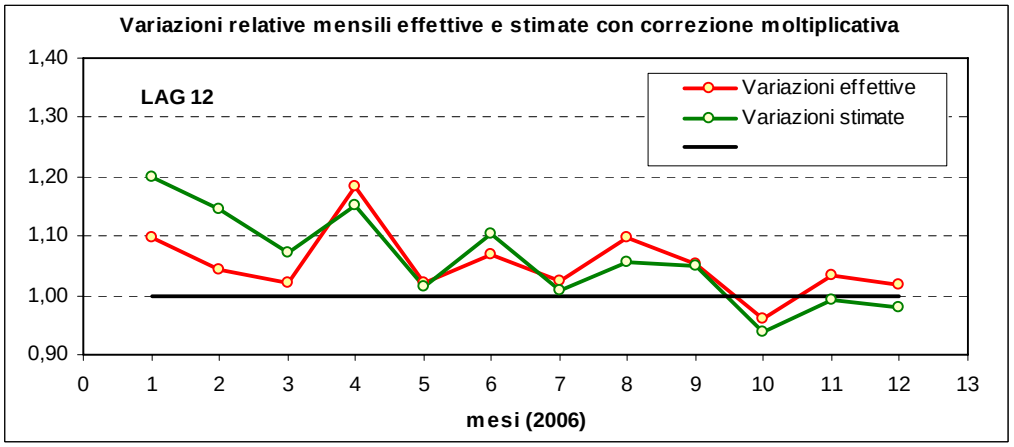


Grafico 5.19 (Regressione separata1 - Correzione moltiplicativa)









Riportiamo infine, nella tabella 5.11 seguente la solita misura della bontà delle stime anticipate delle variazione relative mensili, in cui abbiamo calcolato la media e la varianza delle 12 differenze fra variazioni stimate ed effettive.

Tabella 5.11

Variazioni mensili relative effettive e differenze fra variazioni stimate e variazioni effettive

Mesi	Variazione	Differenze fra variazioni stimate ed effettive ai lag												
2006	effettiva	lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Correzione additiva														
Media	**1,0517**	**0,0705**	**-0,0002**	**-0,0043**	**-0,0113**	**-0,0140**	**-0,0186**	**-0,0235**	**-0,0143**	**-0,0143**	**-0,0196**	**-0,0098**	**0,0012**	**0,0096**
Varianza		**0,0013**	**0,0011**	**0,0019**	**0,0030**	**0,0026**	**0,0035**	**0,0032**	**0,0029**	**0,0042**	**0,0027**	**0,0018**	**0,0023**	**0,0026**
Correzione moltiplicativa														
Media	**1,0517**	**0,0705**	**-0,0030**	**-0,0064**	**-0,0082**	**-0,0083**	**-0,0114**	**-0,0137**	**-0,0095**	**-0,0105**	**-0,0130**	**-0,0079**	**-0,0009**	**0,0074**
Varianza		**0,0013**	**0,0012**	**0,0017**	**0,0016**	**0,0018**	**0,0018**	**0,0011**	**0,0018**	**0,0030**	**0,0022**	**0,0022**	**0,0026**	**0,0026**

5.8. Confronto fra gli otto procedimenti di stima considerati per le presenze negli alberghi

Effettuiamo un confronto sintetico fra gli otto procedimenti di stima considerati nelle sezioni precedenti. I quattro stimatori con cui ottenere le stime preliminari (grezze) dei totali mensili delle presenze nel dominio Alberghi sono stati i seguenti:

- o regressione separata — post-stratificazione: ARTE, ALTRA RISORSA, X=LETTI
- o post-stratificato — post-stratificazione: 1, 2, 3, 4-5 STELLE
- o rapporto separato — post-stratificazione: 1, 2, 3, 4-5 STELLE, X = LETTI
- o regressione separata 1 — post-stratificazione: 1, 2, 3, 4-5 STELLE, X = QUALITA'

Alle 24 stime ottenute con ciascuno di essi abbiamo applicato la correzione di Rao dal lag 0 (nessuna correzione) lag 12, sia nella versione additiva, sia in quella moltiplicativa.

Infine abbiamo calcolato la variazione relativa mensile per i 12 mesi del 2006 rispetto ai corrispondenti mesi del 2005. Abbiamo confrontato le variazioni relative stimate con quelle effettive. Nella tabella 5.12 seguente e nei successivi grafici 5.20 – 5.23 sono riportate le medie e le varianze delle 12 differenze fra le variazioni mensili relative stimate e quelle effettive.

Tabella 5.12

Medie e varianze delle 12 differenze fra variazioni relative mensili effettive e stimate (gen - dic 2006 / gen-dic 2005) senza correzione di Rao (lag 0) e con correzione al lag 1, …, 12

Stime grezze	Tipo di correzione	Sintesi	Differenze fra variazioni stimate ed effettive ai lag												
			lag 0	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7	lag 8	lag 9	lag 10	lag 11	lag 12
Regressione separata: (Arte, Altra risorsa), X=letti	additiva	Media	0,1306	-0,0005	0,0009	0,0033	0,0074	0,0133	0,0192	0,0317	0,0375	0,0382	0,0561	0,0669	0,0717
		Varianza	0,0009	0,0020	0,0033	0,0036	0,0039	0,0030	0,0021	0,0026	0,0040	0,0037	0,0027	0,0031	0,0020
	moltiplicativa	Media	0,1306	-0,0005	0,0022	0,0069	0,0117	0,0149	0,0191	0,0269	0,0310	0,0341	0,0474	0,0588	0,0663
		Varianza	0,0009	0,0016	0,0025	0,0023	0,0030	0,0021	0,0017	0,0024	0,0035	0,0033	0,0036	0,0033	0,0021
Post-stratificato: (1,2,3,4-5 STELLE)	additiva	Media	0,0614	-0,0013	-0,0060	-0,0152	-0,0192	-0,0221	-0,0253	-0,0175	-0,0162	-0,0209	-0,0084	0,0036	0,0103
		Varianza	0,0018	0,0012	0,0018	0,0028	0,0020	0,0021	0,0018	0,0022	0,0039	0,0028	0,0033	0,0046	0,0043
	moltiplicativa	Media	0,0614	-0,0048	-0,0099	-0,0136	-0,0148	-0,0179	-0,0202	-0,0165	-0,0163	-0,0179	-0,0093	0,0006	0,0094
		Varianza	0,0018	0,0015	0,0022	0,0021	0,0022	0,0019	0,0015	0,0025	0,0039	0,0033	0,0040	0,0045	0,0043
Rapporto separato: (1,2,3,4-5 STELLE), X=letti	additiva	Media	0,0955	0,0024	0,0050	0,0071	0,0110	0,0154	0,0183	0,0322	0,0374	0,0368	0,0528	0,0624	0,0680
		Varianza	0,0007	0,0013	0,0025	0,0031	0,0030	0,0028	0,0019	0,0023	0,0039	0,0025	0,0023	0,0030	0,0013
	moltiplicativa	Media	0,0955	0,0017	0,0054	0,0099	0,0147	0,0180	0,0208	0,0298	0,0333	0,0354	0,0477	0,0570	0,0661
		Varianza	0,0007	0,0013	0,0021	0,0020	0,0022	0,0017	0,0011	0,0018	0,0031	0,0025	0,0027	0,0029	0,0015
Regressione separata (1,2,3,4-5 STELLE), X=qualità	additiva	Media	0,0705	-0,0002	-0,0043	-0,0113	-0,0140	-0,0186	-0,0235	-0,0143	-0,0143	-0,0196	-0,0098	0,0012	0,0096
		Varianza	0,0013	0,0011	0,0019	0,0030	0,0026	0,0035	0,0032	0,0029	0,0042	0,0027	0,0018	0,0023	0,0026
	moltiplicativa	Media	0,0705	-0,0030	-0,0064	-0,0082	-0,0083	-0,0114	-0,0137	-0,0095	-0,0105	-0,0130	-0,0079	-0,0009	0,0074
		Varianza	0,0013	0,0012	0,0017	0,0016	0,0018	0,0018	0,0011	0,0018	0,0030	0,0022	0,0022	0,0026	0,0026

Grafico 5.20

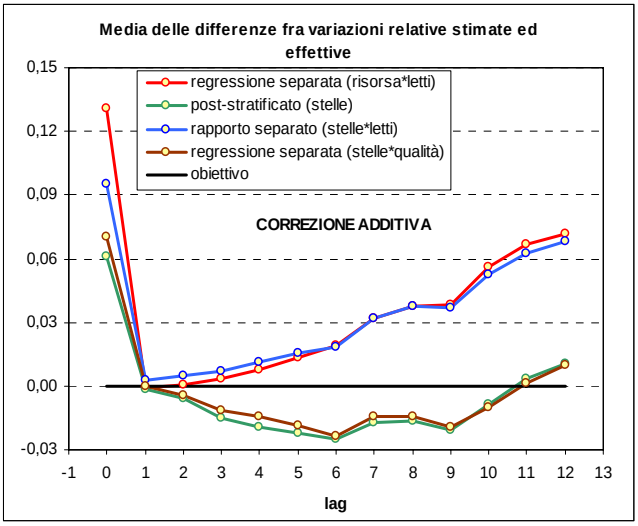


Grafico 5.21

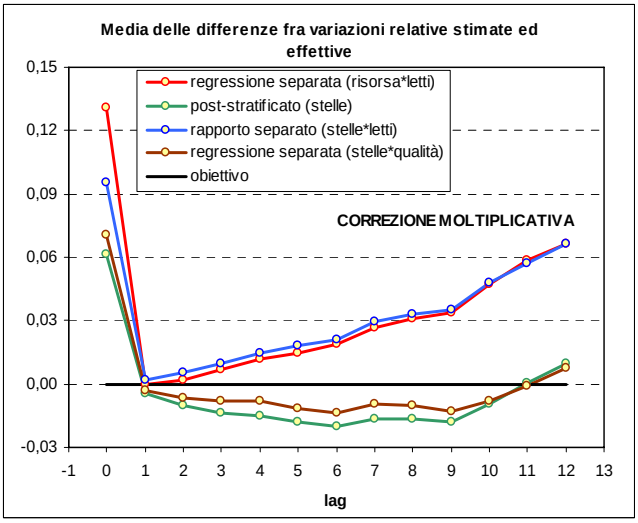


Grafico 5.22

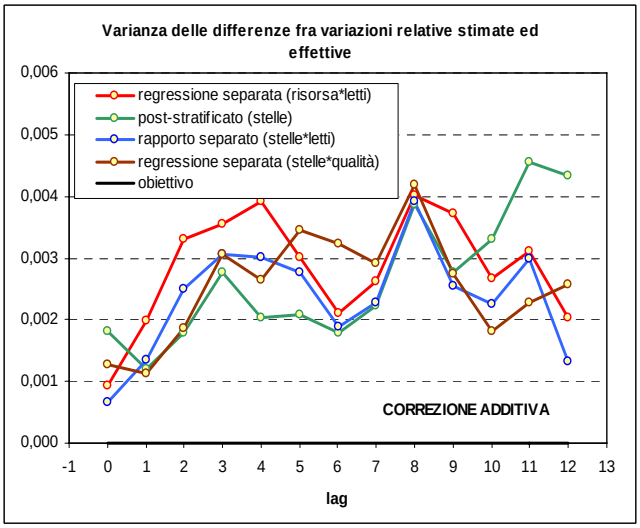
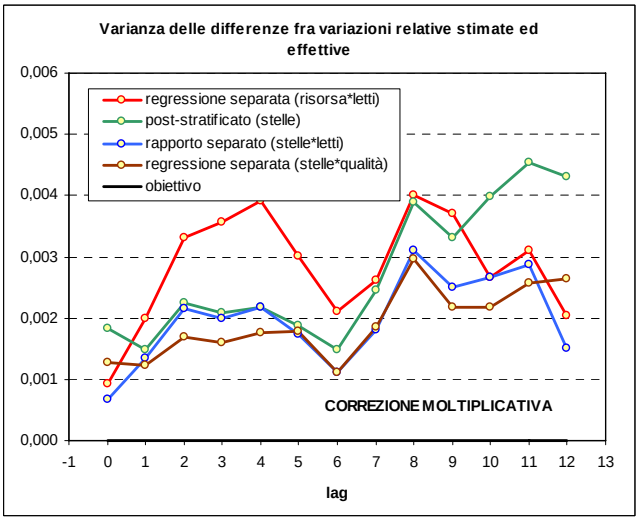


Grafico 5.23



Si nota, negli scarti fra variazioni stimate ed effettive, un comportamento pressoché identico del post-stratificato per stelle e della regressione separata (stelle*qualità): la variabile ausiliaria qualità=bagni/camere, aggiunge un modesto beneficio, una volta presente la categoria dell'albergo.

Un analogo comportamento emerge per la coppia regressione separata (risorsa*letti) e il rapporto separato (stelle*letti); questo fatto è di più difficile interpretazione.

Per tutti i procedimenti di stima delle variazioni relative mensili va però ricordato che, nell'effettuare il rapporto fra presenze stimate del mese m+12 e presenze note del mese m si introduce l'informazione ausiliaria delle presenze dell'anno precedente che, in qualche modo, ha un valore esplicativo delle presenze del mese attuale analogo a quello della dimensione.

Quanto alle varianze, non emergono nette differenze fra gli otto procedimenti di stima: la varianza (dei 12 scarti fra variazioni relative mensili effettive e stimate) risulta però più contenuta nel procedimento che usa lo stimatore di regressione separata (stelle*qualità) e la correzione moltiplicativa.

In definitiva, il procedimento: stima preliminare del totale mensile con regressione separata (stelle*qualità) e aggiustamento di Rao moltiplicativo risulta un buon candidato per la stima anticipata delle presenze mensili nel dominio alberghi e di conseguenza per la stima preliminare delle variazioni relative mensili.

Capitolo 6

6.1. Stima basata sui propensity scores per la sottopopolazione degli esercizi alberghieri

Abbiamo visto che le stime preliminari (grezze) ottenute con gli stimatori di regressione generalizzata fin qui considerati tendono a fornire errori di stima positivi. E' però possibile ottenere stime anticipate delle presenze riducendo, se non proprio eliminando, il bias delle stime preliminari con l'aggiustamento di Rao. Infine, abbiamo visto che possiamo ottenere stime anticipate delle variazioni relative mensili assai precise: l'errore medio di stima delle 12 variazioni relative mensili del 2006 rispetto al 2005 è pressoché nullo quando l'aggiustamento di Rao si effettua al lag 1. Questo errore tende ad allontanarsi da 0 (nella maggior parte dei casi nella direzione positiva) al crescere del lag, perché la capacità di ridurre il bias dell'aggiustamento di Rao diminuisce aumentando il lag.

L'indagine completa sul movimento dei clienti nelle strutture ricettive della provincia di Firenze è in grado di fornire i risultati, stante le attuali condizioni, con circa 5 mesi di ritardo. In una implementazione effettiva dei metodi visti, le stime anticipate dovranno usare un aggiustamento a un lag intorno a 5 mesi. Se poi consideriamo la possibilità di estendere l'applicazione ad altre realtà provinciali e per un certo numero di anni, non possiamo escludere l'utilizzo dell'aggiustamento di Rao a valori del lag di 7 – 8 mesi..

In considerazione di questo limite abbiamo considerato anche un'altra strada che potrebbe fornire stime preliminari migliori: con minor bias e con minor variabilità. Una possibilità è offerta dal metodo che si basa sulla stima della probabilità individuale delle unità della popolazione di appartenere al sottoinsieme turiweb (Biffignandi e Pratesi, 2004).

L'appartenere al sottoinsieme turiweb può dipendere da molti fattori, in particolare da caratteristiche personali del gestore della struttura ricettiva, quali ad esempio l'età, il sesso e l'istruzione. Come abbiamo visto nel cap. 1 gli esercizi di tipo meno tradizionale (probabilmente diretti da titolari più giovani, istruiti, più innovativi) sono più presenti nel sistema turiweb degli esercizi di altre tipologie. Abbiamo anche notato che una resistenza all'investimento nelle nuove tecnologie può dipendere da situazioni di monopolio territoriale. Nella provincia di Firenze, mentre appartengono al sistema turiweb il 31% degli affittacamere, questa percentuale è solo del 22% per gli alberghi. (situazione all'ottobre 2006).

Non sono disponibili informazioni sui gestori, ma è possibile accedere a una serie di dettagliate informazioni sulle attrezzature presenti e i servizi offerti dagli esercizi ricettivi derivanti dalle comunicazioni annuali dei prezzi.

Annualmente le strutture ricettive sono tenute a comunicare i prezzi praticati per le varie forme di offerta di alloggio oltre a un insieme di informazioni sui servizi presenti. Per ciascuna tipologia ricettiva è previsto un modello di rilevazione particolarmente dettagliato. Per circoscrivere la nostra analisi ci siamo limitati agli esercizi alberghieri (alberghi in senso stretto e residenze turistico alberghiere) che presentano una relativa omogeneità nella modalità di offerta di alloggio.

Abbiamo recuperato alcune informazioni dall'archivio delle attrezzature e prezzi, ritenendole indicative della propensione della gestione ad usare la tecnologia internet. La scelta è ragionevole sebbene non sottoposta a un esame dettagliato: altre informazioni, apparentemente distanti da questo obiettivo, potrebbero avere una capacità esplicativa della propensione a partecipare al sistema turiweb, sia singolarmente, sia congiuntamente a quelle considerate.

Dopo un esame delle distribuzioni di frequenza congiunte fra singola covariata e presenza nel sistema turiweb, abbiamo selezionato le due variabili dicotomiche:

- presenza di un sito internet
- servizio di accesso a internet per i clienti in grado di fornire.

Queste due variabili, congiuntamente ad altre che abbiamo utilizzato nei precedenti capitoli, potrebbero avere una capacità esplicativa del meccanismo di selezione e quindi contribuire a specificare un modello adeguato per stimare questo meccanismo.

Per evidenziare la capacità aggiuntiva di queste due variabili nel modellare il meccanismo di selezione abbiamo considerato due modelli logistici: il primo senza e il secondo con le due variabili dicotomiche suddette.

- Modello 1 (con le covariate disponibili nell'archivio della consistenza):
  - stelle
  - qualità della struttura (numero di bagni / numero di camere)
- Modello 2 (con ulteriori due covariate tratte nell'archivio dei prezzi e servizi):
  - stelle
  - presenza di un sito internet
  - servizio di accesso a internet per i clienti
  - qualità della struttura (numero di bagni / numero di camere)

Più formalmente:

M_1: $\text{logit}[\text{prob}(k \in \text{turiweb})|(m, x_k)] = c_m + d \cdot x_k$ (6.1)

M_2: $\text{logit}[\text{prob}(k \in \text{turiweb})|(g, l, m, x_k)] = a_g + b_l + c_m + d \cdot x_k$ (6.2)

dove: c_m per m = 1, 2, 3 , 4 indica il fattore stelle (1, 2, 3, 4-5 stelle)
a_g per g = 0, 1 indica il fattore presenza/assenza di un sito internet
b_l per l= 0, 1 indica il fattore a due livelli presenza/assenza di accesso a internet per i clienti.
d è il coefficiente della covariata quantitativa qualità.

Sono state escluse le interazioni per evitare di trovarsi con celle con frequenze relative 1 o 0. Le covariate si sono tenute anche se non risultavano significative al test. L'unica covariata, sempre significativa, è la categoria dell'albergo (le stelle).

Indicando con $\hat{\pi}_k$ la probabilità stimata con ciascuno dei due modelli, con U l'intera popolazione, con V il sottoinsieme degli esercizi alberghieri del sistema turiweb, con x_k e y_k i posti letto e le presenze del generico albergo, abbiamo considerato i tre stimatori del totale delle presenze:

Stimatore pseudo HT: $$\hat{t}_{pseudoHT} = \sum_V \frac{y_k}{\hat{\pi}_k} \quad (6.3)$$

Stimatore pseudo rapporto $$\hat{t}_{pseudoRap} = \sum_U x_k \cdot \left(\sum_V \frac{y_k}{\hat{\pi}_k} \Big/ \sum_V \frac{x_k}{\hat{\pi}_k} \right) \quad (6.4)$$

Stimatore pseudo stratificato $$\hat{t}_{pseudoStr} = \sum_{h=1}^{h=5} \frac{N_h}{n_h} \sum_{Vh} y_k \quad (6.5)$$

Questo ultimo stimatore è costruito nel modo seguente: V_h è uno dei 5 gruppi delle strutture alberghiere del sistema turiweb ottenuto dai cinquili della distribuzione dei propensity scores (le $\hat{\pi}_k$), così che n_h è pari a n/5 a meno di approssimazioni all'unità. N_h è la dimensione di uno dei 5 gruppi in cui è suddivisa l'intera popolazione nel quale i valori assunti dai propensity scores cadono nell'intervallo individuato dai cinquili.

Sono state ricavate le stime con i tre stimatori (6.3), (6.4), (6.5) per i due modelli logistici M_1 e M_2:

- stimatore pseudo HT (modello logistico M_1)
- stimatore pseudo rapporto (modello logistico M_1)
- stimatore pseudo stratificato (modello logistico M_1)
- stimatore pseudo HT (modello logistico M_2)
- stimatore pseudo rapporto (modello logistico M_2)
- stimatore pseudo stratificato (modello logistico M_2)

Abbiamo preso in considerazione anche il modello M_1, che può sembrare troppo semplice, perché la sua applicazione nello stimare la probabilità di autoselezione non ci costringe ad accedere all'ulteriore archivio delle attrezzature degli alberghi.

I grafici seguenti 6.1 e 6.2 riportano gli errori relativi di stima per il totale delle presenze ottenuti con lo stimatore pseudoHT e pseudo rapporto dove le probabilità di appartenere al sistema turiweb sono stimate con i due modelli M_1 e M_2.

Grafico 6.1

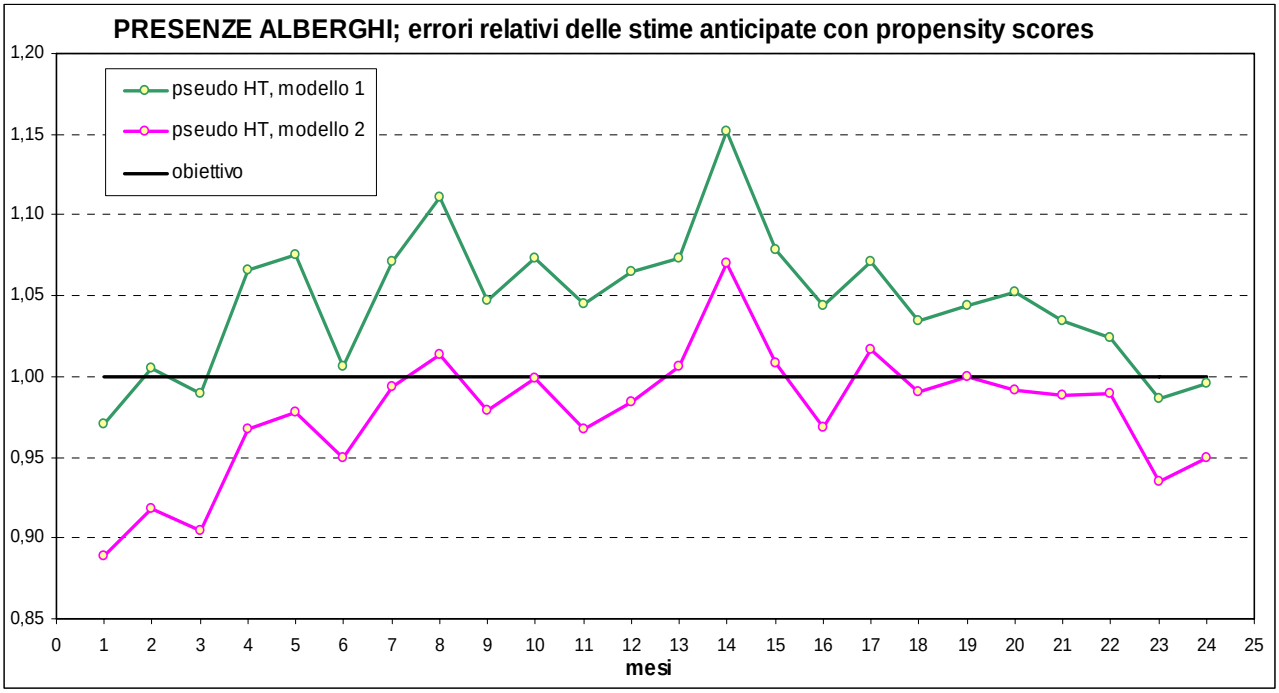
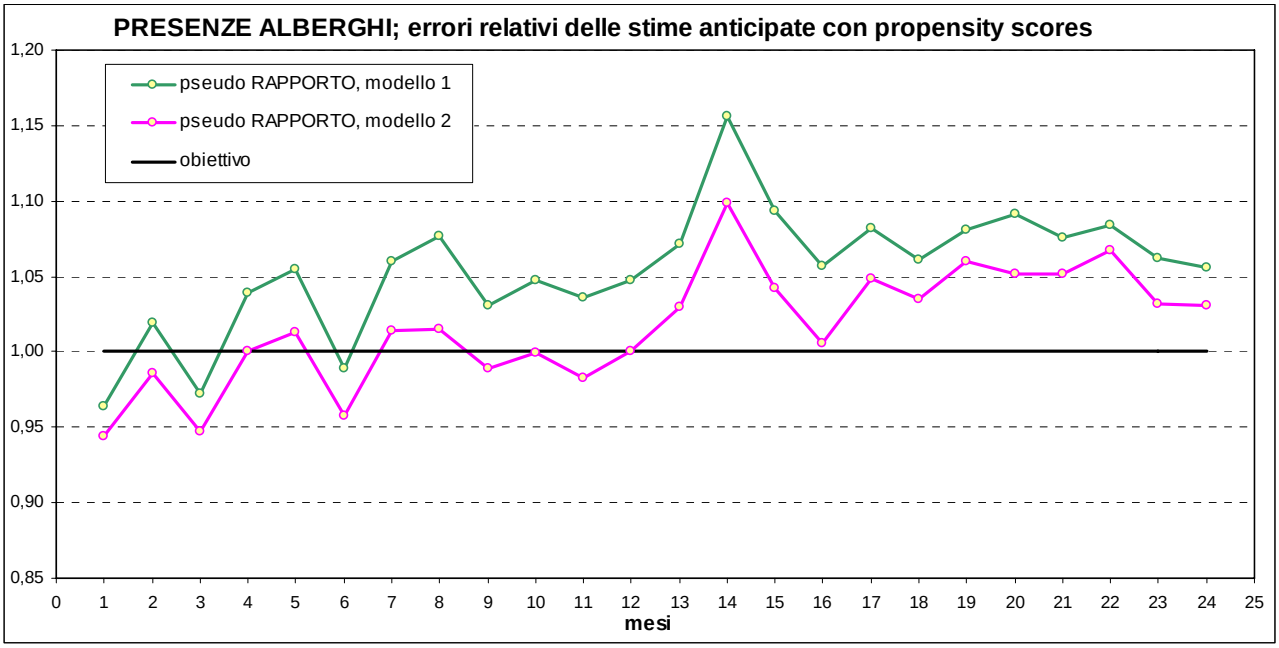


Grafico 6.2



Non riportiamo il grafico delle stime ottenute col terzo stimatore in corrispondenza dei due modelli M_1 e M_2 perché sono molto distanti dall'obiettivo. Quando le probabilità di appartenere al sistema turiweb sono stimate col modello M_2 si ottengono errori di stima relativi assai contenuti. La tabella 6.1 seguente riporta le sintesi statistiche dei 24 errori relativi di stima ottenuti con i tre stimatori (6.3), (6.4), (6.5) per i due modelli logistici M_1 e M_2. In neretto sono evidenziati i valori minimi del bias e della varianza.

Tabella 6.1

Sintesi statistiche dei 24 errori di stima ottenuti con i tre stimatori e propensity scores stimati con i due modelli M_1 e M_2

	Modello logistico 1			Modello logistico 2		
Sintesi	pseudo HT	pseudo RAPPORTO	pseudo STRATIFICATO	pseudo HT	pseudo RAPPORTO	pseudo STRATIFICATO
minimo	0,9705	0,9636	1,2296	0,8886	0,9437	1,2173
massimo	1,1513	1,1556	1,4679	1,0704	1,0985	1,4482
range	0,1808	0,1920	0,2383	0,1818	0,1548	0,2309
media	1,0464	1,0545	1,3488	0,9773	1,0167	1,3130
bias	0,0464	0,0545	0,3488	**-0,0227**	**0,0167**	0,3130
varianza	0,0017	0,0016	0,0040	**0,0015**	**0,0014**	0,0034
mse	0,0038	0,0046	0,1257	0,0020	0,0017	0,1014

Si nota come lo stimatore pseudo stratificato fornisca stime con un bias (nei 24 mesi) del tutto inaccettabile (oltre il 30%), mentre altri forniscono stime assai precise: bias trascurabile e piccola varianza; nel caso dello pseudo HT il bias è addirittura negativo. Gli stimatori pseudo HT e pseudo rapporto con probabilità stimate dal modello logistico M_2 forniscono stime con precisione, in termini di bias e varianza, analoga a quella ottenuta con gli stimatori di regressione visti nel capitolo precedente aggiustati con la correzione di Rao moltiplicativa ai lag 4-5 mesi. Questi due stimatori, possono quindi competere, con i migliori visti in quel capitolo nel caso realistico di anticipare le stime di 4- 5 mesi, senza ricorrere alla correzione di Rao.

Ricordiamo che la correzione di Rao richiede la presenza di una indagine ripetuta sia essa campionaria o censuaria che permetta l'osservazione dell'errore di stima commesso k mesi prima nell'effettuare le stime preliminari.

Nel nostro caso, possiamo comunque considerare preliminari le stime ottenute con lo pseudo HT e lo pseudo rapporto e applicare la correzione di Rao, ripetendo il procedimento di stima delle presenze totali e delle variazioni relative mensili per giungere a un confronto completo con i metodi visti nel capitolo 5.

6.2. Applicazione della correzione di Rao alle stime preliminari ottenute con pseudo HT

La tabella 6.2 seguente riporta le sintesi statistiche dei 24 errori relativi di stima ottenuti con lo stimatore pseudo HT e correzioni di Rao ai lag da 0 (nessuna correzione) a 12.

Tabella 6.2

PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

CORREZIONE ADDITIVA

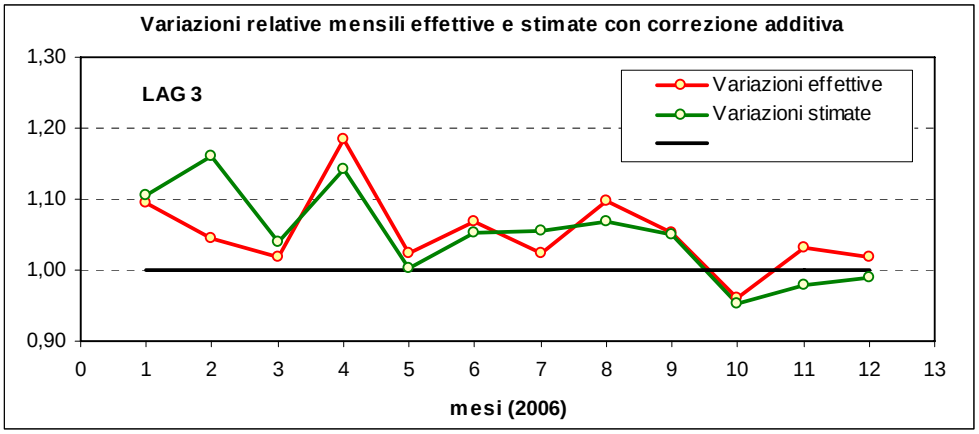
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,8886	0,9499	0,9311	0,9472	0,9360	0,9496	0,9108	0,9216	0,9258	0,8831	0,8897	0,9441	0,9651
massimo	1,0704	1,0635	1,0881	1,1126	1,0737	1,1120	1,0676	1,0912	1,1656	1,1159	1,1406	1,2209	1,1494
range	0,1818	0,1135	0,1569	0,1654	0,1377	0,1624	0,1569	0,1696	0,2398	0,2328	0,2509	0,2769	0,1842
media	0,9773	1,0034	1,0027	1,0061	1,0060	1,0083	1,0111	1,0134	1,0180	1,0163	1,0192	1,0286	1,0290
bias	-0,0227	0,0034	0,0027	0,0061	0,0060	0,0083	0,0111	0,0134	0,0180	0,0163	0,0192	0,0286	0,0290
varianza	0,0015	0,0011	0,0013	0,0013	0,0015	0,0018	0,0017	0,0021	0,0029	0,0035	0,0040	0,0049	0,0033
mese	0,0020	0,0012	0,0013	0,0014	0,0015	0,0019	0,0018	0,0022	0,0033	0,0037	0,0044	0,0057	0,0042

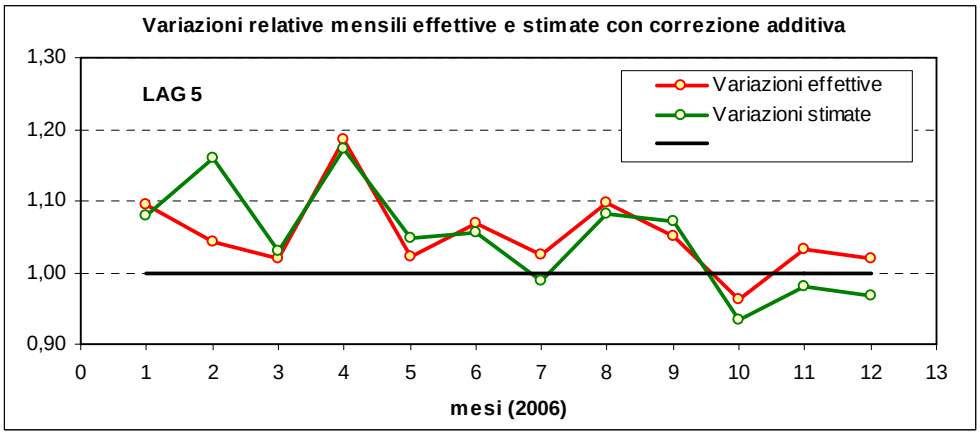
CORREZIONE MOLTIPLICATIVA

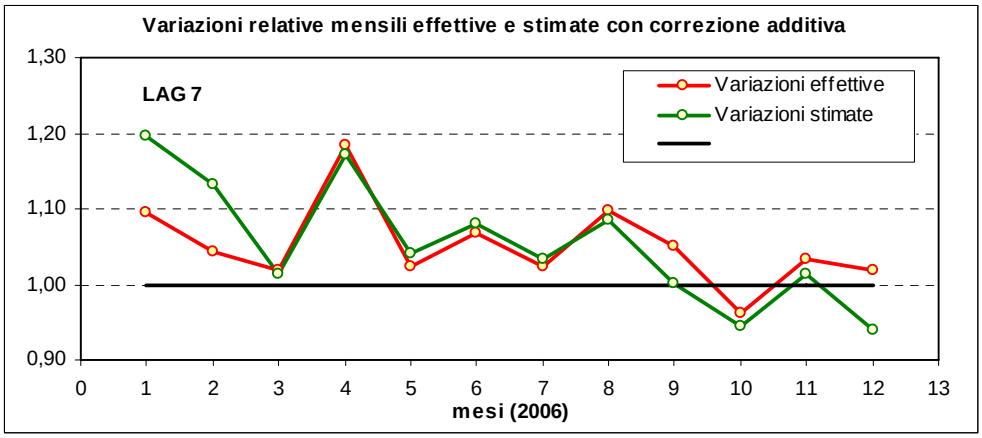
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,8886	0,9418	0,9042	0,9434	0,9254	0,9339	0,9197	0,9231	0,9245	0,8739	0,8875	0,9441	0,9649
massimo	1,0704	1,0698	1,0871	1,1062	1,1008	1,1213	1,1176	1,1409	1,1270	1,1235	1,1129	1,1839	1,1665
range	0,1818	0,1280	0,1829	0,1628	0,1754	0,1875	0,1979	0,2178	0,2025	0,2496	0,2254	0,2398	0,2017
media	0,9773	1,0035	1,0048	1,0094	1,0111	1,0124	1,0161	1,0169	1,0182	1,0188	1,0208	1,0304	1,0346
bias	-0,0227	0,0035	0,0048	0,0094	0,0111	0,0124	0,0161	0,0169	0,0182	0,0188	0,0208	0,0304	0,0346
varianza	0,0015	0,0012	0,0018	0,0019	0,0023	0,0024	0,0029	0,0031	0,0032	0,0037	0,0037	0,0043	0,0042
mese	0,0020	0,0012	0,0018	0,0020	0,0024	0,0026	0,0031	0,0034	0,0035	0,0041	0,0042	0,0053	0,0054

Nei grafici 6.3 e 6.4 che seguono riportiamo il solito confronto fra variazioni relative mensili effettive e stimate. La linea nera indica, al solito la situazione di assenza di variazione relativa fra il mese k e il corrispondente k-12 dell'anno precedente.

Grafico 6.3 (Pseudo HT - Correzione additiva)







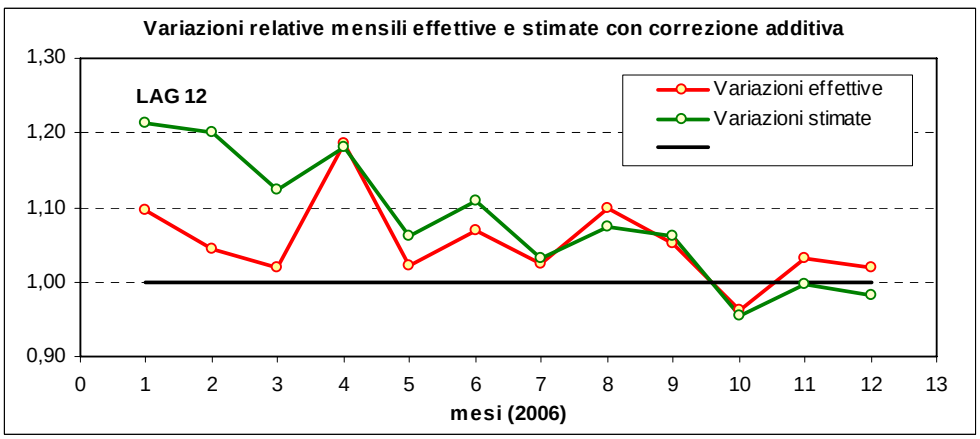
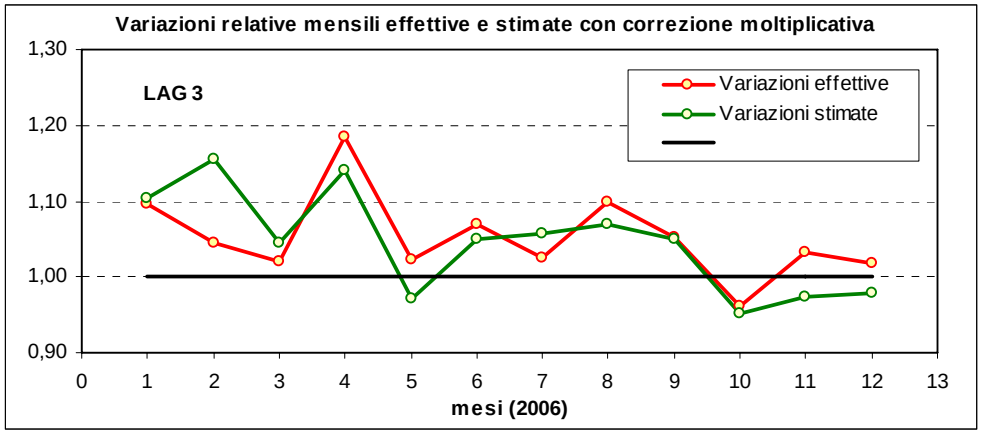
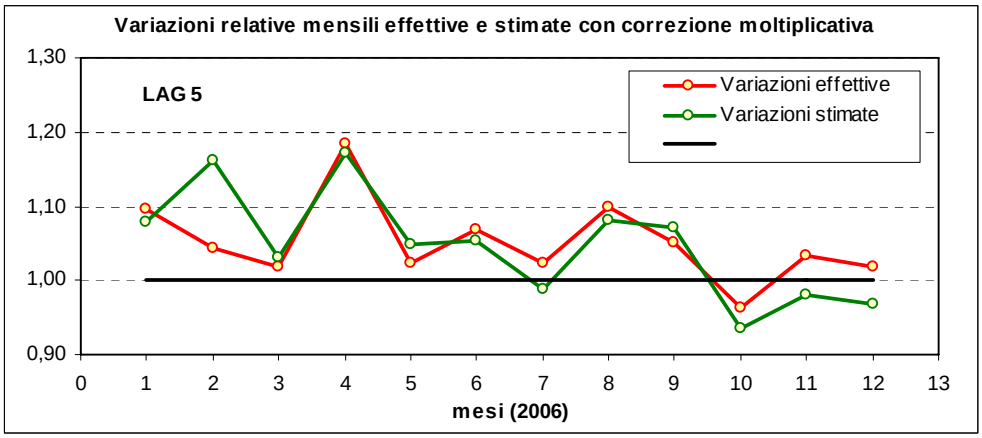
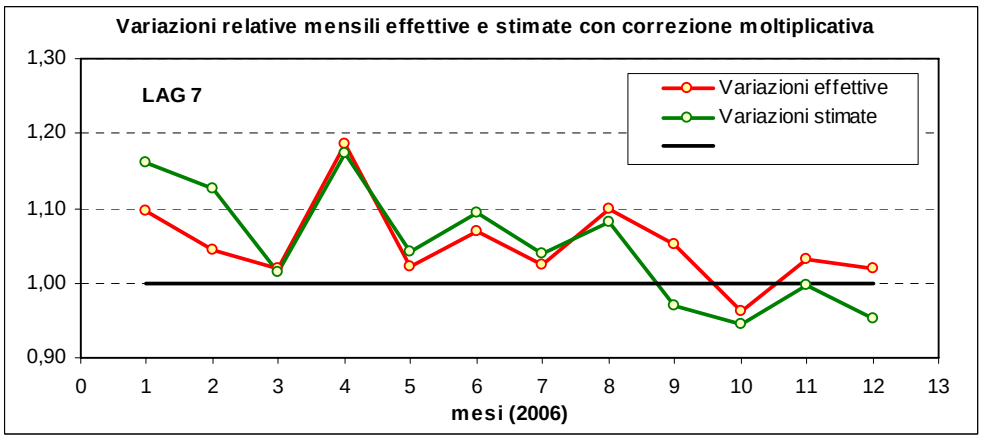
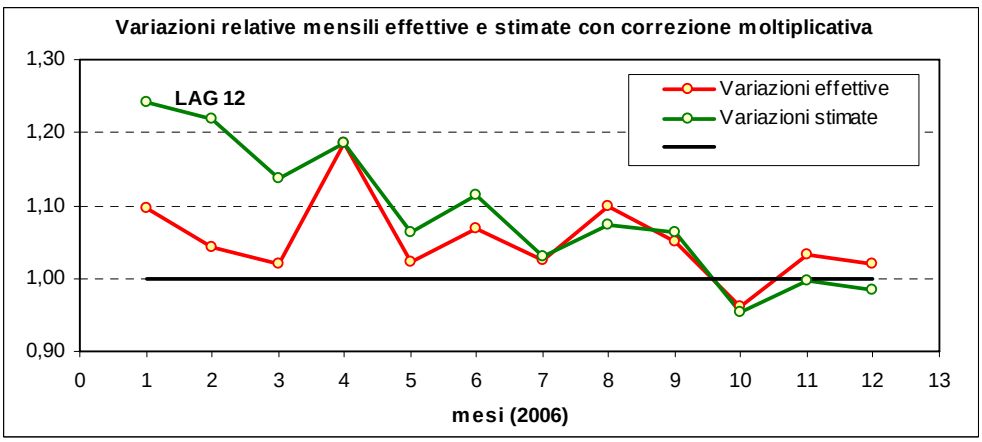


Grafico 6.4 (Pseudo HT - Correzione moltiplicativa)









6.3. Applicazione della correzione di Rao alle stime grezze ottenute con pseudo RAPPORTO

La tabella 6.3 seguente riporta le sintesi statistiche dei 24 errori relativi di stima ottenuti con lo stimatore pseudo RAPPORTO e correzioni di Rao ai lag da 0 (nessuna correzione) a 12.

Tabella 6.3

PRESENZE ALBERGHI - Sintesi statistiche degli errori relativi di stima ai diversi lag di correzione delle stime grezze

CORREZIONE ADDITIVA

LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,9437	0,9371	0,9220	0,9409	0,9431	0,9290	0,9618	0,9505	0,9814	0,9584	0,9466	0,9924	1,0050
massimo	1,0985	1,0657	1,0977	1,1213	1,0995	1,1206	1,0734	1,1020	1,1799	1,0930	1,1037	1,1818	1,1118
range	0,1548	0,1286	0,1757	0,1804	0,1564	0,1917	0,1115	0,1514	0,1984	0,1346	0,1572	0,1894	0,1068
media	1,0167	1,0040	1,0050	1,0110	1,0132	1,0169	1,0235	1,0272	1,0343	1,0347	1,0415	1,0537	1,0581
bias	0,0167	0,0040	0,0050	0,0110	0,0132	0,0169	0,0235	0,0272	0,0343	0,0347	0,0415	0,0537	0,0581
varianza	0,0014	0,0012	0,0016	0,0013	0,0017	0,0016	0,0012	0,0013	0,0020	0,0015	0,0015	0,0021	0,0008
mese	0,0017	0,0012	0,0016	0,0015	0,0019	0,0019	0,0018	0,0020	0,0032	0,0027	0,0032	0,0049	0,0042

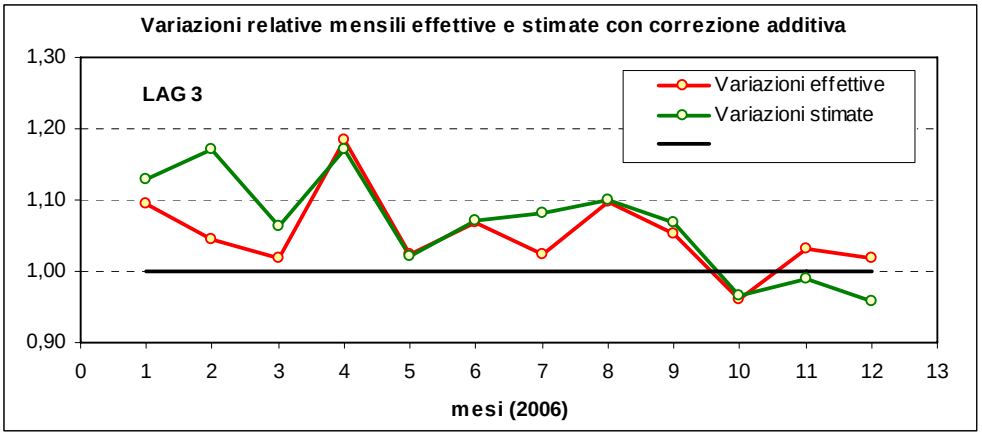
CORREZIONE MOLTIPLICATIVA

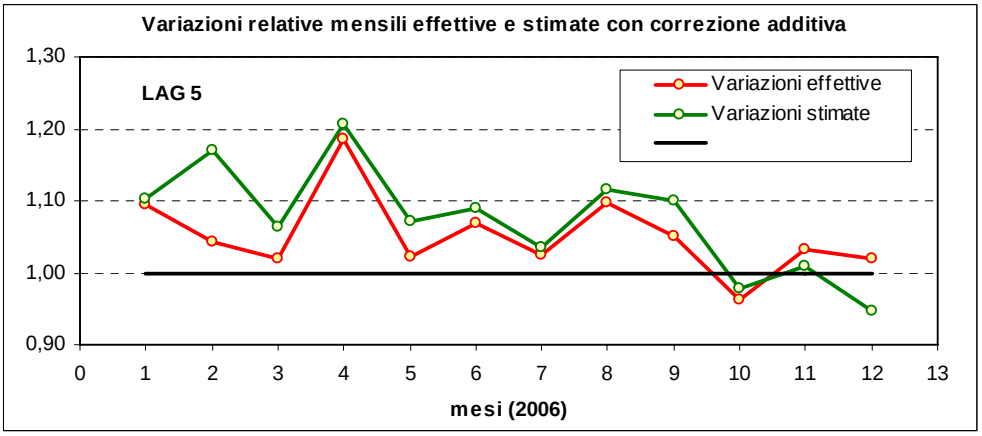
LAG	0	1	2	3	4	5	6	7	8	9	10	11	12
minimo	0,9437	0,9446	0,9153	0,9544	0,9426	0,9654	0,9574	0,9577	0,9720	0,9391	0,9385	0,9924	1,0049
massimo	1,0985	1,0672	1,0977	1,1181	1,0990	1,1113	1,0826	1,0831	1,1477	1,0889	1,0978	1,1601	1,1140
range	0,1548	0,1226	0,1824	0,1637	0,1565	0,1459	0,1252	0,1255	0,1758	0,1498	0,1594	0,1677	0,1091
media	1,0167	1,0045	1,0068	1,0128	1,0161	1,0187	1,0256	1,0283	1,0323	1,0356	1,0413	1,0535	1,0601
bias	0,0167	0,0045	0,0068	0,0128	0,0161	0,0187	0,0256	0,0283	0,0323	0,0356	0,0413	0,0535	0,0601
varianza	0,0014	0,0012	0,0015	0,0013	0,0017	0,0012	0,0012	0,0013	0,0016	0,0016	0,0015	0,0017	0,0010
mese	0,0017	0,0012	0,0015	0,0014	0,0019	0,0015	0,0019	0,0021	0,0027	0,0028	0,0032	0,0046	0,0046

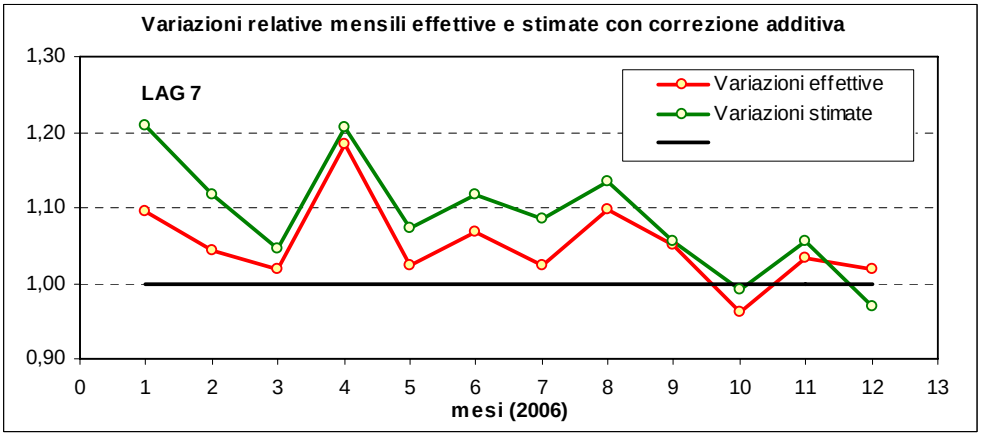
Nei grafici 6.5 e 6.6 che seguono riportiamo il solito confronto fra variazioni relative mensili effettive e quelle stimate, sia adottando la correzione additiva, sia quella moltiplicativa.

La linea nera indica, al solito la situazione di assenza di variazione relativa fra il mese k e il corrispondente k-12 dell'anno precedente.

Grafico 6.5 (Pseudo RAPPORTO - Correzione additiva)







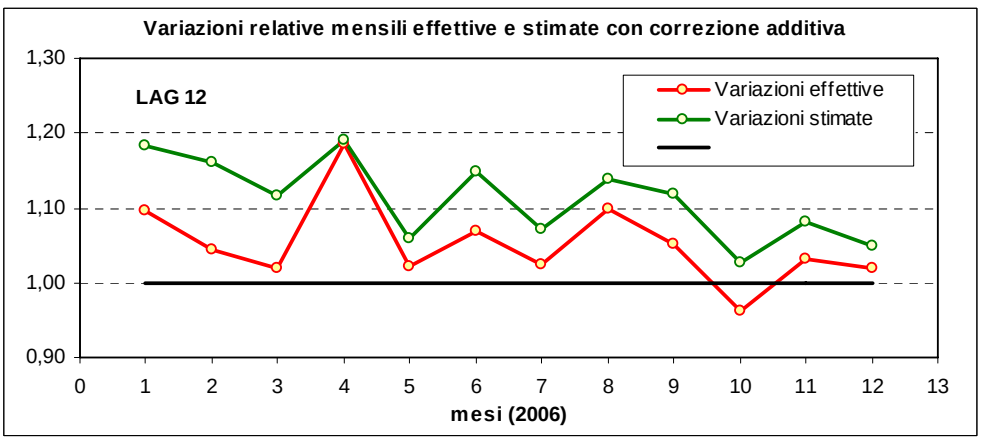
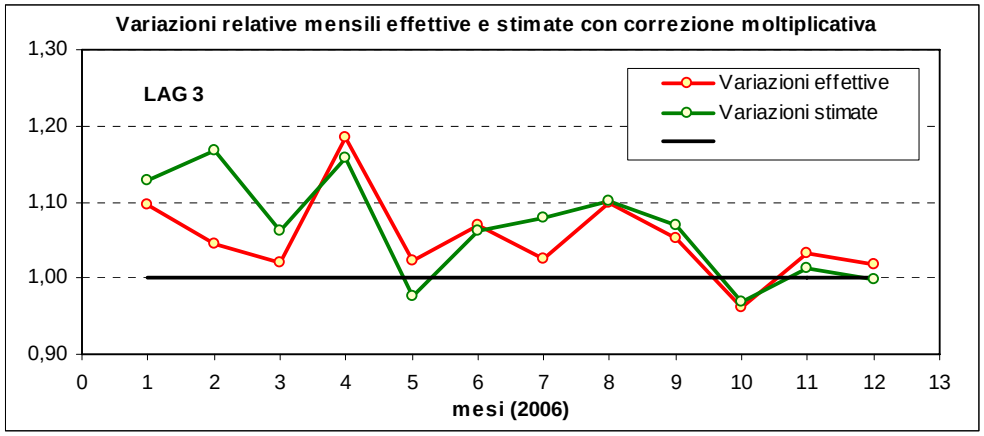
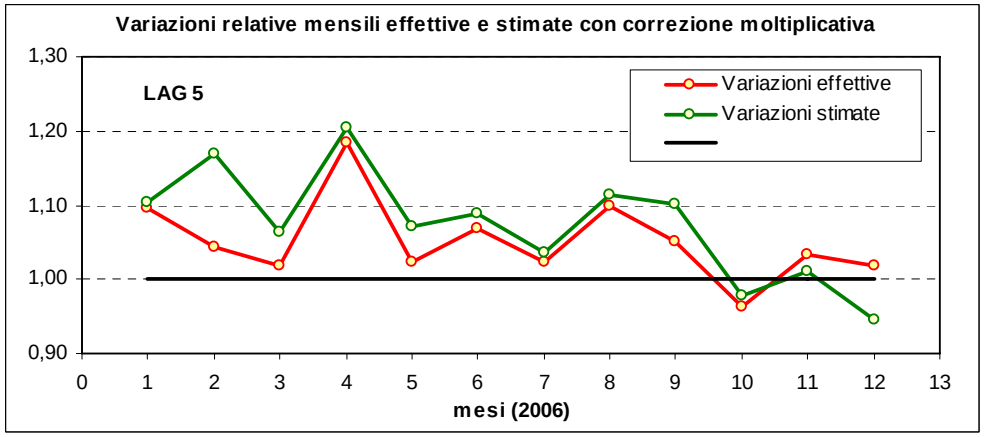
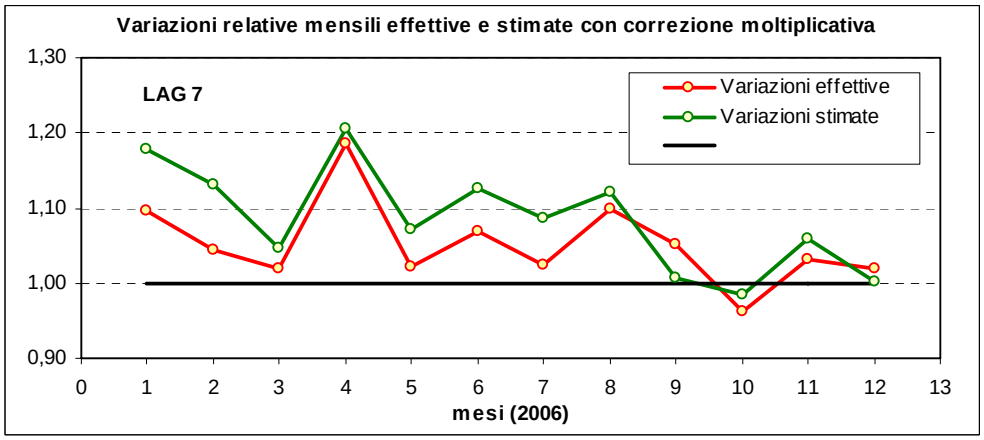
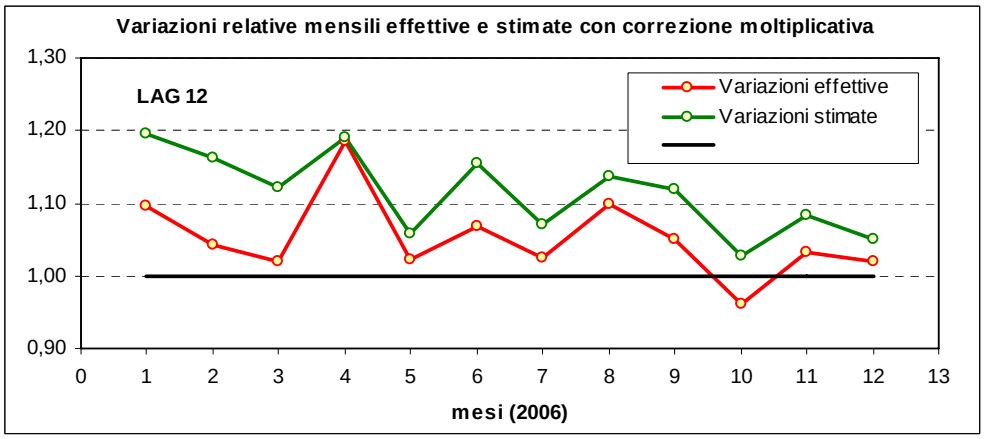


Grafico 6.6 (Pseudo RAPPORTO - Correzione moltiplicativa)









6.4. Confronto fra stimatori di regressione e stimatori propensity scores

E' interessante confrontare sinteticamente i "migliori" procedimenti di stima basati sugli stimatori di regressione generalizzata con i due procedimenti adesso esaminati che si basano sugli stimatori (6.3) e (6.4). Uno dei migliori procedimenti risultava la regressione separata (stelle*qualità) associata alla correzione di Rao moltiplicativa.

Per un confronto sintetico consideriamo la media e la varianza delle 12 differenze fra le variazioni relative mensili effettive e stimate. Con i Grafici 6.7 – 6.10 seguenti si effettua questo confronto.

Grafico 6.7

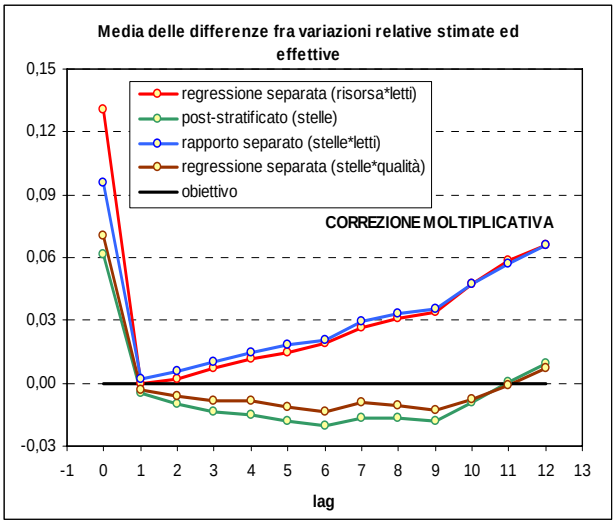


Grafico 6.8

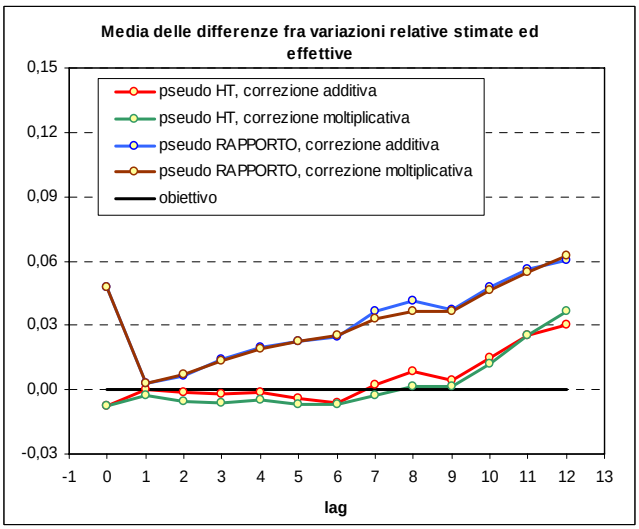


Grafico 6.9

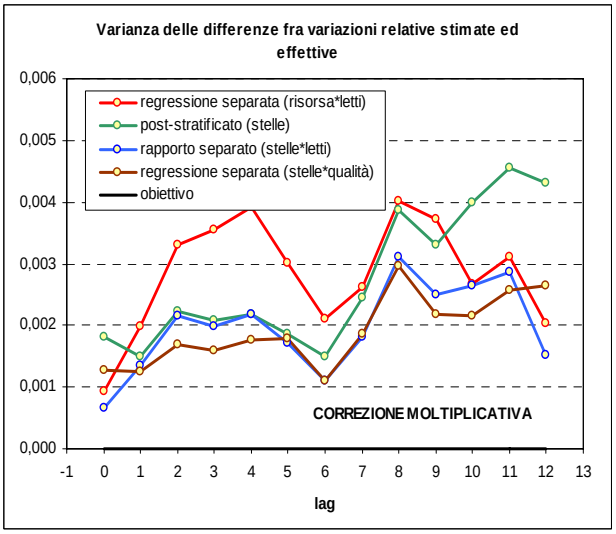
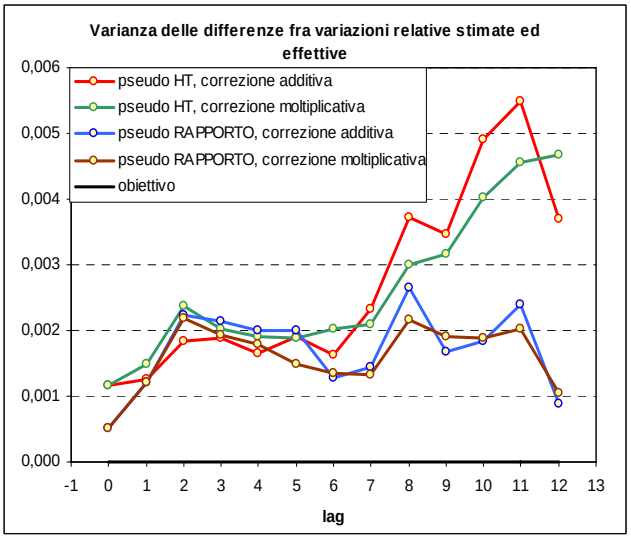


Grafico 6.10



La prima considerazione riguarda la correzione di Rao. Nel procedimento basato sugli stimatori di regressione generalizzata, questa correzione è necessaria perché non sono in grado di eliminare il bias: al lag 0 osserviamo un bias nella stima delle variazioni relative mensile intorno allo 0,06 per il post-stratificato (stelle) e la regressione separata (stelle*qualità), dello 0,13 per la regressione separata (risorsa*letti). Le stime invece ottenute con il metodo dei propensity scores presentano una maggior capacità di contenimento del bias. Con questi stimatori si potrebbero ottenere le stime anticipate delle variazioni relative mensili anche senza applicare la correzione di Rao. Ricordiamo che questa è possibile solo nel caso in cui vi sia una indagine ripetuta.

Una seconda considerazione riguarda il confronto fra pseudo HT e pseudo Rapporto. Ci aspetteremmo dal secondo stime anticipate migliori, in quanto viene inserita anche l'informazione ausiliaria letti, non presente nel modello logistico che stima la probabilità di appartenere al sistema turiweb. Invece, emerge una particolare bontà dello stimatore pseudo HT sia in assenza di correzione di Rao, sia quando questa si estende fino al lag 9. Non emerge una particolare differenza fra l'aggiustamento additivo e quello moltiplicativo.

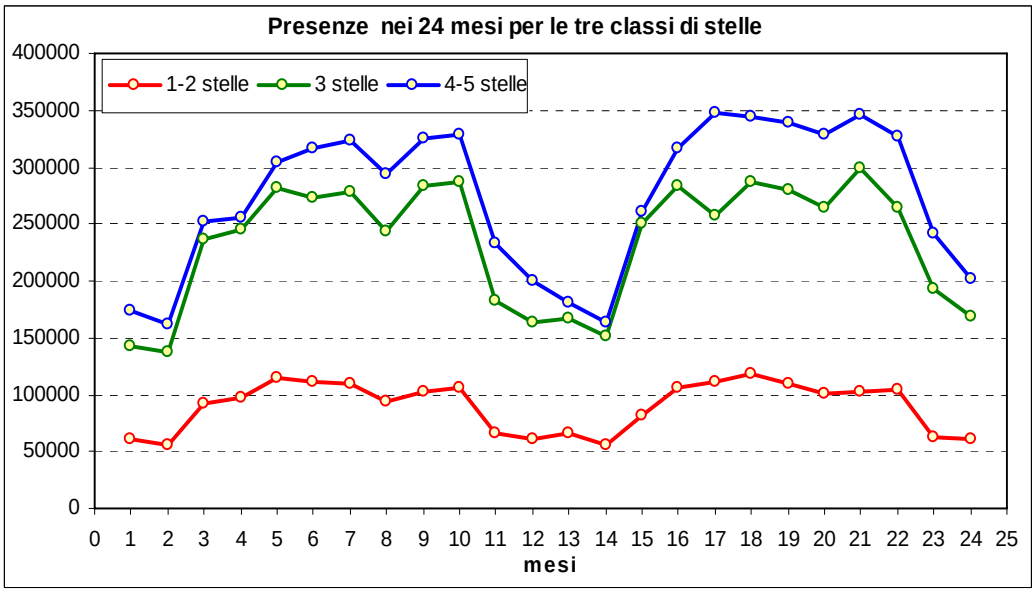
6.5. Un effetto della stagionalità sulla precisione delle stime

Considerando i confronti fra variazioni relative mensili effettive e stimate, si nota, in quasi tutte le prove effettuate, che i maggiori scostamenti si verificano negli ultimi e soprattutto nei primi e mesi dell'anno 2006. Vi possono essere due ragioni.

- Quando gli esercizi ricettivi sono quasi completamente utilizzati (mesi estivi), la relazione fra presenze e dimensione è più forte di quando sono parzialmente utilizzati (mesi invernali).
- Poiché gli esercizi turiweb crescono nel tempo, le stime delle variazioni relative dei primi mesi dell'anno si basano su un numero minore di osservazioni di quelle degli ultimi mesi.

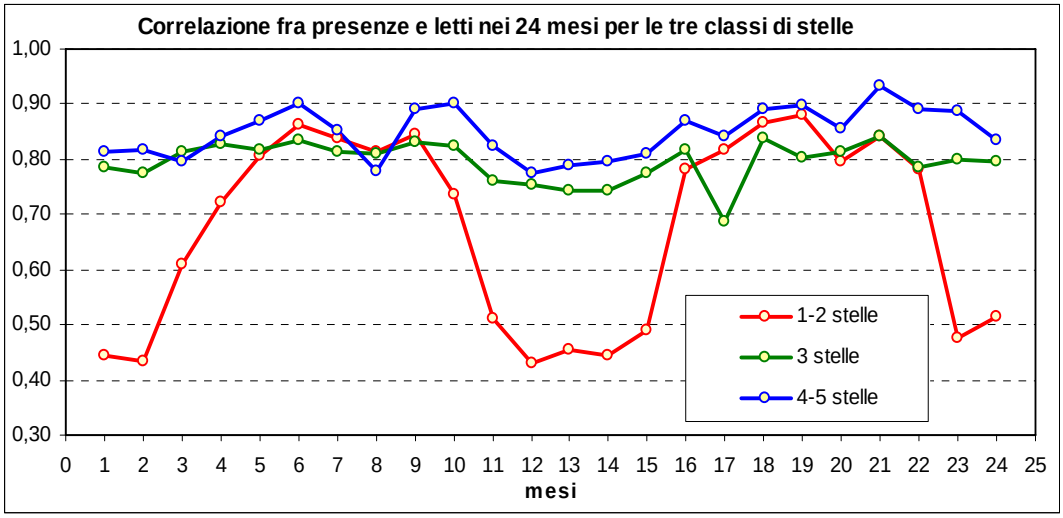
La stagionalità ha un effetto analogo su tutte le categorie alberghiere; le presenze nei mesi di alta stagione sono circa il doppio per le tre classi di stelle che qui consideriamo (Grafico 6.11).

Grafico 6.11



Se però esaminiamo la correlazione fra presenze e dimensione (i letti) il comportamento delle tre classi di stelle si differenzia (Grafico 6.12).

Grafico 6.12



Questo significa che la relazione fra letti e presenze si mantiene assai stretta durante tutto l'anno per gli alberghi di medio-alta categoria: nei mesi di alta stagione questi sono quasi tutti completi, in quelli di bassa stagione quasi tutti occupati per metà; gli alberghi di bassa categoria sono invece quasi tutti completi nel periodo di alta stagione, mentre nel periodo di bassa stagione la loro occupazione è assai differenziata, alcuni alberghi di bassa categoria possono risultare chiusi[1].

Questo fatto spiega, almeno nei casi in cui entra in gioco la relazione con i letti, il motivo della maggior precisione delle stime nei mesi di alta stagione (aprile – ottobre).

[1] Come abbiamo in precedenza segnalato, non abbiamo considerato la chiusura degli esercizi ricettivi perché è una informazione prevalentemente censuaria. Secondo la normativa l'esercizio dovrebbe indicare a ottobre, in occasione della comunicazione dei prezzi, gli eventuali periodi di chiusura per l'anno successivo; però lo stato effettivo si conosce durante lo svolgimento dell'indagine censuaria. Escludere dalla popolazione gli esercizi che hanno dichiarato di esser chiusi poteva condurre al rischio di assumere una popolazione obiettivo più piccola della effettiva. Inoltre i periodi di chiusura possono non coincidere con mesi interi. Comunque il problema meriterebbe un approfondimento.

Capitolo 7

7.1. Il problema della generalizzazione dei risultati

Abbiamo visto nei capitoli precedenti come i diversi metodi considerati siano stati in grado di fornire stime, in alcuni casi anche assai precise, della variazione relativa mensile delle presenze, sia per l'intera popolazione, sia per la sottopopolazione degli esercizi alberghieri, per i mesi del 2006 rispetto ai corrispondenti del 2005.

Sorgono inevitabilmente due domande: La precisione delle stime ottenute sarà tale anche nei mesi futuri? Anche in un'altra provincia? Se non vi è la possibilità di estendere i risultati ottenuti, almeno per la provincia di Firenze nei prossimi mesi, il lavoro svolto è solo un esercizio di verifica a posteriori poco utile per una effettiva applicazione: ci interessa invece applicare le procedure di stima esaminate proprio ai mesi dei prossimi anni, ed eventualmente ad altre province.

Consideriamo le stime ottenute per l'intera popolazione di esercizi ricettivi con il seguente procedimento. Lo stimatore di regressione separata con post-stratificazione per risorsa (Arte, Altra risorsa) e tipologia di esercizio (Alberghi, Altri esercizi), variabile quantitativa posti letto, fornisce la stima preliminare (grezza) del totale mensile. Questa stima viene poi aggiustata con la correzione di Rao moltiplicativa ottenendo la stima anticipata del totale mensile. Il rapporto fra questa stima e il totale noto del corrispondente mese dell'anno precedente conduce alla stima anticipata della variazione relativa mensile. Questo procedimento è risultato uno dei migliori, intendendo con ciò un procedimento che fornisce stime le cui differenze con i valori noti risultano mediamente più piccole di quanto accade con altri procedimenti, nell'insieme delle 12 stime che abbiamo potuto osservare (vedi Grafici 4.15 – 4.18 del cap. 4).

Il procedimento di stima potrebbe esser considerato come un processo che genera la variabile casuale: differenza fra variazione mensile stimata ed effettiva; osserviamo le prime 12 realizzazioni. Possiamo calcolare l'errore quadratico medio (MSE) di queste realizzazioni. Ci piacerebbe disporre di un modello che ci permettesse di descrivere l'MSE in funzione di alcuni aspetti noti di questo processo: la dimensione della popolazione, la dimensione del sottoinsieme delle strutture ricettive del sistema turiweb, il periodo dell'anno, le relazioni fra le variabili ausiliarie disponibili e quella di studio, infine gli algoritmi di stima applicati. Se disponessimo di un tale modello potremmo generalizzare i risultati a situazioni analoghe pur caratterizzate da parametri diversi degli aspetti sopra citati.

Ora i procedimenti applicati iniziano individuando alcuni algoritmi si stima mutuati dall'approccio basato sul disegno; in molti casi tali algoritmi sono gli stessi che si ottengono dall'approccio basato sul modello: rapporto e regressione separati. Tuttavia non siamo nelle condizioni previste dal primo approccio, essendo il campione costituito da un sottoinsieme di volontari; non siamo nemmeno nelle condizioni del secondo approccio, in quanto non possiamo assumere soddisfatta la condizione di indipendenza condizionata del meccanismo di selezione dalla variabile di studio. Sotto queste condizioni vi sono procedure standard per il calcolo dell'incertezza delle stime. Inoltre, nel nostro caso, alle stime ottenute con tali algoritmi viene applicata una ulteriore operazione di aggiustamento che tiene conto dell'errore di stima commesso in precedenti occasioni, errore osservabile grazie alla presenza dell'indagine censuaria. Abbiamo anche visto che la correzione di Rao, riduce, se non proprio elimina il bias, ma fa aumentare la varianza.

7.2. L'incertezza osservata nelle stime anticipate delle variazioni relative mensili

La individuazione di un modello come sopra indicato, ci sembra una operazione ardua. Limitiamoci inizialmente a quantificare l'incertezza osservata e usiamo questa variabilità per giungere a un intervallo di confidenza empirico..

Indichiamo con I_i la variazione relativa mensile effettiva, che abbiamo potuto osservare per i mesi del 2006, data dal rapporto fra il totale delle presenze P_i del mese i il corrispondente totale P_{i-12} dello stesso mese dell'anno precedente:

$$I_i = \frac{P_i}{P_{i-12}} \qquad (7.1)$$

Abbiamo ottenuto la stima anticipata di I_i con:

$$\hat{I}_{i,k} = \frac{\hat{P}_{i,k}}{P_{i-12}} \quad i = 13,\ldots,24; \quad k = 1,\ldots,12 \qquad (7.2)$$

dove $\hat{P}_{i,k}$ indica la stima anticipata del totale (delle presenze) per il mese i ottenuta dalla stima preliminare applicando la correzione di Rao al lag k.

Abbiamo potuto osservare gli errori di stima:

$$\hat{\varepsilon}_{i,k} = \hat{I}_{i,k} - I_i \quad \text{per i = 13, . . ., 24 e k = 1, …, 12} \qquad (7.3)$$

Una stima dell'errore quadratico medio associato al procedimento di stima di I_j (j>i), per un mese successivo a quelli per i quali abbiamo osservato gli errori di stima, potrebbe essere ottenuto con:

$$\hat{MSE}(\hat{\varepsilon}_k) = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{i,k}^2 \qquad (7.4)$$

dove n indica un certo numero di errori di stima osservati al momento in cui si effettua la stima di I_j. Un intervallo di confidenza empirico per il parametro I_j potrebbe essere dato da:

$$CI_k(95\%) = \hat{I}_{j,k} \pm 1{,}96\sqrt{\hat{MSE}(\hat{\varepsilon}_k)} \qquad (7.5)$$

Sorgono alcune questioni:
La prima riguarda l'assunzione che gli errori di stima (7.3) provengano da un processo costante nel tempo e abbiano una distribuzione normale. La seconda riguarda con quali osservazioni stimare l'errore quadratico medio (7.4)

Nel tempo la popolazione evolve e soprattutto il campione dei volontari appartenenti al sistema turiweb cresce: se anche assumiamo per comodità, che gli errori di stima tendano a distribuirsi normalmente, si deve presumere che, crescendo il campione turiweb, gli errori di stima tendano a ridursi per effetto dell'aumento della dimensione campionaria. L'esame dei grafici che riportano il confronto fra i valori I_i e $\hat{I}_{i,k}$ evidenzia questa situazione. Nel momento in cui la dimensione del campione turiweb coincidesse con la dimensione della popolazione, l'errore di stima sarebbe certamente nullo; ma sulla base della (7.5) che dipende dai passati errori di stima osservati, gli assoceremmo un intervallo di confidenza di ampiezza maggiore di 0; questo anche se considerassimo un sistema di ponderazione nella (7.4) con pesi che tengano conto della dimensione campionaria.

La (7.4) e (7.5) sono accettabili nel caso di sostanziale costanza della dimensione del campione turiweb. Quanto al valore di n, cioè a quanti errori di stima passati utilizzare nel calcolo della (7.4) va notato che se n è piccolo lo stimatore (7.4) dell'MSE risulterà instabile perché basato

su poche osservazioni; se n è grande tenderà a sovrastimare l'MSE perché utilizza informazioni "troppo lontane". Se poi il campione turiweb cresce, questa sovrastima sarà ulteriormente accentuata. Probabilmente è meglio sottostare al secondo pericolo piuttosto che al primo.

Dall'esame dei grafici che riportano il confronto fra i valori I_i e $\hat{I}_{i,k}$ risulta abbastanza evidente che gli errori di stima $\hat{\varepsilon}_{i,k} = \hat{I}_{i,k} - I_i$ sono più contenuti nel periodo di alta stagione (aprile – ottobre) e più elevati nel periodo di bassa stagione (gennaio-marzo, novembre-dicembre).

Effettuiamo il calcolo degli intervalli di confidenza (7.5) per due dei procedimenti di stima visti nei capitoli precedenti:

Procedimento 1 (Presenze per l'intera popolazione provinciale degli esercizi ricettivi)
Si ottengono le stime preliminari dei totali mensili con lo stimatore di regressione separata con post-stratificazione per risorsa (Arte, Altra risorsa) e tipologia di esercizio (Alberghi, Altri esercizi), la variabile ausiliaria quantitativa x è costituita dai posti letto. Successivamente si applica la correzione di Rao moltiplicativa ottenendo le stime anticipate dei totali mensili; da queste le stime delle variazioni relative mensili.

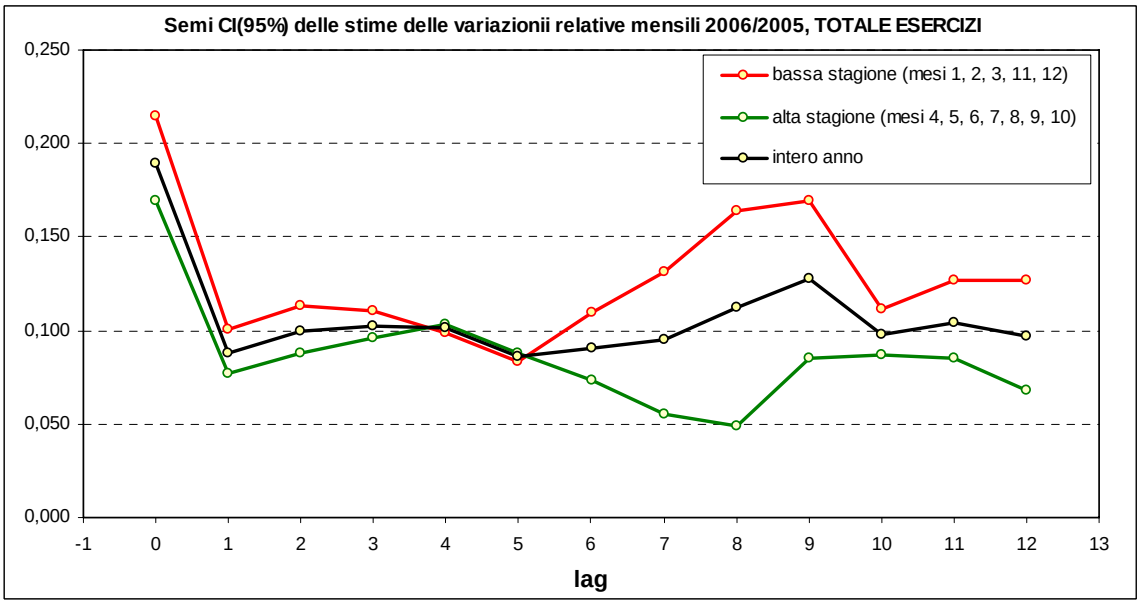
Gli errori di stima (7.3) che abbiamo osservato sono 12, 7 relativi all'alta stagione, 5 alla bassa stagione. Nel grafico 7.1 seguente riportiamo il semi-intervallo di confidenza:

$$\text{semiCI}_k(95\%) = 1{,}96\sqrt{\hat{\text{MSE}}(\hat{\varepsilon}_k)} \qquad (7.6)$$

per il procedimento di stima descritto, per i valori k=0, 1, 2, . . .,12 (i diversi lag di correzione di Rao), e distinguendo fra mesi di bassa e alta stagione.

L' $\hat{\text{MSE}}(\hat{\varepsilon}_k)$ è calcolato con n=12 per l'intero anno, con n=7 e n=5 per i priodi di alta e bassa stagione rispettivamente.

Grafico 7.1



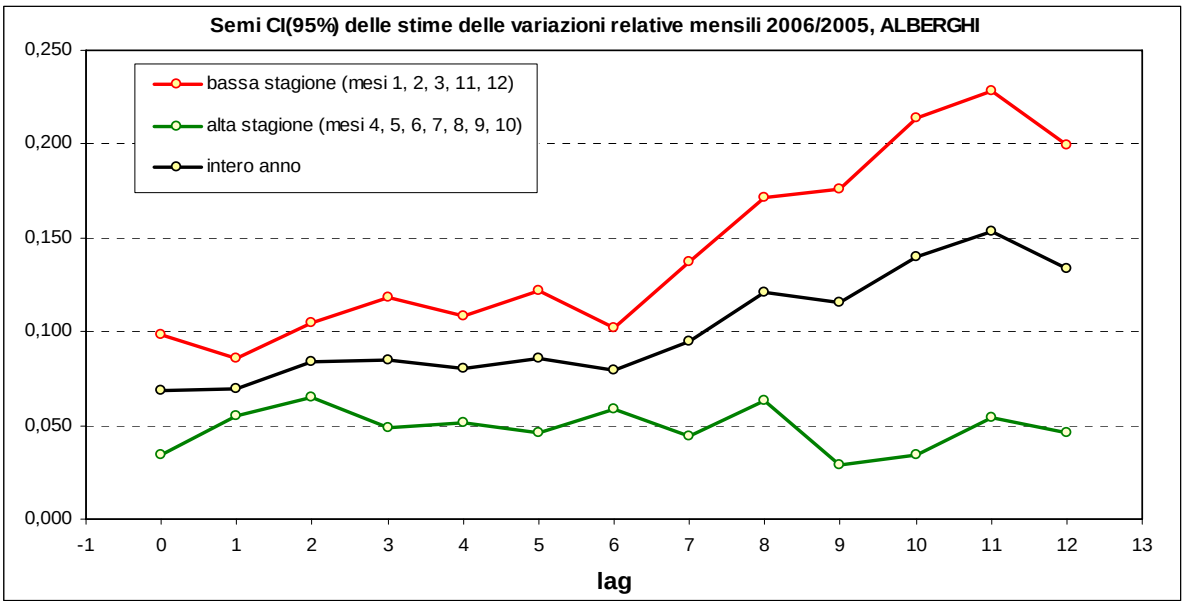
Questo intervallo di confidenza, con i limiti che abbiamo detto, risulta piuttosto ampio per la stima delle variazioni relative mensili. Alla stima della variazione relativa mensile dovremmo aggiungere e togliere circa 10 punti percentuali se applichiamo la correzione di Rao ai lag da 1 a 5. Stime più precise si possono ottenere per le variazioni relative mensili nei mesi di alta stagione nel

caso di lag fra 6 e 12. Avevamo segnalato all'inizio di questo lavoro che il fenomeno ha subito negli anni recenti una evoluzione lenta con variazioni relative mensili di norma inferiori al 10%. L'incertezza delle stime anticipate, così quantificata, risulta dello stesso ordine di grandezza della variazione fisiologica.

Procedimento 2 (Presenze per il dominio Alberghi)
Si ottengono le stime preliminari dei tali mensili con lo stimatore PseudoHT, stimando le probabilità di appartenere al sistema turiweb con un modello logistico. Successivamente si applica la correzione di Rao additiva ottenendo le stime anticipate dei totali mensili; da queste le stime anticipate delle variazioni relative mensili.

Nel grafico 7.2 seguente riportiamo il semi-intervallo di confidenza (7.6) per il procedimento di stima descritto, per i valori k = 0, 1, 2, . . .,12 (i diversi lag di correzione di Rao), e distinguendo, anche in questo caso, fra mesi di bassa e alta stagione.

Grafico 7.2



In questo secondo caso, le stime delle variazioni relative risultano assai precise per i mesi di alta stagione: l'ampiezza del semi-intervallo di confidenza è di 5 punti percentuali, più o meno a tutti i lag di correzione. Nei mesi di bassa stagione, invece, si ottiene una ampiezza del semi-intervallo di 10 punti percentuali per i lag da zero a sei mesi; quando il lag di correzione supera i sei mesi l'incertezza cresce eccessivamente.

Il grafico 7.2 evidenzia di nuovo la bontà del metodo di stima qui ricordato che utilizza lo stimatore PseudoHT. Avevamo evidenziato, nel cap. 6 come le stime preliminari ottenute con questo stimatore potessero essere assunte anche come stime anticipate, senza l'ulteriore applicazione della correzione di Rao che, ricordiamo, richiede la presenza di una indagine periodica. Quanto detto è confermato dal fatto che, nel passare dal lag 0 (assenza di correzione di Rao) al lag 1 non si verifica una riduzione dell'intervallo di confidenza come avviane nel grafico 7.1.

7.3. Stima della varianza dal campione turiweb

Esaminiamo la varianza dello stimatore di regressione separata del totale P delle presenze che si otterrebbe con un disegno stratificato come nel procedimento 1: stratificazione per risorsa (Arte, Altra risorsa) e tipologia di esercizio (Alberghi, Altri esercizi), variabile ausiliaria quantitativa x costituita dai posti letto. Indichiamo con U_g (g =1, 2, 3, 4) i quattro strati; con N_g la dimensione dello strato U_g, con n_g la dimensione del campione $S_g \subseteq U_g$, con $\hat{P}$ lo stimatore di regressione separata adottato nel procedimento 1.

La varianza (approssimata) dello stimatore $\hat{P}$ è data da (Sarndal e altri, 1992):

$$AV(\hat{P}) = \sum_{g=1}^{4} N_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{E_{Ug}}^2 \qquad (7.7)$$

dove:

$$S_{E_{Ug}}^2 = \frac{1}{N_g - 1} \sum_{Ug} E_i^2 \qquad (7.8)$$

con $E_i = P_i - (1, x_i)' \mathbf{B}_g$ per ogni i $\in$ U_g, residui della regressione nello strato U_g, P_i le presenze dell'esercizio i, $(1, x_i)$ il vettore dei regressori per l'esercizio i. $\mathbf{B}_g$ è il vettore dei coefficienti nello strato U_g.

La (7.7) è la varianza (sebbene approssimata) valida se il campione è estratto secondo un disegno stratificato n_g da N_g (g = 1, 2, 3, 4).

Uno stimatore della (7.7) è fornito da:

$$\hat{V}(\hat{P}) = \sum_{g=1}^{4} N_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{ge_{Ug}}^2 \qquad (7.9)$$

dove:

$$S_{ge_{Ug}}^2 = \frac{1}{n_g - 1} \sum_{Sg} g_i^2 e_i^2 \qquad (7.10)$$

$e_i = P_i - (1, x_i)' \hat{\mathbf{B}}_g$ per ogni i $\in$ S_g sono i residui della regressione effettuata con le unità dal campione.

$$g_i = 1 + \frac{n_g (\bar{x}_{Ug} - \bar{x}_{Sg})}{\sum_{Sg} (x_i - \bar{x}_{Sg})^2} (x_i - \bar{x}_{Sg}) \qquad (7.10)$$

sono pesi funzione delle medie della x in U_g e S_g. Siccome di solito i pesi g_i sono prossimi a 1, talvolta si considera una espressione semplificata della (7.9) ponendo $g_i = 1$.

Consideriamo anche le due quantità:

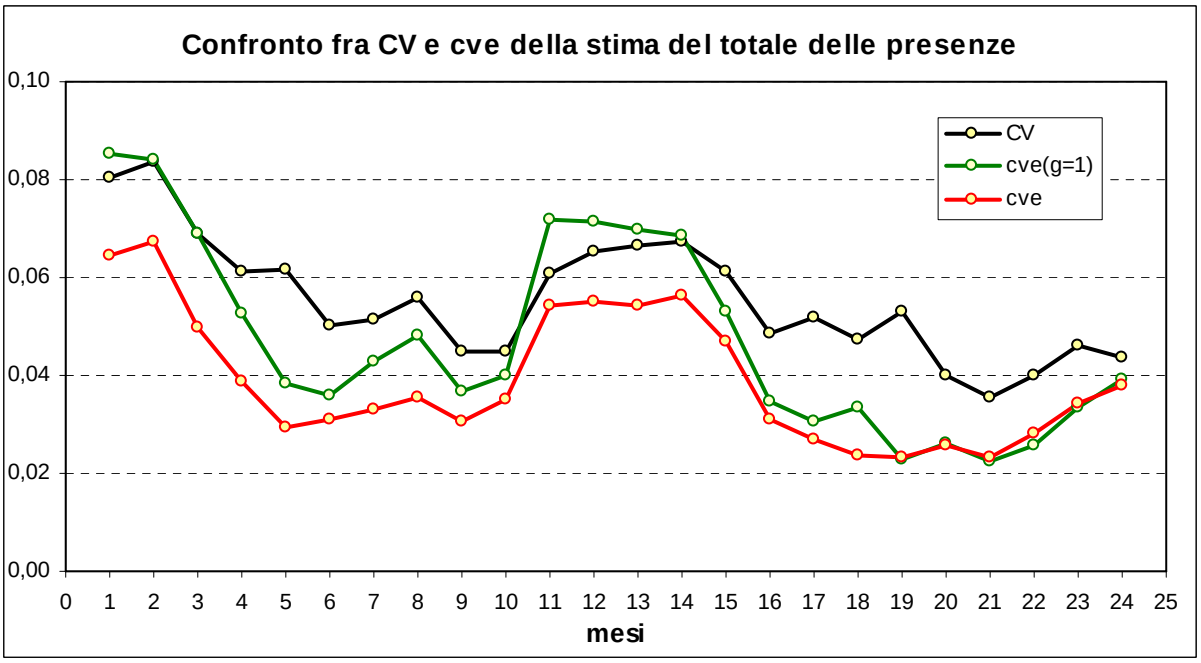
$$CV(\hat{P}) = \frac{\sqrt{AV(\hat{P})}}{P} \qquad (7.11)$$

$$\text{cve}(\hat{P}) = \frac{\sqrt{\hat{V}(\hat{P})}}{\hat{P}} \qquad (7.12)$$

che forniscono il coefficiente di variazione di $\hat{P}$ e una sua stima, rispettivamente. Assumiamo che il procedimento che fornisce le stime anticipate $\hat{P}$ sia corretto. Vediamo che valori assumono le due quantità (7.11) e (7.12) con il nostro campione degli esercizi turiweb.

Nel grafico seguente 7.3 riportiamo tali quantità: per la (7.12) abbiamo considerato i due stimatori della varianza che si ottengono per g_i dato dalle (7.10) e per $g_i = 1$.

Grafico 7.3



I cve ottenuti dal campione turiweb tendono a sottostimare il CV; usando i pesi g_i forniti dalla (7.10), tale sottostima è più accentuata in particolare nel primo anno. E' presente un trend decrescente, dovuto all'aumento della dimensione del campione turiweb e un effetto stagionale: nei mesi di bassa stagione i coefficienti di variazione tendono ad essere più elevati.

Disponendo dell'indagine completa si potrebbe utilizzare CV osservabile alcuni mesi prima, o addirittura 12 mesi prima per fornire una quantificazione dell'incertezza, tenendo conto che quello stimato dal campione turiweb tende a sottostimarlo. Consideriamo la stima della variazione relativa mensile come nella (7.2)

$$\hat{I}_{i,k} = \frac{\hat{P}_{i,k}}{P_{i-12}} \qquad i = 13,\ldots,24; \quad k = 1,\ldots,12$$

dove $\hat{P}_{i,k}$ indica la stima anticipata del totale (delle presenze) per il mese i ottenuta dalla stima preliminare applicando la correzione di Rao al lag k.

Assumiamo l'approssimazione $V(\hat{P}_{i,k}) \cong AV(\hat{P}^{p}_{i-12})$, cioè la varianza della stima anticipata del totale delle presenze nel mese i, sia pressoché uguale alla varianza della stima preliminare dello stesso mese nell'anno precedente. In tal caso:

$$V(\hat{I}_{i,k}) = \frac{V(\hat{P}_{i,k})}{P^2_{i-12}} \cong \frac{AV(\hat{P}^{p}_{i-12})}{P^2_{i-12}} = CV^2(\hat{P}_{i-12}) \qquad (7.13)$$

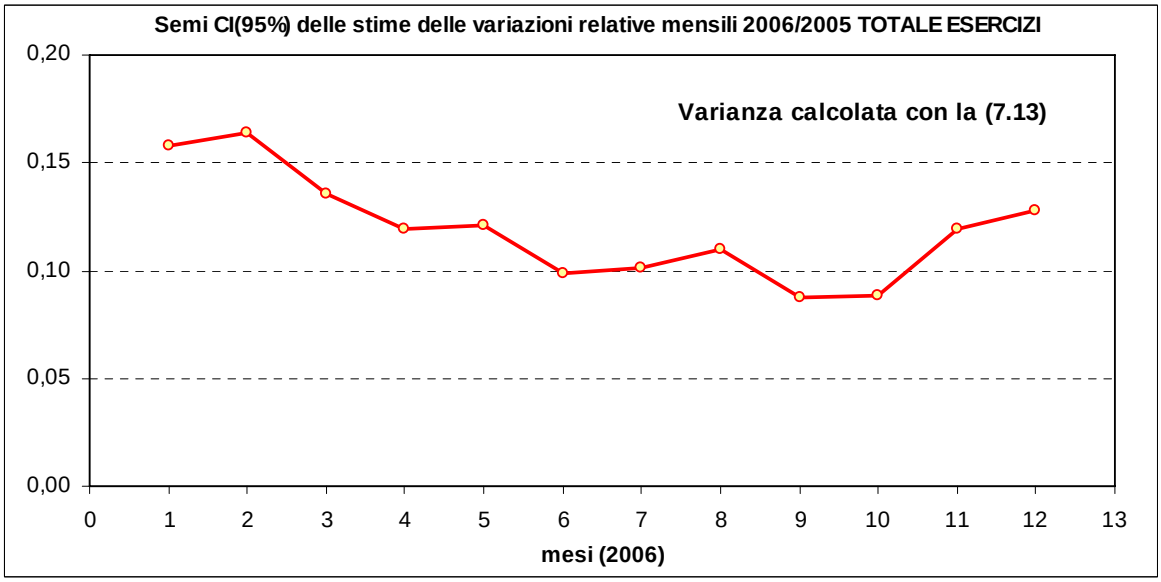
In tal caso un intervallo di confidenza per il parametro I_j potrebbe essere:

$$CI_k(95\%) = \hat{I}_{i,k} \pm 1{,}96 \cdot CV(\hat{P}_{i-12}) \qquad (7.14)$$

Riportiamo nel grafico seguente 7.4 i valori del semi intervallo di confidenza

$$semiCI_k(95\%) = 1{,}96 \cdot CV(\hat{P}_{i-12}) \qquad (7.15)$$

Grafico 7.4



Gli intervalli di confidenza ottenuti con questo secondo metodo risultano più ampi di quelli che abbiamo visto nella sezione 7.2 dove abbiamo utilizzato gli errori di stima delle variazioni relative mensili osservate nei precedenti 12 mesi per calcolare l'MSE della stima attuale della variazione relativa mensile. Va notato che questi ultimi intervalli sono ottenuti calcolando la varianza dello stimatore del totale 12 mesi prima, quando la numerosità del campione turiweb è inferiore a quella su cui si basano le stime. Tuttavia le differenze, soprattutto nei mesi di alta stagione non sono particolarmente grandi.

Va infine notato che questi intervalli di confidenza, hanno un valore indicativo. Quelli ottenuti col metodo della sezione 7.2 si basano sugli errori di stima osservati nei precedenti 12 mesi; quelli ottenuti in questa sezione sulla varianza che lo stimatore di regressione separata avrebbe se il campione fosse selezionato con l'allocazione che realizza il processo di autoselezione. Il calcolo della varianza con la (7.13) dovrebbe condurre a valori più alti perché il campione turiweb che si usa per il calcolo della varianza precede di 12 mesi quello che si usa il calcolo della stima del totale. Risulta quindi plausibile che gli intervalli di confidenza con questo secondo metodo siano più ampi.

Capitolo 8 - Conclusioni

In questo capitolo, effettuiamo alcune considerazioni di sintesi conclusive relative a quattro punti:

- il lavoro svolto,
- i risultati ottenuti,
- i limiti del lavoro,
- possibili approfondimenti e sviluppi.

8.1. Il lavoro svolto

L'obiettivo del lavoro era quello di effettuare stime anticipate del totale delle presenze mensili verificatesi nelle strutture turistico ricettive della provincia di Firenze; in particolare agli utenti delle statistiche sul turismo interessa conoscere tempestivamente la variazione relativa mensile, ovvero la quantità:

$$I_i = \frac{P_i}{P_{i-12}}$$

dove P_i indica il totale delle presenze nel mese i, P_{i-12} il totale delle presenze nel corrispondente mese dell'anno precedente. L'utilità di stime anticipate deriva dal fatto che l'indagine completa richiede molti mesi per fornire i risultati.

Un sottoinsieme di esercizi ricettivi della provincia, partecipando al sistema turiweb, è in grado di trasmettere alla Provincia, pochi giorni dopo la fine del mese, i dati sul movimento che nel mese si sono verificati nella propria struttura. La partecipazione al sistema turiweb è volontaria. Nel tempo sta crescendo il numero dei partecipanti (Cap. 1).

Nel nostro lavoro abbiamo inteso sperimentare la possibilità di effettuare stime anticipate dei totali mensili delle presenze utilizzando il campione delle strutture ricettive del sistema turiweb.

Abbiamo esaminato le due impostazioni dell'inferenza su parametri descrittivi di una popolazione finita e i vincoli che impongono nella selezione del campione per la validità dell'inferenza. Nell'impostazione basata sul disegno si richiede la selezione (e l'osservazione!) di un campione secondo un prefissato disegno. Nella impostazione basata sul modello, la selezione sarebbe trascurabile se si disponesse di un modello esatto per la variabile di studio, purché tale selezione sia indipendente dalla variabile di studio. In pratica per salvaguardarsi dalla malspecificazione del modello è necessario selezionare un campione secondo precisi criteri, ad esempio bilanciato. Nel nostro caso non potevamo agire sulla selezione del campione, e quindi entrambe le impostazioni non fornivano strumenti standard per l'inferenza (Cap. 2).

Abbiamo sperimentato due procedimenti di stima della variazione relativa mensile.

Procedimento 1

Abbiamo effettuato le stime preliminari dei totali mensili applicando alcuni stimatori di regressione generalizzata, mutuati dall'impostazione basata sul disegno, che sfruttassero l'informazione ausiliaria disponibile, nella speranza che la loro efficienza fosse in grado se non di eliminare, almeno di contenere il bias da autoselezione[1]. Una volta ottenute le stime preliminari,

[1] Se intendiamo l'efficienza di uno stimatore come la sua capacità di fornire stime "poco variabili" da un campione all'altro, allora uno stimatore molto efficiente tende a fornire stime vicine al parametro quasi con tutti i campioni, quindi a contenere l'eventuale e ignoto bias da autoselezione.

grazie alla presenza dell'indagine completa, abbiamo potuto osservare gli errori di stima commessi nei mesi precedenti. Sulla base di questi errori di stima abbiamo effettuto una correzione delle stime preliminari (correzione di Rao) ottenendo così le stime anticipate dei totali mensili, ed infine le stime anticipate delle variazioni relative mensili (Cap. 3 e Cap. 4).

Abbiamo applicato il procedimento 1 sia alla stima del totale per l'intera popolazione popolazione degli esercizi ricettivi della provincia, sia al dominio costituito dalle strutture alberghiere (Cap. 5). Per questo dominio disponevamo di una informazione ausiliaria più ricca.

Procedimento 2

Per il dominio alberghi, abbiamo sperimentato inoltre la stima del totale mensili delle presenze utilizzando i propensity scores. Abbiamo stimato le probabilità di appartenere al sistema turiweb con un modello logistico utilizzando un set di covariate che riteniamo abbiano una capacità esplicativa del meccanismo di selezione. Ottenute queste probabilità, abbiamo costruito tre stimatori (pseudo HT, pseudo rapporto, psedo stratificato). I primi due sono analoghi allo stimatore per espansione di Horvitz e Thompson (HT) e allo stimatore rapporto del totale, dove al posto delle probabilità di inclusione del 1° ordine vengono sostituite le probabilità stimate di appartenere al sistema turiweb. Il terzo è uno stimatore mutuato dagli studi osservazionali (Rosenbaum e Rubin, 1984). Con questi stimatori abbiamo ottenuto le stime preliminari dei totali mensili. Abbiamo, anche in questo caso, applicato la correzione di Rao ottenendo le stime anticipate dei totali mensili e quindi le stime anticipate delle variazioni relative mensili. Lo stimatore pseudo HT e pseudo rapporto hanno fornito stime preliminari già molto buone, tali da poter essere usate come stime anticipate anche senza applicare la correzione di Rao che richiede la presenza di una indagine ripetuta (Cap. 6).

8.2. I risultati ottenuti

Per valutare i risultati ottenuti abbiamo confrontato le stime anticipate delle variazioni relative mensili stimate con quelle effettive. Ciò è stato possibile grazie alla presenza dell'indagine completa. Abbiamo lavorato con i dati dell'indagine completa per gli anni 2005 – 2006. Per questi due anni abbiamo avuto a disposizione sia i dati dell'indagine completa sia i dati degli esercizi appartenenti al sistema turiweb. Abbiamo così potuto osservare quanto le stime effettuate con i due procedimenti sopra sintetizzati differissero dai valori ottenibili osservando tutta la popolazione oggetto di studio.

L'esame dei risultati ottenuti conduce alle seguenti considerazioni.

- Gli stimatori di regressione generalizzata tendono a limitare il bias da autoselzione; fra i diversi sperimentati, quelli di regressione separata sono risultati i migliori.
- La correzione di Rao tende a ridurre il bias da autoselezione, ma produce un incremento di varianza. Andando indietro nel tempo il suo beneficio si riduce.
- Per gli esercizi alberghieri, dove abbiamo sperimentato il procedimento 1 e il procedimento 2, è emerso che gli stimatori pseudo HT e pseudo rapporto sono in grado di quasi eliminare il bias da autoselezione potendosi così candidare a strumenti di stima anche in assenza dell'indagine completa. Infine le stime delle variazioni relative mensili ottenute con pseudo HT e correzione di Rao sono risultate molto buone.

Abbiamo riportato in numerosi grafici il confronto fra variazioni relative mensili stimate ed effettive. I risultati sono assai buoni soprattutto per i mesi di alta stagione e vanno migliorando nel tempo per effetto dell'aumento della numerosità del campione turiweb.

Nell'ultimo capitolo (Cap. 7) abbiamo infine proposto due misure empiriche dell'incertezza delle stime: la prima basata sugli errori di stima osservati nei mesi passati, la seconda sulla varianza dello

stimatore calcolata con i dati dell'indagine completa 12 mesi prima. Per stime che utilizzano i propensity scores i risultati sono assai buoni: con un piccolo campione di alberghi autoselezionato si riescono ad ottenere stime delle variazioni relative mensile con una incertezza soddisfacente. Questo è il risultato migliore del lavoro svolto.

8.3. I limiti del lavoro

Confrontando le variazioni relative mensili stimate con quelle effettive abbiamo detto che i risultati sono assai buoni. Abbiamo anche misurato empiricamente quanto le variazioni relative mensili si discostano da quelle stimate. Il problema che non ha trovato soddisfacente soluzione è quello della validità delle inferenze. In sostanza, anche ammettendo che l'applicazione della correzione di Rao conduca a stime pressoché corrette, almeno se il ritardo è di pochi mesi, resta iirisolto il problema della stima della varianza.

I procedimenti di stima che abbiamo applicato e che abbiamo visto fornire determinati errori stima, andranno altrettanto bene nei prossimi mesi? In un'altra provincia?

Il nostro lavoro ha un carattere empirico. La qualità delle stime effettute è valutata sulla base della conoscenza dei parametri che ci proponiamo di stimare. Abbiamo osservato una gerarchia di procedimenti di stima valida per questa popolazione e questo campione autoselezionato. Abbiamo anche ricavato alcune misure dell'incertezza. I metodi da noi utilizzati sono prevalentemente mutuati da quelli applicati per la non risposta nell'ambito dell'inferenza basata sul disegno. Ora non esiste una teoria completamente soddisfacente per risolvere questo problema. Sostanzialmente vi sono due approcci: il primo è quello di usare stimatori di regressione generalizzata nella speranza che, fornendo stime poco variabili da un campione all'altro, siano per ciò poco sensibili all'autoselezione. Il secondo è quello che cerca di stimare la probabilità di risposta e per questa via ricondursi a stimatori del tipo Horvitz e Thompson.

D'altro canto, quando l'obiettivo sono le stime anticipate, ove si dispone di una indagine completa, è possibile osservare gli errori di stima commessi nel passato e quindi applicare alcune forme di correzione, come abbiamo fatto.

Semmai, ci sarebbe piaciuto valutare se l'informazione fornita dall'indagine completa potesse essere sfruttata in modo più efficace, utilizzando qualche modello per le serie storica degli errori di stima che possiamo osservare. La nostra incompetenza dell'analisi delle serie storiche ci ha impedito questo approfondimento.

8.4. Possibili approfondimenti e sviluppi

Abbiamo effettuato stime prevalentemente per i totali mensili delle presenze, in qualche caso per gli arrivi. Abbiamo inoltre effettuato un approfondimento per il dominio alberghi. Si tratta delle informazioni più richieste. L'interesse degli utenti è però rivolto anche a stime per domini territoriali; ad esempio località d'Arte, balneari, ecc. Inoltre interessa una articolazione del totale almeno per le due provenienze della clientela: Italiani, Stranieri.

La stima del totale delle presenze in diversi domini non dovrebbe essere particolarmente critica: abbiamo visto che per il dominio alberghi dove disponevamo, alla fine del 2006 di un campione turiweb di circa 130 unità su una popolazione di 600 si sono potute ottenere stime abbastanza buone. La stima per le due variabili: presenze di clienti italiani, presenze di clienti stranieri, potrebbe essere più difficile per il seguente motivo: le variabili ausiliarie dimensione (letti)

e tipologia esercizio, hanno rilevante capacità esplicativa del volume di clientela assorbito; ma potrebbero avere una scarsa capacità esplicativa delle presenze di una specifica provenienza. Questo tema meriterebbe una esplorazione.

I risultati ottenuti, per la loro natura empirica, richiedono un monitoraggio nel tempo. Individuati i procedimenti di stima (stime preliminari con stimatori di regressione generalizzata, correzione di Rao delle stime preliminari) occorre verificare nel tempo il loro comportamento; non é detto che, evolvendo la popolazione e il campione turiweb, le gerarchie individuate si mantengano nel tempo.

Il procedimento di stima che utilizza i propensity scores si è rivelato molto efficiente. Lo abbiamo sperimentato solo per il dominio alberghi; andrebbe provato anche per l'intera popolazione degli esercizi ricettivi recuperando la maggior informazione possibile che può spiegare il meccanismo di autoselezione.

Infine occorre approfondire la ricerca per utilizzare al meglio l'informazione fornita dalla presenza dell'indagine censuaria. Probabilmente un qualche modello per la serie storica degli errori di stima osservati nel passato potrebbe condurre a una correzione migliore e a una stima più precisa della varianza delle stime anticipate.

Per il procedimento di stima che utilizza i propensity scores occorre ricercare strumenti che permettano di stimare l'incertezza delle stime. Come abbiamo visto questo metodo potrebbe essere applicato anche in assenza dell'indagine completa e quindi configurarsi come uno strumento per la stima di parametri descrittivi della popolazione finita a partire da un campione autoselezionato. Una strada potrebbe essere quella di utilizzare le probabilità di autoselezione stimate e con queste ricondursi a uno schema analogo al disegno di Poisson.

Riferimenti bibliografici

Baldi, C., Ceccato F., Pacini S., Tuzi D., (2005), *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*, Contributi ISTAT.

Barnett, V. (1999). *Comparative Statistical Inference*. John Wiley, Chichester.

Biffignandi, S., Pratesi, M., (2000), *Le indagini via internet sulle imprese: aspetti metodologici e un'analisi dei rispondenti*, in: Tecnologie informatiche e fonti amministrative nella produzione di dati, A cura di C. Filippucci, F. Angeli.

Biffignandi, S., Pratesi, M., (2004), *L'utilizzo dei propensity scores per l'inferenza in indagini Web: potenzialità alla luce di uno studio di simulazione*, in. Tecniche di integrazione di dati, a cura di. Pallara, A., Falorsi, P., Franco Angeli, Milano.

Borges, J. L., (1984), *Del riposo della scienza*, in *Tutte le opere*, Mondadori, Milano.

Cochran, W., (1977), *Sampling Techiques*, New York, John Wiley.

Copas, A. J., Farewell, V. T., (1998), *Dealing with non-ignorable non-response by using an "entusiasm-to-respond" variable*, Journal of Royal Statistics Society, vol. 161, Part 3, pp. 385-396.

Couper, M. P., (2000), *Web Surveys : A review of Issues and Approaches*, The Public Opinion Quarterly, Vol. 64, n°. 4, pp. 464-494.

Cox, B. G., (1995), *Business Survey Methods*, New York, John Wiley.

de Cristofaro, R. (1998). *La Logica della Statistica*. G. Cipparelli, Torino.

Deville, J. C., (1991), *A Theory of Quota Surveys*, Survey Methodololgy, Vol. 17, n° 2, pag. 163-181.

Deville, J. C., Särndal, C. E., (1992), *Calibration Estimators in Survey Sampling*, J. A. S. A**,** 87, pp. 376-382.

Bacchini, F., Falorsi, P.D., Iannaccone, R., (2006), *Shrinkage estimation with preliminary data*, Atti della XLIII Riunione scientifica SIS.

Fabbris, L., a cura di., (1990), *Atti del seminario di studio su "Rilevazione per campione delle opinioni degli italiani"*, Bressanone, 13 set. 1990, SGEditoriali, Padova.

Finney, D. J., (1974), *Problems, data and inference: the address of the President (with Proceedings)*, Journal of Royal Statistics Society, B, vol. 31, pp. 195-233.

Gini, C., Galvani, L., (1929), *Di una applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione (1° dicembre 1921)*, Annali di statistica Serie VI, Vol. IV, ISTAT, Roma.

Giommi, A., (1985), *On Estimation in Nonresponse Situations*, Statistica, Anno XLV, n°. 1.

Giommi, A., (1987), *Nonparametric Methods for Estimating Individual Response Probabilities*, Survey Methodology, Vol. 13, n. 2, pp. 127-134.

Godambe, V. P., (1955), *A Unified Theory of Samplig from Finite Populations*, Journal of the Royal Statistical Society, B, vol. 17, pag. 269-278.

Hansen, H. M., Madow, W. G., Tepping, B. J., (1983), *"An evaluation of model-dependent and probability-sampling inferences in sample surveys*, ***J. A. S. A,*** 78, pp. 776-793.

Isaksson A., Danielsson, S., Forsman, G., (2004), *On the Variability of Estimates Based on Propensity Score Wighted Data from Web Panels*, Procedings Joint Statistical Meeting, Atlanta, A.S.A.

Lessler, J. T., Kalsbeek, W. D., (1992), *Nonsampling Error in Surveys*, New York, John Wiley.

ISTAT, (2001), *Indagini sociali telefoniche, Metodologia ed esperienze della statistica ufficiale*, Metodi e Norme, n° 10, Roma.

Neyman, J., (1934), *"On the Tho Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection."*, Journal of the Royal Statistical Society, A, pp. 558-625.

Pascal, B., (1982), *Pensieri*, trad. it. P. Serini, III° ed., Mondadori, Milano.

R Development Core Team (2007). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J. N. K., Srinath, K. P., Quenneville, B., (1989), *Estimation of level and change using current preliminary data*, in Panel Surveys, (Kasprzyk, Duncan, Kalton G., Singh eds. Wiley New York), pag. 457-485

Rao, J. N. K., (1997), "Developments in sample survey theory: an appraisal", *The Canadian Journal of Statistics*, Vol. 25, n° 1, pp. 1-21.

Rosenbaum, P. R., Rubin, D. B., (1983) *The Central Role of the Propensity Score in Observational Studies for Casual Effects*, Biometrika, 70, pp. 41-55.

Rosenbaum, P. R., Rubin, D. B., (1984) *Reducing Bias in Observational Studies Using Subclassication on the Propensity Score*, J.A.S.A. Vol. 79, n°. 387, pp. 516-524.

Royall, R. M., (1970), On finite population sampling theory under certain linear regression models, Biometrika, vol 57, 2, pag. 377-387.

Särndal, C. E., Swensson, B., (1987), *A General View of Estimation for Two Phases of Selection with Application to Two-Phase Sampling and Nonresponse*, International Statistical Review , 55, 3, pp. 279-294.

Särndal, C. E., Swensson, B., Wretman, J., (1992), *Model Assisted Survey Sampling*, New York, Springer-Verlag.

Särndal, C. E.,Lundström, S., (2005), *Estimation in Surveys with Nonresponse*, Wiley, New York.

Scaffai, G. (1995), *L'efficacia del grafico per lo studio della distribuzione statistica*, Tesi di laurea, Università degli studi di Firenze, a.a. 1994/1995.

Smith, T. M. F., (1983) *On the Validity of Inference from Non-random Samples*, Journal of Royal Statistics Society, vol. 146, Part A, pp. 394-403

Smith, T. M. F., (1984) *Present Position and Potential Developments: Some Personal Views Sample Surveys*, Journal of Royal Statistics Society, vol. 147, Part A, pp. 208-221.

Thompson, M., E. (1997), *Theory of Sample Surveys*, London, Chapman & Hall.

Valliant, R., Dorfman, A. H., Royall, R. M., (2000), *Finite Population Sampling and Inference*, Wiley, New York.

---------- ♣ ----------