rivista di statistica ufficiale

n. 1
2006

Istat
Istituto nazionale di Statistica

rivista di statistica ufficiale

n. 1
2006

Istat

Comitato di redazione della Rivista di Statistica Ufficiale
c/o Dipartimento per la Produzione Statistica ed il Coordinamento Tecnico Scientifico,
Via Cesare Balbo, 16 - 00184 Roma
tel.: 06.46732774 – fax: 06.47888069
e-mail: rivista@istat.it, cadeli@istat.it

rivista di statistica ufficiale

n. 1/2006

Istituto nazionale di statistica
Via Cesare Balbo, 16 - Roma

Coordinamento editoriale:
Servizio Produzione editoriale

Videoimpaginazione:
Raffaella Rose

Copertina:
Maurizio Bonsignori

Stampa:
Rubbettino Industrie Grafiche ed editoriali
Soveria Mannelli (CZ)

Dicembre 2007 - Copie 300

Editoriale

Col numero 1 del 2006 riprende la diffusione quadrimestrale della Rivista di Statistica Ufficiale. Rinnovata nell'organizzazione dei contenuti oltre che nel layout grafico, la pubblicazione torna ad essere stampata e distribuita direttamente dall'Istituto nazionale di statistica, dopo una fruttuosa collaborazione con Franco Angeli che ne ha curato l'edizione per sei annualità.

Un primo esito della riappropriazione del ruolo di editore da parte dell'Istat è la decisione di rendere disponibile on line sul sito web dell'Istat, la Rivista in versione integrale a titolo gratuito. Ogni numero verrà inoltre annunciato con una newsletter ad una platea di lettori interessati o potenzialmente interessati. Tutto ciò dovrebbe garantire una maggiore visibilità al lavoro degli autori e allargare il confronto con altri esponenti del mondo della ricerca.

Ci auguriamo che queste novità siano accolte positivamente. Esse sono il preludio di una fase di transizione verso un prodotto editoriale che sarà radicalmente rinnovato e che verrà posizionato più adeguatamente nello scenario internazionale, per valorizzare con maggiore efficacia la ricerca nell'ambito della statistica ufficiale. Si tratterà di un nuovo spazio di dibattito scientifico al cui interno promuovere il confronto e lo scambio sui risultati e sui progressi che si realizzano nei processi e nell'analisi dell'informazione statistica del nostro Paese. Un'area di discussione aperta ai contributi dei ricercatori e dei tecnici che su questo terreno stanno lavorando, sia negli enti scientifici, sia nelle istituzioni statistiche nazionali e internazionali.

La redazione

La previsione della disoccupazione nelle regioni italiane attraverso il modello stock e flussi. Costruzione del database e primi risultati[1]

Davide Di Laurea[2], Riccardo Gatto[3], Monica Gentile[4], Alessandra Righi[5], Andrea Spizzichino[6], Leonello Tronti[7]

Sommario

Il lavoro presenta le attività di ricerca e previsione svolte per il nuovo modulo di Previsioni regionali del mercato del lavoro del Progetto MARSS (Modello di analisi regionale della spesa sociale). Dopo un inquadramento della letteratura di riferimento e delle principali caratteristiche dell'approccio seguito, viene descritta la fase di predisposizione delle informazioni necessarie all'approccio adottato, ovvero l'elaborazione del database delle serie storiche (di stock e di flusso) relative ai mercati del lavoro regionali dal 1977 al 2003 e degli indicatori rilevanti per l'approccio longitudinale.
Vengono descritti il modello di attrazione-scoraggiamento utilizzato per la previsione delle entrate nelle forze di lavoro e della disoccupazione e le verifiche effettuate su diverse forme funzionali. La capacità predittiva del modello prescelto viene valutata attraverso test condotti sulle tendenze nazionali, ripartizionali e regionali (in particolare su Toscana e Piemonte), relative al periodo 1997-2003. Si presentano infine i risultati delle previsioni della disoccupazione per genere e disaggregazione territoriale entro la serie (tra il 1997 e il 2003), che vengono discussi alla luce delle questioni metodologiche ancora aperte. Conclude il lavoro un'appendice con un esempio di previsione fuori della serie.

Abstract

The paper presents the research activities carried out in the construction of the new forecasting model of Italian regional labour markets in the context of the MARSS Project (Regional Social Expenditure Model of Analysis). After briefly explaining the information

[1] Gli autori ringraziano sentitamente il prof. Michele Bruni dell'Università di Modena e Reggio Emilia per la collaborazione e il sostegno accordati al Gruppo di ricerca, pur restando pienamente responsabili di ogni errore così come delle valutazioni espresse nel lavoro. Sebbene il lavoro sia frutto dell'opera di tutti gli autori, così come l'introduzione e il paragrafo 6, il paragrafo 2 è prevalentemente attribuibile a Leonello Tronti, il paragrafo 3.1 a Riccardo Gatto e Andrea Spizzichino, il paragrafo 3.2 a Davide Di Laurea, il paragrafo 4.1 ad Alessandra Righi e Leonello Tronti, il paragrafo 4.2 ad Alessandra Righi, Leonello Tronti e a Monica Gentile, i paragrafi 4.3 e 5 a Monica Gentile, l'Appendice ad Alessandra Righi e a Monica Gentile. Versioni preliminari di questo lavoro sono state presentate al XIX Convegno Nazionale di Economia del Lavoro, Facoltà di Economia Marco Biagi, Università di Modena e Reggio Emilia, Modena, 23 e 24 Settembre 2004 e alla 92nd International Conference on Applied Econometrics Association, *Labour Market Polices and Unemployment*, Università Napoli Parthenope, Napoli, 1-2 Giugno 2006.

[2] Ricercatore (Istat), e-mail: dilaurea@istat.it.

[3] Ricercatore (Istat), e-mail: rigatto@istat.it.

[4] Ricercatore consulente dell'Irpet Toscana, e-mail: gentile.monica@gmail.com.

[5] Primo ricercatore (Istat), Coordinatore del Gruppo di lavoro Istat sulle "Previsioni del mercato di lavoro del Progetto MARSS", e-mail: righi@istat.it.

[6] Ricercatore consulente dell'Ires Piemonte, e-mail: andrispizz@yahoo.it.

[7] Dirigente di ricerca (Istat), e-mail: tronti@istat.it.

needs that have led to planning the new forecasting model of the labour market, and recalling the literature on workers' flow models and the main characteristics of the chosen model, the paper deals with the database used. This is a reconstruction/alignment of the LFS regional time series by single year of age of the workers, relative to the 1977-2003 period, that can be organised in cohorts. The treatment allows for the computation of many relevant stock and flow indicators.
The paper then describes the forecasting model and the many experiments carried out to model labour market inflows and outflows, based on various functional forms. The forecasting ability is tested within the series for labour force inflows and, consequently, for the level of unemployment per gender, in the whole economy, the two large geographical areas and the regions Toscana and Piedmont. Comments on results and on open issues conclude the paper together with an appendix with an example of forecast outside the series.

Parole chiave: disoccupazione, regioni, flussi, previsione

1. Introduzione

Questo lavoro presenta i risultati delle attività di ricerca e previsione svolte per la predisposizione del modulo di Previsioni regionali del mercato del lavoro nel quadro del Progetto MARSS (Modello di analisi regionale della spesa sociale)[8]. Il Progetto è frutto della collaborazione tra Ires Piemonte, Irpet Toscana e Istat avviata nel 1998, allargata all'Agenzia Umbria Lavoro e ad Abruzzo Lavoro nel 2004 e interrottasi nel 2005.

Allo scopo di soddisfare le esigenze di previsione regionale, il Progetto ha ritenuto di giovarsi della fruttuosa esperienza di previsione, con modelli stock e flussi, della disoccupazione a livello territoriale nazionale e ripartizionale maturata in Italia sin dagli anni '80, verificando se essa poteva essere utilmente estesa anche al livello regionale. Questa scelta, basata su considerazioni di carattere teorico circa la pertinenza dell'approccio di flusso per lo studio dei mercati del lavoro, è divenuta possibile perché si è riusciti a costruire una nuova base di dati di migliore qualità e di maggiore ampiezza temporale rispetto a quelle già in uso.

Dopo un breve inquadramento della letteratura di riferimento e delle principali caratteristiche dell'approccio seguito, vengono presentati i risultati dell'attività di costruzione delle informazioni necessarie a corrispondere all'approccio adottato, e viene quindi sintetizzato il processo di elaborazione che ha portato alla costruzione delle serie storiche, sia di stock che di flusso, relative al mercato del lavoro dal 1977 al 2003 e poi al calcolo degli indicatori relativi all'approccio longitudinale.

Viene quindi descritto il modello di attrazione-scoraggiamento utilizzato per la previsione delle entrate nelle forze di lavoro e della disoccupazione e si presentano le verifiche effettuate su

[8] Il Modello MARSS stima l'evoluzione di medio-lungo periodo a livello regionale della spesa sociale sulla base degli scenari esogeni definiti dalle stime demografiche dell'Istat e dalle proiezioni macroeconomiche disponibili (Casini Benvenuti, Paniccià, 2003). Il livello regionale dell'analisi e la multisettorialità del modello, e quindi l'integrazione fra i diversi settori di intervento pubblico, sono gli elementi di qualificazione di MARSS. Nell'ambito del progetto, il modulo mercato del lavoro occupa una posizione rilevante, in quanto, utilizzando le informazioni sulla popolazione per titolo di studio derivanti dal modulo di previsione dell'istruzione, fornisce l'input per il successivo modulo di previsione della spesa pensionistica.

diverse forme funzionali. Per valutare la capacità predittiva del modello prescelto si sono infatti condotti dei test sulle tendenze nazionali, ripartizionali e regionali relative al periodo 1997-2003. Infine, si presentano i risultati delle previsioni della disoccupazione per genere e disaggregazione territoriale entro la serie (tra il 1997 e il 2003). L'ultimo paragrafo contiene una sintetica valutazione dei risultati sinora ottenuti e una breve trattazione delle questioni ancora aperte. Conclude il lavoro un'appendice sulla previsione delle uscite dall'occupazione con un esempio di previsione fuori della serie.

2. La letteratura di riferimento

La scelta di analizzare i mercati del lavoro attraverso modelli basati sulla considerazione dei flussi delle popolazioni rilevanti (gli occupati, i disoccupati, gli inattivi; le imprese, i posti di lavoro, le *vacancies*) risponde a due esigenze cognitive, distinte ma complementari. La prima è quella di valutare gli effettivi movimenti di quelle popolazioni ad un dettaglio più fine di quello consentito dalla semplice osservazione delle variazioni degli stock, allo scopo di osservare con maggiore precisione l'azione effettiva delle forze dell'offerta e della domanda di lavoro. È, infatti, solo nel quadro di questa cornice concettuale e informativa che è possibile cogliere alcuni aspetti fondamentali del funzionamento del mercato del lavoro, quali: l'attrazione e lo scoraggiamento della partecipazione, la mobilità dei lavoratori tra le imprese, la distribuzione dei cicli di espansione e contrazione occupazionale tra gli effetti di natimortalità delle imprese e le dinamiche delle imprese persistenti, i diversi elementi che influenzano la *matching function* e la curva di Beveridge, ecc. Molti di questi aspetti sono affrontati nei contributi al volume sull'analisi di flusso del mercato del lavoro curato da Ronald Schettkat (1996), mentre i riferimenti fondamentali per l'analisi di flusso del mercato del lavoro italiano sono Contini e Revelli (1987, 1992, 1997) Contini e Trivellato (2005).

La seconda esigenza deriva dal riconoscimento della rilevanza della demografia nella determinazione della quantità e qualità dell'offerta di lavoro e di conseguenza nella generazione degli squilibri che si producono nei diversi segmenti del mercato del lavoro. Questo riconoscimento, peraltro, risulta preliminare alla comprensione delle eterogeneità che caratterizzano i mercati del lavoro locali (ad esempio i mercati regionali in Italia), nella misura in cui questi sono caratterizzati da significative differenze nelle strutture demografiche (ad esempio con riferimento all'età e al genere dei lavoratori, ai tassi di nuzialità e di natalità, alla dimensione delle coorti, ecc.), ma anche nei comportamenti e negli atteggiamenti. Questi aspetti sono infatti alla base dell'eterogeneità nei comportamenti di offerta di lavoro e, attraverso la creazione di frizioni e specificità nei processi di aggiustamento, vanificano un approccio all'equilibrio nei mercati del lavoro basato esclusivamente sul livello dei tassi di salario.

In Italia il tema della segmentazione del mercato del lavoro per età trova la sua prima illustrazione teorica in un volume di Franco Franciosi (1984) in cui l'autore sottolinea la necessità di raggiungere una comprensione delle differenze nelle scelte di offerta di lavoro attraverso la costruzione di "modelli generazionali". Soltanto questi tipi di modelli, infatti, sono in grado di "inquadrare il comportamento dell'offerta di lavoro per specifici segmenti della popolazione in una visione dinamica, capace di tenere conto dell'intero ciclo di vita lavorativa dei soggetti e delle decisioni di offerta riconducibili alle caratteristiche di ciascuna generazione in relazione alle condizioni del mercato del lavoro esistenti al

momento dell'ingresso". Saranno queste considerazioni a guidare Michele Bruni e Franco Franciosi nella prima formulazione del modello stock e flussi di attrazione-scoraggiamento dell'offerta di lavoro, che costituisce il punto di partenza di questo lavoro (Bruni, Franciosi, 1985). Le applicazioni previsive di questo modello, infatti, evidenziano che la variabile generazione possiede la proprietà di catturare un ricco complesso di connotazioni valoriali, culturali e tecnologiche, sociali e comportamentali, e quindi che i modelli demografici, del tipo stock e flussi consentono di analizzare utilmente e prevedere i movimenti degli individui nel mercato del lavoro su orizzonti temporali di diversa ampiezza (da annuali a decennali) (Bruni, Franciosi, 1981 e 1985; Ministero del lavoro e della previdenza sociale, 1987-94).

Successive conferme della rilevanza del rapporto tra demografia e mercato del lavoro si riscontrano nell'ampio studio di Leoni (1987), basato sull'analisi dei profili longitudinali di partecipazione, che mostra come l'offerta di lavoro femminile sia influenzata dai fenomeni demografici in misura molto maggiore che dalle cicliche fluttuazioni delle retribuzioni. Altri contributi si soffermano sulla rigidità dei tassi di disoccupazione rispetto alle retribuzioni reali – un'evidenza che lascia aperto il campo al ruolo delle variabili demografiche. Le stime econometriche disponibili mostrano, nel caso delle regioni italiane, livelli molto modesti di risposta della disoccupazione locale agli shock salariali (Lucifora, Origo, 1999; Fabiani *et al.*, 2001).

Peraltro, lo sviluppo dell'approccio generazionale in Italia è legato anche alla formulazione dell'ipotesi del "gap generazionale" avanzata da Ezio Tarantelli (1986). Secondo quest'ipotesi, la rottura comportamentale dei giovani del '68 deriverebbe dal conflitto, soprattutto formativo e culturale, fra due "cervelli collettivi" compresenti in quegli anni nelle economie avanzate: da un lato quello dei nati prima della Grande Depressione (quarantenni e più), poco adatti alla nuova realtà industriale ma dotati di un elevato controllo istituzionale e burocratico; dall'altro quello delle persone al di sotto dei quarant'anni, più adatte alla nuova realtà industriale del periodo successivo al boom economico e più sviluppate culturalmente, che si trovano a fronteggiare con un eccesso di offerta di credenziali educative un sistema gerarchico formatosi in un periodo di carenza di qualifiche formative. Basandosi su questi risultati, lavori successivi (Brunetta, Turatto, 1992) hanno tentato una nuova sintesi, in chiave di "paradigma generazionale", di diversi aspetti (di carattere sia macro che microeconomico) concernenti l'offerta di lavoro. A livello aggregato, risalta il valore euristico delle dinamiche demografiche, quando queste sono applicate, attraverso modelli stock e flussi, come una lente con cui esaminare i movimenti degli occupati e dei disoccupati. In particolare, i flussi in entrata nella e in uscita dall'occupazione si mostrano contraddistinti da una connotazione generazionale forte, che sostiene la robustezza statistica delle stime previsionali (anche di lungo periodo), basate sulle serie storiche dei flussi per genere ed età.

3. La costruzione della banca dati di stock e di flusso

È sulla base di queste premesse teoriche e di questi risultati empirici che il gruppo di lavoro ha ritenuto di verificare se l'utilizzo del modello di Bruni e Franciosi potesse essere proficuamente esteso anche all'ambito delle previsioni regionali, oltre a quello consueto delle previsioni a livello nazionale e ripartizionale. Per poter procedere in questa direzione si è reso anzitutto necessario costruire una base di dati sia di stock che di flusso sui mercati

del lavoro regionali, costituita da serie storiche abbastanza estese da consentire di valutare la caratterizzazione strutturale delle relazioni riscontrate e la robustezza statistica delle previsioni.

3.1 La costruzione delle serie storiche di stock 1977-2003

A questo fine, attraverso un procedimento di integrazione tra le informazioni raccolte dalla rilevazione sulle Forze di lavoro (FL) e quelle della Popolazione residente per genere, anno di nascita e stato civile (POSAS), si sono quindi prodotte anzitutto, per tutte le regioni italiane, le seguenti nuove serie *degli stock medi annui*, dal 1977 al 2003:

a) occupati, disaggregati per genere, singola classe di età, settori di attività, posizione nella professione e titolo di studio;
b) disoccupati (definizione ILO), distinti per genere, singolo anno di età e titolo di studio;
c) popolazione in età attiva, per genere, singolo anno di età e titolo di studio.

Il database che ne risulta costituisce di per sé un patrimonio informativo assai rilevante per la ricerca sul mercato del lavoro per diversi ordini di motivi. In primo luogo, per il contenuto metodologico avanzato della ricostruzione; in secondo luogo, per l'opera di integrazione di materiali informativi derivanti da differenti basi dati dell'Istat; in terzo luogo, per il livello di disaggregazione e di coerenza delle serie ricostruite; infine, per i nuovi frutti conoscitivi che possono derivare dall'utilizzazione di queste serie per l'analisi e la previsione delle principali grandezze dei mercati del lavoro regionali.

Operando all'interno di un disegno previsivo complesso e integrato, come quello del Progetto MARSS, la ricostruzione delle serie storiche di base ha dovuto rispondere a più obiettivi. Il primo è stato quello di ricostruire serie omogenee degli stock di occupati e disoccupati a livello regionale, utilizzando prevalentemente le informazioni derivanti dalle FL; il secondo, dettato dalla necessità di produrre informazioni di flusso secondo il metodo dei quasi-flussi (utilizzato da Bruni e Franciosi, 1985; Ministero del lavoro, 1987-1994; Brunetta e Turatto, 1992), è stato quello di integrare i dati FL con quelli di una fonte che garantisse una migliore qualità nella disaggregazione della popolazione regionale per singolo anno di età e genere. A questo fine si sono utilizzati i dati provenienti dalla fonte anagrafica POSAS, anche in considerazione del fatto che, per realizzare gli esercizi di previsione è necessario ricorrere a previsioni demografiche delle entrate generazionali nell'età di lavoro, ricavabili dalla previsione ufficiale della popolazione al 2050, definita dall'Istat sulla base della POSAS.

Nell'effettuare la ricostruzione della serie storica delle subpopolazioni dei mercati del lavoro regionali si è cercato di risalire il più possibile all'indietro. Si è però riusciti a farlo soltanto fino al 1977 (non diversamente da quanto realizzato da Tronti, 1997a), perché i microdati FL relativi al periodo 1970-76 sono risultati poco attendibili per quanto riguarda la distinzione degli occupati per posizione nella professione e settore. Si è poi dovuto risolvere il problema di raccordare le serie regionali ricostruite per il periodo 1977-1992 con le serie correnti 1993-2003 delle medie annue della rilevazione trimestrale delle forze di lavoro[9].

Notevoli sono state le difficoltà incontrate nella ricostruzione, a motivo dei cambiamenti nelle rilevazioni FL che hanno determinato, tra l'altro, l'utilizzo di modalità diverse relative

[9] Si pone tuttavia ancora il problema di raccordare le serie complete 1977-2003 con le medie annue della serie attualmente corrente (dal 2004 in avanti) della nuova rilevazione continua delle FL.

ad una stessa variabile o il cambiamento di definizione della variabile stessa. Data la numerosità delle serie regionali da ricostruire, si è scelto un approccio organico generalizzato che potesse applicarsi a tutte le serie. Per questo ci si è orientati verso la metodologia di ricostruzione messa a punto nel 2001 da Gatto, Gennari e Massarelli che, utilizzando strumenti di analisi delle serie temporali, garantisce uniformità di trattamento per tutte le serie, completezza delle informazioni utilizzate, semplicità e velocità di esecuzione. La metodologia originale (Gatto, Gennari, Massarelli, 2001, nel seguito GGM) è stata applicata alle serie storiche a partire dal 1984, mentre una versione più generale è stata applicata per le serie dal 1977 al 1983.

Il metodo GGM è a due passi: il primo consiste nel trattamento dei microdati relativi ai periodi precedenti il cambiamento (metodologico o definitorio), in modo da renderli più omogenei possibile a quelli dei periodi successivi. Si è intervenuti su molti fattori, quali definizioni, piani di compatibilità e metodi di calcolo dei coefficienti di riporto all'universo. Si sono così ricostruite le serie storiche regionali relative al numero degli occupati per genere, settore di attività economica e posizione nella professione e al numero delle persone in cerca di occupazione per genere. Le serie ottenute, però, continuavano a presentare alcuni salti non spiegabili, evidenziando la presenza di effetti statistici residui, che la ricostruzione microfondata non era riuscita ad eliminare.

Si è reso dunque necessario il secondo passo, qualificato come "riallineamento" delle serie, per omogeneizzare definitivamente le serie storiche precedenti e successive al break. L'approccio al riallineamento, operato sulle serie trimestrali, è totalmente macro e la cornice teorica è quella dell'approccio per componenti, in cui si riallinea separatamente il ciclo-trend, la componente stagionale e la componente stocastica di ogni serie. Le serie trimestrali così ottenute sono state quindi utilizzate per calcolare le medie annuali.

Per i dati del periodo 1984-1992, si è utilizzato il metodo descritto e si sono prodotte serie storiche ricostruite fino al dettaglio regionale; per l'occupazione il dettaglio scende al settore di attività economica (a tre modalità: agricoltura, industria e altre attività) e alla posizione nella professione (a due modalità: dipendenti e autonomi), oltre che al genere. Per rispettare la coerenza con le serie ricostruite precedentemente, le serie riallineate nel 2001 sono state usate come vincoli rispetto all'elaborazione attuale. Le differenze tra le serie vincolo e quelle risultanti dall'aggregazione delle serie regionali sono state ridistribuite tra le serie regionali stesse proporzionalmente alla loro dimensione.

Invece, nel caso delle serie del periodo 1977-1983, le differenze tra i metodi di conduzione di indagine hanno reso impossibile una puntuale ricostruzione micro da usare come input per il riallineamento, e si è perciò operato prevalentemente con un riallineamento sugli aggregati. È stata comunque realizzata una ricostruzione micro di tipo definitorio che ha permesso, tra l'altro, di ricostruire all'indietro l'aggregato dei disoccupati secondo la definizione ILO adottata dall'Istat dal 1992 in poi.

Si è inoltre resa necessaria una correzione a posteriori delle serie, in modo da tenere conto delle informazioni di fonte demografica e assicurare la coerenza longitudinale dei dati. Questa operazione ha consentito di ricostruire, in particolare, le strutture per titolo di studio ed età annuali degli aggregati degli occupati e delle persone in cerca di lavoro. Si erano infatti osservate notevoli incongruenze tra classi di età contigue, dovute alla natura campionaria dei dati FL. Per evitare che ciò potesse inficiare le elaborazioni dei tassi di attività, occupazione e disoccupazione, come anche quelle dei quasi-flussi necessari al funzionamento del modello basato sull'approccio generazionale (distorsioni già osservate in

precedenti studi), le serie trimestrali di occupati e disoccupati secondo le caratteristiche oggetto d'interesse sono state riallineate, utilizzando una tecnica che prende spunto dalla metodologia statistica per la stima indiretta di piccole aree, ovvero il metodo *Structure Preserving Estimation* (Rao, 2000).

La ricostruzione delle serie storiche della popolazione in età attiva in linea con quella per la quale l'Istat realizza le previsioni demografiche è stata effettuata utilizzando contemporaneamente dati derivanti dalle FL e dalla POSAS. Per il periodo precedente al 1993, siccome la struttura per età, genere e regione del dataset delle FL (sia in trasversale che per coorte) presentava andamenti non lineari, i totali di popolazione per genere e regione sono stati ricostruiti con un procedimento basato su regressioni lineari fra la serie RTFL (1993-2001) e quella POSAS (1977-2001). Ai totali così ottenuti sono state quindi applicate delle strutture per età calcolate come per gli occupati e i disoccupati. I livelli e i profili dei tassi relativi a tutti gli aggregati del mercato del lavoro calcolati con la popolazione in età di lavoro così ottenuta, sottoposti ad analisi grafica sia di tipo trasversale che per coorte, hanno dato risultati positivi. Ulteriori incongruenze nei flussi annuali, dovute a problemi nei microdati FL (ad es. nei movimenti migratori in uscita nelle età più anziane), sono state eliminate realizzando dei riproporzionamenti con approccio misto (per valori assoluti e per tassi).

3.2 La costruzione delle serie storiche di flusso

I dati così ricostruiti consentono agli utilizzatori di integrare le analisi tradizionali in termini di stock con le analisi sulle transizioni tra le varie condizioni. In assenza di serie omogenee degli effettivi flussi in esame[10], questi possono essere approssimati attraverso i saldi generazionali per anno di calendario e per singolo anno di età degli stock delle subpopolazioni per sesso ricostruite.

Formalmente, per ogni condizione x si ha:

$$x_{t+1,a+1} - x_{t,a} = sx_{t,a} = ex_{t,a} - ux_{t,a} - d \qquad (3.1)$$

dove t indica l'anno di calendario e a l'età delle persone nella condizione x. Con sx è indicato il saldo o variazione degli stock di individui nella condizione x; con d i decessi e con ex e ux rispettivamente le entrate e le uscite dalla condizione x. I saldi sx per ogni anno di età a, possono essere sia positivi sia negativi. Considerando i saldi per tutte le età in ciascun anno t, si conviene di considerare come flusso in entrata la sommatoria dei saldi positivi e come flusso in uscita la sommatoria di quelli negativi. Per ogni subpopolazione, quindi, i flussi sono ottenuti attraverso un semplice schema di calcolo che, partendo dallo stock dell'anno di inizio della serie (1977) e disaggregando le poste relative ai decessi e agli entrati e usciti totali nel corso dell'anno, porta allo stock relativo all'anno successivo. Il procedimento può essere illustrato con un esempio: se nell'anno t la popolazione maschile (femminile) di età a è suddivisa in o occupati, u disoccupati e i inattivi e nell'anno $t+1$ la popolazione maschile (femminile) di età $a+1$ è suddivisa in p occupati, v disoccupati e l

[10] La pubblicazione dei flussi annuali e trimestrali delle FL, interrotta alla fine degli anni '80, è ripresa solo agli inizi degli anni 2000 con la diffusione di microdati riferiti al periodo 1993-94/2002-03.

inattivi, si possono definire:

p - o = entrate nell'occupazione se positive, uscite dall'occupazione se negative;

v - u = entrate nella disoccupazione se positive, uscite dalla disoccupazione se negative;

l - i = entrate nelle non forze di lavoro se positive, uscite dalle non forze di lavoro se negative.

La sommatoria di questi saldi per ciascun anno di calendario e anno di età della popolazione permette la costruzione di flussi in entrata e in uscita specifici per ciascuna subpopolazione del mercato del lavoro considerata (occupati per genere, livello di istruzione, posizione e settore di attività economica; disoccupati o inattivi per genere e livello di istruzione)[11]. Le identità contabili appena esposte mostrano che entrate e uscite dalle forze di lavoro sono la somma dei saldi tra entrate e uscite generazionali e delle transizioni tra forze di lavoro (occupati e disoccupati) e non forze di lavoro. Inoltre, più breve è l'intervallo di tempo considerato, maggiore sarà il peso relativo della componente delle transizioni rispetto a quella generazionale e viceversa, dato che alcune transizioni sono transitorie.

Le serie storiche regionali ricostruite degli stock e dei flussi sono espresse in valori assoluti e sono disponibili per singolo anno di età e genere, a partire da 15 fino a 69 anni. In questo modo, avendo a disposizione dati sia di stock che di flusso, è stato inoltre possibile costruire per tutte le subpopolazioni alcuni indicatori fondamentali per l'analisi del mercato del lavoro, quali le entrate totali, le entrate aggiuntive, le entrate di equilibrio, le uscite totali, i decessi, le età medie all'entrata e all'uscita, le età medie di inizio di alcuni fenomeni e gli indicatori di durata (espressi come rapporti tra i flussi e gli stock iniziali)[12].

Il database è stato quindi notevolmente ampliato grazie alla costruzione dei quasi-flussi e tale organizzazione dei dati determina un arricchimento consistente e coerente del patrimonio informativo che permette nuove analisi della struttura, dell'evoluzione e della dinamica dei singoli mercati del lavoro regionali.

4. Applicazione del modello delle entrate nelle forze di lavoro al livello regionale

Come già sottolineato, il gruppo di lavoro si è posto l'obiettivo di verificare se il tradizionale modello di attrazione/scoraggiamento utilizzato per la previsione della disoccupazione a livello nazionale e ripartizionale (Bruni, 1988) potesse essere adattato per effettuare previsioni anche a livello regionale. Pertanto, una volta costruito il database regionale adeguato, il modello originale è stato sottoposto a una serie di test atti a verificarne l'adattabilità ai nuovi dati e la forma funzionale migliore.

[11] Ovviamente i flussi costruiti attraverso i saldi generazionali annui sono caratterizzati da due limitazioni: a) anzitutto essi sono soltanto un porzione di quelli effettivi, in quanto non comprendono i flussi annui generazionali di pari entità tra entrate e uscite; b) in secondo luogo, essendo costruiti su dati aggregati e non su dati individuali, rappresentano le probabilità di transizione medie generazionali e non le transizioni effettuate da singoli. Va però ricordato che, in assenza di fonti di migliore qualità, poiché il procedimento adottato consente di ottenere comunque un'ampia e articolata informazione sulla dinamica del mercato del lavoro, queste limitazioni appaiono del tutto accettabili.

[12] Tutte le serie storiche elementari e gli indicatori da esse ricavati, per tutte le regioni, sono stati organizzati in un database che è stato collocato su un'area web dedicata, accessibile ai componenti del Progetto MARSS.

4.1 Il modello originario

Se suddividiamo la popolazione in un dato istante e in un dato territorio in sottoinsiemi omogenei, scanditi dalle principali fasi del ciclo vitale, possiamo stabilire tra queste fasi delle relazioni che determinano i flussi in entrata e in uscita dal mercato del lavoro. Partendo dalla domanda di lavoro, possiamo ipotizzare che la quantità di lavoro domandata individui il numero di persone che trovano un lavoro in un certo intervallo di tempo, trascurando i posti che rimangono vacanti. Se si ipotizza un elevato livello di omogeneità sia tra gli occupati che tra i posti di lavoro, è possibile scomporre il flusso di entrata nell'occupazione (*oe*) in due componenti: una componente che riempie i posti di lavoro lasciati liberi dalle persone uscite dall'occupazione e una seconda componente che varia il numero degli occupati sulla base dei processi di creazione/distruzione dei posti di lavoro, ovvero:

$$oe = oa + oq \qquad (4.1)$$

dove: *oa* indica il flusso (positivo o negativo) della domanda aggiuntiva e *oq* quello della domanda di lavoro che riequilibra le uscite assicurando la stazionarietà della popolazione occupata (e, tipicamente, della sua struttura generazionale). Il saldo tra entrate e uscite dall'occupazione è perciò pari alla domanda aggiuntiva, che coincide con l'incremento (o il decremento) dello stock degli occupati. Di conseguenza, l'insieme della domanda di flusso è superiore, uguale o inferiore a quella di equilibrio a seconda che la domanda aggiuntiva sia positiva, nulla o negativa. Si può ipotizzare che la domanda aggiuntiva dipenda dalle variazioni della produzione, del capitale e del progresso tecnico, mentre la domanda di equilibrio generazionale sarà influenzata da molti fattori come la struttura per età e per genere della popolazione, la struttura settoriale dell'economia, il rapporto salario/pensione e le regole di pensionamento.

Nelle fasi di espansione economica verranno creati nuovi posti di lavoro: la domanda totale di nuove entrate nell'occupazione sarà quindi determinata dalla necessità di ricoprire i posti resi vacanti dalle uscite (domanda di equilibrio), nonché dall'aumentata capacità produttiva e occupazionale (domanda aggiuntiva). Di contro, nelle fasi di recessione o di stagnazione, ci sarà una perdita netta di posti di lavoro: la domanda di nuove entrate sarà inferiore o, al più, pari ai posti resi vacanti dall'uscita delle coorti più anziane di lavoratori verso il pensionamento.

Anche la popolazione attiva si può analizzare adottando una metodologia di flusso, secondo la quale, considerando le uscite come esogene, la quantità offerta dipende dal numero degli entrati nelle forze di lavoro nell'intervallo di tempo considerato. Il meccanismo di ingresso può essere modellato attraverso una relazione di attrazione/scoraggiamento *à la* Tella (1964), in cui la principale variabile esplicativa delle entrate nelle forze di lavoro è costituita dalle entrate nell'occupazione.

Sulla base delle considerazioni avanzate più sopra, il modello tradizionale (Bruni, Franciosi 1981 e 1985; ripreso dal Ministero del lavoro, 1987-1994 e da Bruni, Turatto, 1988) tiene però conto anche dell'andamento demografico, che viene modellato mediante l'introduzione nella funzione di una variabile che misura gli entrati nella popolazione in età lavorativa (*pe*). In questo modo, la quantità di lavoro offerta viene rappresentata come una funzione della probabilità di trovare lavoro, e quest'ultima viene considerata indirettamente, attraverso l'interazione tra la domanda di lavoro di flusso, articolata nelle sue componenti e l'evoluzione delle entrate nella popolazione in età di lavoro.

La funzione di attrazione/scoraggiamento originariamente formulata da Bruni e Franciosi è la seguente:

$$fle = \alpha_0 + \alpha_1 oa + \alpha_2 oq + \alpha_3 pe\,. \tag{4.2}$$

4.2 L'adattamento del modello originario

Per verificare la possibilità di utilizzare questo modello a scopi previsivi sono state anzitutto realizzate alcune sperimentazioni sulla specificazione tradizionale. In primo luogo, sulla base di un'ipotesi di comportamenti asimmetrici, si è tentato di scindere la funzione di attrazione da quella di scoraggiamento a seconda del valore, positivo o negativo, della domanda aggiuntiva osservato nelle serie, senza però ottenerne risultati confortanti (forse per la limitata numerosità delle osservazioni a disposizione). Si sono poi fatte sperimentazioni sull'ampiezza della classe di età entrante nelle forze di lavoro per verificare se fosse opportuno limitarsi a considerare la coorte dei quindicenni oppure, dato il progressivo posticipo dell'età di entrata nel lavoro nel corso del tempo, fosse necessario estendere la definizione della variabile anche a classi di età contigue. Le analisi hanno mostrato che non sono intervenute nel tempo sostanziali modificazioni comportamentali e, quindi, che considerare classi più ampie non migliora la significatività del modello. Si sono poi studiati gli effetti dell'azione, non solo contemporanea ma anche temporalmente ritardata, di alcune poste dell'equazione, ma tali effetti non si sono dimostrati significativi. Infine, si è sottoposta a verifica l'ipotesi dell'esistenza di *break* strutturali della funzione di attrazione/scoraggiamento nel lungo periodo esaminato (1977-2003), ottenendo anche in questo caso risultati negativi.

In secondo luogo, al fine di migliorare la performance del modello si sono operati alcuni adattamenti. Anzitutto si sono condotti esperimenti distinti prendendo come variabile target i flussi netti o quelli lordi delle entrate nelle forze di lavoro. Se consideriamo le entrate nelle forze di lavoro come somma della domanda finale di flusso (corrispondente alle entrate nell'occupazione *eo*, che a sua volta sono date dalla somma dei flussi della domanda aggiuntiva *oa* e della domanda di equilibrio generazionale *oq*) e delle entrate nella disoccupazione (*ed*), possiamo scrivere:

$$fle = eo + ed \tag{4.3}$$

dove la notazione *fle* indica le entrate nelle forze di lavoro comprensive dei passaggi intermedi fra disoccupazione e occupazione, ovvero:

$$fle = ed + eo = fle\_netta + (od + do) \tag{4.4}$$

dove *od* rappresenta i flussi dall'occupazione alla disoccupazione, mentre *do* rappresenta i flussi in direzione opposta, dalla disoccupazione all'occupazione.

I test delle stime ottenute al livello nazionale per le entrate nette e per quelle lorde nelle forze di lavoro hanno evidenziato che l'adattamento del modello stimato per i flussi delle entrate nette è notevolmente peggiore di quello per le entrate lorde. Passare dalla stima delle entrate nette a quella delle entrate lorde determina un miglioramento dell'R^2

sostanziale: per la stima del totale nazionale, il coefficiente di determinazione passa da 0,5 a 0,9[13]. Questo risultato offre un'ulteriore conferma all'evidenza, più volte segnalata dalle ricerche sui flussi nel mercato del lavoro e recentemente riconfermata con forza da Contini e Trivellato (2005), che una variazione dell'occupazione è il risultato di un cospicuo processo di mobilità all'interno delle forze di lavoro, con consistenti fenomeni di riallocazione tra occupati e disoccupati[14].

Inoltre, le verifiche della bontà di adattamento del modello di stima delle entrate lorde hanno dimostrato che l'intercetta dell'equazione di stima non è significativa in base al test della t-Student, né per l'Italia, né per le ripartizioni Centro-Nord e Sud-Isole[15]. Quindi, nelle analisi successive si è utilizzata la seguente specificazione:

$$fle^{r}_{t,t+1} = \alpha^{r}_{1} oa^{r}_{t,t+1} + \alpha^{r}_{2} oq^{r}_{t,t+1} + \alpha^{r}_{3} pe^{r}_{t,t+1} + \varepsilon \,, \qquad (4.5)$$

in cui *r* indica la ripartizione geografica o regione. Il modello risulta significativo per tutte le aree geografiche esaminate (tabella 1)[16]. In questa fase dei lavori, ancora esplorativa, il modello è stato testato a livello regionale solo per le due regioni aderenti al Progetto MARSS, ovvero il Piemonte e la Toscana.

Sono stati quindi effettuati test per verificare la significatività della differenza fra i coefficienti stimati per le due ripartizioni considerate[17]: la differenza fra i parametri della domanda di equilibrio α_2 è risultata significativa, ad un livello di confidenza del 95%[18].

Questo risultato evidenzia che nella ripartizione meridionale, dove la componente della domanda aggiuntiva è strutturalmente più debole, la domanda legata al riequilibrio generazionale esercita un effetto di attrazione significativamente più forte che nel Centro-Nord; questo effetto però, come vedremo più avanti, è limitato alla componente maschile.

[13] Si sottolinea che, nonostante le serie storiche si riferiscano ad un periodo piuttosto lungo (26 anni), non si sono riscontrate evidenze di *break* strutturali sulla base del test di Chow, effettuato rispetto ad anni significativi per il verificarsi di importanti cambiamenti normativi relativi al mercato del lavoro o mutamenti nella rilevazione FL.

[14] I dati delle entrate e delle uscite nette dalle forze di lavoro sono stati comunque inseriti nel database del progetto, in modo da consentire agli utilizzatori di sfruttarne tutte le rilevanti potenzialità analitiche.

[15] I valori della t-Student ottenuti sono: -0,09 (Italia), 0,93 (Centro-Nord), -0,34 (Sud-Isole), 0,92 (Piemonte), -0,23 (Toscana).

[16] Poiché non c'è evidenza di autocorrelazione o di eteroschedasticità nei residui, il metodo di stima adottato è quello OLS.

[17] Per applicare tali test si sono stimate contemporaneamente le equazioni (4.5) per il Centro-Nord e per il Sud-Isole, applicando lo stimatore del *Seemingly Unrelated Regressor Model* di Zellner (1962). Le stime ottenute considerando contemporaneamente le equazioni (4.5) per le ripartizioni non vengono riportate poiché uguali fino alla prima cifra decimale a quelle ottenute applicando gli OLS separatamente per ciascuna equazione.

[18] È stato applicato il Wald test, che ha assunto il valore 6,03, che corrisponde ad un *p-value* pari a 0,02.

Tabella 1 - Stima dei parametri dei modelli per area geografica e genere

Area geografica	α_1 (Domanda aggiuntiva)	α_2 (Domanda di equilibrio)	α_3 (Entrate nell'età di lavoro)
	Modello 1 senza distinzione di genere		
Italia	0,82 (20,69)	1,08 (17,62)	0,18 (4,28)
Centro-Nord	0,81 (20,81)	0,94 (17,20)	0,27 (5,57)
Sud-Isole	0,85 (12,08)	1,25 (9,97)	0,17 (2,65)
Piemonte	0,85 (13,22)	0,90 (9,9)	0,37 (3,78)
Toscana	0,90 (13,49)	0,85 (7,4)	0,43 (3,50)
	Modello 2 per genere – Maschi		
Italia	0,35 (11,83)	0,61 (13,34)	0,30 (4,83)
Centro-Nord	0,34 (10,73)	0,52 (11,78)	0,37 (4,97)
Sud-Isole	0,44 (8,36)	0,71 (7,46)	0,34 (3,46)
Piemonte	0,39 (7,29)	0,49 (6,38)	0,47 (2,88)
Toscana	0,35 (4,48)	0,33 (2,42)	0,87 (3,06)
	Modello 2 per genere - Femmine		
Italia	0,41 (10,84)	0,44 (7,58)	0,20 (2,49)
Centro-Nord	0,43 (12,03)	0,38 (7,72)	0,37 (4,13)
Sud-Isole	0,35 (6,04)	0,21 (4,51)	0,21 (1,96)
Piemonte	0,45 (7,62)	0,40 (4,78)	0,54 (2,93)
Toscana	0,45 (12,71)	0,37 (6,08)	0,68 (5,09)

Fonte: elaborazioni su dati Istat.
Nota: in parentesi vengono indicati i corrispondenti valori della t-Student (i valori critici del test al livello del 5% sono +/- 2,07).

Il modello 1 presenta un buon adattamento[19] (tabella 2) per l'Italia e il Centro-Nord, e una *goodness of fit* leggermente peggiore per il Sud-Isole (in termini di R^2 e di distribuzione dei residui), anche se le stime della s*kewness* e della curtosi consentono di affermare che i valori assunti dai momenti campionari di ordine 3 e 4 della distribuzione dei residui non si discostano significativamente dai corrispondenti valori della distribuzione normale. Inoltre, le variabili del modello di regressione risultano generalmente cointegrate, confermando il carattere strutturale delle relazioni proposte[20] (tabella 2).

Si sono poi condotte stime del modello per genere verificando, separatamente per gli uomini e per le donne, la bontà di adattamento ai dati delle seguenti specificazioni (modello 2):

$$fle_{t,t+1}^{m/f,r} = \alpha_1^{m/f,r} oa_{t,t+1}^{r} + \alpha_2^{m/f,r} oq_{t,t+1}^{r} + \alpha_3^{m/f,r} pe_{t,t+1}^{m/f,r} + \varepsilon\,, \qquad (4.6)$$

[19] Per valutare la *goodness of fit* si è calcolato il coefficiente di determinazione R^2, si è applicato il test di Ljung-Box che ha permesso di testare la significatività dell'autocorrelazione dei residui e dei residui al quadrato, e si sono calcolate la *skewness* e la curtosi della serie dei residui, per verificare se la loro distribuzione sia approssimativamente gaussiana. Infine, per individuare gli eventuali casi di regressione spuria, si è applicato il *residual based test* (Engle, Granger, 1987; Engle, Yoo, 1987).

[20] Fa eccezione il caso del Centro-Nord.

Tabella 2 - Bontà di adattamento dei modelli per area geografica e genere

Area geografica	R^2	*Q(20)*	$Q^2(20)$	*Residual based test*
		Modello 1 senza distinzione di genere		
Italia	0,86	17,33	17,75	-4,32
Centro-Nord	0,85	24,79	17,18	-3,15
Sud-Isole	0,72	32,36	10,65	-5,46
Piemonte	0,78	26,78	21,47	-3,35
Toscana	0,85	17,91	8,63	-4,98
		Modello 2 per genere – Maschi		
Italia	0,60	15,80	12,98	-5,86
Centro-Nord	0,49	21,56	9,92	-5,70
Sud-Isole	0,54	53,49	26,85	-4,71
Piemonte	0,58	20,26	26,81	-3,24
Toscana	0,48	20,64	4,34	-5,64
		Modello 2 per genere – Femmine		
Italia	0,68	28,30	10,53	-4,42
Centro-Nord	0,74	32,67	10,53	-4,43
Sud-Isole	0,39	19,71	22,36	-2,06
Piemonte	0,52	16,43	21,99	-3,71
Toscana	0,84	24,54	38,14	-4,48

Fonte: elaborazioni su dati Istat.

Note: R^2 è il coefficiente di determinazione; *Q(20)* è la statistica del test di Ljung-Box applicato considerando 20 *lags* di ritardo nella funzione di autocorrelazione dei residui; $Q^2(20)$ è la statistica del test di Ljung-Box applicato sui residui al quadrato. Il valore critico del test di Ljung-Box al 5% è 31,4, mentre il valore critico al 5% del *residual based test* è -3,6 (Blangiewicz, Charemza, 1990). Si noti che la statistica del *residual based test* per il Piemonte (totale, maschi) non supera il valore critico al livello del 10% (-3.20).

che esprimono la dipendenza delle entrate nelle forze di lavoro maschili/femminili in funzione della domanda aggiuntiva (ipotizzata neutrale rispetto al genere), della domanda di equilibrio generazionale (anch'essa ipotizzata neutrale) e della popolazione entrante nell'età di lavoro per genere. Le due varianti del modello 2 risultano entrambe significative sulla base del test t-Student applicato ai coefficienti (tabella 1), con l'unica eccezione del coefficiente delle entrate nella popolazione in età attiva nel caso delle donne del Mezzogiorno. Questo risultato segnala che, dato il basso tasso di occupazione, gli effetti di attrazione/scoraggiamento della componente femminile delle forze di lavoro meridionali sono legati in misura preponderante all'evoluzione della domanda di lavoro piuttosto che alla dimensione numerica delle coorti di donne che si affacciano all'età di lavoro, e non sembra invece indicare che la specificazione lineare sia errata, dato che il test $Q^2(20)$ rifiuta l'ipotesi di non linearità dei residui. Si tratta, in altre parole, di un risultato che mostra come nel Mezzogiorno l'offerta di lavoro dei due generi risponde a stimoli diversi: nel caso dei maschi (e della media, come già abbiamo notato), l'effetto più forte è quello del riequilibrio generazionale, mentre in quello delle donne prevale l'effetto della domanda aggiuntiva.

Peraltro, in entrambe le ripartizioni i valori dei coefficienti maschili sono più elevati nel caso della domanda di equilibrio generazionale e in quello delle entrate nell'età di lavoro, mentre la partecipazione femminile presenta una reattività superiore nel caso della domanda aggiuntiva – un'evidenza che conferma il carattere prevalentemente addizionale e non sostitutivo dell'offerta di lavoro femminile. I test indicano che la differenza di genere è significativa: nei parametri della domanda di equilibrio per l'intero territorio nazionale; in

quelli della domanda aggiuntiva e della domanda di equilibrio per il Centro-Nord; in quelli della domanda di equilibrio per il Sud e Isole[21].

Le due varianti di genere presentano una *goodness of fit* inferiore rispetto al modello aggregato, e il peggioramento della *performance* si concentra nelle serie del Sud-Isole (tabella 2)[22]. Riguardo ai risultati dell'analisi di cointegrazione, il fatto che nel caso del Centro-Nord il modello risulti stabile per entrambi i sessi ci permette di ipotizzare che il risultato negativo del *residual based test* sulle serie storiche del Centro-Nord senza distinzione di genere vada probabilmente imputato alla ridotta lunghezza delle serie, piuttosto che all'effettiva instabilità del modello.

Infine, le equazioni (4.5) e (4.6) sono state stimate anche sui dati del Piemonte e della Toscana. I risultati sono in linea con quelli ottenuti precedentemente. Il modello con la distinzione di genere presenta una bontà di adattamento inferiore rispetto a quello senza distinzione, così come risulta dall'osservazione dei valori assunti dall'R^2 (tabella 2)[23].

4.3 La capacità previsiva del Modello delle entrate lorde nelle forze di lavoro per genere e area geografica

Una volta verificata la qualità dell'adattamento della specificazione prescelta sul periodo 1977-2003, si sono realizzati alcuni esercizi preliminari sulla capacità previsiva del modello confrontando le previsioni delle entrate nelle forze di lavoro ottenute stimando l'equazione (4.6) con i corrispondenti valori osservati nel periodo (derivanti dalle serie storiche ricostruite)[24]. La domanda aggiuntiva, quella di equilibrio e la popolazione entrante vengono considerate esogene, e ad esse viene attribuito il valore storico (previsione entro la serie), mentre i coefficienti sono stimati ricorsivamente.

Innanzitutto, sono state utilizzate le osservazioni dal flusso 1977-78 fino a quello 1996-1997 per effettuare la previsione h=1,3,5 anni in avanti. Poi il modello è stato stimato nuovamente, usando un'osservazione in più e calcolando un'altra volta la previsione h passi in avanti. L'esercizio previsivo è stato ripetuto fino a che tutte le osservazioni dal 1997 fino al 2003-h sono state utilizzate nella stima. Per valutare l'accuratezza delle previsioni sono state applicate due metriche di errore: il *Mean Percentage Error* (MPE), il *Mean Absolute Percentage Error* (MAPE)[25].

Coerentemente con l'analisi della bontà di adattamento esposta nel paragrafo precedente, la previsione migliore si ha sulle serie senza distinzione per genere dell'Italia,

[21] La statistica di Wald assume i valori rispettivamente: Italia 4,6 (*p-value* 0,03), per il Centro-Nord 3,6 (*p-value* 0,06) e 3,9 (*p-value* 0,05), per il Sud-Isole 3,2 (*p-value* 0,07).

[22] In particolare si nota l'elevato valore del test di Ljung-Box sui residui per gli uomini nel Sud-Isole. Inoltre la statistica del *residual based test* per le donne del Sud-Isole, indica che le variabili non sarebbero cointegrate. Tuttavia, poiché la relazione è stabile per il Sud-Isole, sia per gli uomini che per il totale, si ipotizza che questo risultato sia dovuto più all'insufficiente numerosità della serie che all'effettiva instabilità della relazione.

[23] I residui per gli uomini della Toscana non si distribuiscono secondo la normale, sulla base delle stime della curtosi e della *skewness*; l'autocorrelazione dei residui elevati al quadrato, inoltre, per le donne della Toscana è significativa sulla base del test di Ljung-Box.

[24] Si prevedono i flussi 1997-1998, 1998-1999, 1999-2000, 2000-2001, 2001-2002, 2002-2003.

[25] $MPE = \dfrac{\sum_i \dfrac{(y_i - \hat{y}_i)}{y_i} * 100}{n}$; $MAPE = \dfrac{\sum_i \dfrac{|y_i - \hat{y}_i|}{y_i} * 100}{n}$

del Centro-Nord e delle due Regioni. La *performance* peggiore si ha invece su tutte le serie del Sud-Isole. Il modello, inoltre, presenta una delle sue caratteristiche peculiari già evidenziate in altri studi, ovvero di avere una buona capacità previsiva sul lungo periodo (tabella 3).

Siccome i risultati ottenuti con il modello senza distinzione di genere risultano anche in questo caso migliori di quelli ottenuti con i modelli specificati per i due sessi, si stanno considerando altre varianti dell'equazione (4.6). Si è per questo introdotta una distinzione di genere sia per la domanda di equilibrio, sia congiuntamente per la domanda di equilibrio e per la domanda aggiuntiva. I risultati migliorano, specie nel caso dell'ultima variante considerata, in particolare per le previsioni a livello regionale. Questo risultato potrebbe essere dovuto al fatto che i due segmenti di genere del mercato del lavoro si comportano in modo profondamente diverso. Le stime per la Toscana e il Piemonte, ottenute disaggregando per genere la domanda di equilibrio, permettono un notevole miglioramento della performance previsiva: per i maschi del Piemonte il MAPE diminuisce dal 13,9 al 4,9 per cento e per i maschi della Toscana dall'11,1 al 4,2 per cento.

Tabella 3 - Indicatori dell'accuratezza della previsione della *forza di lavoro entrante* secondo i termini temporali del modello, per area geografica e genere

Area geografica	Modello a 1 anno		Modello a 3 anni		Modello a 5 anni	
	MPE	MAPE	MPE	MAPE	MPE	MAPE
	Modello 1 senza distinzione di genere					
Italia	1,8	2,8	1,6	2,3	2,1	2,4
Centro-Nord	2,7	2,7	3,4	3,4	4,2	4,2
Sud-Isole	-5,9	11,2	-11,4	11,6	-13,7	13,6
Piemonte	1,3	4,6	0,5	5,0	4,5	4,5
Toscana	-3,9	3,9	-6,1	6,1	-5,7	5,7
	Modello 2 per genere – Maschi					
Italia	3,8	6,8	5,5	7,0	9,9	9,9
Centro-Nord	4,5	4,9	5,6	6,2	11,6	11,6
Sud-Isole	-9,0	9,7	-13,7	13,7	-16,1	16,1
Piemonte	-6,1	13,9	-10,4	17,7	4,5	4,5
Toscana	-7,5	11,1	-9,1	13,0	-1,6	3,0
	Modello 2 per genere – Femmine					
Italia	-0,4	7,0	-2,2	8,6	-6,6	6,6
Centro-Nord	-0,1	5,8	0,6	7,2	-4,0	4,0
Sud-Isole	-0,8	13,8	-6,4	11,5	-3,5	16,5
Piemonte	6,2	7,3	7,7	8,1	5,1	5,1
Toscana	-0,8	9,9	-3,8	10,3	-6,9	6,9

Fonte: elaborazioni su dati Istat.
Nota: i valori dei due indicatori per il modello a 5 anni non si differenziano molto perché basati solo su tre osservazioni.

5. La previsione della disoccupazione per genere e area geografica

I risultati delle previsioni delle entrate lorde nelle forze di lavoro vengono a questo punto utilizzati per prevedere lo stock dei disoccupati. Per far questo si utilizza la formula seguente:

$$dis^r_{t+1} = dis^r_t + fle^r_{t,t+1} - oa^r_{t,t+1} - oq^r_{t,t+1} - ud^r_{t,t+1} \quad (5.1)$$

dove *ud* indica il flusso degli usciti dalla disoccupazione tra t e t+1[26] ed in cui

$$fle^r_{t,t+1} = \alpha^r_1 oa^r_{t,t+1} + \alpha^r_2 oq^r_{t,t+1} + \alpha^r_3 pe^r_t . \quad (5.2)$$

La previsione dei disoccupati entro la serie (previsione 1998-2003), calcolata con la formula (5.1), in cui le entrate nelle forze di lavoro sono stimate con il modello 1 presenta, per tutte le subpopolazioni territoriali e di genere, errori più contenuti della previsione delle entrate nelle forze di lavoro. Si tratta di un risultato atteso, in quanto quest'ultima è l'unica variabile stimata, mentre tutte le altre grandezze che entrano nel computo sono poste uguali ai valori storici. I risultati mostrano che il MAPE della previsione dei disoccupati assume valori notevolmente inferiori a quello della previsione delle entrate nelle forze di lavoro (tabella 4). La *performance* dei modelli previsivi della disoccupazione è buona (MAPE al di sotto del 10%) per quasi tutte le aree geografiche e su quasi tutti gli orizzonti temporali. Qualche eccezione si riscontra per i maschi sia del Centro-Nord (h=5), che del Piemonte (h=1, h=3) e della Toscana (h=3)[27].

Tab. 4 - Risultati della previsione della forza lavoro entrante e dei disoccupati secondo i termini temporali del modello, per area geografica e genere

Area geografica	Modello a 1 anno		Modello a 3 anni		Modello a 5 anni	
	MAPE FLE	MAPE Disoccupati	MAPE FLE	MAPE Disoccupati	MAPE FLE	MAPE Disoccupati
			Modello 1 senza distinzione di genere			
Italia	2,8	0,1	2,3	0,8	2,4	1,0
Centro-Nord	2,7	1,6	3,4	2,4	4,1	3,0
Sud-Isole	11,2	2,3	11,6	2,5	13,6	2,6
Piemonte	4,6	2,9	5,0	3,4	4,5	3,6
Toscana	3,8	2,6	6,1	4,1	5,7	3,8
			Modello 2 – Maschi			
Italia	6,8	2,9	7,0	3,3	9,9	5,0
Centro-Nord	4,9	4,6	6,2	6,2	11,6	11,0
Sud-Isole	9,7	2,2	13,7	3,5	16,1	4,2
Piemonte	13,9	10,8	17,7	13,6	4,5	4,6
Toscana	11,1	9,4	13,0	11,8	3,0	3,3
			Modello 2 – Femmine			
Italia	6,9	3,0	8,6	2,7	6,6	2,0
Centro-Nord	5,8	3,1	7,2	4,1	4,0	2,1
Sud-Isole	13,8	2,7	11,5	1,9	16,5	2,7
Piemonte	7,3	3,9	8,1	4,9	5,1	3,4
Toscana	9,9	5,3	10,3	5,9	6,9	3,5

Fonte: elaborazioni su dati Istat.

[26] Si effettua la previsione della disoccupazione per gli anni 1998-1999-2000-2001-2002-2003.

[27] L'errore di previsione rilevato per i maschi del Piemonte può essere attribuito alla non accurata previsione solo di un anno della serie considerata (2001); se tale anno non viene considerato l'errore si riduce sensibilmente: per il Piemonte (maschi) il MAPE decresce per h=1 da 13,9 a 8,1, per h=3 da 17,7 a 9,2.

Infine, in Appendice si riporta un esercizio di previsione dello stock dei disoccupati ai vari livelli territoriali per il 2003 che comporta la previsione dei tassi di uscita dalla occupazione e dalla disoccupazione attraverso modelli ARIMA e l'uso delle previsioni demografiche dell'Istat per le entrate nella forza lavoro.

6. Questioni aperte e conclusioni

A conclusione del lavoro è utile riassumere i risultati raggiunti per valutarne la portata e per identificare i possibili sviluppi della ricerca. Il primo risultato di rilevo è quello della ricostruzione di un database coerente di dati annuali di stock e di flusso a livello regionale, dal 1977 al 2003, della popolazione in età di lavoro per condizione professionale (occupati, in cerca di lavoro, non forze di lavoro), genere, singolo anno di età e titolo di studio (almeno diplomati/non diplomati), che identifica anche la posizione nella professione (dipendente/autonomo) e il macrosettore di attività economica (agricoltura, industria e servizi) degli occupati. Il database si raccorda pienamente con la serie delle medie annue della rilevazione trimestrale delle forze di lavoro (1993-2003). La ricostruzione effettuata non si caratterizza soltanto per l'omogeneità nell'identificazione delle sottopopolazioni che costituiscono il mercato del lavoro (occupati, disoccupati nella definizione ILO, inattivi) lungo tutto l'arco temporale della ricostruzione, ma anche per la coerenza longitudinale delle loro strutture per genere ed età che, oltre ad assicurare una più elevata attendibilità dei dati, consente la costruzione dei dati di flusso (entrate e uscite annuali dalle diverse subpopolazioni) che ne rendono possibile l'utilizzo nel quadro di un modello previsivo stock e flussi dell'offerta di lavoro, fortemente caratterizzato da una visione longitudinale del mercato del lavoro come sequenza di stati (tra gli altri, Tronti, 1997b).

La disponibilità di queste nuove serie storiche omogenee, di stock e di flusso, costituisce un risultato di grande rilevanza in sé, indispensabile per la ricostruzione e la comprensione delle caratteristiche, delle vicende storiche e della performance dei diversi mercati del lavoro regionali che si articolano sul territorio italiano.

Ma il nuovo database ha anche consentito di testare con successo alcune nuove specificazioni del tradizionale modello stock e flussi di previsione della disoccupazione, ideato per primi da Bruni e Franciosi (1985). In particolare, il nuovo database ha consentito di sottoporre l'originario modello di attrazione/scoraggiamento ad un'analisi di cointegrazione, sia nella specificazione consueta (verificando, però, la non significatività dell'intercetta), sia in una nuova versione, con una specificazione della domanda di equilibrio generazionale per genere. Viene confermata la buona capacità previsiva relativa ai flussi in entrata nelle forze di lavoro e, conseguentemente, al numero dei disoccupati per genere a distanza di 1, 3 e 5 anni, sia a livello nazionale, sia per le due ripartizioni (Centro-Nord e Sud-Isole), sia infine per la Toscana e il Piemonte. I risultati sono particolarmente incoraggianti, tanto per la stima degli specifici flussi in entrata nelle forze di lavoro e ancor più per la previsione degli stock di disoccupati, anche a 5 anni di distanza.

Fanno, inoltre, ben sperare anche i risultati dell'esercizio di previsione dei disoccupati fuori dalla serie presentato in Appendice, per il quale sono stati utilizzati vari e diversi set di informazioni, comprese le previsioni demografiche dell'Istat. Per effettuare previsioni fuori della serie si sta ancora perfezionando la stima delle uscite dalle forze di lavoro per verificare sino a che punto queste possano essere integralmente basate su di un modello

(univariato o multivariato) oppure, in considerazione delle rilevanti e ripetute modifiche legislative in corso sulla regolazione dell'età pensionabile che già da diversi anni stanno operando un prolungamento della vita attiva, sia, invece, più opportuno introdurre valutazioni ad hoc dei loro effetti nel corso del tempo.

Per quanto riguarda la prosecuzione dell'attività di ricerca, il gruppo di lavoro intende testare la previsione entro la serie anche su un orizzonte temporale decennale, che dovrebbe dimostrarsi particolarmente idoneo all'utilizzo di un modello longitudinale fortemente determinato dagli effetti di natura demografica. Un altro aspetto più specifico su cui proseguirà la ricerca riguarda l'approfondimento dell'analisi del modello di offerta per il Sud, studiando ulteriori specificazioni che consentano di ottenere previsioni più accurate, con particolare riferimento alla componente femminile. Tra le altre ipotesi, si ipotizza di ampliare la numerosità dei dati a disposizione attraverso l'utilizzo di tecniche di *pool regression* da applicare a gruppi di regioni omogenee. Un ulteriore esercizio di verifica dell'accuratezza previsionale del modello può consistere nel confronto delle sue previsioni con quelle prodotte da uno o più modelli benchmark pubblicati, basati sulla previsione diretta degli stock delle subpopolazioni del mercato del lavoro.

Appendice

Previsione a un anno fuori della serie con l'applicazione di previsioni della popolazione e di un modello ARIMA delle uscite dall'occupazione

Sulla base della specificazione individuata, si è realizzato un esercizio di previsione dello stock dei disoccupati per l'anno 2003 sui dati dell'Italia, delle due ripartizioni e delle regioni Toscana e Piemonte, senza distinzione di genere, non considerando più la domanda di equilibrio e gli usciti dalla disoccupazione come variabili note nel momento in cui si effettua il *forecast* e, quindi, dovendo prevederle.

Per realizzare la previsione del numero dei disoccupati si è applicata l'identità 5.1. Mentre le entrate nelle forze di lavoro sono state, come di consueto, espresse in funzione della domanda aggiuntiva, della domanda di equilibrio e della popolazione entrante dei 15enni. I coefficienti del modello sono stati stimati utilizzando i dati a disposizione fino al 2002. Per le entrante nella popolazione in età di lavoro nel 2003 si è usata la previsione demografica effettuata dall'Istat[28]; la domanda aggiuntiva è stata, invece, trattata come variabile esogena (nell'esercizio, è stato adottato il dato storico 2003). La domanda di equilibrio è stata stimata moltiplicando gli occupati del 2002 per la previsione 2003 del tasso aggregato di uscita dall'occupazione; e analogamente gli usciti dalla disoccupazione sono stati previsti moltiplicando i disoccupati del 2002 per la previsione 2003 del tasso aggregato di uscita dalla disoccupazione.

Le serie dei tassi aggregati di uscita dall'occupazione e dalla disoccupazione sono risultate non stazionarie sulla base dell'*Augmented Dickey Fuller test*[29], con l'unica eccezione del Sud-Isole (tabella A1).

Tab. A1 - Risultati del test ADF(p) sulle serie del tasso aggregato di uscita dall'occupazione e dalla disoccupazione per area geografica

Area geografica	Uscite dall'occupazione			Uscite dalla disoccupazione		
	Senza intercetta	Con intercetta	Trend	Senza intercetta	Con intercetta	Trend
Italia	-0,03 (0,66)	-2,42 (0,15)	-2,23 (0,44)	-0,91 (0,31)	-3,28 (0,02)	-3,19 (0,11)
Nord-Centro	-0,10 (0,63)	-1,94 (0,31)	-1,94 (0,60)	-1,19 (0,20)	-3,15 (0,03)	-3,12 (0,12)
Sud-Isole	-0,46 (0,50)	-3,96 (0,006)	-4,14 (0,01)	-1,00 (0,27)	-3,59 (0,014)	-3,44 (0,07)
Piemonte	-0,58 (0,45)	-1,73 (0,40)	-1,91 (0,61)	-1,20 (0,20)	-3,16 (0,30)	-3,10 (0,13)
Toscana	-0,73 (0,39)	-1,01 (0,73)	-1,73 (0,70)	-0,14 (0,62)	-1,69 (0,42)	-2,08 (0,53)

Fonte: elaborazioni su dati Istat.

Nota: in parentesi sono indicati i *p-value* corrispondenti ai valori assunti dalla statistica test. Le tre colonne si riferiscono alle tre opzioni possibili per effettuare il test: nella prima colonna un'applicazione senza il trend deterministico e senza l'intercetta, nella seconda colonna un'applicazione solo con l'intercetta, nella terza colonna con un trend deterministico.

[28] Si veda il sito www.demo.istat.it

[29] L'ipotesi nulla di non stazionarietà della serie viene rifiutata al livello dell'1%.

Per effettuare le previsioni per il 2003, sono stati applicati i processi ARIMA(p,d,q): nella tabella A2 sono riportati i risultati previsivi migliori, per p = 0, 1, 2, 3 e q = 0, 1, 2, 3, in termini di errore relativo percentuale[30], ottenuti imponendo d = 1 (trasformazione differenze prime) in tutti i casi in cui il test ADF(p) accetta l'ipotesi nulla di non stazionarietà della serie. La tabella mostra che i valori dei tassi di uscita previsti con i modelli ARIMA non sempre sono prossimi a quelli osservati, in particolare per il tasso di uscita dall'occupazione per l'Italia, il Centro-Nord e il Piemonte, e per il tasso di uscita dalla disoccupazione per la Toscana. Questo risultato si spiega, soprattutto, tenendo conto del fatto che nel 2001 si osserva un elevato incremento delle uscite dall'occupazione non confrontabile con i livelli precedenti né con l'anno successivo: il modello ARIMA, infatti, per sua natura, coglie meglio la tendenza di lungo periodo che i picchi congiunturali[31].

Tab. A2 - Errori relativi percentuali corrispondenti alle previsioni dei tassi aggregati di uscita dalla occupazione e dalla disoccupazione per area geografica

Area geografica	Modello	Errore relativo *tuocc*	Modello	Errore relativo *tudisocc*
Italia	ARIMA(0,1,1)	27,4%	ARIMA(3,1,3)	1,6%
Nord-Centro	ARIMA(3,1,2)	24,7%	ARIMA(1,1,3)	1,4%
Sud-Isole	ARIMA(1,0,0)	0,2%	ARIMA(2,1,0)	2,1%
Piemonte	ARIMA(2,1,0)	34,6%	ARIMA(1,1,3)	7,2%
Toscana	ARIMA(0,1,3)	1,0%	ARIMA(0,1,3)	15,5%

Fonte: elaborazioni su dati Istat.

Nonostante ciò, le previsioni della disoccupazione sono accurate: la differenza tra valori osservati e previsti è sempre inferiore al 10%, tranne nel caso del Piemonte dove raggiunge il 12,4% (tabella A3). Tale positivo risultato è dovuto sia all'esiguità dell'impatto dell'errore di stima degli usciti dall'occupazione rispetto allo stock dei disoccupati (si tratta di variabili che, nella media, hanno un ordine di grandezza assai differente), sia al fatto che il modello coglie bene il flusso degli usciti dalla disoccupazione, che costituisce una componente importante della previsione. A livello nazionale la differenza tra valore osservato e valore previsto dei disoccupati del 2003 risulta inferiore all'1%, e parimenti molto contenuti sono anche gli errori per le due ripartizioni territoriali. Per le due regioni, invece, che sono caratterizzate da un numero di disoccupati molto inferiore e per le quali, quindi, un errore assoluto anche piccolo si trasforma in un errore relativo elevato, l'errore di previsione risulta ovviamente più consistente. Ma il risultato ottenuto è comunque molto apprezzabile e certamente migliorabile attraverso un affinamento dei metodi di previsione dei tassi di uscita dall'occupazione e dalla disoccupazione.

[30] $errore = \frac{vstorico - vprevisto}{vstorico} \times 100$

[31] Per verificare l'ipotesi effettuata per spiegare il risultato previsivo sul tasso di uscita dall'occupazione si è realizzata la previsione anche del tasso di uscita del 2001 per l'Italia. Il valore previsto dal Modello ARIMA(0,1,2) corrisponde ad un errore relativo percentuale pari al 6,2%.

Tab. A3 - Errori relativi percentuali corrispondenti alla previsione del numero dei disoccupati e delle entrate nelle forze di lavoro per area geografica

Area geografica	Disoccupazione			Forze di lavoro		
	Valore storico	Valore previsto	Errore relativo %	Valore storico	Valore previsto	Errore relativo %
Italia	2.092.906	2.109.800	-0,8	701.900	822.500	-17,2
Nord-Centro	756.456	743.270	1,7	499.880	548.110	-9,6
Sud-Isole	1.336.450	1.347.100	-0,8	242.600	254.900	-5,1
Piemonte	92.315	80.858	12,4	76.073	72.987	4,1
Toscana	72.660	66.261	8,8	53.707	48.513	9,7

Fonte: nostre elaborazioni su dati Istat.

Riferimenti bibliografici

Blangiewicz M., Charemza W.W. (1990), "Co-integration in small samples: empirical percentiles, drifting moments and customized testing", *Oxford Bulletin of Economics and Statistics*, vol. 52(3), pp. 303-315.

Brunetta R., Turatto, R. (1992), *Disoccupazione, isteresi e irreversibilità*, Etas Libri, Milano.

Bruni M. (1988), "A Stock-Flow Model to Analyse and Forecast Labour Market Variables", *Labour*, vol.2 n.1, Spring, pp.55-116.

Bruni M., Franciosi F.B. (1981), "Una interpretazione in termini di flusso della dinamica delle forze di lavoro", *Economia&Lavoro*, n. 2, Marsilio, Venezia.

Bruni M., Franciosi F.B. (1985), "Scenari alternativi di domanda e di offerta di lavoro: un'analisi in termini di flusso", Ministero del lavoro e della previdenza sociale, *La politica Occupazionale per il Prossimo Decennio*, Roma.

Bruni M., Turatto R. (1988), "Scenari previsti per il quinquennio 1985-1990: l'evoluzione della struttura occupazionale", Ministero del lavoro e della previdenza sociale - CER - Fondazione G. Brodolini, *Rapporto '87. Lavoro e politiche dell'occupazione in Italia*, Poligrafico dello Stato, Roma.

Casini Benvenuti S., Paniccià R. (2003), *A Multiregional Input-Output Model for Italy*, Irpet, Firenze.

Chiarini B. (1989), "Il mercato del lavoro nei modelli macroeconomici. Una rassegna dell'evidenza econometrica dei modelli italiani", *Economia&Lavoro*, XXIII, Marsilio, Venezia.

Contini B., Revelli R. (1987), "The Process of Job Creation and Job Destruction in the Italian Economy", *Labour*, 1, (3), pp. 121-144.

Contini B., Revelli R. (1992), *Imprese, occupazione e retribuzioni al microscopio*, Il Mulino, Bologna.

Contini B., Revelli R. (1997), "Gross flows vs. net flows in the labor market: What is there to be learned?", *Labour Economics*, Vol.4, n.3, pp. 245-263.

Contini B., Trivellato U. (eds.) (2005), *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano*, Il Mulino, Bologna.

Engle R.F., Granger C.W.J. (1987), "Co-Integration and Error Correction: Representation, Estimation and testing", *Econometrica*, n. 55, pp. 251-276.

Engle R.F., Yoo B.S. (1987), "Forecasting and testing in cointegrated systems", *Journal of Econometrics*, vol. 35, pp. 143-159.

Fabiani S., Locarno A., Oneto G., Sestito P. (2001), "The sources of unemployment fluctuations: an empirical application to the Italian case", *Labour Economics*, no.2, May, pp. 259-90.

Franciosi F.B. (1984), *L'offerta di lavoro nell'analisi economica*, FrancoAngeli, Milano.

Fondazione G. Brodolini (1999), "Le previsioni 1998-2000 dell'occupazione per professioni", *Economia&Lavoro*, n. 1, Marsilio, Venezia.

Fondazione G. Brodolini (1999), "Un modello per la previsione della domanda di lavoro e di professioni su base settoriale", *FGB Ricerche n. 136/99*, a cura di Tronti L., Unioncamere-Ministero del lavoro-Commissione Europea, Roma.

Gatto R., Gennari P., Massarelli N. (2001), "La ricostruzione e il riallineamento delle serie storiche delle forze di lavoro 1984–1992", presentato al convegno *Occupazione e disoccupazione in Italia: misura e analisi dei comportamenti*, Murst, Bressanone.

Jarque C.M., Bera A.K. (1980), "Efficient test for normality, homoskedasticity and serial dependence of regression residuals", *Economic letters*, 6, pp.255-259.

Leoni R. (1987), *Le teorie economiche dell'offerta di lavoro*, La Nuova Italia Scientifica, Roma.

Lucifora C., Origo F. (1999), "Alla ricerca della flessibilità: un'analisi della curva dei salari in Italia", *Rivista Italiana degli Economisti*, Anni IV, 1.

Ministero del lavoro e della previdenza sociale (1987-1994), *Rapporto annuale. Lavoro e politiche della Occupazione in Italia*, Poligrafico dello Stato, Roma.

Rao J.N.K. (2000), *Statistical methodology for in direct estimations in small areas*, 39, Eustat.

Schettkat R. (ed.) (1996), *The Flow Analysis of Labour Markets,* Routledge, London and New York.

Tarantelli E. (1986), *Economia politica del lavoro*, UTET, Torino.

Tella A. (1964), "The Relation of Labour to Employment", *Industrial and Labour Force Review*, April, pp. 454-69.

Tronti L. (a cura di) (1997a) "Un modello per la previsione dell'offerta di lavoro", Progetto Strategico CNR *Disoccupazione e basso livello di attività in Italia,* Roma.

Tronti L. (1997b), "Il mercato del lavoro come sequenza di stati. Spunti analitici e ipotesi di intervento", Brunetta R., Vitali L. (a cura di), *Mercato del lavoro: analisi strutturali e comportamenti individual*i, FrancoAngeli, Milano.

Zellner A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias", *Journal of the American Statistical Association*, 57, pp.348-368.

Imputation of Missing Values for Longitudinal Data: an Application to the Italian Building Permits

Fabio Bacchini[1], Roberto Iannaccone[2], Edoardo Otranto[3]

Abstract

In longitudinal data set the missing response can be seen either as a set of item nonresponse in a longitudinal record or, in a cross-section context as a unit nonresponse. For imputation we suggest a simple longitudinal donor technique, choosing the number of adjustment cells and the corresponding loss functions to have results similar to those provided by weighting methods in terms of estimation of the total aggregate of the variable of interest. The choice of the imputation method is made in terms of a cross-validation procedure and then applied to reconstruct the data set of the Italian building permits.

Keywords: Nonresponse; weighting adjustment; imputation; longitudinal data.

1. Introduction

The statistical literature distinguishes the presence of missing response in statistical surveys (either census or sampling surveys) between unit nonresponse (UNR) and item nonresponse (INR). The problem of UNR is generally addressed with weighting procedures, whereas imputation techniques are adopted for INR case.

For longitudinal data set, when the units have to provide information for more than one period, missing values could refer to different points in time. The missing values can be seen as a set of INR's in a longitudinal record, implying that the imputation is the appropriate technique for their estimation. As a cross-section context we can see the missing values as UNR's and a weighting procedure could be used (Kalton, 1986). The choice depends mainly on the purpose of the analysis. When the main interest is for the estimations of aggregate values (as, for example, the total for a variable on some population), weighting or imputation techniques could be in competition. Instead, for estimations in different study domain we have to perform an imputation technique to obtain a complete data set.

In this work we consider the results obtained with a weighting procedure as the benchmark to compare the ones obtained with the imputation procedure extending studies for cross-section cases (see, for example, Haziza et al., 2001, Chen and Shao, 2000, Manzari, 2004).

Actually, the large recent literature about panel data deals with missing values mainly in

[1] Ricercatore (Istat), e-mail: bacchini@istat.it
[2] Ricercatore (Istat), e-mail: iannacco@istat.it
[3] Professore associato (Università degli Studi di Sassari), e-mail: eotranto@uniss.it

terms of relationship between the data generating process and the regression model used (see, for example, Robins et al., 1995, Vella, 1998), whereas applications in which imputation and weighting methods in longitudinal context are compared are neglected.

The data set employed in our experiments is the monthly series of the building permits released by the Italian municipalities for the period 2000-2002 (source: Istat- Istituto Nazionale di Statistica). Istat needs to produce indicators for the building sector both as monthly national aggregate, to fulfill the European Short Time Statistics (STS) regulation, and as yearly disaggregate figures. In this last case only an automatic imputation technique seems appropriate.

After being verified that the missing values generator process is not Missing Completely at Random (MCAR), so that this mechanism in not ignorable in the imputation, we estimate the missing values with a longitudinal imputation method based on the donor. The results have been compared with a traditional cross-section imputation method and a weighting method based on the estimation of the nonresponse probabilities (Little, 1986; Eltinge and Yansaneh, 1997).

It is important to stress that the main purpose of a Statistical Office is to search for a method with a good performance in terms of missing values estimation that can be easily implemented and used in an automatical way. For this reason we exclude from our analysis the imputation model-based techniques, which require frequent checks and updating when new information becomes available.

The evaluation of the method has been conducted by a cross-validation experiment, considering several non response patterns with a similar structure as in the full observed units and imputing the simulated data sets. Several loss functions and criteria are considered to choose the final imputation technique to be adopted for the application on the real data.

The work is structured in five sections. In the first one we briefly describe the survey on building permits. In the second we stress the characteristics of the pattern of the respondents and the MCAR hypothesis is verified with respect to the stratification variables. In the third part we describe the imputation techniques used whereas the fourth part shows the cross-validation experiment. In the fifth part the chosen method is applied to the observed data set and then final remarks follow.

2. The Italian building activity survey

The survey on building activity collects data on building permits for the construction of new residential and not residential buildings and any enlargement of a pre-existing building. It is a monthly census with target population given by the Italian municipalities. If no building permits are released for that month, the municipalities send a form with the indication of null activity. Some of the variables recorded, as the number of new buildings, their useful area or number of dwellings, are required by the European STS regulation (STS-Annex B) and, in particular, the total at national level. Given the non response for some of the municipalities, an estimation of missing values is required.

In the rest of the paper, we deal with only one of the variable (new dwellings) which is the most relevant in the building sector.

First of all, we have divided the n=8,100 municipalities (our statistical population) in two subsets:

- the n_1=160 provinces and non provinces with more than 50,000 citizens, which possess a total

number of citizens equal to 20,998,197 (36.4% of the Italian population);
- the n_2=7,940 non provinces having less than 50,000 citizens, with a total number of 36,691,698 citizens (63.6% of the Italian population).

Considering the number of months in which municipalities are respondent (collaboration months), in the first subset, the number of respondents is rather high (see Table 1); in fact the 85.6% of municipalities (possessing the 91.4% of the total population) has always responded in 2000. This percentage decreases in the following years even if the number of respondents remains quite high.

Table 1 - Distribution of provinces and not provinces for number of collaboration months: 2000-2002

Collaboration	Provinces			Not Provinces		
	2000	2001	2002	2000	2001	2002
0	4	6	7	1,384	1,626	1,875
1	-	-	1	203	196	183
2	-	-	-	136	104	104
3	1	2	2	127	92	113
4	1	2	1	108	106	95
5	1	1	2	136	90	111
6	-	2	5	125	123	113
7	-	1	3	140	141	148
8	1	3	2	177	148	170
9	3	1	3	303	213	279
10	5	4	7	525	307	453
11	7	9	24	1,146	582	871
12	137	129	103	3,430	4,212	3,425
Total	160	160	160	7,940	7,940	7,940

The distribution of the n_2 units belonging to the second subset is concentrated on the tails (Table 1). In the three years from 2000 to 2002, more than 50% of units (respectively 67.9%, 69.1% and 61.9% of the total population) are respondents at least 11 months. On the other side, 1,384 units in 2000, 1,626 in 2001 and 1,875 in 2002, are always non respondent. In the rest of the work we will concentrate on this second subset.

3. The mechanism of missingness

The mechanism of missingness in panel data can be classified according to its dynamics in time. For sake of simplicity, suppose the observational span equal to 3 months and an indicator variable equal to 1 when an unit is respondent at a certain time and 0 if it is not respondent. There are 2^3=8 possible response patterns: 111; 110; 101; 011; 100; 010; 001; 000. Little and David (1983) indicate the cases 110, 100 and 000 as *attrition*, in which 000 represents the UNR case, whereas 110 and 100 are the cases in which the unit collaborates only for the first part of the period. The 101 pattern represents the case of temporary non collaboration. The 011 and 001 patterns are the cases of collaboration with a certain lag. The 010 pattern represents the case in which the unit collaborates with a certain lag and then stops to collaborate. Finally the 111 case represents the full collaboration of the unit.

For the data set on building permits many municipalities show a collaboration at alternate

periods (59.0% of the total of 7,940), whereas the municipalities with full collaboration are 1,910 and the total nonresponse is equal to 922 (table 2).

Table 2 - Characteristic of nonresponse: 2000- 2002

	Frequencies	% Frequencies
Total nonresponse	922	11.6
Attrition	378	4.8
Late entry	46	0.6
Reentry	4,684	59.0
Always respondent	1,910	24.0
Total	7,940	100.0

As well known, the knowledge of the missing value mechanism is negligible if we assume that data are missing completely at random (MCAR) or missing at random (MAR). The difference between the two cases is that the probability that an observation is missing is unrelated to its value or the value of other variables (MCAR) or after controlling for another variable (Little and Rubin, 1987). To test the existence of a MCAR mechanism, we consider two auxiliary variables available for all the municipalities and related to the phenomenon studied: the population at December 31, 1999 (quantitative variable) and the geographical repartition (qualitative variable). Referring to the Italian geographical repartition, Italy is divided in 20 regions, having a certain degree of homogeneity; very often the regions of North-East, North-West, Center, South, Islands, are considered with similar economic and social behavior. So we consider some different geographical subdivisions: starting from five repartitions (1= North-East, 2= North-West, 3= Center, 4= South, 5= Islands), we arrive, by successive aggregations, to two repartitions (combining in all the possible ways the five original repartitions satisfying a neighboring constraint).

To verify the null hypothesis of MCAR mechanism we estimate a logistic regression model for each stratification adopted, in which the response variable z is defined as:

$$z = \begin{cases} 0 & \text{if number of collaboration months} = 12 \\ 1 & \text{otherwise} \end{cases} \qquad (1)$$

Let $\pi(i)$ the probability that the municipality i has collaborated less than 12 months (z=1 in (1)), $\mathbf{x}_r(i)$ a (r-1)-vector of dummies indicating to which of the r repartitions the municipality i belongs, $x_p(i)$ the number of citizens of municipality i. The logistic regression is given by:

$$\log it[\pi(i)] = \log\left(\frac{\pi(i)}{1-\pi(i)}\right) = \beta_0 + x_r(i)'\beta_r + \beta_p x_p(i) \qquad (2)$$

To choose only one classification in terms of geographical repartitions, we use the Schwartz criterion; the choice falls on the logistic model with 3 geographical repartitions (1=North-East, 2=North-West, 3=rest of Italy). We will note, in the successive steps, that

this solution is not the most feasible.

Of course, the hypothesis of MCAR is verified when probability $\pi(i)$ does not depend on the auxiliary variables, or, in statistical terms, when the coefficients $\boldsymbol{\beta}_r$ and β_p are not significantly different by zero. In Table 3 we show the estimation of the parameters in (2) and the corresponding Wald statistics.

In all the cases the hypothesis of MCAR is rejected. In addition, the non response probability is negatively correlated with the municipalities belonging to North-West and North-East, positively correlated with municipalities of Center, South and Islands. This result confirms that the non response is frequent in the third repartition and the negative coefficient of x_p is consistent with this statement, being the Southern cities smaller and less populated with respect to the cities of North-Italy.

Table 3 - Estimated parameters in logistic regression: 2000-2002

	2000		2001		2002	
	Estimate	Wald χ^2	Estimate	Wald χ^2	Estimate	Wald χ^2
β_0	0.433	193.4	-	-	0.346	125.3
β_1	-0.198	34.9	-0.421	161.4	-0.329	96.2
β_2	0.773	515.4	0.819	612.2	0.917	708.5
β_3	-0.00006	242.4	-0.00006	271.3	-0.00004	130.2

The second step is to verify if these results are consistent with the MAR hypothesis, introducing a stratification based on the previous two variables.

3.1 Identification of the adjustment cells

In the previous subsection, we have determined a classification of the units based on the geographical repartition; the other relevant variable is the population. We determine homogeneous classes with the within variance method, independently on the geographical repartition adopted.

Let *SSW* the sum of within variances; the criterion consists in choosing the number of classes *g* with a number of units respectively equal to $n_1 \ldots n_g$, minimizing

$$\frac{SSW}{n-g} \tag{3}$$

We have to determine the values $l_1, l_2, \ldots, l_h$, as classes boundaries. We have applied a grid search procedure: in the first step we have searched the best subdivision in two groups of population, using as splitting value 10,000, 20,000, 30,000 and 40,000. The criterion of minimum within variance is provided by the classes 0-10,000 and 10,000-50,000. Then we split separately the two previous classes to obtain the best subdivision in three classes (with steps of 2,000 units for the class 0-10,000 and 5,000 units for the class 10000-50,000) and the boundaries are 7,000 and 25,000. With similar splitting we obtain successive gains in terms of reduction of within variance for 4, 5, 6 groups. Anyway, the gain of group 6 in

terms of reduction of variance with respect the group 5 is negligible. The final boundaries of the population classes are 3,000, 7,000, 13,000 and 25,000.

Such thin stratifications, combined with the three strata of repartitions, provide strata with a small number of observed data with respect to the missing values. In Table 4 the composition of the 15 strata obtained by 5 population classes and 3 repartition is showed. We indicate with R_1, R_2, R_3 the geographical repartitions obtained in the previous subsection, and with P_1, P_2, P_3, P_4, P_5, the classification of the population. It is possible to note that the municipalities of the repartition 3 (Center-South-Islands) with small population (classes P_1 and P_2) show a large number of missing values (until 79% of the units), so the application of imputation methods will be very difficult.

Table 4 - Percentage of nonresponse for strata

strata	2000	2001	2002
R_1P_1	53.9	39.0	50.2
R_1P_2	50.3	29.2	40.7
R_1P_3	38.0	27.6	40.3
R_1P_4	26.9	26.0	33.7
R_1P_5	26.2	19.0	33.3
R_2P_1	42.3	42.0	49.9
R_2P_2	44.2	35.0	34.5
R_2P_3	28.8	21.2	26.4
R_2P_4	20.9	5.8	9.3
R_2P_5	12.5	4.2	0.0
R_3P_1	77.8	70.6	79.0
R_3P_2	70.6	59.5	70.6
R_3P_3	56.7	50.2	62.6
R_3P_4	46.4	43.5	63.7
R_3P_5	32.9	36.4	59.3

For this reason, we prefer to apply the imputation methods to 10 strata, obtained by the previous five population classes combined with two geographical repartitions (NW+C+S+I against NE). Also for this case, the MCAR hypothesis is rejected (using the model (2) we reject the null hypothesis of no dependence by geographical repartition and population). Anyway, in the final application, we will try to estimate the missing values also with finer classifications (15 and 20 strata).

However for this data set the MAR hypothesis is not fully respected. To verify if the mechanism of missingness is random within 10 strata we estimate models as in (2) for each stratum, using as geographical repartitions the administrative subdivision in regions. We obtain that for the geographical area with small population the mechanism of missingness is not random; similar results are obtained for the other classifications.

4. The imputation methods used

In this section we illustrate two alternative methods to impute the missing values: a longitudinal method, based on a donor of minimum distance, and a traditional cross-section method based on the mean. We do not consider model-based methods because the MAR hypothesis is not verified and because of the need for an automatic procedure always valid (whereas a model-based technique would require frequent adjustments for the availability of new observations).

We evaluate these methods in terms of capability to estimate the total of the monthly new dwellings series for n_2=7,940 municipalities of the second group in Table 1. In addition, these methods are compared with a weighting procedure applied to each month of the period. In the following subsection we describe the methods that have been used.

4.1 Longitudinal imputation: the donor

In longitudinal imputation it is possible to use the information available in several periods for the estimation of missing values. Heeringa and Lepkowski (1986) group the longitudinal imputation in 5 classes: i) direct longitudinal replacement; ii) deterministic imputation of variations; iii) imputation by longitudinal regression; iv) longitudinal hot-deck; v) longitudinal hot-deck of variations.

The method we use is a direct replacement by the minimum distance donor. As well known, the donor method (or nearest neighbor imputation, *NNI*) provides, under certain hypotheses, asymptotically unbiased and consistent estimators of the total, the distribution and the quantiles of the variable (Chen and Shao, 2000). The application of this method is frequent in many Statistical Offices (i.e. Statistics Canada, U.S. Census Bureau) with an increasing development, mainly due to the diffusion of general software for editing and data imputation (see, for example, the results of the Euredit project in Chambers (2001)).

In the simpler cross-section case, let y the variable of interest in a survey conducted on n units, for which only r of them are respondent. In addition, let x an auxiliary variable observed for all the units. For the sake of simplicity, we suppose that $y_{r+1},\ldots,y_n$ are the n-r missing values. The NNI method estimates y_j, $r+1 \le j \le n$, with the value y_i, $1 \le i \le r$, where i is the donor nearest to j in terms of a given distance measured on the auxiliary variable x:

$$| x_i - x_j | = \min_{1 \le k \le r} | x_k - x_j | \qquad (4)$$

If more than one donor satisfies (4) for the same unit j, then we operate with a random extraction among the equal minimum distance units.

In the longitudinal case, considering the year as period of reference, we consider two possible situations. The first one is the most frequent and represents the case in which the municipality is respondent at least in one month of the reference year. In each stratum, the donor i for the non respondent j is selected minimizing:

$$| x_i - x_j | = \min_{1 \le k \le r_h} \sum_{m \in M} | x_k^m - x_j^m |$$

where M is the set of months (not necessarily consecutive) in which j was respondent

during one year and r_h is the number of respondents for all the 12 months in the stratum h.

The second situation happens when the municipality is non respondent in all the months of the same year; in this case the donor is randomly extracted from the respondent units belonging to the stratum. Actually, we have noted in simulation experiments that the previous distance criterion provides biased estimations of this type of non respondents.

In both cases the same donor is used to estimate all missing values of the unit j; this would keep the seasonal profile of the variable studied or, more in general, its autocorrelation.

An alternative strategy is the imputation of the missing values of the unit j with the same donor for the three years. We will call *simultaneous imputation method* this methodology, whereas *separate imputation method* the case in which we impute separately the missing values of the same unit in different years. We will consider both options in the simulation experiment.

4.2 Cross-section Imputation: the mean

Another method very simple to implement and wise, especially when the purpose is the calculation of the mean or totals, is the mean. It is a cross-section imputation, in the sense that we estimate the missing values for a certain month, only using the information about the respondents of that month.

In particular, for each month we calculate the mean of the number of new dwellings for the respondents for each stratum; this mean would be the estimation of all the missing values belonging to that stratum. In formula:

$$y_h^i = \frac{1}{n_h^r} \sum_{j=1}^{n_h^r} y_j$$

Where y_h^i is the *i-th* non respondent in the stratum h for a certain month and n_h^r is the number of respondents in the stratum h in that month. As well known (see, for example, Little and Rubin, 1987), when the adjustment cells are defined, the application of this method is very easy, but reduces the real variance of the variable studied and provides a bias estimation of its distribution.

4.3 Weighting

In the longitudinal case, when the missing values for a unit in some periods can be interpreted as a set of item nonresponse, a weighting method is an appropriate technique for the estimation. In our application, the observed data can be weighted using weights related to the municipalities propensity to respond. The response probability of the municipality m for each month t can be estimated with the logistic model:

$$\log it[\pi(i)] = \log\left(\frac{\pi(i)}{1-\pi(i)}\right) = \beta_0 + x_r(i)'\beta_r + \beta_p x_p(i) + \beta_d x_d(i) \qquad (5)$$

where $\mathbf{x}_r(i)$ contain the dichotomous variables representing the geographical

repartitions as in (2), x_p represents again the population and x_d is a dichotomic variable, equal to 1 if the municipality i is respondent at time t-12, equal to 0 otherwise.

The probabilities estimated in (5) are not used directly as weights of the respondent units, but we adopt the method based on the adjustment cells (Little, 1986, Eltinge and Yansaneh, 1997). If the cells are efficiently constructed, the units belonging to the same class will show homogeneous response probabilities. These classes are detected using the k (k=2,...,n) quantiles of the distribution of $\pi(i)$. Following this procedure, the estimation of the total of the variable Y is given by:

$$\hat{Y}^k = \sum_{h=1}^{k+1} \frac{n_{s_h}}{r_{s_h}} \sum_{i \in s_h} y_i \qquad (6)$$

where n_{s_h} is the number of units of the class s_h and r_{s_h} is the number of respondents of the same class.

Comparing the estimation results for increasing values of k, it is possible to establish the threshold value k^*; for each $k>k^*$ the difference in the estimation of the total is not significant (for the test used, see Eltinge and Yansaneh, 1997)

In the following applications this method is applied for each month of our series.

5. The simulation experiment

To evaluate the performance of the imputation and weighting methods described in the previous section, we have designed a simulation experiment, based on the classification in 10 strata illustrated in section 3. For the weighting procedure, we have identified the homogeneous cells starting from the logit model (5). In this way, all the methods use the same auxiliary information, dealing with it in different ways. This aspect provides a more correct comparison.

5.1 The simulation design

Our simulation experiment is based on the 1,910 municipalities respondents in all the 36 months. We create artificially the holes in this data set in a random way, but respecting the missingness structure in the data set of 7,940 municipalities. We simulate 500 patterns and on these artificial data sets we apply the methods illustrated in the previous section and the calculation of indicators to evaluate the different strategies of imputation.

Table 5 - Panel of municipalities always respondent

Population	North-West	North-East	Center	South-Islands	Total
$x \leq$ 3,000	606	212	39	83	940
3,000 $< x \leq$ 7,000	192	150	45	64	451
7,000 $< x \leq$ 13,000	87	107	31	51	276
13,000 $< x \leq$ 25,000	50	64	17	26	157
$x >$ 25,000	21	20	22	23	86
Total	956	553	154	247	1,910

The panel of the 1,910 selected municipalities show a strong asymmetry with respect to the geographical repartition, noted also in the analysis of collaboration described in section 3; there is a clear majority of municipalities in the North (table 5).

We have used a 2-steps procedure to generate missing values patterns similar to those observed in the full data set: first of all we select randomly a sample of municipalities representing the non respondents and then we select an intra-year non response pattern.

To clarify this procedure, we refer to the distribution in Table 1. In the first step, we select a random sample constituted by 1,450 municipalities (corresponding to the 76% of the municipalities of the panel data set); on these municipalities we generate the missing values pattern. The selection is performed with a selection probability proportional to the missing value probability estimated with model (2).

For the sample selected, we have dropped the data relative to the months of certain years with the purpose to respect the real collaboration in the data set. For example, from 2000 to 2002 the 12.4% of municipalities has collaborated 35 months; in this case we have dropped one month for 234 of the 1,910 municipalities selected, choosing randomly a response profile among the 987 municipalities which have collaborated 35 months.

5.1.1 The evaluation indicators

In this subsection we describe the set of indicators we use to evaluate the performance of the different approaches in the simulation experiment. We have to distinguish the comparison among the three imputation methods (separate longitudinal imputation, contemporaneous longitudinal imputation and mean) and the comparison of the imputation methods with respect to the weighting procedure. In the first case we have a large set of indicators, being possible to evaluate the differences among the estimation of the aggregates (in terms of comparison of means and variances) and in terms of micro-data. In the second case we make the comparison only in terms of estimation of the monthly total or in terms of year-by-year variations and autocorrelations.

More in details, let $\mathbf{y}^{*}=\left(y_1^{*}, y_2^{*},..., y_n^{*}\right)$ the vector containing the true values of the variable *Y* observed on the *n* units and $\overline{\mathbf{y}} = \left(\overline{y}_1 , \overline{y}_2 ,..., \overline{y}_n\right)$ the corresponding vector with the estimated values. A first measure of the goodness of imputation is obtained comparing the means and the variances of the two empirical distributions, given by:

$$F_{\overline{Y}_n}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\overline{Y}_i \le t) \tag{7}$$

and

$$F_{Y_n^*}(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i^* \le t) \tag{8}$$

where $Y_i = \sum_{j=1}^{i} y_j$ and $Y_i^* = \sum_{j=1}^{i} y_j^*$.

The similarity of the two distribution can be measured by the Kolmogorov-Smirnov statistic:

$$d_{KS}(F_{\bar{Y}_n}, F_{\bar{Y}_n^*}) = \max_t(|F_{\bar{Y}_n}(t) - F_{\bar{Y}_n^*}(t)|) \quad (9)$$

In literature there are several methods to compare the goodness of imputation in terms of micro-data, but the presence of a high percentage of observed values equal to 0 makes the methods based on regressions misleading. For this reason we prefer to base the comparison in terms of similarity measures also in this case; in particular, the measure adopted is:

$$d_1(\mathbf{y}^*, \bar{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} |\bar{y}_i - y_i^*| \quad (10)$$

The comparison of results obtained by the weighting procedure and imputation methods are obtained with the Mean Absolute Percentage Error (MAPE). Considering the n_2 units of the original data set, we calculate the following indicator:

$$d_2 = 100 * \frac{|\sum_{i=1}^{n_2} \bar{y}_i - \sum_{i=1}^{n_2} y_i^*|}{\sum_{i=1}^{n_2} y_i^*} \quad (11)$$

The same indicators (and related graphs), generally adopted in a cross-section context, can be used in a longitudinal dimension to compare the methods in terms of year-by-year variations and autocorrelations.

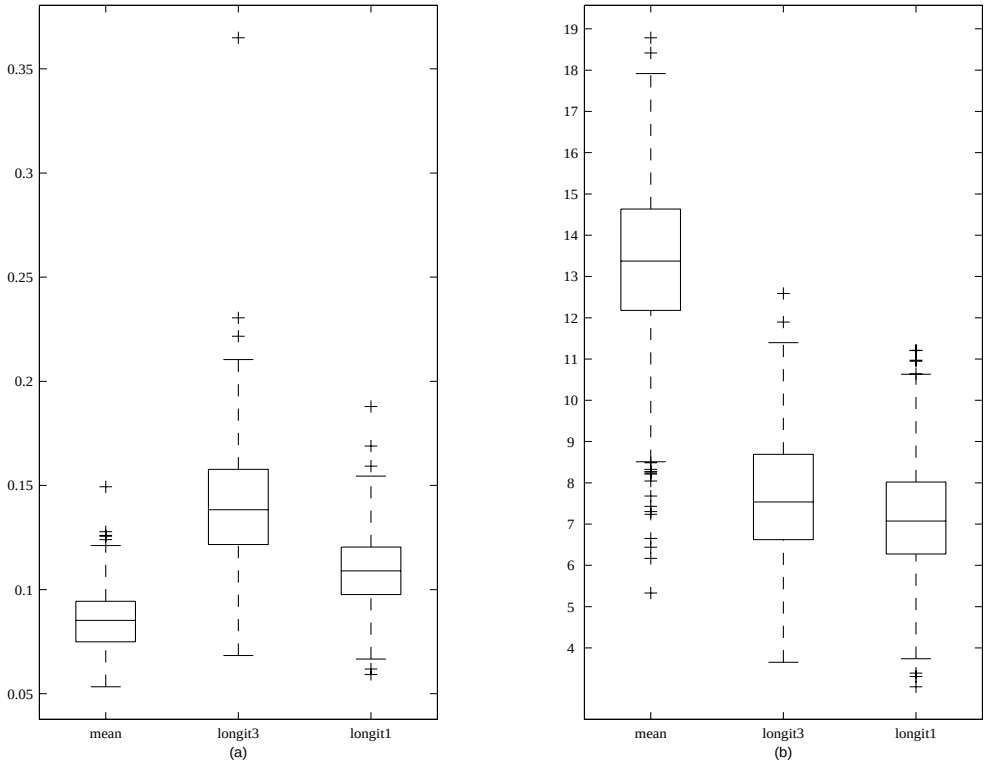
5.2 Results

We synthesize here the results of the comparison made with the previous indicators, evaluated for each month. To favor the analysis and for the sake of space, we prefer to illustrate the results by graphs, using boxplots, in which the means of the 36 values for each simulations are showed.

We indicate with *mean* the imputation method based on the mean, *longit1* the separate longitudinal imputation, *longit3* the simultaneous longitudinal imputation, *weight* the weighting procedure.

In Figure 1 we show the results obtained by the comparison of the means and variances. For each simulation and month we have calculated the absolute differences between the true mean of population and that estimated with the three imputation methods. As known in literature, the *mean* provides a more precise value of the true mean, but reduces the variance artificially. The *longit1* and *longit3* methods provide similar results in terms of variance, whereas the first one shows a better performance in terms of mean estimation.

Figure 1 - Distribution of mean (a) and variance (b) deviations



In Figure 2 we exhibit the boxplots relative to the Kolmogorov-Smirnov statistic. As expected, the *mean* shows better results using the distance d_1, but the bias in estimation variance is dramatic and it is clearly larger than those provided by the donor methods, which produce similar statistics with a smaller variability of the *longit1* method. For this reason, we will not use the method *mean* in the following analysis.

Now, the results obtained with the longitudinal donors are compared with them obtained by the weight method in terms of estimation of the monthly aggregate values.

As previously said, a weighting method needs the identification of adjustment cells, obtained from the quantiles of the response probability. As in other empirical analysis, we have chosen 8 classes, because in this case the estimations seem more stable.

The boxplots of Figure 3 represent the distributions of the distance d_2. The *weight* method show a slightly better behavior with respect to the *longit1* method, whereas the results obtained with *longit3* are the worst in terms of median and for the presence of abnormal values (indicated by a longer boxplot).

The analysis in terms of year-by-year variations and autocorrelations was conducted calculating, for each method, the linear correlation between the estimated and true values. We exhibit the results in Figure 4; they seem acceptable for all the methods with a slight preference for the *longit1* method with respect to the *longit3*.

Figure 2 - Distribution of d_1 (a) and Kolmogorov-Smirnov statistics (b)

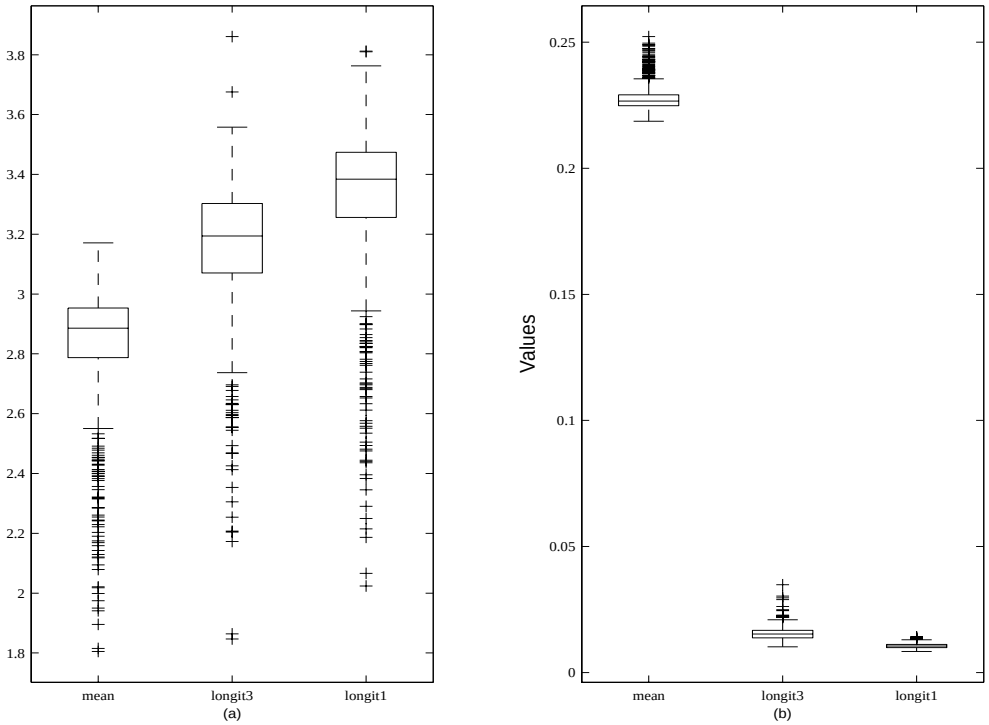


Figure 3 - Distribution of d_2

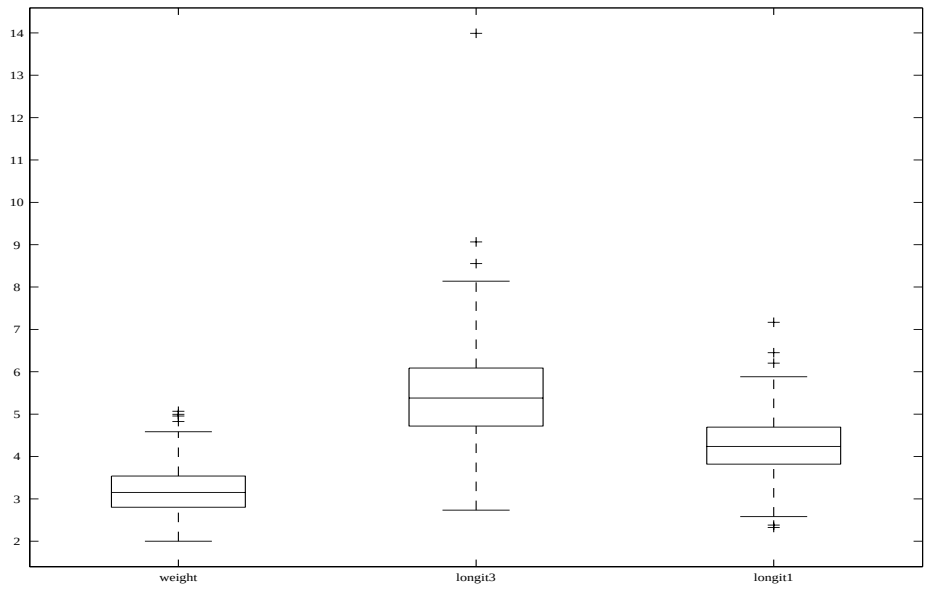
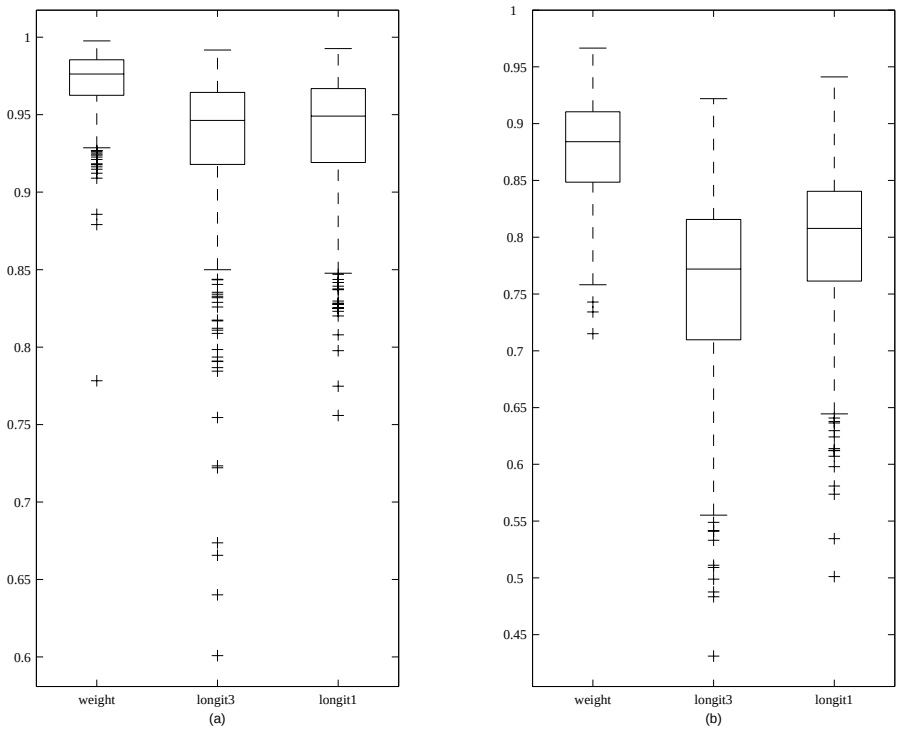


Figure 4 - Distribution of quarterly growth rate (a) and autocorrelation (b)



6. Imputation of the real data set

The simulation experiment previously described has a cross-validation role, in the sense that it suggests to adopt the *longit1* method for the imputation of the real data set, being its results more similar with respect the *weight* method in aggregate terms. We apply this method to estimate the new dwellings missing values for the 7,940 municipalities non provinces with less than 50,000 citizens. In this case, to have empirical evidence that finer stratifications provide worse imputation, as stated in section 3.1, we provide also results with different geographical repartitions, maintaining the same subdivision in terms of population. The alternative are the 2 geographical repartitions used in the simulation experiment, 3 geographical repartitions (North-West, North-East, Rest of Italy, providing 15 strata), 4 geographical repartitions (North-West, North-East, Center, South+Islands, providing 20 strata).

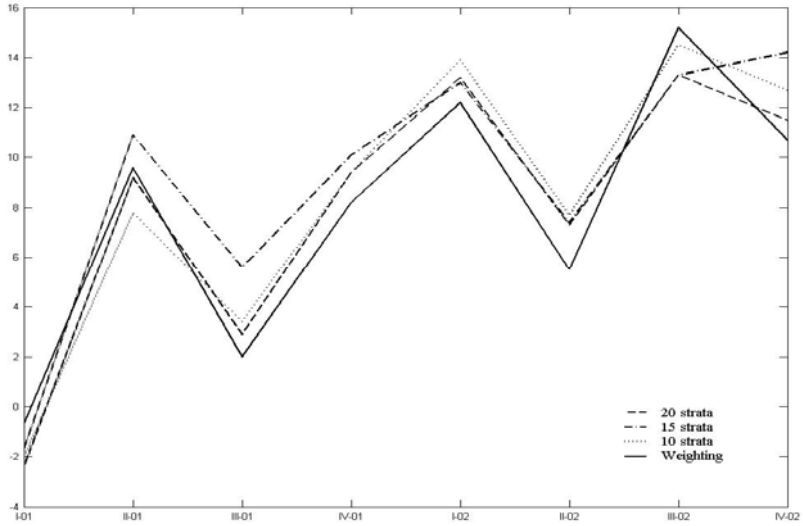
We show in Table 6 the total estimated for each year. The year-by year variations of 2000-2001 show similar results for the 10 strata case (4.7%), the 20 strata case (4.7%) and the weighting methodology (4.9%), whereas for the 15 strata case the variation is 6.2%. In the period 2001-2002 the increase estimated by weighting (10.7%) is small with respect to the imputation methods, which vary between 11.2% and 12.0%. In level terms, the estimation by weighting is greater than that obtained by imputation, but the differences are limited in the case of 10 strata (1.5%, 1.7% and 0.5% respectively for 2000, 2001 and 2002), if compared with the 15 strata (4.3%, 3.0% and 2.0%) and 20 strata cases (3.8%, 3.8% and 3.4%).

Table 6 - Estimated number of dwellings: donor and weighting methods

	Numbers of dwellings				Year growth rate			
	10 strata	15 strata	20 strata	Weight	10 strata	15 strata	20 strata	Weight
2000	144,595	140,562	141,258	146,850	-	-	-	-
2001	151,335	149,322	148,059	153,976	4.7	6.2	4.8	4.9
2002	169,543	167,016	164,628	170,439	12.0	11.8	11.2	10.7

In Figure 5 we show the year-by-year quarterly variations. Now, the variations obtained by imputation methods are similar to those derived by weighting, except for the 15 strata case in the 4-th quarter of 2001. In other terms, the analysis on real data seems to confirm the results of the simulation experiment: the separate longitudinal imputation with 10 strata provides a good imputation technique to reconstruct the data set of new buildings from 2000 to 2002.

Figure 5 - Quarterly growth rate



7. Remarks

In this work we have proposed a method to reconstruct the longitudinal data set of Italian new building dwellings with an imputation technique based on the donor. The data set contains many missing values, so we have proposed an ad hoc procedure, following several criteria. The main steps of this procedure are:

- because MCAR hypothesis is rejected, we have identified a number of strata, based on two auxiliary variables (population and geographical repartition). The choice of the number of strata has followed practical criteria, depending not only on their internal

homogeneity, but also on the number of missing values which they contain;
- we choose the imputation technique with a cross-validation experiment, searching unbiased estimators, providing results similar to those of a weighting technique;
- the time dynamics is evaluated in terms of year-by-year variations and autocorrelations.

The characteristics of the data set have clearly driven our choices, but it is clear that the idea of the selection of strata and cross-validation experiments can be extended to other cases, using also other imputation techniques. Application to the real case supports the Istat decision to use, every year, longitudinal donor techniques to diffuse yearly indicators.

References

[1] Chambers, R. (2001) Evaluation criteria for statistical editing and imputation. *National Statistics Methodological Series*, n. 28, National Statistics, London.

[2] Chen, J. and Shao, J. (2000) Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 113- 131.

[3] Eltinge, J.L. and Yansaneh, I.S. (1997) Diagnostic for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. *Survey Methodology*, 23, 33- 40.

[4] Haziza D., Charbonnier C., Chow O. and Beaumont J.F.(2001). *Construction of imputation cells in the context of the Canadian Labour Force Survey*, Proceedings of Statistics Canada's Symposium 2001.

[5] Heeringa, G.S. and Lepkowski (1986). Longitudinal imputation for the SIPP. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 206- 219.

[6] Kalton, G. (1986) Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, 2, 303- 314.

[7] Little, R.J.A. (1986): Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157, 1986.

[8] Little, R.J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data.* John Wiley & Sons, New York.

[9] Manzari, A. (2004) Combining editing and imputation methods: an experimental application on population census data. *Journal of Royal Statistic Society. Soc. A*, 167, Part 2, 295- 307.

[10] Robins J., Rotnitzky A. and Zhao L. (1995) Analysis of semiparametric regression models for repeated outcomes in presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

[11] Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New York.

[12] Schafer, J.L. (1987) *Analysis of incomplete multivariate data.* Chapman and Hall, New York.

[13] Vella, F. (1998) Estimating models with sample selection bias: a survey. *The Journal of Human Resources*, n. 33, 1, 127-169.

[14] Zhang, L.C. (2001) A method of weighting adjustment for survey data subject to nonignorable nonresponse *Statistics Norway Discussion Papers*, n. 311, Statistics Norway, Kongsvinger.

Matching noise: formalization of the problem and some examples

Mauro Scanu[1]*,Pier Luigi Conti*[2]

Abstract

Statistical matching attempts at producing a unique, synthetic data file, where variables observed in different sample surveys are jointly recorded. Such a file is appropriate for further statistical analysis when the joint probability distribution of the variables of interest in the population coincides with the probability distribution of the same variables in the synthetic data file, or at least when these two distributions are "close enough". The discrepancy between these distributions is called matching noise. In this paper, statistical matching methods based on hot-deck imputation procedures are investigated as a possible cause of matching noise. Two examples when data are generated from uniform and normal distributions are discussed.

Keywords. Data fusion, hot-deck imputation procedures, conditional independence assumption

1. The statistical matching problem

Official statistics produces nowadays many independent sources of data, that are of interest, among others, for econometrics, demography, social statistics, etc.. In several cases, models used in applied statistics are based on many different variables, to be used simultaneously.

The main problem in using the above mentioned sources of data is that each of them refers to a specific sample survey or archive, where only a few variables of interest are observed. Hence, when data coming from different sources are to be used simultaneously in a model, no joint observation of all the variables of interest is available.

Statistical matching techniques attempt at providing a solution to this problem. Their goal consists in producing a unique, synthetic data file, where all the variables of interest are simultaneously present in each record.

Such a goal is pursued by using imputation methods, that essentially fill in missing items in each partially observed record. Of course, this approach is actually successful (and theoretically justified, as well) when the joint probability distribution of the variables of interest in the population coincides with the probability distribution of the same variables in the synthetic (imputed) data file, or at least when these two distributions are "very close".

The discrepancy between the joint distribution of the variables of interest (*a*) in the population, and (*b*) in the synthetic data file is usually referred to as *matching noise* (Paass, 1986). Attempts

1 Ricercatore (Istat), e-mail: scanu@istat.it

[2] Professore ordinario (Università di Roma "La Sapienza"), e-mail: pierluigi.conti@uniroma1.it

at evaluating the "closeness" of the empirical distribution of imputed data to the empirical distribution of "real" data have been performed in the literature (Rodgers and De Vol, 1981; Barr et al, 1981; Rodgers, 1984; Paass, 1986; Barry, 1988). In Rässler (2002) the following normative suggestions for evaluating the quality of a statistical matching procedure are given:

(i) imputed data should coincide with the true, unobserved values;
(ii) the joint distribution of all variables is reflected in the statistically matched file;
(iii) the correlation structure of the variables is preserved;
(iv) the marginal and joint distributions of the variables in the files to match are preserved in the matched file.

These four suggestions seem to be ordered according to their importance: from the reproduction of each single unobserved value to the reproduction of some observed marginal distributions. As a matter of fact, the second suggestion is the most important. It ensures the possibility of having a synthetic complete data file allowing inference on the variables of interest. Actually, if the second suggestion is fulfilled, the matching noise does not exist. On the other hand, the study of the closeness between the two above mentioned distributions is considered very important but difficult, unless the not jointly observed variables in the files to match are independent given the common variables of the files to match. This is the *conditional independence assumption* (CIA, for short).

In this paper, we study the matching noise produced by some commonly used matching techniques under CIA. The basic role of CIA is to simplify the whole analysis by making identifiable the statistical model used for the data at hand. Appropriateness of CIA is discussed in several papers. We quote, among the others, Sims (1972), Rodgers (1984). Although CIA is sometimes unappropriate, we use it merely as a tool to evaluate the presence of matching noise due to imputation. Even in this simplified case, the presence of matching noise can be of great importance.

The paper is organized as follows. In Section 2, the statistical framework for the matching problem is described. In this set up, matching noise is formally defined (Section 3). Than, three usual nonparametric imputation techniques in the "hot-deck family" are discussed, showing if they produce matching noise and how it can be overcome (Sections 4, 5, and 6). Finally, examples when the data sets are generated from a uniform and from a normal distribution are given (Section 7).

2. Statistical framework for matching noise

Let $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ be a $(P+Q+R)$-variate r.v., with

$$\mathbf{X} = (X_1, \ldots, X_P), \mathbf{Y} = (Y_1, \ldots, Y_Q), \mathbf{Z} = (Z_1, \ldots, Z_R)$$

and denote by $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ its joint density function. Let further A and B be two independent samples of size n_A and n_B respectively, generated by $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Assume finally that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is not completely observed in the two samples. Precisely, we suppose that only $(\mathbf{X}, \mathbf{Y})$ is observed in A, and $(\mathbf{X}, \mathbf{Z})$ in B. Hence, $\mathbf{Z}$ is missing in A and $\mathbf{Y}$ is missing in B. As a consequence, the sample data can be written as

$$(\mathbf{x}_a^A, \mathbf{y}_a^A) = (x_{a1}^A, ..., x_{aP}^A, y_{a1}^A, ..., y_{aQ}^A), \quad \text{a}=1,...,n_A$$

$$(\mathbf{x}_{ba}^B, \mathbf{z}_b^B) = (x_{b1}^B, ..., x_{bP}^B, z_{b1}^B, ..., z_{bR}^B), \quad \text{b}=1,...,n_B$$

for samples *A*, *B*, respectively.

The goal is to gain joint information on (**X**, **Y**, **Z**) from the two samples *A*, *B*. As mentioned in Section 1, this problem is generally known as the "statistical matching problem"; see Paass (1986). Our basic assumption is that **Y**, **Z** are independent conditionally on **X**. As written in the introduction, this is the *conditional independence assumption* (CIA).

When the two samples are composed by i.i.d. observations and the density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ belongs to a parametric family of distributions, it is possible to resort to a maximum likelihood approach, whose goal consists in:

1. estimating the marginal distribution of *X* from the overall sample AUB;
2. estimating the conditional distribution of *Y* given *X*, on the basis of the sample *A*;
3. estimating the conditional distribution of *Z* given *X* from the sample *B*.

This procedure is based on the factorization procedure by Rubin (1974). It has been rarely used in practice. In fact, statistical matching has been performed as a particular missing data problem, and has been solved with the reconstruction of complete (synthetic) records by means of *hot-deck* imputation techniques; see, e.g., Little and Rubin (2002). More formally, these methods consist in completing the records of a file (the recipient file, say *A*) by means of the records of the other file (the donor file, say *B*). There are different donor methods, that will be studied afterwards. The final output of this procedure is a new data set $\widetilde{A}$ with records $(\mathbf{x}_a, \mathbf{y}_a, \widetilde{\mathbf{z}}_a)$, a=1,…, n_A, where $\widetilde{\mathbf{z}}_a$ is a **z**-value observed in *B* associated by the imputation technique to record *a* in *A*. The goal of this paper is to study whether these techniques are able to solve the statistical matching problem, i.e. whether the synthetic data set $\widetilde{A}$ is able to give correct inference on the distribution $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

3. Matching noise under CIA: some general aspects

When CIA holds, the joint density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$ can be factorized as:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}).$$

Of course, in the synthetic file $\widetilde{A}$ the (**X**, **Y**)-values are exactly those observed in *A*. Hence, the observed values ($\mathbf{x}_a$, $\mathbf{y}_a$), a=1,…,n_A, are i.i.d. observations from the distribution $f_{\mathbf{XY}}(\mathbf{x},\mathbf{y})=f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$. Let $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})$ be the r.v. representing the donor record from file *B*. Since the observed **x**-values in *A* are generated from **X**, the records $(\mathbf{x}_a, \widetilde{\mathbf{z}}_a)$ in $\widetilde{A}$ are generated from $(\mathbf{X}, \widetilde{\mathbf{Z}})$. The donor procedure works appropriately if the distribution of $(\mathbf{X}, \widetilde{\mathbf{Z}})$ coincides with (is "not too far from") the distribution of (**X**, **Z**). The usual rules on

the factorization of densities lead to:

$$f_{\mathbf{X}\tilde{\mathbf{X}}\tilde{\mathbf{Z}}}(\mathbf{x},\mathbf{t},\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x}) f_{\tilde{\mathbf{Z}}|\mathbf{X}\tilde{\mathbf{X}}}(\mathbf{z} \mid \mathbf{x},\mathbf{t}).$$

Once it is known that $\tilde{\mathbf{X}}$=**t**, $\tilde{\mathbf{Z}}$ and **X** are independent. Furthermore, using the i.i.d. assumption of the observations in *A* and *B*, the distribution of $\tilde{\mathbf{Z}} \mid \tilde{\mathbf{X}}$ coincides with the distribution of **Z**|**X**. Hence:

$$f_{\mathbf{X}\tilde{\mathbf{X}}\tilde{\mathbf{Z}}}(\mathbf{x},\mathbf{t},\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{t}).$$

As a consequence, the synthetic sample data $(\mathbf{x}_a, \tilde{\mathbf{z}}_a)$, a=1,…,n_A, can be considered as observations generated from:

$$f_{\mathbf{X}\tilde{\mathbf{Z}}}(\mathbf{x},\mathbf{z}) = \int f_{\mathbf{X}\tilde{\mathbf{X}}\tilde{\mathbf{Z}}}(\mathbf{x},\mathbf{t},\mathbf{z}) d\mathbf{t} = f_{\mathbf{X}}(\mathbf{x}) \int f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{t}) d\mathbf{t} \qquad (1)$$

If the r.v.s **X**, $\tilde{\mathbf{X}}$ coincide a.s., i.e. if P($\tilde{\mathbf{X}}$ = **X**) = 1, then the two r.v.'s $(\mathbf{X}, \tilde{\mathbf{Z}})$ and (**X**, **Z**) have the same distribution, and there is no matching noise. Clearly, this is equivalent to assume that

$$\mathrm{P}(\tilde{\mathbf{X}} = \mathbf{t} \mid \mathbf{X} = \mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{t} = \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

In all other cases, the two distributions are different. The increase in variability and the (possible) bias due to the presence of the donor distribution $f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x})$ is the *matching noise*. In the next sections, we will illustrate the influence of the matching noise when different donor procedures are considered. This problem has been addressed by many authors (see Sims, 1972; Rodgers, 1984; Paass, 1986; Rässler, 2002, p. 21-22). However, an explicit probabilistic evaluation of the matching noise has never been done. In what follows, the r.v.s **X**, **Y** and **Z** will be assumed absolutely continuous, although many arguments extend to discrete r.v.s.

In subsequent sections, we explicitly evaluate the matching noise for different, widely used imputation techniques. All of them are nonparametric techniques, *i.e.* do not require any special parametric assumption for the density f(**x**, **y**, **z**).

4. Distance hot-deck

Distance hot-deck imputation is probably the most widely used imputation technique. Let D be a $P \times P$ positive definite matrix. Given two P-dimensional vectors **u** and **v**, let d(**u**,**v**) be the corresponding Euclidean distance between the two vectors:

$$d(\mathbf{u}, \mathbf{v}) = \{(\mathbf{u} - \mathbf{v})' D(\mathbf{u} - \mathbf{v})\}^{1/2}.$$

Let further $(\mathbf{x}_a^A, \mathbf{y}_a^A)$ be a fixed record in A. The nearest neighbour record in B to the ath record in A is the vector b^* (=$b^*(a)$) such that:

$$x_{b^*}^B = \arg\min_{1 \le b \le n_B} d(\mathbf{x}_a^A, \mathbf{x}_b^B).$$

Since we are assuming that $\mathbf{X}$ is a continuous r.v., ties do have null probability. Let Γ be the r.v. taking the value b^* for every observed sample, i.e. the "nearest neighbour label" of each record in the sample. Taking into account that the observations in B are i.i.d. r.v.s, Γ is uniformly distributed over 1, …, n_B:

$$P(\Gamma = b) = \frac{1}{n_B}, \quad b=1,...,n_B.$$

In order to evaluate the matching noise, let $\mathbf{\Psi}_b = \mathbf{x}_b^B - \mathbf{x}_a^A$, b=1,…,$n_B$, be the difference between the matching variables with the ath record in A, and:

$$W_b = \mathbf{\Psi}'_b D \mathbf{\Psi}_b = d(\mathbf{x}_a^A, \mathbf{x}_b^B)^2$$

Taking into account that $\mathbf{X}_1^B, ..., \mathbf{X}_{n_B}^B$ is a sample of i.i.d. observations from $f_{\mathbf{X}}(\mathbf{x})$, the sample $\mathbf{\Psi}_1, ..., \mathbf{\Psi}_{n_B}$ is composed by i.i.d. observations with density $f_{\mathbf{\Psi}}(\boldsymbol{\psi})$ and $W_1, ..., W_{n_B}$ is composed by i.i.d. r.v.s with density $f_W(w)$, say. Reorder now the labels b=1,…,n_B so that:

$$W_{n_b:1} \le W_{n_b:2} \le \ldots \le W_{n_b:n_b},$$

$W_{n_b:b}$, b=1,…,n_B being the sample bth *order statistic*.

It is immediate to see that Γ is equal to the label of the W_i observation corresponding to $W_{n_b:1}$. Furthermore, by elementary algebra we obtain

$$P(\Gamma = b, \mathbf{\Psi}_{n_b:1} \le \psi)$$

$$= P(\Gamma = b) P(\mathbf{\Psi}_{n_b:1} \le \psi \mid \Gamma = b)$$

$$= \frac{1}{n_B} P(\boldsymbol{\Psi}_b \le \psi, W_k \ge W_b \ \forall k \ne b \mid \Gamma = b)$$

$$= \frac{1}{n_B} \int_{\Re^P} P(\boldsymbol{\Psi}_b \le \psi, W_k \ge W_b \ \forall k \ne b \mid \Gamma = b, \boldsymbol{\Psi}_b \le \mathbf{x}) f_{\boldsymbol{\Psi}}(\mathbf{x}) d\mathbf{x}$$

$$= \frac{1}{n_B} \int_{(-\infty,\psi]} P(W_k \ge \mathbf{x}' D\mathbf{x} \ \forall k \ne b) f_{\boldsymbol{\Psi}}(\mathbf{x}) d\mathbf{x}$$

$$= \frac{1}{n_B} \int_{(-\infty,\psi]} P(W_k \ge \mathbf{x}' D\mathbf{x})^{n_B - 1} f_{\boldsymbol{\Psi}}(\mathbf{x}) d\mathbf{x},$$

where $(-\infty,\psi]=\{\mathbf{a} \in \Re^P: \mathbf{a} \le \psi\}$ is the orthant of (upper) vertex ψ, and $\boldsymbol{\Psi} \le \psi$ is a componentwise inequality.

Hence, the marginal density of $\boldsymbol{\Psi}_{n_B:1}$ is:

$$f_{\boldsymbol{\Psi}_{n_B:1}}(\psi) = P(W \ge \psi' D \psi)^{n_B - 1} f_{\boldsymbol{\Psi}}(\psi), \quad \psi \in \Re^P. \tag{3}$$

Using the relationship $\widetilde{\mathbf{X}}_a^A = \boldsymbol{\Psi}_{n_B:1} + \mathbf{x}_a^A$, we then have the following result

Proposition 1 The density of $\widetilde{\mathbf{X}}_a^A = \mathbf{X}_{b^*}^B$, given $\mathbf{X}_a^A = \mathbf{x}_a^A$, is equal to

$$f_{\widetilde{\mathbf{X}}_a^A | \mathbf{X}_a^A}(\mathbf{x} \mid \mathbf{x}_a^A) = f_{\boldsymbol{\Psi}_{n_B:1}}(\mathbf{x} - \mathbf{x}_a^A), \quad \mathbf{x} \in \Re^P, \tag{4}$$

where $f_{\boldsymbol{\Psi}_{n_B:1}}(.)$ is given by (3)

Through Equation (1), we obtain the density of $(\mathbf{X}^A, \widetilde{\mathbf{Z}}^A)$ when the distance hot-deck imputation method is used.

It can be shown (Paass, 1985; Cohen, 1991) that distance hot-deck is equivalent to impute missing data through the conditional expectation of **Z** given **X** estimated by the (nonparametric) *kNN* nearest neighbour method, with k=1. This is actually the most important theoretical justification of distance hot-deck. At the same time, it explains why this method produces matching noise, exactly evaluated by (4). Regression equations are able to represent a functional dependence between the expected values of **Z** and **X**. The regressed values $\widetilde{\mathbf{Z}}$ lack of the residual variability of the regression. Adjusting the

regressed values in order to account for the residual variability can reduce the matching noise.

Another approach is suggested by intuition: the matching noise should decrease as n_B increases. This is shown in Proposition 2, under minimal regularity conditions.

Proposition 2 Let ε be a positive number, and let **ε** be a P-dimensional vector with components all equal to ε. Denote further by $(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})$ the open square of sides 2ε centered around $\mathbf{x}_a^A$, and assume that the support of W contains an interval of the form (0, t) for some positive t. Then:

$$\lim_{n_B \to \infty} P\left(\widetilde{\mathbf{X}}_a^A \notin (\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon}) \middle| \mathbf{X}_a^A = \mathbf{x}_a^A\right) = 0 \quad \forall \varepsilon > 0. \tag{5}$$

Proof Denote by $(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c$ the complement of $(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})$. Using relationship (4), we have first

$$P\left(\widetilde{\mathbf{X}}_a^A \notin (\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon}) \middle| \mathbf{X}_a^A = \mathbf{x}_a^A\right) = \int_{(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c} f_{\boldsymbol{\Psi}_{nB:1}}(\mathbf{x} - \mathbf{x}_a^A) d\mathbf{x}$$

$$= \int_{(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c} P\left(W \geq (\mathbf{x} - \mathbf{x}_a^A)' D (\mathbf{x} - \mathbf{x}_a^A)\right)^{n_B - 1} f_{\boldsymbol{\Psi}}(\mathbf{x}) d\mathbf{x}$$

Now the term $(\mathbf{x} - \mathbf{x}_a^A)' D (\mathbf{x} - \mathbf{x}_a^A)$ is strictly positive for every **x** in $(\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c$, from which the inequality

$$P\left(W \geq (\mathbf{x} - \mathbf{x}_a^A)' D (\mathbf{x} - \mathbf{x}_a^A)\right) < 1 \quad \forall \mathbf{x} \in (\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c$$

follows. Hence, we also have

$$\lim_{n_B \to \infty} P\left(W \geq (\mathbf{x} - \mathbf{x}_a^A)' D (\mathbf{x} - \mathbf{x}_a^A)\right)^{n_B - 1} = 0 \quad \forall \mathbf{x} \in (\mathbf{x}_a^A - \boldsymbol{\varepsilon}, \mathbf{x}_a^A + \boldsymbol{\varepsilon})^c$$

from which (5) follows.

5. Random hot-deck

Random hot-deck is probably the simplest imputation technique. It is based on an elementary idea: draw at random a donor from file *B*, for each record in *A*. Hence, the donor density function $f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{u} \mid \mathbf{x})$ is equal to the marginal density $f_{\mathbf{X}}(\mathbf{u})$. As a consequence, the following relationship holds true.

$$f_{\mathbf{X}\tilde{\mathbf{Z}}}(\mathbf{x},\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})\int f_{\mathbf{X}}(\mathbf{u}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z} \mid \mathbf{u}) d\mathbf{u} = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Z}}(\mathbf{z})$$

The positive feature of random hot-deck is that it does not alter the marginal distribution of **Z**. On the other hand, it forces **X** and **Z** to be marginally independent. Hence, random hot-deck does not produce any matching noise for the marginal distribution of **Z**. All the matching noise is due to the difference between the actual joint density $f_{\mathbf{XZ}}(\mathbf{x},\mathbf{z})$ and the product $f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Z}}(\mathbf{z})$. Of course, random hot-deck is noise free if and only if $f_{\mathbf{XZ}}(\mathbf{x},\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Z}}(\mathbf{z})$.

To overcome this problem, in case of categorical variables it is possible to use *conditional random hot-deck* imputation, which is based on imputation classes. More precisely, assume that **X** is categorical, and partition *B* into classes composed by sample units having a common **X** category. Conditional random hot-deck is implemented as follows:

(i) for each record *a* in *A*, the set of donors is composed by the records *b* in *B* such that $\mathbf{x}_b^B = \mathbf{x}_a^A$;

(ii) select at random a donor from the set of donors in (i)

Conditional random hot-deck is defined when **X** is a discrete r.v.. A simple adaptation of the arguments used in Section 3 shows that in this case there is no matching noise. In fact, conditional random hot-deck corresponds to impute missing items through the estimated (in *B*) conditional cumulative distribution function of *Z* given *X*. Given that this procedure fulfills (2), the matching noise is absent provided that each *X* category is observed in *B*, as well as in *A*.

A theoretically well founded extension to the case of a continuous r.v. **X** is still lacking. This will be the subject of a subsequent paper.

6. Rank hot-deck

Although rank hot-deck (Singh et al., 19:93) is not frequently used, it is of some interest to evaluate its matching noise.

For the sake of simplicity, assume that samples *A*, *B* are of the same size: $n_A = n_B = n$, and that the r.v. *X* is univariate (*P*=1). Denote further by

$$X_{n:1}^A,\ldots,X_{n:n}^A, X_{n:1}^B,\ldots,X_{n:n}^B$$

the ordered *X*-values in samples *A*, *B*, respectively. If the *i*-th observation in sample *A* (*B*)

corresponds to the *j*-th ordered *X*-value in *A* (*B*), we will say that its *X*-rank in *A* (*B*) is *j*.

Rank hot-deck consists in matching observations with the same *X*-rank in the two samples. Hence, the donor distribution $f_{\tilde{X}|X}(u \mid x^A_{n:j})$ for the observation of *X*-rank *j* in *A* is the distribution of the observation of *X*-rank *j* in *B*, $f_{X^B_{n:j}}(u)$. This remark leads to the following relationship:

$$f_{\tilde{X}|X}(u \mid x^A_{n:j}) = \frac{n!}{(j-1)!(n-j)!} P(X < u)^{j-1} P(X > u)^{n-j} f_X(u)$$

7. Examples

In the following, two different examples show the effect of the matching noise when distance hot-deck (Section 4) is used. In the first case, *X* is uniformly distributed. In the second example, *X* is normally distributed. The key issue in both the examples is the evaluation of the density $f_{\tilde{\mathbf{X}}|\mathbf{X}}(\mathbf{t} \mid \mathbf{x})$, which is the cause of the matching noise.

Example 1 (Uniform distribution). Let *X* be distributed as a uniform r.v. in the interval (α,β): $X_b \approx$Unif(α,β), *b*=1,…,n_A. Let $X^A_a = x^A_a$ be the value of *X* on the *a*-th unit in *A*. Assessing the matching noise means evaluating the distribution of *X* after matching the *b*-th record in *B* with the *a*-th record in *A*. Consequently $\Psi_b \approx Unif(\alpha - x^A_a, \beta - x^A_a)$, *b*=1,…,$n_B$, with density:

$$f_\Psi(\psi) = \begin{cases} 1/(\alpha-\beta) & \alpha - x^A_a < \psi < \beta - x^A_a \\ 0 & otherwise \end{cases}$$

The distribution function of $W = |\Psi|$ is equal to:

$$\begin{aligned} F_W(w) &= \int_0^w f_W(t)dt \\ &= \int_{-w}^w f_\Psi(t)dt \\ &= \begin{cases} 0 & w < 0 \\ 2w/(\beta-\alpha) & 0 < w < \min \\ \begin{cases} (w + x^A_a - \alpha)/(\beta-\alpha) & x^A_a - \alpha < \beta - x^A_a \\ (w + \beta - x^A_a)/(\beta-\alpha) & x^A_a - \alpha > \beta - x^A_a \end{cases} & \min < w < \max \\ 1 & w > \max \end{cases} \end{aligned}$$

where $\min = \min(x_a^A - \alpha, \beta - x_a^A)$ and $\max = \max(x_a^A - \alpha, \beta - x_a^A)$. Therefore, the density function of W is given by

$$f_W(w) = \begin{cases} 0 & w < 0 \\ 2/(\beta - \alpha) & 0 < w < \min \\ 1/(\beta - \alpha) & \min < w < \max \\ 0 & w > \max \end{cases}$$

and density $X_{b^*}^B \mid X_a^A$ after distance matching is

$$f_{\tilde{\mathbf{X}}_a^A | \mathbf{X}_a^A}(\mathbf{x} \mid \mathbf{x}_a^A) = \begin{cases} 0 & & x < \alpha \\ \left[1 - \dfrac{2\left|x - X_a^A\right|}{\beta - \alpha}\right]^{n_B - 1} \dfrac{1}{\beta - \alpha} & & 0 < \left|x - X_a^A\right| < \min \\ \left[1 - \dfrac{\left|x - X_a^A\right| + \left|\alpha - X_a^A\right|}{\beta - \alpha}\right]^{n_B - 1} \dfrac{1}{\beta - \alpha} & X_a^A - \alpha < \beta - X_a^A & \\ & & \min < \left|x - X_a^A\right| < \max \\ \left[1 - \dfrac{\left|x - X_a^A\right| + \left|\beta - X_a^A\right|}{\beta - \alpha}\right]^{n_B - 1} \dfrac{1}{\beta - \alpha} & X_a^A - \alpha > \beta - X_a^A & \\ 0 & & x > \beta \end{cases}$$

Note that the distribution of the matching noise $X^B|X^A$ is generally not symmetric with respect to X^A. This is true also for a symmetric distribution, as the uniform. Hence, in the average, the matching noise does not reproduce the true sample, i.e. those X^A which have been observed in A.

However, the additional variability due to the matching noise, as well as the bias, decrease when n_B diverges, as a consequence of Proposition 2.

Example 2: the normal distribution Let (X,Y,Z) be a three variate normal r.v., with parameters

$$\begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{bmatrix}.$$

The r.v.s X_a^A, a=1,…,n_A, and X_b^B, b=1,…,n_B, are i.i.d. normal with mean μ_X and

variance σ_X^2. Hence, $\Psi_b \mid X_a^A = x_a^A$ is distributed as $X_b^B - x_a^A$, i.e. it is normal with mean $\mu_X - x_a^A$ and variance σ_X^2. Let *D* be equal to 1, so that *W* is based on the "standard" Euclidean distance.

The distribution function of $\Psi_{n_B:1}$, given $X^A = x_a^A$, is:

$$P\left(\Psi_{n_B:1} \le \psi\right) = \int_{-\infty}^{\psi} P\left(W \ge x^2\right)^{n_B-1} f_\Psi(x)dx$$

$$= \int_{-\infty}^{\psi} P\left(\Psi^2 \ge x^2\right)^{n_B-1} f_\Psi(x)dx$$

$$= \int_{-\infty}^{\psi} \left[P\left(\Psi \ge x\right) + P\left(\Psi \le -x\right)\right]^{n_B-1} f_\Psi(x)dx$$

$$= \int_{-\infty}^{\psi} \left[1 - \Phi\left(\frac{x-(\mu_X - x_a^A)}{\sigma_X}\right) + \Phi\left(-\frac{x+(\mu_X - x_a^A)}{\sigma_X}\right)\right]^{n_B-1} f_\Psi(x)dx,$$

where Φ(.) is the normal standard distribution function. The density function of $\Psi_{n_B:1}$, given $X_a^A = x_a^A$ is then

$$f_{\Psi_{n_B:1}}(\psi) = \frac{\left[1 - \Phi\left(\frac{\psi-(\mu_X - x_a^A)}{\sigma_X}\right) + \Phi\left(-\frac{\psi+(\mu_X - x_a^A)}{\sigma_X}\right)\right]^{n_B-1} \exp\left\{-\frac{\left(\psi-(\mu_X - x_a^A)\right)^2}{2\sigma_X^2}\right\}}{\sigma_X\sqrt{2\pi}}$$

As a result, the joint density of *X* and $\widetilde{Z}$ is:

$$f_{X\widetilde{Z}}(x,z) = f_X(x)\int_{\Re} f_{\Psi_{nB:1}}(t-x) f_{Z|X}(z \mid t)dt$$

$$= f_X(x)\int_{\Re}\left[1 - \Phi\left(\frac{t-\mu_X)}{\sigma_X}\right) + \Phi\left(-\frac{t+\mu_X-2x}{\sigma_X}\right)\right]^{n_B-1} \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left\{-\frac{\left(t-\mu_X\right)^2}{2\sigma_X^2}\right\}$$

$$\times \frac{1}{\sqrt{2\pi\left(\sigma_Z^2 - \frac{\sigma_{XZ}^2}{\sigma_X^2}\right)}} \exp\left\{ - \frac{\left(z - \mu_Z - \frac{\sigma_{XZ}}{\sigma_X^2}\mu_X\right)^2}{2\left(\sigma_Z^2 - \frac{\sigma_{XZ}^2}{\sigma_X^2}\right)} \right\} dt$$

8. Conclusions

In this paper, the problem of the appropriateness of statistically matched files via hot-deck imputation procedures is investigated. More precisely, the focus has been posed on the matching noise, which is the cause of the discrepancy between the data generating model and the imputations generating model. It is claimed that, unless the common variable **X** is categorical and conditional random hot-deck is used, the imputed data file is subject to matching noise. A number of issues deserve further attention and will be studied in further works.

(i) An appropriate distance between distributions should be chosen in order to evaluate more clearly the presence of matching noise.

(ii) It is not always possible to compute explicitly the matching noise, as in the case X is normal (Section 7). In these cases, it is necessary to resort to simulation procedures.

(iii) Appropriate simulations seem the best way to investigate the presence of the matching noise when the previously described hot-deck procedures are constrained (records in B can play the role of donors only once).

(iv) The matching problem is a simplified case of a missing data problem. Even in the case of partially observed data sets, imputations can be subject to the matching noise. However, matching noise should be studied in conjunction with the missing data mechanism. In fact, a large matching noise for an improbable pattern of missing data can have a negligible effect on the overall appropriateness of the imputed file.

(v) The previously described hot deck procedures make use of some nonparametric estimators, as the *kNN* for the distance hot-deck method. It is important to investigate the use of other nonparametric estimators.

References

R.S. Barr, W.H. Stewart, and J.S. Turner (1981). An empirical evaluation of statistical matching methodologies. Technical report, School of Business, Southern Methodist University, Dallas.

J.T. Barry (1993). An investigation of statistical matching. *Journal of Applied Statistics*, 15, 275–283.

M.L. Cohen (1991). Statistical matching and microsimulation models. In *Improving Information for Social Policy Decisions, the Use of Microsimulation Modeling*, volume II. National Academy Press, Technical Papers.

R.J.A. Little, and D.B. Rubin (2002). *Statistical Analysis With Missing Data* (II Ed.). John Wiley & Sons, New York.

G. Paass (1985). Statistical record linkage methodology, state of the art and future prospects. In *Bulletin of the International Statistical Institute*, Proceedings of the 45th Session, volume LI, Book 2.

G. Paass (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In: *Microanalytic Simulation Models to Support Social Fiscal Policy* (Eds. G.H. Orcutt, J. Merz, H. Quinke), pages 401–420. Elsevier Science, Amsterdam.

S. Rässler (2002). *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag, New York.

W.L. Rodgers (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91–102.

W.L. Rodgers, and E. De Vol (1981). An evaluation of statistical matching. In *Proceedings of the American Statistical Association*, Section on Survey Research methods, 128–132.

D.B. Rubin (1974). Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69, 467–474.

C.A. Sims (1972). Comments on: "Constructing a new data base from existing microdata sets: the 1966 merge file", by B.A. Okner. *Annals of Economic and Social Measurements*, 1, 343–345.

A. Singh, H. Mantel, M. Kinack, and G. Rowe (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59–79.

Indicatori di competitività turistica: il quadro teorico e la realtà italiana

Roberto Gismondi[1]

Sommario

L'impulso ispiratore del lavoro è derivato dalla necessità di valorizzare la componente territoriale nelle basi di dati connesse al turismo, tramite l'alimentazione di un database georeferenziato ed il calcolo di indici di attrattività turistica al livello delle 547 circoscrizioni turistiche oggi esistenti. Con riferimento alla scelta delle 34 variabili utilizzate, queste sono state suddivise in 7 tipologie: territorio ed ambiente, dotazione di infrastrutture, attrattive storiche e naturali, altre attrattive, dotazione di posti-letto, profilo economico turistico, domanda turistica finale. Tramite una successiva sintesi in 3 macro-indici di competititività turistica (attrattività, dotazione di posti-letto, impatto turistico), si è pervenuti infine ad un indice globale di attrattività, che ha consentito di segmentare il territorio nazionale in base alle componenti principali - latenti rispetto alla base dati osservata - che sintetizzano la maggior parte dell'informazione ed identificano profili utili al fine di analizzare la turisticità di un sito.Parole chiave: disoccupazione, regioni, flussi, previsione

Abstract

This article has been pushed by the need to turn to account the territorial factor in statistical databases concerned with tourism, on the basis of the building up of an integrated database at the level of 547 tourist territorial bodies, and the calculation of specific tourist attractiveness indexes. With reference to selection of the 34 variables included in analysis, these have been broken down in 7 typologies: territory and environment, infrastructures supply, historical and natural attractions, other attractions, bed-places supply, tourist economic profile, final tourist demand. Through a further synthesis into 3 tourist competitiveness macro-indexes (attractiveness, bed-places supply, tourism impact), we finally got a global attractiveness index, aimed at discriminating national territory on the basis of principal components – latent respect to the observed database – which synthesise the most part of information and identify really relevant local tourist profiles.

1. Introduzione

La valorizzazione del territorio e delle sue risorse rappresenta un obiettivo strategico di primaria importanza per gli analisti e gli operatori del comparto turistico. Lo strumento più utile per perseguirlo consiste nell'accrescere le conoscenze circa i fattori sociali, economici e culturali che tuttora determinano forti asimmetrie nei livelli locali di sviluppo della

[1] Primo ricercatore (Istat), e-mail: gismondi@istat.it.

domanda. Tali fattori dipendono dall'interazione tra un insieme di attività in cui domanda ed offerta reagiscono a stimoli fortemente condizionati dalla componente territoriale.

In tale ottica si parla spesso di *competitività* di una destinazione turistica: in generale, al crescere della competitività dovrebbero associarsi aumenti di produttività e crescita nella ricerca dell'efficienza, secondo una teoria propriamente richiamata in molti contesti finalizzati all'analisi del settore terziario. Essa è da intendersi come "globale", in quanto sintesi di una molteplicità di fattori strategici, esprimibili tramite opportuni indicatori di attrattività turistica, che oltre alla turisticità intrinseca della località dovrebbero sintetizzare anche una serie di altri fattori connessi con lo sviluppo turistico effettivo e potenziale.

In effetti, per chi si appresta ad effettuare studi o a prendere decisioni strategiche nel campo del turismo, è di fondamentale importanza poter disporre di un'ampia gamma di informazioni quantitative e qualitative derivanti dal territorio - o anche georeferenziate - che descrivano al meglio le caratteristiche dei siti d'interesse. Meno evidente e tuttora poco affrontato è il problema della sintesi, tramite opportuni indicatori, del contenuto informativo di questa molteplicità di variabili. Nel dettaglio, l'idea della *georeferenziazione turistica* sottintende un approccio al territorio il più analitico possibile e che si identifica convenzionalmente a livello comunale.

In tale contesto, la *Riforma della legislazione nazionale del turismo*[2] costituisce un atto normativo che attribuisce un'importanza fondamentale alle realtà turistiche locali e, quindi, implicitamente stimola la creazione di un quadro informativo turistico di base il più vasto e dettagliato possibile. In particolare, è esplicitamente richiamato all'attenzione il tema dei *sistemi turistici locali* (*STL*), definiti all'articolo 5 come "…*Contesti turistici omogenei o integrati, comprendenti ambiti territoriali appartenenti anche a regioni diverse, caratterizzati dall'offerta integrata di beni culturali, ambientali e di attrazioni turistiche, compresi i prodotti tipici dell'agricoltura e dell'artigianato locale, o dalla presenza diffusa di imprese turistiche singole o associate*". La precedente definizione evidenzia come:

1. Le aree sistemiche si possono caratterizzare per due peculiarità interne non necessariamente coesistenti: *omogeneità* ed *integrazione*.
2. L'identificazione del sistema non si può basare solo sulla valutazione della presenza di strutture ricettive, ma anche sull'attivazione economica indotta dalla turisticità del sito e dall'*offerta integrata* di beni culturali, ambientali e di attrazioni turistiche.
3. È di fondamentale importanza poter disporre di basi informative molto dettagliate e livello territoriale.

Secondo tale approccio l'attenzione si sposta dal concetto tradizionale di destinazione in quanto *luogo*, delimitato da criteri geografici ed amministrativi ed in cui le attività turistiche si sono sviluppate in misura differenziata, a quello di sistema integrato, nel quale il turismo viene percepito come un prodotto composito. In altri termini, la competitività di una destinazione nasce dall'interazione sinergica tra le risorse attrattive primarie (naturali, umane, artificiali), le infrastrutture che ne agevolano la fruizione, le imprese turistiche, le industrie complementari e di supporto al settore turistico, la popolazione residente e la domanda turistica.

Va precisato, inoltre, che molto spesso in fase di valutazione dell'offerta turistica di un territorio si tende a prendere in considerazione solo le imprese i cui servizi sono destinati

[2] Testo approvato dalla Camera dei deputati il 27 febbraio 2001.

esclusivamente o prevalentemente ai turisti. In realtà, oggetto della produzione turistica è tutto ciò che serve a soddisfare le esigenze del turista stesso. Di conseguenza, occorre ampliare il concetto di *impresa turistica* dalle imprese cosiddette di prima linea[3] a quelle di seconda linea[4].

Per coniugare l'impulso sistemico che anima la nuova normativa con l'esigenza di sedimentare ed interpretare i fattori di attrattività locale, appare di fondamentale importanza poter assegnare ad ogni porzione di territorio un indice di turisticità (o competitività) turistica.

In effetti, i sistemi turistici locali, comunque vengano definiti, devono necessariamente contenere una o più località di elevato interesse turistico, effettivo e/o potenziale. Prima ancora che la legislazione introducesse il concetto di sistema, l'Istat (2000, pagg. 222-232) aveva elaborato una definizione essenzialmente operativa di *STL*. Il concetto ispiratore fondamentale era stato quello della cosiddetta "integrazione della filiera", basata su un'analisi territoriale del grado di sviluppo di alcune attività economiche direttamente od indirettamente legate alla fruizione del territorio in chiave turistica. In tal modo i sistemi turistici locali erano costituiti da quei comuni che rispettano simultaneamente una particolare *condizione di specializzazione* nelle quattro categorie di attività economica ritenute rilevanti nel settore turistico[5].

La letteratura specialistica recente offre una vasta gamma di iniziative assimilabili a quella oggetto d'interesse, che può essere inquadrata nel più ampio palinsesto teorico del *destination management*[6]. Già a metà degli anni '90 una valutazione prevalentemente qualitativa dello sviluppo turistico delle regioni del Mezzogiorno era stata proposto da Costa *et al*. (1996).

Una tecnica di sintesi delle informazioni turistiche territoriali deriva poi dall'indagine campionaria alle frontiere – mirata alla valutazione del turismo internazionale da e verso l'Italia – condotta dall'Ufficio Italiano dei Cambi (1998). In particolare, è stato proposto un *indice sintetico regionale di attrattività turistica* - per quanto circoscritto alle sole regioni del Mezzogiorno - basato sulla stima delle entrate valutarie in tale area e sulla misurazione del livello di allineamento tra risorse disponibili e risorse effettivamente sfruttate.

È però nell'ambito di un successivo progetto strategico sul turismo[7] che è stata condotta un'azione articolata finalizzata ad una prima identificazione di indicatori utili per la georeferenziazione delle località turistiche italiane, con un'applicazione ai comuni del Veneto.

Una tecnica di analisi territoriale, finalizzata alla valutazione della competitività turistica delle regioni e delle province italiane, è stata proposta nel *Decimo Rapporto sul Turismo Italiano*[8]. In tale contesto il criterio di analisi territoriale si era limitato a poche

[3] Ossia quelle che vendono direttamente al turista, la cui attività è in tutto o in parte dipendente dal turismo e che in molti casi sono caratterizzate da una forte connotazione stagionale (ricettività, trasporti, ristorazione, ecc.).

[4] Ossia quelle che non entrano in contatto con il turista, ma i cui beni e servizi vengono venduti alle imprese di primo livello ed impiegate nel loro processo produttivo per soddisfare i bisogni dei turisti.

[5] Servizi di accoglienza di fascia alta: alberghi (codifica ATECO 55.1); servizi di accoglienza di fascia medio-bassa: campeggi ed altri alloggi per brevi soggiorni (55.2); servizi di ristoro e di intrattenimento: ristoranti, bar, discoteche, sale da ballo, *night club* e simili, stabilimenti balneari (55.3, 55.4, 92.34.2); servizi di assistenza al turista: attività delle agenzie di viaggio e turismo, attività delle guide e degli accompagnatori turistici (63.3).

[6] Tamma (2001, pagg. 31-54).

[7] Finanziato dal Consiglio Nazionale delle Ricerche (CNR) nel biennio 1998-1999 e sviluppato in collaborazione con l'Istat. Per dettagli si confronti Greco (1999, pagg. 345-386).

[8] Gismondi (2001, pagg. 101-145).

variabili: un indicatore di penetrazione del mercato turistico, uno di impatto ambientale e due relativi alla composizione della clientela. I risultati emersi confermavano la persistenza del dualismo Nord-Sud e l'eccellenza di province storicamente e strategicamente votate al turismo, sebbene la valutazione di parametri di efficienza svincolati dalle sole quote di mercato (e, quindi, dalla dimensione assoluta del fenomeno turistico) consentisse di individuare contesti territoriali almeno parzialmente inediti.

Uno schema operativo espressamente finalizzato alla selezione di variabili significative per connotare l'attrattività turistica locale del Mezzogiorno è stato successivamente proposto in un recente lavoro da Landi (2003).

Infine, nella medesima prospettiva del suddetto studio CNR-Istat va inquadrata la recente esperienza condotta sui 64 comuni della provincia di Foggia[9], i cui dettagli metodologici sono stati approfonditi da Gismondi e Russo (2004; 2005).

Queste ultime iniziative hanno fortemente ispirato l'approfondimento tematico proposto in questa sede. Infatti, l'obiettivo di fondo è quello di individuare una metodologia d'analisi, generalizzabile a qualunque porzione d territorio, in grado di valorizzare quantitativamente la *turisticità* non solo dei comuni situati in aree turisticamente rilevanti ma anche, e sotto certi versi soprattutto, dei *siti turistici minori*.

Per l'applicazione della suddetta metodologia al territorio italiano, è risultato necessario progettare una base dati che contenesse le informazioni utili allo scopo prefissato ed identificare una tecnica di calcolo di un indicatore territoriale di attrattività turistica.

Riguardo alla scelta delle variabili, l'elemento innovativo è consistito nell'impiego congiunto di variabili tradizionali negli studi di georeferenziazione turistica e di altre legate, invece, a fenomeni culturali locali - la cui incidenza sull'attrattività turistica ha assunto un ruolo significativo solo a partire dagli anni '90 – o di tipo ambientale.

Riguardo alla tecnica di calcolo, si è optato per una tecnica di analisi fattoriale piuttosto nota nel campo della statistica applicata, ma poco sfruttata nelle applicazioni inerenti il turismo. La necessità di ridurre le dimensioni di una base dati integrata la cui disponibilità, per quanto fondamentale, non implica automaticamente una lettura più chiara ed approfondita del potenziale attrattivo di un territorio, rappresenta una tappa fondamentale per gli analisti e consente di identificare le variabili effettivamente rilevanti per determinare la turisticità di un sito. Peraltro, un'applicazione dai contorni metodologici molto simili a quelli qui proposti è stata diffusa, quasi in contemporanea, con riferimento al confronto – su scala internazionale – tra la competitività turistica di 95 stati rappresentativi dei diversi continenti[10].

Lo sviluppo del lavoro segue una traccia delineata dalla necessità di definire innanzitutto il concetto di attrattività turistica (paragrafo 2) e di identificare le variabili utili per alimentare una base dati integrata, ossia fondata su una gamma di informazioni di varia natura, utili per quantificare l'attrattività turistica locale (paragrafo 3).

Successivamente, si introdurrà la metodologia proposta per il calcolo di indici sintetici di competitività turistica (paragrafo 4), i cui risultati operativi saranno illustrati nei paragrafi 5 e, secondo un profilo d'analisi maggiormente interpretativo, nei paragrafi 6 e 7. Conclusioni prospettiche sono raccolte, infine, nel paragrafo 8.

[9] Di Gioia *et al.* (2005).

[10] Gooroochurn e Sugiyarto (2005; pagg. 23-43).

Riguardo alla scelta del livello territoriale a cui riferire l'analisi, si è optato per un approccio per "circoscrizione turistica". Tali circoscrizioni[11] (nel 2003 erano in tutto 547) rappresentano attualmente il livello di dettaglio minimo a cui poter riferire, su scala nazionale, analisi statistiche che necessitino della disponibilità del numero di presenze nelle strutture ricettive a livello locale. Sebbene, quindi, non sia ancora possibile condurre una simile ricerca a livello comunale, si è optato per un approccio innovativo (e, in qualche modo, propositivo per sviluppi futuri) anche per quanto riguarda la specificità locale dei siti analizzati. Peraltro, in alcuni casi tali circoscrizioni si identificano in singole aree comunali – quali Roma, Venezia, Firenze, Napoli, Milano per limitarsi ai casi più rilevanti – il che evidenzia ulteriormente l'utilità di tale approccio.

2. I molteplici profili della attrattività turistica

La necessità di identificare e quantificare la competitività turistica secondo un profilo di analisi multidimensionale deriva dalla complessità del fenomeno turistico, che si traduce in un sistema articolato di interazioni tra offerta e domanda[12].

Se la turisticità di una località rappresenta un concetto poliedrico e sfumato, nonché tuttora almeno parzialmente indeterminato da un punto di vista algoritmico, è fondamentale fin d'ora comprendere che un indice di turisticità, prima ancora di essere definito da un punto di vista matematico, dovrebbe:

- essere ispirato dal concetto di *attrattività complessiva*, ossia che tenga conto sia delle ricchezze potenziali che dell'effettiva risposta della domanda finale misurata in termini di pernottamenti, flussi escursionistici e spesa sostenuta dai visitatori.
- Contribuire a spiegare i motivi per cui una certa località presenta un livello più o meno elevato di competitività turistica: dovrebbe consentire, quindi, una valutazione sintetica di una serie di indicatori non limitati al solo impatto turistico finale, da determinare in base alle considerazioni di cui in seguito.
- Poter avere un'utilizzabilità concreta ai fini della promozione di politiche del turismo locali: ad esempio, consentire l'identificazione di aree potenzialmente turistiche ma attualmente poco premiate dalla domanda finale.
- Infine, essere preferibilmente *adimensionale*, ossia confrontabile indipendentemente dalla diversa dimensione territoriale dei siti analizzati; quindi, le variabili selezionate per l'analisi dovrebbero essere espresse preferibilmente in termini relativi (come percentuali o rapporti rispetto ad indicatori dimensionali quali la superficie o la popolazione residente).

Attualmente non esiste una definizione univoca di *sito turistico*, così come non si dispone di una lista di variabili da misurare che risulti universalmente accettata come quella effettivamente

[11] Nell'ambito della rilevazione mensile sul *Movimento nelle strutture ricettive*, l'Istat si avvale della collaborazione di uffici - operanti a livello provinciale o sub-provinciale in funzione della diversa organizzazione e legislazione turistica territoriale - che assumono denominazioni differenti nei diversi territori di competenza, di cui i principali sono: Aziende di Promozione Turistica (APT), Enti Provinciali per il Turismo (EPT), Aziende Autonome di Cura, Soggiorno e Turismo (AACST). Con il termine di "circoscrizione turistica" si raccoglie in un'unica definizione l'insieme di tali entità territoriali, ognuna delle quali include uno o più comuni. Da alcuni anni l'Istat diffonde i dati sugli arrivi e le presenze nelle strutture ricettive ufficiali anche a livello di singola circoscrizione e con un dettaglio annuale (dati disponibili via internet). I dati provinciali sono riottenibili per somma dei dati relativi alle varie circoscrizioni in cui è divisa la singola provincia.

[12] Costa e Manente (2000, pag. 207).

necessaria per valutare il livello di turisticità. Di conseguenza, manca anche una specifica metodologia in grado di condurre ad un sintetico *indice di turisticità (tourist index)*.

La selezione delle variabili deve essere in grado di esprimere nel modo più completo le potenzialità turistiche del territorio. Il principio che ha guidato la loro selezione è stato quello di scomporre *a priori* il concetto di *turisticità* nelle sue componenti più importanti, il che ha consentito, prima ancora di raccogliere ed elaborare le informazioni, di poter assegnare ad ogni variabile un ben preciso significato e di poterle raggruppare, quindi, in sottoinsiemi omogenei che rappresentino un particolare profilo di analisi dell'attrattività turistica locale.

In altri termini, è stato volutamente capovolto il processo logico, talvolta abusato nel campo dell'analisi statistica multivariata, secondo il quale è soprattutto dopo la fase di raccolta ed elaborazione dei dati che, tramite opportune elaborazioni fattoriali, è possibile reinterpretare *a posteriori* il significato semantico delle variabili originarie. Più nel dettaglio, i passi fondamentali per l'implementazione del modello sono stati i seguenti:

a) definizione del concetto di turisticità di un territorio e del modo in cui quest'ultima possa essere calcolata, stratificando preventivamente i vari aspetti della stessa su due livelli (componenti e sotto-componenti);
b) identificazione e raccolta delle variabili utili per delineare il profilo turistico di ogni località;
c) quantificazione dell'indice di turisticità complessiva e degli indici di livello inferiore più specifici.

Rimandando, per ulteriori dettagli, al paragrafo 3 circa il punto b) ed al paragrafo 4) circa il punto c), riguardo al punto a) il generico concetto di turisticità (tourist - T) di un data località può essere scomposto ad un primo livello in tre componenti:

1) l'attrattività turistica potenziale (*tourist attractiveness* - TA), che rappresenta la dotazione del territorio di tipo strutturale, ambientale, storico-artistico, ecc.;
2) la disponibilità di posti letto per fini turistici (*tourist bed places* - TB), che rappresenta una specifica dotazione del territorio, trattata in modo separato rispetto alle altre variabili strutturali, in quanto più direttamente connessa con il turismo;
3) l'impatto turistico effettivo derivato dalla domanda turistica (*tourist impact* - TI), ovvero dalle presenze registrate e dai visitatori escursionisti transitati, dalla spesa turistica, ecc..

Inoltre, ad un secondo livello di disaggregazione concettuale:

- TA è a sua volta scomponibile in cinque sotto-componenti: territorio ed ambiente (TA1); infrastrutture (TA2); attrattive storiche e naturali (TA3); altre attrattive (TA4); notorietà della località (TA5);
- TB non sarà suddiviso;
- TI, infine, è scomponibile in tre sotto-componenti: profilo economico turistico (TI1); domanda turistica finale (TI2); investimenti turistici (TI3).

In sostanza, il concetto di turisticità può essere suddiviso in 3 componenti, ad un primo livello di disaggregazione, e in 8 sotto-componenti al secondo livello di disaggregazione. Pertanto, gli indicatori cercati saranno complessivamente 12: uno generale (*tourist index* – TI), 3 ad un primo livello di specificità (*tourist attractiveness index* - TAI, *tourist bed-places index* – TBI, *tourist impact index* - TII) e 8 di secondo livello[13] (TAI1, TAI2, TAI3, TAI4, TAI5, TII1, TII2, TII3), mentre TBI è un indicatore tanto di primo quanto di secondo livello.

[13] Va detto che gli indicatori TAI5 e TII3 non sono risultati valutabili, a causa della non disponibilità di dati attendibili sulle variabili da cui dipendono, per cui gli indicatori di secondo livello effettivamente calcolati sono 6.

3. La scelta e la misurazione delle variabili

Sono state identificate in tutto 34 variabili fondamentali e misurabili a livello di circoscrizione turistica, nidificate in 6 sotto-componenti (TAI1, TAI2, TAI3, TAI4, TII1, TII2) e nella componente TBI, così come risulta da quanto riportato nella successiva tabella 3.2.

Tra le variabili utilizzate non sono incluse variabili concettualmente rilevanti, ma attualmente non misurabili (né sensatamente stimabili) su scala nazionale e con il livello di dettaglio richiesto, ossia: il numero di visitatori nelle strutture museali e monumentali, la spesa turistica, il numero di escursionisti e la spesa escursionistica ad essi associata, il peso del valore aggiunto generato dal turismo sul valore aggiunto circoscrizionale, il numero di bandiere blu delle località costiere (che però sarebbero attribuibili, ovviamente, alle sole circoscrizioni dotate di comuni marini), la distanza media dei comuni di ogni circoscrizione dalle principali vie d'accesso (porti e/o aeroporti).

In generale, la difficoltà operativa di maggior rilievo è stata la necessità di raccogliere gran parte delle variabili al livello di singolo comune, per poi procedere alla riaggregazione dei dati comunali in dati circoscrizionali. Tale passaggio è stato effettuato utilizzando somme (ad esempio, il numero complessivo di musei), medie ponderate (le variabili ambientali da TAI1_2 a TAI1_5) o rapporti di composizione (le quote occupazionali da TII1_3 a TII1_6) in funzione della diversa tipologia di variabile trattata.

A causa del diverso livello di aggiornamento delle variabili considerate, si è optato per la raccolta delle informazioni più aggiornate per ognuna di esse: di conseguenza, la base dati non è riferibile ad un unico anno, bensì ad un periodo compreso tra il 2001 ed il 2004.

Nel complesso, l'Istat è stato il bacino informativo più utile e qualitativamente affidabile, laddove il ricorso alla rete internet possa, come forse inevitabile, nascondere insidie relative alla completezza ed alla affidabilità di alcune informazioni raccolte.

È per questa ragione che, laddove possibile, i dati di fonte internet sono stati confrontati con quelli – più aggregati – derivati da altre fonti (ad esempio, la lunghezza delle strade). Peraltro, in alcuni casi (come per diverse variabili ambientali) la non disponibilità equivale alla effettiva non misurabilità della variabile a tale livello di dettaglio.

Delle 34 variabili individuate come concettualmente rilevanti, 10 non sono risultate misurabili a livello di circoscrizione e sono state quindi stimate, secondo quanto indicato nell'ultima colonna della tabella. Tuttavia, la fase di stima si è articolata in modo diverso in funzione della tipologia delle stesse. In particolare:

- con riferimento alle variabili ambientali da TAI1_2 a TAI1_5 i dati, originariamente disponibili a livello provinciale, sono stati distribuiti per circoscrizione in proporzione alla popolazione residente.
- Riguardo alle variabili infrastrutturali, alcune delle quali disponibili solo a livello provinciale, TAI2_2 (per la parte relativa alle strade non comunali) e TAI2_3 sono state distribuite in proporzione alla superficie della circoscrizione, mentre TAI2_4 in proporzione ancora alla popolazione residente.
- Riguardo alla stima del prezzo medio di un albergo (TBI_4), tale variabile, come noto, non è disponibile a nessun livello di aggregazione. In questo caso, la stima si è basata sulla raccolta diretta via internet di circa 13.200 prezzi (media tra prezzo della stanza singola e della stanza doppia con sola prima colazione) distinti per numero di stelle e selezionati in modo da essere rappresentativi per ciascuna delle circoscrizioni considerate. Successivamente, per ogni circoscrizione à stata calcolata la media ponderata dei prezzi medi rilevati per ogni classe di stelle dell'albergo, con pesi proporzionali al numero di posti letto alberghieri disponibili per ogni classe di stelle nella circoscrizione. I risultati ottenuti sono coerenti con

una stima proposta in una recente edizione del *Rapporto sul Turismo Italiano*, basata su un approccio indiretto (confronto "macro")[14] e risultano confermati da un confronto – condotto su un campione di circa mille quotazioni medie scelte casualmente – con quanto derivato dall'annuario degli alberghi italiani edito annualmente dall'ENIT (confronto "micro")[15].

- Riguardo alla stima del numero dei posti letto negli alloggi privati (TBI_3), si è fatto ricorso ai dati sul numero di seconde case destinate ad uso vacanze derivate dal censimento della popolazione e delle abitazioni 1991, aggiornato al 2001 tramite i dati del censimento riferito a tale anno ed il numero di concessioni edilizie a fini residenziali disponibili sempre da fonte Istat. Per ulteriori dettagli, così come per quanto concerne la stima delle presenze negli alloggi privati per motivi di vacanza (TII2_3), si rimanda ad altri lavori specialistici[16].

La variabile TAI2_6 (addetti nel commercio al dettaglio) è stata introdotta come indicatore della densità di punti di vendita commerciali, ed è per questo motivo che è stata inserita tra gli indicatori infrastrutturali.

Riguardo alla variabile TAI1_5, si ricorda che gli articoli 155 e 156 del nuovo codice della strada si riferiscono al rispetto delle norme che limitano l'inquinamento acustico ed atmosferico derivante da mezzi a motore.

Riguardo, invece, alla variabile TAI3_3, si è preferito limitare l'informazione alla quota percentuale dei comuni della circoscrizione con un (o dichiarati) sito UNESCO (dunque, patrimonio dell'umanità), senza ponderare tale dato con la superficie, la popolazione od altre variabili economiche.

Con riferimento agli accessi ferroviari, non essendo risultata disponibile su scala comunale l'informazione circa la presenza od assenza di stazione ferroviaria, si è optato per l'utilizzo della variabile TAI2_5 (numero di addetti nel settore delle ferrovie per 1.000 residenti).

Riguardo, poi, agli attrattori storici e culturali, si è optato per considerare, nell'ambito della smisurata gamma di siti ecclesiali e monumentali esistenti in Italia, quelli ad elevato interesse, sfruttando le informazioni fornite da diverse pubblicazioni specialistiche del Touring Club Italiano[17]. In particolare, nel conteggio del numero di musei (TAI3_2) non sono stati inclusi i circuiti e le aree monumentali[18].

Inoltre, va precisato che le attività economiche relative agli indicatori da TII1_3 a TII1_6 sono le seguenti:

- TII1_3: servizi di accoglienza, ossia alberghi e strutture ricettive extra-alberghiere

[14] Gismondi (2001, pagg. 101-142).

[15] ENIT (2004).

[16] Gismondi (2000, pagg. 87-104); Gismondi e Mirto (2002, pagg. 33-66); Mercury (2005). Va ricordato che l'indagine mensile sul movimento nelle strutture ricettive condotta correntemente dall'*Istat* raccoglie i dati sulle presenze con riferimento ad una sola tipologia di alloggio privato, ossia agli *alloggi privati gestiti in forma imprenditoriale*. I flussi turistici consumati in tale tipologia di alloggio rappresentano solo una quota minimale (inferiore al 10%) del totale delle presenze trascorse nel complesso degli alloggi privati, la cui quota preponderante è determinata dall'utilizzo diretto degli alloggi di proprietà, non oggetto di misurazione. Altre fonti ufficiali che diffondono stime mensili o trimestrali sui pernottamenti negli alloggi privati - ma con un dettaglio territoriale limitato - sono date dall'indagine Istat su "Viaggi e vacanze degli Italiani" e dall'indagine alle frontiere sul turismo internazionale condotta dall'Ufficio Italiano dei Cambi.

[17] Sono state consultate le seguenti pubblicazioni del Touring Club Italiano: *Guida turistica d'Italia* (1996*a*); *Italia da scoprire – viaggio nei centri minori* (1996*b*); *Artigianato, sapori e tradizioni d'Italia* (2002); inoltre, un valido ausilio è stato fornito dalle "guide rosse" regionali.

[18] Si ricorda che, attualmente, il Ministero dei Beni Culturali diffonde i dati sul numero ed i visitatori dei soli musei statali (www.beniculturali.it)

(ATECO 55.1 e 55.2[19]);

- TII1_4: servizi di ristoro, ossia ristoranti, bar e similari (55.3 e 55.4);
- TII1_5: servizi di assistenza al turista, ossia le attività delle agenzie di viaggio e turismo e le attività delle guide e degli accompagnatori turistici (63.3).
- TII1_6: servizi d'intrattenimento, ossia discoteche, sale da ballo, *night club* e simili, parchi divertimenti, attività culturali, altre attività ricreative e d'intrattenimento, stabilimenti balneari (92.3, 92.5, 92.72.1).

Tutte le variabili sono quantitative, ad eccezione delle due variabili dicotomiche associate alla presenza o all'assenza di porto ed aeroporto: le informazioni fornite da tali indicatori sono state sintetizzate in un'unica variabile, pari ad uno nel caso sia presente almeno una delle due possibilità di accesso e a zero altrimenti.

Infine, un ultimo aspetto non banale è rappresentato dalla possibilità di valutare l'insieme delle variabili originarie in termini assoluti o relativi A fini comparativi l'opzione preferibile è sembrata essere la seconda, soprattutto per unità territoriali – quali le circoscrizioni – per le quali la dimensione e la forma del territorio analizzato non sono sempre logicamente relazionabili alla fruibilità turistica dello stesso.

Le graduatorie ottenute sulla base di una tecnica di analisi fattoriale si riferiscono, quindi, alla *attrattività turistica relativa* di ogni circoscrizione, ovvero all'intensità con cui le diverse componenti della turisticità si manifestano nell'unità territoriale in relazione alla sua dimensione territoriale o demografica.

Nella tabella 3.1 è stata riportata la distribuzione regionale delle circoscrizioni turistiche, con le relative incidenze in termini di posti-letto e di presenze nel 2003.

La sintesi regionale, pur evidenziando la sostanziale omogeneità della distribuzione delle circoscrizioni nelle quattro aree geografiche ed il peso considerevole del Nord/est sulle presenze complessive (38,3%), nasconde inevitabilmente le eterogeneità locali e si limita al monitoraggio della capacità ricettiva e della domanda finale limitatamente alle strutture ricettive ufficiali.

[19] Secondo tale classificazione le tre cifre identificano il gruppo di attività economica, le quattro cifre la classe e le cinque cifre la categoria, secondo un livello di dettaglio crescente.

Tabella 3.1: Distribuzione regionale delle circoscrizioni turistiche nel 2003

Regione/area	Ammontari assoluti			Composizioni percentuali		
	Numero	Letti	Presenze	Numero	Letti	Presenze
Piemonte	44	146.685	8.943.998	8,0	3,6	2,6
Valle d'Aosta	14	52.654	3.496.219	2,6	1,3	1,0
Lombardia	66	256.579	25.972.014	12,1	6,3	7,5
Trentino-Alto Adige	54	367.550	39.570.587	9,9	9,0	11,5
Veneto	27	654.041	55.111.931	4,9	16,0	16,0
Friuli-Venezia Giulia	15	150.462	8.863.178	2,7	3,7	2,6
Liguria	15	144.407	14.769.598	2,7	3,5	4,3
Emilia-Romagna	39	398.238	36.621.302	7,1	9,7	10,6
Toscana	65	417.148	36.837.331	11,9	10,2	10,7
Umbria	12	66.369	5.795.242	2,2	1,6	1,7
Marche	27	213.774	13.449.366	4,9	5,2	3,9
Lazio	20	244.054	24.054.701	3,7	6,0	7,0
Abruzzo	26	96.243	7.115.155	4,8	2,3	2,1
Molise	5	11.878	769.334	0,9	0,3	0,2
Campania	29	168.197	19.708.952	5,3	4,1	5,7
Puglia	22	188.111	10.702.634	4,0	4,6	3,1
Basilicata	5	32.595	1.761.639	0,9	0,8	0,5
Calabria	16	193.245	7.333.813	2,9	4,7	2,1
Sicilia	32	139.313	13.152.348	5,9	3,4	3,8
Sardegna	14	158.042	10.383.975	2,6	3,9	3,0
Nord/ovest	139	600.325	53.181.829	25,4	14,6	15,4
Nord/est	135	1.570.291	140.166.998	24,7	38,3	40,7
Centro	124	941.345	80.136.640	22,7	23,0	23,3
Sud/isole	149	987.624	70.927.850	27,2	24,1	20,6
Italia	**547**	**4.099.585**	**344.413.317**	**100,0**	**100,0**	**100,0**

Fonte: Istat, Rilevazioni sulla “capacità delle strutture ricettive” ed il “movimento nelle strutture ricettive”.

Tabella 3.2 - Lista delle variabili che hanno alimentato la base dati per circoscrizione

SIGLA	VARIABILE	DESCRIZIONE	ANNO	DETTAGLIO	FONTE	INDIRIZZO/RILEVAZIONE	Nota
TAI1	**Territorio e ambiente**						
TAI1_1	Superficie in Km^2 (per 1.000 residenti)	Chilometri quadrati di superficie per 1.000 Residenti	2004	Comune	ANCITEL	www.ancitel.it	
TAI1_2	Sostanze inquinanti (per abitante)	Numero di giornate di superamento dei livelli di attenzione per le principali sostanze inquinanti monitorate	2001	Provincia	ISTAT	Osservatorio ambientale sulle città (www.ISTAT.it)	Stima
TAI1_3	Impianti depurazione (% su residenti)	Percentuale di popolazione residente servita da impianti di depurazione delle acque reflue urbane	2001	Provincia	ISTAT	Osservatorio ambientale sulle città	Stima
TAI1_4	Raccolta differenziata (% su residenti)	Percentuale di rifiuti da raccolta differenziata rispetto al totale dei rifiuti urbani raccolti	2001	Provincia	ISTAT	Osservatorio ambientale sulle città	Stima
TAI1_5	Violazione codice strada (per 100.000 veicoli)	Multe per violazione codice strada 155 e 156 (per 100.000 veicoli)	2001	Provincia	ISTAT	Osservatorio ambientale sulle città	Stima
TAI1_6	Comuni in parchi (% su totale)	Quota % dei comuni della circoscrizione inseriti in un parco o riserva naturale (statale o regionale)	2004	Comune	Web	www.parks.it www.naturalia.org	

Tabella 3.2 segue **- Lista delle variabili che hanno alimentato la base dati per circoscrizione**

SIGLA	VARIABILE	DESCRIZIONE	ANNO	DETTAGLIO	FONTE	INDIRIZZO/RILEVAZIONE	Nota
TAI2	**Infrastrutture**						
TAI2_1	Porto-aeroporto	Presenza di porto e/o aeroporto (presenza o assenza)	2003	Comune	ISTAT	Statistiche sui trasporti marittimi. Statistiche sul trasporto aereo. Conto nazionale dei trasporti	
TAI2_2	Km di strade (per 1.000 residenti)	Chilometri di strade (comunali, provinciali, statali, autostrade)	2003	Comune Provincia	ANCITEL Ministero infrastrutture e trasporti	www.ancitel.it Conto nazionale dei trasporti	
TAI2_3	Km di linee autobus… (per 100 Km^2)	Chilometri di linee autobus, tram, filobus, metropolitana, funicolari, piste ciclabili per 100 chilometri quadrati	2001	Provincia	ISTAT	Osservatorio ambientale sulle città	Stima
TAI2_4	Parcheggi (per 100 residenti)	Stalli di sosta in parcheggi (per 100 abitanti)	2001	Provincia	ISTAT	Osservatorio ambientale sulle città	Stima
TAI2_5	Addetti ferrovie (per 1.000 residenti)	Numero di addetti nel settore ferroviario per 1.000 residenti	2001	Comune	ISTAT	Censimento delle imprese	
TAI2_6	Addetti commercio al dettaglio (%)	Quota percentuale degli addetti nel comparto commerciale al dettaglio sul totale degli addetti nella circoscrizione	2001	Comune	ISTAT	Censimento delle imprese	
TAI2_7	Numero ospedali (per 1.000 residenti)	Numero di ospedali per 1.000 residenti	2001	Comune	ISTAT	Censimento delle imprese	

Tabella 3.2 segue - **Lista delle variabili che hanno alimentato la base dati per circoscrizione**

SIGLA	VARIABILE	DESCRIZIONE	ANNO	DETTAGLIO	FONTE	INDIRIZZO/RILEVAZIONE	Nota
TAI3	**Attrattori storici e naturali**						
TAI3_1	Chiese-palazzi-monumenti (per 1.000 residenti)	Numero di chiese storiche, palazzi storici, monumenti per 1.000 residenti	2002	Comune	Web Touring Club Italiano	www.archeologia.puntopartenza.it www.comune.it www.saperviaggiare.it www.metropolis.it	
TAI3_2	Musei (per 1.000 residenti)	Numero di musei statali e non statali per 1.000 residenti	2002	Comune	Web	www.progettoecotur.it/Musei.htm www.comuni-italiani.it	
TAI3_3	Siti UNESCO	Percentuale di comuni della circoscrizione che presentano (o sono) un sito UNESCO	2004	Comune	Web	www.linklavoro.it/index_tempo_libero.asp	
TAI4	**Altri attrattori**						
TAI4_1	Terme (per 1.000 residenti)	Numero di comuni con centri termali per 1.000 residenti	2003	Comune	Web Touring Club Italiano	www.terme.info	
TAI4_2	Fiere (per 1.000 residenti)	Numero di comuni che ospitano fiere per 1.000 residenti	2003	Comune	Web Touring Club Italiano	www.portaleitalia.net www.paesionline.it	
TAI4_3	Mercati-feste (per 1.000 residenti)	Numero di comuni che ospitano mercati o feste per 1.000 residenti	2003	Comune	Web Touring Club Italiano	www.mercatiniditalia.it www.borghitalia.it www.ospitalitalia.it/arte_ricerca.php	
TAI4_4	Porto turistico (per 1.000 residenti)	Numero di posti barca in porti turistici per 1.000 residenti	2002	Comune			

Tabella 3.2 segue **- Lista delle variabili che hanno alimentato la base dati per circoscrizione**

SIGLA	VARIABILE	DESCRIZIONE	ANNO	DETTAGLIO	FONTE	INDIRIZZO/RILEVAZIONE	Nota
TBI	**Infrastrutture turistiche**						
TBI_1	Numero letti negli alberghi (per 1.000 residenti)	Numero di posti-letto negli alberghi per 1.000 residenti	2003	Comune	ISTAT	Capacità delle strutture ricettive	
TBI_2	Numero letti nelle altre strutture (per 1.000 residenti)	Numero di posti-letto nelle strutture ricettive extra-alberghiere per 1.000 residenti	2003	Comune	ISTAT	Capacità delle strutture ricettive	
TBI_3	Numero letti in alloggi privati (per 1.000 residenti)	Numero di posti-letto negli alloggi privati per 1.000 residenti	2001	Comune	ISTAT	Censimento popolazione e abitazioni 1991 e 2001; statistiche sull'attività edilizia	Stima
TBI_4	Prezzo medio albergo	Prezzo medio di un pernottamento in albergo	2002	Circoscrizione	Rapporto turismo italiano WEB	www.venere.com www.travelgo.it Rapporto sul turismo italiano 2003	Stima
TII1	**Profilo economico turistico**						
TII1_1	Valore aggiunto (000 Euro per residente)	Valore aggiunto a prezzi base (al netto SIFIM) per residente in migliaia di euro		Provincia	ISTAT	Conti economici territoriali	Stima
TII1_2	Numero addetti (per 100 residenti)	Numero di addetti nella circoscrizione per 100 residenti	2001	Comune	ISTAT	Censimento delle imprese	
TII1_3	Addetti 55.1 e 55.2 (% su addetti)	Quota % di addetti negli alberghi e nelle strutture ricettive extra-alberghiere sul totale degli addetti	2001	Comune	ISTAT	Censimento delle imprese	

Tabella 3.2 segue **- Lista delle variabili che hanno alimentato la base dati per circoscrizione**

SIGLA	VARIABILE	DESCRIZIONE	ANNO	DETTAGLIO	FONTE	INDIRIZZO/RILEVAZIONE	Nota
TII1_4	Addetti 55.3 e 55.4 (% su addetti)	Quota % di addetti nei bar e nei ristoranti sul totale degli addetti	2001	Comune	ISTAT	Censimento delle imprese	
TII1_5	Addetti 63.3 (% su addetti)	Quota % di addetti nelle attività di agenzia di viaggi e tour operator sul totale degli addetti	2001	Comune	ISTAT	Censimento delle imprese	
TII1_6	Addetti 92.3, 92.5, 92.72.1 (% su addetti)	Quota % di addetti nelle attività ricettive e di intrattenimento sul totale degli addetti	2001	Comune	ISTAT	Censimento delle imprese	
TII2	**Domanda turistica**						
TII2_1	Presenze alberghi (per 1.000 residenti)	Numero di presenze nelle strutture alberghiere per 1.000 residenti	2003	Circoscrizione	ISTAT	Movimento clienti nelle strutture ricettive	
TII2_2	Presenze complementari (per 1.000 residenti)	Numero di presenze nelle strutture ricettive extra-alberghiere per 1.000 residenti	2003	Circoscrizione	ISTAT	Movimento clienti nelle strutture ricettive	
TII2_3	Presenze in alloggi privati (per 1.000 residenti)	Numero di presenze negli alloggi privati per 1.000 residenti	2001	Circoscrizione	ISTAT – Ufficio Italiano Cambi - ENEL	Censimento abitazioni 1991 e 2001; Indagine viaggi e vacanze degli Italiani; Indagine dell'Ufficio Italiano Cambi alle frontiere (www.UIC.it); Consumi di energia elettrica per destinazione d'uso	Stima
TII_4	Indice di stagionalità	Quota percentuale delle presenze consumate tra giugno ed agosto sul totale	2003	Circoscrizione	ISTAT	Movimento clienti nelle strutture ricettive	

4. Definizione analitica e calcolo degli indici di turisticità

Riguardo alla definizione ed al calcolo degli indici di turisticità, e sulla base di quanto introdotto nei paragrafi precedenti, va preliminarmente sottolineata la loro diversa valenza logico-informativa. Infatti, l'indice TAI è un indicatore di *turisticità potenziale*, mentre l'indice TII misura la *turisticità effetti*va indotta da un sito.

Inoltre, l'indice TBI rappresenta un particolare indicatore strutturale che, in un certo senso, fa da *ponte tra la turisticità potenziale e quella effettiva* e che va, quindi, considerato separatamente rispetto alle altre variabili di attrattività, proprio perchè essa non può che riflettere la stessa domanda turistica finale emersa negli ultimi anni precedenti a quello di analisi.

Sebbene, naturalmente, un semplice indice turistico *ex-post* potrebbe basarsi sul solo indicatore TII2 – che in pratica misura l'effettiva efficacia a consuntivo di tutte le risorse attrattive disponibili *ex-ante* nella circoscrizione – il bisogno di considerare separatamente le varie componenti che determinano TI deriva dal fatto che il solo indice TII2 non fornisce informazioni circa le ragioni di un'alta o bassa attrattività turistica, nè tantomeno un profilo qualitativo del sito analizzato.

Ciò premesso, la tecnica di calcolo degli indici di turisticità proposta è la seguente[20]. Si può definire x_{vi} come il valore x che la variabile attiva v (v=1,2,…,V) assume nella circoscrizione i (i=1,2,…,547). La metodologia utilizzata per calcolare per ogni comune l'indice generale di turisticità TI si basa sui passi seguenti.

1. Tutte le variabili x sono preliminarmente standardizzate[21] in modo da essere rese omogenee e confrontabili in termini di valore medio e variabilità. In questo modo saranno disponibili delle nuove variabili standardizzate indicate con z_{vi}.
2. Poiché un indice di turisticità, per definizione, deve crescere al crescere dell'attrattività turistica, è preferibile esprimere tutte le variabili in modo che assumano valori crescenti al crescere della componente di attrattività che esse esprimono; quindi, le variabili TAI1_2 (sostanze inquinanti), TAI1_5 (violazione codice della strada), TBI_4 (prezzo albergo) e TII_4 (indice di stagionalità) sono state preventivamente cambiate di segno. Si suppone in tal modo che il prezzo medio di un pernottamento in albergo e la stagionalità delle presenze siano correlate inversamente con la competitività di una destinazione, ipotesi in realtà non sempre veritiera.
3. Se l'insieme delle 34 variabili attive è stato partizionato in 7 sotto-componenti misurabili, ciascuna basata su V variabili (ad esempio, la sotto-componente TA1 si basa su V=6 variabili, in accordo alla tabella 3.1), un punteggio s (*score*) per la circoscrizione i può essere calcolato in due modi: a) media aritmetica ponderata dei contributi relativi che le V variabili forniscono con riferimento alle prime "componenti principali" (*metodo CP*) estratte dalle 34 variabili attive originarie, dove i pesi sono dati dalle corrispondenti quote di varianza "spiegata" da tali componenti; b) media aritmetica semplice (*metodo MS*) delle V variabili standardizzate. Nel caso di h componenti, se a_{hv} indica la *v-ma* coordinata (v=1,2,…,34) dell'*h-mo* asse fattoriale,

[20] Per un esempio relativo al possibile utilizzo di tecniche diverse da quella qui proposta, si veda Cracolici (2004, pagg. 703-706).

[21] La standardizzazione comporta, come noto, la trasformazione di una variabile x nella nuova variabile $(x-\mu)/\sigma$, dove μ indica la media e σ la deviazione standard di x.

mentre λ_h è la varianza della *h-ma* componente principale, i punteggi ottenuti utilizzando i due metodi alternativi saranno forniti, rispettivamente, dalle formule:

$$\text{a) } s_{CPi} = \left[\sum_{h=1}^{k} \left(\sum_{v=1}^{V} z_{vi}\, a_{hv} \right) \lambda_h \right] \Big/ \left(\sum_{h=1}^{k} \lambda_h \right); \quad \text{b) } s_{MSi} = \sum_{v=1}^{V} z_{vi} / V \qquad (4.1)$$

Nella formula (4.1a), se si pone V=34 la sommatoria in parentesi tonde si riduce semplicemente al valore che la circoscrizione *i-ma* assume in corrispondenza della *h-ma* componente principale. Quindi, per V=34 il punteggio s_{CP} si basa sulla media aritmetica ponderata di ciascuno di tali k valori, con pesi dati dalle varianze di ognuna delle k componenti principali considerate. Da questa considerazione risulta immediatamente valutabile il contributo *additivo* che ogni variabile fornisce alla determinazione dell'indice di turisticità finale. Se V identifica l'insieme delle variabili che determinano uno dei sette indici di attrattività definiti nel paragrafo 2, la media ponderata sarà estesa alle sole variabili inerenti a tale indice e, di conseguenza, l'indice di attrattività globale TI sarà dato dalla somma dei sette indici parziali di attrattività.

4. Al fine di poter trattare con valori variabili tra zero ed uno - il che rende più semplice l'interpretazione dei risultati e consente eventuali confronti nello spazio e nel tempo – con riferimento ad ognuno dei sette indici di attrattività parziale si possono calcolare i punteggi finali S relativi alla circoscrizione *i-ma* che, con i due metodi proposti, saranno dati rispettivamente da[22]:

$$S_{CPi} = \left(s_{CPi} - s_{CP,MIN}\right) / \left(s_{CP,MAX} - s_{CP,MIN}\right); \; S_{SMi} = \left(s_{MSi} - s_{MS,MIN}\right) / \left(s_{MS,MAX} - s_{MS,MIN}\right) \quad (4.2)$$

5. Per ogni circoscrizione, TAI sarà dato dalla media aritmetica semplice dei punteggi finali S relativi agli indici TAI1, TAI2, TAI3, TAI4; TBI sarà identificato direttamente dal suo punteggio S, non essendo scomposto in sotto-componenti; TII sarà dato dalla media aritmetica semplice dei punteggi finali S relativi agli indici TII1 e TII2.

6. Infine, TI deriverà dalla media aritmetica semplice degli indici TAI, TBI e TII calcolati come al punto precedente

Nel prosieguo verranno illustrati i risultati ottenuti utilizzando esclusivamente la metodologia di aggregazione delle variabili originarie basata sul criterio delle componenti principali sintetizzato dalla formula (4.1a). Il ricorso al metodo *CP,* come noto, presenta il notevole vantaggio di basarsi su punteggi finali determinati dalla combinazione di variabili incorrelate (ossia le stesse componenti principali). In effetti, il metodo *MS* può risentire della presenza, quasi inevitabile in pratica, di variabili fortemente interdipendenti, con la conseguente sopravvalutazione di alcuni aspetti alla turisticità inevitabilmente molto legati tra loro. Ad esempio, è comprensibile come tra il numero di presenze negli alberghi e negli esercizi complementari per 1.000 residenti (variabili TII2_1 e TII2_2) ed il numero di posti-letto negli alberghi e negli esercizi complementari per 1.000 residenti (variabili

[22] La trasformazione (4.2) consente anche di uniformare il campo di variazione effettivo delle diverse variabili considerate, riducendo il rischio di sopravvalutare le variabili caratterizzate da una variabilità intrinseca più elevata.

(TBI_1 eTBI_2) si verifichino correlazioni lineari elevate (rispettivamente, 0,76 e 0,92).

Tuttavia, è importante evidenziare come, dal confronto tra le formule (4.1a) e (4.1b), sia evidente il parallelismo tra le due metodologie di sintesi, entrambe basate su medie aritmetiche: nel secondo caso si tratta della media semplice delle variabili originarie standardizzate; nel primo, della media ponderata relativa alle proiezioni delle variabili originarie standardizzate sulle prime nuove *k* componenti principali.

Per quanto riguarda il criterio di calcolo dell'indice sintetico appena descritto, altre tecniche sono state proposte in lavori recenti, peraltro avulsi dal contesto turistico[23], e per analisi comparative in merito si rimanda ai lavori già citati di Gismondi e Russo. In particolare, è stato dimostrato come, almeno con riferimento al contesto in analisi, il calcolo degli indici sintetici TAI, TII e TI tramite medie ponderate basate su ulteriori applicazioni del metodo *CP* agli indici di ordine inferiore (quindi, ad esempio, con riferimento a TAI, agli indici da TAI1 a TAI4) condurrebbe a risultati sostanzialmente analoghi

Tuttavia, è utile chiedersi se e sotto quali condizioni entrambi i suddetti metodi (4.1a) e (4.1b) possano condurre a graduatorie finali simili, nel qual caso evidentemente il ricorso al metodo *MS*, per la sua immediatezza ed intuitività, sarebbe del tutto giustificato. I due metodi tendono a coincidere[24]:

- nel caso banale in cui tutte le variabili di input fossero uguali tra loro o pari ad una costante.
- Se tra le V variabili originarie solo alcune determinano fortemente le disuguaglianze tra i singoli comuni e risultano, quindi, le sole dotate di fortissima variabilità.
- Qualora il punteggio s fosse determinato da poche variabili che presentano valori di z molto più elevati rispetto alle variabili rimanenti.
- Se le variabili originarie sono tutte caratterizzate da basse varianze, ossia risultano molto stabili al variare della circoscrizione turistica considerata.
- Quanto più le variabili originarie sono reciprocamente incorrelate.

Formalmente, si può dimostrare che, in generale, partendo da *V* variabili *x* di tipo quantitativo, la correlazione lineare tra i punteggi ottenuti applicando il metodo delle componenti principali e quelli derivati dall'uso della media aritmetica semplice è data dalla relazione:

$$r(s_{CP}, s_{MS}) = \sum_{h=1}^{k}\sum_{v=1}^{V}\lambda_h^2 a_{hv} \Bigg/ \sqrt{\sum_{h=1}^{k}\lambda_h^3}\sqrt{\sum_{i=1}^{n}\sum_{v=1}^{V}\sum_{w=1}^{V} x_{iv}x_{iw}}\,, \tag{4.3}$$

che tende a crescere quanto più le variabili originarie sono caratterizzate da basse varianze e tanto più la correlazione tra tutte le variabili originarie (che si suppone siano state standardizzate) tenda a zero, nel qual caso si avrà:

$$r(s_{CP}, s_{MS}) \to \sum_{h=1}^{k}\sum_{v=1}^{V} a_{hv} \Big/ \sqrt{2nV}\,. \tag{4.4}$$

[23] Cfr., ad esempio, Giudici e Avrini (2002, pagg. 61-80); Aiello e Attanasio (2004, pagg. 327-338).

[24] Per tali valutazioni e gli sviluppi (4.3) e (4.4) cfr. Gismondi e Russo (2005, *cit*.).

Per quanto riguarda, infine, la scelta del numero *k* di componenti principali da considerare nella formula (4.1a), è utile ricordare che, in base alla metodologia *CP*, la somma delle varianze delle 34 componenti estraibili dalla base dati originaria riproduce esattamente la somma delle varianze di tale base dati, per cui un buon principio consiste nel selezionare le prime *k* componenti tali da rappresentare almeno il 50% della varianza originaria complessiva. Una numerosità *k* elevata indica, da un lato, la coesistenza di molteplici profili semantici sottostanti ai dati osservati; dall'altro, può comportare problemi di interpretazione di tali profili e non consentire una chiave di lettura immediata dei risultati ottenuti.

5. I risultati dell'applicazione

La tabella 5.1 riassume i principali parametri utili per interpretare i risultati dell'analisi in componenti principali[25]. In particolare, con riferimento ad ognuna delle 34 variabili considerate per l'analisi, la tabella fornisce queste informazioni:

- correlazione lineare – o *contributo relativo* - tra l'*h-ma* componente principale e la *v-ma* variabile originaria, per *v*=1,2,…,34, data da $c_{hv} = a_{hv}\sqrt{\lambda_h}$;
- qualità di rappresentazione della variabile *v* rispetto all'insieme delle *k* componenti principali considerate – o *contributo assoluto* -, data dalla somma dei quadrati delle precedenti correlazioni, ossia da $q_v = \sum_{h=1}^{k} c_{hv}^2$. L'indice q_v è compreso tra 0 e 1 ed indica la "quota" della variabile *v* spiegata dall'insieme delle *k* componenti considerate.
- Stima del peso della variabile originaria *v* nella determinazione del punteggio finale, data da: $p_v \approx \left(\sum_{h=1}^{k} \left| a_{hv}\lambda_h \right| \right) \Big/ \left(\sum_{v=1}^{V}\sum_{h=1}^{k} \left| a_{hv}\lambda_h \right| \right)$. Tale quantità esprime una stima del peso sintetico con cui ogni variabile finisce con il contribuire nella media ponderata delle 34 variabili originarie. La somma dei pesi al variare della variabile considerata è pari ad uno[26]. Normalmente tali pesi sono molto simili a quelli ottenibili sulla base dei contributi assoluti di cui al punto precedente, qualora questi fossero tutti rapportati alla loro somma.

Per poter arrivare a spiegare almeno il 50% della variabilità originaria della base dati è stato necessario considerare le prime 7 componenti principali, che spiegano in tutto il 51,5% della varianza. Le prime due componenti risultano molto più rilevanti della altre, perché da sole determinano quasi il 25% della variabilità originaria. Si tratta di un numero di variabili elevato, che indica la coesistenza, nella base dati, di una molteplicità di profili e l'assenza di una variabile (o di un nucleo di poche variabili) in grado di determinare una parte rilevante della variabilità complessiva. Nella formula (4.1a) si è dunque posto *k*=7.

L'identificazione del contenuto semantico delle componenti principali considerate si

[25] Per ulteriori dettagli metodologici, si veda, ad esempio, Fabbris (1997, pagg. 163-213).

[26] Si tratta di una *stima* del peso, meno precisa del coefficiente di correlazione tra ogni variabile e l'indice di turisticità globale (cfr. grafico 5.1), in quanto per garantire che la somma dei pesi sia uguale ad uno è stato necessario valutare gli addendi delle sommatorie in valore assoluto.

basa sull'analisi dei coefficienti di correlazione lineare tra tali componenti e le variabili originarie. Valori elevati e positivi indicano un legame diretto tra tali variabili e la turisticità complessiva, mentre vale il viceversa in presenza di valori molto negativi[27].

Prima componente principale: spiega il 15,3% della varianza e si può identificare con la attrattività turistica effettiva, ossia la dotazione di infrastrutture turistiche e la domanda finale attivata. Infatti, le correlazioni più elevate con questa componente sono state ottenute con riferimento alla disponibilità di posti-letto, all'occupazione attivata nelle strutture ricettive e di ristorazione ed alle presenze ufficiali e negli alloggi privati. Particolarmente rilevanti sono anche la densità di superficie per 1.000 residenti e la lunghezza delle linee di trasporto.

Seconda componente: spiega il 9,5% della varianza e si identifica essenzialmente con l'attivazione economica della circoscrizione (valore aggiunto per 1.000 residenti e quota di addetti sui residenti) e la dotazione di una particolare tipologia di infrastrutture, quelle ospedaliere. Contrastanti sembrano, invece, le indicazioni derivanti dalle variabili ambientali.

Terza componente: spiega il 7,9% della varianza ed è fortemente caratterizzata dalla presenza di punti di accesso (porti, aeroporti) e dall'attivazione occupazionale in attività di intrattenimento e di servizi per il turista (punti vendita commerciali al dettaglio, agenzie di viaggi, porti turistici e, sebbene in forma più contenuta, fiere ed occupati nei servizi di intrattenimento). Queste peculiarità, tuttavia, si associano a scarsa attenzione per l'ambiente, carenza di alcune infrastrutture (parcheggi, linee di trasporto pubblico) e delle attrattive storico-culturali e, infine, da una domanda finale non molto elevata e limitata essenzialmente ai soli alberghi.

Quarta componente: spiega il 5,9% della varianza ed è identificata dalla peculiarità delle circoscrizioni fortemente caratterizzate dall'elevato ricorso agli alloggi privati piuttosto che alle strutture ricettive ufficiali. Rispetto alla prima componente assumono un peso superiore la densità degli ospedali e di chiese, palazzi e monumenti.

Quinta componente: spiega il 4,8% della varianza ed è particolarmente caratterizzata dalla carenza di attrattive storico-culturali (indice TAI3), da una scarsa attivazione occupazionale indotta da attività turistiche e da elevata stagionalità; gli aspetti più positivi, ma che non sembrano avere un particolare effetto sulla turisticità effettiva, sono dati da alcuni indicatori ambientali e dall'elevata dotazione media di strade.

Sesta componente: spiega il 4,4% della varianza ed è la componente delle attrattive culturali (chiese, palazzi e monumenti, musei) – con eccezione dei siti UNESCO – e, sebbene in misura minore, del numero di fiere. Scarse le peculiarità ambientali e le dotazioni infrastrutturali, così come l'intensità di domanda finale e, in particolare, del turismo in alloggi privati.

Settima componente: spiega il 4% della varianza e coagula alcuni indicatori poco rappresentati (o rappresentati solo in parte) dalle componenti precedenti, quali l'elevata stagionalità della domanda, la presenza di porti turistici e, soprattutto, la buona attivazione occupazionale nelle attività di intrattenimento (che presentano la correlazione più elevata proprio con questa componente), mentre a fronte di una buona densità media di strutture complementari si associa un ricorso effettivo quasi esclusivo alle strutture alberghiere.

[27] Si ricorda che tali coefficienti sono compresi tra –1 e +1. La possibilità di poter reinterpretare *a posteriori*, con il metodo *CP*, il significato semantico intrinseco della matrice dei dati originaria, assegnando un significato specifico alle prime componenti derivate dall'analisi fattoriale, appare meno stringente – sebbene sempre utile – rispetto ad altri contesti. Ciò in quanto la stessa selezione *a priori* delle variabili di *input* - come già evidenziato - è stata notevolmente filtrata e si è basata sull'associazione a priori di ogni variabile ad uno specifico aspetto della turisticità.

In sintesi, i fattori latenti che discriminano maggiormente tra diversi livelli di attrattività turistica sono rappresentati da: I) attivazione di offerta e domanda turistica effettiva, II) sviluppo economico complessivo, III) attivazione di attività di intrattenimento e di servizi per il turista ed accessibilità, IV) turismo negli alloggi privati piuttosto che nelle strutture ricettive ufficiali, V) attrattori ambientali (poco connessi con la domanda finale), VI) presenza di attrattori storico-culturali (poco connessi alla domanda finale), VII) elevata stagionalità ed altri indicatori residuali.

Nel complesso, è poco chiaro il ruolo giocato dagli indicatori ambientali, su cui può aver influito anche la necessità di ricorrere a stime per passare dai dati provinciali a quelli circoscrizionali. In effetti, se si considerano i contributi assoluti delle variabili originarie con riferimento alle sette componenti analizzate, si nota come le variabili TAI1_2 (sostanza inquinanti), TAI1_3 (depurazione acque) e TAI1_6 (parchi) siano mal rappresentate e, quindi, o non sono connesse con l'attrattività turistica, oppure sono fortemente correlate con altre variabili più discriminanti che, in un certo senso, le finiscono con incorporare. Un'informazione piuttosto simile deriva dalla colonna dei pesi finali, che conferma anche la scarsa significatività di attrattori come TAI3_3 (siti UNESCO), TAI4_1 (terme) e TAI4_3 (mercati, fiere, feste).

A livello aggregato, il grafico 5.1 - che riporta i valori medi dei contributi assoluti e dei pesi per i 7 indici di turisticità sintetici – conferma la forte rilevanza degli indici TII1, TII2 e TBI (i contributi assoluti medi sono pari rispettivamente a 0,71, 0,64 e 0,63), mentre tra gli indici di attrattività prevale quello legato alla dotazione infrastrutturale (TAI2, 0,50), a fronte della modesta rilevanza degli "altri attrattori" (TAI4, 0,29).

Nel dettaglio (grafico 5.2), gli indicatori di offerta turistica teorica (TBI_1), di attivazione turistica indiretta (TII1_4, addetti 55.3 e 55.4), di domanda finale in alloggi privati (TII2_3) e di offerta in alloggi privati (TBI_3) giocano il ruolo più rilevante (correlazioni con l'indice sintetico di turisticità TI comprese tra 0,79 e 0,68); le presenze negli alberghi sono meno discriminanti di quelle nelle strutture complementari (rispettivamente, 0,43 e 0,63) e, tra le variabili infrastrutturali, primeggiano TAI2_3 (lunghezza delle linee di trasporto pubblico) e TAI2_6 (addetti nel commercio al dettaglio).

Grafico 5.1 - Contributi assoluti e pesi relativi dei 7 gruppi di indicatori

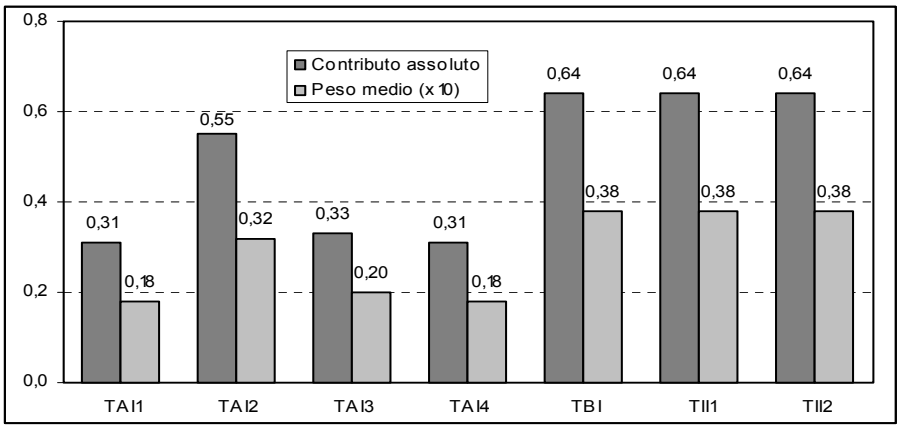
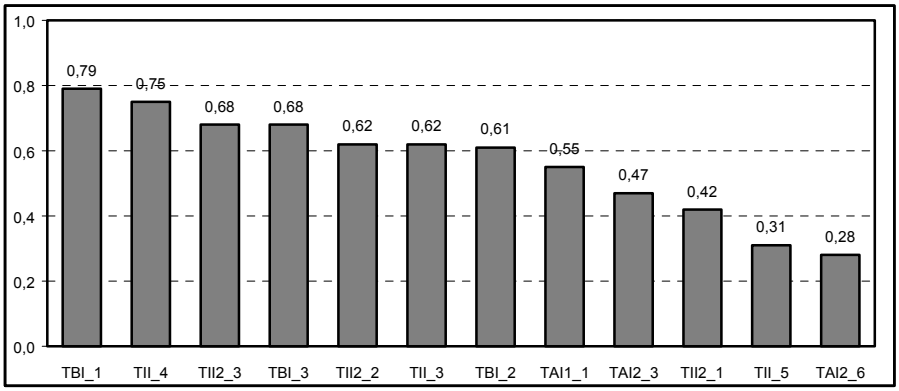


Tabella 5.1 - Risultati del metodo CP ottenuti sulla base degli indicatori relativi

Sigla	Variabili originarie	Correlazioni con le componenti							Contributo assoluto	Stima peso finale
		I (15,3)	II (24,8)	III (32,7)	IV (38,8)	V (43,6)	VI (48,0)	VII (52,0)		
TAI1_1	Superficie	0,698	-0,260	-0,300	0,241	0,160	0,115	-0,193	0,779	0,043
TAI1_2	Sostanze inquinanti	0,010	-0,176	0,128	-0,152	0,399	0,145	-0,183	0,284	0,021
TAI1_3	*Depurazione acque*	*0,284*	*0,103*	*-0,096*	*0,080*	*-0,003*	*0,059*	*-0,077*	*0,116*	*0,016*
TAI1_4	Raccolta differenziata	0,144	0,294	-0,598	0,079	-0,115	-0,242	0,269	0,615	0,033
TAI1_5	Violazione codice strada	-0,114	-0,148	-0,017	0,306	0,480	-0,024	0,219	0,408	0,023
TAI1_6	Parchi	0,140	0,127	0,063	0,282	0,022	-0,306	0,047	0,216	0,019
TAI2_1	Porto-aeroporto	-0,212	0,374	0,497	0,266	0,070	-0,027	-0,116	0,522	0,032
TAI2_2	Km di strade	0,044	-0,268	0,199	0,056	0,488	0,250	0,020	0,418	0,024
TAI2_3	Km di linee autobus, tram, …	0,554	-0,111	-0,329	0,293	0,138	0,284	-0,172	0,643	0,039
TAI2_4	Parcheggi	0,014	0,207	-0,388	-0,005	-0,290	-0,139	0,246	0,358	0,023
TAI2_5	Addetti ferrovie	-0,300	0,254	0,194	0,252	0,162	-0,018	-0,311	0,379	0,030
TAI2_6	Addetti commercio al dettaglio (52)	0,349	-0,214	0,584	-0,002	-0,206	-0,173	-0,061	0,585	0,033
TAI2_7	Numero ospedali	-0,395	0,526	0,127	0,372	0,000	-0,025	0,089	0,596	0,033
TAI3_1	Chiese-palazzi-monumenti	0,191	0,108	0,018	0,388	-0,131	0,516	0,154	0,506	0,027
TAI3_2	Musei	0,229	-0,038	-0,165	0,221	-0,262	0,553	0,144	0,525	0,029
TAI3_3	Siti UNESCO	-0,115	0,173	0,043	0,278	-0,324	-0,047	0,013	0,230	0,019
TAI4_1	*Terme*	*0,077*	*-0,031*	*0,102*	*-0,127*	*-0,239*	*0,028*	*0,248*	*0,153*	*0,015*
TAI4_2	Fiere	-0,102	0,354	0,365	0,294	0,049	0,135	-0,158	0,401	0,028
TAI4_3	Mercati-fiere-feste	0,132	-0,098	-0,094	0,202	0,050	0,306	0,106	0,184	0,018
TAI4_4	Porto turistico (posti barca)	0,155	0,070	0,505	0,023	0,166	-0,091	0,339	0,435	0,025
TBI_1	Numero letti alberghi	0,762	0,380	0,011	-0,290	0,095	0,061	-0,145	0,843	0,039
TBI_2	Numero letti complementari	0,558	0,128	0,143	-0,003	0,158	-0,008	0,360	0,503	0,029
TBI_3	Numero letti alloggi privati	0,720	-0,174	-0,043	0,410	0,065	-0,351	-0,117	0,860	0,040
TBI_4	Prezzo medio albergo	0,129	-0,233	-0,366	-0,107	0,132	-0,161	0,216	0,306	0,026
TII1_1	Valore aggiunto	-0,022	0,863	-0,266	0,023	0,275	-0,013	-0,001	0,892	0,030
TII1_2	Addetti	-0,048	0,843	-0,300	0,052	0,286	-0,036	0,047	0,891	0,033
TII1_3	Addetti 55.1-55.2	0,682	0,288	0,077	-0,367	-0,178	0,220	-0,111	0,781	0,041
TII1_4	Addetti 55.3-55.4	0,817	-0,030	0,174	0,189	-0,161	-0,098	0,137	0,789	0,036
TII1_5	Addetti 63.3	0,264	0,297	0,504	-0,048	-0,255	-0,055	-0,177	0,514	0,032
TII1_6	Addetti 92.3-92.5-92.72.1	0,128	0,127	0,311	0,295	-0,201	0,134	0,347	0,395	0,028
TII2_1	Presenze alberghi	0,418	0,198	0,223	-0,349	0,189	-0,001	0,316	0,521	0,034
TII2_2	Presenze complementari	0,602	0,407	0,048	-0,455	0,059	0,074	-0,149	0,768	0,039
TII2_3	Presenze alloggi privati	0,717	-0,160	-0,007	0,355	0,105	-0,395	-0,116	0,846	0,039
TII2_4	Stagionalità	0,044	0,161	-0,306	0,149	-0,326	-0,018	-0,421	0,428	0,025

Nota: sono evidenziate in neretto le variabili con correlazioni pari ad almeno 0,5; sottolineate quelle con correlazioni inferiori a 0,5 ma superiori a 0,3, in corsivo quelle mal rappresentate su ognuno dei 7 assi fattoriali (basso contributo assoluto). Le cifre in parentesi nella testata indicano le quote cumulate di varianza spiegata dalle componenti.

Grafico 5.2 - Coefficienti di correlazione lineare tra gli indicatori originari e l'indice sintetico di turisticità TI – I primi 12 indicatori

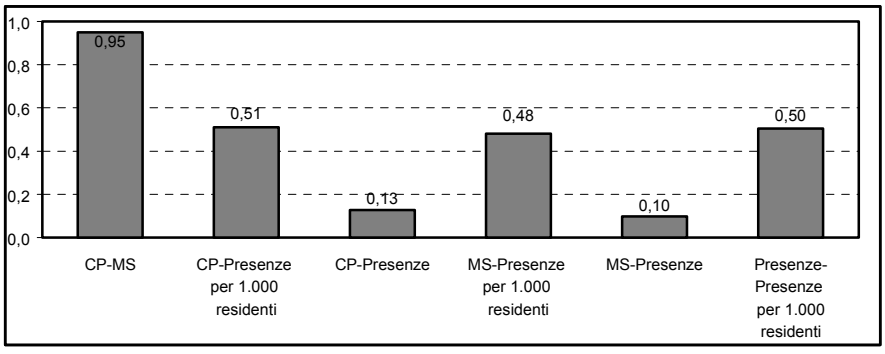


Legenda

TBI_1	Numero letti alberghi	TBI_2	Numero letti complementari
TII1_4	Addetti 55.3-55.4	TAI1_1	Superficie
TII2_3	Presenze alloggi privati	TAI2_3	Km di linee autobus, tram, …
TBI_3	Numero letti alloggi privati	TII2_1	Presenze alberghi
TII2_2	Presenze complementari	TII1_5	Addetti 63.3
TII1_3	Addetti 55.1-55.2	TAI2_6	Addetti commercio al dettaglio (52)

Sebbene, in questo contesto, si siano considerati i soli risultati ottenuti con il criterio di sintesi espresso dalla formula (4.1a), può essere utile riassumere – come fatto nel grafico 5.3 – un'analisi di concordanza tra i risultati ottenuti e quelli ottenibili in base ad altri criteri, sulla base del calcolo dei coefficienti di correlazione lineare tra i punteggi finali attribuiti alle singole circoscrizioni con vari metodi. I metodi *CP* e *MS* sono risultati piuttosto concordanti (correlazione pari a 0,95), mentre ben più contenuta è la concordanza tra le graduatorie ottenibili con i metodi suddetti e la semplice graduatoria derivata dall'ordinamento decrescente delle circoscrizioni in base al rapporto tra presenze nelle strutture ricettive ufficiali e 1.000 residenti (0,53 per *CP* e 0,48 per *MS*). D'altra parte, la similitudine con la graduatoria ottenuta in base alle sole presenze assolute (ossia non rapportate a nessun indicatore dimensionale) è molto più bassa.

Grafico 5.3 - Coefficienti di correlazione lineare tra gli indici di turisticità ottenuti utilizzando le variabili relative



6. Gli indicatori di competitività a confronto

La graduatoria rispetto all'indice di turisticità sintetico TI delle prime 30 circoscrizioni (tabella 6.1) premia la circoscrizione di Limone del Garda (provincia di Brescia), seguita da Medesimo (Sondrio), Corvara in Badia (Bolzano), Aosta e Val Brembana (Bergamo).

Limone del Garda e Madesimo distanziano nettamente la circoscrizione al terzo posto, e solo 6 circoscrizioni (le precedenti più Gran San Bernardo) presentano un indice globale di turisticità superiore a 0,5; l'indice medio di turisticità è pari a 0,182 per l'insieme delle 547 circoscrizioni ed a 0,439 per le prime 30 in graduatoria[28].

Poiché il limite massimo teorico dell'indice è pari ad uno, emerge un quadro che conferma come lo sviluppo turistico locale possa essere perseguito cercando di ottimizzare quantomeno alcuni dei numerosi attrattori introdotti in analisi, risultando irrealistico l'obiettivo di raggiungere i livelli di eccellenza in corrispondenza di tutti i 7 indicatori di primo livello in cui è stata scomposta l'attrattività turistica. Ad esempio, le prime 5 circoscrizioni sono carenti per quanto riguarda TAI3 (attrattori storico-culturali) e TAI4 (altri attrattori), ma presentano livelli medio-alti con riferimento alla maggioranza degli altri indicatori.

Nel complesso, nelle prime 30 circoscrizioni è stato consumato, nel 2003, solo il 5,33% delle presenze (18,4 milioni sui 344,4 complessivi), ad indicare che la graduatoria non premia le dimensioni turistiche assolute in quanto tali, ma consente di identificare profili effettivamente o potenzialmente molto efficienti in chiave turistica in rapporto alla popolazione residente. E' anche per questo motivo che nella graduatoria delle prime 30 circoscrizioni non trovano posto i grandi centri urbani e diversi bacini altamente attrattivi, ma evidentemente poco dotati di alcune specifiche tipologie di attrattività.

Un risultato di particolare rilevanza è dato dal fatto che l'indice sintetico TAI, per quanto più elevato se valutato sull'insieme delle prime 30 circoscrizioni (0,27) oppure sull'insieme complessivo (0,17), appare meno discriminante rispetto agli indici TBI (0,59 contro 0,21) e TII (0,46 contro 0,17). In altri termini, l'indice TAI finisce con includere componenti in realtà non sempre molto variabili da circoscrizione a circoscrizione, o comunque non determinanti ai fini della attrattività turistica globale di un sito, sebbene su tale aspetto – come già ricordato – possa avere influito anche la necessità di stimare a livello di circoscrizione alcuni degli indicatori elementari, disponibili originariamente solo a livello provinciale.

Nelle prime 30 posizioni si collocano 4 circoscrizioni turistiche localizzate nelle province di Aosta e Bolzano e 3 circoscrizioni delle province di Brescia, L'Aquila e Trento. In particolare (tabella 6.2), se si considera esplicitamente la graduatoria delle prime 30 province[29], Padova, Aosta, L'Aquila, Sondrio e Verona si aggiudicano le prime 5 posizioni, in una griglia che premia, in particolare, l'Emilia Romagna (con 5 province), la Lombardia (4 province), il Friuli Venezia Giulia, la Liguria ed il Veneto (3 province).

Nel complesso, le prime 30 province determinano una quota elevata, ma non

[28] E' utile ricordare che le graduatorie derivano da una valutazione *complessiva* del livello di integrazione, in ogni circoscrizione, tra infrastrutture generiche, attrattori storico-naturali-culturali, infrastrutture turistico-ricettive e domanda finale effettiva. Graduatorie evidentemente differenti si sarebbero potute ottenere senza riferire molte delle variabili analizzate ai residenti, o escludendo comunque una serie di indicatori infrastrutturali.

[29] I punteggi delle province sono stati ottenuti per media aritmetica semplice dei punteggi delle relative circoscrizioni. In modo analogo sono stati ottenuti i punteggi delle regioni e delle altre aggregazioni della successiva tabella 6.3.

preponderante delle presenze turistiche ufficiali (il 42%, equivalente a 144,8 milioni di notti), con le punte massime di Bolzano (7,45%), Rimini (4,46%) e Trento (4,03%). Anche in questo caso, l'indice infrastrutturale TAI è risultato poco discriminante, assumendo un valore medio di 0,19, dunque molto simile a quello medio generale (0,17).

L'area Nord/est è presente nelle prime 30 posizioni con 13 circoscrizioni, seguita dal Nord/ovest con 10 e dal Mezzogiorno con 7, mentre il Centro non è rappresentato. La performance non brillante di tale area geografica è confermata dalla tabella 6.3: essa si caratterizza per il punteggio medio di TI più basso (0,158), rispetto a 0,162 del Mezzogiorno, 0,202 del Nord/ovest ed a 0,204 del Nord/est.

Nel dettaglio, le regioni le cui circoscrizioni turistiche hanno registrato i punteggi sintetici TI mediamente più alti sono così state, nell'ordine, Valle d'Aosta, Trentino Alto Adige, Lombardia, Liguria, Friuli Venezia Giulia, Abruzzo e Veneto.

Con riferimento alle tipologie di località, quelle montane e lacuali si collocano saldamente ai primi 2 posti della graduatoria, seguite da quelle di interesse storico-artistico e dalle località marine; meno positiva è la *performance* delle località termali e, come prevedibile, dei capoluoghi di provincia e dei comuni non altrove classificabili (n.a.c.).

Sulla base della tabella 6.4, tra le prime 5 circoscrizioni nella graduatoria dell'indice globale TI solo quella della Val Brembana occupa in una sola occasione una delle prime 5 posizioni nelle graduatorie relative ai 7 indici sintetici di primo livello (quella relativa a TBI), mentre troviamo Medesimo nei primi 5 posti in relazione a 4 indicatori, nonché Aosta, Monte Bianco, Limone del Garda e Corvara in Badia in relazione a 3 indicatori. Per il resto, oltre alla Val Brembana ben altre 18 circoscrizioni occupano per una sola volta[30] uno dei primi 5 posti nelle 7 graduatorie. Questa forte segmentazione è un ulteriore segnale della poliedricità della gamma di variabili considerate per l'analisi, per cui circoscrizioni relativamente poco note possono risultare valorizzate con riferimento ad almeno un gruppo di indicatori.

A livello provinciale, una delle prime 5 posizioni è occupata per 3 volte da Aosta, L'Aquila, Padova e Sondrio, e per 3 volte da Bologna, Bolzano, Varese e Verona. Tra le regioni, primeggia nettamente la Valle D'Aosta (6 volte su 7), seguita dal Trentino Alto Adige e dalla Lombardia (4 volte), mentre del tutto assenti tra le posizioni di élite sono le regione del Centro.

Un'ulteriore sintesi dei risultati è proposta nella tabella 6.5, che riepiloga i punteggi medi relativi ai diversi indici di turisticità per ogni area geografica.

Come già visto, la variabilità tra le aree dell'indice globale TI non è particolarmente elevata (si passa dal massimo di 0,204 nel Nord/est al minimo di 0,158 nel centro).

L'indice di turisticità più stabile al variare delle aree geografiche è TAI2 (infrastrutture) - con un coefficiente di variazione tra aree pari a 0,07 - seguito da TAI4 (altri attrattori); d'altra parte, l'indice TII2 (domanda turistica) è quello più variabile (coefficiente di variazione pari a 0,24), mentre tra gli indici del gruppo TAI la maggiore variabilità relativa spetta a TAI1 (territorio ed ambiente, 0,14).

[30] Si tratta, nel dettaglio, di Altri comuni Varese, Rovigo - località marine, Verona - altri comuni, Terme Euganee, Biella, Caserta, Castel San Pietro Terme, Grado, La Maddalena-Palau, Località lacuali Pinerolo, Località termali Firenze, Pisa, Ravello, Reggio di Calabria, Rivisondoli, Santa Teresa Gallura, Sirmione, Sorrento-Sant'Agnello.

Tabella 6.1 - Le prime 30 circoscrizioni turistiche rispetto all'indice di turistica sintetico TI *(presenze in migliaia)*

CIRCOSCRIZIONE TURISTICA	Provincia	Presenze	Presenze %	TAI1	TAI2	TAI3	TAI4	TBI	TII1	TII2	TAI	TBI	TII	TI	Posiz.
Limone del Garda	Brescia	826	0,24	0,51	0,31	0,00	0,00	0,93	1,00	1,00	0,21	0,93	1,00	0,712	1
Madesimo	Sondrio	124	0,04	0,90	0,34	0,00	0,00	1,00	0,59	0,92	0,31	1,00	0,75	0,688	2
Corvara in Badia	Bolzano	854	0,25	0,39	0,27	0,00	0,00	0,87	0,62	0,91	0,16	0,87	0,77	0,600	3
A.I.A.T. Aosta	Aosta	299	0,09	1,00	1,00	0,00	0,00	0,86	0,46	0,33	0,50	0,86	0,39	0,583	4
Val Brembana	Bergamo	37	0,01	0,66	0,15	0,00	0,00	0,84	0,44	0,61	0,20	0,84	0,52	0,524	5
A.I.A.T. Gran San Bernardo	Aosta	113	0,03	0,51	0,52	0,31	0,20	0,76	0,45	0,35	0,38	0,76	0,40	0,514	6
Rivisondoli	L'Aquila	84	0,02	0,54	0,16	0,00	0,00	0,74	0,42	0,69	0,17	0,74	0,56	0,490	7
A.I.A.T. Monte Bianco	Aosta	715	0,21	0,73	0,75	0,42	0,11	0,59	0,35	0,35	0,50	0,59	0,35	0,480	8
Dolomiti di Brenta-Paganella	Trento	1.078	0,31	0,37	0,19	0,00	0,00	0,68	0,49	0,63	0,14	0,68	0,56	0,458	9
Grado	Gorizia	1.435	0,42	0,42	0,35	0,14	0,74	0,51	0,37	0,53	0,41	0,51	0,45	0,457	10
Selva di Val Gardena	Bolzano	1.003	0,29	0,34	0,15	0,00	0,14	0,58	0,58	0,62	0,16	0,58	0,60	0,446	11
Valle di Fassa	Trento	2.498	0,73	0,38	0,27	0,13	0,06	0,58	0,40	0,56	0,21	0,58	0,48	0,422	12
Roccaraso	L'Aquila	231	0,07	0,49	0,11	0,00	0,00	0,59	0,49	0,54	0,15	0,59	0,52	0,419	13
Amerino - Amelia	Terni	60	0,02	0,36	0,17	0,27	0,53	0,55	0,37	0,36	0,34	0,55	0,36	0,416	14
Santa Teresa Gallura	Sassari	702	0,20	0,31	0,43	0,00	0,64	0,47	0,47	0,36	0,35	0,47	0,42	0,411	15
Località lacuali Pinerolo	Torino	16	0,00	0,42	0,29	0,20	1,00	0,37	0,50	0,21	0,48	0,37	0,35	0,400	16
A.I.A.T. Cogne-Gran Paradiso	Aosta	174	0,05	0,52	0,44	0,21	0,51	0,45	0,36	0,26	0,42	0,45	0,31	0,393	17
APT di Genova - località collinari	Genova	88	0,03	0,32	0,26	0,00	0,25	0,70	0,30	0,23	0,21	0,70	0,27	0,392	18
Ponte di Legno	Brescia	201	0,06	0,62	0,28	0,10	0,00	0,51	0,37	0,43	0,25	0,51	0,40	0,387	19
Sesto	Bolzano	560	0,16	0,52	0,12	0,20	0,12	0,43	0,40	0,54	0,24	0,43	0,47	0,380	20
APT di Padova - altri comuni	Padova	485	0,14	0,32	0,11	0,00	0,00	0,71	0,31	0,28	0,11	0,71	0,29	0,369	21
Vieste	Foggia	1.791	0,52	0,38	0,38	0,05	0,21	0,52	0,30	0,30	0,26	0,52	0,30	0,358	22
Pescasseroli	L'Aquila	241	0,07	0,54	0,27	0,35	0,00	0,36	0,54	0,32	0,29	0,36	0,43	0,358	23
Livigno	Sondrio	829	0,24	0,59	0,38	0,00	0,14	0,39	0,39	0,36	0,28	0,39	0,38	0,349	24
Folgaria-Lavarone-Luserna	Trento	482	0,14	0,33	0,22	0,10	0,09	0,49	0,31	0,43	0,18	0,49	0,37	0,348	25
Sirmione	Brescia	1.004	0,29	0,33	0,19	0,06	0,12	0,36	0,63	0,37	0,18	0,36	0,50	0,346	26
Caltanissetta	Caltanissetta	63	0,02	0,49	0,26	0,00	0,01	0,48	0,34	0,38	0,19	0,48	0,36	0,344	27
APT di Verona - altri comuni	Verona	590	0,17	0,72	0,54	0,38	0,07	0,37	0,36	0,11	0,43	0,37	0,24	0,344	28
Scena	Bolzano	925	0,27	0,33	0,02	0,00	0,00	0,42	0,57	0,47	0,09	0,42	0,52	0,342	29
Otranto	Lecce	851	0,25	0,28	0,35	0,00	0,62	0,33	0,41	0,36	0,31	0,33	0,38	0,341	30
Totale/Media		18.362	5,33	0,49	0,31	0,10	0,17	0,59	0,46	0,46	0,27	0,59	0,46	0,439	
Totale/Media generale (le 547 circoscrizioni)		344.520	100,00	0,30	0,22	0,06	0,08	0,21	0,21	0,14	0,17	0,21	0,17	0,182	
Mediana		-	-	0,46	0,27	0,00	0,08	0,53	0,42	0,37	0,25	0,53	0,41	0,406	
Mediana generale (le 547 circoscrizioni)		-	-	0,30	0,19	0,03	0,05	0,17	0,18	0,11	0,15	0,17	0,14	0,158	

Tabella 6.2 - Le prime 30 province rispetto all'indice di turistica sintetico TI *(presenze in migliaia)*

PROVINCIA	Regione	Presenze	Presenze %	TAI1	TAI2	TAI3	TAI4	TBI	TII1	TII2	TAI	TBI	TII	TI	Posiz.
Padova	Veneto	4.607	1,34	0,26	0,29	0,38	0,09	0,35	0,33	0,18	0,25	0,35	0,26	0,287	1
Aosta	Valle d'Aosta	3.496	1,01	0,41	0,34	0,15	0,16	0,34	0,26	0,23	0,27	0,34	0,25	0,286	2
L'Aquila	Abruzzo	1.493	0,43	0,47	0,20	0,06	0,06	0,34	0,28	0,28	0,20	0,34	0,28	0,274	3
Sondrio	Lombardia	2.309	0,67	0,46	0,18	0,04	0,03	0,33	0,24	0,28	0,18	0,33	0,26	0,256	4
Verona	Veneto	10.667	3,10	0,42	0,42	0,19	0,14	0,25	0,27	0,15	0,29	0,25	0,21	0,254	5
Brescia	Lombardia	7.353	2,13	0,40	0,22	0,06	0,07	0,31	0,31	0,21	0,19	0,31	0,26	0,251	6
Gorizia	Friuli-Venezia Giulia	1.850	0,54	0,36	0,27	0,09	0,29	0,27	0,23	0,21	0,25	0,27	0,22	0,247	7
Terni	Umbria	774	0,22	0,34	0,27	0,12	0,22	0,28	0,23	0,20	0,24	0,28	0,22	0,247	8
Bolzano	Trentino-Alto Adige	25.675	7,45	0,33	0,16	0,05	0,08	0,28	0,30	0,27	0,15	0,28	0,29	0,242	9
Trento	Trentino-Alto Adige	13.895	4,03	0,33	0,19	0,04	0,04	0,31	0,24	0,24	0,15	0,31	0,24	0,232	10
Bergamo	Lombardia	1.358	0,39	0,39	0,16	0,07	0,04	0,30	0,22	0,20	0,17	0,30	0,21	0,226	11
Belluno	Veneto	5.485	1,59	0,29	0,18	0,06	0,05	0,32	0,21	0,19	0,15	0,32	0,20	0,222	12
Avellino	Campania	262	0,08	0,39	0,13	0,08	0,00	0,29	0,30	0,14	0,15	0,29	0,22	0,221	13
Bologna	Emilia-Romagna	3.353	0,97	0,36	0,37	0,15	0,07	0,21	0,23	0,18	0,24	0,21	0,20	0,215	14
Genova	Liguria	3.169	0,92	0,25	0,20	0,02	0,09	0,30	0,21	0,14	0,14	0,30	0,17	0,205	15
Sassari	Sardegna	5.309	1,54	0,29	0,28	0,03	0,28	0,19	0,24	0,12	0,22	0,19	0,18	0,199	16
Cagliari	Sardegna	2.947	0,86	0,23	0,18	0,04	0,05	0,26	0,22	0,19	0,13	0,26	0,21	0,199	17
Caltanissetta	Sicilia	163	0,05	0,29	0,24	0,01	0,01	0,26	0,15	0,22	0,14	0,26	0,19	0,195	18
Ferrara	Emilia-Romagna	2.271	0,66	0,30	0,20	0,08	0,06	0,23	0,22	0,15	0,16	0,23	0,18	0,192	19
Savona	Liguria	6.586	1,91	0,27	0,32	0,05	0,12	0,22	0,21	0,11	0,19	0,22	0,16	0,192	20
Udine	Friuli-Venezia Giulia	5.604	1,63	0,29	0,17	0,04	0,16	0,23	0,24	0,11	0,16	0,23	0,18	0,189	21
Modena	Emilia-Romagna	1.375	0,40	0,32	0,29	0,06	0,08	0,21	0,18	0,16	0,19	0,21	0,17	0,189	22
Cuneo	Piemonte	1.001	0,29	0,42	0,21	0,06	0,07	0,21	0,20	0,11	0,19	0,21	0,15	0,186	23
Ravenna	Emilia-Romagna	6.242	1,81	0,29	0,22	0,05	0,16	0,20	0,29	0,06	0,18	0,20	0,17	0,185	24
Rimini	Emilia-Romagna	15.349	4,46	0,33	0,21	0,05	0,12	0,21	0,26	0,07	0,18	0,21	0,17	0,184	25
Lecce	Puglia	3.235	0,94	0,26	0,18	0,02	0,23	0,21	0,22	0,12	0,17	0,21	0,17	0,184	26
Trieste	Friuli-Venezia Giulia	786	0,23	0,20	0,32	0,08	0,13	0,20	0,23	0,12	0,18	0,20	0,17	0,184	27
Torino	Piemonte	3.561	1,03	0,35	0,22	0,09	0,18	0,17	0,19	0,15	0,21	0,17	0,17	0,184	28
Imperia	Liguria	3.562	1,03	0,27	0,29	0,06	0,13	0,20	0,20	0,12	0,18	0,20	0,16	0,183	29
Varese	Lombardia	1.066	0,31	0,22	0,25	0,23	0,10	0,14	0,30	0,10	0,20	0,14	0,20	0,180	30
Totale/Media		144.802	42,03	0,33	0,24	0,08	0,11	0,26	0,24	0,17	0,19	0,26	0,20	0,217	
Totale/Media generale		344.520	100,00	0,30	0,22	0,06	0,08	0,21	0,21	0,14	0,17	0,21	0,17	0,182	
Mediana		-	-	0,32	0,22	0,06	0,09	0,26	0,23	0,16	0,18	0,26	0,20	0,202	
Media generale		-	-	0,30	0,19	0,03	0,05	0,17	0,18	0,11	0,15	0,17	0,14	0,158	

Tabella 6.3 - I punteggi sintetici per regione, area geografica e tipologia di località turistica *(presenze in migliaia)*

Regione/area/località	Presenze	Presenze %	TAI1	TAI2	TAI3	TAI4	TBI	TII1	TII2	TAI	TBI	TII	TI	Posiz.
Piemonte	8.944	2,60	0,35	0,20	0,06	0,06	0,18	0,17	0,12	0,17	0,18	0,15	0,166	11
Valle d'Aosta	3.496	1,02	0,41	0,34	0,15	0,16	0,34	0,26	0,23	0,27	0,34	0,25	0,286	1
Lombardia	25.972	7,54	0,37	0,21	0,06	0,06	0,24	0,24	0,18	0,18	0,24	0,21	0,210	3
Trentino-Alto Adige	39.571	11,49	0,33	0,17	0,05	0,07	0,29	0,28	0,26	0,15	0,29	0,27	0,239	2
Veneto	55.112	16,00	0,27	0,22	0,11	0,09	0,20	0,21	0,15	0,17	0,20	0,18	0,185	7
Friuli-Venezia Giulia	8.863	2,57	0,26	0,22	0,06	0,16	0,22	0,22	0,14	0,18	0,22	0,18	0,190	5
Liguria	14.770	4,29	0,26	0,29	0,05	0,11	0,23	0,20	0,12	0,18	0,23	0,16	0,191	4
Emilia-Romagna	36.621	10,63	0,32	0,23	0,06	0,08	0,19	0,21	0,11	0,17	0,19	0,16	0,175	10
Toscana	36.837	10,70	0,31	0,21	0,06	0,08	0,17	0,16	0,10	0,16	0,17	0,13	0,154	15
Umbria	5.795	1,68	0,33	0,25	0,06	0,11	0,20	0,18	0,14	0,19	0,20	0,16	0,180	8
Marche	13.449	3,91	0,28	0,17	0,06	0,07	0,21	0,19	0,07	0,14	0,21	0,13	0,163	12
Lazio	24.055	6,98	0,30	0,20	0,06	0,08	0,16	0,18	0,09	0,16	0,16	0,14	0,151	17
Abruzzo	7.115	2,07	0,32	0,21	0,04	0,06	0,23	0,21	0,14	0,16	0,23	0,18	0,189	6
Molise	769	0,22	0,23	0,23	0,02	0,04	0,17	0,15	0,08	0,13	0,17	0,12	0,138	20
Campania	19.709	5,72	0,14	0,23	0,07	0,12	0,15	0,27	0,12	0,14	0,15	0,19	0,162	13
Puglia	10.703	3,11	0,27	0,22	0,04	0,12	0,19	0,15	0,09	0,16	0,19	0,12	0,155	14
Basilicata	1.762	0,51	0,27	0,16	0,10	0,07	0,14	0,21	0,07	0,15	0,14	0,14	0,142	19
Calabria	7.334	2,13	0,27	0,28	0,05	0,07	0,17	0,14	0,06	0,17	0,17	0,10	0,147	18
Sicilia	13.152	3,82	0,23	0,24	0,05	0,09	0,16	0,19	0,12	0,15	0,16	0,15	0,153	16
Sardegna	10.384	3,01	0,25	0,25	0,04	0,18	0,19	0,21	0,11	0,18	0,19	0,16	0,177	9
Nord/ovest	53.182	15,44	0,36	0,23	0,07	0,08	0,23	0,22	0,16	0,18	0,23	0,19	0,202	2
Nord/est	140.167	40,70	0,31	0,20	0,06	0,08	0,24	0,24	0,18	0,16	0,24	0,21	0,204	1
Centro	80.137	23,27	0,30	0,20	0,06	0,08	0,18	0,17	0,10	0,16	0,18	0,13	0,158	4
Sud/isole	70.928	20,59	0,24	0,23	0,05	0,10	0,18	0,20	0,11	0,16	0,18	0,15	0,162	3
Città di interesse storico e artistico	75.680	21,97	0,31	0,21	0,08	0,14	0,21	0,21	0,12	0,19	0,21	0,16	0,188	3
Località collinari	12.377	3,59	0,30	0,19	0,06	0,07	0,19	0,17	0,11	0,16	0,19	0,14	0,163	6
Località lacuali	22.278	6,47	0,35	0,21	0,06	0,08	0,24	0,28	0,19	0,17	0,24	0,23	0,218	2
Località marine	115.920	33,66	0,23	0,27	0,04	0,12	0,19	0,24	0,14	0,17	0,19	0,19	0,183	4
Località montane	46.097	13,38	0,37	0,19	0,07	0,07	0,29	0,23	0,23	0,18	0,29	0,23	0,233	1
Località termali	13.871	4,03	0,27	0,28	0,05	0,12	0,15	0,22	0,08	0,18	0,15	0,15	0,162	7
Località religiose	302	0,09	0,35	0,16	0,04	0,12	0,19	0,20	0,11	0,17	0,19	0,15	0,171	5
Capoluogo di provincia n. a. c.	8.000	2,32	0,28	0,26	0,07	0,1	0,15	0,22	0,1	0,18	0,15	0,16	0,161	8
Comuni n. a. c.	49.888	14,48	0,28	0,16	0,05	0,03	0,17	0,13	0,09	0,13	0,17	0,11	0,137	9
Totale/Media generale	**344.520**	**100,00**	**0,30**	**0,22**	**0,06**	**0,08**	**0,21**	**0,21**	**0,14**	**0,17**	**0,21**	**0,17**	**0,182**	

Tabella 6.4 - Le prime 5 posizioni in graduatoria per i 7 indicatori sintetici di attrattività turistica per circoscrizione, provincia e regione

Indice	Circoscrizioni	Province	Regioni
TAI1	A.I.A.T. Aosta	L'Aquila	Valle d'Aosta
	Medesimo	Sondrio	Lombardia
	A.I.A.T. Monte Bianco	Verbano Cusio Ossola	Piemonte
	APT di Verona - altri comuni	Lodi	Umbria
	Località termali Firenze	Cuneo	Trentino-Alto Adige
TAI2	A.I.A.T. Aosta	La Spezia	Valle d'Aosta
	A.I.A.T. Monte Bianco	Verona	Liguria
	Reggio di Calabria	Bologna	Calabria
	Biella	Oristano	Sardegna
	Pisa	Reggio di Calabria	Umbria
TAI3	APT Terme Euganee	Padova	Veneto
	Altri comuni Varese	Varese	Basilicata
	Ravello	Verona	Campania
	A.I.A.T. Monte Bianco	Aosta	Piemonte
	Castel San Pietro Terme	Bologna	Sardegna
TAI4	Località lacuali Pinerolo	Caserta	Sardegna
	La Maddalena-Palau	Gorizia	Friuli-Venezia Giulia
	Caserta	Sassari	Valle d'Aosta
	Grado	Lecce	Campania
	Santa Teresa Gallura	Terni	Puglia
TBI	Medesimo	Padova	Valle d'Aosta
	Limone del Garda	Aosta	Trentino-Alto Adige
	Corvara in Badia	L'Aquila	Lombardia
	A.I.A.T. Aosta	Sondrio	Liguria
	Val Brembana	Belluno	Abruzzo
TII1	Limone del Garda	Padova	Trentino-Alto Adige
	Sirmione	Brescia	Campania
	Corvara in Badia	Bolzano	Valle d'Aosta
	Sorrento-Sant'Agnello	Varese	Lombardia
	Medesimo	Avellino	Friuli-Venezia Giulia
TII2	Limone del Garda	L'Aquila	Trentino-Alto Adige
	Medesimo	Sondrio	Valle d'Aosta
	Corvara in Badia	Bolzano	Lombardia
	Rivisondoli	Trento	Veneto
	Apt di Rovigo - località marine	Aosta	Abruzzo

Il Nord/est primeggia proprio nelle componenti della turisticità più direttamente connesse con la domanda finale, ossia le infrastrutture turistiche (TBI), il profilo economico-turistico (TII1) e gli indicatori di domanda turistica (TII2), tutte componenti che vedono il Centro in posizione di retroguardia. Il Mezzogiorno, di contro, si posiziona sempre meglio del Centro ad eccezione degli indicatori TAI1 (territorio ed ambiente, rispetto ai quali presenta un notevole *gap* rispetto a tutte le altre aree geografiche) e, con toni peraltro molto sfumati, TAI3 (attrattori storici e naturali).

In sintesi, la chiave di lettura del territorio proposta individua più nel Centro che nel Mezzogiorno un ritardo nell'adeguamento delle proprie infrastrutture e del grado di

attivazione delle funzioni produttive prettamente turistiche ai livelli della domanda potenziale, a cui fa riscontro una domanda effettiva per residente decisamente più contenuta rispetto alle altre aree.

Tabella 6.5 - Variabili originarie ed indicatori sintetici ottenuti con il metodo CP per area geografica

Sigla	Variabile	Media	Area geografica				C.v.
			Nord/ovest	Nord/est	Centro	Sud/isole	
TI	Indice di attrattività turistica	0,182	0,202	0,204	0,158	0,162	0,12
TAI1	Territorio ed ambiente	0,30	0,36	0,31	0,30	0,24	0,14
TAI1_1	Superficie	12,9	16,5	14,9	12,0	8,4	0,24
TAI1_2	Sostanze inquinanti	3,5	4,4	3,3	3,2	3,2	0,14
TAI1_3	Depurazione acque	86,2	89,2	86,1	81,9	87,1	0,03
TAI1_4	Raccolta differenziata	17,9	24,7	18,4	18,4	10,9	0,27
TAI1_5	Violazione codice strada	1,1	1,1	1,2	0,9	1,2	0,11
TAI1_6	Parchi	23,4	30,8	18,7	24,7	19,8	0,20
TAI2	Infrastrutture	0,22	0,23	0,20	0,20	0,23	0,07
TAI2_1	Porto-aeroporto	14,4	9,2	12,8	9,4	24,9	0,46
TAI2_2	Km di strade	4,7	3,6	4,7	5,4	5,1	0,15
TAI2_3	Km di linee autobus, tram, …	36,4	58,4	45,3	26,0	16,6	0,45
TAI2_4	Parcheggi	32,3	41,4	32,9	27,5	27,2	0,18
TAI2_5	Addetti ferrovie	0,2	0,2	0,2	0,2	0,3	0,19
TAI2_6	Addetti commercio al dettaglio (52)	9,9	9,6	9,3	9,7	10,7	0,05
TAI2_7	Numero ospedali	3,1	3,3	3,0	3,2	3,0	0,04
TAI3	Attrattori storici e naturali	0,06	0,07	0,06	0,06	0,05	0,12
TAI3_1	Chiese-palazzi-monumenti	3,0	2,8	4,4	1,8	3,0	0,31
TAI3_2	Musei	9,8	11,6	9,5	9,7	8,3	0,12
TAI3_3	Siti UNESCO	0,11	0,15	0,09	0,14	0,07	0,30
TAI4	Altri attrattori	0,08	0,08	0,08	0,08	0,10	0,10
TAI4_1	Terme	11,4	8,7	11,8	16,6	9,3	0,27
TAI4_2	Fiere	35,2	31,5	39,5	29,7	39,4	0,13
TAI4_3	Mercati-fiere-feste	1,6	2,2	0,6	2,3	1,4	0,42
TAI4_4	Porto turistico (posti barca)	4,1	3,0	3,6	1,2	7,8	0,62
TBI	Infrastrutture turistiche	0,21	0,23	0,24	0,18	0,18	0,13
TBI_1	Numero letti alberghi	178	178	311	88	131	0,47
TBI_2	Numero letti complementari	149	170	216	99	110	0,32
TBI_3	Numero letti alloggi privati	839	1.253	672	578	820	0,31
TBI_4	Prezzo medio albergo	67,5	62,3	61,4	67,8	77,4	0,09
TII1	Profilo economico turistico	0,21	0,22	0,24	0,17	0,20	0,12
TII1_1	Valore aggiunto	19,6	20,9	21,9	18,2	17,4	0,09
TII1_2	Addetti	342,3	363,0	377,3	327,6	303,4	0,08
TII1_3	Addetti 55.1-55.2	4,46	3,67	7,31	2,32	4,39	0,41
TII1_4	Addetti 55.3-55.4	4,48	5,03	4,63	3,91	4,31	0,09
TII1_5	Addetti 63.3	0,25	0,25	0,18	0,14	0,42	0,43
TII1_6	Addetti 92.3-92.5-92.72.1	0,77	0,63	0,79	0,81	0,83	0,10
TII2	Domanda turistica	0,14	0,16	0,18	0,10	0,11	0,24
TII2_1	Presenze alberghi	8.225	6.388	17.227	4.170	5.158	0,64
TII2_2	Presenze complementari	21.526	19.543	45.608	6.104	14.392	0,69
TII2_3	Presenze alloggi privati	24.097	31.449	24.603	16.898	22.773	0,22
TII2_4	Stagionalità	48,2	40,4	47,5	50,8	54,0	0,10

Nota: C.v. indica il coefficiente di variazione tra le aree geografiche.

7. Un'analisi di raggruppamento

Un'ulteriore analisi classificatoria è rappresentata dalla possibilità di raggruppare le circoscrizioni turistiche in sottoinsiemi il più possibile omogenei al loro interno e, di conseguenza, il più possibile diversi tra loro, sulla base dei 7 indicatori sintetici di primo livello identificati con la procedura proposta. A tali gruppi potrà essere poi assegnata *a posteriori* un'identità logico-intepretativa che ne connoti meglio le peculiarità essenziali.

Tale passaggio logico-operativo ha la duplice finalità di (i) proporre una lettura sintetica dei risultati non più basata sul solo indice di turisticità globale TI, ma sull'analisi congiunta delle graduatorie parziali di turisticità, e (ii) verificare se i risultati siano o meno coerenti con quelli derivati dalla turisticità sintetica derivata dal solo indice TI.

La tecnica di *clustering* prescelta è basata sul *metodo di Ward*, che come noto ha l'obiettivo di identificare la suddivisione in k gruppi che massimizza il rapporto $VAR(B)/VAR(T)$, dove le due varianze sono date, rispettivamente, dalla varianza *Between* (cioè tra *clusters*) e la varianza *Total* (cioè dell'intero collettivo) delle variabili considerate, ossia TAI1, TAI2, TAI3, TAI4, TBI, TII1 e TII2.

Come noto, in simili contesti la questione da dibattere riguarda il numero di gruppi in cui risulta opportuno suddividere l'insieme osservato. Allo scopo sono state provate sei iterazioni della procedura, che hanno identificato via via da 2 a 7 sottogruppi. In particolare, l'indice di Calinski e Harabasz (Fabbris, 1997, pag. 337) – basato sul rapporto tra le devianze *between* e *within* divise per i rispettivi gradi di libertà – conferma pienamente l'esistenza di una struttura gerarchica nei dati, diminuendo progressivamente al crescere del numero dei gruppi[31], fornendo tuttavia un'indicazione non conclusiva circa la numerosità ideale dei gruppi. La conferma della preferibilità della suddivisione in 2 gruppi è fornita però dall'indice F di Beale (*Ibidem*, pag. 336), che nel passaggio tra 2 e 3 gruppi risulta pari a 0,1003, quindi, fortemente non significativo[32], a denotare un insufficiente incremento di varianza *between* al crescere del numero di gruppi.

La segmentazione finale in 2 soli gruppi, di semplice ed immediata interpretabilità, evidenzia chiaramente il *gap* tra circoscrizioni ad elevata competitività turistica (gruppo 1) e le rimanenti, tra le quali non sono state identificate altre segmentazioni statisticamente significative.

E' evidente dalla tabella 7.1 come le 63 circoscrizioni appartenenti al primo gruppo siano complessivamente le migliori con riferimento a ciascuno degli indicatori di turisticità sintetici: solo 5 delle prime 63 circoscrizioni nella graduatoria secondo TI non appartengono al gruppo 1, ed in tale gruppo compare solo una circoscrizione penalizzata dalla graduatoria suddetta, ossia Caserta (al 161° posto secondo TI).

Tuttavia, da un'analisi discriminante a posteriori, sulla cui base è stata verificata la significatività dei 7 indici in analisi al fine di spiegare l'appartenenza delle circoscrizioni ai gruppi 1 o 2, la maggiore significatività spetta nettamente, nell'ordine, agli indici TBI (offerta ricettiva) e TII2 (domanda finale), seguiti da TAI4 (altri attrattori) e TII1 (profilo economico-turistico). Quindi, se da un lato l'insieme degli altri attrattori non è risultato ben rappresentato

[31] Nel dettaglio, l'indice C, passando da 2 a 3 gruppi, scende bruscamente da 177,5 a 55,6, per poi continuare a scendere progressivamente fino al valore di 33,7 corrispondente a 7 gruppi.

[32] La funzione test va confrontata con il valore critico della F di Snedecor con k gradi di libertà al numeratore e $k(n\text{-}g)$ al denominatore, dove k è il numero delle variabili (7), n il numero di unità (547) e g il numero di gruppi. Il valore critico di F è pari a 2,01.

sui primi 7 fattori principali considerati per l'analisi (grafico 5.1), dall'altro la "parte" di essi spiegata dall'analisi fattoriale contribuisce sensibilmente all'identificazione delle circoscrizioni eccellenti.

A conferma di quanto già evidenziato nel paragrafo 6 (tabella 7.2), ben 50 delle 63 circoscrizioni del primo gruppo sono localizzate al Nord (30 nel Nord/est) e solo 2 nel Centro (in cui operano 122 circoscrizioni, quindi più delle 105 del Nord/est). La Valle d'Aosta ed il Trentino Alto Adige primeggiano nettamente per la loro quota relativa di circoscrizioni appartenenti al primo gruppo (rispettivamente, il 50,0% ed il 40.7%), mentre tutte le circoscrizioni di Marche, Lazio, Molise, Basilicata e Calabria appartengono al gruppo 2.

Tabella 7.1 - Valore medio degli indici sintetici di turisticità nei due gruppi e significatività delle variabili discriminanti

Gruppo	Numero	TAI1 (1,370)	TAI2 (0,129)	TAI3 (0,847)	TAI4 (3,417)	TBI (6,574)	TII1 (2,494)	TII2 (5,019)	TAI	TBII	TII	TI
1	63	0,44	0,26	0,09	0,16	0,45	0,39	0,37	0,24	0,45	0,38	0,36
2	484	0,28	0,21	0,06	0,07	0,17	0,18	0,11	0,16	0,17	0,15	0,16
Totale	547	0,30	0,22	0,06	0,08	0,21	0,21	0,14	0,17	0,21	0,17	0,18

Nota: sono riportati in parentesi i valori della t di Student associati ai 7 parametri del modello discriminante che spiega a posteriori l'appartenenza all'uno od all'altro gruppo in funzione dei 7 indici in oggetto. I livelli di significatività della t al 90%, 95%, 97,5%, 99,0% e 99,5% sono pari, rispettivamente, a 1,282, 1,645, 1,960, 2,326 e 2,576. Per inciso, nel modello discriminante è significativa anche la costante e la percentuali dei casi di corretta classificazione è estremamente elevata (98,2%).

Tabella 7.2 - Distribuzione del numero di circoscrizioni turistiche e valori medi dell'indice TI per regione e gruppo

	Numero circoscrizioni			Numero circoscrizioni %			Valore medio di TI		
Regione/area	Gruppo1	Gruppo2	Totale	Gruppo1	Gruppo2	Totale	Gruppo1	Gruppo2	Totale
Piemonte	1	43	44	2,3	97,7	100,0	0,400	0,161	0,166
Valle d'Aosta	7	7	14	50,0	50,0	100,0	0,392	0,179	0,286
Lombardia	10	56	66	15,2	84,8	100,0	0,416	0,173	0,210
Trentino-Alto Adige	22	32	54	40,7	59,3	100,0	0,330	0,177	0,239
Veneto	2	25	27	7,4	92,6	100,0	0,357	0,171	0,185
Friuli-Venezia Giulia	2	13	15	13,3	86,7	100,0	0,395	0,159	0,190
Liguria	2	13	15	13,3	86,7	100,0	0,345	0,168	0,191
Emilia-Romagna	4	35	39	10,3	89,7	100,0	0,309	0,159	0,175
Toscana	1	64	65	1,5	98,5	100,0	0,288	0,151	0,154
Umbria	1	11	12	8,3	91,7	100,0	0,416	0,159	0,180
Marche	-	27	27	-	100,0	100,0	-	0,163	0,163
Lazio	-	20	20	-	100,0	100,0	-	0,151	0,151
Abruzzo	3	23	26	11,5	88,5	100,0	0,423	0,158	0,189
Molise	-	5	5	-	100,0	100,0	-	0,138	0,138
Campania	2	27	29	6,9	93,1	100,0	0,250	0,156	0,162
Puglia	2	20	22	9,1	90,9	100,0	0,349	0,135	0,155
Basilicata	-	5	5	-	100,0	100,0	-	0,142	0,142
Calabria	-	16	16	-	100,0	100,0	-	0,147	0,147
Sicilia	1	31	32	3,1	96,9	100,0	0,344	0,147	0,153
Sardegna	3	11	14	21,4	78,6	100,0	0,344	0,132	0,177
Nord/ovest	20	119	139	14,4	85,6	100,0	0,400	0,168	0,202
Nord/est	30	105	135	22,2	77,8	100,0	0,334	0,167	0,204
Centro	2	122	124	1,6	98,4	100,0	0,352	0,155	0,158
Sud/isole	11	138	149	7,4	92,6	100,0	0,349	0,147	0,162
Totale/Media generale	63	484	547	11,5	88,5	100,0	0,358	0,159	0,182

8. Sintesi e conclusioni

L'impulso ispiratore del lavoro è derivato dalla necessità di valorizzare la componente territoriale nelle basi di dati statistici connesse al turismo. Sebbene le terminologie di *sistema statistico georeferenziato*, di *sistema informativo sul turismo* e di *competitività turistica* siano ormai di uso corrente (soprattutto in citazioni di matrice politico-strategica), non esistono in merito definizioni univoche, né piano organici per un'implementazione metodologica ed operativa di tali concetti.

In tale ottica, il problema primario da affrontare ha riguardato la selezione delle variabili da considerare, valutando contestualmente sia la loro eventuale rilevanza teorica, sia l'effettiva reperibilità ed il relativo livello di qualità, nonché il loro trattamento in chiave statistica.

Con riferimento al problema della scelta delle variabili, la loro selezione ha preso spunto da una riconsiderazione strutturale del concetto di *turisticità* (*tourist* - T) di un dato territorio, che è stato preventivamente scomposto ad un primo livello in tre componenti: l'attrattività turistica potenziale in termini ambientali, storico-artistici, ecc. (*tourist attractiveness* - TA); la disponibilità di posti letto per fini turistici (*tourist bed places* - TB); l'impatto turistico effettivo derivato dalla domanda, ovvero presenze e spesa turistica (*tourist impact* - TI).

Ad un secondo livello di disaggregazione concettuale, due delle tre componenti sono state ulteriormente scomposte: TA in cinque sotto componenti (territorio ed ambiente; infrastrutture; attrattive storiche e naturali; altre attrattive; notorietà), TI in tre (profilo economico turistico; domanda turistica finale; investimenti turistici). Questa segmentazione del concetto latente di "turisticità" di un territorio ha comportato un ulteriore valore aggiunto in termini informativi, ovvero la quantificazione di indicatori via via più specifici per singolo livello di disaggregazione.

La base dati è stata implementata a livello di circoscrizione turistica, che rappresenta un aggregato intermedio tra la provincia ed il comune e la disaggregazione territoriale più fine di cui si disponga del dato sulle presenze turistiche annuali

Si è poi optato per uno sviluppo analitico basato sull'utilizzo di indicatori relativi, ossia *adimensionali* rispetto alla superficie o alla popolazione residente nelle singole circoscrizioni, a fine di consentire confronti spaziali e temporali. In tal modo è anche possibile valorizzare le potenzialità di circoscrizioni altrimenti penalizzate da una dimensione territoriale od economica ridotta.

Uno dei risultati fondamentali dell'applicazione è stata l'identificazione di alcune componenti latenti rispetto alla base dati osservata (composta da 34 variabili), che sintetizzano la maggior parte dell'informazione complessiva in essa contenuta ed identificano profili realmente rilevanti secondo cui analizzare la turisticità di un sito.

Poter disporre di una base dati fortemente territorializzata e tramite cui poter derivare strumenti sia per una lettura descrittiva del territorio, sia per un'analisi sintetica di potenzialità turistica presenta diversi vantaggi. Tra questi:

- si tratta di uno strumento di supporto alle decisioni delle amministrazioni locali, soprattutto nella prospettiva di dover pervenire, da parte di tutte le regioni italiane, alla predisposizione di modelli di aggregazione in termini sistemici delle singole realtà locali, così come predisposto dalla citata legge quadro sul turismo con riguardo ai STL.
- Si possono individuare realtà locali ancora poco toccate dalla domanda turistica finale, ma potenzialmente sviluppabili e quindi possibili destinatarie di piani di investimento locale.

- La predisposizione di un modello teorico completo sia per quanto riguarda la scelta delle variabili, sia con riferimento alla metodologia per la loro analisi e sintesi, consente la replicabilità delle elaborazioni nel tempo e nello spazio e la trasferibilità a realtà locali di variegata dimensione (comprensori, distretti del lavoro, aree di censimento, province, ecc.), sebbene si dovrebbe cercare di operare ad un livello di dettaglio almeno comunale.

Un particolare aspetto - di assoluta rilevanza - certamente da migliorare in prospettiva futura, riguarda la completezza e la qualità dei dati raccolti. Mentre i dati relativi all'offerta ed alla domanda turistica in strutture ufficiali sono molto capillari ed aggiornati, quelli concernenti l'attivazione economica diretta ed indiretta – per quanto dettagliati ed affidabili – non sono sempre aggiornabili annualmente; ben più frammentario è il panorama informativo riguardo agli indici di attrattività TAI, mentre le informazioni relative alle strutture ricettive private ed ai prezzi delle strutture ricettive ufficiali sono attualmente frutto esclusivo di stime.

Bibliografia

Aiello P., Attanasio M., "How to Transform a Batch of Single Indicators to Make Up a Unique One?", *Atti della XLII riunione scientifica della Società Italiana di Statistica (Sessioni plenarie e specializzate)*, pagg. 327-338, Cleup, Padova.

Costa P., Manente M. (2000), *Economia del turismo*, Touring Club Italiano, Milano.

Costa P., Gambuzza M., Manente M., Minghetti V. (1996), "Accessibility and Mobility Conditions and Tourist Development. The Case of Southern Italy", *Quaderni del CISET*, 12/96, CISET, Oriago di Mira.

Cracolici M.F. (2004), "Tourist Performance Evaluation: a Novel Approach", *Atti della XLII riunione scientifica della Società Italiana di Statistica*, pagg. 703-706, Cleup, Padova.

Crouch, G., Ritchie B.J.R. (1999), "Tourism, Competitiveness and Social Prosperity", *Journal of Business Research*, vol.44, 3, pagg.137-152.

De Cantis, Olivieri A.M. (2005), "Le fonti statistiche per l'analisi dei mercati turistici sub-regionali", in Giambalvo O., Parroco A.M. (eds.), *Analisi dei mercati turistici regionali e sub-regionali*, pagg.181-188, Cleup, Padova.

Di Gioia L., Gismondi R., Meccariello I., Morelli P., Russo M.A. (2005), *Dal comune turistico al sistema locale di offerta turistica per la provincia di Foggia*, Franco Angeli, Milano.

ENIT (2004), *Annuario alberghi d'Italia 2003*, Enit, Roma.

Fabbris L. (1997), *Statistica Multivariata – analisi esplorativa dei dati*, McGraw-Hill, Milano.

Gismondi R. (2000), "Per una stima del movimento turistico non rilevato: una proposta di integrazione tra fonti", *Nono rapporto sul turismo italiano*, pagg. 87-104, Mercury, Firenze.

Gismondi R. (2001) "Le *performances* del turismo nelle regioni e nelle province", *Decimo rapporto sul turismo italiano*, pagg. 101-142, Touring Club Italiano, Milano.

Gismondi R. (2005) "La competitività dei siti turistici italiani", *Quattordicesimo rapporto sul turismo italiano*, in corso di stampa, Mercury, Firenze.

Gismondi R., Mirto A.P.M. (2000), "Exhaustive Estimation of Tourist Nights Spent in Italy", *Rivista di statistica ufficiale*, 2, pagg. 33-66, Franco Angeli, Milano, 2002.

Gismondi R., Russo M.A. (2004), "Scelta e sintesi di indicatori per l'identificazione dei comuni turistici", *Atti della XLII riunione scientifica della Società Italiana di Statistica*, pagg. 711-714, Cleup, Padova.

Gismondi R., Russo M.A. (2005), "Definizione e calcolo di un indice territoriale di turisticità: un approccio statistico multivariato", in corso di pubblicazione su *Statistica*, Clueb, Bologna.

Giudici P., Avrini P. (2002), "Modelli statistici per la costruzione di indicatori della qualità della vita: aspetti metodologici", *Rivista di statistica ufficiale*, 1, pagg. 61-80, Franco Angeli, Milano.

Gooroochurn N., Sugiyarto G. (2005), "Competitiveness Indicators in the Travel and Tourism Industry", *Tourism Economics*, 11 (1), pagg. 25-43, Londra.

Greco M.A. (1999), "La georeferenziazione dei siti turistici italiani", in M. Colantoni (ed.) *Turismo: una tappa per la ricerca*, pagg. 345-386, Patron Editore, Bologna.

Kozak M., Rimmington M. (1999), "Measuring Tourist Destination Competitiveness: Conceptual Considerations and Empirical Findings", *International Journal of Hospitality Management*, vol.18, 3, pagg. 273-283.

Istat (2000), *Rapporto annuale 1999*, pagg. 222-232, Istat, Roma.

Landi S. (2003), "I sistemi turistici locali per lo sviluppo di turismo ed ospitalità nel Mezzogiorno", *Rapporto di ricerca Confindustria – Comitato Mezzogiorno*, 50, Roma.

Mercury (2005), *Il turismo italiano negli appartamenti – Primo rapporto 2005*, Mercury, Firenze.

Mihalic T. (2000), "Environmental Management of a Tourist Destination: a Factor of Tourist Competitiveness", *Tourism Management*, vol.21, 1, pagg. 65-78.

Tamma, "Aspetti strategici del destination management", in Pechlaner H., Weiermair K. (eds.), *Destination Management – Fondamenti di marketing e gestione delle destinazioni turistiche*, pagg. 31-54, Touring Club Italiano, Milano, 2001.

Touring Club Italiano (1996*a*), *Guida turistica d'Italia*, Touring Club Italiano, Milano.

Touring Club Italiano (1996*b*), *Italia da scoprire – viaggio nei centri minori*, Touring Club Italiano, Milano.

Touring Club Italiano (2002), *Artigianato, sapori e tradizioni d'Italia*, Touring Club Italiano, Milano.

Touring Club Italiano (Anni vari), *Le guide rosse regionali*, Touring Club Italiano, Milano.

Ufficio Italiano dei Cambi (1998), "The Geography of International Tourism Demand in Italy", paper presentato al *4th International Forum on Tourism Statistics*, Copenhagen.

Norme redazionali

La Rivista di Statistica Ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche Istat corredati, a parte, da una nota informativa dell'Autore contenente: appartenenza ad istituzioni, attività prevalente, qualifica, indirizzo, casella di posta elettronica, recapito telefonico e l'autorizzazione alla pubblicazione firmata dagli Autori. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di un referente scelto tra gli esperti dei diversi temi affrontati. Gli originali, anche se non pubblicati, non si restituiscono.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file Template.doc disponibile on line o su richiesta. In base a tali standard la lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 30–35 pagine.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 12 righe); quelli in italiano dovranno prevedere anche un *Abstract* in inglese. La bibliografia, in ordine alfabetico per autore, deve essere riportata in elenco a parte alla fine dell'articolo. Quando nel testo si fa riferimento ad una pubblicazione citata nell'elenco, si metta in parentesi tonda il nome dell'autore, l'anno di pubblicazione ed eventualmente la pagina citata. Ad esempio (Bianchi, 1987, Rossi, 1988, p. 55). Quando l'autore compare più volte nello stesso anno l'ordine verrà dato dall'aggiunta di una lettera minuscola accanto all'anno di pubblicazione. Ad esempio (Bianchi, 1987a, 1987b).

Nella bibliografia le citazioni di libri e articoli vanno indicate nel seguente modo. Per i libri: cognome dell'autore seguito dall'iniziale in maiuscolo del nome, il titolo in corsivo dell'opera, l'editore, il luogo di edizione e l'anno di pubblicazione. Per gli articoli: dopo l'indicazione dell'autore si riporta il titolo tra virgolette, il titolo completo in corsivo della rivista, il numero del fascicolo e l'anno di pubblicazione. Nei riferimenti bibliografici non si devono usare abbreviazioni.

Nel testo dovrà essere di norma utilizzato il corsivo per le parole in lingua straniera e il corsivo o grassetto per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare il Comitato di redazione delle pubblicazioni scientifiche Istat e per inviare lavori: rivista@istat.it. Oppure scrivere a:

Comitato di redazione delle pubblicazioni scientifiche
C/O Carlo Deli (cadeli@istat.it)
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

€ 10,00