Lavoro e disoccupazione: questioni di misura e di analisi

Progetto di ricerca cofinanziato dal Ministero per l'Università
e la Ricerca Scientifica e Tecnologica - Assegnazione: 1998
Coordinatore: Ugo Trivellato

Un modello per la stima di flussi nel mercato del lavoro affetti da errori di classificazione in rilevazioni retrospettive

Francesca Bassi
Dip. di Scienze Statistiche, Univ. di Padova

Working Paper n. 4 ottobre 1998

Unità locali del progetto:

Dip. di Economia Politica, Univ. Di Modena	(coord. Michele Lalla)
Dip. di Economia "S. Cognetti De Martiis", Univ. di Torino	(coord. Bruno Contini)
Dip. Di Statistica, Univ "Ca' Foscari" di Venezia	(coord. Tommaso Di Fonzo)
Dip. di Metodi Quantitativi, Univ. di Siena	(coord. Achille Lemmi)
Dip. di Scienze Statistiche, Univ. di Padova	(coord. Ugo Trivellato)

Dip. di Scienze Statistiche
via S. Francesco 33, 35121 Padova

1. Introduzione

Le informazioni sulla condizione lavorativa vengono usualmente raccolte mediante indagini longitudinali. I dati longitudinali possono essere ottenuti con diverse strategie; tra le più comuni si distinguono le indagini ripetute nel tempo su uno stesso campione di individui (*panel*) e le indagini retrospettive. Ciascun individuo viene classificato, in base ai dati rilevati, in uno degli stati di Occupato (O), Disoccupato (D) e Non appartenente alla forza lavoro (N), per ogni tempo considerato.

La disponibilità di informazioni longitudinali consente di stimare i flussi, i quali risultano essere uno strumento prezioso per l'analisi del mercato del lavoro. Mentre i saldi infatti misurano variazioni nette negli *stock* di occupati e/o disoccupati in determinati istanti temporali, i flussi permettono di analizzare il mercato del lavoro nella sua evoluzione dinamica. Essi consentono di misurare la proporzione di unità che in istanti di tempo successivi permangono nella medesima condizione lavorativa e che, complementarmente, la mutano. Ancora è possibile, ad esempio, valutare in quale misura un aumento della disoccupazione sia dovuto all'aumento del numero di coloro che, essendo precedentemente occupati, hanno perso il lavoro, oppure all'aumento del numero di individui che entrano nel mercato del lavoro alla ricerca di una occupazione. Eventuali errori di classificazione nella condizione lavorativa causano l'osservazione di flussi distorti e di una dinamica nel mercato del lavoro non conforme alla realtà.

La letteratura classica sugli effetti degli errori di misura (o classificazione) sulla stima dei flussi si basa sull'ipotesi che questi siano indipendenti in istanti di tempo successivi provocando l'osservazione di transizioni spurie con un conseguente aumento della mobilità osservata. Questa assunzione non è realistica in molte situazioni nelle quali la natura del disegno di indagine e le strategie di rilevazione suggeriscono piuttosto che esista correlazione tra gli errori di misura (si vedano, ad esempio, Skinner e Torelli, 1993; Singh e Rao, 1995; van de Pol e Langeheine, 1997). In particolare, in indagini retrospettive, numerose evidenze empiriche mostrano che la correlazione tra gli errori è dovuta ad inaccuratezze nel processo di memoria.

Il presente lavoro propone un modello per la correzione da errori di classificazione dei flussi tra condizioni lavorative rilevate mediante indagini retrospettive. Il modello viene specificato nel contesto dell'analisi a classi latenti, più precisamente nell'ambito di un particolare caso dell'approccio LISREL modificato proposto da Hagenaars (1990), che consente di formulare in modo parsimonioso un modello per la dipendenza degli errori di classificazione dal tempo intercorso tra il momento dell'indagine ed il periodo a cui si riferiscono le informazioni rilevate. È così possibile tenere conto dell'effetto memoria sulla rilevazione della condizione lavorativa osservata, ed in particolare del fatto che più lontano nel tempo è l'evento da ricordare maggiori sono le probabilità di fornire risposte inaccurate.

Il modello proposto è applicato per correggere da errori di classificazione correlati i flussi nel mercato del lavoro francese rilevati mediante l'indagine sulle Forze di Lavoro (*Enquéte sur l'emploi*); un'indagine longitudinale annuale con quesiti retrospettivi, per il periodo da marzo 1990 a marzo 1992. Il nostro interesse a modelli per la correzione di flussi con dati raccolti retrospettivamente è stato inoltre stimolato dal fatto che nella Rilevazione Trimestrale delle Forze di Lavoro (principale fonte di informazione sul mercato del lavoro Italiano), a partire dal 1996, nell'indagine di aprile nell'ambito dell'inchiesta comunitaria, è stato inserito un quesito retrospettivo per conoscere la condizione lavorativa nell'aprile dell'anno precedente di ciascuno degli intervistati.

2. L'indagine Francese sulle Forze di Lavoro

L'Indagine Francese sulle Forze di Lavoro (IFFL) raccoglie informazioni sulla condizione lavorativa dei membri, di età superiore a 15 anni, di un campione di famiglie residenti nel Paese. Per il periodo di interesse di questo lavoro, la rilevazione è stata condotta annualmente con disegno longitudinale ruotato: ogni anno un terzo del campione veniva rinnovato.

Le informazioni sulla partecipazione al mercato del lavoro sono state raccolte mediante quesiti retrospettivi con periodo di riferimento i 13 mesi precedenti l'intervista. Ad ogni rispondente è stato chiesto di riportare la propria condizione lavorativa, riempiendo una griglia riassuntiva nella quale egli si è classificato, per ogni mese del periodo di riferimento, in una delle otto categorie seguenti:

1. lavoratore autonomo;
2. lavoratore dipendente a tempo determinato;
3. lavoratore dipendente a tempo indeterminato;
4. disoccupato;
5. in periodo di formazione;
6. studente;
7. in servizio militare;
8. altro (pensionato, casalinga, ecc.).

Ai fini del nostro studio, si sono aggregate le otto categorie negli usuali tre stati di Occupato, Disoccupato e Non appartenente alla forza lavoro. Più precisamente, si sono considerati *O* coloro che si sono classificati in una qualunque delle prime tre categorie; *D* coloro che si classificati nella quarta categoria; *N* tutti gli altri.

La nostra analisi si basa sulle informazioni raccolte nelle due interviste di marzo 1991 e marzo 1992 su un sottocampione di rispondenti: coloro che avevano un'età compresa tra i 18 e i 29 anni nel 1992 e che hanno risposto a tre interviste consecutive (gennaio 1990, marzo 1991 e marzo 1992). Questo sottocampione, privo dunque di mancate risposte, è composto da 5.427 individui.

Le transizioni mensili tra le condizioni lavorative osservate per il nostro campione, nel periodo da marzo 1990 a marzo 1992, presentano alcune evidenze interessanti, prevalentemente legate alle caratteristiche del campione stesso e precisamente al fatto che si tratta di individui di età giovane, molti dei quali iniziano l'attività lavorativa, o iniziano a cercarla, per la prima volta proprio nel periodo di copertura dell'indagine. In particolare:

a) I flussi osservati mostrano di essere caratterizzati da stagionalità, presumibilmente legata al calendario scolastico. Tra i mesi di giugno e luglio, ad esempio, si osserva un proporzione di individui che entrano nel mercato del lavoro superiore alla media osservata negli altri periodi. Viceversa, tra i mesi di agosto e settembre, una proporzione di unità superiore alla media lascia la condizione di occupato; si tratta probabilmente di studenti che lavorano durante l'estate e a settembre riprendono la scuola.
b) L'andamento della distribuzione marginale delle unità nei tre stati *O*, *U* ed *N* da marzo 1990 a marzo 1992 mostra un progressivo ingresso degli individui del campione nel mercato del lavoro: nel gennaio 1989, il 44% delle unità si dichiara occupato o disoccupato, mentre per marzo 1992, questa proporzione sale al 54%.

I periodi di riferimento delle due interviste consecutive di marzo 1991 e marzo 1992, essendo entrambi costituiti dai 13 mesi precedenti il momento dell'indagine, si sovrappongono per il mese di marzo 1991. La condizione lavorativa dei rispondenti in questo mese è richiesta sia nello stesso marzo 1991, che un anno dopo, nel marzo 1992, con quesito retrospettivo. Per il periodo tra febbraio 1991 ed aprile dello stesso anno, è dunque possibile osservare due coppie di flussi nel mercato del lavoro: una basata su informazioni raccolte con

il medesimo questionario e che denomineremo *within wave* (WW), seguendo una terminologia ormai consolidata in letteratura; l'altra basata su informazioni raccolte a distanza di un anno in due interviste distinte, che denomineremo *between waves* (BW).

La doppia informazione sulla posizione nel mercato del lavoro per il mese di marzo 1991 fornisce alcune prime evidenze sugli errori di risposta che gravano sui dati: l'8% dei rispondenti si classifica in uno stato diverso nelle due interviste. Le due coppie di flussi, WW e BW, forniscono poi evidenze del *pattern* di questi errori di misura: le transizioni WW descrivono un mercato del lavoro più stabile delle corrispondenti transizioni BW (tabella 1), il che può essere considerato come una evidenza, sia pure indiretta, della presenza di errori di classificazione correlati nel tempo.

TABELLA 1
Transizioni mensili osservate (%), febbraio - aprile 1991, distinte per tipo

		OO	OD	ON	DO	DD	DN	NO	ND	NN
F-M	WW	98.19	1.67	0.14	9.11	90.65	0.24	0.28	0.11	99.61
	BW	93.17	3.58	3.25	25.18	65.23	9.59	3.75	1.96	94.29
M-A	WW	98.60	1.04	0.36	8.89	90.37	0.74	0.24	0.29	99.47
	BW	93.24	3.33	3.43	25.90	63.79	10.31	3.79	2.07	94.14

In generale, nelle indagini longitudinali con quesiti retrospettivi ci si aspetta che la causa principale degli errori di risposta sia l'effetto memoria degli intervistati. In particolare, è probabile che si verifichino difficoltà a ricordare correttamente eventi e durate se il periodo di riferimento è piuttosto esteso, come nella IFFL. La dinamica del ricordo durante un'intervista retrospettiva e i possibili esiti che questa può avere in termini di errori negli eventi o nelle durate riportate è ampiamente documentata in numerosi studi di psicologia cognitiva (per una rassegna, vedi O'Muircheartaigh, 1996).

Tra gli esiti principali dell'effetto memoria si ha il fatto che i rispondenti tendono ad omettere di riportare eventi avvenuti nel passato e/o a collocarli in modo inesatto nel tempo. Per quel che riguarda i flussi nel mercato del lavoro, spesso accade che soprattutto episodi brevi non siano riportati e che cambiamenti di stato vengano anticipati o posticipati verso gli estremi del periodo di riferimento, causando distorsione nelle durate osservate. In particolare, si parla di effetto "telescopio in avanti" per indicare la tendenza da parte dei rispondenti a spostare verso il momento dell'intervista eventuali cambiamenti di condizione lavorativa e di effetto di "condizionamento" per indicare la propensione a ripetere meccanicamente la medesima risposta per istanti successivi (mesi ad esempio) all'interno di un periodo di riferimento piuttosto lungo.

Le evidenze sull'andamento della distribuzione marginale nei tre stati e delle transizioni osservate nel nostro campione, insieme alle considerazioni appena svolte sul processo di ricordo delle informazioni richieste, inducono ad ipotizzare innanzitutto, che gli errori siano correlati nel tempo. In secondo luogo, appare plausibile che nella IFFL l'effetto memoria agisca in modo tale da diminuire la probabilità di ottenere risposte corrette all'aumentare della distanza temporale intercorrente tra il momento dell'intervista ed il periodo di riferimento delle informazioni richieste. Oltre al suggerimento di indole generale che viene dalla psicologia cognitiva, secondo il quale più lontani nel tempo sono gli eventi, più difficile risulta ai rispondenti ricordarli correttamente, un'indicazione in tal senso viene anche da analisi empiriche sui flussi osservati con la IFFL (Magnac e Visser, 1995).

3. Un modello per errori di classificazione correlati in indagini retrospettive

L'effetto complessivo degli errori di classificazione correlati sui dati della IFFL è quello di osservare un mercato del lavoro più stabile di quanto non sia in realtà. Appare dunque chiaro che eventuali strategie per la correzione dei flussi che si basino sull'assunzione di errori di classificazione indipendenti (ECI) nei diversi istanti temporali, si rivelerebbero inefficaci. Queste strategie, infatti, ipotizzando che gli errori causino l'osservazione di transizioni spurie e una mobilità tra gli stati superiore a quella reale, correggono necessariamente il mercato del lavoro osservato verso una situazione di maggiore stabilità. La direzione di una tale correzione è in contraddizione con le evidenze sugli errori di misura nella IFFL presentate nel paragrafo 2. Quel che è richiesto è la formulazione di un modello per la correzione dei flussi che tenga esplicitamente in considerazione la dipendenza temporale degli errori di misura. Come già anticipato, il modello proposto, e applicato ai flussi osservati con la IFFL, è specificato nell'ambito dell'analisi a classi latenti. La vera condizione lavorativa di ciascun individuo è trattata come variabile non osservabile (latente), la condizione riportata come suo indicatore.

Il modello si compone di due parti: una componente strutturale, che descrive la dinamica delle vere transizioni nel mercato del lavoro secondo una catena di Markov del primo ordine non stazionaria; una componente di misura, che collega ciascuna variabile latente ai suoi indicatori. Sui parametri del modello di misura si impongono vincoli tali da incorporare le informazioni/ipotesi a priori di cui si dispone sul meccanismo di generazione degli errori di classificazione. In particolare, gli errori vengono espressi in funzione del tempo, in modo tale che la probabilità che vengano commessi aumenti con la distanza intercorrente tra il periodo in cui è accaduto l'evento da riportare e il momento in cui viene condotta l'indagine.

Al fine di specificare e stimare il modello proposto, risulta cruciale sfruttare il legame esistente tra modelli a classi latenti e modelli loglineari, ovvero il fatto che qualunque specificazione nell'ambito dell'analisi a classi latenti può essere espressa in termini di modello loglineare con alcune delle variabili, coinvolte nella struttura di associazione, non direttamente osservabili.

3.1. I modelli a classi latenti e l'approccio LISREL modificato

Le catene di Markov del primo ordine sono uno strumento molto usato per l'analisi dei flussi, in particolare nel mercato del lavoro. I modelli a classi latenti markoviani (CLM), introdotti da Wiggins e Poulsen (Wiggins, 1973, Poulsen, 1982) ed estesi e generalizzati da van de Pol e Langeheine (1990), consentono di considerare l'eventuale presenza di errori di classificazione nelle transizioni osservate e di stimare flussi corretti.

Nei modelli a CLM si assume che il vero stato occupato nel mercato del lavoro sia non osservabile, e quindi trattato alla stregua di una variabile latente, mentre lo stato riportato è considerato come indicatore, affetto da errore, della corrispondente variabile latente. Nella loro formulazione più semplice, i modelli a CLM assumono che le vere transizioni (non osservabili per la presenza di errori di misura negli stati) si comportino secondo una catena di Markov del primo ordine. Inoltre, come in tutte le specificazioni standard di modelli a classi latenti, si introduce l'ipotesi di indipendenza locale tra gli indicatori, ovvero si assume che essi siano indipendenti condizionatamente alle variabili latenti. Nel modello CLM con un indicatore per ciascuna variabile non osservabile, l'assunzione di indipendenza locale coincide con ECI nelle diverse occasioni.

Se consideriamo T istanti temporali, possiamo definire $P(Y_1,\dots Y_T)$ la proporzione di unità osservate in una generica cella della tabella di contingenza a T vie. Y_t ($t=1,\dots T$), dunque,

è la variabile che denota lo stato osservato (nel nostro caso *O*, *D* o *N*) alla t-esima occasione; mentre con y_t ($t=1,...,T$) si indica la corrispondente variabile latente.

Definiamo poi per una generica unità della popolazione:

$q_t^{j_t l_t} = P(Y_t = j_t | y_t = l_t)$, la probabilità di risposta, ovvero la probabilità di osservare al tempo *t* lo stato j_t, mentre lo stato vero nel mercato del lavoro è l_t;

e le probabilità, non osservabili, che definiscono la catena di Markov del primo ordine:

$v_1^{l_1} = P(y_1 = l_1)$, che è la probabilità di occupare lo stato iniziale l_1 e

$r_t^{l_t l_{t-1}} = P(y_t = l_t | y_{t-1} = l_{t-1})$, che è la probabilità di transitare da l_{t-1} a l_t.

Con la notazione appena introdotta, un modello a CLM semplice per *T* occasioni temporali è definito dalla espressione seguente:

$$P(Y_1 = j_1, \ldots Y_T = j_t) = \sum_{l_1=1}^{s} \ldots \sum_{l_t=1}^{s} v_1^{l_1} q_1^{j l_1 l_1} \prod_{t=2}^{T} q_t^{l_t j_t} r_t^{l_{t-1} l_t} \qquad (3.1)$$

Nel nostro caso, j_t e l_t variano su *O*, *D* e *N* e dunque *s=3*.

Un modello a CLM può essere definito anche nell'ambito dei modelli loglineari (Haberman, 1979) e, specificatamente, utilizzando i modelli *modified path* di Goodman (1973). I modelli loglineari classici infatti consentono di ipotizzare solo relazioni simmetriche tra variabili categoriali, non è quindi possibile tenere conto di un eventuale ordinamento di causalità tra di esse. Nell'equazione (3.1) è invece implicito un ordine tra le variabili dato, ad esempio, dall'assunzione di markovianità. I modelli *modified path* sono stati proposti non solo per stimare l'intensità dell'associazione tra un insieme di variabili categoriali, ma anche per tenere conto di eventuali informazioni a priori sulle loro relazioni d'ordine. Ciò si ottiene specificando un sistema di equazioni di tipo logit multinomiale. Ciascuna equazione del sistema considera una variabile dipendente in funzione delle variabili esplicative che la precedono nell'ordinamento causale e viene stimata utilizzando i dati contenuti nella sottotabella marginale ottenuta collassando la tabella di contingenza completa sulle variabili che non influenzano direttamente la variabile dipendente considerata nell'equazione.

L'approccio LISREL modificato di Hagenaars (1990) estende i modelli *modified path* di Goodman a situazioni in cui siano presenti anche variabili latenti.

Per *T*, ad esempio, pari a 4, la (3.1) diventa:

$$P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3, Y_4 = j_4) = \sum_{l_1=1}^{3} \sum_{l_2=1}^{3} \sum_{l_3=1}^{3} \sum_{l_4=1}^{3} v_1^{l_1} q_1^{j_1 l_1} q_2^{j_2 l_2} q_3^{j_3 l_3} q_4^{j l_4 l_4} r_2^{l_2 l_1} r_3^{l_3 l_2} r_4^{l_4 l_3} \qquad (3.2)$$

ed in termini di sistema di equazioni logit, nella notazione usata per modelli gerarchici, in cui si specificano solo le interazioni di ordine superiore tra le variabili:

$$\{y_1 y_2, y_2 y_3\} \qquad (3.3.1)$$

$$\{y_1 y_2 y_3, y_3 y_4\} \qquad (3.3.2)$$

$$\{y_1 y_2 y_3 y_4, y_1 Y_1, y_2 Y_2, y_3 Y_3, y_4 Y_4\} \qquad (3.3.3)$$

Ciascuna equazione del sistema (3.3) è specificata su una sottotabella marginale. L'equazione (3.3.1) sulla tabella definita dalle variabili y_1, y_2 e y_3; la (3.3.2) su quella definita da y_1, y_2, y_3 e y_4; la (3.3.3) sulla tabella di contingenza completa.

Le relazioni d'ordine tra variabili latenti ed indicatori implicate dalla (3.2) e dalla (3.3) corrispondono al diagramma della figura 1, in cui le frecce descrivono effetti causali diretti.

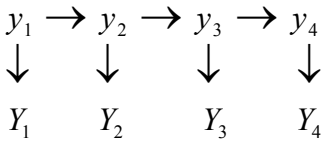

Figura1. Modello a CLM semplice per quattro occasioni temporali

Ciascuna delle probabilità condizionate della (3.2) può essere espressa in funzione dei parametri della equivalente rappresentazione loglineare (3.3). Ad esempio:

$$q_1^{j_1 l_1} = \frac{\exp(\lambda_{l_1}^{y_1} + \lambda_{l_1 j_1}^{y_1 Y_1})}{\sum_{l_1=1}^{3} \exp(\lambda_{l_1}^{y_1} + \lambda_{l_1 j_1}^{y_1 Y_1})}, \tag{3.4}$$

con $\lambda_{l_1}^{y_1}$ e $\lambda_{l_1 j_1}^{y_1 Y_1}$ i parametri di interazione, rispettivamente, del primo e del secondo ordine del modello loglineare (3.3.3).

Dalla (3.4) appare evidente come eventuali vincoli sulle probabilità condizionate possano essere equivalentemente imposti sui parametri della rappresentazione loglineare.

All'interno dell'approccio LISREL modificato, i modelli a CLM semplici possono poi essere agevolmente estesi a comprendere eterogeneità osservata e non (van de Pol e Langeheine, 1990) ed errori di misura correlati nel tempo, mediante l'introduzione di effetti diretti tra gli indicatori (Hagenaars, 1988).

In generale, la rappresentazione dei modelli in termini di prodotto di probabilità condizionate presenta il vantaggio di una maggiore immediatezza interpretativa. Dall'altra parte, l'utilizzo dei modelli loglineari consente specificazioni più parsimoniose. La decomposizione in probabilità condizionate, infatti, implica la stima, per la struttura di associazione tra le variabili coinvolte nella probabilità in questione, di un modello logit (o loglineare) saturo. Viceversa, l'approccio *modified path* di Goodman consente di specificare modelli nei quali alcune interazioni di ordine elevato, considerate superflue, possono essere vincolate a 0.

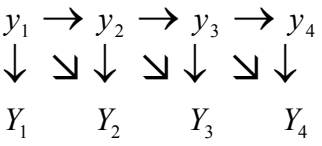

Figura2. Modello a CLM in cui la risposta al tempo t dipende dalla transizione tra t e t-1, per quattro occasioni temporali

Si consideri, a mo' di esempio, un insieme di relazioni casuali tra le variabili leggermente più complicato di quello contenuto nella figura 1. Specificatamente, si ipotizzi che la risposta fornita per l'istante temporale *t* (Y_t) dipenda non solo dallo stato vero in *t* (y_t), ma anche dallo stato per cui si è transitati all'istante *t-1* (y_{t-1}). Il diagramma causale che descrive queste relazioni, per quattro occasioni temporali, è nella figura 2.

In termini di decomposizione in probabilità condizionate, il modello viene definito dalla espressione seguente:

$$P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3, Y_4 = j_4) = \sum_{l_1=1}^{3}\sum_{l_2=1}^{3}\sum_{l_3=1}^{3}\sum_{l_4=1}^{3} v_1^{l_1} q_1^{j_1 l_1} q_2^{j_2 l_1 l_2} q_3^{j_3 l_2 l_3} q_4^{j_4 l_3 l_4} r_2^{l_2 l_1} r_3^{l_3 l_2} r_4^{l_4 l_3} \quad (3.5)$$

dove $q_t^{j_t l_t l_{t-1}}$, per *t=2,3,4*, definisce la probabilità di risposta per l'istante *t*, ovvero la probabilità di riportare lo stato j_t, dato che si è transitati da l_{t-1} e l_t.

La (3.5) implica il seguente sistema di equazioni logit:

$$\{y_1 y_2, y_2 y_3\} \quad (3.6.1)$$
$$\{y_1 y_2 y_3, y_3 y_4\} \quad (3.6.2)$$
$$\{y_1 y_2 y_3 y_4, y_1 Y_1, y_1 y_2 Y_2, y_2 y_3 Y_3, y_3 y_4 Y_4\} \quad (3.6.3)$$

Le interazioni del terzo ordine nella (3.6.3) risultano ridondanti rispetto alle relazioni causali ipotizzate. Nell'ambito dell'approccio LISREL modificato la (3.6.3) può essere sostituita dalla più parsimoniosa espressione:

$$\{y_1 y_2 y_3 y_4, y_1 Y_1, y_1 Y_2, y_2 Y_2, y_2 Y_3, y_3 Y_3, y_3 Y_4, y_4 Y_4\}.$$

3.2. Un modello a classi latenti markoviano per la correzione dei flussi osservati con la IFFL

I modelli a classi latenti markoviani, e le loro estensioni all'interno dell'approccio LISREL modificato, si dimostrano uno strumento particolarmente utile per distinguere il cambiamento osservato in cambiamento reale ed effetto dovuto all'azione degli errori di misura. Il modello LISREL modificato, in generale, è uno strumento estremamente flessibile per descrivere il meccanismo di generazione dei dati raccolti in indagini longitudinali. In particolare, tramite l'imposizione di vincoli sulle probabilità condizionate che costituiscono la sua componente di misura, è possibile incorporare nel modello informazioni a priori sulla natura degli errori, da utilizzare per ottenere flussi corretti.

TABELLA 2
Transizioni trimestrali osservate (%), dicembre 1990 - marzo 1992, distinte per tipo

		OO	OD	ON	DO	DD	DN	NO	ND	NN
D90-M91	WW	94.77	4.25	0.98	24.53	72.40	3.07	0.98	0.66	98.36
	BW	91.50	4.86	3.64	31.60	56.84	11.56	4.40	2.10	93.50
M91-G91	WW	96.03	3.02	0.95	23.21	74.32	2.47	1.28	0.68	98.04
	BW	91.48	4.63	3.89	35.01	54.20	10.79	4.84	2.14	93.02
G91-S91	WW	94.29	3.94	1.77	20.93	78.29	0.78	4.71	2.95	92.34
S91-D91	WW	93.73	4.48	1.79	23.63	74.89	1.48	3.22	1.65	95.13
D91-M92	WW	93.90	4.80	1.30	21.67	76.74	1.59	1.70	0.59	97.71

In un precedente lavoro (Bassi, Torelli e Trivellato, 1998) si è applicata una specificazione *ad hoc* dell'approccio LISREL modificato ai flussi trimestrali osservati con la IFFL e riportati nella tabella 2, allo scopo di correggerli dagli errori di classificazione. Il modello finale contiene una serie di vincoli sulle probabilità condizionate di risposta volte a cogliere il processo di generazione degli errori, in particolare ad incorporare nella procedura di stima restrizioni suggerite dalle considerazioni sul *pattern* degli errori di misura svolte nel paragrafo 2.

In questo lavoro ci si prefigge sempre lo scopo di correggere i flussi trimestrali osservati della tabella 2 mediante un'applicazione opportuna dell'approccio LISREL modificato. Il modello proposto qui, rispetto alle applicazioni precedenti, intende sfruttare la maggiore flessibilità della specificazione in termini loglineari per inserire nella componente di misura una precisa ipotesi sul legame funzionale tra errori di classificazione e lunghezza del periodo di ricordo, così come emerge dalle evidenze sui flussi osservati descritte nel paragrafo 2.

La figura 3 contiene il diagramma causale che descrive le relazioni d'ordine ipotizzate tra le variabili coinvolte. Come è evidente dalla sua ispezione, si assume che i flussi trimestrali per le sei occasioni temporali considerate si comportino secondo una catena di Markov del primo ordine non stazionaria. Per cinque trimestri su sei di dispone di un indicatore per ciascuna variabile latente, mentre per marzo 1991 si hanno due osservazioni indipendenti della condizione lavorativa, dovute al fatto che i periodi di riferimento delle due interviste consecutive del marzo 1991 e del marzo 1992 si sovrappongono. Gli indicatori Y_2, Y_3, Y_4, Y_5 e Y_6 rappresentano lo stato osservato nelle cinque occasioni del periodo di riferimento dell'indagine di marzo 1992 (marzo, giugno, settembre e dicembre 1991 e marzo 1992) mentre, W_1 e W_2 si riferiscono all'intervista precedente (dicembre 1990 e marzo 1991). Si assume, inoltre, che la risposta fornita per l'istante temporale t dipenda dalla vera transizione occorsa tra t e $t+1$, ovvero che vi sia una sorta di effetto telescopio a ritroso; questo è il significato delle frecce oblique nella figura.

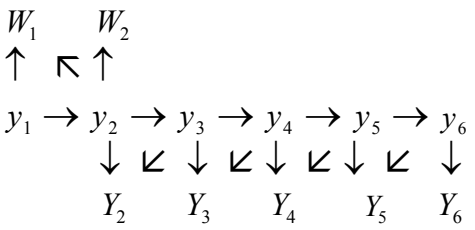

Figura 3. Modello a CLM per sei occasioni temporali

Le relazioni ipotizzate tra le variabili e descritte dal diagramma causale contenuto nella figura 3 possono essere espresse decomponendo la proporzione osservata nella generica cella della tabella di contingenza a sette vie nel seguente prodotto di probabilità condizionate:

$$P(W_1 = k_1, W_2 = k_2, Y_2 = j_2, Y_3 = j_3, Y_4 = j_4, Y_5 = j_5, Y_6 = j_6) =$$
$$= \sum_{l_1=1}^{3}\sum_{l_2=1}^{3}\sum_{l_3=1}^{3}\sum_{l_4=1}^{3}\sum_{l_5=1}^{3}\sum_{l_6=1}^{3} v_1^{l_1} r_2^{l_2 l_1} r_3^{l_3 l_2} r_4^{l_4 l_3} r_5^{l_5 l_4} r_6^{l_6 l_5} q_6^{j_6 l_6} q_5^{j_5 l_5 l_6} q_4^{j_4 l_5 l_5} q_3^{j_3 l_3 l_4} q_2^{j_2 l_2 l_3} z_2^{k_6 l_2} z_1^{j_1 l_1 l_2}$$

dove $z_1^{k_1 l_1 l_2}$ è la probabilità di osservare lo stato k_1 per il mese di dicembre 1990, dato che tra dicembre 1990 e marzo 1991 si è transitati tra lo stato l_1 e lo stato l_2.

In questo lavoro si è preferito invece descrivere le stesse relazioni tra le variabili nell'ambito della più parsimoniosa specificazione loglineare traducendole nel seguente sistema di equazioni logit.

$$\{y_1 y_2, y_2 y_3\} \qquad (3.7.1)$$
$$\{y_1 y_2 y_3, y_3 y_4\} \qquad (3.7.2)$$
$$\{y_1 y_2 y_3 y_4, y_4 y_5\} \qquad (3.7.3)$$

$$\{y_1y_2y_3y_4y_5, y_5y_6\} \quad (3.7.4)$$
$$\{y_1y_2y_3y_4y_5y_6, y_6Y_6, y_6Y_5, y_5Y_5, y_5Y_4, y_4Y_4, y_3Y_3, y_3Y_2, y_2Y_2, y_2W_2, y_2W_1, y_1W_1\} \quad (3.7.5)$$

Si noti che il modello loglineare gerarchico (3.7.5) non contiene i termini di interazione del terzo ordine $\lambda^{y_t y_{t+1} Y_t}_{l_t l_{t+1} j_t}$ (*t=2,3,4,5*) e $\lambda^{y_1 y_2 W_1}_{l_1 l_2 k_1}$, che sarebbero invece implicati da una definizione del modello in termini di probabilità condizionate.

Sui parametri del sistema (3.7) poi si sono introdotti alcuni vincoli al duplice scopo di introdurre ipotesi ed evidenze empiriche sul meccanismo di generazione degli errori di misura (paragrafo 2) e di rendere ancor più parsimonioso il modello stimato con i dati osservati:

1) Per tenere conto del fatto che l'evidenza empirica suggerisce che la probabilità di commettere errori aumenta con la distanza intercorsa tra l'evento da ricordare e il momento dell'intervista, per i parametri di interazione del secondo ordine, $\lambda^{y_t Y_t}_{l_t j_t}$, per *t=2,3,4,5*,6, che descrivono l'associazione tra ciascuna variabile latente ed il suo indicatore, si è specificata la seguente funzione:

$$\lambda^{y_t Y_t}_{l_t j_t} = \mu^{y_t Y_t}_{l_t j_t} + \delta(\mu_{l_t} f(t)) \quad (3.8)$$

dove

$\mu^{y_t Y_t}_{l_t j_t}$ è un parametro che misura l'associazione tra ciascuna variabile latente y_t ed il suo indicatore Y_t, che dipende dalla combinazione stato osservato (j_t) e stato vero (l_t);

δ è una funzione indicatrice che assume valore 1 se $l_t = j_t$, ovvero nel caso in cui lo stato vero nel mercato del lavoro è riportato correttamente, altrimenti, se $j_t \neq l_t$, δ=0;

μ_{l_t} è un fattore di proporzionalità che varia a seconda dello stato vero che stiamo considerando e

è una funzione del tempo *t*, dove *t* indica la distanza dell'evento da ricordare dal momento dell'intervista.

Affinché la probabilità di commettere errori aumenti con la lunghezza del periodo di ricordo, o equivalentemente, la probabilità di riportare correttamente lo stato vero aumenti all'avvicinarsi al momento dell'intervista, *f(t)* deve essere una funzione crescente di *t*. Per il nostro insieme di dati la specificazione *f(t)=exp(t)* ha fornito il migliore adattamento tra un insieme di funzioni alternative tra cui la lineare e la quadratica, in base al confronto dei valori del rapporto di logverosimiglianza L^2.

2) Per rendere più parsimonioso il modello ed assicurarne l'identificabilità; sui parametri sono stati imposti i seguenti ulteriori vincoli:

$\lambda^{Y_2}_{j_2} = \lambda^{Y_3}_{j_3} = \lambda^{Y_4}_{j_4} = \lambda^{Y_5}_{j_5} = \lambda^{Y_6}_{j_6}$;
$\lambda^{y_3 Y_2}_{j_3 l_2} = \lambda^{y_4 Y_3}_{j_4 l_3} = \lambda^{y_5 Y_4}_{j_5 l_4} = \lambda^{y_6 Y_5}_{j_6 l_5}$;
$\mu_{j_1} = \mu_{j_2} = \mu_{j_3} = \mu_{j_4} = \mu_{j_5}$ per j_t=*O,D ed N;*
$\mu^{y_2 Y_2}_{j_2 l_2} = \mu^{y_3 Y_3}_{j_3 l_3} = \mu^{y_4 Y_4}_{j_4 l_4} = \mu^{y_5 Y_5}_{j_5 l_5} = \mu^{y_6 Y_6}_{j_6 l_6}$

che fissano costanti nel tempo i parametri ignoti della (3.8) e l'associazione tra Y_{t-1} e y_t.

3) Inoltre, si è ipotizzato che la probabilità di commettere errori sia la medesima nello stesso mese di anni diversi (dicembre e marzo). Questa assunzione, oltre a rendere il modello ancora più parsimonioso, è anche ragionevole poiché le interviste vengono condotte nello stesso mese di anni consecutivi e l'evidenza empirica mostra che la quantità

di errore commesso è legata alla distanza del momento a cui fa riferimento l'evento dall'intervista, più che al mese di calendario in cui l'evento è accaduto.

Il modello specificato è stato stimato con il programma LEM (Vermunt, 1996) che consente di ottenere stime di Massima Verosimiglianza per i parametri di modelli loglineari e loro estensioni, mediante un'opportuna implementazione dell'algoritmo EM.

3.3. Risultati

I flussi stimati sono contenuti nella tabella 3. Dal confronto con la tabella 2, emerge come le transizioni osservate siano corrette secondo le aspettative sulle conseguenze degli errori di misura in indagini retrospettive. Le transizioni trimestrali riportate per il periodo centrale, da marzo 1990 a dicembre 1991, sono corrette in modo tale che il mercato del lavoro che ne risulta è più dinamico di quello osservato. Il modello specificato stima che non si commettano errori nel riferire lo stato vero nel mercato del lavoro al momento dell'intervista e che l'entità dell'errore di risposta commesso nel ricordare la propria posizione lavorativa nel trimestre precedente sia trascurabile. La conseguenza di ciò è che le transizioni stimate tra dicembre e marzo dei due anni considerati risultano coincidere con quelle osservate. Per gli altri trimestri del periodo di riferimento dell'indagine di marzo 1992, il modello stima che la probabilità di commettere errori nel rispondere aumenti andando a ritroso nel tempo, così come implicato dalla (3.8); l'entità della correzione sui flussi osservati diminuisce dunque avvicinandosi al momento dell'intervista.

TABELLA 3
Transizioni trimestrali stimate (%), dicembre 1990 - marzo 1992

	OO	OD	ON	DO	DD	DN	NO	ND	NN
D90-M91	94.77	4.25	0.98	24.53	72.40	3.07	0.98	0.66	98.36
M91-G91	92.41	4.05	3.54	28.69	62.34	8.97	4.33	2.05	93.62
G91-S91	93.26	4.61	2.13	27.34	70.90	1.76	5.14	3.14	91.37
S91-D91	93.63	4.51	1.86	23.79	73.92	2.29	3.28	1.69	95.03
D91-M92	93.90	4.80	1.30	21.67	76.74	1.59	1.70	0.59	97.71

4. Alcune considerazioni conclusive

In questo studio si correggono i flussi osservati nel mercato del lavoro Francese da errori di classificazione mediante un'opportuna specificazione del modello LISREL modificato. Quest'approccio è stato scelto per la flessibilità da esso consentita nella formulazione di modelli che descrivano relazioni causali tra variabili categoriali, in particolare nella specificazione in termini loglineari, più parsimoniosa della corrispondente decomposizione in probabilità condizionate.

La specificazione del modello come prodotto di probabilità condizionate si presta all'introduzione, mediante vincoli sulle probabilità stesse, di ipotesi a priori sul comportamento dei rispondenti in indagini longitudinali e ad una maggiore facilità di interpretazione del significato dei parametri. Fissando, ad esempio, il valore di alcune probabilità di risposta ad una costante o imponendo che altre probabilità debbano essere uguali tra di loro, è agevole descrivere il meccanismo di generazione dei dati a livello cosiddetto micro, ovvero a livello di comportamento individuale (per una serie di applicazioni di questo approccio, si veda Bassi, Torelli e Trivellato, 1998).

La specificazione in termini loglineari è invece più flessibile ed adatta ad introdurre nel modello ipotesi sui legami funzionali esistenti tra l'insieme di variabili considerate. Per correggere i flussi osservati con un'indagine longitudinale retrospettiva con un periodo di riferimento piuttosto lungo come la IFFL, appare decisivo riuscire a formulare la dipendenza degli errori di risposta dal tempo secondo una precisa relazione matematica come la (3.8) dovuta all'effetto memoria[1].

Il modello proposto corregge le transizioni osservate nel mercato del lavoro Francese nella direzione attesa. Le stime dei parametri sono però fortemente condizionate dalla forma funzionale specifica scelta per la dipendenza delle probabilità di commettere errori dalla lunghezza del periodo di ricordo. Diventa quindi cruciale, per questo tipo di applicazioni, individuare criteri robusti per la valutazione dell'adattamento dei modelli. Questi criteri dovrebbero superare i limiti dei metodi classici impiegati nell'ambito di modelli a classi latenti e loglineari, quali le statistiche X^2 di Pearson e il rapporto di logverosimiglianza L^2, che non sono sempre affidabili in situazioni in cui le tabelle di contingenza a cui si fa riferimento abbiano dimensione elevata e siano sia sparse che sbilanciate, come nel caso di analisi di flussi nel mercato del lavoro.

RIFERIMENTI BIBLIOGRAFICI

F. Bassi, N. Torelli, U. Trivellato (1998), *Data and modelling strategies in estimating labour force gross flows affected by classification errors*, "Survey Methodology", 24, pp.109-122.

L. Goodman (1973), *The analysis of a multidimensional contingency table when some variables are posterior to the others*, "Biometrika", 60, pp.179-192.

S.J. Haberman (1979), *Analysis of Qualitative Data*, Vol.2, Academic Press, New York.

J.A. Hagenaars (1988), *Latent structure models with direct effects between the indicators, local dependence models*, "Sociological Methods and Research", 16, pp.379-405.

J.A. Hagenaars (1990), *Categorical Longitudinal Data: Log-Linear Panel, Trend ad Cohort Analysis*, Sage, Newbury Park.

T. Magnac, M. Visser (1995), *Transition models with measurement errors*, *Working Paper,* Institut National de la Recherche Agronomique (INRA), Paris.

A.C. Singh, J.N.K Rao (1995), *On the adjustment of gross flows estimates for classification errors with application to data from the Canadian Labor Force Survey*, "Journal of the American Statistical Association", 90, pp.1-11.

C. Skinner, N. Torelli (1993), *Measurement error and the estimation of gross flows from longitudinal economic data*, "Statistica", 3, pp.391-405.

[1] Anche il lavoro precedente (Bassi, Torelli e Trivellato, 1998) correggeva i flussi osservati verso un mercato del lavoro più mobile, come atteso dalle considerazioni sulle conseguenze degli errori di misura. Nel modello stimato si era considerata la correlazione tra gli errori di misura introducendo una generica dipendenza delle risposte al tempo t dalla transizione vera sperimentata tra gli istanti t e $t+1$, non si era esplicitamente trattata la dipendenza degli errori di risposta dal tempo intercorso tra intervista ed evento da ricordare.

C. O'Muircheartaigh (1996), *Measurement errors in panel surveys: implications for survey design and for survey instruments*, in "Atti della XXXVIII Riunione Scientifica Società Italiana di Statistica", 1, pp.207-218.

C.S. Poulsen (1982), *Latent Structure Analysis with Choice Modeling Applications*, Ph.D. Dissertation, Wharton School, University of Pennsylvania.

F. Van de pol, R. Langeheine (1990), *Mixed Markov latent class models*, in C.Clogg. (eds.), "Sociological Methodology", Blackwell, New York, pp.213-247.

F. Van de pol, R. Langeheine (1997), *Separating change and measurement error in panel surveys with an application to labour market data*, in L.Lyberg *et al.* (eds.), "Survey Measurement and Process Quality", Wiley, New York, pp.671-688.

J. Vermunt (1997), *Log-linear Event History Analysis: A General Approach with Missing Data, Latent Variables, and Unobserved Heterogeneity*, Sage, Thousand Oaks.

L.M. Wiggins (1973), *Panel Analysis; Latent Probability Models for Attitude and Behavior Change*, Elsevier, Amsterdam.

RIASSUNTO

Si presenta un modello per la correzione dei flussi nel mercato del lavoro da errori di classificazione, quando questi flussi sono osservati mediante indagini longitudinali con quesiti retrospettivi. Il modello consiste in una applicazione dell'analisi a classi latenti, più precisamente, si tratta di una specificazione appropriata del cosiddetto approccio LISREL modificato proposto da Hagenaars. La parametrizzazione in termini loglineari consente di specificare un modello particolarmente parsimonioso e di introdurre nella sua componente di misura una struttura di dipendenza degli errori di risposta dal tempo intercorso tra il momento dell'intervista e l'episodio lavorativo da ricordare. In questo modo è possibile considerare l'effetto memoria sulla condizione lavorativa riportato, ovvero il fatto che, all'allontanarsi nel tempo degli eventi, la probabilità di commettere errori aumenta. Il modello proposto viene applicato per correggere flussi trimestrali osservati con l'indagine Francese sulle forze di lavoro.

SUMMARY

A Model to Estimate Gross Labour Force Flows Affected by Classification Errors in Retrospective Surveys

The paper proposes a model to correct gross flows in the labour market from classification errors, when flows are observed by means of longitudinal surveys with retrospective questions. The model consists in an application of latent class analysis, specifically it is a special case of the modified LISREL approach proposed by Hagenaars. The loglinear parametrization allows one to specify a very parsimonious model and to introduce into its measurement part a dependence of response errors on the time elapsed between the moment of interview and the event to recall. In this way, it is possible to consider the effect of memory decay on reported conditions in the labour market and, specifically, the fact that the longer the

recall period, the higher the probability of making mistakes. The model is applied to correct quarterly gross flows observed by means of the French labour force survey.

Working Papers già pubblicati

1. E. Battistin, A. Gavosto e E. Rettore, *Why do subsidized firms survive longer? An evaluation of a program promoting youth entrepreneurship in Italy*, Agosto 1998.
2. N. Rosati, E. Rettore e G. Masarotto, *A lower bound on asymptotic variance of repeated cross-sections estimators in fixed-effects models*, Agosto 1998.
3. U. Trivellato, *Il monitoraggio della povertà e della sua dinamica: questioni di misura e evidenze empiriche*, Settembre 1998.
4. F. Bassi, *Un modello per la stima di flussi nel mercato del lavoro affetti da errori di classificazione in rilevazioni retrospettive*, Ottobre 1998.

Le richieste di working papers vanno indirizzate alla Sig.ra Patrizia Piacentini, Dipartimento di Scienze Statistiche, via C. Battisti 241-243, 35121 Padova, oppure all'indirizzo e-mail patrizia.piacentini@unipd.it.