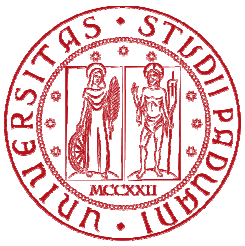


UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN SCIENZE STATISTICHE, ECONOMICHE, FINANZIARIE E AZIENDALI

TESI DI LAUREA

LA MOBILITÀ DEI LAVORATORI PER POSIZIONE PROFESSIONALE E ATTIVITÀ ECONOMICA NELLA RCFL, 2004-2007. ANALISI DELLE INCOERENZE A VARI LIVELLI DI DISAGGREGAZIONE.

THE MOBILITY OF WORKERS BY PROFESSIONAL STATUS AND INDUSTRY IN THE RCFL, 2004-2007. ANALYSIS OF INCONSISTENCIES AT VARIOUS LEVELS OF DISAGGREGATION.

RELATORE: CH.MA PROF.SSA FRANCESCA BASSI

LAUREANDO: PIETRO TIOZZO

ANNO ACCADEMICO 2010-2011

INDICE

INTRODUZIONE

In un periodo in cui il mondo del lavoro presenta situazioni sempre più incerte (il tasso di disoccupazione in Italia a fine 2010 è salito all'8,6%, con quello giovanile che ha superato il 25%, aggiunto ad un numero sempre più elevato di contratti precari) è molto importante svolgere un'analisi longitudinale dell'offerta di lavoro. Monitorare la variazione dello stato occupazionale di un individuo nel tempo, promuovere politiche che favoriscano la crescita del tasso di occupazione, sono alcune possibili azioni che si possono compiere a partire dalle analisi dei dati sul mercato del lavoro. In Italia questi dati provengono da indagini statistiche condotte dall'Istat.

In particolare la Rilevazione Continua sulle Forze di Lavoro (RCFL) è un'importantissima indagine da cui poter ricavare informazioni utili sulla situazione lavorativa nel nostro paese. Essa ha il pregio di fornire dati sezionali (cioè tante unità statistiche intervistate in un determinato trimestre), e offre la possibilità di creare dati di tipo longitudinale (questi stessi soggetti non sono intervistati un'unica volta, ma sono seguiti nel corso del tempo). Quest'ultima caratteristica, in particolare, è di capitale importanza per lo sviluppo di questa tesi, in quanto risulta possibile ricostruire i flussi degli individui nel mercato del lavoro tenendo conto di rilevazioni fatte in diverse occasioni.

La tesi ha come scopo analizzare le incoerenze nelle risposte fornite dagli intervistati nelle diverse occasioni a dodici mesi di distanza rispetto a due variabili chiave che contraddistinguono la loro storia lavorativa: la posizione professionale e la branca di attività economica. È inoltre studiata una terza variabile, detta "Classificazione congiunta dell'attività professionale", costruita a partire dalle risposte fornite alle due variabili chiave.

Il lavoro prende spunto dalla tesi di laurea del 2007 della studentessa Alessandra Padoan, in cui si analizzano le incoerenze di risposta rispetto a queste due variabili a vari livelli di aggregazione, verificando anche se si riscontra un miglioramento significativo man mano che si riduce il numero di

modalità di risposta. il tutto è stato fatto analizzando dati raccolti tra il 1993 e il 2003.

In base alle possibili risposte che il questionario propone in corrispondenza delle domande riguardanti le variabili chiave, l'intervistato sceglie la categoria cui ritiene di appartenere. Essendoci però un alto dettaglio di risposta (13 possibili modalità per la posizione professionale e 12 per l'attività economica), c'è il rischio che questa scelta ad un anno di distanza, pur non essendoci stato nessun cambiamento nella condizione lavorativa, risulti essere diversa: la tesi quindi intende focalizzare il suo interesse sull'incoerenza delle informazioni ricavate dalle matrici di contingenza per lavoratori che hanno affermato di essere continuativamente impiegati in tutto il periodo considerato e che non hanno cambiato lavoro.

Le differenze rispetto al lavoro del 2007 sono però sostanziali: infatti la presente tesi tratta i dati del periodo successivo, dal 2004 al 2007, proprio quando è partita la RCFL, e considera tutti i possibili confronti fra trimestri con distanza temporale pari ad un anno (ad esempio si confrontano le risposte del primo trimestre del 2004 con il primo del 2005, il secondo del 2004 col secondo del 2005 e così via fino al quarto del 2006 con il quarto del 2007), mentre la tesi di Padoan considerava solo il secondo trimestre di ogni anno. Nel periodo studiato da Padoan era inoltre in vigore la RTFL, cioè la Rilevazione Trimestrale delle Forze di Lavoro (la vecchia versione della RCFL, le novità saranno illustrate in seguito). Inoltre, la variabile legata alla posizione professionale con la RCFL è passata da 11 a 13 possibili modalità di risposta. Infine, nel considerare lo studio legato alla variabile "congiunta", si è affrontato anche un nuovo processo di aggregazione proposto da Trivellato *et al*. (2005) in uno studio sul turnover dei lavoratori.

Dopo una presentazione della RCFL, delle tecniche di abbinamento dei record e della definizione del campione, l'analisi viene svolta a partire da panel che confrontano i dati raccolti in due diverse interviste ad un anno di distanza: si vuole valutare le caratteristiche delle incoerenze nelle risposte ai quesiti riguardanti il tipo di attività lavorativa svolta. Viene spontaneo infatti chiedersi se l'alto numero di modalità di risposta possa provocare incertezze

nell'intervistato che potrebbe quindi involontariamente rispondere in maniera diversa. Inoltre, si vuole valutare se l'eventuale aumento di coerenze nelle risposte, che avviene man mano che si aggregano le modalità, è sufficientemente elevato da giustificare la perdita di informazione che avviene per la riduzione di possibili risposte. Questa analisi è svolta attraverso lo studio di una serie di indicatori del grado di accordo delle risposte (in particolare il Kappa di Cohen), attraverso test dipendenti e indipendenti sui Kappa man mano che si effettua il processo di aggregazione, e infine tramite l'uso del modello log-lineare di quasi indipendenza proposto da Goodman (1968).

Dai risultati si nota che, in generale, è più facile rispondere alla domanda relativa alla branca di attività economica rispetto a quella della posizione professionale e della variabile ottenuta congiungendo le modalità di queste. Questa considerazione deriva dal fatto che sia gli indicatori descrittivi sia le analisi fatte con l'indice Kappa di Cohen mostrano un maggior accordo per la branca di attività economica, e anche l'aggregazione delle modalità di risposta per le variabili mostra che per la posizione professionale spesso è necessaria una suddivisione binaria dei lavoratori per raggiungere il più alto e significativo livello di accordo. Infine, il modello di quasi indipendenza mostra (anche al livello di aggregazione più alto possibile per tutte le variabili) che le stime fatte in queste indagini di panel sono affette da errori di misura non casuali.

1-LA DIMENSIONE LONGITUDINALE DELLA RCFL

1.1- La Rilevazione Continua sulle Forze di Lavoro

L'indagine sulle Forze di lavoro è condotta dall'Istat; da essa derivano le stime ufficiali degli occupati, dei disoccupati e delle persone che costituiscono la cosiddetta non forza lavoro. Da questi dati, poi, è possibile calcolare una serie di indicatori che descrivono la situazione del mercato del lavoro, come il tasso di attività, quello di occupazione[1] e di disoccupazione. Questa indagine è condotta in Italia fin dal 1959[2], ma in tutti questi anni è stata modificata molte volte (sia nelle metodologie di rilevazione sia nel contenuto del questionario) per fronteggiare sempre meglio i mutamenti che a loro volta hanno contraddistinto la società e il mondo del lavoro italiano. In particolare, si è assistito ad un fenomeno di invecchiamento della popolazione, a sua volta mitigato da un grande flusso di stranieri verso il nostro paese proprio per motivi legati principalmente al lavoro (oltre ad una notevole mobilità interna, soprattutto dal mezzogiorno verso il settentrione). Da considerare anche l'importanza sempre più elevata nel nostro paese dei posti di lavoro creatisi grazie alla terziarizzazione (a discapito in particolare del settore agricolo) e alla creazione delle più disparate forme di contratto. Infine, si è definito in maniera più dettagliata il concetto di famiglia (chiamata famiglia di "fatto") e dei legami di parentela possibili all'interno della stessa (con la RCFL si è passati da sei a diciassette tipi di legame). Fino a prima del 2004, l'indagine era effettuata con una periodicità trimestrale nei mesi di Gennaio, Aprile, Luglio e Ottobre nella prima settimana del mese che non aveva una festività. Dal 2004 l'indagine ha

[1] Il tasso di attività misura la parte di popolazione che partecipa attivamente al mercato del lavoro. Considera quindi sia gli occupati sia le persone che cercano lavoro. Il tasso di occupazione evidenzia invece la parte di popolazione che lavora. Una crescita del tasso di attività, ad esempio, indica che un maggior numero di persone sono presenti sul mercato del lavoro, a prescindere dal fatto che siano occupate oppure in cerca di lavoro.

[2]Per approfondimenti sull'evoluzione dell'indagine si vedano Di Pietro (1993), Favero e Trivellato (2000), e la documentazione Istat "La nuova rilevazione sulle forze lavoro. Contenuti, metodologie, organizzazione" del 3 Giugno 2004.

assunto carattere continuativo, in linea con quanto stabilito dall'Unione Europea (vedi regolamento n. 577/98), ed ha preso il nome di Rilevazione Continua sulle Forze di Lavoro. Il campione di famiglie viene infatti intervistato lungo tutto l'arco del trimestre di riferimento (quindi in una delle tredici settimane che compongono il trimestre).

Per selezionare le famiglie che fanno parte dell'indagine, in modo simile a come si faceva nella RTFL, la fase di campionamento avviene in due stadi. Nel primo si estraggono i comuni, stratificati entro ciascuna provincia per ampiezza demografica: comuni di tipo A (detti autorappresentativi), cui fanno parte tutti i capoluoghi di provincia e quelli con una popolazione residente superiore ad una soglia prefissata per ciascuna provincia. I restanti comuni fanno parte del gruppo B. Nel campione entrano tutti i comuni di tipo A e un comune di tipo B per ogni strato in cui si è deciso di raggrupparli, selezionato con probabilità proporzionale all'ampiezza demografica. Successivamente, si svolge una seconda stratificazione, di tipo mensile, che consenta a tutti gli strati di essere presenti almeno una volta in ciascun mese del trimestre. I comuni di tipo A di maggiori dimensioni sono presenti ogni settimana, i restanti sono rilevati una volta al mese.

Al secondo stadio, dai comuni estratti, si selezionano casualmente le famiglie. Come definizione di famiglia, l'Istat fornisce la seguente: un nucleo di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti ed aventi dimora nello stesso comune. Sono esclusi i membri permanenti di convivenze (ospizi, istituti religiosi, caserme ecc.), le famiglie residenti che vivono abitualmente all'estero (chi invece è temporaneamente all'estero è considerato, purché ancora iscritto alle anagrafi comunali) e gli stranieri presenti in Italia ma che non hanno la residenza nel nostro paese. Le famiglie sono suddivise in quartine, perché per ogni famiglia estratta se ne estraggono altre tre sostitutive, nel caso la prima esca dal campione per motivi particolari (scomparsa dei componenti o, più frequentemente, motivi legati alla volontà di non rispondere più ai questionari). Questo problema, chiamato attrition, potrebbe portare a risultati inferenziali errati se fosse correlato alle variabili

oggetto di interesse. Nel nostro caso però, poiché le famiglie sostitute sono selezionate casualmente (quindi probabilmente avranno caratteristiche differenti), è difficile pensare che tutte e quattro le famiglie possano essere affette da questo problema.

Le unità del campione (circa 75000 famiglie) seguono uno schema di rotazione del tipo 2-2-2: ogni famiglia è intervistata per due trimestri consecutivi, esce temporaneamente per i due successivi e vi rientra per due ulteriori interviste, dopodiché esce definitivamente dal campione. Le interviste sono quindi effettuate a tre, dodici e quindici mesi di distanza dalla prima. Questo si ottiene grazie alla suddivisione delle famiglie in quattro sezioni di rotazione, che si susseguono come mostrato in tabella 1. La conseguenza principale è che il 50% delle famiglie intervistate a tre e dodici mesi di distanza sono le stesse (al netto delle uscite “forzate” dal campione). Tutto ciò consente di utilizzare i dati campionari per valutare la variazione congiunturale, intesa come la variazione dell'offerta di lavoro rispetto alla rilevazione precedente, e la variazione tendenziale, intesa come la variazione dell'offerta di lavoro rispetto alla rilevazione effettuata nello stesso periodo dell'anno precedente. La tabella riporta un esempio di come si succedono le interviste nei diversi trimestri: il campione intervistato nel primo trimestre del generico anno t è costituito da famiglie che sono intervistate per l’ultima volta, dopodiché escono definitivamente dall’indagine (gruppo A). Il secondo gruppo (gruppo B) è formato da quelle persone che sono intervistate per la terza volta: la seconda intervista risaliva al secondo trimestre dell’anno precedente, e l’ultima sarà svolta il prossimo trimestre. Analogamente per le famiglie del gruppo E (seconda intervista, subito successiva alla prima subita il quarto trimestre dell’anno precedente) e per quelle del gruppo F (costituito dalle famiglie che sono intervistate per la prima volta). In particolare per quest’ultimo gruppo si vede come sono somministrate tutte e quattro le interviste.

TABELLA 1: SCHEMA DI ROTAZIONE DELLE FAMIGLIE NEL CAMPIONE

Sezione di rotaz.	Sequenza di indagini					
	t.I	t.II	t.III	t.IV	T+1.I	t+1.II
A	X					
B	X	X				
C		X	X			
D			X	X		
E	X			X	X	
F	X	X			X	X
G		X	X			X
H			X	X		
I				X	X	
L					X	X

E' così quindi che si possono costruire i dati di panel: per il primo trimestre l'informazione sezionale è fornita dai gruppi A, B, E ed F; quella longitudinale, ad esempio per le famiglie del gruppo F, studiando i trimestri 1, 2 e i primi due dell'anno t+1. I panel quindi permettono al tempo stesso di studiare la dinamica delle scelte di comportamento a livello di unità micro-economica (in particolare per questa tesi, la partecipazione al lavoro) e di controllare, grazie ad esempio tramite modelli econometrici, l'omissione di variabili non osservabili che variano tra le unità ma che sono costanti nel tempo (ad esempio l'abilità) e le fonti di varianza nel fenomeno d'interesse (derivante da fattori permanenti che caratterizzano le unità in esame).

Le interviste sono effettuate da intervistatori di elevato livello professionale che sono selezionati, formati e monitorati dall'Istat (mentre prima della RCFL le interviste dipendevano dai comuni). Questo anche perché molte volte errori nelle indagini erano direttamente imputabili agli intervistatori stessi. I rilevatori eseguono le interviste con l'ausilio del computer, che permette man mano di "costruire" il percorso da far svolgere all'intervistato secondo le risposte che fornisce: questo permette sia una velocizzazione della somministrazione (es. non si chiede da quanto tempo si lavora se in precedenza l'intervistato si è definito disoccupato), sia un maggior controllo (se si danno risposte incoerenti con quanto detto in precedenza, si può

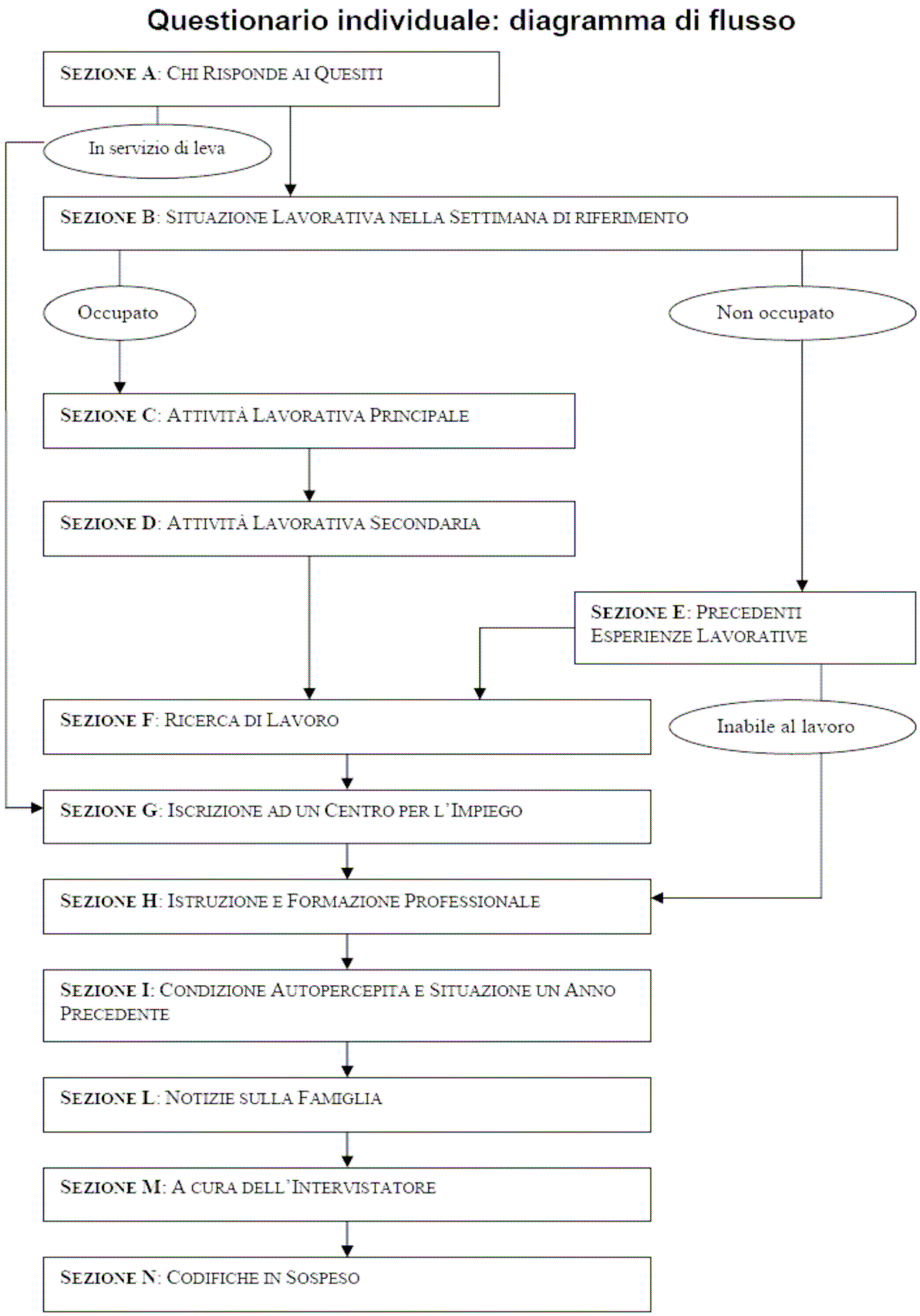
riformulare la domanda). Le domande sono effettuate sia presso l'abitazione delle famiglie (tecnica CAPI, computer assisted personal interviewing) sia telefonicamente (tecnica CATI, computer assisted telephone interviewing). Normalmente, la prima intervista è svolta con la tecnica CAPI (che permette la somministrazione di questionari più lunghi e più complessi, in cui si aiuta il soggetto a comprendere al meglio le domande e lo si stimola a risponderne ad un numero più alto possibile), le successive con quella CATI (che garantisce un certo risparmio, maggiore rapidità, minor resistenza da parte del rispondente che non si sente “invaso” nella sua abitazione, non rende necessaria un’organizzazione capillare del territorio poiché può esserci un unico sistema centralizzato e fa sì che si eviti una certa “interazione” tra intervistante ed intervistato che potrebbe nuocere all’indagine): questa doppia tipologia permette di monitorare meglio il processo di rilevazione attraverso un sofisticato sistema informativo che permette di avere sotto controllo l'andamento della rilevazione e la qualità dei dati raccolti. A questo controllo segue un'accurata fase di verifica a posteriori.

1.2- Il questionario

Il questionario si apre con la scheda generale, rivolta a tutti i componenti della famiglia, in cui si pongono alcune domande (alcune però riservate Istat) relative al sesso, età, provincia di residenza, titolo di studio, relazione di parentela con la persona a cui è intestata la scheda di famiglia, e altre informazioni tutte utili ad identificare in maniera univoca una persona nell'ambito dell'indagine longitudinale (in seguito questo aspetto sarà approfondito nella sezione dedicata alla descrizione di come è stato creato il dataset).

Studiando il questionario, di importanza per la nostra analisi sono le sezioni A (che raccoglie informazioni sulla persona [o sulle persone] della famiglia che risponde ai quesiti), la sezione B (in cui si descrive la situazione lavorativa nella settimana di riferimento), la sezione C (attività lavorativa principale) e tutto l'elenco finale di variabili ricostruite dai curatori dell'indagine (utili anche per verificare come vengono elaborati i dati a partire dalle variabili grezze, cioè proprio quelle cui rispondono le famiglie). Le altre sezioni del questionario di fatto non sono rilevanti per il presente lavoro; tuttavia in seguito è riportato lo schema seguito dagli intervistatori per la somministrazione delle domande.

TABELLA 2: SCHEMA DEL QUESTIONARIO



1.3- Costruzione dei panel tramite abbinamenti

Per poter svolgere indagini panel, è necessario caratterizzare ogni individuo con una precisa chiave identificativa, in modo tale che per i trimestri successivi sia immediatamente possibile riconoscere il record associato ad uno stesso soggetto e poter quindi verificarne le risposte rispetto alle due variabili chiave. Per costruire questa chiave sono disponibili alcune risposte fornite dal soggetto che ne permettono l'identificazione nei diversi trimestri: anno di estrazione, codice della provincia e del comune di residenza, codice della quartina di appartenenza della famiglia, il codice dell'ordine della famiglia nella quartina stessa e il codice identificativo unico e invariato all'interno della famiglia. Quest'ultimo codice è molto importante ai fini dell'abbinamento ed è stato aggiunto solo con la RCFL; prima questa informazione si poteva ricavare solo con algoritmi di tipo probabilistico, mentre con la RCFL la ricostruzione longitudinale avviene con abbinamenti deterministici. L'elenco dettagliato delle domande del questionario, utilizzate per costruire la chiave identificativa, è presente nell'appendice A.

Sono stati creati quindi tutti gli abbinamenti possibili ad un anno di distanza unendo i record che presentano i medesimi valori nella variabile chiave. Teoricamente, il 50% del campione riguardante due interviste svolte a distanza di un anno è abbinabile, ma questa quota non viene mai raggiunta a causa delle uscite dal campione. Queste possono essere dovute a morti, rifiuti, irreperibilità alla seconda intervista, errori di digitazione o vero e proprio cambiamento di valore delle variabili che costituiscono la chiave identificativa (ad esempio se una famiglia cambia comune di residenza, il valore della chiave a sua volta muta e quindi non è più possibile fare l'abbinamento). Di fatto l'abbinamento o la mancata unione tra due record dipende dalla piena o parziale concordanza della variabile identificatrice, tenendo presente che si può andare incontro a due tipi di errori: l'unione di record riguardanti due diverse persone (i falsi positivi) e la mancata unione di due record relativi alla stessa persona (i falsi negativi). Non essendo possibile ridurre al tempo stesso entrambi gli errori, normalmente nelle

indagini panel si tende a diminuire il numero di falsi positivi: essi, infatti, portano a incoerenze nelle variabili d'interesse (per la posizione professionale e il settore di attività economica potrebbero ad esempio sovrastimare il numero di persone che rispondono in maniera errata se in realtà nei due trimestri si considerano due persone diverse), mentre i falsi negativi portano solo a stime meno efficienti (in quanto diminuisce la numerosità campionaria). Per proteggersi quindi dal rischio di falsi positivi, si svolgono ulteriori verifiche sulla bontà degli abbinamenti, in particolare sulle variabili invarianti nel corso del tempo (come il sesso e la data di nascita dell'individuo). Per i dati riguardanti la data di nascita, l'individuo può anche non rispondere (questo accade talvolta quando non è la persona di riferimento a rispondere al questionario), e per questi soggetti non si può stabilire con sicurezza se l'abbinamento sia avvenuto in maniera corretta. Può accadere anche che i dati riguardanti il sesso o la data di nascita nelle due occasioni siano diversi. Una volta definito quindi quali soggetti in particolare entreranno nel campione ai fini dello studio su posizione professionale e settore di attività economica, bisognerà eseguire questi ulteriori controlli sui falsi positivi e i falsi negativi.

1.4- Definizione del campione e delle matrici di contingenza

Una volta effettuati gli abbinamenti dei record riguardanti le stesse persone individuate in indagini svolte a distanza di un anno, è necessario definire le caratteristiche che un soggetto deve avere per fare parte del campione di analisi sulle variabili relative alla posizione professionale e attività economica. Chiaramente il controllo sulla coerenza delle risposte ha senso solo quando si è accertato che l'episodio lavorativo è lo stesso nelle due diverse interviste.

Per prima cosa è necessario eliminare tutti quei soggetti che in almeno una delle due occasioni non sono occupati (in quanto non possono essere oggetto d'interesse, date le variabili da studiare). Per stabilire chi è occupato, si è fatto riferimento alla definizione Istat:

1) OCCUPATO: un individuo risulta occupato se ha più di 15 anni e nella settimana di riferimento dichiara una delle seguenti possibilità:

-aver svolto almeno un'ora in una qualsiasi attività lavorativa che preveda un corrispettivo monetario o in natura;

-aver svolto almeno un'ora di lavoro non retribuito nella ditta di un familiare nella quale collabora abitualmente;

-essere assente dal lavoro (ad esempio per ferie o malattia) con opportune condizioni:

a) la sua assenza non è superiore ai tre mesi, oppure se durante l'assenza continua a percepire almeno il 50% della retribuzione nel caso in cui sia un lavoratore dipendente;

b) nel suo periodo di assenza conserva la sua attività nel caso in cui sia un lavoratore autonomo;

c) la sua assenza non supera i tre mesi nel caso in cui sia un coadiuvante familiare.

2) DISOCCUPATO: un individuo è classificato disoccupato (o in cerca di occupazione) se ha un'età compresa tra i 15 e i 74 anni, non è occupato e dichiara di:
-aver effettuato almeno un'azione attiva di ricerca di lavoro nelle quattro settimane che precedono la settimana di riferimento ed essere disponibile a lavorare (o avviare un'attività autonoma) entro le due settimane successive all'intervista;
-iniziare un lavoro entro tre mesi dalla settimana di riferimento ed essere comunque disponibile a lavorare (o avviare un'attività autonoma) entro le due settimane successive all'intervista, se fosse possibile anticipare l'inizio del lavoro.

3) INATTIVO: un individuo è considerato tale se non appartiene alle forze lavoro, cioè se non è occupato né disoccupato.

L'Istat nel riportare i dati delle interviste presenta anche una variabile, chiamata COND3, con cui afferma a quale di questi tre gruppi appartiene un individuo intervistato. In appendice A si riporta l'elenco delle variabili grezze usate e come sono state trattate per costruire la COND3 a partire dalla definizione di occupato illustrata in precedenza.
In seguito si sono individuati quei soggetti per cui è ragionevole pensare che in entrambe le interviste abbiano risposto in riferimento allo stesso episodio lavorativo. Per stabilire questo, la regola seguita è quella di considerare la data di inizio occupazione dichiarata nella seconda occasione: se essa è antecedente alla data della prima intervista, allora si può concludere che si è di fronte allo stesso episodio lavorativo. Quindi tutti quei soggetti che soddisfano questa condizione, entrano a far parte del campione; sono invece eliminati tutti quei soggetti che dichiarano di aver iniziato il nuovo lavoro (ricordando che sono rimaste solo persone occupate sia alla prima sia alla seconda intervista) dopo la prima intervista. Per tutti quei soggetti che iniziano l'attività lavorativa proprio nel trimestre in cui è svolta la prima intervista, è necessario confrontare la data esatta di inizio lavoro dichiarata

(quindi compreso anche il giorno) con la data dell'intervista stessa: solo se si può stabilire con certezza che quest'ultima è successiva alla prima, il soggetto viene tenuto nel campione. Se invece per colpa di dati mancanti nella data della prima intervista o nella data di inizio lavoro dichiarata nella seconda occasione non è possibile stabilire se i rispondenti avessero o no cambiato lavoro tra le due somministrazioni del questionario, si è deciso che fossero eliminati per evitare ancora una volta i falsi positivi. Questa situazione è comunque molto rara: occorre che una persona abbia trovato lavoro proprio nel trimestre della prima intervista e, normalmente in questi casi, le date vengono ricordate abbastanza facilmente essendo un evento recente. In appendice A sono presenti tutte le variabili usate per fare questo tipo di considerazioni e viene mostrato anche come sono state analizzate.

A questo punto va eseguito l'ultimo controllo sugli abbinamenti tramite lo studio delle variabili invarianti nel tempo (come detto in precedenza sesso e data di nascita dichiarata). Prima considerazione da fare è che per quei soggetti che sono rimasti nel campione sono sempre disponibili per tutti i trimestri i dati riguardanti la data di nascita e il sesso: il campione finale quindi non viene modificato per dati mancanti relativi a queste due variabili. Confrontando invece le risposte errate, dalla tabella 3a si può notare che per tutti e quattro i panel del periodo 2004-2005 i tassi di errore per queste due variabili di controllo sono bassissimi. Nei pochi casi in cui si registra più di un errore, si è andati a verificare altre variabili invarianti (o comunque con un solo possibile tipo di transizione, come il titolo di studio) che risultano spesso congruenti. Tenendo conto che gli errori possono essere dovuti anche ad errate compilazioni da parte dei rilevatori, non sembra essere necessario eliminare questi pochi soggetti: anche se l'abbinamento fosse sbagliato, l'errore praticamente non incide nei risultati per la bassissima frequenza percentuale di persone (dati riscontrabili in tabella 3a).

TABELLA 3a: ERRORI NEGLI ABBINAMENTI PER DATA DI NASCITA E SESSO TRA IL 2004 E IL 2005

TRIMESTRE	***1 ERRORE***		***2 ERRORI***		***3-4 ERRORI***	
	Valori assoluti	Valori percentuali	Valori assoluti	Valori percentuali	Valori assoluti	Valori percentuali
2004-05 1	42	0,17	7	0,025	17	0,08
2004-05 2	61	0,25	6	0,025	14	0,05
2004-05 3	64	0,28	11	0,041	27	0,12
2004-05 4	92	0,39	16	0,061	17	0,06

Diverso il discorso per i due anni successivi, poiché il numero di errori tende a crescere: per quanto riguarda i dati relativi ad un unico errore si nota dalla tabella 3b che la quota percentuale supera sempre l'1% (tranne il primo trimestre) e anche i valori relativi ad almeno tre errori aumentano, arrivando anche allo 0,70%. La quota relativa a due errori, pur aumentando sensibilmente, resta comunque bassa (solo in un caso supera lo 0,28%). Vista la quota percentuale di errori e non sapendo se essi comunque sono dovuti a chi ha compilato il questionario o se sono effettivamente dati da errori nella fase di abbinamento dei record, si è deciso di eliminare quei soggetti che presentavano almeno tre errori, oppure due errori congiuntamente ad errori riguardanti il titolo di studio, nella consapevolezza che comunque il basso numero di soggetti esclusi (non si arriva mai a 200 a fronte di una numerosità campionaria di oltre 21000 persone) in questa fase non va a modificare oltre qualche millesimo i valori degli indicatori che saranno successivamente proposti.

TABELLA 3b: ERRORI NEGLI ABBINAMENTI PER DATA DI NASCITA E SESSO TRA IL 2005 E IL 2007

TRIMESTRE	***1 Errore***		***2 Errori***		***3-4 Errori***	
	Valori assoluti	Valori percentuali	Valori assoluti	Valori percentuali	Valori assoluti	Valori percentuali
2005-06 1	207	0,86	66	0,27	137	0,57
2005-06 2	272	1,14	59	0,25	165	0,69
2005-06 3	226	1,01	50	0,22	150	0,67
2005-06 4	278	1,25	53	0,24	140	0,63
2006-07 1	289	1,31	63	0,28	155	0,70
2006-07 2	336	1,52	82	0,37	155	0,70
2006-07 3	220	1,03	59	0,28	100	0,47
2006-07 4	272	1,24	43	0,20	105	0,48

Individuati i record per cui si può assumere con un buon margine di sicurezza che si stanno studiando risposte relative allo stesso impiego, si procede all'analisi delle incoerenze riguardanti la posizione professionale, l'attività economica e la variabile congiunta dichiarata a distanza di un anno. Sempre in appendice A viene illustrato il modo con cui dalle variabili grezze somministrate nel corso del questionario si è arrivati a ricostruire le variabili POSPRO e CAT12, che sono usate in questa tesi per calcolare gli indicatori proposti.

In base alle risposte date nelle due diverse occasioni, si costruisce la tabella di frequenza a doppia entrata, mostrando in maniera chiara le incoerenze: infatti, tutti i valori al di fuori della diagonale principale rappresentano il numero di persone che, data una particolare risposta al tempo t, rispondono al tempo t+1 con una modalità differente.

TABELLA 4: TABELLA A DOPPIA ENTRATA DELLE FREQUENZE PER UNA GENERICA VARIABILE A 12 MODALITA' DI RISPOSTA

	Mod. 1	*Mod. 2*	*Mod. 3*	*Mod. 4*	*Mod. 5*	*Mod. 6*	*Mod. 7*	*Mod. 8*	*Mod. 9*	*Mod. 10*	*Mod. 11*	*Mod. 12*
Modalità 1	X											
Modalità 2		X										
Modalità 3			X									
Modalità 4				X								
Modalità 5					X							
Modalità 6						X						
Modalità 7							X					
Modalità 8								X				
Modalità 9									X			
Modalità 10										X		
Modalità 11											X	
Modalità 12												X

La tabella mostra come viene costruita la matrice delle risposte al tempo t (catalogate per riga) e al tempo t+1 (catalogate per colonna). I valori X, della diagonale principale, mostrano il numero di persone che dichiarano la stessa caratteristica di impiego ad un anno di distanza.

1.5- Gli errori di misura

La letteratura è unanime nel sottolineare la gravità degli errori di misura (di trascrizione o di codifica) in contesti di dati longitudinali: essi infatti tendono a compensarsi in media nella stima degli stock, ma si possono accumulare inficiando la correttezza delle stime dei flussi.[3] La letteratura indica in particolare tre cause di errori di misura:

1) natura socio-psicologica;
2) disegno dell'indagine e procedura di abbinamento;
3) struttura del questionario e processo di memoria.

Per quanto riguarda la prima causa, la risposta ad una domanda comporta una serie di azioni che, in alcune particolari situazioni, si rivelano essere non così scontate come invece potrebbero sembrare in apparenza: è causa di distorsione di natura socio-psicologica il condizionamento sociale e da reintervista. Il primo tipo di condizionamento è legato alla difficoltà nel rilasciare risposte a domande ritenute socialmente sensibili: si sottostima volontariamente un periodo di disoccupazione, o a celare un'attività "umile" con una simile ma ritenuta più "gratificante". Il secondo tipo di condizionamento si può verificare in persone intervistate più volte, che a domande uguali tendono a dare le stesse identiche risposte della prima intervista per "inerzia" o poco impegno, anche se in realtà si sono verificati dei cambiamenti di stato. Se questi errori non vengono rilevati, possono essere motivo di erronea mobilità (sovrastimandola o in altri casi sottostimandola).
La seconda fonte di errore è data dai falsi positivi: l'abbinamento di record riguardanti persone in realtà diverse può generare mobilità spuria. E' per

[3] Si riporta un esempio tratto da Moriani (1981): due individui, uno permanentemente occupato e l'altro permanentemente disoccupato, vengono intervistati entrambi due volte. Se alla seconda intervista entrambi commettono un errore di classificazione dello stato occupazionale, invertendo i rispettivi ruoli, le stime di stock non ne soffriranno perché gli errori si compenseranno a si continuerà a registrare, nelle due date, lo stesso numero di occupati e disoccupati. Invece le stime dei flussi registreranno due movimenti che in realtà non sono mai avvenuti.

questo che il metodo di creazione del campione è sempre finalizzato a minimizzare i falsi positivi.

L'ultima causa è legata alla struttura del questionario e al processo di memoria indotto dalle domande retrospettive: esse possono risultare oscure, mal formulate o male interpretate, o ancora fornite involontariamente sbagliate da un soggetto che risponde per un'altra persona. C'è da dire comunque che, con il passaggio alla RCFL e alla formazione di intervistatori direttamente a carico dell'Istat, congiuntamente al fatto che è aumentata anche l'"assistenza" al rispondente, questo pericolo si è sensibilmente ridotto.

Gli errori di tipo retrospettivo sono causati da ricordi sbagliati (la persona può involontariamente dimenticare alcuni eventi, come un periodo di occupazione o di disoccupazione molto breve), da errata collocazione temporale di un evento (si ricorda cioè un evento in una data più vicina a quella dell'indagine piuttosto che la data effettiva, o viceversa si ricorda un evento in una data più remota rispetto a quella vera), o ancora dal cosiddetto effetto "ammucchiamento", che genera la collocazione di un evento in particolari date (ad esempio Gennaio è il mese normalmente indicato come quello di inizio di un'attività e Dicembre come quello della fine).

Ulteriore fonte di errore può essere l'eccessivo numero di modalità di risposta per alcune variabili, che porta un soggetto a cambiare risposta (ad esempio una molto simile) nelle varie interviste, pur non essendosi verificato nessun cambiamento di stato. In particolare per la variabile relativa all'attività economica, la codifica può risultare molto difficile, essendoci più di 500 codici tra cui scegliere: sbagliare involontariamente risposta fornendone una molto simile porta ad un aumento in realtà fittizio della mobilità (e questi errori in particolare sono l'oggetto della tesi).

2-ANALISI DESCRITTIVE DEI DATI

Le prime tecniche di analisi descrittive dei dati consistono nel calcolo del tasso di risposte coerenti, del tasso di differenza netta e del tasso di incoerenza.

2.1-Posizione professionale

La tabella 5a mostra la percentuale di risposte coerenti per la variabile posizione professionale nei panel che considerano i trimestri dal 2004 al 2007.

Le 13 possibili modalità di risposta per la posizione professionale sono:

1. Dipendente: Dirigente
2. Dipendente: Quadro
3. Dipendente: Impiegato
4. Dipendente: Operaio
5. Dipendente: Apprendista
6. Dipendente: Lavoratore presso il proprio domicilio per conto di un'impresa
7. Autonomo: Imprenditore
8. Autonomo: Libero professionista
9. Autonomo: Lavoratore in proprio
10. Autonomo: Socio di cooperativa
11. Autonomo: Coadiuvante nell'azienda di un familiare
12. Autonomo: Collaborazione coordinata e continuativa
13. Autonomo: Prestazione d'opera occasionale

Il primo tipo di indicatore è dato dalla semplice proporzione di persone del campione che, a distanza di un anno, rispondono alle domande riguardanti le variabili chiave con la stessa identica modalità. Rispetto alla matrice mostrata in tabella 4, si tratta quindi di calcolare la quantità di persone presenti sulla diagonale principale.

TABELLA 5a: PERCENTUALE DI RISPOSTE COERENTI PER LA POSIZIONE PROFESSIONALE

TRIMESTRE	*NUMEROSITA'*	*% RISPOSTE COERENTI*
2004-05 1	24029	94,87%
2004-05 2	23680	95,38%
2004-05 3	23434	95,15%
2004-05 4	25192	95,31%
2005-06 1	24176	93,43%
2005-06 2	23918	93,43%
2005-06 3	22422	93,53%
2005-06 4	22289	92,82%
2006-07 1	22059	91,02%
2006-07 2	22122	91,26%
2006-07 3	21419	93,33%
2006-07 4	21981	93,54%

Come si può notare, tutti i valori sono compresi tra il 91 e il 95,5%. Il maggior numero di risposte coerenti si ha in generale il primo anno, con valori più bassi nei primi due trimestri del terzo anno. Interessante anche notare che è difficile individuare un trend nel corso del tempo.

Il secondo indicatore descrittivo è il tasso di differenza netta, che mostra l'errore commesso per ogni tipologia di risposta, espresso come differenza tra le proporzioni marginali osservate nelle due occasioni:

$$e_i = \left(\frac{X_{.i} - X_{i.}}{X_{..}} \right) * 100 \qquad i=1,2,3,\ldots$$

dove

$X_{.i}$ rappresenta il totale delle persone che al tempo t+1 rispondono alla domanda relativa alla posizione professionale con la modalità i-esima, analogo del valore $X_{i.}$ per il tempo t

$X_{..}$ rappresenta il numero totale di soggetti che compongono il campione. Questo indicatore può assumere valori compresi tra -100 e 100: man mano che si avvicina a uno di questi due valori, ci si trova in una situazione in cui al tempo t+1 (nel caso il valore sia 100) quasi tutte le persone si identificano con l'i-esima modalità, a differenza del tempo t in cui quasi nessuno aveva fornito questa risposta (il contrario per -100). Più è vicino a zero, più è alto

il numero di persone che nelle due interviste hanno risposto con la stessa modalità. Questo ovviamente non significa che non ci sia incoerenza: potrebbe accadere, ad esempio, che dieci persone al tempo t dicano di essere quadri, e al tempo t+1 impiegati, ma altre dieci che avevano risposto impiegati, l'anno successivo rispondano quadri. Per entrambe le modalità, l'indicatore assume valore zero, ma tutte le persone hanno risposto in maniera incoerente.

Nella tabella 5b vengono riportati i risultati:

TABELLA 5b: TASSO DI DIFFERENZA NETTA PER MODALITA DELLA VARIABILE POSIZIONE PROFESSIONALE

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Dirigente	-0,41	-0,41	-0,42	-0,31	0,21	0,10	0,15	0,06	-0,17	0,09	0,00	-0,05
Quadro	0,53	0,44	0,41	0,29	0,09	0,02	0,09	0,02	0,05	-0,12	-0,08	0,04
Impiegato	0,28	0,08	0,25	0,30	-0,22	0,28	-0,23	-0,12	0,33	0,25	0,15	0,10
Operaio	-0,05	0,19	0,41	0,19	0,09	-0,13	-0,07	0,02	-0,26	0,06	-0,13	-0,22
Apprendista	-0,17	-0,14	-0,06	-0,03	-0,08	-0,10	0,01	-0,07	-0,07	-0,14	-0,13	-0,11
Lavorante a domicilio	0,01	0,01	0,00	0,00	-0,01	0,00	0,00	0,00	-0,01	0,00	-0,02	0,00
Imprenditore	-0,16	-0,11	-0,07	-0,15	-0,14	-0,16	-0,26	-0,22	-0,10	-0,11	-0,12	0,01
Libero professionista	0,01	-0,01	-0,04	0,02	-0,07	-0,06	-0,06	-0,05	0,01	-0,02	-0,03	0,06
Lavorante in proprio	0,22	0,31	0,09	0,21	0,27	0,18	0,33	0,31	0,26	0,23	0,31	0,10
Socio di cooperativa	-0,02	-0,10	-0,06	-0,05	-0,03	-0,03	-0,06	-0,04	0,00	0,03	-0,02	0,01
Coadiuvante familiare	-0,22	-0,27	-0,44	-0,43	-0,16	-0,06	0,05	0,00	0,08	-0,12	0,08	0,07
Co. co. co.	0,01	0,01	-0,03	-0,02	0,07	-0,03	0,04	0,13	-0,10	-0,13	-0,06	-0,02
Prestazione occasionale	-0,03	0,00	-0,05	-0,04	-0,02	0,00	0,00	-0,04	-0,02	-0,03	0,03	0,01

Come si vede dalla tabella, i valori sono tutti molto bassi, in particolare non si vai mai oltre lo 0,5% in valore assoluto. I valori più distanti da zero sono assunti dalla modalità dirigente e da quella quadro. Inoltre si osservano valori leggermente inferiori per i lavori di tipo autonomo.

Il terzo indicatore proposto è il tasso di incoerenza, dato dal rapporto tra le risposte incoerenti osservate e quelle attese in ipotesi di casualità ed indipendenza per una certa modalità di risposta.

L'indicatore è espresso come:

$$I_i = \frac{X_{.i} + X_{i..} - 2X_{ii}}{\frac{1}{X_{..}}\left[X_{.i}\left(X_{..} - X_{i.}\right) + X_{i.}\left(X_{..} - X_{.i}\right)\right]} * 100 \qquad i=1,2,3,\ldots$$

dove

X_{ii} rappresenta la frequenza osservata nella cella della diagonale principale di posto i. Questo tasso assume valori compresi tra 0 (assenza di risposte incoerenti) e 100 (totalità di risposte casuali).

TABELLA 5c: TASSO DI INCOERENZA PER MODALITA' DELLA VARIABILE POSZIONE PROFESSIONALE

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Dirigente	14,82	13,63	14,34	12,54	11,38	9,40	11,02	9,94	16,45	14,61	12,49	8,17
Quadro	10,09	9,94	10,37	10,13	11,79	12,30	11,39	11,65	17,94	16,42	12,79	11,71
Impiegato	5,71	5,08	5,69	5,41	7,96	8,24	8,18	8,81	11,83	10,97	7,82	7,99
Operaio	4,80	4,17	4,51	3,85	6,78	6,34	6,53	7,46	8,91	8,55	6,62	6,79
Apprendista	14,48	14,05	11,77	13,93	24,24	22,29	21,10	27,64	23,58	26,70	26,43	20,39
Lavorante a domicilio	11,77	5,56	15,16	0,00	14,29	0,00	15,80	20,01	50,02	42,87	29,42	31,05
Imprenditore	8,95	8,15	7,10	8,53	12,50	14,18	17,15	17,29	16,21	19,44	17,94	14,14
Libero professionista	5,71	5,22	4,20	4,91	6,83	6,73	6,60	7,35	10,63	11,11	8,34	7,86
Lavorante in proprio	4,39	3,62	3,29	4,03	5,81	6,10	5,74	6,79	8,06	8,19	6,32	6,05
Socio di cooperativa	50,55	63,94	65,92	59,39	74,34	65,31	74,76	73,09	53,34	72,72	67,54	58,36
Coadiuvante familiare	16,98	16,85	19,00	18,84	24,88	25,05	25,24	28,83	29,16	23,49	19,35	21,86
Co. co. co.	21,28	19,06	22,36	20,44	20,80	23,19	20,55	27,57	27,81	31,24	24,80	26,54
Prestazione occasion	26,36	23,76	26,38	28,21	41,75	40,98	38,73	50,74	41,40	41,39	38,09	46,34

Da questa tabella si osserva che il gruppo di lavoratori autonomi ha un tasso di incoerenza leggermente più basso. Particolarmente elevato risulta quello dei soci di cooperativa (che avevano un tasso di incoerenza molto basso) che arriva quasi al 75%: questo cambiamento avviene sia per una serie di errori di misura descritti in precedenza, sia per il numero molto basso di soggetti che rispondono alla domanda sulla posizione professionale proprio con questa modalità: anche piccole variazioni marginali delle frequenze osservate possono quindi generare grandi variazioni percentuali per questo tasso.

2.2-Attività economica

La seconda variabile che viene presa in considerazione è quella legata all'attività economica dei soggetti. Le 12 possibili modalità di risposta per questa variabile sono:

1. Agricoltura, caccia e pesca
2. Industria: energia/estrattiva
3. Industria delle trasformazioni
4. Industria delle costruzioni
5. Altre attività: commercio
6. Altre attività: alberghi e ristoranti
7. Altre attività: trasporti e comunicazioni
8. Altre attività: intermediazione monetaria, finanziaria e attività immobiliari
9. Altre attività: servizi alle imprese e altre attività professionali ed imprenditoriali
10. Altre attività: pubblica amministrazione, difesa ed assicurazioni sociali obbligatorie
11. Altre attività: istruzione, sanità ed altri servizi sociali
12. Altre attività: altri servizi pubblici, sociali e alle persone

TABELLA 6a: PERCENTUALE DI RISPOSTE COERENTI PER L'ATTIVITÁ ECONOMICA

TRIMESTRE	*NUMEROSITA'*	*% RISPOSTE COERENTI*
2004-05 1	24029	89,06%
2004-05 2	23680	92,89%
2004-05 3	23434	92,92%
2004-05 4	25192	97,59%
2005-06 1	24176	95,82%
2005-06 2	23918	95,74%
2005-06 3	22422	95,97%
2005-06 4	22289	95,74%
2006-07 1	22059	93,24%
2006-07 2	22122	93,89%
2006-07 3	21419	95,61%
2006-07 4	21981	95,85%

Rispetto ai dati relativi alla posizione professionale, si registra un aumento del range: il valore più basso si assesta all'89% e quello più alto oltre il 97%. Inoltre dal quarto trimestre del primo anno (quindi nove volte su dodici) le risposte a questa variabile sono costantemente più coerenti (cosa che già si registrava nel periodo studiato da Padoan). Anche per l'attività economica è difficile accorgersi di un eventuale trend nelle percentuali, anche se i risultati degli ultimi trimestri sono sicuramente più alti rispetto a quelli dei primi. Già da questi indicatori descrittivi, come anticipato in introduzione, sembra che in generale sia più facile definire la branca entro cui rientra la propria attività economica piuttosto che la posizione professionale che si svolge.

TABELLA 6b: TASSO DI DIFFERENZA NETTA PER MODALITA' DELLA BRANCA DI ATTIVITA' ECONOMICA

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Agricoltura, caccia e pesca	-0,05	-0,07	-0,09	-0,07	0,03	-0,11	-0,07	-0,03	0,09	0,20	-0,01	0,00
Energia/estrattiva	-0,03	0,02	0,00	0,00	0,00	-0,01	-0,01	-0,03	0,03	-0,03	0,02	0,04
Trasformazioni	0,29	0,14	0,18	0,06	0,02	-0,01	-0,05	0,03	0,12	0,28	-0,17	0,06
Costruzioni	0,10	-0,04	0,06	0,01	-0,09	0,00	-0,03	-0,02	0,15	-0,01	0,00	-0,05
Commercio	-0,39	-0,24	-0,28	-0,01	-0,02	0,01	0,05	-0,04	-0,24	-0,29	0,00	-0,05
Alberghi e ristoranti	-0,04	-0,03	0,09	0,04	0,06	0,10	-0,02	0,01	-0,02	-0,05	-0,03	0,03
Trasporti e comunicazioni	-0,14	0,04	0,14	-0,02	-0,01	0,03	0,08	-0,02	0,00	0,00	-0,01	0,03
Intermediaz. f.,m., e att. Immobiliari	0,00	0,02	-0,03	0,04	0,05	0,05	0,00	0,00	-0,04	0,02	-0,02	0,00
Servizi ad imprese	0,41	0,14	-0,03	0,01	-0,05	-0,11	-0,06	-0,06	-0,08	-0,05	0,03	0,01
Pubblica amministrazione	-0,16	-0,11	0,01	-0,05	0,06	0,04	0,01	0,00	0,03	0,04	0,01	0,02
Istruzione, sanità, serv. sociali	0,04	0,07	-0,11	-0,03	-0,03	0,03	0,08	0,14	0,04	-0,04	0,08	0,04
Altri servizi pubblici	-0,05	0,07	0,06	0,00	-0,03	-0,02	0,02	0,01	-0,07	-0,07	0,09	-0,13

Anche dalla tabella 6b sembra che ci sia maggior accordo nelle risposte date all'attività economica (pur ricordando che valori vicini allo zero di questo tasso sono segnali di buona somiglianza delle distribuzioni marginali, ma questo non basta ad assicurare assenza di errori). Se con la posizione professionale infatti i valori erano compresi tra -0,42% e lo 0,53%, ora i valori estremi sono -0,41% e 0,39%. Il settore che assicura una maggior facilità di risposta sembra essere quello primario (anche perché per questo

c'è solo una modalità di risposta), mentre quello che presenta più difficoltà è il terziario (che fra l'altro ha più opzioni di risposta possibili).

TABELLA 6c: TASSO DI INCOERENZA PER MODALITA' DELLA BRANCA DI ATTIVITA' ECONOMICA

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Agricoltura, caccia e pesca	11,82	7,88	6,74	3,29	5,17	4,68	4,15	4,80	8,08	7,84	5,21	4,65
Energia/estrattiva	20,64	12,42	15,88	3,31	8,47	5,97	5,41	7,76	10,78	8,87	6,42	4,90
Trasformazioni	14,09	9,24	8,55	2,43	5,17	4,88	4,66	4,91	8,76	8,30	5,25	5,10
Costruzioni	13,37	7,74	8,29	2,22	5,04	4,68	4,91	4,61	7,71	6,61	5,25	4,90
Commercio	14,33	9,07	8,35	2,47	5,15	5,68	5,48	5,84	9,50	8,46	5,46	4,89
Alberghi e ristoranti	8,43	5,13	6,30	2,67	4,16	4,84	4,27	4,66	6,28	5,31	4,26	4,37
Trasporti e comunicazioni	13,42	7,54	8,10	2,79	4,55	4,81	3,95	4,73	6,98	6,33	5,18	4,46
Intermediaz. f.,m., e att. Immobiliari	6,05	4,97	4,80	2,32	3,09	3,01	3,07	3,22	4,78	3,79	3,31	2,73
Servizi ad imprese	19,54	13,38	14,06	4,73	6,86	7,41	7,26	6,86	10,39	9,56	7,24	7,35
Pubblica amministrazione	10,58	6,94	7,76	2,98	3,88	4,28	3,95	3,72	6,35	5,73	4,56	4,53
Istruzione, sanità, serv. sociali	5,35	3,87	3,39	1,64	2,32	2,45	2,13	2,67	3,96	3,23	2,77	2,54
Altri servizi pubblici	19,13	11,88	11,93	4,72	7,57	7,22	7,04	8,35	10,14	9,46	6,45	6,48

Dai valori del tasso di incoerenza (tabella 6c) si conferma migliore l'adattamento delle risposte alla seconda variabile considerata: infatti non si raggiungono picchi come quelli della posizione professionale e in generale i risultati di questo tasso sono più bassi. Inoltre, a differenza di quanto avviene per la posizione professionale, il tasso tende a migliorare col passare del tempo, segno di un miglioramento della somministrazione e apprendimento delle domande (cosa che si notava già confrontando la percentuale di risposte coerenti delle tabelle 5a e 6a).

2.3-Classificazione congiunta dell'attività professionale

La terza variabile che viene analizzata è ottenuta come combinazione delle informazioni ricavate dalle due variabili chiave considerate fino a questo momento, raggruppando i lavoratori autonomi in un'unica modalità di risposta, e ripartendo tutti i dipendenti nei 12 settori dell'attività economica. Questa nuova variabile dunque è costituita da 13 possibili risposte:

1. Autonomi
2. Dipendenti nell'agricoltura, caccia e pesca
3. Dipendenti nell'industria: energia/estrattiva
4. Dipendenti nell'industria delle trasformazioni
5. Dipendenti nell'industria delle costruzioni
6. Dipendenti nel commercio
7. Dipendenti in alberghi e ristoranti
8. Dipendenti in trasporti e comunicazioni
9. Dipendenti nell'intermediazione monetaria, finanziaria e attività immobiliari
10. Dipendenti in servizi alle imprese e altre attività professionali ed imprenditoriali
11. Dipendenti nella pubblica amministrazione, difesa ed assicurazioni sociali obbligatorie
12. Dipendenti nell'istruzione, sanità ed altri servizi sociali
13. Dipendenti in altri servizi pubblici, sociali e alle persone

TABELLA 7a: PERCENTUALE DI RISPOSTE COERENTI PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE

TRIMESTRE	*NUMEROSITA'*	*% RISPOSTE COERENTI*
2004-05 1	24029	90,45%
2004-05 2	23680	93,44%
2004-05 3	23434	93,21%
2004-05 4	25192	96,84%
2005-06 1	24176	95,24%
2005-06 2	23918	95,31%
2005-06 3	22422	95,31%
2005-06 4	22289	94,86%
2006-07 1	22059	92,96%
2006-07 2	22122	93,67%
2006-07 3	21419	94,86%
2006-07 4	21981	95,21%

Come ci si poteva aspettare, le percentuali di risposte coerenti per la variabile congiunta stanno sempre tra le due "primitive" che la determinano.

Il range di valori è compreso tra il 90% e il 97% circa. Anche qui è difficile scorgere la presenza di trend nelle percentuali.

TABELLA 7b: TASSO DI DIFFERENZA NETTA PER MODALITA' DELLA CLASSIFICAZIONE CONGIUNTA DELL' ATTIVITA' PROFESSIONALE

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Autonomi	-0,19	-0,17	-0,59	-0,44	-0,09	-0,16	0,04	0,09	0,12	-0,15	0,20	0,25
Agricoltura, caccia e pesca	-0,06	-0,08	-0,01	0,02	0,09	0,00	-0,05	-0,08	0,03	0,11	-0,06	-0,06
Energia/estrattiva	0,00	0,03	-0,01	0,00	0,00	-0,01	-0,02	-0,03	0,02	-0,03	0,01	0,04
Trasformazioni	0,24	0,14	0,27	0,21	0,05	-0,11	-0,04	-0,05	0,03	0,30	-0,02	0,02
Costruzioni	0,07	0,02	0,09	0,04	-0,09	0,06	-0,04	-0,03	0,05	0,04	-0,04	-0,10
Commercio	-0,17	-0,01	0,00	0,06	0,06	0,05	0,04	0,05	-0,24	-0,21	-0,11	0,01
Alberghi e ristoranti	0,04	0,00	0,15	0,08	0,05	0,07	-0,04	0,05	0,05	0,05	-0,04	-0,01
Trasporti e comunicazioni	-0,08	0,03	0,14	0,00	-0,02	0,02	0,12	-0,01	0,04	0,01	-0,01	0,04
Intermediaz. f.,m., e att. Immobiliari	0,03	-0,02	-0,05	0,02	0,02	0,05	0,00	-0,02	-0,02	0,02	-0,02	0,00
Servizi ad imprese	0,26	0,18	0,03	0,06	-0,02	0,01	-0,05	-0,06	-0,03	-0,05	0,01	0,01
Pubblica amministrazione	-0,16	-0,10	0,02	-0,04	0,06	0,01	0,00	-0,04	0,02	0,03	0,00	-0,01
Istruzione, sanità, serv. sociali	-0,02	0,02	-0,09	-0,04	-0,01	0,03	0,11	0,15	0,02	-0,04	0,01	-0,05
Altri servizi pubblici	0,04	-0,03	0,06	0,04	-0,10	0,00	-0,07	-0,03	-0,10	-0,07	0,06	-0,12

Il range del tasso di differenza netta di questa variabile (tabella 7b) è compreso tra il -0,59% e lo 0,26%. Rispetto alle due variabili chiave tende a spostarsi leggermente verso valori negativi. La categoria autonomi è la modalità che presenta una maggiore differenza nei vari panel (e questo è in linea con quanto affermato in precedenza poiché la variabile posizione professionale assume valori più alti del settore economico).

TABELLA 7c: TASSO DI INCOERENZA PER MODALITA' DELLA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE

	2004-05 1	2004-05 2	2004-05 3	2004-05 4	2005-06 1	2005-06 2	2005-06 3	2005-06 4	2006-07 1	2006-07 2	2006-07 3	2006-07 4
Autonomi	5,94	5,54	5,25	4,78	4,89	4,64	5,22	5,97	5,94	5,54	5,25	4,78
Agricoltura, caccia e pesca	15,93	15,18	12,19	11,21	13,77	12,68	12,15	12,46	15,93	15,18	12,19	11,21
Energia/estrattiva	10,12	8,18	6,12	5,06	8,77	5,81	5,77	7,92	10,12	8,18	6,12	5,06
Trasformazioni	8,17	8,12	5,55	4,98	4,97	4,54	4,38	4,79	8,17	8,12	5,55	4,98
Costruzioni	11,44	8,52	8,39	7,50	7,05	6,87	7,20	7,79	11,44	8,52	8,39	7,50
Commercio	13,67	11,49	8,19	7,11	8,20	8,37	7,91	8,52	13,67	11,49	8,19	7,11
Alberghi e ristoranti	9,73	8,10	7,14	8,08	8,94	8,03	8,58	9,76	9,73	8,10	7,14	8,08
Trasporti e comunicazioni	8,60	7,40	6,46	5,59	5,52	5,83	4,07	5,03	8,60	7,40	6,46	5,59
Intermediaz. f.,m., e att. Immobiliari	4,88	4,08	3,68	3,59	3,30	3,32	3,34	3,85	4,88	4,08	3,68	3,59
Servizi ad imprese	15,19	13,20	9,63	10,63	9,04	10,67	10,49	9,86	15,19	13,20	9,63	10,63
Pubblica amministrazione	6,35	6,08	4,51	4,66	4,04	4,36	4,03	3,74	6,35	6,08	4,51	4,66
Istruzione, sanità, serv. sociali	7,49	3,61	3,27	2,99	2,75	2,93	2,84	3,43	7,49	3,61	3,27	2,99
Altri servizi pubblici	13,56	11,18	9,28	8,96	10,55	8,58	8,85	11,23	13,56	11,18	9,28	8,96

I risultati della tabella 7c sembrano mostrare un miglioramento sia nei valori che si riferiscono ai lavoratori autonomi (rispetto ai valori della tabella 5c), sia rispetto ai corrispondenti valori della variabile riguardante la branca di attività economica. Analizzando quindi le tre variabili rispetto al tasso di incoerenza, sembrerebbe che il provare a somministrare un'unica domanda che sintetizzi la posizione professionale e l'ambito di attività economica, possa ridurre l'incertezza che si ha nell'affrontare quesiti riguardanti la propria attività lavorativa (pur ricordando che la percentuale di risposte coerenti si piazza tra le due variabili grezze).

C'è inoltre da sottolineare che, confrontando i risultati con quelli della tesi di Padoan, c'è un miglioramento di tutti gli indicatori: questo testimonia la bontà dei miglioramenti dati dal passaggio da RTFL a RCFL grazie alle modifiche descritte in precedenza. A questo proposito comunque è stato proposto un'ulteriore analisi in appendice B.

3-ANALISI DELLA CONCORDANZA

3.1-Il coefficiente Kappa di Cohen

Una volta calcolate le misure descrittive, il passo successivo è dato dalla quantificazione del grado di accordo riscontrabile tra le risposte date a distanza di un anno.
Essendo le nostre variabili oggetto di studio di tipo categoriale, per valutare il grado di accordo si calcola il coefficiente Kappa di Cohen, che serve a misurare l'accordo osservato eccedente quello che potrebbe essere dovuto al caso.
Il coefficiente viene calcolato nel seguente modo:

$$K = \frac{p_o - p_c}{1 - p_c}$$

dove
p_0 rappresenta la proporzione di coloro per i quali si osserva l'accordo (corrispondente ai valori mostrati in tabella 5a, 6a e 7a)
p_c la proporzione di coloro dai quali ci si attende un accordo dovuto al caso (dato dal prodotto delle marginali sul totale degli intervistati).
Il Kappa dunque rappresenta lo scarto presente tra la proporzione di accordi osservata e quella attesa, dividendolo per la massima differenza possibile non causale. Il coefficiente può assumere valori compresi tra -1 e 1, anche se di fatto ha senso solo per valori positivi, visto che si vuole valutare l'accordo osservato a fronte di quello casuale. In particolare se il coefficiente assume valore pari a 0, significa che il numeratore dell'indicatore è 0, e quindi le risposte date sono tutte casuali (p_o coincide con p_c) e l'accordo è dovuto al caso. Se il coefficiente raggiunge valore 1, si è nel caso in cui sia il numeratore che il denominatore sono uguali a 1, e quindi p_c è uguale a 0: c'è quindi perfetta concordanza tra le risposte. In questo caso quindi i totali marginali di riga e colonna della matrice di contingenza coincidono e i tutti i

valori delle celle al di fuori della diagonale principale sono uguali a 0. Infine il caso di valore -1 del coefficiente significa perfetta discordanza.
Per valutare tutti i valori del coefficiente Kappa, sono state proposte molte griglie di valutazione, le più comuni sono quelle di Fleiss (1981) e di Landis e Koch (1977).
Fleiss ripartisce i valori del Kappa nel modo seguente:

- <0,4 rappresenta una riproducibilità marginale
- 0,4-0,75 rappresenta una buona riproducibilità
- >0,75 rappresenta un'eccellente riproducibilità.

Landis e Koch definiscono la seguente partizione:

- <0,01 Accordo nullo
- 0,01-0,2 Accordo scarso
- 0,21-0,4 Accordo modesto
- 0,41-0,6 Accordo moderato
- 0,61-0,8 Accordo sostanziale
- >0,8 Accordo quasi perfetto.

Oltre a calcolare il coefficiente Kappa, si sono sviluppati alcuni test per verificare se, in gruppi indipendenti, la differenza assunta da questo indicatore fosse significativa o no. La significatività può essere calcolata tramite un test bilaterale:

$$\begin{cases} H_0 : K_1 = K_2 \\ H_1 : K_1 \neq K_2 \end{cases}$$

oppure tramite quello unilaterale:

$$\begin{cases} H_0 : K_1 = K_2 \\ H_1 : K_1 > K_2 \end{cases}$$

Ovviamente il test può essere anche visto come differenza tra i due Kappa, ed essa viene posta uguale (ipotesi nulla) o diversa o maggiore (l'alternativa) da 0. La letteratura[4] dimostra che la variabile aleatoria costituita dalla differenza di Kappa indipendenti si distribuisce asintoticamente (la numerosità campionaria deve essere almeno pari a 100), sotto l'ipotesi nulla, come una Normale di media 0 e varianza data dalla somma delle varianze dei Kappa. La stima della varianza della statistica Kappa è data da

$$\hat{\sigma}_k^2 = \frac{p_0(1-p_0)}{N(1-p_c)}$$

Pertanto si può verificare la significatività del test confrontando il valore della statistica Z^{oss} con i quantili teorici della Normale standard, con:

$$Z^{oss} = \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma}_{k1}^2 + \hat{\sigma}_{k2}^2}}.$$

I sottogruppi che si sono selezionati per svolgere questo test sono dati dalla disaggregazione in base alle variabili che misurano il sesso, l'età, il livello di istruzione e in base al confronto delle risposte date direttamente in prima persona (self) piuttosto di quelle date da un altro componente della famiglia (proxy). Nella tabella 8 vengono riportati i valori del coefficiente registrati nei vari panel.

[4] In particolare si veda Cohen J. (1960), e Landis J.R., Koch G.G. (1977).

TABELLA 8: COEFFICIENTE KAPPA DI COHEN

TRIMESTRE	POSIZIONE PROFESSIONALE	ATTIVITA' ECONOMICA	CLASSIFICAZIONE CONGIUNTA
2004-05 1	0,9328	0,8747	0,8873
2004-05 2	0,9393	0,9187	0,9217
2004-05 3	0,9363	0,9192	0,9201
2004-05 4	0,9383	0,9724	0,9628
2005-06 1	0,9137	0,9522	0,9439
2005-06 2	0,9137	0,9515	0,9448
2005-06 3	0,9145	0,9540	0,9450
2005-06 4	0,9051	0,9513	0,9396
2006-07 1	0,8818	0,9229	0,9174
2006-07 2	0,8852	0,9304	0,9258
2006-07 3	0,9117	0,9499	0,9398
2006-07 4	0,9144	0,9527	0,9439

I dati riguardanti il grado di accordo (tabella 8) confermano quanto emerso già nell'analisi del capitolo due: nei primi tre panel il coefficiente relativo alla posizione professionale assume risultati maggiori rispetto a quello dell'attività economica, che però si conferma più facilmente interpretabile dal quarto trimestre del periodo 2004-05 in poi, con la variabile congiunta che assume valori sempre compresi tra quelli delle altre due variabili grezze. Come detto, ora si vuole valutare se è possibile individuare dei gruppi all'interno dei campioni che per certe caratteristiche (come ad esempio il livello di istruzione) sono portati a rispondere meglio alle domande, cioè se più facilmente definiscono allo stesso modo l'attività lavorativa svolta ad un anno di distanza, e se questo eventuale scarto è statisticamente significativo o semplicemente dovuto al caso, utilizzando il test proposto in precedenza.

3.2-Disaggregazione in base a chi risponde

Un primo test sulle differenze dei Kappa può essere svolto sui sottocampioni formati, rispettivamente, da quei soggetti che rispondono per se stessi (self) e da quelle persone per cui risponde un altro componente della famiglia (proxy). In questo caso si è scelto il test unilaterale destro, cioè con ipotesi alternativa data dalla differenza tra i due Kappa e posta maggiore di 0; quindi i quantili di riferimento sono 1.29, 1.64 e 2.33

TABELLA 9: COEFFICIENTE KAPPA CALCOLATO IN BASE A CHI RISPONDE E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
	SELF	PROXY	$H_0:K_s=K_p$ $H_1:K_s>K_p$	SELF	PROXY	$H_0:K_s=K_p$ $H_1:K_s>K_p$	SELF	PROXY	$H_0:K_s=K_p$ $H_1:K_s>K_p$
2004-05 1	0,94	0,92	5,61***	0,87	0,88	-0,95	0,89	0,88	2,58***
2004-05 2	0,95	0,93	7,58***	0,92	0,92	-1,23	0,93	0,92	2,64***
2004-05 3	0,95	0,93	6,08***	0,92	0,92	-1,17	0,92	0,92	1,1
2004-05 4	0,95	0,93	5,36***	0,97	0,97	0,91	0,97	0,96	3,65***
2005-06 1	0,92	0,9	6,01***	0,95	0,96	-2,36	0,95	0,94	2,40***
2005-06 2	0,93	0,9	7,90***	0,95	0,95	1,28	0,95	0,94	5,83***
2005-06 3	0,93	0,9	6,40***	0,95	0,95	-0,32	0,95	0,94	4,56***
2005-06 4	0,92	0,89	6,63***	0,95	0,95	-0,86	0,95	0,93	0,1
2006-07 1	0,89	0,87	5,11***	0,92	0,92	-1,08	0,92	0,91	2,71***
2006-07 2	0,89	0,87	4,08***	0,93	0,93	0,12	0,93	0,92	4,04***
2006-07 3	0,92	0,91	2,64***	0,95	0,95	-0,97	0,94	0,94	2,12**
2006-07 4	0,93	0,9	8,61***	0,96	0,95	2,71***	0,95	0,93	6,79***
****significatività per α=0,1***			*****significatività per α=0,05***			******significatività per α=0,01***			

I dati della tabella 9 mostrano di fatto quanto già detto in precedenza, e cioè che la variabile relativa alla posizione professionale è quella che presenta più difficoltà di interpretazione: infatti si può notare che per essa il Kappa è sempre più alto per il gruppo formato da quelle persone che rispondono per se stesse rispetto a coloro che "sono risposti" da un altro familiare. Inoltre questa differenza è sempre significativa al 99%. Questo non accade invece per le risposte sulla branca di attività economica, in cui invece la differenza è spesso negativa. Solo in un caso è significativa, proprio nell'ultimo trimestre di confronto: questo significa che di fatto non c'è differenza nelle risposte fornite direttamente dalla persona oggetto di studio e da quelle fornite da altri membri della famiglia, sintomo di relativa facilità di apprendimento della domanda e conseguente facilità di fornire la risposta esatta. Infine, la variabile congiunta presenta quasi sempre differenze positive e significative: i gruppi quindi forniscono risposte diverse, e quindi si sente ancora l'influenza della confusione dovuta al definire la posizione professionale di un altro individuo.

3.3-Disaggregazione in base al sesso

La seconda disaggregazione è stata fatta rispetto al sesso degli individui. In questo caso si è scelto di studiare un test bilaterale, in quanto ovviamente non si può stabilire a priori se un gruppo è più abile dell'altro a fornire le risposte (a differenza del test precedente dove si può supporre che le risposte fornite in prima persona siano migliori di quelle fornite da terzi); quindi i quantili di riferimento sono 1.64, 1.96 e 2.57. Il test è dato dalla differenza tra il Kappa dei maschi e quello delle femmine.

TABELLA 10: COEFFICIENTE KAPPA CALCOLATO IN BASE AL SESSO E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
	MASCHI	FEMMINE	$H_0:K_m=K_f$ $H_1:K_m\neq K_f$	MASCHI	FEMMINE	$H_0:K_m=K_f$ $H_1:K_m\neq K_f$	MASCHI	FEMMINE	$H_0:K_m=K_f$ $H_1:K_m\neq K_f$
2004-05 1	0,9349	0,9256	2,72***	0,8631	0,8864	-5,33***	0,8802	0,8924	-2,90***
2004-05 2	0,9465	0,924	6,68***	0,9137	0,9228	-2,49**	0,9218	0,9206	0,33
2004-05 3	0,9432	0,9215	6,28***	0,9128	0,9259	-3,59***	0,9174	0,9207	-0,90
2004-05 4	0,9426	0,9281	4,47***	0,9718	0,972	-0,09	0,9635	0,9599	1,44
2005-06 1	0,9198	0,8984	5,51***	0,9485	0,9558	-2,60***	0,943	0,9425	0,16
2005-06 2	0,92	0,8984	5,53***	0,9484	0,9539	-1,93*	0,944	0,9433	0,23
2005-06 3	0,9194	0,9012	4,54***	0,9515	0,9559	-1,53	0,9451	0,9421	0,94
2005-06 4	0,9105	0,8902	4,82***	0,948	0,954	-2,03**	0,9404	0,9353	1,53
2006-07 1	0,8862	0,8676	3,994***	0,917	0,9286	-3,15***	0,9155	0,9163	-0,207
2006-07 2	0,8884	0,8734	3,282***	0,9253	0,9351	-2,79***	0,9238	0,9252	-0,382
2006-07 3	0,9129	0,9052	1,878*	0,9452	0,955	-3,28***	0,9362	0,9424	-1,855*
2006-07 4	0,9192	0,9014	4,402***	0,9499	0,9548	1,395	0,945	0,9394	1,734*
***significatività per α=0,1**			****significatività per α=0,05**			*****significatività per α=0,01**			

Dalla tabella 9 si può notare che il gruppo formato dai maschi ha più facilità a rispondere alla domanda relativa alla posizione professionale (solo in un caso la differenza è significativa solo al 90%), ma questo risultato tende a rovesciarsi se si considera la variabile relativa all'attività economica. Infine, la variabile ottenuta dalle modalità delle due precedenti, tende a stabilire che non c'è una sostanziale differenza nei Kappa dei due gruppi, anche se in tre casi su dodici la differenza è significativa (solo una volta comunque al 99%).

3.4-Disaggregazione in base al livello di istruzione

In seguito si sono analizzati e confrontati i Kappa dei gruppi formati da chi possiede al massimo la licenza elementare, da chi possiede la licenza media (inferiore o superiore) e chi possiede almeno una laurea dei titoli post-diploma. In questo caso si è svolto un test di tipo unilaterale.

TABELLA 11a: COEFFICIENTE KAPPA CALCOLATO IN BASE AL LIVELLO DI ISTRUZIONE E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$
2004-05 1	0,9371	0,9327	0,78	0,882	0,8694	2,02**	0,9053	0,8823	4,04***
2004-05 2	0,9409	0,9397	0,22	0,9216	0,9168	0,79	0,9221	0,9215	0,12
2004-05 3	0,9296	0,9364	-1,14	0,9302	0,9164	2,31**	0,922	0,9185	0,67
2004-05 4	0,9348	0,9396	-0,84	0,9733	0,9717	0,43	0,9626	0,9629	-0,09
2005-06 1	0,8928	0,9141	-2,80	0,955	0,9518	0,83	0,9512	0,9433	1,96**
2005-06 2	0,9149	0,9132	0,25	0,9522	0,9507	0,37	0,9477	0,944	0,88
2005-06 3	0,9052	0,9143	-1,24	0,9579	0,9528	1,29*	0,9456	0,9451	0,11
2005-06 4	0,898	0,904	-0,77	0,9485	0,9509	-0,57	0,9377	0,9401	-0,52
2006-07 1	0,9029	0,8803	2,85***	0,9176	0,922	-0,82	0,9165	0,9167	-0,04
2006-07 2	0,8703	0,8877	-2,00	0,9321	0,9284	0,75	0,9333	0,9246	1,77**
2006-07 3	0,9308	0,9104	3,03***	0,9478	0,9493	-1,06	0,9365	0,9396	-0,49
2006-07 4	0,9026	0,915	-1,59	0,956	0,9513	0,88	0,9435	0,9435	0,00
***significatività per α=0,1** ***significatività per α=0,05** ***significatività per α=0,01**									
1=gruppo formato dalle persone con un titolo superiore al diploma									
2=gruppo formato dalle persone in possesso di licenza media inferiore o superiore									

Dai risultati si può notare che, di fatto, il livello di istruzione porta ad un miglioramento generale nella concordanza delle risposte, ma questo miglioramento raramente risulta essere significativo: è sufficiente un titolo di licenza media (inferiore o superiore) per rispondere in maniera corretta alle domande. Solo in otto casi su trentasei c'è una differenza significativa nei Kappa dei due gruppi. A questo punto è interessante verificare se in generale si assiste ad uno scarto positivo almeno tra licenza media ed elementare. I risultati sono riportati in tabella 11b

TABELLA 11b

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$	1	2	$H_0:K_1=K_2$ $H_1:K_1>K_2$
2004-05 1	0,9327	0,9049	5,06***	0,8694	0,8716	-0,29	0,8823	0,8826	-0,04
2004-05 2	0,9397	0,9118	5,20***	0,9168	0,9083	1,53*	0,9215	0,9178	0,59
2004-05 3	0,9364	0,9141	4,15***	0,9164	0,9103	1,10	0,9185	0,9153	0,49
2004-05 4	0,9396	0,91	5,71***	0,9717	0,9703	0,45	0,9629	0,9529	2,07**
2005-06 1	0,9141	0,8892	4,33***	0,9518	0,9348	2,86***	0,9433	0,9222	3,26***
2005-06 2	0,9132	0,8842	4,85***	0,9507	0,9414	1,65**	0,944	0,9346	1,58*
2005-06 3	0,9143	0,888	4,26***	0,9528	0,9452	1,36*	0,9451	0,9292	2,52***
2005-06 4	0,904	0,8802	3,84***	0,9509	0,943	1,35*	0,9401	0,924	2,40***
2006-07 1	0,8803	0,841	5,52***	0,922	0,9139	1,10	0,9167	0,9075	1,22
2006-07 2	0,8877	0,8385	6,91***	0,9284	0,923	0,81	0,9246	0,9079	2,29**
2006-07 3	0,9104	0,8798	4,76***	0,9493	0,9445	1,06	0,9396	0,935	0,94
2006-07 4	0,915	0,8867	4,72***	0,9513	0,9493	0,48	0,9435	0,9357	1,23
***significatività per α=0,1** ****significatività per α=0,05** *****significatività per α=0,01**									
1=gruppo formato dalle persone in possesso di licenza media inferiore o superiore									
2=gruppo formato dalle persone in possesso di licenza elementare									

Dai dati della tabella 11 si può notare che effettivamente avere almeno un grado di istruzione dato da una licenza media permette di rispondere in maniera più coerente alla posizione professionale, e questo conferma quanto detto in precedenza: sembra che questa domanda risulti più insidiosa, mentre anche l'avere solo la licenza elementare permette di rispondere coerentemente (molti test non sono significativi per $\alpha=0,05$) alla domanda relativa all'attività economica. La variabile congiunta come al solito presenta risultati "intermedi", e in sei casi la differenza tra i due gruppi risulta essere significativa.

3.5-Disaggregazione in base all'età

Infine si è compiuta la disaggregazione per età, svolgendo un test bilaterale. Per l'età (si è considerata l'età della prima intervista) si sono costituite 4 fasce: le persone fino ai 29 anni, le persone tra i 30 e i 39 anni, quelle tra i 40 e i 59, e le persone con almeno 60 anni di vita. Dai risultati si nota che in generale, col passare dell'età, le persone tendono a rispondere in maniera più coerente (e la differenza risulta quasi sempre essere significativa): questo si verifica per tutte e tre le variabili considerate. Non si nota inoltre un grosso cambiamento tra le ultime due fasce di età. Con questo si può dedurre che gli anni di esperienza aiutano ad affrontare meglio il questionario: infatti chi ha più anni probabilmente è da molto tempo che ha assunto la sua posizione lavorativa attuale e quindi sa definire meglio il tipo di attività lavorativa che svolge (tabelle 12a, 12b e 12c).

TABELLA 12a: COEFFICIENTE KAPPA CALCOLATO IN BASE ALL'ETA' (FINO A 29 ANNI E DAI 30 AI 39) E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
			$H_0:K_1=K_2$			$H_0:K_1=K_2$			$H_0:K_1=K_2$
	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$
2004-05 1	0,9066	0,9384	-5,55***	0,8489	0,8728	-3,28***	0,8435	0,8817	-5,24***
2004-05 2	0,909	0,9441	-6,16***	0,9159	0,9124	0,597	0,9023	0,9167	-2,36**
2004-05 3	0,9086	0,9386	-5,14***	0,9023	0,9148	-2,02**	0,8956	0,9143	-2,95***
2004-05 4	0,9143	0,9368	-4,04***	0,9711	0,9712	-0,03	0,9479	0,9603	-2,82***
2005-06 1	0,8738	0,9057	-4,6***	0,9423	0,9479	-1,14	0,9242	0,9381	-2,53**
2005-06 2	0,8528	0,9144	-8,49***	0,9298	0,9479	-3,44***	0,9084	0,9404	-5,45***
2005-06 3	0,8594	0,9104	-7,27***	0,9368	0,9471	-1,96**	0,9123	0,938	-4,26***
2005-06 4	0,8386	0,9048	-8,28***	0,9353	0,943	-1,41	0,9053	0,9303	-3,92***
2006-07 1	0,8208	0,88	-6,97***	0,8977	0,9165	-2,8***	0,8717	0,9069	-4,81***
2006-07 2	0,8288	0,8791	-5,97***	0,9006	0,9238	-3,51***	0,8759	0,9191	-6,02***
2006-07 3	0,8687	0,9053	-4,78***	0,9289	0,9413	-2,13**	0,9069	0,9282	-3,26***
2006-07 4	0,8572	0,9043	-5,98***	0,9345	0,9445	-1,78*	0,9106	0,9328	-3,47***
***significatività per α=0,1**			****significatività per α=0,05**			*****significatività per α=0,01**			
1=gruppo di persone fino ai 29 anni									
2=gruppo di persone tra i 30 e i 39 anni									

TABELLA 12b: COEFFICIENTE KAPPA CALCOLATO IN BASE ALL'ETA' (DAI 30 AI 39 ANNI E DAI 40 AI 59) E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
			$H_0:K_1=K_2$			$H_0:K_1=K_2$			$H_0:K_1=K_2$
	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$
2004-05 1	0,9384	0,9357	0,75	0,8728	0,8814	-1,74*	0,8817	0,898	-3,45***
2004-05 2	0,9441	0,9439	0,06	0,9124	0,9214	-2,14**	0,9167	0,929	-3,02***
2004-05 3	0,9386	0,9411	-0,69	0,9148	0,9229	-1,96**	0,9143	0,9269	-3,04***
2004-05 4	0,9368	0,9452	-2,41**	0,9712	0,9739	-1,13	0,9603	0,9684	-2,95***
2005-06 1	0,9057	0,9248	-4,52***	0,9479	0,9563	-2,62***	0,9381	0,9509	-3,71***
2005-06 2	0,9144	0,9275	-3,18***	0,9479	0,9586	-3,33***	0,9404	0,9558	-4,54***
2005-06 3	0,9104	0,9281	-5,02***	0,9471	0,9605	-4,03***	0,938	0,9552	-4,82***
2005-06 4	0,9048	0,9183	-3***	0,943	0,959	-4,67***	0,9303	0,9515	-5,64***
2006-07 1	0,88	0,8948	-2,94***	0,9165	0,9317	-3,62***	0,9069	0,9318	-5,72***
2006-07 2	0,8791	0,899	-3,92***	0,9238	0,9398	-3,95***	0,9191	0,9383	-4,63***
2006-07 3	0,9053	0,922	-3,61***	0,9413	0,9574	-4,45***	0,9282	0,9505	-5,63***
2006-07 4	0,9043	0,9297	-5,66***	0,9445	0,9594	-4,31***	0,9328	0,9554	-6,05***
***significatività per α=0,1**			****significatività per α=0,05**			*****significatività per α=0,01**			
1=gruppo di persone tra i 30 e i 39 anni									
2=gruppo di persone tra i 40 e i 59 anni									

TABELLA 12c: COEFFICIENTE KAPPA CALCOLATO IN BASE ALL'ETA' (DAI 40 AI 59 ANNI E DAI 60 IN SU) E TEST SUI GRUPPI

TRIMESTRE	POSIZIONE PROFESSIONALE			ATTIVITA' ECONOMICA			CLASSIFICAZIONE CONGIUNTA		
			$H_0:K_1=K_2$			$H_0:K_1=K_2$			$H_0:K_1=K_2$
	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$	1	2	$H_1:K_1\neq K_2$
2004-05 1	0,9357	0,9188	1,64	0,8814	0,8705	0,96	0,898	0,911	-1,32
2004-05 2	0,9439	0,9326	1,63	0,9214	0,924	-0,29	0,929	0,9235	0,6
2004-05 3	0,9411	0,924	1,62	0,9229	0,9469	-2,98***	0,9269	0,9311	-0,46
2004-05 4	0,9452	0,9096	3,71***	0,9739	0,9606	2,06**	0,9684	0,9302	4,48***
2005-06 1	0,9248	0,9112	1,42	0,9563	0,9537	0,37	0,9509	0,9397	1,22
2005-06 2	0,9275	0,9105	1,41	0,9586	0,9426	2,01**	0,9558	0,9305	3,64***
2005-06 3	0,9281	0,8934	3,13***	0,9605	0,9621	-0,23	0,9552	0,9517	0,9
2005-06 4	0,9183	0,8895	2,52**	0,959	0,9453	1,64	0,9515	0,9408	3,31***
2006-07 1	0,8948	0,8742	1,36	0,9317	0,9113	1,62	0,9318	0,9297	0,73
2006-07 2	0,899	0,865	2,84***	0,9398	0,9257	1,55	0,9383	0,9305	0,87
2006-07 3	0,922	0,9046	1,64	0,9574	0,9587	-0,19	0,9505	0,9517	-0,17
2006-07 4	0,9297	0,9011	2,87***	0,9594	0,9618	-0,38	0,9554	0,9408	1,65
***significatività per α=0,1**			****significatività per α=0,05**			*****significatività per α=0,01**			
1=gruppo di persone tra i 40 e i 59 anni									
2=gruppo di persone di almeno 60 anni									

4-AGGREGAZIONI GERARCHICHE ED ANALISI DELLA CONCORDANZA

In questo capitolo si affronta il tema dell'aggregazione di una o più modalità di risposta delle variabili chiave finora considerate: questo per verificare se è possibile individuare una specifica struttura di incoerenza. Può accadere infatti che due risposte date a distanza di un anno siano diverse per una certa somiglianza delle modalità di risposta: applicando dei pesi al calcolo del coefficiente Kappa di Cohen si vuole verificare se, aggregando man mano diverse modalità di risposta, si coglie un aumento significativo di concordanza. Si è inoltre testata la significatività dell'eventuale miglioramento che giustifichi la perdita di informazioni conseguente al processo di aggregazione. Si entra dunque in un'ottica in cui si ammette che non tutti gli errori hanno la stessa gravità: infatti quelli commessi in corrispondenza di modalità di risposta diverse ma affini si possono considerare meno gravi. Per esprimere questa diversa gravità si assegnano pesi diversi alle celle della matrice di contingenza, sulla base delle aspettative di concordanza o discordanza che il ricercatore può sospettare a priori.

I pesi w_{ij} possono assumere valori 0 o 1, e sono tali per cui il loro uso implica l'aggregazione delle frequenze relative a modalità di risposta affini: il peso assumerà valore 1 in prossimità di celle in cui c'è un perfetto accordo (questo vale per le celle della diagonale principale, in cui appunto si esprime un perfetto accordo tra la risposta data al tempo t e quella data al tempo t+1) e in quelle di incrocio tra modalità diverse ma affini (che in questa fase di aggregazione sono quindi considerate come perfetti accordi). Il peso assumerà invece valore 0 per quelle celle di incrocio tra modalità diverse e non considerate affini.

Detto questo, si esprime ancora il Kappa di Cohen come

$$K_w = \frac{p_o - p_c}{1 - p_c},$$

con:

$$p_0 = \sum_{i=1}^{r}\sum_{j=1}^{c} w_{ij}\, p_{ij}$$

$$p_c = \sum_{i=1}^{r}\sum_{j=1}^{c} w_{ij}\, p_{i.}\, p_{.j}$$

dove

r e c sono, rispettivamente, il numero di righe e di colonne della tabella di contingenza

p_{ij} è la proporzione osservata nella cella (ij)

$p_{i.}p_{.j}$ è la proporzione attesa sotto l'ipotesi di casualità, ottenuta come prodotto delle proporzioni marginali della i_esima riga per la j_esima colonna.

A differenza di quanto visto nel capitolo tre, questo test non viene svolto su campioni indipendenti, perché prima si era suddivisa la popolazione in modo tale che non ci fosse nessun elemento di un sottocampione che facesse parte anche di un altro (es. nessun maschio poteva entrare nel gruppo delle femmine e viceversa), mentre qui la popolazione di un gruppo è la medesima dell'altro gruppo su cui si va a fare il test di aggregazione: per questo il test non può essere come il precedente perché si deve tenere conto della covarianza dei due Kappa confrontati.

Il test che si vuole verificare ora è del tipo

$$\begin{cases} H_0 : K_1 = K_2 \\ H_1 : K_1 \neq K_2 \end{cases}$$

che, ancora, può essere visto come differenza dei due Kappa posta uguale o diversa da 0. Detto in precedenza che la differenza di due Kappa indipendenti

si distribuisce normalmente, nel caso di differenza tra due Kappa dipendenti la statistica test di riferimento[5] è data da:

$$W^{oss} = \frac{\left(\hat{K}_1 - \hat{K}_2\right)^2}{\hat{\sigma}^2_{k1} + \hat{\sigma}^2_{k2} - 2\hat{\sigma}_{k1k2}}$$

La statistica test, sotto l'ipotesi nulla, ha una distribuzione χ^2 con il numero di gradi di libertà dato dal numero di aggregazioni imposte passando dal primo al secondo sistema di pesi.

Le matrici di varianza e covarianza dei Kappa pesati sono state calcolate nel seguente modo[6]:

data una tabella di contingenza con I righe e J colonne, sia

$$p = \left(\frac{n_{ij}}{n_{..}}\right)$$

il vettore di dimensione (rx1) con r=IxJ, n_{ij} il numero di osservazioni presente nella generica cella (ij) e $n_{..}$ il numero totale di osservazioni della tabella. Una stima consistente della matrice di covarianza per p è data dalla matrice V(p) di dimensione (rxr) tale che

$$V(p) = \frac{1}{n}\left(D_p - pp'\right)$$

dove D_p è la matrice diagonale (rxr) sulla cui diagonale si trovano gli elementi del vettore p.

Sia $F_1(p)$, $F_2(p)$,...,$F_u(p)$ un insieme di u funzioni di p che permettono di calcolare il coefficiente Kappa considerando gli u insiemi di pesi, che mostrano i vari livelli di aggregazione delle risposte. F è definita come

[5] Si veda, oltre a quantocitato in nota 4, H.X. Barnhart e J.M. Williamson (2002) e M. Banerjee, M. Capozzoli, L. McSweeney e D. Sinha (1999).

[6] Si veda G.G. Koch *et al.* (1977).

$$F(p) = K = \frac{\sum_i^n w_i p_i - \sum_i^n w_i \left(\sum_j^n a_{ij} p_j \right) \left(\sum_j^n b_{ij} p_j \right)}{1 - \sum_i^n w_i \left(\sum_j^n a_{ij} p_j \right) \left(\sum_j^n b_{ij} p_j \right)}$$

con w_i gli elementi della matrice (rx1) dei pesi, a_i e b_i gli elementi di due matrici A e B di dimensione (rxr) e tali che $\left(\sum_j^n a_{ij} p_j \right)$ e $\left(\sum_j^n b_{ij} p_j \right)$ determinino le distribuzioni marginali al tempo t e t+1. Si assume che ogni funzione F(p) ammetta derivate parziali continue almeno fino al secondo ordine rispetto agli elementi del vettore p: una stima consistente della matrice di covarianza di F (e quindi dei Kappa gerarchici) è data dalla matrice di dimensioni (uxu) $V_F = H[V(p)]H$ □, dove

$$H = \left[\frac{\partial F(x)}{\partial x \mid x = p} \right]$$

è la matrice (uxu) delle derivate parziali prime in p della funzione F.

4.1-Posizione professionale

Per la variabile posizione professionale l'Istat propone come possibile alternativa direttamente l'aggregazione binaria dipendenti/autonomi; in questa tesi viene anche considerata una disaggregazione intermedia a 7 modalità di risposta:

1. Dirigente, quadro ed impiegato
2. Operaio ed apprendista
3. Lavoratore presso il proprio domicilio per conto di un'impresa
4. imprenditore, libero professionista e lavoratore in proprio
5. Socio di cooperativa e collaborazione coordinata e continuativa
6. Coadiuvante nell'azienda di un familiare
7. Prestazione d'opera occasionale

Le 3 diverse matrici di pesi pertanto saranno:

A) Una matrice diagonale di 13 righe e 13 colonne con la diagonale principale costituita di elementi tutti pari a 1 (per la variabile inizialmente considerata a 13 modalità come presentata nel questionario)

B) Per l'aggregazione a 7 modalità di risposta una matrice di pesi del tipo

	Dirigente	**Quadro**	**Impiegato**	**Operaio**	**Apprendista**	**Lavorante a domicilio**	**Imprenditore**	**Libero professionista**	**Lavorante in proprio**	**Socio di cooperativa**	**Coadiuvante familiare**	**Co. co. co.**	**Prestazione occasionale**
Dirigente	1	1	1	0	0	0	0	0	0	0	0	0	**0**
Quadro	1	1	1	0	0	0	0	0	0	0	0	0	**0**
Impiegato	1	1	1	0	0	0	0	0	0	0	0	0	**0**
Operaio	0	0	0	1	1	0	0	0	0	0	0	0	**0**
Apprendista	0	0	0	1	1	0	0	0	0	0	0	0	**0**
Lavorante a domicilio	0	0	0	0	0	1	0	0	0	0	0	0	**0**
Imprenditore	0	0	0	0	0	0	1	1	1	0	0	0	**0**
Libero professionista	0	0	0	0	0	0	1	1	1	0	0	0	**0**
Lavorante in proprio	0	0	0	0	0	0	1	1	1	0	0	0	**0**
Socio di cooperativa	0	0	0	0	0	0	0	0	0	1	0	1	**0**
Coadiuvante familiare	0	0	0	0	0	0	0	0	0	0	1	0	**0**
Co. co. co.	0	0	0	0	0	0	0	0	0	1	0	1	**0**
Prestazione occasionale	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**

C) Per l'aggregazione a 2 modalità di risposta una matrice di pesi del tipo

	Dirigente	**Quadro**	**Impiegato**	**Operaio**	**Apprendista**	**Lavorante a domicilio**	**Imprenditore**	**Libero professionista**	**Lavorante in proprio**	**Socio di cooperativa**	**Coadiuvante familiare**	**Co. co. co.**	**Prestazione occasionale**
Dirigente	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Quadro	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Impiegato	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Operaio	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Apprendista	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Lavorante a domicilio	1	1	1	1	1	1	0	0	0	0	0	0	**0**
Imprenditore	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Libero professionista	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Lavorante in proprio	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Socio di cooperativa	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Coadiuvante familiare	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Co. co. co.	0	0	0	0	0	0	1	1	1	1	1	1	**1**
Prestazione occasionale	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Nelle tabelle 13a e 13b vengono quindi riportati i valori assunti dal Kappa nelle tre diverse aggregazioni per ogni panel studiato e i valori del test W^{oss}.

TABELLA 13a: VALORI DEL COEFFICIENTE KAPPA PER LA POSIZIONE PROFESSIONALE IN BASE AD OGNI LIVELLO DI AGGREGAZIONE

TRIMESTRE	13 MODALITA'	7 MODALITA'	2 MODALITA'
2004-05 1	**0,9328**	**0,9541**	**0,9593**
2004-05 2	**0,9393**	**0,9576**	**0,9616**
2004-05 3	**0,9363**	**0,9529**	**0,9565**
2004-05 4	**0,9383**	**0,9572**	**0,9597**
2005-06 1	**0,9137**	**0,9381**	**0,9512**
2005-06 2	**0,9137**	**0,9385**	**0,953**
2005-06 3	**0,9145**	**0,9378**	**0,9478**
2005-06 4	**0,9051**	**0,9289**	**0,9403**
2006-07 1	**0,8818**	**0,9173**	**0,9407**
2006-07 2	**0,8852**	**0,9203**	**0,9447**
2006-07 3	**0,9117**	**0,9373**	**0,9485**
2006-07 4	**0,9144**	**0,9378**	**0,9522**

TABELLA 13b: VALORI DELLA STATISTICA TEST W^{oss} PER LA POSIZIONE PROFESSIONALE E LIVELLO DI SIGNIFICATIVITA'

TRIMESTRE	H_0:K13=K7	H_0:K7=K2
2004-05 1	**319,05*** **	**13**
2004-05 2	**266,42*** **	**8,18**
2004-05 3	**228,3*** **	**5,91**
2004-05 4	**294*** **	**3,41**
2005-06 1	**353,33*** **	**58,93*** **
2005-06 2	**360,09*** **	**78,11*** **
2005-06 3	**311,29*** **	**31,83**
2005-06 4	**305,36*** **	**35,92**
2006-07 1	**488,66*** **	**125,01*** **
2006-07 2	**487,92*** **	**140,61*** **
2006-07 3	**317,99*** **	**35,06**
2006-07 4	**307,1*** **	**62,8*** **
***significatività per α=0,1**	****significatività per α=0,05**	*****significatività per α=0,01**

Come si può notare dalla tabella 13b, il test che eguaglia 13 e 7 modalità di risposta porta sempre a rifiutare l'ipotesi nulla, anche con un livello di confidenza del 99% (il test ha 8 gradi di libertà, con un quantile teorico pari a 20,09). Quindi concludiamo che la perdita di informazioni nella riduzione del numero di modalità è giustificata da un aumento significativo del coefficiente Kappa.

Diverso il discorso per quanto riguarda il passaggio da 7 a 2 modalità di risposta: in questo caso il test ha 28 gradi di libertà, e i quantili di riferimento sono per α=0.1, 0.05 e 0.01, rispettivamente 37.92, 41.34 e 48.28: in quest'ultiomo caso l'ipotesi nulla viene rifiutata in cinque panel, per il resto la differenza non è così elevata da giustificare un ulteriore aggregazione e, quindi, un'ulteriore perdita di informazione. Si può quindi concludere che, nella maggior parte dei casi, per la variabile posizione professionale, i migliori risultati si raggiungono presentando 7 possibili modalità di risposta. Da sottolineare che nella tesi di Padoan i risultati suggerivano il passaggio a 2 modalità, ma in quella tesi il test era fatto direttamente da 11 (il numero di possibili risposte nella RTFL) a 2 modalità, senza considerare il caso intermedio[7]. Anche con i dati di questa tesi il test K13=K2 verrebbe sempre rifiutato, quindi anche qui tra 13 e 2 modalità si sceglierebbe la seconda alternativa; i dati però dimostrano che nella maggior parte dei casi analizzati è sufficiente fermarsi a 7 modalità di risposta. L'Istat tra 13 e 2 modalità di risposta suggerisce come migliore quest'ultima, cosa che conferma quanto detto per i dati usati, anche se forse il livello a 7 modalità proposto in questa tesi è ancora migliore.

[7] Si veda anche Bassi, Padoan e Trivellato (2008) in cui si analizzano i dati provenienti dalla vecchia RTFL e in cui si analizza anche il passaggio da 11 a 6 modalità di risposta.

4.2-Attività economica

Per la variabile posizione professionale vengono proposti tre diversi livelli di aggregazione: il primo è costituito da 6 modalità di risposta ed è formato da:

1. Agricoltura
2. Industria in senso stretto (energia estrattiva e delle trasformazioni)
3. Industria delle costruzioni
4. Commercio
5. Servizi (alberghi e ristoranti, trasporti e comunicazioni, intermediazione monetaria-finanziaria e attività immobiliari, servizi alle imprese)
6. Pubblica amministrazione (pubblica amministrazione, istruzione, sanità e servizi sociali, altri servizi pubblici)

Con il secondo livello di aggregazione si passa a 5 modalità di risposta, in cui si uniscono le ultime 2 modalità dell'aggregazione a 6 risposte (servizi e pubblica amministrazione che formano le "altre attività"). Infine è possibile aggregare le modalità fino a 3 diversi gruppi: agricoltura, industria e altre attività (compreso il commercio).

Le matrici dei pesi usate per stimare i coefficienti Kappa sono costruite con la stessa logica illustrata in precedenza, per poter così riprodurre le diverse aggregazioni delle risposte.

Nelle tabelle 14a e 14b vengono quindi riportati i valori assunti dal Kappa nelle diverse aggregazioni per ogni panel studiato e i valori del test W^{oss}.

TABELLA 14a: VALORI DEL COEFFICIENTE KAPPA PER LA BRANCA DI ATTIVITA' ECONOMICA IN BASE AD OGNI LIVELLO DI AGGREGAZIONE

TRIMESTRE	12 MODALITA'	6 MODALITA'	5 MODALITA'	3 MODALITA'
2004-05 1	0,8747	0,884	0,8843	0,8858
2004-05 2	0,9187	0,9239	0,9261	0,9243
2004-05 3	0,9192	0,924	0,9272	0,9291
2004-05 4	0,9724	0,9745	0,9765	0,9755
2005-06 1	0,9522	0,9552	0,9565	0,956
2005-06 2	0,9515	0,9551	0,9565	0,956
2005-06 3	0,954	0,9567	0,9578	0,9579
2005-06 4	0,9513	0,954	0,9559	0,9561
2006-07 1	0,9229	0,9254	0,9262	0,9211
2006-07 2	0,9304	0,9329	0,9334	0,9287
2006-07 3	0,9499	0,9524	0,9542	0,9532
2006-07 4	0,9527	0,9559	0,9578	0,9563

TABELLA 14b: VALORI DELLA STATISTICA TEST W^{oss} PER LA BRANCA SI ATTIVITA' ECONOMICA E LIVELLO DI SIGNIFICATIVITA'

TRIMESTRE	H_0:K12=K6	H_0:K6=K5	H_0:K5=K3
2004-05 1	84,71***	0,08	0,77
2004-05 2	42,78***	5,84	1,82
2004-05 3	37,59***	11,88	1,95
2004-05 4	20,23**	12,73	2,01
2005-06 1	23,06**	3,36	0,24
2005-06 2	32,16***	3,88	0,24
2005-06 3	19,7**	2,48	0,01
2005-06 4	18,29*	6,37	0,04
2006-07 1	11,24	0,75	13,83
2006-07 2	12,3	0,34	13,1
2006-07 3	15,28	5,38	0,81
2006-07 4	22,56**	6,37	2,04

**significatività per α=0,1 **significatività per α=0,05 ***significatività per α=0,01*

Dai dati in tabella si può notare che l'ipotesi di uguaglianza tra il coefficiente Kappa calcolato su 12 e 6 modalità viene rifiutata 4 volte se $\alpha=0,01$ (la statistica test ha 10 gradi di libertà e il quantile di riferimento è 23,21), 8 volte per $\alpha=0,05$ (il quantile vale 18,31), e un'altra volta se $\alpha=0,1$ (il quantile teorico è 15,99). Tre volte su dodici il test verrebbe accettato. Si può concludere, quindi, che in generale è ragionevole ridurre da 12 a 6 il numero di modalità. Il passaggio da 6 a 5 volte invece viene sempre accettato (anche per $\alpha=0,1$: in questo caso i gradi di libertà sono 12 e il quantile teorico vale 18,55). Anche la perdita di informazioni data dal passaggio da 5 a 3 modalità di risposta non sarebbe giustificata da un aumento significativo del coefficiente: anche in questo caso infatti per tutti i campioni si accetta l'ipotesi nulla anche per un $\alpha=0,1$: il test ha 9 gradi di libertà, e il quantile teorico vale 14,68.

Se per la posizione professionale dunque si poteva essere indecisi se considerare come migliore l'aggregazione a 7 o a 2 modalità di risposta, per l'attività economica si può affermare che c'è un'indecisione tra il livello a 12 o a 6 possibili risposte, pur propendendo maggiormente per questa seconda possibilità (in quanto l'ipotesi di uguaglianza si rifiuta nel 75% dei casi). L'Istat inoltre suggerisce come migliore proprio quella a 12 modalità di risposta.

Si può notare ancora, come già visto in precedenza, che l'attività economica è più facilmente interpretabile della posizione professionale, in quanto per quest'ultima spesso si deve arrivare fino ad una classificazione binaria delle possibili risposte per raggiungere il più alto (e significativo) livello di accordo.

4.3-Classificazione congiunta dell'attività professionale

Per quanto riguarda la variabile ottenuta dalla congiunzione di posizione professionale e attività economica, le possibili aggregazioni sono date dal gruppo di autonomi e dei dipendenti che sono raggruppati nello stesso modo che è stato fatto per la classificazione della branca di attività economica, per cui le possibili modalità di risposta scendono progressivamente da 13 a 7, 6 e 4. A queste viene aggiunta un nuovo livello di aggregazione a 4 modalità di risposta, come si può vedere dalla tabella 15:

TABELLA 15: LIVELLI DI AGGREGAZIONE PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE

<table>
<tr><th>13 Modalità</th><th>7 Modalità</th><th>6 Modalità</th><th>4 Modalità (a)</th><th>4 Modalità (b)</th></tr>
<tr><td>Autonomi</td><td>Autonomi</td><td>Autonomi</td><td>Autonomi</td><td>Autonomi</td></tr>
<tr><td>Agricoltura</td><td>Agricoltura</td><td>Agricoltura</td><td>Agricoltura</td><td>Agricoltura</td></tr>
<tr><td>Energia/estrattiva</td><td rowspan="2">Industria in senso stretto</td><td rowspan="2">Industria in senso stretto</td><td rowspan="3">Industria</td><td rowspan="6">Industria e servizi privati</td></tr>
<tr><td>Trasformazioni</td></tr>
<tr><td>Costruzioni</td><td>Costruzioni</td><td>Costruzioni</td></tr>
<tr><td>Commercio</td><td>Commercio</td><td>Commercio</td><td rowspan="6">Servizi</td></tr>
<tr><td>Alberghi/ristoranti</td><td rowspan="4">Servizi</td><td rowspan="5">Altri servizi</td></tr>
<tr><td>Trasporti/comunicazioni</td></tr>
<tr><td>Intermediari</td></tr>
<tr><td>Servizi alle imprese</td></tr>
<tr><td>Pubblica amministrazione</td><td rowspan="3">Pubblica amministrazione</td><td rowspan="3">P. a. e servizi sociali</td></tr>
<tr><td>Istruzione, sanità e servizi sociali</td></tr>
<tr><td>Altri servizi pubblici</td></tr>
</table>

TABELLA 16a: VALORI DEL COEFFICIENTE KAPPA PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE

TRIMESTRE	13 MODALITA'	7 MODALITA'	6 MODALITA'	4 (a) MODALITA'	4 (b) MODALITA'
2004-05 1	0,8873	0,8989	0,908	0,9207	0,9315
2004-05 2	0,9217	0,9305	0,9349	0,9404	0,9442
2004-05 3	0,9201	0,9263	0,9326	0,9395	0,9404
2004-05 4	0,9268	0,9642	0,9564	0,9661	0,9669
2005-06 1	0,9439	0,9465	0,9494	0,9515	0,9536
2005-06 2	0,9448	0,9484	0,9508	0,9519	0,9522
2005-06 3	0,945	0,9473	0,9479	0,9494	0,9504
2005-06 4	0,9396	0,9416	0,9423	0,9452	0,9467
2006-07 1	0,9174	0,9206	0,9225	0,9234	0,9241
2006-07 2	0,9258	0,9296	0,9324	0,9338	0,9342
2006-07 3	0,9398	0,9421	0,9443	0,945	0,9483
2006-07 4	0,9439	0,9475	0,9515	0,9533	0,9555

TABELLA 16b: VALORI DELLA STATISTICA TEST W^{oss} PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE E LIVELLO DI SIGNIFICATIVITA'

TRIMESTRE	H_0:K13=K7	H_0:K7=K6	H_0:K6=K4a	H_0:K6=K4b
2004-05 1	**157,38*** **	**46,63*** **	**28,18*** **	**35,16*** **
2004-05 2	**144,03*** **	**39,6*** **	**25,44*** **	**32,21*** **
2004-05 3	**95,69*** **	**49,39*** **	**27,73*** **	**33,52*** **
2004-05 4	**11,8**	**6,37**	**2,92**	**15,71**
2005-06 1	**23,14** **	**6,52**	**8,38**	**10,77**
2005-06 2	**37,55*** **	**21,82** **	**10,13**	**12,27**
2005-06 3	**18,24***	**1,14**	**7,27**	**8,66**
2005-06 4	**13,75**	**1,35**	**8,43**	**9,92**
2006-07 1	**23,19** **	**10,01**	**6,5**	**7,54**
2006-07 2	**34,18*** **	**16,11**	**7,15**	**9,33**
2006-07 3	**16,96***	**12,48**	**4,24**	**10,55**
2006-07 4	**30,57*** **	**26,7*** **	**6,39**	**11,48**

***significatività per α=0,1** ****significatività per α=0,05** *****significatività per α=0,01**

Come si vede, l'uguaglianza tra 13 e 7 modalità di risposta è quasi sempre rifiutata: il test ha 10 gradi di libertà, e i quantili di riferimento sono 15.99, 18.31 e 23.21: solo in due casi l'ipotesi non viene rifiutata neanche con un α=0,1. L'ipotesi di uguaglianza tra 7 e 6 modalità di risposta invece è rifiutata 5 volte: il test in questo caso ha 12 gradi di libertà e i quantili teorici valgono 18.55, 21.03 e 26.22. Tra 6 e 4 si rifiuta solo nei primi tre panel (il test ha 9 gradi di libertà e per α=0,01 il quantile di riferimento vale 21.67).

Per questa variabile quindi il percorso che da 13 possibili risposte arriva alla modalità 4(a) o 4(b) non permette di stabilire facilmente quale sia la modalità migliore: in tre casi si dovrebbero proporre solo 4 possibili risposte (fra l'altro la modalità 4b mostra un Kappa sempre maggiore della modalità 4a),

in altri due casi 6, in altri cinque casi 7 modalità di risposta e in due casi tutte e 13.

5-ANALISI DELLA STRUTTURA DELLE INCOERENZE: IL MODELLO DI QUASI INDIPENDENZA

L'ultimo capitolo della tesi è dedicato allo studio dei pattern di incoerenza tra le modalità di risposta a vari livelli di disaggregazione. In generale l'utilizzo di modelli log-lineari si rivela un buono strumento per questo scopo (Agresti, 1990), e in particolare il modello di quasi indipendenza è adatto a valutare la casualità delle risposte incoerenti (Hagenaars, 1990). Questo modello è usato per valutare se, lasciando a parte le celle della diagonale principale (quindi nella nostra indagine quelle in cui si osservano frequenze di risposte coerenti), le restanti celle mostrano particolari e significativi pattern di associazione nella tabella che è stata troncata (Goodman, 1968). Alle celle della diagonale principale (e agli incroci tra modalità affini man mano che si aggregano le risposte) è imposta frequenza nulla e lo studio viene svolto sulle celle rimanenti.
L'espressione del modello log-lineare di quasi indipendenza per una matrice di dimensioni *IxI* viene specificata come:

$$\log F_{ij} = \mu + \mu_i + \mu_j + \mu_{ij}$$

dove
F_{ij} rappresenta la frequenza attesa in una generica cella (ij) della tabella di contingenza
μ è il livello medio di tutte le celle di frequenza
μ_i e μ_j rappresentano, rispettivamente, gli effetti dell'i-esima riga e della j-esima colonna
μ_{ij} rappresenta l'ulteriore effetto per le celle della diagonale principale (il parametro vale 0 per quelle fuori della diagonale).
In questo caso, assumere la quasi indipendenza implica che le incoerenze nelle risposte sono indipendenti; rifiutare il modello viceversa implica che le incoerenze hanno un andamento non casuale, seguendo specifici schemi di

associazione tra le modalità di risposta. Una volta stimato il modello, quindi, si analizzeranno i residui per capire per quali coppie di modalità delle variabili chiave finora considerate l'associazione risulta particolarmente forte: questa potrebbe giustificare la scelta di aggregare le modalità con un grado associazione più alto (in valore assoluto), valutandone la bontà imponendo frequenze nulle anche nelle celle corrispondenti e ristimando poi il modello. Per tutte queste operazioni è stato usato il programma LEM, un programma per l'analisi di dati categoriali (Vermunt, 1997).
Gli indicatori di adattamento ai dati calcolati per il modello sono:

1. la statistica Pearson chi-square: $X^2 = \dfrac{\sum_i \left(n_i - \hat{m}_i \right)^2}{\hat{m}_i}$

   dove n_i sono le frequenze osservate e $\hat{m}_i$ quelle stimate dal modello

2. la statistica Likelihood-ratio chi square: $L^2 = 2\sum_i n_i \log\left(\dfrac{n_i}{\hat{m}_i}\right)$

   Entrambe le statistiche si distribuiscono, sotto l'ipotesi nulla di casualità, come un chi quadro, con un numero di gradi di libertà pari a [(R-1)x(C-1)-k], con k pari al numero di celle che vengono di volta in volta bloccate

3. l'indice BIC: $BIC = L^2 - df \log N$

   dove df sono i gradi di libertà calcolati per le prime due statistiche e N è la numerosità del campione.

5.1-Posizione professionale

TABELLA 17a: INDICI DI ADATTAMENTO DEL MODELLO DI QUASI INDIPENDENZA PER LA POSIZIONE PROFESSIONALE

Trimestre	*Numero categorie*	*df*	*X-square*	*L-square*	*Bic*
2004-05 1	13	131	1605,216	1322,332	-100,735
	7	115	578,1034	412,8647	-747,142
	2	59	211,949	169,4481	-425,686
2004-05 2	13	131	1313,079	1054,109	-265,374
	7	115	472,8095	323,2563	-835,068
	2	59	151,5035	125,7969	-468,474
2004-05 3	13	131	1217,077	969,0728	-349,042
	7	115	443,1081	345,0174	-812,106
	2	59	119,4997	106,5324	-487,122
2004-05 4	13	131	1206,349	1056,499	-271,092
	7	115	458,7933	329,7851	-835,657
	2	59	181,0257	141,174	-456,749
2005-06 1	13	131	1576,72	1322,124	-0,0739
	7	115	422,8885	345,0982	-815,61
	2	59	119,7637	129,1644	-466,329
2005-06 2	13	131	1712,49	1447,825	127,0211
	7	115	593,7876	420,1801	-739,304
	2	59	169,1865	144,5872	-450,279
2005-06 3	13	131	1542,921	1273,218	-39,1134
	7	115	640,515	410,1236	-741,923
	2	59	205,0926	181,3797	-409,67
2005-06 4	13	131	1564,503	1262,495	-49,0574
	7	115	465,2891	312,8394	-838,523
	2	59	178,1084	129,2948	-461,404
2006-07 1	13	131	2200,207	1697,748	387,5543
	7	115	727,1333	441,8844	-708,285
	2	59	165,5789	139,2375	-450,85
2006-07 2	13	131	2471,715	1917,634	607,0667
	7	115	788,5216	519,6365	-630,861
	2	59	133,8714	133,828	-456,427
2006-07 3	13	131	1705,441	1333,084	26,7479
	7	115	474,1818	340,2269	-806,557
	2	59	149,5046	138,8523	-449,498
2006-07 4	13	131	1625,438	1324,003	14,2735
	7	115	540,1607	381,9055	-767,857
	2	59	150,3888	138,1288	-451,749

Tutti i valori del X^2 e del L^2 sono significativi per α=0,01 ad ogni livello di aggregazione (con 131, 115 e 59 gradi di libertà i quantili di riferimento valgono, rispettivamente, 171.57, 153.19 e 87.17): è evidente quindi che persiste una dipendenza significativa (e quindi non casuale) tra risposte non

coerenti, e questo accade perfino con l'ultimo livello di aggregazione che suddivide i lavoratori in autonomi e dipendenti.

Anche se l'adattamento non risulta mai accettabile, l'indicatore BIC raggiunge il suo valore minimo in corrispondenza del modello "migliore"[8]: si può notare che, per tutti e 12 i campioni considerati, il valore minimo è raggiunto nel caso in cui si considerano 7 modalità di risposta. Questo è abbastanza in linea con quanto già visto nel capitolo quattro, in cui l'aggregazione di modalità per lo studio del coefficiente Kappa indicava che in 7 casi su 12 era proprio il livello a 7 modalità di risposta ad essere il migliore.

A questo punto è utile studiare i residui standardizzati del modello che rappresentano la dipendenza eccedente a quella dovuta al caso: più sono elevati (in valore assoluto), maggiore è la differenza tra il valore stimato sotto l'ipotesi di casualità e il valore osservato. Nello specifico, stime positive dei residui indicano che il modello sottostima l'associazione esistente tra le due modalità di risposta considerate, stime negative evidenziano una sovrastima dell'associazione.

I calcoli dei residui fanno riferimento all'aggregazione che suddivide gli occupati in dipendenti e autonomi, per capire dove sono le associazioni che persistono anche al livello di risposta meno dettagliato e che quindi impediscono di attribuire le incoerenze osservate ad un numero troppo elevato di modalità di risposta. La tabella 17b, per ogni campione considerato, illustra gli accoppiamenti tra modalità per cui si è osservato un residuo molto grande o molto piccolo (sono stati scelti i valori 2 e -2 come riferimento) in seguito alla stima del modello di quasi indipendenza, e i nuovi valori degli indicatori illustrati in precedenza in seguito alla stima del nuovo modello con le ulteriori restrizioni evidenziate dall'analisi dei residui. In questa nuova situazione non sarà necessario valutare il valore del BIC, perché si è alla presenza di modelli che presentano un buon adattamento dei dati.

[8] In presenza di modelli alternativi che non presentano un buon adattamento dei dati, il BIC premette di scegliere quello migliore, non quello corretto (Hagenaars, 1990).

TABELLA 17b: ASSOCIAZIONI POSITIVE E NEGATIVE MAGGIORI DI |2| RESIDUE AL MODELLO DI QUASI INDIPENDENZA PER LA POSIZIONE PROFESSIONALE A 2 MODALITA' DI RISPOSTA E LIVELLO DI SIGNIFICATIVITA'

Trimestre	*Modalità*	*Associazioni*		*X-square*	*L-square*	*Df*
		>2	<-2			
2004-05 1	1	7 8		29,96 (0,97)	32,67 (0,93)	46
	2					
	3		9 10			
	4	9	8			
	5	11				
	6					
	7	1 2				
	8	1 2	4			
	9					
	10					
	11					
	12	3				
	13					
2004-05 2	1	8		44,62 (0,69)	46,65 (0,61)	50
	2	8				
	3	12				
	4	10	8 12			
	5					
	6					
	7					
	8	1 2				
	9					
	10	4				
	11					
	12					
	13					
2004-05 3	1	7		35,65 (0,94)	38,21 (0.89)	50
	2	8				
	3					
	4		8			
	5					
	6					
	7	1				
	8	2	4			
	9					
	10					
	11					
	12	3	4			
	13	5				
2004-05 4	1			40,3957 (0,86)	43,07 (0,78)	51
	2					
	3					
	4	10	8			
	5	11				
	6					
	7	1				

	8 9 10 11 12 13	1 2 3	4			
2005-06 1	1 2 3 4 5 6 7 8 9 10 11 12 13	8 8 2 3 4 3	 8 4 3 	45,69 (0,65)	49,66 (0,49)	50
2005-06 2	1 2 3 4 5 6 7 8 9 10 11 12 13	7 8 8 9 1 1 2 4 	 8 4 	45,94 (0,59)	43,91 (0,68)	49
2005-06 3	1 2 3 4 5 6 7 8 9 10 11 12 13	7 8 1 2 4 6 3 5	 9 8 4 2 	31,26 (0,96)	31,58 (0,96)	47
2005-06 4	1 2 3 4 5 6 7 8 9 10	7 8 8 11 1 2 4 	 8 4 	61,31 (0,11)	58,44 (0,17)	49

	11 12 13	 5				
2006-07 1	1 2 3 4 5 6 7 8 9 10 11 12 13	 7 8 9 1 2 	 8 4 	47,87 (0,64)	50,69 (0,53)	52
2006-07 2	1 2 3 4 5 6 7 8 9 10 11 12 13	10 8 12 11 1 2 4 	 10 8 12 4 3 	37,6 (0,83)	38,71 (0,8)	47
2006-07 3	1 2 3 4 5 6 7 8 9 10 11 12 13	8 7 8 9 2 2 3 5 	 8 4 	54,94 (0,26)	57,29 (0,19)	49
2006-07 4	1 2 3 4 5 6 7 8 9 10 11 12 13	 8 12 2 3 5	 8 	61,91 (0,19)	67,16 (0,09)	53

Come già anticipato, si è sempre dinanzi a modelli che presentano un buon adattamento dei dati. In particolare le coppie di categorie che si son dovute più volte bloccare perché presentavano residui molto elevati sono: libero professionista con le prime 4 categorie di lavoro dipendente, imprenditore e quadro, lavoratore in proprio e operaio, prestazione d'opera occasionale e apprendista. Da questo sembra che i residui si concentrino in particolare nelle relazioni tra le più alte classi delle categorie dipendenti/autonomi. Questo potrebbe essere perché usualmente le cariche più alte, anche se si tratta di lavoratori dipendenti, hanno condizioni di lavoro più flessibili, così che talvolta esse possono confondere la loro posizione con una dell'altro gruppo. Un'altra possibile spiegazione è che in Italia, così come in altri paesi europei, la suddivisione dei lavoratori in dipendenti o autonomi è diventata troppo rigida, e non è adatta a cogliere le nuove forme non-standard di occupazione (Burchell, Deakin e Honey, 1999).

5.2-Attività economica

TABELLA 18a: INDICI DI ADATTAMENTO DEL MODELLO DI QUASI INDIPENDENZA PER L'ATTIVITA ECONOMICA

Trimestre	*Numero categorie*	*df*	*X-square*	*L-square*	*Bic*
2004-05 1	12	109	1518,9787	1352,576	-253,0912
	6	89	831,9902	752,0183	-323,0044
	5	65	380,0318	332,6517	-145,7262
	3	47	234,7747	210,7883	-263,3014
2004-05 2	12	109	920,1542	874,7818	-223,1083
	6	89	656,3206	574,2646	-433,1547
	5	65	273,1516	221,5504	-322,1777
	3	47	187,1849	163,4873	-309,9148
2004-05 3	12	109	799,7429	750,3248	-346,427
	6	89	539,4501	505,5479	-402,8938
	5	65	319,9006	251,1325	-389,9651
	3	47	175,7442	157,161	-315,7503
2004-05 4	12	109	326,7276	336,5815	-635,2046
	6	89	274,3847	266,7465	-768,0553
	5	65	195,8965	178,9934	-479,7349
	3	47	136,8808	135,1891	-341,1222
2005-06 1	12	109	493,5549	484,8304	-615,3192
	6	89	327,7301	304,9784	-593,3089
	5	65	180,9266	146,4437	-509,6089
	3	47	142,2417	119,8941	-354,4823
2005-06 2	12	109	490,4882	491,0686	-607,9115
	6	89	306,4526	299,2852	-598,0472
	5	65	141,8522	135,6242	-519,7309
	3	47	108,1293	106,4752	-367,397
2005-06 3	12	109	403,4757	386,2093	-705,7307
	6	89	272,612	252,356	-639,228
	5	65	125,679	122,8482	-528,3087
	3	47	95,163	86,0169	-384,8196
2005-06 4	12	109	425,6379	432,6186	-658,6729
	6	89	305,628	300,739	-590,3155
	5	65	133,2761	135,1795	-515,5907
	3	47	85,784	86,2886	-384,2683
2006-07 1	12	109	638,6857	597,6726	-492,4883
	6	89	468,684	421,4532	-468,6782
	5	65	238,6044	201,9825	-448,1135
	3	47	152,152	146,0819	-323,9875
2006-07 2	12	109	648,028	622,166	-468,3057
	6	89	471,4664	432,6959	-457,6893
	5	65	227,763	196,252	-454,0293
	3	47	160,3893	151,0103	-319,1931
2006-07 3	12	109	459,9139	449,2034	-637,7483
	6	89	323,7097	309,7006	-577,8104
	5	65	176,0229	161,3054	-486,8786
	3	47	130,7095	125,3406	-343,345
2006-07 4	12	109	381,4805	359,44	-730,3347
	6	89	269,4836	242,1027	-647,7134
	5	65	118,449	96,1432	-553,7225
	3	47	79,2471	69,6879	-400,215

Anche per la branca di attività economica, tutti i valori del X^2 e del L^2 sono significativi per α=0,01 ad ogni livello di aggregazione (con 109, 89, 65 e 47 gradi di libertà i quantili di riferimento valgono, rispettivamente, 146.26, 122.94, 94.42 e 72.44): persiste una dipendenza significativa (e quindi non casuale) tra risposte non coerenti, e questo accade anche all'ultimo livello di aggregazione a tre modalità di risposta.

Analizzando i valori assunti dall'indicatore BIC, si nota che per i primi quattro panel il valore minimo è raggiunto in corrispondenza del livello di aggregazione a 6 modalità di risposta, come veniva suggerito anche dallo studio del coefficiente Kappa. Nei trimestri successivi il valore minimo del BIC viene sempre raggiunto in corrispondenza del livello di aggregazione a 12 modalità di risposta: quest'ultima evidenza è confermata altre tre volte dallo studio del Kappa, e anche ulteriori tre volte se il test del coefficiente Kappa fosse svolto con un livello di significatività del 99%.

TABELLA 18b: ASSOCIAZIONI POSITIVE E NEGATIVE MAGGIORI DI |2| RESIDUE AL MODELLO DI QUASI INDIPENDENZA PER LA BRANCA DI ATTIVITA' ECONOMICA A 3 MODALITA' DI RISPOSTA E LIVELLO DI SIGNIFICATIVITA'

Trimestre	*Modalità*	*Associazioni*		*X-square*	*L-square*	*Df*
		>2	<-2			
2004-05 1	1	6 10 12	7	39,79 (0,07)	40,82 (0,06)	28
	2	8				
	3		10 12			
	4	9	5			
	5	3	1			
	6	1				
	7	2 4	1			
	8	4				
	9					
	10	1	3			
	11					
	12	1				
2004-05 2	1	6 10 12	9	43,77 (0,12)	44,53 (0,11)	34
	2	10				
	3		10			
	4	8				
	5		4			
	6					
	7	4				

	8 9 10 11 12	4 2 4 1				
2004-05 3	1 2 3 4 5 6 7 8 9 10 11 12	10 12 7 11 1 4 1 1	 10 1 1 3 3	35,21 (0,41)	38,67 (0,27)	34
2004-05 4	1 2 3 4 5 6 7 8 9 10 11 12	6 10 11 7 5 4 2 1	5 9 10 3 	37,15 (0,37)	34,09 (0,51)	35
2005-06 1	1 2 3 4 5 6 7 8 9 10 11 12	6 12 12 10 1 2 4 1	 5 	51,18 (0,09)	47,27 (0,25)	38
2005-06 2	1 2 3 4 5 6 7 8 9 10 11 12	10 7 8 2 	9 10 5 3 	41,52 (0,36)	46,1 (0,20)	39
2005-06 3	1 2	4 12		55,16 (0,07)	50,08 (0,15)	41

	3 4 5 6 7 8 9 10 11 12	 1 1 2 1				
2005-06 4	1 2 3 4 5 6 7 8 9 10 11 12	 1 2 	 10 	62,68 (0,03)	63,34 (0,03)	44
2006-07 1	1 2 3 4 5 6 7 8 9 10 11 12	10 12 12 3 2 4 1 4 1	 1 4 1 3 	41,43 (0,18)	40,82 (0,2)	34
2006-07 2	1 2 3 4 5 6 7 8 9 10 11 12	4 12 8 3 1 4 1 2 1 1	 5 1 4 3 3	45,59 (0,06)	47,37 (0,04)	32
2006-07 3	1 2 3 4 5 6 7 8 9	12 9 3 2 	 5 	44,73 (0,28)	46,99 (0,21)	40

	10 11 12	1	3			
2006-07 4	1 2 3 4 5 6 7 8 9 10 11 12	12 2 4 1		43,35 (0,46)	39,56 (0,62)	43

Con la nuova classificazione l'ipotesi di casualità degli errori viene sempre accettata per α=0,05, tranne che nel panel che fa riferimento al quarto trimestre del periodo 2005-06. In questo caso è più difficile interpretare i dati rispetto a quanto fatto per la posizione professionale, comunque sembra che il modello di quasi indipendenza sottostimi, ad esempio, l'associazione agricoltura con pubblica amministrazione e altri servizi pubblici, industria delle costruzioni con servizi alle imprese, e trasformazioni con il commercio. Sovrastima, invece, l'associazione tra industria delle trasformazioni e pubblica amministrazione e difesa, industria delle costruzioni e commercio.
Tuttavia è difficile attribuire questi risultati solo all'inconsistenza delle risposte fornite ad un anno di distanza, perché sembra che esse siano condizionate da errori non casuali.

5.3-Classificazione congiunta dell'attività professionale

TABELLA 19a: INDICI DI ADATTAMENTO DEL MODELLO DI QUASI INDIPENDENZA PER LA CLASSIFICAZIONE CONGIUNTA

Trimestre	*Numero categorie*	*df*	*X-square*	*L-square*	*Bic*
2004-05 1	13	131	1391,0837	1261,454	-59,945
	7	111	900,2832	812,7442	-306,9147
	6	87	600,4167	521,2214	-356,3491
	4a	69	382,7373	356,6619	-339,3422
	4b	69	286,1528	301,3479	-394,6563
2004-05 2	13		926,0672	878,609	-440,8735
	7		749,9508	661,9588	-456,0761
	6		511,8967	457,331	-418,9666
	4a		350,265	338,4826	-356,512
	4b		256,8712	244,9137	-450,0809
2004-05 3	13		882,1524	840,0181	-478,0965
	7		656,5518	633,1523	-483,7234
	6		503,5873	461,6837	-413,7054
	4a		310,9533	305,5713	-388,7028
	4b		272,083	301,3402	-392,9339
2004-05 4	13		357,9224	374,4654	-801,1781
	7		333,0188	323,7271	-953,1255
	6		277,3176	249,1307	-632,5518
	4a		226,305	203,9576	-495,3078
	4b		207,2645	199,1685	-500,0969
2005-06 1	13		531,8807	511,8309	-810,3672
	7		428,3503	387,0432	-733,2927
	6		365,8224	320,3287	-557,7724
	4a		267,3087	247,9621	-448,4629
	4b		184,1244	174,0345	-522,3905
2005-06 2	13		494,6146	488,8965	-831,9071
	7		383,9065	360,046	-759,1082
	6		277,0829	261,0138	-616,1611
	4a		208,4036	194,17	-501,5205
	4b		196,237	191,8786	-503,8118
2005-06 3	13		421,0629	404,0945	-908,237
	7		324,1212	313,2842	-798,6914
	6		266,4089	258,8504	-612,698
	4a		198,1048	199,196	-492,032
	4b		147,437	156,1432	-535,0849
2005-06 4	13		402,6884	427,1456	-884,4065
	7		332,8857	348,0095	-763,3057
	6		276,7604	284,4403	-586,5906
	4a		203,411	209,6851	-481,1325
	4b		169,631	167,211	-523,6066
2006-07 1	13		686,7271	649,5229	-660,6704
	7		538,8884	512,8899	-597,2739
	6		455,0589	422,4572	-447,6712
	4a		311,5354	296,1243	-393,9775
	4b		205,293	212,2901	-477,8118
2006-07 2	13		582,5234	586,6031	-723,9639
	7		462,9407	459,0986	-651,3818
	6		363,9091	359,9762	-510,4004
	4a		250,3121	257,5802	-432,7185
	4b		157,2268	165,6833	-524,6154

2006-07 3	13		505,4334	479,2626	-827,0738
	7		396,7824	379,5832	-727,3126
	6		291,1461	287,3596	-580,2073
	4a		213,2783	212,4741	-475,5952
	4b		181,8458	174,6053	-513,465
2006-07 4	13		413,5514	396,7527	-912,9766
	7		325,3704	308,3483	-801,4224
	6		238,0894	233,2602	-636,56
	4a		170,9889	168,1026	-521,7548
	4b		117,4676	117,8751	-571,9823

Anche per questa variabile si conclude che ad ogni livello di associazione le incoerenze nelle risposte date ad un anno di distanza seguono un pattern non casuale. Dai risultati si può notare che l'indice BIC dal primo trimestre del periodo 2005-06 indica come migliore sempre il livello a 13 modalità di risposta, segno che, dato questo studio di associazione tra le incoerenze, questa variabile ricostruita descrive bene la situazione lavorativa dei rispondenti in quanto nella maggior parte dei casi i migliori risultati del modello di quasi-indipendenza sono ottenuti senza fare alcuna aggregazione.

TABELLE 19b E 19c: ASSOCIAZIONI POSITIVE E NEGATIVE MAGGIORI DI |2| RESIDUE AL MODELLO DI QUASI INDIPENDENZA PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE: MODALITA' 4a E 4b DI RISPOSTA E LIVELLO DI SIGNIFICATIVITA'

Trimestre	*Modalità*	*Associazioni*		*X-square*	*L-square*	*Df*
		>2	<-2			
2004-05 1	1	2 7 12 13	4 6	45,64 (0,18)	49,64 (0,1)	38
	2	11 13	8			
	3	9				
	4	6	7 11 12 13			
	5	1 2	6			
	6	4	1 2			
	7	1				
	8	3 5				
	9	1				
	10					
	11	2 5	4			
	12	1	4			
	13	1				
2004-05 2	1	2 7 12	6	53,68 (0,2)	63,03 (0,05)	46
	2	11	6 10			
	3	11				
	4	6	1 7 11 12			
	5	1				
	6	4	1 5			

	7					
	8	5				
	9					
	10	3				
	11	2				
	12	1	4			
	13		4			
2004-05 3	1	2 7	4 11	60,82 (0,08)	61,18 (0,08)	47
	2	11 13	6			
	3	8				
	4	6	7 13			
	5	11				
	6	4	1 2			
	7	1				
	8					
	9	5				
	10	4	2			
	11	2				
	12	1	4			
	13					
2004-05 4	1		11	66,42 (0,14)	68,27 (0,11)	55
	2	11				
	3	8				
	4	6 10				
	5	11				
	6	4	5			
	7					
	8					
	9					
	10	3				
	11	2 3 5				
	12	1				
	13		4			
2005-06 1	1	2 5 12	6	52,24 (0,35)	59,48 (0,15)	49
	2					
	3	8 13				
	4	6	1 12			
	5	11	6			
	6	4	1			
	7					
	8					
	9					
	10	4				
	11	2 3				
	12	1	4			
	13	1	4			
2005-06 2	1	2 12	6	68,09 (0,09)	69,63 (0,07)	54
	2	5 11				
	3	8 9				
	4	6	11 12			
	5	11				
	6	4				
	7					
	8					
	9					

	10 11 12 13	3 3 1				
2005-06 3	1 2 3 4 5 6 7 8 9 10 11 12 13	2 12 1 5 13 6 10 11 3 1 	6 12 1 12 2 	65,12 (0,14)	68,23 (0,09)	54
2005-06 4	1 2 3 4 5 6 7 8 9 10 11 12 13	7 12 11 8 9 11 6 1 4 4 5 3 	6 1 12 1 	78,39 (0,0133)	83,63 (0,0046)	53
2006-07 1	1 2 3 4 5 6 7 8 9 10 11 12 13	7 12 1 11 13 11 13 6 1 4 3 2 5 1 1	6 1 6 1 4 	60,51 (0,13)	63,06 (0,08)	49
2006-07 2	1 2 3 4 5 6 7 8 9 10 11	7 12 5 11 6 1 13 4 2 3	4 12 6 1 5 4	59,42 (0,15)	64,84 (0,0643)	49

	12	1	5			
	13	1	4			
2006-07 3	1	7 12	6	50,8 (0,56)	51,99 (0,51)	53
	2	1 13				
	3	13				
	4	6	1			
	5					
	6	4				
	7	3				
	8	3				
	9					
	10					
	11	2	4			
	12	1	5			
	13	3				
2006-07 4	1	2 7		67,13 (0,25)	73,69 (0,1102)	60
	2					
	3					
	4	6	13			
	5	11				
	6	4	1			
	7					
	8	3				
	9					
	10					
	11					
	12	1				
	13					

Trimestre	*Modalità*	*Associazione*		*X-square*	*L-square*	*Df*
		>2	<-2			
2004-05 1	1	6 12	9 10 11	57,14 (0,06)	64,16 (0,01)	42
	2	4 11	8 10			
	3					
	4	2	12			
	5					
	6	1	11			
	7	12				
	8	11				
	9					
	10	1 2	11			
	11	8 9	1 4 6			
	12	1 7	6			
	13	10				
2004-05 2	1	6	9 11	57,27 (0,22)	62,47 (0,11)	50
	2	3 4 11	10			
	3	11				
	4	2				
	5					
	6					
	7					
	8					
	9					

	10		2			
	11	8 9 10	1 6			
	12	1 7 10	4			
	13					
2004-05 3	1	2 6	10 11	60,06 (0,05)	60,95 (0,05)	44
	2	4 11				
	3					
	4					
	5					
	6	1				
	7					
	8	11				
	9					
	10	11	1 2			
	11	2 3 8 10	6 7			
	12	1 10	5 6 8			
	13	8 9	6			
2004-05 4	1	6	10 11	59,8 (0,31)	70,19 (0,08)	55
	2	11				
	3	13				
	4	2				
	5					
	6					
	7	12				
	8					
	9					
	10	11				
	11	3 8 9 10	1			
	12	10				
	13					
2005-06 1	1	12	10 11	52,83 (0,56)	58,73 (0,34)	55
	2	4 6				
	3	13				
	4	2				
	5	11				
	6					
	7					
	8					
	9					
	10	13				
	11	3 10	1			
	12	10	6			
	13					
2005-06 2	1	12	11	104,89 (0)	104,11 (0)	59
	2	11	10			
	3					
	4					
	5					
	6					
	7					
	8	11				
	9					
	10	11				
	11	3 10				
	12	10				

	13	9				
2005-06 3	1	12		75,68 (0,07)	80,68 (0,04)	59
	2	5	12			
	3	13				
	4					
	5	11				
	6		12			
	7					
	8					
	9					
	10	11	2			
	11	8				
	12					
	13	9				
2005-06 4	1		11	84,47 (0,01)	85,39 (0,01)	58
	2	11				
	3	11 13				
	4					
	5					
	6		12			
	7					
	8	13				
	9					
	10	12	2			
	11	3 10				
	12					
	13		6			
2006-07 1	1	6	11	82,35 (0,01)	89,97 (0)	53
	2	13	10			
	3	13				
	4	2				
	5	1				
	6		13			
	7	11				
	8					
	9					
	10	12	2			
	11	2 3 5	1			
	12					
	13		6			
2006-07 2	1	6	11	72,38 (0,07)	74,58 (0,05)	56
	2	5				
	3	11				
	4					
	5					
	6					
	7					
	8					
	9					
	10	11	2			
	11	3 10	6			
	12	7	5			
	13	9	6			
2006-07 3	1		10 11	55,6 (0,45)	59,92 (0,3)	55
	2					

	3	13				
	4	2				
	5					
	6	1				
	7					
	8	11				
	9	11				
	10	12	1			
	11	2 10	4			
	12	1				
	13	3				
2006-07 4	1	2	10 11	58,31 (0,64)	60,85 (0,55)	63
	2					
	3					
	4					
	5					
	6					
	7					
	8					
	9					
	10	11				
	11	10				
	12					
	13	8				

Come accaduto per la branca di attività economica, con la nuova classificazione l'ipotesi di casualità degli errori viene quasi sempre accettata per $\alpha=0,05$. Per la classificazione 4a solo nel panel che fa riferimento al quarto trimestre del periodo 2005-06 c'è ancora un'evidenza di pattern non casuali negli errori, mentre per la classificazione 4b il rifiuto dell'ipotesi di casualità degli errori avviene anche in altre due occasioni.

Dalle tabelle si può notare anche che, per la parte dei dipendenti, molte delle associazioni già notate nello studio della branca di attività economica si ripresentano anche con questa variabile congiunta. Inoltre si può notare che c'è una sottostima dell'associazione tra gli autonomi e i dipendenti nell'agricoltura e una sovrastima di quella tra autonomi e i dipendenti nella pubblica amministrazione.

Come per la branca di attività economica, rispetto alla posizione professionale è più difficile spiegare il perché di queste associazioni non casuali che affliggono le risposte date a distanza di un anno.

APPENDICE A

Le variabili studiate per stabilire se in due diversi trimestri due record fanno riferimento alla stessa persona sono:

- Anno di estrazione della famiglia campione (variabile ricostruita ANNOES)
- codice della provincia e del comune di residenza (variabili ricostruite CODPRO e CODCOM)
- Codice della quartina di appartenenza della famiglia (variabile ricostruita CODQUA)
- Codice dell'ordine della famiglia nella quartina stessa (variabile ricostruita CODFAM)
- Codice identificativo unico ed invariato all'interno della famiglia (variabile ricostruita INDIV)
- Inoltre, per stabilire con maggior precisione quali possono essere i falsi positivi all'interno del gruppo degli occupati, si sono studiate anche le incoerenze delle variabili relative al sesso dell'individuo, data di nascita e titolo di studio (variabili grezze SG11, SG19 e SG24).

La variabile ricostruita COND3 (che suddivide gli intervistati in occupati, persone in cerca di occupazione e in inattivi) assegna al rispondente lo stato di occupato (e quindi una persona cui vengono somministrate le domande della sezione C del questionario) se soddisfa uno e uno solo dei seguenti criteri:

1. Alla domanda B1="*LA SCORSA SETTIMANA Lei ha svolto almeno un'ora di lavoro? Consideri il lavoro da cui ha ricavato o ricaverà un guadagno o il lavoro non pagato solo se effettuato abitualmente presso la ditta di un familiare*" risponde in maniera affermativa.
2. Dopo aver risposto negativamente a B1 alla domanda B2="*Sempre nella settimana che va "DA LUNEDI'.… A DOMENICA.…" Lei aveva comunque un lavoro dal quale era assente (ad esempio,per ridotta attività dell'impresa, per malattia, per vacanza, per cassa integrazione guadagni)?*" risponde in maniera affermativa e non risponde a tutte le altre domande della sezione B: in questo caso l'Istat assegna al

rispondente lo stato di occupato, poiché non ha nessun elemento per definirlo disoccupato

3. Dopo aver riposto negativamente alla domanda B1 e positivamente alla B2 (da ora in poi queste due condizioni non verranno più ripetute perché sempre necessarie), alla domanda B3=”*Qual è il motivo principale per cui non ha lavorato in quella settimana?*” fornisce una delle seguenti modalità di risposta: sciopero, maltempo, malattia/problemi di salute personali, ferie, festività nella settimana, orario variabile o flessibile (es. riposo compensativo), assenza obbligatoria per maternità.
4. Se alla domanda B4 (in cui si dichiara se si svolge un lavoro alle dipendenze o si dichiara l’attività autonoma) dice di fornire una collaborazione coordinativa e continuata o di fornire prestazioni d’opera occasionale, si ritiene occupato se alla domanda B5= “*E’ assunto con un contratto di lavoratore dipendente?*” risponde in maniera affermativa
5. Se alla domanda B4 dice di essere dipendente o coadiuvante nell’azienda di un familiare, è ritenuto occupato se alla domanda B6=”*Questo periodo di assenza dal lavoro durerà meno o più di tre mesi, da quando è iniziato a quando terminerà*?” risponde in maniera affermativa; se risponde negativamente è considerato occupato solo se si definisce un coadiuvante familiare (alla domanda B4) e alla domanda B9=”*Come coadiuvante familiare percepisce una retribuzione*” risponde positivamente, oppure se alla domanda B7=”*Questo periodo di assenza è retribuito almeno in parte?*” afferma di ricevere almeno il 50% della retribuzione
6. Se alla domanda B4bisβ=”*Il contratto di prestazione d’opera occasionale prevede l’obbligo di applicazione della ritenuta d’acconto. Per questo lavoro Le viene applicata una ritenuta d’acconto?*” risponde di sì e alla domanda B10=”*In questo periodo di assenza Lei ha un contratto o un accordo verbale con il datore di lavoro?*” risponde affermativamente
7. Se alla domanda B4 risponde di essere imprenditore, libero professionista o lavoratore in proprio, alla domanda B11=”*La Sua attività lavorativa è momentaneamente sospesa (ad esempio, per aggiornamento professionale, per ristrutturazione dei locali, per chiusura stagionale) o è definitivamente conclusa?*” risponde che è solo momentaneamente conclusa.

Fanno parte del campione quei soggetti che, oltre ad essere occupati in entrambe le occasioni di intervista fatte a distanza temporale di un anno, hanno lo stesso episodio lavorativo. Per capire chi non corrisponde a questo profilo, si sono usati i seguenti criteri di eliminazione:

1. Sono stati eliminati quei soggetti che hanno trovato lavoro nell'anno t+1: alla domanda C55α="*In che anno ha cominciato a lavorare per il datore di lavoro attuale?*" (se dipendente) rispondono appunto indicando l'anno della seconda intervista oppure, se lavoratori autonomi, rispondono alla domanda analoga (C55β), con l'anno della seconda intervista o scegliendo la risposta "non sa" (in un'ottica di eliminazione di potenziali falsi positivi). Ovviamente per questo criterio si considerano le risposte fornite alla seconda intervista
2. Se alla domanda c55α o c55β (sempre in occasione della seconda intervista) rispondono indicando proprio l'anno della prima intervista e alla domanda C57="*Si ricorda il mese?"* rispondono indicando un mese successivo a quello della data della prima intervista (indicata nella sezione A del questionario) o, ancora, dichiarando di non ricordarlo
3. Se nella seconda intervista dichiarano di aver trovato lavoro proprio nell'anno e nel mese della prima intervista, sono stati tenuti solo quei soggetti che ricordano anche il giorno di inizio lavoro (C58="*Si ricorda il giorno? ")* e, ovviamente, quest'ultimo è precedente al giorno della prima intervista
4. Infine, come già descritto nel primo capitolo della tesi, vengono considerate alcune variabili invarianti nel tempo (sesso e data di nascita) e una che può variare solo in un certo modo (il titolo di studio), e si confrontano i valori assunti per queste variabili nelle due diverse occasioni di intervista: se esse presentano un certo tipo di incongruenze (già illustrato nella tesi) si è deciso di eliminarle per coprirsi dal rischio di falsi positivi.

Le variabili chiave della tesi sono quelle relative alla posizione professionale (variabile ricostruita POSPRO), la branca di attività economica (variabile ricostruita CAT12) e la variabile data dalla classificazione congiunta dell'attività professionale.
La prima viene catalogata come segue:

- Si somministra la domanda C1=”*Lei svolge:*”, e le possibili risposte sono “un lavoro alle dipendenze” o tutte le possibili voci relative alle posizioni autonome già elencate nel capitolo 2: se risponde alla domanda proprio con la modalità che indica una posizione dipendente, si somministra la domanda C9=”*Lei è:*” e si sceglie una delle sei possibili voci (dirigente, quadro, impiegato, operaio, apprendista o lavoratore presso il proprio domicilio per conto di un’impresa); altrimenti, il lavoratore viene classificato con una posizione autonoma (imprenditore, libero professionista, lavoratore in proprio, coadiuvante nell’azienda di un familiare o socio di cooperativa)
- Se alla domanda C1 ha risposto con la modalità “coadiuvante familiare”, viene in seguito somministrata la domanda C1Aα=”*Lei è assunto con un contratto di lavoro alle dipendenze”*, cui il soggetto risponde indicando una delle sei modalità della domanda C9 o ribadendo ancora di essere un coadiuvante nell’azienda di un familiare. Per chi alla domanda C1 risponde “coadiuvante familiare”, la variabile POSPRO assume la modalità fornita alla domanda C1Aα
- Infine, per chi alla domanda C1 risponde “socio di cooperativa”, viene somministrata la domanda C1αβ=”*Lei è assunto con un contratto di lavoro alle dipendenze o con un contratto di collaborazione continuata e continuativa?*”: se risponde “no”, la variabile POSPRO assume modalità “socio di cooperativa”, se risponde “sì, contratto di collaborazione coordinata e continuativa”, POSPRO assume modalità “collaborazione coordinata e continuativa”, e se risponde “sì, contratto di lavoro alle dipendenze” risponde anche alla domanda C9 e la POSPRO assumerà valore corrispondente alla risposta data proprio al quesito C9

Da sottolineare che queste ultime due distinzioni possono essere fatte solo dal primo trimestre del 2005, in quanto questa sezione del questionario è stata modificata nei 4 trimestri del 2004 ed è arrivata a questa forma definitiva solo a partire dal 2005.

Per quanto riguarda invece la variabile CAT12 relativa alla branca di attività economica, la classificazione Istat è molto più facile da ricostruire in quanto fa di fatto riferimento alla domanda del questionario C16=”*Parola chiave e codifica dell’attività economia*”, che ha

ben 514 modalità di risposta. Automaticamente poi la risposta viene ricodificata in una delle 12 possibili modalità della variabile CAT12.

La variabile data dalla classificazione congiunta dell'attività professionale invece non è direttamente riscontrabile dai dato forniti dall'Istat, in quanto non è presente tra le variabili ricostruite. Essa è stata costruita assegnandole valore 1 se alla variabile POSPRO l'intervistato risponde con una delle 7 modalità che contraddistinguono i lavoratori autonomi (dalla settima alla tredicesima modalità di risposta illustrata nel capitolo 2 della tesi); se invece la variabile POSPRO assegna all'individuo una risposta relativa ad una posizione dipendente, la classificazione congiunta assume valore compreso tra 2 e 13, cioè il valore assunto dalla variabile CAT12 a cui si somma 1 (poiché il primo valore è già assunto dai lavoratori autonomi).

APPENDICE B

Confrontando i risultati di questa tesi con quelli ottenuti nel 2007 da Padoan, si nota che essi evidenziano un migliore adattamento dei dati: sembra quindi che con l'avvento della RCFL si sia notevolmente abbassata la frequenza con cui si verificano errori di misura.

Una spiegazione di questo può essere data dal fatto che col passaggio dalla RTFL (metodo di indagine usato da Padoan) alla RCFL è cambiato il metodo di indagine: oltre a tutte le migliorie ottenute (ad esempio, come già visto nel capitolo uno della tesi, il fatto che dal 2004 gli intervistatori sono formati direttamente dall'Istat), sono cambiate anche le domande stesse del questionario.

La prima grande differenza viene dalla definizione stessa di occupato, di basilare importanza perché tutte le persone che non fanno parte di questo gruppo non sono prese in considerazione per lo studio delle incoerenze. Infatti, se con l'avvento della RCFL per essere considerato occupato un soggetto deve soddisfare moltissimi criteri (illustrati in appendice A), Padoan stabilisce se una persona è occupata ispirandosi ai principi dell'ILO (International Labour Organization), che definisce occupati "*gli individui di quindici anni o più che dichiarano tale condizione oppure che, pur dichiarando uno stato diverso, affermano di aver lavorato almeno un'ora nella settimana di riferimento*".

Questa domanda sembra molto soggettiva: infatti è un individuo che dichiara autonomamente di essere occupato oppure no, senza che vengano imposti dei requisiti (come, ad esempio, il guadagno di un salario) e senza che venga fornita una definizione di occupato. La risposta quindi si basa molto sull'auto-percezione, lasciando molto spazio all'interpretazione personale (fra l'altro molte persone rispondono anche per i propri familiari, con conseguente aumento della distorsione data dalla libera interpretazione dello stato di occupato). Viene spontaneo pensare che i criteri più rigorosi introdotti dalla RCFL inducano una persona a riflettere maggiormente sulla propria posizione lavorativa, permettendo così al compilatore di stabilire con maggior sicurezza se l'intervistato è effettivamente occupato oppure no. Questo ha permesso quindi di catalogare come disoccupati o non forza lavoro molti soggetti che invece, basandosi sulla propria percezione, credevano, sbagliando, di essere occupati a tutti gli effetti. Questo si può notare anche dalla

tabella in basso che mostra come si riduce il numero medio di soggetti che compongono il campione passando dalla RTFL alla RCFL.

Con la RCFL quindi diminuisce molto il numero di soggetti che non hanno una percezione chiara circa l'essere occupato, e probabilmente questi stessi soggetti a loro volta non riuscivano a dare risposte coerenti circa la loro posizione professionale e branca di attività economica. Per questo la proporzione di soggetti che danno risposte incoerenti diminuisce con la nuova metodologia di indagine.

Una seconda differenza è data dal fatto che nella presente tesi, nel considerare quei casi in cui nella seconda intervista si dichiara di aver trovato lavoro proprio nell'anno e nel mese della prima intervista, si è guardato anche il giorno esatto, e sono stati tenuti solo quei soggetti che lo ricordano e che è precedente al giorno della prima intervista. Nel lavoro di Padoan invece, anche per una maggior penuria di informazioni, si sono tenuti tutti quei soggetti che hanno trovato lavoro proprio nell'anno e nel mese della prima intervista, senza andare a guardare anche il giorno, che invece potrebbe evidenziare la presenza di soggetti che trovano lavoro dopo la prima intervista.

Infine, in questa tesi è stato fatto un ulteriore studio su alcune variabili di controllo (sesso, data di nascita e livello di istruzione) che non era stato fatto in precedenza e che ovviamente ha portato ad ulteriori esclusioni di potenziali falsi positivi con conseguente aumento di coerenza nei risultati.

Non essendo purtroppo disponibile nella tesi di Padoan il numero di soggetti che man mano sono stati eliminati applicando i vari criteri di selezione del campione, ma solo la numerosità finale, non si possono fare confronti per capire quanto incidano nella definizione del campione i criteri di selezione corretti o aggiunti con il passaggio dalla RTFL alla RCFL. L'unico confronto che si può fare è tra le numerosità finali: dalla tabella si può notare che effettivamente la numerosità dei campioni studiati nel periodo della RTFL è più alta (mediamente di circa 2000 unità), e questo testimonia ulteriormente il fatto che probabilmente i campioni studiati da Padoan contengano un maggior numero di falsi positivi che aumentano la quota di risposte incoerenti.

TESI PADOAN		TESI TIOZZO	
TRIMESTRE	***NUMEROSITA'*** [9]	***TRIMESTRE***	***NUMEROSITA'***
1993-1994 2	26219	2004-05 1	24029
1994-1995 2	22154	2004-05 2	23680
1995-1996 2	25694	2004-05 3	23434
1996-1997 2	25739	2004-05 4	25192
1997-1998 2	25166	2005-06 1	24176
1998-1999 2	24779	2005-06 2	23918
1999-2000 2	25240	2005-06 3	22422
2000-2001 2	24778	2005-06 4	22289
2001-2002 2	25141	2006-07 1	22059
2002-2003 2	26026	2006-07 2	22122
		2006-07 3	21419
		2006-07 4	21981
MEDIA	25094	***MEDIA***	23060

A questo punto si è svolta un'ulteriore analisi di sensibilità, andando a prendere i panel del periodo 2004-2007 che si possono costruire basandosi sulle risposte fornite in interviste fatte a tre mesi di distanza, per vedere se le stime ottenute cambiano molto rispetto a quelle ottenute dalle interviste svolte a distanza di un anno. Dalla tabella uno del capitolo uno della tesi, si può vedere immediatamente che, come per le interviste fatte ad un anno di distanza, metà della popolazione può ancora fare parte del campione. Per il processo abbinamento dei record e di selezione del campione sono state fatte le stesse identiche scelte già spiegate in appendice A (ovviamente correggendo solo quei passaggi in cui si implica la distanza temporale di un trimestre anziché di un anno). In seguito sono mostrate, per ogni trimestre, la numerosità campionaria, la percentuale di risposte coerenti e il valore del coefficiente Kappa per tutte e tre le variabili chiave.

[9] Per la numerosità dei campioni della tesi di Padoan si sono sommati gli individui del gruppo A1 (coloro che in entrambe le interviste dichiaravano la stessa identica data di inizio lavoro) e quelli del gruppo A2 (chi non ricordava perfettamente la data, ma che si poteva assumere aver trovato l'occupazione di riferimento per la seconda intervista prima della data della prima intervista). Anche se si facessero i confronti solo con il panel A1, la media sarebbe comunque più alta di 1098 soggetti.

TABELLA 20a: PERCENTUALE DI RISPOSTE COERENTI E VALORE DEL COEFFICIENTE KAPPA PER LA POSIZIONE PROFESSIONALE. STIME OTTENUTE DA INTERVISTE SVOLTE A DISTANZA DI 3 MESI

TRIMESTRE	*NUMEROSITA'*	*% RISPOSTE COERENTI*	*KAPPA*
2004 1-2	21350	96,33%	0,9518
2004 2-3	21688	96,86%	0,9589
2004 3-4	23749	96,38%	0,9524
2004 4-2005 1	23993	96,91%	0,9594
2005 1-2	22970	97,64%	0,9689
2005 2-3	23636	97,47%	0,9666
2005 3-4	23532	97,35%	0,9648
2005 4-2006 1	22216	95,88%	0,9456
2006 1-2	23769	93,55%	0,9153
2006 2-3	22107	94,66%	0,9297
2006 3-4	24107	95,75%	0,9438
2006 4-2007 1	22824	95,97%	0,9467
2007 1-2	23482	96,24%	0,9505
2007 2-3	21191	96,14%	0,9489
2007 3-4	22915	96,03%	0,9472

TABELLA 20b: PERCENTUALE DI RISPOSTE COERENTI E VALORE DEL COEFFICIENTE KAPPA PER LA BRANCA DI ATTIVITA' ECONOMICA. STIME OTTENUTE DA INTERVISTE SVOLTE A DISTANZA DI 3 MESI

TRIMESTRE	*NUEROSITA'*	*% RISPOSTE COERENTI*	*KAPPA*
2004 1-2	21350	86,98%	0,8507
2004 2-3	21688	86,93%	0,8505
2004 3-4	23749	86,86%	0,85
2004 4-2005 1	23993	98,59%	0,9838
2005 1-2	22970	98,78%	0,986
2005 2-3	23636	98,83%	0,9867
2005 4-5	23532	98,60%	0,9841
2005 4-2006 1	22216	97,55%	0,972
2006 1-2	23769	95,12%	0,9444
2006 2-3	22107	96,56%	0,9608
2006 3-4	24107	97,32%	0,9694
2006 4-2007 1	22824	97,47%	0,9711
2007 1-2	23482	97,21%	0,9682
2007 2-3	21191	97,62%	0,9729
2007 3-4	22915	97,47%	97,12

TABELLA 20c: PERCENTUALE DI RISPOSTE COERENTI E VALORE DEL COEFFICIENTE KAPPA PER LA CLASSIFICAZIONE CONGIUNTA DELL'ATTIVITA' PROFESSIONALE. STIME OTTENUTE DA INTERVISTE SVOLTE A DISTANZA DI 3 MESI

TRIMESTRE	*NUMEROSITA'*	*% RISPOSTE COERENTI*	*KAPPA*
2004 1-2	21350	89,34%	0,875
2004 2-3	21688	89,32%	0,8745
2004 3-4	23749	89,18%	0,8735
2004 4-2005 1	23993	97,90%	0,9752
2005 1-2	22970	98,12%	0,9779
2005 2-3	23636	97,95%	0,976
2005 3-4	23532	97,81%	0,9745
2005 4-2006 1	22216	96,78%	0,9622
2006 1-2	23769	94,77%	0,9388
2006 2-3	22107	96,06%	0,9539
2006 3-4	24107	96,59%	0,9602
2006 4-2007 1	22824	96,80%	0,9626
2007 1-2	23482	96,75%	0,962
2007 2-3	21191	96,84%	0,963
2007 3-4	22915	96,72%	0,9619

Dai risultati si può notare che la percentuale di risposte coerenti e il valore del coefficiente Kappa sono leggermente più alti. A differenza delle risposte fornite a distanza di un anno, infatti, la minore distanza temporale aiuta a ricordare meglio la risposta data in precedenza. Questo minore scarto temporale evidenzia inoltre la presenza di soggetti che nelle interviste svolte a distanza di un anno si definiscono in entrambe le occasioni occupati, mentre la risposta intermedia è disoccupato o occupato da poco, con una durata dell'episodio lavorativo dichiarata inferiore del caso dell'intervista fatta a distanza di un anno. Questo potrebbe essere in particolare il caso di lavori occasionali o di tipo stagionale, svolti sempre a servizio dello stesso datore di lavoro. Se l'intervista viene effettuata proprio nel trimestre in cui un individuo sta svolgendo questo lavoro (ad esempio nel periodo estivo o invernale per i lavoratori stagionali), sicuramente sarà catalogato come occupato, e alla domanda "*da quando lavori con questo datore?*", potrebbe rispondere indicando l'anno in cui ha avuto la prima esperienza lavorativa, anche se poi ci sono state delle interruzioni (es. il periodo invernale per un bagnino). Se il confronto viene fatto ad un anno di distanza, l'intervista è svolta sempre nel trimestre in cui si lavora (ad esempio, per un bagnino, confrontando il terzo trimestre dell'anno t con il terzo dell'anno t+1), e quindi l'individuo viene considerato ancora occupato. Se invece il confronto è fatto con le risposte fornite il trimestre successivo

a quello in cui si è lavorato (sempre nel caso del bagnino, il periodo ottobre-dicembre), probabilmente nella seconda intervista l'individuo dichiara di essere disoccupato, perché la stagione lavorativa è finita, uscendo quindi dalla popolazione di interesse.

Pur non potendo verificare la stessa cosa con i campioni analizzati da Padoan, questo mostra che le stime sono sensibili (nell'ordine di qualche punto percentuale) alle scelte che vengono fatte nel momento in cui si uniscono i record derivanti da diverse interviste e nella fase di selezione del campione, e quindi si può concludere che lo studio ha comunque possibili sviluppi metodologici ed analitici, soprattutto avendo a disposizione questionario e dati provenienti dalla RTFL per fare ulteriori verifiche ed analisi di sensibilità.

BIBLIOGRAFIA

Agresti A. (1990), "Categorical data analysis", *New York: Wiley.*

Banerjee M., M. Capozzoli, L. McSweeney, D. Sinha (1999), "Beyond Kappa: a review of interrater agreement measures", *The Canadian journal of Statistics, 27: 3-23.*

Barnhart H.X., J.M. Williamson (2002), "Weighted least-squares approach for comparig correlated Kappa", *Biometrics 58: 1012-1019.*

Bassi F., A. Padoan, U. Trivellato (2008), "Inconsistencies in reported employment characteristics among employed stayers", *IZA, discussion paper 3908.*

Bound J., C. Brown, N. Marthiowetz (2001), "Measurement error in survey data". In J.J. Heckman e D. Leamer, *Handbook of Econometrics 5*, 59. *Amsterdam, North Holland: 3705-3843.*

Burchell B., S. Deakin, S. Honey (1999), "The employment status of individuals in non-standard employment", *EMAR publications 6.*

Cohen J. (1960), "A coefficient of agreement for nominal scales", *Educational and Psychological measurement, 20: 37-46.*

Cohen J. (1972), "Weighted Chi square: an extension of the Kappa method", *Educational and Psychological measurement, 32: 61-74.*

Di Pietro E. (1993),*"* La nuova indagine ISTAT sulle forze di lavoro. Economia e Lavoro*"*

Goodman L.A. (1968), "The analysis of cross-classified data: independence, quasi-independence and interaction in contingency tables", *Journal of the American Statistical association, 63: 1019-1131.*

Goodman L.A. (1969), "On the measurement of social mobility: an index of status presistence", *American Sociological review, 34: 831-850.*

Hagenaars J.A. (1990), "Categorical longitudial data", *Sage publications.*

Istat (2002), "Le matrici di transizione della rilevazione trimestrale sulle forze lavoro. Nota Metodologica", *Collana approfondimenti, Roma*.

Istat (2004), "La Nuova Rilevazione sulle Forze Lavoro. Contenuti, metodologie, organizzazione".

Landis J.R., G.G. Koch (1977), "The measurement of observer agreement for categorical data", *Biometrics, 33: 159-174.*

Landis J.R., G.G. Koch, J.L. Freeman, D.H. Freeman, R.G. Lehnen (1977),"A general methodology for the analysis of experiments with repeated measurement of categorical data", *Biometrics, 33: 133-158.*

Moriani C. (1981), "Forze di lavoro e flussi di popolazione", *Supplemento al bollettino mensile di statistica, Istat, 15.*

Padoan A. (2007), "La mobilità dei lavoratori per posizione professionale e attività economica nella RTFL, 1993-2003. Analisi delle incoerenze a vari livelli di disaggregazione", *Università degli studi di Padova, facoltà di scienze statistiche.*

Paggiaro A. e N. Torelli (2002), "Una procedura generalizzata per l'abbinamento esatto di record". In F. Camillo e G. Tassinaro, *Data mining, web mining e CRM, metodologie, soluzioni e prospettive*. Franco Angeli, 26-38.

Tanner M.A., M.A. Young (1985), "Modelling agreement among raters", *Journal of the American Statistical association, 80: 175-180.*

Trivellato U., A. Paggiaro, R. Leombruni, S. Rosati (2005), "La dinamica recente della mobilità dei lavoratori, 1998-2003", in B. Contini e U. Trivellato, *Eppur si muove. Dinamiche e persistenze nel mercato del lavoro italiano,* il Mulino: 271-323.

Vermunt J.K. (1997), "LEM: a general program for the analysis of categorical data", *Department of Methodology and Statistics, Tilburg university.*