



WHIP Technical Report no. 1/2010

Note di pubblicazione di WHIP v. 3.2

Roberto Leombruni, Roberto Quaranta, Claudia Villosio

Note di pubblicazione di WHIP v. 3.2

Roberto Leombruni¹, Roberto Quaranta², Claudia Villosio³

In una conferenza pubblica tenuta a Roma il 4 Maggio 2010, presso la sede centrale dell'INPS, è stato presentato il rilascio della versione 3.2 di WHIP. Rispetto al precedente rilascio (WHIP v. 2.3), l'attuale versione presenta alcune importanti novità:

- l'archivio anagrafico degli individui è stato ricostruito interamente tenendo conto della revisione e dell'aggiornamento dell'archivio dei dati anagrafici fornito dall'INPS (vedi paragrafo 2);
- sono state apportate delle modifiche nelle scelte di pubblicazione degli episodi e degli individui (vedi paragrafo 1);
- gli archivi relativi ai rapporti di lavoro dipendente (RL e RL_annui_[1985:1999]) sono ora aggiornati al 2004. Di conseguenza non sono stati pubblicati gli archivi sulle retribuzioni (Retribuzioni_[2000:2004]), che coprivano gli anni in cui nel rilascio precedente, mancando le informazioni sulle imprese, non era stata effettuata la ricostruzione longitudinale dei rapporti di lavoro;
- gli archivi contenenti le informazioni annue sui rapporti di lavoro dipendente RL_annui_[1985:1999] sono stati rinominati in RL_[1985:2004] e sono state pubblicate delle nuove variabili: `sett_retribuite`, `gg_retribuite`, `tipo_contratto`, `id_impresa` e `episodi_whip` (per i dettagli su queste nuove variabili si veda il paragrafo 4);
- l'archivio che contiene le informazioni sulle prestazioni non pensionistiche erogate dall'INPS è stato rinominato da `Indennita` ad `Indennita_annui`, in questo modo è reso più esplicito che il dataset contiene per ogni individuo le informazioni sulle prestazioni di cui è titolare consolidate anno per anno;
- sono stati corretti alcuni errori nella procedura, in particolare nel calcolo della retribuzione unitaria dei part-timers, che nella versione precedente risultava sovrastimata (vedi paragrafo 3).

In seguito all'aggiornamento della base dati si è reso necessario ricreare i codici identificativi degli individui (variabile `id_individuo`) e dei rapporti di lavoro dipendente (`id_rl`), non vi è quindi più nessuna relazione con gli identificativi della versione precedente.

¹ Università di Torino e Laboratorio R. Revelli; roberto.leombruni@unito.it

² Collegio Carlo Alberto; roberto.quaranta@carloalberto.org

³ Laboratorio R. Revelli; claudia.villosio@laboratoriorevelli.it

1. La popolazione WHIP

1.1. Scelte di pubblicazione degli individui e degli episodi

La banca dati WHIP trae origine da una fonte di natura amministrativa, basata sui dati registrati negli archivi gestionali dell'INPS e raccolti per ottemperare all'espletamento di specifici obblighi legislativi in materia previdenziale, assicurativa e assistenziale. Per consentire analisi empiriche su vari temi relativi al mercato del lavoro – in particolare su occupazione e mobilità del lavoro, retribuzioni, dinamiche e demografia di impresa – le informazioni registrate negli archivi gestionali dell'INPS vengono ricodificate, arricchite, sottoposte ad algoritmi di pulizia per migliorarne la qualità e pubblicate in WHIP.

La prima questione che sorge per un utilizzo a fini di ricerca di dati amministrativi è la corretta identificazione della popolazione di riferimento. Gli archivi gestionali dell'INPS da questo punto di vista hanno una rappresentatività piena solo per l'occupazione dipendente nel settore privato non agricolo dell'economia; per il lavoro autonomo del commercio e dell'artigianato; per il lavoro c.d. parasubordinato. Questi episodi di lavoro vengono definiti come “episodi WHIP”.

L'INPS gestisce però anche alcune assicurazioni non pensionistiche nel settore pubblico (ad es. i dipendenti dello Stato a tempo determinato che beneficiano della copertura assicurativa contro la disoccupazione); parte del settore agricolo; infine, le attività dei lavoratori dipendenti di imprese, (ad esempio le Ferrovie dello Stato e le Poste), passate nel corso degli anni Novanta, interamente all'INPS da gestioni alternative in virtù della loro trasformazione in società per azioni. Per tutte queste attività di lavoro non sono attualmente disponibili studi che garantiscano la corretta identificazione della popolazione di riferimento, la loro rappresentatività e utilizzabilità a fini di ricerca. Per questi motivi tali attività di lavoro non vengono classificati come “episodi WHIP”⁴.

La “Popolazione WHIP” è costituita quindi da tutte e sole le persone, indipendentemente dalla nazionalità e dal luogo di residenza, che hanno avuto almeno un “episodio WHIP” in età maggiore ai 15 anni. Sono in altre parole incluse tutte e sole le persone con età superiore ai 15 anni che nel periodo, che attualmente va dal 1985 al 2004, hanno avuto un'occupazione come dipendenti privati nei settori dell'industria, delle costruzioni e dei servizi⁵; come titolari di attività di lavoro artigiano; come commercianti; come parasubordinati.

Per le persone identificate come appartenenti alla popolazione WHIP vengono pubblicate *tutte* le informazioni sugli episodi di lavoro disponibili in INPS e le informazioni sull'eventuale presenza di episodi identificati come “episodi non WHIP”. Viene inoltre pubblicato l'eventuale pensionamento e gli eventuali periodi nei quali hanno beneficiato di prestazioni sociali, quali gli assegni di disoccupazione o l'indennità di mobilità.

Questa è una grossa novità rispetto alla precedente versione della banca dati, nella quale venivano esclusi del tutto gli episodi classificati come “non WHIP”, dando origine a “buchi” nella carriera delle persone che avevano entrambi i tipi di episodi di lavoro.

Per conciliare l'esigenza di svolgere analisi su attività di lavoro per le quali la rappresentatività è adeguata e l'esigenza di avere una rappresentazione il più possibile completa per gli individui inclusi nel campione, nell'archivio contenente le informazioni sui rapporti di lavoro dipendente (RL, RL_[1985:2004]) è stata creata una nuova variabile denominata

⁴ Sono in fase di sviluppo studi che permetteranno l'inclusione di alcuni di queste tipologie di rapporti di lavoro dipendente all'interno della popolazione di riferimento di WHIP.

⁵ La selezione, così come per la versione 2.3, viene fatta sulla base del codice a tre digit delle attività economiche corrispondenti alla codifica ISTAT Ateco1981.

“*episodi_whip*”. Tale variabile assume valore 1 per tutti i rapporti di lavoro riconducibili al settore privato, perché l’informazione sul settore è presente ed è corretta, oppure in caso di informazione incongruente o mancante quando è stato possibile attribuire in modo univoco il rapporto di lavoro al settore privato, mentre assume valore 0 quando un rapporto di lavoro non è riconducibile alla popolazione WHIP.

Inoltre, in linea con quanto riporta la definizione di “Popolazione WHIP”, le persone che non hanno episodi di lavoro nel settore privato, pensionamento o che hanno solo episodi precedenti il 1985 non sono più presenti nell’archivio anagrafico (*Anagrafica_individui*), mentre gli episodi di pensionamento con decorrenze precedenti il 1985 non sono più presenti nell’archivio Pensioni.

Questa è un’altra differenza rispetto al precedente rilascio: l’anagrafica della versione 2.3 conteneva tutti gli individui che hanno avuto accesso ai servizi erogati nel corso degli anni dall’INPS e non solo quelli la cui carriera lavorativa è presente in WHIP, come è invece il caso della versione 3.2.

Infine sono stati esclusi da WHIP gli individui che risultano avere nella loro carriera ricostruita un numero di rapporti di lavoro dipendente ritenuto anomalo, secondo la seguente condizione: numero medio annuale di rapporti di lavoro, con l’esclusione del primo anno di osservazione, maggiore di 6, oppure numero di rapporti di lavoro nel primo anno di osservazione maggiore di 9. In totale vengono esclusi 92 individui.

La tabella 1.1 confronta le numerosità negli archivi per le due versioni 2.3 e 3.2 di WHIP *Full Edition*. Per quanto riguarda il nuovo criterio di pubblicazione dei rapporti di lavoro dipendente, possiamo osservare nella versione 3.2 un numero maggiore di rapporti di lavoro pubblicati, ma quelli identificati dal valore 1 della variabile *episodi_whip* sono quantitativamente simili alla versione precedente (tab. 1.1). Le differenze maggiori in termini di episodi pubblicati si osservano negli archivi che contengono gli episodi di pensionamento e le indennità non pensionistiche. L’archivio delle pensioni ora contiene 83.090 episodi, mentre la versione 2.3 conteneva le informazioni sulle pensioni erogate dall’INPS nell’arco di quasi cinquant’anni, un numero pari a 427.531, ma poche erano quelle riconducibili a individui per i quali era pubblicata una carriera di lavoro. Discorso analogo per l’archivio che contiene le informazioni sulle indennità non pensionistiche, dove anche per questo caso nell’attuale versione sono pubblicati solo gli episodi che hanno come titolare un individuo la cui carriera di lavoro è presente in WHIP.

Tabella 1.1. Numero di individui e numero di episodi in WHIP Full Edition

	Versione		
	2.3		3.2
Individui			
Anagrafica_Individui	748.521	345.230	
Episodi			
Autonomi	97.221	96.154	
RI	544.219	862.074	
RI annui			di cui episodi_whip = 1
1985	102.413	120.174	114.511
1986	109.417	120.656	114.402
1987	117.727	124.540	118.051
1988	122.159	128.812	122.048
1989	125.826	132.199	125.456
1990	129.303	136.175	128.963
1991	128.919	136.404	128.645
1992	127.097	135.104	126.964
1993	119.965	127.819	119.826
1994	118.685	126.882	118.625
1995	121.079	129.746	120.915
1996	122.192	131.353	122.164
1997	122.350	132.117	122.511
1998	125.106	133.665	124.840
1999	131.834	141.219	132.524
2000		150.436	141.505
2001		156.339	147.339
2002		161.495	151.123
2003		164.711	154.365
2004		166.404	155.501
Retribuzioni			
2000	152.353	<i>Adesso RL</i>	<i>Adesso RL</i>
2001	158.606	<i>Adesso RL</i>	<i>Adesso RL</i>
2002	163.144	<i>Adesso RL</i>	<i>Adesso RL</i>
2003	165.978	<i>Adesso RL</i>	<i>Adesso RL</i>
2004	163.699	<i>Adesso RL</i>	<i>Adesso RL</i>
Indennita_annui	96.858	86.209	
Parasubordinati			
1996	9.037	9.020	
1997	11.174	11.154	
1998	12.523	12.666	
1999	13.235	13.473	
2000	12.708	13.073	
2001	13.732	14.476	
2002	16.222	17.598	
2003	18.603	20.703	
2004	16.212	18.448	
Pensioni	427.531	83.090	

2. Anagrafica

2.1. Ricostruzione e aggiornamento dell'archivio anagrafico degli individui

L'archivio anagrafico contiene le informazioni degli individui la cui carriera lavorativa è stata ricostruita e pubblicata in WHIP.

Con il rilascio della versione 3.2, questo archivio è stato interamente sostituito per tenere conto dell'aggiornamento e della revisione della procedura di identificazione degli individui che l'ultima fornitura INPS contiene. Di conseguenza, sono stati ricostruiti dei nuovi codici individuali (variabile: `id_individuo`). Inoltre, con la modifica dei criteri di selezione e di pubblicazione degli individui e degli episodi⁶, il nuovo rilascio ha comportato anche la pubblicazione di informazioni relative ad un numero di individui inferiore rispetto alla precedente versione.

L'esigenza di ricostruire l'attuale archivio e di assegnare nuovamente i codici identificativi alle persone (variabile `id_individuo`)⁷, nonostante il periodo coperto sia il medesimo per entrambe le versioni di WHIP, è dovuta a due motivi: da una parte, la versione precedente pubblicata non si basava esclusivamente sulle informazioni anagrafiche fornite dall'INPS, in quanto l'anagrafica fornita non copriva tutti gli anni per i quali erano stati forniti gli archivi con gli episodi di lavoro; dall'altra, il nuovo archivio anagrafico fornito dall'INPS ha un contenuto informativo diverso dal precedente, principalmente per l'attività di certificazione dei codici fiscali individuali che l'Istituto ha svolto negli ultimi anni.

Quando la versione 2.3 è stata rilasciata, nell'aprile del 2005, le informazioni individuali erano state estratte da un archivio anagrafico aggiornato solo fino al 2000, mentre le informazioni disponibili negli archivi gestionali coprivano un periodo più lungo, l'ultimo anno era il 2004, e includevano individui titolari di episodi ancora non presenti nell'archivio anagrafico. Per colmare questa lacuna, l'archivio anagrafico è stato pubblicato ricavando le informazioni anagrafiche mancanti dagli archivi gestionali: questo ha significato integrare il 5% di individui su un totale di tutti i presenti in anagrafica⁸, il 10% se si considera solo gli individui con episodi di lavoro tra il 1985 e il 2004 (vedi paragrafo 1, tab. 1.1). Con la versione 3.2, invece, l'archivio anagrafico si basa interamente sull'ultima fornitura inviata dall'INPS, aggiornato al 2006.

La seconda questione ha a che fare con errori di registrazione dei codici identificativi (e in particolare del codice fiscale), che rendono necessaria una tabella che associ a tutti i codici individuali presenti negli archivi gestionali il codice univoco presente nell'archivio anagrafico. Può succedere infatti che negli archivi gestionali vi siano degli episodi il cui titolare è unico, ad esempio tutti gli episodi appartengono all'individuo con codice individuale A, ma che per errore sono riferiti a codici identificativi diversi, ad esempio al codice A e al codice AA. I fattori che portano a determinare queste situazioni sono molteplici: i più comuni sono gli errori che si generano durante la compilazione dei dati anagrafici o del codice fiscale di un individuo. Ad esempio, se per errore il codice fiscale dell'individuo "QRNRRTC17L219M" viene riportato in modo incompleto o errato in uno degli archivi gestionali o nell'archivio anagrafico, ad esempio come "QRNRRTC17I219M", se

⁶ Si vedano le linee guida sulle scelte di pubblicazione nel paragrafo 1.

⁷ Per semplificare riportiamo il seguente caso ad esempio: entrambe le versioni pubblicate, la 3.2 e la 2.3, contengono un valore della variabile `id_individuo` pari a 1, questo è riferito nella versione 3.2 ad una persona di genere maschile, nata nel 1964, mentre nella versione 2.3, ad una donna, nata nel 1952.

⁸ L'anagrafica della versione 2.3 contiene tutti gli individui che hanno avuto accesso ai servizi erogati nel corso degli anni dall'INPS e non solo quelli la cui carriera lavorativa è presente in WHIP, come è invece il caso della versione 3.2: se in precedenza gli individui pubblicati erano 748.521, adesso il numero di individui è pari a 345.230 (vedi paragrafo 1, tab. 1.1).

non riconosciuto risulteranno registrate informazioni riferite a due persone: alcuni episodi saranno intestati al codice fiscale “QRNRRTC17L219M”, altri a “QRNRRTC17I219M”.

L’archivio anagrafico che viene fornito dall’INPS è quindi strutturato per contenere anche le informazioni che consentono l’identificazione e la tracciabilità nei casi in cui queste vengono corrette o modificate. Per consentire una corrispondenza tra il codice identificativo presente nell’archivio anagrafico e i codici identificativi associati agli episodi di lavoro registrati negli archivi gestionali, l’archivio anagrafico viene strutturato dall’INPS duplicando le informazioni degli individui identificati da corrispondenze problematiche, al codice (definito *cf_univoco*) che identifica in modo univoco l’individuo sono associati i diversi codici identificativi generati, per errore, durante la registrazione delle informazioni negli archivi gestionali (definito *cf_gestionale*). Tra queste due chiavi vi possono essere relazioni uno a uno o uno a molti a seconda del grado di correttezza delle informazioni riportate. Quando i codici identificativi dell’individuo sono registrati in modo coerente l’individuo è identificato sempre e solo da una sola coppia di codici (e conterranno il medesimo valore), mentre se questa situazione non si verifica e si riconosce l’errore, per poter assegnare i dati registrati all’individuo che ha il codice identificativo non corretto si generano tutte le volte delle corrispondenze che duplicano le informazioni anagrafiche. In più, alcune correzioni possono portare l’INPS a sostituire i codici identificativi errati con quelli corretti nell’archivio anagrafico o negli archivi gestionali.

Le due anagrafiche fornite dall’INPS sono quindi diverse, non solo per una questione di aggiornamento, ma perché diverse sono le procedure che le hanno generate. L’attuale archivio è il prodotto di operazioni non sistematiche di correzione, sostituzione e abbinamento di codici individuali eseguite nel momento in cui si manifestano delle incoerenze nella prassi gestionale e amministrativa dell’INPS⁹. Il precedente archivio è stato invece il risultato di un mix di procedure: a quella operata dall’INPS è stata associata una procedura di identificazione *ex post*, applicata all’interno della filiera di produzione di WHIP, che correggeva e identificava gli individui in modo sistematico sulla base delle similarità tra le informazioni anagrafiche associate ai diversi codici (nome, cognome, luogo di nascita, data di nascita)¹⁰.

Essendo diversi i livelli di aggiornamento e le procedure di identificazione delle corrispondenze tra codici, possono verificarsi diversi casi di incoerenze tra le anagrafiche. Da un lato, una fornitura può contenere dei codici identificativi, e quindi degli individui, che nell’altra non sono presenti, per il diverso aggiornamento o perché nel periodo trascorso tra un’estrazione e l’altra l’INPS ha sostituito alcuni codici identificativi. Dall’altro possono esservi casi in cui in una anagrafica due codici individuali sono associati allo stesso individuo e nell’altra no.

Dopo una scelta attentamente valutata, agevolata anche dalla constatazione che negli anni questi errori sono andati attenuandosi e allo stesso tempo i controlli dell’INPS sono stati più efficienti, è stato deciso di adottare come base per la costruzione a regime di WHIP solo l’archivio anagrafico fornito dall’INPS.

⁹ Un caso tipico di emersione di un errore di questo tipo, avviene quando i lavoratori verificano la loro situazione contributiva o quando vi sono richieste di pensionamento da parte degli individui: in questi casi può avvenire che all’individuo risultino dei periodi contributivi in meno rispetto a quelli che dovrebbero essere registrati, e che quelli mancanti siano registrati sotto un altro codice fiscale, errato. A questo punto, tutti i contributi verranno assegnati allo stesso individuo, ma non verranno modificati i codici gestionali.

¹⁰ Procedura che veniva sottomessa in modo automatico sui dati anagrafici non criptati presso l’INPS.

2.2. Analisi e confronti tra le forniture

Le analisi svolte hanno riguardato la verifica della qualità e della coerenza del contenuto informativo della nuova fornitura, sia come dato a se stante sia rispetto a quanto presente nella fornitura precedente, e si è verificata la capacità di identificare gli individui rispetto a quanto prodotto dalla procedura utilizzata per la costruzione dell'archivio della versione precedente.

Nella tabella 2.1 si riportano le distribuzioni di frequenza delle lunghezze dei codici identificativi registrati nelle due anagrafiche¹¹. La nuova fornitura ha il 44% di record in più, dai 771.189 dell'archivio precedente a 1.112.315 di quella attuale. Il `cf_univoco`¹² ha sempre una lunghezza pari 16 digit, la lunghezza del `cf_gestionale` è di 16 digit nel 92% dei casi, di 11 digit nell'8% e di una lunghezza inferiore nello 0.14% dei casi. Per confronto, nella vecchia fornitura erano presenti codici solo di lunghezza 11, 15 e 16 digit e rispettivamente nel 6.59%, nello 0.19% e nel 93.25% dei casi.

Tabella 2.1. Consistenza e lunghezza dei codici fiscali del precedente e del nuovo archivio anagrafico inviato da INPS

Lunghezza	3.2			2.3 (pre integrazione)		
	<code>cf_gestionale</code>	%	<code>cf_univoco</code>	%	<code>cf_gestionale</code>	%
6	1.077	0,10	-	-	-	-
8	101	0,01	-	-	-	-
9	312	0,03	-	-	-	-
11	87.635	7,88	-	-	50.604	6,56
15	1.988	0,18	-	-	1.467	0,19
16	1.021.202	91,81	1.112.315	100,00	719.118	93,25
Totale	1.112.315	100,00	1.112.315	100,00	771.189	100,00

Rispetto all'archivio fornito in precedenza, l'anagrafica attuale contiene tutti i record degli individui presenti negli archivi dell'INPS ai quali è stato possibile attribuire un codice fiscale completo (`cf_univoco`), ossia, per le persone fisiche, un codice fiscale composto da una stringa alfanumerica di 16 caratteri: 3 caratteri per il cognome, 3 per il nome, 5 per la data di nascita e il genere, 4 per il comune di nascita e 1 per il codice di controllo.

Da questo primo confronto, emerge anche che la procedura di identificazione dell'INPS ha potenzialmente il vantaggio di recuperare delle informazioni dai codici gestionali che hanno lunghezze al di sotto degli 11 digit, operazione che la procedura precedente non permetteva.

Le correzioni eseguite dall'INPS nell'ultima fornitura si possono quantificare guardando quanti sono i casi nei quali ad un `cf_univoco` sono stati attribuiti più `cf_gestionale`: circa l'86% dei record ha un `cf_univoco` uguale al `cf_gestionale`, mentre il 14% dei record presenta delle corrispondenze diverse (tab. 2.2).

¹¹ Il confronto è svolto sui dati "grezzi", ossia allo stato di fornitura e non sugli archivi ricostruiti.

¹² Nella nuova anagrafica il codice fiscale che identifica in modo univoco l'individuo è stato definito `cf_univoco`, mentre l'altra chiave, presente in entrambe le anagrafiche, riporta tutti i codici fiscali così come sono stati registrati negli archivi gestionali dell'INPS (versamento dei contributi, erogazione pensione...) è stata definita `cf_gestionale`.

Tabella 2.2. Corrispondenze tra Cf_univoco e Cf_gestionale, versione 3.2

Corrispondenze	Numero di casi	%
Diverse	160.682	14,45
Uguali	951.633	85,55
Totale	1.112.315	100,00

Nonostante questo considerevole numero di correzioni, le nostre analisi hanno quantificato un 2% di casi, pari a 21.624 codici fiscali, in cui le corrispondenze risultavano problematiche:

- le corrispondenze tra le due chiavi identificative (cf_univoco/cf_gestionale) non sempre rispettano almeno una di queste proprietà:
 - **“uno a uno”**, ad un solo cf_gestionale corrisponde un solo cf_univoco, da cui discende che l’individuo è identificato da una sola coppia di codici (cf_gestionale/cf_univoco) e nell’archivio è presente solamente un record con le informazioni anagrafiche;
 - **“molti a uno”**, ad un cf_univoco corrispondono più cf_gestionale, da cui consegue che le correzioni apportate portano ad avere diversi cf_gestionale attribuiti a un solo cf_univoco: l’individuo è identificato da più coppie cf_gestionale/cf_univoco e nell’archivio sono presenti un numero di record pari alle coppie create con le informazioni anagrafiche ripetute;
- in presenza di più corrispondenze (e quindi di correzioni) non sempre cf_univoco compare anche come cf_gestionale, essendo quest’ultimo la chiave che “dialoga” con gli archivi gestionali.

Successivamente, le due anagrafiche originali sono state normalizzate per individuo e le corrispondenze problematiche, precedentemente riscontrate, sono state corrette in ordine alle proprietà riportate in precedenza.

Per entrambe, la normalizzazione è consistita nella creazione di due archivi: un archivio primario con le tutte le informazioni anagrafiche, organizzato per individuo e una chiave identificativa univoca (cf_univoco); un archivio secondario con tutte le corrispondenze (cf_gestionale a cf_univoco) individuate, in cui tutti i possibili codici fiscali presenti negli archivi gestionali sono stati raccordati alla chiave identificativa univoca dell’archivio primario.

Le due anagrafiche normalizzate sono state confrontate rispetto al numero di individui presenti e alle consistenze per coorte, successivamente è stata verificata l’affidabilità delle due procedure nell’identificare gli individui.

Il risultato di questo processo di normalizzazione ha prodotto un’anagrafica nuova con 1.110.624 corrispondenze cf_gestionale – cf_univoco (tab. 2.3) e 951.094 individui (cf_univoco) il 33% in più rispetto alla versione precedente, da 714.153 individui si passa a 951.094 (tab. 2.3), mentre il numero di individui ricostruiti per abbinamento, dato dal rapporto tra il numero di individui e il totale dei codici identificativi registrati, se prima era pari allo 0.93, ora è circa lo 0.86.

Tabella 2.3. Risultato delle anagrafiche normalizzate

Anagrafiche normalizzate	Codice univoco (a)	Codici gestionali (b)	(a)/(b)
2.3	714.153	771.189	0.93
3.2	951.094	1.110.624	0.86

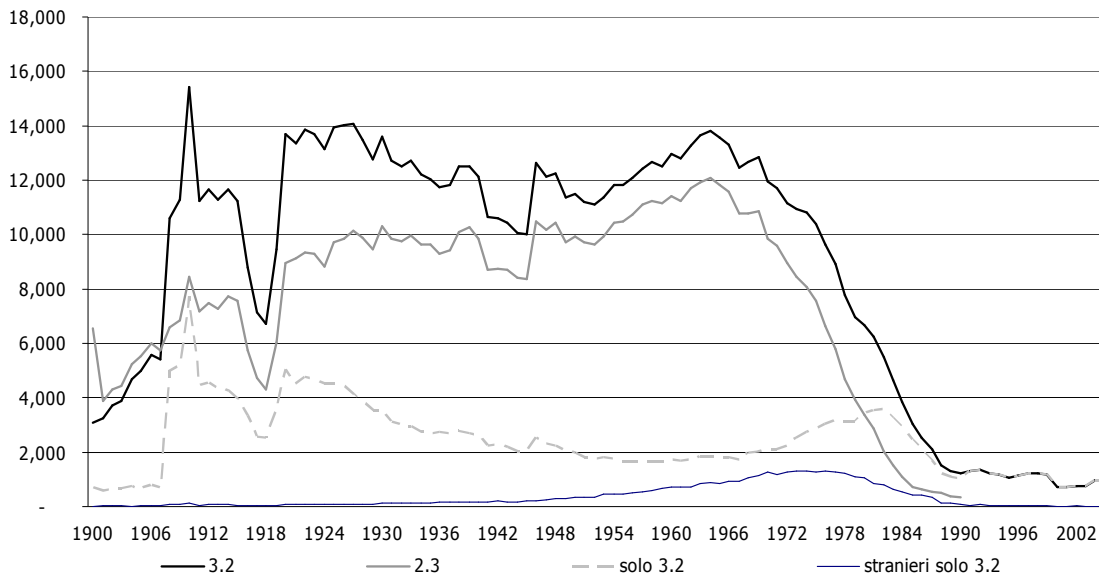
Per quanto riguarda il confronto tra gli individui, rimandando alla fine del paragrafo il confronto sull'identificazione degli individui, in figura 2.1 è possibile vedere le differenze nelle consistenze di coorte.

A parte le coorti dei nati tra il 1900 e il 1907, dove la vecchia anagrafica ha un numero maggiore di individui (l'idea che ci sia stata una sorta di censura non è stata provata, per questi individui è stato controllato l'anno di nascita registrato nel codice fiscale - dato non criptato - e risulta coerente con il dato presente nel campo "anno di nascita"), la nuova anagrafica ha sempre un numero maggiore di individui nati negli anni successivi.

Parte della differenza è attesa, poiché la nuova fornitura ha più individui in quanto più aggiornata:

- entrano nella selezione coloro che accedono al mercato del lavoro per la prima volta dopo il 1998 (coorti più giovani, o coloro che sono occupati nelle gestioni assorbite dall'INPS nell'ultimo periodo, come i dipendenti delle Ferrovie dello Stato, dei telefoni e così via);
- gli individui stranieri occupati con lavori in nero che emergono dalle regolarizzazioni (susseguites per tutti gli anni Novanta fino all'ultima del 2002). Nella vecchia anagrafica gli individui nati all'estero erano il 4% (31.000), mentre nella nuova saliamo al 7% (67.000). Questo aspetto spiega più della metà della differenza che osserviamo tra le coorti dal 1960 fino 1980;
- a questo si aggiungono altri individui, che prima non erano presenti negli archivi gestionali, e che entrano negli archivi INPS al momento in cui percepiscono una pensione non diretta (ad esempio una pensione di reversibilità).

Figura. 2.1. Consistenze delle coorti – Individui presenti nelle due anagrafiche



Gli individui presenti solo nella nuova anagrafica e nati nelle coorti tra il 1908 e il 1948 sono 208.121, di questi solo il 13% ha episodi registrati negli archivi gestionali e solo un numero esiguo fa parte della popolazione di riferimento della base dati WHIP: 300 individui con episodi di lavoro tra il 1985 e il 2004.

Discorso analogo per gli individui “bambini”, ossia con età inferiore ai 15 anni, circa 20.000 individui: di questi solo 82 hanno episodi di lavoro dipendente tra il 1985 e il 2004, 49 hanno episodi di lavoro autonomo, altri, in numero esiguo, hanno una data di nascita errata e successiva alla data dell’episodio di lavoro osservato.

La tabella 2.4 mostra i risultati dell’unione tra i codici identificativi presenti nella vecchia anagrafica e in quella nuova. Dall’unione emerge che nella nuova anagrafica vi sono 381.347 codici identificativi non presenti in quella precedente, il 33% sul totale di tutti i codici. D’altra parte, però, mancano 41.920 codici che erano stati selezionati nella vecchia fornitura. In termini di individui, questi 41.920 codici identificativi corrispondono a 31.718 persone, di cui solo 13.256 risultano avere episodi negli archivi gestionali: nella maggior parte dei casi sono individui presenti nell’archivio delle pensioni, 9.084, il 92% dei quali con episodi di pensionamento precedente al 1985, mentre le persone presenti nelle altre gestioni sono decisamente meno (tab. 2.5). Gli individui che effettivamente verrebbero a mancare dovessimo scegliere la nuova fornitura senza l’integrazione dei codici non presenti sarebbero 4.174. Questi codici sono associati ad episodi gestionali che molto probabilmente sono stati sostituiti dall’INPS, con altri codici nel frattempo corretti e inclusi nella nuova anagrafica e nei nuovi aggiornamenti degli archivi gestionali.

I codici identificativi presenti solo nel nuovo archivio, invece, sono associati ad individui presenti negli archivi gestionali in anni successivi alla data di estrazione della vecchia anagrafica (dopo il 2000). Tra questi, una quota non trascurabile, il 14%, riguarda individui nati all’estero.

Tabella 2.4. Risultato dell'unione di vecchi e nuovi identificativi

Esito abbinamento	%	
Presenti in entrambi	729.351	63.28
Presenti solo nella 2.3	41.920	3.64
Presenti solo nella 3.2	381.347	33.09
Totale	1.152.618	100

Tabella 2.5. Individui non presenti nella nuova fornitura con episodi negli archivi gestionali

Archivio	Prima e dopo il periodo di copertura di WHIP			
	Precedente il 1985		Dopo 1985	
	Tutti	A 16 digit	Tutti	A 16 digit
Pensioni	8.369	1.137	715	630
Dipendenti	-		2.860	2.367
Autonomi	-		1.100	935
Parasubordinati	-		17	17
Prestazioni non pensionistiche	-		195	195
Totale	8.369	1.137	4.887	4.174

L'analisi seguente mostra le differenze che scaturiscono dalle due differenti procedure che identificano gli individui. Da quanto emerge nei risultati riportati nella tabella 2.3, il numero maggior di individui ottenuti dall'abbinamento di più codici gestionali è prodotto dalla nuova procedura, ma per rendere il confronto più omogeneo, eliminando i disturbi generati dal differente grado di aggiornamento dei due archivi e, come è stato riscontrato, dalla mancata presenza nel nuovo archivio dei codici presenti solo nella precedente anagrafica, deve essere condotto solo sui codici presenti sia in entrambe le anagrafiche fornite sia negli archivi gestionali.

Dai 729.351 codici "comprensenti" si scende ai 641.136 codici "comprensenti e attivi" (tab. 2.6), di questi la quota di codici abbinati è rispettivamente nella vecchia procedura del 6%, dando origine a 18.528 individui su un totale di 621.636, e il 4% nella nuova, ricostruendo 11.244 individui su un totale di 629.463. La differenza è di circa 8.000 individui.

Tabella 2.6. Codici fiscali e individui

	Ricostruzione 2.3			Ricostruzione 3.2		
	Codici	%	Individui	Codici	%	Individui
Codici comprensenti	729.351					
Codici comprensenti Attivi	641.136					
Univoci	603.378	94		618.219	96	
<i>di cui individui</i>			603.378			618.219
Abbinati	37.758	6		22.917	4	
<i>di cui individui</i>			18.258			11.244
Totale	641.136	100	621.636	641.136	100	629.463

Sui codici “compresenti e attivi” la coerenza delle due procedure è stata verificata guardando al risultato dell’abbinamento prodotto da una procedura rispetto all’altra e viceversa: le due ricostruzioni sono coerenti tra loro se entrambe identificano gli individui nello stesso modo, ovvero attribuendogli gli stessi codici gestionali. I casi possibili sono stati riassunti nell’esempio della tabella 2.7¹³: il risultato delle due procedure di ricostruzione è coerente solo nel primo caso, mentre negli altri casi le due ricostruzioni portano ad un risultato incoerente.

¹³ Nel caso 1 le ricostruzioni sono coerenti in quanto entrambe attribuiscono i tre codici gestionali ad un solo individuo (nell’esempio identificato da A) e creano così una corrispondenza multipla codice gestionale/individuo. Nel caso 2 i tre codici gestionali vengono abbinati ad un solo individuo dalla procedura 1 (individuo A), mentre la 2 non identifica un solo individuo ma due (individuo A e B), uno attraverso il primo codice gestionale, l’altro creando una corrispondenza multipla con i due codici restanti (individuo B). Nel caso 3 la procedura 1 identifica un individuo per ognuno dei tre codici gestionali (individuo A, B e C), mentre la 2 ne identifica uno solo (individuo A) attribuendogli tutti e tra i codici gestionali. Infine, nel caso 4 le due procedure identificano entrambe due individui. La procedura 1 identifica un individuo abbinando i primi tre codici gestionali (individuo A) e il secondo individuo (D) viene identificato attribuendo una corrispondenza univoca con il codice gestionale ZZCCSS45SSSSWW. La procedura 2 identifica un individuo abbinando gli ultimi tre codici gestionali (individuo F) e il secondo individuo (C) viene identificato attribuendo una corrispondenze univoca con il codice gestionale F4UOGP39J10X4DK9.

Tabella 2.7. Casi possibili di coerenza/incoerenza di due procedure di identificazione degli individui

Caso	Ricostruzione 1	Codici gestionali	Ricostruzione 2	Esito
1	A	F4UOGP39J10X4DK9	A	Coerente
		N4UOGP39J10X4356		
		M4UOGP39J10X4DK9		
2	A	F4UOGP39J10X4DK9	— A	Non coerente
		N4UOGP39J10X4356	}	
		M4UOGP39J10X4DK9		
3	A	— F4UOGP39J10X4DK9	}	Non coerente
	B	— N4UOGP39J10X4356		
	C	— M4UOGP39J10X4DK9		
4	A	F4UOGP39J10X4DK9	— C	Non coerente
		N4UOGP39J10X4356	}	
		M4UOGP39J10X4DK9		
	D	— ZZCCSS45SSSSWW	F	

Il risultato del confronto è mostrato nella tabella 2.8: sul totale degli individui identificati attraverso l’abbinamento di più codici gestionali dalla vecchia procedura (18.258) solo il 37% risulta coerente, caso 1, rispetto alla nuova, mentre nella nuova il 62%. Gli individui ricostruiti in modo coerente in entrambe le procedure sono circa 7.000, mentre gli individui ricostruiti in modo non coerente sono 11.583 dalla vecchia procedura e 4.315 individui dalla nuova procedura.

La non coerenza nella maggior parte dei casi è dovuta al fatto che il codice gestionale in una procedura è abbinato e nell’altra no (come nell’esempio del caso 2, mentre in pochissimi altri casi è a sua volta abbinato ad altri codici, esempio casi 3 e 4): questo avviene rispettivamente nel 98% dei casi nella nuova e nel 96% nella vecchia.

Tabella 2.8. Confronto sulla coerenza degli individui identificati per abbinamento tra vecchia e nuova ricostruzione

Esito	Ricostruzione individui			
	2.3 rispetto alla 3.2		3.2 rispetto alla 2.3	
		%		%
Coerente	6.945	37	6.929	62
Non coerente	11.583	63	4.315	38
Totale	18.258	100	11.244	100

	Ricostruzione individui non coerente rispetto alla versione			
	3.2		2.3	
	Individui	%	Individui	%
Univoco	11.394	98	4.135	96
Abbinato	189	2	180	4
Totale	11.583	100	4.315	100

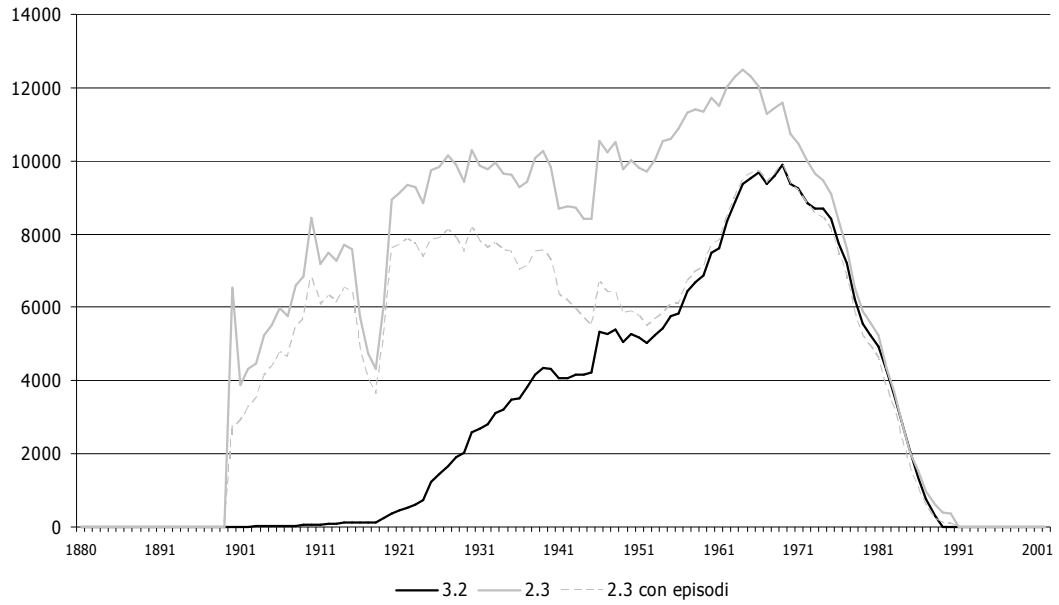
Se osserviamo le storie lavorative degli individui che sono stati identificati dalla vecchia procedura attraverso un abbinamento che è risultato non coerente rispetto alla nuova, vediamo come in termini relativi il rischio che le due procedure falliscano è simile, ma è più alto in termini assoluti per la precedente procedura. Degli 11.583 individui abbinati dalla precedente procedura solo la metà, 6.285 persone, ha episodi di lavoro registrati tra il 1985 e il 2004: il 53% (3.343 persone) risulta sempre negli archivi gestionali con il medesimo codice identificativo, mentre gli individui che hanno più codici registrati negli archivi gestionali, 2.942 individui, il 7% risulta avere degli episodi sovrapposti (in uno stesso periodo risulta essere lavoratore autonomo e lavoratore dipendente, o avere più attività di lavoratore autonomo o più rapporti di lavoro) nella carriera lavorativa. Viceversa, tra i 4.315 individui identificati nella nuova procedura per abbinamento e non coerenti rispetto alla vecchia, ve ne sono 2.629 con episodi di lavoro: di questi il 28% (726 individui) non cambia codice gestionale, mentre 1903 individui cambiano codice gestionale e il 6% ha periodi di sovrapposizione negli episodi di lavoro.

2.3. Differenza tra le anagrafiche pubblicate nella Full Edition

In figura 2.2 vediamo il confronto delle consistenze di coorte delle anagrafiche rilasciate della versione 3.2 e della versione 2.3, considerando tutti gli individui pubblicati e solo gli individui che hanno avuto episodi pubblicati (versione “2.3 con episodi”). Entrambe le versioni pubblicate vengono prima normalizzate per individuo, seguendo le procedure descritte in precedenza. All’interno dell’archivio vi è un solo record per ogni persona: la chiave che permette di identificare l’individuo e di collegare le informazioni anagrafiche a tutti gli altri archivi è `id_individuo`.

Le differenze che emergono sono da attribuirsi principalmente alle scelte e alle modifiche che sono state introdotte nella nuova versione: in linea con quanto riportato nella definizione di “Popolazione WHIP”, la nuova anagrafica non contiene più gli individui privi di episodi di lavoro tra il 1985 e il 2004 o esclusivamente con episodi di pensionamento.

Figura 2.2. Consistenze delle coorti, individui presenti nella versione 3.2 e 2.3



3. Rapporti di lavoro

3.1. Ricostruzione rapporti di lavoro

Gli archivi dei rapporti di lavoro dipendente rappresentano il contributo di maggior valore della procedura di costruzione di WHIP rispetto a quanto incluso nei dati sorgente forniti dall'INPS¹⁴: in particolare, nella fase in cui avviene l'identificazione longitudinale dei rapporti di lavoro, che porta alla stima delle date di inizio e di fine dei rapporti di lavoro e alla normalizzazione dei dati dei moduli contributivi in dati annui relativi al rapporto di lavoro.

I record forniti dall'INPS sono organizzati per dichiarazioni contributive prodotte a favore di un lavoratore, utilizzate dall'INPS per il calcolo delle prestazioni pensionistiche e non pensionistiche che possono spettare all'assicurato. Non esiste però una relazione uno a uno tra gli episodi di contribuzione dichiarati e i rapporti di lavoro instaurati dalla persona (la relazione è del tipo molti a molti). Inoltre, non è registrata esplicitamente una data di inizio del rapporto di lavoro, e la data di cessazione – seppur presente – è di qualità carente, sia in termini di valori mancanti che di valori scorretti. I passi della procedura dedicati all'identificazione longitudinale dei rapporti di lavoro correggono queste carenze e permettono di creare le variabili `inizio_rl` e `fine_rl` (per i dettagli su queste variabili consultare i metadati on line).

Con il rilascio della versione 3.2, le informazioni sui rapporti di lavoro sono state ricostruite ed estese dal 1999 al 2004, di conseguenza gli archivi sui rapporti di lavoro sostituiscono gli archivi sulle retribuzioni (Retribuzioni_[2000:2004]), che coprivano gli anni in cui nella versione 2.3 non era stata possibile la ricostruzione del rapporto di lavoro per mancanza dei dati di impresa.

Inoltre, nella nuova *release* della base dati WHIP è stata recuperata l'informazione sul settore di attività che risultava mancante per alcuni rapporti di lavoro presenti negli anni 1985 e 1986¹⁵. Tale mancanza non consentiva l'identificazione precisa dei rapporti di lavoro rappresentativi della popolazione di riferimento della base dati WHIP: questo produce nella versione attuale la pubblicazione di un numero maggiore di rapporti di lavoro rispetto alla precedente, sia per i settori totalmente coperti dalla fonte originaria (identificati dal valore 1 della variabile `episodi_whip`), sia per i settori coperti solo parzialmente (`episodi_whip` = 0).

Le figure seguenti, fig. 3.1-3.3, mostrano una sintesi del confronto tra le due versioni ricostruite delle dimensioni più influenzate dal processo di identificazione longitudinale dei rapporti di lavoro¹⁶: occupazione media, numero di associazioni, numero di separazioni e *gross worker turnover* (gwt)¹⁷. Da tale confronto emergono risultati che confermano la bontà della ricostruzione della versione 3.2, mentre le differenze osservate sono spiegate dalle scelte e dalle modifiche, migliorative, che sono state introdotte.

¹⁴ Per una descrizione della procedura di costruzione degli archivi si rimanda alla documentazione presente su sito WHIP.

¹⁵ Parallelamente la stessa operazione ha consentito il recupero anche dell'informazione sulla dimensione media di impresa mancante nella versione precedente (si veda il paragrafo 3.3).

¹⁶ Tutte le statistiche sono basate sulle informazioni dei rapporti di lavoro identificati come appartenenti alla "popolazione WHIP", ossia dove la variabile `episodi_whip` assume valore pari a 1. Per maggiori dettagli consultare i metadati on line e il paragrafo 1 sulle linee guida alle scelte di pubblicazione di WHIP.

¹⁷ Il *gross worker turnover* (gwt) è definito come rapporto tra la somma di associazioni e di separazioni avvenute nel periodo e lo stock di lavoratori. Per maggiori dettagli sulla costruzione di questo indicatore e sulla determinazione del numero di associazioni e separazioni si rimanda al capitolo di R. Leombruni e R. Quaranta "Mobilità dei lavoratori in Italia, 1985-1996: andamento ciclico e cambiamenti strutturali", in B. Contini (a cura di), **Osservatorio sulla mobilità del lavoro in Italia**, Il Mulino, Bologna, 2002.

Per quanto riguarda l'andamento dell'occupazione media dei lavoratori dipendenti, nei dati si osservano valori simili in entrambe le versioni negli anni dal 1987 al 1999, mentre sono presenti valori differenti solo ad inizio periodo, tra il 1985 e il 1986, come conseguenza del recupero delle informazioni utilizzate per l'identificazione dei settori di attività dei rapporti di lavoro presente nella nuova versione: questo ha reso possibile la pubblicazione di un numero di rapporti di lavoro prima non selezionati.

Nella versione 3.2 si osservano più associazioni e separazioni ad inizio periodo, tra il 1985 e il 1986, per effetto della presenza di rapporti di lavoro prima non presenti, mentre i valori più bassi, soprattutto per le separazioni, osservati tra il 1998 e il 1999 sono dovuti, da un parte, ad un miglioramento della procedura di identificazione longitudinale dei rapporti di lavoro, dall'altra, all'effetto dell'aggiornamento e al prolungamento delle date di cessazione di alcuni rapporti di lavoro in precedenza fatti cessare prima per mancanza di informazioni.

Infine, il *gwt* ha un andamento identico lungo tutto il periodo sovrapposto, quello che va dal 1987 al 1999; i valori maggiori nei primi due anni sono dovuti ad un aumento del numero di associazioni e separazioni in misura superiore all'aumento dello stock di occupati. L'andamento crescente del *gwt* iniziato dopo la crisi del biennio 1992-1993, toccando nel 1993 il valore minimo (56%), sembra arrestatosi nel 2001, con un turnover pari al 78%, mantenendosi successivamente tra il 2002 e il 2004 attorno al 77%, risultato, questo, decisamente più basso rispetto alle stime fatte sulle versioni provvisorie successive alla 2.3.

Figura. 3.1. Andamento dell'occupazione media dei lavoratori dipendenti. WHIP Full Edition

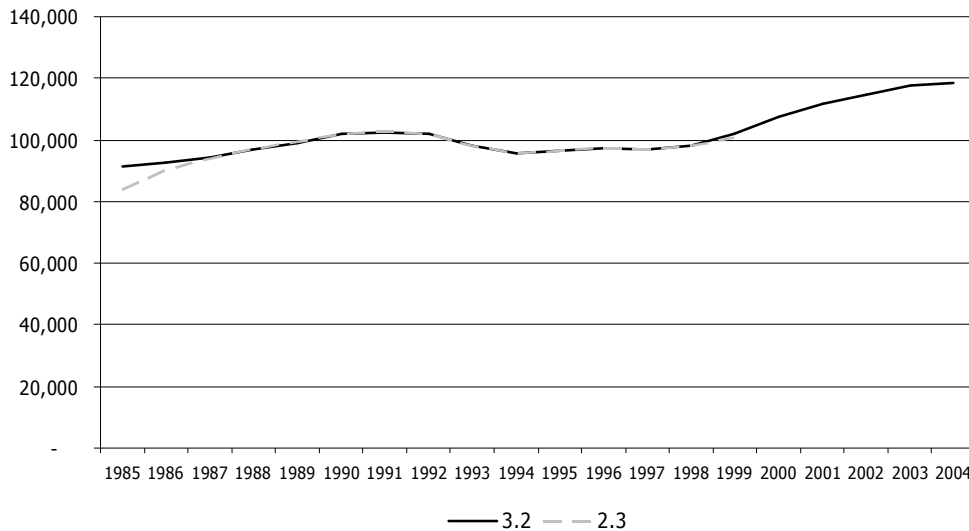
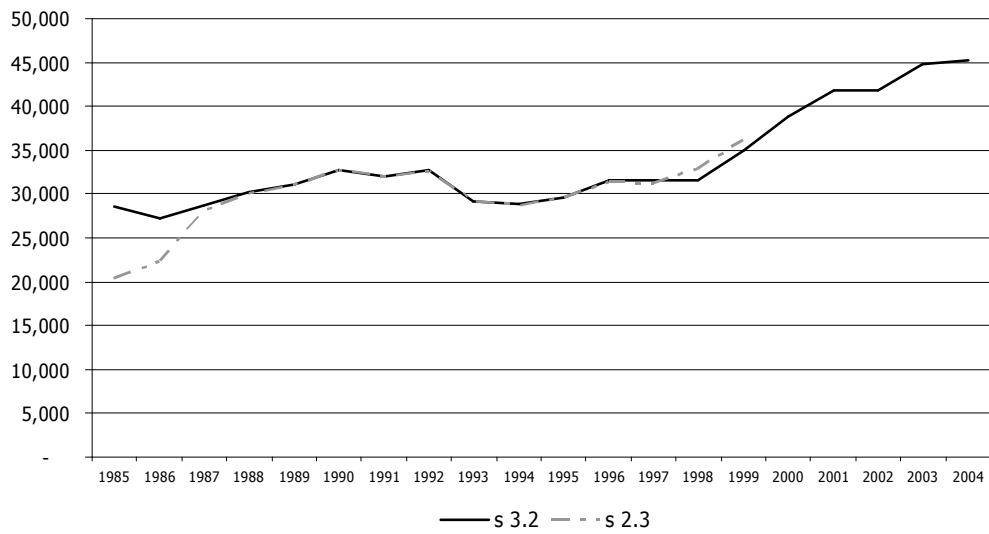
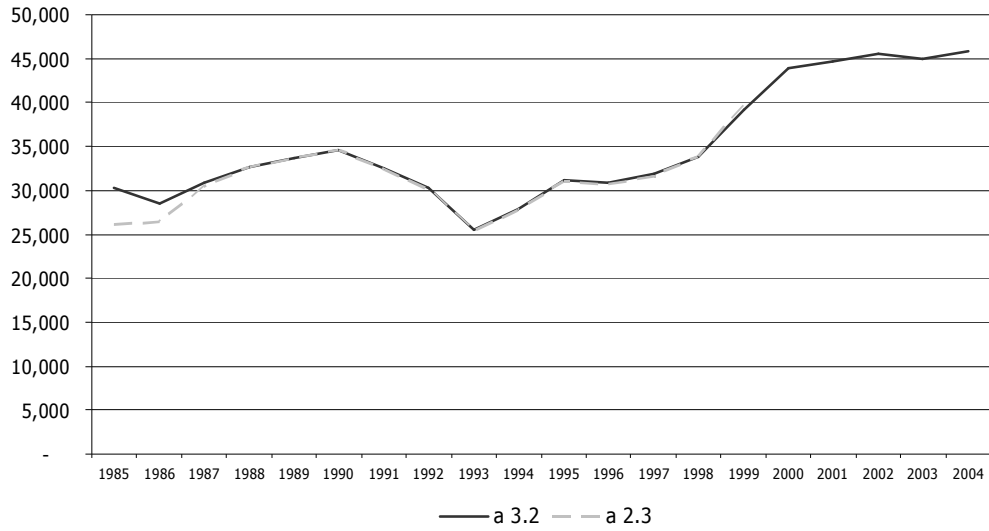
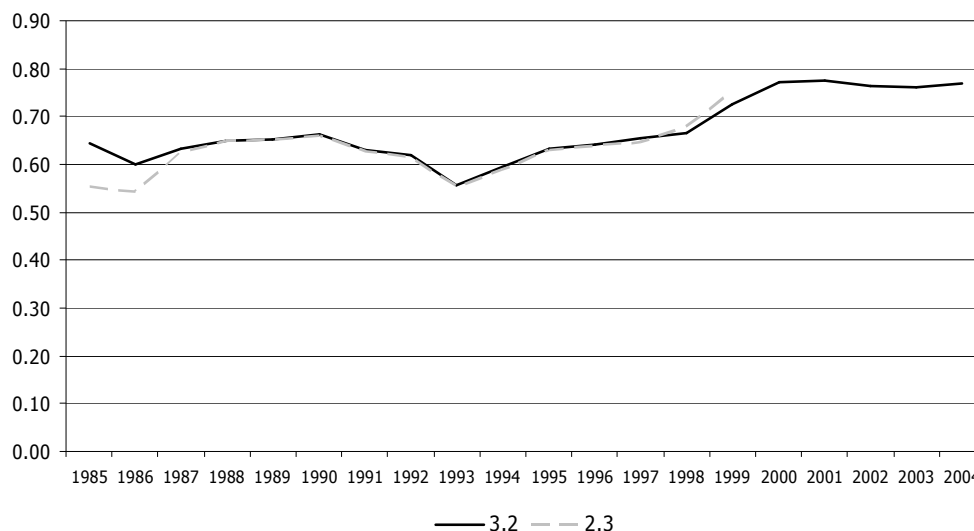


Figura 3.2. Numero di associazioni e separazioni
WHIP Full Edition



Nota: a = numero di associazioni, s = numero di separazioni.

**Figura 3.3. GWT, attuale e precedente versione
WHIP Full Edition**



3.2. Correzione delle giornate e delle settimane dei part-timers

In concomitanza con il nuovo rilascio si è provveduto alla correzione di un errore nella procedura riscontrato negli archivi degli episodi di lavoro dipendente della versione 2.3, nella quale veniva sottostimato il totale delle settimane e delle giornate lavorate (e quindi sovrastimate le retribuzioni unitarie) dei soggetti che durante l'anno avevano una trasformazione del contratto di lavoro da tempo pieno a tempo parziale, o viceversa.

Gli archivi che contengono i dettagli annuali dei rapporti di lavoro dipendente (archivi *RL_anno*) sono strutturati in modo da avere un record per ciascun rapporto attivo nell'anno. L'informazione presente in WHIP per un rapporto di lavoro che nel corso dell'anno modifica, senza interrompersi, una o più caratteristiche, quali la qualifica (ad es. passare da operaio a impiegato, variabile *qualifica*), o la durata dell'orario di lavoro (ad es. passare da tempo parziale a tempo pieno, variabile *ptime*), è quella relativa all'ultima modalità osservata in ordine di tempo; mentre le informazioni riguardanti la retribuzione, il trattamento di fine rapporto, le giornate e le settimane lavorate vengono consolidate.

Nella precedente versione, ogni rapporto di lavoro interessato da una trasformazione da tempo pieno a tempo parziale, o viceversa, riportava correttamente consolidata il valore della retribuzione annuale percepita e il trattamento di fine rapporto, mentre il numero di giornate e di settimane effettive lavorate si riferivano *solo* al periodo di lavoro che intercorreva prima della trasformazione: risultavano così un numero inferiore di settimane e di giornate lavorate rispetto a quante effettivamente svolte durante l'anno. Per queste persone, di conseguenza, si produceva una sovrastima delle retribuzioni unitarie (settimanali o giornaliere).

Come si può osservare nella tabella 3.1, dove sono riportati i valori medi della retribuzione, delle settimane e dei giorni lavorati, distintamente per i rapporti di lavoro a tempo pieno e parziale, e per quelli trasformati durante l'anno, la retribuzione media calcolata su tutto il periodo è molto simile in entrambe le versioni per tutte le tipologie di contratto, mentre il numero medio di giornate e di settimane lavorate è simile solo se si considerano i rapporti di lavoro che non hanno subito modifiche nel corso dell'anno. Per i rapporti di lavoro che hanno subito una trasformazione durante

l'anno i valori medi riportati dei giorni e delle settimane effettive lavorate nella versione precedente sono inferiori di circa il 60% rispetto alla versione attuale.

Sebbene la correzione apportata interessi solo un numero esiguo di rapporti di lavoro, il cui peso sul periodo 1985-1999 è solo dello 0,7%, la quota di episodi corretti risulta ben più elevata (7%) se si considerano soltanto i rapporti di lavoro a tempo parziale, mentre la quota di correzioni sui rapporti a tempo pieno è dello 0,3% (tab. 3.1).

L'influenza del mancato consolidamento delle giornate e delle settimane lavorate è evidente se guardiamo ai differenziali retributivi tra lavoratori a tempo pieno e a tempo parziale: sebbene il livello della retribuzione media settimanale e giornaliera sia sostanzialmente rimasta immutata, tra le due versioni (tab. 3.1) i livelli medi di retribuzione settimanale/giornaliera tra i lavoratori a tempo pieno e a tempo parziale differiscono: nella versione precedente erano i lavoratori a tempo parziale ad avere una retribuzione media settimanale o giornaliera superiore, nella versione attuale, quella corretta, questo dato si modifica considerevolmente (tab. 3.2)¹⁸.

Tabella 3.1. Retribuzioni medie, numero medio di giornate e settimane lavorate e consistenze dei rapporti di lavoro per tipologia di rapporto di lavoro e trasformazione

	Consistenze		Retribuzione media		Numero medio di settimane lavorate (sett_effettive)		Numero medio di giornate lavorate (gg_effettive)	
	2.3	3.2	2.3	3.2	2.3	3.2	2.3	3.2
Tutti	121,605	122,696	11,504	11,533	38	38	218	220
Tempo pieno	114,000	115,066	11,982	12,014	40	40	225	227
Tempo parziale	7,605	7,630	5,109	5,125	20	21	110	115
Non trasformati								
Tempo pieno	113,633	114,700	11,993	12,026	40	40	225	227
Tempo parziale	7,104	7,127	4,894	4,913	20	20	112	113
Trasformati								
in tempo pieno	366	366	8,841	8,837	13	36	73	204
in tempo parziale	502	503	7,996	7,993	14	33	78	185

Note: Statistiche sul periodo coperto da entrambe le banche dati - anni 1985 - 1999

¹⁸ Nonostante la correzione apportata, l'analisi delle retribuzioni e dei differenziali di reddito dei lavoratori dipendenti rimane comunque complessa soprattutto per i rapporti di lavoro a tempo parziale, in quanto i dati a disposizione non contengono le ore lavorate, ma il numero di giornate e di settimane lavorate, le quali vengono registrate dall'INPS secondo le regole amministrative specifiche, ossia viene considerata lavorata la giornata e la settimana in cui sia stata retribuita almeno un'ora di lavoro.

**Tabella 3.2. Retribuzione settimanale e giornaliera
per tipologia di rapporto di lavoro**

	Retribuzione settimanale		Retribuzione giornaliera	
	2.3	3.2	2.3	3.2
Tutti	290	285	55	52
Tempo pieno	290	288	54	52
Tempo parziale	306	258	61	51

Note: Statistiche sul periodo coperto da entrambe le banche dati - anni 1985 - 1999

3.3. Dimensione di impresa

Nella nuova *release* della base dati WHIP è presente l'informazione sulla dimensione di impresa nella quale il lavoratore dipendente è occupato per gli anni dal 1985 al 2002.

Rispetto alla precedente versione sono state recuperate le informazioni sulla dimensione media di impresa per gli anni dal 1985 al 1986: si registra ancora una copertura dei rapporti di lavoro non del tutto completa, dove la dimensione di impresa è al momento presente nell'83% di rapporti di lavoro nel 1985 e per il 93% nel 1986. Con l'aggiornamento delle informazioni sui rapporti di lavoro al 2004, il dato sulla dimensione di impresa è adesso presente fino al 2002, mentre non è al momento disponibile per gli anni dal 2003 al 2004, in quanto l'informazione non è presente negli archivi sorgente.

4. Le nuove variabili presenti

La versione 3.2 contiene alcune variabili non presenti nella versione precedente.

4.1 Variabile "episodi_whip"

È possibile che nell'archivio WHIP siano presenti alcuni rapporti di lavoro dipendente non riconducibili al settore privato. Tali rapporti di lavoro sono identificati dalla variabile `episodi_whip = 0`. La variabile assume valore 1 per tutti i rapporti di lavoro riconducibili al settore privato, perché l'informazione sul settore è presente ed è corretta, oppure in caso di informazione incongruente o mancante quando è stato possibile attribuire in modo univoco il rapporto di lavoro al settore privato.

Per i dettagli su questa variabile si veda la sezione 1 sulla popolazione WHIP.

4.2 Variabile "tipo_contratto"

Identifica la tipologia contrattuale con il quale il lavoratore è occupato:

- 10= contratto indeterminato
- 21= stagionale
- 22= tempo determinato
- 23= formazione lavoro
- 24= interinale
- 25= contratti di apprendistato

L'informazione sul tipo di contratto "stagionale" (codice 21), "tempo determinato" (codice 22), "interinale" (codice 24) è presente in WHIP a partire dal 1998, nei primi due casi poiché l'informazione è presente nei dati INPS sorgente solamente a partire da tale data; nel caso del contratto interinale poiché è stato istituito nel 1998.

Quindi per gli anni precedenti al 1998 è possibile identificare con esattezza solamente i contratti di apprendistato (codice 25) e di formazione lavoro (codice 23), mentre quelli stagionali e tempo determinato sono indistinguibili all'interno della categoria "contratto indeterminato" (codice 10).

4.3 Variabile “gg_retribuite” e “sett_retribuite”

La variabile indica il numero di giornate e di settimane per i quali è stata corrisposta una retribuzione soggetta a contribuzione obbligatoria nell'anno in corso. Sono considerate retribuite le settimane contenenti almeno una giornata retribuita.

4.4 Variabile “id_impresa”

La procedura di costruzione di WHIP costruisce un identificativo di impresa a partire dai dati sorgente di INPS. Per l'utilizzo di tale variabile è importante tenere in conto le seguenti precisazioni:

- i dati sorgente non sono organizzati per impresa (come viene normalmente intesa da un punto di vista economico) ma per posizione assicurativa (una sorta di codice che registra i versamenti dei contributi previdenziali da parte dell'impresa), e ogni impresa (soprattutto le più grandi) possono avere più posizioni assicurative non necessariamente riferite ai diversi stabilimenti (ad esempio un'impresa potrebbe decidere di versare i contributi relativi agli operai su una posizione assicurativa e quelli relativi agli impiegati su un'altra). Le procedure di costruzione di WHIP ricostruiscono le imprese aggregando le varie posizioni assicurative.
- Nel periodo i codici identificativi delle posizioni assicurative possono cambiare e può cambiare la composizione stessa dell'impresa per scorpori, fusioni e altre trasformazioni societarie, che la procedura di costruzione di WHIP tiene conto nell'identificare il rapporto di lavoro (si veda a questo proposito la descrizione della procedura di costruzione WHIP sul sito). Per questo motivo l'identificativo di rapporto di lavoro "ricuce" eventuali trasformazioni amministrative che non hanno effetto sulla natura economica della relazione lavoratore-impresa ma solo su quella giuridica.

In tutti i casi di trasformazione dell'impresa (come di flussi spuri) che non intaccano la continuità del rapporto di lavoro, l'identificativo del rapporto di lavoro non cambia, ma può cambiare quello di impresa.