DOCUMENTI ISTAT

n. 4/2007

Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770

V. Spinelli

Istat

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Le collane esistenti presso l'ISTAT - *Rivista di Statistica Ufficiale*, *Contributi ISTAT* e *Documenti ISTAT* - costituiscono strumenti per promuovere e valorizzare l'attività di ricerca e per diffondere i risultati degli studi svolti, in materia di statistica ufficiale, all'interno dell' ISTAT, del SISTAN, o da studiosi esterni.

La *Rivista di Statistica Ufficiale* accoglie lavori che hanno come oggetto la misurazione dei fenomeni economici, sociali, demografici e ambientali, la costruzione di sistemi informativi e di indicatori, le questioni di natura metodologica, tecnologica o istituzionale connesse al funzionamento dei sistemi statistici e al perseguimento dei fini della statistica ufficiale.

I lavori pubblicati in *Contributi ISTAT* sono diffusi allo scopo di stimolare il dibattito intorno ai risultati preliminari di ricerca in corso.

I *Documenti ISTAT* forniscono indicazioni su linee, progressi e miglioramenti di prodotto e di processo che caratterizzano l'attività dell'Istituto.

Il Comitato di redazione esamina le proposte di lavori da pubblicare nelle tre collane sopra indicate. Quelli pubblicati nei Contributi ISTAT e nei Documenti ISTAT sono valutati preventivamente dai dirigenti dell'Istituto, mentre i lavori pubblicati nella Rivista di Statistica Ufficiale sono subordinati al giudizio di referee esterni.

Direttore responsabile della Rivista di Statistica Ufficiale: Patrizia Cacioli

Comitato di Redazione delle Collane Scientifiche dell'Istituto Nazionale di Statistica

Coordinatore: Giulio Barcaroli

Membri:

Corrado C. Abbate	Rossana Balestrino	Giovanni A. Barbieri
Giovanna Bellitti	Riccardo Carbini	Giuliana Coccia
Fabio Crescenzi	Carla De Angelis	Carlo M. De Gregorio
Gaetano Fazio	Saverio Gazzelloni	Antonio Lollobrigida
Susanna Mantegazza	Luisa Picozzi	Valerio Terra Abrami
Roberto Tomei	Leonello Tronti	Nereo Zamaro

Segreteria: Gabriella Centi, Carlo Deli e Antonio Trobia

Responsabili organizzativi per la *Rivista di Statistica Ufficiale*: Giovanni Seri e Carlo Deli

Responsabili organizzativi per i *Contributi ISTAT* e i *Documenti ISTAT*: Giovanni Seri e Antonio Trobia

ISTAT – Produzione libraria e centro stampa

DOCUMENTI ISTAT

n. 4/2007

Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770

V. Spinelli(*)

(*) ISTAT - Servizio Istituzioni pubbliche e private

Contributi e Documenti Istat 2007

Istituto Nazionale di Statistica
Servizio Produzione Editoriale

Produzione libraria e centro stampa:
Carla Pecorario
Via Tuscolana, 1788 - 00173 Roma

Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770

1 Introduzione

Questo report è stato predisposto allo scopo di fornire una descrizione puntuale del processo di trattamento preliminare degli archivi dei modelli di dichiarazione fiscale 770 (M-770). Le considerazioni espresse nel documento rappresentano la sintesi delle esperienze maturate nell'ambito del progetto *"Trattamenti Monetari Non Pensionistici"* ($TMNP$) (ci si riferisca ai lavori [CO00], [CO02] per avere un quadro definitorio e normativo completo del progetto), svolto nel Servizio *"Istituzioni Pubbliche e Private"* (SIP).

In particolare vengono descritti i risultati relativi alla prima parte del processo di produzione, quello in cui gli archivi elettronici acquisiti dall'Istat vengono trasformati in basi dati utili per l'avvio della produzione delle informazioni statistiche[1].

Nel documento si utilizzano le convenzioni e le codifiche riportate nelle specifiche tecniche che descrivono gli archivi amministrati acquisiti ed utilizzati nel progetto $TMNP$. Si rimanda a tali specifiche qualora si fosse interessati ad un approfondimento dei temi descritti in queste pagine.

1.1 Definizioni preliminari

Nell'ambito del progetto $TMNP$, si possono individuare i seguenti attori principali del flusso informativo rappresentato dagli archivi M-770:

- *"Agenzia delle Entrate"* ([AGEN]), svolge tutte le funzioni ed i compiti ad essa attribuiti dalla legge in materia di entrate tributarie e diritti erariali, al fine di perseguire il massimo livello di adempimento degli obblighi

[1] L'impostazione dell'intero processo di lavorazione e l'elaborazione degli anni $2001 - 2003$ si deve attribuire esclusivamente all'autore. L'elaborazione ed i risultati relativi all'anno 2004 si devono attribuire al lavoro congiunto dell'autore con Lamberto Pizzicannella.

fiscali. Per questo motivo detiene gli archivi amministrativi relativi alle dichiarazioni fiscali che sono di interesse per questo documento. L'Agenzia delle Entrate è il soggetto istituzionale a cui ci si è rivolti per l'acquisizione degli archivi M-770.

- *"Sogei"* ([SOGE]) è la società che supporta tecnologicamente l'Agenzia delle Entrate (*http* : *//www.sogei.it/attivita/index.htm*). La Sogei acquisisce ed elabora per conto dell'Agenzia delle Entrate gli archivi M-770, quindi è il soggetto di riferimento per la definizione ed attuazione delle procedure di acquisizione degli archivi M-770 in Istat.
- *"Servizio SIP"* è la struttura organizzativa in cui il progetto $TMNP$ si svolge. Il servizio SIP ha acquisito gli archivi M-770 in formato grezzo dalla Sogei, ed ha implementato una sequenza di procedure automatiche per renderli fruibili ai processi di produzione statistica.

I modelli di dichiarazione fiscale 770 [M770] sono annuali ed è necessario distinguere tra:

- *Anno di dichiarazione*: indica l'anno in cui è compilata e presentata la dichiarazione da parte del sostituto d'imposta.
- *Anno di riferimento o imposta*: indica l'anno a cui si riferiscono i dati inseriti nella dichiarazione.

L'anno di dichiarazione e l'anno di riferimento differiscono di un anno (es. anno di dichiarazione 2002 implica che l'anno di riferimento sia 2001, e viceversa). Nel seguito si farà riferimento all'anno di dichiarazione.

1.2 Stato dell'arte

Le considerazioni presentate in questo lavoro si riferiscono ai risultati ottenuti nel trattamento di archivi M-770 relativi a quattro annualità:

- L'anno 2001 è stato acquisito dalla Sogei come un archivio in formato binario e l'elaborazione informatica è stata completata.
- Gli anni $2002-2003$ sono stati acquisiti dalla Sogei come archivi in formato testo e sono stati anch'essi elaborati completamente.
- L'anno 2004, acquisito in formato testo compresso, è attualmente in fase di elaborazione.

Questo significa che le tabelle presentate in questo report non sempre riportano gli indicatori relativi al 2004, inoltre l'archivio relativo all'anno 2001 ha subito un trattamento iniziale diverso da quello relativo agli anni $2002-2003$, come spiegato nel paragrafo 2. Nelle tabelle presentati in questo documento, i dati mancanti sono stati indicati con *"-"* e significa che il valore non è ancora disponibile (elaborazioni ancora in essere) ovvero che non è definito per la cella considerata.
Come sperimentazione iniziale, il servizio SIP ha acquisito un campione dell'archivio 770 relativo all'anno 2000, la cui dimensione è di circa $22,000$

aziende. I risultati del'elaborazione di questo archivio non vengono descritti in questo report poichè non omogenei (in dimensione ed ampiezza di variabili) e quindi anche non confrontabili con quelli degli archivi successivamente acquisiti.

Fino al 2001 vi era una sola versione del modello 770 che conteneva 22 quadri, mentre dal 2002 vi sono due versioni del modello: ordinario e semplificato. Il modello ordinario contiene 13 quadri, mentre quello semplificato ne contiene solo 5. Ai fini del progetto $TMNP$ sono stati considerati solo gli archivi relativi al modello semplificato, perché contiene i dati relativi al lavoro dipendente ed autonomo.

Allo stato attuale non esiste una struttura standard degli archivi fissata in base ad un protocollo di trasmissione dei dati dalla Sogei all'Istat. Questo ha comportato la definizione e la implementazione di processi autonomi di armonizzazione e trasformazione degli archivi in ingresso. Le procedure relative a questi processi sono l'oggetto di questo report.

2 Archivi Iniziali

Gli archivi M-770 vengono acquisiti dalla *Sogei* con cadenza annuale. Gli archivi inviati da *Sogei* si trovano in uno stato grezzo, in quanto provengono dagli archivi di *"raccolta telematica delle dichiarazioni"* della *Sogei* stessa, e non hanno subito alcun trattamento preliminare.

Ogni anno il modello di dichiarazione 770 viene aggiornato, quindi anche le basi dati che si riferiscono a questo modello non sono costanti nel tempo.

Questa situazione di continua evoluzione delle informazioni presenti negli archivi informatici ha portato alla definizione in Sogei di un tracciato record, valido per le operazioni di rilascio degli archivi verso l'esterno, con una struttura che possiamo definire: *"tracciato variabile con campi a serrare"*. Questa struttura permette di rappresentare tracciati diversi nello stesso archivio e di rappresentare solo i campi effettivamente valorizzati nella dichiarazione.

Anche i supporti informatici utilizzati per il trasferimento degli archivi da *Sogei* ad Istat sono cambiati nel tempo. In particolare, la trasmissione degli archivi avviene attualmente su supporti ottici (DVD) contenenti un numero variabile di file di testo compressi, mentre precedenti acquisizioni sono avvenute utilizzando supporti magnetici. La tabella 1 illustra il passaggio da supporti magnetici (nastri $IBM3490$) a supporti ottici.
La colonna "Totale file" della tabella 1 indica il numero effettivo di file che si ottengono dopo lo scarico dal supporto ottico ovvero magnetico.
Gli archivi M-770 sono archivi amministrativi e sono quindi soggetti a ben precise dinamiche di popolamento come nel caso degli archivi *DM10* dell'Inps [DM10]. La tabella 2 fornisce una prima approssimazione sulle dinamiche di popolamento valide per gli archivi M-770, anche se non è stato ancora fatto uno studio puntuale in Istat su questo argomento. In particolare la tabella 2 mostra alcuni punti importanti:

Tabella 1. Prospetto riassuntivo della trasmissione dei Modelli 770

Anno di dichiarazione	Totale file	Totale DVD	Totale nastri
2001	28	0	28
2002	36	36	0
2003	36	36	0
2004	261	5	0

Tabella 2. Prospetto riassuntivo degli archivi iniziali

Anno di dichiarazione	Dimensione dati (MB)	Numero di record
2001	69,293	5,146,903
2002	93,707	51,609,229
2003	94,400	51,990,437
2004	130,828	64,694,126

- L'archivio relativo al 2001 ha dimensioni diverse da quelle degli anni successivi: il numero di record è pari a $\frac{1}{10}$ rispetto all'anno 2002, mentre la dimensione complessiva è pari a $\frac{2}{3}$ rispetto sempre al 2002. Queste differenze possono essere spiegate sia con il diverso modo di rappresentare i dati nell'archivio e sia con il miglioramento del livello di popolamento degli archivi passando da un anno all'altro.
- L'archivio relativo all'anno 2004, rispetto all'anno 2003, presenta la stessa discontinuità descritta per l'anno 2001. In questo caso il fenomeno è spiegabile dicendo che dall'anno di dichiarazione 2004 vengono acquisiti i quadri relativi al lavoro autonomo, non presenti nelle precedenti acquisizioni.

In alcuni casi (es. anni 2002 e 2003) è stato necessario un ulteriore scarico per integrare informazioni mancanti (sezione anagrafica, frontespizio) come evidenziato in tabella 3.

Tabella 3. Prospetto riassuntivo delle trasmissioni integrative

Anno di dichiarazione	Totale file dati	Totale file anagrafica
2001	0	0
2002	1	36
2003	1	36
2004	0	0

Nella tabella 4 vengono descritti la distribuzione, come numero di record, dei vari quadri per ogni anno.
La colonna "Altro" della tabella 4 indica tutti i record presenti nell'archivio Sogei ma che non sono oggetto di elaborazione.

Tabella 4. Prospetto riassuntivo dei quadri acquisiti

Anno di dichiarazione	Frontespizi	Lavoro Dipendente	Autonomi	Altro
2002	1,863,384	49,745,832	0	13
2003	1,914,331	50,076,096	0	10
2004	3,784,447	48,156,395	12,753,212	72

2.1 Conversione Preliminari per Archivi binari

I file inviati da Sogei sono file di testo (quindi in formato ASCII [ASCII] per ambiente DOS) con una eccezione per l'anno 2001 che sono file binari (formato EBCDIC [EBCD]). Per questo motivo è stato necessario per l'anno 2001 un passo preliminare di conversione da file binario a file di testo [CONV]. La tecnica di conversione utilizzata deriva dal programma "dd" (presente nella suite GNU/fileutils), per la definizione del vettore di transcodifica di un byte in formato EBCDIC ad uno in formato ASCII. Oltre alla conversione di formato, il programma "ebc2asc" effettua una prima pulizia dei codici EBCDIC che non hanno generato un valore nella codifica ASCII stampabile, come definito dalla funzione "isprint()" (presente in "<ctype.h>") del linguaggio C. L'azione di pulizia consiste nella sostituzione dei caratteri cosí individuati con il carattere standard "ASCII BLANK" (codifica decimale 032/esadecimale 20). Infine, il programma "ebc2asc" è definito sulle caratteristiche fisiche dei file binari SOGEI (es. block size = 28200 byte, length record = 14100 byte, $RECFM = FB$).

L'ipotesi di fondo su cui si basa questa tecnica di conversione è quella che i dati sono rappresentati in formato *"DISPLAY"*, ma non sempre questa ipotesi risulta essere confermata. Esistono alcune variabili numeriche che sono memorizzate in formati compressi (*es. COMP, COMP1*) e quindi non passano il filtro della funzione isprint(). Dalle analisi effettuate, le variabili che presentano questo inconveniente non sono rilevanti ai fini del progetto $TMNP$, quindi si è proceduto comunque alla conversione anche se vi è stata una perdita di informazione (stimata al 10% - 20%) dell'archivio di partenza. Qualora l'archivio relativo all'anno di dichiarazione 2001 venga utilizzato per scopi statistici diversi da quelli del progetto $TMNP$ è necessario creare un programma di conversione, preferibilmente scritto in linguaggio COBOL, che permetta di convertire correttamente in formato testo tutte le variabili non in formato DISPLAY.

2.2 Conversioni Preliminari per Archivi di testo

Gli archivi relativi agli anni $2002 - 2004$, che non sono sottoposti al filtro del programma *"ebc2asc"*, devono comunque passare una prima fase di conversione e filtraggio analoga a quella del paragrafo 2.1. Una analisi dei file provenienti da *Sogei* ha evidenziato che vi sono sostanzialmente due filtri da eseguire:

- *File testo DOS*: in ambiente Unix-Linux i file di testo provenienti da ambiente DOS, sono caratterizzati dalla presenza di alcuni caratteri non più necessari. In particolare, i caratteri da eliminare sono: 'CTRL-M' alla fine di ogni riga e 'CTRL-Z' alla fine del file. Questa operazione può essere effettuata con il comando "dos2unix" [DOS2].
- *Caratteri non stampabili*: i caratteri con accenti non sono sempre correttamente convertiti nel passaggio da codifiche diverse (es. *EBCDIC* ed *ASCII*); questo potrebbe comportare la presenza di caratteri di fine riga (valore esadecimale '0') in posizioni non corrette. La presenza di questo tipo di problema è generalmente legata alla trascrizione di parole non italiane (es. nomi di persone ed indirizzi).

Per evitare che si verifichino casi di righe troncate per la presenza dei caratteri non stampabili è stato implementato un apposito programma che rimuove queste anomalie. Il programma si basa sull'ipotesi che una qualsiasi riga dell'archivio *M*-770 deve iniziare con la sequenza di caratteri "IST", come da specifiche di tracciato rilasciate da *Sogei*. Questo significa che tutte le righe che non soddisfano questa condizione devono essere considerate troncate, come precedentemente descritto, e quindi "unite" alla riga immediatamente precedente. I risultati di queste operazioni preliminari di pulizia fisica degli archivi sono sintetizzati in tabella 5.

Tabella 5. Tabella riassuntiva sui risultati dei primi controlli

Anno di dichiarazione	Righe interrotte	Caratteri sporchi
2002	3	84,153
2003	0	83,517
2004	8	95,201

Dalla tabella 5, si osserva che il problema delle righe interrotte è sicuramente meno importante di quello dei caratteri sporchi. Questo aspetto quantitativo non deve però far dimenticare che entrambe le situazioni possono determinare situazioni di blocco dei programmi durante il trattamento dei dati.

Inoltre la dimensione del fenomeno è sostanzialmente legata in modo lineare alle dimensioni dell'archivio considerato, e questo implica che non ci si aspetta che il fenomeno esploda nei prossimi anni acquisendo archivi che sono maggiormente popolati.

3 Tracciato Archivi dei Modelli 770

I dati rappresentati negli archivi *M*-770 sono legati al modello di dichiarazione e quindi all'anno considerato. I cambiamenti principali noti allo stato attuale, sono i seguenti:

- *Anno 2002*: il modello 770 nel 2002 si divide in due versioni: semplificato ed ordinario. Nella versione semplificata vi sono tutte le informazioni di interesse per il progetto $TMNP$.
- *Anno 2006*: il modello semplificato 770 nel 2006 ha la parte relativa ai trattamenti assistenziali e previdenziali Inps molto ridotta, poichè nel gennaio 2005 è entrata in vigore la dichiarazione mensile *E-Mens* che viene inviata direttamente all'Inps; in ogni caso la parte previdenziale Inps del modello 770 contiene informazioni sufficenti per creare un legame con gli archivi *E-Mens* [MENS]. L'archivio relativo all'anno 2006 non è ancora disponibile, ma questi cambiamenti ci permettono di dire che il processo di elaborazione descritto in questo documento deve essere integrato con la parte relativa all'elaborazione degli archivi *E-Mens*.

Questi sono i cambiamenti della struttura della dichiarazione. Le modifiche minori (es. inserimento/modifica/cancellazione di singole variabili) sono presenti ogni anno. Gli archivi provenienti dalla Sogei sono caratterizzati da un tracciato record variabile con campi a serrare che offrono un doppio vantaggio:

- *Variabilità dei M-770*: ogni anno i modelli di dichiarazione vengono aggiornati e questi aggiornamenti possono avere un impatto notevole sulla struttura degli archivi.
- *Standardizzazione delle forniture dei dati*: la progettazione dei processi automatici di trattamento degli archivi M-770 richiede dati statici nel tempo, anche con un tracciato non fisso ma con delle strutture riconoscibili in un opportuno linguaggio regolare (es. espressioni regolari).

Queste considerazioni hanno portato alla definizione di uno schema di tracciato in cui vi sono parti variabili e parti fisse e, per ridurre l'occupazione di spazio, i singoli campi sono presenti solo se valorizzati.

Un modello M-770 si compone di un insieme di prospetti/quadri, ognuno dei quali generano nell'archivio M-770 un distinto tracciato. Ogni tracciato può avere sia una parte fissa sia una parte variabile, e si collega con gli altri tracciati attraverso alcune variabili presenti nella sola parte fissa (es. identificativo telematico). Ai fini del progetto $TMNP$, sono stati presi in considerazioni solo alcuni prospetti e quindi il numero di possibili tracciati record che sono stati analizzati sono i seguenti:

- *Record di tipo A*: è presente una sola volta nella fornitura (primo record del primo file), ed indica le coordinate della fornitura stessa (es. identificativo, progressivo, data di creazione). Per questo motivo viene ignorato durante le elaborazioni. Qualora la fornitura è divisa in blocchi, i record di tipo A sono presenti in ogni blocco.
- *Record di tipo B*: descrive le informazioni relative al sostituto d'imposta e al rappresentante firmatario della dichiarazione. Questo tipo di record viene comunemente indicato come frontespizio della dichiarazione e lo si può considerare come una anagrafica dell'azienda.

- *Record di tipo G*: descrive i dati relativi alla certificazione del lavoro dipendente (anagrafica, dati fiscali, dati previdenziali ed assistenza fiscale).
- *Record di tipo H*: è l'analogo del record G per la certificazione del lavoro autonomo, provvigioni e redditi diversi.
- *Record di tipo Z*: è l'analogo del record A per la chiusura della fornitura (ultimo record dell'ultimo file), contenente i dati di sintesi (es. numero delle righe per ogni tipo record).

Tutti i record considerati hanno una parte fissa, mentre solo i record G ed H hanno anche una parte variabile definita dall'insieme dei campi non posizionali presenti nel prospetto relativo. Dalle successive analisi verranno ignorati i record di tipo A e Z, ininfluenti ai fini del trattamento dell'archivio.

3.1 Struttura della parte fissa

Esiste una sezione iniziale comune a tutti i tipi di record considerati che viene utilizzata per creare il legame tra i record presenti nella fornitura. In particolare vengono utilizzati i primi 36 caratteri di questa sezione comune con la seguente struttura:

- *Identificativo di fornitura*: ha un lunghezza di 8 caratteri e si compone di tre valori costanti "IST", "7" e "2004". Il primo valore segna l'inizio di ogni riga e viene usato nella procedura di recupero delle righe troncate descritto nel paragrafo 2.2; il terzo valore indica l'anno della dichiarazione (il valore 2004 si riferisco all'ultimo anno in lavorazione). Il secondo valore indica che si tratta di un modello 770; è necessario osservare che solo nella fornitura dell'anno 2001 vengono differenziati vari tipi di modelli anche se tutti hanno la stessa struttura dati: "U" Unico persone fisiche, "5" modello 750, "6" modello 760, "8" modello 760BIS e "7" modello 770.
- *Progressivo di fornitura*: indica un progressivo nella sequenza dei file della fornitura, ma è una informazione non necessaria.
- *Identificativo telematico* ($IdTel$): è una sequenza alfanumerica di 25 caratteri, che serve per identificare univocamente i record relativi ad una singola dichiarazione. Esso viene attribuito univocamente in base all'ordine di ricezione della dichiarazione stessa da parte del sistema informatico dell'Agenzia delle Entrate. Questo valore è fondamentale per differenziare due dichiarazioni della stessa azienda nello stesso anno.
- *Tipo record*: è un campo di un solo carattere che permette di individuare il tipo di record considerato (es. *"B", "G" ed "H"*).

Il campo *"tipo record"* viene utilizzato per leggere ed interpretare correttamente le informazioni contenute nella restante parte del record.

3.2 Struttura della parte variabile

I processi di lavorazione degli archivi filtrano i record di interesse sulla base del campo *"tipo record"*. I casi che si possono presentare sono sostanzialmente

di due tipi: solo tracciato fisso (tipo record="B") ovvero fisso+variabile (tipo record="G","H"). Il caso più semplice è quello a tracciato solo fisso dove ogni variabile è caratterizzato da una terna di informazioni *(posizione iniziale, lunghezza, tipo di variabile)* note a priori. Questo significa che il programma che si occupa di estrarre le informazioni relative al frontespizio di una azienda legge in modo sequenziale le variabili relative (circa 200). Il caso del tracciato variabile necessita di considerazioni più approfondite. Le variabili normalmente valorizzate nei quadri "G" ed "H" costituiscono una percentuale esigua rispetto a tutte quelle presenti sul modello, ed è questo il motivo che ha portato alla definizione di un tracciato variabile con campi a serrare; questa organizzazione implica che una variabile è presente nell'archivio solo se essa è effettivamente valorizzata nel record considerato. Questa impostazione implica che non sia nota a priori la sequenza di variabili presenti su ogni record e quindi è stato necessario identificare ogni variabile tramite una quadrupla *(sezione,riga,colonna,valore)* con il seguente significato:

- *Sezione*: il quadro relativo al lavoro dipendente e/o autonomo è diviso in sezioni. Ogni sezione è codificata in 2 caratteri (es. "DA","DC","AU"), la cui decodifica è fornita con le specifiche della fornitura dati dalla *Sogei*.
- *Riga*: le informazioni relative ad una persona fisica, sia essa lavoratore dipendente sia autonomo, possono richiedere più fogli della dichiarazione per essere descritte (es. cambi di qualifica lavorativa durante l'anno). Per distinguere ogni foglio della dichiarazione di un singolo dipendente viene utilizzato un campo riga. Questo campo è lungo 3 caratteri.
- *Colonna*: ogni variabile presente nella dichiarazione, tranne per il frontespizio, è stata opportunamente numerata e questo numero viene definito come colonna. Questo campo è lungo 3 caratteri.
- *Valore*: i valori effettivi indicati nella compilazione del modello 770 vengono riportati nel campo valore. L'interpretazione di questo valore dipende dalle altre tre componenti. La lunghezza di questo campo è pari a 16 caratteri.

La coppia $(Sezione, Colonna)$ individua l'etichezza di una variabile, mentre $Valore$ indica l'effettivo valore assunto dalla variabile nel record considerato. Infine, le quadruple descritte sono presenti a partire dalla colonna 273 di ogni record di tipo G ed H.

Infine, le quadruple descritte sono presenti a partire da una specifica colonna di ogni record di tipo G ed H; questa colonna dipende dal tracciato e quindi dall'anno considerato (es. per l'anno 2004 si considera la colonna 273).

4 Strutturazione degli Archivi 770

Gli archivi forniti dalla *Sogei* non sono immediatamente utilizzabili in un processo di produzione e/o analisi statistica. Per rendere fruibile l'archivio M-770 ai ricercatori statistici è quindi necessario definire una struttura di

archivio diversa da quella acquisita dalla *Sogei* ed implementare una procedura di conversione di archivi. A partire dalla struttura delle informazioni fornite da Sogei, è stato disegnato uno schema relazionale basato sulle seguenti caratteristiche:

- *Omogeneità delle tabelle*: l'archivio sorgente è un insieme eterogeneo di tracciati non fissi.
- *Stabilità delle informazioni*: ogni anno la struttura informativa dei modelli è rivista e ció incide sulla quantità e la natura delle informazioni in esso presenti in ciascun anno.

Il modello relazionale degli archivi M-770 definito ed utilizzato nel servizio SIP, si basa su alcuni assunti:

- Ogni dichiarazione è univocamente identificata dal campo $IdTel$: cosí nel caso di dichiarazioni multiple (aziende che presentano più dichiarazioni relative allo stesso anno) si hanno valori diversi di $IdTel$ e quindi ogni $IdTel$ è attribuibile ad una sola azienda.
- In ogni dichiarazione deve essere presente uno ed un solo frontespizio, mentre le sezioni relative al lavoro dipendente ovvero autonomo sono presenti in numero variabile (in termini relazionali la notazione $[0, n]$ indica questo tipo di relazione).
- Ogni sezione ha una struttura informativa variabile nel corso degli anni.

L'ultima considerazione impone la definizione di una struttura flessibile, modificabile in parte o del tutto nel corso degli anni.

Il modello proposto è semplice ma allo stesso tempo si è dimostrato flessibile per le elaborazioni statistiche. Le tabelle possono essere descritte con un approccio top-down:

- *Ogni anno* esprime un gruppo di tabelle.
- *Ogni gruppo di tabelle* è composto da una tabella frontespizio, una tabella anagrafica per i percettori di somme, una tabella per il lavoro dipendente e una tabella per il lavoro autonomo (presente a partire dall'anno 2004).
- *La tabella frontespizio* rappresenta le informazioni che descrivono l'azienda che presenta una dichiarazione, quindi può considerarsi in prima approssimazione come una anagrafica aziende per l'anno considerato. La tabella è strutturata in una parte fissa ed una variabile:
  - *La parte fissa* è comune a tutti gli anni ed è composta dalle variabili: *anno, modello, idtel, codice fiscale azienda, divisa,(flag di compilazione), flag di validazione del codice fiscale azienda*. Il campo *divisa* indica la valuta in cui è espressa la dichiarazione; questa indicazione è utile per il periodo in cui era possibile avere dichiarazioni sia in euro sia in lire. *I flag di compilazione* indicano se ci troviamo in presenza di dichiarazioni particolari (correttiva, integrativa ed eventi eccezionali). *Il flag di validazione* è presente a partire dall'anno 2004.

  - *La parte variabile* si articola nella lista dei campi presenti nella dichiarazione dell'anno considerato. I nomi dei campi sono stati codificati in modo omogeneo con il tracciato della Sogei, indicando la posizione nel tracciato fisso a cui si riferisce (es. *D*0447 è la variabile del frontespizio che parte da posizione 447).
- *La tabella anagrafica* contiene i dati identificativi dei percettori di somme presenti nei modelli di dichiarazione. L'acquisizione dei quadri relativi al lavoro autonomo ha comportato la presenza nell'anagrafica di informazioni relative sai a persone fisiche sia ad altri soggetti; quindi le tabelle anagrafiche sono state strutturate in modo diverso a seconda degli anni di riferimento:
  - *Anni* 2001-2003: in questi anni sono presenti solo lavoratori dipendenti, quindi persone fisiche; ogni persona fisica viene descritta con il seguente schema: *codice fiscale, codice fiscale corretto, cognome, nome, dettagli sulla nascita (comune, provincia, data), dettagli sulla residenza (comune, provincia, CAP, indirizzo)*. Nel caso vi siano cambiamenti di residenza nel corso dell'anno, sono presenti più record nella stessa tabella ma non è specificato il periodo in cui è da considerarsi effettiva la singola residenza. Infine, la determinazione delle variabili *(sesso,etá)*, non è possibile in automatico quando sia nel campo codice fiscale, sia in quello corretto è presente una partita IVA.
  - *Anno* 2004: nella stessa tabella sono descritti sia le persone fisiche sia quelle giuridiche; i due casi hanno una parte comune costituita dalla coppia *(codice fiscale, flag di validazione)*. A questa parte comune si affiancano due sezioni valorizzate in modalitá esclusiva:
    - *persone fisiche*: informazioni analoghe a quelle valide per gli anni 2001-2003.
    - *persone giuridiche*: è composta da *denominazione, forma giuridica, domicilio fiscale (comune, provincia, indirizzo)*.
- *La tabella del lavoro dipendente* si articola anch'essa in una parte fissa ed una variabile:
  - *La parte fissa* è comune a tutti gli anni ed è composta dalle variabili: *anno, idtel, codice fiscale dipendente, progressivo di dichiarazione*. il campo idtel serve per risalire al frontespizio della dichiarazione ed a raggruppare tutte le pagine di una stessa dichiarazione relative ai lavoratori dipendenti. il progressivo di dichiarazione specifica quale foglio di dichiarazione per il singolo dipendente si considera. Nella maggior parte dei casi esiste un solo foglio per ogni dipendente, quindi il valore è quasi sempre posto uguale ad 1.
  - *La parte variabile* si struttura in gruppi di variabili. Questi gruppi possono essere considerati abbastanza stabili nel tempo, ma la loro composizione è sicuramente mutevole. I gruppi di variabili presenti (A=anagrafica dipendente, B=dati fiscali, C=dati previdenziali ed assistenziali, D=dati relativi all'assistenza fiscale) seguono i raggruppamenti indicati sul modello *M*-770. In ogni raggruppamento, le variabili

sono identificate dalla sigla del raggruppamento (DA, DB, DC, DD) e dal progressivo indicato nel modello di dichiarazione (es. $DB001$ indica generalmente i redditi per cui è possibile fruire delle deduzioni).

- *La tabella del lavoro autonomo* ha una struttura simile a quella del lavoro dipendente, con una parte fissa ed una variabile. La parte fissa si struttura come quella del lavoro dipendente, mentre la parte variabile è più semplice[2] (sempre rispetto a quella del lavoro dipendente).

4.1 Tabelle fisiche

L'insieme delle tabelle relative al M-770 possono essere viste in una struttura gerarchica che ripropone tutti i concetti fin qui espressi:

- *Archivio*: tutte le tabelle relative al M-770 hanno il prefisso **M770**.
- *Anno di dichiarazione*: ogni tabella ha l'esplicita indicazione dell'anno di dichiarazione sia nel nome e sia nel tracciato (es. **2001, 2002, 2003, 2004**).
- *Sezione*: ogni sezione e/o quadro di cui si compone la dichiarazione 770, e che viene considerata in questo progetto, ha una codifica univoca (es. **B, G, H**).
- *Gruppo di variabili*: le variabili di ogni sezione hanno un prefisso ben preciso (es. **DA, DB, DC, DD**) ed un progressivo (es. **000, 001, 002**)

Tutti i codici utilizzati per specificare questa gerarchia sono state mutuati direttamente dagli archivi Sogei, al fine di avere sempre una tracciabilità con gli archivi di origine.
Analoghe considerazioni possono essere fatte con altri archivi amministrativi gestiti nell'ambito del progetto $TMNP$; in particolare si possono considerare gli archivi mensili DM10 dell'Inps che presentano una struttura a quadri.

4.2 Vincoli di Integrità

L'insieme delle tabelle risultanti risultano essere legati tra loro da pochi vincoli di integrità referenziale, in quanto esse risultano essere l'immagine più fedele possibile agli archivi di origine, preservando le informazioni contenute. Questo significa che non vengono risolti le condizioni di duplicazione/mancanza delle dichiarazioni in quanto questi controlli vengono demandati alla fase di controllo e validazione statistica degli archivi. In definitiva queste tabelle rappresentano le informazioni ancora ad un livello di semilavorato, in cui si ritengono effettuati solo alcuni controlli: eliminazione di errori sintattici, ricostruzione della sequenza logica delle variabili, formattazione dei valori secondo formati

[2] Esiste un solo gruppo di variabili identificate dalla sigla di raggruppamento AU e da un indice progressivo come riportato nel modello di dichiarazione (es. AU016 indica la causale dell'erogazione delle somme per l'anno 2004).

uniformi (es. i campi data hanno tutti lo stesso formato *"YYYY-MM-DD"* alla fine del processo di lavorazione).

Esistono alcune variabili che permettono di legare tra loro gruppi di variabili e si possono considerare il punto di partenza delle successive elaborazioni statistiche

- *Identificativo telematico*: questa variabile lega un singolo record nella tabella frontespizio a tutti i record di dettaglio presenti nelle altre tabelle (lavoro dipendente ed autonomo)
- *Numero di dichiarazione*: questo numero permette di ordinare i vari fogli di dichiarazione utilizzati per descrivere un dipendente nell'ambito della stessa dichiarazione.

5 Processo di trattamento degli archivi

Il processo di trattamento degli archivi *M*-770 ha lo scopo di trasformare gli archivi amministrativi acquisiti dalla *Sogei*, come descritti nel paragrafo 3, applicando i filtri preliminari descritti nel paragrafo 2.2, per ottenere archivi strutturati secondo il modello delineato nel paragrafo 4. Per ottenere questo risultato è stata definita una procedura a due passi:

- nel primo vengono verticalizzate tutte i microdati presenti nell'archivio.
- nel secondo viene ricostruita la sequenza delle variabili tenendo conto di eventuali salti, incoerenze e duplicazioni dei microdati memorizzati.

Se consideriamo lo schema relazionale descritto come un linguaggio regolare, gli archivi Sogei possono essere considerati come frasi del linguaggio da riconoscere e convertire. Questo implica che il processo di verticalizzazione assume il ruolo di uno scanner, mentre il processo di ricostruzione della sequenza (orizzontalizzazione) assume il ruolo del compilatore (da lista unidimensionale a matrice bi-dimensionale). Il formato verticalizzato permette, inoltre, di gestire in modo più efficace le variabili concatenate che vengono rappresentati su più campi.

5.1 Processo di verticalizzazione

Gli archivi *M*-770 acquisiti dalla *Sogei*, anche se riportano esplicitamente solo i campi effettivamente valorizzati, utilizzano sempre record di lunghezza fissa (l'effettiva lunghezza dipende dalle specifiche tecniche allegate in ogni singola fornitura annuale) e questo comporta la presenza di lunghe sequenze nulle negli archivi (sequenza nulla $\equiv$ sequenza di caratteri BLANK) dove potenzialmente vi dovrebbero essere microdati[3]. Un altro modo di vedere questo

[3] Il processo di verticalizzazione è significativo solo per i record di tipo G ed H. I record di tipo B sono giá rappresentati a tracciato fisso negli archivi *Sogei*,

fenomeno è quello di considerare questi archivi come delle matrici aventi per righe le singole dichiarazioni individuali e per colonne i dati da inserire nella dichiarazione. Queste matrici mostrano una struttura di *matrice sparsa*, quindi aventi una percentuale poco significativa di elementi non nulli rispetto alle dimensioni della matrice stessa.

Questa osservazione è sufficente per giustificare l'adozione di un formato verticalizzato nei primi passi della procedura di elaborazione degli archivi M-770. Durante il processo di verticalizzazione vengono effettuate alcune operazioni di semplificazione dei formati adottati nel tracciato record degli archivi acquisiti.

In particolare, la struttura informativa degli archivi acquisiti dalla *Sogei* definisce un elevato numero di formati per ogni tipologia di variabili (es. vi sono almeno tre formati differenti per rappresentare le variabili di tipo data) che possono essere ridotti a quattro tipologie di riferimento: stringa di caratteri a lunghezza variabile, numeri interi, numeri reali con virgola mobile e date con formato *YYYY-MM-DD*.

Nella struttura verticalizzata, ogni riga rappresenta una microdato non nullo nell'archivio acquisito M-770. Il tracciato record verticalizzato è costituito da 7 coordinate che rappresentano *"la posizione logica"* all'interno della struttura originaria, piú un campo per rappresentare il microdato.

- *Anno, IdTel, codice fiscale dipendente*: è il nucleo primario delle coordinate, avente lo scopo di separare le singole dichiarazioni relative alle singole persone fisiche.
- *Progressivo* - fornisce un indice all'interno della dichiarazione considerata identificabile tramite (anno, $IdTel$).
- *Tipo* - indica il prospetto e/o quadro che stiamo considerando.
- *Riga, colonna* - indica le coordinate del microdato come facente parte di una struttura matriciale.
- *Valore* - indica il valore vero e proprio della variabile così individuata.

L'ordine in cui vengono inseriti i record nell'archivio verticalizzato deriva direttamente dalla sequenza con cui sono memorizzati i microdati nell'archivio acquisito.

5.2 Processo di ricostruzione della sequenza

Il processo di ricostruzione ha come ingresso l'archivio verticalizzato e come uscita un archivio avente la struttura descritta nel paragrafo 4.1. In questa fase vengono risolti i problemi relativi alla ricostruzione delle sequenze di variabili che definiscono i prospetti relativi al lavoro dipendente ed autonomo. Ogni

---

quindi il processo che prepara i dati del frontespizio filtra i record B e separa i campi presenti mantenendo la sequenza originale. Questo implica che il processo di verticalizzazione e successiva ricostruzione della sequenza non è significativo in questo caso.

record presente nell'archivio verticalizzato rappresenta un valore non nullo nel modello M-770. Per le esigenze del progetto $TMNP$, non tutte le 7, descritte nel paragrafo 5.1, coordinate sono necessarie ovvero giocano lo stesso ruolo:

- *(Anno, IdTel)* individuano la singola dichiarazione nell'anno di dichiarazione considerato.
- *Codice fiscale dipendente* individua la parte della dichiarazione relativa alla singola persona fisica.
- *Progressivo* - ai fini del progetto $TMNP$, non risulta necessario fornire informazioni sull'ordinamento delle persone fisiche all'interno delle singole dichiarazioni, quindi nell'attuale versione degli archivi questa coordinata viene sostanzialmente ignorata.
- *(Tipo, colonna)* definiscono la posizione del valore all'interno del record che si sta ricostruendo, e quindi il nome della variabile.
- *Riga* definisce l'n-esimo record relativo alla singola persona fisica.

La creazione di un tracciato orizzontale avviene attraverso un processo di ricostruzione della sequenza corretta dei valori che compongono ogni record; questa sequenza non sempre è deducibile dalla lettura sequenziale dei valori in quanto vi sono salti, sia di riga sia di colonna, non facilmente interpretabili. La ricostruzione del record finale, relativo ad un singolo lavoratore (dipendente e/o autonomo) viene fatta selezionando le righe dell'archivio verticalizzato che presentano una uniformità di coordinate (in particolare indice di riga e di colonna).

Example 1. sia la sequenza di coordinate (riga,colonna): .... $(1, 22)$ $(1, 28)$ $(2, 28)$ $(2, 29)$ $(2, 30)$ $(2, 31)$ $(2, 32)$ $(1, 53)$ .... $(1, 97)$ $(2, 1)$ $(2, 2)$ .....
Si possono individuare tre cambi di sequenza:

- $(1, 28)$ $(2, 28)$: la stessa variabile (come posizione di colonna, ma non di valore) viene attribuita a due sequenze (righe) diverse.
- $(2, 32)$ $(1, 53)$: da una sequenza si passa ad una precedente ma non si ricomincia dall'inizio.
- $(1, 97)$ $(2, 1)$: finisce una sequenza ed inizia quella successiva.

Solo il terzo caso rappresenta una situazione legale, mentre le prime due pongono problemi di ricostruzione.

Qualunque sia la procedura automatizzata di ricostruzione della sequenza, essa deve sostanzialmente ispirarsi ad uno dei seguenti approcci:

- *Locale*: in ogni passo della procedura esiste al piú una sequenza su cui si lavora. Ogni salto di sequenza (cambio dell'indice di riga) determina un nuovo record. Questa è una soluzione molto semplice ed efficente ma poco raffinata quando si hanno salti di sequenza non ordinati come nell'esempio descritto.

- *Semi locale*: in ogni passo della procedura vi sono un numero limitato di sequenze su cui si lavora contemporaneamente. Ogni sequenza puó essere chiusa, e quindi generare un nuovo record, in modo indipendente dalle altre sequenze in lavorazione. Questo approccio si ispira sempre ad una ricerca locale delle seguenze, ma riesce a recuperare più situazioni dubbie rispetto all'approccio precedente avendo una visione più ampia (un numero maggiore di sequenze lavorate contemporaneamente).
- *Globale*: tutte le sequenze relative ad una dichiarazione vengono considerate aperte. Quando si chiude una dichiarazione tutte le sequenze collegate si chiudono. Questa soluzione è sicuramente la più efficace tra quelle considerate ma risulta essere onerosa quando si considerano dichiarazioni relative ad aziende con molti dipendenti (es. l'Inps è sostituto d'imposta per i suoi pensionati).

Attualmente è stata implementata una soluzione *semi locale* con un buffer di ricostruzione di grandezza 3.

L'ipotesi di fondo che si considera sempre valida è quella che le coordinate sono sempre significative quindi non sono presenti valori negativi ovvero più grandi del massimo consentito nel singolo quadro (ogni coordinata riga, colonna ha un ben preciso intervallo di definizione del tipo $[1, N]$). Quando questa ipotesi viene smentita, il record nell'archivio verticalizzato viene scartato poichè non è stato possibile definire alcuna regola per il recupero delle coordinate non valide; infatti qualunque ipotesi non ha retto alla verifica sperimentale sui dati. Si deve osservare che il recupero di questi casi risulta essere quasi impossibile a questo livello di lavorazione dell'archivio poichè non si possono fare ipotesi sul contenuto della variabile (es. tipo di dato, valore assunto) e quindi assegnarlo alla cella più probabile. Questo tipo di errori tendono a ridursi con il susseguirsi delle forniture (miglioramento delle procedure di inserimento dati) e riguardano un numero molto ridotto di casi se visti in termini percentuali.

Un altro tipo di problema da affrontare durante questa fase è relativo alla presenza di campi nel modello di dichiarazione che possono assumere due valori contemporaneamente (es. durata temporale espressa in numero di anni e mesi). Questi casi vengono rappresentati nell'archivio verticalizzato fornendo coordinte di colonna alfanumeriche (es. A64, B64), ma che devono essere inserite nella stessa cella della variabile 64 del relativo quadro. La soluzione adottata è stata quella di inserire questi valori separandoli con opportuni separatori di sottocampo. Questi separatori sono stato scelti in modo da poter distinguere i due valori (es. (34) [2] indica un periodo di 34 anni e 2 mesi). Questo tipo di campi non sono frequenti ma si è osservato che non sempre i valori inseriti sono significativi e rispettano il vincolo di massimo due valori.

Esiste una situazione di *"collisione delle coordinate"*[4] che rappresenta una situazione legale e quindi si deve gestire correttamente. Negli archivi *Sogei*,

[4] Questa situazione si verifica quando due record consecutivi, nel file verticalizzato, sono identici ad eccezione del campo valore.

relativi all'anno 2004, esistono campi aventi lunghezza superiore a quella massima prevista[5] (16 caratteri). Per poter rappresentare eventuali valori che superano questa soglia sono state definite le "variabili concatenate": la sequenza piú lunga di 16 caratteri viene divisa in sottosequenze di 15 caratteri e ad ognuna delle quali viene aggiunto il carattere *"+"*. Solo la prima sottosequenza non presenta questo carattere aggiuntivo, ed è quindi lunga 16 caratteri, mentre l'ultima sottosequenza è di lunghezza arbitraria (da 1 a 15 caratteri)[6]. Ognuna delle sottosequenze calcolate viene memorizzata nell'archivio *Sogei* con le stessa struttura *(sezione, riga, colonna)*.
Alla fine di questo processo di ricostruzione si ottengono gli archivi sequenziali divisi per tipologia di record (B, G ed H), come descritto nel paragrafo 4.

Tabella 6. Prospetto riassuntivo degli archivi ricostruiti

Anno di dichiarazione	Frontespizio	Lavoro Dipendente	Lavoro Autonomo
2001	1,628,313	45,114,541	0
2002	1,863,384	56,601,669	0
2003	1,914,331	52,361,853	0
2004	3,784,447	53,138,122	12,848,219

Nella tabella 6 vengono indicati le dimensioni degli archivi (espresso come numero di record totali) risultanti dal processo di ricostruzione descritto. Il confronto con la tabella 2 del paragrafo 2 evidenzia che il numero di record iniziali è inferiore a quello degli archivi ricostruiti. Questo è coerente con la funzione di ricostruzione che separa i valori posti su uno stesso record in ingresso, mettendoli su più record in uscita.

6 Raffinamento degli archivi

Gli archivi derivati dal processo di ricostruzione non possono ancora essere considerati pronti per essere resi disponibili a successive elaborazioni statistiche. L'osservazione di alcuni indicatori fornisce una prima chiave di lettura dei possibili raffinamenti da effettuare.
La tabella 7 illustra la situazione dei record distinti presenti nei quadri B, G ed H. I record mancanti, rispetto alla tabella 6, forniscono un indicatore della percentuale di duplicazioni presenti negli archivi Sogei acquisiti. Queste

[5] Attualmente i casi possibili sono relativi a campi alfanumeri, ed in particolare *"località di residenza estera"* ($AUXXX013$), e *"via e numero civico"* ($AUXXX014$).

[6] Es. la sequenza *"01234567890123456789012345678901234 5"*, lunga 36 caratteri, viene divisa in tre sottosequenze: *"0123456789012345"*, *"+678901234567890"* e *"+12345"*

Tabella 7. Prospetto dei record distinti nelle sezioni B, G ed H

Anno di dichiarazione	Frontespizio	Lavoro Dipendente	Lavoro Autonomo
2001	1,628,313	31,439,417	0
2002	1,863,361	43,653,289	0
2003	1,713,204	42,242,188	0
2004	3,784,447	53,113,167	12,816,439

Tabella 8. Percentuali dei record distinti nelle sezioni B, G ed H

Anno di dichiarazione	Frontespizio	Lavoro Dipendente	Lavoro Autonomo
2001	100.0%	69.7%	0.0%
2002	$\sim$ 100.0%	77.1%	0.0%
2003	89.5%	80.7%	0.0%
2004	100.0%	99.9%	99.7%

duplicazioni non possono essere rilevati durante il processo di ricostruzione in quanto è un processo che ha una visione ristretta alla singola dichiarazione (insieme di record legati tramite la variabile $IdTel$); questo significa che le duplicazioni presenti in punti diversi dell'archivio possono essere rilevati solo dopo la ricostruzione di tutti i record, quando si ha una visione completa. I valori percentuali presenti nella tabella 8 sono il risultato del confronto tra la tabella 7 con l'universo dei record B, G ed H espresso in tabella 6. Queste percentuali indicano che nel corso degli anni vi è un progressivo miglioramento del processo di acquisizione degli archivi.

La natura/causa di queste duplicazioni non è sempre ben chiara ma si possono ipotizzare sostanzialmente due eventi:

- Errori di trasmissione dei dati nel passaggio da Sogei ad Istat.
- Dichiarazione inviate più volte dalle aziende per sbaglio.

A questo livello, si definisce duplicazione di record quando si hanno record dell'archivio perfettamente uguali, quindi quando anche la variabile $IdTel$ coincide. Questa situazione è compatibile sicuramente con la prima ipotesi e meno con la seconda poichè due invii generano comunque valori di $IdTel$ diversi. A valle di questo controllo è naturale procedere alla identificazione di situazioni di duplicazione logica e non fisica delle informazioni. Questo concetto di duplicazione può presentarsi in vari modi ma che sono riconducibili tutti alla stessa domanda:

> Quando una stessa azienda (identificata da codice fiscale e/o partita iva) invia più dichiarazioni (più record presenti nella tabella frontespizio), quali dichiarazioni dobbiamo considerare valide e quali trascurare?

Per rispondere a questa domanda è necessario illustrare alcune possibili cause che spiegano la presenza di dichiarazioni multiple:

- *Errori*: una azienda invia la stessa dichiarazione in due momenti diversi pur non essendoci alcun motivo. In questo caso le due dichiarazioni differiscono solo per la variabile $IdTel$.
- *Split dimensionale*: le aziende molto grandi non riescono a inserire tutte le dichiarazioni relative ai propri dipendenti, e quindi decide di frazionare l'unica dichiarazione in vari pezzi caratterizzati da valori di $IdTel$ diversi. I diversi invii differiscono tra loro non solo per la variabile $IdTel$, ma anche per il dettaglio dipendenti riportato.
- *Dichiarazioni integrative*: è possibile inviare delle integrazioni alla dichiarazione già fatta qualora sia necessario inserire informazioni non fornite in precedenza ed i termini di presentazione sono scaduti.
- *Dichiarazioni correttive*: qualora ci si accorge di aver fornito dati non veri, è possibile inviare nuovamente la dichiarazione con le modifiche qualora non siano scaduti i termini della presentazione.
- *Eventi eccezionali*: in presenza di eventi eccezionali è possibile avere delle proroghe/deroghe nella compilazione del modello.

La prima condizione può essere trattata in modo abbastanza semplice qualora le dichiarazioni sono identiche a meno della variabile $IdTel$. Nel secondo caso possono essere considerate dichiarazioni distinte nella tabella G. Gli ultimi tre casi possono essere trattati in linea di principio basandosi su alcune variabili che indicano il tipo di dichiarazione (correttiva/integrativa/evento eccezionale) considerata.

Tabella 9. Prospetto dei record multipli nella sezione B

Anno di dichiarazione	Integrative	Correttive	Eventi Eccezionali	I+C+E
2001	16,280	21,267	1,497	38869
2002	19,729	26,671	715	47083
2003	20,172	22,408	920	43461
2004	32,595	15,751	2,696	51,002

Nella tabella 9 si ha un termine di paragone del fenomeno delle dichiarazioni multiple rispetto all'universo descritto nella colonna Frontespizio della tabella 6. La colonna (I+C+E) indica il numero di dichiarazioni in cui sia presente almeno uno dei tre tipi di evento, quindi misura il grado di correlazione degli eventi. Attualmente non è stata implementata alcun procedura automatica per il trattamento delle dichiarazioni multiple utilizzando queste variabili. Inoltre non si hanno studi sulla reale affidabilità di queste variabili, per poterle utilizzare in un processo di controllo e correzione.

Una delle prime considerazioni da fare quando si entra nel merito dei valori espressi nella dichiarazione, riguarda la determinazione dell'unità di misura degli importi.

Tabella 10. Prospetto delle divise/valuta nella sezione B

Anno di dichiarazione	Dichiarazioni in Lire	Dichiarazioni in Euro
2001	1,625,823	2,490
2002	1,669,783	193,578
2003	0	1,713,204
2004	0	3,784,447

Dalla tabella 10 si ricava che fino al 2002 è stato possibile scegliere quale valuta adottare nella dichiarazione; questo significa che prima di un trattamento integrato delle dichiarazioni relative a questo periodo, è necessario ricondurre tutti gli importi ad una stessa unità di misura. Anche in questo caso è opportuno fare valutazioni sulla reale corrispondenza tra la valuta indicata nel frontespizio e gli importi reali presenti negli altri quadri; qualsiasi processo di valutazione di importi anomali risente pesantemente di unità di misura sbagliate in quanto i valori espressi in Lire e quelli espressi in Euro differiscono di 3 ordini di grandezza.

7 Conclusione

I processi descritti rappresentano una prima fase di lavorazione degli archivi *M*-770, fase in cui si pone l'accento sulla ricostruzione dei legami tra variabili di uno stesso quadro e sulla separazione dei record in tabelle aventi tracciati omogenei. La sequenza dei paragrafi ha lo scopo di evidenziare come qualsiasi ipotesi sulle dinamiche esistenti negli archivi devono essere verificate anno per anno in quanto non sono costanti ovvero sempre predicibili. Questo implica che difficilmente uno stesso processo viene eseguito inalterato per due anni consecutivi.

Se lo scenario è questo allora la progettazione degli archivi finali deve essere fatta tenendo conto di questa variabilità. Gli archivi finali possono essere considerati definitivi per quanto riguarda il processo di acquisizione e trattamento informatico, ma sono suscettibili di ulteriori trattamenti al fine della produzione statistica a regime. Questo implica che per valorizzare statisticamente le basi amministrative, trattate informaticamente nel modo descritto, altri passi di elaborazione metodologico-statistica devono essere fatti. Ad esempio, nei vari anni disponibili, le unità di misura ed i valori fuori dominio devono essere trattati in modo da uniformarli tra loro.

Tra questi passi quello più importante è quello che consente di distinguere la dichiarazione principale da quelle secondarie.

Contributi ISTAT(*)

1/2002 - Francesca Biancani, Andrea Carone, Rita Pistacchio e Giuseppina Ruocco - *Analisi delle imprese individuali*

2/2002 - Massimiliano Borgese .-.*Proposte metodologiche per un progetto d'indagine sul trasporto aereo alla luce della recente normativa comunitaria sul settore*

3/2002 - Nadia Di Veroli e Roberta Rizzi - *Proposta di classificazione dei rapporti di lavoro subordinato e delle attività di lavoro autonomo: analisi del quadro normativo*

4/2002 – Roberto Gismondi – *Uno stimatore ottimale in presenza di non risposte*

5/2002 – Maria Anna Pennucci – *Le strategie europee per l'occupazione dal Libro bianco di Delors al Consiglio Europeo di Cardiff*

1/2003 – Giovanni Maria Merola – *Safety Rules in Statistical Disclosure Control for Tabular Data*

2/2003 – Fabio Bacchini, Pietro Gennari e Roberto Iannaccone – *A new index of production for the construction sector based on input data*

3/2003 – Fulvia Cerroni e Enrica Morganti – *La metodologia e il potenziale informativo dell'archivio sui gruppi di impresa: primi risultati*

4/2003 – Sara Mastrovita e Isabella Siciliani – *Effetti dei trasferimenti sociali sulla distribuzione del reddito nei Paesi dell'Unione europea: un'analisi dal Panel europeo sule famiglie*

5/2003 – Patrizia Cella, Giuseppe Garofalo, Adriano Paggiaro, Nicola Torelli e Caterina Viviano – *Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità*

6/2003 – Enrico Grande e Orietta Luzi – *Metodologie per l'imputazione delle mancate risposte parziali: analisi critica e soluzioni disponibili in Istat*

7/2003 – Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Indagine sperimentale sui posti di lavoro vacanti*

8/2003 – Mario Adua – *L'agricoltura di montagna: le aziende delle donne, caratteristiche agricole e socio-rurali*

9/2003 – Franco Mostacci e Roberto Sabbatini – *L'euro ha creato inflazione? Changeover e arrotondamenti dei prezzi al consumo in Italia nel 2002*

10/2003 – Leonello Tronti – *Problemi e prospettive di riforma del sistema pensionistico*

11/2003 – Roberto Gismondi – *Tecniche di stima e condizioni di coerenza per indagini infraannuali ripetute nel tempo*

12/2003 – Antonio Frenda – *Analisi delle legislazioni e delle prassi contabili relative ai gruppi di imprese nei paesi dell'Unione Europea*

1/2004 – Marcello D'Orazio, Marco Di Zio e Mauro Scanu – *Statistical Matching and the Likelihood Principle: Uncertainty and Logical Constraints*

2/2004 – Giovanna Brancato – *Metodologie e stime dell'errore di risposta. Una sperimentazione di reintervista telefonica*

3/2004 – Franco Mostacci, Giuseppina Natale e Elisabetta Pugliese – *Gli indici dei prezzi al consumo per sub popolazioni*

4/2004 – Leonello Tronti – *Una proposta di metodo: osservazioni e raccomandazioni sulla definizione e la classificazione di alcune variabili attinenti al mercato del lavoro*

5/2004 – Ugo Guarnera – *Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi: il softaware* Quis

6/2004 – Patrizia Giaquinto, Marco Landriscina e Daniela Pagliuca – *La nuova funzione di analisi dei modelli implementata in Genesees v. 3.0*

7/2004 – Roberto Di Giuseppe, Patrizia Giaquinto e Daniela Pagliuca – *MAUSS (Multivariate Allocation of Units in Sampling Surveys):*

un software generalizzato per risolvere il problema dell' allocazione campionaria nelle indagini Istat

8/2004 – Ennio Fortunato e Liana Verzicco – *Problemi di rilevazione e integrazione della condizione professionale nelle indagini sociali dell'Istat*

9/2004 – Claudio Pauselli e Claudia Rinaldelli – *La valutazione dell'errore di campionamento delle stime di povertà relativa secondo la tecnica Replicazioni Bilanciate Ripetute*

10/2004 – Eugenio Arcidiacono, Marina Briolini, Paolo Giuberti, Marco Ricci, Giovanni Sacchini e Giorgia Telloli – *Procedimenti giudiziari, reati, indagati e vittime in Emilia-Romagna nel 2002: un'analisi territoriale sulla base dei procedimenti iscritti nel sistema informativo Re.Ge.*

11/2004 – Enrico Grande e Orietta Luzi **–** *Regression trees in the context of imputation of item non-response: an experimental application on business data*

12/2004 – Luisa Frova e Marilena Pappagallo **–** *Procedura di now-cast dei dati di mortalità per causa*

13/2004 – Giorgio DellaRocca, Marco Di Zio, Orietta Luzi, Emanuela Scavalli e Giorgia Simeoni **–** *IDEA (Indices for Data Editing Assessment): sistema per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*

14/2004 **–** Monica Pace, Silvia Bruzzone, Luisa Frova e Marilena Pappagallo **–** *Review of the existing information about death certification practices, certificate structures and training tools for certification of causes of death in Europe*

15/2004 – Elisa Berntsen – *Modello Unico di Dichiarazione ambientale: una fonte amministrativa per l'Archivio delle Unità Locali di Asia*

16/2004 – Salvatore F. Allegra e Alessandro La Rocca – *Sintetizzare misure elementari: una sperimentazione di alcuni criteri per la definizione di un indice composto*

17/2004 – Francesca R. Pogelli – *Un'applicazione del modello "Country Product Dummy" per un'analisi territoriale dei prezzi*

18/2004 – Antonia Manzari – *Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi*

19/2004 – Claudio Pauselli – *Intensità di povertà relativa: stima dell'errore di campionamento e sua valutazione temporale*

20/2004 – Maria Dimitri, Ersilia Di Pietro, Alessandra Nuccitelli e Evelina Paluzzi – *Sperimentazione di una metodologia per il controllo della qualità di dati anagrafici*

21/2004 – Tiziana Pichiorri, Anna M. Sgamba e Valerio Papale – *Un modello di ottimizzazione per l'imputazione delle mancate risposte statistiche nell'indagine sui trasporti marittimi dell'Istat*

(*) ultimi cinque anni

22/2004 – Diego Bellisai, Piero D. Falorsi, Annalisa Lucarelli, Maria A. Pennucci e Leonello G. Tronti – *Indagine pilota sulle retribuzioni di fatto nel pubblico impiego*

23/2004 – Lidia Brondi – *La riorganizzazione del sistema idrico: quadro normativo, delimitazione degli ambiti territoriali ottimali e analisi statistica delle loro caratteristiche strutturali*

24/2004 – Roberto Gismondi e Laura De Sandro – *Provisional Estimation of the Italian Monthly Retail Trade Index*

25/2004 – Annamaria Urbano, Claudia Brunini e Alessandra Chessa – *I minori in stato di abbandono: analisi del fenomeno e studio di una nuova prospettiva d'indagine*

26/2004 – Paola Anzini e Anna Ciammola – *La destagionalizzazione degli indici della produzione industriale: un confronto tra approccio diretto e indiretto*

27/2004 – Alessandro La Rocca – *Analisi della struttura settoriale dell'occupazione regionale: 8° Censimento dell'industria e dei servizi 2001 7° Censimento dell'industria e dei servizi 1991*

28/2004 – Vincenzo Spinelli e Massimiliano Tancioni – *I Trattamenti Monetari non Pensionistici: approccio computazionale e risultati della sperimentazione sugli archivi INPS-DM10*

29/2004 – Paolo Consolini – *L'indagine sperimentale sull'archivio fiscale modd.770 anno 1999: analisi della qualità del dato e stime campionarie*

1/2005 – Fabrizio M. Arosio – *La stampa periodica e l'informazione on-line: risultati dell'indagine pilota sui quotidiani on-line*

2/2005 – Marco Di Zio, Ugo Guarnera e Orietta Luzi – *Improving the effectiveness of a probabilistic editing strategy for business data*

3/2005 – Diego Moretti e Claudia Rinaldelli – *EU-SILC complex indicators: the implementation of variance estimation*

4/2005 – Fabio Bacchini, Roberto Iannaccone e Edoardo Otranto – *L'imputazione delle mancate risposte in presenza di dati longitudinali: un'applicazione ai permessi di costruzione*

5/2005 – Marco Broccoli – *Analisi della criminalità a livello comunale: metodologie innovative*

6/2005 – Claudia De Vitiis, Loredana Di Consiglio e Stefano Falorsi – *Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro*

7/2005 – Edoardo Otranto e Roberto Iannaccone – *Continuous Time Models to Extract a Signal in Presence of Irregular Surveys*

8/2005 – Cosima Mero e Adriano Pareto – *Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie*

9/2005 – Filippo Oropallo – *Enterprise microsimulation models and data challenges*

10/2005 – Marcello D' Orazio, Marco Di Zio e Mauro Scanu – *A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study*

11/2005 – Stefania Macchia, Manuela Murgia, Loredana Mazza, Giorgia Simeoni, Francesca Di Patrizio, Valentino Parisi, Roberto Petrillo e Paola Ungaro – *Una soluzione per la rilevazione e codifica della Professione nelle indagini CATI*

12/2005 – Piero D. Falorsi, Monica Scannapieco, Antonia Boggia e Antonio Pavone – *Principi Guida per il Miglioramento della Qualità dei Dati Toponomastici nella Pubblica Amministrazione*

13/2005 – Ciro Baldi, Francesca Ceccato, Silvia Pacini e Donatella Tuzi – *La stima anticipata OROS sull'occupazione. Errori, problemi della metodologia attuale e proposte di miglioramento*

14/2005 – Stefano De Francisci, Giuseppe Sindoni e Leonardo Tininini – *Da Winci/MD: un sistema per data warehouse statistici sul Web*

15/2005 – Gerardo Gallo e Evelina Palazzi – *I cittadini italiani naturalizzati: l'analisi dei dati censuari del 2001, con un confronto tra immigrati di prima e seconda generazione*

16/2005 – Saverio Gazzelloni, Mario Albisinni, Lorenzo Bagatta, Claudio Ceccarelli, Luciana Quattrociocchi, Rita Ranaldi e Antonio Toma – *La nuova rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*

17/2005 – Maria Carla Congia – *Il lavoro degli extracomunitari nelle imprese italiane e la regolarizzazione del 2002. Prime evidenze empiriche dai dati INPS*

18/2005 – Giovanni Bottazzi, Patrizia Cella, Giuseppe Garofalo, Paolo Misso, Mariano Porcu e Marianna Tosi – *Indagine pilota sulla nuova imprenditorialità nella Regione Sardegna. Relazione Conclusiva*

19/2005 – Fabrizio Martire e Donatella Zindato – *Le famiglie straniere: analisi dei dati censuari del 2001 sui cittadini stranieri residenti*

20/2005 – Ennio Fortunato – *Il Sistema di Indicatori Territoriali: percorso di progetto, prospettive di sviluppo e integrazione con i processi di produzione statistica*

21/2005 – Antonella Baldassarini e Danilo Birardi – *I conti economici trimestrali: un approccio alla stima dell'input di lavoro*

22/2005 – Francesco Rizzo, Dario Camol e Laura Vignola – *Uso di XML e WEB Services per l'integrazione di sistemi informativi statistici attraverso lo standard SDMX*

1/2006 – Ennio Fortunato – *L'analisi integrata delle esigenze informative dell'utenza Istat: Il contributo del Sistema di Indicatori Territoriali*

2/2006 – Francesco Altarocca – *I design pattern nella progettazione di software per il supporto alla statistica ufficiale*

3/2006 – Roberta Palmieri – *Le migranti straniere: una lettura di genere dei dati dell'osservatorio interistituzionale sull'immigrazione in provincia di Macerata*

4/2006 – Raffaella Amato, Silvia Bruzzone, Valentina Delmonte e Lidia Fagiolo – *Le statistiche sociali dell'ISTAT e il fenomeno degli incidenti stradali: un'esperienza di record linkage*

5/2006 – Alessandro La Rocca – *Fuzzy clustering: la logica, i metodi*

6/2006 – Raffaella Cascioli – *Integrazione dei dati micro dalla Rilevazione delle Forze di Lavoro e dagli archivi amministrativi INPS: risultati di una sperimentazione sui dati campione di 4 province*

7/2006 – Gianluca Brogi, Salvatore Cusimano, Giuseppina del Vicario, Giuseppe Garofalo e Orietta Patacchia – *La realizzazione di Asia Agricoltura tramite l'utilizzo di dati amministrativi: il contenuto delle fonti e i risultati del processo di integrazione*

8/2006 – Simonetta Cozzi – *La distribuzione commerciale in Italia: caratteristiche strutturali e tendenze evolutive*

9/2006 – Giovanni Seri – *A graphical framework to evaluate risk assessment and information loss at individual level*

10/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Feasibility studies for the coverage of public institutions in sections N and O*

11/2006 – Diego Bellisai, Annalisa Lucarelli, Maria Anna Pennucci e Fabio Rapiti – *Quarterly labour cost index in public education*

12/2006 – Silvia Montagna, Patrizia Collesi, Florinda Damiani, Danila Fulgenzio, Maria Francesca Loporcaro e Giorgia Simeoni – *Nuove esperienze di rilevazione della Customer Satisfaction*

13/2006 – Lucia Coppola e Giovanni Seri – *Confidentiality aspects of household panel surveys: the case study of Italian sample from EU-SILC*

14/2006 – Lidia Brondi – *L'utilizzazione delle surveys per la stima del valore monetario del danno ambientale: il metodo della valutazione contingente*

15/2006 – Carlo Boselli – *Le piccole imprese leggere esportatrici e non esportatrici: differenze di struttura e di comportamento*

16/2006 – Carlo De Greogorio – *Il nuovo impianto della rilevazione centralizzata del prezzo dei medicinali con obbligo di prescrizione*

1/2007 – Paolo Roberti, Maria Grazia Calza, Filippo Oropallo e Stefania Rossetti – *Knowledge Databases to Support Policy Impact Analysis: the EuroKy-PIA Project*

2/2007 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, e Marina Sorrentino – *Production of job vacancy statistics: coverage*

3/2007 – Carlo Lucarelli e Giampiero Ricci – *Working times and working schedules: the framework emerging from the new Italian lfs in a gender perspective*

4/2007 – Monica Scannapieco, Diego Zardetto e Giulio Barcaroli – *La Calibrazione dei Dati con R: una Sperimentazione sull'Indagine Forze di Lavoro ed un Confronto con GENESEES/SAS*

5/2007 – Giulio Barcaroli e Tiziana Pellicciotti – *Strumenti per la documentazione e diffusione dei microdati d'indagine: il Microdata Management Toolkit*

6/2007 – AA.VV. – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 1ª giornata*

7/2007 – Raffaella Cianchetta, Carlo De Gregorio, Giovanni Seri e Giulio Barcaroli – *Rilevazione sulle Pubblicazioni Scientifiche Istat*

8/2007 – Emilia Arcaleni, e Barbara Baldazzi – *Vivere non insieme: approcci conoscitivi al Living Apart Together*

9/2007 – Corrado Peperoni e Francesca Tuzi – *Trattamenti monetari non pensionistici metodologia sperimentale per la stima degli assegni al nucleo familiare*

10/2007 – AA.VV – *Seminario sulla qualità: l'esperienza dei referenti del sistema informativo SIDI - 2ª giornata*

11/2007 – Leonello Tronti – *Il prototipo (numero 0) dell'Annuario di statistiche del Mercato del Lavoro (AML)*

12/2007 – Daniele Frongia, Raffaello Martinelli, Fernanda Panizon, Bruno Querini e Andrea Stanco – *Il nuovo Sistema informatico Altri Servizi. Progetto di reingegnerizzazione dei processi produttivi delle indagini trimestrali di fatturato degli altri servizi*

13/2007 – Carlo De Gregorio, Stefania Fatello, Rosanna Lo Conte, Stefano Mosca, Francesca Rossetti – *Sampling design and treatment of products in Istat centralised CPI surveys*

Documenti ISTAT(*)

1/2002 – Paolo Consolini e Rita De Carli - *Le prestazioni sociali monetarie non pensionistiche: unità di analisi, fonti e rappresentazione statistica dei dati*

2/2002 – Stefania Macchia - *Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività economiche*

3/2002 – Maria De Lucia - *Applicabilità della disciplina in materia di festività nel pubblico impiego*

4/2002 – Roberto Gismondi, Massimo Marciani e Mauro Giorgetti - *The italian contribution towards the implementantion of an european transport information system: main results of the MESUDEMO project*

5/2002 – Olimpio Cianfarani e Sauro Angeletti - *Misure di risultato e indicatori di processo: l'esperienza progettuale dell'Istat*

6/2002 – Riccardo Carbini e Valerio De Santis – *Programma statistico nazionale: specifiche e note metodologiche per la compilazione delle schede identificative dei progetti*

7/2002 – Maria De Lucia – *Il CCNL del personale dirigente dell'area 1 e la valutazione delle prestazioni dei dirigenti*

8/2002 – Giuseppe Garofalo e Enrica Morganti – *Gruppo di lavoro per la progettazione di un archivio statistico sui gruppi d'impresa*

1/2003 – Francesca Ceccato, Massimiliano Tancioni e Donatella Tuzi – *MODSIM-P:Il nuovo modello dinamico di previsione della spesa pensionistica*

2/2003 – Anna Pia Mirto – *Definizioni e classificazioni delle strutture ricettive nelle rilevazioni statistiche ufficiali sull'offerta turistica*

3/2003 – Simona Spirito – *Le prestazioni assistenziali monetarie non pensionistiche*

4/2003 – Maria De Lucia – *Approfondimenti di alcune tematiche inerenti la gestione del personale*

5/2003 – Rosalia Coniglio, Marialuisa Cugno, Maria Filmeno e Alberto Vitalini – *Mappatura della criminalità nel distretto di Milano*

6/2003 – Maria Letizia D'Autilia – *I provvedimenti di riforma della pubblica amministrazione per l'identificazione delle "Amministrazioni pubbliche" secondo il Sec95: analisi istituzionale e organizzativa per l'anno 2000*

7/2003 – Francesca Gallo, Pierpaolo Massoli, Sara Mastrovita, Roberto Merluzzi, Claudio Pauselli, Isabella Siciliani e Alessandra Sorrentino – *La procedura di controllo e correzione dei dati Panel Europeo sulle famiglie*

8/2003 – Cinzia Castagnaro, Martina Lo Conte, Stefania Macchia e Manuela Murgia – *Una soluzione in-house per le indagini CATI: il caso della Indagine Campionaria sulle Nascite*

9/2003 – Anna Pia Maria Mirto e Norina Salamone – *La classificazione delle strutture ricettive turistiche nella normativa delle regioni italiane*

10/2003 – Roberto Gismondi e Anna Pia Maria Mirto – *Le fonti statistiche per l'analisi della congiuntura turistica: il mosaico italiano*

11/2003 – Loredana Di Consiglio e Stefano Falorsi – *Alcuni aspetti metodologici relativi al disegno dell'indagine di copertura del Censimento Generale della Popolazione 2001*

12/2003 – Roberto Gismondi e Anna Rita Giorgi – *Struttura e dinamica evolutiva del comparto commerciale al dettaglio: le tendenze recenti e gli effetti della riforma "Bersani"*

13/2003 – Donatella Cangialosi e Rosario Milazzo – *Fabbisogni formativi degli Uffici comunali di statistica: indagine rapida in Sicilia*

14/2003 – Agostino Buratti e Giovanni Salzano – *Il sistema automatizzato integrato per la gestione delle rilevazioni dei documenti di bilancio degli enti locali*

1/2004 – Giovanna Brancato e Giorgia Simeoni – *Tesauri del Sistema Informativo di Documentazione delle Indagini (SIDI)*

2/2004 – Corrado Peperoni – *Indagine sui bilanci consuntivi degli Enti previdenziali: rilevazione, gestione e procedure di controllo dei dati*

3/2004 – Marzia Angelucci, Giovanna Brancato, Dario Camol, Alessio Cardacino, Sandra Maresca e Concetta Pellegrini – *Il sistema ASIMET per la gestione delle Note Metodologiche dell'Annuario Statistico Italiano*

4/2004 – Francesca Gallo, Sara Mastrovita, Isabella Siciliani e Giovanni Battista Arcieri – *Il processo di produzione dell'Indagine ECHP*

5/2004 – Natale Renato Fazio e Carmela Pascucci – *Gli operatori non identificati nelle statistiche del commercio con l'estero: metodologia di identificazione nelle spedizioni "groupage" e miglioramento nella qualità dei dati*

6/2004 – Diego Moretti e Claudia Rinaldelli – *Una valutazione dettagliata dell'errore campionario della spesa media mensile familiare*

7/2004 – Franco Mostacci – *Aspetti Teorico-pratici per la Costruzione di Indici dei Prezzi al Consumo*

8/2004 – Maria Frustaci – *Glossario economico-statistico multilingua*

9/2004 – Giovanni Seri e Maurizio Lucarelli – *"Il Laboratorio per l'analisi dei dati elementari (ADELE): monitoraggio dell'attività dal 1999 al 2004"*

10/2004 – Alessandra Nuccitelli, Francesco Bosio e Luciano Fioriti – *L'applicazione RECLINK per il record linkage: metodologia implementata e linee guida per la sua utilizzazione*

1/2005 – Francesco Cuccia, Simone De Angelis, Antonio Laureti Palma, Stefania Macchia, Simona Mastroluca e Domenico Perrone – *La codifica delle variabili testuali nel 14° Censimento Generale della Popolazione*

2/2005 – Marina Peci – *La statistica per i Comuni: sviluppo e prospettive del progetto Sisco.T (Servizio Informativo Statistico Comunale. Tavole)*

3/2005 – Massimiliano Renzetti e Annamaria Urbano – *Sistema Informativo sulla Giustizia: strumenti di gestione e manutenzione*

4/2005 – Marco Broccoli, Roberto Di Giuseppe e Daniela Pagliuca – *Progettazione di una procedura informatica generalizzata per la sperimentazione del metodo Microstrat di coordinamento della selezione delle imprese soggette a rilevazioni nella realtà Istat*

5/2005 – Mauro Albani e Francesca Pagliara – *La ristrutturazione della rilevazione Istat sulla criminalità minorile*

6/2005 – Francesco Altarocca e Gaetano Sberno – *Progettazione e sviluppo di un "Catalogo dei File Grezzi con meta-dati di base" (CFG) in tecnologia Web*

* ultimi cinque anni

7/2005 – Salvatore F. Allegra e Barbara Baldazzi – *Data editing and quality of daily diaries in the Italian Time Use Survey*

8/2005 – Alessandra Capobianchi – *Alcune esperienze in ambito internazionale per l'accesso ai dati elementari*

9/2005 – Francesco Rizzo, Laura Vignola, Dario Camol e Mauro Bianchi – *Il progetto "banca dati della diffusione congiunturale"*

10/2005 – Ennio Fortunato e Nadia Mignolli – *I sistemi informativi Istat per la diffusione via web*

11/2005 – Ennio Fortunato e Nadia Mignolli – *Sistemi di indicatori per l'attività di governo: l'offerta informativa dell'Istat*

12/2005 – Carlo De Gregorio e Stefania Fatello – *L'indice dei prezzi al consumo dei testi scolastici nel 2004*

13/2005 – Francesco Rizzo e Laura Vignola – *RSS: uno standard per diffondere informazioni*

14/2005 – Ciro Baldi, Diego Bellisai, Stefania Fivizzani, Annalisa Lucarelli e Marina Sorrentino – *Launching and implementing the job vacancy statistics*

15/2005 – Stefano De Francisci, Massimiliano Renzetti, Giuseppe Sindoni e Leonardo Tininini – *La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT*

16/2005 – Ennio Fortunato e Nadia Mignolli – *Verso il Sistema di Indicatori Territoriali: rilevazione e analisi della produzione Istat*

17/2005 – Raffaella Cianchetta e Daniela Pagliuca – *Soluzioni Open Source per il software generalizzato in Istat: il caso di PHPSurveyor*

18/2005 – Gianluca Giuliani e Barbara Boschetto – *Gli indicatori di qualità dell'Indagine continua sulle Forze di Lavoro dell'Istat*

19/2005 – Rossana Balestrino, Franco Garritano, Carlo Cipriano e Luciano Fanfoni – *Metodi e aspetti tecnologici di raccolta dei dati sulle imprese*

1/2006 – Roberta Roncati – *www.istat.it (versione 3.0) Il nuovo piano di navigazione*

2/2006 – Maura Seri e Annamaria Urbano – *Sistema Informativo Territoriale sulla Giustizia: la sezione sui confronti internazionali*

3/2006 – Giovanna Brancato, Riccardo Carbini e Concetta Pellegrini – *SIQual: il sistema informativo sulla qualità per gli utenti esterni*

4/2006 – Concetta Pellegrini – *Soluzioni tecnologiche a supporto dello sviluppo di sistemi informativi sulla qualità: l'esperienza SIDI*

5/2006 – Maurizio Lucarelli – *Una valutazione critica dei modelli di accesso remoto nella comunicazione di informazione statistica*

6/2006 – Natale Renato Fazio – *La ricostruzione storica delle statistiche del commercio con l'estero per gli anni 1970-1990*

7/2006 – Emilia D'Acunto – *L'evoluzione delle statistiche ufficiali sugli indici dei prezzi al consumo*

8/2006 – Ugo Guarnera, Orietta Luzi e Stefano Salvi – *Indagine struttura e produzioni delle aziende agricole: la nuova procedura di controllo e correzione automatica per le variabili su superfici aziendali e consistenza degli allevamenti*

9/2006 – Maurizio Lucarelli – *La regionalizzazione del Laboratorio ADELE: un'ipotesi di sistema distribuito per l'accesso ai dati elementari*

10/2006 – Alessandra Bugio, Claudia De Vitiis, Stefano Falorsi, Lidia Gargiulo, Emilio Gianicolo e Alessandro Pallara – *La stima di indicatori per domini sub-regionali con i dati dell'indagine: condizioni di salute e ricorso ai servizi sanitari*

11/2006 – Sonia Vittozzi, Paola Giacchè, Achille Zuchegna, Piero Crivelli, Patrizia Collesi, Valerio Tiberi, Alexia Sasso, Maurizio Bonsignori, Giuseppe Stassi e Giovanni A. Barbieri – *Progetto di articolazione della produzione editoriale in collane e settori*

12/2006 – Alessandra Coli, Francesca Tartamella, Giuseppe Sacco, Ivan Faiella, Marcello D'Orazio, Marco Di Zio, Mauro Scanu, Isabella Siciliani, Sara Colombini e Alessandra Masi – *La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane*

13/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Intrastat*

14/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: rilevazione Extrastat*

15/2006 – Ersilia Di Pietro – *Le statistiche del commercio estero dell'Istat: comparazione tra rilevazione Intrastat ed Extrastat*

16/2006 – Fabio M. Rapiti – *Short term statistics quality Reporting: the LCI National Quality Report 2004*

17/2006 – Giampiero Siesto, Franco Branchi, Cristina Casciano, Tiziana Di Francescantonio, Piero Demetrio Falorsi, Salvatore Filiberti, Gianfranco Marsigliesi, Umberto Sansone, Ennio Santi, Roberto Sanzo e Alessandro Zeli – *Valutazione delle possibilità di uso di dati fiscali a supporto della rilevazione PMI*

18/2006 – Mauro Albani – *La nuova procedura per il trattamento dei dati dell'indagine Istat sulla criminalità*

19/2006 – Alessandra Capobianchi – *Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa*

20/2006 – Francesco Altarocca – *Gli strumenti informatici nella raccolta dei dati di indagini statistiche: il caso della Rilevazione sperimentale delle tecnologie informatiche e della comunicazione nelle Pubbliche Amministrazioni locali*

1/2007 – Giuseppe Stassi – *La politica editoriale dell'Istat nel periodo 1996-2004: collane, settori, modalità di diffusione*

2/2007 – Daniela Ichim – *Microdata anonymisation of the Community Innovation Survey data: a density based clustering approach for risk assessment*

3/2007 – Ugo Guarnera, Orietta Luzi e Irene Tommasi – *La nuova procedura di controllo e correzione degli errori e delle mancate risposte parziali nell'indagine sui Risultati Economici delle Aziende Agricole (REA)*

4/2007 – Vincenzo Spinelli – *Processo di Acquisizione e Trattamento Informatico degli Archivi relativi al Modello di Dichiarazione 770*