Istat
Istituto Nazionale di Statistica

Indagine sulla struttura delle retribuzioni 2002 (SES3)

Anno 2002

Descrizione del file*

Dicembre 2008

INDICE

1. Introduzione 3
2. Descrizione delle variabili 3
3. Metodologia statistica per la tutela della riservatezza 6
4. Analisi del contenuto informativo 8
5. Riferimenti bibliografici 12
6. Contatti 13

* Traduzione dall'inglese di Francesca Scambia

1. Introduzione

Nella raccolta dei dati l'Istat deve garantire che l'informazione statistica non sia rilasciata quando informazioni riservate relative ad un singolo rispondente possano essere dedotte in modo troppo accurato. E' stata messa a punto una metodologia statistica per la riduzione del rischio di violazione della riservatezza dei rispondenti. Sono state individuate le unità a rischio di re-identificazione. Sono state quindi modificate unicamente le variabili identificative o riservate. Solo i valori delle unità a rischio di re-identificazione sono state perturbate.

L'utente del file di microdati per la ricerca deve rispettare le condizioni stabilite nel contratto di rilascio.

Gli effetti della metodologia per la limitazione del rischio nei dati statistici saranno presentati in questo documento al fine di favorirne l'uso corretto. Nel paragrafo 2 si forniscono informazioni utili circa lo status di ciascuna variabile nel file di microdati per scopi scientifici. Gli effetti principali della metodologia per la limitazione del rischio nei dati statistici sono illustrate nel paragrafo 3. Il paragrafo 4 evidenzia alcuni vincoli generati dalla metodologia di riduzione del rischio.

2. Descrizione delle variabili

Questo paragrafo riassume i cambiamenti delle variabili relative al file che può essere analizzato nei Safe Centre dell'Istat o di Eurostat. Tali cambiamenti sono necessari al fine di proteggere la riservatezza dei rispondenti.

Le etichette e i nomi delle variabili seguono il Regolamento della Commissione 1916/2000. Le stesse etichette sono utilizzate anche nel documento MFR_SES_NotaMetodologica_2002.pdf.
Il file di microdati può essere fornito in formato fisso (ogni variabile è posizionata in corrispondenza di una colonna ed ha lunghezza fissa) oppure in formato tab-delimited.
La prima riga del file di microdati in formato tab-delimited contiene i nomi delle variabili.
Il separatore decimale è ".".
La prima riga del file di microdati in formato tab-delimited contiene i nomi delle variabili.

Tavola 1 – Lo status[1] di ciascuna variabile nel file di microdati per la ricerca scientifica della SES 2002 italiana

Variabile	Nome della variabile	Tipo	Status
A.1.1	Collocazione geografica	Obbligatorio	Invariato
A.1.2	Dimensione dell'impresa	Obbligatorio	Variato
A.1.3	Principale attività economica	Obbligatorio	Variato
A.1.4	Forma di controllo economico e finanziario	Obbligatorio	Invariato
A.1.5	Esistenza di contratti collettivi di lavoro	Obbligatorio	Invariato
A.1.6	Numero totale di dipendenti nell'unità locale	Facoltativo	Rimosso
A.1.7	Mercato principale per i prodotti dell'impresa	Facoltativo	Invariato
A.1.8	Dimensione del gruppo di imprese	Facoltativo	Invariato
A.1.9	Paese di residenza dell'ente di controllo del gruppo di imprese	Facoltativo	Invariato
A.4.1	Coefficiente di riporto all'universo dell'impresa	Obbligatorio	Rimosso
B.2.1	Sesso	Obbligatorio	Invariato
B.2.2	Età del dipendente	Obbligatorio	Variato
B.2.3	Professione	Obbligatorio	Invariato
B.2.4	Mansioni direttive o di supervisione	Facoltativo	Invariato
B.2.5	Più elevato livello di istruzione e formazione completa raggiunto	Obbligatorio	Invariato
B.2.6	Anzianità di servizio nell'impresa	Obbligatorio	Invariato
B.2.7	Tempo pieno o tempo parziale	Obbligatorio	Invariato
B.2.7.1	Parte di tempo pieno	Obbligatorio	Invariato
B.2.8	Tipo di contratto di lavoro	Obbligatorio	Invariato
B.2.9	Cittadinanza	Facoltativo	Invariato
B.2.10	Copertura da parte di un piano governativo per promuovere l'occupazione	Facoltativo	Invariato
B.2.11	Periodo totale d'interruzioni di carriera-mesi	Facoltativo	Invariato
B.3.0	Media della retribuzione oraria lorda nel mese di riferimento	Obbligatorio	Variato
B.3.1	Totale retribuzioni lorde nel mese di riferimento	Obbligatorio	Variato
B.3.1.1	Retribuzioni relative alle ore di straordinario	Obbligatorio	Invariato
B.3.1.2	Retribuzione speciale per lavoro a turno	Obbligatorio	Invariato
B.3.2	Retribuzioni totali lorde annuali nell'anno di riferimento	Obbligatorio	Variato
B.3.2.1	Numero di settimane a cui si riferisce la retribuzione lorda annuale	Obbligatorio	Invariato
B.3.2.2	Totale annuale dei premi	Obbligatorio	Invariato
B.3.2.2.1	Premi ordinari non corrisposti ogni periodo	Facoltativo	Invariato
B.3.2.2.2	Premi annuali basati sulla produttività	Facoltativo	Invariato

[1] Lo status si riferisce **solo** all'applicazione di una qualsiasi metodologia di riduzione del rischio di re-identificazione.

B.3.2.2.3	Premio annuale di compartecipazione agli utili	Facoltativo	Invariato
B.3.3	Contributi sociali e tasse pagate dal datore di lavoro a nome del lavoratore alle autorità governative durante il mese di riferimento	Facoltativo	Invariato
B.3.3.1	Contributi sociali obbligatori	Facoltativo	Invariato
B.3.3.2	Tasse e imposte	Facoltativo	Invariato
B.3.4	Numero di ore retribuite durante il mese di riferimento	Obbligatorio	Invariato
B.3.4.1	Numero di ore di straordinario retribuite nel mese di riferimento	Obbligatorio	Invariato
B.3.5	Giorni di assenza annuali	Obbligatorio	Invariato
B.3.5.1	Numero di giorni di ferie	Obbligatorio	Invariato
B.3.5.1.1	Diritto alle ferie o numero di giorni	Obbligatorio	Invariato
B.3.5.2	Giorni di assenza per malattia annuali	Facoltativo	Invariato
B.3.5.2.1	Giorni di assenza per malattia annuali pagati del datore di lavoro	Facoltativo	Invariato
B.3.5.2.2	Giorni di assenza per malattia annuali non pagati dal datore di lavoro	Facoltativo	Invariato
B.3.5.3	Giornate di addestramento professionale annuali	Facoltativo	Invariato
B.3.6	Stima annuale delle retribuzioni annuali in natura	Optional	Invariato
B.4.2	Coefficiente di riporto all’universo del dipendente	Obbligatorio	Invariato

Nel prossimo paragrafo saranno forniti maggiori dettagli sulla perturbazione delle variabili.

3. Metodologia statistica per la tutela della riservatezza

3.1 Scenari di violazione della riservatezza

Seguendo la struttura gerarchica del file di microdati SES, sono stati adottati due scenari di violazione della riservatezza: uno per le imprese ed uno per i dipendenti

3.2. Modifiche del contenuto informativo delle variabili

3.2.1 Soppressione della variabili

I seguenti tipi di variabili sono stati soppressi dal file dei microdati da rilasciare.

1. Identificativi diretti (nome, indirizzo, etc).
2. Variabili che non erano state osservate nell'indagine italiana (la maggioranza delle variabili facoltative nell'indagine europea; per esempio la cittadinanza)
3. Variabili chiave considerate troppo dettagliate:
    a. Numero di dipendenti dell'unità locale.

Al fine di mantenere la struttura del file di microdati, l'operazione di soppressione è consistita nell'applicazione del valore "OPT" per ciascun valore di ciascuna variabile soppressa.

3.2.2 Ricodifica globale

Per il rilascio del file di microdati per scopi scientifici della SES 2002 italiana, sono state ricodificate le seguenti variabili:

1. Numero di dipendenti dell'impresa (variabile A1.2) è stata aggregata in 4 categorie: *E10 – 49, E50 – 249, E250-999, E1000+*.
2. Le divisioni NACE 10-14 sono state aggregate insieme in una nuova categoria chiamata *R10*. Le divisioni NACE 15-16 sono state aggregate in una nuova categoria chiamata *R15*.
3. L'età (variabile B2.2) è stata ricodificata in 14-19, 20-29, 30-39, 40-49, 50-59, 60+. Queste categorie sono state chiamate rispettivamente "*20)","[20-30)", "[30-40)"," [40-50)","[50-60)"* e "*[60"*.

3.3. Protezione delle imprese

La protezione delle imprese a rischio di re-identificazione è stata raggiunta attraverso una procedura di ricodifica globale. La procedura di ricodifica globale è stata applicata alla variabile *SizeEnt*. La variabile che ne è risultata è stata chiamata *Size.* La ricodifica

globale è stata effettuata rispetto alle frequenze della popolazione. Questo significa che nessuna delle combinazioni della popolazione di *NACE*, *NUTS* e *Size* contiene un numero molto ridotto di unità. I dettagli sul tipo di ricodifica e della frequenza corrispondente sono riportati nella Tavola2.

3.4. Protezione dei dipendenti

Solo le variabili *AnnualEarnings* e *MontlyEarnings* sono state modificate e solo per i record dei dipendenti considerati a rischio di re-identificazione. La protezione della variabile *AnnualEarnings* è stata ottenuta utilizzando una tecnica di perturbazione controllata. La perturbazione dei dipendenti a rischio di re-identificazione è stata realizzata attraverso un modello di regressione vincolata. La variabile *MonthlyEarmings* è stata modificata in modo proporzionale.

4. Analisi del contenuto informativo

4.1 Qualità dell'informazione sulle imprese

SizeEnt è stata ricodificata inizialmente in 4 classi. Quindi, sono state create classi dimensionali maggiormente aggregate per garantire la protezione delle imprese. In particolare, sono state aggregate le sole classi dimensionali che contengono casi singoli o doppi nel campione e nella popolazione. La tavola che segue presenta le frequenze campionarie risultanti per la variabile *SizeEnt*.

Tavola 2 – Distribuzione di frequenza delle classi dimensionali nel file di microdati per scopi scientifici della SES2002

SizeEnt	**Frequenza del file di microdati per scopi scientifici**
E10-49	4794
E1000	213
E250-999	1112
E50-249	2322
E10-49_E50-249	53
E10-49_E50-249_E250-999_E1000	13
E250-999_E1000	152
E50-249_E250-999	152
E50-249 _ E250-999 _ E1000	6

4.2 Qualità dell'informazione sui dipendenti

Sono stati modificati solo i record dei dipendenti a rischio di re-identificazione; ciò vuol dire che solo lo 0.39% dei record dei dipendenti sono stati modificati. La tavola che segue presenta statistiche riassuntive della perturbazione relativa assoluta (in percentuale) introdotta nei record a rischio di re-identificazione sulla variabile *AnnualEarnings*.

Tavola 3 – Statistiche riassuntive della perturbazione relativa assoluta dei record a rischio di re-identificazione (variabile *AnnualEarmings*).

Min	**Q1**	**Mediano**	**Medio**	**Q3**	**Max**
0.12	3.99	10.82	14.91	20.73	100

Coerenza

Poichè la *MonthlyEarnings* è stata modificata proporzionalmente rispetto alla *AnnualEarnings*, la loro coerenza è preservata automaticamente.

La variabile *Average gross hourly earnings in the representative month* è stata calcolata come rapporto rispetto alla *MonthlyEarnings* perturbata, quindi la loro coerenza è mantenuta.

Le variabili relative al tempo (numero di ore lavorate, giorni di assenza, etc) non sono state per nulla modificate, pertanto la loro coerenza interna è stata preservata.

I valori della variabile *MonthlyEarnings* risultano ancora essere in tutti i casi maggiori di quelli della *Special payment for shift work*. Il loro coefficiente di correlazione è rimasto invariato (0.03) dopo la perturbazione.

I valori di *MonthlyEarnings* risultano ancora in tutti i casi essere maggiori di *Earnings related to overtime*. Il loro coefficiente di correlazione è rimasto invariato (0.14) dopo la perturbazione.

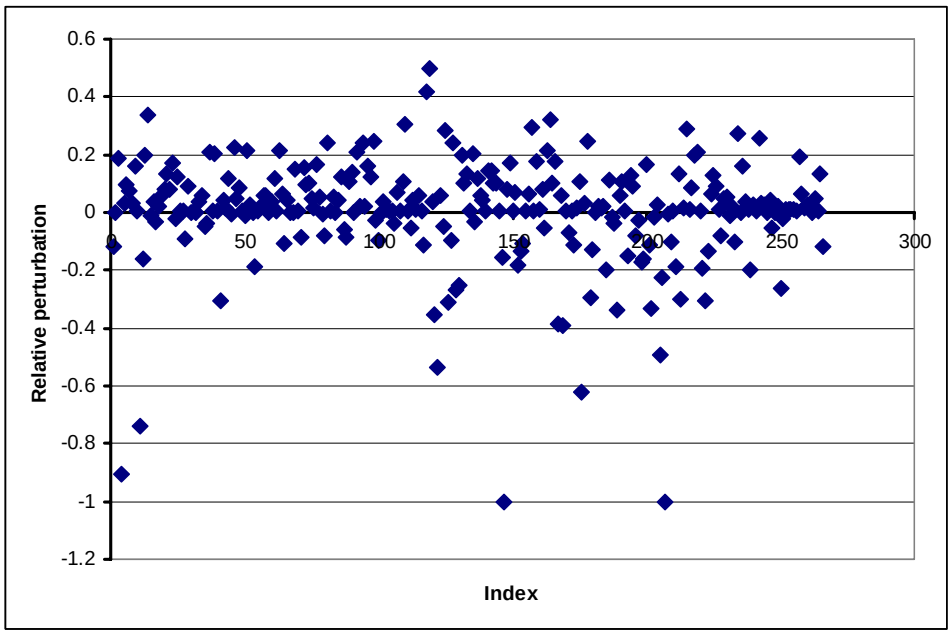
I valori di *AnnualEarnings* sono risultati ancora maggiori di *Total annual bonuses*. Il loro coefficiente di correlazione è rimasto invariato (0.59) dopo la perturbazione.

I valori di *AnnualEarnings* sono risultati essere sempre maggiori di quelli di *Annual bonuses based on productivity*. Il loro coefficiente di correlazione è cambiato da 0.41 (valore originale) a 0.40 (valori perturbati).

Confronto fra medie

Il file di microdati contiene 26295 combinazioni delle variabili *Nace*, *Nuts*, *SizeEnt*, *Gender*, *Age*, *FtPt*, *ManPos*, *Occup*. Solo 263 (1%) di queste sono state modificate in seguito all'applicazione del metodo di perturbazione. Il grafico seguente illustra la distribuzione dei cambiamenti relativi nelle medie ponderate della variabile *AnnualEarnings*[2] per tutte le 263 combinazioni che sono state modificate dalla procedura di perturbazione.

[2] Ovviamente, le medie delle stesse combinazioni sono state modificate considerando anche la variabile *MonthlyEarnings*.



La media di queste perturbazioni relative è uguale a 0.005.

Le modifiche delle medie ponderate[3] delle variabili di *AnnualEarnings* e *MonthlyEarnings* sono state confrontate con tutte le combinazioni di ciascuna delle variabili *Nace*, *Nuts*, *Size*, *Gender*, *Age* ed *Occup*. Le prossime tavole presentano la perturbazione relativa (in percentuale) delle medie rispetto alle categorie di queste variabili.

Tavola 4 – Perturbazioni relative per ciascuna modalità della variabile *Nace*.

Nace	*AnnualEarnings*	*MonthlyEarnings*	*Nace*	*AnnualEarnings*	*MonthlyEarnings*
R10	0.00	0.00	*R37*	0.00	0.00
R15	0.01	0.01	*R40*	0.00	0.00
R17	0.01	0.01	*R41*	0.01	0.01
R18	0.00	0.00	*R45*	0.00	0.00
R19	0.00	0.00	*R50*	0.00	0.00
R20	0.00	0.00	*R51*	0.01	0.01
R21	0.01	0.00	*R52*	0.00	0.00
R22	0.01	0.03	*R55*	0.01	0.00
R23	0.02	0.02	*R60*	0.00	0.00
R24	0.01	0.00	*R61*	0.01	0.01
R25	0.00	0.00	*R62*	0.00	0.00
R26	0.00	0.00	*R63*	0.00	0.00
R27	0.00	-0.01	*R64*	0.00	-0.01
R28	0.00	0.00	*R65*	0.01	0.00
R29	0.00	-0.01	*R66*	0.02	-0.06
R30	0.00	0.00	*R67*	0.00	0.00
R31	0.00	0.00	*R70*	0.00	0.00
R32	0.01	0.00	*R71*	0.00	0.00
R33	0.01	0.02	*R72*	0.01	0.02
R34	0.01	0.01	*R73*	0.00	0.00
R35	0.01	0.00	*R74*	0.01	-0.01
R36	0.00	0.00			

[3] Sono stati utilizzati i coefficienti di riporto all'universo.

Tavola 5 – Perturbazioni relative per ciascuna modalità della variabile *Nuts*.

Nuts	*AnnualEarnings*	*MonthlyEarnings*
ITC	0.01	0.00
ITD	0.01	0.00
ITE	0.00	0.00
ITF	0.00	0.00
ITG	0.01	0.01

Tavola 6 – Perturbazioni relative per ciascuna modalità della variabile *Gender*.

Gender	*Annual earnings*	*Monthly earnings*
F	-0.01	-0.01
M	0.01	0.01

Tavola 7 – Perturbazioni relative per ciascuna modalità della variabile *SizeEnt*.

SizeEnt	*AnnualEarnings*	*MonthlyEarnings*
E10-49	0.00	0.00
E10-49_E50-249	0.00	0.00
E10-49_E50-249_E250-999_E1000	0.00	0.00
E1000	-0.04	-0.05
E250-999	0.10	0.08
E250-999_E1000	0.02	0.02
E50-249	0.00	0.00
E50-249_E250-999	0.00	0.00
E50-249_E250-999_E1000	0.00	0.00

Tavola 8 – Perturbazioni relative per ciascuna modalità della variabile di *Age*.

AGE	*AnnualEarnings*	*MonthlyRarnings*
AGE1	0.00	0.00
AGE2	0.01	0.00
AGE3	-0.02	-0.02
AGE4	0.01	0.00
AGE5	0.13	0.10
AGE6	-1.22	-1.14

Tavola 9 – Perturbazioni relative per ciascuna modalità della variabile *Occup*.

Occup	*AnnualEarnings*	*MonthlyEarnings*	*Occup*	*AnnualEarnings*	*MonthlyEarnings*
I11	-1.73	-1.77	*I51*	0.00	0.00
I12	0.92	0.83	*I52*	0.01	0.01
I13	0.24	0.18	*I61*	0.00	0.00
I21	-0.89	-1.00	*I71*	0.00	0.00
I22	0.25	0.24	*I72*	0.00	0.00
I23	-7.26	-6.96	*I73*	0.00	0.00
I24	-2.86	-2.46	*I74*	0.00	0.00
I31	0.05	0.04	*I81*	0.00	0.00
I32	0.00	0.00	*I82*	-0.01	-0.01
I33	-0.06	-0.08	*I91*	0.00	0.00
I34	0.01	0.01	*I92*	0.00	0.00
I41	0.01	0.00	*I93*	0.00	0.00

5. Note per un'adeguata analisi statistica del file di microdati per scopi di ricerca scientifica [4]

A livello europeo, l'indagine sulla struttura delle retribuzioni è particolarmente rilevante poiché fornisce informazioni specifiche sui differenziali delle retribuzioni in base alle caratteristiche demografiche e professionali dei dipendenti da una parte e dell'unità di produzione dall'altra. Ne consegue che anche piccole variazioni degli indicatori possono essere fondamentali.

Nel paragrafo precedente sono stati presentati gli effetti principali della metodologia di riduzione del rischio di violazione della riservatezza sia sulle imprese che sui dipendenti. Come già evidenziato, se si considerano le singole variabili categoriche, gli effetti delle modifiche sui valori di *AnnualEarnings* e *MonthlyEarnings* appaiono irrilevanti: la differenza dei valori di *AnnualEarnings* e *MonthlyEarnings* nel file di microdati per scopi di ricerca scientifica rispetto al file di microdati non perturbato è minima.

L'attribuzione dell'impresa alla classe dimensionale citata nel Implementation Regulation for SES, se confrontata con la stratificazione originale, non è possibile per 376 imprese (su un totale di 8817 imprese nel campione). Le nuove classi dimensionali permettono solo un confronto parziale tra le grandi imprese (con più di 250 dipendenti) e imprese medio-piccole (con meno di 250 dipendenti). In effetti per 171 imprese non è possibile distinguere tra grandi e medie-piccole imprese.

Ne consegue che l'utente deve tener conto degli effetti della metodologia di riduzione del rischio di violazione della riservatezza nei casi elencati nella Tavola 11. Se si effettua l'analisi ad una combinazione troppo dettagliata di variabili di stratificazione, gli indicatori che ne risultano potrebbero differire in modo significativo da quelli pubblicati a livello nazionale o anche da quelli contenuti nel data base New Cronos. Comunque, queste differenze tra i file di microdati per scopi scientifici e il file di microdati non perturbati possono verificarsi solo quando le variabili *AnnualEarnings* e *MonthlyEarnings* sono coinvolte nell'analisi e solo nelle corrispondenti categorie di NACE e NUTS..

La tavola 11 contiene la lista delle combinazioni di codici NACE e NUTS 1 per le quali non si possono costruire gli indicatori statistici per tutte le classi dimensionali originali (e in particolare per le classi "imprese con meno di 250 dipendenti" e "imprese con più di 250 dipendenti").

[4] Questo paragrafo è stato redatto da S.Cardinaleschi

Tavola 11 – Lista delle combinazioni NACE 2 cifre e NUTS 1 per le quali non possono essere costruiti gli indicatori relativi alla classe dimensionale "più di 250 dipendenti"

Nace 2 cifre	NUTS1
30	ITC
10	ITG
10	TOTAL
17	ITF
20	ITE
22	ITF
23	ITG
24	ITF
26	ITF
26	ITG
27	ITG
29	ITF
29	ITG

Nace 2 cifre	NUTS1
30	ITF
33	ITF
35	ITG
37	ITE
37	TOTAL
40	ITG
50	ITD
50	ITG
61	ITD
62	ITD
62	ITF
67	ITD

Nota: la prima riga (combinazione NACE 30 e NUTS1 ITC) si riferisce alla classe dimensionale "meno di 250 dipendenti"

6. Riferimenti bibliografici

1. Ichim, D. (2008) *Controlled model-based disclosure limitation of business microdata*. XLIV Riunione Scientifica della Società Italiana di Statistica, Università della Calabria, 25 - 27 giugno 2008.
2. Ichim, D., Franconi, L. (2007) Disclosure scenario and risk assessment: Structure of Earnings Survey, *Eurostat Monographs on statistical data confidentiality*, available at http://epp.eurostat.ec.europa.eu/portal/page?_pageid=3154,70730193,3154_70730647&_dad=portal&_schema=PORTAL

7. Contatti

La responsabile della SES può essere contattata all'indirizzo: cardinal@istat.it.

La responsabile delle "Metodologie statistiche per la tutela della riservatezza" può essere contattata all'indirizzo: ichim@istat.it.

Curatori dei capitoli

Il documento è stato redatto da Daniela Ichim.