Istat
Istituto Nazionale di Statistica

Rilevazione sulla formazione del personale nelle Imprese

anno 2005

Descrizione del file

14 ottobre 2011

INDICE

1. Introduzione

I dati sezionali riguardanti la *Rilevazione sulla formazione del personale nelle Imprese* (usando l'acronimo inglese, Continuous Vocational Training Survey) per l'anno 2005 sono esaminati allo scopo di definire il grado di dettaglio del rilascio per finalità di ricerca. La necessità di tutela della riservatezza discende da vincoli di legge e da obblighi assunti verso i rispondenti. Seguendo l'Handbook on Statistical Disclosure Control (AA.VV., 2010), l'intrusione in informazioni non pubbliche si concretizza quando il singolo record è correttamente assegnato ad un record contenuto in un file esterno a disposizione dell'intrusore. La valutazione del rischio si fonda sulla definizione di uno scenario idoneo a definire la quantità di informazioni di cui egli disponga. Le unità statistiche oggetto dell'analisi sono le imprese, condizionatamente alla stratificazione di interesse. Nelle sezioni successive vengono brevemente discusse le ipotesi alla base delle analisi condotte nonché alcuni aspetti di metodo ed una sintesi dei principali risultati.

2. Descrizione delle variabili

La riduzione del rischio di violazione della riservatezza è stata perseguita tramite perturbazione aleatoria dei dati. I dettagli sono esposti nel paragrafo 3.2. Per il significato delle variabili si rinvia al tracciato record ed al questionario di indagine.

3. Metodologia statistica per la tutela della riservatezza

3.1 Valutazione del Rischio

Poiché "all units are unique (rare) with respect to a small set of quantitative variables" (AA.VV., 2010) sembra rilevante il concetto di intrusione inferenziale, ossia quanto attiene alle informazioni che possono essere dedotte con elevato grado di attendibilità dalle proprietà statistiche dei dati rilasciati. È parso ragionevole uno scenario di identificazione univariato in ragione della facilità di reperimento, sulla base dell'informazione esterna d'azienda, di indicatori espressione del concetto di dimensione aziendale. Mentre la stratificazione d'interesse è definita da NACE 20 e sei classi di addetti (10-19, 20-49, 50-249, 250-499, 500-999, 1000 e oltre), le unità da proteggere sono state enucleate condizionatamente alla NACE 30. Facendo uso delle tecniche consuete per i problemi di classificazione non supervisionata, individuate le unità statistiche appartenenti a sottoinsiemi costituiti da un numero di osservazioni inferiore ad una prefissata soglia, tutti i record di strato sono perturbati se almeno uno risulta sensibile al rischio di intrusione.

3.2 Una strategia di protezione

In ordine al mantenimento delle principali caratteristiche rilevanti per i fruitori, sono state considerate le seguenti priorità:

- quantità artificiali espresse in numeri interi e coerenti con i rispettivi range,
- accurata riproduzione di tutti i rapporti caratteristici a livello di singolo record,
- approssimativa conservazione delle somme pesate di strato,

Si tratta di scelte riconducibili al valore informativo riconosciuto ai rapporti (la proposta di Eurostat [2010] verte sulla possibilità di consentirne il rilascio, in via esclusiva, in luogo dei dati originali) e alla natura quantitativa dei caratteri rilevati. Il dettaglio informativo del dataset perturbato, definito dalla stratificazione secondo NACE 20 e sei classi di addetti, è maggiore di quello proposto da Eurostat (2010). Il secondo requisito tra quelli precedentemente menzionati, ossia l'accurata riproduzione dei rapporti, impone che tutte le variabili di una certa unità statistica siano perturbate nella stessa proporzione. Il primo vincolo implica a sua volta che il rumore moltiplicativo sia generato tenendo conto dei campi di variazione ammissibili. Il terzo richiede che, all'interno di ciascuno strato, la distorsione indotta dal condizionamento dei rumori moltiplicativi ai range sia globalmente limitata. Poiché una misura di protezione può essere definita dall'errore assoluto relativo (ad es. Foschi e Liseo, 2010) dovuto alla sostituzione dei dati originali con gli omologhi perturbati, occorre garantire che l'entità del disturbo aleatorio sia significativa. Posti w il peso di espansione, x la variabile osservata di scenario e z una qualsiasi altra variabile in numeri interi, x^* e z^* le omologhe quantità perturbate, s l'indice di uno strato di almeno due unità, un semplice metodo di protezione può essere espresso dalle relazioni:

$$m_{xs} \equiv \sum_{i \in s} w_{is} x_{is} \Big/ \sum_{i \in s} w_{is}\,, \qquad \varepsilon_{is} = \alpha_s \cdot \left(x_{is} / m_{xs} - 1 \right), \qquad \alpha_s \in (0,1)$$

$$w_{is}^* = \max\left[w_{is} / (1+\varepsilon_{is}),\, 1 \right], \qquad x_{is}^* = \left\lfloor m_{xs} \cdot (1+\varepsilon_{is}) + 0.5 \right\rfloor, \quad z_{is}^* = \left\lfloor z_{is} \cdot (x_{is}^* / x_{is}) + 0.5 \right\rfloor$$

In parole, ciascuna intensità dello strato s^{mo} viene moltiplicata per un rumore tale che:

- la variabile di scenario assuma valori ammissibili,
- la discrepanza tra somme di strato e omologhi valori perturbati sia limitata.

Il parametro α_s governa il trade-off tra protezione ed accuratezza (massimizzate agli estremi dell'intervallo unitario, rispettivamente da $\alpha_s \to 0$ e $\alpha_s \to 1$) nello strato s^{mo}. Negli strati di una sola unità statistica si possono adottare quali rumori moltiplicativi i rapporti tra medie pesate relative ad una stratificazione meno sottile (ovvero intensità artificiali dell'intervallo ammissibile) e valori osservati per la variabile di scenario, modificando i coefficienti di espansione (ove il vincolo di pesi non inferiori ad uno sia soddisfatto) in modo da mantenere sostanzialmente immutati i totali di popolazione. Per i rumori negli

strati di una sola unità si considerano NACE 6 e sei classi di addetti.

4. Sintesi dei risultati

Le ascisse dei diagrammi in appendice indicano le classi di attività economica e addetti dello strato di riferimento nonché le rispettive numerosità. Nel calcolo delle statistiche, i record recanti i codici di questionario relativi alle risposte non dovute sono stati ignorati. Il grafico 1 evidenzia i valori medio, minimo e massimo riferiti ai rapporti tra le somme pesate di strato sintetiche e originali. La figura 2 permette di cogliere in dettaglio il mantenimento delle somme pesate per ognuna delle sessantuno variabili lungo l'asse delle ordinate: i cerchi blu individuano i casi in cui i rapporti esprimono una deviazione percentuale assoluta inferiore al 10%, e i triangoli rossi quelli in cui la variabile di riferimento non costituisce oggetto di rilevazione. In questo modo la figura 2 fornisce indicazione della qualità dei dati perturbati per ogni variabile in ciascuno strato. Le frequenze relative degli errori di segno su tutti i coefficienti di correlazione al variare degli strati sono esposte graficamente nel diagramma 3. Con significato analogo a quanto illustrato circa il secondo grafico, i segni riprodotti correttamente per le principali correlazioni sono esposti nella figura 4.

5. Riferimenti bibliografici

AA.VV. *Handbook on Statistical Disclosure Control*. CENtre of EXcellence for Statistical Disclosure Control, 2010. http://neon.vb.cbs.nl/casc/handbook.htm. 18/08/2010.

Eurostat, Directorate F, Unit F-4 (2010), *CVTS 2005 anonymisation criteria*, Doc TF CVTS4/10/02.

Foschi F., Liseo B. (2010), *Artificial Continuous Data for SDC.* Contributi Istat 5.

6. Contatti

Per i dati dell'Indagine CVTS: **cvts@istat.it** (DICS/DCSP).

Per la tutela della riservatezza: **rilascio.microdati@istat.it** (DIQR/DCIQ)

7. Appendice

Nelle pagine seguenti, alcuni grafici illustrano il trade-off tra protezione e accuratezza:

Figura 1

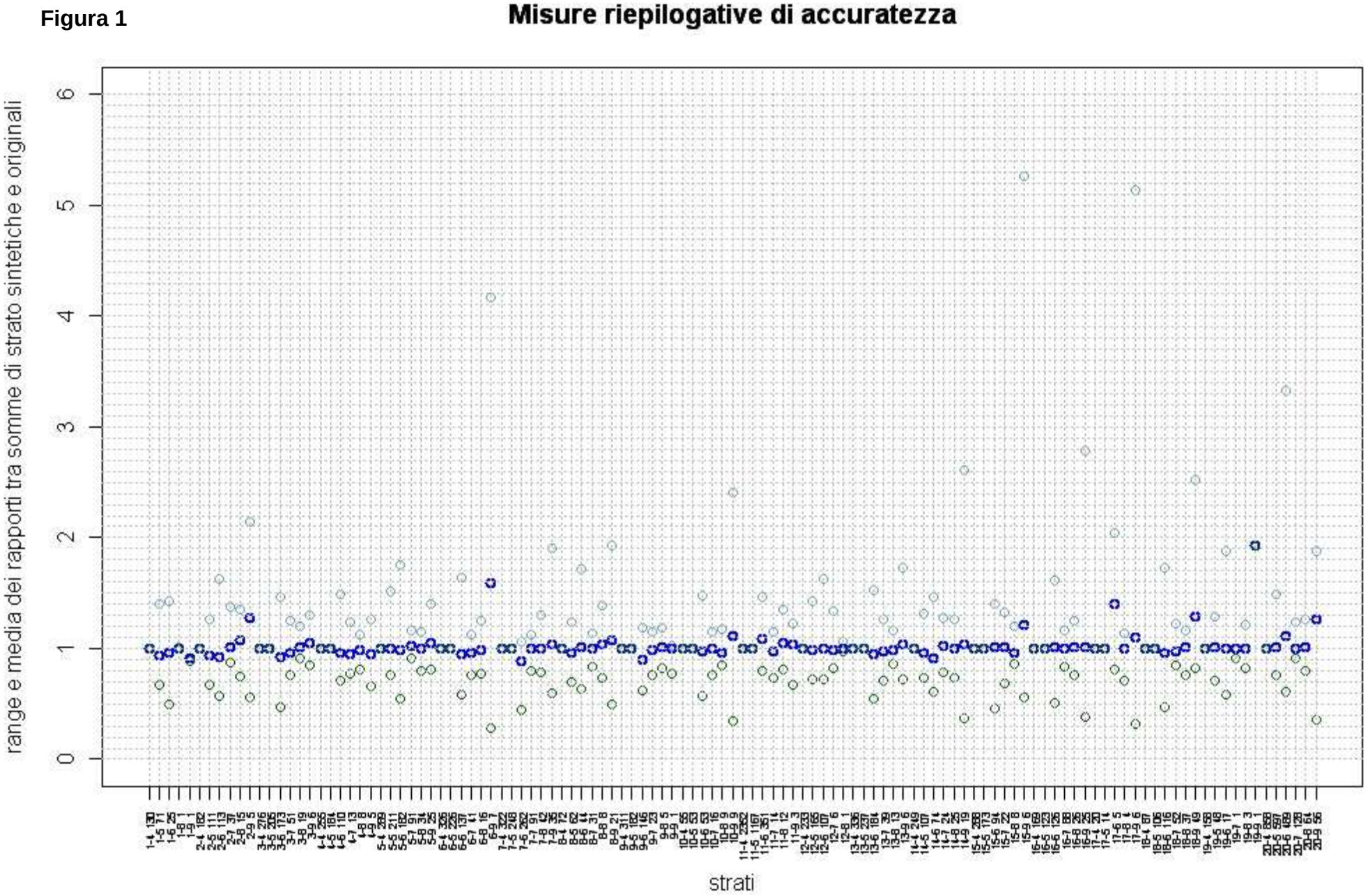

Misure riepilogative di accuratezza
range e media dei rapporti tra somme di strato sintetiche e originali
0
1
2
3
4
5
6
strati

Figura 2

Accuratezza: scarti inferiori al 10% tra totali di sintesi e originali

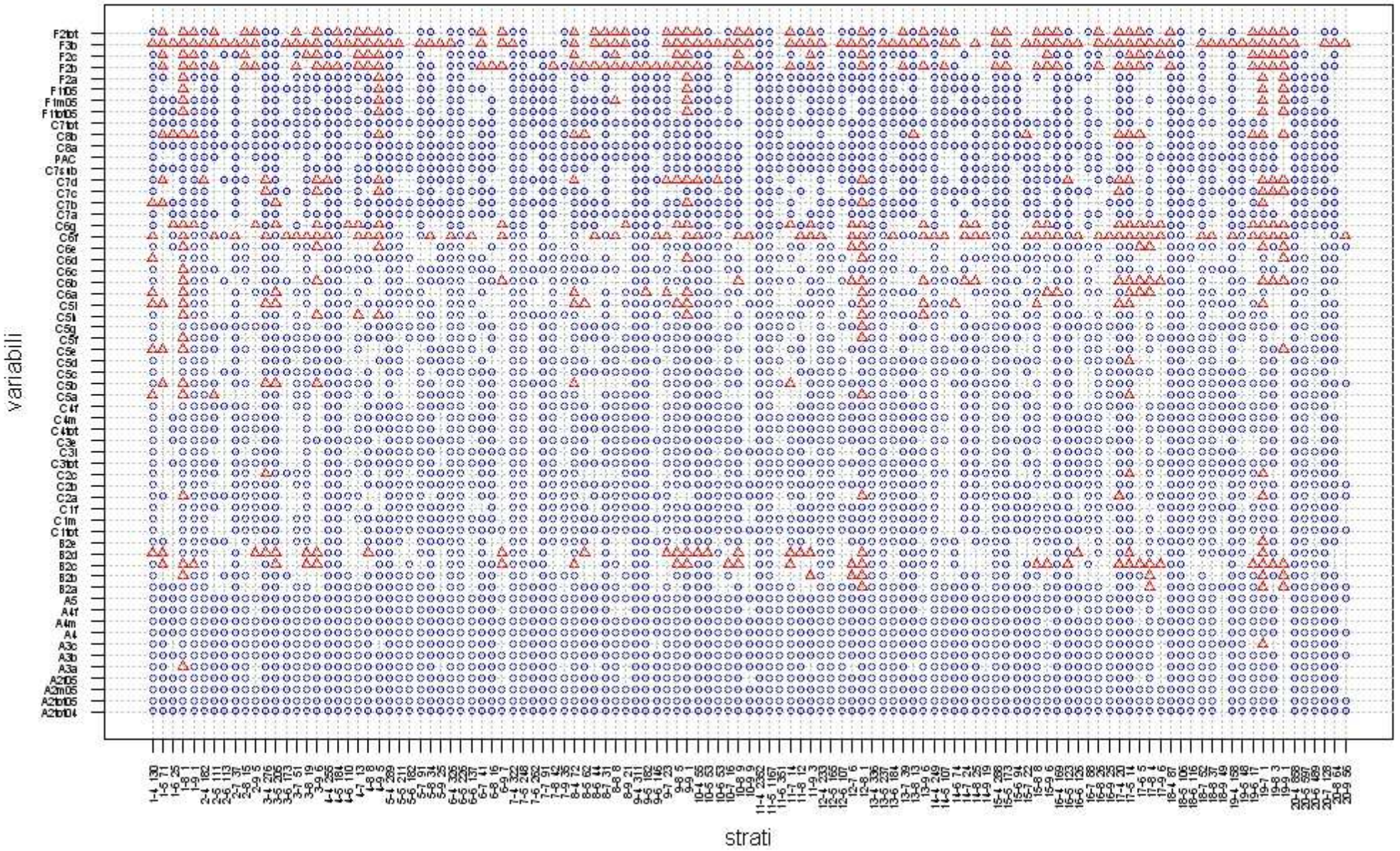


Figura 3

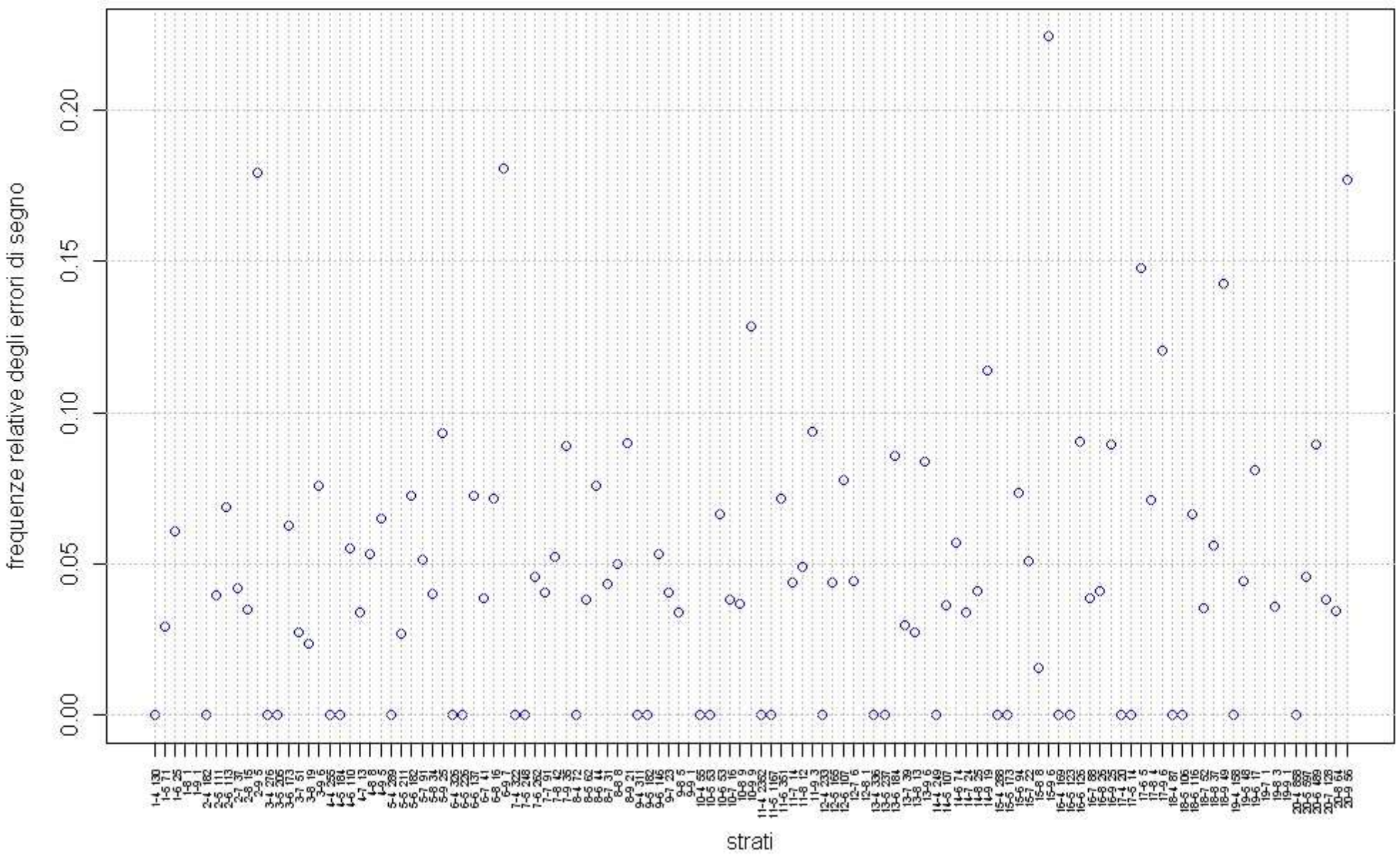
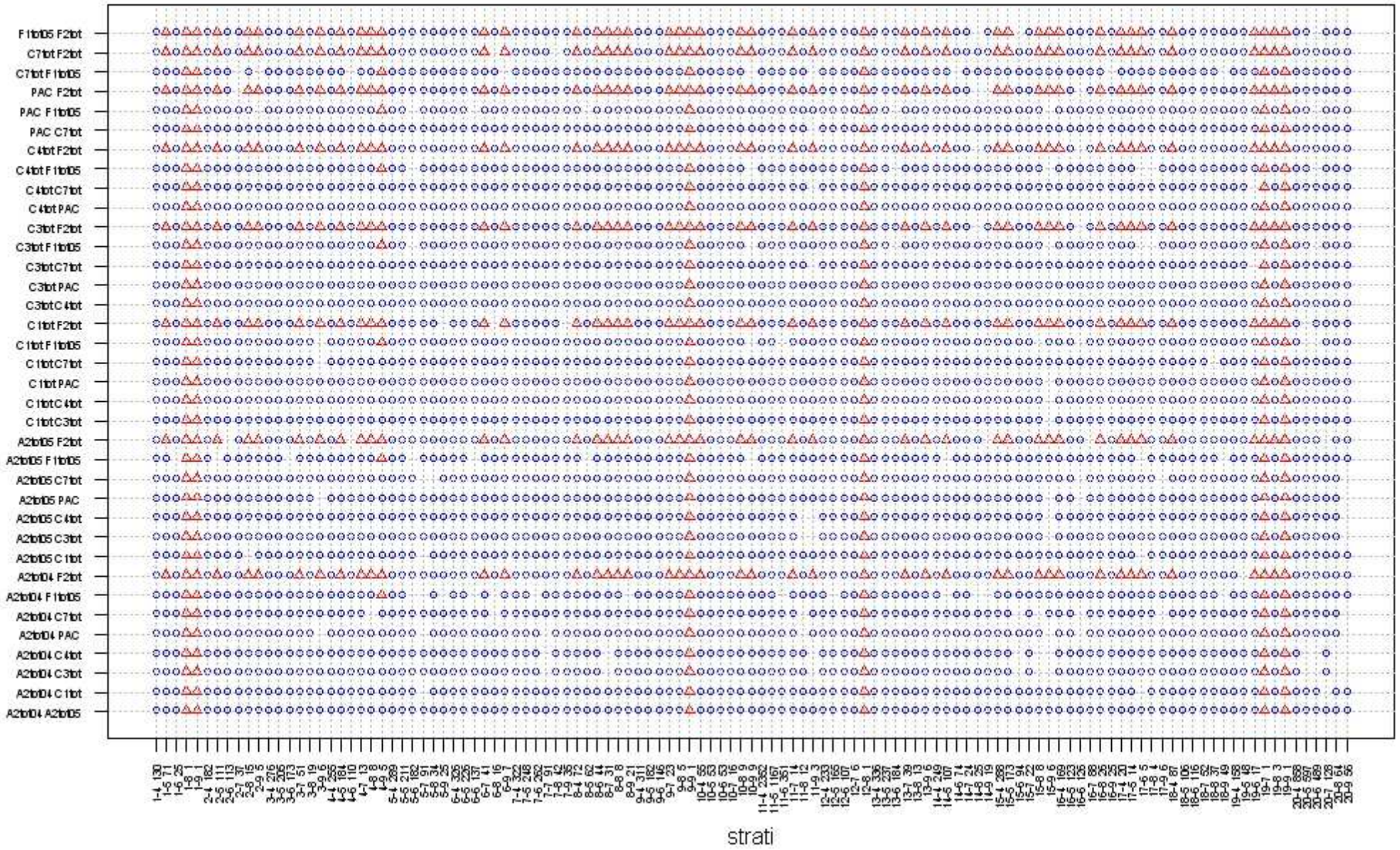

Confronto tra segni delle correlazioni originali e sintetiche
frequenze relative degli errori di segno
0.00
0.05
0.10
0.15
0.20
strati

Figura 4

Segni corretti per le principali correlazioni sintetiche



Curatore

Flavio Foschi (DIQR/DCIQ)