

**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**

*Versione 2.0*

**Autori**

*Giorgio Della Rocca (\*)*

*Marco Di Zio (\*)*

*Orietta Luzi (\*)*

*Giorgia Simeoni (\*)*

(\*) ISTAT - Servizio MTS

(\*\*) ISTAT - Servizio PSM

## ***Sommario***

Nel lavoro vengono descritti gli indicatori per la valutazione degli effetti su dati si indagine dovuti all'applicazione di procedure di controllo e correzione disponibili nel software IDEA (Indices for Data Editing Assessment). Alcuni degli indicatori consentono la valutazione dell'impatto del processo di controllo e correzione sui dati originali in termini di effetto sui dati elementari, sulle distribuzioni e sulle relazioni fra variabili. Inoltre, IDEA consente il calcolo e la visualizzazione degli indicatori di valutazione per la fase di revisione previsti in SIDI, consentendo anche la predisposizione del file da utilizzare come input per SIDI stesso. Nel documento sono inoltre illustrate le caratteristiche tecniche e operative dello strumento.

## ***Abstract***

In the contribution the indicators for evaluating the effects on survey data of editing and imputation processes available in the software IDEA (Indices for Data Editing Assessment) are described. Some indicators allow the evaluation of the impact of the editing and imputation process on original survey data in terms of effect on micro data, distributions and relations among variables. Furthermore, IDEA allows the computation and the analysis of the set of indicators (relating to the editing and imputation process) required by the SIDI system (Sistema Informativo di Documentazione delle Indagini). In particular, IDEA produces a file containing all the information on the required quality indicators to be used as input for SIDI. In the contribution the technical and operational characteristics of the tool are also described.

# INDICE

## 1. INTRODUZIONE

### 1.1. LO STRUMENTO IDEA: ASPETTI GENERALI

## 2. GLI INDICATORI DI VALUTAZIONE

### 2.1. INDICATORI DI VALUTAZIONE A LIVELLO DI VARIABILI

#### 2.1.1. *Variabili categoriche*

- 2.1.1.1. Modifica dei valori elementari
- 2.1.1.2. Modifica delle distribuzioni marginali
- 2.1.1.3. Modifica delle distribuzioni congiunte
- 2.1.1.4. Modifica delle relazioni (associazioni) doppie

#### 2.1.2. *Variabili numeriche continue*

- 2.1.2.1. Modifica dei valori elementari
- 2.1.2.2. Modifica delle distribuzioni marginali e degli aggregati
- 2.1.2.3. Modifica delle relazioni doppie

### 2.2. INDICATORI DI VALUTAZIONE A LIVELLO COMPLESSIVO: INDICATORI SIDI

## 3. IL SISTEMA

### 3.1. ASPETTI GENERALI

### 3.2. DATI DI INPUT

### 3.3. INDICATORI DI VALUTAZIONE A LIVELLO DI VARIABILI

#### 3.3.1. *Variabili categoriche nominali o ordinali*

#### 3.3.2. *Variabili numeriche continue*

### 3.4. INDICATORI SIDI

## 4. ASPETTI TECNICI: REQUISITI HARDWARE E SOFTWARE

## RIFERIMENTI BIBLIOGRAFICI

## 1. INTRODUZIONE

In ogni processo di indagine le informazioni fornite dalle unità contattate vengono sottoposte a diverse elaborazioni (si parla in genere di *trattamento dei dati*) volte a rendere i dati rilevati adeguati alle analisi statistiche obiettivo della rilevazione. La misurazione degli effetti di tali elaborazioni sulla struttura e sulle proprietà statistiche dei dati osservati rappresenta un problema di importanza centrale. Tale misurazione può avere diversi obiettivi: *documentare* gli effetti delle elaborazioni sui dati originali, *monitorare* tali effetti nel tempo, *ottimizzare* i costi, i tempi di elaborazione e la qualità dei risultati finali attraverso l'analisi e la rimozione di eventuali inefficienze del processo, ecc.

Nel campo della Statistica Ufficiale, una fase di elaborazione di dati particolarmente critica è rappresentata dal *processo di controllo e correzione/imputazione*, in cui i dati osservati e registrati vengono sottoposti a elaborazioni volte a individuare e rimuovere dai dati quegli errori non campionari (incluse le mancate risposte) che danno luogo a incoerenze di tipo logico, matematico, statistico. In generale, attraverso il processo di procedura di controllo e correzione si cerca di rendere i dati completi e coerenti rispetto a prefissati criteri di accuratezza. E' noto infatti che, nonostante l'utilizzo di approcci alla rilevazione e alla registrazione dei dati volti a garantire massima accuratezza nei dati registrati (ad esempio attraverso l'uso di tecniche di rilevazione Computer Aided) questi restano generalmente affetti da errori di diversa natura (valori anomali, valori mancanti, valori incoerenti, ecc.). Inoltre, le diverse tipologie di errore possono contaminare in diverso modo e con diverse incidenze i parametri obiettivo della rilevazione.

Nel seguito del lavoro, con l'espressione *procedura di controllo e correzione*<sup>1</sup> (C&C) s'intende un insieme integrato e complementare di metodologie, ciascuna finalizzata al trattamento di un particolare tipo di unità/variabili/errori non-campionari, ma aventi il comune obiettivo di produrre dati finali completi e coerenti (a livello micro e/o aggregato). Il problema della valutazione di una procedura di C&C può essere affrontato secondo due approcci distinti (Granquist, 1997; Di Zio et al., 2002):

- misurare gli *effetti* della procedura di C&C su un insieme di dati iniziali a fini di documentazione, monitoraggio, miglioramento della procedura di C&C;
- valutare la *qualità* della procedura di C&C in termini di capacità di individuare correttamente gli errori e/o ripristinare i valori veri posseduti dalle unità rilevate.

E' evidente che il secondo tipo di valutazione può essere effettuato solo conoscendo, per ogni unità rilevata, i corrispondenti dati *veri*: un approccio poco costoso (anche se non privo di controindicazioni) per porsi in questa situazione teorica è quello della *simulazione*. Tale approccio è

---

<sup>1</sup> Il termine correzione in questo caso include anche il concetto di *imputazione* delle mancate risposte.

applicabile, ad esempio, utilizzando il software E.S.S.E. (*Editing Systems Standard Evaluation*) (Barcaroli et al., 2001) per la generazione di errori e di mancate risposte parziali in dati completi e coerenti. Indicatori di prestazione utilizzabili in questo ambito sono discussi ad esempio in Chambers (2001).

Il primo approccio è tipico invece di situazioni operative in cui, dato un insieme di dati grezzi, sia stata applicata ad essi una procedura di C&C e sia stato ottenuto il corrispondente insieme di dati *puliti* (cioè coerenti e privi di valori mancanti): in questa situazione, l'obiettivo è generalmente quello di misurare l'entità delle variazioni prodotte sui dati iniziali dalla procedura di C&C utilizzata. In questo caso è necessario misurare l'entità di tali variazioni a livello sia micro (dati elementari) sia macro (distribuzioni marginali e congiunte, stime, relazioni fra variabili).

La disponibilità di queste misure può essere utilizzata a scopi diversi, a seconda della fase del processo di C&C in cui si opera: *in fase di test della procedura*, per verificare la presenza di inefficienze su particolari unità, variabili, strati di analisi ecc. ed essere in grado di apportare le opportune modifiche a livello di strategia complessiva di C&C o di singole metodologie, di gerarchie fra variabili o tipologie di errori ecc.; *in corso di applicazione ai dati correnti*, per verificare e monitorare il corretto procedere delle elaborazioni effettuate sui dati; *in fase di rilascio dei risultati*, per documentare l'entità delle variazioni prodotte sui dati originali dalla specifica procedura di C&C adottata; *in fase di analisi*, per monitorare gli effetti della procedura nel tempo o su specifiche sottopopolazioni, oppure confrontarli con quelli ottenuti da altre indagini.

Nel presente lavoro si illustrano le caratteristiche metodologiche e le funzionalità tecniche dello strumento IDEA (*Indices for the Data Editing Assessment - Indicatori per la valutazione degli effetti di procedure di controllo e correzione dei dati e per il calcolo degli indicatori SIDI*).

Nel paragrafo 1.1 sono descritte le problematiche generali che possono essere affrontate attraverso IDEA nell'area della valutazione di procedure di C&C. Nel paragrafo 2 sono descritti i criteri di valutazione ed i corrispondenti indicatori disponibili in IDEA. Nel paragrafo 3 è descritta la struttura del software e la modalità di utilizzo delle diverse funzioni in esso disponibili. Il paragrafo 4 contiene istruzioni di tipo tecnico su requisiti hardware e software e sulle modalità di installazione per l'utilizzo di IDEA.

### **1.1. Lo strumento IDEA: aspetti generali**

IDEA è stato realizzato con l'obiettivo di rendere possibile il confronto fra due opportuni insiemi di dati, contenenti informazioni sugli stessi fenomeni osservati sulle stesse unità, al fine di misurarne le differenze secondo i due approcci illustrati nell'introduzione:

- 1) valutare gli effetti di procedure di C&C su un insieme di dati grezzi;
- 2) valutare la qualità di una procedura o di un metodo di C&C.

### 1) *Valutare gli effetti di procedure di C&C su un insieme di dati grezzi*

In questo contesto, il problema consiste nel valutare l'impatto di una procedura di C&C ponendo a confronto un insieme di dati *grezzi* e il corrispondente insieme di dati *puliti* ottenuto mediante applicazione della procedura di C&C oggetto di valutazione. IDEA consente di effettuare tale valutazione a due diversi livelli:

1. a livello 'alto' (attraverso l'analisi del complesso delle variabili/unità rilevate), per la documentazione o il monitoraggio dell'impatto della procedura di C&C sui dati nel complesso;
2. a livello 'basso' (attraverso l'analisi delle singole variabili osservate o di sottogruppi di esse), per la documentazione o il monitoraggio degli effetti della procedura di C&C sulle caratteristiche micro/macro dei singoli item.

Nel primo caso, l'obiettivo è di effettuare una valutazione *complessiva* degli effetti della procedura di C&C (Fortini et al., 1999; Fortini et al., 2000) al fine di monitorare e documentare nel tempo tali effetti e rendere anche possibile il confronto fra indagini diverse (nel tempo o in certo periodo di riferimento). In questo caso, gli indicatori standard calcolati in IDEA sono quelli gestiti nel Sistema Informativo di Documentazione delle Indagini SIDI (Brancato et al., 2001; Simeoni, 2001). Più precisamente, IDEA risponde alla doppia esigenza di calcolare gli indicatori standard di SIDI sia per fini di analisi statistica dei dati, sia per fini strettamente operativi: IDEA produce infatti in output il file richiesto come input di SIDI per l'aggiornamento degli indicatori di qualità sul processo di revisione dei dati. In questo senso, IDEA rappresenta uno strumento operativo per la standardizzazione e la semplificazione del processo di calcolo e aggiornamento degli indicatori SIDI da parte dei responsabili di indagine (Della Rocca et al., 2003).

Nel secondo caso, cioè a livello di singole variabili (o sottoinsiemi di esse), lo strumento può essere utilizzato in generale per i seguenti obiettivi di valutazione:

- i) valutare gli effetti statistici della procedura di C&C a livello micro e macro, generalmente al fine di documentare la qualità dei dati iniziali su specifici fenomeni (attraverso l'analisi del tipo e dell'ammontare di errori riscontrati e trattati);
- individuare inefficienze:
  - o della procedura di C&C o di singole sottofasi di essa (da modificare in fase di disegno oppure in vista delle successive ripetizioni dell'indagine);
  - o di altre fasi dell'indagine (ad esempio, errori sistematici dovuti al questionario);
- ii) valutare gli effetti su prefissate variabili dovuti a cambiamenti nella strategia di C&C.

Al fine di consentire queste analisi, sono stati implementati in IDEA alcuni semplici indicatori che tengono conto dei seguenti aspetti principali:

1. differenze prodotte dalla procedura di C&C sui microdati, in termini di *numero* di valori diversi e/o di *entità* delle differenze;
2. differenze prodotte nelle distribuzioni marginali e congiunte delle variabili osservate.
3. differenti prodotte sulle relazioni fra variabili.

Per ogni variabile, o sottogruppo di variabili, gli indicatori possono essere calcolati utilizzando tutte le unità del campione oggetto di analisi, oppure un sottoinsieme di esso. In quest'ultimo caso, il sottoinsieme è costituito dalle unità per le quali la procedura di C&C ha prodotto una modifica nel valore originale della/e variabile/i considerata/e (*sottoinsieme dei dati modificati*). In questo modo è possibile effettuare un'analisi più approfondita e dettagliata. Inoltre, restringere l'attenzione ai soli dati modificati può essere vantaggioso, soprattutto nei casi in cui la percentuale di tali valori è bassa rispetto al totale del campione osservato: in questi casi una valutazione che tenga conto di tutto il campione potrebbe diluire eccessivamente gli effetti della procedura di C&C, e non permettere di evidenziare eventuali effetti distorsivi sulle variabili o sulle loro distribuzioni marginali/congiunte.

Gli indicatori disponibili in IDEA sono inoltre calcolabili separatamente per *strati* o *domini* opportunamente definiti.

Si sottolinea come per valutazioni su specifiche metodologie/fasi del processo di C&C (ad esempio tecniche localizzazione degli errori casuali, imputazione delle mancate risposte parziali, trattamento dei casi influenti ecc.) è sufficiente confrontare i dati in input alla sottofase con quelli ottenuti in seguito all'applicazione dello specifico metodo di trattamento oggetto di valutazione.

Un problema aperto, non affrontato in IDEA ma che merita di essere menzionato riguarda il problema della valutazione degli effetti del C&C sulle stime obiettivo dell'indagine. In IDEA è possibile valutare esclusivamente le differenze esistenti fra "stime" puntuali di valori (sostanzialmente medie e totali, ponderati o non) prima e dopo il processo di C&C. Misure di distanza sintetiche fra tali valori (Chambers, 2001) saranno rese disponibili in successive versioni dello strumento. Rimane insoluto il problema della misurazione delle componenti non campionarie della varianza delle stime dovute alla presenza di errori/mancate risposte parziali e al processo di trattamento, per il quale sono necessari approcci più complessi. Fra i metodi sviluppati in letteratura per stimare correttamente la precisione delle stime tenendo conto delle componenti non campionarie, i più efficaci sono tecniche di ricampionamento (vedere fra gli altri, Lee et. al., 2001; Rao, 2001; Beaumont et al., 2002), e l'imputazione multipla (Rubin, 1987; Schafer, 1997).

## 2) *Valutare la qualità di procedure di C&C*

Relativamente all'uso degli indicatori di IDEA per valutare la *qualità* di una procedura (o di un singolo metodo) di C&C, in questo caso il confronto viene effettuato fra coppie di insiemi di dati contenenti stesse informazioni osservate su uno stesso campione di unità, relative alla situazione

‘vera’ ed alla corrispondente situazione ottenuta mediante applicazione di una procedura/metodo di C&C. Questo tipo di valutazione scaturisce evidentemente da situazioni sperimentali in cui l’obiettivo della valutazione è verificare la capacità di una procedura di controllo/correzione di identificare/ripristinare correttamente gli errori presenti in un insieme di dati di cui si conosce la situazione *vera*. Un tipico modo di procedere in questo caso consiste nel simulare errori/mancate risposte in un certo insieme di dati completi e coerenti secondo prefissati modelli e incidenze, quindi applicare la procedura di controllo e correzione da valutare e procedere alla valutazione a livello micro e/o macro attraverso il confronto fra dati iniziali veri e dati finali editati e imputati. Una tipica applicazione di questo approccio corrisponde all’obiettivo di valutare comparativamente le prestazioni di più metodi concorrenti a fronte di un certo insieme di dati/variabili/errori/meccanismi di errore o di mancate risposte. E’ questo il caso delle analisi effettuate nel corso del progetto europeo Euredit (<http://www.cs.york.ac.uk/euredit/>).

Alcuni degli indicatori previsti in IDEA sono utilizzabili anche per questo tipo di valutazione, in particolare quelli che consentono valutazioni a livello aggregato. E’ evidente che, quando si vuole misurare la qualità di una procedura di C&C, la valutazione è efficace se viene effettuata considerando, per ogni variabile, il sottoinsieme dei *dati modificati*, cioè le sole unità in cui la procedura di C&C ha prodotto una modifica del valore originale. In questo modo è possibile misurare in dettaglio le eventuali distorsioni prodotte dalla procedura di C&C rispetto ai dati veri.

### 3) Altre valutazioni

Gli indicatori disponibili in IDEA sono in realtà utilizzabili tutte le volte che si vogliano misurare le differenze esistenti in due insiemi di dati, relativi alle stesse variabili e unità, legati temporalmente, territorialmente o rispetto al loro contenuto. Ad esempio, potrebbe essere di interesse verificare l’entità delle differenze statistiche a livello di distribuzioni, relazioni semplici o multivariate, stime di aggregati semplici ecc. in due diverse onde di un panel o in due successive ripetizioni di un’indagine, oppure confrontare le distribuzioni di uno stesso fenomeno rilevato in due indagini diverse (ad esempio, un censimento e un’indagine corrente). Allo stesso modo, potrebbe essere di interesse valutare l’entità di queste differenze in diversi domini (ripartizioni o altri tipi di sottopopolazioni).

## 2. GLI INDICATORI DI VALUTAZIONE

In questo paragrafo sono illustrati gli indicatori di prestazione attualmente disponibili nel software IDEA: nel paragrafo 2.1 sono descritti gli indicatori per la valutazione degli effetti a livello di variabile/gruppi di variabili, nel paragrafo 2.2 sono discussi gli indicatori standard previsti in SIDI.



## 2.1. Indicatori di valutazione a livello di variabili

Gli indicatori presentati in questo paragrafo sono ovviamente distinti per tipologia di variabili oggetto di analisi (*categoriche nominali*, *categoriche ordinali* o *numeriche continue*). E' evidente che a seconda della natura delle variabili devono essere definite diversi tipi di misure.

Inoltre, sono stati definiti diversi tipi di indicatori a seconda del tipo di valutazione (a livello micro/aggregato) e del tipo di differenze fra valori di una o più variabili negli insiemi di dati a confronto oggetto di misurazione. In particolare, in IDEA è possibile effettuare valutazioni rispetto ai seguenti criteri:

- 1) grado di *modifica dei valori iniziali* di Y (in termini sia di *quantità* di valori modificati dalla procedura, sia, laddove possibile, di *entità* delle modifiche apportate);
- 2) grado di *modifica della distribuzione* marginale e dei *principali aggregati* statistici della distribuzione di Y;
- 3) grado di *modifica delle relazioni* fra Y e altre variabili;

Nei paragrafi che seguono, per ogni tipologia di variabili sono illustrati gli indicatori attualmente disponibili in IDEA separatamente per ogni criterio di valutazione adottato. E' evidente che, a seconda della coppia di insiemi di dati posta a confronto, e a seconda che la valutazione sia effettuata, per una certa variabile, considerando tutti i dati o il sottoinsieme dei dati modificati, il significato degli indicatori va opportunamente interpretato.

### 2.1.1. Variabili categoriche

#### 2.1.1.1. Modifica dei valori elementari

A seconda del tipo di variabile (categoriche nominali o ordinali), il grado di modifica dei valori originali di una certa variabile Y viene misurato in IDEA mediante diversi tipi di indicatori.

##### b. Per variabili sia nominali che ordinali

###### 1) *Percentuale di Imputazione:*

$$D_1 = \frac{\sum_{i=1}^n w_i I(Y_i, Y_i^*)}{\sum_{i=1}^n w_i} \times 100$$

dove  $Y_i$  e  $Y_i^*$  sono rispettivamente la modalità della variabile Y nel data set di riferimento e in quello finale, e inoltre  $I(Y_i, Y_i^*) = 1$  se  $Y_i \neq Y_i^*$  e 0 altrimenti.

L'indicatore assume il suo minimo ( $\mathbf{D}_1=0$ ) in caso di uguaglianza fra tutte le modalità iniziali e le corrispondenti modalità finali di Y, mentre assume il valore massimo ( $\mathbf{D}_1=1$ , massima dissimilarità) nel caso in cui tutti i valori iniziali sono diversi dai corrispondenti valori finali.

Nel caso di confronto grezzi/puliti, questo indicatore dà un'idea immediata sia della qualità dei dati iniziali, sia dell'impatto sui dati della procedura di C&C in termini di quantità di valori modificati.

Nel caso di confronto veri/puliti, questo indicatore dà un'idea immediata della capacità della procedura di C&C di individuare correttamente gli errori e/o ripristinare correttamente i valori veri.

## 2) Percentuale di Imputazione netta

$$\mathbf{D}_{1i} = \frac{\sum_{i=1}^{n_k} w_i I(Y_i, Y_i^*)}{\sum_{i=1}^{n_k} w_i} \times 100$$

dove  $n_k$  è il numero di valori di Y modificati da *blank* a un altro codice appartenente al dominio di Y, e  $w_i$  è il peso di ogni unità campione.  $\mathbf{D}_{1i}$  assume significato diverso a seconda che il valore *blank* sia o meno ammissibile per Y:

- se il valore *blank* non è nel dominio di Y,  $\mathbf{D}_{1i}$  misura il *tasso di Imputazione netta*, cioè il numero di mancate risposte (risposte dovute ma non fornite) cui è stato assegnato un valore nel dominio di Y.
- se il *blank* è nel dominio di Y,  $\mathbf{D}_{1i}$  misura la porzione di risposte modificate da questa modalità a una delle altre modalità del dominio di Y.

Questo indicatore assume il suo minimo [0] se nessun valore di Y ha subito questo tipo di modifica.

In entrambe le situazioni, l'indicatore fornisce un'indicazione sull'ammontare di casi modificati in seguito a errori di percorso del questionario.

Nel caso  $\mathbf{D}_{1i}$  sia calcolato sul sottoinsieme dei valori modificati dalla procedura di controllo e correzione, l'indicatore misura l'entità della sola componente di imputazione netta rispetto al totale dei valori di Y modificati dalla procedura di C&C.

### 3) Percentuale di Modificazione

$$\mathbf{D_{1m}} = \frac{\sum_{i=1}^{n_k} w_i I(Y_i, Y_i^*)}{\sum_{i=1}^{n_k} w_i} \times 100$$

Questo indicatore assume il suo minimo, 0, se nessun valore di Y è stato modificato da una modalità ad una modalità diversa nel dominio della variabile.

Nel caso si effettui il calcolo su tutte le unità campione,  $\mathbf{D_{1m}}$  misura l'entità della componente di modifica da valore a valore rispetto al totale delle osservazioni  $n$ .

Nel caso si effettui il calcolo sul sottoinsieme dei valori modificati,  $\mathbf{D_{1m}}$  misura l'entità della sola componente di modifica da valore a valore rispetto al complesso dei valori di Y cambiati dalla procedura di C&C.

### 4) Percentuale di Cancellazione

$$\mathbf{D_{1c}} = \frac{\sum_{i=1}^{n_k} w_i I(Y_i, Y_i^*)}{\sum_{i=1}^{n_k} w_i} \times 100$$

Questo indicatore assume il suo minimo, 0, se nessun valore di Y è stato modificato da valore non blank e non mancante a valore blank (che corrisponde in questo caso ad una modalità ammissibile per Y).

Nel caso si effettui il calcolo su tutte le unità campione,  $\mathbf{D_{1c}}$  misura l'entità della componente di assegnazione della modalità “risposta non dovuta” a risposte effettivamente fornite (*cancellazione*) rispetto al totale delle osservazioni  $n$ .

Nel caso si effettui il calcolo sul sottoinsieme dei valori modificati,  $\mathbf{D_{1c}}$  misura l'entità della sola componente di cancellazione rispetto al totale dei valori modificati dalla procedura di C&C.

5) Per ogni variabile Y, IDEA fornisce la corrispondente *matrice di transizione*, cioè una tabella di contingenza ottenuta incrociando le modalità di Y nei due data set posti a confronto. Ogni cella [i,j] della tabella contiene la frequenza di casi passati dalla modalità della riga i (modalità posseduta da Y nel data set iniziale) alla modalità della colonna j (modalità posseduta da Y dopo la procedura di C&C). Le frequenze sulla diagonale principale rappresentano i casi in cui non vi è alcuna

differenza fra le categorie di Y nei due data set posti a confronto. Nel caso di confronto grezzi/puliti, tali frequenze indicano il numero di cambiamenti prodotti dalla procedura di C&C. L'analisi di tali cambiamenti consente di identificare eventuali effetti distorsivi della procedura di C&C.

6) Rappresentazione grafica delle distribuzioni di frequenza di Y nei due data set posti a confronto, per una valutazione esplorativa complessiva delle modifiche subite dai dati elementari.

b. Solo per variabili ordinali

Per questo tipo di variabili, in IDEA sono disponibili i seguenti *indici di dissomiglianza relativi*:

1) *Indice relativo di dissomiglianza complessivo*

$$\mathbf{D}_2 = \frac{I}{m \times \sum_{i=1}^n w_i} \sum_{i=1}^n w_i d(Y_i, Y_i^*) \times 100$$

Dove  $w_i$  sono gli eventuali pesi campionari,  $n$  è il numero di osservazioni e la metrica a blocchi  $d(.,.)$  è definita come:

$$d(Y_i, Y_i^*) = \begin{cases} 0 & \text{se } Y_i = Y_i^* \\ |Y_i - Y_i^*| & \text{se } Y_i \neq Y_i^* \text{ e } Y_i, Y_i^* \neq \text{blank} \\ m & \text{se } Y_i \neq Y_i^* \text{ e } Y_i = \text{blank o } Y_i^* = \text{blank} \end{cases}$$

con  $m = (\max_Y - \min_Y) + 1$  se la modalità *blank* è nel dominio di Y, mentre  $m = (\max_Y - \min_Y)$  se la modalità *blank* non è nel dominio di Y, e  $\max_Y, \min_Y$  sono rispettivamente la modalità inferiore e superiore di Y. Anche questo indicatore varia tra 0 (uguaglianza fra le modalità prima/dopo di Y) e 1 (massima dissimilarità) (Chambers, 2001).  $\mathbf{D}_2$  quantifica in modo sintetico non solo quanti valori sono diversi nei due data set, ma anche l'entità delle differenze esistenti fra essi (nel caso di confronto grezzi/puliti, l'entità delle modifiche determinate dal processo di C&C).

2) *Indice relativo di dissomiglianza netto*

$$\mathbf{D}_{2n} = \frac{I}{m \times \sum_{i=1}^n w_i} \sum_{i=1}^n w_i d(Y_i, Y_i^*) \times 100$$

definito esattamente come **D<sub>2</sub>** nel caso in cui la modalità *blank* non è ammissibile per Y, ma calcolato escludendo dal calcolo tutti i record in cui la variabile Y assume appunto valore *blank* (evidentemente nel file grezzo). I *blank* in questo caso rappresentano mancate risposte (quindi valori effettivamente da imputare), e se le mancate risposte sono molte l'indice **D<sub>2</sub>** potrebbe dare una misura distorta dell'impatto della procedura di C&C sui dati (dal momento che in questi casi il valore della distanza assume il massimo, cioè appunto  $w(Y_i, Y_i^*) = \max_Y - \min_Y$ ). Di conseguenza, per dare allo statistico la possibilità di valutare l'impatto della procedura *al netto delle mancate risposte*, per variabili in cui la modalità *blank* non è ammissibile vengono calcolati **entrambi** gli indici **D<sub>2</sub>** e **D<sub>2n</sub>**.

### 2.1.1.2. Modifica delle distribuzioni marginali

#### Per variabili sia nominali che e ordinali

In questo caso la valutazione del grado di modifica della distribuzione di Y viene effettuata in IDEA utilizzando le seguenti misure.

1) *Indici di dissomiglianza* (Leti, 1983):

$$\begin{aligned}
 - \quad \mathbf{I}_{m1} &= \frac{1}{2} \sum_{k=1}^K \left| f_{Y_k} - f_{Y_k^*} \right| \\
 - \quad \mathbf{I}_{m2} &= \left\{ \frac{1}{2} \sum_{k=1}^K \left| f_{Y_k} - f_{Y_k^*} \right|^2 \right\}^{\frac{1}{2}}
 \end{aligned}$$

dove  $f_{Y_k}$ ,  $f_{Y_k^*}$  sono rispettivamente le frequenze marginali della variabile Y nel file iniziale e in quello ottenuto dopo la fase di C&C<sup>2</sup>. Entrambi gli indicatori assumono valore 0 nel caso di uguaglianza fra le distribuzioni iniziale e finale, e valore massimo 1 nel caso di massima dissimilarità fra esse (caso in cui in ciascuna distribuzione tutte le unità presentano una stessa modalità che però è diversa per le due distribuzioni).

2) *Matrici di transizione prima/dopo*

Per ogni variabile Y selezionata, viene costruita una matrice doppia di contingenza fra le modalità di Y nei due data set posti a confronto: le frequenze interne della matrice poste al di fuori della diagonale principale descrivono gli “spostamenti” delle unità

---

<sup>2</sup> Nel computo delle differenze  $\left| f_{Y_k} - f_{Y_k^*} \right|$ , le modalità su cui è effettuato il calcolo sono quelle *teoriche*, cioè tutte le modalità ammissibili per Y fornite dall'utente. Questo al fine di evitare di non considerare nell'indice modalità presenti solo nell'uno o nell'altro dei file messi a confronto.

rispetto alle modalità di Y (inclusi i valori mancanti) nei due data sets. Nel caso di confronto fra dati grezzi e dati puliti, queste matrici sono molto utilizzate sia al fine di valutare l'entità delle trasformazioni determinate dalla procedura di C&C sulle distribuzioni marginali di ogni variabile Y, ma anche al fine di individuare eventuali difetti nella procedura di C&C che determinano spostamenti sistematici.

- 3) Rappresentazione grafica delle distribuzioni di frequenza di Y nei due data set posti a confronto, per una valutazione esplorativa complessiva delle modifiche subite dalle distribuzioni marginali.

### 2.1.1.3. Modifica delle distribuzioni congiunte

Per variabili sia nominali che ordinali

Nel caso bivariato, siano Y ed X due variabili categoriche e siano  $f_{yx}$ ,  $\tilde{f}_{yx}$  le frequenze della tabella a doppia entrata costruita sulla base delle rispettive modalità nei due data set di confronto. Il grado di modifica delle relazioni doppie fra Y ed X viene misurata mediante i seguenti indici descrittivi:

$$\begin{aligned} \text{i.} \quad & \text{semplice:} \quad \mathbf{I}_{j1} = \frac{1}{2} \sum_y \sum_x |f_{yx} - \tilde{f}_{yx}| \\ \text{ii.} \quad & \text{quadratico:} \quad \mathbf{I}_{j2} = \left\{ \frac{1}{2} \sum_y \sum_x |f_{yx} - \tilde{f}_{yx}|^2 \right\}^{\frac{1}{2}} \end{aligned}$$

Entrambi gli indicatori variano in  $[0,1]$ , e possono essere estesi a qualunque  $k$ -upla di variabili ( $2 \leq k \leq 5$ ) per analisi su distribuzioni multiple.

Si noti che il valore di questi indici può essere relativamente basso per una combinazione di variabili che assume poche modalità rispetto ad una combinazione di variabili che assume molte modalità. Di conseguenza, questi indici sono molto utili quando si vogliano ad esempio confrontare i risultati prodotti da diversi metodi di C&C per lo stesso insieme di variabili (cioè per lo stesso numero di variabili e modalità delle variabili stesse).

Nel caso di confronto grezzi/puliti,  $\mathbf{I}_{j1}$  e  $\mathbf{I}_{j2}$  forniscono un'informazione sintetica sull'impatto delle modifiche apportate ai dati sulle associazioni multiple fra le variabili di interesse.

Nel caso in cui gli indici siano calcolati sul sottoinsieme dei dati modificati, per ogni  $k$ -upla di variabili vengono utilizzate le unità in cui almeno una delle variabili stesse assume valore diverso nei due data set posti a confronto.

#### 2.1.1.4. Modifica delle relazioni (associazioni) doppie

Per variabili sia nominali che ordinali

Il grado di modifica delle relazioni (associazioni) fra le variabili categoriche  $Y$  ed  $X$  sottoposte a controllo e correzione viene misurata mediante il *coefficiente di contingenza di Cramer* (eventualmente ponderato) (Leti, 1983):

$$\mathbf{Cramer} = \left\{ \frac{\chi^2}{n \min(r-1, c-1)} \right\}^{1/2}$$

dove  $\chi^2$  è l'indice quadratico di contingenza basato sugli scarti tra le frequenze della tabella a doppia entrata costruita utilizzando le modalità di  $Y$  e  $X$ , e le corrispondenti frequenze teoriche di indipendenza,  $n$  è la numerosità della tabella,  $r$  e  $c$  sono la numerosità delle categorie rispettivamente di  $Y$  e  $X$ . *Cramer* assume i valori da 0 (indipendenza) ad 1 (associazione completa).

In pratica, per qualunque coppia di variabili nel sottoinsieme di variabili selezionato, in IDEA vengono pertanto calcolate le *matrici dei coefficienti di contingenza del Cramer* fra esse nei due data set posti a confronto.

Nel caso di confronto grezzi/puliti, tali matrici forniscono informazioni sull'impatto delle modifiche apportate ai dati sulle associazioni fra le variabili soggette a C&C.

Nel caso in cui l'indice sia calcolato sul sottoinsieme dei dati modificati, per ogni coppia di variabili vengono utilizzate le unità in cui almeno una delle variabili stesse è stata modificata dalla procedura di C&C.

## 2.1.2. Variabili numeriche continue

### 2.1.2.1. Modifica dei valori elementari

Diverse misure sono previste per misurare il grado di modifica dei valori elementari di una variabile numerica continua Y.

1) Gli indicatori definiti per le variabili categoriche:  $\mathbf{D_I}$  (*Percentuale di Imputazione*),  $\mathbf{D_{Ii}}$  (*Percentuale di Imputazione netta*),  $\mathbf{D_{Im}}$  (*Percentuale di Modificazione*),  $\mathbf{D_{Ic}}$  (*Percentuale di Cancellazione*) (vedi paragrafo 2.1.1.1). Nella definizione degli indicatori il valore *blank* corrisponde ai valori della variabile uguali a 0.

2) Indici di distanza:

i. *Semplice*:  $\mathbf{D_{L1}} = D_L^1(Y_i, Y_i^*) = \frac{\sum_{i=1}^n w_i |Y_i - Y_i^*|}{\sum_{i=1}^n w_i}$

ii. *Quadratica*:  $\mathbf{D_{L2}} = D_L^2(Y_i, Y_i^*) = \left\{ \frac{\sum_{i=1}^n w_i (Y_i - Y_i^*)^2}{\sum_{i=1}^n w_i} \right\}^{\frac{1}{2}}$

iii. *In  $L^\infty$* :  $\mathbf{D_{Linf}} = D_L^\infty(Y_i, Y_i^*) = \frac{\max_i \{w_i |Y_i - Y_i^*|\}}{\sum_{i=1}^n w_i}$

dove  $w_i$  sono gli eventuali pesi campionari. Tutti questi indicatori assumono ovviamente valore minimo (0) solo nel caso in cui tutti i valori sono uguali fra loro nei due insiemi di dati posti a confronto. E' evidente che tanto maggiore è la distanza fra i valori di Y nei due data set (in termini di numero di valori diversi e/o entità delle differenze), tanto maggiore è il valore assunto dagli indici.

In caso di confronto fra dati grezzi e puliti, nell'interpretazione dei risultati si tenga conto del fatto che questi indicatori risentono della presenza nei dati iniziali di valori anomali o (nel caso degli indici ponderati) di valori influenti, cui risultano associati addendi molto grandi.



- 3) Un secondo approccio alla valutazione è basato su un'analisi di tipo regressivo. Dati i valori  $Y_i$  e  $Y_i^*$  della variabile  $Y$  nei due data set a confronto, viene costruito il modello senza intercetta:

$$Y = \beta \times Y^*$$

Per la valutazione dell'accostamento fra valori  $Y$  e  $Y^*$  sono usate le seguenti misure:

1.  $N$  (numero di osservazioni su cui è stimato il modello)
2. Coefficiente di regressione  $\beta$
3. Indici  $R^2$  e  $R^2$  corretto
4. RMSE (*Root Mean Square Error*)

E' evidente che valori di  $R^2$  vicini ad 1 indicano un effetto contenuto della procedura di C&C in termini di entità delle variazioni prodotte sui dati grezzi.

Nel caso di confronto veri/puliti, valori di  $R^2$  vicini ad 1 indicano una buona capacità della procedura di C&C di "recuperare" i valori veri dei dati errati/mancanti.

Per una migliore analisi dei dati IDEA produce lo *scatter plot* della regressione effettuata.

- 4) Rappresentazioni grafiche interattive dei dati di  $Y$  nei due data set posti a confronto mediante strumenti SAS Insight, in particolare box plot e distribuzione dei dati.

### 2.1.2.2. Modifica delle distribuzioni marginali e degli aggregati

Nel caso di variabili numeriche continue, il grado di modifica delle distribuzioni semplici iniziali di una certa variabile  $Y$  viene valutata attraverso i seguenti tipi di indicatori.

- 1) *Distanza di Kolmogorov-Smirnov* ( $N\_KS$ )

$$KS(F_{Y_n^R}, F_{Y_n^F}) = \max_t |F_{Y_n}(t) - F_{Y_n^*}(t)|$$

dove

$$F_{Y_n}(t) = \frac{\sum_{i=1}^n w_i I(Y_i \leq t)}{\sum_{i=1}^n w_i}, F_{Y_n^*}(t) = \frac{\sum_{i=1}^n w_i I(Y_i^* \leq t)}{\sum_{i=1}^n w_i}$$

e  $w_i$  sono gli eventuali pesi campionari. Tale indicatore assume evidentemente valore zero solo quando le due distribuzioni sono identiche.

- 2) Tabelle contenenti i valori delle principali statistiche univariate e dei principali aggregati della distribuzione di Y nei file iniziale e finale (*Totale, Media, Mediana, valori minimo e massimo, 1° quartile, 3° quartile, deviazione standard, numero di osservazioni e numero di valori missing*).
- 3) Analisi di tipo regressivo analoga a quella effettuata per la verifica del grado di *modifica dei valori elementari* (vedi paragrafo 2.1.2.1, punto 3).
- 4) Analisi grafica interattiva tramite box plot e distribuzioni di SAS Insight analoga a quella prevista per la verifica del grado di *modifica dei valori elementari* (vedi paragrafo 2.1.2.1, punto 4).

### **2.1.2.3. Modifica delle relazioni doppie**

Per qualunque sottoinsieme selezionato di variabili, IDEA fornisce informazioni sull'impatto delle modifiche apportate ai dati sulle relazioni bivariate fra le variabili soggette a una certa procedura di C&C attraverso i seguenti indici:

- *indice di correlazione del Pearson* calcolato fra coppie di variabili in ciascuno dei due data set posti a confronto;
- *covarianze* fra tali variabili in ciascuno dei due data set posti a confronto.

Nel caso in cui l'indice sia calcolato sul sottoinsieme dei dati modificati, per ogni coppia di variabili vengono utilizzate le unità in cui almeno una delle variabili stesse è stata modificata dalla procedura di C&C.

## **2.2. Indicatori di valutazione a livello complessivo: indicatori SIDI**

IDEA fornisce un insieme di valori utili per la valutazione globale degli effetti della procedura di imputazione sul complesso dei dati. In questo caso, IDEA consente non solo di visualizzare gli indicatori sulla fase di revisione richiesti da SIDI, ma anche di produrre in output un file che costituisce l'input per SIDI: a partire dai valori contenuti in questo output, SIDI procede al calcolo e alla memorizzazione degli indicatori di qualità della fase di Revisione. Questi ultimi sono poi visualizzabili nel sistema di interrogazione SIDI-TOP. In SIDI-TOP sono disponibili diverse funzionalità per l'analisi temporale e territoriale degli indicatori attraverso rappresentazioni grafiche e tabellari di tali indicatori che ne consentono, ad esempio, il monitoraggio nel tempo e il confronto con altre indagini.

I valori che IDEA calcola confrontando dati grezzi e puliti possono essere ricondotti a 3 tipi:

- Indicatori sull'ammontare di dati sottoposti alla procedura di imputazione.
- Indicatori per la valutazione complessiva degli effetti della procedura di imputazione.
- Indicatori di sintesi sulla distribuzione del tasso di imputazione per variabile e per record.

Tali indicatori possono essere calcolati sia non ponderati sia ponderati utilizzando una variabile di ponderazione che, in genere, coincide con i coefficienti di riporto all'universo.

Seguono le definizioni dei singoli indicatori raggruppati nelle 3 suddette tipologie.

### **1) Indicatori sull'ammontare di dati sottoposti alla procedura di imputazione**

- *Totale Record*: numero complessivo di unità osservate. Tale indicatore viene fornito per ripartizione geografica oltre che per il totale Italia
- *Totale Variabili*: numero delle variabili contenute nel data-set originario (dati grezzi)
- *Totale Variabili Soggette a Imputazione*: numero delle variabili correggibili durante la fase di revisione automatica. Si escludono alcune tipologie di variabili, come ad esempio i codici identificativi delle unità, che non sono soggette ad imputazione

### **2) Indicatori per la valutazione complessiva degli effetti della procedura di imputazione**

#### ***Definizioni dei valori assoluti necessari al calcolo degli indicatori:***

- *Valori Passibili di Imputazione*: totale record moltiplicato per il totale delle variabili soggette a imputazione
- *Valori Imputati*: conteggio dei valori sui quali hanno agito le regole d'imputazione, e che quindi sono stati modificati dalla procedura d'imputazione
- *Valori Modificati da Codice a Codice Diverso*: conteggio dei valori trasformati dalle procedure di imputazione da un codice (valore non blank) a un codice diverso
- *Valori Modificati da Blank a Codice*: conteggio dei valori trasformati dalle procedure di imputazione da blank (valore mancante) a un codice
- *Valori Modificati da Codice a Blank*: conteggio dei valori trasformati dalle procedure di imputazione da codice a blank (valori cancellati)
- *Valori Non Imputati*: conteggio dei valori non trasformati dalle procedure di imputazione
- *Valori Blank Non Imputati*: conteggio dei valori blank che rimangono blank dopo l'implementazione delle procedure di imputazione

- *Valori Non Blank Non Imputati*: conteggio dei valori il cui codice rimane immutato dopo l'implementazione delle procedure di imputazione

***Indicatori:***

- *Tasso di Imputazione* =  $[\text{Valori Imputati} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Modificazione* =  $[\text{Valori Modificati da Codice a Codice Diverso} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Imputazione Netta* =  $[\text{Valori Modificati da Blank a Codice} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Cancellazione* =  $[\text{Valori Modificati da Codice a Blank} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Non Imputazione* =  $[\text{Valori Non Imputati} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Valori Blank Immutati* =  $[\text{Valori Blank Non Imputati} / \text{Valori Passibili di Imputazione}] \times 100$
- *Tasso di Valori Non Blank Immutati* =  $[\text{Valori Non Blank Non Imputati} / \text{Valori Passibili di Imputazione}] \times 100$

*Componenti Percentuali del Tasso di Imputazione*

- *Percentuale di Modificazione* =  $[\text{Valori Modificati da Codice a Codice Diverso} / \text{Valori Imputati}] \times 100$
- *Percentuale di Imputazione Netta* =  $[\text{Valori Modificati da Blank a Codice} / \text{Valori Imputati}] \times 100$
- *Percentuale di Cancellazione* =  $[\text{Valori Modificati da Codice a Blank} / \text{Valori Imputati}] \times 100$

*Componenti Percentuali del Tasso di Non Imputazione*

- *Percentuale di Valori Blank Immutati* =  $[\text{Valori Blank Non Imputati} / \text{Valori Non Imputati}] \times 100$
- *Percentuale di Valori Non Blank Immutati* =  $[\text{Valori Non Blank Non Imputati} / \text{Valori Non Imputati}] \times 100$

I precedenti indicatori vengono calcolati per ripartizione geografica e per il totale Italia.

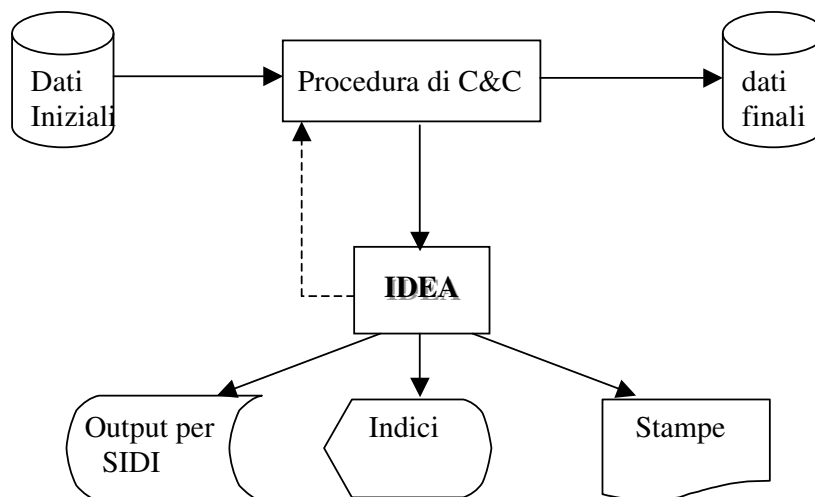
**3) Indicatori di sintesi sulla distribuzione del tasso di imputazione per variabile e per record**

- *Primo Quartile del Tasso di Imputazione per Variabile*: valore del tasso di imputazione per variabile che lascia a sinistra il 25% delle variabili ordinate in senso crescente rispetto al tasso stesso. Per tasso di imputazione per variabile si intende il rapporto tra il numero di valori imputati relativi ad una singola variabile ed il relativo numero massimo di imputazioni possibili (Totale record).
- *Terzo Quartile del Tasso di Imputazione per Variabile*: valore del tasso di imputazione per variabile che lascia a sinistra il 75% delle variabili ordinate in senso crescente rispetto al tasso stesso. Per tasso di imputazione per variabile si intende il rapporto tra il numero di valori imputati relativi ad una singola variabile ed il relativo numero massimo di imputazioni possibili (Totale record).
- *Numero di Variabili con Tasso di Imputazione > al 5%*: conteggio delle variabili che presentano un tasso di imputazione per variabile superiore al 5%.
- *Numero di Variabili con Tasso di Imputazione > al 2%*: conteggio delle variabili che presentano un tasso di imputazione per variabile superiore al 2%.
- *Primo Quartile del Tasso di Imputazione per Record*: valore del tasso di imputazione per record che lascia a sinistra il 25% dei record ordinati in senso crescente rispetto al tasso stesso. Per tasso di imputazione per record si intende il rapporto tra il numero di valori imputati relativi ad un singolo record ed il relativo numero massimo di imputazioni possibili (Totale variabili soggette a imputazione).
- *Terzo Quartile del Tasso di Imputazione per Record*: valore del tasso di imputazione per record che lascia a sinistra il 75% dei record ordinati in senso crescente rispetto al tasso stesso. Per tasso di imputazione per record si intende il rapporto tra il numero di valori imputati relativi ad un singolo record ed il relativo numero massimo di imputazioni possibili (Totale variabili soggette a imputazione).
- *Numero di Record con Tasso di Imputazione > al 5%*: conteggio dei record che presentano un tasso di imputazione per record superiore al 5%.
- *Numero di Record con Tasso di Imputazione > al 2%*: conteggio dei record che presentano un tasso di imputazione per record superiore al 2%.

Per gli indicatori ponderati viene anche fornito il *numero totale di osservazioni non ponderate*.

### 3. IL SISTEMA

Il software IDEA costituisce un ambiente integrato per la valutazione degli effetti di procedure di C&C mediante opportuni indicatori di valutazione. L'utilizzo di IDEA nell'ambito di un processo di valutazione di procedure di C&C è schematizzato nella Figura 1.



**Figura 1 – Utilizzo del software IDEA per la valutazione di procedure di C&C**

Nei paragrafi seguenti sono illustrate le caratteristiche generali del software e le funzionalità dei singoli moduli che lo compongono.

#### 3.1. Aspetti generali

Il software IDEA è stato sviluppato completamente in SAS. Al fine del calcolo degli indici, compresi quelli relativi a SIDI, è possibile utilizzare solo data set SAS. Per file con formati diversi (ad esempio, .xls, .dbf, .txt ecc.) si possono utilizzare le funzioni di import e di export native del sistema SAS.

Il sistema è in grado di calcolare gli indici di prestazione sulle variabili selezionate nei due data set prescelti (file *originale* e corrispondente file *corretto*) se e solo se:

- i. i due data set hanno la stessa struttura,
- ii. i nomi delle variabili sono uguali nei due archivi;
- iii. i due data set hanno una sola variabile di accoppiamento (campo chiave).

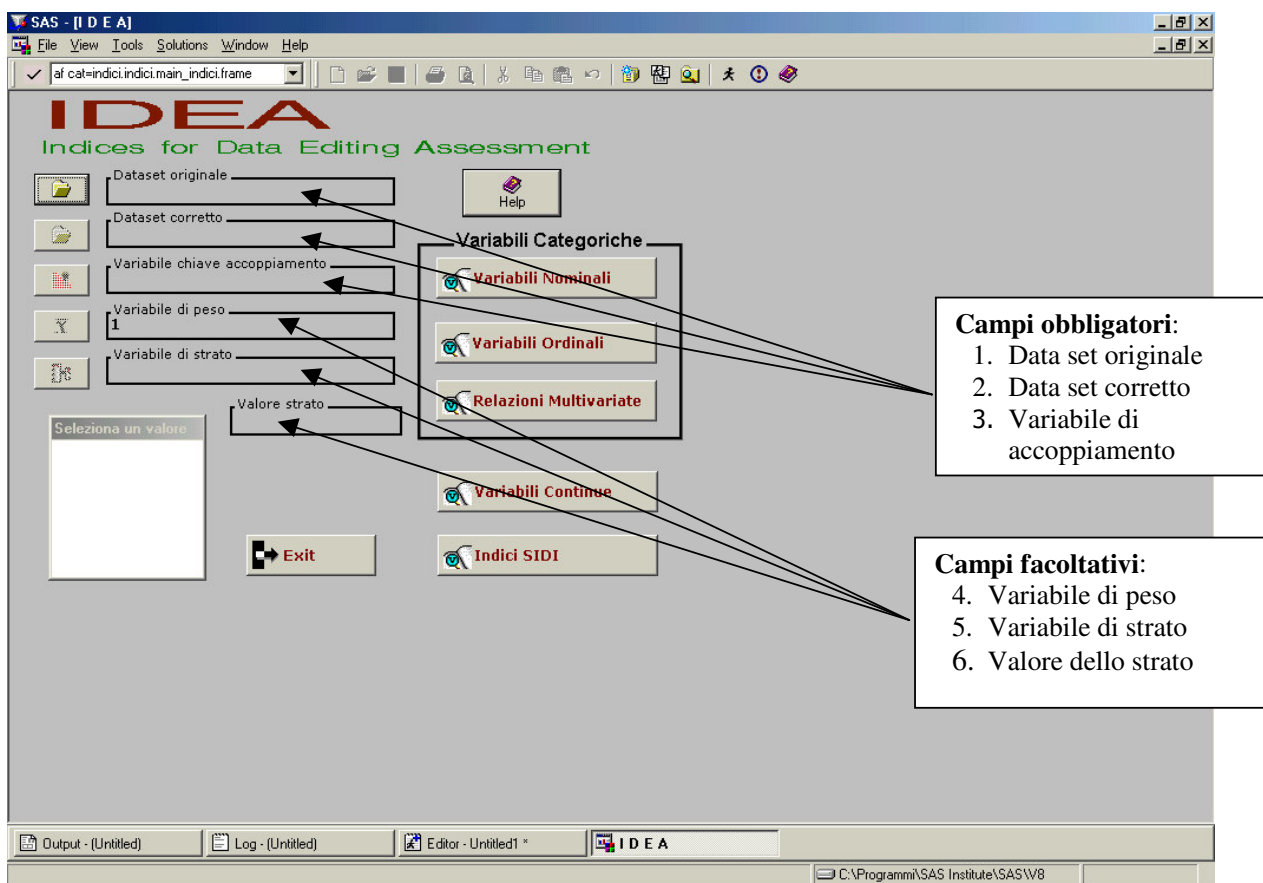
Il sistema controlla automaticamente queste caratteristiche e non permette di procedere alle fasi successive della procedura di valutazione in caso di incompatibilità.

Per quanto riguarda la parte relativa a SIDI si ricorda che il data set originale deve contenere una sola variabile corrispondente alla *ripartizione geografica*, con modalità corrispondenti ai 5 livelli numerici *nord-ovest=1, nord-est=2, centro=3, sud=4, isole=5*.

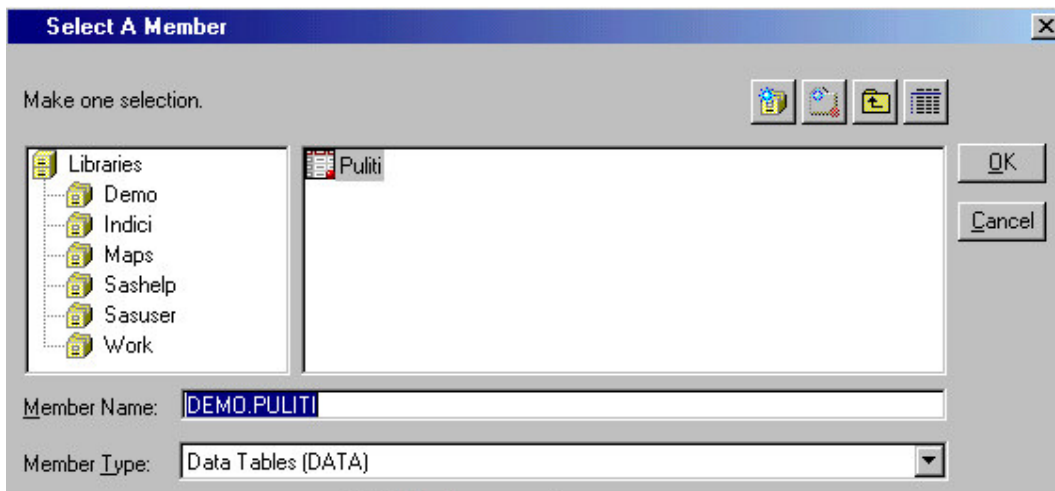
### 3.2. Dati di input

Nella figura 2 sono evidenziati i campi obbligatori e quelli facoltativi presenti nella prima maschera di IDEA per la selezione dei dati da sottoporre a elaborazione. Tali campi vanno riempiti seguendo la sequenza impostata. Se si sceglie di calcolare gli indici per strato si ricorda che ogni strato deve essere elaborato singolarmente.

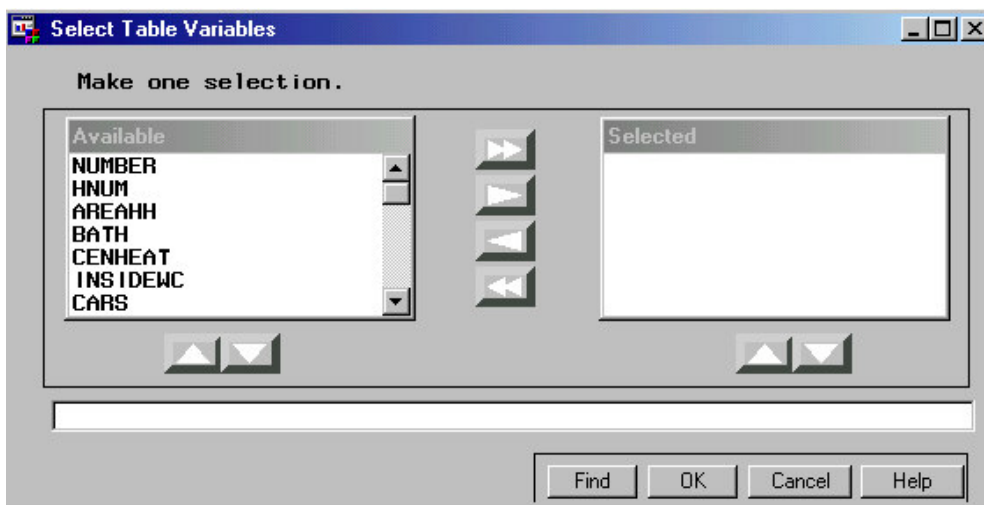
Le maschere riportate nelle Figure 3 e 4 (*maschera per la selezione della libreria e dei data set e maschera per la selezione della chiave identificativa dei record*) vengono dichiarate una sola volta all'inizio della procedura e sono comuni a tutte le fasi successive della procedura stessa. In particolare, la maschera per la selezione della libreria e dei data set consente anche di definire nuove librerie.



**Figura 2 - Maschera per la selezione dei dati in input**



**Figura 3 - Maschera per la selezione della libreria e dei data set.**



**Figura 4 - Maschera per la selezione della chiave identificativa dei record**

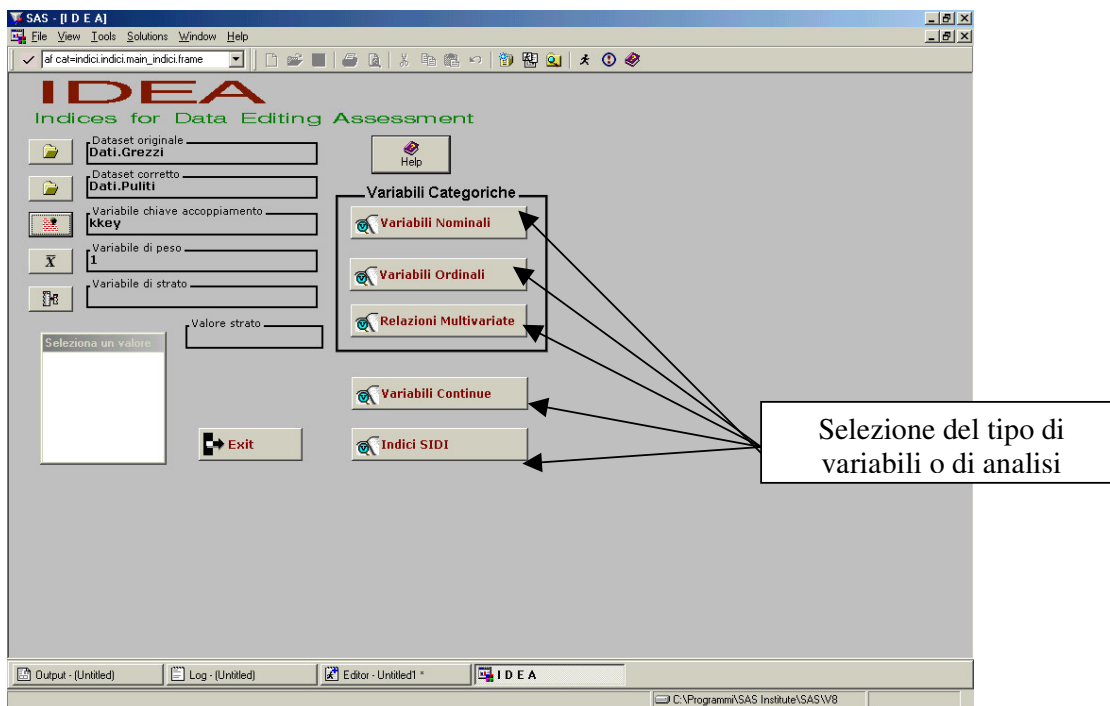
### **3.3. Indicatori di valutazione a livello di variabili**

Una volta riempiti i campi relativi alla scelta dei data set e agli altri campi obbligatori e/o facoltativi, si può passare al calcolo degli indici a livello di variabile per tipologia di variabili (nominali, ordinali e continue) o di analisi (analisi delle relazioni multivariate per variabili categoriche).

A questo scopo, è necessario selezionare il tipo di variabili/analisi per le quali si intende procedere alla valutazione, come indicato nella figura seguente.

A ciascuna selezione corrisponde la disponibilità di diversi indicatori di valutazione.

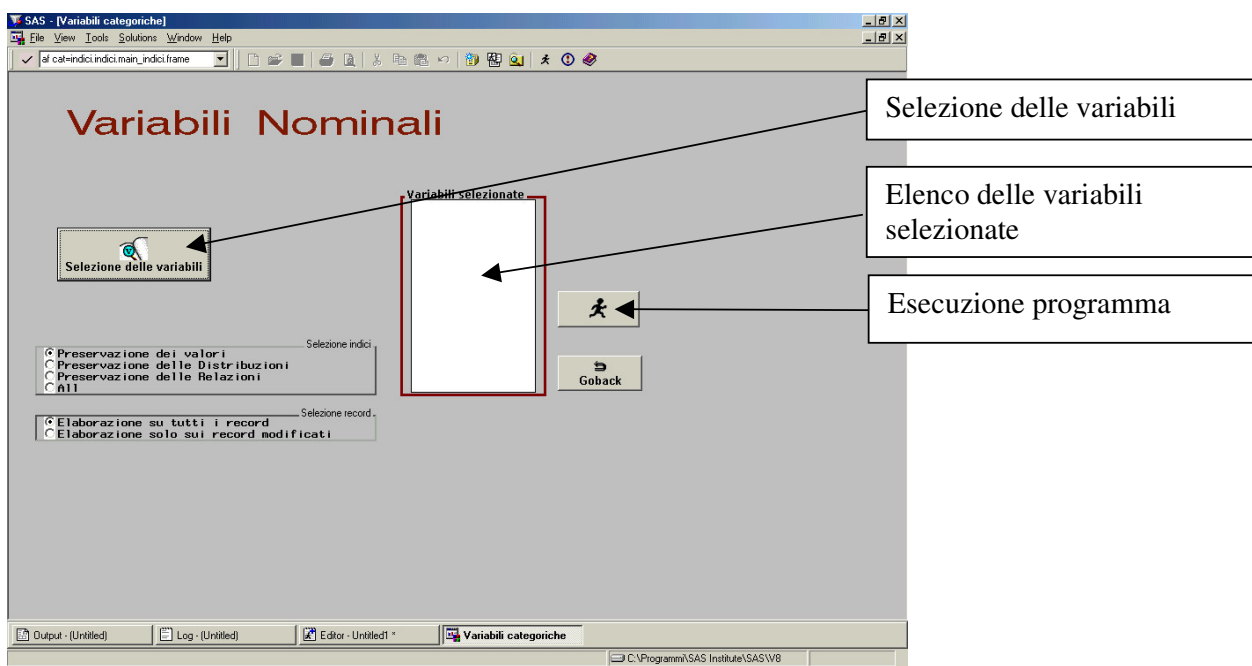




**Figura 5 – Maschera per la selezione del tipo di variabili/analisi**

### 3.3.1. Variabili categoriche nominali o ordinali

Nel caso delle variabili categoriche (di tipo sia nominale che ordinale o per analisi multivariate), la selezione del sottoinsieme di variabili oggetto di valutazione e del tipo di indicatori da calcolare viene effettuata attraverso la maschera seguente.



**Figura 6 - Maschera di scelta delle variabili e dei relativi indici**

Una volta selezionate le variabili di interesse, viene visualizzata una maschera (Figura 7) nella quale sono evidenziati i domini osservati per le variabili selezionate. Qualora tali domini divergano dai domini teorici, l'utente può modificare i valori con la seguente sintassi:

- il valore missing (punto) va inserito per primo seguito da una virgola;
- i range dei domini vanno separati dai due punti (minimo:massimo). Sono ammissibili anche modalità con valori negativi;
- più range non consecutivi vanno ordinati in senso crescente e separati da una virgola.

Una volta selezionate le variabili e definiti i corrispondenti range teorici è possibile effettuare il calcolo degli indicatori, criterio per criterio, su tutti i record o solo sul sottoinsieme dei modificati (Figura 8). L'opzione **All** calcola tutti gli indicatori disponibili e li mostra sequenzialmente a video.

**Dominio osservato delle variabili selezionate**  
**Modificare se necessario**

	Variabili selezionate	Distribuzione Reale/Teorica
1	BATH	1:3
2	CENHEAT	1:3
3	INSIDEWC	1:3
4	AGE	0:2 , 5:9 , 11:13 , 16:69 , 71:72 , 74:78 , 80:87 , 89:91 , 95:95

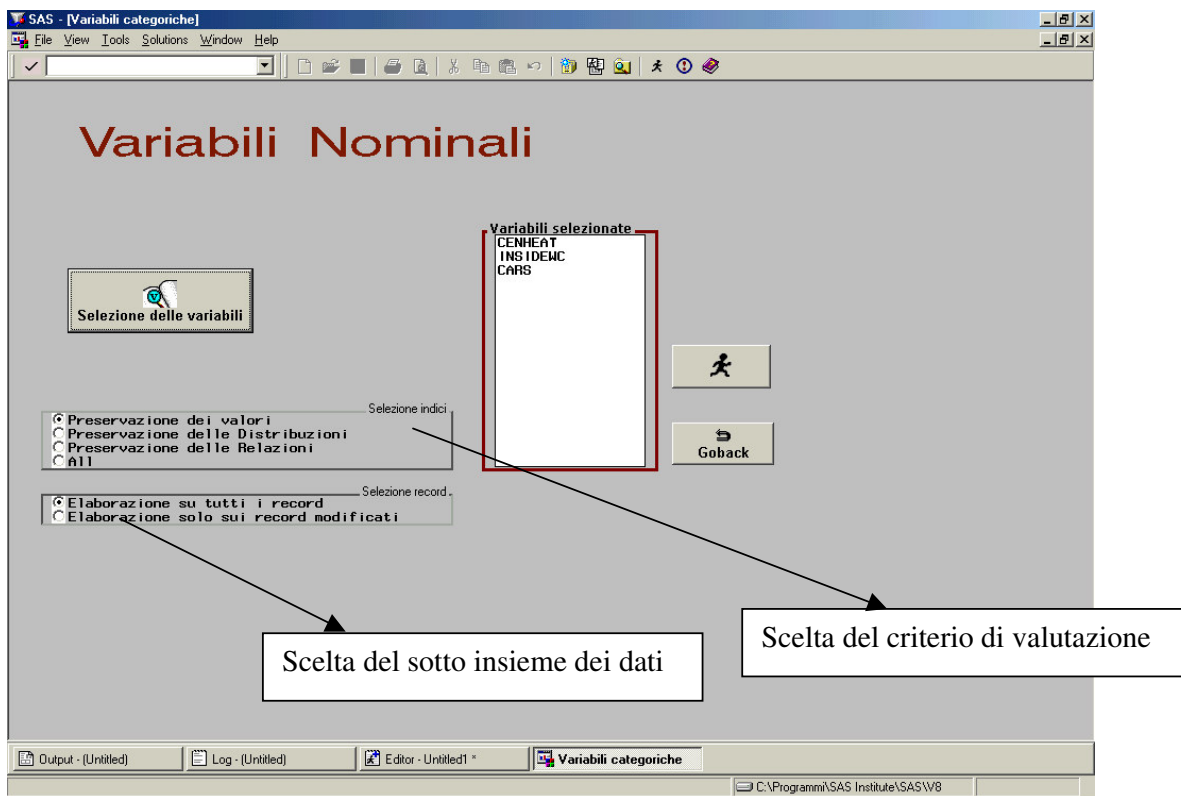
**Esempio:** -9;-9,-6;-2,0;14,16:16

**Goback**

**Distribuzione Reale / Teorica**

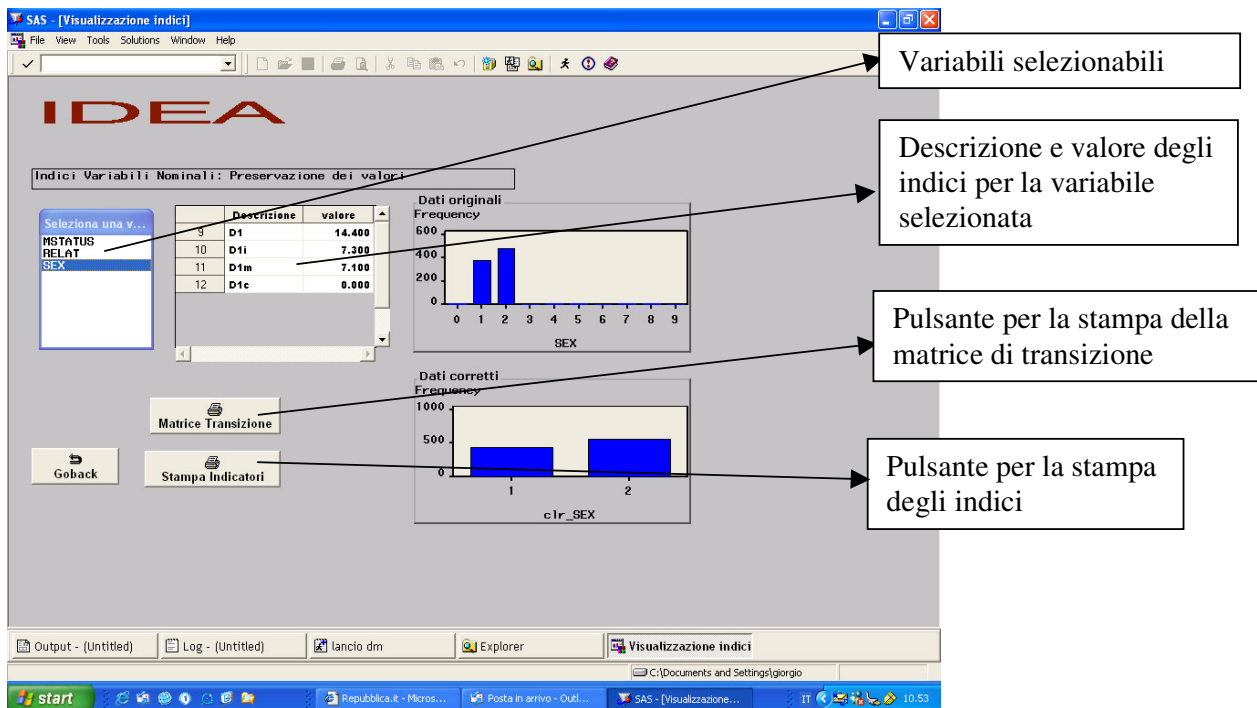
Output - (Untitled) Log - (Untitled) Editor - Untitled1 \* Modifica range C:\Programmi\SAS Institute\SAS\W8

**Figura 7 – Maschera per la definizione dei range delle variabili**



**Figura 8 – Maschera per la scelta del criterio di valutazione e del tipo di elaborazione**

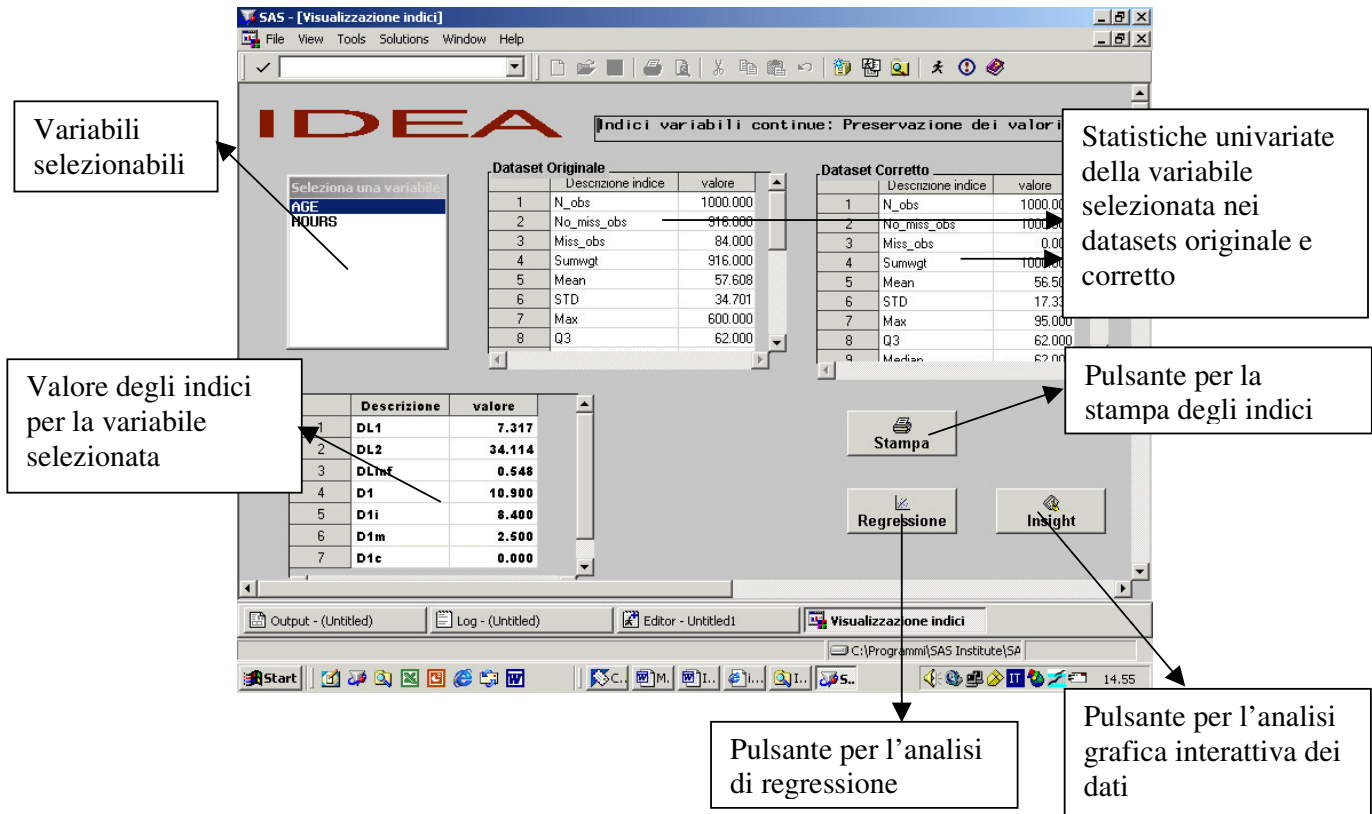
Nella maschera che segue è riportato un esempio di output della procedura di calcolo degli indicatori per variabili di tipo nominale.



**Figura 9 – Esempio di output della procedura**

### 3.3.2. Variabili numeriche continue

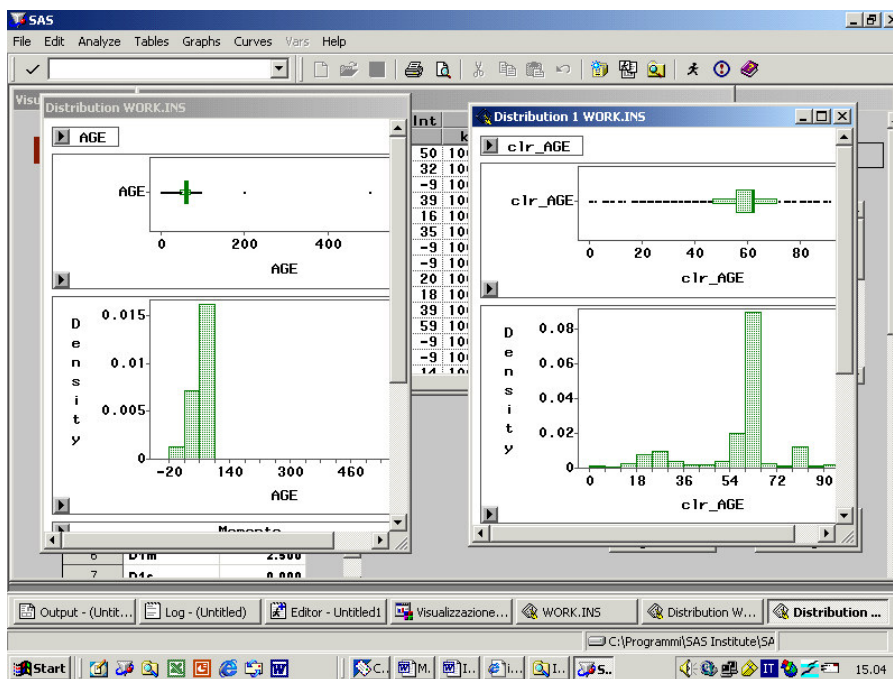
Per quanto riguarda la valutazione per variabili di tipo continuo si può far riferimento alla prima parte del punto 3.3.1, in quanto la procedura è del tutto simile. Diverse invece sono in questo caso le maschere relative alla visualizzazione degli indici calcolati, che sono illustrate di seguito.



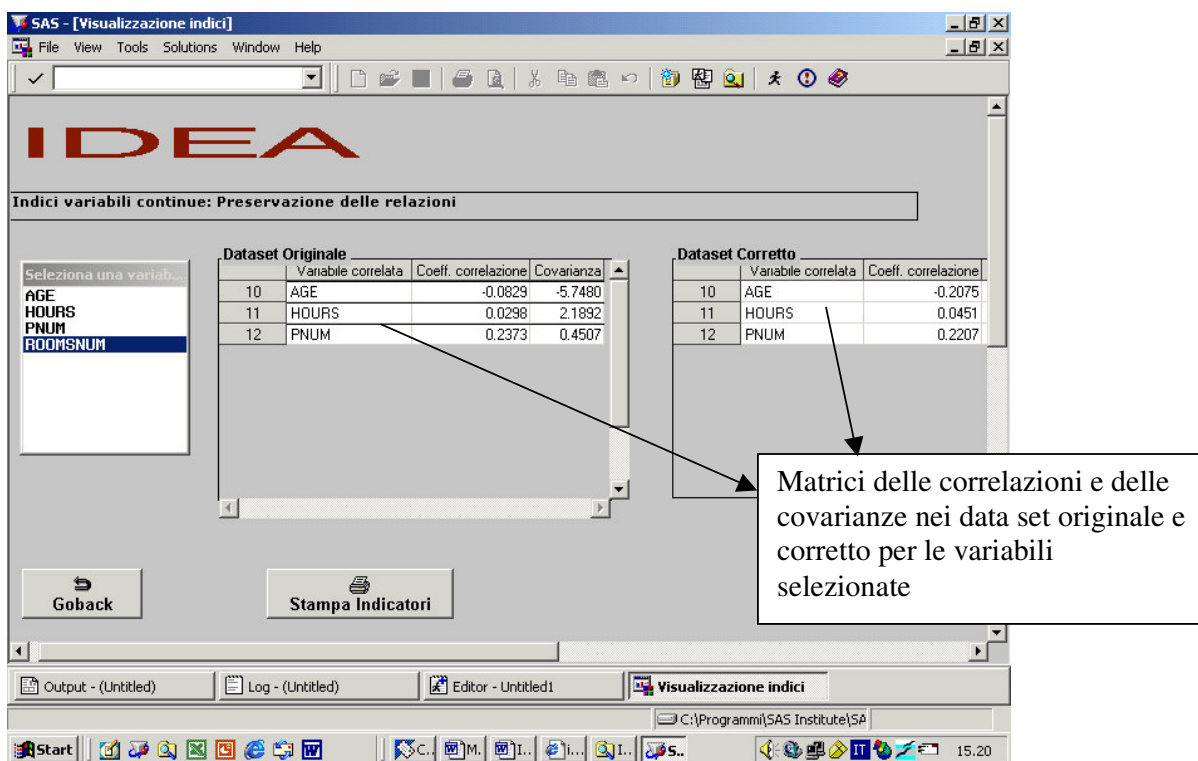
**Figura 10 - Esempio di output per gli indici relativi alla modifica dei valori, delle distribuzioni e degli aggregati in variabili numeriche continue.**

Nella figura 11 sono mostrati i grafici interattivi automaticamente prodotti da IDEA una volta selezionata l'opzione Insight. Si tratta di box plot e istogrammi per la rappresentazione della distribuzione marginale della variabile selezionata. E' necessario che Insight venga chiuso prima di proseguire con altre analisi in IDEA.

Nella figura 12 è invece mostrato il risultato dell'analisi dei dati per la verifica del grado di preservazione delle relazioni fra variabili.



**Figura 11 – Esempio di rappresentazioni grafiche interattive con SAS Insight automaticamente prodotte da IDEA**



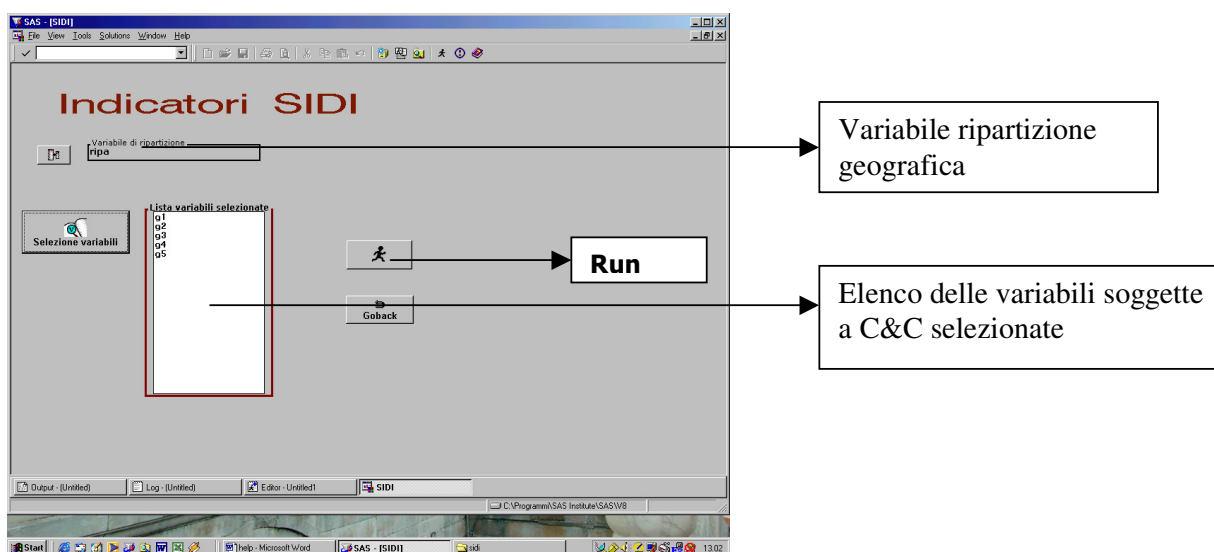
**Figura 12 – Esempio di output per gli indici relativi alla modifica delle relazioni in variabili numeriche continue.**

### 3.4. Indicatori SIDI

Per quanto riguarda il calcolo degli indicatori SIDI, si ricorda che la definizione dei file e della chiave di accoppiamento è del tutto analoga alla prima parte del punto 3.3.1 con la possibilità di selezionare la variabile peso per il calcolo di indicatori ponderati.

Una volta definiti i file da utilizzare nelle elaborazioni e la variabile di accoppiamento, selezionando l'opzione "**Indici SIDI**" nella maschera mostrata in Figura 5, viene visualizzata una maschera di lavoro (Figura 13), in cui è obbligatorio selezionare:

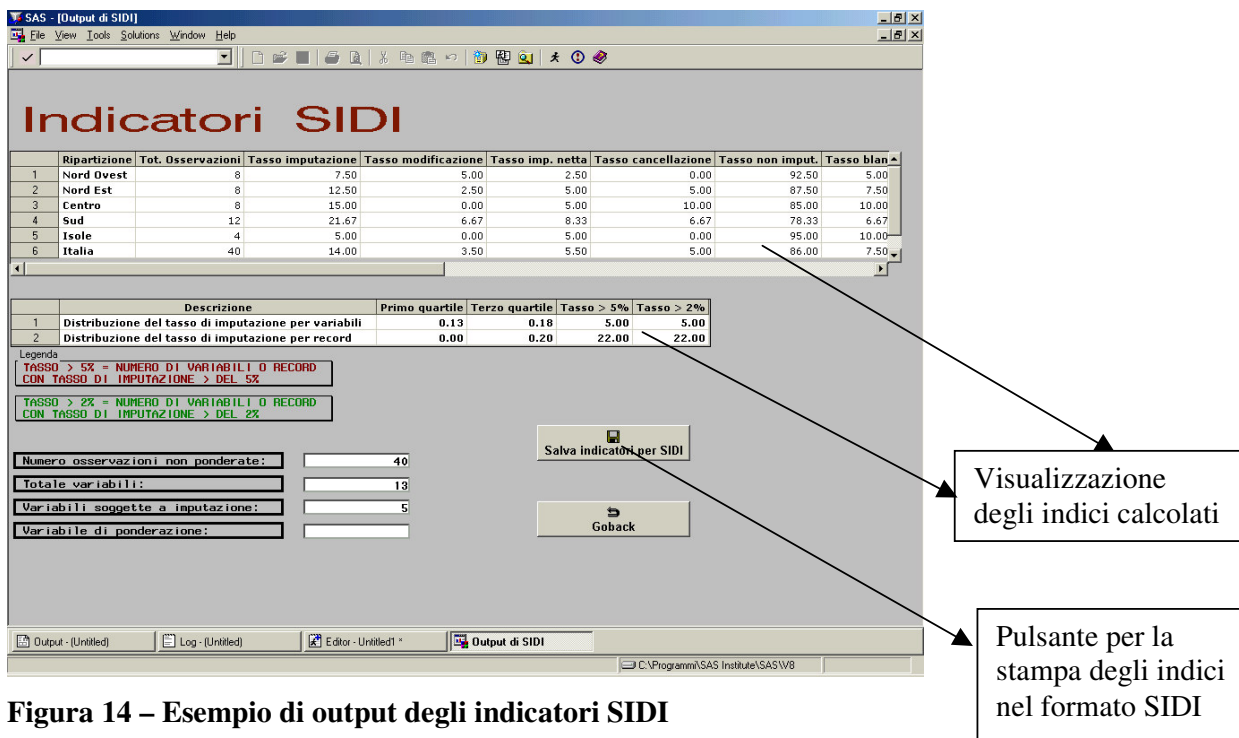
1. la variabile relativa all'informazione sulle ripartizioni geografiche (**Variabile di ripartizione**);
2. le variabili soggette a C&C nell'indagine (**Seleziona variabili**).



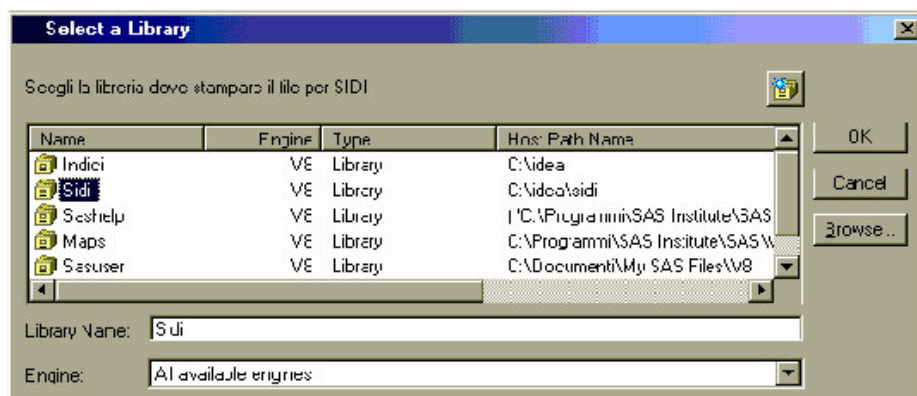
**Figura 13 – Maschera per la selezione delle variabili per il calcolo degli indicatori SIDI**

Una volta completata la fase di impostazione dei parametri e di dichiarazione delle variabili, si può eseguire l'elaborazione (**Run**), ottenendo i risultati illustrati nella maschera in Figura 14.

Se si vuole produrre il file per il calcolo degli indicatori di revisione per SIDI direttamente nel formato di import necessario per SIDI bisogna selezionare il pulsante **Salva Indicatori per SIDI**: una schermata successiva chiederà di selezionare la libreria in cui si vuole memorizzare il file, che sarà quindi archiviato nel *path* prescelto con il seguente nome: *nome\_del\_dataset\_corretto\_w.txt* nel caso di indicatori ponderati, *nome\_del\_dataset\_corretto\_now.txt* nel caso di indicatori non ponderati, come mostrato nella maschera in Figura 15.



**Figura 14 – Esempio di output degli indicatori SIDI**



**Figura 15 – Selezione del path per il salvataggio del file di input per SIDI**

#### 4. ASPETTI TECNICI: REQUISITI HARDWARE E SOFTWARE

L'attuale versione di IDEA è interamente sviluppata usando SAS SYSTEM v8.x, per cui i requisiti hardware e software necessari per l'installazione del prodotto sono i seguenti:

- per quanto riguarda i requisiti hardware, essi sono gli stessi necessari per l'utilizzo di SAS SYSTEM v8.x. L'ammontare di memoria disco necessaria per IDEA è di circa 2 Mb. L'ulteriore memoria necessaria è evidentemente dipendente dalla dimensione dei data set da elaborare;
- è richiesto un sistema operativo WINDOWS 95, WINDOWS 98 or WINDOS NT 4.0 o superiori;



3. per quanto riguarda i requisiti software, è necessario che siano installati i moduli del SAS v8.x **SAS/BASE** e **STAT**.

## **RIFERIMENTI BIBLIOGRAFICI**

- Agresti A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.
- Barcaroli G., D'Aurizio L. (1997). Evaluating Editing Procedures: the Simulation Approach, *UN/ECE Work Session on Statistical Data Editing*, Prague 1997.
- Barcaroli G., Della Rocca G., Di Zio, M., Luzi O., Manzari A., Seeber A.C. (2001). *E.S.S.E. User Manual*, Documento interno ISTAT.
- Beaumont J.-F., Mitchell C. (2002) The System for Estimation of Variance due to Nonresponse and Imputation (SEVANI), *Proceedings of the Statistics Canada Symposium 2002, Modeling Survey Data for Social and Economic Research* (to appear).
- Brancato G., Fanfoni L., Fortini M., Scanu M., Signore M. (2001) Il sistema SIDI: uno strumento generalizzato per il controllo di qualità delle indagini ISTAT, *Scritti di Statistica Economica*, n.7, CD-ROM.
- Chambers R. (2001). *Evaluation Criteria for Statistical Editing and Imputation* – T001.05, EUREDIT report ([www.cs.york.ac.uk/euredit/](http://www.cs.york.ac.uk/euredit/)).
- Della Rocca G., Luzi O., Scavalli E., Signore M., Simeoni G. (2003) Evaluating, monitoring and documenting the effects of editing and imputation in ISTAT surveys, *UN/ECE Work Session on Statistical Data Editing*, Madrid, 2003.
- Di Zio M., O. Luzi, Manzari A. (2002). Evaluating Editing and Imputation Processes: the Italian Experience, *UN/ECE Work Session on Statistical Data Editing*, Helsinki, 2002.
- Fortini M., Scanu M. and Signore M. (1999). Measuring and Analysing the Data Editing Activity in ISTAT Information System for Survey Documentation, *UN/ECE Work Session on Statistical Data Editing*, Rome, 1999.
- Fortini M., Scanu M. and Signore M. (2000). Use of indicators from data editing for monitoring the quality of the survey process: the Italian information system for survey documentation (SIDI), *Statistical Journal of the United Nations ECE*, n.17, pp. 25-35.
- Granquist L. (1997). An overview of methods of Evaluating Data Editing procedures, In *Statistical Data Editing, Methods and Techniques*, Vol. II Statistical Standard and Studies N. 48, UN/ECE 112-123.
- Lee H., Rancourt E., Särndal C.-E. (2001) *Variance Estimation from Survey Data under Single Imputation*, in Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A. (eds), *Survey Nonresponse*, New-York:John Wiley&Sons, Inc., pp. 315-328.



- Leti G. (1983), *Statistica descrittiva*, Il Mulino, Bologna.
- Rao, J.N.K. (2001). Variance Estimation in the Presence of Imputation for Missing Data. *Proceedings of the Second International Conference on Establishment Surveys (ICESII)*, pp. 599-608.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New York.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.
- Simeoni, G. (2001) *Gestione indicatori di qualità in SIDI: descrizione delle funzioni di caricamento dati e calcolo indicatori*, Documento interno ISTAT.