

**LA NUOVA FUNZIONE DI ANALISI DEI MODELLI
IMPLEMENTATA IN GENESEES V. 3.0**

di

PATRIZIA GIAQUINTO, MARCO LANDRISCINA, DANIELA PAGLIUCA ^(*)

^(*) Patrizia Giaquinto ha redatto il paragrafo 2, Marco Landriscina il paragrafo 5 e Daniela Pagliuca ha redatto i rimanenti paragrafi.

Introduzione

Il presente documento ha la finalità di descrivere una nuova funzione implementata nel software Genesees v3.0, che costituisce l'ultima versione del software Genesees (Generalised Software for Sampling Estimates and Errors in Surveys).

La versione precedente – la v2.0 – era nata basandosi su diverse procedure SAS, sviluppate in Istat dai ricercatori esperti Piero Demetrio Falorsi e Stefano Falorsi, per il calcolo dei pesi finali, delle stime e degli errori campionari utilizzando la teoria degli stimatori di regressione generalizzata e per la presentazione sintetica degli stessi errori di campionamento.

Tali procedure, dal punto di vista dell'architettura e degli algoritmi utilizzati, hanno costituito la base delle funzioni di *Riponderazione* e di *Stime ed Errori campionari*, già disponibili nella versione 2.0 di Genesees e ricollocate nella versione 3.0.

Il software generalizzato Genesees v3.0 è stato realizzato all'interno di un progetto di sviluppo nell'ambito della direzione DCMT e costituisce una attività programmata dall'unità MTS/F, che si occupa dello sviluppo di software generalizzati per la produzione statistica, la cui responsabile è Daniela Pagliuca. L'unità afferisce al servizio MTS “Servizio metodologie, tecnologie e software per la produzione statistica”, il cui responsabile è Giulio Barcaroli.

L'attività, svolta in collaborazione con l'unità PSM/A costituisce l'ultimo passo di un progetto di sviluppo che è partito con l'obiettivo di ottimizzare le procedure SAS iniziali - implementando i controlli necessari per l'esecuzione e sviluppando una interfaccia user-friendly per consentire agli utenti un'interazione di tipo avanzato – e si è concluso con lo sviluppo di un nuovo modulo. L'unità PSM/A, di cui è responsabile Stefano Falorsi, appartiene al servizio PSM “Progettazione e supporto metodologico nei processi di produzione statistica”, di cui è invece responsabile Piero Demetrio Falorsi.

Rispetto alla versione 2.0, il software Genesees v3.0 comprende difatti un modulo aggiuntivo: la funzione *Analisi dei Modelli*, che agevola l'utente nella rappresentazione sintetica degli errori campionari, permettendo l'analisi esplorativa dei dati, comprensiva di rappresentazione grafica, per individuare ed eventualmente eliminare i valori considerati *estremi* rispetto al modello.

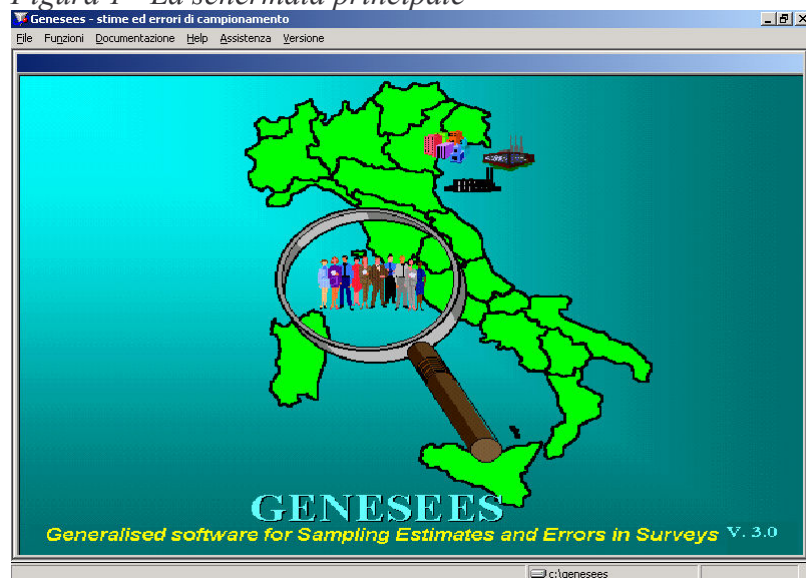
Il lavoro è strutturato come segue: nel paragrafo 1 viene presentato il software Genesees v3.0 e si fornisce una breve descrizione delle funzioni del software. Successivamente si procede con la trattazione nel dettaglio della funzione *Analisi dei Modelli*, a partire dal paragrafo 2 in cui si illustra la metodologia alla base della funzione. Al lavoro è infine allegata una appendice, utile per coloro che vogliano consultare i documenti tecnici progettuali.

1. Il software Genesees v3.0: un insieme di funzioni

Prima di descrivere la nuova funzione *Analisi dei Modelli* è opportuno fornire una descrizione del software nel suo insieme.

Genesees v3.0 viene attivato aprendo la schermata principale (cfr. *figura 1*):

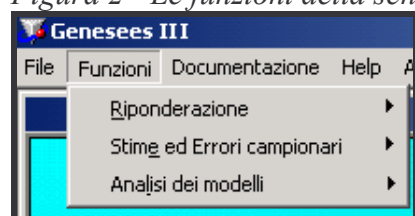
Figura 1 - La schermata principale



L'opzione **Funzioni** del menu principale fornisce la possibilità di accedere alle tre funzionalità principali implementate nel software (cfr. figura 2):

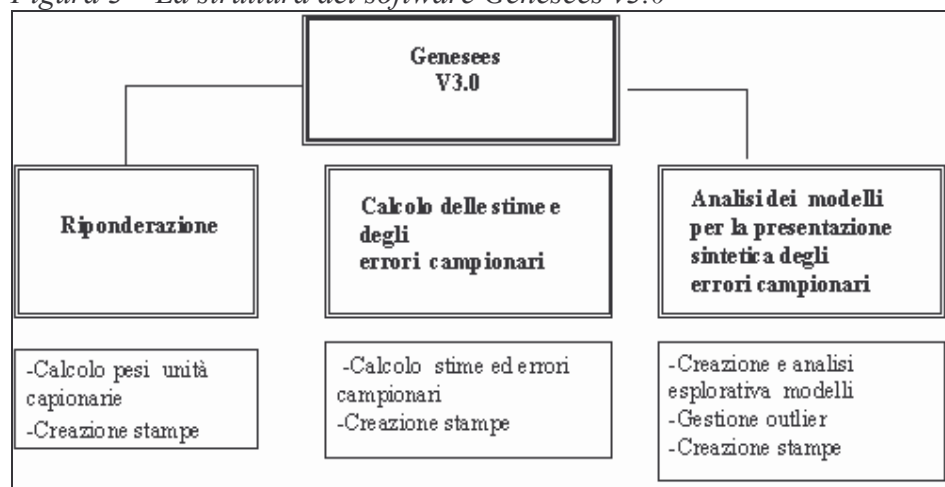
- Riponderazione
- Stime ed Errori campionari
- Analisi dei Modelli

Figura 2 - Le funzioni della schermata principale



Il software Genesees v3.0 è strutturato come mostrato nella successiva figura 3:

Figura 3 – La struttura del software Genesees v3.0



Genesees è un software sviluppato utilizzando il SAS SYSTEM v.8. per Microsoft Windows, ovvero un package di uso generale che incorpora statistiche e procedure di analisi dei dati. Per

utilizzare Genesee è necessario che sia installato il sistema SAS versione 8 ed in particolare i moduli: SAS Language and Macro-facility, SAS IML Language, SAS STAT, SAS GRAPH.

Lo spazio sul disco fisso necessario per l'installazione è di circa 4 MB ed è consigliabile una memoria di almeno 64 MB. Il tempo d'esecuzione della procedura è legato, ovviamente, alla velocità del processore installato e alla dimensione e complessità dei dati da elaborare.

Il software è disponibile anche effettuando il **download** :

- via internet (per utenti esterni all'Istat):
<http://www.istat.it/Metodologi/index.htm> (selezionare “Metodi e Software per indagini statistiche”).
- via intranet (per utenti Istat):
<http://intranet/> (selezionare: “Prodotti e Applicazioni on-line. Software Generalizzati” e da qui selezionare “MTS-F: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)”).

Propedeutica all'installazione del software è – ovviamente - quella del SAS v.8.

Sia il CD-ROM che il download del software permettono di ottenere un file compresso.

Il presente documento si riferisce in particolare alla funzione *Analisi dei Modelli*, mentre le funzioni di *Riponderazione* e *Stime ed Errori campionari* sono ampiamente descritte nei rispettivi manuali utente. A tal proposito si evidenzia che i manuali delle prime due funzioni, riferiti a Genesee v3.0, sono disponibili via intranet/internet (Pagliuca, 2004). I due manuali saranno a breve pubblicati nella nuova collana Tecniche e Strumenti, mentre il manuale utente della funzione *Analisi dei Modelli* è tuttora in corso di stesura e verrà pubblicato anche esso nella stessa collana.

La funzione di Riponderazione

La funzione di *Riponderazione* è applicabile in tutti i casi in cui esistono informazioni ausiliarie, espresse in termini di totali noti di variabili, definite appunto “ausiliarie”, legate alle variabili di interesse.

Essa è finalizzata al calcolo dei pesi finali da attribuire alle unità campionarie, sulla base di totali noti delle variabili ausiliarie e dei valori assunti da queste nel campione estratto.

Il contesto metodologico nel quale la funzione è stata concepita è quello degli stimatori di calibrazione (*calibration estimators*); tale teoria consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di calibrazione (Deville, Särndal, 1992).

La funzione di Stime ed Errori campionari

Lo scopo principale delle indagini campionarie è quello di fornire le stime di alcuni parametri descrittivi dell'intera popolazione, o di sottopopolazioni predefinite, dalla quale il campione viene estratto.

La funzione per il calcolo delle stime e degli errori campionari è finalizzata al calcolo delle stime e degli errori di campionamento e produce, per ciascuna sottopopolazione di interesse: le stime oggetto di indagine e i corrispondenti errori di campionamento assoluti, relativi, e gli intervalli di confidenza; le principali statistiche che forniscono informazioni sull'efficienza della strategia di campionamento utilizzata (effetto del disegno ed effetto dello stimatore); i modelli di regressione per la presentazione sintetica degli errori di campionamento.

Anche tale funzione fa riferimento alla teoria degli stimatori di calibrazione (Deville, Särndal, 1992).

La funzione *Analisi dei Modelli*

La funzione di *Analisi dei modelli* nasce come estensione rispetto a quanto già implementato in Genesees v2.0 e aiuta l'utente a determinare la migliore rappresentazione sintetica degli errori campionari.

Tale funzione permette infatti di costruire i modelli per la presentazione sintetica degli errori di campionamento, come già era previsto nella versione 2.0 di Genesees, ma permette in aggiunta di analizzare la validità di tali modelli, in modo semplice ed interattivo, grazie anche al supporto di alcune funzionalità grafiche. L'utente viene in tal modo agevolato nell'individuazione di alcuni valori giudicati *estremi* rispetto al modello scelto e può procedere alla determinazione di un nuovo modello, che non tenga in considerazione tali valori estremi.

Le principali caratteristiche metodologiche di Genesees sono contenute nei lavori seguenti: Falorsi e Falorsi (1995), Falorsi e Falorsi (1997), Falorsi e Falorsi (1998), Falorsi e Rinaldelli (1998), Falorsi, Pagliuca e Scepi (1999), Falorsi, Pagliuca e Scepi (2000), Pagliuca e Righi (2002), De Vitiis e Pagliuca (2003).

2. Cenni metodologici relativi alla funzione *Analisi dei modelli*

Nel momento in cui vengono pubblicati i risultati di un'indagine campionaria, per valutare la variabilità delle stime prodotte diventa necessario affiancare a ciascuna stima il corrispondente errore campionario; in tal modo però, le tavole di pubblicazione risulterebbero praticamente illeggibili, data la numerosità delle stime.

Per ovviare a tale difficoltà, sono stati concepiti dei modelli matematici in grado di rappresentare in maniera essenziale i suddetti errori di campionamento. Le metodologie impiegate a riguardo sono due: una basata sull'*effetto del disegno di campionamento* (Verma, Scott e O'Muircheartaigh, 1980; Verma, 1982; Wolter, 1985) ed una basata sui *modelli regressivi* (Russo, 1987).

La funzione di *Analisi dei Modelli* adotta questa ultima metodica.

In questo paragrafo vengono forniti i principi fondamentali della teoria sottostante (per una ampia trattazione di tale metodologia è possibile leggere l'*appendice A5* del Manuale Utente della funzione di *Stime ed Errori campionari* (Pagliuca 2004b)).

Siano $X_1, X_2, \dots, X_i, \dots, X_t$ t parametri di interesse, le cui stime sono indicate con $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_i, \dots, \hat{X}_t$.

Si definiscono, inoltre, le varianze delle stime: $\sigma^2(\hat{X}_1), \sigma^2(\hat{X}_2), \dots, \sigma^2(\hat{X}_i), \dots, \sigma^2(\hat{X}_t)$ e gli errori relativi $e^2(\hat{X}_1), e^2(\hat{X}_2), \dots, e^2(\hat{X}_i), \dots, e^2(\hat{X}_t)$, in cui:

$$e^2(\hat{X}_i) = \sigma^2(\hat{X}_i) / X_i^2. \quad i = 1, \dots, t \quad [1]$$

La metodologia basata sui *modelli regressivi* si fonda sulla idea che per poter presentare in maniera concisa gli errori campionari è fondamentale individuare un'opportuna relazione matematica, che esprima gli errori in funzione delle stime di riferimento.

L'ipotesi basilare del metodo utilizzato è differente a seconda che le variabili oggetto di stima siano qualitative o quantitative.

Nel primo caso, e cioè per variabili qualitative di cui si vogliano stimare le frequenze assolute o relative, sussistono fondamenti teorici per affermare che l'andamento dell'errore relativo dipende solo dal valore delle X_i . In altri termini la funzione esplicativa degli errori campionari è decrescente al crescere delle stime stesse.

Se per le variabili qualitative sussiste tale fondamento teorico, per le stime di totali di variabili quantitative, invece, il criterio adottato è di tipo empirico ed è fondato su risultati sperimentali, che evidenziano che l'errore assoluto di un totale appare essere una funzione crescente del totale stesso.

Indicando per comodità espositiva con \hat{X} la stima del generico parametro X_i , in linea generale si parte da un modello teorico di riferimento del tipo:

$$e^2(\hat{X}) = f(X, \vartheta_1, \dots, \vartheta_s, c), \quad [2]$$

in cui $\vartheta_1, \dots, \vartheta_s$ costituiscono dei parametri incogniti, c indica la componente accidentale.

Nella pratica viene utilizzato il modello operativo

$$\hat{e}^2(\hat{X}) = f(\hat{X}, \vartheta_1, \dots, \vartheta_s, c), \quad [3]$$

in cui occorre stimare i parametri $\vartheta_1, \dots, \vartheta_s$.

La stima dei parametri si ottiene scegliendo un numero congruo di stime (minore o uguale al numero totale t); ovviamente, la rappresentatività del modello aumenta in base al numero delle stime esaminate.

Si perviene in tal modo alla stima $\hat{e}^{*2}(\hat{X})$:

$$\hat{e}^{*2}(\hat{X}) = f(\hat{X}, \hat{\vartheta}_1, \dots, \hat{\vartheta}_s), \quad [4]$$

e da questa all'errore campionario:

$$\hat{e}^*(\hat{X}) = \sqrt{\hat{e}^{*2}(\hat{X})} \quad [5]$$

Per quanto riguarda la scelta del modello matematico, per la stima delle frequenze nel caso di variabili qualitative il seguente modello è idoneo a rappresentare il legame esistente fra le stime e gli errori campionari (e definisce dunque la f nella [2]):

$$\hat{e}^2(\hat{X}) = \exp(\alpha + \beta \ln \hat{X} + v), \quad [6]$$

che può essere linearizzato considerando il logaritmo di entrambi i membri della [6] e giungendo al modello equivalente:

$$\log\left(\hat{e}^2(\hat{X})\right) = \alpha + \beta \ln \hat{X} + v.$$

Esso esprime la nota relazione regressiva classica fra $\hat{e}^2(\hat{X})$ e \hat{X} e pertanto si può giungere alla stima dei parametri α e β mediante il metodo dei minimi quadrati ottenendo il modello stimato:

$\hat{e}^2(\hat{X}) = \exp(a + b \ln \hat{X} + v)$, nel quale a e b rappresentano rispettivamente le stime dei parametri α e β e v indica il residuo.

Nel caso di stime di totali di variabili quantitative, invece, il modello utilizzato è del seguente tipo:

$$\sigma(\hat{X}) = \gamma + \lambda \hat{X} + \delta \hat{X}^2 + v \quad [7]$$

dove $\sigma(\hat{X})$ indica questa volta l'errore assoluto della stima \hat{X} del totale, γ, λ, δ i parametri da stimare per il modello e v è la componente casuale.

Il modello stimato corrispondente in questo caso è:

$$\hat{\sigma}(\hat{X}) = g + l \hat{X} + d \hat{X}^2 + u \quad [8]$$

in cui, analogamente al caso precedente, l , d e g rappresentano le stime dei parametri e u costituisce il residuo.

Si noti che il modello in questo caso viene ottenuto, come detto in precedenza, sulla base di considerazioni di tipo empirico, per cui è anche possibile utilizzare dei modelli alternativi al [7] e selezionare quello che fornisce il migliore adattamento ai dati osservati. Per lo stesso motivo, in questo caso, è preferibile utilizzare un numero elevato di stime (il numero totale e non un suo sottoinsieme), poiché altrimenti non verrebbe garantito un buon livello di adattamento anche per le stime non incluse in quelle disponibili sulla base delle quali si stimano i parametri.

Per quanto concerne l'adattamento del modello scelto ai dati osservati, è opportuno fare uso di uno strumento di misura che verifichi l'accostamento ai dati reali ottenuto tramite la funzione interpolatrice adottata (retta di regressione nel caso di variabili qualitative e parabola nel caso delle quantitative). E' prassi comune impiegare allo scopo l'indice di determinazione R^2 , ottenuto come rapporto fra la devianza spiegata dal modello e la devianza totale.

Nella fattispecie, per il modello [6], o l'equivalente modello linearizzato, l'indice di determinazione è dato da:

$$R^2 = \frac{\text{Dev}(\hat{e}^{*2}(\hat{X}))}{\text{Dev}(\hat{e}^2(\hat{X}))} \text{ nel caso di variabili qualitative,}$$

e per il modello [6] da:

$$R^2 = \frac{\text{Dev}(\hat{\sigma}^*(\hat{X}))}{\text{Dev}(\hat{\sigma}(\hat{X}))} \text{ nel caso di variabili quantitative.}$$

L'indice R^2 varia fra 0 e 1, raggiungendo l'estremo superiore (1) nel caso in cui la variabilità totale è interamente spiegata dalla variabilità del modello, dando origine ad una perfetta rappresentazione dei dati tramite funzione interpolata. Nei casi reali, ovviamente, tale limite non è mai raggiunto, ma valori prossimi a 0,8 o a 0,9 sono indicativi di una buona rappresentatività.

3. La funzione *Analisi dei Modelli* di Genesees v3.0

La funzione *Analisi dei Modelli* permette una analisi esplorativa dei risultati prodotti dalla funzione *Stime ed Errori campionarie* di Genesees. E' da osservare che l'elaborazione dei modelli [6] e [7] descritti nel paragrafo 2 era già implementata nella versione precedente di Genesees; sulla base delle stime e degli errori campionari, con la versione 2 era già possibile ottenere alcune tabelle, contenenti informazioni sui modelli, che rappresentavano un output della funzione *Stime ed Errori campionari*. Anche nella versione 3 di Genesees i modelli sono ottenibili come stampe di output della funzione di *Stime ed Errori campionari*. Tali modelli possono però non essere soddisfacenti e potrebbero richiedere una ulteriore analisi, per verificare se sia possibile ottenere un miglior adattamento. Il problema dell'analisi dei modelli si riconduce dunque a quello generale dell'adattamento di un modello di regressione ad una serie di dati; l'analisi può essere agevolata ricorrendo ad una rappresentazione grafica, in cui i dati osservati costituiscono una nuvola di punti ed il modello è una curva interpolatrice, ovvero una retta nel caso del modello [6] linearizzato ed una parabola nel caso del modello [7]). Un R^2 basso può essere determinato sia da una nuvola di punti molto dispersa che dalla presenza di uno o più *valori estremi* che potrebbero inficiare la validità del modello. Nel primo caso risulta difficile la scelta di un modello che rappresenti in maniera adeguata i dati osservati; nel secondo caso la presenza di uno o più *valori estremi* potrebbe inficiare la validità del modello, dove con *valori estremi* (o *outlier*) non si intendono *outlier* rispetto al valore delle stime - ovvero quei valori la cui esclusione dal campione potrebbe determinare un forte impatto sulle stime - ma si intendono piuttosto quei valori che si discostano dagli altri in riferimento al modello di regressione e che, presentando errori campionari estremi, potrebbero determinare uno spostamento della curva di regressione.

Per ottenere una buona presentazione sintetica tramite modelli di regressione, potrebbe essere indispensabile l'eliminazione di alcuni valori estremi. Per consentire l'eliminazione di uno o più punti dalla nuvola di punti a cui adattare il modello, è necessario stabilire una soglia al di sopra della quale considerare un punto come *valore estremo*. Tale questione indubbiamente non ha una soluzione generale.

Allo scopo di agevolare l'utente, si è pertanto inserito in Genesees v3.0 la funzione *Analisi dei Modelli*, supportata da visualizzazioni grafiche, che aiuta l'utilizzatore ad individuare gli eventuali *valori estremi* da eliminare.

Il software permette infatti di visualizzare graficamente la nuvola di punti e le curve di regressione [6] e [7] e permette anche di calcolare e visualizzare graficamente i residui standardizzati, per ciascun dominio di stima pianificato.

La visualizzazione grafica dei residui rende possibile anche l'identificazione della soglia oltre cui considerare un punto come *outlier* e Genesees permette infatti l'eliminazione degli *outlier* sulla base dei valori di tali soglie.

La scelta delle soglie, così come l'eliminazione dei *valori estremi*, può effettuarsi in modo differenziato per i diversi domini di stima pianificati. E' utile chiarire a tal fine che con *dominio di stima pianificato* si intendono le sottopopolazioni per le quali le indagini forniscono le stime.

Una volta eliminati i *valori estremi*, è possibile stimare nuovamente i parametri del modello e verificare se si è conseguito un miglioramento nell'adattamento.

In conclusione la versione 3 di Genesees permette dunque di supportare il processo di produzione statistica attuato nelle indagini campionarie, per la realizzazione delle importanti fasi di attività che riguardano il calcolo dei pesi, delle stime e degli errori campionari e la loro rappresentazione sintetica; l'utente ha anche a disposizione uno strumento di analisi esplorativa che consente di scegliere il modello più adeguato per la pubblicazione degli errori campionari in modo sintetico.

4. L'input della funzione *Analisi dei Modelli* di Genesees v3.0

La funzione *Analisi dei Modelli* rappresenta l'ultimo passo effettuabile in un processo di calcolo dei pesi, delle stime e degli errori campionari e ovviamente implica che le fasi precedenti siano state ultimate.

In termini pratici, per utilizzare tale funzione è necessario che sia disponibile un *data-set* Sas, *wtotale* che corrisponde al *data-set* di output della funzione di *Stime ed Errori campionari*¹. Tale funzione permette infatti, a partire dai valori osservati delle variabili di interesse, di ottenere un primo *data-set* utilizzabile per le elaborazioni successive, contenente le stime e gli errori campionari per ciascuno strato (*wstrato*) e da questo un secondo *data-set* contenente le stesse informazioni per ciascun *dominio di stima* (*wtotale*) (*capitolo 2, sezione II*, Pagliuca, 2004b)

I *domini di stima pianificati* in Genesees si ottengono come unione di *strati* ed, in tal modo, gli *strati* rappresentano la minima partizione atta a determinare il livello di interesse al quale riportare le stime finali (provincia, regione,...). Le variabili che servono a definire i *domini di stima non pianificati* vengono invece dette *sottoclassi* e servono dunque a definire partizioni della popolazione rispetto alle quali interessano le stime finali che non rappresentano unione di *strati*. In altri termini i *domini di stima non pianificati* sono sottoinsiemi della popolazione caratterizzati dal fatto che non tutte le unità di uno stesso strato appartengono allo stesso sottoinsieme della partizione.

Il *data-set wtotale* contiene le stime di interesse riferite ad una combinazione di variabili: il singolo record individua infatti una singola stima, il cui identificativo è individuato dalle variabili *domst*, *xvariabil*, *modalità*, *xsottocla*, *modscl* descritte in tabella 1 (tale tabella riporta anche altre informazioni del *data-set wtotale*, utili per la funzione *Analisi dei modelli*).

Le variabili *xvariabil* e *xsottocla* rappresentano le variabili di interesse e di sottoclasse e sono definite entrambe da un numero progressivo, mentre le variabili *labvar* e *labsottocla* sono descrittive e mettono in corrispondenza i numeri progressivi con i nomi delle variabili originarie definite dall'utente.

Infine le variabili *STIMA* ed *ERREL* sono le variabili che rappresentano la stima e l'errore relativo e che sono alla base dei modelli [6] e [7] (cfr. *paragrafo 2*).

¹ *Wtotale* (*data-set* sas v.8) corrisponde al file *wtotale.sas7bdat* del *file manager* di Windows o al *data-set* Sas *libname.totale*, dove con *libname* si è indicato il nome di una *libreria* dell'ambiente Sas, associata al *data-set*. Per semplificare l'esposizione successiva si farà riferimento ai *data-set* solo con il nome, senza l'estensione del file o la libreria di riferimento.

Tabella 1 – Informazioni del data-set WTOTALE utili per la funzione Analisi dei modelli

nome variabile	significato della variabile del data-set	
domst o domstn	dominio pianificato	codice di dominio pianificato (domst compare nel caso di stime per variabili qualitative; domstn compare, invece, nel caso di stime per variabili quantitative)
xvariabil	Progressivo identificativo della variabile di interesse	numero progressivo della variabile d'interesse
labvar	Nome della variabile di interesse scelto dall'utente	nome della variabile di interesse
modalita	modalità variabile	modalità della variabile di interesse (se la variabile è qualitativa e valore 1 nel caso in cui la variabile è quantitativa)
xsottocla	Progressivo identificativo della variabile di sottoclasse	numero progressivo di riferimento della variabile di sottoclasse (pari a 0 se le informazioni sull'osservazione non considerano la suddivisione in sottoclassi)
labsottocla	Nome della variabile di sottoclasse scelto dall'utente	nome della variabile di sottoclasse (se xsottocla è pari a "0" allora labsottocla non contiene alcuna informazione)
ERREL	Errore relativo	Errore relativo (o Coefficiente di variazione)
STIMA	Stima	Stima del totale

Wtotale è dunque un insieme di stime e di errori, riferiti a ciascun dominio pianificato, per ogni sottoclasse all'interno del dominio pianificato e per ciascuna delle variabile di interesse.

5. Presentazione delle maschere della funzione

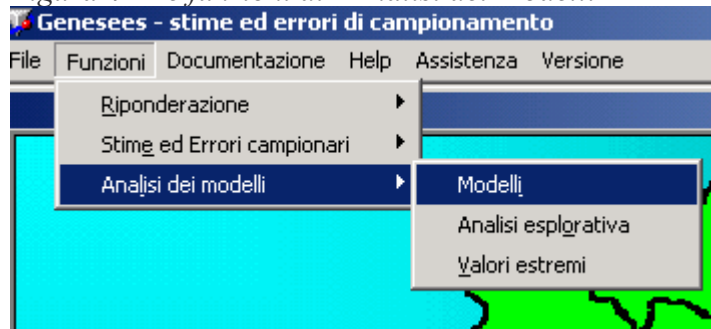
Tramite la funzione *Analisi dei modelli* si accede ad un area interattiva di Genesee v3.0, che permette di realizzare le fasi necessarie per giungere alla rappresentazione sintetica degli errori campionari. Il primo passo da compiere in tale processo è costituito dall'elaborazione del modello matematico che si ritiene rappresentativo per le osservazioni del *data-set wtotale* (e quindi per le stime di riferimento). Successivamente è possibile visualizzare graficamente sia dette osservazioni, raggruppate per *dominio di stima pianificato*, che la funzione matematica interpolatrice, così come è possibile prendere visione dei parametri stimati del modello e del valore dell'indice di determinazione R^2 ottenuto, i quali sono riportati in apposita tabella. L'ultima fase riguarda l'individuazione dei valori estremi, che possono quindi essere eliminati dal *data-set* delle osservazioni, al fine di raggiungere un grado di adattamento migliore del modello. A tal fine, si deve rielaborare il modello e ripetere l'analisi esplorativa (in forma di rappresentazione grafica o osservazione del nuovo valore di R^2).

Ognuna delle suddette operazioni può essere effettuata più volte, scegliendo di elaborare in alternativa il primo o il secondo modello (rispettivamente descritti dalla [6] e dalla [7]) ed eliminando o reinserendo osservazioni. E' possibile poi stampare i risultati ottenuti in formato txt o excel, oltre che mostrarli direttamente su video.

I file corrispondenti, nei tre formati indicati, sono disponibili al termine dell'elaborazione nella cartella di lavoro indicata dall'utente (cartella di output della funzione *Analisi dei Modelli*).

Di seguito si effettua una panoramica dell'ambiente interattivo della funzione *Analisi dei modelli*, illustrandone il funzionamento e fornendo una sorta di manuale utente sintetico.

Figura 4 - Le funzioni di “Analisi dei Modelli”

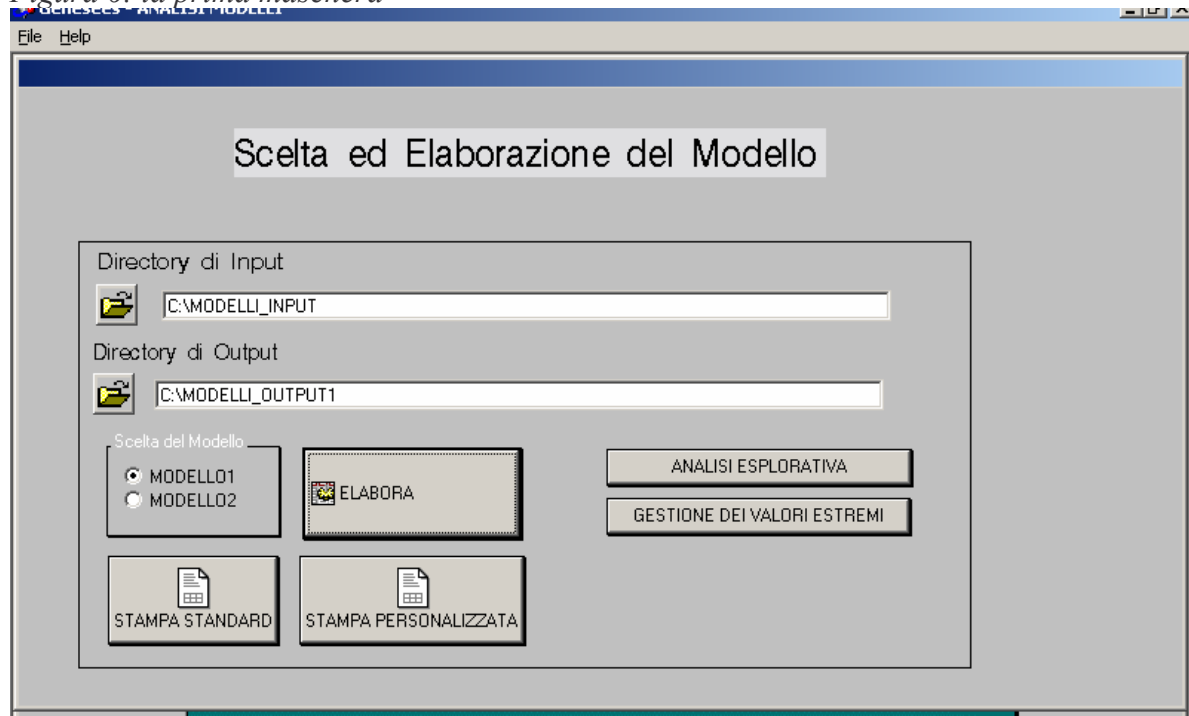


La funzione consta di tre maschere principali, corrispondenti alle tre voci del menu attivato da “Analisi dei Modelli” (figura 4). La prima maschera permette la elaborazione dei modelli e la produzione delle stampe. La seconda maschera consente la visualizzazione grafica del *data-set* *wtotale*, raggruppando le osservazioni per *dominio di stima pianificato* e permettendo una facile individuazione di valori estremi eventualmente presenti. In particolare, essa attiva anche la visualizzazione della tabella contenente i parametri stimati del modello utilizzato e il valore dell'indice R^2 ad esso relativo, nonché la visualizzazione di una tabella di frequenze, in cui i residui standardizzati originati dal modello vengono riportati suddivisi in classi prefissate di valori.

Tramite la terza maschera l'utente può indicare con un flag, che viene memorizzato sul *data-set* *totale*, le osservazioni relative ai valori estremi individuati, per escluderli da una nuova elaborazione del modello.

5.1 La prima maschera

Figura 6: la prima maschera



Nella prima maschera si definisce l'ambiente di lavoro, scegliendo la cartella di input e la cartella di output. Nella cartella di input deve essere presente il *data-set* *wtotale*.

Nella cartella di output verranno memorizzati i *data-set* creati dalla funzione e le stampe prodotte. La scelta effettuata viene considerata il default nel seguito dell'elaborazione, in tutte le maschere successive quindi il campo relativo alle cartelle risulta già impostato. Si noti che, nel momento in cui si desidera produrre una stampa personalizzata, come descritto successivamente, è necessario che nella cartella di output sia presente il *data-set* *nuove_stime*. Opportuni messaggi di errore o di warning guidano l'utente in tali operazioni, pertanto i vari bottoni della maschera si attivano solo se vengono rispettati i presupposti richiesti dalle operazioni stesse.

I due bottoni "Analisi esplorativa" e "Gestione dei valori estremi" permettono di passare rispettivamente alla seconda e terza maschera.

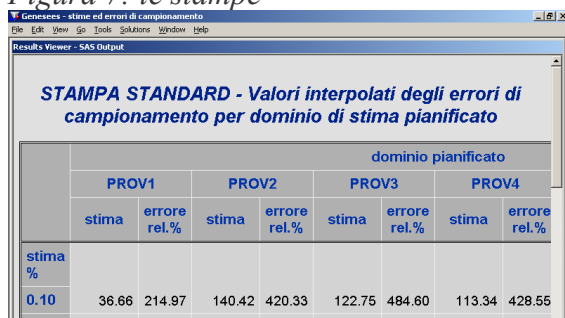
○ La scelta ed elaborazione del modello

Definito l'ambiente di lavoro si può scegliere, tra i due, il modello che interessa ed avviare l'elaborazione (bottone "elabora"). La scelta del modello e il pulsante "elabora" saranno attivi solo se sono state scelte correttamente le cartelle di lavoro. Un opportuno messaggio informerà l'utente se il modello scelto è stato elaborato correttamente oppure no. In caso di esito positivo sulla cartella di output sono stati scritti una serie di *data-set* e si saranno attivati i bottoni per la produzione delle stampe.

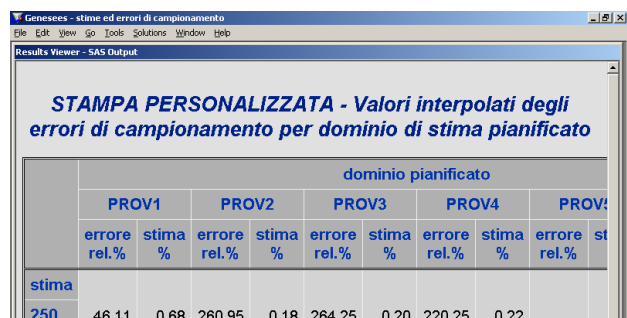
○ Le stampe

Se l'elaborazione termina correttamente si attivano i bottoni per le stampe. I due bottoni "stampa_standard" e "stampa_personalizzata" propongono a video delle stampe in formato html

Figura 7: le stampe



		dominio pianificato							
		PROV1		PROV2		PROV3		PROV4	
		stima	errore rel.%	stima	errore rel.%	stima	errore rel.%	stima	errore rel.%
stima	%								
0.10		36.66	214.97	140.42	420.33	122.75	484.60	113.34	428.55

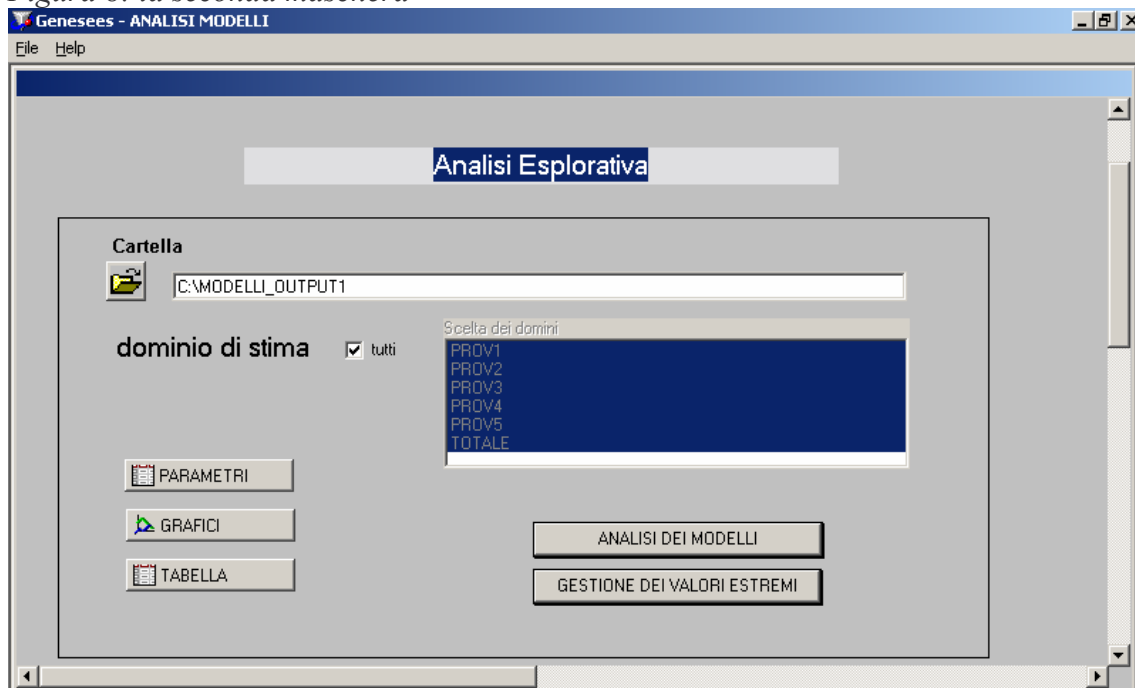


		dominio pianificato									
		PROV1		PROV2		PROV3		PROV4		PROV5	
		errore rel.%	stima %	errore rel.%	stima %	errore rel.%	stima %	errore rel.%	stima %	errore rel.%	stima %
stima											
250		46.11	0.68	260.95	0.18	264.25	0.20	220.25	0.22		

e producono alcuni file informativi nella cartella di output. Per informazioni aggiuntive sulle stampe prodotte si può leggere il successivo *paragrafo 6*.

5.2 La seconda maschera: l'analisi esplorativa

Figura 8: la seconda maschera



Questa maschera permette di selezionare i domini su cui effettuare l'analisi, utilizzando una scroll list che elenca quelli presenti nel *data-set wtotale*.

Una volta scelti i domini di analisi si hanno 3 possibilità:

- 1) Parametri: è possibile visualizzare i parametri di regressione calcolati dal modello
- 2) Grafici: è possibile ottenere una visualizzazione grafica delle osservazioni suddivise per domini
- 3) Tabella: è possibile visualizzare la tabella di frequenze dei residui standardizzati

Sulla base dei grafici, eventuali *valori estremi* vengono facilmente individuati ed è possibile procedere alla loro eliminazione grazie alla terza maschera.

I due bottoni "Analisi dei Modelli" e "Gestione dei valori estremi" permettono di passare rispettivamente alla prima e terza maschera.

- Parametri

Figura 9: i parametri di regressione

PARAMETRI di REGRESSIONE Visualizzazione				
	DOMST	R2	A	B
1	PROV1	82.336935635	7.3068580241	-1.603761739
2	PROV2	84.038584283	11.0443956	-1.652831356
3	PROV3	71.03368924	11.3576462	-1.705013222
4	PROV4	84.832302197	10.870890296	-1.682833523
5	PROV5	89.716823474	11.951418205	-1.762212066
6	TOTALE	88.356373313	12.75056064	-1.72983581

E' una visualizzazione del *data-set* contenente i parametri i regressione calcolati dal modello selezionato, suddivisi per dominio di stima pianificato.

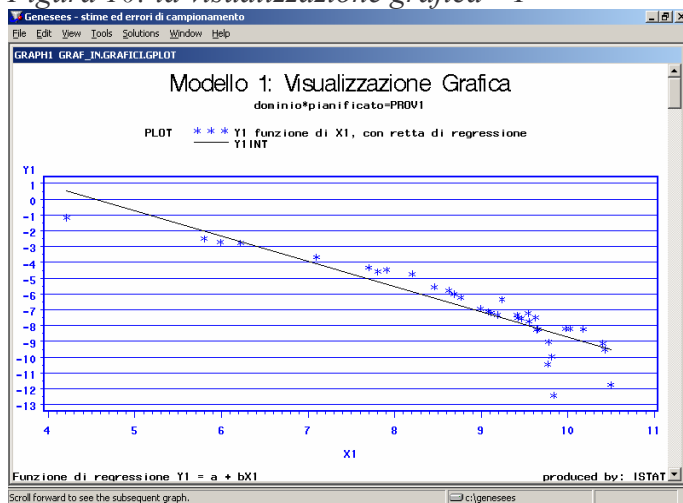
- Grafici

I grafici mostrano, per ciascun dominio di stima pianificato, le osservazioni e le relative curva di regressione ottenute mediante i modelli [6] o [7] (primo grafico) e mostrano anche i residui standardizzati (secondo grafico).

L'eventuale eliminazione dei *valori estremi* è facilmente verificabile sulla base del grafico dei residui standardizzati.

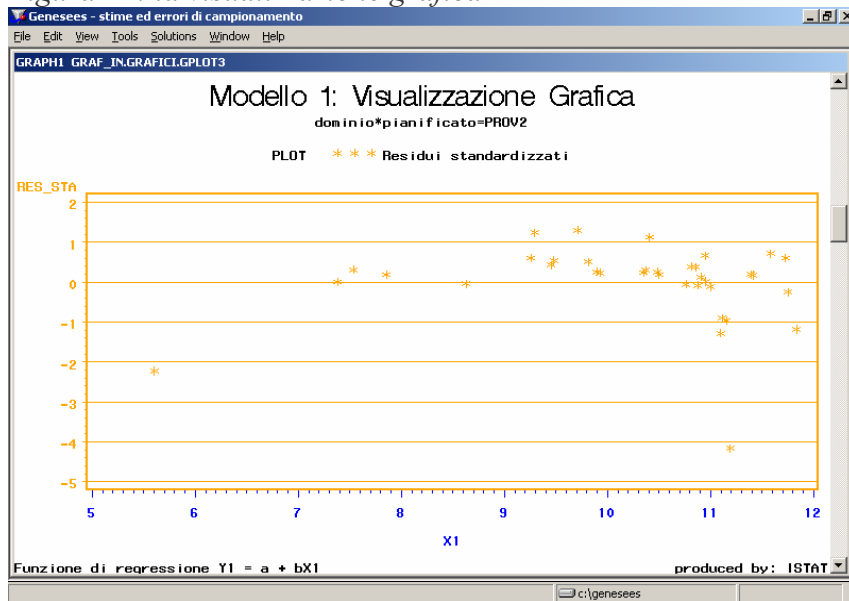
La figura mostra un output di Genesees, derivante da una specifica applicazione del modello [6]. In ascissa sono riportati i valori delle stime e in ordinata sono riportati contemporaneamente i valori del logaritmo delle stime degli errori e il modello di regressione stimato. Dalla figura è facile osservare come siano immediatamente identificabili i punti più lontani dalla retta di regressione.

Figura 10: la visualizzazione grafica – I



Viene anche presentata la visualizzazione dei residui standardizzati; in base all'analisi dei residui è facile individuare *valori estremi*, caratterizzati da residui standardizzati distanti da zero.

Figura 11: la visualizzazione grafica - II



- **La tabella di frequenza dell'analisi esplorativa**

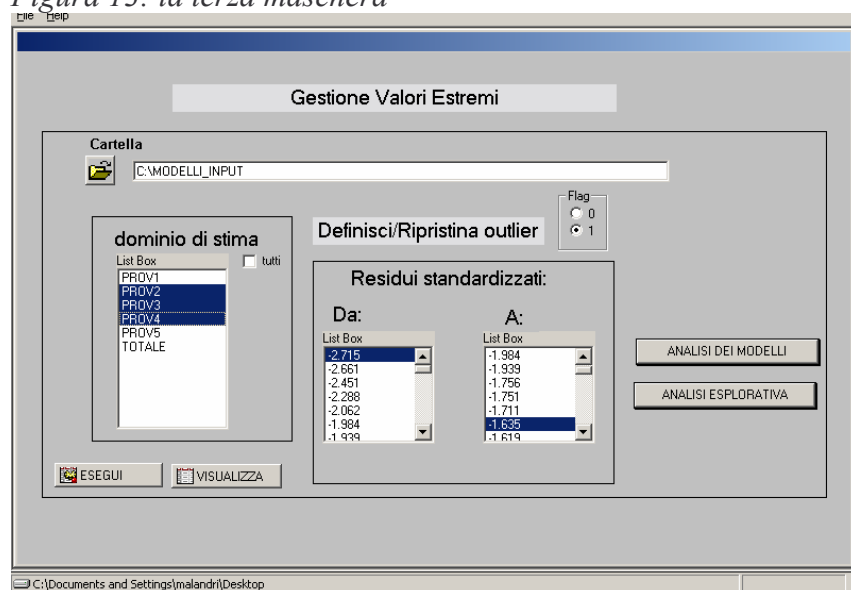
E' una tabella di frequenza che presenta, per ogni dominio di stima pianificato, il numero dei residui standardizzati che ricadono in classi di valori predeterminati.

Figura 12: la tabella di frequenze

TABELLA di FREQUENZA Visualizzazione			
	DOMST	classe	frequenza
1	PROV1	0 =< abs(RES_STA)< 1	26
2	PROV1	1 =< abs(RES_STA)< 2	8
3	PROV1	2 =< abs(RES_STA)< 2.5	1
4	PROV1	2.5 =< abs(RES_STA)< 3	1
5	PROV2	0 =< abs(RES_STA)< 1	24
6	PROV2	1 =< abs(RES_STA)< 2	11
7	PROV2	2.5 =< abs(RES_STA)< 3	1
8	PROV3	0 =< abs(RES_STA)< 1	25
9	PROV3	1 =< abs(RES_STA)< 2	9
10	PROV3	2 =< abs(RES_STA)< 2.5	2
11	PROV4	0 =< abs(RES_STA)< 1	22
12	PROV4	1 =< abs(RES_STA)< 2	13

5.3 La terza maschera: gestione dei valori estremi

Figura 13: la terza maschera



La cartella selezionata deve contenere il *data-set* *wtotale*. Si selezionano/deselezionano i domini di stima di interesse, si definisce il range di valori estremi che si vogliono eliminare/reinserire per un'ulteriore elaborazione del modello.

Si pone flag = 1 per eliminare il valore dall'elaborazione del modello, si pone flag = 0 per reinserire il valore nell'elaborazione del modello.

La selezione si effettua con riferimento ai residui standardizzati.

Il bottone "esegui" modifica effettivamente il *data-set* *wtotale*, il bottone "visualizza" mostra i record selezionati senza modificare il *data-set* (figura 14).

Il *data-set* "outlier", presente sulla cartella di output, mostra tutte le osservazioni di *wtotale* da considerare come *valori estremi*, aggiornato all'ultima elaborazione del modello.

I due bottoni "Analisi dei Modelli" e "Analisi esplorativa" permettono di passare rispettivamente alla prima e seconda maschera.

Figura 14 : Visualizzazione valori estremi

	MODSCL	MODALITA	OSSERVAZ	UP	UF	COMUNI
1	2	0	500	154	204	154
2	2	1	500	154	204	154
3	2	0	500	5	173	5
4	2	0	500	101	163	101

6. Le stampe di output

La funzione *Analisi dei modelli* permette la creazione di due stampe, la stampa *Standard* e la *Personalizzata*, che possono ottenersi con riferimento ad entrambi i modelli [6] o [7] (cfr. paragrafo 2).

La stampa denominata *Standard* corrisponde alle stampe 5b o 7b ottenibili dalla funzione di *Stime ed Errori campionari* (per approfondimenti si legga il capitolo 6 in (Pagliuca 2004b)). Ovviamente la funzione *Analisi dei modelli* permette di ottenere tali stampe sulla base dei modelli ulteriori che l'utente genera, rispetto a quelli ottenibili tramite la funzione di *Stime ed Errori campionari*, nel tentativo di migliorare l'adattamento del modello ai dati, eliminando alcuni valori reputati *estremi*.

La stampa *Standard* considera alcuni valori di stima percentuale crescenti e, in corrispondenza di ogni stima, vengono calcolati i valori assoluti delle stime e gli errori relativi percentuali, calcolati secondo i modelli [6] o [7], con riferimento a tutti i diversi domini di stima pianificati ed al totale generale.

La stampa *Standard* riferita al modello [6] è relativa al modello regressivo per la presentazione sintetica degli errori delle stime di frequenze e riporta dunque gli errori relativi percentuali, calcolati secondo il modello [6], con riferimento a percentuali definite, percentuali che vanno dallo 0.1% fino al 50% della popolazione campionaria. In particolare la prima colonna riporta le *stime %*, ovvero una serie di frequenze percentuali che si ritengono utili per l'utente. Tali stime, espresse in termini di frazioni percentuali della stima del totale della popolazione, corrispondono alle frazioni 0,1%, 0,5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%. Sulla base di queste percentuali, vengono dunque presentati le stime assolute e gli errori relativi percentuali.

La stampa *Standard* riferita al modello [7] è relativa al modello regressivo per la presentazione sintetica degli errori delle stime di totali di variabili quantitative e riporta dunque gli errori relativi percentuali calcolati secondo il modello [7], sempre con riferimento ad un insieme predefinito di valori tipici di stime percentuali, che in questo caso variano dallo 0.01% fino al 50%: sono presentati gli errori, calcolati secondo il modello, con riferimento alle percentuali 0,01% 0,02% 0,03% 0,04% 0,05% 0,1%, 0,5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%. Sulla base di queste percentuali, vengono dunque presentati le stime assolute e gli errori relativi percentuali.

Le informazioni contenute nella *Stampa Standard* permettono di calcolare l'errore relativo di una qualsiasi stima di frequenza assoluta con riferimento ad un certo dominio di stima pianificato, o per approssimazione (cercando il livello di stima che più si avvicina a quello di interesse) o mediante una semplice espressione matematica, tale per cui, su una qualsiasi retta, dati due punti distinti si può facilmente trovare un punto intermedio.

Per facilitare ulteriormente l'utente, si è voluto comunque aggiungere la stampa *Personalizzata*. Tale stampa permette di ottenere gli errori, calcolati tramite i modelli, relativi a stime che l'utente sceglie. Per far ciò l'utente deve creare un *data-set* (*nuove_stime*) in cui memorizza tutte le stime di interesse.

APPENDICE : DOCUMENTI DI PROGETTAZIONE DELLA FUNZIONE *ANALISI DEI MODELLI* DI GENESEES V3.0

FASE 1 - Analisi relativa alla prima fase progettuale

La funzione *Analisi dei modelli* sugli errori campionari” deve permettere:

- 1) la creazione dei modelli per la presentazione sintetica degli errori campionari secondo due modelli stabiliti (denominati Modello (1) e (2) nel seguito e corrispondenti ai [6] e [7] del *paragrafo 2*)
- 2) la visualizzazione grafica dei dati reali (gli errori) e dei dati interpolati in funzione delle stime
- 3) l’eliminazione di eventuali outlier
- 4) la creazione di modelli utente e la possibilità di eseguire i passi 2 e 3 anche per questi

L’utente può scegliere tra tre selezioni iniziali:

- 1) Modello (1)
- 2) Modello (2)
- 3) Modello Utente (attualmente non implementato, *rappresenta una possibile estensione futura*)

- Modello (1) -

I ciclo :

- Punto1 (programma1.sas)

L’utente seleziona la cartella di input.

Data-set di input WTOTALE

(se non esiste nella cartella di input → messaggio di errore)

Dal punto (b) in poi, se X e Y sono parametriche, il programma è lo stesso utile per il Modello Utente

- (a) Il programma trasforma le variabili secondo il modello stabilito
- (b) Si calcola il modello1 di interpolazione $Y=A+BX$
- (c) Si calcola l’ R^2
- (d) Si memorizzano i parametri del modello (A,B, R^2) in un *data-set* (MODEL)
- (e) Si calcolano i residui standardizzati
- (f) Si memorizzano le informazioni del primo ciclo in un *data-set* utile per i plot (INFO)
- (g) Si calcola la tabella di frequenza
- (h) Si memorizzano i dati della tabella di frequenza (FREQUE)

Data-set di output MODEL, INFO, FREQUE

MODEL,
INFO e
FREQUE

- Punto2

Per studiare il problema l’utente può scegliere tra diversi tipi di analisi

♦ Dal data-set INFO: 2 grafici

- a) X1 con Y1 e Y1INT (ovvero $\log(\text{stima})$ e $\log(\text{errel}^2)$)
- b) X1 con il residuo standardizzato

I grafici si possono effettuare per uno o più domstima e si devono memorizzare.

- ♦ **Dal data-set MODEL: visualizzazione parametri A B e R2 di modelli**
- ♦ **Dal data-set FREQUE: visualizzazione della tabella di frequenze sui residui**

L'utente non elimina i record corrispondenti agli outlier in modo definitivo in *WTOTALE*, ma li evidenzia con un flag.

La funzione di eliminazione degli outlier può effettuarsi con i residui standardizzati: (ad esempio: if abs(res_sta) < valore utente (da INFO))

Viene creato il *data-set* OUTLIER con le righe “eliminate” (segnalate con il flag) in *WTOTALE* definitivo, che rimane inalterato.

II ciclo

L'input è a questo punto il *data-set* *WTOTALE* che ha gli outlier segnalati.

Punto 1: creazione di nuovi MODEL, INFO, FREQUE senza considerare i valori segnalati come outlier

- **Modello (2)** -

I ciclo :

- **Punto1** (programma2.sas)

L'utente seleziona la cartella di input.

Dal punto (b) in poi,
se X e Y e Z sono
parametriche, il
programma è lo stesso
utile per il Modello
Utente

Data-set di input WTOTALE

(se non esiste nella cartella di input → messaggio di errore)

- (a) Il programma trasforma le variabili secondo il modello stabilito
- (b) Si calcola il modello2 di interpolazione $Z=AX+BY+C$
- (c) Si calcola l'R2
- (d) Si memorizzano i parametri del modello (A,B,C, R2) in un *data-set* (MODEL2)
- (e) Si calcolano i residui standardizzati
- (f) Si memorizzano le informazioni del primo ciclo in un *data-set* utile per i plot (INFO2)
- (g) Si calcola la tabella di frequenza
- (h) Si memorizzano i dati della tabella di frequenza (FREQUE2)

MODEL2,
INFO2 e
FREQUE2

Data-set di output MODEL2, INFO2, FREQUE2

- **Punto2**

Per studiare il problema l'utente può scegliere tra diversi tipi di analisi

- ♦ **Dal data-set INFO2: 2 grafici**

- I grafici si possono effettuare per uno o più domstima e si devono memorizzare:

- L'utente non elimina i record corrispondenti agli outlier in modo definitivo in *WTOTALE*, ma li evidenzia con un flag.

Viene creato il *data-set* OUTLIER con le righe “*eliminate*” (segnalate con il flag) in *WTOTALE* definitivo, che rimane inalterato.

L'input è a questo punto il *data-set* *WTOTALE* che ha gli outlier segnalati.

Punto 2: creazione di un nuovo OUTLIER

```
graph TD
    subgraph Parte1 [Parte 1]
        direction TB
        subgraph TopRow
            direction LR
            DB1[(errori.Model)]
            DB2[(Errori.Info)]
            DB3[(Errori.frequ)]
        end
        F1((1)  
Esegui modello (1o2)  
considerando o meno  
gli outlier (Flag=0/1))
        F2((2)  
Ripristino o definisco  
outlier, sulla base di  
un intervallo  
(Flag/noflag))
        F3((3)  
Effettuo analisi  
(grafici, tabelle,  
parametri))
        Box1[Cataloghi?]
        
        F1 --> DB1
        F1 --> DB2
        F1 --> DB3
        DB1 --> F2
        DB2 --> F2
        DB3 --> F2
        F2 --> F3
        F3 -.-> Box1
    end
    Parte1 --> Parte2[Parte 2]
    Parte2 -.-> DB4[(Errori.interp)]
```

FASE 1: interfacce da implementare

Maschera di partenza

Analisi modelli

(1) Modelli

(2) Analisi esplorativa

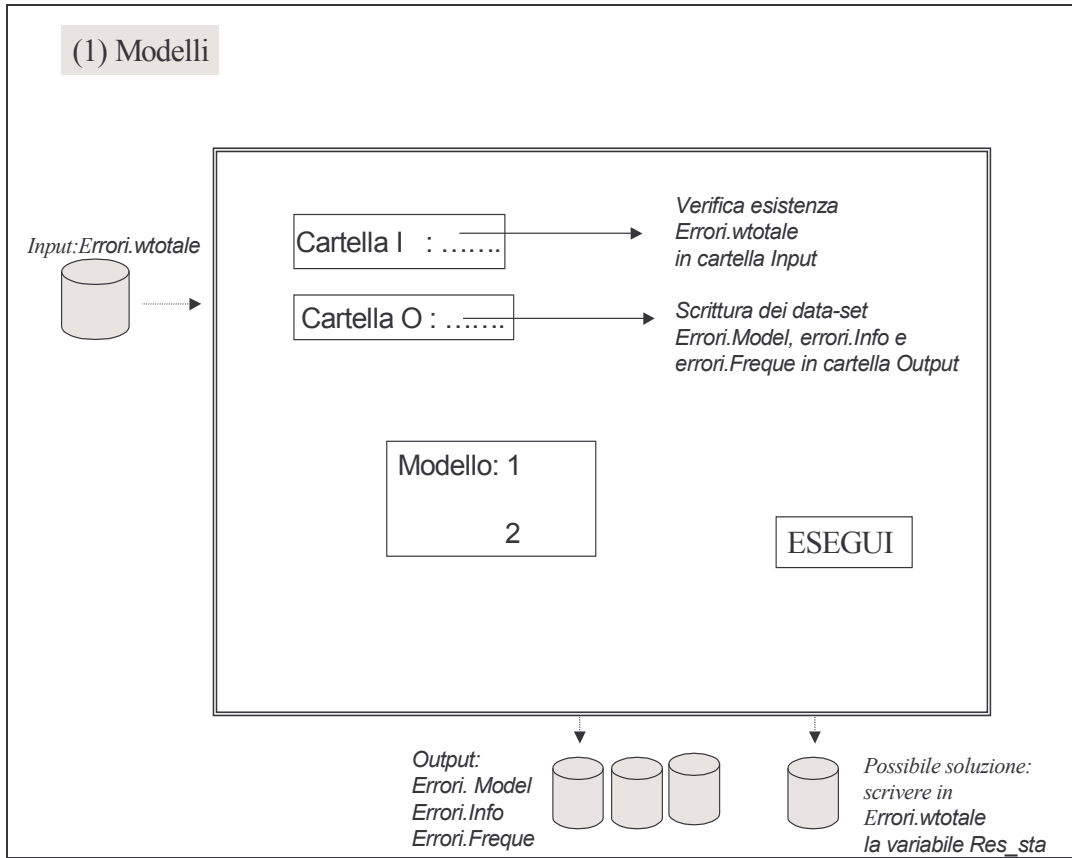
(3) Outlier

(1) Modelli : richiama la Maschera 1

(2) Analisi esplorativa : richiama la Maschera 2

(3) Outlier : richiama la Maschera 3

Maschera 1



INPUT: Errore.wtotale

OUTPUT: a) Errori.model, Errori.Info, Errori.Freque
b) Scrittura in Totale di Res sta

Il modello viene calcolato sui dati di Errori.WTOTALE, escludendo i dati considerati outlier, ovvero quelli che – tramite la Voce 3 – presentano il flag =1.

(un utente può sempre ripristinare il flag successivamente, riponendolo a 0, utilizzando la Voce 3).

Modifiche da fare al calcolo del modello (programma.sas):

Creazione del *data-set* UNO da Errori.wtotale

a) In UNO portare le variabili chiave che definiscono il singolo record, ovvero fare la keep delle variabili xvariabil e xsottocla (invece di variabil e sottocla di Totale), di modalita e modscl.

b) Sempre in UNO, selezionare solo i record dove il flag è diverso da 1 (ovvero, oltre a non considerare quelli che attualmente non considera - con varianza negativa o nulla - eliminare anche gli outlier).

Attenzione: le variabili chiave dovranno essere memorizzate anche in Errori.INFO.

Tramite queste, definire se utile e/o necessario portare la variabile `res_sta` (memorizzata in `errori.info`) anche in `TOTALE` (questa soluzione faciliterebbe la Voce (3), che legge solo `wtotale`)

Da questa maschera dovrà essere consentito passare anche alle funzioni (2) e (3) senza tornare al menu principale: attivare la cartella di input della Voce (2) con quella di output di questa Voce e la cartella di input della Voce (3) con quella di input di questa Voce.

Maschera 2

(2) Analisi esplorativa

Verifica esistenza errori.Model,
error.Info e errori.Freque in cartella
Input

Input
Errori.Model
Errori.Info
Errori.Freque

Cartella: _____

dominio di stima

tutti ☒
da: ____ a: ____

Parametri

☒

Grafici

☒

Tabella

☒

ESEGUI

NO Output (ci sarebbe se si memorizzasse il catalogo dei grafici)

INPUT Errori.Model, Errori.Info, Errori.Freque OUTPUT: nessuno (verificare!)

Funzione solo di esplorazione; prende in input i *data-set* Errori.Model, Errori.Info ed Errori.Freque, scritti tramite la Voce (1).

L'utente può richiederla quando vuole e la fa prendendo come input una cartella su cui sono stati memorizzati i dati di output della Voce (1).

Non si attiva solo dopo l'esecuzione del modello, ma è l'utente che sceglie la cartella di input dei dati da analizzare.

Può analizzare anche un solo dominio di stima pianificato, o un gruppo di domini, o tutti. Legge i *data-set* che hanno tutti i domini, ma elabora solo quelli selezionati.

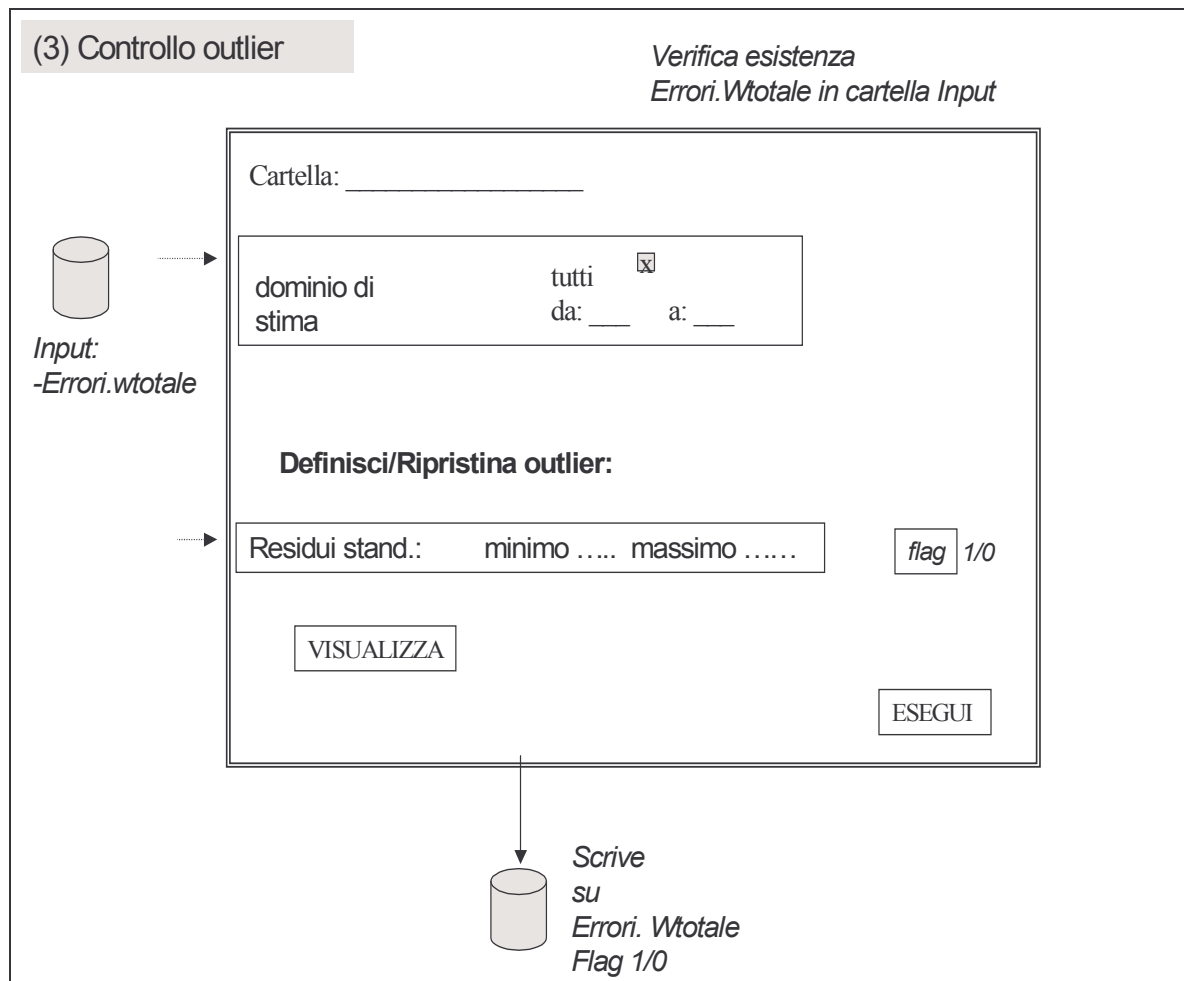
Se l'utente chiede "Parametri, verificare l'esistenza di Errori.Model;

se vuole "Grafici" verificare l'esistenza di Errori.Info;

se vuole la Tabella verificare l'esistenza di Errori.Freque

Da questa maschera dovrà essere consentito passare anche alle funzioni (1) e (3):
non attivare in automatico le cartelle

Maschera 3



Funzione di definizione e ripristino outliers.

INPUT e OUTPUT : Errori.wtotale

L'utente può richiederla quando vuole e la fa prendendo come input la cartella dove è Errori.wtotale. Viene variato il valore della variabile outlier da 0 ad 1 o viceversa, sulla base della selezione del dominio di stima pianificato (uno o un gruppo o tutti) e dei valori che assume la variabile Res-sta. Comunque, scrive sul *data-set wtotale* che ha tutti i domini.

Il bottone Visualizza consente di visualizzare le osservazioni del *data-set wtotale* corrispondenti ai domini di stima e ai residui standardizzati scelti.

Con un unico bottone per il flag posso definire un outlier o ripristinare un dato, scegliendo 0 o 1.

Da questa maschera dovrà essere consentito passare anche alle funzioni (1) e (2): attivare la cartella di input della Voce (1) con quella di input di questa Voce e non attivare in automatico la cartella della Voce (2).

FASE II - Analisi relativa alla seconda fase progettuale

Parte A (stampa 5b)

La funzione si attiva a partire dal menu, a partire dalla voce (1)

-Analisi Modelli/

Modelli

Stampe/

Stampa(5b o 7b)

Stampa Personalizzata

Analisi Esplorativa

Outlier

La funzione deve riprodurre la stampa 5b e 7b già prodotta da Geneseees, contenente i valori interpolati degli errori di campionamento per dominio di stima pianificato.

Il calcolo tiene anche conto dei nuovi valori dei parametri di regressione ottenuti grazie all'eliminazione degli outlier nel *data-set* di input (Modelli I).

Parte B

La funzione si attiva a partire dal menu

-Analisi Modelli/

Modelli

Stampe/

Stampa(5b o 7b)

Stampa Personalizzata

Analisi Esplorativa

Outlier

Si da scelta all'utente di partire da un *data-set* in cui memorizza i valori assoluti delle stime, per ciascun dominio di stima pianificato prescelto.

L'utente crea perciò un *data-set* a cui associare i parametri di regressione ed applicare l'algoritmo noto. E' obbligatorio il nome del *data-set*: *nuove_stime*

Ipotesi progettazione futura

- Modello Utente -

(Funzione generale che permetterebbe di costruire anche i modelli (1) e (2))

I ciclo :

L'utente seleziona

- 1) la cartella di input
- 2) il *data-set* di input

Scelto il modello l'utente ha a disposizione una **funzione di trasformazione delle variabili**; deve essere possibile selezionare una o due variabili X e Y generiche e ottenere:

- selezione di una variabile: $X^{**0.5}$ X^2 X^3 $1/X$ $\log(X)$ $\exp(X)$
- selezione di 2 variabili: $X*Y$ X/Y $X-Y$ $X+Y$

Deve essere possibile ottenere ad esempio, da X e Y :

X^2/Y ovvero

- a) *step 1: da X si ottiene e memorizza X^2*
- b) *step 2: si moltiplica X^2 per Y*



La selezione delle variabili per il modello (trasformate o originarie) consente di scegliere fino a tre variabili.

Da qui come Modello (1) o (2) → program1.sas o programma2.sas dal punto (b)

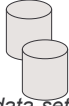
se l'utente ha scelto 2 variabili → $Y=AX+B$

se l'utente ha scelto 3 variabili → $Z=AX+BY+C$

- L'utente potrebbe inserire percentuali di stima da interfaccia, invece che tramite data-set come ora previsto

Parte II

Input
Errori. Model


data-set
utente

Cartella: _____

Nome data-set totali: _____

(Verifica esistenza
errori. Model e data-set
dei totali per ciascun
dominio di stima)

dominio di stima

tutti ☒
da: ____ a: ____

Perc.Stime

Nome file Output: _____

(in automatico scrive sia un data-set
che un excel)

Bibliografia

Deville J. C., Särndal C. E. (1992), Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, vol. 87, pp. 367-382.

De Vitiis, C., Pagliuca, D., 2003, *La presentazione sintetica degli errori campionari e l'analisi grafica degli outlier nel software Genesees*, Atti del Convegno Intermedio "Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia" della Società Italiana di Statistica (su CD-ROM).

Falorsi, P.D., Falorsi, S., 1995, *Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese*, Rapporto di ricerca CON.PRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, n. 13.

Falorsi, P.D., Falorsi, S., 1997, *The Italian Generalized Package for Weighting Persons and Families: Some Experimental Results with Different Non-Response Models*, Statistics in Transitions Journal of the Polish Statistical Association, vol. 3, n. 2.

Falorsi, P. D., Falorsi S., 1998, *The Italian generalized estimation package: some experimental results for estimation on households suveys with different non response mechanism*, Quaderni di Ricerca, ISTAT, n.4, pp.63-94.

Falorsi, S., Rinaldelli, C., 1998, *Un Software generalizzato per il calcolo delle stime e degli errori di campionamento*, Statistica Applicata, vol. 10, n. 2 , pp. 217-234.

Falorsi, S., Pagliuca, D., Scepi, G., 1999, *Generalised Software for Sampling Errors – GSSE*, Proceedings of the Seminar on Exchange of Technology and Know-How (ETK 99), held in Prague, Czech Republic on the 13-15 October 1999, pp. 169-175.

Falorsi, S., Pagliuca, D., Scepi, G., 2000, *Generalised Software for Sampling Errors – GSSE*, Research in Official Statistics - ROS, vol. 3, n. 2, pp. 89-108.

Pagliuca, D., Righi, P., 2002, *Genesees v1.0*, Proceedings of the Conference CompStat 2002 – Short Communications and Posters, Berlin August 24-th to August 28th 2002 (disponibile su CD-ROM).

Pagliuca D. (a cura di), 2004, *Funzioni di Genesees* (1.“Riponderazione” e 2.“Stime ed Errori campionari”), Manuali Utente e Aspetti Metodologici, disponibili via internet (per utenti esterni all'Istat): <http://www.istat.it/Metodologi/index.htm> (selezionare “Metodi e Software per indagini statistiche”) oppure via intranet (per utenti Istat): <http://intranet/> (selezionare: “Prodotti e Applicazioni on-line. Software Generalizzati” e da qui selezionare “MTS-F: Software Generalizzati per la Produzione Statistica (Area Download e Informazioni)”.

Pagliuca D. (a cura di), 2004b, *Manuale Utente della funzione di Stime ed Errori campionari del software Genesees v.3.0*, Collana Tecniche e Strumenti, ISTAT, n.2.

Russo A., 1987, *Sulla Presentazione degli Errori di Campionamento mediante Modelli. Il Metodo dei Modelli Regressivi*, Quaderni di Discussione, ISTAT, n. 87, 04.

Verma, V., Scott, C., O'Muircheartaigh, C., 1980, *Sample Designs and Sampling Errors of the World Fertility Survey*, Journal of the Royal Statistical Society A, vol. 143, Part. 4, pp. 431-473.

Verma, V., 1982, *The Estimation and Presentation of Sampling Errors*, Technical Bulletins, World Fertility Survey, New York.

Wolter, K. M., 1985 *Introduction to variance estimation*. Springer-Verlag. New York.