Alcuni metodi di imputazione delle mancate risposte parziali per dati quantitativi. Il software *QUIS*.

Autore

Ugo Guarnera (*)

(*) ISTAT - Servizio MTS
e-mail: guarnera@istat.it

Sommario

Questo contributo contiene una descrizione di alcuni metodi di trattamento delle mancate risposte parziali per dati quantitativi che sono stati implementati nel software *QUIS*. Tra questi, i primi tre sono metodi di imputazione singola, mentre l'ultimo è l'*imputazione multipla* proposta da Rubin. Ad eccezione del donatore di minima distanza, tutti i metodi descritti si basano sull'ipotesi di normalità dei dati. Si è cercato pertanto di illustrare vantaggi e svantaggi derivanti da tale assunzione e, più in generale, dall'utilizzo di un modello parametrico esplicito. Si sono inoltre messe in rilievo le caratteristiche dei diversi metodi in termini di robustezza rispetto all'allontanamento dalla normalità. L'ultima parte del lavoro è dedicata ad una descrizione delle modalità pratiche di utilizzo del software *QUIS*.

Abstract

This paper contains the description of some methods for the treatment of partial non-response in quantitative data. The first three methods are single imputation methods, whereas the last one is Rubin *multiple imputation*. Apart from the Nearest Neighbour Donor, all of the methods rely on the normality assumption. Thus we have tried to highlight advantages and drawbacks related to the normality hypothesis and, more in general, to the use of some parametric explicit model. Also, the features of the different methods have been described in term of robustness with respect to departures from normality. Finally in the last section of the paper some technical detail is given on the software *QUIS*.

Indice

1. Introduzione

L'analisi e il trattamento dei dati incompleti costituisce un problema di grande interesse nell'ambito della ricerca sulla qualità dei dati in un'indagine statistica. L'incompletezza dei dati è riconducibile essenzialmente a due cause principali:

a) la *mancata risposta totale* (*unit non-response)* che si riferisce alla totale assenza di informazione su talune unità del campione;

b) la *mancata risposta parziale* (*MRP* nel seguito) che ha luogo quando, per una data unità, non sono disponibili i valori di alcune variabili. In questo lavoro ci occuperemo soltanto del secondo aspetto.

Le strategie per affrontare il problema della MRP sono essenzialmente quattro:

1) fornire all'utente l'insieme di dati incompleto
2) scartare tutte le unità con valori mancanti
3) usare tecniche di riponderazione
4) produrre valori "artificiali" con i quali rimpiazzare i dati mancanti.

1) La prima opzione, che potrebbe essere considerata la più "onesta", delega essenzialmente all'utente la decisione sull'approccio da seguire nel trattamento delle *MRP*, consentendogli di adattare la strategia alle sue specifiche esigenze. D'altro canto questa scelta esclude la possibilità di utilizzare metodi e software standard (sviluppati cioè per data-set completi) presupponendo una dimestichezza dell'analista con strumenti inferenziali anche sofisticati, specificamente sviluppati per il trattamento dei dati incompleti.

2) La seconda strategia, adottata frequentemente, consiste nel basare le analisi di interesse sulle sole osservazioni complete. Sfortunatamente questo approccio produce serie distorsioni nelle stime a meno che i valori mancanti non costituiscano un campione casuale dell'intero insieme di dati (meccanismo di mancata risposta *Missing Completely at Random* o *MCAR*) o, in altre parole, la non-risposta sia indipendente da tutte le variabili di interesse. Inoltre, anche qualora il meccanismo di mancata risposta sia *MCAR*, il non utilizzo di tutta l'informazione disponibile comporta una diminuzione della precisione delle stime (aumento degli errori standard).

3) La terza strategia, usualmente adottata per la mancata risposta totale, si basa sull'utilizzo di variabili ausiliarie mediante le quali vengono definite celle di riponderazione per compensare la

mancata risposta. Sebbene semplice in linea di principio, questo approccio presenta il non trascurabile svantaggio di richiedere la definizione di un diverso insieme di pesi per ogni variabile soggetta a *MRP*, rendendo peraltro problematica la stima di parametri legati alla relazioni di interdipendenza tra le variabili.

4) Il quarto approccio al trattamento della *MRP*, cui ci si riferisce comunemente con il nome di *imputazione* è probabilmente il più popolare. La sostituzione dei valori mancanti con valori prodotti "artificialmente" consente di riprodurre un data-set completo sul quale, utilizzando strumenti standard, diverse analisi possono essere effettuate in modo consistente. Uno degli aspetti più critici di questo approccio consiste tuttavia nel fatto che, una volta prodotto un data-set "rettangolare" completo, i valori imputati tendono ad essere considerati come valori effettivamente osservati. La componente della variabilità delle stime associata alla non-risposta e al suo trattamento viene in tal modo trascurata, con un conseguente effetto di sottostima degli errori standard.

Diversi approcci sono stati proposti in letteratura per affrontare il problema dell'impatto della mancata risposta sulla precisione delle stime nelle indagini statistiche. Il più popolare è probabilmente quello dell'*Imputazione Multipla*, introdotta da Rubin (1978, 1987) che consiste essenzialmente nella ripetizione del processo di imputazione *m* volte e, conseguentemente, nella generazione di un insieme di *m* data-set completi. Rubin (1987) ha dimostrato che, purchè il processo di imputazione soddisfi alcuni specifici requisiti, inferenze valide possono essere ottenute svolgendo analisi indipendenti sui singoli data-set completati e combinando opportunamente i risultati. Mediante questo metodo, che trova la sua collocazione più naturale in un contesto Bayesiano, si perviene a stime corrette dei parametri e dei relativi errori standard. Di contro, l'interpretabilità da un punto di vista frequentista richiede che siano soddisfatte delle ipotesi sulle caratteristiche del processo di imputazione difficilmente verificabili in generale, ed inoltre la necessità di gestire un certo numero di data-set per ogni insieme di dati incompleto da trattare, può costituire un limite da un punto di vista operativo specialmente in presenza di grosse moli di dati.

Un modo tradizionale di classificare i metodi di imputazione è quello di distinguerli in *parametrici* e *non parametrici* a seconda che essi facciano uso o meno di un modello esplicito. Alla prima categoria appartiene, ad esempio, l'imputazione mediante regressione, mentre alla seconda l'imputazione con donatore casuale o con donatore di minima distanza. I metodi di imputazione più diffusi, in ogni caso, si basano sull'assunzione che il meccanismo di mancata risposta sia almeno *Missing at Random* (*MAR*), cioè la mancata risposta, condizionatamente ai dati osservati, non dipenda dai valori mancanti (Little e Rubin, 2002); sotto questa ipotesi, infatti, è possibile effettuare

inferenze valide sui parametri di interesse senza specificare un modello per la non risposta. E' evidente che la validità di tale assunzione non è verificabile a meno che non si disponga di opportuni *training set* costruiti mediante ritorno alle unità non rispondenti.

In questo lavoro verranno decritti sinteticamente alcuni metodi per il trattamento delle mancate risposte parziali per dati quantitativi che sono stati implementati dal Servizio Studi dell'ISTAT nel software applicativo ***QUIS*** (***Quick Imputation System***) in ambiente *SAS*. Nella sez. 2. si illustreranno tre metodi di imputazione singola:

1) la *regressione basata sull'algoritmo EM*

2) *l'imputazione da donatore di minima distanza*

3) il *predictive mean matching multivariato.*

In particolare, per quest'ultimo metodo, per il quale non erano disponibili software applicativi nell'ambito di package statistici di uso comune, ci si soffermerà più in dettaglio sugli aspetti computazionali. Nella sez. 3 verranno fatti alcuni cenni all'*imputazione multipla* di Rubin che in *QUIS* è disponibile sotto forma di una *macro SAS* in cui è implementato l'algoritmo generale illustrato da Schafer (1997). La sez. 5. infine, contiene alcune conclusioni generali ed indicazioni di carattere pratico.

2. Metodi di imputazione singola

Nella descrizione che segue indicheremo con $\boldsymbol{Y}$ la matrice $n \times p$ dei dati (osservazioni x variabili) che si avrebbe in assenza di mancata risposta e con $y_{ij} \equiv \{\boldsymbol{Y}\}_{ij}$ il valore della variabile Y_j sulla i-esima unità. Indicando con $\boldsymbol{Y}_{obs}$ l'insieme dei valori osservati e con $\boldsymbol{Y}_{mis}$ l'insieme dei valori mancanti, potremo pertanto scrivere: $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$. Denoteremo inoltre con $\boldsymbol{R}$ la matrice di variabili indicatrici definita da:

$$\boldsymbol{R}_{ij} = \begin{cases} 1 & \text{se } y_{ij} \text{ è osservato} \\ 0 & \text{se } y_{ij} \text{ è mancante} \end{cases}$$

Gli elementi della matrice $\boldsymbol{R}$ vanno intesi come variabili aleatorie; il meccanismo di mancata risposta è definito attraverso la specificazione di un modello per $\boldsymbol{R}$: $f(\boldsymbol{R} \mid \boldsymbol{Y}, \xi)$ dove ξ è un insieme di parametri.

2.1 Verosimiglianza per dati incompleti

Molti metodi di trattamento dei dati incompleti si basano sull'assunzione di un modello esplicito e sulla stima di massima verosimiglianza (*MLE*) dei suoi parametri. In linea di principio non ci sono differenze tra i metodi di stima per dati completi e quelli per dati incompleti. In entrambi i casi occorre massimizzare una funzione di verosimiglianza rispetto ai parametri del modello. Tuttavia, nel caso di dati incompleti è necessario fare delle ipotesi sul meccanismo di mancata risposta affinché sia lecito effettuare inferenze sui parametri del modello assunto per i dati senza specificare esplicitamente un modello di non-risposta. Effettivamente l'ipotesi alla base dei metodi più comuni di imputazione delle *MRP* è, come già accennato nell'introduzione, l'ipotesi *MAR*. Nell'ambito di un approccio parametrico, è facile rendersi conto, esplicitando le distribuzioni di probabilità in gioco, della necessità di tale ipotesi ai fini della *ignorabilità* del meccanismo di mancata risposta. In accordo con le notazioni già introdotte, sia $f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} \mid \theta)$ la distribuzione di probabilità congiunta di $\boldsymbol{Y}_{obs}$ e $\boldsymbol{Y}_{mis}$ (θ è un insieme di parametri). Chiameremo *Verosimiglianza a Dati Osservati* una funzione $L(\theta \mid \boldsymbol{Y}_{obs})$ di θ, proporzionale alla densità marginale $f(\boldsymbol{Y}_{obs} \mid \theta) \equiv \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} \mid \theta) d\boldsymbol{Y}_{mis}$ ottenuta integrando la densità $f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} \mid \theta)$ rispetto ai valori mancanti. Un meccanismo di mancata risposta è ignorabile se è possibile effettuare inferenze valide sui parametri θ sulla base (esclusivamente) della verosimiglianza a dati osservati $L(\theta \mid \boldsymbol{Y}_{obs})$. Per comprendere come questa definizione richieda che il meccanismo di mancata risposta sia *MAR*, si consideri la distribuzione congiunta dei dati $\boldsymbol{Y}$ e delle variabili indicatrici $\boldsymbol{R}$ definite sopra:

$$(1) \qquad f(\boldsymbol{Y}, \boldsymbol{R} \mid \theta, \xi) = f(\boldsymbol{Y} \mid \theta) f(\boldsymbol{R} \mid \boldsymbol{Y}, \xi)$$

I dati effettivamente osservati consistono in realtà nei valori delle variabili $\boldsymbol{Y}_{obs}$, $\boldsymbol{R}$ cosicchè in generale, le inferenze sui parametri (θ, ξ) dovrebbero basarsi sulla verosimiglianza:

$$(2) \qquad L(\theta, \xi \mid \boldsymbol{Y}_{obs}, \boldsymbol{R}) \propto \int f(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis} \mid \theta) f(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \xi) d\boldsymbol{Y}_{mis}$$

ottenuta integrando la (1) sui valori mancanti. Si vede dunque che, se la mancata risposta $\boldsymbol{R}$, non dipende dai valori mancanti $\boldsymbol{Y}_{mis}$ condizionatamente a quelli osservati $\boldsymbol{Y}_{obs}$ cioè se :

$$(3) \qquad f(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}, \xi) = f(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \xi)$$

allora dalla (2) si ottiene:

$$(4) \qquad L(\theta, \xi \mid \boldsymbol{Y}_{obs}) \propto f(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \xi) f(\boldsymbol{Y}_{obs} \mid \theta) \propto L(\theta \mid \boldsymbol{Y}_{obs})$$

cioè i valori di θ che massimizzano la (2) sono gli stessi che massimizzano $L(\theta \mid \boldsymbol{Y}_{obs})$ ossia le inferenze su θ possono basarsi sulla sola verosimiglianza a dati osservati ignorando il meccanismo di mancata risposta. E' importante sottolineare che questo risultato dipende strettamente dall'assunzione (3) che, in effetti, è la definizione di *MAR*.

2.2 L'algoritmo EM

Si è detto che l'ignorabilità del meccanismo di mancata risposta è alla base di tutte le procedure che stimano i parametri θ di una distribuzione di probabilità utilizzando soltanto la verosimiglianza a dati osservati $L(\theta \mid \boldsymbol{Y}_{obs})$. Tuttavia, anche quando si possa ritenere ragionevole tale assunzione, la massimizzazione di $L(\theta \mid \boldsymbol{Y}_{obs})$ si presenta in generale come un problema di non facile soluzione. A meno che i valori mancanti non si presentino in particolari configurazioni (come nel caso di *pattern monotoni* (Little e Rubin, 2002) infatti, la verosimiglianza a dati osservati è una funzione complicata dei parametri, e raramente le equazioni di massima verosimiglianza possono essere risolte in modo analitico. L'algoritmo *EM* (*Expectation-Maximization*), è un metodo che consente, attraverso un procedimento iterativo, di effettuare le stime di massima verosimiglianza dei parametri in presenza di dati incompleti, riconducendo il problema ad un problema standard di stima per dati completi. Partendo da una stima iniziale $\theta^{(0)}$ dei parametri (*starting guess*), l'algoritmo consiste, ad ogni iterazione t del procedimento, nella applicazione dei seguenti due passi:

i) ***E-step*** calcolo del valore atteso $H(\theta,\theta^{(t)})$ della verosimiglianza $L(\theta \mid \mathbf{Y}_{obs})$ rispetto alla distribuzione dei dati mancanti condizionatamente ai dati osservati e alle stime correnti dei parametri $\theta^{(t)}$:

$$H(\theta,\theta^{(t)}) \equiv \int L(\theta \mid \mathbf{Y}_{obs},\mathbf{Y}_{mis})\, f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs},\theta^{(t)})\, d\mathbf{Y}_{mis}$$

ii) ***M-step*** massimizzazione di $H(\theta,\theta^{(t)})$ rispetto a θ.

L'algoritmo genera una successione $\{\theta^{(t)}\}_{t=1,2,..}$ che, sotto alcune ipotesi di regolarità, si dimostra (Dempster et al., 1977) convergere alla stima di massima verosimiglianza di θ.

2.3 Algoritmo EM per la famiglia esponenziale. Distribuzione normale multivariata

Quando il modello di probabilità assunto per i dati completi appartiene alla *famiglia esponenziale regolare* (Mardia, 1979), la verosimiglianza a dati completi, basata su *n* osservazioni, può essere scritta come una funzione lineare di statistiche sufficienti per i parametri del modello (Schafer, 1997):

$$(5) \qquad L(\theta \mid \mathbf{Y}) = \zeta(\theta)^T \mathbf{T}(\mathbf{Y}) + n \cdot \lambda(\theta) + c$$

dove $\zeta(\theta)$ sono i parametri in forma canonica, $\mathbf{T}(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}),\ldots, T_r(\mathbf{Y}))^T$ un insieme di *r* statistiche sufficienti, $\lambda(\theta)$ un opportuna funzione dei parametri e *c* una costante. Inoltre ciascuna statistica $T_l(\mathbf{Y})$ è una funzione additiva del tipo: $T_l(\mathbf{Y}) = \sum_{i=1}^{n} h_l(y_i)$, $(l = 1,..r)$ dove le h_l sono un opportuno insieme di funzioni. Essendo la (5) lineare nelle statistiche sufficienti $T_l(\mathbf{Y})$, il passo di *Expectation* dell'algoritmo *EM* si riduce al calcolo dei valori attesi $E(T_l(\mathbf{Y}) \mid \mathbf{Y}_{obs},\theta)$. Inoltre, poiché le stime di massima verosimiglianza per dati completi si ottengono come soluzione dell'*equazione dei momenti*:

(6) $E\,(\boldsymbol{T}(\boldsymbol{Y}) \mid \boldsymbol{Y}_{obs}, \theta) = \boldsymbol{t}$

dove $\boldsymbol{t}$ è la realizzazione osservata di $\boldsymbol{T}$ (Cox e Hinkley, 1974), nel caso di dati incompleti le *MLE* possono essere effettuate (*M-step*) risolvendo una equazione analoga alla (6), in cui la statistica $\boldsymbol{t}$ è sostituita dal suo valore atteso condizionato ai valori osservati $\boldsymbol{Y}_{obs}$ calcolato nell'*E-step*.

Nel caso della distribuzione normale, l'equazione (6) può essere risolta in forma chiusa fornendo una formula ricorsiva esplicita per il calcolo delle *MLE*. In pratica, occorre calcolare quantità del tipo: $E\,(y_{ij} \mid \boldsymbol{y}_{i(obs)}, \theta)$ o $E\,(y_{ij}\, y_{ik} \mid \boldsymbol{y}_{i(obs)}, \theta)$ con $i=1,...,n$ e $j,k=1,\ldots,p$, in cui $\boldsymbol{y}_{i(obs)}$ indica l'insieme delle variabili osservate nell'unità *i*. Ciò può essere fatto agevolmente con l'ausilio della famiglia di operatori di *sweep* (*sweep operator*), che essenzialmente consiste in un insieme di trasformazioni che consentono di esprimere i parametri delle distribuzioni condizionate di una normale multivariata in funzione dei parametri della distribuzione congiunta. Una descrizione dettagliata delle proprietà dello sweep operator e del suo utilizzo nell'algoritmo *EM* per dati normali si può trovare in Schafer (1997).

2.4 Algoritmo EM ed imputazione

Nel contesto dell'imputazione delle *MRP* l'algoritmo *EM* è utilizzato per stimare i parametri del modello mediante il quale vengono imputati i dati mancanti. Una volta che le *MLE* θ^* sono state calcolate, l'imputazione può essere svolta essenzialmente in due modi diversi: i dati mancanti possono essere imputati con i valori attesi condizionati $E\,(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \theta^*)$ oppure generati dalla distribuzione condizionata $P\,(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \theta^*)$. Il metodo migliore dipende dagli specifici obiettivi di ricerca. In particolare, date le proprietà di ottimalità del primo metodo come metodo di previsione, il suo utilizzo è preferibile se si vogliono stimare quantità univariate lineari nei dati come medie o totali; al contrario, qualora si sia interessati alla stima di parametri distribuzionali legati alle relazioni di interdipendenza tra le variabili, l'imputazione mediante valori attesi condizionati potrebbe comportare forti distorsioni, e il secondo metodo è preferibile (Kalton e Kasprzyk, 1986). Per l'implementazione pratica dei due metodi nel caso di dati normali, è conveniente avvalersi ancora dello *sweep operator* mediante il quale, a partire dalle stime $\theta^* \equiv (\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, possono essere calcolate le stime (di massima verosimiglianza) dei parametri delle distribuzioni condizionate corrispondenti ai vari *pattern* di valori mancanti.

Vale la pena osservare che, poiché i valori attesi condizionati $E\,(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs}, \theta^*)$ nel caso di dati normali, sono come è noto funzioni lineari di $\boldsymbol{Y}_{obs}$, i parametri delle distribuzioni condizionate

P ($\boldsymbol{Y}_{mis}$ | $\boldsymbol{Y}_{obs}$, θ^*) potrebbero essere correttamente stimati mediante regressioni lineari sulla base del sottoinsieme dei dati completamente osservati. Un tale procedimento tuttavia appare più inefficiente comportando la stima di un insieme di parametri diverso per ogni configurazione di valori mancanti. Inoltre, nella procedura di stima, non verrebbe utilizzata tutta l'informazione disponibile.

E' utile sottolineare infine che il procedimento di stima dei parametri di una distribuzione di probabilità in presenza di dati incompleti mediante l'algoritmo *EM* non è equivalente al procedimento iterativo consistente nell'imputazione dei dati mancanti mediante regressioni basate sulle stime correnti e ricalcolo delle stime sulla base del data-set completo così ottenuto. Ciò è dovuto al fatto che, in generale, le statistiche sufficienti di cui si calcola il valore atteso nell'*E-step*, non sono funzioni lineari dei dati.

2.5 Metodi non parametrici. Il donatore di minima distanza

Il trattamento dei dati incompleti mediante l'assunzione di un modello parametrico presenta alcuni importanti vantaggi come la possibilità di formalizzare le relazioni di interdipendenza delle variabili di analisi mediante espressioni matematiche esplicite (ad es. relazioni lineari) e la possibilità di affrontare con una certa generalità il problema della variabilità delle stime associata alla non-risposta (*imputazione multipla*). Tuttavia la validità di tutte le analisi inferenziali eseguite nell'ambito di tale approccio dipende in modo cruciale dalla specificazione del modello. In linea di principio sarebbe pertanto auspicabile che ogni strategia adottata in tale contesto includesse un'analisi di sensitività rispetto alla "mispecificazione" del modello. Nel caso di variabili quantitative, la distribuzione multivariata di gran lunga più utilizzata per la modellizzazione dei dati è la distribuzione normale. Questa infatti possiede la comoda proprietà di avere distribuzioni condizionate che sono funzioni lineari delle variabili condizionanti, cosicché l'imputazione delle mancate risposte parziali può essere effettuata agevolmente mediante regressione lineare. Inoltre, per la distribuzione normale sono stati sviluppati specifici strumenti inferenziali per dati incompleti (come ad esempio alcune implementazioni dell'algoritmo *EM*) di (relativamente) facile utilizzo. L'assunzione di normalità tuttavia, raramente risulta adeguata per dati provenienti da indagini economiche: frequentemente infatti, i dati presentano distribuzioni empiriche asimmetriche o a carattere semi-continuo (cioè con concentrazioni di osservazioni su taluni valori, tipicamente "zeri"). Mentre il primo problema può talvolta essere trattato con opportune trasformazioni di variabili (ad es. trasformazioni di Box e Cox (1964)), il secondo costituisce spesso un ostacolo

insormontabile all'assunzione di normalità. In tutti quei casi in cui la specificazione di un modello adeguato risulti problematica o, anche qualora esso possa considerarsi soddisfacente, per il modello assunto non siano disponibili strumenti d'analisi per dati incompleti, può essere conveniente ricorrere a metodi di imputazione non parametrici. Tra i metodi più diffusi e di semplice realizzazione nell'ambito di questa famiglia vi sono i cosiddetti metodi *hot-deck*. Con questo termine ci si riferisce ad una classe di metodi che effettuano l'imputazione dei valori mancanti prelevandoli da unità, appartenenti allo stesso insieme di dati, in cui tali valori sono osservati. Il più semplice metodo di questo tipo è l'imputazione mediante valori estratti casualmente (con o senza ripetizione) dall'insieme dei rispondenti (*donatori*). Una versione più sofisticata del metodo consiste nel definire sottoinsiemi di unità (*classi di imputazione*) mediante variabili ausiliarie (tipicamente categoriche o "categorizzate") e, per ogni record con valori mancanti (*ricevente*), limitare la scelta del donatore a quelle unità che appartengono alla stessa classe del ricevente; naturalmente in questo caso si pone il problema della selezione delle variabili da utilizzare per definire le classi di imputazione: è evidente che la scelta va indirizzata verso quelle variabili che si suppone abbiano un forte potere predittivo nei confronti delle variabili da imputare. E' importante osservare che l'utilizzo di classi di imputazione presuppone implicitamente un modello statistico in cui sono incluse tutte le interazioni tra le variabili usate per definire le classi (Little, 1988). Non sempre questa risulta essere la strategia migliore: in taluni casi infatti, potrebbe essere più efficiente un approccio che contempli l'impiego di un insieme di covariate come regressori in un modello che consideri soltanto gli effetti principali.

Un altro approccio molto comune consiste nell'introdurre un concetto di "somiglianza" tra le unità, basato su un'opportuna funzione di distanza, definita mediante un sottoinsieme di M variabili $Y_{j_1},\ldots,Y_{j_M}$ ($M<p$), opportunamente riscalate, dette *variabili di matching*. Se indichiamo con $d(u_a, u_b)$ la distanza tra le unità u_a e u_b, e se $Y_j \quad j \in \{1,..,p\}$ è una variabile non osservata nell'unità u_a, il valore da imputare y_{aj} viene prelevato dall'unità, tra quelle in cui Y_j è osservata, che minimizza la funzione $d(u_a,\cdot)$. In presenza di più variabili non osservate $Y_1,\ldots,Y_k$, con $k<p$, può essere scelto un donatore diverso per ogni variabile (*imputazione sequenziale*) o un unico donatore per tutte le variabili (*imputazione congiunta*); in quest'ultimo caso ovviamente la ricerca è ristretta a quei record in cui $Y_1,...,Y_k$ sono tutte osservate. Del metodo appena descritto, detto del *donatore di minima distanza* (*Nearest Neighbor Donor o NND*) sono conosciute diverse varianti: il criterio della minima distanza può essere reso infatti più flessibile consentendo la scelta del donatore in modo casuale tra i g record più vicini (dove g è fissato a priori) oppure tra tutti quelli che sono ad una distanza dal ricevente inferiore ad una certa soglia prefissata; inoltre può essere introdotta una "penalizzazione" per i record già utilizzati come donatori al fine di sfavorire l'eccessivo riutilizzo di

taluni record. Anche i criteri per la definizione del "serbatoio" dei donatori sono molteplici: ad esempio potrebbero essere accettati come donatori solo quei record in cui sono osservate tutte le variabili, o quei record che soddisfano un certo insieme di vincoli logici o aritmetici definiti dal ricercatore. Inoltre possono essere esclusi dal serbatoio le unità che presentano valori anomali di alcune variabili (*outliers*). Per quanto riguarda la metrica utilizzata per definire la "similitudine" delle unità mediante le variabili di matching $Y_{j_1},\ldots,Y_{j_M}$, scelte comuni della funzione di distanza sono:

i) $d_2(u_a, u_b) = \sqrt{\sum_{m=1}^{M}\left(y_{aj_m} - y_{bj_m}\right)^2}$ (*distanza euclidea*)

ii) $d_1(u_a, u_b) = \sum_{m=1}^{M}\left|y_{aj_m} - y_{bj_m}\right|$ (*distanza di Manhattan*)

iii) $d_\infty(u_a, u_b) = \max_{m=1,\ldots,M}\left|y_{aj_m} - y_{bj_m}\right|$ (*distanza min-max*)

Nelle *i*)-*iii*) y_{aj_m} (y_{bj_m}) rappresenta il valore che assume la variabile Y_{j_m} sull'unità u_a (u_b) $m = 1,\ldots,M$. E' utile ribadire inoltre che è sempre preferibile standardizzare preliminarmente le variabili di matching al fine di evitare che i diversi contributi alla funzione complessiva di distanza dipendano eccessivamente dalla "scala" delle corrispondenti variabili.

2.6 Il Predictive Mean Matching

Nonostante il metodo del donatore di minima distanza sia da molto tempo largamente impiegato, specialmente nel contesto della Statistica Ufficiale, le sue proprietà statistiche rimangono tuttora in gran parte inesplorate. Chen e Shao (2000) hanno dimostrato che nel caso di una variabile continua Y con valori mancanti, il metodo del donatore di minima distanza basato su una covariata X sempre osservata fornisce stimatori asintoticamente corretti e consistenti (rispetto ad un modello sottostante) di alcuni parametri come medie o totali sotto opportune condizioni di regolarità per il valore atteso condizionato $E(Y \mid X)$. Tuttavia analoghi risultati per casi più generali, come quello dell'utilizzo di più covariate come variabili di matching, non sono al momento disponibili. La scelta di questo tipo di metodo in luogo, ad esempio, dell'impiego di un modello parametrico esplicito è spesso motivata dalla difficoltà di includere nel modello un gran numero di variabili, e conseguentemente, di dover stimare un gran numero di parametri. D'altra parte, anche

qualora il metodo del donatore di minima distanza possa essere considerato in qualche modo equivalente (almeno asintoticamente) all'imputazione dalla distribuzione predittiva dei valori mancanti condizionatamente all'insieme di variabili di matching osservate, l'impiego simultaneo di un gran numero di covariate con diverso potere predittivo potrebbe rivelarsi altamente inefficiente.

Il *predictive mean matching* (*PMM*), proposto da Rubin (1986) nel contesto del matching statistico (D'orazio et al., 2002), è un metodo di imputazione che, collocandosi ad un livello intermedio tra la famiglia dei metodi parametrici e quella dei metodi non parametrici, conserva alcune caratteristiche del donatore di minima distanza, pur facendo uso di un modello parametrico esplicito. Per descrivere il metodo, iniziamo con il considerare il caso in cui solo una delle variabili $Y_1,..,Y_p$, ad esempio la prima, è affetta da mancata risposta. Per ogni unità u_i in cui la variabile Y_1 non è osservata, un valore "intermedio" y_{i1}* (media predittiva) può essere determinato mediante regressione di Y_1 su $Y_2,..,Y_p$. Al fine di preservare le caratteristiche distribuzionali nel data-set completato, dovrebbe essere aggiunto un residuo a ciascuna media predittiva y_{i1}*. Questo potrebbe essere fatto, con procedimento sostanzialmente analogo a quello descritto nel contesto più generale dell'imputazione mediante algoritmo *EM*, semplicemente generando da una distribuzione normale univariata a media 0 e varianza uguale alla varianza residua stimata dalla regressione; tale procedimento tuttavia si basa fortemente sull'ipotesi di normalità, ed in particolare su quella di omoschedasticità dei residui. Un modo di rendere il metodo meno legato all'assunzione di normalità consiste nel aggiungere ogni residuo "prelevandolo" dal rispondente con media predittiva più vicina a quella del non-rispondente in esame (David et al., 1986); alternativamente la media predittiva può essere utilizzata come "variabile di matching" per selezionare il record da cui imputare il valore di *Y*. Questa seconda opzione, ha il vantaggio di produrre valori imputati che sono sicuramente ammissibili (in quanto effettivamente osservati su qualche unità); inoltre, poiché la media predittiva è utilizzata esclusivamente per determinare l'abbinamento di un record ricevente con un donatore, il metodo è probabilmente meno sensibile ad eventuali allontanamenti dalla normalità ed in particolare più robusto rispetto alla presenza di eteroschedasticità. In questo lavoro ci riferiremo al metodo del *PMM* soltanto in questa seconda accezione. Sebbene a rigore l'utilizzo della media predittiva come variabile di matching non sia equivalente all'introduzione di una metrica nello spazio dei regressori $Y_2,..,Y_p$ (IR^{p-1}), ad eccezione del caso banale $p=2$ in cui il metodo è identico al *NND*, il predictive mean matching può essere visto come un particolare metodo di imputazione basato sul donatore di minima distanza; tuttavia a differenza di quanto avviene per i metodi che utilizzano distanze "proprie" (come quella euclidea o quella di Manhattan), nel *PMM* l'influenza delle covariate sulla selezione dei record donatori dipende automaticamente dal diverso potere predittivo nei confronti della variabile da imputare. Supponiamo ora che tra le *p* variabili di analisi,

le prime k (Y_1,..,Y_k) siano affette da mancata risposta e supponiamo inoltre che la mancata risposta possa manifestarsi su di esse solo "in blocco", cioè le variabili Y_1,..,Y_k sono o tutte osservate o tutte mancanti. Per semplicità di notazione poniamo $\boldsymbol{Y} = (\tilde{\tilde{\boldsymbol{Y}}}, \tilde{\boldsymbol{Y}})$ con $\tilde{\tilde{\boldsymbol{Y}}} = (Y_1,..,Y_k)$ e $\tilde{\boldsymbol{Y}} = (Y_{k+1},..,Y_p)$. Si potrebbe allora applicare il metodo appena descritto ad ogni componente Y_j (j=1,..,k) di $\boldsymbol{Y}$ separatamente, utilizzando in modo sequenziale le medie predittive per selezionare un donatore da cui prelevare il valore da imputare. Il procedimento tuttavia comporterebbe in generale la scelta di un diverso donatore per ogni variabile Y_j e produrrebbe quindi delle distorsioni nella stima delle relazioni di associazione tra le variabili $\tilde{\tilde{\boldsymbol{Y}}}$. Se tali associazioni sono oggetto di studio, è preferibile abbinare un unico donatore ad ogni ricevente, ricorrendo alla regressione multivariata di $\tilde{\tilde{\boldsymbol{Y}}}$ su $\tilde{\boldsymbol{Y}}$ per determinare, per ogni record ricevente u_i, il vettore delle medie predittive y_{i1}*,.. ,y_{ik}*, mediante le quali definire la funzione di distanza. Come già notato nel paragrafo precedente, nel calcolo della distanza tra le unità, è opportuno tener conto delle diverse scale di variabilità delle variabili di matching impiegate. Nel caso del *PMM*, un'opportuna funzione di distanza può essere definita nel modo seguente: sia $\boldsymbol{y}_a^* = (y_{a1}^*,.. ,y_{ak}^*)^T$ il vettore (k *x 1*) ottenuto trasponendo il vettore delle medie predittive relative all'unità u_a calcolate mediante regressione di $\tilde{\tilde{\boldsymbol{Y}}}$ su $\tilde{\boldsymbol{Y}}$; sia inoltre $\boldsymbol{\Sigma}_{\tilde{\tilde{\boldsymbol{Y}}}\cdot\tilde{\boldsymbol{Y}}}$ la matrice di varianze e covarianze residua stimata dalla regressione. La distanza $d_{Mah}(u_a, u_b)$ tra le unità u_a e u_b può allora essere definita come:

$$(7) \qquad d_{Mah}(u_a, u_b) = (\boldsymbol{y}_a^* - \boldsymbol{y}_b^*)^T \boldsymbol{\Sigma}_{\tilde{\tilde{\boldsymbol{Y}}}\cdot\tilde{\boldsymbol{Y}}}^{-1} (\boldsymbol{y}_a^* - \boldsymbol{y}_b^*)$$

La distanza (7), nota come distanza di Mahalanobis, tiene automaticamente conto delle diverse "scale" delle variabili; nel caso di indipendenza delle variabili Y_1,.. ,Y_k condizionatamente a $\tilde{\boldsymbol{Y}}$, ad esempio, la matrice $\boldsymbol{\Sigma}_{\tilde{\tilde{\boldsymbol{Y}}}\cdot\tilde{\boldsymbol{Y}}}^{-1}$ è diagonale ed è facile rendersi conto che la (7) si riduce alla distanza euclidea ordinaria purchè per ogni $i = 1,\ldots,n$ i valori y_{i1}*,.. ,y_{ik}* siano stati preliminarmente standardizzati dividendoli per i corrispondenti elementi sulla diagonale di $\boldsymbol{\Sigma}_{\tilde{\tilde{\boldsymbol{Y}}}\cdot\tilde{\boldsymbol{Y}}}$ (varianze residue).

Consideriamo ora il caso generale di mancate risposte che possono manifestarsi su arbitrari sottoinsiemi di tutte le variabili di analisi $Y_1,\ldots, Y_p$, cioè il caso in cui in un insieme di dati siano presenti diversi *pattern* di valori mancanti. E' possibile allora applicare il procedimento appena descritto stimando tanti modelli di regressione quanti sono i diversi pattern, per determinare le medie predittive da utilizzare nel calcolo della distanza. Più precisamente, se al solito $Y_1,\ldots,Y_p$ sono

le variabili di analisi, ogni pattern $\mathcal{P}$ di mancata risposta è associato univocamente a una bipartizione dell'insieme di indici: $\{1,2,\ldots,p\}$ in due sottoinsiemi di indici $O(\mathcal{P})$ e $M(\mathcal{P})$ corrispondenti rispettivamente all'insieme $\boldsymbol{Y}_{O(\mathcal{P})}$ delle variabili osservate e a quello $\boldsymbol{Y}_{M(\mathcal{P})}$ delle variabili non osservate per il dato pattern. Il metodo del predictive mean matching richiede allora la stima dei parametri della distribuzione condizionata $P(\boldsymbol{Y}_{M(\mathcal{P})} \mid \boldsymbol{Y}_{O(\mathcal{P})})$. Come già accennato alla fine del par. *2.4*, la stima di un insieme di parametri per ogni pattern $\mathcal{P}$ mediante modelli di regressione effettuati sulla base dei dati completamente osservati non è in generale la strategia più efficiente: come già per il metodo completamente basato sull'utilizzo di un modello esplicito descritto sempre nel par. *2.4*, sembra dunque più vantaggioso stimare mediante l'algoritmo *EM* un unico insieme di parametri relativo alla distribuzione congiunta delle variabili $Y_1,\ldots,Y_p$ e ricavare i parametri delle distribuzioni condizionate corrispondenti ai diversi pattern di mancata risposta attraverso l'utilizzo degli *sweep operator*. I passi necessari per l'implementazione della procedura descritta sono:

1. Stimare i parametri $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ della distribuzione congiunta delle variabili $Y_1,\ldots,Y_p$ mediante algoritmo *EM*.
2. Per ogni pattern $\mathcal{P}$ di mancata risposta:

   i) stimare con l'ausilio degli *sweep operator* i parametri della distribuzione condizionata $P(\boldsymbol{Y}_{M(\mathcal{P})} \mid \boldsymbol{Y}_{O(\mathcal{P})})$ come funzioni dei parametri stimati $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$;

   ii) utilizzando i parametri stimati in *i*) calcolare, per ogni unità u_b senza mancate risposte la media predittiva $\boldsymbol{y}_{b\,M(\mathcal{P})}^{*}$ delle variabili $\boldsymbol{Y}_{M(\mathcal{P})}$ condizionatamente ai valori assunti dalle variabili $\boldsymbol{Y}_{O(\mathcal{P})}$:

   $$\boldsymbol{y}_{b\,M(\mathcal{P})}^{*} \equiv E\,(\boldsymbol{Y}_{M(\mathcal{P})} \mid \boldsymbol{y}_{b\,O(\mathcal{P})})\ ;$$

   iii) procedere come in *ii*) per calcolare la media predittiva $\boldsymbol{y}_{a\,M(\mathcal{P})}^{*}$ di ogni unità u_a con pattern di mancata risposta $\mathcal{P}$;

   iv) per ogni unità u_a con pattern di mancata risposta $\mathcal{P}$ selezionare un donatore $u^*(u_a)$ in modo da minimizzare la distanza di Mahalanobis definita in (7):

   $$(8) \qquad u^*(u_a) \equiv \min_{u_b \in D} (\boldsymbol{y}_{a\,M(\mathcal{P})}^{*} - \boldsymbol{y}_{b\,M(\mathcal{P})}^{*})^{T} \boldsymbol{\Sigma}_{\boldsymbol{Y}_{M(\mathcal{P})}\cdot\boldsymbol{Y}_{O(\mathcal{P})}}^{-1} (\boldsymbol{y}_{a\,M(\mathcal{P})}^{*} - \boldsymbol{y}_{b\,M(\mathcal{P})}^{*})$$

   dove D indica l'insieme dei donatori (record senza mancate risposte) e $\boldsymbol{\Sigma}_{\boldsymbol{Y}_{M(\mathcal{P})}\cdot\boldsymbol{Y}_{O(\mathcal{P})}}$ la matrice di varianze e covarianze residua dalla regressione di $\boldsymbol{Y}_{M(\mathcal{P})}$ su $\boldsymbol{Y}_{O(\mathcal{P})}$;

v) imputare i valori delle variabili $\boldsymbol{Y}_{M(\varphi)}$ nell'unità u_a con i valori che tali variabili assumono su $u^*(u_a)$.

Da un punto di vista pratico l'algoritmo di stima dei parametri e di ricerca del donatore mediante media predittiva può essere reso più efficiente ordinando i dati per pattern di mancata risposta: in tal modo infatti, si può evitare di ripetere più volte il calcolo dei parametri delle varie distribuzioni condizionate mediante *sweep operator*.

3. Imputazione Multipla

Tutti i metodi illustrati finora sono metodi di *imputazione singola*, ossia sono basati sulla sostituzione dei valori mancanti con un unico insieme di valori "plausibili". Come è stato già accennato nell'introduzione, il maggior difetto di questo approccio consiste nel fatto che tipicamente, le analisi che vengono effettuate sui data-set completati non tengono conto della mancata risposta come sorgente di incertezza, tendendo a considerare tutti i dati (anche quelli imputati) come se fossero stati effettivamente osservati. L'*imputazione multipla* è essenzialmente un metodo Monte Carlo che consente di effettuare un'ampia classe di analisi inferenziali in presenza di non-risposta mediante analisi standard su diversi data-set completi. In questo paragrafo descriveremo brevemente i principi su cui si fonda il metodo e accenneremo alle possibili applicazioni facendo riferimento in particolare al caso di dati normali; una descrizione dettagliata si può trovare nel testo di Rubin (1987).

Con l'imputazione multipla, ai valori mancanti $\boldsymbol{Y}_{mis}$ sono associati un certo numero m di insiemi di valori artificiali $\boldsymbol{Y}^{(1)}_{mis}$, $\boldsymbol{Y}^{(2)}_{mis}$, …, $\boldsymbol{Y}^{(m)}_{mis}$ in modo da ottenere m data-set completi che possono essere analizzati con metodi standard per dati completi. La variabilità dei risultati che si ottengono con le m analisi indipendenti effettuate, riflette l'incertezza associata alla mancata risposta e, combinata con la (eventuale) componente di variabilità di origine campionaria, può fornire una misura complessiva di incertezza nelle inferenze sui parametri di interesse.

Da un punto di vista Bayesiano l'imputazione multipla può essere vista come una procedura di generazione di m realizzazioni indipendenti dalla distribuzione predittiva a posteriori dei dati mancanti condizionatamente ai dati osservati $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$. Le procedure più comuni sono basate sull'assunzione di un modello parametrico esplicito per i dati completi e di un'opportuna distribuzione a priori (generalmente scelta non informativa). In questo contesto, la distribuzione predittiva a posteriori dei dati mancanti può essere espressa come:

$$(9) \qquad P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}) = \int P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs},\theta)P(\theta \mid \boldsymbol{Y}_{obs})d\theta$$

cioè come la media della distribuzione predittiva condizionata $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta)$ di $\boldsymbol{Y}_{mis}$ dati i parametri θ, rispetto alla distribuzione a posteriori dei parametri "a dati osservati" $P(\theta \mid \boldsymbol{Y}_{obs})$. Per poter generare dalla distribuzione $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs})$ sarebbe pertanto sufficiente essere in grado di generare da ciascuna delle due distribuzioni di probabilità che compaiono nella (9) sotto il segno di integrale. Mentre la generazione dei dati mancanti $\boldsymbol{Y}_{mis}$ dalla $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta)$ non presenta particolari difficoltà, la distribuzione a posteriori a dati osservati $P(\theta \mid \boldsymbol{Y}_{obs})$ è in generale intrattabile. Si è soliti pertanto ricorrere a tecniche di tipo MCMC (Markov Chain Monte Carlo) che consentono di ricondurre il problema della generazione dalla distribuzione $P(\theta|\boldsymbol{Y}_{obs})$ a quello più semplice della generazione dalla distribuzione "a dati completi" $P(\theta|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$. Uno dei procedimenti più comuni consiste nel seguente schema iterativo (Shafer, 1997):

- dato un insieme di valori correnti $\theta^{(t)}$ per i parametri, generare dalla distribuzione predittiva condizionata $P(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}, \theta^{(t)})$ un insieme di valori dei dati mancanti $\boldsymbol{Y}_{mis}^{(t+1)}$ (***I-Step***);
- condizionatamente a $\boldsymbol{Y}_{mis}^{(t+1)}$, generare dalla distribuzioni a posteriori a dati completi $P(\theta \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t+1)})$ un nuovo insieme di parametri $\theta^{(t+1)}$ (***P-step***).

Lo schema descritto è noto in letteratura come *Data Augmentation* (Schafer,1997). Esso fornisce, a partire da un insieme di parametri iniziali $\theta^{(0)}$, una successione aleatoria $\{\theta^{(t)}, \boldsymbol{Y}_{mis}^{(t)}\}_{t=1,2,\ldots}$ la cui distribuzione stazionaria è $P(\boldsymbol{Y}_{mis}, \theta| \boldsymbol{Y}_{obs})$ (Tanner e Wong, 1987). In particolare le sottosuccessioni $\{\theta^{(t)}\}_{t=1,2,\ldots}$ e $\{\boldsymbol{Y}_{mis}^{(t)}\}_{t=1,2,\ldots}$ hanno come distribuzioni stazionarie $P(\theta \mid \boldsymbol{Y}_{obs})$ e $P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ rispettivamente.

Nel caso in cui il modello assunto per i dati completi sia normale multivariato si ha: $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $(\boldsymbol{Y}_{mis}, \boldsymbol{Y}_{obs}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dove $\boldsymbol{\mu}$ è il vettore p-dimensionale delle medie e $\boldsymbol{\Sigma}$ la matrice $p \times p$ di varianze e covarianze. Il passo di *I-step* non presenta particolari difficoltà: in pratica, avendo supposto indipendenti le n osservazioni del data-set da analizzare, esso si riduce, per ogni ciclo t e per ogni record incompleto u_i, alla generazione dalla distribuzione di probabilità condizionata $P(\boldsymbol{Y}_{M(\mathcal{P})} \mid \boldsymbol{y}_{iO(\mathcal{P})}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ dove, in accordo con le notazioni del paragrafo precedente, $\mathcal{P}$ è il *pattern* di mancata risposta del record u_i e $O(\mathcal{P})$, $M(\mathcal{P})$ sono i relativi insiemi di indici che si riferiscono alle variabili osservate e non osservate rispettivamente. Il passo di *I-step* in sostanza è equivalente all'imputazione con residuo nel metodo basato sull'algoritmo *EM* con la differenza che nelle

distribuzioni condizionate utilizzate compaiono i valori correnti dei parametri anziché le loro stime di massima verosimiglianza. Per quanto riguarda il passo di *P-step*, un'appropriata distribuzione a priori $\pi(\theta)$ per i parametri $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ può essere costruita a partire dalla distribuzione di Wishart inversa (Mardia, 1979) che è la distribuzione coniugata naturale per la funzione di verosimiglianza relativa a dati normali multivariati (Box e Tiao, 1992). Più precisamente $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ può essere specificata assumendo che la distribuzione condizionata di $\boldsymbol{\mu}$ data $\boldsymbol{\Sigma}$ sia normale multivariata e che $\boldsymbol{\Sigma}$ abbia distribuzione di Wishart inversa:

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}_0, \tau^{-1}\boldsymbol{\Sigma}) \tag{10}$$

$$\boldsymbol{\Sigma} \sim W^{-1}(q, \Lambda) \tag{11}$$

dove $\boldsymbol{\mu}_0$, τ, q, Λ sono parametri noti detti *iperparametri*. Se non sono disponibili informazioni a priori sui parametri, una appropriata distribuzione a priori non informativa si ottiene ponendo:

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2} \tag{12}$$

dove $|\boldsymbol{\Sigma}|$ indica il determinante di $\boldsymbol{\Sigma}$. La (12) si ottiene dalle (10), (11) nel limite, $\tau \to 0$, $q \to -1$, $\Lambda^{-1} \to 0$; il fatto che essa non dipenda da $\boldsymbol{\mu}$ equivale ad assumere che $\boldsymbol{\mu}$ abbia una distribuzione (impropria) uniforme su tutto lo spazio p-dimensionale. Moltiplicando la (12) per la funzione di verosimiglianza relativa alla distribuzione multinormale basata sui dati $\boldsymbol{Y}^{(t)} \equiv (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)})$ si ottiene finalmente la distribuzione a posteriori a dati completi:

$$P(\boldsymbol{\mu}^{(t)} \mid \boldsymbol{\Sigma}^{(t)}, \boldsymbol{Y}^{(t)}) = N(\bar{\boldsymbol{y}}^{(t)}, n^{-1}\boldsymbol{\Sigma}^{(t)}) \tag{13}$$

$$P(\boldsymbol{\Sigma}^{(t)} \mid \boldsymbol{Y}^{(t)}) = W^{-1}(n\text{-}1, (n\boldsymbol{S}^{(t)})^{-1})$$

dove n rappresenta la numerosità del data-set, e $\bar{\boldsymbol{y}}^{(t)}$, $\boldsymbol{S}^{(t)}$ sono rispettivamente la media e la matrice di varianze e covarianze campionarie calcolate sulla base dei dati $\boldsymbol{Y}^{(t)}$. Il *P-step* per il modello multinormale alla t-esima iterazione consiste dunque nella generazione dei parametri $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ dalla (13).

Come ultima osservazione sul metodo della *data augmentation* è utile sottolineare che, affinché si possa considerare ogni insieme di valori $\boldsymbol{Y}_{mis}^{(t)}$ come generato dalla distribuzione predittiva

$P(\boldsymbol{Y}_{mis} \mid \boldsymbol{Y}_{obs})$ è necessario che t sia sufficientemente grande da garantire la stazionarietà; per la generazione di data-set multipli, si è soliti pertanto considerare solo valori di t maggiori di un valore prefissato t_0 (*burn-in period*) che si ritiene abbastanza elevato. Inoltre la richiesta di indipendenza per le imputazioni multiple effettuate suggerisce di non utilizzare data-set ottenuti da iterazioni consecutive dello schema descritto. Al contrario, è conveniente sottocampionare dalla successione $\{\theta^{(t)}, \boldsymbol{Y}_{mis}^{(t)}\}_{t= t_0,\, t_0+1,\ldots}$ estraendo da essa a passo costante K, dove K è abbastanza grande da poter considerare trascurabile la dipendenza tra gli insiemi estratti.

Una volta ottenuto mediante imputazione multipla un insieme di m data-set completi, è semplice effettuare inferenze su un parametro di interesse Q della popolazione di riferimento combinando opportunamente i risultati delle m analisi effettuate sui singoli data-set completati. Supponiamo che $\hat{Q} = \hat{Q}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ sia una stima puntuale a dati completi per Q, cioè la stima che si otterrebbe se tutti i dati fossero osservati. Sia inoltre $U = U(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ la varianza associata alla stima $\hat{Q}$ e supponiamo che possa considerarsi valida l'approssimazione:

$$\hat{Q}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \approx E(Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \qquad (14)$$
$$U(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}) \approx V(Q \mid \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$$

con $E(Q \mid \cdot)$, $V(Q \mid \cdot)$ rispettivamente media e varianza a posteriori di Q. Per il t-esimo data-set completato, potranno essere calcolate le quantità $\hat{Q}^{(t)} = \hat{Q}(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)})$ e $U^{(t)} = U(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis}^{(t)})$. La stima puntuale di Q basata sull'imputazione multipla sarà allora:

$$\bar{Q} = \frac{1}{m}\sum_{t=1}^{m} \hat{Q}^{(t)} \qquad (15)$$

e la varianza associata può essere calcolata come:

$$T = \bar{U} + (1+m^{-1})B \qquad (16)$$

dove:

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m} U^{(t)} \qquad (\textit{varianza within})$$

$$B = \frac{1}{m-1}\sum_{t=1}^{m} (\hat{Q}^{(t)} - \bar{Q})^2 \qquad (\textit{varianza between}).$$

Mentre da un punto di vista Bayesiano, le (15), (16) possono essere interpretate, almeno entro i limiti di validità delle (14), come approssimazioni Monte Carlo della media e della varianza a posteriori di *Q*, la giustificazione delle stesse formule da un punto di vista frequentista richiede maggior cura. Difficoltà ancora maggiori sorgono allorquando le inferenze sui parametri di una popolazione finita non sono basate su un modello di superpopolazione sottostante, ma su un disegno campionario (Cochran,1977). Argomenti teorici ed euristici a favore della validità delle regole (15), (16) in vengono forniti da Rubin (1987) che introduce a tal uopo la nozione di *imputazione propria*. In questa sede non verrano analizzati questi aspetti.

4. Il software QUIS

4.1 Generalità

Il software *QUIS* (*Quick Imputation System*) disponibile al momento in versione prototipale, è un'insieme di programmi che implementano altrettanti metodi di imputazione di mancate risposte parziali. Il software è stato interamente sviluppato nell'ambito del sistema SAS SYSTEM V8, un package statistico di uso comune che incorpora numerose procedure di analisi dei dati. Da un punto di vista funzionale è costituito da alcune routine (macro SAS) scritte utilizzando i moduli SAS Base e SAS IML. L'interfaccia utente, che si avvale di maschere interattive per passare ai programmi i parametri necessari, è stata sviluppata nel linguaggio SCL (*Screen Control Language*), un linguaggio per la creazione di applicazioni interattive in ambiente SAS. I requisiti hardware minimali per la installazione e l'utilizzo di QUIS sono una memoria RAM di almeno 64 Mbyte e spazio di memoria su disco fisso di almeno 5 Mbyte. La velocità di esecuzione dipende ovviamente dalle caratteristiche del processore e dalla mole di dati da processare. L'utilizzo di QUIS richiede che sia installato il SAS in versione V8.1 o superiori ed in particolare che siano disponibili i moduli SAS Base, SAS STAT , SAS IML, SAS AF. I dati da processare devono essere forniti sotto forma di data-set SAS; se sono disponibili in altri formati comuni (fogli di lavoro Excel, tabelle Access, file testo, ecc) possono essere trasformati nel formato idoneo mediante gli opportuni moduli di importazione nativi di SAS.

All’avvio del programma appare la seguente schermata:

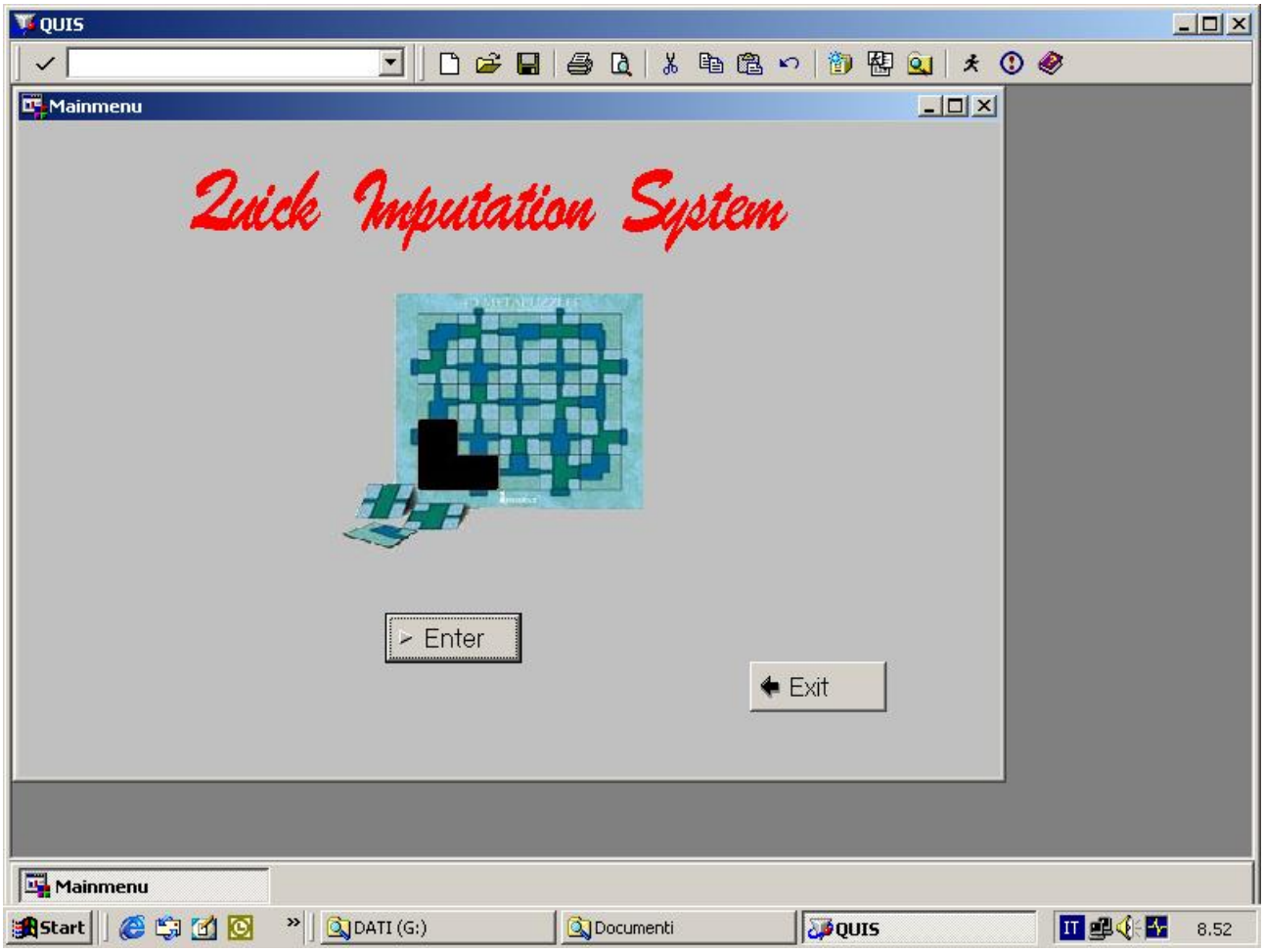


Fig.1 *Schermata di ingresso*

da questa come dalle altre finestre è possibile disporre dei tasti di gestione e della linea di comando SAS cosicché, ad esempio, l’utente può controllare il corretto funzionamento di programmi aprendo la finestra di “*log*”, o visualizzare librerie e data-set con il tasto “*explorer*”. Il bottone “*Exit*” chiude l’applicazione mentre il bottone “Enter” attiva la finestra principale:

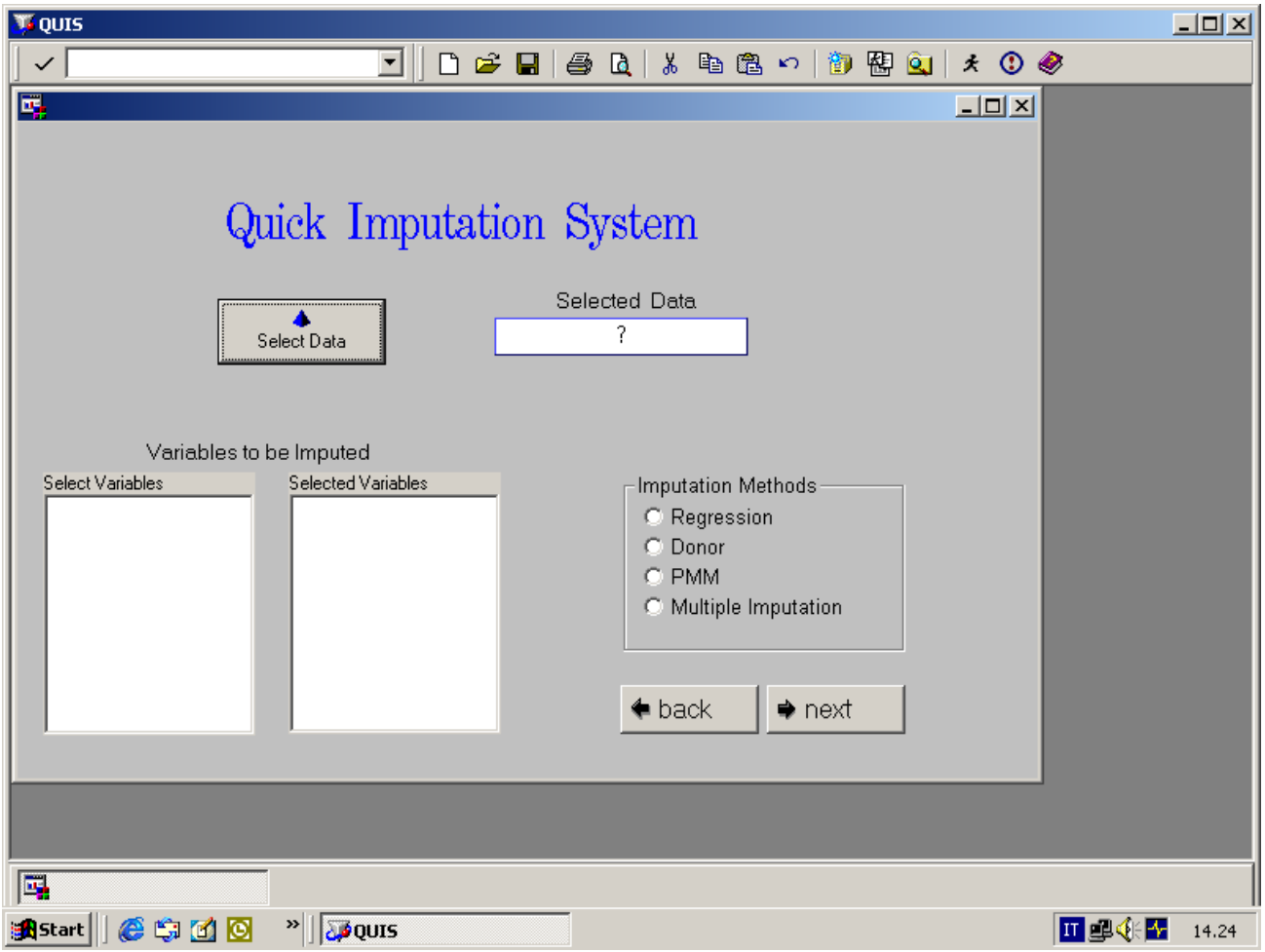


Fig.2 *Maschera principale*

La maschera principale è quella in cui l'utente seleziona il data-set incompleto da analizzare. Questa operazione viene effettuata tramite il bottone "*Select Data*" che attiva una finestra di esplorazione delle librerie definite dall'utente. La selezione di una libreria determina la visualizzazione di tutti i data-set in essa contenuti. Tra questi può essere scelto il data-set da analizzare. Dalla stessa finestra inoltre possono essere definite nuove librerie da assegnare a specificati *path*; l'operazione è sostanzialmente analoga a quella che si effettua normalmente in ambiente SAS :

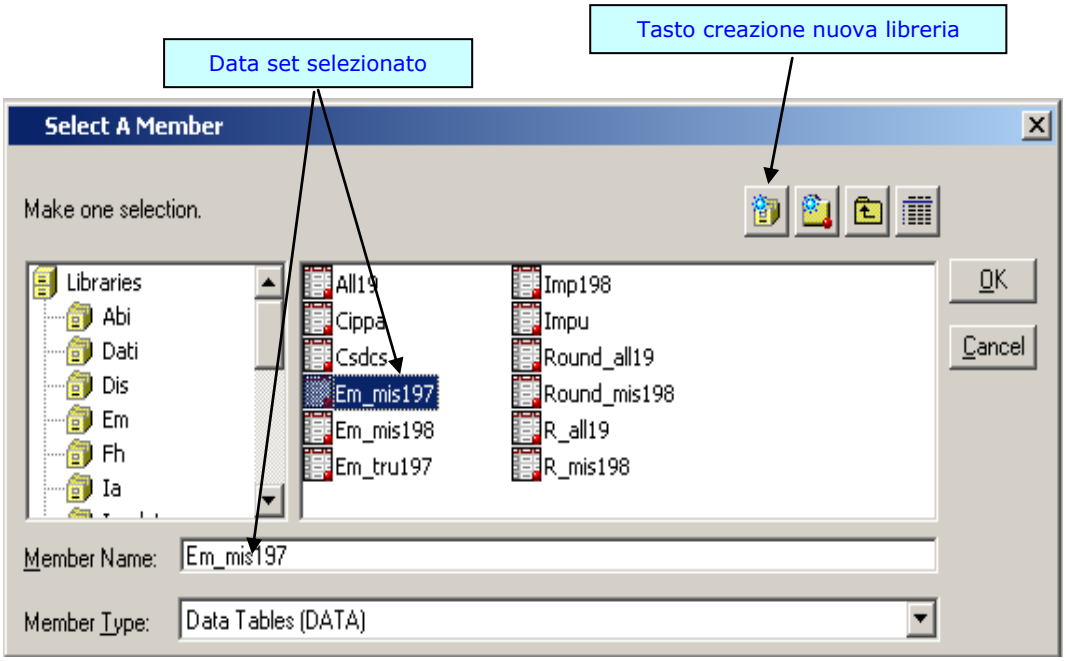


Fig.3 *Selezione dei dati*

Nel momento in cui con il tasto "OK" si conferma la selezione del data-set, viene visualizzato in un'apposita form il corrispondente nome a due livelli (*libname.membername*). Inoltre, nelle finestra "*Select Variables*" appaiono i nomi delle variabili contenute nel data-set. Le variabili che devono essere imputate, ovvero, anche se non presentano valori mancanti, devono essere incluse tra quelle di analisi nei metodi di imputazione che fanno uso di modello esplicito, possono quindi essere selezionate (o deselezionate) e i nomi corrispondenti vengono elencati nella finestra contigua "*Selected Variables*". E' opportuno ricordare che i metodi implementati in *QUIS* prevedono il trattamento esclusivamente di variabili quantitative; variabili di tipo diverso non vengono accettate dal programma. Una volta selezionate le variabili di analisi, l'utente deve specificare il metodo di imputazione tra quelli elencati nell'area "*Imputation Method*"; con il tasto "*Next*" attiverà quindi la finestra corrispondente al metodo selezionato.

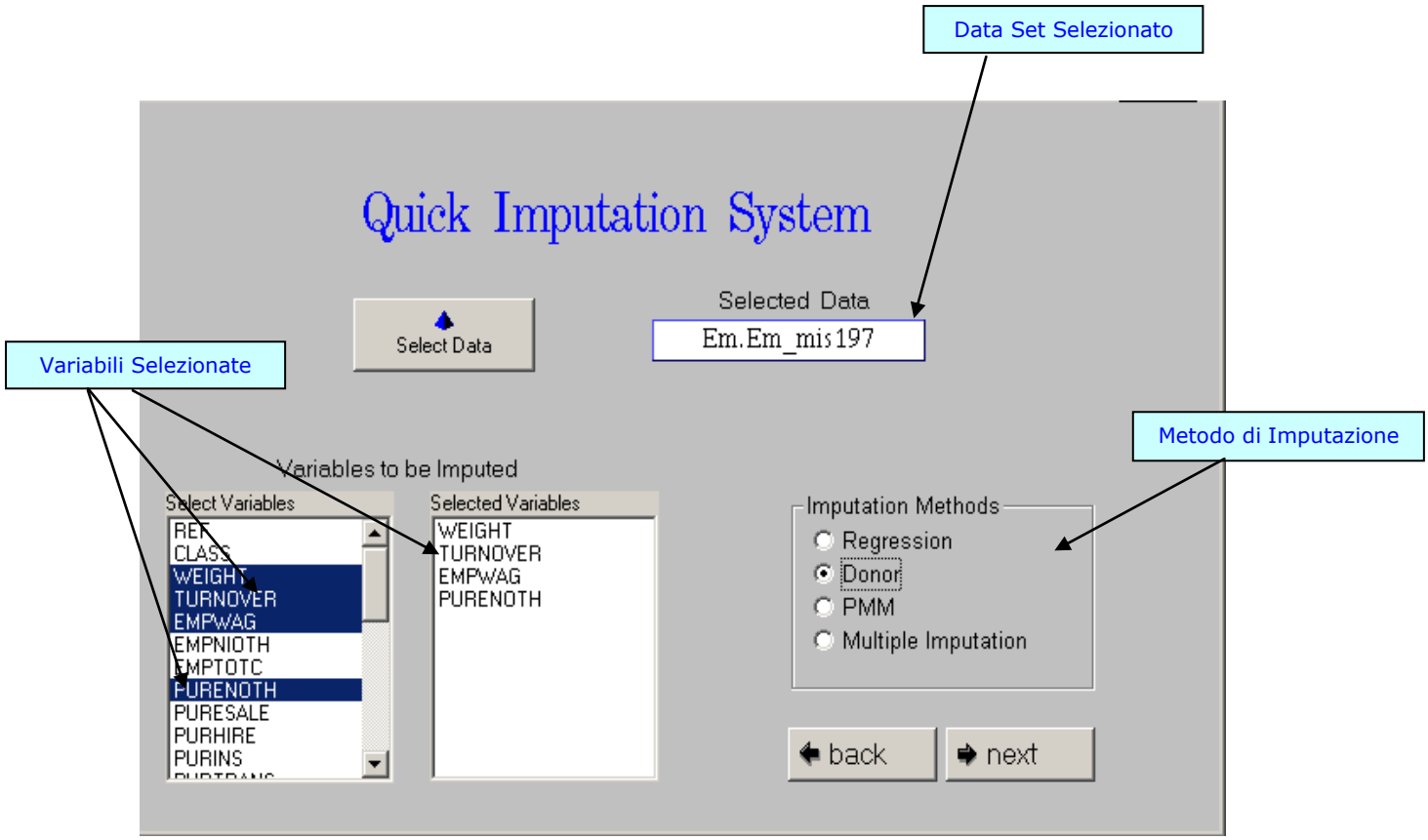


Fig.4 *Selezione delle variabili e del metodo di imputazione*

Il tasto "*Next*" attiverà a questo punto la finestra corrispondente al metodo selezionato. I metodi implementati sono stati illustrati in generale nei paragrafi 2- 3. Nei paragrafi che seguono saranno descritti alcuni dettagli operativi.

4.2 Regressione mediante algoritmo EM

L'impiego dell'algoritmo *EM* per l'imputazione dei dati mancanti è stato descritto nei paragrafi 2.2-2.4. *QUIS* utilizza una versione dell'algoritmo implementata da Paul D. Allison, University of Pennsylania , e disponibile su Web come *Macro* SAS (http://www.ssc.upenn.edu/~allison/). Il programma di Allison è in realtà finalizzato all'imputazione multipla di dati multinormali, e si serve dell'*EM* per inizializzare i valori dei parametri del modello nella Data Augmentation. E' possibile tuttavia utilizzare la macro anche per effettuare l'imputazione singola dal modello stimato con l'algoritmo *EM*, scegliendo una delle due opzioni descritte nel par. *2.4*. La regola di arresto dell'algoritmo consiste nell'interrompere il ciclo iterativo quando due iterazioni successive determinano variazioni delle stime inferiori allo 0.1% per tutti i parametri. E' tuttavia possibile impostare un numero massimo di iterazioni consentite oltre il quale il ciclo viene comunque interrotto (il valore di default è 20, il valore massimo è 200), al fine di evitare tempi di esecuzione

troppo lunghi nei casi critici in cui la convergenza è troppo lenta. Tale valore può essere direttamente digitato dall'utente nell'apposito campo. Gli altri parametri da impostare sono:

- libreria a cui assegnare il data-set di output (bottone "*Select Library*");
- nome del data-set di output;
- tipo di imputazione (con o senza residuo);

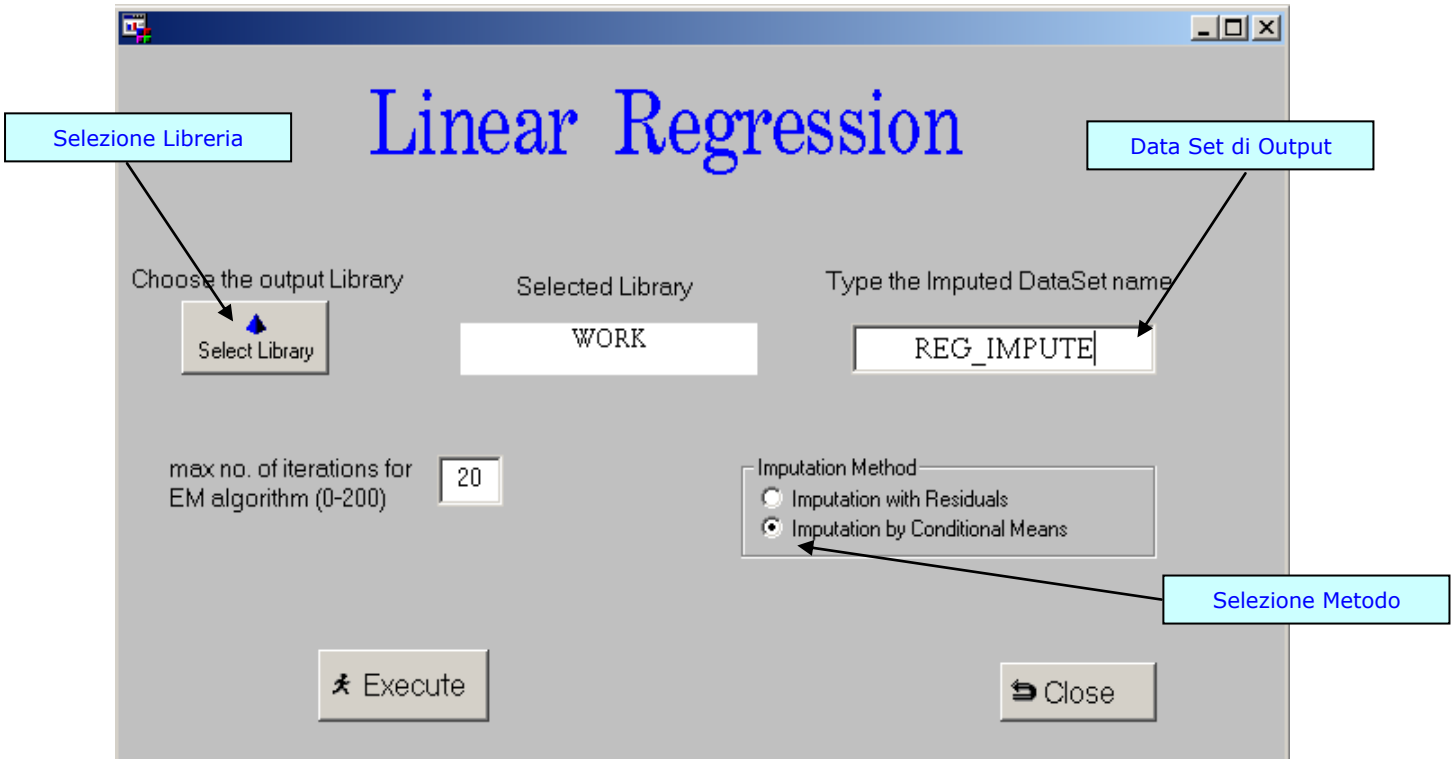


Fig.5 *Maschera per l'Imputazione con Regressione Lineare*

4.3 Donatore di minima distanza

La versione del metodo del donatore di minima distanza implementato in *QUIS* consente tre possibilità nella scelta della metrica (par. *2.5*): Euclidea, Manhattan e Min-Max; l'utente effettua la scelta dalla maschera "*Nearest Neighbour Donor*". Nella stessa maschera si possono definire le variabili di matching selezionandole dalla lista contenuta nella finestra "*Select Variables*"; i nomi delle variabili selezionate appariranno nella finestra "*Selected Variables*". E' importante osservare che l'insieme delle variabili di matching può contenere alcune o tutte le variabili da imputare: se in un dato record ricevente una variabile di matching non è osservata, essa semplicemente non verrà presa in considerazione nel calcolo della funzione di distanza complessiva. Nel caso in cui non

venga specificata alcuna variabile di matching, il programma seleziona un record in modo casuale dal serbatoio dei donatori. Quest'ultimo è costituito a sua volta da tutti e soli i record in cui sono osservate tutte le variabili di matching e le variabili da imputare. Nel caso in cui, per un dato ricevente, esistano più record a distanza minima, il donatore viene selezionato in modo casuale tra questi.

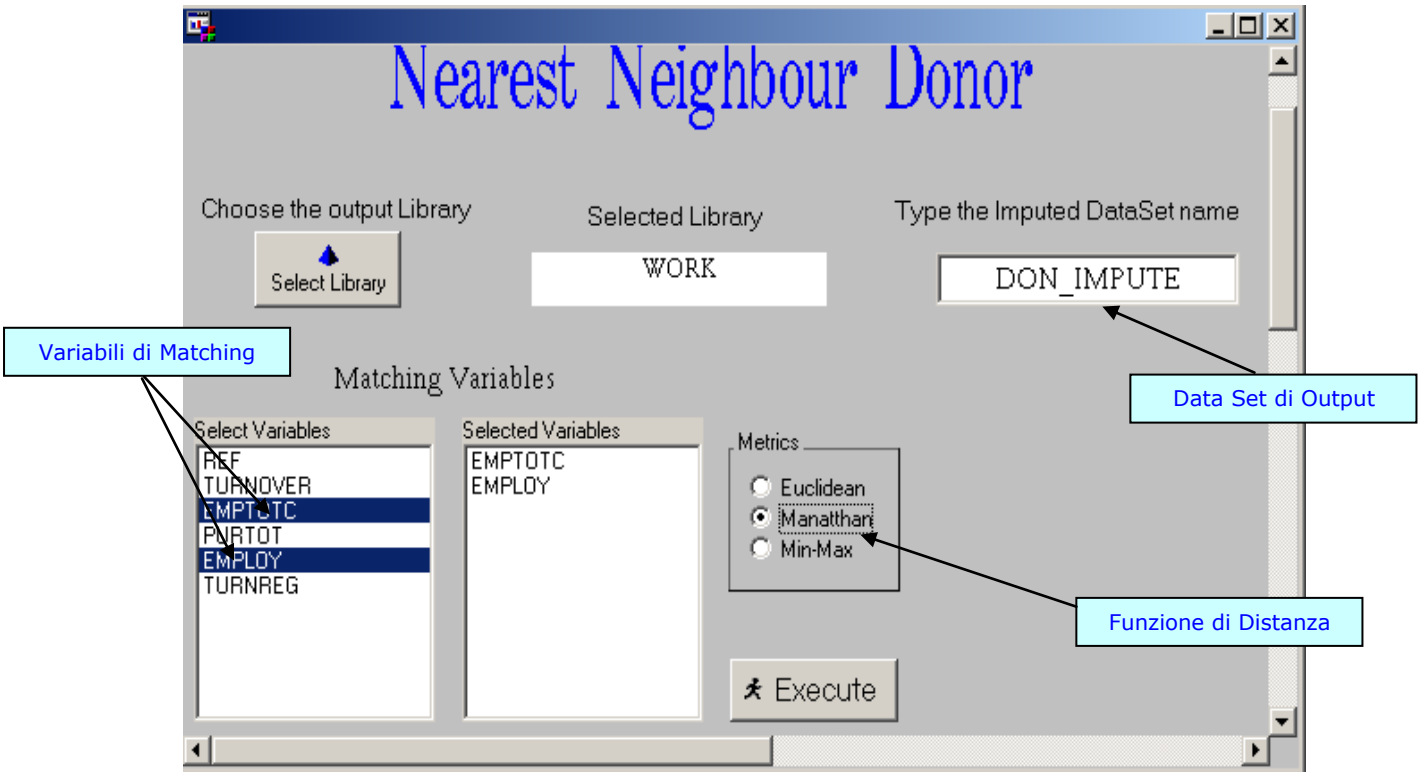


Fig.6 *Maschera per il donatore di minima distanza*

4.4 Predictive Mean Matching

Per questo metodo l'utente deve solo definire, oltre a libreria e nome del data-set di output, il numero massimo di iterazioni consentite per l'algoritmo *EM* (par. *5.1*). Anche in questo caso, come già per il metodo *NND*, se vengono individuati più donatori a distanza minima da un dato ricevente, ne viene scelto uno in modo casuale.

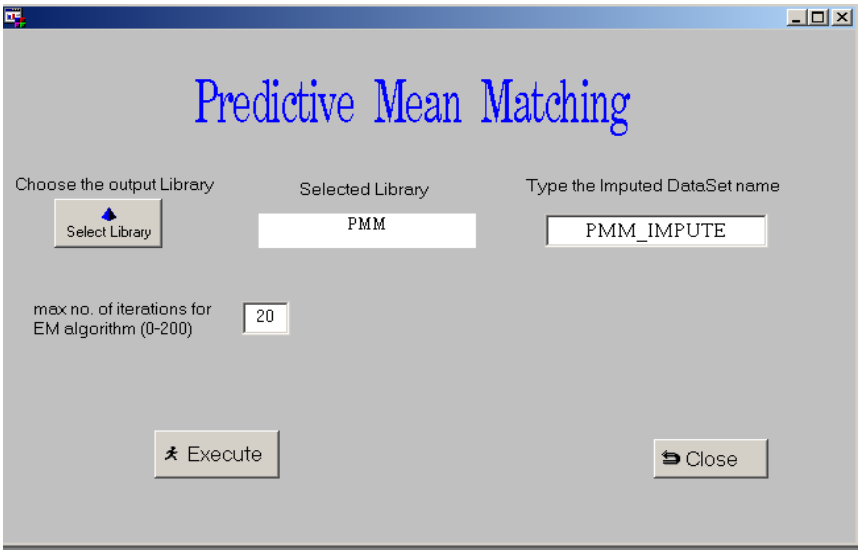


Fig.7 *Maschera perPredictive Mean Matching*

4.5 Imputazione Multipla

Questo metodo è implementato in *QUIS* attraverso la già citata Macro SAS di P.J. Allison. L'imputazione multipla, come è stato ampiamente spiegato, produce, a partire da un data-set con valori mancanti, più data-set completi. Il numero di imputazioni da effettuare (default=5, max=50) è definito dall'utente digitandolo nell'apposito campo. Inoltre devono essere specificati il numero massimo di iterazioni per l'*EM* (par. *2.4*) e per la *Data Augmentation* (sez. 3.). Il programma produce in output un unico data-set costruito concatenando i diversi data-set imputati. La variabile aggiuntiva "DSNUM", che assume valori interi da 1 al numero di imputazioni effettuate, identifica ciascuna imputazione. In tal modo l'utente può svolge le analisi inferenziali di interesse su un solo data-set anziché gestirne tanti quante sono le imputazioni.

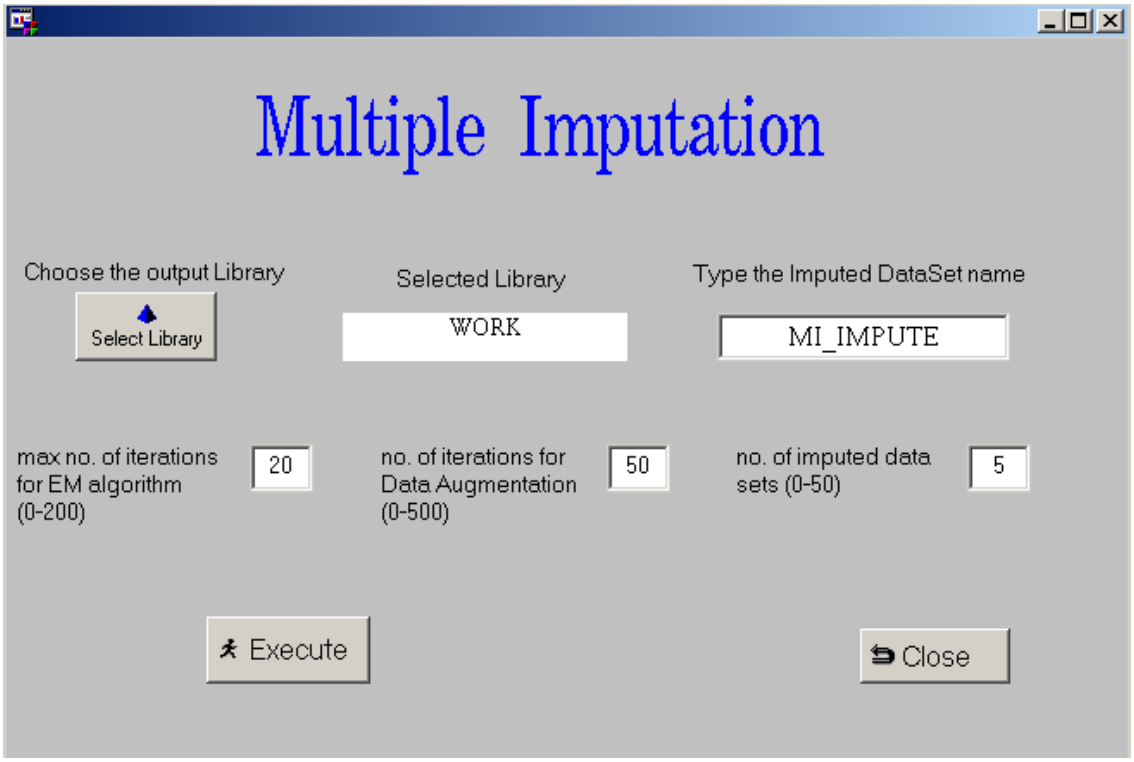


Fig.8 *Maschera perl'Imputazione Multipla*

4.6 Output

I programmi che implementano i diversi metodi vengono posti in esecuzione dalle corrispondenti maschere, mediante il bottone "Execute". La terminata esecuzione viene segnalata dall'attivazione di una finestra con il relativo messaggio, dalla quale, inoltre, si chiede all'utente se egli desideri visualizzare il data-set di output. In caso di risposta affermativa (tasto "*Yes*") viene attivata una finestra nella quale possono essere selezionate le variabili di interesse per visualizzarne i valori nel data-set imputato.

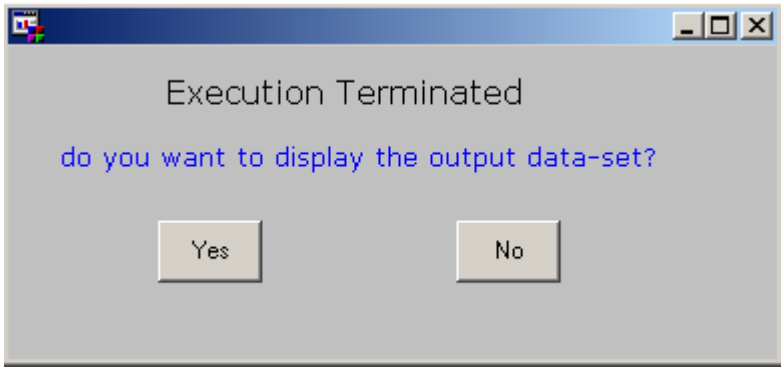
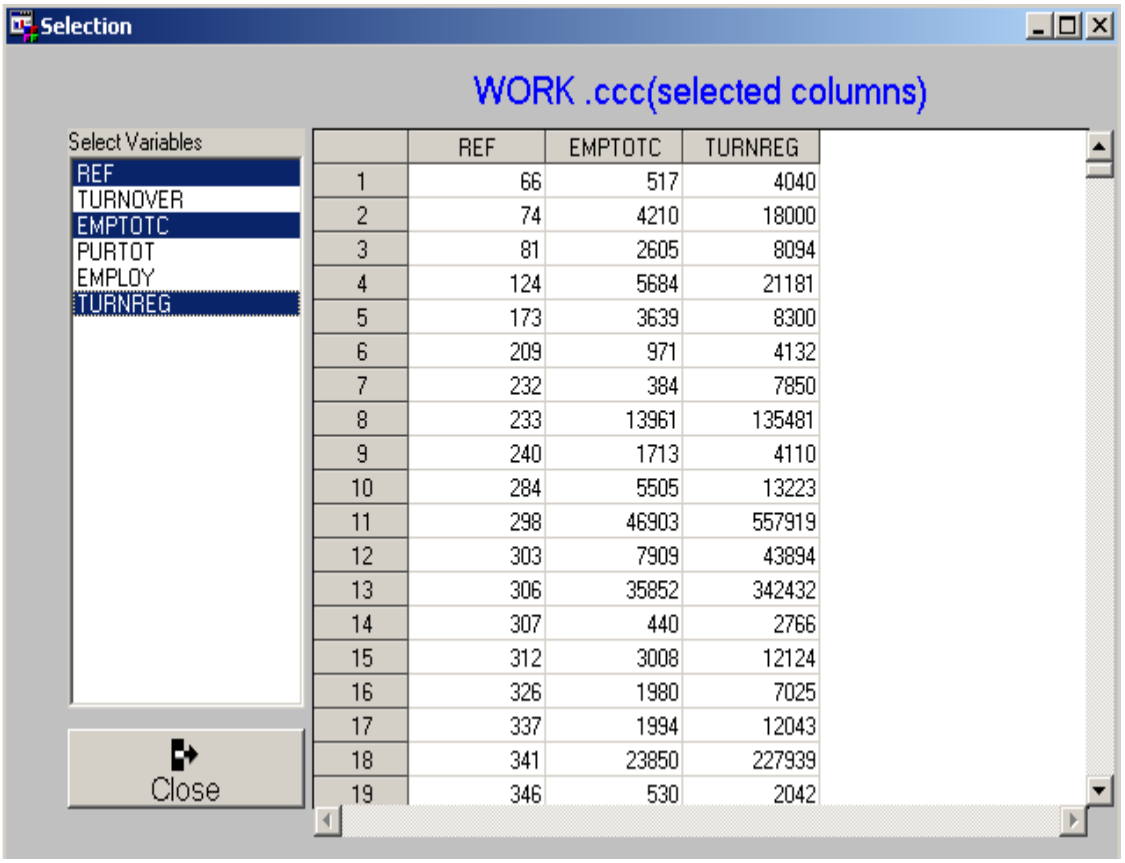


Fig.9 *Messaggio di terminata esecuzione*



	REF	EMPTOTC	TURNREG
1	66	517	4040
2	74	4210	18000
3	81	2605	8094
4	124	5684	21181
5	173	3639	8300
6	209	971	4132
7	232	384	7850
8	233	13961	135481
9	240	1713	4110
10	284	5505	13223
11	298	46903	557919
12	303	7909	43894
13	306	35852	342432
14	307	440	2766
15	312	3008	12124
16	326	1980	7025
17	337	1994	12043
18	341	23850	227939
19	346	530	2042

Fig.10 *Tavola di visualizzazione de data-set di output*

5. Considerazioni conclusive

L'analisi dei dati incompleti, ed in particolare il trattamento delle MRP, è un problema complesso. Con la realizzazione di questo prototipo, cui seguiranno versioni più curate, si è voluto mettere a disposizione dell'utente, anche non in possesso di conoscenze specifiche, uno strumento di semplice utilizzo che gli consenta di applicare alcune delle tecniche più diffuse per il trattamento dei valori mancanti. Tra i metodi inclusi, la *Regressione mediante EM* e l'*Imputazione Multipla* si basano fortemente sull'ipotesi di normalità dei dati, e il *Predictive Mean Matching*, pur essendo probabilmente più robusto rispetto all'allontanamento da tale ipotesi, ne fa comunque uso. E' chiaro pertanto che l'applicazione di queste tecniche potrà fornire risultati soddisfacenti solo qualora preliminarmente venga effettuata una valutazione dell'adeguatezza della distribuzione normale per il "fitting" dei dati da analizzare. D'altra parte il metodo del Donatore di Minima Distanza, che non presuppone l'assunzione di un modello esplicito, richiede uno studio preliminare delle relazioni di associazioni tra le variabili di analisi al fine di una scelta opportuna delle variabili di matching. E'

evidente inoltre che in taluni casi, una buona strategia potrebbe consistere nell'applicazione combinata di diverse tecniche. Ad esempio, nel caso di un'indagine economica sulle imprese, alcune variabili principali (come *Fatturato*, *Numero di Addetti*, *Spese Contributive*, ecc.) potrebbero essere imputate ipotizzando per esse un modello parametrico esplicito e, una volta imputate queste, per un altro insieme di variabili (come *Spese Assicurative*, *Tasse su Immobili*, ecc.), potrebbe essere usata la tecnica del donatore di minima distanza. (Pannekoek, 2002). In alcune situazioni infine, potrebbe essere opportuno applicare uno stesso metodo a domini di studio diversi separatamente, il che corrisponde ad assumere modelli statistici (impliciti o espliciti) differenti per differenti sottoinsiemi di unità.

Con questa breve digressione si è voluto sottolineare come l'applicazione di ciascuno dei metodi descritti costituisca solitamente soltanto una fase dell'intero processo di analisi dei dati incompleti. Si è pertanto ritenuto che l'implementazione di questi metodi in un ambiente elaborativo (SAS) largamente diffuso in ISTAT, e all'interno del quale spesso si sviluppano le altre fasi del processo, presentasse una qualche utilità.

Bibliografia

Box, G.E.P., Tiao, G.C. (1992) *Bayesian Inference in Statistical Analysis*, Wiley & Sons, New York.

Box, G. E. P. and Cox, D. R. (1964) An Analysis of Transformations**,** *Journal of the Royal Statistical Society***,** 211-243.

Chen J., Shao J. (2000) Nearest Neighbour Imputation for Survey Data, *Journal of Official Statistics*, **16**, 113-131.

Cochran W.J. (1977) *Sampling Techniques*. Wiley & Sons, New York.

Cox, D. R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman & Hall, London.

David, M., Little, R.J.A., Samuel, M.E., and Triest, R.K. (1986) Alternative Methods of CPS Income Imputation, *Journal of the American Statistical Assocition*, **81**, 29-41.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38.

D'Orazio, M., Di Zio, M., Scanu, M. (2002) Statistical Matching and Official Statistics, *Quaderni di Ricerca ISTAT*, **1**, 5-24.

Kalton, G. and Kasprzyk, D. (1986) The treatment of Missing Survey Data. *Survey Methodology*, Vol. 12, **1**, 1-16.

Little, R.J.A. (1988) Missing Data Adjustments in Large Survey, *Journal of Business & Economic Statistics*, **6**, 287-295.

Little, R.J.A., Rubin, D.B. (2002) *Statistical Analysis with Missing Data*. Wiley & Sons, New York.

Mardia, K.V., Kent, J.T. and Bibby J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Pannekoek, J. (2002) Multivariate Regression and Hot Deck Imputation Method, *EUREDIT Deliverable 5.1.1.*

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.*.Wiley & Sons, New York.

Schafer J. L., (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Tanner, M.A. and Wong, W.H. (1987) The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, **82**, 528-550.