Valutazione comparativa di alcuni metodi di imputazione singola delle mancate risposte parziali per dati quantitativi

Antonia Manzari (*)

(*) ISTAT - Servizio MTS

e-mail: manzari@istat.it

Riassunto

La mancata risposta parziale è generalmente trattata dagli Istituti nazionali di statistica mediante tecniche di imputazione singola. Particolarmente critico è il caso dell'imputazione di variabili di reddito, in termini sia di analisi e modellizzazione del meccanismo di mancata risposta, sia di preservazione delle proprietà statistiche delle distribuzioni delle variabili oggetto di imputazione. In questo lavoro vengono presentati i risultati di uno studio sperimentale condotto in ISTAT al fine di valutare comparativamente l'accuratezza di diverse tecniche di imputazione singola di variabili di reddito.

INDICE

1 INTRODUZIONE

In ogni tipologia di indagine (sia censuaria sia campionaria) si può presentare l'impossibilità di ottenere tutte le informazioni dalle unità di rilevazione configurandosi il fenomeno che in letteratura è noto come *mancata risposta* o *incompletezza dei dati*.

E' pratica comune distinguere tra:

a) *mancata risposta totale* (*MRT* nel seguito) che si verifica quando nessuna risposta è disponibile per una data unità di rilevazione;
b) *mancata risposta parziale* (*MRP* nel seguito) che ha luogo quando alcune risposte non sono disponibili per una data unità di rilevazione.

La mancata risposta totale può verificarsi, ad esempio, perché l'unità selezionata risulta impossibile da contattare o si rifiuta di partecipare all'indagine. La mancata risposta parziale può verificarsi, ad esempio, per una dimenticanza dell'intervistatore nel porre una domanda o nel registrare una risposta, per incapacità o rifiuto dell'intervistato a fornire una determinata informazione o perché alcune risposte sono state cancellate in fase di *editing* in quanto ritenute errate.

La presenza di mancate risposte provoca dei problemi in fase di analisi dei dati. Tali problemi sono principalmente relativi alla:

1. perdita di efficienza delle stime causata dalla riduzione della dimensione campionaria dei dati completi (gli errori standard sono più elevati, gli intervalli di confidenza sono più ampi e quindi la potenza dei test statistici si riduce);
2. possibile distorsione nelle stime in presenza di una mancata risposta sistematica (i rispondenti sono sistematicamente diversi dai non rispondenti);
3. maggiore difficoltà incontrata per effettuare le analisi sui dati incompleti (i data set incompleti richiedono metodi complessi per la stima dei parametri che potrebbero non essere disponibili nei software statistici solitamente utilizzati per l'analisi dei data set completi).

Diversi metodi sono stati proposti per eliminare o ridurre i problemi associati ai dati incompleti. Generalmente i metodi utilizzati per il trattamento delle *MRT* sono differenti da quelli utilizzati per il trattamento delle *MRP*.

Le *MRT* sono comunemente trattate mediante procedure di riponderazione aventi l'obiettivo di aumentare i pesi di alcune unità rispondenti in modo da rappresentare i non rispondenti (Kalton e Kasprzyk, 1986). Nel caso di *MRT* la riponderazione è un aggiustamento globale che tenta di compensare simultaneamente tutti i valori mancanti (per ciascuna unità rispondente si calcola un

nuovo peso campionario). In linea teorica questo approccio è utilizzabile anche per la compensazione delle *MRP* ma ha l'inconveniente di richiedere tanti insiemi di pesi campionari quante sono le variabili di interesse affette da *MRP* rendendo problematica l'esecuzione di qualunque tipo di analisi multivariata dei dati.

Tra le strategie utilizzate in presenza di *MRP*, particolarmente diffuse sono quelle che consentono di utilizzare gli strumenti statistici standard comunemente disponibili per l'analisi dei dati completi. Tra queste ricordiamo *l'analisi dei casi completi*, *l'analisi dei casi disponibili* e *l'imputazione singola*.

L'analisi dei casi completi (*complete-case* analysis) consiste nell'effettuare le analisi sulle sole unità con osservazioni complete. Si osservi che questa strategia produce serie distorsioni nelle stime a meno che i casi completi 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 probabilità che un valore sia mancante deve essere indipendente dai dati osservati e dai dati non osservati. Inoltre, anche qualora il meccanismo di mancata risposta sia *MCAR*, la riduzione delle unità su cui effettuare le analisi comporta una perdita della precisione delle stime.

L'analisi dei casi disponibili consiste nell'usare il maggiore sottoinsieme di casi disponibili per la stima di parametri distinti (*available-case* analysis). Il maggiore inconveniente di questo approccio riguarda la possibile inconsistenza dei risultati di differenti analisi condotte sullo stesso insieme di dati incompleti a causa del set di variabili usate nelle analisi (ad esempio la stima della matrice di varianza e covarianza potrebbe non essere definita positiva).

L'approccio dell'*imputazione singola* è sicuramente quello più diffuso nella pratica corrente. Esso consiste nell'assegnazione di un valore plausibile a ciascun valore mancante in modo da ottenere un data set completo su cui differenti analisi, effettuate utilizzando strumenti statistici standard, producono risultati consistenti. E' proprio questa caratteristica che rende l'approccio dell'imputazione singola la scelta d'elezione per gli Istituti nazionali di statistica obbligati a fornire data set completi e coerenti (nel senso di rispetto di regole di compatibilità) per uso pubblico. Anche l'approccio dell'imputazione *singola* può presentare degli inconvenienti: ad esempio non è garantito che le stime ottenute dal data set dei dati imputati siano meno distorte di quelle ottenibili dal data set incompleto o che le distribuzioni marginali e congiunte delle variabili rispecchino quelle dell'ipotetico data set completo (molto dipende dalla tipologia di mancata risposta, dalla procedura di imputazione utilizzata, dal tipo di stima). L'inconveniente più evidente è però quello connesso alla non conoscenza dei valori mancanti ed alla applicazione automatica di metodi per l'analisi dei dati completi al data set dei dati imputati: quando i valori imputati sono trattati come se fossero stati effettivamente osservati, la variabilità del meccanismo di non-risposta e la variabilità

aggiuntiva dovuta alle imputazioni non sono tenute in considerazione causando una sottostima degli errori standard delle stime (per una rassegna dei metodi proposti in letteratura per stimare correttamente la varianza delle stime in presenza di imputazione singola si veda ad esempio Rao, 1996).

Diversi metodi di imputazione singola sono stati proposti per l'integrazione delle *MRP* (una descrizione delle caratteristiche più importanti dei metodi comunemente utilizzati insieme ai vantaggi e agli inconvenienti relativi al loro uso è presente in Kalton e Kasprzyk, 1982).

Qualunque sia il metodo prescelto, imputare significa assegnare un valore a ciascun valore mancante. I valori imputati sono generalmente delle stime ottenute mediante una modellizzazione esplicita o implicita delle informazioni disponibili. Un modello esplicito è ad esempio alla base dell'imputazione mediante *regressione* mentre non è altrettanto esplicito il modello che nell'imputazione con il *donatore di minima distanza* (*nearest-neighbour donor*) pone la variabile da imputare in relazione con le variabili ausiliarie utilizzate per l'identificazione del donatore.

In linea di principio ogni analisi effettuata su dati incompleti richiederebbe la specificazione sia della distribuzione di probabilità (modello) dei dati sia della distribuzione di probabilità (meccanismo) della mancata risposta. Per poter effettuare inferenze valide sui parametri del modello assunto per i dati (parametri di interesse) senza specificare esplicitamente il meccanismo di mancata risposta è necessario fare l'assunzione che il meccanismo di mancata risposta sia *Missing at Random* (*MAR*), cioè che la mancata risposta, condizionatamente ai dati osservati, non dipenda dai valori mancanti (Little e Rubin 2002). Si osservi che il meccanismo *Missing Completely at Random* (*MCAR*) che si verifica quando la mancata risposta non dipende dai valori dei dati, mancanti o osservati, è un caso speciale della *MAR*. Se oltre all'assunzione *MAR* si verifica che i parametri del modello dei dati sono distinti dai parametri del meccanismo di mancata risposta, nel senso che lo spazio congiunto dei parametri è il prodotto di due differenti spazi dei parametri (*distinctness*), il meccanismo di mancata risposta è detto *ignorabile* e quindi inferenze non distorte possono essere ottenute sulla base dei soli dati osservati. Nella maggior parte delle applicazioni pratiche la *distinctness* è sempre verificata pertanto la condizione rilevante è la sola assunzione *MAR*. Ecco perché alla base dei metodi più comuni di imputazione delle *MRP* vi è l'assunzione che il meccanismo di mancata risposta sia *MAR*.

Questo lavoro descrive uno studio di valutazione comparativa della qualità di cinque metodi di imputazione singola per variabili quantitative basato su dati reali dall'indagine European Community Household Panel (ECHP). Tre metodi effettuano l'imputazione all'interno di classi individuate dalla costruzione di *alberi di regressione*, il quarto metodo consiste nell'imputazione

con *donatore di minima distanza* (*nearest-neighbour*) mentre il quinto è *l'imputazione da regressione* con residuo casuale.

Lo studio di valutazione descritto in questo lavoro rappresenta una delle prime attività di sperimentazione volte, nell'ambito del progetto EU-SILC, alla individuazione della/e tecnica/che di imputazione dei valori mancanti nelle variabili importi di reddito rilevate dall'indagine. Il Regolamento dell'indagine EU-SILC prevede, infatti, che i valori mancanti nelle variabili importi di reddito devono essere imputati al fine di salvaguardare il massimo dell'informazione rilevata ed evitare che una unità sia esclusa dall'analisi a causa dell'incompletezza dell'informazione fornita, lasciando ai Paesi che partecipano all'indagine la facoltà di scegliere la tecnica di imputazione. I dati utilizzati nelle sperimentazioni sono quelli relativi all'indagine ECHP per l'affinità che essi presentano con le caratteristiche dei dati EU-SILC.

Nella Sezione 2 sono illustrati i metodi a confronto, nella Sezione 3 è descritto lo studio di valutazione, nella Sezione 4 sono presentati i risultati, nella Sezione 5 sono presentati i risultati di analisi aggiuntive. Infine, nella Sezione 6 sono riportate le conclusioni insieme ad alcune considerazioni.

2 METODI DI IMPUTAZIONE

La qualità dei valori imputati, e di conseguenza delle stime ottenute dal data set imputato, dipende fortemente dalla procedura di imputazione utilizzata. E' generalmente riconosciuto che, per predire il valore da imputare, l'utilizzo delle informazioni disponibili sull'unità non rispondente migliora la qualità delle imputazioni in quanto riduce la distorsione dovuta alla non risposta e la variabilità dovuta all'imputazione (Little, 1988; Kalton e Kasprzyk, 1986; Kovar e Whitridge, 1995). Inoltre le misure di associazione (es: covarianze e coefficienti di regressione) tra la variabile affetta da mancata risposta e altre variabili completamente osservate (ausiliarie) risultano non distorte se le variabili ausiliarie sono utilizzate per predire i valori da imputare (Kalton e Kasprzyk, 1982).

Un approccio comunemente usato per cercare di estrarre i valori imputati dalla distribuzione predittiva dei valori mancanti condizionata ai valori osservati è quello di utilizzare le cosiddette *classi di imputazione*. Le classi di imputazione sono sottoinsiemi di unità (rispondenti e non rispondenti) formate in base ai valori di variabili completamente osservate in modo da rendere i valori delle variabili da imputare omogenei all'interno delle classi. Per ciascun non rispondente (*ricevente*) la scelta del rispondente (*donatore*) da cui prelevare i valori è limitata a quelle unità che

appartengono alla stessa classe del ricevente. La costruzione delle classi di imputazione pone il problema sia della selezione delle variabili da utilizzare per definire le classi sia della specificazione delle combinazioni di valori che le definiscono.

Un altro approccio molto comune per predire i valori da imputare condizionatamente ai valori di variabili ausiliare consiste nell'introdurre un concetto di somiglianza tra le unità, basato su un'opportuna funzione di *distanza*, definita sulle variabili ausiliarie. Per ciascun ricevente la scelta del donatore da cui prelevare i valori è limitata a quelle unità che minimizzano la funzione di distanza (*nearest-neighbors)*.

Un terzo approccio, anch'esso molto diffuso, consiste nel regredire la variabile da imputare su un insieme di variabili ausiliarie mediante un modello parametrico esplicito e di utilizzare l'equazione di regressione per predire i valori da imputare (*imputazione da regressione*).

Per tutti i tre metodi (o meglio, le tre classi di metodi) sopra menzionati, che differiscono essenzialmente per il modo in cui sono utilizzate le variabili ausiliarie, si pone il problema della selezione delle variabili ausiliarie da utilizzare. La scelta dovrebbe essere indirizzata verso quelle variabili che spiegano la variabilità delle variabili da imputare (per ridurre la variabilità delle imputazioni) tanto più se sono associate alla mancata risposta (per rendere plausibile l'ipotesi *MCAR* all'interno delle classi ossia per selezionare un donatore che abbia risposte simili al non rispondente) e se saranno successivamente inserite nel modello utilizzato per analizzare i dati (per ottenere stime non distorte delle misure di associazione).

La scelta del metodo da applicare dipende da una serie di fattori tra i quali ricordiamo la natura delle variabili da imputare (qualitative o quantitative), l'esistenza o meno di un modello parametrico che spieghi la relazione tra le variabili, il numero e il livello di dettaglio delle variabili ausiliarie, la numerosità delle unità. A seconda delle situazioni potrebbero essere considerate opportune delle applicazioni combinate di diversi metodi. Ad esempio, l'imputazione da regressione potrebbe essere utilizzata per imputare i valori di alcune variabili continue per le quali risulta adeguata l'ipotesi di un modello parametrico esplicito per descrivere la relazione con le variabili ausiliarie, mentre l'imputazione con il donatore di minima distanza potrebbe essere utilizzata per imputare altre variabili per le quali le relazioni con le variabili ausiliarie non sono facilmente formalizzabili mediante modelli parametrici. Oppure si potrebbero individuare delle classi omogenee rispetto ai valori delle variabili da imputare e utilizzare l'imputazione con il donatore di minima distanza all'interno delle classi.

2.1 Metodi basati sugli alberi di regressione (tree-based methods)

La regressione e la classificazione ad albero sono delle applicazioni dei metodi di *segmentazione binaria* o *partizione ricorsiva* (Sonquist, Baker e Morgan, 1973; Breiman *et al*., 1984). Il loro obiettivo è classificare i dati in termini dei valori di un insieme di variabili esplicative **X** (indipendenti) in sottoinsiemi omogenei rispetto ai valori di una variabile risposta Y (dipendente). La classificazione è ottenuta mediante la partizione progressiva dell'insieme di dati, costituiti dalle coppie (**x**, y), in sottoinsiemi disgiunti, detti nodi, attraverso una sequenza di suddivisioni binarie sulla base di condizioni lineari sui valori delle variabili esplicative. Il criterio per la scelta, a ciascun nodo, della migliore partizione, è basato essenzialmente su una valutazione dell'omogeneità dei valori della variabile risposta *all'interno* dei nodi generati da ogni possibile partizione. I nodi sono pertanto caratterizzati da omogeneità crescente rispetto alla variabile risposta. Le misure di omogeneità si differenziano a seconda che la variabile risposta sia quantitativa (alberi di regressione) o qualitativa (alberi di classificazione). Si distingue tra nodi *non terminali*, cioè nodi che saranno ulteriormente bipartiti in due nodi discendenti e nodi *terminali*, cioè nodi che non sono ulteriormente suddivisi.

Nel contesto dell'imputazione delle *MRP*, gli alberi di regressione/classificazione sono impiegati per ottenere una selezione automatica delle classi di imputazione: i rispondenti sono utilizzati per costruire un albero che "spiega" la distribuzione della variabile risposta in termini dei valori delle variabili esplicative; successivamente i nodi terminali sono usati come classi di imputazione e i valori da imputare per i non rispondenti classificati in un determinato nodo sono estratti, secondo uno specificato metodo di imputazione, dai rispondenti classificati nello stesso nodo.

Nel presente lavoro le imputazioni basate sugli alberi di regressione sono state effettuate mediante il software *Weighted Automatic Interaction Detection* (*WAID*) (Chambers *et al*, 2001; De waal *et al*, 2001). Il software *WAID* è stato sviluppato nell'ambito del progetto europeo AutImp (http://www.cbs.nl/en/services/autimp/autimp.htm) e consente di costruire alberi di regressione/classificazione e di effettuare imputazioni di variabili quantitative e qualitative all'interno delle classi individuate dagli alberi costruiti.

Nel software *WAID* gli alberi sono costruiti utilizzando variabili esplicative che possono essere di tipo esclusivamente qualitativo (le variabili quantitative devono essere preliminarmente raggruppate in classi). La misura di omogeneità utilizzata per la costruzione degli alberi di regressione è la somma pesata dei quadrati dei residui (WSSR) definita rispetto ad una misura di posizione dei valori della variabile risposta nel nodo:

$$WSSR_k = \sum_{i=1}^{n_k} w_i (y_i - \bar{y}_{wk})^2$$

ove w_i è il peso dell'i-esima unità nel nodo k e

$$\bar{y}_{wk} = \sum_{i=1}^{n_k} w_i y_i \Big/ \sum_{i=1}^{n_k} w_i$$

è la media pesata della variabile risposta nel nodo k.

Differenti sistemi di pesi possono essere selezionati in modo da ottenere una misura di posizione robusta rispetto alla presenza di valori anomali.

Il principio usato per la costruzione dell'albero di regressione è di seguito sintetizzato. A parte il nodo iniziale, il nodo prescelto per la suddivisione (split) è quello con il più grande valore di WSSR. Supponiamo di avere a disposizione un insieme di s variabili esplicative, la miglior suddivisione per ciascuna variabile X_m, m=1,…,s è definita dal sottoinsieme di valori di X_m che genera nodi con valori di WSSR minore di quelli ottenuti con qualunque altra suddivisione dei valori di X_m. La miglior suddivisione complessiva è la migliore (in termini di minimizzazione del valore di WSSR) tra quelle definite migliori per ciascuna X_m. Il processo di crescita dell'albero si interrompe nel momento in cui intervengono una o più regole di arresto. Le regole di arresto riguardano fondamentalmente il numero minimo di unità presenti all'interno di un nodo e la numerosità massima dei nodi terminali (dimensione dell'albero).

Tre metodi di imputazione per variabili quantitative sono implementati nel software *WAID*: l'imputazione con *selezione casuale del donatore*, l'imputazione con *donatore di minima distanza* e l'imputazione con la *media*.

Nell'imputazione con *selezione casuale del donatore* (*Tree-RS* nel seguito), per ciascun non rispondente classificato in un determinato nodo terminale, si estrae casualmente un donatore dallo stesso nodo e si usa il suo valore della variabile risposta per imputare il corrispondente valore mancante del non rispondente.

Nell'imputazione con *donatore di minima distanza* (*Tree-NN* nel seguito), per ciascun non rispondente classificato in un determinato nodo terminale prima si calcola la distanza totale tra esso e ciascun rispondente entro il nodo. Successivamente si estrae il donatore che minimizza la funzione di distanza totale e si usa il suo valore della variabile di risposta per imputare il corrispondente valore mancante del non rispondente. Se più di un rispondente minimizza la funzione di distanza totale, il metodo effettua una selezione casuale tra quelli con distanza minima. La distanza totale è la somma delle distanze rispetto a tutte le variabili esplicative che definiscono l'albero. Per ciascuna variabile esplicativa la funzione di distanza è pari a 0 se il valore della

variabile esplicativa nel non rispondente è uguale al valore posseduto dal rispondente, ed è uguale a 1 altrimenti.

Nell'imputazione con la *media* (*Tree-Mean* nel seguito), per ciascun non rispondente classificato in un determinato nodo terminale, si imputa la media aritmetica della variabile risposta calcolata su tutti i rispondenti presenti nel nodo.

2.2 Imputazione con donatore di minima distanza (nearest-neighbour)

L'imputazione con *donatore di minima distanza* (*nearest-neighbors*) (*NN* nel seguito) seleziona il donatore da cui prelevare i valori tra le unità più "vicine" al ricevente. La vicinanza è definita in termini di una misura di distanza multivariata tra il donatore e il ricevente basata sulle variabili ausiliarie **X**. Le variabili da imputare e le variabili ausiliarie possono essere di natura sia quantitativa che qualitativa.

La funzione di distanza totale è data dalla somma delle distanze elementari rispetto a tutte le variabili ausiliarie. Differenti funzioni di distanza sono state proposte in letteratura (es: Sande, 1979). Si osservi che le distanze elementari sono generalmente calcolate su variabili ausiliarie preliminarmente standardizzate al fine di evitare che i diversi contributi alla funzione di distanza totale dipendano eccessivamente dall'unità di misura delle variabili. Nel caso in cui, per un dato ricevente, esistano più unità con la stessa distanza totale minima, il donatore viene selezionato in modo casuale tra queste.

Un sistema di pesi può essere utilizzato per assegnare diversa importanza alle singole variabili nel computo della distanza totale.

Un fattore di penalizzazione associato a ciascun utilizzo di uno stesso donatore può essere introdotto per ridurre il molteplice utilizzo di un donatore (ed evitare una distorsione nella distribuzione finale causata dalla sovrarappresentazione delle risposte provenienti da uno stesso donatore).

L'intero insieme di dati può essere considerato come un'unica classe di imputazione oppure è possibile suddividere le unità in strati distinti, definiti dai valori di variabili qualitative, all'interno dei quali eseguire l'imputazione *NN*. In questo caso per ciascun ricevente classificato all'interno di uno strato si seleziona il donatore più vicino all'interno dello stesso strato (analogamente a quanto effettuato dal metodo *Tree-NN*).

Può capitare che nel record ricevente una variabile ausiliaria, utilizzata per definire la funzione di distanza, presenti valore mancante. In questo caso la variabile viene esclusa dal computo della

distanza totale. In pratica, la distanza elementare per una variabile ausiliaria con valore mancante non può essere calcolata e si assume quindi che il suo contributo alla distanza totale sia zero. In questo modo si ammette l'assenza del valore per le variabili utilizzate per definire la funzione di distanza, ossia per le variabili utilizzate per la selezione del record donatore. Si osservi che il valore mancante non è ammesso per le variabili ausiliarie utilizzate per definire gli strati, ossia per le variabili utilizzate per costruire i serbatoi dei donatori: le variabili di stratificazione devono essere osservate su tutte le unità (riceventi e donatori). E' inoltre opportuno osservare che una variabile ausiliaria può avere valore mancante solo nei record riceventi e non nei record che compongono il serbatoio dei donatori: i donatori sono tutti e soli i record in cui sono osservate tutte le variabili ausiliarie (usate per definire la distanza o gli strati) e le variabili da imputare.

Nel presente lavoro il metodo di imputazione *NN* è stato applicato mediante procedure SAS realizzate dal Dott. Ugo Guarnera.

2.3 Imputazione da regressione con residuo casuale

Il metodo dell'*imputazione da regressione* prevede la specificazione di un modello esplicito di regressione che spieghi la relazione tra le variabili ausiliarie **X** e la variabile da imputare Y ed utilizza l'equazione di regressione $E(\mathrm{Y}) = f(\mathbf{X})$ per predire i valori da imputare. La variabile da imputare deve essere di natura quantitativa mentre le variabili ausiliarie possono essere sia quantitative che qualitative (nell'ultimo caso si utilizzano delle variabili *dummies*).

Il generico valore imputato può essere direttamente il valore atteso condizionato (*predicted value*) o il valore atteso condizionato aumentato di un residuo casuale, estratto per riflettere l'incertezza nel valore atteso condizionato, (*imputazione da regressione con residuo casuale* o *stocastic regression imputation, SRI* nel seguito). Diversi modi sono stati suggeriti in letteratura per la determinazione dei residui (Kalton e Kasprzyk, 1986; Little, 1988). L'aggiunta di un residuo è effettuata quando si vuole preservare la distribuzione e la variabilità dei dati imputati. A questo riguardo si osservi che la scelta di aggiungere o no un residuo al valore atteso condizionato dipende dalle analisi che devono essere condotte sui dati imputati. Se l'obiettivo è quello di stimare quantità lineari dei dati come medie o totali, l'aggiunta di un residuo potrebbe causare una perdita di precisione nelle stime (aumento della variabilità). Al contrario se l'obiettivo è stimare parametri distribuzionali o parametri legati alle relazioni di interdipendenza tra le variabili, il solo valore atteso condizionato potrebbe comportare forti distorsioni e quindi l'aggiunta di un residuo è, in questi casi, la scelta obbligata.

La *regressione con residuo casuale* è un metodo frequentemente utilizzato per l'imputazione delle variabili quantitative. In particolare è il metodo adottato dalla Banca d'Italia per l'imputazione dei dati mancanti dell'indagine campionaria sui bilanci delle famiglie italiane (Banca d'Italia, 1993, pag.19) e da Eurostat per l'imputazione dei dati mancanti delle variabili di reddito dell'indagine ECHP (Eurostat, 2002).

Nel presente lavoro il metodo di imputazione *SRI* è stato applicato mediante il modulo Impute del software *IVEware*. *IVEware* è un *freeware* (http://www.isr.umich.edu/src/smp/ive/) sviluppato dal *Survey Research Center, Institute for Social Research dell'University of Michigan*, composto da Macro SAS e procedure C e FORTRAN, che effettua l'imputazione *singola* o *multipla*[1] dei valori mancanti, per variabili sia quantitative sia qualitative, usando il metodo delle regressioni sequenziali (SRMI) descritto nel lavoro di Raughunatan *et al*. (2001).

Il metodo SRMI effettua l'imputazione multivariata (numero di variabili da imputare ≥ 2) dei valori mancanti mediante una sequenza di modelli di regressione multipla univariata (approccio "*variable by variable*"). Ciascun modello descrive la relazione tra la variabile da imputare con tutte le altre variabili e utilizza sia i valori osservati sia i valori imputati nei passi precedenti. Il singolo modello è selezionato separatamente senza bisogno di definire un modello multivariato per l'intero insieme dei dati di cui, però, si assume l'esistenza. A seconda della natura della variabile da imputare è selezionato un differente modello di regressione tra quelli disponibili. Ad esempio per le variabili continue si usa il modello di regressione lineare normale (la variabile deve essere preventivamente trasformata se non è normale nella scala originaria, e dopo l'imputazione si procede alla trasformazione inversa per ottenere i valori nella scala originaria) mentre per le variabili binarie si usa il modello di regressione logistico.

Si osservi che il metodo SRMI è un metodo ideato per eseguire l'imputazione *multipla* dei dati che può, a detta degli autori, essere usato per eseguire anche l'imputazione *singola* (basta prendere una sola esecuzione della procedura). Le imputazioni ottenute sono definite come estrazioni dalla distribuzione predittiva a posteriori dei valori mancanti, condizionatamente ai valori osservati,

[1] L'Imputazione Multipla è un metodo per l'analisi di dati incompleti che si basa su m imputazioni ripetute estratte dalla distribuzione predittiva a posteriori dei dati mancanti condizionatamente ai dati osservati, o su una sua approssimazione. Ciascun valore mancante è sostituito da m ($m>2$) valori in modo da ottenere m insiemi di dati completi. Ciascun *data set* completo è analizzato con procedure standard per dati completi. I risultati delle m analisi sono successivamente combinati in un'analisi finale secondo le formule di stima riportate in Rubin (1987). L'obiettivo è ottenere stime puntuali e intervallari dei parametri che tengano in considerazione oltre alla variabilità campionaria anche la variabilità del meccanismo di mancata risposta e la variabilità aggiuntiva dovuta alle imputazioni.

specificata dal modello di regressione con una distribuzione a priori non informativa dei parametri del modello. Utilizzare il metodo SRMI per eseguire l'imputazione *singola* dei valori mancanti di una sola variabile Y dovrebbe essere equivalente ad effettuare una imputazione da regressione con residuo casuale ove il modello di regressione utilizzato è quello specificato per la variabile Y. La differenza con una imputazione da regressione "standard" risiede nel fatto che i parametri che compaiono nella distribuzione predittiva a posteriori dei valori mancanti, condizionatamente ai valori osservati, sono a loro volta generati dalla propria distribuzione a posteriori a dati osservati anziché essere semplicemente le stime di massima verosimiglianza.

Il software *IVEware* è stato selezionato perché è quello utilizzato da Eurostat per eseguire l'imputazione singola delle variabili di reddito dell'indagine ECHP (Eurostat, 2002).

3 STUDIO DI VALUTAZIONE

Questa Sezione descrive lo studio predisposto ed eseguito per la valutazione comparativa dell'accuratezza dei cinque metodi descritti nella Sezione 2 nell'imputazione di una variabile quantitativa.

Quantificare l'accuratezza di una procedura di imputazione significa misurare la vicinanza tra i valori "veri" effettivamente posseduti dalle unità statistiche e i valori imputati con la procedura di imputazione. Poiché i valori "veri" corrispondenti ai valori mancanti non sono generalmente disponibili, lo studio è stato condotto simulando i valori mancanti: ciascun metodo di imputazione è stato applicato ad un insieme di valori reali osservati (provenienti da unità rispondenti) e artificialmente resi mancanti mediante un meccanismo *MAR*. Successivamente opportuni indicatori di accuratezza sono stati calcolati confrontando i valori imputati con i corrispondenti valori osservati.

Lo studio si compone di diverse fasi di seguito brevemente descritte:

1. *Selezione dell'insieme dei dati e delle variabili*.
2. *Generazione dei valori mancanti*.
3. *Applicazione dei metodi di imputazione*.
4. *Calcolo degli indicatori*.

3.1 Selezione dell'insieme dei dati e delle variabili

Obiettivo della sperimentazione è confrontare l'accuratezza dei metodi nell'imputazione di una variabile quantitativa di reddito. Sono stati utilizzati i dati reali (finali) italiani provenienti da una fase dell'indagine ECHP.

La variabile PI111 *Reddito da lavoro dipendente (annuale netto)* è stata selezionata come variabile da imputare (variabile di risposta, Y). Per ciascun valore della variabile PI111 una variabile indicatrice informa se il valore presente nel data set è un valore effettivamente osservato o se è un valore imputato da Eurostat mediante la procedura di imputazione eseguita a livello centralizzato per i dati provenienti da tutti i Paesi membri (Eurostat, 2002).

Inizialmente sono stati selezionati 6457 record aventi un valore diverso da zero per la variabile di risposta PI111 (sono stati esclusi i record aventi PI111 uguale a zero in quanto tale valore significava che il reddito non era stato percepito dall'unità). Per 697 record (10.8%) il valore della variabile PI111 era un valore imputato da Eurostat mentre per i rimanenti 5760 record il valore della variabile PI111 era un valore effettivamente osservato. Poiché si vuole valutare la capacità dei metodi di imputazione di ripristinare i valori veri, solo il set di 5760 record è stato considerato valido per lo studio e sottoposto alla procedura di generazione dei valori mancanti descritta nella sottosezione 3.1.

La variabile di risposta PI111 è ottenuta dall'aggregazione di differenti componenti. Eurostat procede ad imputare le componenti delle variabili di reddito in diversi passi, alcune componenti sono imputate individualmente, altre sono raggruppate insieme prima dell'imputazione. La procedura di imputazione delle componenti del *Reddito da lavoro dipendente* prevede che la trasformata logaritmica degli ammontari di reddito sia imputata mediante il software *IVEware* utilizzando un set di dodici variabili ausiliarie (Eurostat, 2002, pp. 80-81). Le variabili ausiliarie utilizzate da Eurostat sono: *Reddito familiare netto mensile* (in classi) (RANGREV, 9 classi), *Regione* (REGION, 12 modalità), *Numero di componenti intervistati occupati* (NBWCL, 3 modalità), *Età in anni* (in classi) (AGECLA, 6 classi), *Sesso* (GENDER, 2 modalità), *Titolo di studio* (EDUC, 5 modalità), *Attività lavorativa principale* (OCCUP, 5 modalità), *Settore di attività economica* (INDUST, 5 modalità), *Numero di lavoratori nell'Ente o Azienda* (ENTSIZ, 8 modalità), *Posizione nella professione* (STEMP, 5 modalità), Ore lavorate a settimana (in classi) (WORKHH, 5 classi), *Ruolo gerarchico* (SUPERV, 3 modalità). Per la descrizione dettagliata delle modalità e della codifica adottata si veda Eurostat, 2001 e 2002.

Poiché la selezione delle variabili ausiliarie da utilizzare nelle procedure di imputazione, che richiede un'analisi dei dati oltre alla conoscenza del fenomeno e degli obiettivi dell'indagine, non è

oggetto di questo studio, si è preferito avvalersi del lavoro svolto dagli esperti Eurostat. Pertanto, le dodici variabili ausiliarie selezionate da Eurostat per l'imputazione delle componenti della variabile *Reddito da lavoro dipendente* sono state utilizzate come variabili ausiliarie (**X**) nei metodi di imputazione posti a confronto.

3.2 Generazione dei valori mancanti

Per ciascuna delle 6457 unità del data set iniziale, la variabile indicatrice di mancata risposta R_i (i=1,…,6457) informa se il valore della variabile Y è stato osservato o no:

$$R_i \begin{cases} 0 \text{ se } Y_i \text{ è osservato} \\ 1 \text{ se } Y_i \text{ è non osservato (e quindi è un valore imputato)} \end{cases}$$

Questa informazione consente di verificare l'assunzione MCAR rispetto alle 12 variabili ausiliarie. A tal fine si è provveduto a regredire R_i sulle 12 variabili ausiliarie mediante un modello di regressione logistica ed a selezionare il migliore sottoinsieme di variabili tramite una procedura di selezione *backward* (livello di significatività del Wald chi-square per la rimozione delle variabili dal modello = 0.05). Sei variabili sono state trattenute nel modello: RANGREV, REGION, OCCUP, INDUST, STEMP e SUPERV. La Tabella 1 riporta i gradi di libertà, la statistica di Wald e i p-value per ciascuna delle sei variabili:

Tabella 1. Analisi delle variabili trattenute nel modello

	DF	**Wald Chi-Square**	**Pr>Chi-Square**
RANGREV	8	59.0731	<.0001
REGION	11	73.5059	<.0001
OCCUPCLA	4	11.1910	0.0245
INDUST	4	13.0630	0.0110
STEMP	4	156.9542	<.0001
SUPERV	2	14.0101	0.0009

Il risultato ottenuto fornisce indicazioni sulla esistenza di una relazione tra la probabilità di osservare una mancata risposta per la variabile Y e i valori assunti dalle sei variabili trattenute nel modello, in altre parole, c'è evidenza che la probabilità di osservare una mancata risposta "non è indipendente" dai valori osservati per le sei variabili ausiliarie trattenute nel modello.

Nella generazione dei valori mancanti da utilizzare per lo studio di valutazione, si è voluto tenere conto dell'indicazione riscontrata nei dati reali e si è provveduto a simulare un meccanismo di mancata risposta, per i valori della variabile Y, che dipendesse dai valori delle sei variabili ausiliarie della Tabella 1. Il meccanismo di mancata risposta *MAR* simulato nello studio è di seguito descritto.

Per le 5760 unità selezionate, la probabilità di mancata risposta per la variabile Y è stata specificata mediante una funzione logistica delle 33 variabili dummies (**Z**) definite per le sei variabili ausiliarie della Tabella 1:

$$(1) \qquad P(R_i = 1 \mid \mathbf{Z}) = \frac{\exp\left(\beta_0 + \sum_{m=1}^{33} \beta_m z_{im}\right)}{1 + \exp\left(\beta_0 + \sum_{m=1}^{33} \beta_m z_{im}\right)}$$

Nella (1) β_0 indica l'intercetta, β_m indica il coefficiente della generica variabile dummy (z_m) mentre z_{im} indica il valore assunto dalla variabile dummy z_m nell'unità i.

Il processo di generazione dei valori mancanti è stato infine realizzato mediante la selezione di un campione di 634 unità (corrispondenti all'11% di 5760) con una probabilità di selezione proporzionale alla probabilità di mancata risposta, per la variabile Y, stimata con il modello (1)[2]. Per ciascuna unità inclusa nel campione selezionato si è proceduto a cancellare il valore osservato per la variabile Y, simulando quindi una mancata risposta. In questo modo il data set complessivo di 5760 unità è stato diviso in due gruppi, quello dei rispondenti composto da 5126 unità e quello dei non rispondenti (artificiali) composto da 634 unità.

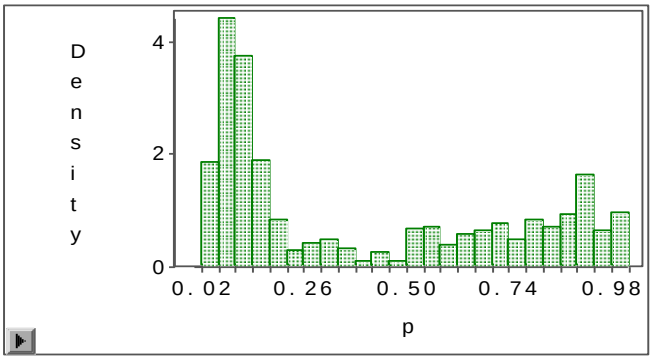
I valori dei coefficienti β_0, β_1, …, β_{33} utilizzati nella funzione (1) sono stati scelti in modo tale da ottenere una proporzione di unità con probabilità di mancata risposta elevata avendo cura che la distribuzione della stessa risultasse differenziata per i due gruppi di rispondenti e non rispondenti ma che contemplasse, in entrambi i gruppi, valori bassi e valori alti (vedi Figura 1). Il razionale di questa scelta risiede nel volere evitare che una eccessiva omogeneità rispetto alla probabilità di mancata risposta (solo bassa per i rispondenti e solo elevata per i non rispondenti) si riflettesse nei valori delle variabili ausiliarie rendendo l'imputazione una operazione di "estrapolazione". In pratica, le stime dei coefficienti β_0, β_1, …, β_{33} ottenute dalla regressione logistica di R_i sulle sei variabili ausiliarie della Tabella 1 sulle 6457 unità del data set iniziale, sono state utilizzate come

[2] La selezione delle 634 unità è stata effettuata utilizzando la *PROC SURVEYSELECT* del SAS v. 8.1 secondo lo schema del campionamento con probabilità proporzionali ad una data misura d'ampiezza.

valori iniziali nella funzione (1) e successivamente modificati in modo da ottenere le distribuzioni in Figura 1.

Figura 1. Distribuzione della probabilità di mancata risposta per i rispondenti e per i non rispondenti

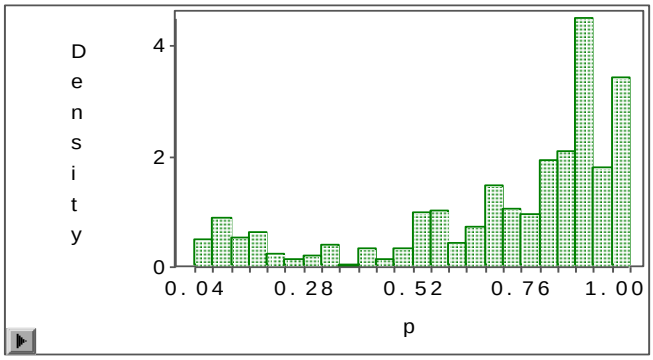
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	0.4018	Sum	2059.7301
Std Dev	0.3403	Variance	0.1158
Skewness	0.4945	Kurtosis	-1.4265
USS	1421.0396	CSS	593.3985
CV	84.6827	Std Mean	0.0048

Quantiles			
100% Max	1.0000	99.0%	0.9955
75% Q3	0.7503	97.5%	0.9910
50% Med	0.1978	95.0%	0.9644
25% Q1	0.0998	90.0%	0.9089
0% Min	0.0219	10.0%	0.0630
Range	0.9781	5.0%	0.0573
Q3-Q1	0.6505	2.5%	0.0522
Mode	0.1091	1.0%	0.0474

Non rispondenti



Moments			
N	634.0000	Sum Wgts	634.0000
Mean	0.7218	Sum	457.6475
Std Dev	0.2669	Variance	0.0712
Skewness	-1.1795	Kurtosis	0.2744
USS	375.4417	CSS	45.0928
CV	36.9751	Std Mean	0.0106

Quantiles			
100% Max	1.0000	99.0%	0.9995
75% Q3	0.9089	97.5%	0.9982
50% Med	0.8320	95.0%	0.9945
25% Q1	0.5987	90.0%	0.9837
0% Min	0.0474	10.0%	0.1978
Range	0.9525	5.0%	0.1192
Q3-Q1	0.3102	2.5%	0.0911
Mode	0.9089	1.0%	0.0630

3.3 Applicazione dei metodi di imputazione

I valori mancanti generati nel passo precedente sono stati imputati mediante i metodi descritti nella Sezione 2.

3.3.1 Metodi basati sugli alberi di regressione (Tree-RS, Tree-NN e Tree-Mean)

Il software *WAID* obbliga ad utilizzare esclusivamente variabili ausiliarie di tipo qualitativo. Pertanto tutte le variabili sono state considerate strettamente qualitative. Inoltre, poiché l'obiettivo era imputare esclusivamente la variabile di risposta PI111, i valori mancanti presenti nelle variabili

ausiliarie sono stati codificati secondo la codifica adottata da Eurostat (*WAID* non ammette mancate risposte per le variabili che non devono essere imputate) e trattati come una modalità.

Nel processo di costruzione dell'albero di regressione è stato utilizzato il criterio dei minimi quadrati ordinari (OLS) che consiste nell'attribuire all'i-esima unità nel nodo k un peso unitario w_i=1. In questo modo la misura di posizione dei valori della variabile Y nel nodo k è la media aritmetica calcolata sui rispondenti.

Il numero minimo di unità presenti all'interno di un nodo non può essere utilizzato come regola di arresto nel processo di costruzione di un albero di regressione (può essere utilizzato solo per la costruzione di alberi di classificazione). Pertanto, l'unica regola di arresto utilizzata è stata la numerosità massima dei nodi terminali (dimensione dell'albero). Nel presente lavoro sono state considerate tre differenti dimensioni (10, 30, 40) e differenti imputazioni sono state eseguite per ciascuna dimensione. Si è voluto in tal modo rilevare le eventuali differenze nelle prestazioni delle procedure di imputazione al variare delle dimensioni dell'albero.

3.3.2 Imputazione con donatore di minima distanza (NN)

Tutte le variabili ausiliarie sono state considerate qualitative (analogamente a quanto effettuato nei metodi basati sugli alberi di regressione) secondo la codifica adottata da Eurostat. Pertanto la distanza elementare tra l'unità u_i e l'unità u_j è stata definita uguale a:

$$D_q(u_i, u_j) = \begin{cases} 0 & if \quad x_i = x_j \\ 1 & if \quad x_i \neq x_j \end{cases}$$

Un peso unitario è stato assegnato a ciascuna variabile nel computo della distanza totale.

Nessun fattore di penalizzazione né variabili di strato sono stati utilizzati.

Si osservi che il metodo *NN* implementato nello studio considera l'intero insieme di dati come un'unica classe di imputazione diversamente dal metodo *Tree-NN* che effettua l'imputazione con *donatore di minima distanza* all'interno delle classi (nodi terminali) individuate dalla struttura ad albero.

3.3.3 Imputazione da regressione con residuo casuale (SRI)

Il logaritmo naturale della variabile risposta PI111 è stato imputato usando il modello di regressione lineare normale. La trasformazione logaritmica è stata necessaria per "normalizzare" la distribuzione dei valori (la distribuzione della variabile PI111 nella scala originaria si discostava dalla normalità). Dopo l'imputazione si è proceduto ad effettuare la trasformazione inversa per ottenere i valori nella scala originaria. Tutte le variabili ausiliarie sono state considerate qualitative (con la codifica adottata da Eurostat) ad eccezione della variabile *Età in anni* che non è stata raggruppata ma è stata considerata quantitativa (variabile AGE).

3.4 Calcolo degli indicatori

L'accuratezza di ciascun metodo di imputazione è stata valutata misurando l'effetto prodotto sui valori originali, sulla distribuzione marginale dei valori originali e su alcuni parametri della distribuzione dei valori originali. A tal fine i seguenti indicatori sono stati calcolati sul sottoinsieme di valori imputati (n=634):

<u>Preservazione dei valori originali</u>

- *Deviazione assoluta media* tra i valori imputati Y'_i e i corrispondenti valori originali Y_i:

$$diff_Y = \Sigma_i^n |Y'_i - Y_i| / n$$

  $diff_Y$ è un indicatore della distanza tra i valori imputati e i valori originali. Assume valore minimo (0) solo nel caso in cui i valori imputati sono uguali ai valori originali. Tanto maggiore è il valore assunto dall'indice, tanto maggiore è la distanza fra i valori Y' e Y in termini di numero di valori diversi e/o entità delle differenze.
- *Coefficiente di regressione* (*coeff*) e *Indice di determinazione* (R^2) ottenuti dalla regressione dei valori originali Y_i sui valori imputati Y'_i secondo il modello lineare $Y = \beta Y' + \varepsilon$. Se il valore di *coeff* (stima di β) è prossimo ad 1, indicando che il metodo non introduce distorsioni sistematiche nelle imputazioni, e il valore di R^2 è prossimo ad 1, indicando che la maggior parte della variabilità dei valori veri è spiegata dai valori imputati, il metodo di imputazione preserva i valori originali.

Preservazione della distribuzione marginale dei valori originali

- *Distanza di Kolmogorov-Smirnov* tra la funzione di ripartizione empirica dei valori imputati Y'_i e la funzione di ripartizione empirica dei valori originali Y_i:

$$d_{KS}[F_n(Y'), F_n(Y)] = \max |F_n(Y') - F_n(Y)|$$

  Questo indicatore assume valore zero solo quando le due distribuzioni sono identiche.

Preservazione dei parametri della distribuzione dei valori originali

- *Differenza assoluta* tra la *media* calcolata sui valori imputati e la *media* calcolata sui valori originali:

$$diff_m = \left|\overline{Y'} - \overline{Y}\right|.$$

  $diff_m$ è una misura della distanza tra la *media* dei valori imputati e la *media* dei valori originali.
- *Differenza assoluta* tra la *deviazione standard* calcolata sui valori imputati e la *deviazione standard* calcolata sui valori originali:

$$diff_\sigma = \left|\sigma_{Y'} - \sigma_Y\right|.$$

  $diff_\sigma$ è una misura della distanza tra la *deviazione standard* dei valori imputati e la *deviazione standard* dei valori originali.

Per avere una indicazione del segno della distorsione sono state calcolate anche le corrispondenti differenze semplici:

$$dist_m = \overline{Y'} - \overline{Y}$$

$$dist_\sigma = \sigma_{Y'} - \sigma_Y$$

Per tenere conto della variabilità del meccanismo di mancata risposta, e cercare quindi di evitare gli effetti dovuti alla selezione di un particolare campione di valori da cancellare, la sequenza dei processi *generazione dei valori mancanti - applicazione dei sistemi di imputazione - calcolo degli indicatori* è stata replicata cinque volte. Ad ogni replica un nuovo campione di 634 valori è stato cancellato dal data set complessivo di 5760 valori (in Appendice A sono riportate le distribuzioni dei valori originali della variabile risposta nei rispondenti e nei non rispondenti per ciascuna delle cinque repliche). Successivamente i valori cancellati sono stati imputati mediante le procedure di imputazione a confronto. Infine, per ciascun campione e ciascun metodo di imputazione sono stati calcolati i valori degli indicatori (riportati in Appendice B). Dalle Tabelle in Appendice B si evince che i valori degli indicatori (specialmente quelli relativi alla preservazione degli aggregati)

mostrano una variabilità non trascurabile sui campioni di non rispondenti. Questa variabilità è dovuta sia al meccanismo di mancata risposta sia alla casualità delle imputazioni. Una valutazione comparativa più accurata potrebbe essere effettuata tenendo in maggiore considerazione le due fonti di variabilità (elevando il numero di campioni ed eseguendo ripetizioni delle imputazioni con componente casuale per ciascun campione). Si ritiene comunque che i valori medi degli indicatori sui cinque campioni diano utili indicazioni sull'accuratezza comparativa dei metodi.

4 RISULTATI

La Tabella 2 riporta, per ogni metodo di imputazione, i valori medi degli indicatori sui cinque campioni (i valori degli indicatori calcolati per ciascuna replica e ciascun metodo di imputazione sono riportati in Appendice B).

Tabella 2. Valori medi (sui cinque campioni) degli indicatori di accuratezza dei metodi di imputazione per la variabile *Reddito da lavoro dipendente*

Metodo	Nodi	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$
Tree-RS	10	9330.119	0.827	0.691	0.052	400.124	603.205
	30	8541.762	0.859	0.729	0.045	291.306	666.634
	40	8351.999	0.859	0.734	0.046	222.564	959.462
Tree-NN	10	7527.030	0.875	0.761	0.046	435.616	1269.516
	30	7409.695	0.881	0.771	0.044	390.729	868.541
	40	7397.979	0.876	0.774	0.040	528.126	933.071
Tree-Mean	10	6708.599	1.000	0.813	0.199	377.710	5261.845
	30	6270.942	0.979	0.828	0.156	298.013	3865.014
	40	6196.396	0.971	0.831	0.151	233.706	3405.954
NN	-	7364.444	0.904	0.773	0.048	317.448	1305.832
SRI	-	10936.856	0.688	0.639	0.181	942.275	5315.717

I valori riportati nella Tabella 2 indicano che l'accuratezza del metodo di imputazione da *regressione con residuo casuale* (*SRI*) è inferiore a quella degli altri metodi rispetto a tutti i criteri di valutazione considerati.

Riguardo ai metodi *basati sugli alberi di regressione* (*Tree-RS, Tree-NN e Tree-Mean*), i risultati ottenuti non evidenziano un metodo superiore agli altri perché i valori degli indicatori di accuratezza variano a seconda del criterio considerato: il metodo *Tree-Mean* è il migliore in termini di preservazione dei valori individuali (*diff, coeff,* R^2); i metodi *Tree-NN* e *Tree-RS* sono i migliori rispetto alla preservazione della distribuzione marginale (d_{KS}); i metodi *Tree-RS* e *Tree-Mean* mostrano l'accuratezza migliore nella preservazione della media ($diff_m$); infine rispetto alla preservazione della variabilità ($diff_\sigma$), il metodo *Tree-RS* sembra migliore del *Tree-NN* per dimensioni ridotte dell'albero.

I risultati provenienti da differenti dimensioni dell'albero indicano che solo per il metodo *Tree-Mean* si osserva una accuratezza crescente all'aumentare delle dimensioni per tutti i criteri considerati. L'accuratezza degli altri metodi cresce all'aumentare delle dimensioni rispetto alla preservazione dei valori individuali e della distribuzione marginale ma andamenti diversificati sono osservati rispetto alla preservazione dei parametri. In particolare, l'accuratezza del metodo *Tree-NN* presenta un andamento non monotono rispetto alla preservazione di entrambi i parametri mentre l'accuratezza del metodo *Tree-RS* cresce rispetto alla preservazione della media e si riduce rispetto alla preservazione della variabilità.

La struttura ad albero non migliora necessariamente l'accuratezza delle imputazioni quando il metodo con *donatore di minima distanza* è utilizzato: il metodo *NN* è più accurato del metodo *Tree-NN* rispetto alla preservazione di valori individuali e della media ma risulta meno accurato rispetto alla preservazione della variabilità (non si osservano differenze di rilievo tra i valori dell'indicatore d_{KS}).

Riguardo al segno della distorsione nella stima della media, i valori dell'indicatore $dist_m$, riportati nelle Tabelle in Appendice B, mostrano segni positivi e segni negativi con una prevalenza di quelli positivi (media dei dati imputati maggiore della media dei corrispondenti dati osservati).

Infine, riguardo al segno della distorsione nella stima della variabilità dei dati, i valori dell'indicatore $dist_\sigma$, riportati nelle Tabelle in Appendice B, mostrano segni positivi e segni negativi con una prevalenza di quelli negativi (variabilità dei valori imputati inferiore alla variabilità dei corrispondenti valori osservati). Fa eccezione il metodo *SRI* per il quale il segno dell'indicatore $dist_\sigma$ è sempre positivo indicando un aumento della variabilità dei valori imputati rispetto a quella dei valori osservati.

5 RISULTATI DI ANALISI AGGIUNTIVE

5.1 Metodo *NN* "*modificato*"

Il metodo *NN* (eseguito mediante procedure SAS) ammette, a differenza del metodo *Tree-NN* (incorporato in *WAID*), la presenza dei valori mancanti nelle variabili ausiliarie del ricevente: la/e variabile/i con valore mancante sono escluse dal computo della distanza totale (vedi sottosezione 2.2). Inoltre, la maggior flessibilità del codice SAS consente di definire funzioni di distanza elementare adeguate alla natura delle variabili.

Nei dati utilizzati per lo studio si rileva che:

- alcune variabili ausiliarie presentano dei valori mancanti (RANGREV, EDUC, OCCUP, INDUST, ENTSIZ, STEMP, SUPERV);
- alcune variabili ausiliarie sono di natura qualitativa ordinale codificate con i valori interi da 1 al numero delle classi (EDUC) e quantitativa (AGE);
- alcune variabili ausiliarie sono di natura qualitativa ordinale semicontinua (assumono la modalità "non ammissibile" (o "NA") con una probabilità non nulla e sono ordinali altrimenti) (RANGREV, ENTSIZ, WORKHH).

La frequenza dei valori mancanti e dei valori non ammissibili sul totale dei valori (5760) è riportata, per ciascuna variabile ausiliaria, nella Tabella 3:

Tabella 3. Frequenza dei valori mancanti e non ammissibili per variabile ausiliaria

Variabile	Valore mancante		Valore non ammissibile	
	Frequenza	%	Frequenza	%
RANGREV	171	2.97	8	0.14
EDUC	2	0.03	-	-
OCCUP	112	1.94	535	9.29
INDUST	23	0.40	535	9.29
ENTSIZ	45	0.78	2218	38.51
STEMP	2	0.03	26	045
WORKHH	-	-	562	9.76
SUPERV	9	0.16	776	13.47

Poiché le frequenze riportate nella Tabella 3 sono non trascurabili, si è provveduto ad apportare alcune modifiche al programma SAS che effettua l'imputazione con *donatore di minima distanza*. Nella versione *modificata* del metodo *NN* (*NN-mod* nel seguito) i valori mancanti delle variabili ausiliarie non sono stati codificati, causando l'esclusione della variabile nel computo della distanza totale, e differenti funzioni di distanza elementare sono state definite in base alla natura delle variabili.

Nel metodo *NN-mod* la distanza elementare tra l'unità u_i e l'unità u_j definita per le variabili qualitative (REGION, NBWCL, GENDER, OCCUP, INDUST, STEMP, SUPERV) è:

$$D_q(u_i, u_j) = \begin{cases} 0 & if \quad x_i = x_j \\ 1 & if \quad x_i \neq x_j \end{cases}$$

Si osservi che la modalità "non ammissibile" è trattata come una qualunque modalità.

La distanza elementare definita per la variabile quantitativa (AGE) e la variabile qualitativa ordinale (EDUC) è:

$$D_1(u_i, u_j) = \left| x_i - x_j \right|.$$

La funzione di distanza elementare definita per le variabili qualitative ordinali semicontinue (RANGREV, ENTSIZ, WORKHH) è riportata nella Tabella 4:

Tabella 4. Matrice delle distanze per una variabile semicontinua

u_i	u_j	
	"NA"	*X*
"NA"	0	1
X	1	$D_1(u_i, u_j)$

La funzione di distanza definita nella Tabella 4 consente di usare una metrica di tipo qualitativo per alcune coppie di unità ed una metrica di tipo quantitativo per altre coppie di unità. In particolare:

- se sia il ricevente che il donatore presentano X="*NA"*, la distanza elementare è zero;
- se per il ricevente X="*NA"* mentre per il donatore X="valore", la distanza elementare è uno;
- se per il ricevente X="valore", e per il donatore X="*NA"* la distanza elementare è uno;
- se sia il ricevente che il donatore presentano X="valore", la distanza elementare è $|x_i-x_j|$.

Inoltre, per ottenere distanze elementari standardizzate (che assumono valori nell'intervallo [0,1]), la distanza $D_1(u_i, u_j)$ è stata calcolata sui valori delle variabili divisi per il corrispondente range (*max-min*).

I valori medi (sui cinque campioni) degli indicatori ottenuti applicando il metodo *NN-mod* sono riportati nella Tabella 5:

Tabella 5: Valori medi (sui cinque campioni) degli indicatori di accuratezza del metodo di imputazione *NN-mod* per la variabile *Reddito da lavoro dipendente*

Metodo	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$
NN-mod	7201.257	0.943	0.791	0.052	339.790	2077.385

Il metodo *NN-mod* appare lievemente più accurato del metodo *NN* solo nella preservazione dei valori individuali mentre una minore accuratezza è osservata rispetto agli altri criteri.

5.2 Imputazione da regressione con residuo casuale mediante PROC REG

Il metodo di imputazione da *regressione con residuo casuale* è stato applicato anche utilizzando la procedura REG del SAS (*SRI-ProcReg* nel seguito). Analogamente a quanto effettuato nella procedura *SRI* (che utilizza il software *IVEware*), la trasformata logaritmica della variabile risposta PI111 è stata regredita sulle variabili dummies definite per tutte le variabili ausiliarie ad eccezione

della variabile AGE (che è stata considerata quantitativa). Il valore atteso condizionato aumentato di un residuo casuale è stato utilizzato come valore da imputare. Il residuo è stato estratto da una variabile normalmente distribuita con media nulla e varianza uguale alla varianza residua stimata dal modello di regressione.

I valori medi (sui cinque campioni) degli indicatori ottenuti applicando il metodo *SRI-ProcReg* sono riportati nella Tabella 6:

Tabella 6. Valori medi (sui cinque campioni) degli indicatori di accuratezza del metodo di imputazione *SRI-ProcReg* **per la variabile *Reddito da lavoro dipendente***

Metodo	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$
SRI-ProcReg	10857.639	0.700	0.639	0.176	998.189	4498.319

L'imputazione da *regressione con residuo casuale* eseguita in modalità "standard" (*SRI-ProcReg*) appare lievemente più accurata dell'imputazione da *regressione con residuo casuale* eseguita mediante il software *IVEware* (*SRI*) rispetto alla preservazione dei valori individuali e della distribuzione marginale. Per quanto riguarda la preservazione dei parametri, si osserva una minore accuratezza per la media (998.189 vs 942.275) ed una maggiore accuratezza per la variabilità dei dati (4498.319 vs 5315.717).

6 CONCLUSIONI

In generale, per l'insieme di dati selezionato e il meccanismo di mancata risposta simulato, i risultati ottenuti mostrano che nessuno dei metodi usati per imputare la variabile di reddito può essere preferito agli altri rispetto a tutti i criteri di valutazione considerati: la scelta del metodo di imputazione dovrebbe essere "guidata" dalle principali proprietà statistiche che si desidera siano preservate dai dati finali. Il metodo *Tree-Mean* dovrebbe essere preferito quando si desidera preservare i valori individuali e la media mentre i metodi basati sul donatore (di minima distanza o selezionato casualmente) dovrebbero essere preferiti quando si desidera preservare la distribuzione e la variabilità dei dati. La preferenza di una proprietà rispetto all'altra dipende a sua volta dall'uso che si intende fare dei dati finali e quindi dalle analisi che si intende eseguire sugli stessi. Se, ad esempio, le imputazioni sono effettuate per produrre stime di aggregati, la preservazione dei valori individuali ha scarsa rilevanza. Se invece le imputazioni sono effettuate per ottenere dati finali completi da rilasciare per uso pubblico o da usare per lo sviluppo di modelli di previsione, la preservazione dei valori individuali assume grande rilevanza.

Con il software utilizzato (*WAID*) per eseguire i metodi di imputazioni basati sugli alberi di regressione, non è stato possibile utilizzare il numero minimo di unità presenti all'interno di un nodo come regola di arresto nel processo di costruzione dell'albero. L'assenza di questa regola ha prodotto nodi terminali ove il numero di rispondenti era a volte estremamente esiguo (in alcune repliche è stato osservato un nodo contenente 1 sola unità rispondente) dando luogo quindi a situazioni potenzialmente critiche (ricordiamo che i rispondenti classificati in un nodo sono utilizzati per imputare i valori dei non rispondenti classificati nello stesso nodo) che possono ripercuotersi sui risultati (l'eccessivo utilizzo di un donatore potrebbe provocare una distorsione nella distribuzione finale in quanto le risposte provenienti dal donatore risulterebbero sovrarappresentate). In questo lavoro non è stato verificato se esistevano non rispondenti classificati nei nodi con numerosità esigua ma in un contesto operativo è buona norma evitare che tali situazioni si verifichino. In ogni caso, il processo di costruzione dell'albero rappresenta l'aspetto critico di questa tipologia di metodi e particolare attenzione deve quindi essere dedicata ad esso.

I risultati ottenuti sconsigliano l'utilizzo del metodo di imputazione da *regressione con residuo casuale* per imputare i valori mancanti della variabile *Reddito da lavoro dipendente*. In generale i metodi di imputazione da *regressione* consentono di ottenere buoni risultati a condizione che il modello di regressione sia ben specificato e che le variabili ausiliarie siano fortemente correlate alla variabile da imputare. Per verificare tali condizioni sono stati utilizzati i risultati ottenuti dall'applicazione *SRI-ProcReg*. La rappresentazione grafica dei residui del modello sul valore atteso condizionato ottenuto dal modello ha mostrato, per ciascuna delle cinque regressioni eseguite sui 5126 rispondenti, lo stesso andamento: una disposizione dei valori dei residui sbilanciata verso i valori negativi (tra gli scarti grandi in valore assoluto prevalgono gli scarti negativi su quelli positivi). L'asimmetria negativa della distribuzioni dei residui indica una inadeguatezza del modello lineare per spiegare la relazione tra le variabili ausiliarie e la variabile da imputare (che ricordiamo, è il logaritmo della variabile risposta). Inoltre, i valori dei coefficienti di correlazione multipla (R^2) ottenuti dalle regressioni utilizzate per imputare, oscillano nell'intervallo [0.46, 0.48] indicando che l'iperpiano di regressione spiega meno della metà della variabilità della variabile risposta, in altre parole, uno scarso adeguamento del modello ai dati. Ricordiamo che il modello e le variabili ausiliarie sono stati selezionati in quanto utilizzati da Eurostat nella procedura di imputazione (delle componenti della variabile *Reddito da lavoro dipendente*) eseguita a livello centralizzato per i dati provenienti da tutti i Paesi membri. Risultati migliori potrebbero essere ottenuti mediante una specificazione del modello e delle variabili che tenga in maggior conto le peculiarità dell'insieme di dati analizzato.

Il metodo con *donatore di minima distanza* appartiene alla classe dei metodi di imputazione che non usano un modello esplicito per definire la relazione tra la/le variabile/i da imputare e le variabili ausiliarie ed è quindi, in generale, più robusto, rispetto alle errate specificazioni del modello, dei metodi basati su modelli espliciti quali, ad esempio, l'imputazione da regressione (come si evince dai risultati). Si consideri però che il metodo con *donatore di minima distanza* richiede la definizione di un concetto di vicinanza e la scelta della metrica inserisce un elemento di soggettività nella procedura di imputazione che può influire sui risultati, come evidenziato dal confronto tra l'applicazione *NN* e l'applicazione *NN-mod*. Questo confronto mostra, inoltre, come l'utilizzo di una funzione di distanza più idonea alla natura delle variabili ha effetti differenti a seconda del criterio considerato e quindi non migliora necessariamente l'accuratezza delle imputazioni.

RINGRAZIAMENTI

Un caloroso ringraziamento al dott. Ugo Guarnera per aver realizzato e reso disponibili i programmi SAS utilizzati per eseguire i metodi *NN* e *NN-mod* e alla Dott.ssa Orietta Luzi per i preziosi suggerimenti forniti.

BIBLIOGRAFIA

BANCA D'ITALIA (1993) I Bilanci delle Famiglie Italiane nell'Anno 1991, *Supplementi al Bollettino Statistico, Note Metodologiche e Informazioni Statistiche*, anno III, n.44.

Breiman. L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and Regression Trees*, Wadsworth International, Belmont, CA.

Chambers, R., Hoogland, J., Laaksonen, S., Mesa, D.M., Pannekoek, J., Piela, P., Tsai, P. and De Waal, T. (2001) *The AUTIMP Project: Evaluation of Imputation Software*. Research Paper 0122, Statistics Netherlands 2001.

De Waal, T., De Waard, J., Plomp, R. (2001) Manual WAID (4.1), Statistics Netherlands.

Eurostat (2001) Doc.PAN 166/2001-12. ECHP UDB *Description of variables. Data Dictionnary, Codebook and Differences between Countries and Waves*.

Eurostat (2002) Doc.PAN 164/2002-12. *Imputation of income in the ECHP*.

Kalton, G. e Kasprzyk, D. (1982) Imputing for Missing Survey Responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 22-31.

Kalton, G. e Kasprzyk, D. (1986) The treatment of Missing Survey Data. *Survey Methodology*, Vol. 12, No. 1, pp. 1-16.

Kovar, J,G., Whitridge, P.J. (1995) Imputation of Business Survey Data in *Business Survey Methods*, John Wiley & Sons, New York.

Little, R.J.A. (1988) Missing-Data Adjustments in Large Survey, *Journal of Business & Economic Statistics*, Vol. 6, No. 3, 287-295.

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

Rao, J.N.K. (1996) On variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association*, 91, 499-520.

Raughunathan T.E., Lepkowski J.M., Van Hoewyk J. and Solenberger P. (2001) A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, June 2001, vol. 27, No.1, pp85-95

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

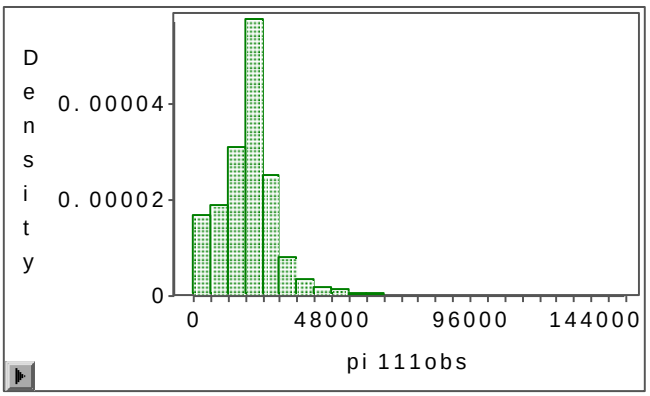
Sande, G. (1979) Numerical edit and imputation. Paper presented to the International association for Statistical computing, 42nd Session of the International Statistical Institute.

Sonquist, J.A., Baker,E.L., Morgan, J.N. (1973) *Searching for Structure* (rev. ed.), Institute for Social Research, University of Michigan, Ann Arbor.

APPENDICE A

Distribuzione dei valori originali della variabile *Reddito da lavoro dipendente* per i rispondenti e per i non rispondenti nella replica 1.

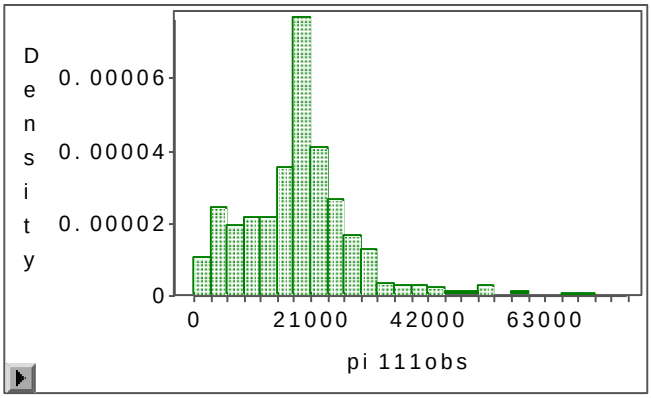
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	19559.6537	Sum	100262785
Std Dev	11024.3745	Variance	121536832
Skewness	1.9323	Kurtosis	11.8046
USS	2.584E+12	CSS	6.229E+11
CV	56.3628	Std Mean	153.9801

Quantiles			
100% Max	149000.000	99.0%	57850.0000
75% Q3	24000.0000	97.5%	44600.0000
50% Med	19450.0000	95.0%	36500.0000
25% Q1	13200.0000	90.0%	30000.0000
0% Min	180.0000	10.0%	5850.0000
Range	148820.000	5.0%	3126.0000
Q3-Q1	10800.0000	2.5%	2000.0000
Mode	18000.0000	1.0%	1020.0000

Non rispondenti

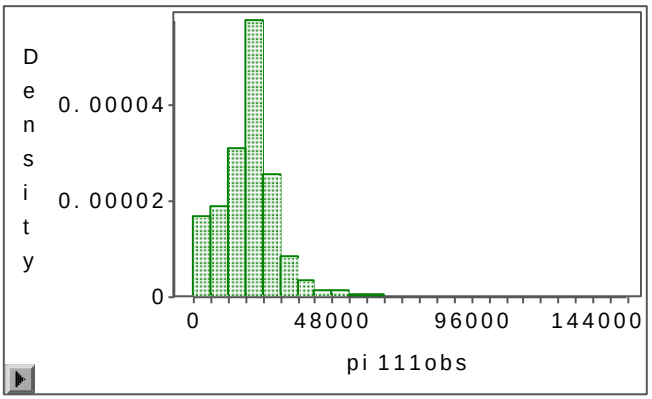


Moments			
N	634.0000	Sum Wgts	634.0000
Mean	19465.8991	Sum	12341380.0
Std Dev	11074.9056	Variance	122653534
Skewness	1.3884	Kurtosis	4.1744
USS	3.179E+11	CSS	7.764E+10
CV	56.8939	Std Mean	439.8407

Quantiles			
100% Max	77200.0000	99.0%	58200.0000
75% Q3	23600.0000	97.5%	51000.0000
50% Med	19200.0000	95.0%	39200.0000
25% Q1	12900.0000	90.0%	30485.0000
0% Min	460.0000	10.0%	5550.0000
Range	76740.0000	5.0%	3600.0000
Q3-Q1	10700.0000	2.5%	2300.0000
Mode	18000.0000	1.0%	1400.0000

Distribuzione dei valori originali della variabile *Reddito da lavoro dipendente* per i rispondenti e per i non rispondenti nella replica 2.

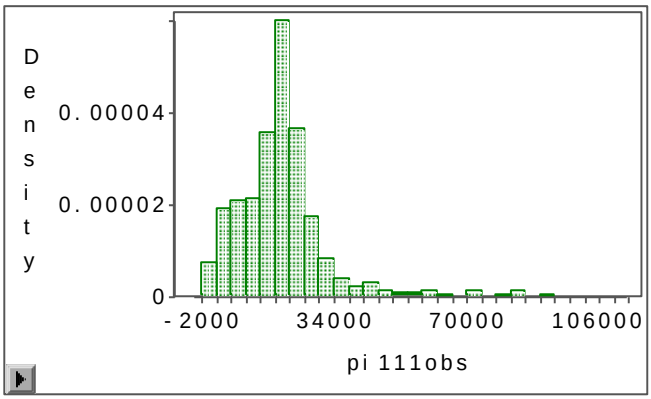
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	19444.2837	Sum	99671398.0
Std Dev	10599.6015	Variance	112351551
Skewness	1.6944	Kurtosis	10.8932
USS	2.514E+12	CSS	5.758E+11
CV	54.5127	Std Mean	148.0472

Quantiles			
100% Max	149000.000	99.0%	54500.0000
75% Q3	24000.0000	97.5%	43940.0000
50% Med	19400.0000	95.0%	36000.0000
25% Q1	13250.0000	90.0%	30000.0000
0% Min	180.0000	10.0%	5910.0000
Range	148820.000	5.0%	3200.0000
Q3-Q1	10750.0000	2.5%	2000.0000
Mode	18000.0000	1.0%	1100.0000

Non rispondenti

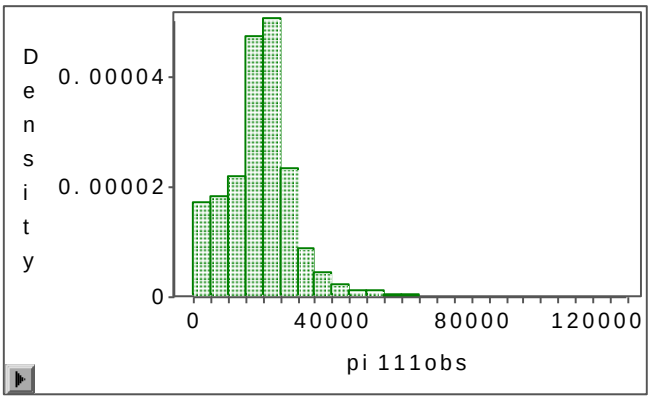


Moments			
N	634.0000	Sum Wgts	634.0000
Mean	20398.6861	Sum	12932767.0
Std Dev	14007.7430	Variance	196216864
Skewness	2.2954	Kurtosis	8.5811
USS	3.880E+11	CSS	1.242E+11
CV	68.6698	Std Mean	556.3185

Quantiles			
100% Max	112600.000	99.0%	82000.0000
75% Q3	23500.0000	97.5%	63600.0000
50% Med	19400.0000	95.0%	45600.0000
25% Q1	12700.0000	90.0%	32000.0000
0% Min	200.0000	10.0%	5200.0000
Range	112400.000	5.0%	3000.0000
Q3-Q1	10800.0000	2.5%	1800.0000
Mode	16800.0000	1.0%	1000.0000

Distribuzione dei valori originali della variabile *Reddito da lavoro dipendente* per i rispondenti e per i non rispondenti nella replica 3.

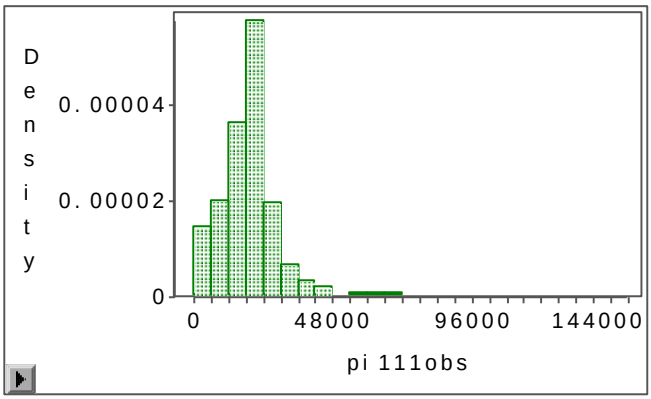
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	19528.9263	Sum	100105276
Std Dev	10834.6355	Variance	117389325
Skewness	1.6074	Kurtosis	8.2204
USS	2.557E+12	CSS	6.016E+11
CV	55.4799	Std Mean	151.3300

Quantiles			
100% Max	123000.000	99.0%	56400.0000
75% Q3	24000.0000	97.5%	44750.0000
50% Med	19450.0000	95.0%	36560.0000
25% Q1	13150.0000	90.0%	30200.0000
0% Min	180.0000	10.0%	5760.0000
Range	122820.000	5.0%	3200.0000
Q3-Q1	10850.0000	2.5%	2000.0000
Mode	18000.0000	1.0%	1050.0000

Non rispondenti

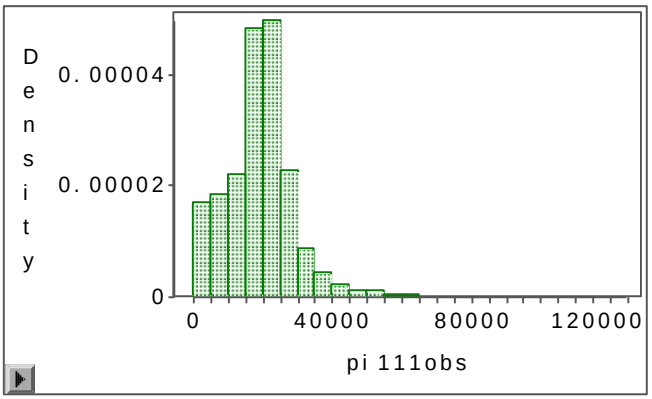


Moments			
N	634.0000	Sum Wgts	634.0000
Mean	19714.3360	Sum	12498889.0
Std Dev	12498.4184	Variance	156210462
Skewness	3.2210	Kurtosis	22.7446
USS	3.453E+11	CSS	9.888E+10
CV	63.3976	Std Mean	496.3756

Quantiles			
100% Max	149000.000	99.0%	69700.0000
75% Q3	23400.0000	97.5%	55600.0000
50% Med	19200.0000	95.0%	38700.0000
25% Q1	13850.0000	90.0%	30000.0000
0% Min	240.0000	10.0%	6700.0000
Range	148760.000	5.0%	3215.0000
Q3-Q1	9550.0000	2.5%	2100.0000
Mode	18000.0000	1.0%	1300.0000

Distribuzione dei valori originali della variabile *Reddito da lavoro dipendente* per i rispondenti e per i non rispondenti nella replica 4.

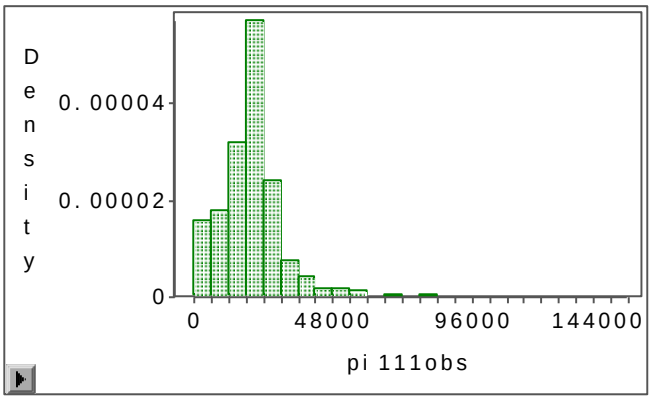
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	19479.9311	Sum	99854127.0
Std Dev	10875.3341	Variance	118272891
Skewness	1.6788	Kurtosis	8.6093
USS	2.551E+12	CSS	6.061E+11
CV	55.8284	Std Mean	151.8984

Quantiles			
100% Max	123000.000	99.0%	57850.0000
75% Q3	24000.0000	97.5%	44600.0000
50% Med	19400.0000	95.0%	36456.0000
25% Q1	13200.0000	90.0%	30000.0000
0% Min	180.0000	10.0%	5820.0000
Range	122820.000	5.0%	3200.0000
Q3-Q1	10800.0000	2.5%	2000.0000
Mode	18000.0000	1.0%	1100.0000

Non rispondenti

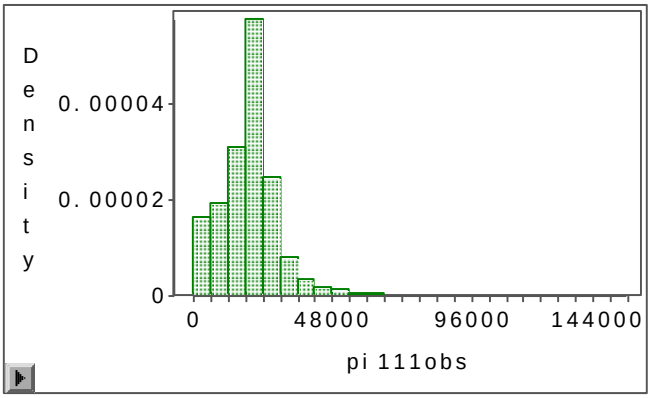


Moments			
N	634.0000	Sum Wgts	634.0000
Mean	20110.4700	Sum	12750038.0
Std Dev	12195.6163	Variance	148733056
Skewness	2.9344	Kurtosis	22.3220
USS	3.506E+11	CSS	9.415E+10
CV	60.6431	Std Mean	484.3498

Quantiles			
100% Max	149000.000	99.0%	63000.0000
75% Q3	24100.0000	97.5%	51000.0000
50% Med	19500.0000	95.0%	39800.0000
25% Q1	13600.0000	90.0%	30800.0000
0% Min	250.0000	10.0%	6000.0000
Range	148750.000	5.0%	3300.0000
Q3-Q1	10500.0000	2.5%	2400.0000
Mode	18000.0000	1.0%	1000.0000

Distribuzione dei valori originali della variabile *Reddito da lavoro dipendente* per i rispondenti e per i non rispondenti nella replica 5.

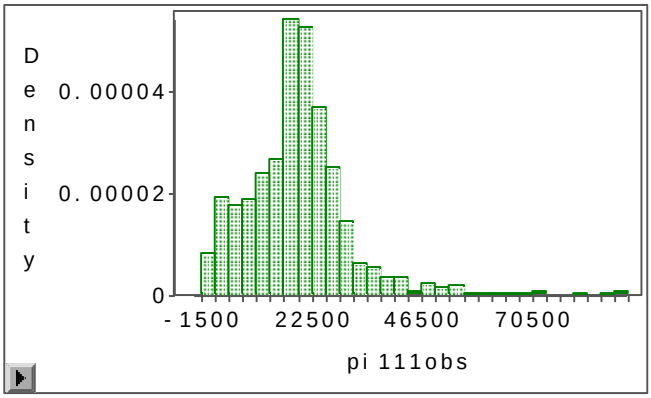
Rispondenti



Moments			
N	5126.0000	Sum Wgts	5126.0000
Mean	19501.3067	Sum	99963698.0
Std Dev	10869.5942	Variance	118148079
Skewness	1.8797	Kurtosis	11.6775
USS	2.555E+12	CSS	6.055E+11
CV	55.7378	Std Mean	151.8183

Quantiles			
100% Max	149000.000	99.0%	57400.0000
75% Q3	24000.0000	97.5%	44600.0000
50% Med	19400.0000	95.0%	36400.0000
25% Q1	13200.0000	90.0%	30000.0000
0% Min	180.0000	10.0%	6000.0000
Range	148820.000	5.0%	3200.0000
Q3-Q1	10800.0000	2.5%	2000.0000
Mode	18000.0000	1.0%	1100.0000

Non rispondenti



Moments			
N	634.0000	Sum Wgts	634.0000
Mean	19937.6451	Sum	12640467.0
Std Dev	12244.5194	Variance	149928257
Skewness	1.7816	Kurtosis	6.7544
USS	3.469E+11	CSS	9.490E+10
CV	61.4141	Std Mean	486.2920

Quantiles			
100% Max	91000.0000	99.0%	69700.0000
75% Q3	24300.0000	97.5%	52000.0000
50% Med	19300.0000	95.0%	40500.0000
25% Q1	12700.0000	90.0%	31200.0000
0% Min	250.0000	10.0%	5000.0000
Range	90750.0000	5.0%	2800.0000
Q3-Q1	11600.0000	2.5%	1440.0000
Mode	18000.0000	1.0%	1000.0000

APPENDICE B

Valori degli indicatori calcolati per ciascun campione di non rispondenti (e valore medio sui cinque campioni) e ciascun metodo di imputazione.

Tree-RS

Nodi	replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
	1	9023.505	0.839	0.728	0.063	427.183	32.970	427.183	32.970
	2	9200.640	0.868	0.690	0.054	116.338	1779.550	-116.338	-1779.550
10	3	9591.503	0.797	0.662	0.073	747.062	274.040	747.062	-274.040
	4	9346.342	0.822	0.670	0.033	152.945	87.466	-152.945	87.466
	5	9488.606	0.807	0.703	0.035	557.091	841.999	557.091	841.999
	media	**9330.119**	**0.827**	**0.691**	**0.052**	**400.124**	**603.205**	**292.411**	**-218.231**
	1	8338.030	0.861	0.744	0.032	159.377	378.280	-159.377	378.280
	2	8944.349	0.869	0.746	0.060	590.162	1195.556	590.162	-1195.556
30	3	8103.021	0.856	0.719	0.058	459.159	1201.743	459.159	-1201.743
	4	8592.257	0.841	0.698	0.036	187.336	41.396	-187.336	41.396
	5	8731.155	0.866	0.736	0.041	60.495	516.196	60.495	-516.196
	media	**8541.762**	**0.859**	**0.729**	**0.045**	**291.306**	**666.634**	**152.621**	**-498.764**
	1	8622.287	0.833	0.742	0.036	65.707	1385.650	65.707	1385.650
	2	8836.292	0.878	0.745	0.050	369.576	1344.954	369.576	-1344.954
40	3	8052.468	0.849	0.706	0.063	310.923	973.359	310.923	-973.359
	4	8239.516	0.847	0.712	0.050	271.724	665.996	271.724	-665.996
	5	8009.431	0.887	0.767	0.030	94.888	427.350	-94.888	-427.350
	media	**8351.999**	**0.859**	**0.734**	**0.046**	**222.564**	**959.462**	**184.608**	**-405.202**

Tree-NN

Nodi	replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
	1	7299.358	0.873	0.784	0.047	485.727	197.901	485.727	-197.901
	2	7665.306	0.954	0.789	0.046	214.432	2879.189	-214.432	-2879.189
10	3	8168.590	0.768	0.702	0.068	1227.057	2025.194	1227.057	2025.194
	4	7016.385	0.886	0.762	0.027	217.511	305.372	-217.511	-305.372
	5	7485.513	0.895	0.770	0.041	33.352	939.922	33.352	-939.922
	media	**7527.030**	**0.875**	**0.761**	**0.046**	**435.616**	**1269.516**	**262.839**	**-459.438**
	1	7222.675	0.883	0.785	0.044	120.155	58.361	120.155	-58.361
	2	7555.390	0.927	0.787	0.030	135.465	1768.792	-135.465	-1768.792
30	3	7725.757	0.796	0.724	0.073	1086.161	1270.390	1086.161	1270.390
	4	7144.587	0.911	0.780	0.033	422.735	698.940	-422.735	-698.940
	5	7400.068	0.888	0.780	0.041	189.128	546.221	189.128	-546.221
	media	**7409.695**	**0.881**	**0.771**	**0.044**	**390.729**	**868.541**	**167.449**	**-360.385**
	1	7410.587	0.852	0.779	0.046	548.429	652.266	548.429	652.266
	2	7446.850	0.941	0.799	0.032	253.715	1910.521	-253.715	-1910.521
40	3	7750.274	0.799	0.722	0.055	879.328	1360.420	879.328	1360.420
	4	6958.218	0.917	0.791	0.033	440.278	653.321	-440.278	-653.321
	5	7423.964	0.869	0.780	0.035	518.882	88.827	518.882	-88.827
	media	**7397.979**	**0.876**	**0.774**	**0.040**	**528.126**	**933.071**	**250.529**	**-127.997**

Tree-Mean

Nodi	replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
	1	6423.571	0.988	0.839	0.197	88.317	4093.487	88.317	-4093.487
	2	7078.717	1.049	0.812	0.200	349.424	6925.977	-349.424	-6925.977
10	3	6698.496	0.965	0.790	0.199	607.156	5453.849	607.156	-5453.849
	4	6635.713	1.008	0.796	0.189	550.975	5044.055	-550.975	-5044.055
	5	6706.499	0.988	0.829	0.208	292.677	4791.857	292.677	-4791.857
	media	**6708.599**	**1.000**	**0.813**	**0.199**	**377.710**	**5261.845**	**17.550**	**-5261.845**
	1	6183.429	0.979	0.839	0.144	102.882	3036.372	-102.882	-3036.372
	2	6484.976	1.011	0.837	0.153	42.294	4864.323	42.294	-4864.323
30	3	6337.018	0.949	0.801	0.186	681.371	4197.445	681.371	-4197.445
	4	6154.658	0.988	0.818	0.137	244.316	3922.134	-244.316	-3922.134
	5	6194.630	0.967	0.845	0.161	419.202	3304.796	419.202	-3304.796
	media	**6270.942**	**0.979**	**0.828**	**0.156**	**298.013**	**3865.014**	**159.134**	**-3865.014**
	1	6166.201	0.945	0.835	0.144	97.399	1814.837	97.399	-1814.837
	2	6412.539	1.005	0.839	0.167	104.447	4626.159	104.447	-4626.159
40	3	6251.338	0.952	0.803	0.155	510.63	3953.246	510.630	-3953.246
	4	6120.574	0.979	0.824	0.132	123.411	3495.625	-123.411	-3495.625
	5	6031.328	0.972	0.853	0.159	332.642	3139.903	332.642	-3139.903
	media	**6196.396**	**0.971**	**0.831**	**0.151**	**233.706**	**3405.954**	**184.341**	**-3405.954**

NN

replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
1	7490.666	0.864	0.773	0.046	303.95	274.546	303.950	274.546
2	7572.391	0.956	0.793	0.039	526.76	2326.463	-526.760	-2326.463
3	7107.584	0.881	0.761	0.060	379.098	1056.089	379.098	-1056.089
4	7336.954	0.895	0.762	0.046	50.043	1278.473	50.043	-1278.473
5	7314.623	0.925	0.778	0.050	327.39	1593.588	-327.390	-1593.588
media	**7364.444**	**0.904**	**0.773**	**0.048**	**317.448**	**1305.832**	**-24.212**	**-1196.013**

SRI

replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
1	11082.266	0.655	0.626	0.191	1075.716	6533.137	1075.716	6533.137
2	11760.988	0.629	0.598	0.175	1640.762	6962.975	1640.762	6962.975
3	10973.877	0.684	0.644	0.156	1404.702	4938.856	1404.702	4938.856
4	10691.521	0.707	0.619	0.216	129.134	4687.238	-129.134	4687.238
5	10175.626	0.766	0.710	0.169	461.061	3456.379	461.061	3456.379
media	**10936.856**	**0.688**	**0.639**	**0.181**	**942.275**	**5315.717**	**890.621**	**5315.717**

NN-mod

replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
1	7196.418	0.915	0.802	0.044	113.386	786.939	-113.386	-786.939
2	7514.353	0.996	0.799	0.046	757.694	3656.974	-757.694	-3656.974
3	6775.940	0.943	0.786	0.066	66.028	2740.162	-66.028	-2740.162
4	7133.674	0.900	0.772	0.043	15.39	1135.627	15.390	-1135.627
5	7385.899	0.959	0.794	0.062	746.454	2067.225	-746.454	-2067.225
media	**7201.257**	**0.943**	**0.791**	**0.052**	**339.790**	**2077.385**	**-333.634**	**-2077.385**

SRI-ProcReg

replica	$diff_Y$	*coeff*	R^2	d_{KS}	$diff_m$	$diff_\sigma$	$dist_m$	$dist_\sigma$
1	11766.447	0.612	0.590	0.167	1628.401	7500.077	1628.401	7500.077
2	11387.051	0.681	0.637	0.148	1845.865	4590.221	1845.865	4590.221
3	10088.218	0.726	0.655	0.159	1094.131	3129.700	1094.131	3129.700
4	10636.130	0.736	0.645	0.202	162.593	3552.929	162.593	3552.929
5	10410.348	0.743	0.668	0.203	259.955	3718.668	259.955	3718.668
media	**10857.639**	**0.700**	**0.639**	**0.176**	**998.189**	**4498.319**	**998.189**	**4498.319**