1-2004

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

TECNICHE E STRUMENTI

CONCORD V. 1.0

Controllo e correzione dei dati

Manuale utente
e aspetti metodologici

ISTAT

SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

CONCORD V. 1.0

Controllo e correzione dei dati

Manuale utente
e aspetti metodologici

A cura di: Ercole Riccini Margarucci

Ha collaborato: Giulio Barcaroli

Per chiarimenti sul contenuto
della pubblicazione rivolgersi a:
Istat, Servizio Metodologie, tecnologie
e software per la produzione statistica
Tel. 06 4673.2323
Fax 06 4673.2955
E-mail *mts-f@istat.it*

CONCORD V. 1.0
Controllo e correzione dei dati
Manuale utente e aspetti metodologici

Istituto nazionale di statistica
Via Cesare Balbo, 16 - Roma

Coordinamento editoriale:
Andrea Braghin, Piero Crivelli
Servizio Produzione libraria
Via Tuscolana, 1788 - Roma

Videoimpaginazione e progetto grafico:
Carlo Calvaresi, Antonio Maggiorani

Stampa digitale:
Istat - Produzione libraria e centro stampa

Marzo 2004 –

Indice

Presentazione

Il problema dell'individuazione e del trattamento degli errori non campionari presenti nei dati rilevati ha sempre rivestito un'enorme importanza nel processo di produzione delle informazioni proprio delle rilevazioni statistiche, in special modo di quelle condotte dagli Istituti di statistica nazionali o da altri Enti operanti nel campo delle statistiche ufficiali.

Gli obiettivi della fase di controllo e correzione sono, in generale, quelli di garantire una maggiore qualità delle stime prodotte (qualità intesa come accuratezza), ma anche la presentabilità dei dati elementari una volta che questi vengano diffusi all'utenza esterna.

Storicamente, questo secondo obiettivo è stato, all'inizio, quello dominante: lo sforzo era concentrato nel far sì che i microdati messi a disposizione del pubblico non rivelassero incoerenze al loro interno, indice di carenze nelle modalità di rilevazione, registrazione e trattamento dei dati dell'indagine. Ciò si traduceva nel fatto che le attività di controllo e correzione, sia quelle di tipo interattivo che automatico, erano finalizzate all'individuazione ed alla risoluzione delle incoerenze, più che alla individuazione ed eliminazione degli errori che di tali incoerenze erano la causa. Nel caso delle procedure automatiche, gli strumenti che permettevano di raggiungere tale obiettivo erano costituiti da programmi per l'applicazione di regole "deterministiche", del tipo "se nei dati è presente una certa incoerenza, allora agisci nel senso di modificare il valore di una determinata variabile assegnandole un determinato valore". Da un punto di vista

qualitativo, pesanti limiti caratterizzavano tale approccio, che ciò nonostante ha dominato incontrastato i processi di trattamento dei dati negli Istituti di statistica fino alla metà degli anni '70.

Nel 1976 Fellegi ed Holt proposero una metodologia completamente diversa (Fellegi, Holt 1976), adottata immediatamente all'interno di Statistics Canada, e poi successivamente da altri Istituti compreso il nostro. La metodologia in questione si basa su un approccio non deterministico o probabilistico: l'obiettivo non è più solo, e non tanto, quello di eliminare in qualsivoglia modo le incoerenze dai dati, bensì, sulla base delle incoerenze riscontrate, individuare le variabili che più probabilmente sono errate, e procedere a correggerle assegnando loro i valori verosimilmente più vicini al valore vero, producendo in tal modo un incremento dell'accuratezza delle stime prodotte e diffuse.

Unitamente al nuovo approccio metodologico, si fece strada l'idea che, data la complessità dei necessari algoritmi per la localizzazione degli errori e per la loro correzione, per poterlo applicare era necessario disporre di strumenti software di tipo generalizzato, utilizzabili cioè in indagini diverse, senza essere costretti a sviluppare ogni volta ex-novo il codice necessario. Vennero quindi prodotti negli anni '70 e '80 sistemi come CANEDIT (Statistics Canada), AERO (Istituto di statistica ungherese), DIA (Istituto di statistica spagnolo), DISCRETE (U.S. Bureau of the Census) per le variabili categoriche, GEIS (Statistics Canada) e SPEER (U.S. Bureau of the Census) per quelle di tipo continuo.

A partire dalla metà degli anni '80 l'ISTAT avviò una fase di ricognizione e sperimentazione delle nuove tecniche e del relativo software. In seguito a tale fase, portata avanti congiuntamente dal settore studi metodologico e da quello informatico, si decise di procedere allo sviluppo autonomo di un sistema improntato alla metodologia Fellegi-Holt, applicabile alle variabili categoriche. Tale sistema, chiamato SCIA (Sistema Controllo e Imputazione Automatici), venne applicato per la prima volta in occasione del Censimento delle Abitazioni e della Popolazione del 1991, ed immediatamente dopo alla Indagine sulle Forze di Lavoro in occasione

della ristrutturazione avvenuta nel 1992. Le prime applicazioni evidenziarono, accanto ai vantaggi attesi, alcuni limiti sia della metodologia (in particolare, la sua non immediata applicabilità in situazioni caratterizzate dalla presenza di errori sistematici), sia dello strumento (rigidità di utilizzo, difficoltà di applicabilità nel caso di questionari complessi con elevato numero di regole di coerenza). L'analisi di tali limiti produsse da una parte la definizione di una metodologia mista che contempla l'utilizzo dell'approccio probabilistico per il trattamento degli errori casuali e di quello deterministico per il trattamento degli errori sistematici; dall'altra, il disegno e l'implementazione di un software che permettesse l'applicabilità di tale metodologia, integrando diversi strumenti nel frattempo prodotti all'interno dell'Istituto: oltre a SCIA (migliorato ed arricchito rispetto alla prima versione), GRANADA per la localizzazione degli errori sistematici mediante l'applicazione di regole deterministiche del tipo se-allora, e RIDA, per l'imputazione mediante donatore delle variabili contenenti errori o mancate risposte parziali.

Tale software è, per l'appunto, CONCORD (CONtrollo e CORrezione dei Dati), che si propone come uno standard in primo luogo per le indagini sulle famiglie, in cui sono preponderanti le variabili di tipo categorico, ma anche per quelle sulle imprese caratterizzate da variabili di tipo quantitativo continuo, limitatamente in quest'ultimo caso alla possibilità di applicare il solo approccio deterministico.

Ovviamente, CONCORD non esaurisce le necessità in questo campo. Ad esempio, il trattamento degli errori individuabili solo dall'esame delle incoerenze riscontrabili tra diverse unità di analisi, anziché di quelle interne ad una stessa unità, è gestibile in modo ottimale facendo ricorso ad un approccio ancora diverso, di tipo "data driven", ideato ed implementato ancora una volta dai ricercatori canadesi. Ancora, è auspicabile l'estensione dell'approccio probabilistico al trattamento congiunto di variabili categoriche e continue, così come l'utilizzo di moduli per l'individuazione ed il trattamento di valori "outlier".

Infine, nuove tecniche basate sull'utilizzo di reti neurali potrebbero essere utilizzate sia in fase di localizzazione degli errori che in fase di imputazione.

L'attività di ricerca in questo campo, spesso portata avanti attraverso la cooperazione internazionale (Nazioni Unite, Comunità Europea), è continua, e produce un incessante adeguamento dei metodi e delle tecniche. Uno dei vantaggi di CONCORD è che la sua architettura modulare permetterà l'inserimento dei nuovi strumenti che via via si renderanno disponibili, garantendo in tal modo la possibilità di utilizzare i metodi più avanzati.

Giulio Barcaroli

Responsabile del Servizio
Metodologie, tecnologie e
software per la produzione statistica

Introduzione

Il desiderio di ogni statistico è quello di poter utilizzare dati provenienti da fonti che possano fornire informazioni esatte fin dall'origine. Le indagini campionarie che si basano su pochi quesiti e si avvalgono di strumenti di registrazione controllata vengono spesso incontro a questa necessità.

Rimangono però indagini basate su questionari complessi, tipo quelle censuarie, che necessitano di controlli sui dati e di eventuali correzioni e che devono avvalersi di programmi spesso sofisticati e di difficile messa a punto.

Poter fornire un software generalizzato, adeguabile facilmente a questionari diversi, basato su molteplici ma identiche e validate metodologie di correzione, e soprattutto residente su personal computer, è stata l'idea che ha determinato lo sviluppo di CONCORD.

Il seguente manuale guida gli utenti, che devono fare uso del software CONCORD, nella scelta dell'approccio di correzione più adatto alla loro problematica e nell'utilizzo dei moduli e delle relative funzioni.

Punti di forza di CONCORD sono:
1. la possibile integrazione tra i diversi metodi di correzione, tutti residenti contemporaneamente nello stesso software;
2. il fatto che sia stato sviluppato per personal computer con sistema operativo Windows, sistema capillarmente diffuso;
3. la filosofia di software totalmente generalizzato, quindi adatto ad ogni indagine.

La versione attuale di CONCORD permette di impostare un piano di controllo e, eventualmente, di correzione dei dati di una qualsiasi indagine,

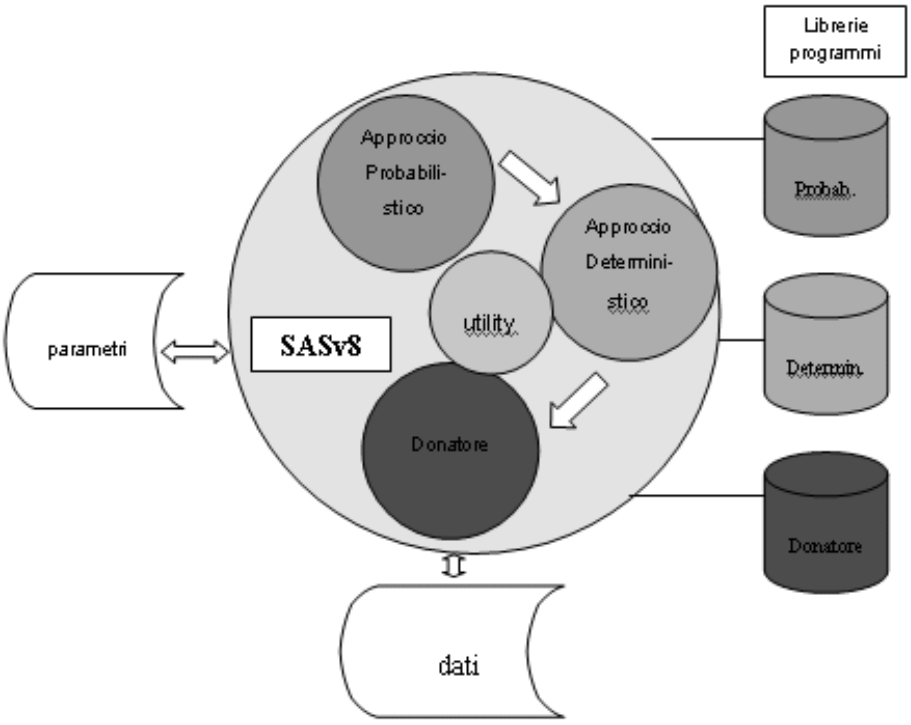
purché registrati in un file ASCII in tracciato a record singolo e campi fissi. È possibile scegliere l'approccio, probabilistico o deterministico, con il quale effettuare la localizzazione degli errori; le correzioni possono essere effettuate, oltre che con i due approcci suddetti, anche con il metodo del donatore.

Il software permette l'integrazione tra i tre metodi in modo gerarchico, passando cioè dall'approccio probabilistico a quello deterministico e da questo alla correzione tramite donazione, utilizzando sempre le stesse definizioni dei campi del record e delle regole di controllo. Le caratteristiche essenziali di CONCORD sono la flessibilità, essendo adatto a tutti i tracciati record purché a campi fissi, il supporto metodologico e la semplicità d'uso corredata da programmi di utilità che agevolano l'utente nella gestione dei dati e da un help in linea completo e immediato.

CONCORD, come mostrato in figura 1, è formato da un modulo centrale, che interfaccia l'utente verso i programmi residenti in librerie esterne fornendo loro parametri, sotto forma di file, e mandandoli in esecuzione quando necessario.

Il modulo centrale controlla e "colloquia" con l'utente eliminando, quando possibile, gli errori formali e di sintassi prima dell'esecuzione dei programmi, che "interpretano" i parametri forniti e si adeguano così alle varie esigenze. Le frecce indicano l'andamento dell'integrazione tra i diversi approcci.

Figura 1 L'impostazione di CONCORD



1. Cenni metodologici

La fase di controllo e correzione dei dati consiste nell'individuazione e nel trattamento degli errori (in senso generale, comprensivi cioè delle mancate risposte parziali) presenti nei dati raccolti mediante una certa indagine, allo scopo di garantire risultati finali qualitativamente migliori.

In generale, diciamo che una certa variabile rilevata in una data unità statistica è affetta da errore quando il suo valore non corrisponde al valore vero che essa presenta in quella unità. È chiaro che la presenza di errori, di qualunque natura essi siano, può provocare distorsioni nelle distribuzioni delle variabili investigate, nelle stime finali dei dati (totali, medie, ecc.), e in tutte le analisi statistiche effettuate sui dati non corretti (Masselli et al 1992).

La localizzazione delle risposte errate in un certo insieme di dati statistici è basata su diversi tipi di controlli, che possono essere classificati in tre categorie principali:

1. controlli di coerenza: verificano che prefissate combinazioni di valori assunti da variabili rilevate in una stessa unità soddisfino certi requisiti (regole di compatibilità);
2. controlli di validità o di range: verificano che i valori assunti da una data variabile siano interni all'intervallo di definizione della variabile stessa;
3. controlli statistici: utilizzati al fine di isolare quelle unità statistiche che presentano, per alcune delle variabili in esse contenute, valori che si discostano in modo significativo dai valori che le stesse variabili assumono nel resto delle unità campionarie o rispetto ad una rilevazione precedente.

Gli edit di coerenza o consistenza vengono utilizzati per la costruzione dei cosiddetti piani di compatibilità[1]. Più rigorosamente, si definisce piano di compatibilità un insieme di vincoli (edit) non ridondanti e non contraddittori, che devono essere contemporaneamente soddisfatti da ogni unità statistica affinché l'informazione corrispondente possa essere considerata corretta.

Il controllo effettuato sui dati mediante un piano di compatibilità è di tipo intra-unità se utilizza le sole informazioni fornite da ogni singola unità osservata, è di tipo inter-unità quando i dati relativi ad una certa osservazione vengono confrontati con informazioni prodotte da altre osservazioni della stessa popolazione.

Gli edit componenti un piano di compatibilità possono essere distinti in:
1. regole formali, che derivano dalla struttura del modello, cioè direttamente dalle norme di compilazione e dai "percorsi interni" (salti) del modello;
2. regole sostanziali, che derivano da considerazioni di tipo statistico-matematico, o da conoscenze specifiche a priori del fenomeno oggetto di rilevazione.

È chiaro che la natura degli edit (sia formali che sostanziali) di un piano di compatibilità è strettamente dipendente dal tipo di variabili (qualitative o quantitative) oggetto di verifica. Mentre nel caso di variabili qualitative, infatti, tali edit hanno la forma di relazioni logiche tra le variabili, nel caso di variabili quantitative le regole di compatibilità sono espresse in forma di relazioni statistico/matematiche (equazioni o disequazioni lineari, rapporti, ecc.).

Una volta individuati i record i cui valori violano uno o più vincoli del piano di compatibilità, il problema diventa la localizzazione delle variabili i cui valori devono essere considerati errati ed in quanto tali da sottoporre ad un passo di correzione.

Sia il problema della localizzazione dei record errati, sia quello dell'individuazione delle variabili che, per ogni record errato, sono da considerarsi responsabili della violazione di una o più regole di compatibilità, possono essere risolti adottando un approccio di tipo interattivo oppure automatico.

[1] Oppure di *incompatibilità* qualora gli edit rappresentino condizioni di incoerenza

Nel caso dell'editing automatico, si deve distinguere il caso in cui si utilizzi software specificamente sviluppato per una data tipologia di rilevazioni, oppure generalizzato, cioè immediatamente adattabile a diverse tipologie di indagine.

Nell'ambito dell'editing di tipo automatico possiamo ulteriormente distinguere a seconda che per la costruzione della procedura di editing si adotti un approccio di tipo deterministico oppure probabilistico (Barcaroli et al 1999).

La fase di applicazione delle regole di dominio, di compilazione e di compatibilità ai dati grezzi non può che essere compiuta in modo deterministico:

per ogni record, o per gruppi di record, vengono applicate tali regole che, se verificate, segnalano sicuramente la presenza di errori.

Ad esempio:

SE (sesso = maschio E professione = casalinga) ALLORA sussiste incompatibilità x

Una regola di questo tipo non individua, di per sé, l'errore che ne causa l'attivazione: infatti, il valore non vero può celarsi in una o nell'altra delle variabili, o in entrambe.

È nella fase di localizzazione degli errori che diviene decisivo il tipo di approccio adottato. Nell'approccio deterministico, ad ogni situazione di incompatibilità segue, contestualmente, l'indicazione delle variabili che debbono considerarsi errate, e, in quanto tali, da imputare. Nell'esempio considerato avremo, per ipotesi:

SE (sesso = maschio E professione = casalinga) ALLORA sesso ← femmina

Il che significa che, se in un record è attivata la condizione di incompatibilità "maschio/casalinga", la regola indica l'azione da effettuare per correggere l'errore, che consiste nell'imputare la modalità femmina alla variabile sesso.

Generalizzando, una volta attivate, mediante le regole di compatibilità, una o più condizioni di errore in un dato record, sono determinate a priori le azioni da intraprendere per riportare il medesimo record in una situazione di correttezza.

Le procedure deterministiche sono generalmente costituite da regole di imputazione deterministica (R.I.D.) del tipo:

SE [incompatibilità] ALLORA [localizzazione e correzione errore]

La condizione di incompatibilità esprime delle relazioni inammissibili intercorrenti tra due o più variabili; la localizzazione consiste nell'indicazione di quali variabili considerare errate, ed eventualmente di quali valori assegnare per correggerle.

Un record, durante l'esecuzione della procedura di correzione, potrà causare l'attivazione delle regole in corrispondenza delle quali è verificata la parte SE: in tal caso saranno modificate le variabili indicate nella parte ALLORA assegnando loro valori predefiniti o scelti in altro modo

Al contrario di quello precedente, l'approccio probabilistico non prevede la definizione a priori, per ogni situazione di errore, dell'elenco delle azioni da intraprendere per eliminare gli errori dai dati: l'esperto statistico deve limitarsi a definire le situazioni di errore, demandando ad un prefissato algoritmo il compito di riportare il record ad una situazione di correttezza.

L'approccio probabilistico ha il suo riferimento nella cosiddetta metodologia Fellegi-Holt (Fellegi e Holt 1976).

Un piano probabilistico è composto, da regole di incompatibilità, che seguendo la terminologia di Fellegi e Holt, vengono chiamate edit in forma normale. Un edit in forma normale è costituto dalla congiunzione di due o più condizioni sui valori di variabili del record: l'edit è attivato da un dato record quando sono verificate simultaneamente tutte le condizioni in esso definite. La parte SE di una R.I.D. (cioè quella che esprime la situazione di errore) può corrispondere a uno o più edit in forma normale.

L'algoritmo che elimina gli errori provvede a determinare, per ogni record e per ogni situazione di errore, le variabili da modificare in modo da avere la certezza di eliminare gli errori individuati e, soprattutto, di non introdurne altri nel record, minimizzando nel contempo il numero di variabili modificate.

Gli edit in forma normale definiti dall'esperto, gli edit espliciti, sono sufficienti ad individuare la presenza di errori all'interno dei record di un file, ma non a garantire una imputazione di valori corretta ed ottimale. Infatti,

la scelta di quali variabili modificare e di quali nuovi valori assegnare, è condizionata dai vincoli di correttezza (non introdurre nuovi errori nel record) e di minimalità (modificare il minor numero possibile di variabili). A tal fine, occorre considerare anche i cosiddetti edit impliciti, derivabili da quelli espliciti ed individuare così l'insieme minimo e completo degli edit.

La metodologia di Fellegi-Holt prevede che, una volta definiti gli edit espliciti, questi siano analizzati sia per scoprire la presenza di contraddizioni e/o ridondanze che per derivare tutti gli edit impliciti in essi contenuti.

La fase dell'analisi e della derivazione degli edit, produce un insieme di regole che ha le seguenti caratteristiche:
1. è minimale, privo cioè di edit ridondanti;
2. è corretto, privo di edit tra loro contraddittori;
3. è completo, in quanto contiene esplicitamente tutti gli edit definiti implicitamente all'interno di quelli iniziali.

La derivazione degli edit impliciti nell'ambito della metodologia Fellegi-Holt rappresenta un'operazione altamente critica: infatti la generazione degli edit impliciti richiede un numero di operazioni che è esponenziale rispetto al numero di edit espliciti. Spesso la derivazione degli edit impliciti risulta impossibile; in questo caso si ricorre ad euristiche che permettono di limitare a priori il numero delle operazioni necessarie e alla partizione dell'insieme iniziale di edit suddividendo la fase di correzione in tante sottofasi quanti sono i sottoinsiemi di edit cosi definiti.

Quali sono i vantaggi e gli svantaggi dei due diversi approcci? Molto schematicamente, possiamo ascrivere ai vantaggi del metodo deterministico:
- la completa applicabilità: una procedura deterministica è sempre applicabile ai dati una volta tradotte le regole di imputazione deterministica in istruzioni di un programma;
- l'efficienza elaborativa: il tempo necessario per eseguire il programma che traduce la procedura deterministica è lineare rispetto al numero di regole e al numero di record;
- l'orientabilità degli effetti: lo statistico può orientare i risultati dell'applicazione della procedura deterministica definendo in modo opportuno la parte imputazione di ogni regola, e la sequenza di queste nel piano.

Questo ultimo elemento è di una certa importanza: ad esempio, sulla base della fiducia che lo statistico nutre rispetto alla correttezza delle variabili, egli può implicitamente stabilire una gerarchia tra queste, orientando la modifica verso quelle che egli ritiene meno affidabili.

Tra gli svantaggi ed i limiti del deterministico citiamo:

- la mancata garanzia di correttezza dei record alla fine della fase di correzione e la conseguente necessità di cicli di controllo e correzione;
- la mancata garanzia di minimizzazione dei cambiamenti;
- l'introduzione di distorsioni nelle distribuzioni e la perdita di variabilità.

In caso di errori sistematici, l'approccio deterministico si rivela, nella maggior parte dei casi, il più adatto, soprattutto nel passo di localizzazione degli errori. L'applicazione del probabilistico, al contrario, rischia di introdurre nuove distorsioni nei dati, qualora non si pesino opportunamente le variabili per tener conto della sistematicità di tali errori.

I vantaggi dell'approccio probabilistico, speculari ai limiti di quello deterministico, sono:

- la correttezza finale dei record sottoposti a correzione;
- la minimalità del cambiamento (assicurata dallo stesso algoritmo che provvede a trovare il numero minimo di variabili, l'insieme minimale, da modificare);
- un migliore mantenimento della distribuzione congiunta delle variabili.

Una volta che siano state individuate le variabili affette da errore che hanno causato l'attivazione delle incompatibilità, oppure i cui valori sono stati giudicati outlier, occorre procedere alla fase di imputazione di tali variabili, onde rimuovere gli errori, cercando di ripristinare i valori veri. I possibili metodi per l'imputazione sono numerosi (Kovar e Whitridge 1995).

Tra questi citiamo:

- imputazione da valore prefissato: nella parte ALLORA della regola si definisce la variabile errata da correggere e viene al contempo indicato il valore da assegnare a tale variabile;
- imputazione da serie storica: per variabili che tendono ad essere stabili nel tempo, in caso di imputazione viene riproposto il valore disponibile nel periodo immediatamente precedente. Come varian-

te, tale valore viene "aggiustato" per tener conto del trend della serie storica relativa alla variabile (solo per variabili quantitative);

- imputazione del valor medio: alla variabile viene imputato il valor medio calcolato sui dati disponibili, o in un opportuno strato di questi (è un metodo che può essere utilizzato solo per le variabili quantitative). Lo svantaggio è che in tal modo viene introdotta una seria distorsione nella distribuzione della variabile, creando un picco artificiale in corrispondenza del suo valor medio;
- imputazione sequenziale da donatore "hot deck": in una data variabile il valore errato viene sostituito dal valore corrispondente della ultima unità rispondente. Con questo metodo è estremamente importante l'ordinamento cui è sottoposto il file oggetto della correzione: le variabili di ordinamento sono quelle rispetto alle quali è assicurata la minima distanza tra il record ricevente e quello donatore. Un possibile aspetto negativo di tale metodo è nel fatto che uno stesso donatore può essere utilizzato più volte, tante quanto la dimensione di un insieme di record adiacenti che necessitano di correzione: ciò può creare picchi artificiali nei valori della variabile;
- imputazione dal più vicino donatore: la differenza col metodo precedente consiste nel fatto che il donatore è scelto in modo tale che una qualche misura della distanza tra esso ed il ricevente risulti minima. In genere, la distanza scelta non è di tipo spaziale, ma una misura multivariata basata sui dati disponibili: per tale ragione, il metodo è più appropriato per le variabili quantitative. Tra i vantaggi, citiamo quello relativo al mantenimento ottimale delle distribuzioni multivariate originali. Lo svantaggio è comune al metodo precedente: uno stesso donatore può essere utilizzato più volte; esiste però la possibilità di limitare questo svantaggio, ponendo un tetto al numero di volte che uno stesso record può essere utilizzato come donatore, oppure introducendo nella funzione di distanza una funzione di penalizzazione che tiene conto del numero di volte che un dato record è già stato utilizzato come donatore;
- imputazione da regressione: per l'imputazione di una data variabile viene utilizzato il valore fornito da una funzione di regressione che fa uso di una o più variabili ausiliarie. La variabile da imputare deve essere quantitativa, mentre le variabili indipendenti possono essere

continue o discrete. Il metodo assicura buoni risultati sotto due condizioni: (i) alta correlazione tra variabile da imputare e variabili ausiliarie e (ii) disponibilità di valori corretti delle variabili ausiliarie per tutti i (o per gran parte dei) record. Un caso particolare è dato dall'imputazione da rapporto (ratio) in cui viene considerata la relazione tra la variabile da imputare ed una variabile ausiliaria con essa altamente correlata: in tal caso entrambe devono essere di tipo continuo. Questo metodo si rivela adeguato nei casi in cui la variabile da imputare è affetta da un errore di tipo stocastico, oppure sistematico ma il cui andamento è legato alla variabile ausiliaria.

La definizione, lo sviluppo e la messa a punto di una procedura automatica per il controllo e la correzione dei dati dovrebbero essere finalizzati a far sì che questa:

- localizzi ed elimini il maggior numero di errori possibile;
- non introduca distorsioni nei dati.

Tra i due approcci descritti in precedenza, è quello probabilistico l'unico in grado di assicurare questo tipo di risultato, almeno in una situazione di tipo "ideale", tale cioè che la tipologia degli errori presenti nei dati sia di carattere stocastico, o quantomeno che la componente sistematica negli errori sia trascurabile. Se ciò non avviene, se cioè gli errori sistematici sono presenti in quantità tale da non poter essere considerati trascurabili, deve essere introdotta una specifica componente deterministica nella procedura, dato che è dimostrato che l'approccio probabilistico non è adatto al trattamento di tali errori, ma anzi è suscettibile di introdurre ulteriori distorsioni nei dati.

La soluzione ottimale dovrebbe prevedere il trattamento congiunto in un unico passo di entrambe le tipologie di errore (Barcaroli 1998). Nella pratica questo non è possibile, non disponendosi ancora di implementazioni degli opportuni algoritmi. In fase di disegno della procedura complessiva occorre quindi:

a) prevedere la massimizzazione del ricorso all'approccio probabilistico, disegnando in primo luogo un piano di compatibilità che ricalchi i principi della metodologia Fellegi-Holt;

b) individuare quindi le eventuali componenti sistematiche dell'errore e

prevedere, come eccezione, l'applicazione di procedure deterministiche per la loro rimozione.

In prospettiva, qualora si possa intervenire sul processo di raccolta e registrazione dei dati, e si abbia quindi la possibilità di rimuovere le cause che producono gli errori sistematici, occorre procedere in tal senso, al fine di minimizzare e, al limite, eliminare il ricorso a passi di tipo deterministico (che sono comunque suscettibili di introdurre distorsioni addizionali nei dati).

Tutto ciò implica che la fase di messa a punto delle procedure non è finalizzata solo ad una ottimizzazione della procedura probabilistica ideata nella fase di disegno (verifica della completezza e correttezza del piano di compatibilità), ma anche all'individuazione della componente sistematica degli errori (per lo sviluppo di passi deterministici), ed alla identificazione delle cause di tali errori (per la loro rimozione dal processo produttivo).

Il software CONCORD (CONtrollo e CORrezione dei Dati) permette di applicare sia l'approccio probabilistico che quello deterministico, quest'ultimo integrato col metodo del donatore, mediante una metodologia la cui sequenza di passi è contenuta nella figura 1.1.

Nel software sono infatti disponibili tre diversi moduli, sviluppati a suo tempo indipendentemente presso l'ISTAT:

- il modulo SCIA (Sistema Controllo ed Imputazione Automatici), che permette l'applicazione integrale della metodologia Fellegi-Holt per la localizzazione degli errori e per la loro imputazione, limitatamente alle variabili categoriche;
- il modulo GRANADA (Gestione delle Regole per l'ANAlisi dei DAti), che permette di effettuare la localizzazione deterministica degli errori mediante l'applicazione di regole del tipo SE-ALLORA a variabili sia di tipo categorico che continuo;
- il modulo RIDA (Ricostruzione Informazioni con Donazione Automatica) che permette l'imputazione mediante donatore sia di variabili categoriche che continue.

Nella metodologia proposta, mediante l'utilizzo del modulo SCIA[2] in CONCORD è possibile effettuare le operazioni 1 e 2, illustrate in figura 1.1, di definizione ed esecuzione del passo probabilistico della procedura complessiva di controllo e correzione.

Il passo di definizione prevede:

- l'indicazione degli edit che rappresentano le regole, formali e sostanziali, individuabili a partire dal questionario e dalla conoscenza relativa ai fenomeni indagati (insieme iniziale di edit);
- la generazione dell'insieme minimale di edit, ottenuto da quello iniziale mediante un processo di eliminazione degli edit ridondanti e contraddittori;
- la generazione dell'insieme completo di edit, ottenuto da quello minimale mediante generazione di tutti gli edit impliciti, quelli cioè logicamente contenuti negli edit iniziali, la cui esplicitazione è fondamentale ai fini della correttezza della localizzazione degli errori.

Il passo di esecuzione prevede l'applicazione dell'insieme completo di edit così ottenuto all'insieme dei dati da trattare. Ciò produce un insieme di statistiche (record esatti e record errati; distribuzione degli edit per frequenza di attivazione; variabili per frequenza di imputazione) il cui esame da parte dello statistico (operazione 3: analisi dei risultati) permette l'individuazione di eventuali errori sistematici.

Qualora questi esistano, l'utilizzo congiunto dei moduli GRANADA e RIDA permette di effettuare le operazioni 4 e 5 di definizione ed esecuzione del passo deterministico.

GRANADA[3] permette di definire le regole di tipo SE-ALLORA già introdotte. Tenendo conto che la parte SE di tali regole esprime la stessa condizione di errore definita in un corrispondente edit del passo probabilistico, esiste la possibilità di importare tutte le regole già definite mediante SCIA, inizializzando in tal modo il modulo deterministico. L'utente non dovrà far altro che scegliere quali regole mantenere, e per

[2] SCIA è stato sviluppato da E. Riccini Margarucci, F. Silvestri e P. Floris

[3] GRANADA è stato sviluppato da E.Riccini Margarucci, P.Floris, R.Ciacci e T.Buglielli

queste indicarne la parte ALLORA, che corrisponde alla localizzazione deterministica dell'errore.

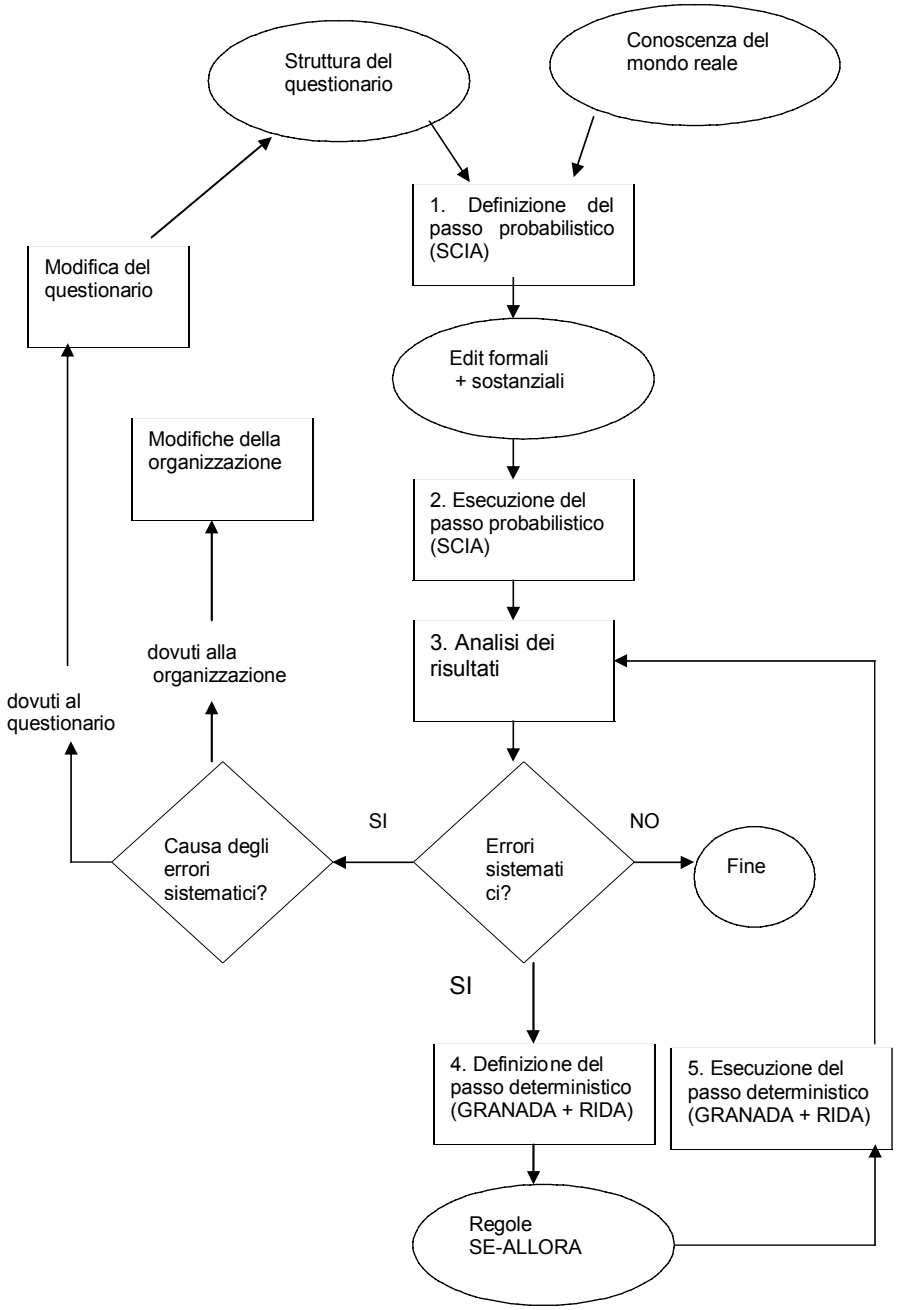
A questo punto, applicando le regole così definite, è possibile bipartire l'insieme iniziale dei dati in due sottoinsiemi, quello dei dati esatti e quello dei dati contenenti errori.

GRANADA consentirebbe anche di imputare direttamente le variabili giudicate errate, indicando il valore puntuale da assegnare; da un punto di vista statistico questa operazione è però da evitare, o quantomeno da ridurre al minimo, in quanto può comportare rilevanti distorsioni delle distribuzioni originali. È bene quindi limitare l'uso di GRANADA all'impostazione di caratteri di controllo nelle variabili giudicate errate, caratteri che verranno utilizzati dal modulo RIDA per riconoscere i valori da imputare.

Mediante RIDA[4] la correzione si esegue prelevando i nuovi valori da un record corretto simile al record errato (Abbate 1996). La similitudine si calcola utilizzando alcune variabili, dette di "match", scelte sulla base della loro correlazione con la variabile da correggere. Questo metodo presuppone che le variabili utilizzate per calcolare la distanza fra record errato e donatore siano corrette. Per la ricerca del donatore si procede a confrontare il record errato con tutti i record esatti, scegliendo quello con distanza minima. Le variabili, utilizzate per individuare la similitudine fra i record, si distinguono in variabili di strato e variabili di match. Le variabili di strato si utilizzano per limitare la ricerca all'interno di sottoinsiemi di record che presentano uguali valori di tali variabili. Le variabili di match si utilizzano per calcolare la funzione di distanza mista per tutti i record dello strato. Il donatore prescelto è quello più vicino al record errato, cioè quello con distanza minima.

[4] RIDA è stato sviluppato da G.Massimini, T.Buglielli e R.Colosi

Figura 1.1 - La metodologia per la messa a punto della procedura di controllo e correzione mediante i diversi moduli di CONCORD



2. L'installazione

Sintesi: *in questo capitolo vengono riportati i requisiti hardware e software richiesti dal software CONCORD e la procedura d'installazione.*

2.1. Requisiti hardware e software

CONCORD è un sistema sviluppato utilizzando il *SAS SYSTEM v 8.1*, il *Cobol*, il *Fortran* e il *C* per il sistema operativo Microsoft Windows. Per quanto riguarda SAS, è stato utilizzato sia il linguaggio del SAS base che l'SCL, linguaggio di programmazione particolarmente adatto alla creazione di applicazioni interattive, sviluppate tramite il software SAS/AF. Per utilizzare CONCORD è dunque necessario che sia installato il sistema SAS versione 8 ed in particolare i moduli:

- SAS base e core
- SAS FS
- SAS GRAPH

Tutti gli altri programmi che fanno parte di CONCORD sono in versione eseguibile. Lo spazio sul disco fisso necessario per l'installazione è di circa *6 MB*.

È necessaria una memoria RAM di *64 MB* o superiore.

Il tempo d'esecuzione dei diversi programmi della procedura è legato, ovviamente, alla velocità del processore installato e alla grandezza e complessità dei dati.

Per l'installazione è necessario avere il programma WINZIP, per poter espandere il file CONCORD.zip (*vedi paragrafo 2.2*).

2.2. La procedura dI installazione ed avvio

1. Espandere il file CONCORD.zip nella cartella **c:\concord** eseguendo le seguenti indicazioni:
   - eseguire il programma di estrazione Winzip;
   - selezionare "I agree" in modo che venga mostrato il contenuto del file;
   - selezionare "Extract";
   - nella successiva maschera, nel campo "Extract to", scrivere **c:\concord** come indicato in figura 2.1 e poi premere il bottone "Extract".

Figura 2.1: Maschera d'utilizzo di WinZip

Il programma crea la nuova cartella "concord", o la utilizza se preesistente, ed estrae i file contenuti nello zip file. Fare attenzione che non venga creata dal programma di estrazione una sottocartella concord in **c:\concord**.

2. Eseguire il programma **installa.bat**, con doppio click del mouse sul file, eventualmente scrivendo, se richiesto, i riferimenti se il SAS è residente in una cartella diversa da quella consigliata di default ("Programmi\SAS Institute\SAS\V8\sas.exe").

Viene eseguito prima un programma che copia il file di configurazione dal SAS nella cartella c:\concord aggiungendo in fondo al file di configurazione alcuni parametri; viene poi eseguito in batch il SAS, che importa il catalogo dei programmi dal file "sascat.dat".

L'esecuzione del programma **installa.bat** deve essere effettuata anche nel

caso di installazioni successive alla prima. è possibile l'avvio della procedura cliccando due volte sull'icona di collegamento a CONCORD cambiando prima eventualmente i riferimenti al SAS se quest'ultimo è stato caricato in una cartella diversa da quella di default[5].

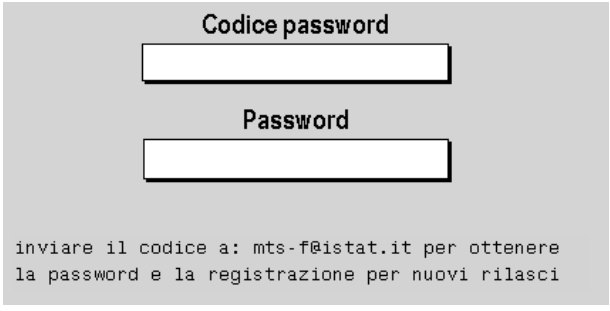
Le suddette istruzioni di installazione sono contenute anche nel file read.me. Una volta installato il software, la cartella "c:\concord" conterrà due cartelle "Esempio" e "help". La cartella "Esempio" contiene dei file con alcuni dati di esempio.

La cartella "help" contiene i file help-on-line; tali file possono essere stampati con la seguente procedura:

- attivare "Internet Explorer" ;
- scegliere "File" e "Open";
- da "browse" selezionare la cartella "c:\concord\help" e il file "lancio.htm" e poi "Ok";
- da "File" scegliere "Print";
- selezionare "Print all linked document";
- l stampare con "Ok".

Alla prima esecuzione di CONCORD viene mostrata una maschera (*vedi figura 2.2*) con evidenziato un codice nel campo "Codice per password"; questo codice deve essere trasmesso via e-mail all'indirizzo scritto nella maschera per poter ottenere la password di esecuzione.

Figura 2.2: Maschera per la password



[5] Ciò è possibile cliccando col pulsante destro sull'icona e scegliendo l'opzione "Collegamenti"

Nella e-mail di richiesta devono essere indicati:
1. nome e cognome della persona richiedente;
2. ente o istituto di appartenenza;
3. settore interno;
4. campo di potenziale applicazione del sistema.

Una volta in possesso della password eseguire di nuovo CONCORD e scriverla nel campo "Password" in modo che il sistema passi alle maschere successive.

Questa procedura deve essere effettuata solo alla prima installazione e non ai successivi rilasci del software.

Il software deve essere *sempre* installato con la procedura suddetta e *mai* copiato da un computer all'altro. Ogni installazione deve avere la sua diversa password di registrazione.

3. L'utilizzo del software

Sintesi: *Il capitolo descrive in modo dettagliato l'utilizzo dell'interfaccia del software CONCORD. I paragrafi aiutano l'utente ad utilizzare la schermata principale e le varie funzioni del software.*

3.1. I dati input del software

I dati di input per CONCORD, che chiameremo "*dati grezzi*", sono i dati provenienti dai piani di registrazione dei questionari delle indagini.

CONCORD elabora record registrati in ASCII a lunghezza fissa con caratteri di tipo testo.

Nel caso che un piano di registrazione preveda più tipi record, e che tipi record diversi debbano essere sottoposti a controllo e correzione tramite CONCORD, sarà necessario trasformare in un solo record tutti i record del questionario, o estrarre i vari tipi record in tanti file e sottoporre ogni file ad elaborazione separata; ovviamente, in quest'ultimo caso, ogni file necessiterà della propria descrizione dei campi da trattare.

Deve essere, quindi, cura dell'utente sottoporre all'elaborazione solamente i record da trattare, poiché non è possibile nel caso di registrazione multi-record dei dati, l'elaborazione di record con tracciati diversi.

3.2. La schermata principale

Il software viene attivato tramite l'icona del programma presente nella cartella "c:\concord" dopo aver effettuato l'installazione *(vedi § 2.2)*.

Con l'avvio della procedura, si apre la schermata principale *(vedi figura 3.1)*.

Figura 3.1 - La schermata principale

Nella schermata principale viene presentato un menu nel quale compaiono le seguenti voci:

- *Progetto*, per aprire o chiudere un progetto o uscire dal software;
- *Definizioni*, per entrare nella fase di definizione;
- *Funzioni*, per attivare una delle possibili funzioni;
- *Analisi*, per analizzare i risultati di elaborazioni effettuate;
- *Utilità*, per attivare un programma di utilità;
- *Log*, per visualizzare la finestra di log del SAS;
- *History*, per visualizzare i passi effettuati nei vari progetti;
- *Help*, per mostrare l'Help on line.

3.3. Uso della schermata principale - il progetto

Una delle prime scelte possibili, una volta attivato CONCORD, è il "*progetto*". Con questa scelta possiamo uscire dal software oppure scegliere tra una nuova elaborazione o una elaborazione precedentemente effettuata. Per "*progetto*" in CONCORD si intende il nome della cartella nella quale risiedono tutti i file e dataset generati dalle varie funzioni, e che è anche il nome con il quale viene assegnata la libreria SAS al momento dell'esecuzione.

Dal menu Progetto con "*Nuovo*", sempre attivo, si definisce un nuovo progetto scegliendo una cartella (directory), che può essere creata utilizzando l'apposito simbolo dopo aver selezionato il percorso. 

Figura 3.2 - Il Progetto

Il nome della nuova cartella deve rispettare la sintassi dei nomi di libreria SAS e cioè al massimo otto caratteri di cui il primo alfabetico e senza caratteri speciali (*es. redditi, forzelav, ecc.*). Scelto il nome del progetto si deve scegliere il tipo di correzione e, a conferma avvenuta, viene assegnato il progetto come libreria al SAS, scritto un record di progetto nel dataset "*metadati*" nella cartella "c:\concord", e registrati nella cartella di progetto tutti i dataset e i file necessari all'esecuzione dei vari programmi del sistema. Il nome del progetto scelto sarà mostrato nel titolo di tutte le maschere principali.

Con "*Apri*", attivo solo se precedentemente è stato elaborato un progetto, cioè se esiste almeno una osservazione nel dataset *"metadati"* nella cartella "c:\concord", si sceglie un progetto tra quelli che vengono mostrati, corrispondente a una cartella (directory). È possibile rimuovere un progetto, dopo averlo scelto, cliccando sul tasto con il simbolo di cancellazione a fianco del nome del progetto e, dopo conferma, il nome del progetto viene eliminato dal dataset *"metadati"* e da *"history"* lasciando inalterato il contenuto della cartella relativa.

Con "*Chiudi*", attivo quando un progetto è stato scelto, si chiude il pro-

getto in corso e si aggiorna automaticamente il record corrispondente nel dataset "*metadati*".

In una stessa cartella, e quindi nello stesso progetto, possono coesistere i vari tipi di correzione: probabilistica, deterministica o tramite donatore.

Si può passare *da un tipo all'altro di correzione* chiudendo e riaprendo il progetto con lo stesso nome e scegliendo il tipo di correzione opportuno.

Dopo avere scelto il progetto è necessario impostare il tipo di correzione che si vuole effettuare sui dati scegliendo uno dei tre approcci possibili:

- Probabilistica
- Deterministica
- Donatore

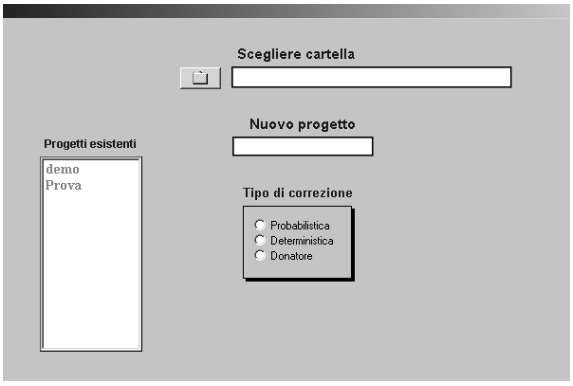


Figura 3.3 - La scelta dell Progetto

4. L'approccio di correzione probabilistica

Sintesi: *sono brevemente descritte, la metodologia e le varie fasi, da eseguire in modo gerarchico dall'utente, per la correzione effettuata secondo l'approccio probabilistico.*

Con questo metodo si ottiene la correzione di un file di dati con l'approccio di tipo probabilistico. Da Barcaroli et al (1999) sono di seguito descritti i principi su cui la tecnica si basa.

L'*approccio probabilistico* non impone la necessità di definire a priori, per ogni situazione di errore, l'elenco delle azioni da intraprendere per eliminare gli errori dai dati: l'esperto statistico deve limitarsi a definire le situazioni di errore, demandando ad un prefissato algoritmo il compito di riportare il record ad una situazione di correttezza.

L'approccio probabilistico, sviluppato in CONCORD, ha il suo punto di riferimento nella cosiddetta *metodologia Fellegi-Holt*, esposta nell'articolo "A systematic approach to automatic edit and imputation" di I.Fellegi e D.Holt, pubblicato nel 1976 sul Journal of the American Statistical Association.

Un piano probabilistico è composto da regole di incompatibilità, che seguendo la terminologia di Fellegi e Holt, sono chiamate *edit in forma normale*. Un edit in forma normale è costituto dalla congiunzione di due o più condizioni sui valori di variabili del record: l'edit è attivato, in un dato record, quando sono verificate simultaneamente tutte le condizioni in esso definite. La parte SE di una regola di incompatibilità, cioè quella che esprime la situazione di errore, può corrispondere a uno o più edit in forma normale.

L'algoritmo che elimina gli errori provvede a determinare, per ogni record e per ogni situazione di errore, quali variabili modificare, in modo da avere la certezza di eliminare gli errori individuati e, soprattutto, di non introdurne altri nel record, minimizzando nel contempo il numero di variabili modificate.

Gli edit in forma normale definiti dall'esperto, detti edit espliciti, sono sufficienti ad individuare la presenza di errori all'interno dei record di un file, ma non a garantire una imputazione di valori corretta ed ottimale. Infatti, la scelta di quali variabili modificare e di quali nuovi valori assegnare, è condizionata dai vincoli di correttezza - non introdurre nuovi errori nel record - e di minimalità - modificare il minor numero possibile di variabili. A tal fine, occorre considerare anche i cosiddetti edit impliciti, derivabili da quelli espliciti ed individuare così l'insieme minimo e completo degli edit.

La metodologia di Fellegi-Holt prevede che, una volta definiti gli edit espliciti, questi siano analizzati, sia per scoprire la presenza di contraddizioni e/o ridondanze, che per derivare tutti gli edit impliciti in essi contenuti.

La fase dell'analisi e della derivazione degli edit, produce un insieme di regole che ha le seguenti caratteristiche:

- è minimale, privo cioè di edit ridondanti;
- è corretto, privo di edit tra loro contraddittori;
- è completo in quanto contiene, in forma esplicita, tutti gli edit che sono implicitamente definiti all'interno di quelli iniziali.

La derivazione degli edit impliciti nell'ambito della metodologia Fellegi-Holt, rappresenta un'operazione altamente critica: infatti la generazione degli edit impliciti richiede un numero di operazioni che è esponenziale rispetto al numero di edit espliciti. Talvolta, la derivazione degli edit impliciti risulta impossibile, anche ricorrendo ad euristiche che permettono di limitare a priori il numero delle operazioni necessarie; in questo caso si può ripartire l'insieme iniziale di edit suddividendo la fase di correzione in tante sottofasi quanti sono i sottoinsiemi di edit cosi definiti.

Prima fase, nella stesura del piano di incompatibilità, è quella della definizione delle variabili, eventualmente delle liste di variabili, e delle regole di incompatibilità.

4.1. La fase di definizione

Figura 4.1 - Funzioni di definizione

4.1.1. Definizione delle variabili

Propedeutica ad ogni altra funzione è la definizione delle variabili, cioè dei campi del record da sottoporre a controllo ed eventuale correzione. Nell'approccio probabilistico sono trattate variabili *categoriche o qualitative* con valori codificati da 0 a 9999. Per definire le variabili è necessario avere il tracciato record con indicate le posizioni iniziali e le lunghezze di ogni variabile da trattare.

Nei record di input le variabili da trattare devono essere completate con 0 a sinistra del valore per tutta la lunghezza del campo; quindi i campi più lunghi di un carattere devono essere completati con zeri iniziali.

È necessario definire solo le variabili che vogliamo sottoporre a controllo; il resto del record verrà ricopiato automaticamente senza modifiche.

Il numero massimo di variabili definibili è 500.

Per ogni variabile è necessario indicare:

- il nome, univoco e lungo al massimo 6 caratteri (es.: ETA, Q12, COL1_6); il nome deve essere univoco anche rispetto ai nomi di eventuali LISTE *(vedi § 4.1.2);*
- la posizione, cioè la colonna del tracciato record dove inizia la variabile, in un valore compreso tra 1 e 5000;
- la lunghezza della variabile, in un valore compreso tra 1 e 4: una variabile non può superare i 4 caratteri di lunghezza.

Esempi di variabili:

ETÀ	posizione 1,	lunghezza 1;
SESSO	posizione 5,	lunghezza 1;
REDDIT	posizione 345,	lunghezza 3.

Il programma controlla che le posizioni insieme alle lunghezze non si sovrappongano.

Il nome, la posizione, la lunghezza e i domini di una variabile possono essere modificati scegliendo la variabile da modificare, modificandola dove necessario e selezionando "add/mod" dopo la modifica.

La cancellazione di una variabile si effettua scegliendola dalla lista e selezionando "delete"; con la cancellazione si perdono anche i relativi domini *(vedi sotto)*.

Con "*clear*" si annullano le modifiche che riguardano la variabile.

Definita una variabile bisogna passare all'inserimento dei valori dei domini di una variabile selezionando "*DOMINI* ".

Per dominio di una variabile si intendono i valori che la variabile può assumere, ossia le possibili risposte ad un quesito del questionario dell'indagine in oggetto, per un massimo di *700*; l'eventuale ammessa assenza di risposta "*blank*" va indicata con "B". Nella fase di inserimento della variabile, scritti il nome la posizione e la lunghezza, bisogna definire i valori del dominio, espressi in forma di intervallo "da a".

Se per esempio una qualsiasi variabile può' assumere i valori 1,5,8,9 ed è ammessa la risposta mancante:

- scrivere 1 nella zona da e 1 nella zona a e dare invio o "add/mod";
- scrivere 5 nella zona da e 5 nella zona a e dare invio o "add/mod";
- scrivere 8 nella zona da e 9 nella zona a e dare invio o "add/mod";
- scrivere b nella zona da e b nella zona a e dare invio o "add/mod";

Il programma controlla che il dominio di una variabile non si sovrapponga con gli altri domini della stessa variabile, e non superi la lunghezza definita per la variabile stessa.

Per *aggiungere* ulteriori domini ad una variabile:

- selezionare la variabile verranno evidenziati i valori del dominio;
- selezionare "DOMINI ";
- aggiungere i domini necessari nel modo sopraindicato.

Per *modificare* il dominio di una variabile:

- selezionare la variabile verranno evidenziati i valori del dominio;
- selezionare "DOMINI ";
- scegliere il dominio da modificare;
- riscrivere il dominio e premere invio o "add/mod".

Per *eliminare* un dominio di una variabile:

- selezionare la variabile;
- selezionare "DOMINI";
- scegliere il dominio da cancellare;
- premere "delete".

Con "*clear*" si annullano le modifiche del dominio selezionato.

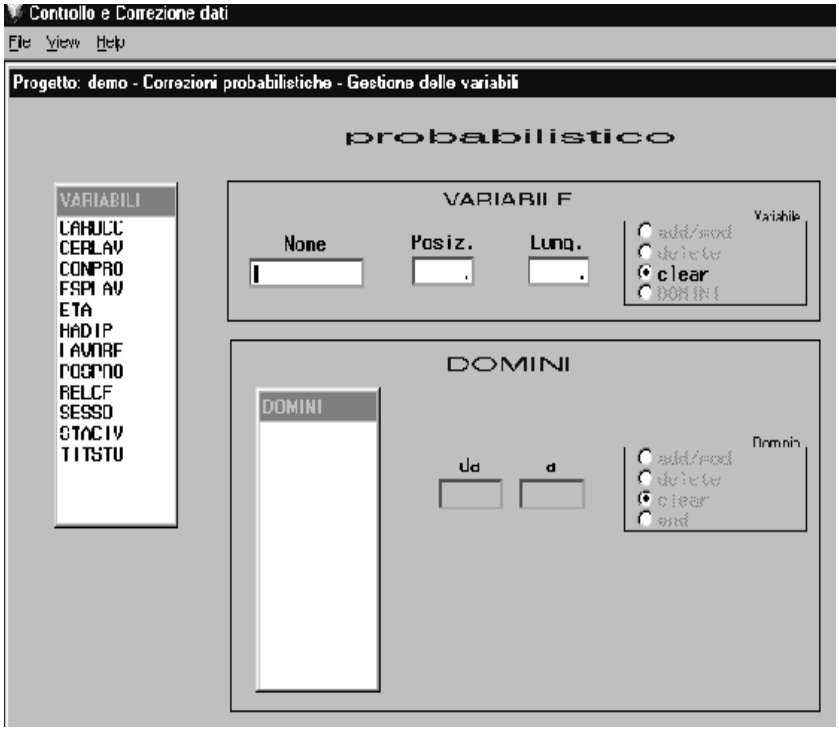
Terminato l'inserimento o la modifica dei domini la variabile viene aggiornata scegliendo "*end*".

Dal menu a tendina "*File*" *(vedi figura 4.2)*:

- Con "save and exit" si crea o aggiorna nella cartella di progetto il file esterno "VARDOM.dat" che conterrà la descrizione formattata delle variabili;
- Con "exit" si può uscire senza salvare le modifiche effettuate nelle variabili;
- Con "import" si importano variabili da un file esterno con tracciato formattato, uguale a quello del file "VARDOM.dat", da scegliere in una cartella;
- Con "export per deterministico" si registra nella cartella di progetto, attuando così una prima integrazione tra gli approcci probabilistico e deterministico, la descrizione delle variabili sul file esterno "fvardom.dat" in formato adatto all'approccio deterministico.

"View" visualizza tutte le variabile preesistenti ordinandole per posizione.

Figura 4.2 La gestione delle variabili



4.1.2. Definizione delle liste di variabili

Le liste di variabili , utili per agevolare la scrittura delle regole, sono insiemi di variabili la cui risposta nel questionario può mancare o non mancare in funzione del valore di un'altra variabile.

Tipicamente appartengono ad una lista tutte le variabili di una sezione del questionario che deve essere compilata o no sulla base di un quesito "*filtro*" precedente, oppure sezioni di questionario che ammettono molte risposte multiple.

Scrivendo una sola regola, che indica incompatibilità tra una variabile e la variabile di lista, saranno automaticamente generate, nella fase di controllo *(vedi § 4.2.1)*, *tante regole* quante sono le variabili indicate nella lista se la lista è stata definita ***OR***, o una regola che comprende tutte le variabili della lista se questa è stata definita *AND*.

Esempio:
poniamo che se l'età è minore di 14 anni non deve essere compilata una sezione del questionario che comprende sei variabili; dovremmo scrivere sei regole: una tra ETA e VAR1, un'altra tra ETA e VAR2, …e infine una tra ETA e VAR6; invece creando una lista che chiameremo "LISETA" comprendente le variabili precedentemente definite VAR1, VAR2... VAR6 potremo scrivere una sola regola (cfr. definizione delle regole) tra ETA e LISETA:

ETA(0-13) LISETA<)

Se la variabile LISETA è stata dichiarata OR, definizione più comune per le variabili lista, saranno automaticamente generate nella fase di controllo delle regole, sei regole di incompatibilità, una tra la variabile ETA e la variabile VAR1, una tra la variabile ETA e la variabile VAR2, e così via.

Se invece la variabile LISETA è stata dichiarata AND verrà automaticamente generata, nella fase di controllo delle regole, una sola regola di incompatibilità, tra variabile ETA e tutte le variabili elencate in LISETA.

Una lista può comprendere al massimo 100 variabili.

Per *definire* una lista:

- scrivere un nome univoco, anche rispetto alle variabili, lungo al massimo *6* caratteri;
- scegliere "*O*" per liste in *OR* oppure "*A*" per liste in *AND*.

Vengono mostrate *tutte le variabili precedentemente definite la cui risposta può mancare*, cioè che ammettono "B" nel dominio *(vedi § 4.1.1)*.

Per *aggiungere* variabili ad una lista:

- scegliere con il mouse le variabili da inserire nella lista; le variabili così scelte faranno parte della lista e verranno automaticamente spostate dal listbox "scegli variabile" al listbox "variabili in lista" *(vedi figura 4.3)*.

Per *cancellare* variabili da una lista:

- le variabili possono essere cancellate da una lista nella quale erano state precedentemente inserite, scegliendole con il mouse dal listbox "variabili in lista"; viene compiuta automaticamente un'operazione inversa alla precedente;
- alla fine premere "*add/mod*" nella zona lista per aggiornare la lista.

Per *modificare* una lista:

- scegliere la variabile nel listbox "LISTE" e poi eseguire le operazioni sopra descritte;
- alla fine premere "*add/mod*" nella zona lista per aggiornare la lista.

Per *cancellare* una lista:

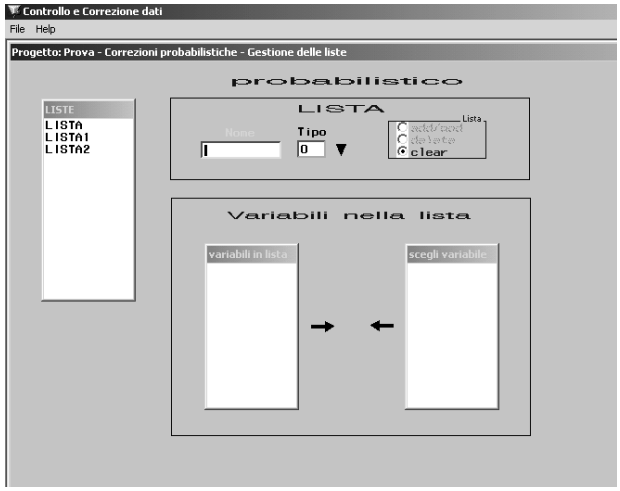
- scegliere la variabile nel listbox "LISTE";
- premere "*delete*".

Con "*clear*" si annullano le modifiche alla lista.

Dal menu a tendina "file" sono possibili le seguenti scelte:

- Con "*exit*" si esce senza memorizzare alcuna operazione effettuata;
- Con "*save and exit*" vengono memorizzate le modifiche e registrati, con apposito formato, nella cartella di progetto i file esterni "STRUTT.dat" per le liste definite in "*OR*", e "LISTE.dat" se esistono liste definite in "*AND*";
- Con "*import*" è possibile, dopo avere scelto la cartella tramite apposita maschera, caricare le liste da file esterni, scritti in formato apposito, del tipo "STRUTT.dat" e "LISTE.dat".

Figura 4.3 La gestione delle liste



4.1.3. Definizione delle regole di incompatibilità

Le regole di incompatibilità o "*edit*" scritte dall'utente, per una massimo di 2.000, sono chiamate "*insieme minimale*" e descrivono l'*incompatibilità* tra *le variabili* considerate nei loro *sottodomini*.

L'esperto dell'indagine deve definire, quindi, quali sono le *incompatibilità* possibili nelle risposte ai quesiti del questionario e trascriverle sotto forma di regola, mettendo, se gli è più consono, in forma negativa o "*diverso da*" una risposta possibile o compatibile.

Ad esempio, se in un questionario lo stato civile "non libero" è compatibile solo con l'età maggiore di 14 anni e dobbiamo scrivere una regola per controllare le risposte, diremo:

se stato civile non libero ed età > 14 anni

per considerare compatibili le risposte a quesiti, trasformandola poi in incompatibilità scrivendo:

se stato civile non libero ed età 0-13 anni.

Nell'approccio probabilistico, vengono trattate solo variabili di tipo categorico, anche se codificate numeriche, con soli controlli di uguaglianza, e non sono possibili operazioni di tipo aritmetico tra le variabili; inoltre bisogna considerare sottoinsiemi del dominio di una variabile poiché, se ne considerassimo l'intero dominio, la regola sarebbe sempre verificata qualsiasi fosse il valore della variabile a confronto.

Non bisogna scrivere regole per controllare il dominio delle variabili, dato che questo controllo viene effettuato e corretto automaticamente dal programma.

Se una regola inizia con un asterisco "*", questa viene considerata un *commento*.

Una regola può contenere un massimo di *16* variabili, e ogni variabile della regola può avere al massimo *100* valori di sottodominio.

Ogni regola deve essere così interpretata:
***Se** è vero V1(sd1 o ..sdn) **ed** è vero V2(sd1 o ..sdn) ...**ed** è vero Vn(sd1 o ..sdn) **allora** l'incompatibilità **è** vera.*

Le variabili all'interno di una regola esprimono *contemporaneità* (AND) mentre i valori dei sottodomini tra parentesi esprimono *alternatività* (OR).

Esempio:
consideriamo che se l'età (descritta dalla variabile ETA) è minore di 14 anni e lo stato civile (descritto dalla variabile STACIV) è libero (risposta 1 al quesito relativo, che ammette le risposte possibili 1-5) la risposta è esatta;
in forma negativa diremo se se l'età è minore di 14 anni e lo stato civile è "diverso da" libero "1" allora l'incompatibilità è verificata;

scriveremo quindi la regola:

ETA(0-13) STACIV<1)

oppure

ETA<14-110) STACIV(2,3,4,5)

Quindi:

- ( ) così si indicano i sottodomini "uguali a";
- <) così si indicano i sottodomini "diversi da";
- da-a indica un intervallo di valori compresi gli estremi;
- la "," separa i valori del sottodominio;

Per *inserire* una regola:

- scegliere una variabile tra quelle mostrate;
- scegliere il simbolo "*eq*" per uguale (o "*ne*" per diverso);
- scegliere il valore tra i domini mostrati ed eventualmente modificarlo con le frecce e, quando il valore è quello voluto, premere "*store*" per memorizzare il dominio nella regola;
- scegliere un nuovo valore tra i domini oppure un'altra variabile;
- così via fino a scrivere tutta la regola.

Alla fine scegliere "*add/mod*";

In caso di errore usare il tasto "undo".

Con "clear" si annulla l'intera operazione.

Per *commentare* una regola:

- Con "*remark*" si commenta una regola scelta o si inserisce un commento;

Per *cancellare* una regola:

- selezionare la regola da cancellare e click su "delete";

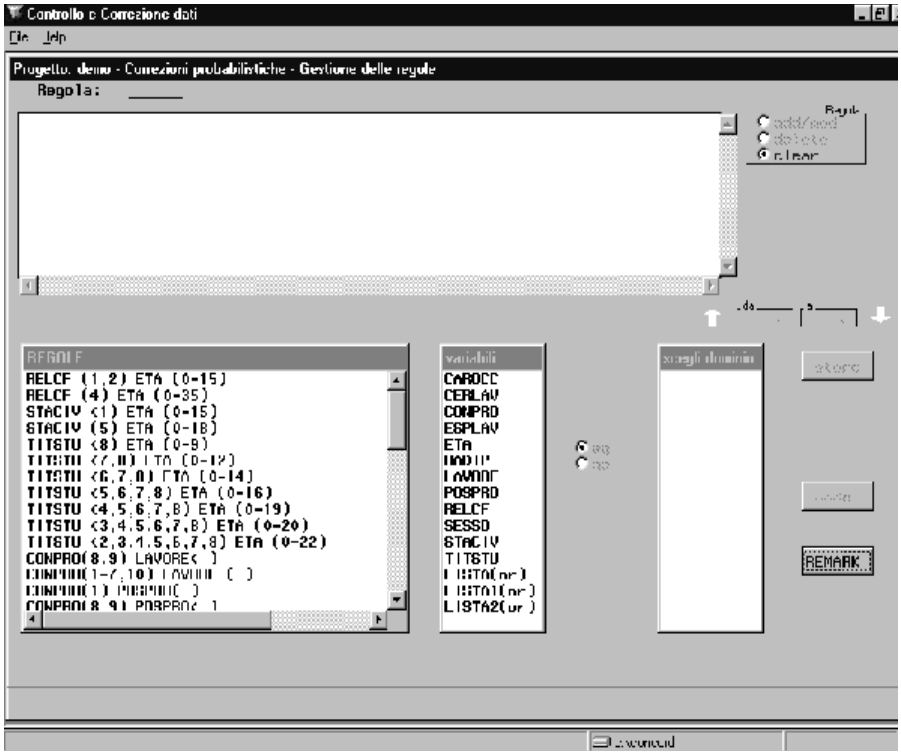
Per *modificare* una regola:

- selezionare la regola da modificare;
- effettuare le modifiche direttamente sulla riga senza aiuto da parte dell'interfaccia; ricordiamo che le regole subiranno un controllo per verificarne l'esattezza con apposita funzione *(cfr. § 4.2.1)*;
- dopo la modifica click su "add/mod".

Da menu a tendina "*file*":

- Con "*exit*" si esce senza memorizzare le operazioni effettuate;
- Con "*save and exit*" si memorizzano le operazioni effettuate e si crea o aggiorna, nella cartella di progetto, il file esterno "REGOLE.dat";
- Con "*import*" si importano regole dal file esterno "REGOLE.dat", dopo aver scelto la cartella con apposita maschera, e si sostituiscono sempre eventuali regole esistenti; in questo caso "*exit*" **non** annulla l'operazione;
- Con "*export per deterministico*", funzione attiva solo dopo aver effettuato il controllo delle regole, si registrano nella cartella di progetto, le regole sul file esterno "fregout.dat", e le variabili sul file esterno "fvardom.dat", nei formati adatti al passaggio deterministico.

Figura 4.4 La maschera per la gestione delle regole



4.2. Le funzioni

4.2.1. Controllo delle regole

Eseguita la fase di definizione delle variabili, delle liste di variabili e delle regole di incompatibilità, si deve scegliere la funzione di "*Controllo delle regole*".

La fase di controllo delle regole scritte dall'utente, ossia dell'insieme minimale, verifica eventuali errori di sintassi, incongruenze e ridondanze delle regole, accorpa regole che hanno le stesse variabili con domini diversi solo per un valore, esplode le liste creando tante regole quante sono le variabili nella lista se definita in "OR", segnala le modifiche effettuate, fermando il processo se esistono regole contraddittorie, cioè regole in contrapposizione tra loro, o regole che per accorpamenti successivi

rimangono con una sola variabile, o regole che per una variabile hanno indicato tutto il dominio.

Figura 4.5 Le funzioni per l'approccio probabilistico

Gli errori di sintassi nelle regole si possono verificare se le regole vengono corrette, o se vengono eliminate o modificate variabili o liste, mentre le incompatibilità , le ridondanze e le incoerenze sono sempre possibili durante la scrittura delle regole stesse, e verificabili solo con un processo iterativo.

Alla fine del passaggio di controllo delle regole viene mostrata la lista delle segnalazioni di errori o avvertimenti, quale ad esempio "*edit ridondante*". Dalla maschera è possibile uscire con "F3" e, se il passaggio è andato a buon fine, sarà possibile eseguire la funzione di "*check dei dati*", la "*derivazione dell'insieme completo*" e, nel modulo di definizione delle regole, la funzione di "*export per deterministico*" *(vedi § 4.1.3)*.

In questa fase vengono registrati tra gli altri i seguenti file esterni:

- "TABVARF.dat" che contiene, per ogni variabile, il dominio definito in classi derivate dall'insieme delle regole che la trattano;
- "MINICE.dat" con la matrice derivante dell'insieme minimo in binario;
- "MINSET.dat" con le regole trasformate in forma leggibile da "MINICE.dat". Notare che le regole sono spesso trasformate in forma affermativa;
- "SERICE.dat" per il trattamento della matrice nei passi successivi;
- "SYSCON.dat" con i messaggi del passaggio di controllo;
- "regole_da_minset.dat" per il riciclo del passo di controllo per eventuali ulteriori accorpamenti.

Le segnalazioni di errore o di avvertimento, riferite al file "MINSET.dat", comprendono nella numerazione anche i commenti.

È importante e opportuno riciclare il passo di controllo ponendo in input il file generato "regole_da_minset.dat", rinominandolo "REGOLE.dat"; questo riciclo va effettuato fino a che non si verifichino più ulteriori accorpamenti delle regole derivate dalla precedente esecuzione della funzione di controllo delle regole.

Dopo il passo di controllo, nel quale vengono eseguiti i programmi "contreg.exe" e "genreg.exe", avendo i dati di input a disposizione è opportuno, prima di eseguire la fase di derivazione degli edit impliciti, eseguire la fase di controllo o "CHECK" dei dati di input o dati grezzi.

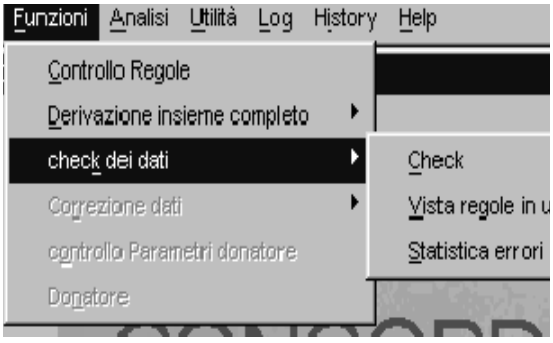
4.2.2. Check dei dati

Selezionando la funzione "check dei dati" sono possibili tre scelte:

1. CHECK

La scelta della funzione di controllo o "*Check*" dei dati di input, o dati grezzi, dovrebbe essere eseguita prima della fase di derivazione dell'insieme completo, per eliminare dall'insieme degli edit espliciti o "*insieme minimale*" scritti dall'utente, eventuali errori di definizione delle variabili, sia nei domini che rispetto al tracciato di registrazione, o errori di scrittura delle regole dal punto di vista logico.

Figura 4.6 La maschera per la funzione di check dei dati



Il check dei dati separa i dati grezzi in esatti ed errati, verificando per i record di input, l'incompatibilità descritta nelle regole dell'insieme minimo tramite il file esterno codificato "MINICE.dat" scritto nella fase di controllo delle regole.

Durante la fase di check, per ogni record in esame se anche una sola regola di incompatibilità è vera il record viene registrato tra gli errati, e viene registrato un record nel file "ERRORI.dat" con il numero della regola errata; in questo caso si può vedere la regola corrispondente al numero tramite la funzione "Vista delle regole in uso" *(vedi punto 2)*.

A fine passaggio di check, segnalato da apposito messaggio, vengono mostrati i contatori dei record letti, esatti ed errati e una tabella *(vedi figura 4.7)* con elencate le regole violate e gli eventuali dati fuori dominio. Tutti i riferimenti numerici delle regole vanno fatti con il file "MINSET.dat" *(vedi punto 2, "Vista delle regole in uso")* ove sono elencate e numerate anche le regole generate automaticamente, nella fase di controllo delle regole, che si riferiscono al "fuori dominio" delle variabili.

L'analisi di questi risultati permette all'utente di verificare gli eventuali errori di definizione e permette, inoltre, di verificare se i dati di input sono affetti da errori sistematici che dovrebbero essere eliminati, con approccio diverso, prima della correzione probabilistica.

Infatti è importante controllare attentamente le regole attivate con troppa frequenza, indice di errore di definizione, o di registrazione, o di risposta ad un quesito affetto da errore sistematico.

Una distribuzione casuale, "*a pioggia*", delle regole violate indica che il piano di check è ben strutturato, e che i dati non sono affetti da errori non stocastici.

Nel passo di check, che esegue il programma esterno "genckf", vengono registrati nella cartella di progetto i seguenti file:

- "SYSCHK.dat" con i messaggi e i valori dei contatori relativi ai record esatti ed errati;
- "LISCHK.dat" con la tabella riassuntiva delle regole verificate; queste regole sono descritte sul file "MINSET.dat", e la numerazione è quella riportata sul file "LISCHK.dat"; se il numero si riferisce a un "(VALORE FUORI DOMINIO)" la variabile indicata nella regola di "MINSET.dat" ha avuto per n volte un valore non ammissibile;
- "ESATTI.dat" con i record esatti;
- "ERRATI.dat" con i record errati;
- "ERRORI.dat" parallelo ai record errati con i numeri delle regole errate per ogni record;

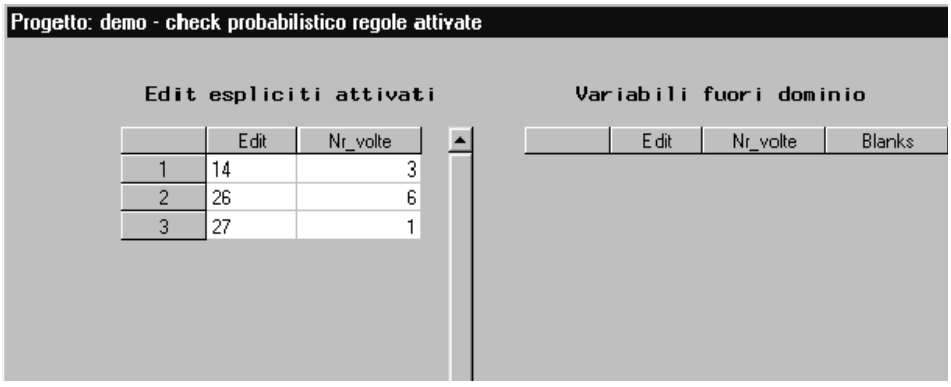
- "FREQUEN.dat" con la frequenza dei casi per ogni dominio di "TABVARF.dat".

Il file esterno di input per il passo di check, deve essere scelto tramite l'apposita maschera evidenziata al momento della scelta della funzione. Una volta scelta la cartella e selezionato il file con "*apri*", o cliccando due volte sul nome del file stesso, viene aggiornato, nella cartella di progetto, il file "tabfile.prm" che contiene il nome dei file che saranno utilizzati nel passo di check e nel successivo passo d'imputazione.

I record del file di input devono essere registrati in ASCII e con lunghezza e campi fissi *(vedi cap.3)*.

Si ricorda che per eseguire il check dei dati non è necessario aver effettuato la funzione di "*derivazione delle regole implicite*".

Figura 4.7 La tabella delle regole attivate nel check dei dati



2. VISTA REGOLE in USO

La scelta di questa funzione mostra le regole, numerate, registrate sul file esterno "MINSET.dat". Dalla tabella si può uscire con il tasto funzione "F3".

3. STATISTICA ERRORI

I contatori e la tabella degli edit espliciti attivati vengono mostrate automaticamente alla fine dell'esecuzione del programma di check dei dati.

È comunque possibile rivederli scegliendo la funzione "*Statistica degli errori*" che mostra, in successione, prima i contatori dei record esatti ed errati e poi la tabella degli edit espliciti attivati con le variabili fuori dominio *(vedi figura 4.7)*

4.2.3. La derivazione degli edit impliciti

La scelta della funzione "*Esegue derivazione*", da "*Derivazione insieme completo*", effettua la derivazione delle eventuali regole implicite dalle regole esplicite scritte dall'utente, e sottoposte al passo di controllo detto "*insieme minimale*", ed è indispensabile per un corretto passo di correzione dei dati errati o imputazione.

Figura 4.8 La funzione di derivazione dell'insieme completo

La derivazione è un programma che per ogni variabile, detta "*generatrice*", verifica se le *n* regole che la contengono, raggruppate a 2,3,...a *n* coprono l'intero dominio della variabile, e in questo caso, se le altre variabili contenute nelle regole generano una nuova regola che viene detta "*implicita*". Se questa regola non risulta compresa nelle altre regole dell'insieme minimo viene detta "*essenzialmente nuova*", diventa parte dell'insieme delle regole, e rientra nel passo che viene riciclato.

Non è possibile quantificare a priori il tempo, talvolta anche di molte ore, necessario alla derivazione dell'insieme completo.

In teoria per ogni variabile sarebbe necessario un numero di combinazioni pari a 2^n-*n-1*, ove *n* è il numero delle regole che contengono la variabi-

le in esame; e questo moltiplicato per tutte le variabili definite dall'utente che ad *n* chiudono il dominio. È stato necessario introdurre dei tagli, alcuni suggeriti dalla stessa metodologia di Fellegi-Holt, altri sviluppati con algoritmi che permettono di ridurre il numero delle combinazioni, quali ad esempio le "classi di equivalenza" che raggruppano in una stessa classe tutte le regole che, per la variabile in esame, hanno lo stesso dominio.

In questa fase vengono registrati nella cartella di progetto i seguenti file:

- "MAXICE.dat" che contiene la matrice dell'insieme completo in forma binaria;
- "GENER.dat" con la storia di ogni regola nella generazione;
- "COMPLETO.dat" con le regole in forma leggibile dall'insieme completo;
- "LISGEN.dat" con la descrizione per ogni regola generata della variabile;
- "SYSDER.dat" con i messaggi del passo di derivazione.

Solo dopo il passo di derivazione, nel quale vengono eseguiti i programmi esterni "derivaz", "decoreg" e "gengen", è possibile effettuare la fase di correzione.

Se il passo di derivazione si ferma per "*edit degenere*" o "*edit contraddittori*", bisogna riferirsi al file "GENER.dat" e al file "COMPLETO.dat" per capire quali regole sono tra loro incompatibili considerando che "GENER.dat" contiene un record per ogni regola con numeri indicanti:

- la variabile generatrice (primo numero nella riga);
- le regole che hanno contribuito alla generazione della regola in esame (numeri successivi);
- -9999 che chiude l'insieme delle regole generatrici.

I numeri di regola negativi (es.: -34) indicano che la regola è stata cancellata nei cicli di derivazione, e quindi bisogna considerare le regole successive al -9999 che hanno derivato la regola in esame e così via. Le regole esplicite hanno il record corrispondente su "*GENER.dat*" con tutti 0.

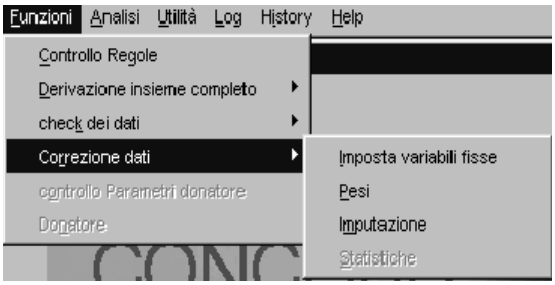
Se il passo di derivazione è andato a buon fine viene resa possibile la funzione "*Correzione dati*".

Scegliendo "*display derivazione*" si ha una lista dei passi effettuati durante la derivazione.

Scegliendo "*insieme completo*" vengono mostrate tutte le regole, numerate, che formano l'insieme delle regole sia esplicite che implicite.

4.2.4. La correzione dei dati

Figura 4.9 La funzione di correzione dei dati



4.2.4.1. Impostazione delle variabili fisse

Questa funzione, propedeutica alla fase d'imputazione, permette di condizionare la correzione di singole variabili con l'impostazione di alcuni parametri.

Infatti è possibile:

- alterare il peso con il quale una variabile può essere scelta per fare parte dell'insieme minimo delle variabili da cambiare;
- assegnare ad una variabile un'altra variabile di confronto;
- definire una variabile come variabile di strato per il cambio del serbatoio dei donatori;
- definire per una variabile l'imputazione da valori prefissati.

Per ogni variabile selezionata è possibile:

- definire una variabile di match, cioè una variabile che dovrà avere lo stesso valore sia nel record errato che nel record donatore, se il programma di correzione decide di correggere la variabile selezionata; per impostare una variabile di match selezionare una variabile tra quelle mostrate e questa verrà scritta nella zona "*Match*";
- cambiare la fissità: con questo parametro è possibile alterare il peso, di norma uguale per tutte le variabili, con il quale esse vengono scelte dall'algoritmo di formazione dell'insieme minimo. Dell'insieme

minimo fanno parte le variabili che devono essere cambiate in modo che il record errato passi tutte le regole di controllo. Impostando per una variabile una fissità da 1 a 8, forziamo il programma di correzione a cercare un insieme minimo di variabili che comprenda la suddetta variabile solo se non è stato trovato un insieme minimo di variabili con fissità minore. La fissità scelta è relativa a tutte le altre fissità, per cui è inutile impostare la fissità di una variabile a 1 e la fissità di un'altra variabile a 3, poiché questa verrà considerata dal programma a 2. La fissità 9 impedisce il cambiamento del valore della variabile e, di conseguenza, questa non può entrare negli insiemi minimi se è da correggere, producendo record "*incorreggibili senza insieme minimo*" nel caso che la variabile fissata a 9 abbia un valore fuori dominio;

- definirla "*chiave*", cioè variabile di stratificazione per il cambio del serbatoio dei record donatori. Se una variabile è definita chiave si presuppone che i record, esatti ed errati, siano ordinati in modo ascendente per detta variabile, e il rinnovo del serbatoio dei record esatti avviene al cambio del valore di una delle suddette variabili sui record errati; il programma domanda all'utente se deve effettuare l'eventuale ordinamento dei dati esatti ed errati, se esistono variabili definite chiave, scelta da evitare se i dati di input del check sono già ordinati. Le variabili chiave devono avere fissità 9 e quindi sono variabili i cui valori devono essere sicuramente esatti;
- definirla "*marginale*" in questo caso una variabile definita marginale viene corretta con l'imputazione sequenziale, seguendo la distribuzione marginale delle frequenze dei dati grezzi, frequenze che possono essere modificate con l'impostazione dei pesi *(vedi sotto)*.

Per ogni variabile impostata è possibile aggiungere note o commenti.

Da menu a tendina "*File*":

- Con "*exit*" si esce senza salvare le modifiche.
- Con "*save and exit*" si esce dalla funzione e si salvano le modifiche anche sul file esterno "VARFIX.dat".
- Con "*import*" si importano variabili da un file esterno formattato del tipo "VARFIX.dat".

4.2.4.2. Impostazione dei pesi

Con questa funzione è possibile variare la distribuzione dei pesi, quale risulta dal passaggio di check, per i domini di variabili opportunamente scelte, in modo che sia applicata, in casi particolari, una distribuzione scelta dall'utente.

Il programma di correzione, nel caso debba imputare una variabile con un valore che debba essere forzato perché non trovato nel serbatoio donatori, oppure se la variabile è stata definita *"marginale" (vedi § 4.2.4.1)*, prenderà i valori da imputare dalla distribuzione delle frequenze ottenute nel passaggio di check e registrate sul file "FREQUEN.dat".

Con la funzione *"Pesi"* è possibile modificare questa distribuzione all'interno dei valori del dominio di una variabile. Il programma mostra l'intera distribuzione delle frequenze nella *"Lista totale dei pesi in %"*.

Per creare una variabile peso:

- Scegliere una variabile tra quelle mostrate; viene mostrato l'intero dominio della variabile con i pesi relativi in percentuale;
- Scegliere il dominio da modificare;
- Scrivere il nuovo peso in percentuale e dare invio;
- Modificare un altro dominio in modo che la somma dei pesi sia sempre 100;
- Alla fine *"add/mod"*.

Per cancellare una variabile peso:

- sceglierla e poi *"delete"*.

Da menu a tendina *"File"*:

- Con *"exit"* si esce senza salvare le modifiche;
- Con *"save and exit"* si esce e si salvano le modifiche sul file esterno "PESI.dat";
- Con *"import"* si importano pesi da un file esterno "PESI.dat" selezionato dall'utente.

4.2.4.3. Impostazione dei parametri di esecuzione dell'imputazione

Con questa funzione si impostano i parametri che condizionano l'intera esecuzione del passo di imputazione, contrariamente alle impostazioni del paragrafo 4.2.4.1, che riguardano le singole variabili.

I parametri sono:

- Tipo di imputazione:
  impostato all'esecuzione di tutti e tre i tipi di imputazione; con questo parametro si può ridurre la sequenza dei tentativi di imputazione alle sole imputazioni allargata e sequenziale, oppure ancora più restrittivamente, alla sola imputazione sequenziale. Cambiare il tipo di imputazione può servire a diminuire il tempo di esecuzione del passaggio di correzione in fase di messa a punto del progetto.
- Statistiche:
  impostato a SI, scelta consigliata. Con NO si elimina la produzione delle statistiche di correzione.
- Numero record serbatoio donatori:
  impostato a 2.000; può essere diminuito; notare che il serbatoio dei donatori viene rinnovato al cambio del valore delle eventuali variabili *chiave*.
- Numero massimo di donazioni:
  impostato a 99999 (numero di donazioni senza limite). Si può diminuire e serve a limitare le donazioni che un singolo record esatto può effettuare.

Con 'Save and exit' si registrano i parametri, nella cartella di progetto nel file "PARM.dat", e si esegue il programma di correzione.

Se sono state definite variabili fisse di tipo "*chiave*" il programma, prima dell'esecuzione del passo di imputazione, domanda all'utente se deve ordinare il file degli ESATTI ed il file degli ERRATI, ordinamento che può essere evitato se i dati sono già ordinati per le suddette variabili.

4.2.4.4. IMPUTAZIONE

La fase di correzione dei dati, o imputazione, trasforma i record errati, output della fase di controllo o check, in record corretti, utilizzando l'algoritmo del minimo cambiamento, una delle basi della metodologia di Fellegi-Holt.

Per ogni record errato, tramite l'insieme completo "MAXICE.dat", generato nella fase di derivazione degli edit impliciti, vengono verificate le regole di incompatibilità, e si cerca il numero minimo di variabili che, modificate con i valori presi da un serbatoio di donatori e tentando di prendere sempre il record esatto più somigliante, rendono corretto il record errato in esame.

Il serbatoio dei record donatori è costituito dal numero di record esatti, output del passaggio di check , determinato da un parametro di imputazione, per un massimo di 2000 record. Questo serbatoio, che viene rinnovato solo se esistono variabili definite *"chiave"*, è l'unica fonte che fornisce i valori per la correzione delle variabili che fanno parte dell'insieme minimo di variabili da correggere.

Per prendere il record esatto più somigliante, il programma cerca la distanza minima tra il record da correggere e i record del serbatoio degli esatti, che viene rinnovato a rottura di strato determinato dalle variabili definite *chiave*, come segue: viene tentata prima l'imputazione ristretta, viene cioè cercato tra gli esatti un record con i valori delle variabili da non modificare uguali a quelle del record da correggere; se non è possibile trovare detto record nel serbatoio degli esatti, si tenta l'imputazione allargata, cioè trovare un record tra gli esatti con i valori "possibili" nelle variabili da non modificare, e infine, se anche questo tentativo d'imputazione non è riuscito, il record viene imputato sequenzialmente, una variabile da correggere alla volta, ricercando tra gli esatti un record compatibile, e se non trovato, forzando random tra i valori possibili, il valore esatto.

Il tempo necessario alla correzione dei record dipende dalla grandezza dell'insieme completo, dal tipo di imputazione prescelto, e ovviamente dal numero dei record errati.

I file utilizzati dal programma "genimpn2" sono:

- "PARM.dat" con i parametri di imputazione prescelti;
- "TABVARF.dat" con i domini separati in classi per ogni variabile;
- "VARFIX.dat" con le variabili definite fisse;
- "MAXICE.dat" insieme completo in forma binaria;
- "SYSCHK.dat" dal check con contatori esatti/errati;
- "ESATTI.dat" dal check record esatti;
- "ERRATI.dat" dal check record errati;
- "CORRETTI.dat" record errati corretti;
- "INCORRETTI.dat" eventuali record non correggibili perché fissati in modo errato;
- "SYSIMP.dat" con i seguenti contatori:
  a) *"Imput. congiunta ristretta"*, numero dei record corretti con l'imputazione ristretta;

b) *"Imput. congiunta allargata"*, numero dei record corretti con l'imputazione allargata;
c) *"Imput. Sequenziale"*, numero dei record corretti con l'imputazione sequenziale;
d) *"Forzature"*, totale delle forzature effettuate nell'imputazione sequenziale;
e) *"Incorretti"*, eventuali record incorretti - derivazione errata, MAXICE non valido;
f) *"Incorretti senza ins.min."*, eventuali record incorretti derivati da aver fissato a 9 variabili con valori errati;
g) *"Corretti da incorretti"*, record corretti con insiemi minimi alternativi risultanti ad aver fissato a 9 molte variabili.

- "STATIS.dat" statistiche di correzione;
- "FREQUEN.dat" output del check con frequenze dei dati grezzi per le variabili da correggere con forzature;
- "PESI.dat" eventuale diversa distribuzione delle frequenze per alcuni domini per le variabili da correggere con forzature.

4.3. Analisi dei risultati

Eseguita la fase di imputazione, CONCORD permette un'analisi dei risultati tramite funzioni specifiche.

4.3.1. Tavole di verifica

Viene proposta una maschera *(vedi figura 4.11)* nella quale scegliere, con doppio click del mouse, fino a tre variabili, tra quelle mostrate, che identifichino *in modo univoco* i record del file dati prima e dopo la correzione.

Figura 4.10 Le funzioni di analisi dei risultati

Nel caso non esistano, tra quelle mostrate, variabili che identifichino univocamente i record del file dati, ma siano registrati sui dati grezzi valori adatti allo scopo, è possibile definire, fino ad un massimo di tre, nuove

variabili con posizione e lunghezza. In questo caso il sistema nominerà automaticamente le variabili con key1, key2 e key3.

Queste variabili servono ad accoppiare il file dei dati grezzi, input della fase di check, con un file ottenuto dai dati corretti e dai dati esatti in modo da poter evidenziare le variazioni subite. Per questo motivo, i record devono essere identificabili in tutti gli archivi.

Se non esistono valori, e quindi variabili, nel file dei dati grezzi, che identifichino in modo univoco i record da sottoporre a check e correzione, è possibile utilizzare un programma di utilità, *(vedi § 7.1.2)* per creare una variabile di numerazione adatta allo scopo e numerare i record del file di input prima del passaggio di check.

Esempio:sia dato un file con i seguenti record:

1___5____0____5_

record 1 0100100111222345

record 2 0100200304222354

record 3 0100300302232322

record 4 0100400102043421

con definite variabili a posizione 6 lunga 3, posizione 9 lunga 1, 10 lunga 2 e 15 lunga 2; non esistendo variabili che identifichino in modo univoco il record possiamo però definirne una con posizione 3 e lunghezza 3.

Se i dati invece fossero:

1___5____0____5_

record 1 0100100111222345

record 2 0100100304222354

record 3 0100300302232322

record 4 0100300102043421

non avremmo nessuna zona dei record che li possa identificare univocamente; dobbiamo perciò, prima dei passaggi di check e correzione, numerare con l'apposita funzione nei programmi di utilità (vedi 7.1.2), i record a posizione 17 per 1 carattere e, una volta effettuati i passaggi di controllo e correzione, definire una variabile a posizione 17 lunga 1 che serva all'accoppiamento dei dati per le tavole di verifica.

Le tavole di verifica sono ottenute con procedure SAS quali SORT, MERGE e FREQ e in output si avranno la distribuzione delle frequenze per tutte le variabili definite, e la distribuzione incrociata di tutte le variabili definite tra il file dei dati grezzi, prima delle fasi di controllo e correzione, e il file dei record esatti più quello dei record corretti.

Dal menu a tendina *"File"*:

"Save and run": esegue l'elaborazione attraverso una macro SAS che sviluppa gli incroci per tutte le variabili ed esce.

"Exit": esce senza elaborare.

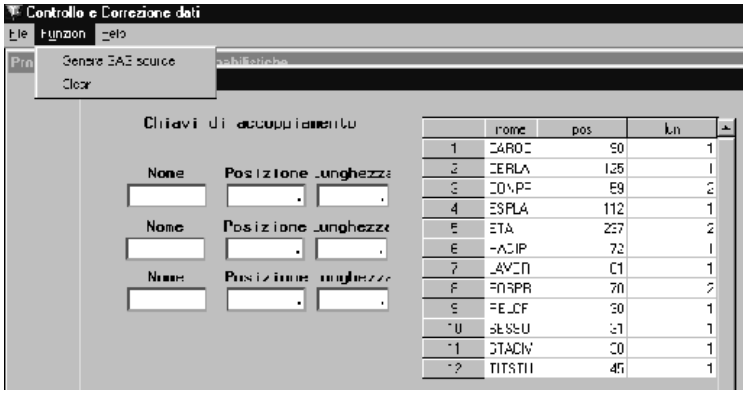
"Run source file": esegue l'elaborazione dei programmi generati dall'apposita funzione, sotto descritta, ed esce. Questa scelta è possibile solo se è stato generato il sorgente SAS.

Dal menu *"Funzioni"*:

"Genera SAS source": dopo aver indicato le variabili è possibile generare, nella cartella di progetto, i sorgenti dei programmi SAS necessari all'elaborazione. Questa opzione permette all'utente di modificare, eventualmente, i sorgenti SAS per ottenere una elaborazione personalizzata, eliminando gli incroci per le variabili non interessanti.

"Clear": annulla tutte le variabili inserite ripristinando i campi.

Figura 4.11 La maschera per le funzioni di verifica



4.3.2. Grafici

Con questa funzione è possibile ottenere, a video, dei grafici.

Sono possibili le seguenti scelte:

- Distribuzioni dei dati:
  viene proposta una maschera nella quale bisogna selezionare lavariabile e il tipo di istogramma desiderato.
  Viene mostrata la distribuzione di frequenza, per la variabile selezionata, del valori dei dati grezzi e dei dati dopo la correzione.
- Frequenza errori
  Viene mostrata la distribuzione di frequenza delle regole di incompatibilità verificate.

4.4. Errori di esecuzione

I seguenti errori bloccano i programmi durante l'esecuzione.

PROGRAMMA - *errore* -	Causa
CONTREG (controllo delle regole)	
Matrice troppo larga:	Superati i limiti di memoria
GENREG (controllo delle regole)	
L'ampiezza dei domini > 1500 caratteri:	Superati i limiti di memoria.
Manca il file delle regole:	Non esiste il file"REGOLE.dat".
VARIABILI DA GESTIRE > 500: Superato il limite delle variabili.	
MATRICE TROPPO GRANDE:	Più di 2.000 edit espliciti.
Variabile xx nella regola yy errata:	La variabile non è definita.
Regola con una sola variabile :	Una regola deve avere al minimo 2 variabili.
edit con variabili uguali (lista errata?):	Nella lista esiste una variabile presente anche nella regola.
Valore vv della variabile xx nella regola yy errata:	Valore fuori dominio.
Edit contraddittori:	Regole in contraddizione tra loro.

DERIVAZ (derivazione edit impliciti)

TROPPI EDIT DELLA STESSA CLASSE EQUIVALENZA:	Il numero di edit ha superato i limiti previsti
TROPPE CLASSI DI EQUIVALENZA:	Il numero delle classi ha superato i limiti previsti.
TROPPI EDIT GENERATI:	Il numero di edit ha superato i limiti previsti.
EDIT DEGENERE, *edit generati da edit contraddittori,* *edit contraddittori* *edit xx degenere per unione uguali a meno di uno:*	*L'edit, indicato con xx, che risulta dalla generazione ha una sola variabile. Nella matrice iniziale - minimal set - ci sono edit generati da regole contraddittorie; è possibile risalire agli edit espliciti la cui unione ha formato un edit degenere (con una sola variabile) tramite i file "GENER.dat" e "COMPLETO.dat" ove è indicata sia la variabile generatrice che gli edit uniti tra loro (gli edit tra parentesi su "COMPLETO.dat" o con -9999 su "GENER.dat" sono stati eliminati, perché ridondanti, da nuovi edit) (vedi § 4.2.3.)*

GENCKF (check dei dati grezzi)

MANCA IL FILE DELLE VARIABILI:	Non esiste il file "VARDOM.dat".
MANCA IL FILE DELLE REGOLE:	Non esiste il file "REGOLE.dat".
NON ESISTE IL FILE DATI INPUT:	Non esiste il file "DATI.dat".
RECORD 5.000 CARATTERI	Il record in input > limite previsto.

GENIMPN2 (imputazione)

solo forzature per strato xx senza esatti:	Per lo strato xx non esistono esatti nel serbatoio. Diminuire le variabili strato.
MANCA FILE DELLE VARIABILI:	Non esiste il file "VARDOM.dat".
Troppi domini. Matrice insufficiente:	Superati i limiti di programma.
TROPPI EDIT. Matrice insufficiente:	Superati i limiti di programma.
MANCA INSIEME COMPLETO *-fatta la derivazione?*	Non esiste il file MAXICE.
TROPPI EDIT. Matrice insufficiente:	Superati i limiti di memoria del programma.

TROPPI CAMPI CHIAVE:	Sono state definite più di 5 variabili chiave.
RECORD > 5.000 CARATTERI:	Il record input troppo lungo.
NON ESISTE IL FILE ERRATI:	Non è stata effettuata la fase di check o i dati sono tutti esatti.
NON ESISTE IL FILE ESATTI:	Tutti i dati dalla fase di check sono errati.

4.5. L'integrazione verso l'approccio deterministico

Le funzioni che permettono l'integrazione automatica, dall'approccio di correzione probabilistica all'approccio di correzione deterministica, si trovano nella definizione delle variabili e nella definizione delle regole. In ognuna di queste due fasi di definizione è disponibile dal menu a tendina una funzione di "export verso deterministico" che permette di generare automaticamente i file ASCII, formattati come necessario, per i programmi eseguiti nell'approccio deterministico.

Nel capitolo 9 - ESEMPIO di APPLICAZIONE - sono elencati ed eseguiti tutti passi necessari alla completa integrazione tra i tre metodi di correzione.

DEFINIZIONE delle VARIABILI *(vedi § 4.1.1):*

Usando la funzione di "export" nella definizione delle variabili si otterrà un file formattato con le stesse variabili pronte per essere utilizzate dal modulo deterministico e, particolarmente, con il tipo di variabile impostato ad alfabetico "A".

DEFINIZIONE delle REGOLE *(vedi § 4.1.3)*:

Usando la funzione di "export" da *"File"* nella definizione delle regole *(vedi figura 4.12)*, funzione attiva solo dopo avere effettuato il controllo delle regole, si otterranno invece:

- il file *"fvardom.dat"* formattato con le stesse variabili pronte per essere utilizzate dal modulo deterministico e, particolarmente, con il tipo di variabile impostato ad alfabetico "A";
- il file *"fliste.dat"* formattato delle liste di valori, delle variabili probabilistihe che superano 5 definizioni da-a, con le variabili lista numerate L00001, L00002 ecc. e i valori espressi in forma di lista dei domini;

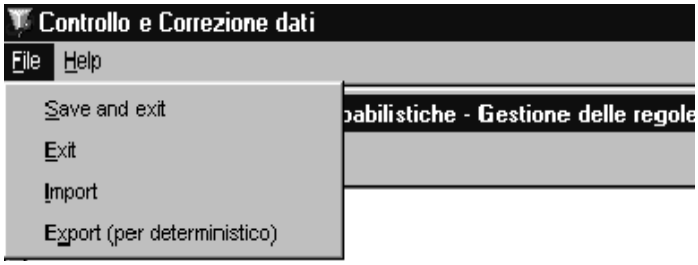
- il file *"fregout.dat"* formattato per il deterministico, sia delle regole di incompatibilità probabilistiche trasformate in testo per l'approccio deterministico, sia di regole, appositamente aggiunte, che servono a controllare se i valori delle variabili sono fuori dominio; ricordo che queste ultime regole non esistono nell'approccio probabilistico.

Per sfruttare l'integrazione automatica, una volta utilizzata la suddetta funzione dalla definizione delle regole nell'approccio probabilistico, basterà, quindi:

- chiudere il progetto probabilistico;
- riaprirlo scegliendo l'approccio deterministico.

Poi, nelle funzioni di definizione delle variabili, liste e regole di controllo si dovranno importare i relativi file dalla cartella di progetto e salvarli con la scelta di *"Save and exit"*.

Figura 4.12 La scelta di export per l'integrazione verso deterministico



5. L'approccio di correzione deterministica

Sintesi: *sono brevemente descritte la metodologia e le varie fasi, da eseguire in modo gerarchico dall'utente, per la correzione da effettuare secondo l'approccio deterministico.*

Le regole di controllo, o edit, in un piano di compatibilità possono essere distinte in:

- regole formali, che derivano dalla struttura del modello, cioè direttamente dalle norme di compilazione e dai "percorsi interni", o "salti", del modello;
- regole sostanziali, che derivano da considerazioni di tipo statistico o matematico, o da conoscenze specifiche a priori del fenomeno oggetto di rilevazione.

È chiaro che la natura degli edit, sia formali che sostanziali, di un piano di compatibilità è strettamente dipendente dal tipo di variabili, qualitative o quantitative, oggetto di verifica. Mentre nel caso di variabili qualitative, infatti, tali edit hanno la forma di relazioni logiche tra le variabili, nel caso di variabili quantitative le regole di incompatibilità possono essere anche espresse in forma di relazioni *statistico/matematiche.*

Una volta individuati i record i cui valori violano uno o più vincoli del piano di compatibilità, il problema diventa la localizzazione delle variabili i cui valori devono essere considerati errati e, in quanto tali, da sottoporre ad un passo di correzione.

Sia il problema della localizzazione dei record errati, sia quello dell'individuazione delle variabili che, per ogni record errato, sono da considerarsi

responsabili della violazione di una o più regole di compatibilità, possono essere risolti adottando un approccio di tipo *deterministico*.

Per ogni record, o per gruppi di record, vengono cioè applicate regole che, se verificate, segnalano sicuramente la presenza di errori.

Ad esempio:

SE (sesso = maschio E professione = casalinga) ALLORA sussiste incompatibilità x.

Una regola di questo tipo non individua, di per sé, l'errore che ne causa l'attivazione: infatti, il *valore non vero* può celarsi in una o nell'altra delle variabili, o in entrambe.

Nell'approccio deterministico, una situazione di incompatibilità è seguita, contestualmente, dall'indicazione delle variabili che debbono considerarsi errate, e, in quanto tali, da imputare. Nell'esempio considerato avremo, per ipotesi:

SE (sesso = maschio E professione = casalinga) ALLORA sesso ← femmina

il che significa che, se in un record è attivata la condizione di incompatibilità *"maschio e casalinga"*, nella regola stessa viene indicata l'azione da effettuare per correggere l'errore, che consiste nell'imputare la modalità *"femmina"* alla variabile sesso.

Generalizzando, una volta attivate, mediante le regole di compatibilità, una o più condizioni di errore in un dato record, sono determinate a priori le azioni da intraprendere per riportare il medesimo record in una situazione di correttezza.

Le procedure deterministiche sono generalmente costituite da regole di imputazione deterministica, dette R.I.D., del tipo:

SE (incompatibilità) ALLORA (localizzazione e correzione errore)

La condizione di incompatibilità esprime relazioni inammissibili intercorrenti tra due o più variabili; la localizzazione consiste nell'indicazione di quali variabili considerare errate e quali valori assegnare per correggerle. Un record, durante l'esecuzione della procedura di correzione, potrà causare l'attivazione delle regole in corrispondenza delle quali è verificata la

parte SE: in tal caso, eventualmente, saranno modificate le variabili indicate nella parte ALLORA assegnando loro valori predefiniti o scelti in altro modo. In CONCORD le regole di imputazione deterministica possono essere del tipo:

SE (incompatibilità) [ALLORA (localizzazione e correzione errore)].

Può essere, cioè, definita la sola incompatibilità applicando, eventualmente, la localizzazione e correzione dell'errore.

5.1. La fase di definizione

Figura 5.1 - Le funzioni di definizione

5.1.1. Definizione delle variabili

Vengono trattate sia variabili qualitative che quantitative registrate in un file in ASCII con caratteri di tipo testo su record di lunghezza fissa. Per definire le variabili è necessario avere il piano di registrazione, del questionario in esame, e il tracciato record nel quale siano indicate le posizioni iniziali e le lunghezze dei vari campi.

È possibile definire le sole variabili che vogliamo sottoporre a controllo per un massimo di 500; il resto del record verrà ricopiato senza modifiche.

Per ogni variabile è necessario indicare:

- nome: univoco e lungo al massimo 6 caratteri (esempio ETA,Q12,COL1_6,...); il nome deve essere univoco anche rispetto ai nomi di eventuali liste *(vedi § 5.1.2)*;
- tipo: "N" per variabile numerica, "A" per variabile alfanumerica;
- posizione: cioè la colonna del tracciato record dove inizia la variabile, in un valore compreso tra 1 e 9999;

- lunghezza della variabile: da 1 a 18 caratteri per variabile numerica; da 1 a 20 caratteri per variabile alfabetica;
- numero decimali: da 0 a 4 se la variabile è numerica (0 per default).

Esempi di variabili:

ETA	*Alfabetica, posizione 1,*	*lunghezza 1;*
SESSO	*Alfabetica, posizione 5,*	*lunghezza 1;*
REDDIT	*Numerica, posizione 345,*	*lunghezza 7, decimali 0;*

Le posizioni e lunghezze si possono *sovrapporre* permettendo la ridefinizione. I dati sui record riferiti ad una variabile numerica devono essere allineati a destra e completati con zeri iniziali; una variabile definita numerica *"N"* permette algoritmi di tipo aritmetico nelle regole di controllo e di correzione.

Nel riquadro variabile:
"add/mod" salva le modifiche di una variabile;
"clear" annulla tutte le modifiche di una variabile.

Da menu a tendina *"File"*:
"View" mostra tutte le variabili ordinate per posizione;
"Exit" esce senza salvare le modifiche;
"Save and exit" salva le modifiche sul file esterno formattato *"fvardom.dat"*;
"Import" legge le variabili da un file esterno formattato *"fvardom.dat"*.

5.1.2. Definizione delle variabili di lista

Le variabili lista, create per agevolare la scrittura delle regole, sono variabili associate a liste di valori, oppure associate ad un file esterno, ordinato in senso ascendente per la zona del record interessata ai valori da prendere in considerazione.

Definendo una variabile tipo lista possiamo poi, nelle regole, confrontarla con qualsiasi variabile di tipo alfanumerico *"A"*, e il confronto verrà automaticamente effettuato con tutti i valori della lista.

La variabile lista deve essere inserita con il:

- NOME, univoco e lungo al massimo 6 caratteri univoco anche rispetto ai nomi delle variabili (esempio ETA, Q12, COL1_6, ecc...).

Se la variabile è associata ad un file che contiene i valori allora:

- nome del file compreso il percorso;
- posizione, posizione di partenza sul record della zona contenente il valore;
- lunghezza, lunghezza del campo;

e fare attenzione che il file sia ordinato in senso ascendente per queste posizioni, altrimenti si avrà un messaggio di errore in esecuzione della fase di controllo.

Esempio di variabile lista associata ad un file esterno:
Nome lista: LISCOM; File di riferimento: "C:\fdl\comuni.dat"; Posizione: 3; Lunghezza: 6

la variabile lista si chiama "LISCOM" e i valori di lista si trovano registrati nel file "C:\fdl\comuni.dat" nelle posizioni 3-8 di ogni record.

Nelle regole potremo scrivere:
[se] COMUNE = LISCOM e in esecuzione la variabile COMUNE sarà confrontata con tutti i record del file "C:\comuni\comuni.dat" da posizione 3 a 8;

Se la variabile lista non è associata ad un file che contiene i valori allora:
inserimento della variabile lista:

- scrivere il nome della lista;
- inserire i valori associati scrivendo nel relativo campo della zona "VALORE" il valore e poi invio o "add/mod";
- "add/mod" nella zona "LISTA";

variazione del nome della lista:

- scegliere la variabile lista;
- scrivere il nuovo nome;
- "add/mod" nella zona "LISTA";

in cancellazione della lista:

- scegliere la lista;
- "delete" nella zona "LISTA".
- "clear" nella zona "LISTA" annulla tutte le variazioni.

Attenzione! Modificando o cancellando il nome di una lista si perdono tutti i valori associati; inoltre bisogna considerare anche le regole in cui essa

figurava per evitare errori o inesattezze quali valori inesistenti, lista inesistente, valori diversi da ecc.. Quindi è necessario ricontrollare le regole.

Valori di una variabile lista:
inserimento dei valori:

- scrivere nel campo il valore per un massimo di 20 caratteri;
- invio oppure "add/mod" nella zona "VALORE";

variazione dei valori:

- scegliere il valore da modificare;
- scrivere il nuovo valore per un massimo di 20 caratteri;
- "add/mod" nella zona "VALORE";

cancellazione dei valori:

- scegliere la variabile lista;
- scegliere il valore da modificare;
- "delete" nella zona "VALORE".

"clear" annulla tutte le operazioni effettuate.

Alla fine delle operazioni *"add/mod"* nella zona *"LISTA"* e il nome della lista viene mostrato nell'apposita listbox;

Da menu a tendina *"File"*:
"Save and exit" salva, nella cartella di progetto, i nomi delle liste sul file *"fvardom.dat"* e salva gli eventuali valori associati nel file *"fliste.dat"*;
"Exit" esce senza salvare le modifiche;
"*Import*" importa le variabile da un file esterno *"fvardom.dat"* e *"fliste.dat"* scelti con apposita maschera.

5.1.3. Definizione delle regole di incompatibilità

Le regole di controllo definite per l'approccio deterministico, similmente a quelle dell'approccio probabilistico, sono regole di *incompatibilità*; bisogna, quindi, descrivere una condizione che, se verificata, determina un errore.

Una regola di incompatibilità è formata da un'espressione di variabili, costanti e operatori, unite da *"and"* o *"or"* e racchiuse, se necessario, tra vari ordini di parentesi "( )".

La scrittura di una regola di incompatibilità segue, praticamente, la sintassi di uno statement *"IF"*, non nidificato, del SAS base con *omessa* la *"IF"* iniziale.

Le regole di incompatibilità che iniziano con un "#" sono considerate commenti.

Gli operatori che, in una regola, mettono in relazione tra loro variabili e costanti sono di tipo:

a) operatore *aritmetico*:
   1) ****** indica l'elevazione a potenza;
   2) indica prodotto;
   3) **/** indica rapporto;
   4) **+** indica somma;
   5) indica sottrazione.

b) operatore *logico*:
   1) oppure **gt** per indicare "maggiore di";
   2) **<** oppure **lt** per indicare "minore di";
   3) **=** oppure **eq** per indicare "uguale a";
   4) **ne** per indicare "diverso da";
   5) **ge** per indicare "non minore di";
   6) **le** per indicare "non maggiore di".

c) operatore *booleano*:
   1) **&** oppure **and** per indicare contemporaneità;
   2) **|** oppure **or** per indicare alternatività.

Gli operatori, nello sviluppo di una regola, vengono considerati nell'ordine di priorità di descrizione, quindi verranno considerati prima gli operatori aritmetici e, nel loro ambito, prima la potenza, poi il prodotto ecc., poi quelli logici e, nel loro ambito, prima il ">" poi "<" e così via e, per ultimi, gli operatori booleani.

Si possono usare indifferentemente i simboli doppi, quali ">" o "gt", "&" o "and", ecc.

Esempi di regole:
1) # ------- esempio di regola ---------- (commento)
2) sesso = 2 and profes = list1
(se sesso uguale a 2 e professione uguale a un valore contenuto nella lista list1 allora il record è errato)

3) ateco ne listac and ((totdip > 1200 & salar < 1200000) or (totdip < 1201 & salar < 1100000))
(se ateco diverso dai valori della lista listac e, il totale dei dipendenti è maggiore di 1200 e il salario minore di 1200000, oppure il totale dei dipendenti è minore di 1201 e il salario minore di 1100000 allora il record è errato).

Per inserire una regola:

- è possibile scrivere direttamente nell'area apposita dando poi invio;

oppure:

- scegliere la variabile tra quelle mostrate o il simbolo "(";
- scegliere l'operatore necessario tra quelli mostrati;
- scegliere una seconda variabile o la parentesi o una costante (fare attenzione che se la costante si riferisce ad una variabile di tipo "A" - alfanumerica- deve essere inclusa tra singoli apici, ad esempio 'costantè) e così via;
- alla fine scegliere *"add"*;

La regola viene subito controllata sintatticamente e, se non ci sono errori, aggiunta alla fine dell'insieme di regole, altrimenti viene segnalato in chiaro l'errore.

Attenzione! Per inserire una nuova regola *dopo una determinata regola* selezionare quest'ultima con un *solo* click del mouse;

Per modificare una regola:

- scegliere la regola con *doppio click* del mouse;
- modificare la regola evidenziando con il mouse la zona da variare;
- alla fine *"modify"*; la regola viene controllata e, se non ci sono errori, sostituita.

Attenzione! La modifica potrebbe influenzare eventuali regole di correzione associate.

Per cancellare una regola:

- selezionare la regola da cancellare e poi *"delete"*.

Attenzione! Verranno cancellate anche le eventuali regole di correzione associate.

"clear" annulla le modifiche della regola.

Da menu a tendina *"File"*:
"Save and exit" registra, nella cartella di progetto, le operazioni effettuate sul file esterno *"fregout.dat"* che conterrà le regole in chiaro, e sul file esterno *"fregole.dat"* che conterrà invece le regole codificate sotto forma di matrice, interpretabile dai moduli del programma di controllo deterministico; inoltre registra il file esterno *"FSCREEN.dat"* che può essere utilizzato per il *controllo interattivo dei dati* con la funzione gestione dati *(vedi § 7.1.1);*
"Exit" esce senza salvare le modifiche;
"Import" importa le regole da un file esterno *"fregout.dat"* scelto con apposita maschera.

5.1.4. Definizione delle regole di correzione

Nell'approccio deterministico è fondamentale poter far seguire alla condizione *"IF"* anche il relativo *"THEN"*, con il quale assegnare valori ad una o più variabili indicate dall'utente.

Quindi una regola di correzione è *l'assegnazione di un'espressione aritmetica o di una costante o di una variabile ad una data variabile, al verificarsi di un errore determinato da una regola di incompatibilità.*

Scegliendo questa funzione appare una maschera che elenca, nel listbox *"REGOLE" (vedi figura 5.2)*, le regole di incompatibilità definite nella funzione precedente:

Per inserire una regola di correzione:

- selezionare nella zona suddetta, con un click del mouse, la regola di incompatibilità per la quale effettuare la correzione dalla lista delle regole;
- selezionare la variabile da modificare tra quelle mostrate nel listbox "Variabili"; il simbolo di assegnazione "=" verrà inserito automaticamente dal programma;
- scrivere la regola utilizzando variabili ed operatori.

Esempi di regole di correzione:

1. *eta = 1998 - datana (che significa l'età sarà uguale a 1998 - data di nascita);*
2. *ateco = '10101' (che significa mettere il valore 10101 nella variabile ateco).*

La regola viene controllata sintatticamente al momento dell'inserimento e la numerazione associata è automatica.

È possibile inserire *più correzioni* per una stessa incompatibilità, in questo modo si possono cambiare i valori di *più variabili* in funzione di un singolo errore.

Una variabile corretta da una regola di correzione *rientra* nei controlli successivi nel passo di check.

Per modificare una regola di correzione:
- selezionare la regola di correzione da modifica;
- modificare la regola;
- premere *"add/mod"*.

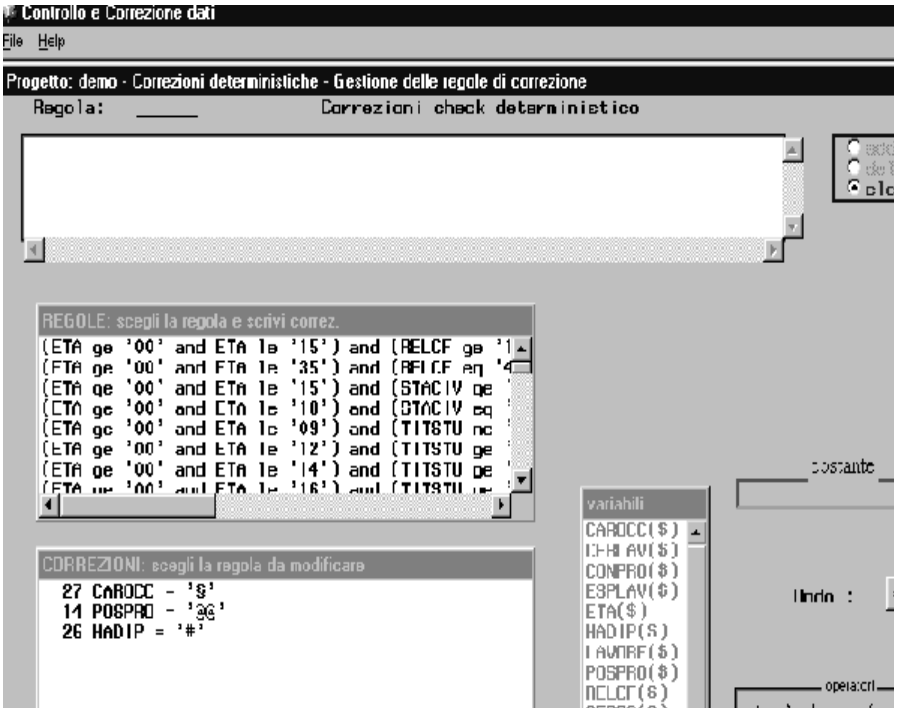
Per cancellare una regola di correzione:
- selezionare la regola da cancellare;
- premere *"delete"*.

"*clear*" annulla le modifiche della regola.

Da menu a tendina *"File"*:

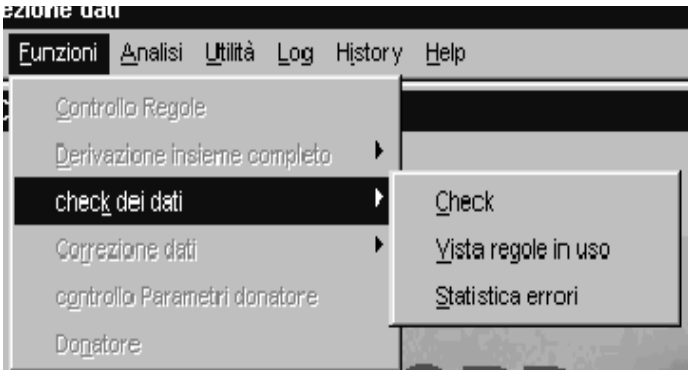
"Save and exit" registra le operazioni effettuate sul file esterno *"fdetout.dat"*, che conterrà le regole in chiaro, e sul file esterno *"fregdet.dat"*, che conterrà invece le regole codificate sotto forma di matrice interpretabile dai moduli del programma da correzione deterministico; i file verranno registrati nella cartella di progetto;
"Exit" esce senza salvare le modifiche;
"Import" importa le variabili dal file esterno "fdetout.dat" scelto con apposita maschera.

Figura 5.2 - La maschera per la gestione delle regole di correzione



5.2. Le funzioni

Figura 5.3 La scelta della funzione per il check dei dati



5.2.1. Check e correzione dei dati

Sono possibile tre scelte:

1) CHECK

Il check e l'eventuale correzione dei dati, eseguibile dopo la fase di definizione delle variabili, liste, regole di incompatibilità e regole di correzioni, separa i dati grezzi, scelti con apposita maschera dall'utente, in dati esatti e dati errati, verificando per i record di input, l'incompatibilità descritta nelle regole relative.
Scegliendo la funzione *"Check"*, il programma chiede di selezionare il file esterno di input per il passo di check.
Una volta selezionato con *"apri"*, o cliccando due volte sul nome del file scelto, viene generato nella cartella di progetto il file *"tabgran.prm"* che contiene il nome dei file che saranno utilizzati nel passo di check, e viene eseguito il programma di controllo e contemporanea correzione.
I record del file di input devono essere di lunghezza fissa.
Nel passaggio di check se anche una sola incompatibilità è vera, e non esistono correzioni per quella incompatibilità, il record viene registrato nel file degli errati *"ferrati.dat"* e il numero della regola, che ha determinato l'errore, viene registrato nel file *"ferrori.dat"*.
Se invece esistono regole di correzioni per l'incompatibilità verificatasi, i record verranno registrati nel file *"fcorretti.dat"*.

In questo passo, che esegue il programma *"granada.exe"*, vengono usati i seguenti file esterni i cui nomi sono registrati nel file *"tabgran.prm"*:

- "fvardom.dat"-variabili;
- "fliste.dat" - liste di valori;
- "fregole.dat" - regole di incompatibilità in codice;
- "flischk.dat" - lista delle regole di incompatibilità verificate;
- "fregerr.dat" - parallelo a "ferrati.dat" contenente per ogni record errato i numeri delle regole che hanno determinato l'errore;
- "fsyschk.dat" - con i contatori;
- "fdati.dat" - input dati grezzi;
- "fesatti.dat" - dati esatti;
- "ferrati.dat" - dati errati non corretti;

- "fregdet.dat" - regole di correzione in codice;
- "fregerc.dat" - parallelo a "fcorretti.dat";
- "fcorrett.dat" - dati corretti;
- "fregout.dat" - regole di incompatibilità.

Alla fine del passaggio vengono mostrati i contatori dei record letti, esatti, errati e corretti e la lista delle regole di controllo attivate, con il relativo numero di regola e quante volte ogni regola è stata attivata.

2) VISTA REGOLE in IN USO
La scelta della funzione *"Vista delle regole in uso"* mostra le regole numerate registrate sul file esterno *"fregout.dat"*.
Se una regola di incompatibilità è corredata da una o più regole di correzione, regole che sono registrate nel file esterno *"fdetout.dat"*, queste ultime vengono mostrate precedute da un *"THEN"* sotto la relativa regola di incompatibilità.

3) STATISTICA ERRORI
I contatori e la lista delle regole di incompatibilità attivate, vengono mostrate automaticamente alla fine dell'esecuzione del programma di check e correzione dei dati.
È possibile rivedere la lista delle regole di incompatibilità attivate, che sono registrate sul file esterno *"flischk.dat"* nella cartella di progetto, scegliendo la funzione *"Statistica degli errori"*.
Il file esterno *"fsyschk.dat"* contiene la lista dei contatori dei record letti, esatti, errati e corretti dell'ultimo passaggio di check effettuato.

5.3. Analisi dei risultati

5.3.1. Grafici

Poiché, nell'approccio di correzione deterministico l'imputazione delle variabili è determinata dall'utente, e quindi non è molto significativo vedere la distribuzione delle variabili prima e dopo la correzione, che inoltre può anche mancare, Concord mette a disposizione solamente i grafici.

Sono possibili le seguenti scelte:
- Distribuzioni dei dati:
  viene proposta una maschera nella quale bisogna selezionare la

variabile e il tipo di istogramma desiderato.
Viene mostrata la distribuzione i frequenza, per la variabile selezionata, del valori dei dati grezzi, prima della correzione, e dei dati esatti più i dati corretti, dopo la correzione.

- Frequenza errori:
  Viene mostrata la distribuzione di frequenza delle regole di incompatibilità verificate.

5.4. Errori di esecuzione

Errore	Causa
Errore apertura file xxx:	Il file xxx non esiste.
Lista non ordinata	Il file delle liste non è ordinato per posizione e lunghezza e vengono mostrati i due valori della lista non ordinati.
liste troppo numerose	I valori di lista superano i 10000.

5.5. L'integrazione verso la correzione con donatore

L'integrazione, dall'approccio di correzione deterministica alla correzione effettuata con il metodo del donatore, si attua utilizzando le regole di correzione deterministica.

Vediamo come è possibile applicare l'integrazione tra i due metodi.

Dopo aver effettuato il passaggio di controllo deterministico *"senza"* correzioni, bisogna analizzare i risultati ottenuti e decidere quali sono le variabili che vogliamo, successivamente, sottoporre a correzione tramite donatore.

Per queste sole variabili scriviamo le regole di correzione che, in questo caso, possiamo definire *"regole di impostazione dei caratteri di riconoscimento"* poiché serviranno solamente a cambiare il valore delle variabili da correggere in caratteri specifici *(vedi § 6.1.1)*.

Una volta scritte le regole di correzione, possiamo nuovamente effettuare il passaggio di controllo e, questa volta di correzione deterministica o,

se vogliamo, di impostazione, e otterremo il file dei dati *"corretti"* o, meglio, *"flaggati"*, nel quale le variabili da sottoporre successivamente a correzione con donazione saranno chiaramente individuabili.

Chiudiamo l'approccio deterministico e lo riapriamo con il donatore.

Nella definizione delle variabili di correzione *(vedi § 6.1.1)* scegliamo, tramite l'apposita listbox, le variabili da correggere tra quelle mostrate, che sono tutte le variabili definite nell'approccio di correzione deterministico e che, a loro volta, possono essere state migrate dall'approccio probabilistico, completando così l'integrazione tra i tre metodi.

Esempio:
Dopo aver effettuato un passaggio di check deterministico dei dati decidiamo di correggere le variabili CAROCC , POSPRO e HADIP.

Decidiamo di assegnare alla prima il carattere "§", alla seconda i caratteri "@@" e alla terza il carattere "#".

Attenzione! Il carattere scelto deve essere dato per tutta la lunghezza della variabile; potremmo assegnare qualsiasi carattere, anche semplici lettere maiuscole o minuscole.

Scriveremo allora le seguenti regole di correzione:
CAROCC = '§'
POSPRO = '@@'
HADIP = '#'

selezionando opportunamente le regole di incompatibilità in funzione delle quali vogliamo che vengano corrette le variabili.

Eseguiamo il passaggio di controllo e correzione.

Alla fine del passaggio il file "fcorretti.dat" avrà le variabili impostate ai caratteri prescelti e potrà essere usato per la correzione tramite donatore (vedi cap.6).

6. Le correzioni tramite donatore

Sintesi: *viene descritto il metodo e le varie fasi, da eseguire in modo gerarchico dall'utente, per la correzione effettuata secondo il metodo del donatore*

Le funzioni di CONCORD, nell'approccio di correzione tramite donatore, sono uguali a quelle di RIDA (Ricostruzione delle Informazioni con Donazione Automatica).

Con questo metodo si ottiene la correzione di un file di dati, sempre registrato in ASCII, tramite la tecnica del donatore.

Verranno di seguito descritti sia i principi su cui la tecnica si basa, che i passi che l'utente deve eseguire per rendere operativo il sistema.

Rappresentazione dei dati.

Sia data una matrice di dati *X*, formata da *n* unità e *k* variabili di tipo qualsiasi. Le unità rappresentano i vettori-riga, le variabili i vettori-colonna. Le variabili sono di tipo qualsiasi.

Dal punto di vista della archiviazione elettronica della informazione, la matrice dei dati X è contenuta in un file, costituito da un insieme di record, ognuno rappresentante una unità, e contenente un numero di campi pari al numero di variabili (da ora in poi useremo il termine record o unità come sinonimi). Un insieme di campi al limite anche uno solo, consente di identificare in modo univoco il record-unità ed è detto chiave o identificativo del record.

Dividiamo in due gruppi le variabili:

- variabili affette da errore (in numero di h<k);
- variabili esatte (in numero di k-h).

Supponiamo di sottoporre ad un processo di controllo ogni record, in modo che ognuno degli h campi corrispondenti alle variabili affette da errore contenga o un flag di errore o un valore esatto. Il file risulta diviso in due:

- insieme dei record totalmente esatti;
- insieme dei record che presentano almeno un flag di errore.

Costruzione della metrica delle distanze.

Proponiamoci ora di misurare la distanza tra due unità, rispetto alle variabili esatte. A questo scopo è necessario introdurre una metrica per ogni tipologia di variabile *(si veda Abbate 1996)*.

Sia quindi *d* la distanza tra due unità, misurata rispetto ai seguenti tipi di variabile:

a) Variabile qualitativa sconnessa.
Si pone:
d=0 se le unità presentano la stessa modalità,
d=1 se la modalità è diversa.

Formalmente: $X1 = X2 \Rightarrow d=0$, $X1 \neq X2 \Rightarrow d=1$

b) Variabile ordinata con m modalità.
Si pone:
d=0 se sulle due unità è stata rilevata la stessa modalità;
d=1 se le modalità sono adiacenti;
d=2 se tra di esse ce n'è una sola, e così via fino a d=m-1, se le due modalità sono agli estremi opposti.
Per rendere d variabile tra 0 e 1, essa viene divisa per il suo massimo m-1.

Formalmente: $X1 = X2 \Rightarrow d=0$, $X1 = r$, $X2 = s$ $(r \neq s) \Rightarrow d = \frac{|r-s|}{m-1}$

c) Variabile qualitativa telescopica.

Tali variabili sono rappresentabili tramite un insieme di gruppi primari di livello 1, contenenti ognuno più sottogruppi di livello 2.

Ogni sottogruppo di livello 2 contiene più sottogruppi di livello 3 e così via fino ad un sottogruppo di livello j, contenente modalità non ulteriormente scomponibili in sottogruppi, che sono al livello più basso j+1.

Una modalità siffatta può essere codificata con g gruppi di bit, ognuno dei quali è dimensionato in modo da poter rappresentare tutti i sottogruppi relativi a quel livello. Poniamo:

d=0 se le due unità presentano stessa modalità,
d=1 se le due modalità diverse sono nello stesso sottogruppo di livello j,
d=2 se esse sono in gruppi differenti di livello j, ma nello stesso sottogruppo di livello j-1,
d=3 se sono in gruppi differenti di livello j-1, ma nello stesso sottogruppo di livello j-2 e così via fino ad un massimo di d=j+1 se le due modalità sono in gruppi primari diversi di livello 1

Rendiamo la distanza variabile tra 0 ed 1 dividendola per il suo massimo pari a j+1. Sia r il livello più alto a partire dal quale si riscontra una differenza tra X1 ed X2, r assume quindi valori tra 1 e j+1.

Formalmente: $X1 = X2 \Rightarrow d=0$, $X1 \neq X2 \Rightarrow d=\frac{|j+2-r|}{j+1}$

In CONCORD, come in RIDA, questo tipo di distanza è utilizzato nel caso particolare che sia sufficiente una sola cifra per rappresentare ogni livello. Date quindi due generiche modalità di una variabile di tipo telescopico, esse distano 0 se tutte le cifre sono uguali, 1 se solo l'ultima è diversa, 2 se sono diverse soltanto l'ultima e la penultima e così via;

d) Variabile quantitativa.

Sia X1 il valore assunto dalla variabile X nella prima unità, X2 nella seconda. Poniamo d=| X1- X2 |. La distanza può essere resa variabile tra 0 e 1 dividendola per il suo massimo, pari alla differenza tra i valori massimo (Xmax) e minimo (Xmin) della variabile X presenti nel file.

Formalmente: $X1 = X2 \Rightarrow d=0$, $X1 \neq X2 \Rightarrow d=\frac{|X1 - X2|}{X\max - X\min}$

Nella versione attuale, il valore assoluto della differenza tra X1 e X2 è diviso per X1+1, misurando uno scostamento relativo rispetto ad X1 (la scelta di X1+1 serve per evitare un denominatore degenere, nel caso che sia X1=0). È evidente che la scelta di una distanza siffatta privilegia l'importanza della variabile quantitativa, in particolare se il valore di X2 risultasse molto distante da quello di X1.

Formalizzazione della funzione di distanza mista ponderata.

Assegnata una matrice di dati, presentante k-h variabili non affette da errore, definiamo distanza mista ponderata D tra due generiche unità una espressione del tipo:

$$D = \sum_{i=1}^{r} W_i D_i$$

dove Di è la distanza tra le due unità rispetto alla variabile i, misurata con una delle espressioni di cui sopra e Wi è un numero reale positivo che rappresenta l'importanza assegnata alla variabile i nel calcolo della distanza. Le r variabili sono scelte tra le k-h quelle non affette da errore. L'attuale versione accetta solo numeri naturali per Wi. Chiamiamo variabili di accoppiamento o di matching le r variabili scelte per il calcolo della distanza.

Scelta dell'unità donatrice.

Data un'unità affetta da errore nella variabile k si vuole trovare l'unità esatta posta alla distanza minima. Essa è detta unità donatrice, perché il valore della variabile k relativo ad essa è "donato" all'unità affetta da errore. L'insieme della unità tra le quali è scelta l'unità donatrice è detto serbatoio dei donatori. Il serbatoio dei donatori può essere costruito in due modi:

1) selezionando le unità esatte rispetto alla sola variabile k;
2) selezionando le unità esatte rispetto a tutte le variabili.

Nel primo caso si usa un diverso serbatoio per ogni variabile errata; nel secondo caso si utilizza un serbatoio unico per tutte le variabili affette da errore. La prima procedura è utile quando si desidera disporre di serbatoi di donatori relativamente numerosi per ogni variabile da correggere. Questa scelta deve essere effettuata e realizzata prima di utilizzare CONCORD. La scelta dell'unità donatrice è ulteriormente affinabile scegliendo, nell'insieme delle variabili non affette da errore e non usate come variabili di accoppiamento, delle variabili dette di strato. Dopo aver formato il serbatoio dei donatori in uno dei due modi di cui sopra, si seleziona l'unità donatrice tra quelle che inoltre, rispetto alle variabili di strato, presentano le stesse modalità dell'unità affetta da errore. L'uso di variabili di strato implica l'accettazione della possibilità di non avere donatori idonei per quell'unità.

Funzione di distanza mista ponderata corretta.

Possiamo introdurre un perfezionamento alla distanza mista ponderata sopra introdotta, per penalizzare l'unità del serbatoio che è già stata utilizzata nella donazione. Ridefiniamo la distanza D come:

$$D = \sum_{i=1}^{r} W_i D_i + kp$$

dove k è il numero di volte per cui l'unità è stata precedentemente utilizzata, p è un fattore di penalità. Questa espressione più completa è adottata da CONCORD, che richiede che p sia un numero intero.

Ponderazione delle variabili di matching.

Sono molte le tecniche possibili di ponderazione delle variabili di matching. Le applicazioni finora realizzate nell'interno dell'istituto hanno utilizzato il criterio del χ^2 . Esso si applica nel seguente modo (Abbate 1996):

1) si misura la connessione tra la variabile affetta da errore e quelle esatte tramite l'indice χ^2 . Il valore dell'indice dipende dal numero di celle della tabella di contingenza. Poiché bisogna confrontare il valore dei χ^2 ottenuti, per renderli confrontabili occorre o riclassificare in modo opportuno almeno la variabile da correggere, se di tipo quantitativo, in modo da ottenere tabelle di contingenza di dimensioni omogenee, oppure dividere direttamente il valore del χ^2 per il numero di gradi di libertà, che è pari al prodotto tra il numero delle righe e delle colonne della tabella di contingenza diminuiti entrambi di uno;

2) l'utilizzatore del metodo deve esaminare criticamente i valori di χ^2 così ottenuti, eventualmente divisi per il numero dei gradi di libertà: le variabili non affette da errore che presentano il valore più alto sono le migliori candidate ad essere variabili di strato, quelle con valore immediatamente inferiore possono diventare variabili di matching. L'utilizzatore del metodo deve usare i valori come supporto a una decisione che tiene anche conto della sua conoscenza dell'indagine.

La scelta delle variabili di strato deve tenere conto anche del fatto che all'aumentare del loro numero, aumenterà la selettività nell'ambito del serbatoio dei donatori, ma aumenterà anche la probabilità di non trovare il donatore.

Nelle applicazioni finora realizzate è stata sempre impiegata una sola variabile di strato.

Modalità di utilizzo.

L'utente deve preparare 3 i seguenti file:

1) File contenente i record errati. In esso le variabili affette da errore debbono contenere un carattere di errore ripetuto per tutta la lunghezza del campo.

2) File contenente i record esatti, costituenti il serbatoio dei potenziali donatori.

3) File dei parametri. Viene generato automaticamente da CONCORD dopo aver definito le variabili da correggere, con il carattere di errore particolarmente importante perché solo i campi in cui esso è presente sono soggetti a correzione, le variabili di strato e di matching. Per ogni variabile occorre specificare la posizione iniziale, la lunghezza, il tipo (obbligatorio per le variabili di matching, al fine della scelta della funzione di distanza da adottare) e il peso. Le variabili quantitative possono essere riclassificate, specificando l'estremo superiore di ogni classe. Si possono poi inserire i parametri U,R,L. Essi sono, rispettivamente, il numero massimo di volte che la stessa unità può essere utilizzata come donatrice, il fattore moltiplicativo che penalizza l'uso ripetuto dello stesso donatore e la massima distanza a cui può essere considerato un donatore. Vale la stessa avvertenza formulata a proposito dell'uso degli strati: l'uso dei parametri U e L implica la possibilità di non riuscire a trovare il donatore. I parametri U, R, L possono mancare: questo implica che non si pone alcun limite alla possibilità di riutilizzare lo stesso donatore ed esso può essere scelto anche molto distante rispetto all'unità donatrice.

Dopo la corretta esecuzione dei vari passi, l'utente dovrà provvedere a fondere in un file unico il file dei record esatti, quello dei corretti e quello eventuale degli incorretti.

Se è stato prodotto il file dei record incorretti, nel file unico creato dall'utente i record non corretti conterranno ancora il carattere di errore: essi debbono essere corretti con una tecnica alternativa.

6.1. La fase di definizione

Figura 6.1 La gestione delle variabili per donatore

La prima funzione possibile, una volta scelto in Concord il metodo di correzione tramite donatore, è quella di definizione. Bisogna definire le variabili di correzione, di strato, di match e infine i parametri di esecuzione.

6.1.1. Definizione variabili di correzione

Le variabili di correzione sono quelle che individuano i valori da correggere sui record errati e sono dette anche variabili di tipo "*A*". Tra i record errati vengono selezionati quelli che hanno un campo, determinato dalla posizione e lunghezza della variabile, contenente gli stessi caratteri di individuazione della variabile di correzione e questi caratteri vengono sostituiti, con diverse modalità, con dati presi dai record donatori.

Esempio:
desideriamo correggere la variabile alfanumerica A01 posizione 30, lunghezza 3 individuata nei record errati dai caratteri BBB, sostituendovi il valore del campo corrispondente di un record donatore individuato in funzione della distanza.

Deve esistere *almeno* una variabile di tipo "A".

Utilizzando l'integrazione tra i metodi, è' possibile scegliere le variabili da correggere tra quelle mostrate nella listbox "variabili" con un doppio click sul tasto sinistro del mouse; in questo caso il programma numera

automaticamente la variabile, ne definisce la posizione, la lunghezza, il tipo (alfanumerico) e la correzione (1 - spostamento), lascia vuoto il carattere di correzione, che deve essere inserito poi dall'utente nella tabella[5], e indica la variabile di provenienza associata.

In alternativa, per ogni variabile è necessario indicare:

- Numero: A*xx* - il numero univoco deve iniziare sempre con la lettera A e *xx* deve assumere i valori nel range 01-99 per un massimo di 24 (esempio A01,A12,A40…);
- Posizione: indica la colonna del tracciato record dove inizia la variabile;
- Lunghezza: indica la lunghezza della variabile e per le variabili numeriche (tipo N e C) può essere nella forma *"i.d"* dove *i* è la parte intera e *d* il numero dei decimali;
- Tipo: il tipo della variabile che può essere:
  X variabile alfanumerica;
  N variabile numerica;
  T variabile telescopica (es.: una variabile che indica l'attività economica);
  C variabile di classificazione.
- Correzione: il tipo di correzione da effettuare:
  1 la correzione consiste nello spostamento della corrispondente variabile del record donatore e, in questo caso, la variabile non può essere di classificazione (tipo C);
  2 la correzione è il risultato di un calcolo (definito dalle variabili "F" ed "E" vedi appresso) e in questo caso il tipo della variabile deve essere numerico (tipo N) o di classificazione (tipo C); se è di classificazione il risultato del calcolo viene sostituito con il valore della classe secondo i parametri della variabile "C" corrispondente con la lunghezza della variabile di imputazione "A".

---

[5] Per inserire o modificare variabili nella tabella posizionare il cursore sulla relativa tabella e utilizzare il tasto destro del mouse.
Per inserire una variabile:
scegliere *"Add row"* e riempire i campi della riga *"new"*;
ripetere *"Add row"* per una seconda variabile e così via.
Alla fine *"Commit new row"*.
Per modificare una variabile:
posizionare il cursore sul numero della riga della tabella da modificare e premere due volte il tasto sinistro del mouse fino ad evidenziare tutta la riga;
spostarsi poi sulla cella da modificare e variarla;
così via per tutte le modifiche.
Alla fine delle modifiche, per assicurarsi che anche l'ultima modifica sia stata memorizzata, con il tasto destro del mouse scegliere *"Add row"* e poi *"Cancel row edit"*.

- Carattere: rappresenta un singolo carattere, che verrà automaticamente esteso per tutta la lunghezza della variabile, che individua nei record errati la variabile sulla quale effettuare la correzione. Se il carattere manca viene assunto per default blank. Se il carattere è "*" la correzione della variabile viene effettuata su tutti i record.

Esempio:
desideriamo correggere la variabile di classificazione A01 posizione 30, lunghezza 3 individuata nei record errati dai caratteri BBB, sostituendovi il valore della terza classe descritta nella variabile "C" e ottenuto dal calcolo, descritto nella variabile "F", sui dati del record esatto determinato in funzione della distanza.

viene generato il seguente record nel file "dvardom.dat" nella cartella di progetto:

.V A01 P=30 L=3 T=C X=2 B

Se la variabile di correzione è di tipo C (classificata) deve esistere almeno una variabile associata di classificazione come sotto descritta.

VARIABILI "C" DI CLASSIFICAZIONE

Queste variabili devono essere inserite manualmente nella tabella.

Per ogni variabile bisogna indicare:
- Riferimento: deve essere uguale alla variabile di imputazione con tipo "C" (esempio C01, C06 ecc.);
- Classi: fino a 7 valori che rappresentano l'estremo superiore delle classi. Se i valori sono più di 7 scriverli su un'altra riga con stesso riferimento.

Esempio (riferito alla variabile "A" precedente):
A01 1 9.99 50 99 499 999 5000
A01 100000

vengono generati i seguenti due record nel file "dvardom.dat" nella cartella di progetto:
.C A01 1 9 50 99 499 999 5000
.C A01 100000

Se nella variabile da correggere "A" il tipo di correzione è 2, ossia calcolo, deve esistere almeno una variabile associata per formula di calcolo con le seguenti caratteristiche:

VARIABILI "F" PER FORMULE DI CALCOLO

Queste variabili devono essere inserite manualmente nella tabella.

Per ogni variabile indicare:

- Riferimento: deve essere uguale alla variabile di imputazione con tipo correzione 2 (esempio A01, A06, ecc.);
- Formula: formula di calcolo del valore da imputare nella variabile Axx:: può contenere gli operatori aritmetici;
  l'ordine di esecuzione delle operazioni è:
  ( ) ^ per elevamento a potenza- / * + -
  gli operandi possono essere costanti o variabili identificate dal riferimento;
  possono essere usate parentesi per modificare l'ordine di esecuzione delle operazioni.

Esempio:

E01/(D01+D02)

Che significa che una volta determinato il record esatto in funzione della distanza, verrà calcolata prima la somma delle variabili D01 e D02 e poi il rapporto. Viene generato il seguente record nel file "dvardom.dat" nella cartella di progetto:

.F A01 E01/(D01+D02)

La variabile Exx e Dxx sono variabili di calcolo sotto descritte:

VARIABILI "*E*" PER IL CALCOLO.

Queste variabili devono essere inserite manualmente nella tabella.

Indicano le variabili che sono utilizzate per la correzione di variabili A*xx* con tipo correzione 2 nelle formule di calcolo descritte nelle variabili di tipo *"F"*.

Per ogni variabile è necessario indicare:

- Numero che può essere:
  E*xx* - il numero, univoco e riferito a quello della formula descritta

nella variabile "F", deve iniziare sempre con la lettera E se il valore è da cercare sul record errato e xx deve assumere i valori nel range 01-99;

D*xx* - il numero, univoco e riferito a quello della formula descritta nella variabile "F", deve iniziare sempre con la lettera D se il valore è da cercare sul record donatore e *xx* deve assumere i valori nel range 01-99.

- Posizione: indica la colonna del tracciato record dove inizia la variabile.
- Lunghezza: indica la lunghezza della variabile e per le variabili numeriche (tipo N) può essere nella forma *"i.d"* dove *i* è la parte intera e *d* il numero dei decimali.
- Tipo: tipo della variabile: N per numeri, C per le classificazioni e in questo caso la lunghezza deve essere un numero intero.
  Anche per le variabili di calcolo di tipo C dovranno essere scritte le relative variabili di classificazione *(vedi variabili tipo "C" di classificazione)*.

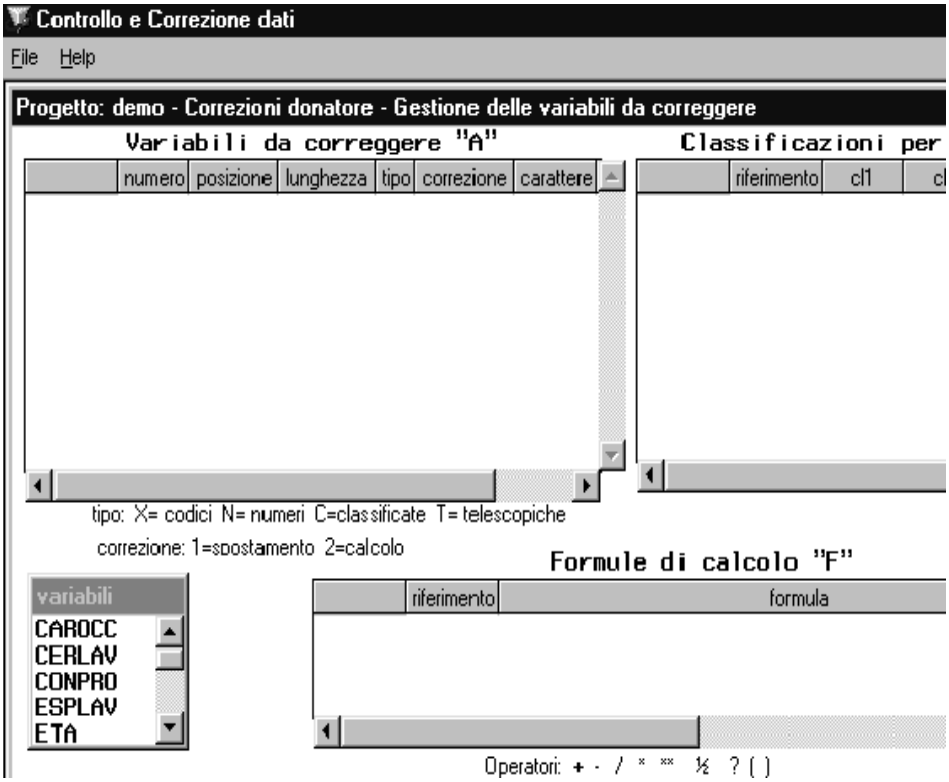
"Import" permette di importare da un file esterno formattato del tipo *"dvardom.dat"* variabili precedentemente definite.

Alla fine delle operazioni di definizione delle variabili da correggere, dal menu a tendina *"File"* è possibile scegliere:

"Save and exit" per memorizzare le operazioni effettuate nel relativo dataset;

"Exit" per uscire senza salvare.

Figura 6.2 La gestione delle variabili da correggere "A" per donatore



6.1.2. Definizione variabili di strato

Indicano le variabili di stratificazione che si utilizzano per individuare gruppi di record donatori simili fra loro e sono chiamate variabili *"S"*.

Lo strato è importante perché definisce la parte del serbatoio dei record donatori ove effettuare la ricerca del record esatto con distanza minima rispetto al record con variabili da correggere.

Deve essere presente almeno uno strato.

Sfruttando l'integrazione tra i metodi, è possibile scegliere la variabile di strato tra quelle mostrate nella listbox *"variabili"* con un doppio click sul tasto sinistro del mouse; in questo caso il programma numera automati-

camente la variabile, ne definisce la posizione, la lunghezza, il tipo (alfanumerico) e la variabile di provenienza associata.

Per ogni variabile bisogna indicare:

- *Numero:* S*xx* - il numero univoco deve iniziare sempre con la lettera S e *xx* deve assumere i valori nel range 01-99 per un massimo di 24 variabili (esempio S01,S12,…S83…).
- *Posizione:* indica la colonna del tracciato record dove inizia la variabile.
- *Lunghezza:* indica la lunghezza della variabile; per le variabili di tipo C può essere nella forma *"i.d"* dove *"i"* è la parte intera e *"d"* il numero dei decimali e sia per i che d la lunghezza massima è 15.
- *Tipo:* tipo della variabile:
  X variabile alfanumerica;
  T variabile telescopica (es.: una variabile che indica l'attività economica);
  C variabile di classificazione.

Esempio:
S01 posizione 30 lunghezza 3 tipo C : viene generato il seguente record nel file "dvardom.dat" nella cartella di progetto:

.V S01 P=30 L=3 T=C

Se il tipo della variabile è C (classificata) deve esistere almeno una variabile associata di classificazione con le seguenti caratteristiche:

VARIABILI "C" DI CLASSIFICAZIONE

Queste variabili devono essere inserite manualmente nella tabella.

Per ogni variabile bisogna indicare:

- Riferimento: deve essere uguale alla variabile di strato con tipo "S" (esempio S01, S06 ecc);
- Classi: fino a 7 valori che rappresentano l'estremo superiore delle classi. Se i valori sono più di 7 scriverli su un'altra variabile con stesso riferimento.

Esempio (riferito alla variabile "A" precedente):
S01 1 9.99 50 99 499 999 5000
S01 100000

vengono generati i seguenti due record nel file "dvardom.dat" nella cartella di progetto:

.C S01 1 9 50 99 499 999 5000

.C S01 100000

"Import" permette di importare da un file esterno formattato variabili precedentemente definite.

Alla fine delle operazioni di definizione delle variabili di strato, dal menu a tendina *"File"* è possibile scegliere:

"Save and exit" per memorizzare le operazioni effettuate sul relativo dataset.

"Exit" per uscire senza salvare.

6.1.3. Definizione delle variabili di match

Le variabili di match sono utilizzate per calcolare la distanza tra il record da correggere e i record che fanno parte del serbatoio dei donatori.

Le variabili di match sono dette variabili di tipo "*M*". Le variabili di match possono mancare; in questo caso il serbatoio dei donatori deve essere mirato alle correzioni da effettuare poiché ogni record si trova, almeno inizialmente, alla distanza minima.

Per ogni variabile di match è necessario indicare:

- *Numero:* M*xx* - il numero, univoco, deve iniziare sempre con la lettera M e *xx* deve assumere i valori nel range 01-24 (esempio M01, M02, M03…).
- *Posizione:* indica la colonna del tracciato record dove inizia la variabile.
- *Lunghezza:* indica la lunghezza della variabile; per le variabili di tipo C,N può essere nella forma *"i.d"* dove *i* è la parte intera e *d* il numero dei decimali e sia per i che d la lunghezza massima è 15.
- *Tipo:* tipo della variabile:
  X variabile alfanumerica.
  N variabile numerica.
  T variabile telescopica (es.: una variabile che indica l'attività economica).
  C variabile di classificazione.

- *Peso* con il quale considerare la variabile di match rispetto alle altre; se non specificato si assume 1.

Esempio:

M01 posizione 30, lunghezza 3, tipo C peso 5: viene generato il seguente record nel file "dvardom.dat" nella cartella di progetto:

.V M01 P=30 L=3 T=C W=5

Se il tipo della variabile è C (classificata) deve esistere almeno una variabile associata di classificazione con le seguenti caratteristiche:

VARIABILI "C" DI CLASSIFICAZIONE

Queste variabili devono essere inserite manualmente nella tabella.

Indicare per ogni variabile di classificazione:

- *Riferimento:* deve essere uguale alla variabile di match con tipo "C" (esempio M01, M06 ecc.);
- *Classi:* fino a 7 valori che rappresentano l'estremo superiore delle classi. Se i valori sono più di 7 scriverli su un'altra variabile con stesso riferimento.

Esempio (riferito alla variabile "M" precedente):

M01 1 9.99 50 99 499 999 5000

M01 100000

vengono generati i seguenti due record nel file "dvardom.dat" nella cartella di progetto:

.C M01 1 9 50 99 499 5000

.C M01 100000

"Import" permette di importare da un file esterno formattato variabili precedentemente definite.

Alla fine delle operazioni di definizione delle variabili di match, dal menu a tendina *"File"* è possibile scegliere:

"Save and exit" per memorizzare le operazioni effettuate sul file relativo dataset;

"Exit" per uscire senza salvare.

6.1.4. Definizione parametri di impostazione per donatore

Queste variabili devono essere inserite manualmente nella tabella.

La definizione dei parametri di impostazione serve a condizionare, con parametri specifici, la fase di imputazione per tutti i record da correggere.

Vediamo in dettaglio i vari parametri:

Parametro "U".

Rappresenta il fattore di penalizzazione per i record donatore già utilizzati, in modo che detti record, rientrino nella donazione solo se non esistono altri record con distanza calcolata più bassa.

Il valore deve essere intero.

Se non esiste si assume 0.

Parametro "R"

Rappresenta il numero massimo di volte in cui uno stesso record donatore può essere utilizzato.

Se non esiste si assume 0 ovvero riutilizzo illimitato.

Il valore deve essere intero.

Parametro "L"

Rappresenta il limite massimo della distanza tra il record donatore e il record da correggere.

Il valore deve essere intero.

Se non esiste si assume 0 ossia qualsiasi distanza .

Parametro "D".

Rappresenta la distanza minima di accettazione per un donatore.

Il valore deve essere intero.

Se non esiste si assume 0 ossia la distanza accettabile è sempre uguale a zero.

Il calcolo della distanza (D) tra due variabili X,Y dipende dal tipo di variabile:

- se tipo X allora D(X,Y)=0 se X=Y, D(X,Y)=1 se X?Y;
- se tipo C allora D(X,Y)= |Classe(X) - Classe(Y)| / Numero delle classi - 1;
- se tipo N allora
  D(X,Y)= |(X - Y) / MAX (|MIN(Errati) - MAX(Esatti)| , |MAX(Errati) - MIN(Esatti)|);
- se tipo T allora
  D(X,Y)=la posizione del primo carattere diverso a partire da sinistra / lunghezza della variabile.

La distanza (D) tra due record K,L è definita:

$$D<K,L> = (Somma(P(I) * D(I)<K,L>)) + R<L> * U$$

in cui P(I) rappresenta il peso della variabile(I), R<L> rappresenta il numero delle volte in cui è stato utilizzato il donatore L, e U è il fattore di penalizzazione.

"Import" permette di importare da un file esterno formattato variabili precedentemente definite.

Alla fine delle definizioni dei parametri dal menu a tendina *"File"* è possibile scegliere:
"Save and exit" per memorizzare le operazioni effettuate sul relativo dataset;

"Exit" per uscire senza salvare.

Per inserire o modificare variabili nella datatable posizionare il cursore sulla relativa tabella e utilizzare il tasto destro del mouse.

6.2. Le funzioni

6.2.1. Controllo delle variabili per donatore

Scegliendo questa funzione si esegue la fase di controllo delle variabili di strato, match, correzione e dei parametri definiti dall'utente tramite le apposite funzioni.

Viene prima registrato nella cartella di progetto il file *"dvardom.dat"* dai

dataset specifici, contenente le variabili come definite dall'utente, e successivamente eseguito il programma di controllo.

Se durante la fase di controllo vengono rilevati errori di definizione delle variabili, o lacune nelle definizioni, verrà mostrato un messaggio di avvertimento ed evidenziata la lista delle variabili errate con gli errori relativi.

Se il controllo non rileva errori viene resa possibile la funzione di correzione.

6.2.2. Correzione

Con questa funzione si eseguono i passi di correzione tramite donatore.

Viene mostrata una maschera dalla quale, cliccando sull'icona relativa, è possibile selezionare il file di input contenente i record corretti donatori.

Selezionare poi il file degli errati da correggere cliccando sull'icona relativa.

Scegliere infine il tipo di messaggistica che verrà visualizzata durante l'elaborazione:

- Per singolo passo di elaborazione: all'esecuzione di ogni programma viene mostrato un messaggio di informazione.
- Solo se si verifica un errore; in questo caso viene mostrato il codice e il tipo di errore.

Con il bottone *"Esegui"* o con *"Save and exit"*, dal menu a tendina, vengono eseguiti i passi di elaborazione.

I file generati, nella cartella di progetto, dai vari programmi durante la correzione tramite donatore sono individuabili dal nome *"donax.ttt"* ove *"x"* è il passo eseguito e *"ttt"* e il tipo di file.

Il primo passo è l'esecuzione del programma *dona2* per la creazione di un file dei record da correggere o dei record esatti con le sole variabili di interesse: estrae le variabili dal file degli errati o dei donatori utilizzando le schede parametro di tipo strato e match, mettendo nel file di output prima le variabili di strato e poi le variabili di match; se le variabili sono di tipo t=c al valore delle variabili sostituisce la classe di appartenenza.

Input:

1. file dei record errati o esatti;
2. schede parametro.

Output:

1. file dei record errati o esatti secondo il seguente tracciato:
   - 01-08 progressivo record;
   - 09 in poi valore delle variabili di strato;
     valore delle variabili di match;
2. scheda parametro contenente le informazioni:
   - 01-08 lunghezza record input;
   - 09-17 numero record errati/esatti;
   - 18 blank;
   - 19-26 lunghezza complessiva delle variabili di strato.

Il secondo passo consiste nell'esecuzione del programma *dona3* per il calcolo delle distanze: il programma calcola a parità di strato, la distanza di ogni record errato con tutti i record donatori tenendo presente i parametri:

limite del riutilizzo, dal parametro R se esiste;

fattore di penalizzazione, dal parametro U se esiste;

Viene scelto come donatore il record con distanza minima, sempre tenendo conto del limite della distanza dal parametro L, se detto parametro esiste.

Input:

1. file ridotto dei record errati ordinato per il valore delle variabili di strato;
2. file ridotto dei record esatti ordinato per il valore delle variabili di strato;
3. file dei parametri tipo record *"m"*;
4. file dei parametri tipo record *"i"*.

Output:

1. file di associazioni tra record errati e donatori contenente:
   - 01-08 numero progressivo record errato;
   - 09-16 numero progressivo record esatto o 0 (donatore non trovato);
   - 17 – distanza tra i due record;

2. file per le statistiche sull'esito della correzione che riporta per ogni strato le seguenti informazioni:
   (a) valore dello strato;
   (b) numero di record da correggere;
   (c) numero di donatori disponibili nello strato;
   (d) numero di donatori utilizzati 1 volta;
   (e) numero di donatori utilizzati 2 volte;
   (f) numero di donatori utilizzati 3 volte;
   (g) numero di donatori utilizzati 4-9 volte;
   (h) numero di donatori utilizzati 10-99 volte;
   (i) numero di donatori utilizzati 100o+ volte;
   (j) numero di donatori con distanza d=0;
   (k) numero di donatori con 0 < d < 1;
   (l) numero di donatori con 1 <= d < 10;
   (m) numero di donatori con 10<= d < 100;
   (n) numero di donatori con 100<= d < 1000;
   (o) numero di donatori con 1000<= d < 10000;
   (p) numero di donatori con 10000<= d;
   (q) numero di record non corretti per limite della distanza;
   (r) numero di record non corretti per limite del riutilizzo.

Il terzo passo consiste nellè l'esecuzione del programma *dona4* che esegue la correzione.

Input:

1. file dei record esatti;
2. file dei record errati;
3. file delle associazioni *"dista.ord"* tra i record errati e i record donatori, ordinato per numero progressivo del record errato;
4. file parametri tipi record *"a"* e *"d"*;
5. file delle schede parametro *"scheda.ord"* contenente la lunghezza record e numero dei record errati ed esatti.

Output:

1. file *"dona4.cor"* con i record corretti;
2. file *"dona4.noc"* con i record non corretti, cioè i record per i quali, nel file delle associazioni, il numero progressivo del donatore è uguale a 0;
3. file *"dona4.wrg"* con eventuali errori di esecuzione.

Il file *"dona4.prm"* contiene l'elenco di tutti i file.

Viene infine eseguito il programma *dona5* per le statistiche di correzione.

Input:

1. il file di statistiche prodotto dal programma dona3.

Output:

1. il file "report.lst" tabulato a livello di strato con le seguenti informazioni:
   - distribuzione per strato / totale;
   - valore dello strato xxxxxxxxxx;
   - donatori utilizzati:
   - 1 volta, 2 volte, 3 volte, 4-9 volte, 10-99 volte, 100 o più volte, totale;
   - donatori utilizzati a distanza:
   - d=0 0<d<1 1<=d<10 10<=d<100 100<=d<1000 1000<=d=10000 d>=10000;
   - totale correzioni xxxxxxxx[6] ;
   - donatore non trovato xxxxxxxx di cui non nello strato s1[7] ;
   - per limite distanza xxxxxxx;
   - per limite riutilizzo xxxxxxx;
   - donatori utilizzati xxxxxxxx;
   - rapporto di utilizzo xx.xxx[8].

Con il bottone *"Annulla"* o con *"Exit"*, dal menu a tendina, si abbandona l'elaborazione.

[6] Numero di record errati per strato e complessivo

[7] Numero di record errati per strato e totale nel caso in cui il numero donatori è = 0

[8] Totale correzioni / donatori utilizzati

7. Funzioni comuni ai tre approcci

7.1. I programmi di utilità

CONCORD è corredato con numerose funzioni di utilità per aiutare l'utente nella manipolazione dei dati, funzioni che possono essere usate anche indipendentemente dall'approccio di correzione scelto. Però non tutte le funzioni sono immediatamente disponibili, ma lo diventano secondo l'approccio e lo stato d'avanzamento del progetto.

Figura 7.1 Le funzioni di utilità

7.1.1 Gestione dei dati

La funzione permette la gestione di un qualsiasi file dati scelto dall'utente ma con possibilità diverse secondo l'approccio in atto.

APPROCCIO PROBABILISTICO

Con questa funzione, che utilizza la proc FSEDIT attiva se il relativo modulo SAS è stato installato, si gestiste un file scelto dall'utente tramite apposita maschera, con le variabili predefinite secondo l'approccio in atto.

È possibile, quindi, registrare dati per un file di prova o modificare, nei soli campi definiti, dati già registrati.

La funzione scrive, nella cartella di progetto, un sorgente SAS e lo esegue in modo assolutamente trasparente per l'utente.

APPROCCIO DETERMINISTICO

Con questa funzione, che utilizza la proc FSEDIT attiva se il relativo modulo SAS è stato installato, è possibile *il controllo interattivo dei dati* compilando il sorgente SAS che viene automaticamente generato nella funzione *"regole"* al momento del salvataggio con *"Save and exit"*.

Per poter effettuare il controllo interattivo dei dati:

- scegliere il file da controllare con l'apposita maschera: il programma avverte l'utente che è possibile il controllo interattivo;
- spostare il cursore sul terzo punto della maschera *"Edit Program Statement and compile"* e dare invio;
- nella maschera successiva scrivere una riga ***%include*** *(vedi esempio più avanti)* per poter compilare il file *"fscreen.dat"* che viene generato automaticamente, nella cartella di progetto, al momento del *"save and exit"* nella funzione *"regole"*; oppure non scrivere nulla se non si desidera il controllo;
- premere F3 per compilare ed attivare il modulo con il controllo interattivo.

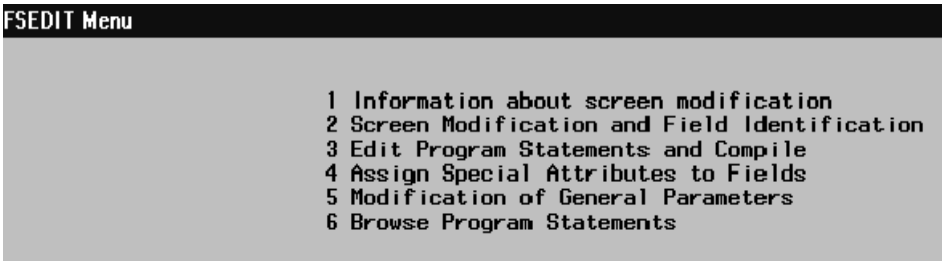
Per esempio: se il progetto risiede nella cartella "d:\prova" la riga ***%include*** *da scrivere deve essere (attenzione al ;):*

%include 'd:\prova\fscreen.dat';

se durante il controllo interattivo viene trovato un errore nei dati, il programma segnala, nella zona messaggi del SAS, il numero della la regola attivata e la regola stessa.

Il file *"fscreen.dat"* è un sorgente scl SAS che, volendo, può essere modificato dall'utente.

Figura 7.2 La maschera per il controllo interattivo dei dati nell'approccio deterministico



7.1.2. Numera

Con questa funzione, utile per creare una variabile che identifichi in modo univoco i record come è necessario per le tavole di verifica nell'analisi dei dati (*vedi §4.3.1*), si registra su un file di output, copiato da un file scelto dall'utente in input, un campo che contiene una numerazione progressiva:

- scrivere il nome del file di input comprensivo di path;
- scrivere il nome del file di output comprensivo di path;
- definire la posizione e la lunghezza del numeratore.

Esempio:
digitiamo posizione 100 e lunghezza 5 e otterremo sui record di output a posizione 100 unnumero progressivo 00001, 00002, 00003 ecc. sul primo, secondo, terzo, ...n record;

se la posizione eccede la lunghezza massima del record di input, il record in output viene allungato.

7.1.3 Ordina

Con questa funzione, utilissima poiché WINDOWS non mette a disposi-

zione programmi di sort efficienti, si ordina un file scelto in input, che verrà ricopiato ordinato nel file di output definito dall'utente:

- scegliere prima il file di input;
- scegliere la directory (cartella) di output;
- scrivere il nome del file di output;
- definire le chiavi di sort con nome (qualsiasi), posizione e lunghezza; per ogni chiave tramite il tasto destro del mouse scegliere *"Add row"* e riempire i campi; alla fine *"Commit new row"* sempre con il tasto destro del mouse;
- *"Esegui"*: il programma controlla la definizione dei campi ed esegue il sort, oppure segnala eventuali errori.

7.1.4. Unisci

Con questa funzione, utile ad esempio quando si deve ottenere il file dei dati puliti come insieme dei dati esatti e dei dati corretti, si copiano due o più file esterni insieme su un nuovo file:

- scegliere i file di input;
- scegliere il file di output o scriverlo nell'area di output;
- con conferma si esegue la copia.

7.1.5. Browser

Con questa funzione si visualizza un file esterno secondo la definizione delle variabili attuale oppure scegliendo a piacere campi del record.

È una funzione, simile ad un text-editor, molto utile poiché permette di vedere solo le parti del record che interessano e solo i record del file che interessano.

È necessario:

- scegliere il file di input;
- scegliere:
  a. le singole variabili da mostrare, prese dalla definizione delle variabili, selezionando le singole variabili, o *"tutte"* le variabili con l'apposito bottone di check-control;
  b. oppure inserire righe nella tabella *"campi da visualizzare"* (con il tasto destro del mouse *"add rows"* e alla fine degli inserimenti

"commit new rows") definendo posizione, lunghezza ed eventuali valori di filtro. In questo caso ogni blank eventuale nel campo valore deve essere sostituito dal carattere ~(alt126). Esempio: se dobbiamo estrarre i record contenenti *" 12 3 "* a posizione 1 lunga 6 dobbiamo scrivere ~12~3~ nella zona *"valore"* relativa.

- *"Zoom"* se vogliamo che i campi ci vengano mostrati in colonne.
- Inserire l'eventuale valore *"dal record"* e *"al record"* per saltare record iniziali o per selezionare solo alcuni record del file.

Dal menu con *"Save and run"* si esegue il programma.

7.1.6. Genera data set

La funzione, possibile dopo la fase di correzione o di check deterministico genera, nella cartella di progetto, il data set SAS *"puliti"* seguendo la definizione delle variabili. Il data set viene formato dall'unione dei file esatti e corretti.

7.1.7. Copia dati

Con questa funzione è possibile la copia, selettiva, totale o parziale, di dati da file esterno su altro file esterno, tramite parametri.

Attivata la funzione occorre:

- scegliere il file di input;
- definire il file di output;
- scegliere il numero di record da copiare oppure *"all"* per copiare l'intero file;
- eventualmente inserire, nella tabella, i parametri di selezione con posizione, operatore (eq, ne, lt, gt) e stringa del valore da ricercare considerando che i blank nella stringa vanno indicati con il simbolo @. Sono ammessi più parametri di selezione.

7.1.8. Genera un file ascii da un data set SAS

Con questa funzione si sceglie, tramite l'apposita maschera, da una cartella di input un dataset SAS da convertire e un file di output (se il file di output non esiste copiare un qualsiasi file della cartella scelta in output con il tasto destro del mouse).

Viene generato, dal dataset SAS di input, un file ASCII a formato fisso con i campi di lunghezza calcolata sulla massima grandezza dei valori delle singole variabili.

Ad esempio: se la variabile "tasso" del dataset contiene i valori 20.5 100 e -6.18 il campo di output sarà di 6 posizioni: 4 per la parte intera, compreso il segno in prima posizione, e 2 per la parte decimale e i record del file conterranno:

002050

010000

-00618

Viene contemporaneamente generato, nella cartella di output, il file *"vardomsas.dat"* con il nome della variabile troncato a sei posizioni, la posizione del campo sul record, la lunghezza e il tipo (A,N).

7.2. Help

CONCORD mette a disposizione dell'utente un help in linea che può essere visto tramite il browser delle pagine html (Interner Explorer, Netscape, ecc.) attivo.

Selezionando dal menu-bar la funzione di *"Help"* viene mostrato a sinistra l'indice dell'help ove è possibile selezionare l'argomento che interessa.

L'help può essere stampato scegliendo da *"Risorse del computer"* la sottocartella *"Help"* nella cartella *"C:\concordconcord"* e il file *"lancio.html"* e stamparlo tramite Internet Explorer con *"Print"* da *"File"* selezionando *"Print all linked documents"*,

7.3. History

Il sistema registra cronologicamente nella cartella "C:\concordconcord" un record sul data set *"history"* in caso di:

- apertura di un progetto;
- chiusura di un progetto;
- importazione di un file esterno;
- esecuzione di una funzione che richiede la scelta di un file da parte dell'utente.

Scegliendo *"Progetto"* oppure *"All"* viene mostrata una tabella con le informazioni registrate fino a quel momento:

- Il campo *"tipo"* della tabella riporta le tre tipologie di correzione;
- Il campo *"step"* riporta il modulo in esecuzione;
- Il campo *"azione"* riporta *"esistente"* per un progetto già in uso, *"nuovo"* per un progetto definito per la prima volta, e *"Import"* per un'azione di caricamento da un file esterno il cui nome viene registrato nel campo *"file"*.

La tabella, selezionando *"Progetto"* dal menu, può contenere il solo progetto in corso; oppure tutti i progetti esistenti selezionando *"All"*.

"Clear" cancella tutte le informazioni presenti in tabella, per il solo progetto in corso scegliendo *"solo progetto"*, o per tutti i progetti selezionando *"tutti i progetti"*. Fare attenzione a *"clear"*: poiché alcuni programmi fanno riferimento ai dati presenti nel dataset *"History"* in fase di esecuzione, questa funzione dovrebbe essere usata solo *al termine del progetto*.

7.4. Log

Selezionando *"Log"* dal menu-bar è possibile passare alla finestra di log del SAS. Per tornare poi all'applicazione puntare con il mouse in un punto qualsiasi della maschera di CONCORD.

8. Esempio di applicazione

I dati di esempio sono registrati nella sottocartella *"Esempio"* in C:\Concord.

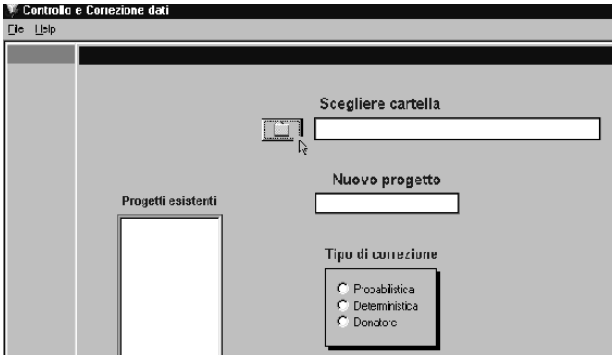
Dopo aver attivato Concord è possibile eseguire l'intero sistema di correzione attraverso i seguenti passi:

1. Dalla maschera iniziale scegliere un nuovo progetto *(vedi figura 8.1)*.

Figura 8.1

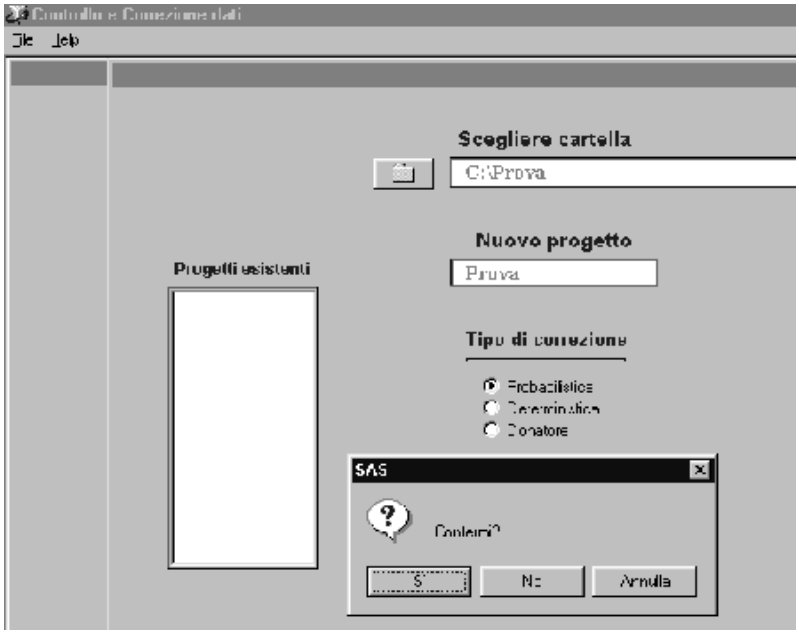
2. Scegliere o generare una cartella, ad esempio *"Prova" (vedi figura 8.2)*.

Figura 8.2



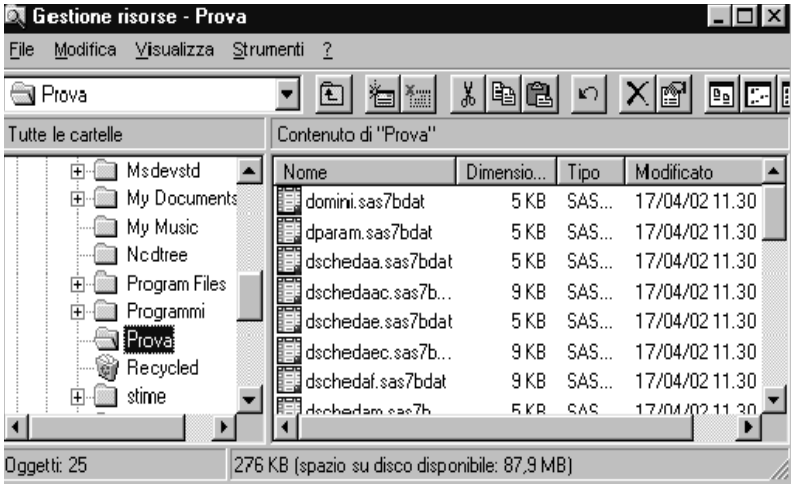
3. Scegliere il tipo di correzione probabilistica e confermare *(vedi figura 8.3)*

Figura 8.3



nella cartella *"Prova"* verranno generati i dataset SAS che verranno utilizzati nelle successive elaborazioni *(vedi figura 8.4)*.

Figura 8.4



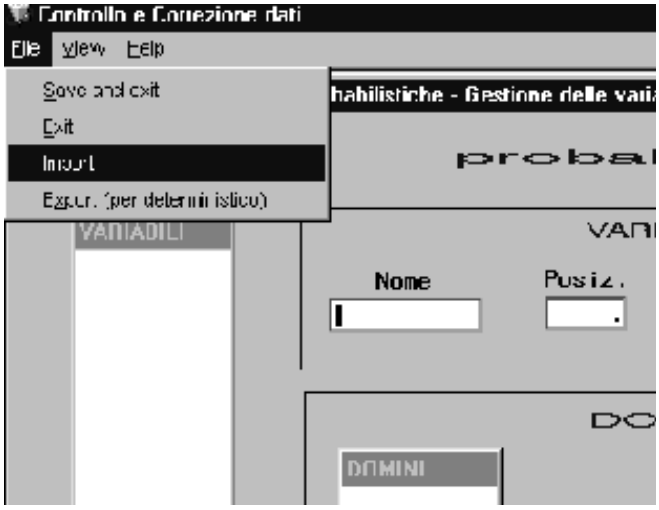
4. Dalla maschera principale di Concord scegliere *"Definizioni variabili" (vedi figura 8.5)*

Figura 8.5

Viene mostrata la maschera di gestione delle variabili con tutti i campi vuoti.

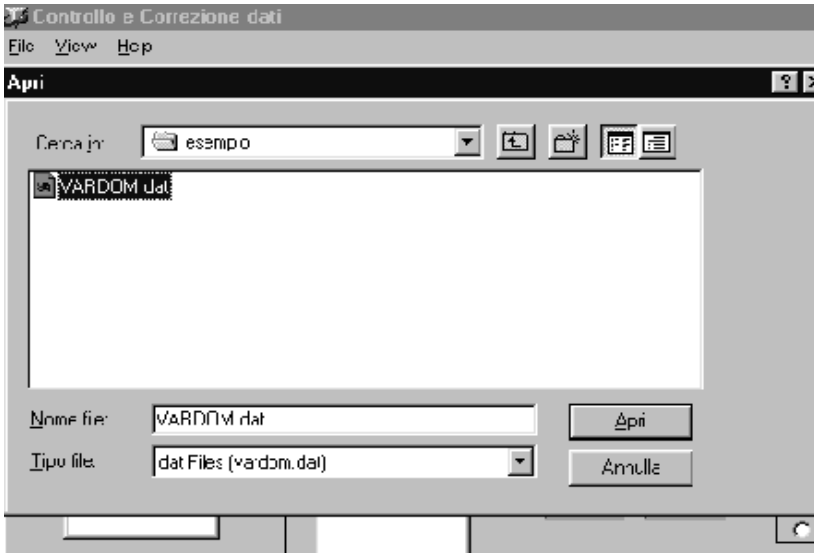
5. Da *"File"* scegliere *"Import" (vedi figura 8.6).*

Figura 8.6



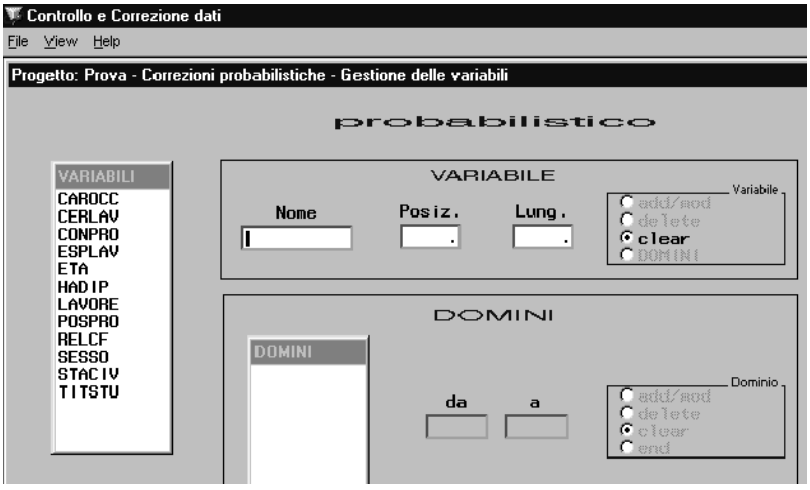
6. Selezionare la sottocartella *"Esempio"* di Concord e scegliere il file *"VARDOM.dat"* e poi *"Apri" (vedi figura 8.7).*

Figura 8.7



Vengono caricate e mostrate le variabili *(vedi figura 8.8).*

Figura 8.8

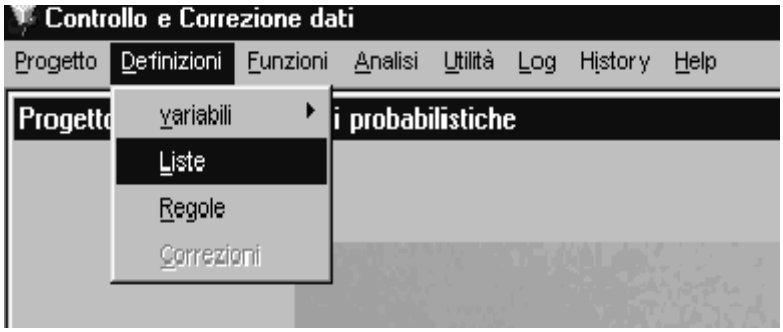


È possibile vedere tutte le variabili con "View" o selezionandole con il mouse.

7. Da *"File"* scegliere *"Save and exit"*. Il programma memorizza le variabili a rende possibili ulteriori scelte in "Definizioni" nella maschera principale.

8. Selezionare *"Liste" (vedi figura 8.9).*

Figura 8.9



Il programma mostra la maschera vuota delle liste.

9. Da *"File"* scegliere import, scegliere la cartella *"Esempio"* di C:\Concord e il file *"STRUTT.dat"* e poi *"Apri" (vedi figura 8.10).* Viene caricato il file delle liste in "*OR*".

Figura 8.10

10. Scegliere il file *"LISTE.dat"* per le liste in *"AND"*.

Dopo l'operazione di *"Import"* dei file contenenti le liste dalla cartella *"Esempio"* nella maschera delle liste per l'approccio probabilistico sono mostrate le tre variabili di lista *(vedi figura 8.11)*; selezionandole con il mouse è possibile vedere le variabili associate ad ogni variabile lista.

Figura 8.11

11. Da *"File"* scegliere *"Save and exit"*. Le liste vengono memorizzate e il programma torna alla maschera iniziale di Concord.

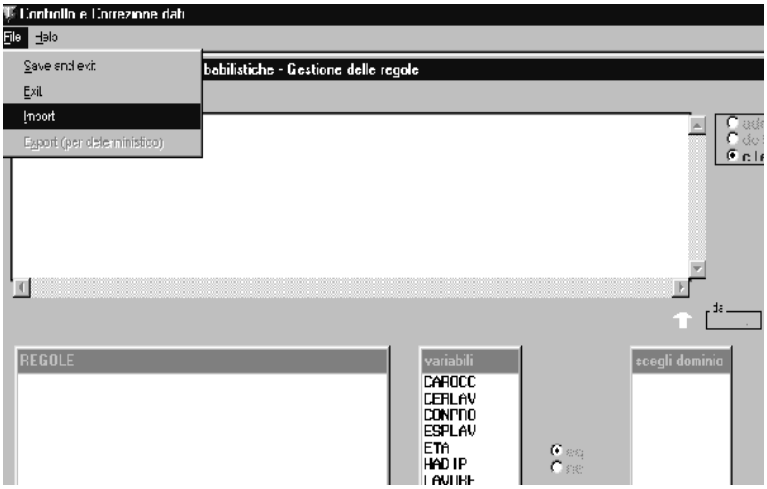
12. Selezionare *"Regole" (vedi figura 8.12)*.

Figura 8.12

Viene mostrata la maschera vuota delle regole di incompatibilità.

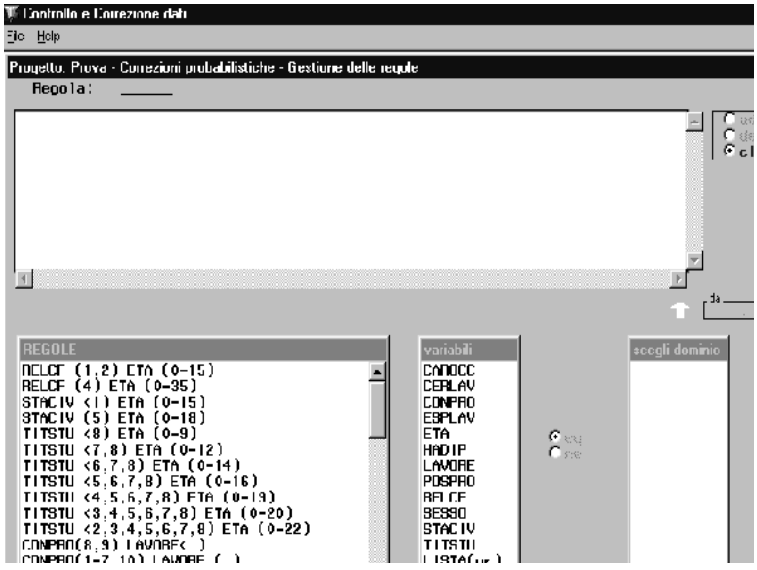
13. Da *"File"* scegliere *"Import" (vedi figura 8.13).*

Figura 8.13



14. Con la stessa procedura seguita per le variabili e le liste selezionare il file *"REGOLE.dat"* dalla cartella *"C:\Concord\Esempio"* e caricare le regole di incompatibilità. A fine caricamento le regole verranno mostrate nella listbox *"REGOLE" (vedi figura 8.14).*

Figura 8.14



15. Da *"File"* scegliere *"Save and exit"*; il programma torna alla maschera iniziale rendendo possibili la funzione per il controllo delle regole.

16. Selezionare *"Controllo Regole" (vedi figura 8.15).*

Figura 8.15

Viene eseguito il programma esterno di controllo come evidenziato dalla maschera SAS *(vedi figura 8.16).*

Figura 8.16

Al termine dell'esecuzione del programma appare una maschera con un messaggio che evidenzia l'esistenza di un edit ridondante *(vedi figura 8.17)*.

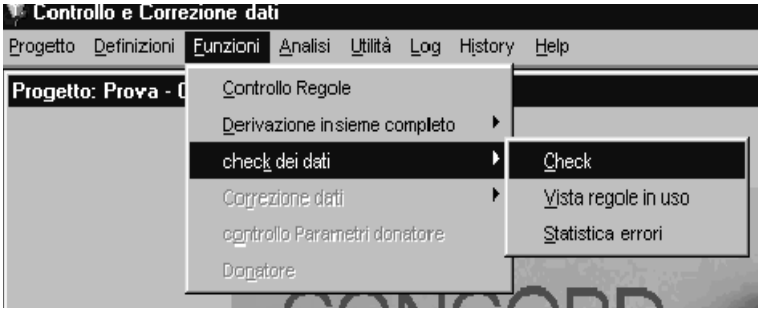
In questo caso il messaggio è solo di avvertimento e sblocca le successive funzioni; se invece fosse stato rilevato un errore nel controllo delle regole di incompatibilità le ulteriori funzioni, quali il controllo (check) dei dati e la derivazione degli edit impliciti, sarebbero stata rese impossibili fino all'eliminazione degli errori e al completamento della fase di controllo.

Figura 8.17

Un messaggio informa che il programma di controllo è terminato.

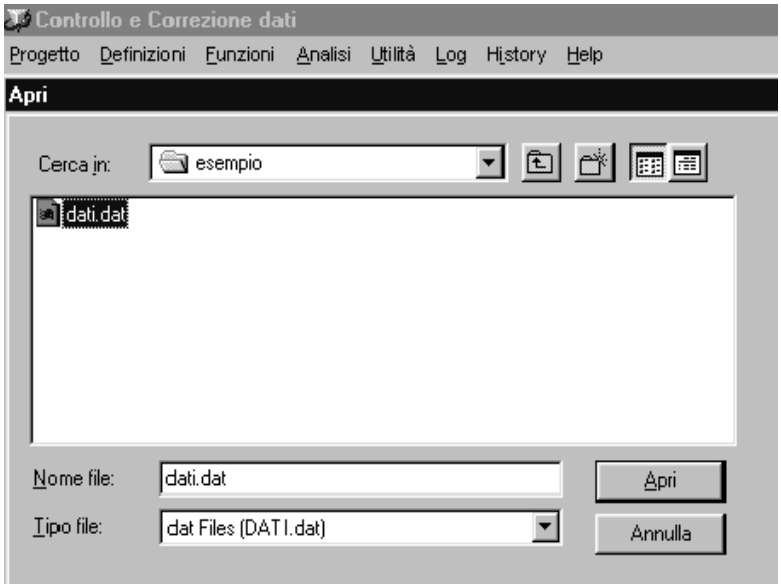
17. Selezionare *"check dei dati"* e *"Check" (vedi figura 8.18)*.

Figura 8.18



Nella maschera scegliere dalla cartella *"esempio"* il file *"dati.dat"* e poi *"Apri" (vedi figura 8.19).*

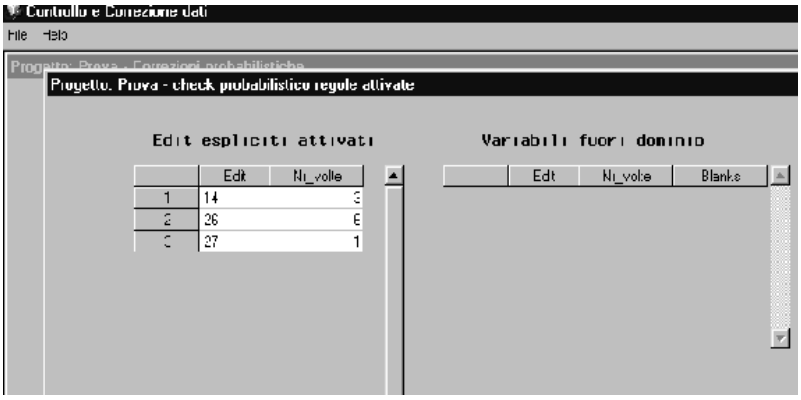
Figura 8.19



Viene eseguito il programma esterno di controllo dei dati e al termine mostrati i totali dei dati esatti e dei dati errati.

18. Da *"File"* usciamo con *"End"* e possiamo vedere la distribuzione degli edit errati ed eventuali variabili fuori dominio *(vedi figura 8.20).*

Figura 8.20



19. Prima di eseguire la correzione dei dati è necessario eseguire la derivazione degli edit impliciti. Da *"Funzioni"* scegliamo *"Derivazione insieme completo"* e *"Esegue derivazione" (vedi figura 8.21).*

Figura 8.21

La derivazione termina con il messaggio *"Derivazione regole terminata"*

Figura 8.22

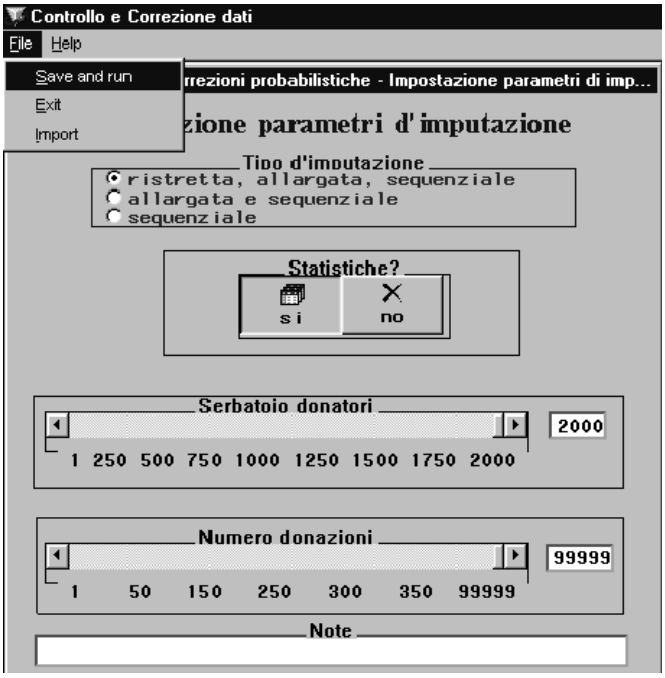
da cui usciamo con *"OK" (vedi figura 8.22).*

20. Passiamo alla fase di correzione dei dati errati selezionando da *"Funzioni"* la riga del menu a tendina *"Correzione dati"* e dal menu a bandiera *"Imputazione"* come mostrato in figura 8.23.

Figura 8.23

Viene mostrata la maschera *(vedi figura 8.24)* per la scelta dei parametri di esecuzione della fase di correzione dei dati.

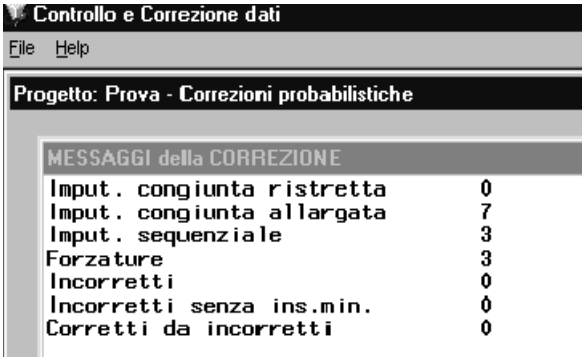
Figura 8.24



Lasciamo i parametri impostati di default e da *"File"* scegliamo *"Save and run"*.

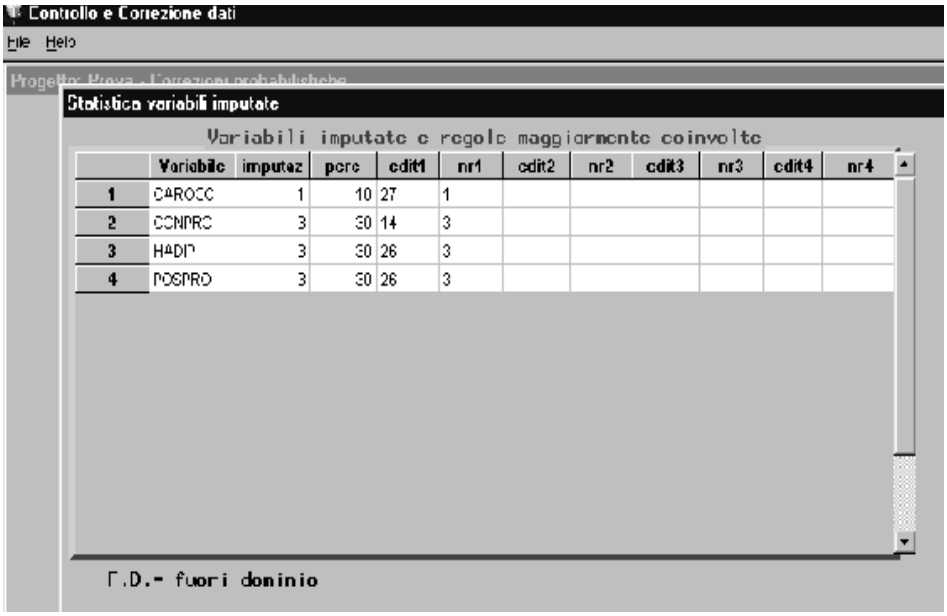
Viene eseguito il programma di correzione che termina mostrando la maschera *(vedi figura 8.25)* dei totali dei tipi di imputazione eseguita.

Figura 8.25



Usciamo con *"File"* e *"End"* e vediamo la maschera che mostra le variabili imputate e gli edit implicati nella correzione *(vedi figura 8.26)*.

Figura 8.26



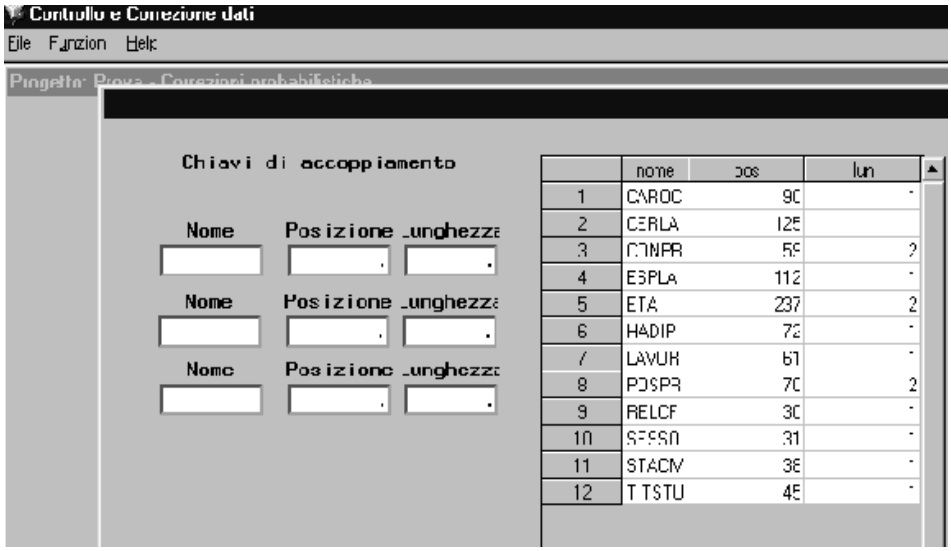
Ovviamente è possibile modificare i parametri ed eseguire di nuovo la fase d'imputazione.

Al termine e buon fine della fase di imputazione dei dati errati sono rese possibili le funzioni per le analisi dei risultati *(vedi figura 8.27)*.

Figura 8.27

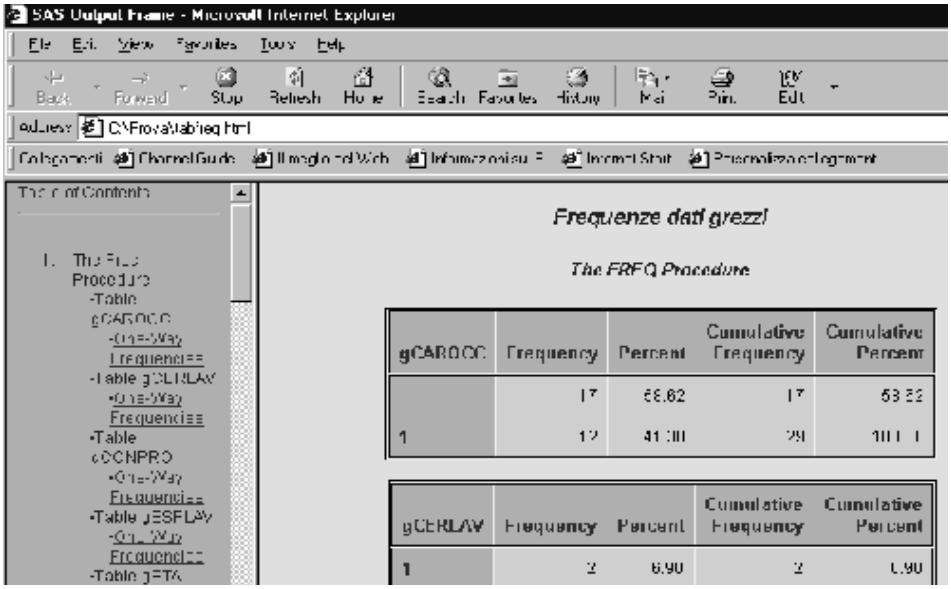
21. Da *"Analisi"* scegliere *"Tavole di verifica"* e viene mostrata la seguente maschera *(vedi figura 8.28)* per la selezione delle variabili di accoppiamento dei dati prima e dopo la correzione.

Figura 8.28



Poiché nessuna variabile individua in modo univoco il singolo record, non è possibile selezionare alcuna variabile tra quelle mostrate. Se osserviamo però i dati notiamo che dal primo al quarto carattere di ogni record è registrato un numero progressivo. Allora posizioniamo il cursore sul primo campo *"Posizione"* e scriviamo 1 e sul primo campo *"Lunghezza"* scriviamo 4 e poi da *"File"* scegliamo *"Save and run"*. Viene eseguita una procedura SAS che termina con un tabulato *(vedi figura 8.29)* dal quale è possibile verificare la distribuzione incrociata delle singole variabili imputate. Eventualmente riferirsi al file *"tabfreq.html"* generato nella cartella *"prova"*.

Figura 8.29

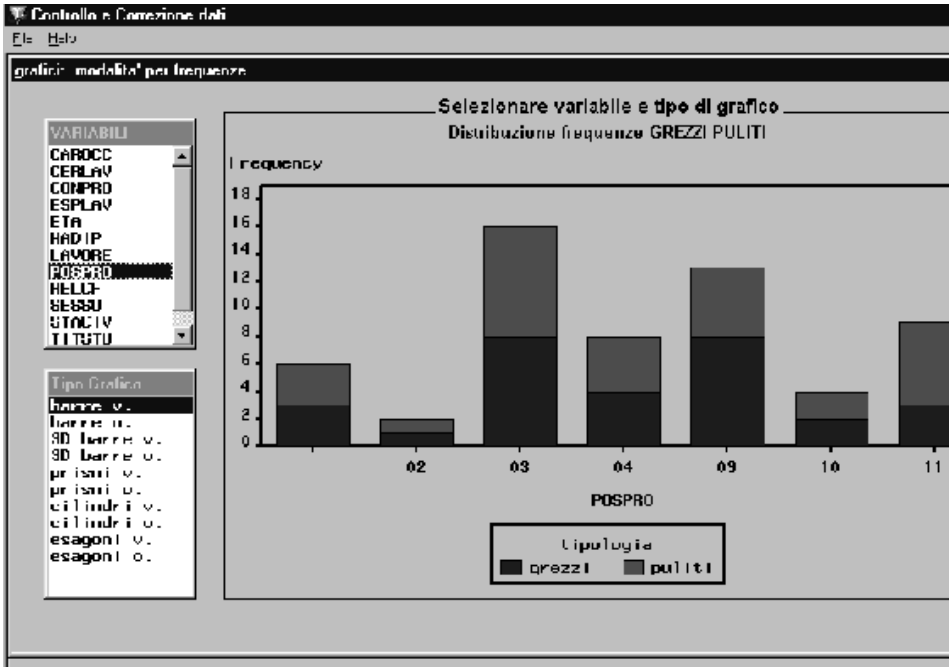


22. Da *"Analisi"* scegliendo *"Grafici"* e *"Distribuzione dati"* viene mostrata la maschera *(vedi figura 8.30)* per la scelta di una variabile e del tipo di grafico ottenendo immediatamente la distribuzione dei dati grezzi e puliti per la variabile selezionata.

Figura 8.30

Nella maschera seguente *(vedi figura 8.31)* è stata scelta la variabile *"POSPRO"* e il tipo di grafico istogramma a barre verticali.

Figura 8.31

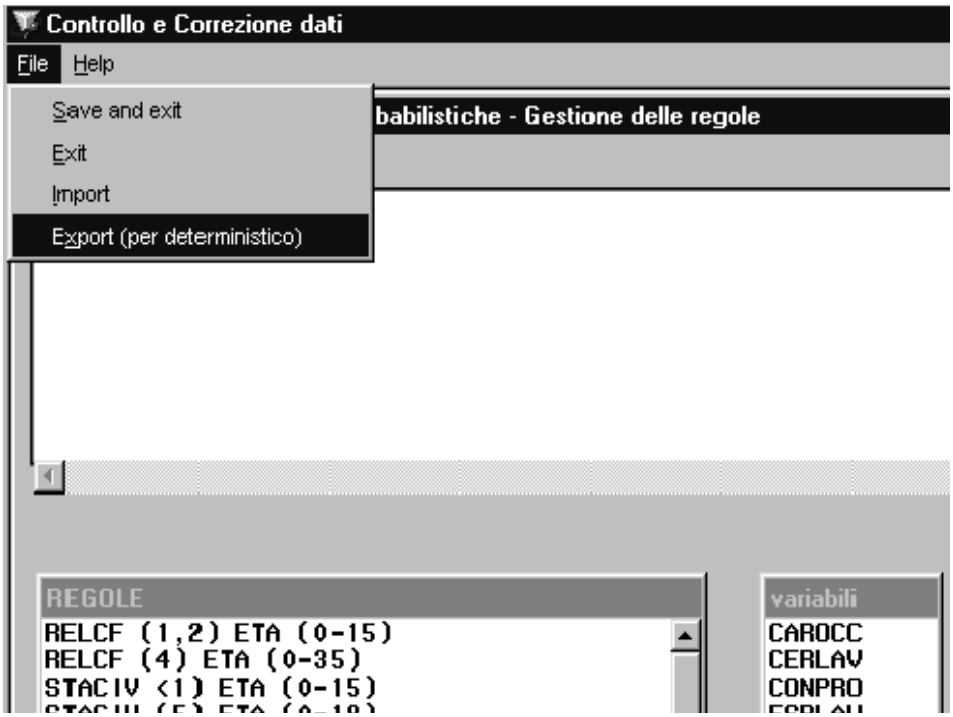


23. Da *"Analisi"* scegliendo *"Grafici"* e *"Frequenze errori" (vedi figura 8.30)* otteniamo un grafico delle frequenze degli edit espliciti.

A questo punto abbiamo praticamente visto tutte le funzioni dell'approccio probabilistico e possiamo passare alle funzioni di integrazione dall'approccio probabilistico all'approccio deterministico.

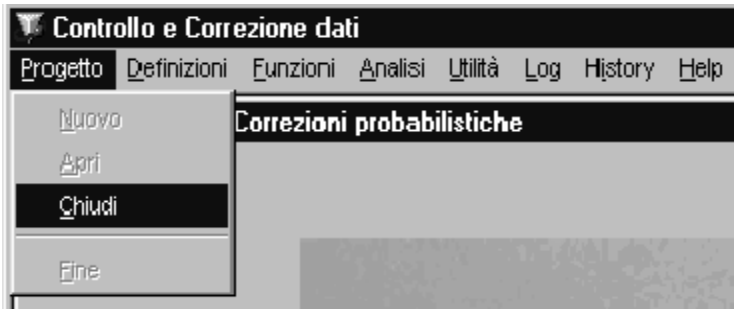
24. Tornando dalla maschera principale alla definizione delle regole di incompatibilità vediamo *(vedi figura 8.32)* che nel menu a tendina è attiva la funzione *"Export (per deterministico)"* e selezionandola facciamo eseguire un programma che scrive le regole deterministiche partendo da quelle probabilistiche.

Figura 8.32



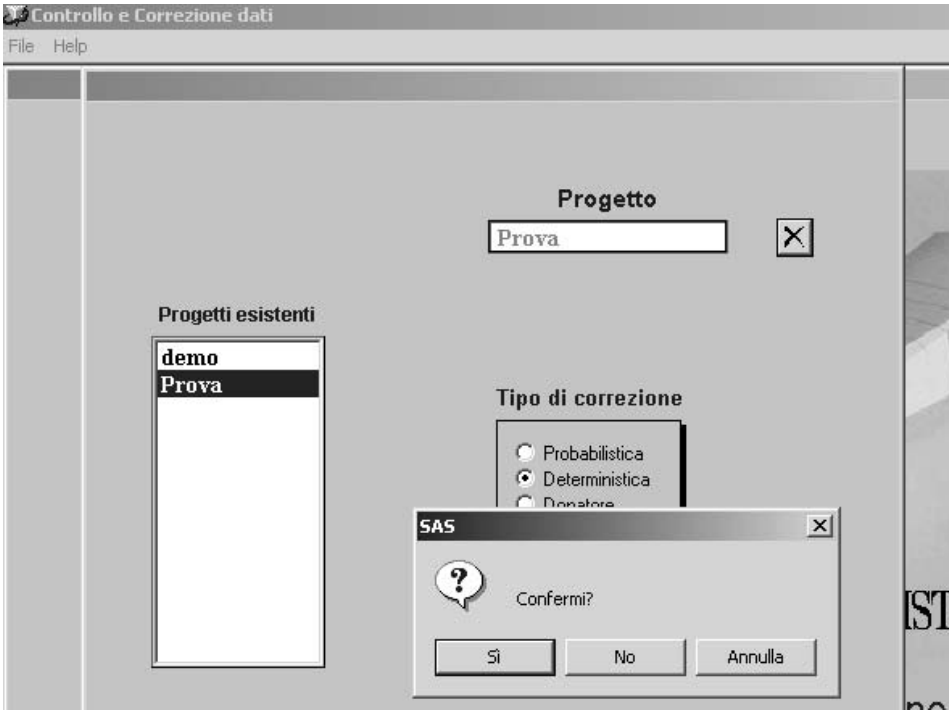
25. Chiudiamo il progetto *"prova" (vedi figura 8.33).*

Figura 8.33



26. Riapriamo il progetto *"Prova"* con tipo correzione *"deterministico" (vedi figura 8.34).*

Figura 8.34



27. Da *"Definizioni"* selezioniamo *"Variabili"*; viene mostrata la maschera per la gestione delle variabili per l'approccio deterministico con elencate le variabili create con la funzione di *"Export"* dall'approccio probabilistico.

28. Nella maschera di gestione delle variabili da *"File"* selezioniamo *"Save and exit"*. Il programma memorizza le variabili, rende possibile la scelta di ulteriori definizioni e torna alla maschera iniziale.

29. Da *"Definizioni"* selezioniamo *"Liste"*; viene mostrata la maschera per la gestione delle liste per l'approccio deterministico con elencate le variabili di lista con i relativi valori create con la funzione di *"Export"* dall'approccio probabilistico.

30. Nella maschera di gestione delle liste da *"File"* selezioniamo *"Save and exit"*. Il programma memorizza le liste e torna alla maschera iniziale.

31. Da *"Definizioni"* selezioniamo *"Regole"*; viene mostrata la maschera per la gestione delle regole di incompatibilità per l'approccio deterministico con elencate le regole create con la funzione di *"Export"* dall'approccio probabilistico.

32. Nella maschera di gestione delle regole di incompatibilità da *"File"* selezioniamo *"Save and exit"*. Il programma memorizza le regole dopo averle controllate, cosa resa evidente dallo scorrimento della riga nella listbox delle regole, e torna alla maschera iniziale.

33. Da *"Definizioni"* selezioniamo *"Correzioni"*; viene mostrata la maschera per la gestione delle regole di correzione per l'approccio deterministico con la listbox delle regole correzioni vuota.

34. Da *"File"* nella maschera di gestione delle regole di correzione selezioniamo *"Import" (vedi figura 8.35)*.

Figura 8.35

Viene mostrata la maschera per la scelta del file da importare. Andare nella cartella *"Esempio"* e selezionare il file *"FDETOUT.dat"*, come mostrato in figura 8.36 e poi *"Apri"*.

Figura 8.36

Vengono caricate tre regole di correzione nella listbox *"CORREZIONI"* della maschera di gestione delle regole di correzione dell'approccio deterministico.

35. Nella maschera di gestione delle regole di incompatibilità da *"File"* selezioniamo *"Save and exit"*. Il programma memorizza le regole di correzione dopo averle controllate, cosa si nota dallo scorrimento della riga nella listbox delle correzioni e torna alla maschera iniziale.
36. Nella maschera principale da *"Funzioni"* selezioniamo *"check dei dati"* e *"Check" (vedi figura 8.37)*.

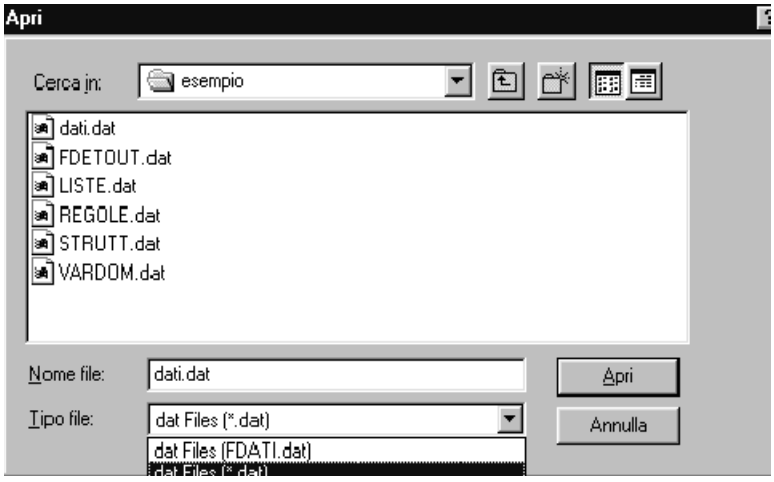
Figura 8.37

Viene mostrata la maschera *(vedi figura 8.38)* per la scelta del file di dati da controllare.

Scegliere la cartella *"Esempio"* attivare come *"Tipo file"* i *"data Files [*.dat]"*, selezionare il primo file dell'elenco *"dati.dat."* e poi *"Apri"* oppure doppio click del testo sinistro del mouse sul nome del file.

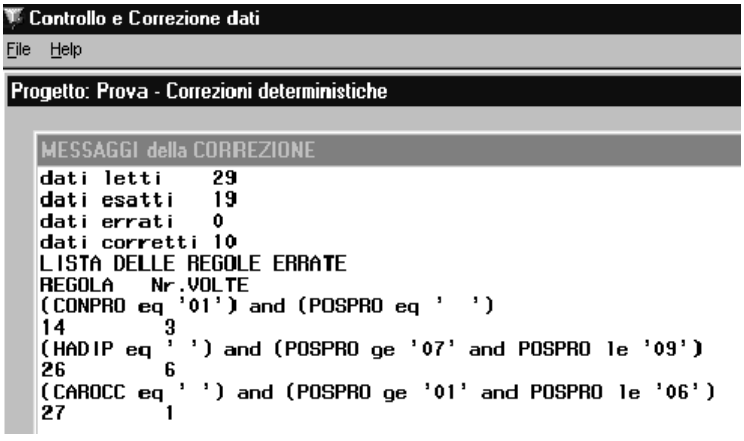
Viene eseguito il programma di controllo e correzione deterministico.

Figura 8.38



Al termine dell'esecuzione vengono mostrati *(vedi figura 8.39)* i totali dei record letti, esatti, errati e corretti e le regole attivate con relative regole di correzione.

Figura 8.39



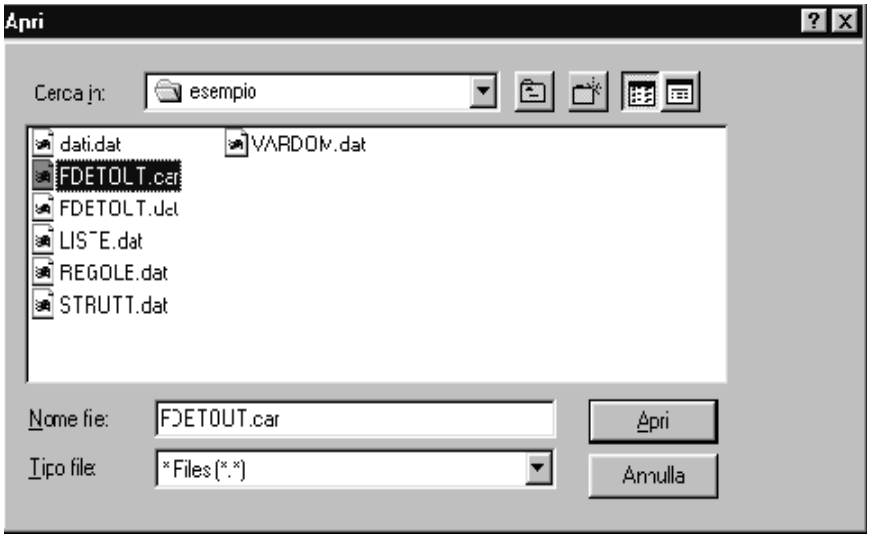
Terminata l'esecuzione del passo di controllo e correzione è possibile controllare i risultati, in modo simile all'approccio probabilistico. I file nella cartella *"Prova"* sono individuabile dal nome che inizia con *"f"* come mostrato in figura.

Possiamo passare alle correzioni dei dati tramite donatore.

37. Per sfruttare però l'integrazione tra il metodo deterministico e il sistema di correzione tramite donatore prima di chiudere il progetto e riaprirlo con correzioni tramite donatore bisogna ripetere i passi dal punto 33 in poi.

Al punto 33 selezioniamo la definizione *"Correzioni"* e al punto 34 da *"File"* eseguiamo l' *"Import"* dalla cartella "Esempio" del file *"FDETOUT.car"*, attivando come *"Tipo file" "*.Files(*.*)"*, invece del file *"FDETOUT.dat"*, come mostrato nella figura 8.40.

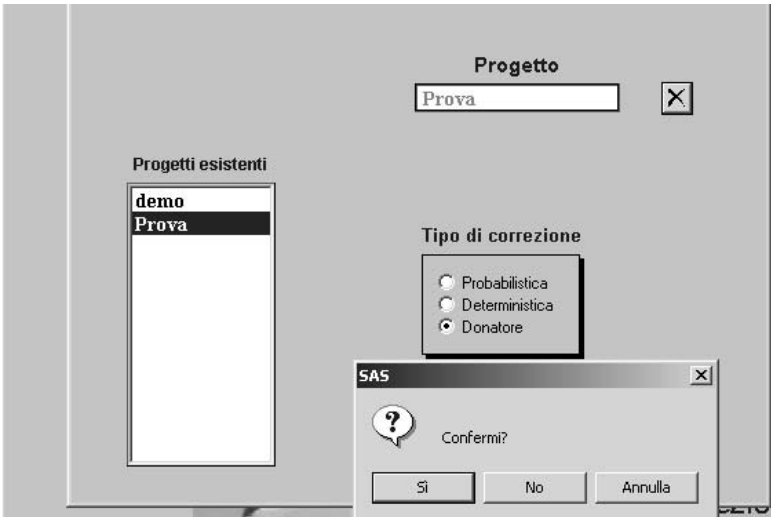
Figura 8.40



Il file *"FDETOUT.car"* contiene tre regole di correzione che impostano particolari caratteri nelle variabili che devono essere corrette al verificarsi delle regole di incompatibilità. Salvare le correzioni e ripetere la funzione di check dei dati che produrrà ovviamente gli stessi risultati ma con dati corretti in modo diverso.

38. Eseguiti i passi suddetti possiamo chiudere il progetto e riaprirlo scegliendo e confermando il tipo di correzione *"Donatore" (vedi figura 8.41)*.

Figura 8.41



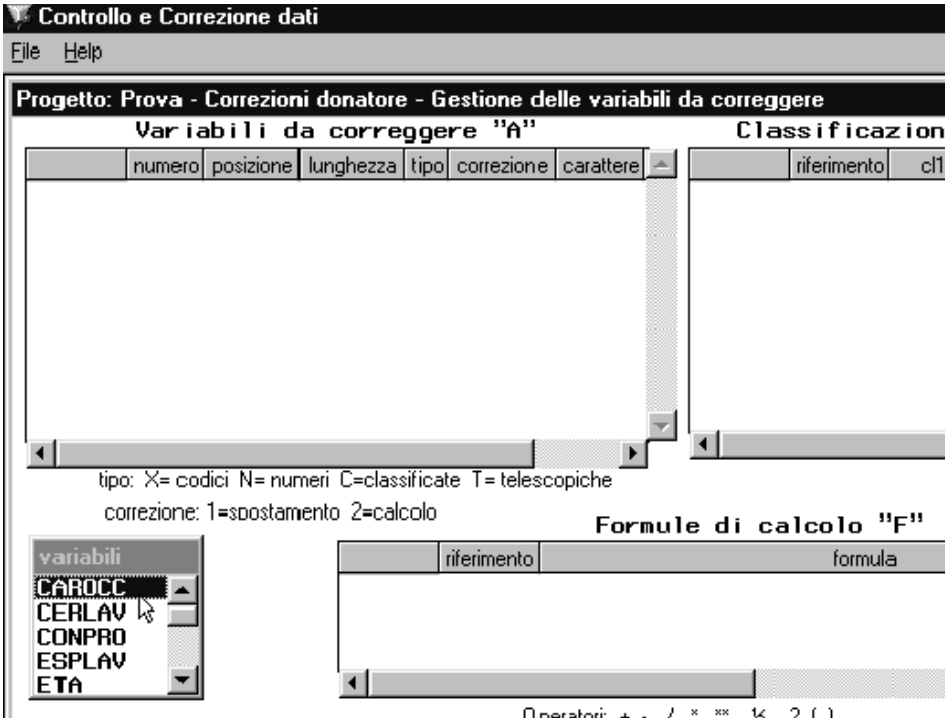
Vengono rese disponibili le definizioni dei vari tipi di variabili.

39. Nella maschera principale *(vedi figura 8.42)* dal menu a tendina scegliamo *"Definizioni"* e *"Variabili"*.

Figura 8.42

Viene mostrata la maschera *(vedi figura 8.43)* per la gestione delle variabili da correggere con evidenziate le variabili definite nell'approccio deterministico a loro volta importate dal probabilistico.

Figura 8.43

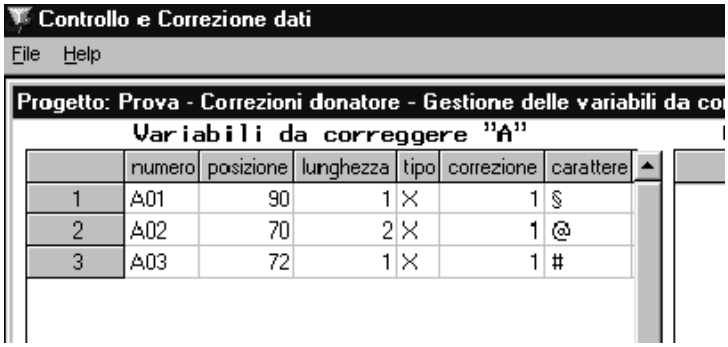


40. Da *"File"* selezioniamo *"Import"* e nella cartella *"Esempio"* selezioniamo, con doppio click del mouse o con *"Apri"*, il file *"DVARDOM.dat"*. Verranno caricate nella tabella *"Variabili da correggere (A)"* tre righe riferite alle variabili CAROCC, POSPRO e HADIP.

La stessa cosa poteva farsi con un doppio click sulla relativa variabile nella listbox *"variabili" (vedi figura 8.43)* e scrivendo il carattere relativo nel campo *"carattere"* di ogni riga dopo averla resa modificabile con doppio click.

Ad operazione conclusa la tabella appare come in figura 8.44.

Figura 8.44



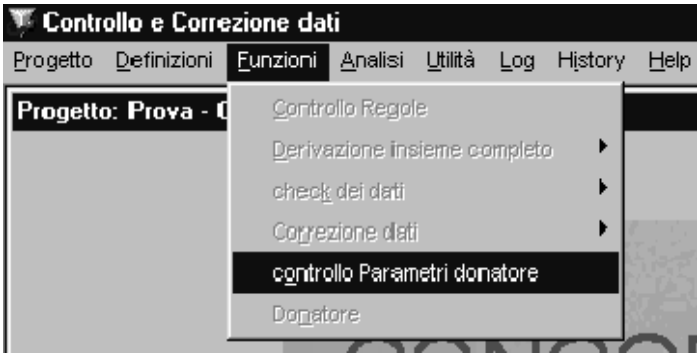
41. Salviamo le operazioni effettuate sulla tabella da *"File"* con *"Save and exit"*. Il programma torna alla maschera principale.

42. Ripetiamo la funzione di *"Import"* nella definizione delle variabili di strato e delle variabili di match selezionando sempre il file *"DVARDOM.dat"* dalla cartella *"Esempio"*.

43. Definite le tre variabili suddette possiamo procedere al *"Controllo parametri donatore"* selezionando l'apposita funzione, ora attiva.

Viene eseguito il controllo dei parametri e resa eseguibile l'esecuzione della funzione *"Donatore"* per la correzione dei dati.

Figura 8.45



44. Selezioniamo la funzione *"Donatore"* e viene mostrata una maschera nella quale è necessario completare i riferimenti al file dei dati esatti e al file dei dati da correggere.

Per il file dei dati esatti o donatori scegliamo dalla cartella *"Prova"* il file *"fesatti.dat"*.

Per il file dei dati da correggere scegliamo dalla cartella *"Prova"* il file *"fcorrett.dat"*. Ambedue i file provengono dal passaggio deterministico precedente.

45. Scelti i due file è possibile poi tramite *"Esegui"* eseguire i vari passi di correzione al termine dei quali verrà mostrata la lista con i dettagli delle correzioni effettuate.

Appendice

La metodologia Fellegi-Holt

Tre sono i criteri fondamentali per l'imputazione delle variabili qualitative alla base della metodologia proposta da Fellegi e Holt[9]:

1. in ogni record i dati devono soddisfare tutte le regole di validità e incompatibilità, cambiando il meno possibile il valore dei campi;
2. le regole di imputazione devono essere derivate dalle regole di controllo, senza esplicita specificazione;
3. le distribuzioni di frequenza marginali e congiunte devono essere mantenute il più possibile.

EDIT IN FORMA NORMALE

Distinguiamo gli edit logici, riguardanti le variabili qualitative, dagli edit aritmetici, riguardanti le variabili quantitative.

DEFINIZIONE: un *edit logico* esprime una condizione di inaccettabilità su una data combinazione di valori di due o più variabili

Un edit può essere formalizzato come l'applicazione di una funzione f a sottoinsiemi dei domini di n variabili:

$$f(A_1^0, A_2^0, \ldots, A_n^0)$$

[9] Quanto segue è una sintesi dell'articolo "A Systematic Approach to Automatic Edit and Imputation" di I.Fellegi e D.Holt pubblicato sul Journal of the American Statistical Association (marzo 1976)

dove:

A_i^0 : sottoinsieme del dominio della variabile i-esima

f : funzione logica che connette i vari A_i^0 mediante gli operatori logici di intersezione ($\cap$) e unione ($\cup$)

Un record **a** è errato se:

$$\mathbf{a} \in f(A_1^0, A_2^0, \ldots, A_n^0)$$

Applicando ripetutamente alla f la legge distributiva otteniamo:

$$f(A_1^0, A_2^0, \ldots, A_n^0) =$$

$$(A^1_{i_1} \cap A^1_{i_2} \cap \ldots \cap A^1_{m_i}) \cup (A^2_{j_1} \cap A^2_{j_1} \cap \ldots \cap A^2_{m_j}) \cup \ldots \cup (A^r_{k_1} \cap A^r_{k_2} \cap \ldots \cap A^r_{m_k})$$

Possiamo dire che un record è errato se appartiene ad almeno uno dei termini a secondo membro. Definiamo come *"edit in forma normale"* ognuno di tali termini.

DEFINIZIONE: un *edit in forma normale* è un edit logico in cui l'unico operatore ammesso è quello di intersezione

In simboli:

$$\bigcap_{i \in s} A_i^*$$

Ogni edit logico, di qualsiasi forma, può sempre essere tradotto in una serie di edit in forma normale. Consideriamo, ad esempio, la seguente regola (di compatibilità):

"Se una persona ha età inferiore a 16 anni, oppure frequenta una scuola elementare, allora non può essere capo-famiglia, ed il suo stato civile deve essere celibe o nubile"

Questa regola può essere convertita in una serie di edit in forma normale attraverso i seguenti passi[10]:

formalizzazione:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})] \rightarrow [\ (\neg\ \text{Capo-famiglia}) \cap (\text{Celibe/Nubile})\ ]$$

2. *traduzione in regola di incompatibilità*:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})]\ \cap \neg\ [\ (\neg\ \text{Capo-famiglia}) \cap \text{Celibe/Nubile})\ ] = \text{errore}$$

3. *semplificazione*:

$$[\ (\text{Età} < 16) \cup (\text{Scuola Elementare})]\ \cap [\ (\text{Capo-famiglia}) \cup (\neg\ \text{Celibe/Nubile})\ ] = \text{errore}$$

4. *applicazione della legge distributiva*:

$$[\ (\text{Età} < 16) \cap (\text{Capo-famiglia})] \cup$$
$$[\ (\text{Età} < 16) \cap (\neg\ \text{Celibe/Nubile})] \cup$$
$$[\ (\text{Scuola Elementare}) \cap (\text{Capo-famiglia})] \cup$$
$$[\ (\text{Scuola Elementare}) \cap (\neg\ \text{Celibe/Nubile})] = \text{errore}$$

I quattro termini nell'ultima espressione sono altrettanti edit in forma normale.

L'INSIEME COMPLETO DEGLI EDIT

DEFINIZIONE: gli edit in forma normale specificati direttamente dallo statistico sono detti *edit espliciti*.

Un record che non attiva alcun edit esplicito si dice corretto, e non necessita di alcuna modifica. Al contrario, un record che attiva almeno un edit esplicito si dice errato, e necessita della modifica di almeno una variabile.

[10] Oltre agli operatori di intersezione e unione, facciamo uso anche di quelli di negazione (¬) e implicazione (→)

Mentre gli edit espliciti sono necessari e sufficienti per determinare la correttezza di un record, essi non sono sufficienti per una sua ottimale correzione.

DEFINIZIONE: chiamiamo *edit implicito* un edit logicamente contenuto negli edit espliciti.

La funzione degli edit impliciti, considerati congiuntamente con gli edit espliciti, è quella di permettere la correzione ottimale di un record errato.

DEFINIZIONE: l'*insieme completo* degli edit è dato dall'unione degli edit espliciti e di quelli impliciti.

Per eseguire in modo ottimale il passo di scelta delle variabili da imputare, e di determinazione del range di valori imputabili, è necessario preventivamente generare l'insieme completo di edit.

Consideriamo il seguente esempio.

Supponiamo che un record contenga tre variabili, di cui siano definiti i seguenti domini:

VARIABILI	*DOMINI*
Età (ETA)	0-14, 15-99
Stato civile (STACIV)	celibe, coniugato, separato, divorziato, vedovo
Relazione con il capo famiglia (RELCF)	capofamiglia, coniuge, altro

Siano stati definiti i seguenti edit in forma normale espliciti, esprimenti condizioni di incompatibilità:

I. (ETA = 0-14) ∩ (STACIV = coniugato, separato,divorziato,vedovo)

II. (STACIV = celibe, separato, divorziato, vedovo) ∩ (RELCF = coniuge)

Possiamo riscriverli come condizioni di compatibilità nel seguente modo:

(ETA = 0-14) → (STACIV = celibe)

(STACIV = celibe, separato, divorziato, vedovo) → (RELCF ≠ coniuge)

Poiché la conseguenza della prima implicazione è contenuta nella premessa della seconda, possiamo derivare che:

(ETA = 0-14) → (RELCF ≠ coniuge)

relazione che, opportunamente ritradotta in forma normale, diventa:

III. (ETA = 0-14) ∩ (RELCF = coniuge)

Questo terzo edit era implicitamente contenuto nei primi due.

Supponiamo ora di considerare il seguente record:

(ETA = 0-14) ∩ (STACIV = coniugato) ∩ (RELCF = coniuge)

Questo record attiva gli edit I e III.

Per correggere il record, ricerchiamo l'insieme minimo di variabili che *copra* tutti gli edit attivati (espliciti e impliciti) dal record in questione. Nel nostro caso verifichiamo che la variabile ETA è presente sia nel primo che nel terzo edit attivato. Per disattivare tali edit è sufficiente assegnare a ETA un valore interno all'*intersezione dei complementi* dei valori che compaiono negli edit attivati o attivabili:

$$(\neg\ 0\text{-}14) \cap\ (\neg\ 0\text{-}14) = 15\text{-}99$$

Assegnando il valore 15-99 alla variabile ETA, il record può dirsi corretto, in quanto non attiva alcun edit: nel far ciò abbiamo tenuto conto del principio del minimo cambiamento, in quanto abbiamo modificato una sola variabile.

Se in questo processo di ricerca dell'insieme minimale di variabili da imputare non avessimo tenuto conto dell'edit implicito, avremmo considerato il solo edit I: per disattivarlo, avremmo potuto scegliere di imputare sia ETA che STACIV. Se avessimo scelto STACIV, che compare anche nell'edit II, avremmo constatato che l'intersezione del complemento dei relativi valori è l'insieme vuoto ∅:

$$\neg\ (\text{coniugato, separato, divorziato, vedovo}) \cap \neg\ (\text{celibe, separato, divorziato, vedovo}) =$$
$$= \text{celibe} \cap \text{coniugato} = \varnothing$$

L'impossibilità di trovare dei valori imputabili a STACIV tali da correggere il record deriva dal fatto che STACIV non è contenuto nell'edit III, implicito, attivato dai valori delle variabili ETA e RELCF. La conseguenza di carattere generale è che *la non considerazione degli edit impliciti non permette di definire sempre insiemi minimi di variabili da imputare che siano in grado di riportare il record in una situazione di correttezza.*

LEMMA: dati s edit e_r e n variabili, per ogni arbitraria variabile i, un edit

$$\mathbf{e_i^* : \bigcap_{j=1}^{n} A_j^*}$$ si dice generato dagli s edit se e solo se

$$\begin{cases} A_j^* = \bigcap_{r \in s} A_j^r & j = 1,2,\ldots,n \quad i \neq j \\ A_i^* = \bigcup_{r \in s} A_i^r \end{cases}$$

In altri termini, fissata una variabile i (detta *generante*), il corrispondente A_i^* sarà ottenuto come *unione* degli A_i^r, mentre ogni altro A_j^* sarà ottenuto come *intersezione* degli A_j^r.

DEFINIZIONE: Un edit generato si dice *edit implicito essenzialmente nuovo* se e solo se:

1. A_i^* coincide col dominio della variabile i;
2. A_i^r ogni è non vuoto ed è un sottoinsieme proprio del dominio della variabile i;

Consideriamo il seguente esempio. Siano dati gli edit:

I. (ETA = 0-14) ∩ (RELCF = qualsiasi) ∩ (STACIV ≠ celibe)

II. (ETA=qualsiasi) ∩ (RELCF = coniuge) ∩ (STACIV = celibe, separato, divorziato, vedovo)

Se fissiamo ETA come variabile generante otteniamo:

(ETA=qualsiasi) ∩ (RELCF = coniuge) ∩ (STACIV = separato, divorziato, vedovo)

che è ridondante rispetto al secondo edit.

Fissando invece RELCF otteniamo:

(ETA=0-14) ∩ (RELCF = qualsiasi) ∩ (STACIV = separato, divorziato, vedovo)

che è ridondante rispetto al primo edit.

Infine, scegliendo STACIV come variabile generante:

(ETA=0-14) ∩ (RELCF = coniuge) ∩ (STACIV = qualsiasi)

che è un edit implicito essenzialmente nuovo.

DEFINIZIONE: Un edit generato da due o più edit tra loro contraddittori (inconsistenti) è detto *edit degenere*

Consideriamo il seguente esempio:

I. (ETA = 0-14) ∩ (STACIV ≠ celibe)
II. (ETA = 15-99) ∩ (STACIV ≠ celibe)

Assumendo ETA come campo generante, otteniamo l'edit esplicito

III. (ETA = qualsiasi valore) ∩ (STACIV ≠ celibe) = (STACIV ≠ celibe)

che ci dice che sono errati tutti i valori di STACIV diversi da celibe, il che chiaramente contraddice la definizione del dominio della variabile STACIV. L'edit III è un edit degenere, ed in quanto tale può essere generato solo da edit tra loro contraddittori.

I seguenti teoremi e corollari assicurano che, *avendo a disposizione l'insieme completo di edit, un qualsiasi record errato è sempre correggibile, e lo è in modo ottimale.*

Sia $\mathbf{\Omega}$ l'insieme completo di edit, e sia $\mathbf{\Omega}_k$ un sottoinsieme tale da coinvolgere le prime k variabili (con l'esclusione, quindi, di tutti gli edit in cui compaiano le variabili k+1, k+2, ... , n).

TEOREMA 1: se gli a_i^0 sono possibili valori per le prime k-1 variabili, e se questi valori soddisfano tutti gli edit in $\mathbf{\Omega}_{k-1}$, allora esiste un qualche valore a_k^0 tale da soddisfare tutti gli edit in $\mathbf{\Omega}_k$.

La ripetuta applicazione del teorema 1 permette di conseguire il seguente

COROLLARIO 1:

se un record ha n variabili, di cui le prime k-1 hanno valori a_i^0 (i=1,2,...,k-1) tali che tutti gli edit in $\mathbf{\Omega}_{k-1}$ sono soddisfatti, allora esistono valori a_i^0 (i=k,k+1,...,n) tali da soddisfare tutti gli edit in $\mathbf{\Omega}$.

Ed inoltre:

COROLLARIO 2:
se un record ha n variabili, di cui un sottoinsieme s ha la proprietà che almeno uno dei valori a_i ($i \in s$) compare in ogni edit attivato dal record, allora esistono dei valori a_i^0 ($i \in s$) tali che, assieme agli a_i ($i \notin s$) fanno si che il record soddisfi tutti gli edit.

METODI DI IMPUTAZIONE

La metodologia prevede, per ogni record errato:

1. l'identificazione dell'*insieme minimo di variabili da modificare*;
2. per ogni variabile rientrante nell'insieme minimo, la *determinazione dell'insieme di valori attribuibili,* e *imputazione* di uno tra questi.

Per quanto riguarda il punto 1, ricordiamo che l'insieme minimo di variabili da imputare è costituito da quell'insieme di variabili che *"coprono"* tutti gli edit attivati dal record e che risulta essere di dimensione minima.

Per quanto concerne il punto 2, sono proposti due metodi, entrambi di tipo *hot deck*, consistenti nell'imputare in una variabile del record corrente (ricevente) il valore della stessa variabile in un record (donatore) scelto tra quelli esatti. I metodi in questione sono:

- metodo dell'imputazione sequenziale;
- metodo dell'imputazione congiunta.

METODO 1: IMPUTAZIONE SEQUENZIALE

Consideriamo un record errato di cui sia già stato identificato un insieme minimo di k variabili da imputare. Il metodo consiste nell'imputare dapprima la k-esima variabile, e poi, sequenzialmente, le variabili k-1,k-2,...,1.

Consideriamo tutti gli M edit in cui

- è presente la variabile k;
- non sono presenti le variabili 1,2,...,k-1.

Tra questi, consideriamo solo gli M' edit in cui non sono presenti gli edit sicuramente disattivati dai valori correnti delle variabili k+1, k+2, ... , n: gli M' edit

sono quelli che possono essere attivati o meno in funzione dei valori della sola variabile k. Se vogliamo che il record soddisfi tali edit, il valore da assegnare alla variabile k deve soddisfare la condizione:

$$a_k^0 \in \bigcap_{r=1}^{M'} \overline{A_r^k}$$

cioè deve appartenere all'insieme intersezione dei complementi dei valori indicati per la variabile k in tutti gli M' edit: tale insieme non è mai vuoto per il teorema 1.

Lo stesso procedimento viene iterato per le variabili k-1, k-2, ...1, fino all'esaurimento dell'insieme minimo di variabili da imputare.

Consideriamo il seguente esempio, con 5 variabili:

VARIABILI	***DOMINI***
SESSO	maschio, femmina
ETA' (ETA)	0-14,15-16,17-99
STATO CIVILE (STACIV)	celibe, coniugato, separato, divorziato, vedovo
RELAZIONE COL CAPOFAMIGLIA (RELCF)	moglie, marito, figlio, altro
LIVELLO D'ISTRUZIONE (ISTRUZ)	nessuno, elementare, secondario, post-secondario

L'insieme (completo) degli edit è il seguente:

e_1 : (SESSO=maschio) $\cap$ (RELCF=moglie)
e_2 : (ETA=0-14) $\cap$ (STACIV≠celibe)
e_3 : (STACIV≠coniugato) $\cap$ (RELCF=moglie,marito)
e_4 : (ETA=0-14) $\cap$ (RELCF=moglie,marito)
e_5 : (ETA=0-16) $\cap$ (ISTRUZ=post-secondaria)

Sia dato il seguente record:

VARIABILE	***VALORE***
SESSO	maschio
ETA	12
STACIV	coniugato
RELCF	moglie
ISTRUZ	elementare

Il record attiva gli edit e_1 , e_2 , e_4 . Nessuna singola variabile *"copre"* i tre edit. Tre coppie di variabili coprono gli edit attivati: (SESSO, ETA), (ETA, RELCF) e (STACIV, RELCF). Supponiamo di scegliere la coppia (SESSO, ETA): la dimensione s dell'insieme è pari a 2.

Sia ETA la variabile k-esima (k=2). Consideriamo tutti gli edit che contengono ETA ma non SESSO (la variabile k-1=1):

e_2 : (ETA=0-14) ∩ (STACIV≠celibe)

e_4 : (ETA=0-14) ∩ (RELCF=moglie,marito)

e_5 : (ETA=0-16) ∩ (ISTRUZ=post-secondaria)

L'edit e_5 è sempre soddisfatto per qualsiasi valore di ETA dal momento che nel record il valore di ISTRUZ è *"elementare"*. Per calcolare i valori imputabili ad ETA dobbiamo quindi considerare solo A_2^2 e A_2^4 :

$$a_2^* \in \overline{A_2^2} \cap \overline{A_2^4} \equiv \overline{(0-14)} \cap \overline{(0-14)} = (15\text{-}99)$$

cercheremo quindi un record donatore con un valore di ETA compreso tra 15 e 99: supponiamo 22.

Passiamo ora variabile SESSO (k-1=1). Solo l'edit la contiene, quindi:

$$a_1^* \in \overline{A_1^1} \equiv \overline{\text{maschio}} = \text{femmina}$$

Essendo unico, il valore *"femmina"* è direttamente imputato alla variabile SESSO. Il record corretto sarà quindi il seguente:

VARIABILE	***VALORE***
SESSO	femmina
ETA	22
STACIV	coniugato
RELCF	moglie
ISTRUZ	elementare

METODO 2: IMPUTAZIONE CONGIUNTA

Per un dato record errato siano state definite le k variabili da imputare. Si considerino gli M" edit con le k variabili

$$e_r : \bigcap_{i=1}^{n} A_i^r \quad (r=1,2,...,M")$$

dove $a_i^0 \in A_i^r$ (i=k+1,k+2,...,n). Sono gli edit in cui sono presenti le k variabili, e dove le variabili k+1, k+2, ..., n hanno nel record valori interni agli A_i^r: sono cioè gli edit attivabili o meno in funzione dei valori che si danno alle k variabili.

Si considerino gli insiemi

$$A_i^* = \bigcap_{r=1}^{M"} A_i^r \quad (i=k+1, k+2, ... ,n)$$

Se scegliamo un qualsiasi record, tra quelli esatti, i cui valori delle variabili k+1, k+2, ... ,n siano interni agli insiemi così definiti, i valori di tale record nelle variabili 1,2,...,k sono attribuibili in blocco al record errato corrente, in quanto costituiscono una combinazione che sicuramente garantisce che tutti gli M" edit siano soddisfatti (cioè disattivati). Per tale motivo non c'è alcun bisogno di calcolare l'insieme dei valori attribuibili alle k variabili dell'insieme minimo.

Riprendiamo in considerazione l'esempio visto per l'imputazione sequenziale: siano ancora SESSO ed ETA le variabili dell'insieme minimo: queste due variabili sono presenti negli edit e_1 , e_2 , e_4 ed e_5 . Quest'ultimo è soddisfatto comunque per il valore di ISTRUZ. Restano:

e_1 : (SESSO=maschio) $\cap$ (RELCF=moglie)
e_2 : (ETA=0-14) $\cap$ (STACIV$\neq$celibe)
e_4 : (ETA=0-14) $\cap$ (RELCF=moglie,marito)

È questo l'insieme M" di edit. Si determinano gli insiemi di valori per le variabili k+1, k+2, ..., n, cioè per STACIV (3), RELCF (4) e ISTRUZ (5):

A_3^* = coniugato, separato, divorziato, vedovo

A_4^* = moglie $\cap$ (moglie, marito) = moglie

A_5^* = qualsiasi valore

A questo punto, tra i record esatti viene ricercato un donatore che abbia i valori di STACIV e RELCF interni agli insiemi così determinati, ed i relativi valori di SESSO ed ETA vengono attribuiti al record errato corrente.

Bibliografia

ABBATE C., BOVE G., CRESCENZI F. (1992) - *"Metodi statistici multivariati per la ricostruzione dell'informazione mancante"*, in *Avanzamenti metodologici e statistiche ufficiali*, *Atti delle prime giornate di studio SIS-ISTAT*, Roma 13-14 dicembre 1992

ABBATE C., GIOMMI A. (1993) - *"Metodi di ponderazione e di correzione di dati elementari"*, *Atti del Convegno "La qualità dell'informazione statistica e la qualità industriale"*, SIS-ISTAT-AICQ, Roma 10 maggio 1991

ABBATE C., SCHIEVANO R. (1993) - *"Efficacia dell'imputazione da donatore con distanza minima"*, *Atti del Convegno SIS*, Sanremo 1993

ABBATE C. (1996) - *"La completezza delle informazioni e l'imputazione da donatore con distanza mista minima"*, Quaderni di Ricerca ISTAT 1996

BARCAROLI G. (1992) - *"An integrated system for edit and imputation of data in the Italian Statistical Insitute"*, *Survey and Statistical Computing*, pp.167-177

BARCAROLI G. (1993) - *"Un approccio logico formale al problema del controllo e della correzione dei dati statistici"*, *Quaderni di ricerca ISTAT* n.9/1993

BARCAROLI G., CECCARELLI C., LUZI O. (1995) - *"An edit and imputation system of quantitative variables based on macroediting techniques"*, *Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol (UK), 1-4 Aprile 1995, pp.12-17

BARCAROLI G., CECCARELLI C., LUZI O., MANZARI A., RICCINI MARGARUCCI E., SILVESTRI F. (1995) - *"The Methodology of Editing and Imputation of Qualitative Variables implemented in SCIA"*, *Documento interno ISTAT*.

BARCAROLI G., LUZI O. (1995) - *"Sistema generalizzato per l'editing e l'imputazione di variabili quantitative (GEIS)"*, *Quaderni di ricerca ISTAT* n.1/1995 (nuova serie)

BARCAROLI G. (1998) - *"La correzione probabilistica dei dati: il trattamento congiunto degli errori di rilevazione casuali e sistematici mediante l'applicazione del teorema di Bayes alla metodologia Fellegi-Holt"*, *Statistica Applicata vol. 10 n.2/1998*

BARCAROLI G., D'AURIZIO L., LUZI O.,MANZARI A., PALLARA A. (1999) - *"Metodi e software per il controllo e la correzione dei dati"*, *Documenti ISTAT* n.1/1999

COTTON C. (1991) - *"Functional description of the generalized edit and imputation system"*, Statistics Canada, Business Survey Methods Division, July 25

DAVILA H. E. (1992) - "The Hidiroglou-Berthelot Method", in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992

ENGSTROM P., ANGSVED C. (1994) - *"A description of a geographical Macromacro-editing application"*, *Statistical Commission and Economic Commission for Europe-Conference of European Statisticians*, Cork, Ireland, 17-20 October 1994

FELLEGI I.P., HOLT D. (1976) - *"A systematic approach to edit and imputation"*, *Journal of the American Statistical Association,* vol.71, pp.17-35

FORD B.L. (1983) - *"An overview of Hot-deck procedures"* in *Incomplete data in sample survey*, vol.1, pg. 191, Academic Press, New York

GARCIA RUBIO E., VILLAN CRIADO I. (1988) - *"Sistema DIA, Sistema de deteccion y imputacion automatica de errores para datos cualitativos"*, Istituto Nacional de Estadistica, Madrid, 1988

GRANQUIST L. (1992a) - *"A Review of methods for rationalizing the editing of survey data"*, in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992

GRANQUIST L. (1992b) - *"The Aggregate Method"*, in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992

GRANQUIST L. (1992c) - *"The Top-Down Method"*, in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992

GRANQUIST L. (1992d) - *"On the need for generalized numeric and imputation system"*, in *Statistical Data Editing Methods and Techniques,* United Nations, Vol. I, February, 1992

GRANQUIST L. (1995) - *"Improving the traditional editing process"*, in *Business Survey Methods*, John Wiley and sons

GRANQUIST L.(1995) - *"An overview of methods of evaluating editing processes"*, *Conference of European Statisticians* (Athens, Greece, 6-9 November), Working Paper n. 3

GRENLESS J.S., REECE W.S., ZIESCHANG K.D. (1982) - *"Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed"*, *Journal of the American Statistical Association*, 77, pp 251-261

HAWKINS D. M. (1974) - *"The Detection of Errors in Multivariate Data Using Principal Components"*, *Journal of the American Statistical Association*, Vol. 69. No 346

HIDIROGLOU M.A., BERTHELOT J.M.(1986) - *"Statistical Editing and Imputation for Periodic Business Surveys"*, *Survey Methodology*, June 1986, vol.12, N.1, pp.73-83

KALTON G. and., KASPRZIK D. (1986) - *"The treatment of missing survey data", in "Survey methodology"*, 12, 1, Statistics Canada

KOVAR J.G., MacMILLIAN J.H., WHITRIDGE P. (1988) - *"Overview and strategy for the generalized edit and imputation system"*, Statistics Canada, Methodology Branch, April 1988(updated February 1991)

KOVAR J.G., WHITRIDGE P. (1995) - *"Imputation of business survey data"*, in *Business Survey Methods*, John Wiley and sons

LATOUCHE M., BERTHELOT J.M. (1992) - *"Use of Score Function to Prioritize and Limit Recontacts in Editing Business Surveys"*, *Journal of Official Statistics*, Vol.8, No.3, Part II.

LUZI O., CECCARELLI C. (1997) - *"Le componenti principali nello studio dell'editing multivariato"*, *Atti della XXXV Riunione Scientifica della Società Italiana di Economia, Demografia e Statistica*

LUZI O. (1996) - *"Applicabilità ed impatto potenziale dei metodi per l'editing di dati quantitativi basati sugli approcci del Macroediting e dell'Editing Selettivo"*, in corso di pubblicazione sulla collana *Contributi ISTAT*

MASSELLI M., SIGNORE M., PANIZON F. (1992) - *"Il sistema di controllo della qualità dei dati"* in *Manuale di tecniche di indagine*, Vol.6 ISTAT

NORDBOTTEN S.(1995) - "Editing statistical records by neural networks", *Conference of European Statisticians*, Athens, Greece, 6-9 November, Working Paper n. 40

WINKLER W.E. (1994) - *"SPEER Edit System"*, computer system and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington D.C., USA.

Collana - TECNICHE E STRUMENTI

Volumi pubblicati

1 - 2004 CONCORD V. 1.0 - Controllo e correzione dei dati
Manuale utenti e aspetti metodologici

dati forniti su floppy
dati forniti su cd-rom