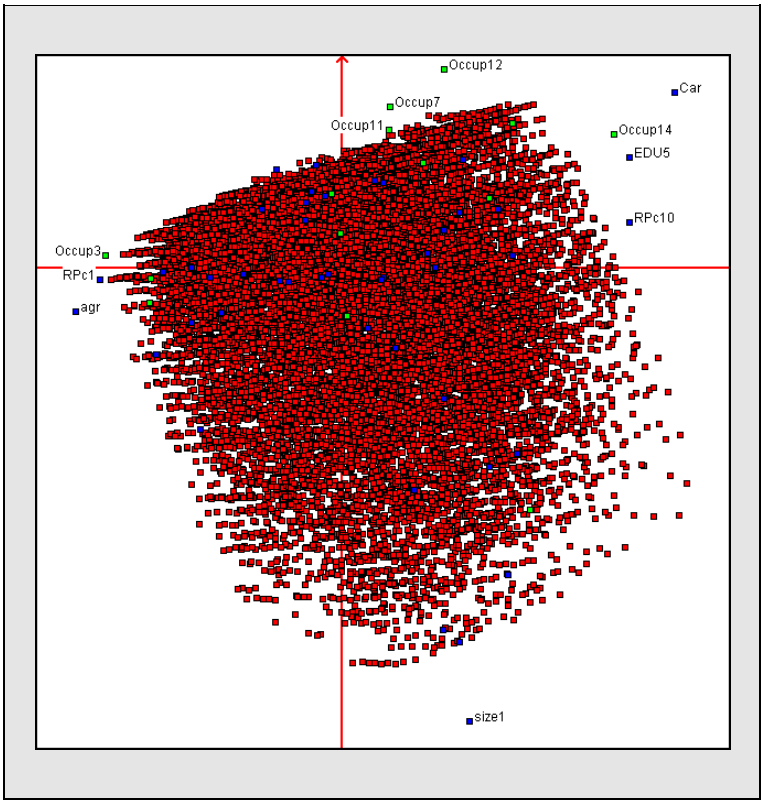
Silvio Griguolo

ADDATI per Windows

Un pacchetto per l'analisi esplorativa dei dati

(versione 6.0 – Febbraio 2008)

Guida all'uso



Università IUAV di Venezia - Dipartimento di Pianificazione

SILVIO GRIGUOLO

ADDATI

Un pacchetto per l'analisi esplorativa dei dati

Guida all'uso

(ver. 6.0 – Febbraio 2008)

Istituto Universitario di Architettura di Venezia
Dipartimento di Pianificazione

INDICE

Cap 1. - Installazione e Generalità

1.1 – Installazione e configurazione del pacchetto

L'ultima versione di ADDATI per Windows (ADDAWIN) è sempre disponibile, come file zippato, dalla pagina

http://cidoc.iuav.it/addawin.html

dalla quale si possono anche scaricare dei Data Set di prova.

È sufficiente creare una directory vuota per il programma ed espandervi il file compresso, senza modificare il Registro di Windows. **Evitare di collocare il programma sul "desktop"**, operazione che può creare inconvenienti. Meglio avere un buon controllo dell'albero delle direttorie e della collocazione dei file nel proprio computer.

Se si sta aggiornando o sostituendo una versione già installata si possono tranquillamente sovrascrivere i file interessati dall'aggiornamento.

La struttura del pacchetto è multilingue. Al momento può essere usato in inglese ed italiano, ma altre lingue potrebbero essere aggiunte con facilità. I file necessari sono tutti contenuti nella cartella dove il file di installazione viene espanso; d'ora in avanti, ci riferiremo ad essa come "l*a cartella di ADDATI*".

> ***Nota**: Si può cambiare la lingua in qualsiasi momento mentre si usa il pacchetto: stringhe alfanumeriche ed help vengono immediatamente presentati nella nuova lingua.*

Terminata l'installazione conviene creare sul desktop un collegamento al programma di gestione ADDAWIN.EXE e **subito configurarlo** (vedi più avanti).

Quando viene mandato in esecuzione, ADDATI per WINdows legge dal file di inizializzazione ADDATI.INI (contenuto nella cartella di installazione) alcuni parametri che determinano il suo modo di funzionamento pre-definito. Subito dopo l'installazione conviene configurarlo modificando tali parametri secondo le proprie esigenze. Si può farlo editando il file ADDATI.INI (che è un file di testo) con l'editor interno di ADDATI, scegliendo dal Menu l'opzione ***File→Edita/Mostra file di testo***; oppure usando l'opzione di Configurazione del pacchetto: ***File→Configura***.

1.1.1 - Il file di inizializzazione ADDATI.INI

Il file INI ha un contenuto simile, ma non necessariamente eguale, a quello mostrato qui sotto. Ogni riga è formata da una **parola-chiave** (**che non va mai modificata**) seguita da un parametro (modificabile) usato per inizializzare il programma.

```
LANGUAGE                ITALIANO
HELPTYPE                WINHELP
MISSING_VALUE_CODE      -9999
INVALID_VALUE_CODE      -998
DEFAULT_N_CLASSES       6
DEFAULT_THRESHOLD_TYPE  EQUAL_FREQUENCY
```

Alcuni brevi commenti:

- La parola-chiave LANGUAGE specifica la lingua con la quale il pacchetto parte. Per ora i valori possibili sono '`ITALIANO`' e '`ENGLISH`'.
- La parola-chiave HELPTYPE specifica il formato desiderato per il file di Help. I valori possibili sono '`WINHELP`' (il file di aiuto ha la classica estensione .HLP) o '`HTMLHELP`' (l'estensione del file è .CHM). Le diverse versioni di Windows accettano entrambi i formati, tranne Windows Vista che ha abbandonato il formato .HLP ed accetta solo file di Help di tipo .CHM. Chi usa Vista deve necessariamente scegliere quest'ultimo, a meno che non decida di installare WINHELP.
- MISSING_VALUE_CODE introduce il valore interpretato come *valore mancante* per tutti quei DataSet **che non ne forniscano uno loro specifico** nel file di documentazione. Il valore mancante dev'essere un valore che certamente non si confonde con i valori validi delle variabili.
- INVALID_VALUE_CODE è un valore usato internamente dal programma per marcare i valori non validi (fuori codifica, ecc.) delle variabili. Anch'esso non deve confondersi con i valori validi, e va dunque modificato ove necessario.
- DEFAULT_N_CLASSES è il numero di classi utilizzato come default nel calcolo delle distribuzioni di quelle variabili quantitative per le quali non ne venga specificato uno più appropriato nel file di documentazione, o durante l'esecuzione.
- DEFAULT_THRESHOLD_TYPE specifica come vadano costruite le classi delle distribuzioni: può valere 'EQUAL_INTERVALS' (intervalli di uguale ampiezza), o 'EQUAL_FREQUENCY' (intervalli *equi-frequenti*, cioè *quantili*, determinati automaticamente). Si applica a tutte le variabili prive di opzioni specifiche. Per qualsiasi variabile l'opzione può essere cambiata durante l'esecuzione, e si possono altresì fissare delle soglie liberamente stabilite dall'utente.

Ricorda

*I programmi di Analisi Multivariata, come ACOMP (Analisi in Componenti Principali) o NONGER (Classificazione non gerarchica**), non ammettono valori mancanti nelle tavole di dati loro sottoposte.** I record che manchino del valore di qualche variabile da utilizzare nell'analisi (sia attiva che supplementare) **vengono esclusi.***

1.1.2 - L'operazione di Configurazione

Scegliendo dal Menu l'opzione ***File→Configura*** appare il dialogo di figura 1.1, che permette di modificare i parametri di default, la lingua ed il formato del file di Help. L'editor di testo, indicato a solo scopo informativo, è quello interno di ADDATI e non è modificabile. Esso è stato concepito apposta per il pacchetto, e presenta troppi vantaggi per cambiarlo.

Si ricordi che i parametri inseriti saranno usati per tutti i DataSet che non ne forniscano altri loro specifici nel file di documentazione, o durante l'esecuzione.

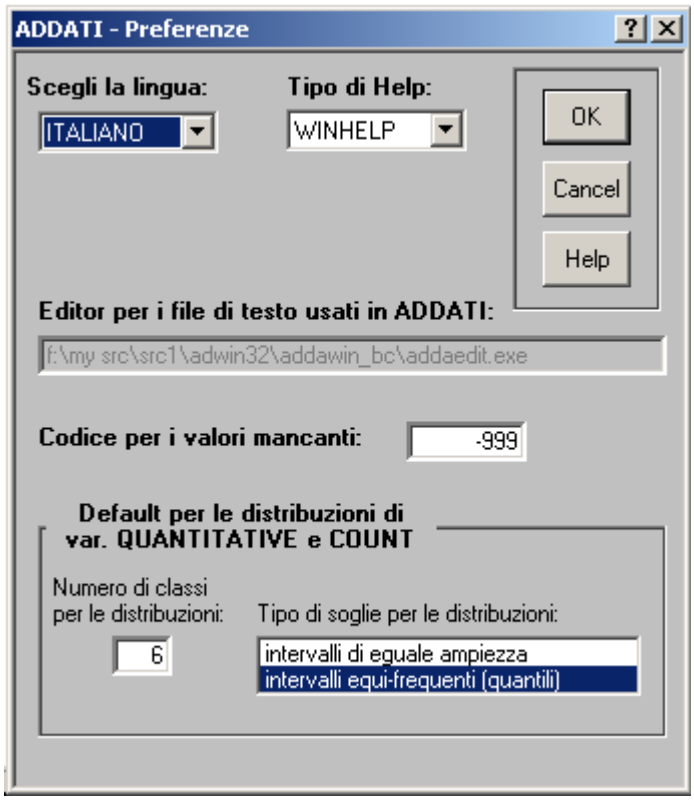


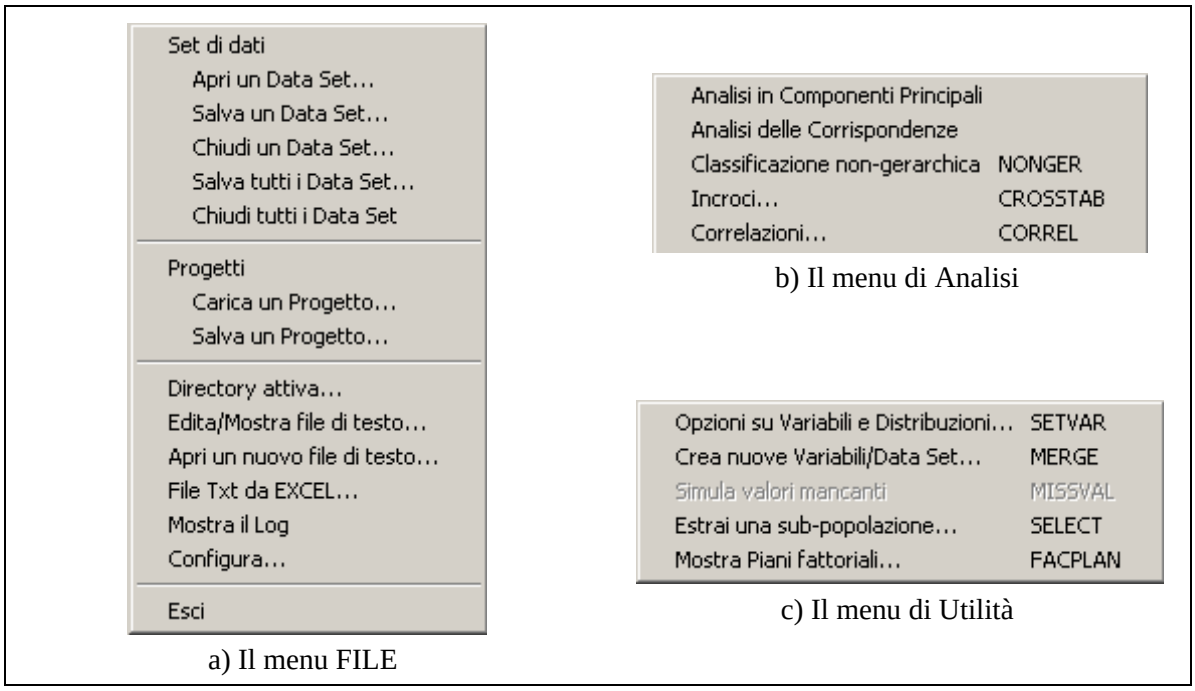
Fig. 1.1 – Il dialogo di configurazione

***Nota**: L'icona di ADDATI è un po' strana ed ha una storia curiosa. Una volta (tanto tempo fa...) ho detto scherzando a mia figlia, che aveva allora 10 anni: "Perché non mi fai un'icona per ADDATI?". E lei di rimando: "Ma cosa fa ADDATI?". Già, cosa fa ADDATI?*

Io, un po' incerto: "Mah, aiuta a scoprire delle cose, a rispondere a degli interrogativi...". La sera mi ha presentato *l'icona inclusa nel pacchetto, fatta con l'editor di immagini del compilatore Watcom* C. Forse non è particolarmente bella, ma con tutti quei punti interrogativi interpreta bene gli obiettivi essenzialmente esplorativi del pacchetto...

1.2 – Generalità sul funzionamento

Il Menu di ADDATI offre i tre sottomenu mostrati qui sotto.



a) Il menu FILE

b) Il menu di Analisi

c) Il menu di Utilità

Figura 1.2 - I tre Menu di ADDATI

Alcune voci restano disabilitate fino a che non sia stato caricato almeno un DataSet. Le operazioni permesse dai tre menu verranno trattate nei capitoli seguenti.

1.2.1 - I DataSet

Un DataSet è formato

- da un file di dati;
- da un file di documentazione che ne descrive il contenuto;
- da una *etichetta* (o *label*, o *nome*) che lo identifica distinguendolo dagli altri DataSet caricati.

Si possono caricare simultaneamente più DataSet, anche se operazioni che ne utilizzano più d'uno (come la creazione di nuovi DataSet ottenuti copiando o ricodificando variabili esistenti, o costruendone di nuove in base a diverse opzioni) possono ovviamente riguardare solo DataSet che descrivano le stesse unità statistiche.

I file di dati, ed i file di documentazione associati, sono file di testo. La loro struttura è descritta in dettaglio nel capitolo 3.

I file di dati

Ogni record consiste di un insieme di *variabili* (o *indicatori*), che descrivono le unità statistiche (individui, famiglie, unità amministrative, ecc). Le estensioni usuali (ma non obbligatorie) per questi file sono **.DAT** o **.CSV (C**omma-**S**eparated **V**alues**)**.

I valori possono essere separati da spazi, virgole, punti e virgola, oppure non avere separatori, nel qual caso ciascuna variabile deve occupare esattamente le medesime colonne in tutti i record per essere distinguibile dalle altre.

Il formato senza separatore è accettato solo in lettura per file provenienti da altre fonti: tipicamente, da Istituti Centrali di Statistica e quando si tratti di grandi inchieste al massimo livello di disaggregazione. In scrittura invece ADDATI usa sempre un separatore.

I file di documentazione

Il contenuto di un file di dati è descritto dal file di documentazione ad esso associato, che ha di solito (ma non necessariamente) lo stesso nome ed estensione **.TXT**. Esso può essere preparato usando un normale editor di testo (come il Blocco Note o l'editor interno di ADDATI), e seguendo le regole descritte nel seguito. E' previsto (ma non ancora implementato) un aiuto limitato per la sua creazione.

I file di dati prodotti o modificati all'interno del programma (copiando, calcolando o ricodificando variabili; scegliendo un sottoinsieme di casi in base ad una condizione; ecc) vengono salvati insieme al loro file di documentazione, prodotto in modo automatico.

1.2.2 - L'help in linea

I vari dialoghi che controllano il funzionamento del programma includono un bottone di HELP che offre una descrizione generale sugli scopi ed il funzionamento del dialogo.

Per un ***aiuto specifico*** sull'utilizzo di qualche controllo (bottoni, finestre di editing, ecc.), si trascini su di esso il '**?**' localizzato nell'angolo in alto a destra della finestra di dialogo, o si prema il tasto **F1**.

1.2.3 - Le informazioni sempre disponibili

In generale tutti i dialoghi includono dei controlli, come quelli mostrati nella figura 1.3 qui sotto, che permettono di esaminare rapidamente le caratteristiche dei DataSet caricati.

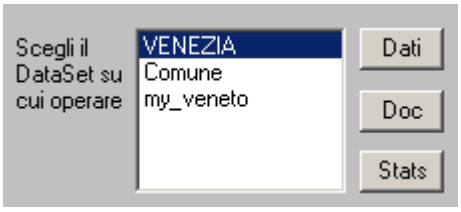


Figura 1.3 - I pulsanti di consultazione presenti in tutti i dialoghi

Si possono visualizzare i dati, la loro documentazione e le statistiche elementari relative a tutte le variabili del DataSet selezionato. La cosa è spesso di grande utilità per orientare le decisioni su come riempire il dialogo.

Le informazioni richieste vengono visualizzate nell'editor interno di ADDATI.

1.2.4 - L'editor interno ADDAEDIT

ADDATI usa un suo editor interno (ADDAEDIT.EXE, che risiede nella cartella di ADDATI) per tutte le operazioni su file di testo. L'editor è stato scritto appositamente, e si integra molto bene con il programma. Ad esempio, dopo che un DataSet è stato caricato, e dunque la sua documentazione è nota ad ADDATI, una richiesta di esaminare i dati (avanzata premendo il pulsante **Dati**) lancia l'editor, che apre una finestra di visualizzazione del tipo mostrato nella figura 1.4 qui sotto.

La barra inferiore della finestra mostra sempre la riga, la colonna, il numero d'ordine ed il nome della variabile sulla quale il cursore è posizionato. Se la variabile è di tipo categoriale, viene anche mostrato il significato del codice sotto il cursore. Le informazioni vengono automaticamente aggiornate cliccando su un altro campo, o spostando il cursore con le frecce.

Inoltre, durante l'esecuzione di alcuni comandi ADDATI può aprite l'editor visualizzando la parte del file di output che l'utente deve esaminare per decidere come proseguire l'analisi.

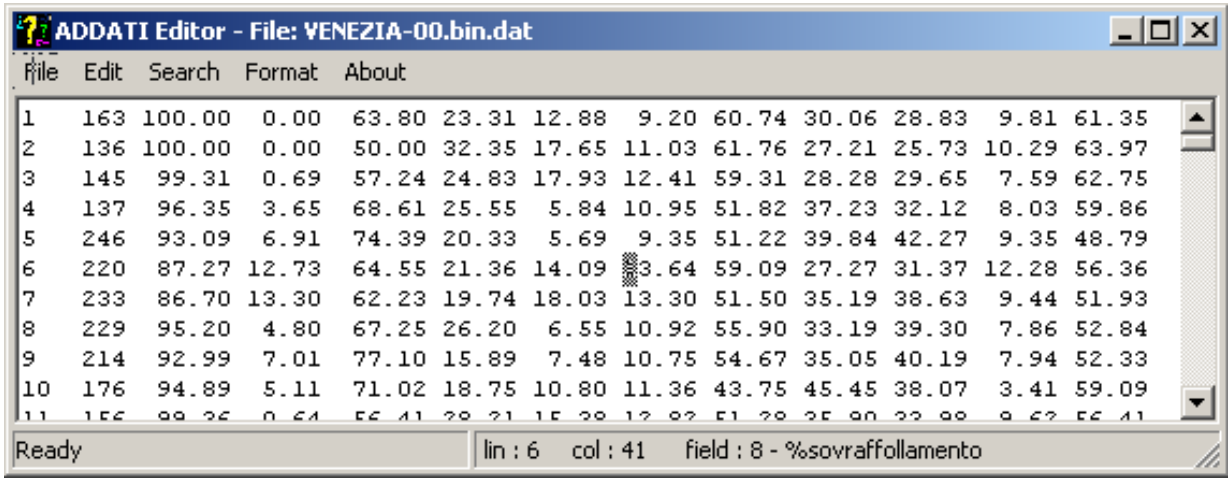


Figura 1.4 – La visualizzazione dei dati di un DataSet nell'editor interno

Nel Capitolo 2 verranno esposte alcune definizioni e nozioni statistiche elementari, necessarie per la comprensione delle operazioni disponibili in ADDATI, e per la comprensione dei risultati. Tali operazioni verranno illustrate nei capitoli successivi.

Cap 2. – Nozioni introduttive

Questo è solo un manuale d'uso: non intende essere un testo teorico di statistica multivariata, né potrebbe esserlo. Per una comprensione approfondita di come operano le tecniche di Analisi Fattoriale e Classificazione è opportuno consultare qualche testo specifico. Poiché ADDATI si ispira alla Scuola Francese di Analisi dei Dati (Analyse des Données), si consigliano i numerosi titoli pubblicati in Francia specialmente da Dunod.

Questo capitolo introduce alcune nozioni statistiche elementari, particolarmente rilevanti in una prima ispezione di una base di dati territoriali, e per capire le funzioni incluse nel Menu di Utilità.

Ulteriori elementi di linguaggio, ed alcune altre definizioni preliminari alle Analisi Multivariate (Analisi in Componenti Principali, Analisi delle Corrispondenze e Classificazione) verranno invece trattate nei capitoli 6 e 7.

2.1 – Analisi multivariate e scale di misura

Quando si voglia eseguire un'analisi, in particolare un'analisi multivariata, l'operazione più importante e delicata è certamente la preparazione della tavola dei dati.

Si può trattare di una ***tavola di descrizione***, le cui righe rappresentano unità statistiche (aree geografiche, imprese, famiglie, ecc.) descritte da alcuni indicatori (le colonne della tavola) direttamente osservati o costruiti opportunamente, sia di tipo quantitativo che qualitativo (ad es., un insieme di variabili socio-economiche, o demografiche, ecc.).

In generale, vanno incluse nella tavola solo le variabili che rappresentano, nel modo più appropriato e completo possibile, ***i caratteri delle unità statistiche che l'analista ritiene rilevanti per la particolare analisi da condurre***. Nulla di meno, anche se spesso è necessario un compromesso per via dell'inadeguatezza dell'informazione disponibile, ma anche nulla di più, dato che l'inclusione di variabili scarsamente pertinenti (a meno di non trattarle come supplementari) può distorcere il risultato in modo indesiderato ed imprevedibile. La scelta delle variabili costituisce dunque un'assunzione sostanziale che richiede riflessione e consenso: è ben noto come la percezione di un problema sia raramente la medesima per tutti gli attori coinvolti.

Questi metodi si possono utilizzare anche per trattare tavole di altro tipo: ad esempio, una tavola di alternative di scelta valutate secondo un insieme di criteri, con l'obiettivo di ordinarle secondo la loro utilità globale. L'aspetto comune è la multi-dimensionalità della descrizione: le ***unità statistiche*** (a volte chiamate nel seguito anche ***oggetti*** o ***individui***) sono considerate secondo una pluralità di ***attributi*** o ***caratteri***, tra i quali si suppone esista un insieme di relazioni a priori ignote, e caratteristiche del particolare contesto dell'analisi. Il percorso di analisi esplora questa rete di relazioni, limitandosi però a quelle lineari, e riduce la multidimensionalità del fenomeno, lasciando cadere in modo ottimale solo una piccola parte dell'informazione. Le unità statistiche vengono poi aggregate in classi secondo la loro somiglianza, definita globalmente tenendo conto di tutti i caratteri elementari.

Le variabili utilizzate possono essere misurate secondo diverse scale, e va scelto un percorso di analisi appropriato per la scala adottata. In particolare, se la scala di misura non è la stessa per tutte

le variabili, esse vanno prima sottoposte ad una opportuna ***ricodifica*** che le riporti ad una medesima scala (***Utilità→Crea nuove Variabili/DataSet***)). Ciò non è ancora sufficiente per una impostazione corretta dell'analisi: se si tratta di variabili categoriali, il numero e la frequenza delle loro categorie andrebbero per quanto possibile bilanciati, allo scopo di evitare la dominanza statistica di una variabile sulle altre e la conseguente diversa influenza sui risultati dell'analisi.

	TIPO DI VARIABILI		
	QUANTITATIVE	**QUALITATIVE**	
		ORDINALI	NOMINALI
Hanno senso operazioni aritmetiche sui valori?	si	no	no
i valori sono ordinati?	si	si	no
tipo di valori	numerici	codici alfa-numerici	codici alfa-numerici

Tabella 2.1 - Le scale di misura per le variabili.

La tabella 2.1 elenca i tipi di scala. È importante riconoscere quella usata per ciascuna variabile, in modo da scrivere correttamente il file di documentazione ed impostare correttamente l'analisi.

È altresì importante individuare se le unità statistiche analizzate, ciascuna delle quali è descritta **da una riga** della tavola dei dati (cioè da un record del DataSet) siano **unità elementari direttamente osservate** (individui, famiglie, ecc.) o siano invece **unità più complesse,** costruite ***aggregando*** unità elementari sottogiacenti, direttamente osservate. È questo il caso di unità amministrative (Comuni, Sezioni Censuarie...) la cui descrizione è ottenuta aggregando i dati elementari dei Fogli di Famiglia. Trattandosi di descrizioni di tipo diverso, richiedono spesso diversi metodi di elaborazione.

2.1.1 - I tipi di variabili

Per poterle trattare correttamente ADDATI ha bisogno di conoscere la scala di misura delle variabili incluse nel DataSet che sta caricando. L'informazione va inserita nel file di documentazione, dove per ogni variabile va specificato uno dei seguenti quattro tipi: QUANTITATIVE, CATEGORIAL, COUNT o ID.

2.1.2 - Variabili di tipo QUANTITATIVE

Esse assumono **valori numerici** espressi in una unità di misura appropriata, oppure a-dimensionali. **Su tali valori ha senso compiere delle operazioni aritmetiche**.

Il reddito pro-capite, la popolazione, il tasso di attività di un comune, la superficie di un alloggio sono esempi di variabili quantitative. La variabile assume un valore numerico in corrispondenza ad ogni unità statistica (comuni, sezioni censuarie, famiglie, individui, ecc.): nei calcoli i valori di queste variabili vengono *ponderati* con il peso associato all'unità statistica cui si riferiscono (il peso di ciascun caso rispecchia la sua importanza relativa). Vengono calcolate ed utilizzate nell'analisi la ***media*** (ponderata) di ciascuna variabile quantitativa (calcolata sull'insieme delle unità statistiche) e la sua *deviazione standard* (o *scarto quadratico medio*), definite più avanti.

2.1.3 - Variabili di tipo CATEGORIAL (o qualitative)

Assumono un insieme *limitato* di valori (detti ***categorie*** o ***modalità***), rappresentati da codici opportuni. Anche se spesso si usano come codici dei numeri, il loro significato non è numerico e nessuna operazione aritmetica ha significato, a parte il contare le unità in ciascuna categoria. Si possono ulteriormente distinguere le variabili categoriali in ***ordinali*** e ***nominali***.

Variabili categoriali ordinali

Sono di solito ottenute ricodificando delle variabili quantitative, allo scopo di riportarle alla stessa scala di altre variabili categoriali contenute nella medesima tavola di dati. L'operazione causa una perdita d'informazione che bisogna cercare di minimizzare, ma non cambia molto il comportamento qualitativo delle unità statistiche e la struttura delle loro somiglianze.

Come esempio, si consideri il tasso di attività (rapporto tra individui attivi e popolazione totale) dell'insieme dei comuni di una regione: esso può assumere *teoricamente* un qualsiasi valore tra 0 e 100%, anche se l'effettivo intervallo di variazione è di solito molto più ristretto: si tratta dunque di una variabile continua. Se si vuole utilizzarlo in una tavola che includa altri caratteri socio-economici dei comuni considerati, ***osservati alla scala categoriale***, il tasso di attività va ricodificato: l'intervallo 0-100 va suddiviso in opportuni sotto-intervalli (classi).

La tabella 2.2 mostra un esempio. L'intervallo 0-100 è suddiviso in quattro ***intervalli di eguale ampiezza***: ne risultano quattro classi con tasso di attività crescente, contraddistinte dai codici numerici da 1 a 4. Tutti i comuni che ricadono in una classe sono contrassegnati dal medesimo codice, e dunque le loro differenze vanno perdute dopo la conversione alla scala categoriale.

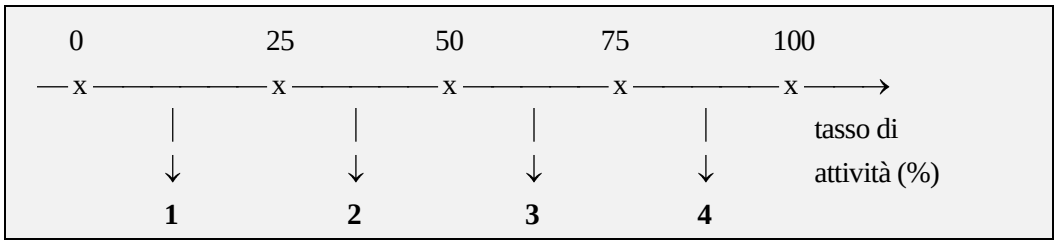


Tabella 2.2 - Ricodifica di una variabile quantitativa in qualitativa con quattro categorie.

La ricodifica precedente risulta inappropriata in molti casi. Ad esempio, se i valori del tasso di attività variano di fatto tra il 17 ed il 45 per cento (cioè se il Comune meno attivo della Regione ha un tasso del 17%, mentre il più attivo presenta un tasso del 45%) la terza e la quarta classe risultano vuote. È allora opportuno decidere la nostra regola di ricodifica **dopo aver determinato** i valori minimo e massimo *effettivamente assunti* dalla variabile in questione. Se si vogliono ancora quattro classi corrispondenti ad intervalli di eguale ampiezza, si arriva alla ricodifica della tabella 2.3.

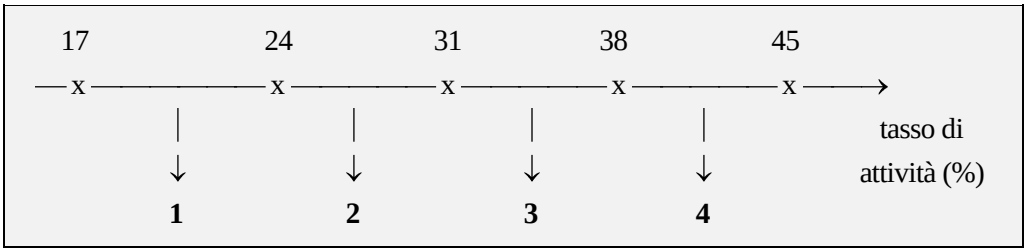


Tabella 2.3 - L'esempio precedente, con un intervallo di valori [17..45] per cento.

Supponiamo ora che la maggior parte dei comuni abbia un tasso di attività tra 25 e 36%. Essi cadono nella seconda e nella terza classe, mentre le due classi estreme risulterebbero quasi vuote. Ciò *appiattisce* la descrizione del sistema (almeno, quella offerta da questa variabile). Viene perduta una parte consistente di informazione, come le differenze che esistono tra i (molti) comuni inclusi nella classe 2, o tra quelli in classe 3.

Si può mostrare che ***la perdita di informazione è minima quando i valori di soglia sono fissati in modo da ottenere classi all'incirca equi-numerose (quantili)***. Tale operazione può essere fatta automaticamente in ADDATI, quando si calcolano distribuzioni o si ricodificano variabili.

Anche se vengono usati codici numerici, le etichette '**1**'...'**4**' usate per le categorie non hanno un significato numerico: '4' non è il doppio di '2', ed i codici '***a***'...'***d***' andrebbero altrettanto bene. Tuttavia, quando una variabile quantitativa viene ricodificata almeno l'ordine dei suoi valori viene salvato: '2' non è il doppio di '1', ma rappresenta certamente un tasso di attività superiore. Per tale ragione, la variabile categoriale prodotta dalla ricodifica è detta ***ordinale***.

Variabili categoriali nominali

Esse non implicano alcun ordine sottogiacente: i codici sono solo delle etichette associate a comportamenti diversi, ***senza alcun ordine***. Spesso si usano i numeri come codici, ma si tratta di un fatto senza alcun significato particolare. Variabili a due valori (ad esempio, '*si*' e '*no*', o *'presente'* e *'assente'*) sono nominali, ma non è l'unico caso.

Si pensi ad esempio di contrassegnare con i codici [1..n] i diversi tipi di coltivazione: si può definire per ogni area geografica una variabile *coltivazione prevalente* con valore da 1 a n. Si tratta di una variabile nominale, che non ha alcun ordine inerente. Come altro esempio, si può considerare nominale la variabile che rappresenta, per ogni unità statistica, il numero della classe cui quell'unità è stata assegnata da una procedura di classificazione.

È chiaro che i codici usati per le variabili categoriali non possono essere sottoposti a trattamenti numerici: almeno, non direttamente. Operazioni come il calcolo della media o della deviazione standard non hanno senso per questa scala di misura. Una variabile categoriale va prima sottoposta ad una ***ricodifica binaria***: essa viene *sostituita da tante nuove variabili quante sono le sue categorie*, ciascuna delle quali vale 1 se l'unità statistica assume la categoria associata, 0 altrimenti. Poiché un caso può ricadere in una sola categoria, solo una delle variabili binarie ottenute ricodificando in tal modo una variabile categoriale può valere 1, mentre tutte le altre varranno 0. Questo tipo di codifica è anche noto come codifica in ***forma disgiuntiva completa***.

La tabella 2.4 mostra i valori ottenuti ricodificando il tasso d'attività della tabella 2.3. Poiché i casi sono raggruppati in quattro classi (categorie), sono necessarie quattro variabili binarie.

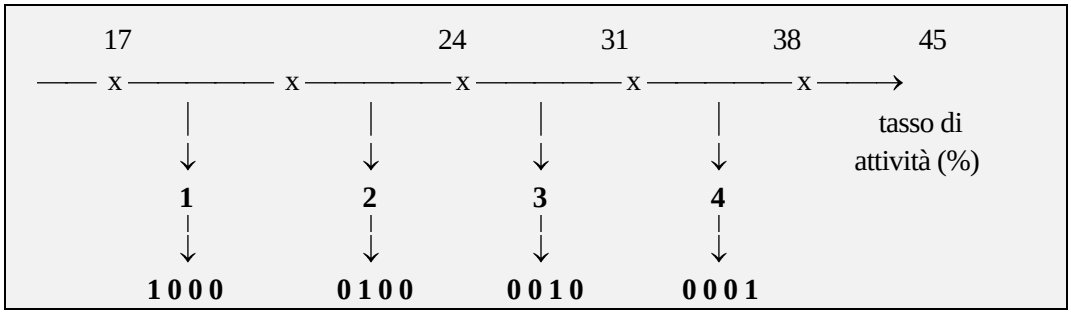


Tabella 2.4 - L'esempio di tabella 2.3, seguito da una ricodifica in forma binaria. Vengono create quattro variabili binarie: quella corrispondente alla categoria assunta vale 1.

2.1.4 - Variabili di tipo COUNT (o di conteggio)

Assumono anch'esse valori numerici come le variabili di tipo QUANTITATIVE, ma solo interi, non decimali. Si tratta di variabili che contano delle unità statistiche (individui, famiglie, alloggi...). Ad esempio, per i Comuni di una Regione variabili come la popolazione di età inferiore a 15 anni, la popolazione con licenza elementare o gli alloggi occupati sono di tipo COUNT. Molte delle operazioni che si possono fare su variabili QUANTITATIVE valgono anche per quelle di tipo COUNT, eppure la differenza è evidente.

In generale, prima di elaborarle variabili di questo tipo andranno convertite in percentuali su un totale opportun . Ha poco senso fare la media **dei valori assoluti** dei laureati nei Comuni di una Regione (cioè dei valori COUNT), perché il numero assoluto dei laureati è grosso modo proporzionale alla popolazione del Comune, e solo in second'ordine dipende dalle differenze socio-economiche esistenti tra i Comuni. Invece, avrà senso calcolare la media **delle percentuali** di laureati (sulla popolazione totale del Comune). Ponderando tali valori con la popolazione del Comune si otterrà la percentuale media **vera** di laureati nella Regione.

	Popolazione < 15 anni	**Popolazione 15-64 anni**	**Popolazione 65+ anni**
Comune1	1000	3000	1500
.....	.....	.....	.....
Comune i	15000	30000	15000
.....	.....	.....	.....
Comune n	10000	40000	20000

Tabella 2.5 - Un esempio di incrocio che produce una descrizione con variabili COUNT

Si consideri la tabella 2.5, che rappresenta un DataSet che descrive n Comuni di una Regione dividendo la popolazione in tre classi d'età. I valori mostrati **contano** gli individui in ciascuna classe d'età, e sono dunque variabili di tipo COUNT. D'altra parte la tavola si può pensare come il risultato dell'incrocio, calcolato su tutti gli individui della Regione (dunque su di un insieme di unità elementari sottogiacenti), tra due variabili categoriali: il Comune di residenza, con n categorie, e la classe d'età con tre categorie. In generale è così: le frequenze sulle categorie di una variabile categoriale, calcolate al massimo livello di disaggregazione, diventano delle variabili COUNT nella descrizione delle unità aggregate di livello superiore.

La distinzione tra variabili QUANTITATIVE e COUNT non è troppo critica, e a volte i due tipi possono confondersi nella pratica senza troppi guai. A volte però non è così, ed è importante che la differenza sia chiara..

2.1.5 - Variabili di tipo ID (identificatori delle unità statistiche)

Assumono anch'esse valori diversi sulle diverse unità statistiche, e dunque le chiamiamo variabili, ma sono di tipo piuttosto particolare. Sono delle **etichette** (label), cioè dei **nomi associati alle unità statistiche**: un numero d'ordine, il Codice ISTAT, il nome del Comune sono degli identificatori. Il codice cartografico che consente un'operazione di ***join*** in ArcView tra il DataSet statistico e lo shapefile che localizza spazialmente le unità statistiche è un ID.

Come a volte una variabile può essere dichiarata sia di tipo QUANTITATIVE che COUNT, anche un ID può a volte essere dichiarato come una variabile di tipo CATEGORIAL. È questione di significato e di convenienza. Un ID può avere una corrispondenza 1 a 1 con le unità statistiche (ad

esempio, ogni Comune ha un codice ISTAT diverso), ma a volte si dichiarano come ID semplicemente variabili sulle quali non si intende eseguire operazioni matematiche. Ad esempio, sul DataSet di tutti i Comuni d'Italia, un campo che contenga il nome della Regione può essere dichiarato come ID. Si tratta tuttavia di una variabile con un certo numero di categorie (i cui codici sono i nomi delle Regioni) e quindi può (dovrebbe) essere dichiarata come CATEGORIAL. La differenza sta nel fatto che indicandola come ID nel file di documentazione non si può usarla in distribuzioni ed incroci, mentre la cosa è possibile se la si dichiara CATEGORIAL. Tutto qui.

2.2 - Standardizzazione (normalizzazione) di variabili continue

Vale la pena di inserire una variabile in un'analisi solo se essa consente di discriminare significativamente il comportamento delle unità statistiche che descrive. Essa deve dunque assumere valori sufficientemente diversi sui diversi casi. Se fosse costante (assumesse cioè lo stesso valore per tutte le unità statistiche), essa non sarebbe di alcun interesse e sarebbe opportuno rimuoverla dall'analisi, per la quale costituirebbe un appesantimento inutile. Considerato che l'analisi s'interessa alla distribuzione dei valori della variabile sull'insieme delle unità statistiche, i comuni concetti di ***media*** e di ***scarto quadratico medio*** rivestono grande utilità.

Sia I l'insieme delle unità statistiche ed indichiamo con x_i il valore assunto dalla variabile x sull'unità i**.** Ad ogni unità è associato un ***peso***, che rappresenta la sua *importanza* nella particolare analisi che si sta conducendo sulla variabile x. **Si tenga presente che la trattazione di variabili diverse può richiedere una diversa ponderazione delle unità statistiche**: ADDATI consente una ponderazione diversa per ciascuna variabile.

Sia m_i il peso assegnato all'unità i. I pesi vengono ***normalizzati*** in ADDATI mediante la trasformazione:

$$m_i \Leftarrow m_i / M$$

dove $M = m_1 + \ldots + m_n$ rappresenta la somma dei pesi di tutte le unità. Una volta normalizzati, i pesi rappresentano **delle frazioni percentuali** che assommano ad 1: ciascuno di essi misura in termini percentuali l'importanza dell'unità associata, cioè la quota-parte dell'intero sistema che essa rappresenta (per fissare le idee, si pensi ai Comuni di una Regione…).

La ***media ponderata*** della variabile x sull'insieme I è data dalla ben nota formula

$$\bar{x} = \text{media}(x) = m_1 * x_1 + m_2 * x_2 + \ldots + m_n * x_n = \Sigma_i \, m_i * x_i \qquad (2.1)$$

dove il contributo di ciascuna unità alla media dipende anche dal suo peso (si ricordi che i pesi sono supposti normalizzati, cioè la loro somma vale 1). Come caso particolare, quando tutte le unità hanno lo stesso peso i pesi normalizzati risultano essere:

$$m_1 = m_2 = \ldots = m_n = 1/n$$

e la (2.1) diventa la ***media semplice***:

$$\bar{x} = (x_1 + x_2 + \ldots + x_n) / n$$

La media è una ***misura di tendenza centrale:*** i valori assunti della variabile sono distribuiti intorno ad essa.

La ***varianza*** di una variabile è una misura della *dispersione* dei suoi valori. Essa è data dalla

$$\textbf{\textit{var}}(x) = \sigma^2(x) = m_1 * d_1^2 + \ldots + m_n * d_n^2$$

dove $d_i = x_i - \bar{x}$ è lo scarto tra il valore che la variabile assume nell'unità *i* e la sua media. La varianza è ottenuta sommando i quadrati di tutte queste differenze, ponderando ciascuna con il peso dell'unità considerata.

Lo ***scarto quadratico medio*** (o ***deviazione standard***) di una variabile è semplicemente la radice quadrata della sua varianza.

$$\textbf{\textit{stdev}}\,(x) = \sqrt{\sigma^2(x)}$$

Poiché i valori numerici assunti da una variabile dipendono dall'unità di misura utilizzata (e così le differenze rispetto alla media, la varianza e lo scarto quadratico medio), è opportuno un cambiamento di scala che riporti tutte le variabili alla medesima rilevanza. A tale scopo il valore x_{ij} assunto dalla variabile *j* nell'unità *i* viene ***standardizzato*** (si usa anche il termine ***normalizzato***), vale a dire trasformato nel modo seguente:

$$x_{ij} \Leftarrow \frac{x_{ij} - \bar{x}_j}{stdev(x_j)}$$

Prima i valori assoluti della variabile sono sostituiti dalle differenze rispetto alla sua media, poi la scala è corretta dividendo per lo scarto quadratico medio. Ciò elimina l'effetto dell'unità di misura sui valori della variabile. La nuova variabile ottenuta mediante questa trasformazione viene ad avere **per costruzione** media 0 e varianza (e scarto quadratico medio) pari ad 1.

Nota *Quando ADDATI carica un DataSet, vengono immediatamente calcolate e memorizzate media, deviazione standard, minimo, massimo ed altre informazioni statistiche su tutte le variabili di tipo QUANTITATIVE e COUNT. Tali valori vengono automaticamente aggiornati se l'utente cambia i parametri del loro calcolo (ad esempio, il modo di ponderare le variabili, o gli intervalli per il calcolo delle distribuzioni).*

*Media, deviazione standard e somma totale delle variabili QUANTITATIVE o COUNT possono essere usate nelle regole di costruzione di nuove variabili (**Utilità→Nuove Variabili/DataSet**) o nelle condizioni che definiscono un segmento di popolazione.*

*L'opzione **ACOMP** (**Analisi→Analisi in Componenti Principali**) del Menu di ADDATI standardizza automaticamente (come default) le variabili continue in ingresso, così da attribuire a ciascuna di esse la medesima importanza nell'analisi. Viene cioè **diagonalizzata** la matrice delle correlazioni tra le variabili. L'utente può comunque scegliere di diagonalizzare la matrice delle varianze-covarianze.*

2.2.1 - Esempio 1 - Indicatori di tendenza centrale e dispersione

Si ipotizzino tre comuni che abbiano la popolazione, il numero degli attivi ed il tasso di attività mostrati nella tabella 2.6.

	Popolazione pop_i	Attivi att_i	tasso attività $t_i = att_i / pop_i$	scarto	scarto al quadrato	peso m_i
Comune 1	40.000	12.000	0.30	–.06	.0036	0.27
Comune 2	100.000	40.000	0.40	+.04	.0016	0.667
Comune 3	10.000	2.000	0.20	–.16	.0256	0.067
Totale	150.000	54.000				1.0

Tabella 2.6 - Dati esemplificativi su tre comuni.

Sarebbe corretto calcolare il tasso di attività medio del sistema dei tre Comuni come segue:

$$\bar{t} = \frac{0.30 + 0.40 + 0.20}{3} = 0.30 \ ?$$

Ha senso cioè fare una ***media semplice***? Bisogna considerare che:

- i Comuni hanno diversa popolazione, dunque **danno contributi diversi** al calcolo delle statistiche relative all'intero sistema;
- ogni Comune costituisce una quota parte del sistema e contribuisce ad esso in proporzione alla sua popolazione.

m_i = peso del Comune i = pop_i/popolazione totale = $pop_i/(pop_1+pop_2+pop_3)$

La somma dei pesi così definiti vale uno**:** $m_1 + m_2 + m_3 = 1$

La ***media (ponderata)*** è un indicatore della tendenza centrale di una distribuzione:

media: $\bar{t} = \sum_i m_i * t_i = m_1 * t_1 + m_2 * t_2 + m_3 * t_3$ = 0.27∗0.30 + 0.667∗0.40 + 0.067∗0.20 = **0.36**

Anche nel calcolo di tutte le altre statistiche va tenuto conto del peso.

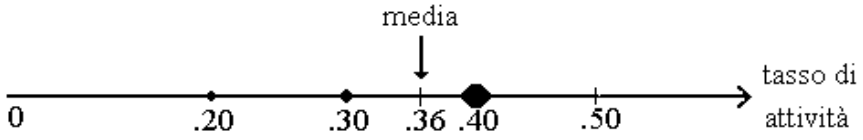


Figura 2.1 - I valori del tasso d'attività dei tre comuni rappresentati su di un asse. La grandezza di ogni punto è proporzionale alla popolazione. La media si colloca nel baricentro del sistema dei tre punti ponderati.

La varianza del tasso d'attività sull'insieme dei tre comuni vale

Varianza(t) = $\boldsymbol{\sigma}^2(t)$=$\sum_i mi(t_i - \bar{t})^2$ = 0.27∗0.0036+0.67∗0.0016+0.067∗0.0256 = 0.0072

deviazione standard = σ = $\sqrt{.0072}$ = .085.

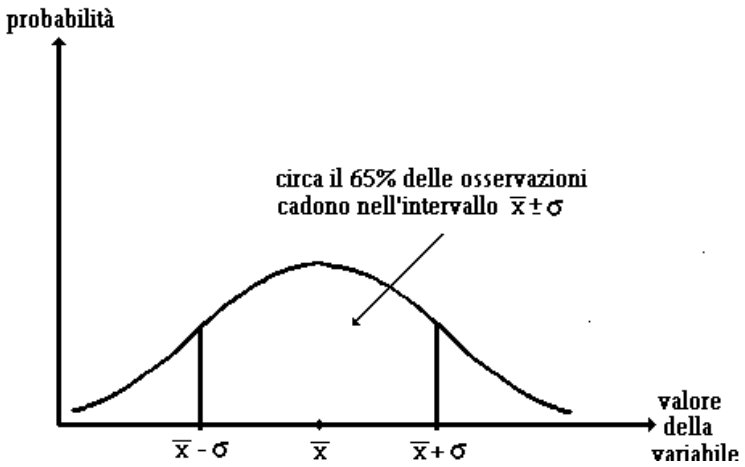


Figura 2.2 - La distribuzione *normale* (o *gaussiana*) ed il significato di **σ.**

La ***deviazione standard*** (o ***scarto quadratico medio***) σ (sigma) misura la dispersione della distribuzione. La figura 2.2 mostra la distribuzione nel caso di una variabile *normale* ed il significato dello scarto quadratico medio.

2.2.2 - Esempio 2 – Variabili quantitative: normalizzazione

La tabella 2.7 si riferisce ad una città divisa in cinque quartieri Q_i (i = 1...5), ciascuno con la sua popolazione pop_i. Poiché la popolazione totale pop_{tot} è di 187,234 persone, il generico quartiere *i* viene ad avere un peso m_i, riportato in tabella, pari a

$$p_i = \frac{pop_i}{pop_{tot}}$$

Si sono rilevate in ciascun quartiere le quattro variabili seguenti, ritenendole complessivamente sufficienti alla costruzione di un ***indicatore*** che misuri il livello di ***benessere globale*** del quartiere:

status: % di imprenditori, liberi professionisti, dirigenti ed impiegati sulla popolazione attiva

dipl: percentuale di laureati o diplomati sulla popolazione

pov: percentuale di poveri (non è qui necessario specificare come siano definiti) sulla popolazione

affoll: indice di affollamento globale (pop_i/stanze totali occupate nel quartiere)

	pop_i	m_i	$status_i$	$dipl_i$	pov_i	$affoll_i$
Q1	28125	.148	18.8	8.0	3.86	1.4
Q2	35853	.188	11.6	2.7	8.24	1.8
Q3	36169	.190	12.9	4.1	2.28	1.7
Q4	30329	.159	60.2	31.9	0.24	0.9
Q5	60028	.315	23.9	6.8	2.84	1.4
media			24.52	9.69	3.49	1.45
σ			16.26	9.83	2.52	0.29

Tabella 2.7 - Indicatori di benessere in cinque quartieri

Si osserva immediatamente che:

- Q4 appare il quartiere migliore, con i livelli più alti di *status* e *dipl*, più bassi di *pov* e *affol*.
- È difficile valutare i valori assoluti delle variabili nei vari quartieri; ha invece senso confrontare ciascuno di essi con la media globale, per capire quali quartieri assumano valori inferiori o superiori alla media relativa all'intero sistema.
- Le medie e le deviazioni standard sono diverse da variabile a variabile.

Se x_{ij} è il valore (riportato nella tabella) assunto dalla variabile *j* nel quartiere *i*, operiamo la trasformazione

$$y_{ij} \leftarrow x_{ij} - \overline{x_j} \qquad (2.2)$$

dove $\overline{x_j}$ è la media della variabile *j* ed y_{ij} misura lo scarto tra il valore della variabile nel quartiere *i* e la sua media globale.

- La (2.2) assegna a y_{ij} valori negativi in tutti i quartieri con valore x_{ij} ***sotto la media*** $\overline{x_j}$, valori positivi se x_{ij} è sopra la media.
  In particolare, y_{ij} = 0 se x_{ij} è eguale alla media.
- Le due variabili hanno ancora la stessa dispersione: $\sigma(y) = \sigma(x)$ ma la *y* ha media nulla (cioè è ***centrata*** sull'origine)
- Ora **il segno** della *y* ci dice subito se il valore originale *x* stava sopra o sotto la media.

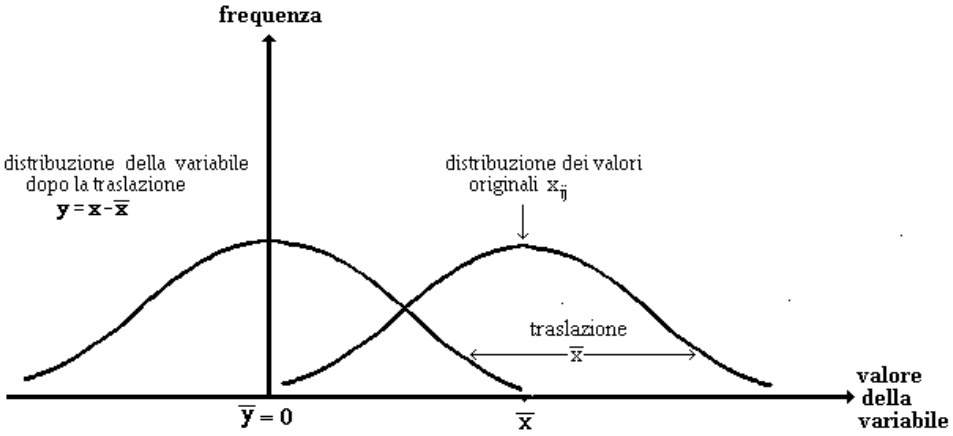


Figura 2.3 - Distribuzione dei valori di *x* prima della traslazione e distribuzione della corrispondente variabile centrata *y*.

Ma come si fa a stabilire rapidamente se un valore di y, cioè uno scarto rispetto alla media, è grande o piccolo? Bisognerebbe confrontarlo con gli scarti degli altri quartieri, e la cosa andrebbe fatta variabile per variabile, visto che esse hanno varianza diversa.

Ad esempio, la tabella 2.7 mostra che la dispersione dello *status* è molto superiore a quella di *affoll*; di conseguenza, uno scarto assoluto di 3.0 rispetto alla media è enorme se si verifica per l'affollamento, è molto meno rilevante per lo status.

Per poter confrontare gli scarti rispetto alla media di variabili diverse **conviene riportarle tutte alla stessa dispersione**, dividendo gli scarti y di ciascuna variabile per la sua deviazione standard σ. In tal modo, più alta è la dispersione (cioè il σ) della variabile, più l'entità degli scarti sulla media

viene ridimensionata. In altri termini, si usa la σ di ciascuna variabile come unità di misura per esprimerne la dispersione.

Si può dimostrare facilmente che la variabile z_j che così si ottiene ha media 0 e varianza 1. Si esprime tale fatto scrivendo $z_j(0,1)$.

$$z_{ij} = \frac{y_{ij}}{\sigma_j} = \frac{x_{ij} - \overline{x}}{\sigma_j}$$

La z_j si dice *normalizzata* o *standardizzata*.

Un insieme di variabili normalizzate sono subito comparabili, poiché hanno la stessa media (pari a 0) e la stessa deviazione standard (pari ad 1), a prescindere dalla distribuzione originaria di ciascuna di esse.

Nella letteratura anglo-sassone i valori z, che rappresentano scarti dalla media espressi assumendo la σ come unità di misura, sono noti come ***z-scores***.

	pop_i	p_i	$status_i$	$dipl_i$	Pov_i	$affoll_i$
Q1	28125	.148	-0.35	-0.17	0.15	-0.17
Q2	35853	.188	-0.79	-0.71	1.88	1.21
Q3	36169	.190	-0.71	-0.57	-0.48	0.86
Q4	30329	.159	2.19	2.26	-1.29	-1.90
Q5	60028	.315	-0.04	-0.29	-0.40	-0.17
media			0	0	0	0
σ			1	1	1	1

Tabella 2.8 - La tabella 2.7 espressa in termini di *z-scores*

2.2.3 - Esempio 3 - Misura dell'associazione tra variabili continue

Dall'esame della tabella 2.8 emerge con evidenza l'esistenza di un insieme di relazioni tra le variabili:

- quartieri con *status* sopra la media (z-score positivo) tendono ad avere anche *dipl* sopra la media, *pov* e *affoll* sotto la media;
- quartieri con *status* sotto la media (z-score negativo) tendono ad avere anche *dipl* sotto la media, mentre *pov* e *affoll* sono sopra la media;

La cosa ammette delle eccezioni, ma solo per quartieri con z-scores prossimi a zero, cioè con comportamento vicino alla media.

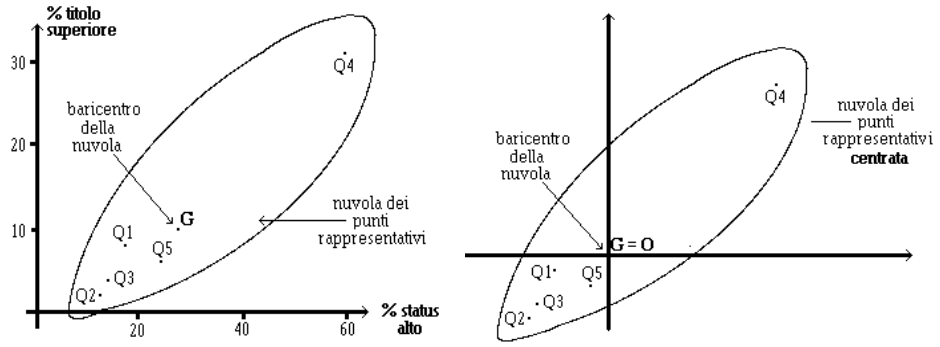
Riportiamo i valori delle due variabili *status* e *affoll* sui due assi di un piano cartesiano: poiché ogni quartiere è descritto da una coppia di tali valori, riportati nella tabella 2.7, esso è univocamente rappresentato da un punto nel piano (*status*, *affoll*). Viceversa, ogni punto del piano individua una coppia di tali valori (le sue coordinate) e dunque un possibile quartiere. In realtà, poiché *status* ed *affoll* non possono essere negative, la collocazione dei punti rappresentativi si limita al primo quadrante.

La figura 2.4a mostra la localizzazione dei punti rappresentativi dei cinque quartieri: si parla di ***nuvola di punti***, la cui distribuzione sul piano è rappresentata schematicamente dall'ellissoide in figura.

Sebbene il termine *nuvola di punti* appaia esagerato quando i punti sono così pochi, nelle analisi reali essi sono molti di più e la nuvola è spesso molto densa.

I punti-quartiere sono dispersi attorno al ***centro di gravità*** **G** della nuvola, le cui coordinate sono i valori medi delle variabili (calcolati ponderando con la popolazione). **G** rappresenta la ***tendenza centrale*** del sistema, cioè il ***comportamento medio*** dell'insieme dei cinque quartieri.

La figura 2.4b rappresenta la stessa nuvola ***centrata***: si sono usati come coordinate gli scarti delle variabili rispetto alla loro media invece che i loro valori iniziali. In pratica, si è operata una traslazione che ha portato l'origine del sistema di coordinate a coincidere con **G**. La trasformazione ha lasciato inalterate la forma della nuvola, la sua dispersione, le distanze tra i punti quartiere.



a) b)

Figura 2.4 - La nuvola di punti (2.4a), e la stessa nuvola centrata (2.4b)

In modo analogo si potrebbe usare qualunque coppia di variabili scelte tra le quattro osservate.

La ***covarianza*** tra due variabili continue *x* e *y* misura la concordanza del loro modo di variare:

$$\mathrm{cov}(x,y) = \sum_i m_i (x_i - \bar{x})(y_i - \bar{y}) \qquad (2.3)$$

Ogni unità statistica *i* dà un contributo *positivo* alla costruzione della covarianza se entrambe le variabili presentano in essa scarti rispetto alla media dello stesso segno (cioè le variabili sono entrambe sopra o entrambe sotto la media, come *status* e *dipl*). Il contributo è *negativo* quando gli scarti hanno segno opposto (ad esempio, *status* e *affoll*).

Si noti che il contributo di ciascuna unità *i* va ponderato con il suo peso m_i.

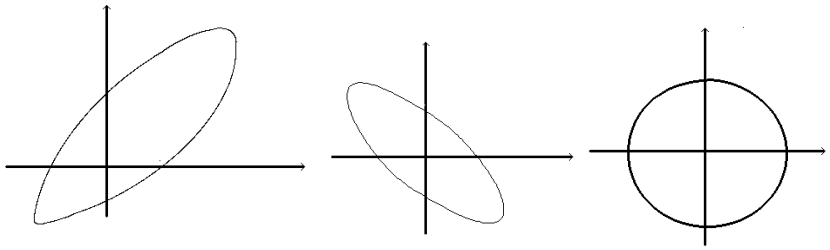
Se le due variabili sono ***centrate*** (cioè se $\bar{x} = 0$ e $\bar{y} = 0$) la (2.3) diventa

$$\mathrm{cov}(x,y) = \sum_i m_i x_i y_i$$

Si possono dare i seguenti casi:

- ***covarianza elevata*** e ***positiva***: le due variabili tendono ad assumere ***insieme*** valori sopra la loro media, oppure sotto. La nuvola ha la forma mostrata in fig. 2.5a.
- ***covarianza elevata*** e ***negativa***: le due variabili tendono a presentare ***scarti di segno opposto*** rispetto alla media (fig. 2.5b).

- ***covarianza trascurabile (prossima a zero)***: scarti positivi di una variabile risultano associati a scarti sia positivi che negativi dell'altra, più o meno in egual misura (fig. 2,5c).



a) b) c)

Figura 2.5 - Diversi tipi di associazione tra variabili continue

In particolare:

- La covarianza di una variabile ***con se stessa*** è null'altro che la sua varianza:

$$\mathrm{cov}(x,x) = \sum_i m_i (x_i - \bar{x})(x_i - \bar{x}) = \sum_i m_i (x_i - \bar{x})^2 = \mathrm{var}(x)$$

- Il valore della covarianza tra due variabili ***non è interpretabile in modo immediato***. Infatti, esso dipende dalla dispersione degli scarti, misurata dalle deviazioni standard σ delle variabili. ***I valori possono cambiare anche solo perché si cambia l'unità di misura***, pur rimanendo immutati i caratteri della relazione descritta.
- Per eliminare l'effetto della diversa dispersione si possono ***standardizzare*** entrambe le variabili prima di calcolarne la covarianza, che viene dunque computata a partire dagli z-scores.

Il valore che si ottiene è indipendente da σ_x e σ_y, poiché entrambe le variabili hanno varianza unitaria dopo la normalizzazione. Tale valore è detto ***correlazione*** tra le variabili:

$$corr(x,y) = \sum_i m_i \frac{(x_i - x)}{\sigma_x} \frac{(y_i - y)}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$$

E' facile mostrare che $-1 \le corr \le 1$.

- Se la correlazione è prossima a 1 le due variabili assumono ***insieme*** valori molto alti o rispettivamente molto bassi (ciascuna rispetto alla propria media).
- Se la correlazione è prossima a -1 le due variabili si muovono in modo opposto: quando l'una assume valori maggiori della sua media, l'altra assume valori minori, e viceversa.
- Se la correlazione è prossima a zero le due variabili non sono significativamente associate.

Nei primi due casi l'informazione apportata dalla seconda variabile ripete in larga misura quella già ricavabile dalla prima.

La tabella 2.9 mostra le *correlazioni* ($*1000$) tra le quattro variabili di tabella 2.7.

- La matrice è simmetrica, cioè per ciascuna coppia di variabili *(i, j)* risulta *corr*(*i*,*j*) = *corr*(*j*,*i*): la correlazione è una proprietà *della coppia* di variabili, misurate su un dato insieme di unità statistiche (quello che a volte chiamiamo il "*contesto*", costituito nel nostro caso dai cinque quartieri), *indipendente dall'ordine in cui le variabili o le unità vengono considerate*.
- La correlazione di una variabile con se medesima è ovviamente 1 (indicato come 1000 nella tabella 2.9).

- *Status* e *dipl* presentano una forte correlazione positiva (= 0.984) e sono entrambe correlate negativamente con *affoll* (le correlazioni valgono -0.948 e -0.915 rispettivamente). *Affoll* e *pov* sono tra loro correlate positivamente, anche se in modo meno forte (= 0.748).
- La forte correlazione tra *status*, *dipl* e *affoll* mostra che l'informazione apportata dalla prima variabile è *quasi ripetuta* dalle altre due. I valori di correlazione non altissimi di *pov* con le altre variabili mostrano la parziale originalità (indipendenza) dell'informazione apportata da questa variabile.

	Status	***Dipl***	***Pov***	***Affoll***
Status	1000	984	-671	-948
Dipl	984	1000	-643	-915
Pov	-671	-643	1000	748
Affoll	-948	-915	748	1000

Tabella 2.9 - Le correlazioni (∗1000) tra le quattro variabili.

2.3 - Tipi di tavole

Oltre a riconoscere la scala di misura in cui ciascuna variabile è espressa è importante anche riconoscere il tipo di tavola che si sta analizzando, dato che tavole di tipo diverso richiedono in generale l'uso di diverse tecniche statistiche. Ci limitiamo qui a distinguere due tipi.

2.3.1 - Tavola di descrizione

Ha tante righe quante sono le unità statistiche (ad esempio, i Comuni) e tante colonne quante sono le variabili da analizzare. Il DataSet di input può includere altre variabili che descrivono le unità, ma non è detto che tutte siano attinenti all'analisi da svolgere: la scelta di quali vadano effettivamente considerate nell'analisi è delicata. Inoltre, va ancora una volta ribadito che le variabili si riferiscono in generale a diversi aspetti e sono misurate su scale diverse; vanno dunque ridotte ad una scala di misura comune prima di poterle analizzare congiuntamente. Una generica *cella* della tavola, localizzata all'incrocio tra la riga i e la colonna j, rappresenta il valore assunto dalla variabile *j* sull'unità *i*.

2.3.2 - Tavola di contingenza (o di conteggio)

È ottenuta contando unità statistiche elementari dello stesso tipo (individui, famiglie, imprese, ecc.) secondo la combinazione dei valori di due ***variabili categoriali*** (o anche più di due, e si otterrà allora una tavola d'incrocio con più di due dimensioni). Come esempio si può riconsiderare la tabella 2.5, oppure la tabella 2.10, che mostra una tavola di contingenza ottenuta incrociando la classe d'età del capofamiglia con la dimensione del nucleo familiare per tutte le famiglie del Centro Storico di Venezia.

Nella tabella 2.10 le unità elementari contate sono le famiglie. Ciascuna è assegnata ad una cella a seconda del numero dei suoi componenti e della classe d'età del CapoFamiglia. Si tratta di due variabili categoriali, entrambe ottenute ricodificando due variabili quantitative sottogiacenti (la

numerosità del nucleo familiare e l'età del CF). Il totale di una colonna conta tutte le famiglie il cui capofamiglia sta in una data classe d'età, indipendentemente dal numero dei componenti. Il totale di una riga conta tutte le famiglie di una data dimensione, a prescindere dall'età del capofamiglia. Questi valori sono noti come i (valori) **marginali** della tavola.

Ricorda

Ogni tavola per la quale abbia senso sommare i valori di una riga o di una colonna può essere pensata come una tavola di contingenza

	< 35 anni	35-55 anni	> 55 anni	totale
1-2 comp.	1341	2852	9689	13882
3 comp.	1680	2924	3521	8125
4 comp.	1325	3693	1749	6767
> 4 comp.	1066	2526	1628	5220
totale	5412	11995	16587	33994

Tabella 2.10 - Un esempio di tavola di contingenza che incrocia la classe d'età del CF e la dimensione del nucleo familiare per tutte le famiglie del Centro Storico di Venezia.

Ricorda

Una tavola formata da variabili di tipo QUANTITATIVE e/o CATEGORIAL con due categorie (che si possono trattare come quantitative), va sottoposta ad un'Analisi in Componenti Principali (ACOMP). Si vedano in proposito la sezione 6.2 e le seguenti.

Variabili di tipo COUNT che costituiscano una o più tavole di contingenza affiancate, o di tipo CATEGORIAL, vanno sottoposte ad un'Analisi delle Corrispondenze (ACORR). Il programma stesso si occupa della ricodifica automatica in forma disgiuntiva completa nel caso di variabili categoriali.

Cap 3. - Il Menu FILE

Consiste delle voci mostrate nella figura 3.1. Il loro uso è illustrato in dettaglio in questo capitolo.

La voce ***Configura*** (seconda dal basso) serve a riconfigurare il programma. Va chiamata immediatamente dopo l'installazione, o in seguito quando sia necessario. È già stata illustrata nel Cap. 1, cui si rimanda.

Figura 3.1 – Il Menu FILE

Scelta della Directory di lavoro

La scelta della ***Directory attiva*** è di solito la prima operazione da fare dopo aver lanciato ADDATI. Quando si parte la directory corrente è quella di installazione del pacchetto, ma è più conveniente scegliere invece come attiva quella che contiene i dati sui quali si deve lavorare.

L'opzione apre un dialogo che permette di scegliere la directory. La scelta viene confermata da un messaggio che appare nella finestra principale di ADDATI. Se un'operazione successiva lamenta di non poter aprire dei file di input o cose simili, è probabile che questo passo sia stato omesso e che la directory di lavoro non sia quella voluta.

La directory scelta viene memorizzata per le sessioni future del programma.

L'operazione successiva è l'apertura dei Set di Dati sui quali si intende lavorare.

3.1 – Set di Dati: descrizione, apertura, chiusura

La prima sezione del Menu FILE riguarda l'apertura e la chiusura di DataSet.

È già stato detto che un DataSet è l'insieme di un **file di dati** (in formato opportuno), **un file di documentazione** che ne descrive il contenuto, ed una **etichetta** (o label, o nome) che lo identifica distinguendolo dagli altri DataSet eventualmente caricati.

3.1.1 - I file di dati

Ogni record consiste di un insieme di variabili (o indicatori) che descrivono una unità statistica (persone, famiglie, unità amministrative...). Le variabili possono essere separate da spazi, virgole, punti e virgola, oppure non avere separatori, nel qual caso ciascuna variabile deve occupare esattamente le medesime colonne in tutti i record per essere distinguibile dalle altre.

Il formato senza separatore è accettato solo in lettura. Si tratta di solito di file molto grandi, al massimo livello di disaggregazione, provenienti da Istituti Centrali di Statistica o prodotti da inchieste molto impegnative. In scrittura, invece, ADDATI usa sempre un separatore.

3.1.2 - I file di documentazione

Il contenuto di un file di dati è descritto dal file di documentazione ad esso associato. Esso ha lo stesso nome del file di dati che descrive, e di solito estensione **.TXT** (non obbligatoria, ma vantaggiosa). Esso può essere preparato usando un normale editor di testo (come il Blocco Note o l'editor interno di ADDATI) e seguendo l'insieme di regole specificate qui di seguito. È previsto (ma non ancora implementato) un aiuto limitato per la sua creazione.

I file di dati prodotti o modificati all'interno del programma (copiando, calcolando o ricodificando variabili; scegliendo un sottoinsieme di casi in base ad una condizione; ecc) vengono salvati insieme ad un file di documentazione prodotto in modo automatico.

3.1.3 - Struttura del file di documentazione di una Data Set

Viene caricato insieme ai dati e fornisce le informazioni necessarie al trattamento delle variabili.

I caratteri '#' o '**;**' (punto e virgola), ovunque si trovino nel record, iniziano un **commento** che dura fino a fine riga ed è **ignorato dal programma**. È opportuno abbondare in commenti per rendere chiaro (all'utente umano) tutto ciò che sia necessario.

Ogni record diverso da un commento consiste di una parola-chiave seguita da uno o più valori che la specificano. Le parole-chiave sono le seguenti (**in rosso e grassetto**):

ARCVIEW_FORMAT (opzionale ed obsoleto) **Specifica se il file dei dati in input è scritto in un formato che lo rende immediatamente caricabile come file di testo da Arc/View (o da EXCEL). In tal caso, il file è facilmente utilizzabile per operazioni di join con i file .DBF che accompagnano gli shapefile cartografici.**

Il file dei dati deve avere un primo record (**ignorato da ADDATI**) con le intestazioni delle colonne (nomi delle variabili) tra doppi apici, ed usare la virgola (COMMA) come separatore di campo.

La parola-chiave ARCVIEW_FORMAT è ancora accettata nei file di documentazione dei DataSet da caricare. Invece, nel salvataggio viene richiesto di enumerare esplicitamente i caratteri del DataSet da scrivere. Infatti, il separatore potrebbe essere il punto e virgola invece che la virgola, il primo record delle intestazioni potrebbe esserci o meno...

Attenzione!

Se viene dichiarato il formato ARCVIEW ma manca invece nel file dei dati il record d'intestazione con i nomi delle variabili, ADDAWIN interpreterà il primo record di dati come se fosse quello d'intestazione, ignorandolo. Ci si troverebbe dunque con un record in meno.

Esempio:

ARCVIEW_FORMAT **YES** # oppure **NO**, che è il default quando manca l'istruzione.

HEADINGS **Serve a specificare se sia presente nel file dei dati un primo record-intestazione che contiene i nomi (brevi) delle variabili tra doppi apici, separati da virgole. Tale record è di solito utile per la piena compatibilità con ARCVIEW o EXCEL.**

HEADINGS YES # oppure NO, che è il default se manca l'istruzione.

N_UNITS (numero delle unità statistiche, opzionale)

Il programma è in grado di determinare il numero dei casi in automatico, leggendo il file. L'istruzione N_UNITS serve quando si voglia caricare solo parte del file, ed è seguita dal numero dei record da caricare. Ad esempio:

N_UNITS 3500 # per caricare solo i primi 3500 record

MISSING_VALUE (codice per i valori mancanti, opzionale)

All'avvio del programma viene caricato dal file di configurazione ADDATI.INI un codice standard da usare per rappresentare i valori mancanti. Tale codice generale può venire modificato editando il file ADDATI.INI, oppure durante l'esecuzione scegliendo dal Menu l'opzione ***File→ Configura*** per aprire il dialogo di configurazione. **Il codice di valore mancante letto da ADDATI.INI viene usato per tutti i Data Set che non ne dichiarino uno loro specifico**, il che può essere fatto inserendo nel file di documentazione l'istruzione

MISSING_VALUE **codice**

dove 'codice' è il valore numerico da usare per i dati mancanti di quel Data Set. Ad esempio:

MISSING_VALUE **-999** # in questo DataSet, e solo in esso, -999 è assunto come dato mancante

FIELD_DELIMITER (separatore di campo utilizzato nel file dei dati)

Il numero delle variabili da caricare è determinato dal programma in base al contenuto del file di documentazione.

Nel record, i valori delle variabili occupano campi separati mediante spazi (SPACE), virgole (COMMA), o punti e virgola (SEMICOLON). Sono questi i soli separatori ammessi quando il Data Set venga salvato, mentre **solo in lettura** è anche ammessa l'assenza di qualsiasi separatore (FIELD_DELIMITER NONE). In quest'ultimo caso tutti i record devono avere esattamente la medesima lunghezza, ed i valori di una variabile si riconoscono dalla loro posizione nel record.

Esempio di istruzione:

FIELD_DELIMITER SPACE # valori possibili: SPACE, COMMA, SEMICOLON, NONE

Seguono poi tanti blocchi introdotti dalla parola-chiave **VARIABLE** quante sono le variabili da caricare.

L'ordine delle parole-chiave è vincolante. Nomi ed etichette devono stare tra doppi apici se includono spazi interni.

VARIABLE nome_esteso [nome_breve] tipo_var [ncat]

'**nome_esteso**' è il nome della variabile. Va incluso tra doppi apici se contiene spazi interni.

'**nome_breve**' è una versione più breve del nome, **opzionale**. Viene usato per intestare le colonne nel caso di salvataggio in formato ARCVIEW o che chieda l'inserimento degli headings. Viene soprattutto usato nelle analisi fattoriali per evitare di rendere troppo confuse le proiezioni sui piani fattoriali. Se manca, viene usato il nome esteso troncato a 12 caratteri, chiedendone conferma all'analista.

'**tipo_var**' identifica il tipo di variabile e può valere CATEGORIAL , QUANTITATIVE, ID o COUNT.

- CATEGORIAL: la variabile è categoriale
- QUANTITATIVE: la variabile è quantitativa, ma **non conta delle unità**. Può essere una variabile costruita, una percentuale, ecc.
- COUNT: conta qualche tipo di unità (persone, famiglie..): assume valori interi. Ad es. la popolazione, il numero dei laureati, il numero degli alloggi prima del 1945, ecc.
- ID: si tratta di un identificatore dell'unità statistica, da non trattare numericamente

'**ncat**' e' il numero delle categorie, **presente (ed obbligatorio) solo per le variabili di tipo** CATEGORIAL

Per pura comodità dell'utente, la parola-chiave 'VARIABLE' può essere seguita, senza spazio di separazione, da un numero d'ordine: quando trova 'VARIABLE23' o 'VARIABLE52' il programma la interpreta come 'VARIABLE' ed ignora il numero.

Variabili di tipo QUANTITATIVE o COUNT

La riga 'VARIABLE' **può** essere seguita da altre righe, tutte opzionali, introdotte dalle parole-chiave CLASSES, THRESHOLDS, WEIGHT e DECIMALS (quest'ultima solo nel caso QUANTITATIVE). Se mancano, vengono applicati i valori di default specificati nel file ADDATI.INI, visualizzabili (e modificabili) dal dialogo di configurazione.

CLASSES è seguito dal numero delle classi da utilizzare nel calcolo della distribuzione della variabile.

THRESHOLDS può essere seguito da:

- 'EQUAL_WIDTH' se il range di valori della variabile va diviso in intervalli di eguale ampiezza;
- 'EQUAL_FREQUENCY' per costruire quantili, cioè classi più o meno di eguale numerosità;
- una sequenza di valori di soglia fissati dall'analista, separati da spazi, coerenti con il numero delle classi richieste. Se c'è disaccordo i valori delle soglie prevalgono, e definiscono implicitamente il numero delle classi.

WEIGHT seguito dal nome della eventuale variabile da usare come peso quando vengono calcolate le statistiche della variabile corrente. Se manca, tutti i casi assumono peso uguale.
La ponderazione delle unità nel calcolo delle statistiche di ciascuna variabile può essere modificata durante l'esecuzione.

DECIMALS specifica il numero di decimali (il default e' 0) **da usare se il DataSet viene salvato** (non quelli che la variabile ha nel file dei dati in input, riconosciuti automaticamente).

Nelle statistiche (media, deviazione standard) il numero dei decimali viene aumentato di uno.

Esempio:

```
CLASSES      8              # vengono richieste 8 classi per le distribuzioni
THRESHOLDS   EQUAL_FREQUENCY  # si vogliono otto classi all'incirca equifrequenti (quantili).
WEIGHT       POPOLAZIONE  # i valori della variabile vanno ponderati con la popolazione
DECIMALS     0            # i valori della variabile vengono scritti senza decimali
```

Variabili di tipo CATEGORIAL

La dichiarazione della variabile è seguita da tanti record **quante sono le sue categorie**. Ogni record ha la forma

VAL code "label"

dove "**code**" è il codice di una categoria, e "**label**" è la sua etichetta. I doppi apici sono necessari solo se la label include degli spazi.

Attenzione: **il codice viene trattato come una stringa ALFANUMERICA, non un numero: "2" e "02" non sono equivalenti.**

La parola-chiave WEIGHT è ammessa anche per le variabili categoriali o binarie, perché può aver senso quando le unità misurate dalla variabile categoriale abbiano un peso (ad es, unita' amministrative, pesate con la loro popolazione o superficie), oppure quando si tratti di un'inchiesta e si voglia applicare alle unità statistiche il loro fattore di campionamento.

L'istruzione seguente è presente solo nel caso in cui FIELD_DELIMITER sia NONE, e serve ad informare il programma sulla posizione di ciascuna variabile nel record:

START_LEN start len

dove '**start**' è (a partire da 1) la colonna (byte) del record dove inizia la variabile e '**len**' è la sua lunghezza di campo, cioè il numero di caratteri che essa occupa.

Un identificatore ID viene trattato come una stringa alfanumerica ed il programma in prima lettura determina automaticamente la sua massima lunghezza.

La parola-chiave '**SKIP**' serve a **saltare variabili** che non si vogliano caricare. La forma è

SKIP **n**

ed il significato dipende dal separatore di campo usato.

- se viene usato un separatore di campo (SPACE, COMMA o SEMICOLON), n è **il numero di campi** da ignorare.
- se FIELD_DELIMITER=NONE, n è **il numero di caratteri** da ignorare in lettura.

Ovviamente una istruzione 'SKIP' dev'essere opportunamente collocata tra blocchi 'VARIABLE'.

3.1.4 - Un esempio di file di documentazione

```
# Il file < COMUNE.DAT > riporta un estratto delle informazioni inclusa nei Fogli di Famiglia di un
# comune dell'Emilia-Romagna, Censimento 1991.
# I record, ciascuno dei quali descrive una famiglia e l'alloggio, sono 11770. Le variabili sono 26.

FIELD_DELIMITER  SPACE

### ALCUNE VARIABILI CHE DESCRIVONO L'EDIFICIO

# Per la variabile seguente viene fornito un solo nome, usato sia come lungo che come breve.
VARIABLE1  "Località"  CATEGORIAL 3
VAL  1  "LOC_centro"
VAL  2  "LOC_nucleo"
VAL  3  "LOC_sparso"

# per la variabile seguente, come in altri casi, vengono forniti sia un nome lungo che uno breve
VARIABLE2  "Struttura portante"  "struttura" CATEGORIAL 5
VAL  1  "STR_cem_pta"          # cemento, p.t. aperto
VAL  2  "STR_cem_ptc"          # cemento, p.t. chiuso
VAL  3  "STR_mattoni"          # pietra,mattoni
VAL  4  "STR_altro"            # altro tipo
VAL  5  "STR_nonind"           # non individuato

VARIABLE3  "alloggi nel fabbricato" "alloggi" CATEGORIAL 7
VAL  1  "ALL1"          # un alloggio
VAL  2  "ALL2"          # due alloggi
VAL  3  "ALL3-4"        # 3-4 alloggi
VAL  4  "ALL5-8"        # 5-8 alloggi
VAL  5  "ALL9-15"       # 9-15 alloggi
VAL  6  "ALL16-30"      # 16-30 alloggi
VAL  7  "ALL>30"        # >30 alloggi

VARIABLE4  "numero piani"  "piani"  CATEGORIAL  5
VAL  1  PIA1            # un piano
VAL  2  PIA2            # due piani
VAL  3  PIA3-5          # 3-5 piani
VAL  4  PIA6-10         # 6-10 piani
VAL  5  PIA>10          # >10 piani

VARIABLE5  "ascensore" CATEGORIAL 2
VAL  1  ASC_si          # presente
VAL  2  ASC_no          # assente

VARIABLE6  "epoca di costruzione" "epcostr" CATEGORIAL 6
VAL  1   EPO<1919       #"prima del '19"
VAL  2   EPO19-45       #"'19-'45"
VAL  3   EPO46-60       #"'46-'60"
VAL  4   EPO61-71       #"'61-'71"
VAL  5   EPO72-81       #"'72-'81"
VAL  6   EPO>81         #"post 81"
```

```
VARIABLE7  "figura proprietario" "figura" CATEGORIAL  9
VAL  1  PRO_persona          # persona fisica
VAL  2  PRO_bank_ass         # banca o assicurazione
VAL  3  PRO_immob            # impresa immobiliare
VAL  4  PRO_altraimp         # altra impresa
VAL  5  PRO_coop             # cooperativa
VAL  6  PRO_EnteTerr         # Ente Territoriale
VAL  7  PRO_EntePrev         # Ente Previdenziale
VAL  8  PRO_IACP             # IACP
VAL  9  PRO_altro            # altro

# ALCUNE VARIABILI CHE DESCRIVONO L'ALLOGGIO, PIU' IL TITOLO DI GODIMENTO

VARIABLE8  "titolo di godimento"  "titgod" CATEGORIAL  3
VAL  1  GOD_proprieta        # "proprietà,usufrutto,riscatto"
VAL  2  GOD_affitto          # "affitto, subaffitto"
VAL  3  GOD_altro

VARIABLE9  "stanze ad uso abit. o promiscuo" "STAnze_ab"  QUANTITATIVE

VARIABLE10  "stanze esclusiv. altro uso"  "STAnze_al"  QUANTITATIVE

VARIABLE11  cucina  CATEGORIAL 5
VAL  1   CUC2+               # "piu' di una"
VAL  2   CUC1_stanz          # "una,stanza"
VAL  3   CUC1_nstanz         # "una,non stanza"
VAL  4   CUC_angcott         # "angolo cottura"
VAL  5   CUC_no              # "manca"

# Per questa variabile vengono fissati dei parametri specifici per il controllo della distribuzione
VARIABLE12  "superficie"  SUPerf    QUANTITATIVE
CLASSES  12
THRESHOLDS  EQUAL_FREQUENCY

VARIABLE13  "acqua potabile" ACQua CATEGORIAL 4
VAL  1  ACQ_int              # "interna"
VAL  2  ACQ_est              # "esterna"
VAL  3  ACQ_pozzo            # "pozzo"
VAL  4  ACQ_no               # "assente"

VARIABLE14  "impianti igienici"  "WC"  CATEGORIAL  5
VAL  1  WC_2+                # "due o +"
VAL  2  WC_1compl            # "uno completo"
VAL  3  WC_1incomp           # "uno, senza sciacquone"
VAL  4  WC_est               # "esterno"
VAL  5  WC_no                # "assente"

VARIABLE15  "vasca o doccia" "BAGno" CATEGORIAL 3
VAL  1  BAG_1                # "una"
VAL  2  BAG2+                # "due+"
VAL  3  BAG_no               # "assente"
```

```
VARIABLE16  riscaldamento  RIScald  CATEGORIAL  5
VAL  1  RIS_centr              # "centralizzato"
VAL  2  RIS_auton              # "autonomo"
VAL  3  RIS_stufet             # "stufe, copertura totale"
VAL  4  RIS_stufep             # "stufe, copertura parziale"
VAL  5  RIS_no                 # "assente"

VARIABLE17  "acqua calda"  acqua_CLD  CATEGORIAL  2
VAL  1  CLD_si                 # "presente"
VAL  2  CLD_no                 # "assente"

VARIABLE18  "telefono"  CATEGORIAL  2
VAL  1  TEL_si                 # "presente"
VAL  2  TEL_no                 # "assente"

####  ALCUNE VARIABILI RELATIVE ALLA FAMIGLIA

VARIABLE19  "num. componenti"  NCOMP  QUANTITATIVE

VARIABLE20  sessoCF   CATEGORIAL  2
VAL  1  SEX_m                  # "maschio"
VAL  2  SEX_f                  # "femmina"

VARIABLE21  EtaCF   QUANTITATIVE

VARIABLE22  "stato civile CF"   "STCiv"   CATEGORIAL  6
VAL  1  STC_celnub             # "celibe, nubile"
VAL  2  STC_coniug             # "coniugato"
VAL  3  STC_sepfat             # "separato di fatto"
VAL  4  STC_sepleg             # "separato legalm."
VAL  5  STC_div                # "divorziato"
VAL  6  STC_ved                # "vedovo"

VARIABLE23  "livello istruz CF"  "titSTUd"  CATEGORIAL  4
VAL  1  STU_laudipl            # "laurea-diploma"
VAL  2  STU_licmed             # "lic.media"
VAL  3  STU_licelem            # "lic.elem."
VAL  4  STU_no                 # "nessun titolo"

VARIABLE24  "cond.prof.CF"  "CondPRof"  CATEGORIAL  8
VAL  1  CPR_occup              # "occupato"
VAL  2  CPR_disocc             # "disoccupato"
VAL  3  CPR_cerca              # "in cerca prima occ."
VAL  4  CPR_casal              # "casalinga"
VAL  5  CPR_stud               # "studente"
VAL  6  CPR_pens               # "pensionato"
VAL  7  CPR_leva               # "militare leva"
VAL  8  CPR_altro              # "altro"
```

```
VARIABLE25  "posiz. prof. CF"  "POSprof"  CATEGORIAL  15
VAL  01 POS_dirig          # "dirigente"
VAL  02 POS_quadro         # "quadro"
VAL  03 POS_impieg         # "impiegato"
VAL  04 POS_interm         # "intermedio"
VAL  05 POS_operaio        # "operaio"
VAL  06 POS_altrodip       # "altro dipendente"
VAL  07 POS_apprend        # "apprendista"
VAL  08 POS_domic          # "lavorante a domic."
VAL  09 POS_milit          # "militare carriera"
VAL  10 POS_imprend        # "imprenditore"
VAL  11 POS_auton          # "lav.autonomo"
VAL  12 POS_libprof        # "libero prof."
VAL  13 POS_soc_coop       # "socio cooper."
VAL  14 POS_coadiuv        # "coadiuvante"
VAL  15 POS_nonocc  #      "non occupato"

VARIABLE26  "Numero occupati"  "noccupati"  QUANTITATIVE
```

3.1.5 - Apri un DataSet

Alla richiesta di caricare un **DataSet** appare il dialogo seguente:

Con i bottoni '**Sfoglia**' vengono visualizzati i file con estensione '.DAT' o 'CSV' (o, rispettivamente, '.TXT' se si tratta del file di documentazione) presenti nella directory attiva. Si può cambiare cartella cercando il file di dati (o di documentazione) desiderato.

Il bottone '**Mostra**' permette di esaminare il contenuto del file di dati. Poiché il file non è stato ancora caricato in ADDATI, la sua documentazione (nomi e tipo delle variabili, ecc.) non è ancora nota al programma; ADDAEDIT non è ancora in grado di mostrare il nome della variabile sotto il cursore.

Il bottone '**Edita**' apre e permette di esaminare e **di modificare** il file di documentazione, prima di caricarlo.

Il bottone '**Crea**' faciliterà la creazione di un file di documentazione (l'opzione non è ancora attiva).

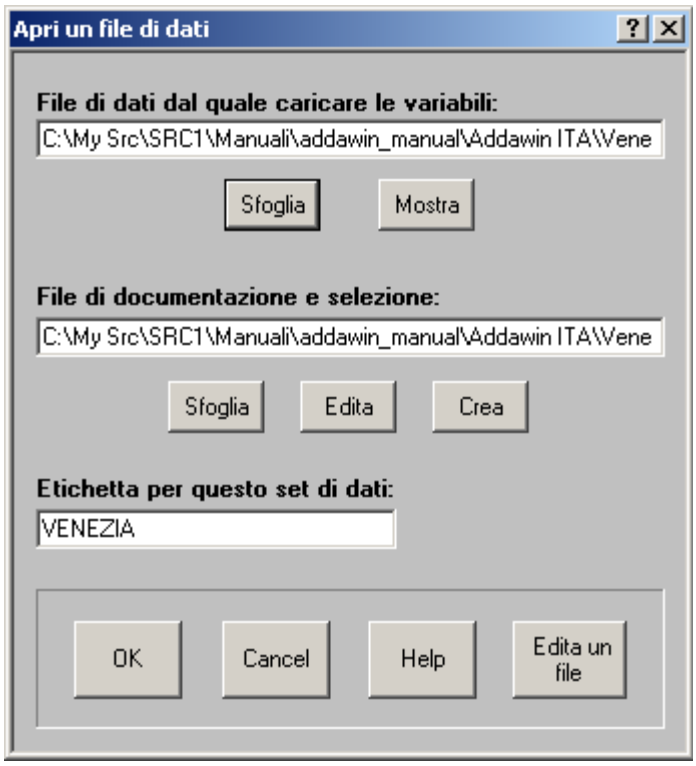


Figura 3.2 – Il dialogo per caricare un DataSet

Se nella directory attiva esiste un file .TXT con lo stesso nome di quello inserito per il file di dati, esso viene automaticamente proposto come il file di documentazione associato. Il nome proposto può essere modificato inserendone o scegliendone uno qualsiasi (ovviamente, deve essere il file di documentazione appropriato per if file di dati che si intende caricare...).

Insomma, non è rilevante il nome, bensì il suo contenuto. Anche le estensioni indicate sono solo dei consigli.

Etichetta (label) del DataSet - Per ogni DataSet che si voglia aprire viene chiesta un'etichetta che lo identifichi in tutte le operazioni successive. Il programma ne propone una, derivata dal nome del file dei dati. Si può confermarla o modificarla.

Si prema poi 'OK'. Il DataSet viene caricato. Gli eventuali errori incontrati vengono elencati in un file che si può visualizzare per effettuare le necessarie correzioni.

Nota *In caso di errore grave (ad esempio se un record è più corto di quanto dichiarato, impedendo un'identificazione sicura delle variabili) il record viene eliminato.*
In caso di errore meno grave (ad es., valori mancanti o fuori codifica) viene usato il codice di valore mancante specificato nel file di documentazione; se questo manca, si usa il codice di valore mancante generale, dichiarato nel file di inizializzazione del pacchetto. In questo caso il record viene caricato.
Anche se vengono segnalati errori il DataSet viene in generale caricato. ***Va dunque chiuso*** *prima di ricaricarlo dopo aver eseguite le correzioni necessarie.*

3.1.6 - Salva/Chiudi un DataSet

Il dialogo in figura 3.3 appare quando si chieda di chiudere e/o salvare su file un DataSet aperto: si può trattare sia di un nuovo DataSet prodotto durante la sessione di lavoro (ad es., ricodificando variabili esistenti), che di un DataSet **caricato da file**, che può essere riscritto in formato diverso, ad esempio cambiando il separatore di campo o il codice per i valori mancanti.

L'operazione è definita dalla combinazione delle spunte nelle due caselle **Salva** e **Chiudi,** ed eseguita con il bottone '**OK**'.

Quando un DataSet viene chiuso la sua label identificativa viene cancellata dalla lista dei DataSet attivi. In caso di salvataggio, ADDATI scrive su file anche la documentazione appropriata.

Una volta salvato o chiuso il DataSet indicato, il dialogo rimane aperto per permettere operazioni

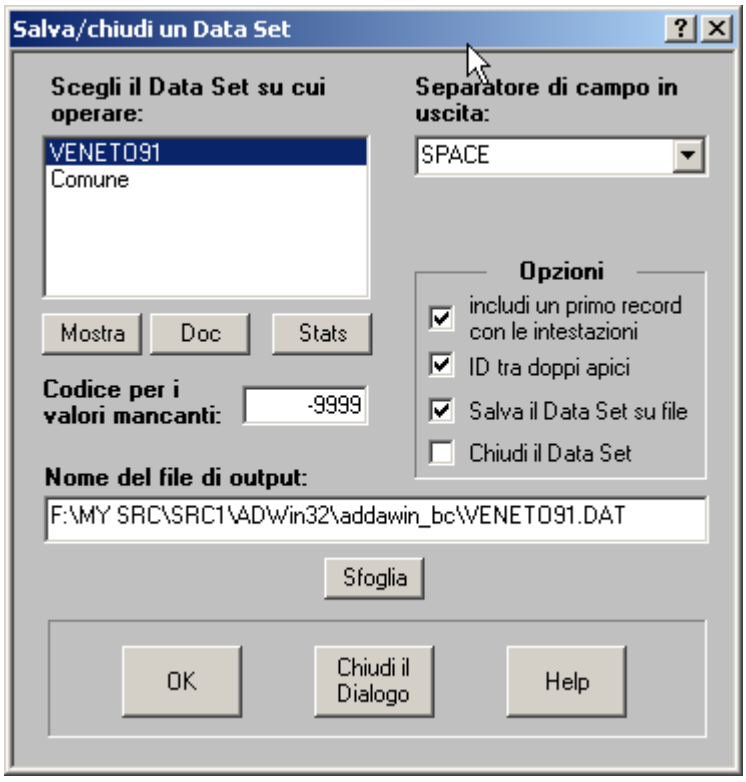


Figura 3.3 – Il dialogo per il salvataggio di un DataSet

analoghe sugli altri DataSet ancora aperti. Il bottone **Chiudi**' chiude il dialogo.

Nota *Quando si salva individualmente un Data Set, le* ***caratteristiche correnti*** *delle sue variabili (numero di classi indicate per le distribuzioni, ponderazione, quantili, soglie, ecc.), se diverse da quelle definite nel file generale di configurazione,* ***vengono memorizzate*** *sul file di documentazione e mantenute quando il Data Set viene ricaricato. Vengono cioè salvate le modifiche introdotte nell'esplorazione delle caratteristiche delle variabili.*

Ricorda *Scegliendo invece l'opzione "**File→Salva tutti i DataSet**", vengono registrati su file solo i* nuovi DataSet *creati durante la sessione di lavoro corrente. Detto altrimenti: le modifiche ai parametri introdotte durante l'esplorazione delle distribuzioni delle variabili vengono mantenute solo se il DataSet interessato viene salvato individualmente (con* File→Salva un Data Set*).*

Separatore di campo - Nel file di uscita le variabili vengono incolonnate e separate con il separatore di campo specificato in questo controllo.

ID tra doppi apici **-** È possibile richiedere che gli Identificatori vengano inclusi tra doppi apici, e vengano dunque considerati come stringhe alfanumeriche se il file viene caricato in Excel o ArcView. Ciò può essere utile se un indicatore deve essere usato per un'operazione di **join**, se nel costruire il file di attributi in ArcView esso è stato dichiarato come stringa. La scrittura tra doppi apici è necessaria per l'utilizzo del DataSet in ADDATI **quando gli indicatori includano spazi interni**.

Aggiunta di un primo record con i nomi delle variabili Il record extra di intestazione, con i nomi brevi delle variabili, è utile nel caso di un successivo caricamento in Excel o ArcView.

Salvataggio in formato adatto per Excel/ArcView **–** Per produrre un file facilmente caricabile da **ArcView o Excel come file di testo**, conviene scriverlo in formato CSV (Comma-Separated Values). Si scelga come separatore di campo la virgola (COMMA) e si spuntino le due caselle '***includi un primo record con le intestazioni***' e '***ID tra doppi apici***'**.** Conviene inoltre assegnare al file di dati che si sta scrivendo estensione .CSV.

Bottoni 'Mostra', 'Doc' e 'Stat'

'**Mostra**' e '**Doc**' mostrano il file di dati e la sua documentazione **nella forma in cui verranno salvati**, che dipende dalla combinazione delle opzioni impostate dall'utente in questo dialogo o in precedenza (codice di valore mancante, separatore di campo, formato CSV per Excel/ArcView, ecc.).

'**Stat**' calcola e mostra le statistiche elementari di tutte le variabili presenti nel file, usando per le distribuzioni e l'eventuale ponderazione i valori correnti dei parametri. Il file di testo visualizzato, contenente le distribuzioni, può essere salvato con un nome opportuno.

La finestra di figura 3.4 è stata ottenuta con il bottone '**Mostra'** dopo aver spuntato la visualizzazione con la virgola come separatore, il primo record con i nomi, la richiesta di indicatori tra doppi apici.

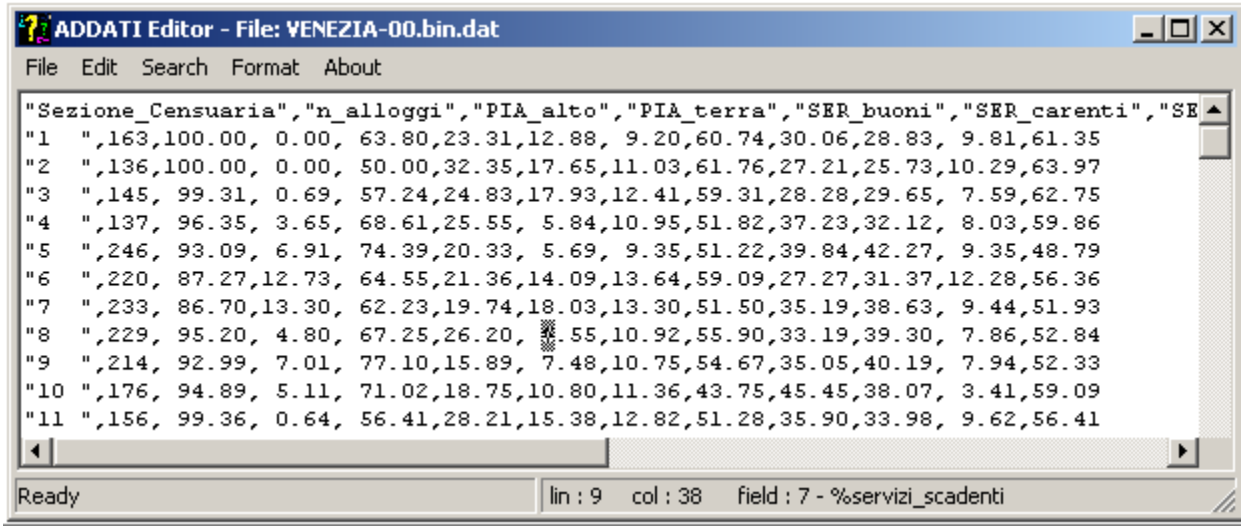


Figura 3.4 – La visualizzazione di un DataSet nel formato scelto per la scrittura

Codice per i valori mancanti - Quando si evidenzia il DataSet su cui operare, questo Edit Box mostra **il suo codice** di valore mancante. Il valore può essere modificato in scrittura.

3.2 – Progetti – Apri e Salva

Se si vuol chiudere una sessione di lavoro per continuare più tardi, e si hanno numerosi DataSet aperti, si può **salvare il progetto**. Ricaricando il file di progetto si eviterà di dover ricaricare i DataSet uno per uno cercandoli nelle cartelle dove si trovano.

Tutto qui. Nessun vantaggio se c'è un solo DataSet aperto: si impiega lo stesso tempo per aprire un progetto o un DataSet. Si fa prima se i DataSet sono più d'uno, e stanno magari in cartelle diverse.

Sia all'apertura che al salvataggio del progetto viene visualizzato un dialogo come quello mostrato nella figura 3.5 qui sotto, che permette di scegliere il file da aprire o da salvare. L'estensione di default per i file di progetto è '.APF' (Addati Project File).

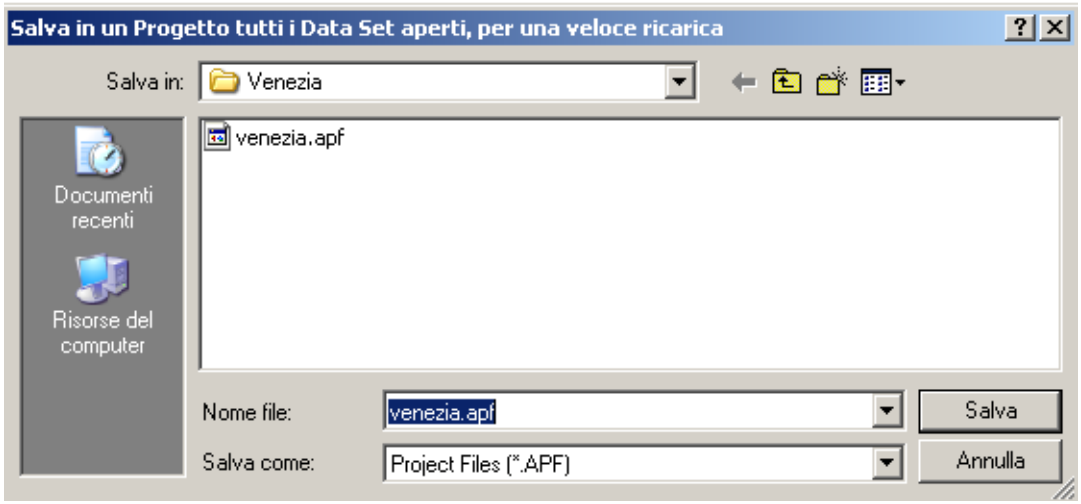


Figura 3.5 – Salvataggio di un insieme di DataSet come Progetto.

3.3 – Le altre opzioni del Menu FILE

La scelta della Directory attiva e l'opzione di Configurazione sono già state trattate.

3.3.1 - Edita/Mostra file di testo

Questa voce evoca un dialogo per la scelta del file di testo da editare (figura. 3.6).

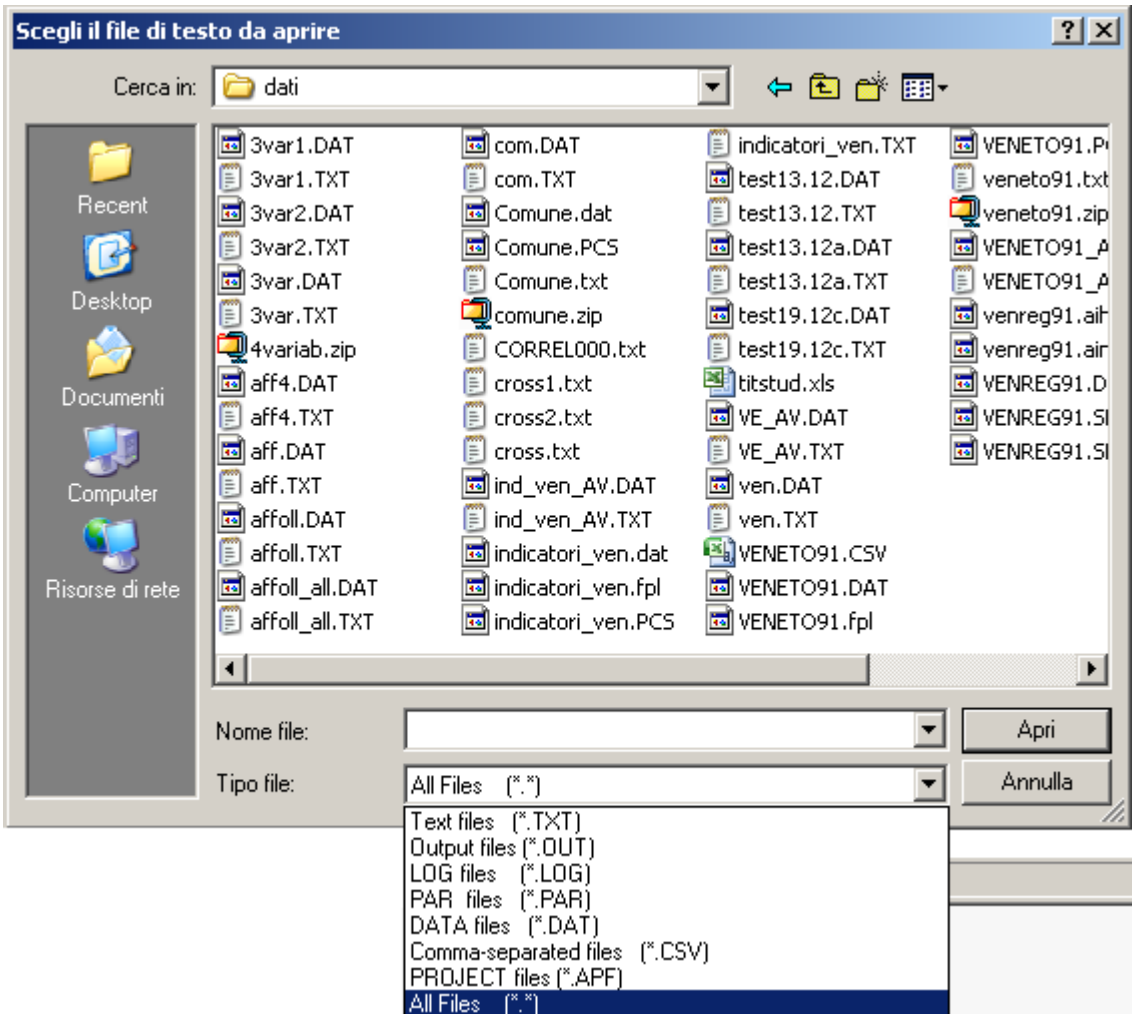


Figura 3.6 - Scelta del file di testo da editare

I file di testo solitamente usati da ADDATI hanno le estensioni mostrate in figura: file di tipo **.TXT** (generici file di testo, o documentazioni di file di dati), file con estensione '**.OUT**' (file prodotti dai programmi di analisi), '**.LOG**' (file LOG di ADDATI), '**.PAR**' (file di parametri, che l'utente riempie appropriatamente per controllare l'esecuzione di alcuni programmi); '.**DAT**' o '**.CSV**' (estensioni comuni per i file di dati). Selezionare l'estensione per visualizzare tutti i file di quel tipo, poi scegliere quello desiderato.

Se si digita "**NEW**" viene aperto un file vuoto.

Il file selezionato viene caricato in **ADDAEDIT**, l'editor interno di **ADDATI**.

ADDAEDIT accetta file molto grandi, e mostra sulla barra di stato la linea e la colonna (carattere) dove sta il cursore.

Inoltre, se il file che si sta editando è un file di dati che usa lo spazio come separatore delle variabili, viene mostrato anche il numero d'ordine del campo sul quale si trova il cursore. I file di dati vengono automaticamente riconosciuti come tali se hanno l'estensione '.DAT' o '.CSV'. Si può comunque forzare la visualizzazione del numero del campo scegliendo la voce "Show Field #" nel menu FILE dell'editor.

Si possono aprire più istanze dell'editor per visualizzare altrettanti file e confrontarne il contenuto. Ad esempio, un file di dati e la sua documentazione; oppure un file di dati e l'uscita di un'analisi, per controllare il comportamento di qualche unità statistica particolare…

ADDATI ed **ADDAEDIT** sono applicazioni indipendenti, ed **ADDATI** può essere usato senza restrizioni anche quando l'editor è aperto.

ADDAEDIT offre alcuni altri vantaggi, legati alla sua seppur limitata capacità di dialogare con **ADDATI**.

3.3.2 - Conversione di file da EXCEL

Questa voce delmenu aiuta l'importazione di file di dati da **EXCEL** in **ADDATI**.

ADDATI richiede file di dati in formato testo. Essi devono contenere esclusivamente dati (nei file .CSV è solo tollerato un record-intestazione con i nomi delle variabili, che viene ignorato), mentre le informazioni complementari (nomi e tipo delle variabili, ecc.) stanno nel file di documentazione associato. Per tale ragione, i file di **EXCEL** devono essere convertiti in formato testo, elimando ogni cosa diversa dai dati.

Il foglio elettronico di EXCEL che contiene i dati va salvato come **file di testo, usando come separatore la tabulazione (file .TXT) o la virgola (file .CSV – Comma-Separated Values)**. Si tratta di due delle opzioni offerte da **EXCEL** quando si scelga "**Save as**" (figura 3.7). **EXCEL** offre anche la possibilità di esportare file di testo che usano come separatore lo spazio (file .PRN: è l'opzione precedente nella figura 3.7), il che sarebbe proprio quel che serve ad **ADDATI**. Purtroppo, tale opzione limita a 240 caratteri (se ricordo bene) la lunghezza dei record che vengono scritti, il che è spesso insufficiente e costringe a lunghe e noiose operazioni di editing. Se invece si usa come separatore la tabulazione (**o meglio ancora la virgola)** l'inconveniente non sussiste.

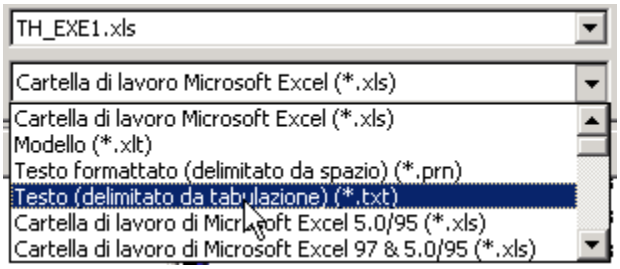


Figura 3.7 - L'opzione **EXCEL** per salvare file di testo con il tab come separatore di campo.

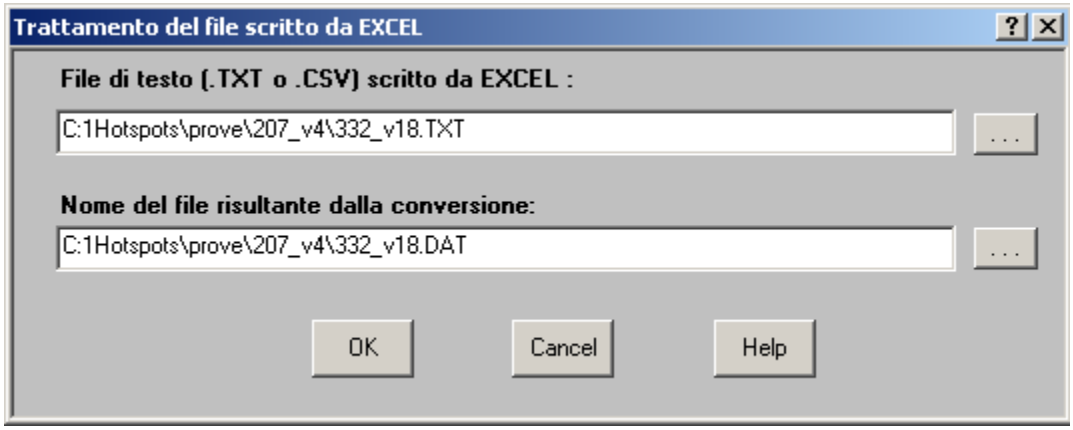


Figura 3.8 - Importazione di dati da **EXCEL**: il dialogo iniziale

Esportati i dati da **EXCEL** (ad esempio come file .TXT), si scelga l'opzione "***File TXT da EXCEL***" dal menu FILE di **ADDATI**. Si apre il dialogo di figura 3.8: si inserisca nell'edit box superiore il nome del file salvato da **EXCEL**, o si sfogli per puntare ad esso. Cliccando sulla casella di editing inferiore, **ADDATI** propone come output del processo di conversione un file con lo stesso nome, ma con l'estensione cambiata in .DAT. Si procede in modo analogo nel caso di file .CSV.

Quando si preme OK appare il messaggio seguente:

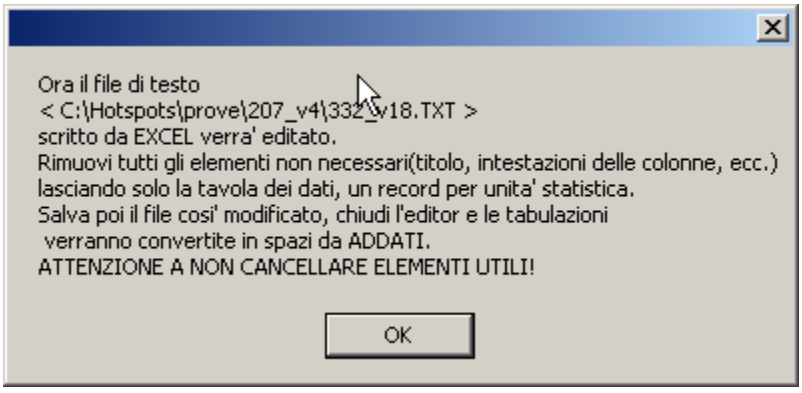


Figura 3.9 – Il messaggio che precede l'apertura nell'editor del file da ripulare

Si prema OK. **ADDATI** carica il file scritto da **EXCEL** nell'editor per permettere all'utente di eliminare il superfluo (documentazione, etichette, ecc.) **lasciando solamente i dati**.

Si salvi poi il file modificato (eventualmente cambiandone il nome, se necessario) **e si chiuda l'editor**: **ADDATI** converte le tabulazioni in spazi, inserisce il codice di valore mancante quando sia necessario (ad esempio, se trova due tabulazioni consecutive senza un valore valido tra di esse), sostituisce con degli underscore gli spazi che trova ***all'interno di uno* stesso *campo***, generando così dei nomi validi. I valori vengono poi incolonnati per permetterne un facile esame. Tutti gli errori trovati durante l'operazione (valori mancanti, record con un numero di campi diverso da quello che il programma si aspetta, ecc.) vengono segnalati.

Attenzione a non cancellare dal file elementi essenziali. In particolare, conviene portare il cursore all'inizio di una nuova linea dopo l'ultimo record, per assicurarsi che questo venga effettivamente terminato con la coppia di caratteri (ASCII 13 e 10) che rappresentano il fine record in Windows.

Attenzione anche a non inserire qualche spazio nella nuova linea, altrimenti essa verrà interpretata da **ADDATI** come un record senza campi, generando un errore.

Attenzione

- Conviene eliminare già da **EXCEL** le eventuali colonne vuote eventualmente inserite come separatori tra blocchi di variabili.
- Si eviti in **EXCEL** ogni ambiguità tra 0 e valore mancante. Il valore 0 va sempre inserito, mentre una cella vuota va intesa come valore mancante.
- L'ultima colonna può porre dei problemi: se contiene delle celle vuote, **EXCEL** tronca la scrittura alla cella precedente e non inserisce una tabulazione finale. Come conseguenza, **ADDATI** trova un campo di meno, e segnala un errore.

  L'ultima colonna va allora controllata con attenzione in **EXCEL**, inserendo (si può farlo con una formula) uno zero oppure il codice di valore mancante nelle celle vuote.

Il file **EXCEL** può anche essere salvato in **formato CSV** (Comma-Separated Values), usando la virgola come separatore. In tal caso ci può essere anche un primo record di intestazione con i nomi delle variabili, purché compresi tra doppi apici. Il formato viene capito da **ADDATI** dichiarandolo opportunamente nel file di documentazione. È anche il formato-testo accettato da ArcView.

Ovviamente, in ogni caso il file .DAT che si ottiene va accompagnato dal necessario file di documentazione, che va scritto prima di poter caricare il DataSet in **ADDATI**.

Nota *Come si è detto, iìl file salvato da EXCEL e sottoposto alla routine di conversione può essere di tipo* ***.TXT*** *o* ***.CSV****. I due tipi di file vengono trattati in modo diverso.*

File che usano la tabulazione come separatore (.TXT)

Viene scritto un file in uscita per il quale viene proposta l'estensione .DAT. Le tabulazioni sono convertite in spazi, inserendo ove necessario i codici di valore mancante.

File che usano la virgola come separatore (.CVS)

Se l'estensione del file di output è anch'essa .CSV*, viene mantenuto il primo record con le intestazioni delle colonne; viene mantenuta la virgola come separatore; vengono inseriti i codici di valore mancante necessari.*

Se invece l'estensione del file di output è .DAT (o qualsiasi altra) ***il primo record con le intestazioni delle colonne viene rimosso****; le virgole sono sostituite con degli spazi; i codici di valore mancante sono inseriti ove necessario.*

Eventuali colonne vuote presenti in EXCEL ed esportate vengono eliminate automaticamente nel processo di conversione.

3.3.3 - Mostra il LOG

Viene mostrato il contenuto del file di LOG dove **ADDATI** registra copia dei messaggi mostrati nella finestra principale.

Cap 4. - Il Menù di Utilità

Il Menù di Utilità è costituito da poche voci, che permettono però di fare molte cose.

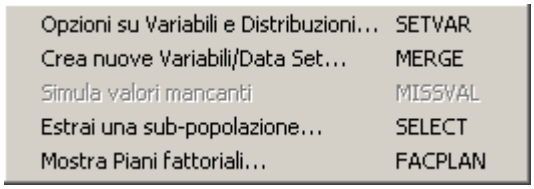


Figura 4.1 - Il Menù di Utilità

Le operazioni possibili sono le seguenti.

- Opzione **SETVAR**: esplorazione delle caratteristiche delle variabili di ciascun DataSet caricato, cambiando a volontà, individualmente per ogni variabile, i parametri che ne controllano la distribuzione: numero delle classi, soglie che le definiscono, ponderazione. Ciò permette di esaminare in dettaglio le distribuzioni, mettendo facilmente in evidenza concentrazioni di valori, possibili *outliers* (cioè valori abnormi, per i quali si può sospettare un errore di qualche tipo).

  Si possono anche calcolare le distribuzioni **limitandole ad un segmento di popolazione** individuato da una condizione particolare. Ad esempio, solo le famiglie al di sopra di un certo reddito; solo gli individui residenti in un certo Comune o Provincia, ecc.
- Opzione **MERGE**: creazione di nuovi DataSet composti da variabili copiate da DataSet esistenti, o da nuove variabili create mediante operazioni matematiche o condizioni logiche a partire da variabili esistenti, oppure ottenute ricodificando variabili esistenti, sia quantitative che categoriali.
- Opzione **SELECT**: costruzione di un nuovo DataSet ottenuto selezionando tra tutte le unità statistiche di un DataSet solo quelle che soddisfano una condizione posta dall'utente.
- Opzione **FACPLAN**: visualizzazione delle proiezioni sui principali piani fattoriali delle unità statistiche e delle variabili. Viene utilizzata dopo un'Analisi Fattoriale o una Classificazione, per l'esame dei risultati.

4.1 – Il calcolo delle distribuzioni

I parametri di default utilizzati per il calcolo delle distribuzioni delle variabili (numero delle classi, soglie, ponderazioni...) sono letti dal file di inizializzazione del programma, e sono modificabili dal dialogo di Configurazione. Essi si applicano a tutte le variabili **per le quali non vengano indicati parametri specifici** nel file di documentazione.

Durante l'esecuzione questo dialogo permette di cambiare - per singole variabili o per gruppi di variabili - i parametri che controllano il calcolo delle statistiche, verificando immediatamente gli effetti sulle distribuzioni che ne risultano. Le distribuzioni secondo tali parametri si possono poi evocare anche da tutti gli altri dialoghi.

Le modifiche restano valide durante la sessione di lavoro, e possono essere consolidate nel file di documentazione se il DataSet viene salvato su disco.

La figura 4.2 illustra il significato dei vari controlli.

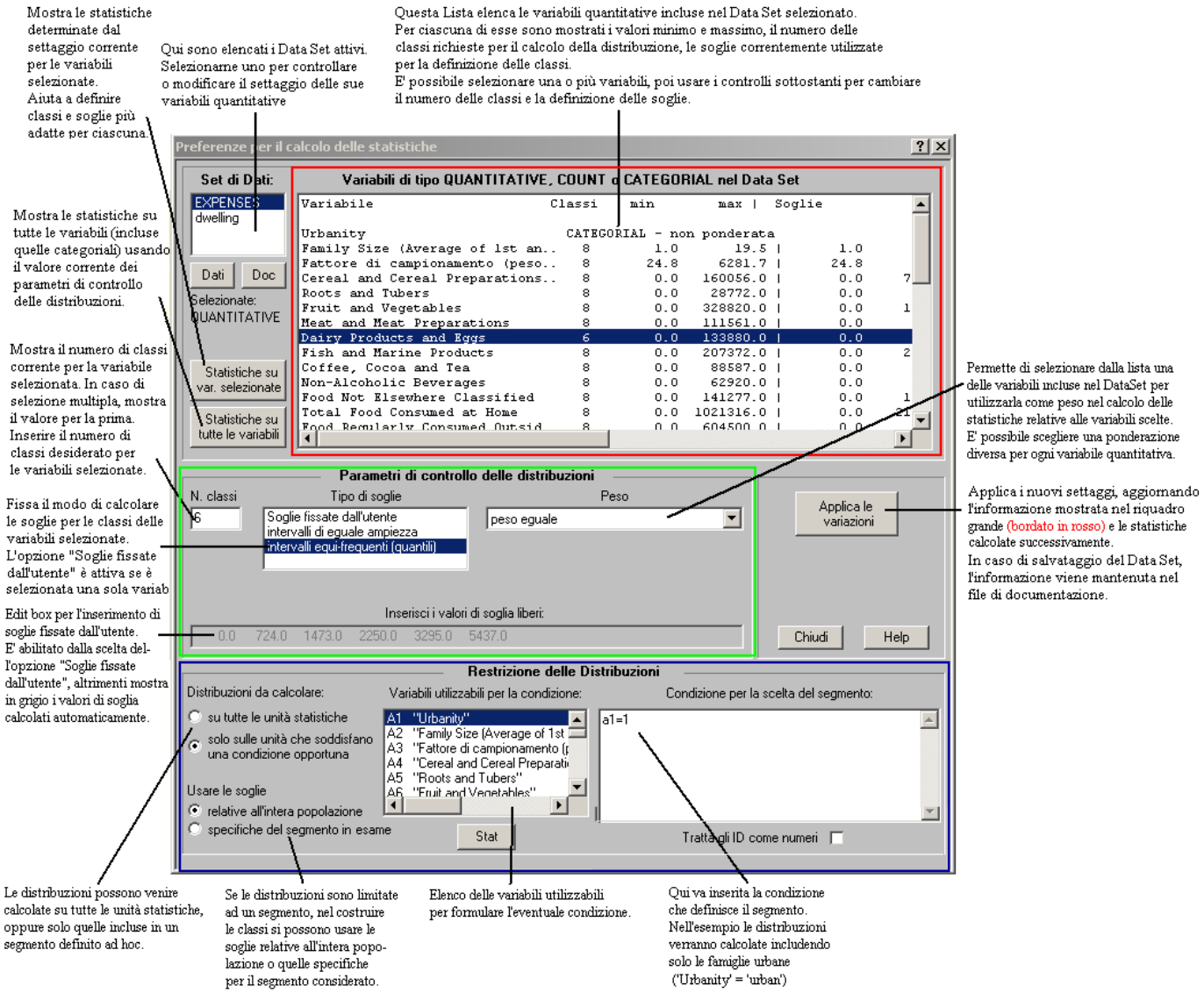


Figura 4.2 - Il dialogo relativo alle operazioni su variabili e distribuzioni, commentato.

Riquadro in rosso: i valori di minimo, massimo e le soglie riportate per le distribuzioni delle variabili sono sempre quelli **relativi all'intera popolazione.**

Riquadro in verde: mostra le preferenze correnti per la variabile (o le variabili) selezionate nella lista riquadrata in rosso.
Le preferenze possono essere modificate separatamente per ciascuna variabile, confermando ogni modifica con il bottone "***Applica le variazioni***". Le opzioni inserite non sono attive finché non vengano confermate.

Riquadro in blu: permette di limitare ad un particolare segmento di popolazione le distribuzioni calcolate. La sintassi per la condizione che individua il segmento è quella comunemente usata in ADDATI.

4.1.1 - Un esempio

Si fa riferimento ad un'inchiesta sui consumi di 39615 famiglie nelle Filippine, condotta dal governo con la collaborazione della World Bank. Le diverse voci di consumo sono espresse in valuta locale ed in valore assoluto. In alternativa, potrebbe essere più utile esprimerle come **percentuali** sulla spesa totale, o come spesa media pro-capite...si tratta di indicatori più chiaramente interpretabili, che si possono costruire scegliendo la voce *Utilità→Crea nuove Variabili/DataSet*.

La tabella 4.1 riporta la distribuzione del consumo di carne e prodotti derivati per l'insieme delle famiglie campionate. Sono state richieste sei classi equi-frequenti, cioè la distribuzione della variabile in sestili. La spesa media globale risulta di 8135 pesos: sono riportate le soglie che definiscono i sestili.

Esaminando la distribuzione si vede che il sesto di famiglie che spende meno per la carne spende una somma compresa tra 0 e 1138 pesos, con una media di 513 pesos circa; il secondo sestile va da 1139 a 2655 pesos, e così via fino all'ultimo sestile, costituito da coloro che spendono di più per la carne, con una spesa di almeno 15.154 pesos. Il massimo risulta di 111.561 pesos, il che è un po' sospetto: si tratta di un valore molto grande rispetto ai consumi medi. Potrebbe valere la pena di indagare ulteriormente isolando con l'opzione SELECT il record interessato, o magari tutte le famiglie con una spesa superiore ad una soglia molto grande, e considerare poi il valore delle altre variabili, per cercare di capire se si tratti di un errore.

```
--------------------------------------------------------------------------------
VARIABLE 7: Spesa per Carne e prodotti derivati
--------------------------------------------------------------------------------
LE DISTRIBUZIONI SONO CALCOLATE SU TUTTE LE UNITA' STATISTICHE.
Casi validi: 39615
Media = 8135.401   Dev.std. = 9096.959   Minimo = 0   Massimo = 111561
Distribuzione in 6 classi - Soglie: intervalli equi-frequenti (quantili)
Peso applicato: Nessuno
     ****************************************
     *MIN CLASSE*      MEDIA * N. DI *  %DI   *
     * (incluso)*     CLASSE * UNITA *  UNITA *
     ****************************************
CL01 *         0 *     512.85 *  6648 * 16.78 * ************************************
CL02 *      1139 *    1863.90 *  6608 * 16.68 * ************************************
CL03 *      2656 *    3790.15 *  6603 * 16.67 * ************************************
CL04 *      5066 *    6789.38 *  6578 * 16.60 * ************************************
CL05 *      8793 *   11635.12 *  6595 * 16.65 * ************************************
CL06 *     15154 *   24325.88 *  6583 * 16.62 * ************************************
     ****************************************
```

Tabella 4.1 - Distribuzione in sestili della spesa per carne e derivati

La tabella 4.2 è stata ottenuta inserendo una condizione, allo scopo di calcolare la distribuzione della spesa per carne e derivati **per le sole famiglie rurali**. Si è richiesto di costruire le classi **usando le soglie di equi-frequenza relative all'intera popolazione.**

Emerge con evidenza come, per quanto riguarda il consumo di carne e derivati, le famiglie rurali si concentrino nelle classi di spesa (relative all'intera popolazione) più basse. La spesa media per famiglia rurale è di 4469 pesos, contro gli 8135 pesos che costituivano la spesa media globale.

Chiedendo l'effettiva equifrequenza, cioè **la costruzione delle classi usando le soglie proprie del segmento**, il confronto dovrebbe focalizzarsi sulla variazione delle soglie, e risulterebbe certo meno immediato.

```
-----------------------------------------------
VARIABLE 7: Spesa per Carne e prodotti derivati
-----------------------------------------------
LE DISTRIBUZIONI SONO CALCOLATE SULLE 16091 UNITA' CHE SODDISFANO LA CONDIZIONE SEGUENTE:
" a1=2 "  ( "Località"="rurale" )
Le classi sono definite dalle soglie di equifrequenza relative all'intera popolazione
Casi validi: 16091
Media = 4468.845   Dev.std. = 5602.514   Minimo = 0   Massimo = 87100
Distribuzione in 6 classi - Soglie: intervalli equi-frequenti (quantili)
Peso applicato: Nessuno
     ****************************************
     *MIN CLASSE*   MEDIA    * N. DI * %  DI  *
     * (incluso)*   CLASSE   * UNITA * UNITA  *
     ****************************************
CL01 *        0   *     526.38 *   4390 *   27.28 * ************************************
CL02 *     1139 *    1847.75 *  3956 *  24.59 * ********************************
CL03 *     2656 *    3738.77 *  3176 *  19.74 * ***************************
CL04 *     5066 *    6586.22 *  2261 *  14.05 * ********************
CL05 *     8793 *   11414.39 *  1458 *   9.06 * ************
CL06 *    15154 *   22211.37 *   850 *   5.28 * *******
     ****************************************
```

Tabella 4.2 - Distribuzione delle **famiglie rurali** sui sestili relativi all'intera popolazione

A completamento del raffronto, la tabella 4.3 mostra la distribuzione della spesa per carne delle famiglie urbane, che si concentrano invece nelle classi più elevate di spesa (la media è 10643 pesos per famiglia).

```
-----------------------------------------------
VARIABLE 7: Spesa per Carne e prodotti derivati
-----------------------------------------------
LE DISTRIBUZIONI SONO CALCOLATE SULLE 23524 UNITA' CHE SODDISFANO LA CONDIZIONE SEGUENTE:
" a1=1 "  ( "Urbanity"="urban" )
Le classi sono definite dalle soglie di equifrequenza relative all'intera popolazione
Casi validi: 23524
Media = 10643.416   Dev.std. = 10119.521   Minimo = 0   Massimo = 111561
Distribuzione in 6 classi - Soglie: intervalli equi-frequenti (quantili)
Peso applicato: Nessuno
     ****************************************
     *MIN CLASSE*   MEDIA    * N. DI * %  DI  *
     * (incluso)*   CLASSE   * UNITA * UNITA  *
     ****************************************
CL01 *        0 *     486.53 *  2258 *   9.60 * ***************
CL02 *     1139 *    1887.97 *  2652 *  11.27 * ******************
CL03 *     2656 *    3837.77 *  3427 *  14.57 * ************************
CL04 *     5066 *    6895.79 *  4317 *  18.35 * ******************************
CL05 *     8793 *   11697.77 *  5137 *  21.84 * ***********************************
CL06 *     15154 *    24639.39 *   5733 *   24.37 * ****************************************
     ****************************************
```

Tabella 4.3 - Distribuzione delle **famiglie urbane** sui sestili relativi all'intera popolazione.

Nota *Le distribuzioni calcolate qui sopra* ***non sono corrette****. Almeno, non nel senso degli obiettivi dell'inchiesta.*

Nel calcolarle, le unità (famiglie) ***non sono state ponderate****. Nel calcolo tutte le famiglie sono considerate equivalenti. I risultati ottenuti valgono allora esclusivamente per l'insieme delle famiglie incluse nel campione, non per il Paese (Filippine) nel suo complesso.*

Il file include anche una variabile detta ***fattore di campionamento*** *(in inglese,* ***'sampling'*** *o* ***'raising' factor****), che misura la* ***rappresentatività*** *di ciascuna famiglia campionata. Il campione è stato scelto in modo da rappresentare al meglio le caratteristiche dell'insieme delle famiglie del Paese (per Regioni amministrative, strati sociali, ecc.). Per cogliere le situazioni meno frequenti (ad es., le caratteristiche delle Regioni più piccole)in modo affidabile è stato allora necessario estrarre per esse un numero* ***percentualmente più elevato*** *di famiglie. In breve,* ***ogni famiglia del campione rappresenta in generale un diverso numero di 'vere' famiglie del Paese con quelle caratteristiche****.*

Se avessimo usato tale valore – il fattore di campionamento, per l'appunto - come peso, avremmo restituito ad ogni famiglia il suo livello di rappresentatività reale, ottenendo risultati validi (almeno approssimativamente) per le Filippine nel loro insieme.

Ricorda *Le diverse opzioni possibili nel calcolo delle distribuzioni forniscono uno strumento potente per farsi un'idea su come vari una variabile sull'insieme delle unità statistiche.*

In generale non basta il calcolo di una sola distribuzione. Per esplorare come i valori delle variabili siano distribuiti è in generale necessario eseguire più prove.
Conviene dapprima calcolare alcune distribuzioni in ***intervalli di uguale ampiezza e con un alto numero di classi****,in modo da individuare in dettaglio sia gli intervalli nei quali cade la maggior parte dei casi che quelli, magari estremi, caratterizzati solo da pochi casi (possibili outliers).*

Solo successivamente si arriverà ad una distribuzione mirata, fornendo delle soglie opportune in modo che nessuna classe risulti troppo vuota o troppo piena, ma evitando però di conglobare in intervalli di grande ampiezza eventuali outliers, come può succedere quando si chiedano classi di eguale frequenza (quantili) senza un'indagine esplorativa preliminare.

4.2 – La costruzione di nuove variabili/DataSet

Scegliendo questa opzione appare il dialogo di figura 4.3. Si vuole **creare un nuovo DataSet** inserendo in esso sia variabili prese pari pari da DataSet esistenti (che debbono ovviamente descrivere le stesse unità statistiche), sia nuove variabili, costruite nei diversi modi di cui diremo.

In ADDATI **non è possibile modificare direttamente le variabili di un DataSet caricato**: bisogna crearne un altro, nel quale si trascrivono le variabili che si vogliono mantenere immutate, e si creano le nuove variabili che si desidera inserire. Esso può poi essere salvato con la sua documentazione, che il programma scrive automaticamente.

Un DataSet già aperto può invece venire salvato con un formato diverso da quello che aveva al caricamento. Ad esempio, cambiando il separatore di campo o il codice di valore mancante, oppure mettendo gli identificatori tra doppi apici, aggiungendo o eliminando il primo record delle intestazioni di colonna, ecc.

In alto a sinistra nella figura 4.3 appaiono, riquadrate in giallo, due finestre. Quella di sinistra elenca tutti i DataSet caricati: per ciascuno di essi si possono visualizzare i dati, la documentazione e le distribuzioni di tutte le variabili usando i tre bottoni a lato della finestra.

Vanno selezionati dalla finestra di sinistra i DataSet che contengono le variabili necessarie per le operazioni che si vogliono compiere; essi vengono trascritti nella finestra di destra, ed a ciascuno di essi viene assegnata una lettera identificativa. Cliccando su uno di essi, l'elenco delle sue variabili appare nella finestra riquadrata in verde.

Al DataSet che risulterà dall'operazione va assegnato un nome, che va inserito nella finestra riquadrata in rosso in figura 4.3.

La finestra intestata ***Output Data Set*** elenca le variabili del DataSet di output man mano che vengono definite.

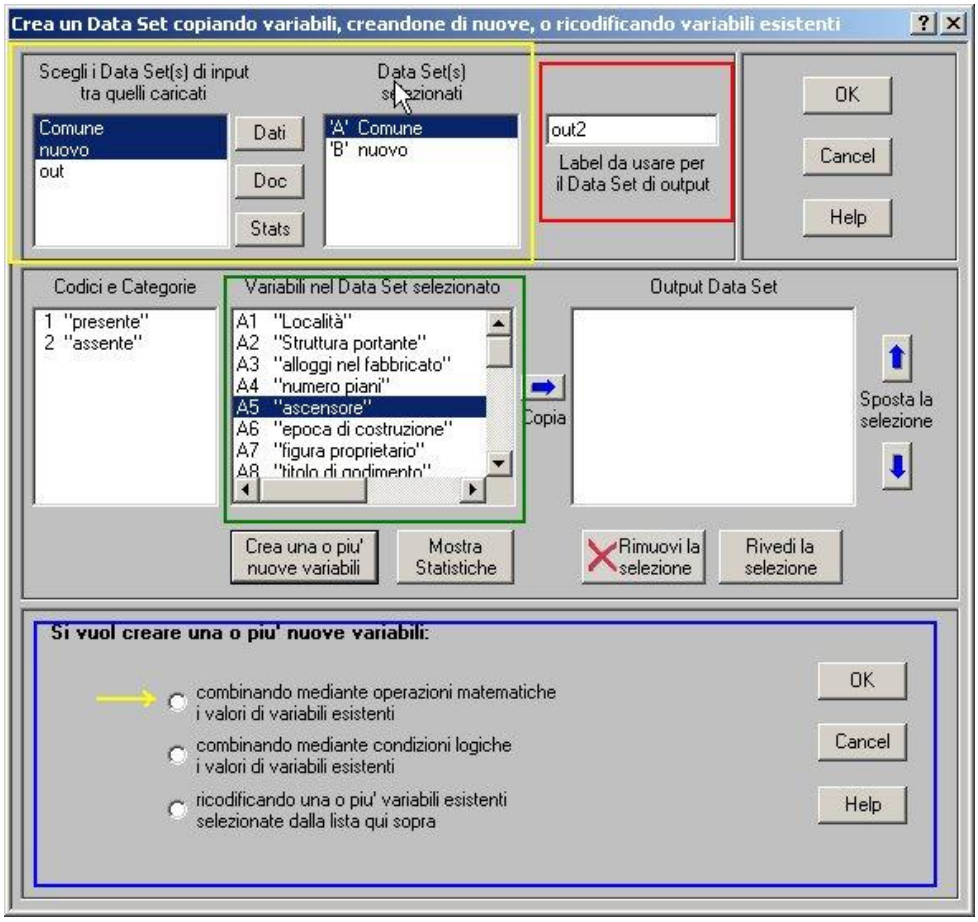


Figura 4.3 – Il dialogo per la costruzione di nuove variabili/DataSet.

Per **trascrivere variabili esistenti** basta selezionarle nella lista riquadrata in verde e poi cliccare sul bottone ***Copia***. Sono ammesse scelte multiple, usando insieme al mouse i tasti CTRL o SHIFT.

Si possono copiare variabili da ciascuno dei DataSet presenti nella finestra di destra, riquadrata in giallo. Le variabili trascritte possono essere eliminate, o se ne può cambiare l'ordine (che è quello nel quale saranno trascritte nel DataSet di output.

Per aggiungere nuove variabili create allo scopo, va premuto il bottone ***Crea nuove variabili***, sotto la lista riquadrata in verde. Sotto appaiono allora le tre opzioni possibili riquadrate in blu in figura 4.3, non visibili all'apertura del dialogo. Esse verranno ora illustrate in dettaglio.

4.2.1 - Costruzione di nuove variabili mediante operazioni matematiche

L'operazione riguarda un insieme di variabili selezionate, che debbono essere tutte di tipo **QUANTITATIVE** o **COUNT**. La scelta è indicata con la freccia gialla nella figura 4.3.

Per questa operazione si possono selezionare dalla finestra in alto a sinistra fino a nove DataSet, che contengono le variabili che si vogliono utilizzare e vengono trascritti nella finestra di destra. Selezionandone uno in tale finestra, le sue variabili appaiono nella finestra riquadrata in verde.

Una volta che si cominci a scegliere le variabili da trascrivere, ricodificare, o da utilizzare per crearne di nuove, **i DataSet selezionati non possono più venire cambiati** pena l'imprevedibilità dei risultati, o il crashing del programma.

Ogni variabile è individuata dalla lettera del DataSet che la contiene, seguita dal suo numero d'ordine in esso: '**A1**' indica il primo campo (variabile) del primo DataSet, '**B7**' indica il settimo campo del secondo DataSet, e così via. L'utente userà questi simboli quando inserisce nell'apposita casella di Edit l'espressione matematica che calcola il valore di una nuova variabile a partire da quelli di alcune variabili esistenti.

La casella di Edit (figura 4.4) appare spuntando il bottone radio indicato dalla freccia gialla in figura 4.3, e premendo OK.

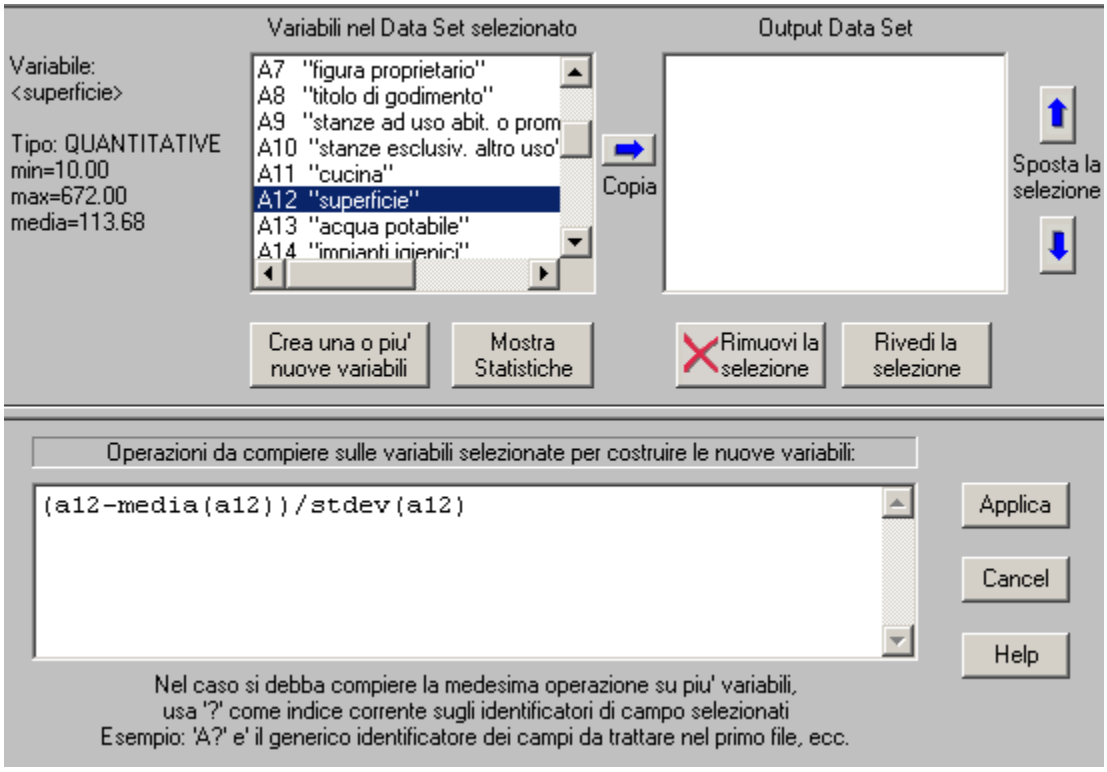


Figura 4.4 - L'espressione inserita in questo esempio standardizza la superficie (variabile A12).

L'espressione inserita può includere:

- **Numeri** ed **identificatori di variabili** (come ‘A12’ o ‘B7’);
- **operatori aritmetici** (+, -, *, / per le quattro operazioni aritmetiche, ' ^ ' per l'elevazione a potenza);
- **funzioni** ('**sqrt**' per radice quadrata, '**log**' per il logaritmo decimale, '**ln**' per il logaritmo naturale, '**exp**' per l'esponenziale);
- **le tre funzioni: 'SOMMA', 'MEDIA' e 'STDEV'** (precalcolate), che rappresentano rispettivamente la somma, la media e la deviazione standard della variabile cui si riferiscono, calcolate su tutte le unità statistiche (non su di un segmento particolare). MEDIA e STDEV sono ponderate se alla variabile è associato un peso.

Si possono usare parentesi rotonde per incrementare la priorita`. Ad uno stesso livello di parentesi un'espressione è valutata da sinistra a destra rispettando l’ordine di precedenza seguente: prima vengono calcolate le funzioni ln, log, exp e sqrt; poi l'elevazione a potenza ^, poi le moltiplicazioni e divisioni (nell'ordine in cui si presentano), ed infine le somme e sottrazioni.

Esempi

1. **sqrt((a1+a2)/2)** calcola, record per record, la radice quadrata della media aritmetica dei valori delle due prime variabili del primo Data Set, assegnando il risultato alla variabile da calcolare.
2. **(B3-MEDIA(B3)) / STDEV(B3)**
   standardizza la variabile B3 (terza variabile del secondo dei DataSet selezionati), assegnando alla nuova variabile, per ogni unità statistica, il valore di z-score che risulta dal calcolo. Nell'esempio mostrato in figura 4.4 viene chiesta la standardizzazione della variabile A12 (superficie dell'alloggio).
3. Si supponga che il Data Set descriva tutti i Comuni di un Paese (un record per Comune), e che la terza variabile rappresenti il numero dei residenti laureati, mentre la decima rappresenta la popolazione totale. L'espressione:

   (A3/A10)*100

   converte il numero assoluto di laureati (che è strettamente correlato alla dimensione del Comune) nella percentuale calcolata su tutti i residenti, interpretabile con maggior chiarezza.

Estensione a più variabili

Per ripetere la medesima procedura di calcolo su più di una variabile, si selezionano dalla lista tutte le variabili alle quali si vuole applicare la definizione e si usa il wildchar '?' al posto del numero della variabile.

Si supponga che in ciascun Comune A3 sia il numero dei laureati, A4 il numero dei diplomati, A5 il numero dei residenti forniti di licenza media....fino ad A9 che rappresenta il numero degli analfabeti. Per convertire tutti questi valori assoluti in percentuali si selezionano dalla lista tutte le variabili da A3 ad A9 (sulle quali bisogna operare), poi si inserisce

(A?/A10)*100

cioè la stessa espressione di prima, usando però un punto interrogativo al posto del ‘3’. Questo permette l'estensione della definizione a più variabili.

Il programma aprirà altri dialoghi per raccogliere informazioni sulle nuove variabili, che serviranno per trattarle e documentarle correttamente.

> ***Nota*** *La routine che controlla la correttezza dell'espressione richiede che essa* ***effettivamente includa qualche operazione*** *da eseguire.*
>
> *Se si vuole ad esempio includere nel Data Set che si sta costruendo una colonna costituita da un numero fisso (ad es. una colonna di 10), oppure la media globale di una variabile, espressioni come '10' o 'media(a12)' non vengono accettate.*
>
> *Conviene in tal caso 'ingannare' il programma* ***simulando un'operazione fittizia****: ad esempio, inserendo '****10 * 1****' e '****media(a12) * 1.0****'.*

Il dialogo "Definizione di nuove variabili"

Ad un certo punto della procedura di costruzione di una nuova variabile mediante operazioni matematiche viene presentato un dialogo come quello di figura 4.5, che mira a definirne le caratteristiche.

L'Edit Box in alto mostra la definizione inserita dall'utente, o adattata nel caso sia stata fornita una definizione contenente degli wildchars '?', da applicare a più variabili. Si può controllarla, ma anche modificarla.

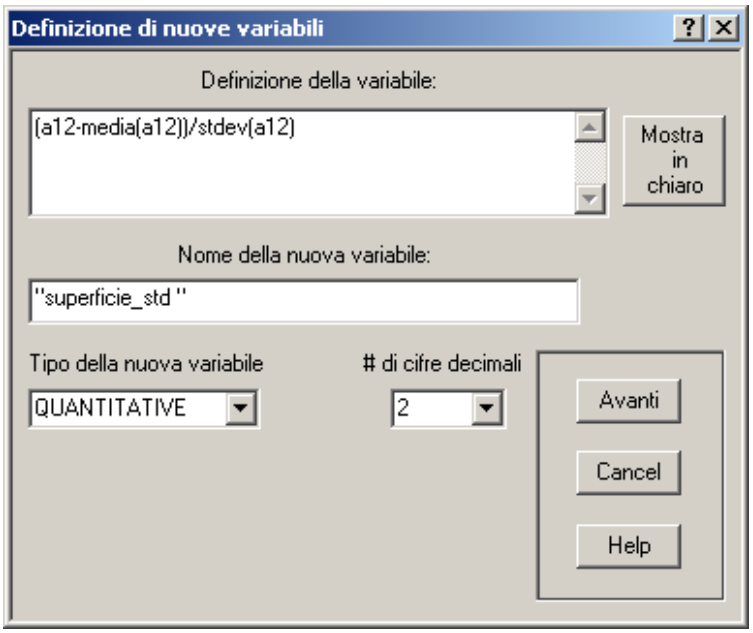


Figura 4.5 – Il dialogo che si apre per definire le caratteristi-che di una nuova variabile

Se le variabili da trattare sono marcate nella lista delle variabili del Data Set considerato, i loro nomi vengono presentati in sequenza nella casella di editing "**Nome della nuova variabile**" e si possono modificare per formare i nomi delle nuove variabili che si stanno creando. Se nessuna variabile è stata selezionata esplicitamente, e si è solo inserita la condizione che definisce quella nuova che si sta creando, viene presentato tra doppi apici un nome vuoto, da riempire.

La variabile risultante dal calcolo è di solito di **tipo** QUANTITATIVE, ma in qualche caso potrebbe essere di tipo COUNT (ad esempio, se si sommano due variabili di tipo COUNT, come "laureati" e "diplomati"). Una casella apposita consente di fissarne il tipo, mentre quella a fianco permette di stabilire il numero dei suoi decimali (si sceglierà 0 se il risultato è un intero).

Il bottone **Avanti** passa alla variabile successiva (se ne è stato impostato un gruppo), o chiude il Dialogo tornando a quello principale.

Il bottone **Cancel** interrompe l'operazione cancellando la definizione della variabile corrente, che viene ignorata. Se è stato ultilizzato il '?' per definire simultaneamente molte variabili, si passa alla successiva.

4.2.2 Costruzione di una nuova variabile mediante condizioni logiche

La figura 4.6 evidenzia la scelta di costruire una nuova variabile CATEGORIALE (il risultato di questa operazione ***è sempre una variabile CATEGORIALE***) mediante condizioni logiche poste sui valori di una o più variabili esistenti.

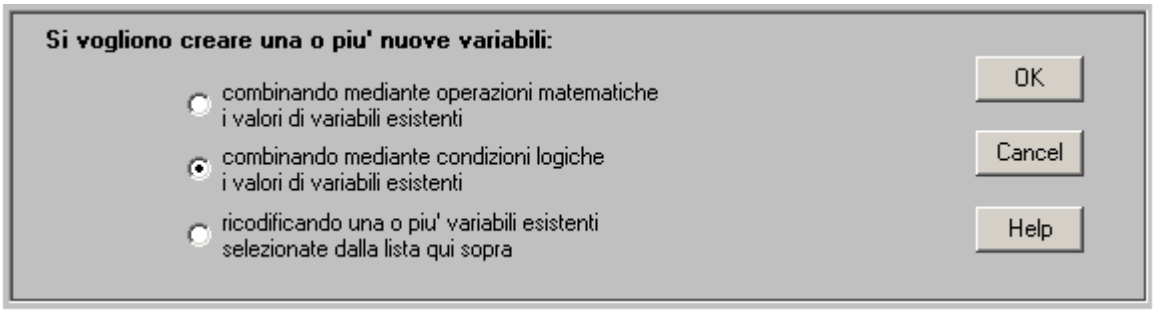


Figura 4.6 – Le tre opzioni per la costruzione di nuove variabili.

In **ciascuna riga** della casella di edit di figura 4.7, che appare quando si preme il tasto OK, va inserita **una regola** che associ un valore ben definito della nuova variabile ad una data combinazione di valori di alcune variabili esistenti.

Esempio

Si voglia creare una nuova variabile che combini il titolo di godimento dell'alloggio (variabile A8: 1=proprietà; 2=affitto) con la sua superficie (variabile A12, quantitativa, mq). Le istruzioni

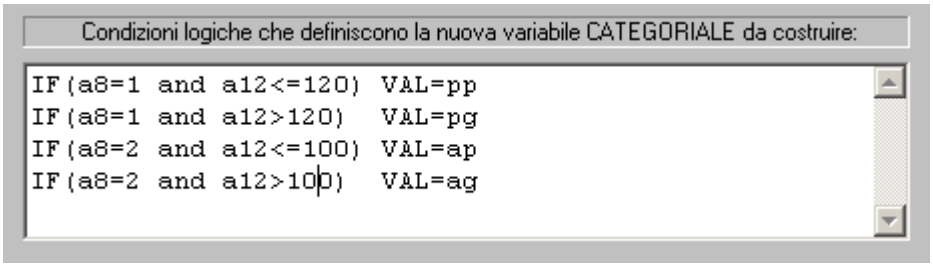


Figura 4.7 – Esempio di regole

mostrate definiscono una nuova variabile con quattro categorie rappresentate dai codici "pp", "pg", "ap", "ag" (stanno per **p**roprietà-**p**iccolo, **p**roprietà-**g**rande, **a**ffitto-**p**iccolo...). I codici indicati verranno usati per le categorie della variabile nel nuovo DataSet, e consolidati su file quando il DataSet venga salvato (potrebbero invece essere "1", "2", "3", "4", più comuni ma magari in questo caso meno chiari). Si noti che il valore di soglia scelto per valutare l'alloggio come piccolo o grande è di 100 mq nel caso dell'affitto, di 120 mq nel caso della proprietà.

Una volta inserito l'insieme delle regole, si clicchi sul bottone ***Applica*** per aprire un dialogo che consente di inserire il nome della nuova variabile e le label delle sue categorie. I codici, cioè i valori che verranno usati nel nuovo DataSet, sono i valori assegnati a VAL nelle regole.

Definizione di nuove variabili mediante condizioni logiche – La sintassi

Va digitato un insieme di regole di costruzione che debbono definire esaustivamente tutte le categorie della nuova variabile che si sta creando. La figura 4.8 mostra un esempio.

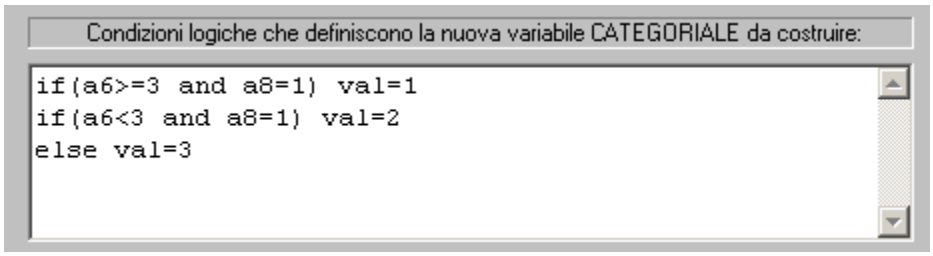


Figura 4.8 – La sintassi delle condizioni logiche

La struttura è la seguente:

IF (condizione1) VAL = "c1"

IF (condizione2) VAL = "c2"

IF (condizione3) VAL = "c3"

.....

ELSE VAL = "cn"

- **IF**, **VAL** ed **ELSE** sono parole-chiave (maiuscolo e minuscolo sono equivalenti).
- **La riga ELSE è opzionale**. Fornisce un codice residuale, da usare quando nessuna delle condizioni già elencate si verifichi.
- Se nessuna condizione si verifica, e non è stato fornito un valore ELSE, viene usato il codice di default per i valori mancanti (missing_value).
- Le **condizioni** (che per chiarezza si raccomanda di scrivere tra parentesi) sono esattamente del tipo usato per la selezione di un insieme di unità statistiche. Possono essere anche molto complesse, e coinvolgere più variabili.
- I codici assegnati per la variabile che si sta costruendo (indicati sopra con "**c1**"..."**cn**") sono **sempre letti come stringhe alfanumeriche**, anche se si tratta, come d'uso, dei valori ***apparentemente*** numerici "1", "2" ecc. I doppi apici sono dunque opzionali e si possono omettere.

Attenzione!

Almeno per ora, la coerenza delle condizioni poste non viene controllata. Se per errore due condizioni sono parzialmente sovrapposte, cioè risultano entrambe vere per qualche combinazione delle variabili di input, scatta la prima. In altri termini, per ogni unità statistica le diverse condizioni vengono valutate nella sequenza in cui sono state fornite, e la prima che viene trovata vera determina il valore assegnato alla variabile che si sta costruendo.

Altri e più interessanti esempi di costruzione di nuove variabili mediante la formulazione di condizioni logiche verranno proposti nel capitolo 5, dopo aver trattato gli Incroci tra variabili categoriali.

4.2.3 - Ricodifica di variabili

Scegliendo questa opzione (la terza tra quelle possibili per costruire nuove variabili), appare un dialogo, diverso a seconda che la variabile da ricodificare sia quantitativa o categoriale.

Ricodifica di variabili quantitative

Il dialogo di figura 4.9 viene presentato per ciascuna delle variabili **QUANTITATIVE** o **COUNT** da ricodificare. La ricodifica produce una variabile **categoriale**, ottenuta ***dividendo opportunamente in classi*** i valori della variabile di partenza.

Il dialogo mostrato ha funzioni esplorative e serve ad orientare l'utente mostrando l'esito di diverse scelte possibili nell'eseguire la ricodifica. Una volta deciso come farla, segue un altro dialogo per la definizione dei codici e delle label delle categorie che verranno create.

Nel dialogo vanno scelti

- il numero delle classi desiderate (cioè delle categorie della variabile che si sta costruendo);
- il modo di fissare le soglie;
- l'eventuale peso da applicare nel calcolo delle distribuzioni esplorative.

Scegliendo l'opzione '*Soglie fissate dall'utente'* nella finestra '*Tipo di soglie*' **viene abilitata** la finestra di editing più in basso, che mostra le '*soglie correnti*". Si possono allora inserire direttamente le soglie desiderate, spesso determinate come esito di alcune prove preliminari con soglie automatiche per controllare la distribuzione.

Viene proposto un nome per la nuova variabile che si sta creando: si può accettarlo o modificarlo.

Riempito il dialogo, il bottone "*Applica modifiche - Mostra Statistiche*" mostra la distribuzione della nuova variabile che risulterebbe dalla scelta, se confermata. Quando si è soddisfatti dell'esito, il bottone "*Codici e nomi*" apre un nuovo dialogo per accettare o modificare i codici ed i nomi che vengono automaticamente proposti per la variabile categoriale che si sta creando.

Se si stanno ricodificando più variabili, selezionate nella lista nella finestra di dialogo principale, il bottone '**Ignora variabile**' cessa il trattamento della variabile corrente e passa alla successiva; il bottone '**Cancel**' chiude invece il dialogo, ignorando tutte le variabili non ancora trattate.

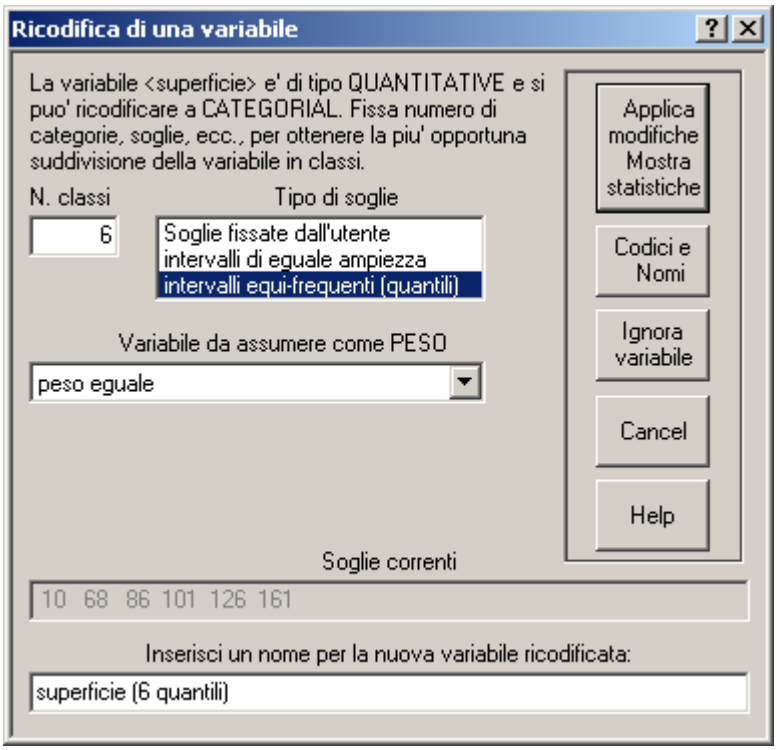


Figura 4.9 – Il dialogo per la ricodifica di una variabile quantitativa

Ricorda: *Le modifiche apportate in questo dialogo ai parametri distributivi (numero di classi, tipo di soglie, ecc.) della variabile da ricodificare hanno scopo puramente esplorativo e* ***non vengono mantenuti****. Per ottenere delle modifiche permanenti usare l'apposito dialogo* ***Utilità → Opzioni su variabili e Distribuzioni***

Ricodifica di variabili categoriali

Il dialogo di figura 4.10 viene presentato per ciascuna delle variabili **categoriali** da ricodificare.

La ricodifica **aggrega opportunamente le categorie della variabile di partenza** e produce una nuova variabile categoriale. Si può usare l'opzione anche solo per modificare i codici di una variabile categoriale esistente, e/o le etichette che le identificano, senza cambiare le categorie.

Nella lista a sinistra vengono presentate le categorie della variabile da ricodificare. All'inizio sono tutte, poi restano man mano quelle non ancora assegnate ad una classe target.

La lista centrale mostra le nuove classi ottenute aggregando le categorie di partenza. Evidenziando una classe (nella figura 4.10 è evidenziata la Classe1) la lista di destra visualizza le categorie già assegnate ad essa.

Si procede selezionando a sinistra una o più categorie da aggregare. Al centro si evidenzia la classe (una già esistente, o quella nuova) alla quale vanno assegnate. Premendo la freccia si effettua l'assegnazione. Si continua fino a che la lista di sinistra non sia vuota, cioè tutte le categorie siano state assegnate.

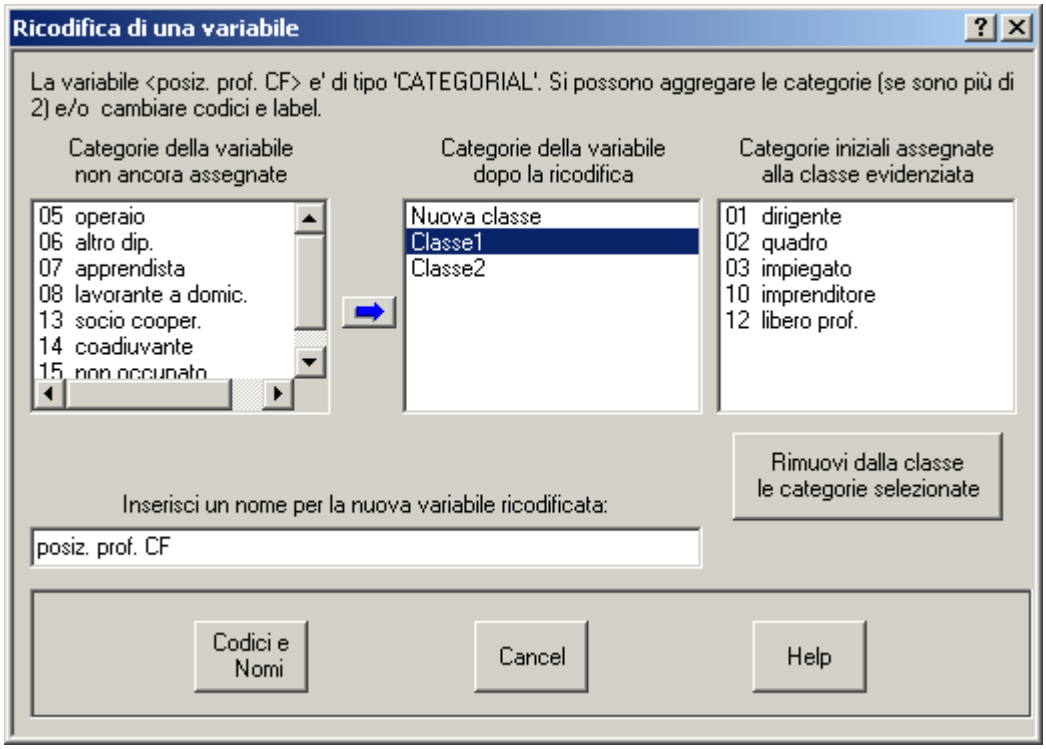


Figura 4.10 – Il dialogo presentato per la ricodifica di una variabile categoriale

La lista di destra mostra le categorie assegnate alla classe selezionata nella lista centrale. Si possono marcare delle categorie e rimuoverle usando l'apposito bottone: esse ritorneranno nella lista di sinistra e sarà possibile riassegnarle ad un'altra classe.

Completata l'assegnazione, si clicchi sul bottone "***Codici e nomi*** " per passare a definire i codici e le etichette della nuova variabile aggregata.

L'assegnazione dei codici e delle label alle categorie

Una volta scelto come ricodificare la variabile, premendo il bottone ***Codici e Nomi*** appare un dialogo come quello di figura 4.11 per l'inserimento dei codici e delle label delle nuove categorie. Lo stesso succede se si ricodifica una variabile quantitativa, ma il dialogo è in quel caso un po' più semplice.

Nella finestra di sinistra vengono proposti, tra doppi apici, dei codici e dei nomi per le nuove categorie. I codici sono numeri da ***1*** a ***n*** se vi sono ***n*** nuove categorie, ma i doppi apici servono a ricordare che essi vengono in realtà trattati come entità alfanumeriche.

L'utente può accettare i codici (l'uso di codici numerici è molto comune), ma è invece opportuno adattare le label: 'Classe1' ecc. sono denominazioni troppo generiche, meglio sostituirle con altre più aderenti al particolare significato di ciascuna categoria.

La finestra di destra è informativa, e mostra le categorie originali assegnate alla classe sotto il cursore.

Anche se non è sempre strettamente necessario mantenere i doppi apici, vi sono situazioni in cui va fatto, quindi meglio tenerli.

Definiti codici e label, il bottone OK conferma e passa a trattare l'eventuale variabile successiva.

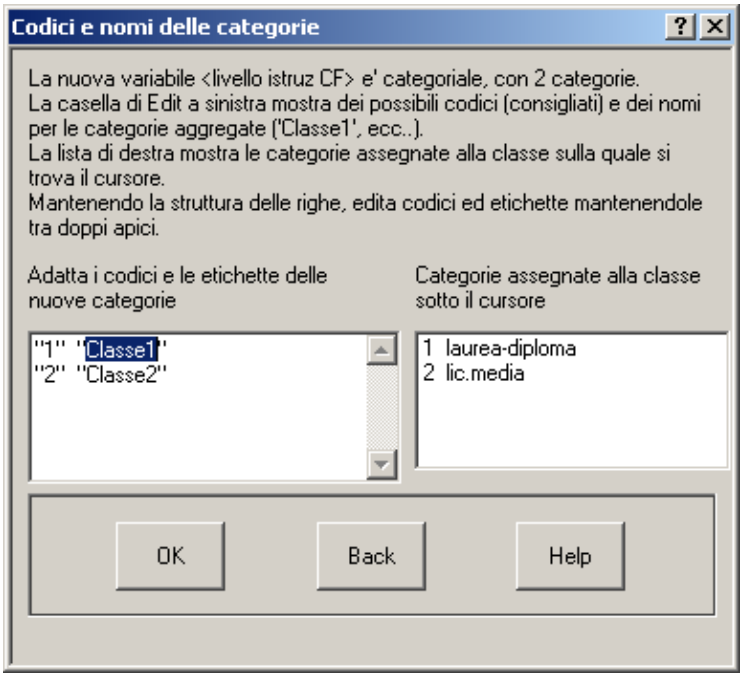


Figura 4.11 – Codici e nomi delle nuove categorie

4.3 - Estrazione di un sottoinsieme di unità statistiche

L'estrazione di un sottoinsieme opportuno di unità statistiche (segmento di popolazione) si rende necessario:

- quando si voglia creare un nuovo DataSet selezionando solo le unità statistiche che hanno determinate caratteristiche (utilità **SELECT**);
- quando **si voglia restringere qualche operazione** ad un segmento opportunamente definito della popolazione considerata. La cosa è possibile sia nel calcolo di distribuzioni, che per gli incroci o le correlazioni tra gruppi di variabili.

  Ad esempio, si può voler limitare i calcoli alle sole famiglie con cinque o più componenti, o ai soli Comuni con meno di 20000 abitanti, ecc. Anche se il loro uso è meno frequente, è tuttavia possibile imporre condizioni molto più complesse, che combinano i valori di più variabili.

4.3.1 - Il caso SELECT

La figura 4.12 mostra un esempio del dialogo. Sono caricati tre DataSet, mostrati nelle due liste in alto a sinistra, che descrivono le 39615 famiglie delle Filippine di cui si è già detto. Il DataSet <**dwelling**> descrive le abitazioni ed i beni mobili; <**hh**> (households) descrive le famiglie; <**expenses**> descrive le voci di spesa di ciascuna famiglia in grande dettaglio. Le variabili sono molte e si è preferito dividerle in quattro file per una maggior maneggevolezza (il quarto, che descrive la composizione del reddito, non viene qui considerato).

Si vogliono qui estrarre dal Data Set <**expenses**> i record relativi a tutte le famiglie con più di cinque componenti **e** con il Capo Famiglia di età compresa tra 30 e 50 anni (inclusi). Nell'esempio, le due variabili ***numero di componenti*** ed ***età CF*** , usate per delimitare il segmento e dette perciò ***variabili di filtro***, sono prese dai due DS <***dwelling***> e <***hh***>.

Tutti i DataSet caricati vengono mostrati sia nella lista intestata "***Data Set di Input***" che in quella intestata "***Data Set(s) di filtro***". Dalla prima si sceglie **un solo** DataSet, cioè quello dal quale verranno estratti i record (*unità statistiche*, o *casi*) d'interesse: in questo esempio è <EXPENSES>. Nella seconda si selezionano i Data set che contengono le ***variabili di filtro***, da usare per costruire la condizione che individua il segmento da estrarre.

> ***Nota*** *Quasi sempre le variabili di filtro (utilizzate per formulare la condizione) stanno anch'esse nel DataSet di input, dal quale verranno scelti i record da trascrivere: DataSet di input e di filtro in tal caso coincidono. Ma non sempre è così, come mostra questo esempio più generale..*

Il programma trascrive i DS di filtro, selezionati dall'utente, nella lista in alto a destra. A ciascuno di essi viene assegnata **una lettera** che ne contraddistingue le variabili. Selezionandone uno, le sue variabili vengono listate nella finestra di sinistra, insieme al loro identificatore. Selezionando una variabile se ne visualizzano alcune informazioni sommarie.

Viene abilitata la casella di editing in basso, dove va inserita la condizione che individua il segmento voluto. Essa può essere digitata (o modificata) direttamente, oppure ci si può avvalere di una forma d'inserimento facilitata dai bottoni disponibili per i diversi operatori.

Un doppio click su una variabile, o su una categoria, ne comanda l'inserimento nella posizione del cursore. **Provare**. Comunque, una volta capita la sintassi, il modo più rapido è digitare la condizione direttamente.

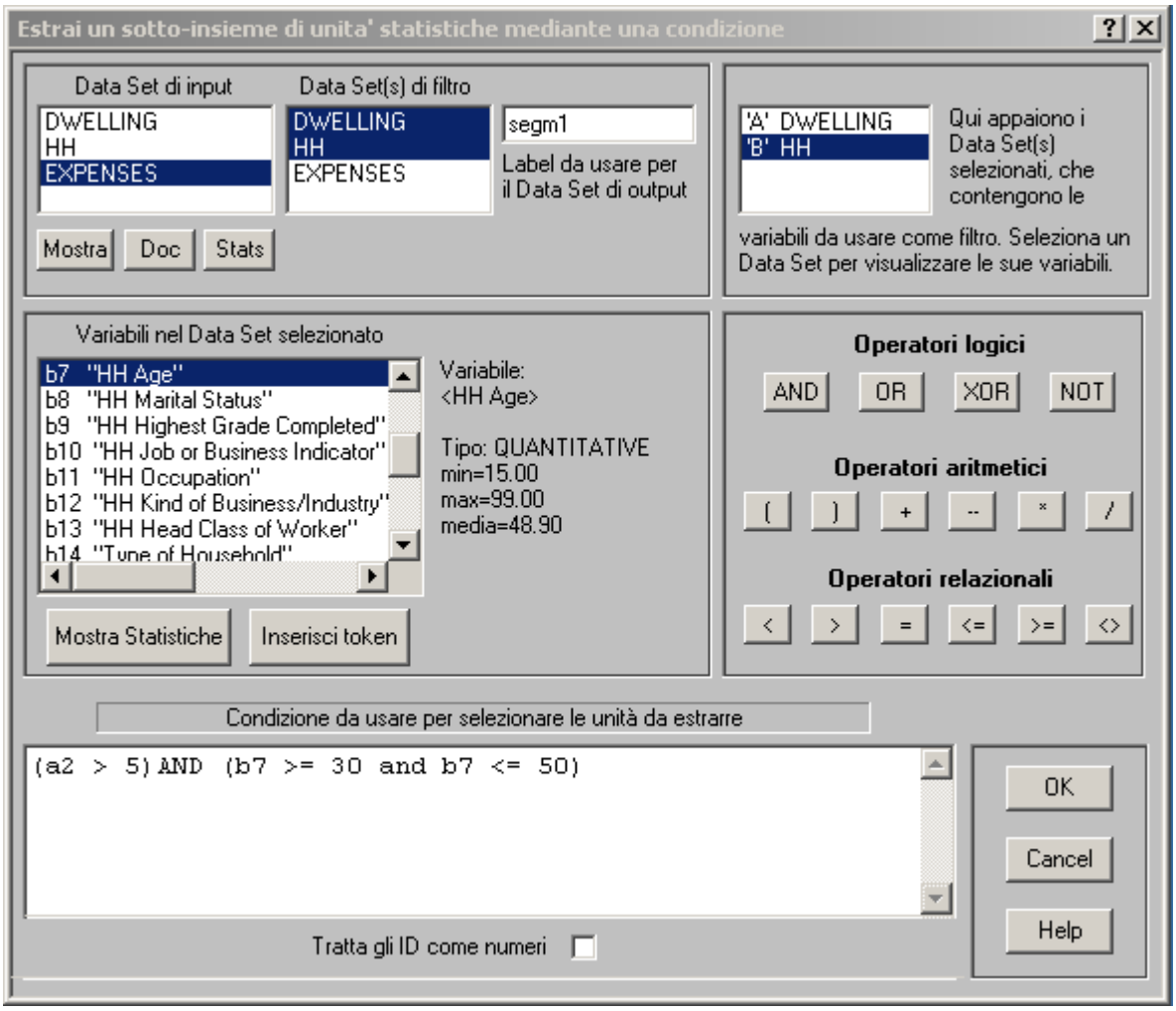


Figura 4.12 - Il dialogo per la estrazione di un sottoinsieme di unità statistiche

Nota *Quando si inizi ad inserire la condizione, la scelta dei DataSet* ***di input e di filtro viene disabilitata****. Una loro ulteriore modifica, infatti, invaliderebbe gli identificatori già usati per le variabili, con risultati imprevedibili. Per prevenire possibili errori le due liste vengono allora bloccate.*
Se a questo punto si vogliono cambiare i Data Set di input o di filtro bisogna chiudere il dialogo con il bottone ***Cancel****, riaprirlo e rifare la selezione.*

Il bottone **OK** comanda il controllo della condizione inserita e la sua esecuzione. Gli eventuali errori vengono segnalati e l'utente può correggerli. Quando la correttezza della condizione è confermata il nuovo DataSet viene creato **in memoria** ed aggiunto all'elenco dei DS aperti. Esso può essere utilizzato in tutte le operazioni successive. Il numero dei casi estratti è comunicato in un messaggio che appare nella finestra principale.

Si ricordi **che il DataSet creato va salvato su file se si vuole conservarlo alla chiusura del programma** ed utilizzarlo in una sessione di lavoro successiva.

Il paragrafo successivo illustra la sintassi della condizione.

4.3.2 - Delimitazione di un sottoinsieme di unità statistiche: sintassi della condizione

La sintassi è la medesima nei vari casi in cui possono venire imposte condizioni, e cioè:

- per estrarre i record relativi a un sottoinsieme di unità statistiche (opzione SELECT);
- per limitare ad un segmento di popolazione le distribuzioni delle variabili (opzione SETVAR), gli INCROCI tra variabili categoriali (opzione CROSSTAB), o le correlazioni tra variabili quantitative (opzione CORREL);
- per la costruzione di variabili definite da un insieme di **condizioni logiche** (opzione MERGE).

> ***Nota*** *L'esame di una condizione produce sempre un **valore di verità** (VERO o FALSO, record incluso nel segmento o escluso...). Diverso è invece il caso della costruzione di una nuova variabile mediante un'espressione matematica (MERGE): qui il **risultato dell'espressione** (che non è una condizione!) è il **valore numerico** da assegnare alla nuova variabile.*

Le condizioni usano le funzioni e gli operatori aritmetici utilizzati per la costruzione di variabili quantitative mediante operazioni matematiche (caso MERGE), aggiungendovi gli operatori relazionali '<', '>', '=', '>=', '<=', '<>', e gli operatori logici '**AND**', '**OR**', '**XOR**' e '**NOT**'.

Gli operatori

Nel caso **SELECT** il Data Set di input e quelli *di filtro* usati per la selezione (che possono coincidere) debbono descrivere le medesime unità statistiche, nello stesso ordine. Ad esempio, se le unità sulle quali si opera la selezione sono i 300 Comuni di una Regione, ciascun Data Set deve contenere esattamente 300 records ed in tutti i Data Set l'i-esimo record deve descrivere il medesimo Comune i.

Bisogna specificare la **CONDIZIONE LOGICA** (di solito ottenuta combinando un insieme di condizioni elementari) che le unità da selezionare debbono soddisfare.

La condizione **identifica** le combinazioni accettabili dei valori delle variabili di selezione. Essa fa uso dei seguenti **tokens (oggetti) elementari**:

- **numeri** (1, -3.2, 300...) che vengono caricati come valori reali;
- **identificatori di campo**, che consistono di una lettera ed un numero intero: la lettera (a partire da '**A**') individua il DataSet che contiene la variabile considerata, e l'intero è il numero d'ordine della variabile nel DataSet;
- **gli operatori aritmetici** ' **+ - * /** ' (operazioni aritmetiche) e ' ^ ' (elevazione a potenza). Essi debbono stare tra i numeri o gli identificatori di campo sui quali operano.
- gli operatori **somma**, **media** e **stdev ;**
- **gli operatori relazionali** ' < > = >= <= <> ' (l'ultimo simbolo significa 'diverso da'); essi debbono stare tra i due termini sui quali operano: numeri, identificatori di campo, valori di variabili tipo ID (stringhe) o gruppi più complessi che includono almeno un operatore aritmetico e vanno pre-valutati;
- **gli operatori logici** ' **AND OR XOR NOT** '.

Si possono usare parentesi rotonde per cambiare la priorità delle operazioni. Ad un medesimo livello di parentesi, un'espressione è valutata da sinistra a destra secondo le **regole di priorità seguenti**:

- vengono valutati prima gli operatori aritmetici, poi quelli di relazione ed infine quelli logici;
- '*****' e '**/**' sono valutati prima di '**+**' e '**-**' ;
- **NOT** è valutato con precedenza sugli altri operatori logici.

Gli **operatori aritmetici agiscono su quantità reali** (numeri, contenuti di campi o risultati di operazioni aritmetiche precedenti) e producono un valore numerico decimale (reale).

Gli **operatori di relazione confrontano valori numerici** (numeri forniti dall'utente, contenuto di campi, risultati di operazioni aritmetiche precedenti) e producono un valore di verita` (VERO se la condizione è soddisfatta, FALSO altrimenti).

Si possono utilizzare in una condizione anche **variabili di tipo ID**, **i cui valori debbono essere inclusi tra apici o doppi apici** e vengono interpretati come stringhe.

Gli **operatori logici** operano su **due valori di verità** già calcolati e generano un altro valore di verità. Fa eccezione '**NOT**', che agisce su un solo termine, il quale convenzionalmente segue l'operatore. Si ricordi che:

AND	restituisce VERO solo quando entrambi i termini sono VERI;
OR	restituisce VERO quando **almeno uno** dei termini è VERO;
XOR (**OR esclusivo**)	restituisce VERO quando **uno e solo uno** dei suoi termini è VERO;
NOT	restituisce VERO quando agisce su FALSO, e viceversa.

Vengono controllate la sintassi dell'espressione digitata dall'utente e l'accettabilità degli indicatori di campo. Viene poi caricato un caso dal DataSet di input e da ciascuno dei DataSet di filtro: i campi inclusi nella condizione vengono caricati **come numeri reali o stringhe alfanumeriche**, ed il loro valore viene assegnato alle espressioni dove figurano come termini. Viene calcolato **il valore di verità globale della condizione**: solo se essa è VERA il caso letto dal DataSet di input viene trascritto (caso SELECT) o tenuto in considerazione per il calcolo delle distribuzioni (SETVAR), degli incroci (CROSSTAB) o delle correlazioni (CORREL).

L'operazione viene ripetuta sulle unità statistiche successive, fino ad esaurire il Data Set di input.

Esempio 1 – Condizioni su variabili quantitative

La condizione **' a3 <> b1 and a4/a5 = .5 '** seleziona le unità per le quali il terzo campo del primo DataSet è diverso dal primo campo del secondo Data Set, e **simultaneamente** il rapporto tra i campi 4 e 5 del primo Data Set vale esattamente 0.5.

Le formulazioni

'NOT(a3 = b1) AND a5 = 2*a4'

oppure

'(a3 < b1 OR a3 > b1) AND a5/a4 = 2'

sono esattamente equivalenti a quella considerata.

Esempio 2

Si supponga che l'identificatore di campo **a12** (dodicesimo campo del primo Data Set) rappresenti la variabile *superficie dell'alloggio*.

La condizione **'a12 >= media(a12)'** estrae tutti gli alloggi con superficie **non minore** della superficie media.

Esempio 3 – Condizioni su variabili ID

Si supponga che in un DataSet che descrive i Comuni italiani **il campo 3 sia una variabile di tipo ID** (identificatore), che fornisce il codice ISTAT del Comune. Ad esempio, sia "05023001" il codice del Comune di AFFI: Comune 001 della provincia 023 (Verona), nella Regione 05 (Veneto). La condizione:

a3 > "05023001" and a3 < "05024001"

selezionerà tutti e soli i Comuni della provincia di Verona (023).

Attenzione!

*I valori delle variabili dichiarate come ID (IDentificatori) vengono caricate come **stringhe alfanumeriche**. Il confronto tra esse, se richiesto in una condizione come quella dell'esempio precedente, viene eseguito in **ordine lessicografico**, del tipo dell'ordine alfabetico in un dizionario. Qui però l'ordine dei caratteri è determinato dal loro codice ASCII, un valore tra 0 e 255 che caratterizza ciascun carattere. Nell'esempio precedente la stringa "05023001" **precede** la stringa "05024001" perché il carattere '3', in quinta posizione nella prima stringa e che ha codice ASCII 51, precede il carattere '4', in quinta posizione nella seconda stringa, il cui codice ASCII è 52. In ordine lessicografico le cifre 0-9 mantengono il loro ordine, e precedono tutti i caratteri alfabetici, sia maiuscoli che minuscoli..*

Esempio 4 – Condizioni su variabili categoriali

I codici delle variabili categoriali, desunti dal file di documentazione, ***sono memorizzati come stringhe, nell'ordine in cui vengono dichiarati***. Il loro confronto in una condizione avviene secondo quest'ordine: **i codici dichiarati prima precedono quelli dichiarati dopo**.

Ad esempio, si supponga che la variabile A1 sia la ***località***, con tre categorie dichiarate come

```
VAL   "1"   centro    # 6000 alloggi
VAL   "2"   nucleo    # 1000 alloggi
VAL   "3"   sparso    # 2000 alloggi
```

I doppi apici non sono necessari, ma vengono inseriti per ricordare che si tratta di stringhe. Dopo il '#' che introduce un commento, è stata scritta la frequenza di ciascuna categoria, per un controllo del risultato..

La condizione " **a1 < 3** "

estrarrà 7000 record, cioè tutti quelli nei quali la ***località*** vale '1' o '2', **ma non perché <u>i numeri</u> 1 e 2 sono minori del numero 3 (non si tratta di numeri!), bensì perché le categorie codificate come 1 e 2 precedonoil codice 3 nella dichiarazione**.

Se i codici della variabile località fossero stati dichiarati nell'ordine seguente

```
VAL   "3"   sparso    # 2000 alloggi
VAL   "1"   centro    # 6000 alloggi
VAL   "2"   nucleo    # 1000 alloggi
```

dichiarazione evidentemente equivalente per significato alla precedente, la stessa condizione " **a1 < 3** " **non avrebbe estratto alcun record, perchè non esiste alcuna categoria che preceda la categoria con codice 3.**

Cap 5. – Associazioni tra coppie di variabili

In questo capitolo descriveremo brevemente gli strumenti usati per valutare la forza dell'associazione esistente tra coppie di variabili quantitative o, rispettivamente, categoriali.

Nel primo caso (variabili quantitative) verranno calcolate delle **correlazioni** , definite nel capitolo 2; nel secondo caso (variabili categoriali) degli **incroci**.

Si tratta delle ultime due voci del Menu di Analisi.

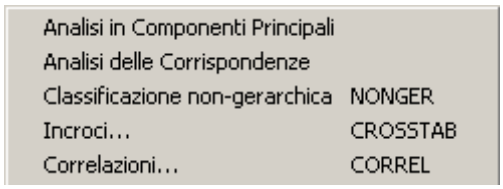


Figura 5.1 - Il Menu di Analisi

5.1 – Variabili quantitative: calcolo delle correlazioni

La figura 5.2 mostra il dialogo di controllo. In alto a sinistra va selezionato il DataSet che contiene le variabili sulle quali si vuole operare, e la lista delle sue variabili appare a destra.

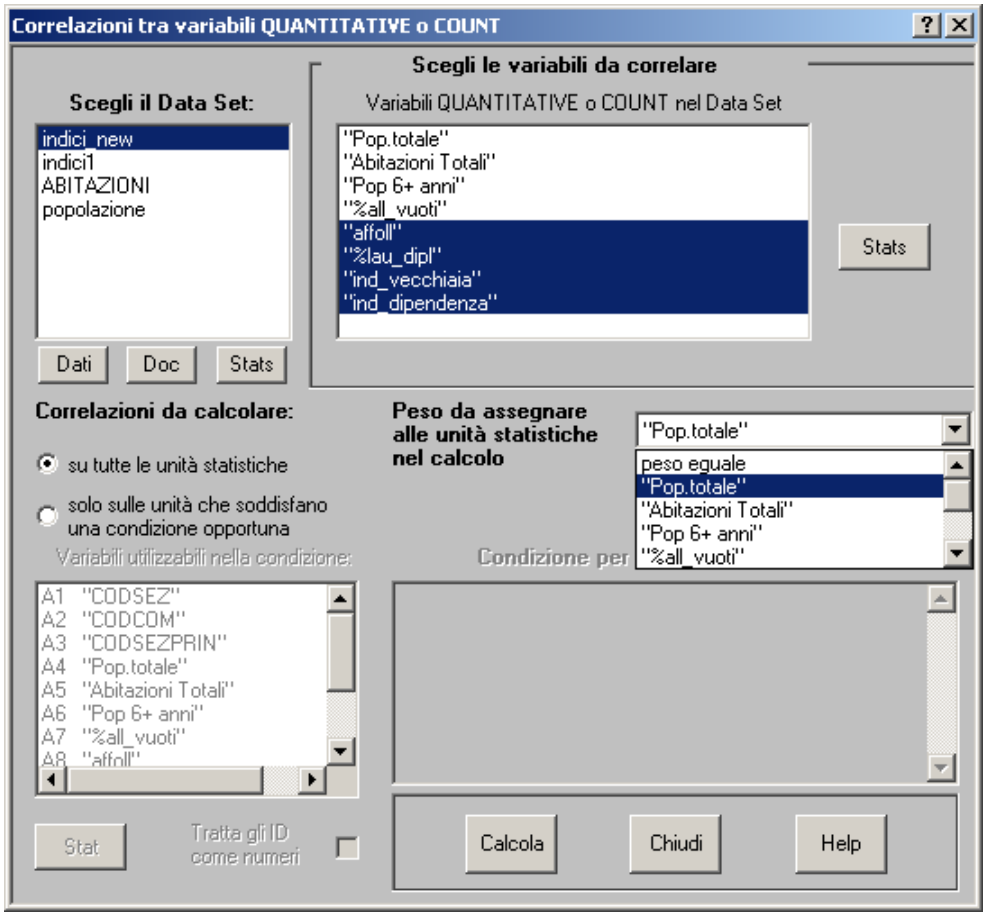


Figura 5.2 – Il dialogo per il calcolo delle correlazioni

Il DataSet selezionato – indici_new – contiene alcune variabili osservate sulle 3323 Sezioni Censuarie della Provincia di Venezia (le unità statistiche). Non vengono mostrate tutte le variabili del DataSet, ma solo di quelle tra le quali ha senso calcolare delle correlazioni, cioè quelle di tipo QUANTITATIVE e COUNT. Nella figura appaiono selezionate quattro variabili.

Le unità possono venire ponderate con i valori di una variabile selezionata dal box apposito (anche qui, le variabili candidate a fungere da peso sono solo quelle quantitative). Il peso scelto viene applicato alle unità **per il calcolo di tutte le correlazioni** richieste. Se per qualche coppia di variabili l'analista ritiene di ponderare le unità in modo diverso, deve calcolare la correlazione in una diversa sessione.

Al solito, il bottone Dati visualizza il contenuto del DataSet prescelto, ed i due bottoni Stat visualizzano le statistiche elementari per tutte le variabili oppure, rispettivamente, per le sole variabili selezionate nella lista a destra.

Il calcolo delle correlazioni può includere tutte le unità statistiche, oppure venire ristretto ad un sottoinsieme di esse, definito da una condizione opportuna (la sintassi è la solita). Ciò consente di calcolare le relazioni tra le variabili di interesse su segmenti diversi della popolazione (redditi alti o bassi; famiglie urbane o rurali, ecc.) confrontando i risultati.

L'output che si ottiene (si può salvarlo come file di testo, e tenerlo aperto per confrontarlo con altri output) è diverso a seconda che nessuna delle variabili da correlare presenti dati mancanti, oppure qualcuna di esse ne abbia.

5.1.1 - L'output quando le variabili non hanno dati mancanti

Se nessuna delle variabili per le quali è richiesto il calcolo di correlazioni ha dei dati mancanti, l'output è semplicemente quello mostrato nella tabella 5.1

Data Set < indici_new > - CORRELAZIONI TRA VARIABILI
LE CORRELAZIONI SONO CALCOLATE SU TUTTE LE UNITA' STATISTICHE.
Unità statistiche ponderate con la variabile < Pop.totale >.
Nessuna variabile considerata per le correlazioni ha valori mancanti

Variabile	media	dev.standard
" affoll "	0.644	0.091
" % lau_dipl "	21.418	11.587
"ind_vecchiaia "	136.022	93.368
"ind_dipendenza"	38.615	8.027

***** MATRICE DELLE CORRELAZIONI *****

	affoll	%lau_dipl	ind_vecchiaia	ind_dipendenza
affoll	1.0000	**-0.4638**	-0.1849	-0.1027
%lau_dipl	**-0.4638**	1.0000	0.3314	-0.0575
ind_vecchiaia	-0.1849	0.3314	1.0000	**0.3622**
ind_dipendenza	-0.1027	-0.0575	0.3622	1.0000

Tabella 5.1 – L'output quando non vi siano valori mancanti

L'uscita elenca media e deviazione standard per ciascuna delle variabili considerate, poi viene scritta la matrice delle correlazioni che, com'è noto, è simmetrica. I termini sulla diagonale principale, che valgono sempre 1, rappresentano le correlazioni di ciascuna variabile con se stessa.

I valori delle correlazioni non appaiono qui molto elevati. Sono stati sottolineati i valori maggiori, che mostrano una debole correlazione positiva tra gli indici di vecchiaia e di dipendenza come, forse in misura superiore, ci si poteva attendere. Un'altra correlazione positiva (0.3314) lega l'indice di vecchiaia e la percentuale di laureati/diplomati, ed una correlazione negativa (-0.4638) appare tra % di laureati/diplomati ed indice di affollamento.

Ricorda!

Nell'esempio della tabella 5.1 le unità statistiche sulle quali sono state calcolate le correlazioni **sono le Sezioni Censuarie, non gli individui o le famiglie**. Ogni considerazione interpretativa dovrà dunque riferirsi alle Sezioni, non alle famiglie sottostanti, per le quali non si è in possesso di una descrizione diretta.

Ad esempio, la correlazione positiva tra indice di vecchiaia e laureati/diplomati non significa necessariamente che i vecchi abbiano un titolo di studio superiore, anzi in generale è proprio il contrario. Si può solo dire che ci si può attenede che le Sezioni Censuarie che hanno un indice di vecchiaia superiore alla media abbiano anche una percentuale superiore alla media di laureati/diplomati: si pensi ai Centri Storici dei capoluoghi, con molti anziani, pochi bambini e mediamente con presenza di strati sociali più abbienti, con molti titoli di studio superiori legati anche alla concentrazione di servizi.

5.1.2 - L'output in presenza di dati mancanti

Se qualcuna delle variabili coinvolte nelle correlazioni presenta valori mancanti le cose si complicano. Non è infatti possibile calcolare il contributo alla covarianza (ed alla correlazione) delle unità statistiche per le quali manchi il valore anche solo di una delle due variabili.

Poiché ogni correlazione considera ***una coppia*** di variabili, media e deviazione standard di ciascuna varieranno in generale a seconda delle unità statistiche prese in considerazione nel calcolo, che dipendono anche dai valori mancanti dell'altra variabile della coppia. Insomma, per ciascuna correlazione ogni variabile presenterà una media ed una deviazione standard in generale diverse, che dovranno essere usate per il calcolo di quella particolare correlazione.

La tavola di output del programma è in questo caso del tipo mostrato in tabella 5.2.

Vengono mostrate le correlazioni tra quattro variabili, osservate su 195 Paesi. Il loro significato è il seguente:

- **EDGDP-2005%** è la spesa complessiva per l'istruzione nel 2005, espressa come percentuale del GDP (Gross Domestic Product – Prodotto Interno Lordo del Paese);
- **EDGOV-2005%** è la stessa spesa, espressa però come percentuale della spesa governativa complessiva, cioè del bilancio dello Stato;
- **EDGOV1-2005%** è la spesa per l'istruzione primaria, come % della spesa governativa totale;
- **EDGOV1-2005%** è la spesa per l'istruzione secondaria, come % della spesa governativa totale.

Per alcuni Paesi può mancare il valore di qualche variabile, e dunque essi non possono essere considerati nel calcolo di correlazioni che coinvolgono quella variabile.

La tabella specifica l'eventuale ponderazione applicata ed include, oltre alla matrice delle correlazioni, altre due matrici: una fornisce il numero di unità statistiche incluse nel calcolo di ciascuna correlazione, l'altra i valori medi (ponderati) di ciascuna variabile, tanti quante sono le correlazioni calcolate.

Si considerino i valori validi nella tabella 5.2: sulla diagonale principale appaiono quelli presenti per ciascuna variabile presa singolarmente, cioè quelli utilizzati nel calcolo della sua correlazione con se stessa, che vale sempre 1. Si vede che EDGDP-2005% ha 153 valori validi, EDGOV-2005% ne ha 134, ecc. La cella (1,2) della matrice fornisce le unità prese in considerazione nel calcolo della correlazione tra queste due variabili: sono 133, una di meno dei valori validi della seconda variabile. Significa ovviamente che per tutti i Paesi che hanno un valore valido di EDGOV-2005% è valido anche il valore di EDGDP-2005%, tranne uno.

La matrice che segue (medie delle variabili) mostra la media di ciascuna variabile calcolata sulle sole unità statistiche prese in considerazione per ciascuna correlazione.

Data Set < education > - CORRELAZIONI TRA VARIABILI

LE CORRELAZIONI SONO CALCOLATE SU TUTTE LE UNITA' STATISTICHE.

Unità statistiche ponderate con la <Popolazione Totale 2005 (milioni)>.

*** **MATRICE DELLE CORRELAZIONI** ***

	EDGDP-2005%	EDGOV-2005%	ED1GOV-2005%	ED2GOV-2005%
EDGDP-2005%	1.0000	0.4357	0.0821	-0.0717
EDGOV-2005%	0.4357	1.0000	0.4457	-0.4913
EDGOV1-2005%	0.0821	0.4457	1.0000	-0.7080
EDGOV2-2005%	-0.0717	-0.4913	-0.7080	1.0000

***** VALORI VALIDI PER COPPIA DI VARIABILI CORRELATE *****

	EDGDP-2005%	EDGOV-2005%	ED1GOV-2005%	ED2GOV-2005%
EDGDP-2005%	153	133	120	115
EDGOV-2005%	133	134	107	102
EDGOV1-2005%	120	107	120	115
EDGOV2-2005%	115	102	115	115

***** MEDIE DELLE VARIABILI (leggere per righe) *****

Ogni riga dà le medie di una variabile nelle correlazioni con le variabili di colonna

	EDGDP-2005%	EDGOV-2005%	ED1GOV-2005%	ED2GOV-2005%
EDGDP-2005%	3.5903	3.5892	3.4399	3.2599
EDGOV-2005%	13.5639	13.5636	13.3376	14.0663
EDGOV1-2005%	37.5921	37.6597	37.5921	39.2767
EDGOV2-2005%	37.6308	37.7412	37.6308	37.6308

Tabella 5.2 – Esempio di output quando vi siano dati mancanti

5.2 – Variabili categoriali: gli Incroci

Lo scopo è di valutare la forza dell'associazione tra due variabili categoriali.

Esempio Si considerino le famiglie residenti in un Municipio dell'Honduras, Ocotepeque (un po' di dati esotici...) alla data del Censimento. Calcoliamo l'incrocio tra:

- ***l'età del CF***, ricodificata nelle tre categorie '<= 35 anni'; '36-54 anni'; '>54 anni'.
- ***il numero dei componenti la famiglia,*** ricodificato nelle quattro categorie seguenti: 1-3 componenti, 4-5 componenti, 6-7 componenti, 8 o più componenti.

L'obiettivo qui potrebbe essere di osservare in che modo il ciclo di vita della famiglia (espresso da una prima fase nella quale aumenta il numero dei componenti, seguita da una fase di stabilità e poi da una fase di declino) sia legato all'età del CF.

La tavola, che ha tante righe quante sono le categorie della prima variabile ed un numero di colonne pari alle categorie della seconda, viene riempita scorrendo tutte le unità statistiche e per ciascuna di esse incrementando di 1 il contenuto della cella che corrisponde alla combinazione dei valori delle due variabili incrociate.

	1-3 comp	**4-5 comp**	**6-7 comp**	**8+ comp**	
< 36 anni	199	280	166	60	705
36-54 anni	150	245	219	223	837
> 54 anni	274	163	132	148	717
	623	688	517	431	2259

Tabella 5.3 - Incrocio tra le variabili età del CF e numerosità del nucleo familiare

L'**incrocio** (o ***tavola di conteggio***, o ***tavola di contingenza***) che si ottiene conta in quante famiglie si presenti ciascuna combinazione delle categorie delle due variabili incrociate.

Nell'esempio,

- la cella (1,1) mostra che vi sono 199 famiglie di 1-3 componenti, e CF con meno di 36 anni;
- la cella (3,1) mostra che vi sono 274 famiglie di 1-3 componenti, e CF anziano (più di 54 anni).

L'ultima riga e l'ultima colonna sono dette ***marginali*** della tabella e danno la ***distribuzione*** (cioè il numero di casi in ciascuna categoria) di ciascuna delle due variabili incrociate.

Le unità statistiche corrispondenti ai marginali possono essere distribuite in molti modi sulle celle. Il numero di celle il cui valore può essere fissato arbitrariamente, rispettando però i valori marginali, rappresenta i ***gradi di libertà*** della tabella. Una tabella con **n** righe e **p** colonne ha (**n**-1)(**p**-1) gradi di di libertà.

La ***tavola di conteggio*** viene trasformata in una ***tavola di frequenze percentuali*** dividendo tutti i suoi valori per il totale delle unità contate, che nel nostro caso è 2259.

	1-3 comp	4-5 comp	6-7 comp	8+ comp	$f_{i.}$
< 36 anni	0.088	0.124	0.073	0.027	0.312
36-54 anni	0.066	0.108	0.097	0.099	0.371
> 54 anni	0.121	0.072	0.058	0.066	0.317
$f_{.j}$	0.276	0.305	0.229	0.191	1.0

Tabella 5.4 – Le frequenze percentuali ***osservate***

Ogni cella riporta la ***percentuale*** delle famiglie nelle quali si osserva la combinazione di caratteri che distingue la cella.

Esempio: Le famiglie con 1-3 componente e CF < 36 anni costituiscono l'8.8% del totale.

Ogni valore marginale rappresenta la frequenza di una categoria di una variabile, ***a prescindere dal valore assunto dall'altra variabile***. Così, le famiglie con 1-3 componenti, a prescindere dall'età del CF, sono il 27.6% di tutte le famiglie residenti; le famiglie con CF minore di 36 anni sono il 31.2%.

Ci proponiamo di verificare se esistono ***associazioni privilegiate*** tra le categorie delle due variabili.

- Il fatto che una famiglia abbia pochi componenti (da 1 a 3) ***rende più probabile*** il fatto che il CF risulti anziano?
- Il fatto che una famiglia abbia 8 o più componenti ***rende meno probabile*** che il CF sia anziano?

dove il raffronto è con le frequenze che ci ***aspetteremmo se non sapessimo nulla*** dell'altra variabile.

In altri termini: la conoscenza del valore assunto da una delle due variabili ***apporta qualche informazione*** sull'altra, nel senso che ne altera ***la distribuzione attesa***?

Per rispondere al quesito dobbiamo calcolare i valori delle frequenze di cella che ci aspetteremmo ***nell'ipotesi che le due variabili incrociate siano indipendenti***.

Indipendenti significa che la frequenza delle categorie di una variabile ***rimane la stessa*** qualunque sia il valore assunto dall'altra. Ad esempio, se le variabili fossero indipendenti la percentuale di CF anziani dovrebbe essere il 31.7%, e quella di CF giovani il 31.2%, ***qualunque sia la dimensione della famiglia***. Si tratta delle frequenze che si osservano sull'intera popolazione di famiglie. Se invece ciò non succede, le variabili hanno un certo ***grado di associazione***, tanto maggiore quanto più la tavola ***osservata*** si scosta da quella ***attesa nell'ipotesi di indipendenza***.

Dobbiamo dunque costruire la tavola che ci aspetteremmo se le variabili fossero indipendenti, confrontarla con quella osservata e definire una misura dello scostamento globale tra le due tavole.

5.2.1 - Costruzione della tavola attesa nell'ipotesi di indipendenza

Prendiamo come esempio la cella (1,4).

La frequenza delle famiglie con 8 o più componenti è $f_{.4} = 0.191$ (cioè il 19.1% di tutte le famiglie). Tra queste quante avranno il CF con meno di 36 anni? ***Se le due variabili sono indipendenti*** la probabilità di avere CF giovane ***deve essere la medesima*** che si riscontra per la popolazione intera, cioè $f_{1.} = 0.312$ (il 31.2%). Dunque, il 31.2% del 19.1% di tutte le famiglie. Detta f^{e}_{14} la frequenza attesa (l'apice 'e' sta per '***expected***') della cella (1,4),si ha

$f^{e}_{14} = f_{1.} * f_{.4} = 0.312*0.191 = 0.060$

cioè il 6%, valore abbastanza diverso dalla frequenza effettivamente osservata, che dalla tavola 5.3 risulta essere il 2.7%.

Date le distribuzioni marginali delle variabili, le famiglie con 8 o più componenti e CF giovane dovrebbero essere il **6.0% nell'ipotesi di indipendenza**, mentre risultano essere solo il 2.7%. Dunque le due modalità (8+ comp., < 36 anni) risultano meno associate di quanto potremmo aspettarci se le due variabili fossero indipendenti. Allora d'accordo, presentano un'associazione. Ma quanto forte è?

La tabella 5.5 mostra le frequenze attese nell'ipotesi che le variabili siano indipendenti: la ***frequenza attesa*** nella generica cella (i,j) è $f^{e}_{ij} = f_{i.} * f_{.j}$, che va confrontata con la frequenza osservata f^{o}_{ij}

	1-3 comp	**4-5 comp**	**6-7 comp**	**8+ comp**	$\mathbf{f_{i.}}$
< 36 anni	0.086	0.095	0.071	0.060	0.312
36-54 anni	0.102	0.113	0.085	0.071	0.371
> 54 anni	0.087	0.097	0.073	0.061	0.317
$\mathbf{f_{.j}}$	0.276	0.305	0.229	0.191	1.0

Tabella 5.5 - Tavola delle percentuali di cella **attese** $f^{e}_{ij} = f_{i.} * f_{.j}$

Il confronto tra celle omologhe delle due tavole mostra le differenze.

La tabella 5.6 mostra le frequenze percentuali di riga attese, calcolate dividendo i valori in ciascuna riga della 5.5 per il totale (***o marginale***) della riga.

	1-3 comp	**4-5 comp**	**6-7 comp**	**8+ comp**	$\mathbf{f_{i.}}$
< 36 anni	0.276	0.305	0.229	0.191	0.312
36-54 anni	0.276	0.305	0.229	0.191	0.371
> 54 anni	0.276	0.305	0.229	0.191	0.317
$\mathbf{f_{.j}}$	0.276	0.305	0.229	0.191	1.0

Tabella 5.6 - Distribuzioni percentuali di riga per la ***tavola attesa***

La distribuzione per numerosità ***attesa*** delle famiglie risulta la stessa per le diverse età del CF.

La tabella 5.7 qui sotto riporta il risultato dello stesso calcolo eseguito sulla tavola osservata.

	1-3 comp	4-5 comp	6-7 comp	8+ comp	$f_{i.}$
< 36 anni	0.282	0.397	0.234	0.087	0.312
36-54 anni	0.178	0.291	0.262	0.267	0.371
> 54 anni	0.382	0.227	0.183	0.208	0.317
$f_{.j}$	0.276	0.305	0.229	0.191	1.0

Tabella 5.7 - Distribuzioni percentuali di riga per la ***tavola osservata***

Nell'ipotesi di indipendenza tra le variabili (tavola attesa) le famiglie con CF nelle diverse fasce d'età, riportate sulle tre righe, sono distribuite percentualmente allo stesso modo sulle quattro classi di numerosità. La distribuzione percentuale di ogni riga coincide con la distribuzione percentuale globale della variabile ***età del CF***.

La tavola osservata è significativamente diversa da quella attesa, e questo indica la presenza di *associazioni privilegiate* tra alcune categorie delle due variabili.

5.2.2 - Definizione di un indicatore che misuri lo scostamento tra le tavole

Un tale indicatore:

- deve tener conto di tutte le celle;
- deve tener conto del fatto che una medesima differenza assoluta è ***meno significativa*** quando la frequenza attesa è più elevata. Ad es., la differenza tra un 3% atteso ed un 6% osservato e più rilevante di quella tra un 50% atteso ed un 53% osservato, nonostante si tratti sempre in assoluto del 3%.
- Si assume come ***contributo della cella (i,j)*** all'indicatore di scostamento il valore $\frac{(f_{ij}^{o} - f_{ij}^{e})^2}{f_{ij}^{e}}$
- L'indicatore di scostamento globale tra le tavole (detto chi-quadro) è la somma dei contributi di tutte le celle:

  $\chi^2 = \mathrm{N} * \sum_{ij} \frac{(f_{ij}^{o} - f_{ij}^{e})^2}{f_{ij}^{e}}$ dove N è il totale degli effettivi nella tavola e la somma è estesa a tutti i valori di *i* e *j*.
- L'elevazione al quadrato esalta l'importanza degli scarti maggiori e rende positivi i contributi di tutte le celle, a prescindere dal segno della differenza $f_{ij}^{o} - f_{ij}^{e}$.
- Il chi-quadro non è mai negativo, ed è nullo ***solo se la tavola osservata coincide esattamente con quella attesa***.

Data una tavola d'incrocio, il valore del suo χ^2 (e dei gradi di libertà della tavola) permette, per confronto con una opportuna tavola di distribuzione teorica, di calcolare la probabilità che le due variabili incrociate siano indipendenti.

IN PRATICA: ***fatto un incrocio, si esamina la probabilità di indipendenza tra le variabili calcolata dal programma. Se essa risulta molto bassa, e dunque la probabl'associazione risulta abbastanza forte, si interpreta l'incrocio.***

La tabella 5.8 mostra l'aspetto di un incrocio calcolato da ADDATI. Ogni cella (i,j) contiene tre valori:

- il ***numero assoluto*** $\mathbf{x_{ij}}$ delle unità statistiche assegnate alla cella;
- la ***percentuale sul totale di riga*** $\mathbf{x_{ij}/x_{i.}}$, dove $\mathbf{x_{i.}}$ è il totale della riga i;
- la ***percentuale sul totale di colonna*** $\mathbf{x_{ij}}$ /$\mathbf{x_{.j}}$, dove $\mathbf{x_{.j}}$ è il totale della colonna j;

L'interpretazione confronta le distribuzioni percentuali delle righe (o delle colonne) con la corrispondente distribuzione marginale.

	1-3 comp	**4-5 comp**	**6-7 comp**	**8+ comp**	
< 36 anni	**199** 28.2 31.9	**280** 39.7 40.7	**166** 23.5 32.1	**60** 8.5 13.9	705 100.0 31.2
36-54 anni	**150** 17.9 24.1	**245** 29.3 35.6	**219** 26.2 42.4	**223** 26.6 51.7	837 100.0 37.1
> 54 anni	**274** 38.2 44.0	**163** 22.7 23.7	**132** 18.4 25.5	**148** 20.6 34.3	717 100.0 31.7
	623 27.6 100.0	688 30.5 100.0	517 22.9 100.0	431 19.1 100.0	2259 100.0 100.0

Tabella 5.8 – Formato di una tavola d'incrocio

Nota: ***Campione e universo- Il significato del χ^2***

Quando si ha a che fare con un campione, il χ^2 permette di valutare la probabilità che le differenze esistenti tra la tavola di incrocio osservata e quella attesa nell'ipotesi di indipendenza siano una mera conseguenza di fluttuazioni aleatorie intervenute nella scelta del campione stesso.

*Tuttavia, se si lavora sull'intera popolazione (l'**universo**), come spesso si fa, non si dovrebbe parlare di* probabilità che le variabili siano indipendenti *(nell'universo). Non essendoci un campione, non esiste alcuna incertezza connessa alla sua estrazione e non c'è alcuna inferenza da trarre: o la tavola osservata coincide con quella attesa (e le due variabili sono indipendenti), oppure no. In pratica le due tavole non coincideranno mai esattamente, tuttavia le differenze possono essere causate solo dal comportamento particolare o inatteso di poche unità e risultare tanto piccole da poterle considerare non significative.*

Quel che ci interessa capire è se sotto l'apparente associazione tra due variabili vi sia qualche ragione strutturale, le cui cause vanno indagate, oppure se si tratti solo di variazioni, da ritenere casuali, nel comportamento di poche unità.

*Possiamo sempre pensare alla nostra popolazione come ad un campione estratto da un ipotetico sovra-universo **nel quale le due variabili incrociate conservino le loro distribuzioni marginali**.*

*Possiamo allora ricorrere al χ^2 per decidere se la differenza globale tra la **tavola osservata** (cioè l'incrocio che abbiamo eseguito sul nostro universo, visto ora come un campione) e **quella attesa** (cioè la tavola che osserveremmo nel sovra-universo in caso di indipendenza) sia solo frutto della casualità dell'estrazione e non abbia alcun presumibile significato strutturale.*

5.2.3 – Calcolo di incroci

Si carichi in ADDAWIN il **DataSet Comune**, incluso tra quelli forniti come esempio, che riporta alcune informazioni relative agli 11770 Fogli di Famiglia di un Comune dell'Emilia-Romagna. Il dato è **al massimo livello di disaggregazione**, dunque molto ricco.

Si scelga dal Menu di Analisi la voce ***Incroci***, e si incrocino le due variabili < *titolo di studio CF* > con quattro categorie, ottenute aggregando opportunamente quelle riportate nel DataSet, e < vasca o doccia >, con tre categorie. La figura 5.3 mostra il dialogo nel quale sono state scelte le variabili da incrociare, e la tabella 5.9 riporta l'incrocio ottenuto, che va interpretato e commentato.

Anche nel caso degli incroci il dialogo mostra la lista completa dei DataSet caricati. Selezionandone uno, le sue variabili appaiono ripetute nelle due liste in alto a destra. Come per le correlazioni, anche qui vengono mostrate solo le variabili che si possono selezionare: in questo caso, solo quelle categoriali. Dalla lista di sinistra si seleziona la variabile da disporre sulle righe della tavola d'incrocio, mentre le categorie di quella selezionata dall'altra tavola andranno sulle colonne.

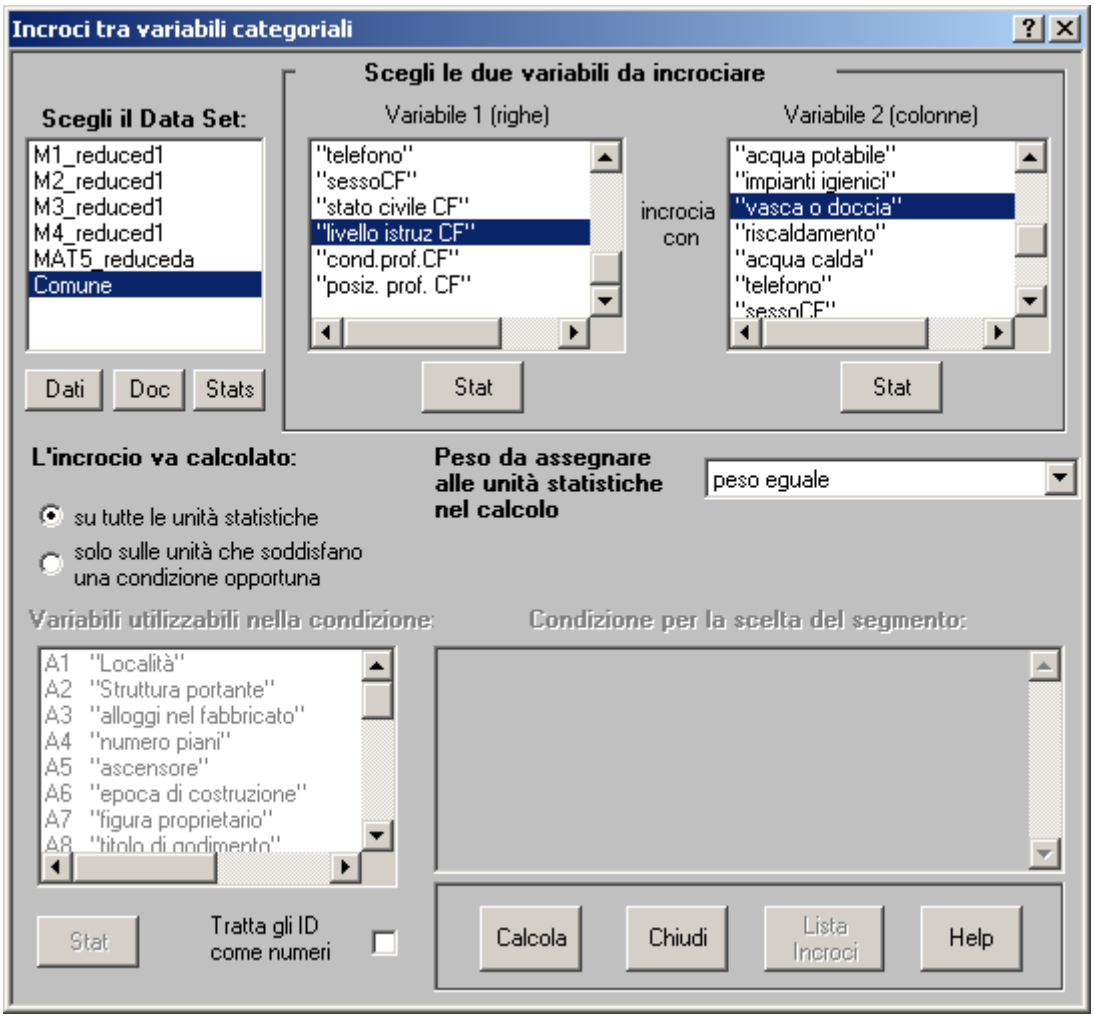


Figura 5.3 – Il dialogo per il calcolo degli incroci

L'incrocio può essere ristretto ad un sottoinsieme di unità statistiche usando i controlli nella parte inferiore del dialogo, proprio come per le correlazioni. Si possono calcolare tutti gli incroci

desiderati (usando il bottone ***Calcola*** dopo aver fissato le variabili e l'eventuale condizione).

Gli incroci fatti sono memorizzati: è possibile elencarli con il bottone ***Lista Incroci***, riesaminarli in qualunque momento, eliminare quelli che proseguendo l'analisi non servano più, salvarli su disco.

5.2.3 – Un esempio di interpretazione e commento

Nel commentare un incrocio si consiglia di seguire per quanto possibile il seguente schema interpretativo.

1. Valutare il livello di associazione tra le due variabili, espresso dal chi-quadro e dalla probabilità d'indipendenza, calcolati dal programma. Se il chi-quadro è molto basso non c'è associazione e l'interpretazione si può omettere. Se è elevato, vanno individuate le celle che corrispondono a categorie delle due variabili molto più o molto meno associate di quanto ci si aspetterebbe nell'ipotesi d'indipendenza.
2. Illustrare il significato dei valori di cella **in termini quantitativi**, basandosi sulle **percentuali** (di riga, di colonna) e **non sui valori assoluti** che possono ingannare. Il confronto va fatto con le percentuali attese nell'ipotesi di indipendenza, rappresentate dalla distribuzione percentuale della riga (o della colonna) marginale. Le considerazioni svolte devono basarsi esclusivamente **su quanto emerge dai dati**, cioè sul contenuto della tavola d'incrocio.
   Vanno evitate espressioni vaghe come "la maggior parte", "i più" o simili, che non permettono di capire a quali valori ci si riferisca esattamente. Vanno invece sempre citati i valori numerici (percentuali) appropriati.
3. Se sono emerse delle associazioni significative (sia positive che negative) tra alcune categorie delle due variabili, si può cercare di formulare delle ipotesi sulle ragioni (demografiche, economiche, sociali, di mercato...) che le hanno causate. Non sempre ciò è possibile o facile, ma ipotizzare delle ragioni può suggerire ulteriori indagini volte ad approfondirne la fondatezza. Una ricerca che parte da un incrocio spesso continua con altri, sia sull'intera popolazione statistica che su segmenti particolari, suggeriti dall'interpretazione dei precedenti.

   Attenti però a non formulare ipotesi del tutto cervellotiche, come a volte accade...meglio allora lasciar perdere.

L'interpretazione va condotta confrontando le distribuzioni percentuali delle righe (o delle colonne) con la corrispondente distribuzione marginale.

La probabilità di ottenere una tavola come quella osservata se le variabili fossero indipendenti è bassissima, molto prossima allo zero (di fatto, l'ordine di grandezza è di 10 elevato alla -107). L'associazione tra le due variabili è dunque molto forte.

La cosa si giustifica considerando le nostre 11770 unità statistiche (famiglie) come un campione estratto da un universo molto più numeroso, nel quale le due variabili siano indipendenti. La probabilità di estrarre **un particolare campione** che presenti una tavola d'incrocio tra le variabili come quella osservata sarebbe praticamente nulla. Minore è la probabilità d'indipendenza tra le variabili (nel senso anzidetto) maggiore riteniamo essere il loro livello di associazione.

Incrocio tra < **livello istruz CF** > e < **vasca o doccia** >.
L'INCROCIO E' CALCOLATO SU TUTTE LE UNITA' STATISTICHE.
Casi validi: 11770 - Casi esclusi per valori mancanti o invalidi: 0.

	* una *	due+ *	assen *	
	* *	*	te *	
	* 1351 *	1231 *	10 *	2592
laurea-d*	52.1 *	47.5 *	0.4 *	100.0
iploma *	17.2 *	32.4 *	7.2 *	22.0
	* 1752 *	856 *	11 *	2619
lic.medi*	66.9 *	32.7 *	0.4 *	100.0
a *	22.4 *	22.6 *	8.0 *	22.3
	* 3253 *	1319 *	66 *	4638
lic.elem*	70.1 *	28.4 *	1.4 *	100.0
. *	41.5 *	34.8 *	47.8 *	39.4
	* 1482 *	388 *	51 *	1921
nessun t*	77.1 *	20.2 *	2.7 *	100.0
itolo *	18.9 *	10.2 *	37.0 *	16.3
	7838	3794	138	11770
	66.6	32.2	1.2	100.0
	100.0	100.0	100.0	100.0

OGNI ENTRATA DELLA TAVOLA CONTIENE NELL'ORDINE GLI EFFETTIVI CONTATI, LA PERCENTUALE SUL TOTALE DI RIGA, LA PERCENTUALE SUL TOTALE DI COLONNA

CHI2 = 479.16 GRADI DI LIBERTA` = 6

LA PROBABILITA` CHE LE VARIABILI INCROCIATE SIANO INDIPENDENTI NON E' MAGGIORE DI 6.914E-107

Tabella 5.9 – L'output di un incrocio

Associazioni significative tra le variabili:

- nell'ipotesi d'indipendenza, la percentuale attesa di famiglie con due o più bagni dovrebbe essere del 32.2% (valore marginale **in viola,** corrispondente all'intera popolazione) qualunque sia il titolo di studio del CF. Si nota invece che per i CF con laurea o diploma la percentuale effettivamente osservata è del 47.5%, e scende al 32.7%, 28.4% e 20.2% per i CF con titolo via via inferiore. Ciò mostra che titoli di studio superiori sono in generale associati a migliori dotazioni fisico-funzionali dell'alloggio, e quantifica tale fatto.
- Simmetricamente, la percentuale attesa di famiglie con una sola vasca o doccia, che è del 66.6% (marginale di riga in rosso), sale invece (come valore osservato) al 77.1% per i senza titolo, scendendo poi al 70.1, 66.9 e 52.1 per i titoli via via superiori.
- Le percentuali relative alla colonna "vasca o doccia assente" mostrano anch'essi l'andamento prevedibile rispetto al titolo, ma le frequenze sono molto basse. L'andamento qualitativo è chiaro, ma i valori quantitativi sono presumibilmente poco affidabili per l'incidenza di possibili fluttuazioni aleatorie.
- Ragionando invece sulle percentuali di colonna mostrate **in verde**, possiamo rilevare che **tra le 3794 famiglie con doppio bagno** la percentuale di CF con laurea-diploma è del 32.4%, sensibilmente maggiore del 22% atteso nell'ipotesi di indipendenza tra le variabili; quella dei

CF con licenza media (22.6%) corrisponde al valore atteso (22.3%). La percentuale dei CF senza titolo tra le famiglie con doppio bagno è del 10.2%, sensibilmente inferiore al 16.3% atteso.

Si stia bene attenti a non confondere il significato della percentuale sul totale di riga con quello della percentuale sul totale di colonna. Ad esempio, la cella (1, 2) mostra che il 47.5% dei CF con laurea-diploma ha un'abitazione con doppio bagno, mentre il 32.4% delle abitazioni con doppio bagno è abitata da CF con laurea-diploma. Sono cose diverse.

Ipotesi interpretative: sono soggettive, e dipendono dalle conoscenze e convinzioni personali dell'analista; le considerazioni precedenti invece, basate esclusivamente sui dati, sono oggettive.
La forte associazione tra titolo di studio e dotazione fisico-funzionale dell'alloggio che emerge dall'incrocio è probabilmente conseguenza della maggiore disponibilità economica dei CF con titolo di studio alto, che trascina l'esigenza di migliori condizioni abitative. Può contare però anche il fatto, verificabile con incroci opportuni, che la mancanza di titolo ed il titolo elementare caratterizzano i CF più anziani (fatto verificabile con un incrocio EtaCF/titolodi studio), i quali stanno mediamente in abitazioni più vecchie (cosa verificabile con un incrocio EtàCF/epoca di costruzione) costruite secondo standard abitativi meno esigenti e solo parzialmente ristrutturate. I CF giovani invece hanno preso casa da poco, ed abitano mediamente in abitazioni di più recente costruzione.

Non sempre le ipotesi interpretative sono così ovvie, e facilmente verificabili. Si cerchi di riflettere e di proporre, se ci si riesce. Meglio comunque evitare di dire sciocchezza tanto per dire qualcosa...

5.2.4 - Uso di Incroci nella definizione di nuove variabili mediante condizioni logiche

È a volte necessario costruire un nuovo indicatore che sintetizzi l'apporto informativo di più variabili esistenti, **combinandone i valori** in modo opportuno. Ne risulterà una nuova variabile CATEGORIALE, ciascuna categoria della quale rappresenta un insieme di opportune combinazioni dei valori delle variabili di partenza, che possono essere sia categoriali che quantitative.

La definizione può basarsi sui valori di un numero qualsiasi di variabili, comunque è difficile che se ne usino più di tre… Se ne viene usata una sola, l'operazione equivale ad una ricodifica della variabile considerata. **Nel caso più frequente la definizione della nuova variabile parte dai valori di due variabili esistenti**. Se esse sono categoriali, è utile ragionare sul loro incrocio per fissare le regole di costruzione.

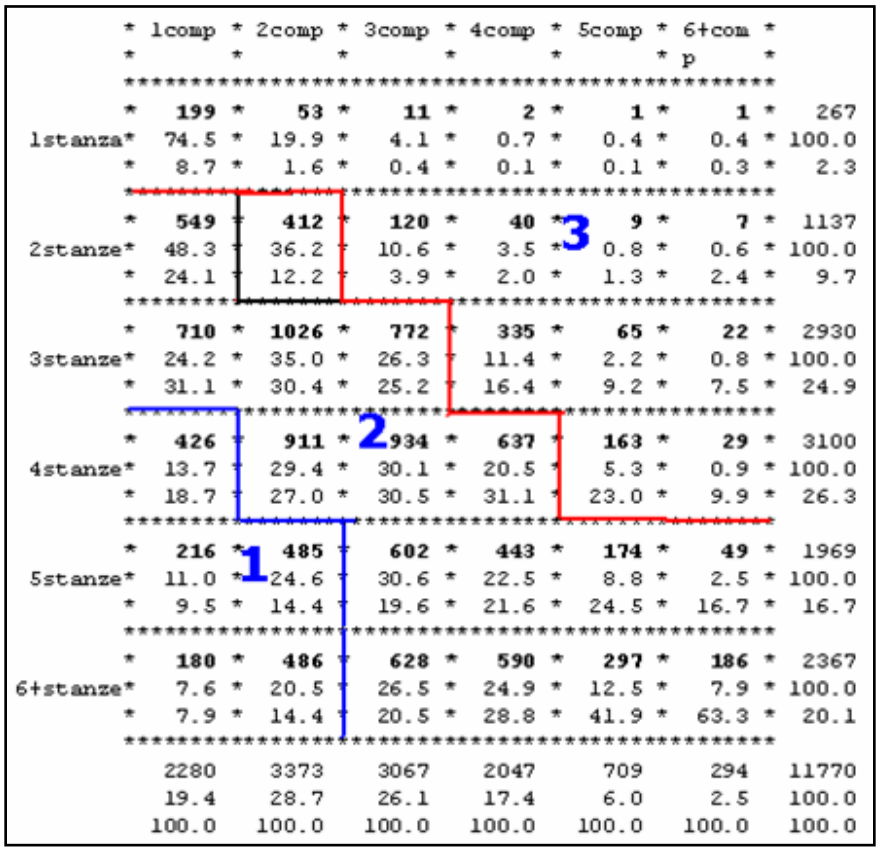
Vediamo tre esempi.

Esempio 1: l'affollamento

La tabella 5.10 mostra l'incrocio tra il numero delle stanze (righe) ed il numero dei componenti la famiglia (colonne). L'incrocio è esteso agli 11770 fogli di famiglia del Comune a noi ormai ben noto.

Le variabili, inizialmente quantitative, sono state preventivamente ricodificate come categoriali accorpando nella sesta categoria dell'una tutti gli alloggi con più di cinque stanze, e nella sesta categoria dell'altra tutte le famiglie di più di cinque persone.

La tavola fornisce informazioni molto dettagliate su come famiglie con uno stesso numero di componenti si distribuiscano su alloggi con diverso numero di stanze. Tale distribuzione, **frutto di osservazione ed a priori non prevedibile**, è l'esito di un complesso intreccio di fattori. Ci limitiamo a citarne alcuni.



	1comp	2comp	3comp	4comp	5comp	6+com p	
	199	**53**	**11**	**2**	**1**	**1**	267
1stanza	74.5	19.9	4.1	0.7	0.4	0.4	100.0
	8.7	1.6	0.4	0.1	0.1	0.3	2.3
	549	**412**	**120**	**40**	**9**	**7**	1137
2stanze	48.3	36.2	10.6	3.5	0.8	0.6	100.0
	24.1	12.2	3.9	2.0	1.3	2.4	9.7
	710	**1026**	**772**	**335**	**65**	**22**	2930
3stanze	24.2	35.0	26.3	11.4	2.2	0.8	100.0
	31.1	30.4	25.2	16.4	9.2	7.5	24.9
	426	**911**	**934**	**637**	**163**	**29**	3100
4stanze	13.7	29.4	30.1	20.5	5.3	0.9	100.0
	18.7	27.0	30.5	31.1	23.0	9.9	26.3
	216	**485**	**602**	**443**	**174**	**49**	1969
5stanze	11.0	24.6	30.6	22.5	8.8	2.5	100.0
	9.5	14.4	19.6	21.6	24.5	16.7	16.7
	180	**486**	**628**	**590**	**297**	**186**	2367
6+stanze	7.6	20.5	26.5	24.9	12.5	7.9	100.0
	7.9	14.4	20.5	28.8	41.9	63.3	20.1
	2280	3373	3067	2047	709	294	11770
	19.4	28.7	26.1	17.4	6.0	2.5	100.0
	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Tabella 5.10 – Costruzione di un indicatore di affollamento a tre categorie.

La distribuzione (in quello specifico Comune) del patrimonio abitativo per numero di stanze rappresenta un vincolo all'offerta: per ciascun tipo di alloggio, per quanto alta sia la richiesta, resta il limite della disponibilità. La distribuzione delle famiglie per numerosità, e la loro capacità economica, rappresentano altrettanti vincoli. Il trade-off tra il vantaggio di poter disporre di un maggior spazio abitativo e la rinuncia ad altri tipi di consumo che famiglie con determinate caratteristiche trovano ragionevole è un altro fattore. La distribuzione per età delle famiglie, con un numero di nuclei anziani più o meno elevato che spesso continuano ad abitare il medesimo alloggio anche dopo l'uscita dei figli, incrementando così le situazioni di sottoutilizzo, ne rappresenta un altro...si potrebbe continuare. Si tratta di condizioni **che caratterizzano il particolare contesto di analisi**.

Non possiamo in generale aspettarci di osservare una tavola caratterizzata dalle medesime distribuzioni in situazioni diverse. Le considerazioni sviluppate dall'analista, e l'eventuale delimitazione delle situazioni abitative da considerare come sovraffollamento, sono dunque strettamente dipendenti dal particolare contesto dell'analisi (nel nostro caso, il Comune considerato). Non sono di solito generalizzabili ad altri contesti, se non in termini qualitativi.

Si consideri ad esempio la seconda colonna della tavola. Essa mostra come le 3373 famiglie di due persone si distribuiscono tra i vari tipi di alloggio.

Se confrontiamo le percentuali osservate ed attese nell'ipotesi di indipendenza, vediamo che solo l'1.6% delle famiglie di due persone abitano in alloggi **di una stanza**, contro un valore atteso del 2.3%. Tali alloggi appaiono dunque *meno preferiti* rispetto ad una distribuzione *indifferente:*

possiamo ipotizzare che le famiglie che li abitano si sentano in condizioni di svantaggio rispetto alle altre famiglie dello stesso tipo, e non siano soddisfatte.

Al contrario, possiamo notare che le famiglie di due persone tendono a concentrarsi negli alloggi di due, tre e quattro stanze più di quanto ci si aspetterebbe nell'ipotesi d'indipendenza: in quelli di tre stanze, ad esempio, abita il 30.4% delle famiglie considerate, contro un valore atteso del 24.9%. Possiamo assumere tale dimensione dell'alloggio come *la norma*, o ***lo standard***, per famiglie di due persone.

Ovunque? È questa la *norma* assoluta? Difficile pensarlo. Se andassimo in India, o in Africa, o magari anche solo in un Comune piuttosto diverso, certamente i parametri di giudizio dovrebbe essere riaggiustati. Assumiamo quei casi come *affollamento standard* non in base ad un giudizio di merito (perché ci pare che tre stanze siano *giuste* per due persone), ma perché ce lo dicono i dati osservati: **in quel Comune** le famiglie di due persone tendono a concentrarsi in alloggi di 2-4 stanze. Le famiglie che stanno in una sola stanza costituiscono una fascia svantaggiata (in ***sovraffollamento***), mentre quelle che abitano in cinque o più stanze costituiscono una fascia minoritaria in ***sottoutilizzo*** (14.4% osservato in 6+ stanze, contro il 20.1% atteso).

Considerazioni analoghe si possono sviluppare per le altre colonne, che rappresentano la situazione abitative delle famiglie di diverse dimensioni. In generale, ogni colonna può essere suddivisa in due o tre fasce in base alla distribuzione di frequenza delle famiglie sui tipi di alloggio. Le celle più numerose rappresentano la situazione ***di normalità nel contesto***, corrispondenti a livelli d'affollamento che possiamo ***assumere come standard***. I gruppi di celle meno numerose rappresentano famiglie in situazione diversa dalla norm, che possiamo assumere come ***sovraffollamento* o *sottoutilizzo***. Le tre fasce che si vengono ad individuare, marcate nella figura con i numeri 1 (sottoutilizzo), 2 (normalità) e 3 (sovraffollamento) sono frutto di una ripartizione della tavola operata dall'analista in base alla sua valutazione dei dati osservati. Ci sono elementi di arbitrarietà: ad esempio, la cella (2, 2) è stata assegnata alla categoria 2 (standard), ma un altro analista potrebbe assegnarla alla 1 (sovraffollamento) operando il taglio nel modo indicato dalla linea in nero.

Nel nostro esempio, ogni famiglia viene assegnata ad una categoria di una variabile categoriale affollamento a tre categorie, definita ripartendo la tavola d'incrocio considerata come mostrato dalle linee rosse. La variabile categoriale va costruita assegnando ogni combinazione stanze-componenti ad una delle sue tre categorie.

Ciò viene fatto inserendo in un'apposita casella di edit una lista di regole.

Nella tabella a fianco si è indicata con a1 la variabile stanze e con a2 la variabile componenti (si suppone di lavorare su di un file che contiene solo tali due variabili, o che comunque le ha per prime). Le regole elencate assegnano alla nuova variabile categoriale *livello di affollamento* un valore ***val*** in corrispondenza *ad ogni possibile combinazione* (stanze, componenti).

```
if(a1=1)val=3
if(a1=2 and a2>2)val=3
if(a1=3 and a2>3)val=3
if(a1=4 and a2>4)val=3
if(a1=2 and a2<3)val=2
if(a1=3 and a2<4)val=2
if(a1=4 and a2=1)val=1
if(a1=4 and (a2=2 or a2=3 or a2=4))val=2
if(a1=5 and a2<3)val=1
if(a1=6 and a2<3)val=1
if(a1=5 and a2>2)val=2
if(a1=6 and a2>2)val=2
```

A scanso di inconvenienti, se le regole sono complicate come in questo caso conviene scriverle in un editor esterno (ad esempio il Blocco Note di Windows) e poi incollarle nell'apposita

casella di Edit di ADDATI.

Si tenga presente che per ogni unità statistica esaminata il programma esamina le regole in sequenza, e quando trova che la condizione introdotta dall'*if* è vera, assegna alla nuova variabile il corrispondente valore di *val*.

Se l'utente sbaglia la formulazione della sequenza, e fornisce due regole non incompatibili, cioè che possono essere simultaneamente vere per qualche unità statistica, viene sempre assegnato il valore-obiettivo specificato dalla prima. Alla seconda regola non ci si arriva mai.

Si tenga ancora presente che di solito le regole possono essere formulate in molti modi distinti ma equivalenti. Per l'esempio considerato, le regole elencate nella tabella a fianco vanno altrettanto bene.

```
if(a1=4 and a2=1) val=1
if(a1=5 and a2<3) val=1
if(a1=6 and a2<3) val=1
if(a1=2 and a2<3) val=2
if(a1=3 and a2<4) val=2
if(a1=4 and a2>1 and a2<5) val=2
if(a1=5 and a2>2) val=2
if(a1=6 and a2>2) val=2
else val=3
```

Esempio 2: Costruzione di un indicatore di qualità dei servizi interni

Si voglia sintetizzare in un nuovo indicatore l'informazione apportata dalle variabili:

Impianto di riscaldamento (5 categorie: 1=centralizzato, 2=autonomo, 3=stufe in tutto l'alloggio, 4=stufe in parte, 5=riscald. assente);

vasca o doccia (1=una, 2=due o più, 3=nessuna)

Anche qui conviene incrociare le due variabili d'interesse per vedere le relazioni tra le loro categorie e tener conto, nel definire il nuovo indicatore, della frequenza nelle singole celle.

Esaminato l'incrocio a fianco, si può decidere ad esempio di costruire il nuovo indicatore con quattro categorie, come mostrato nella figura.

	una	due+	assente	
centralizzato	689	170	3	862
	79.9	19.7	0.3	100.0
	8.8	4.5	2.2	7.3
autonomo	5724	3460	22	9206
	62.2	37.6	0.2	100.0
	73.0	91.2	15.9	78.2
stufe tot	875	99	54	1028
	85.1	9.6	5.3	100.0
	11.2	2.6	39.1	8.7
stufe parz	505	58	52	615
	82.1	9.4	8.5	100.0
	6.4	1.5	37.7	5.2
assente	45	7	7	59
	76.3	11.9	11.9	100.0
	0.6	0.2	5.1	0.5
	7838	3794	138	11770
	66.6	32.2	1.2	100.0
	100.0	100.0	100.0	100.0

1 ottimo 3 carente

2 buono 4scadente

Regole per la definizione dell'indicatore

if (a1=2 and a2=2) val=1
if ((a1=1 and a2<3) or (a1=2 and a2=1)) val=2
if (a1=3 and a2<3) val = 3
else val=4

dove a1 (righe) è '*impianto di riscaldamento*'

a2 (colonne) è '*vasca o doccia*'

Tabella 5.11 – Costruzione di un indicatore del livello di servizi interni (4 categorie)

Esempio 3 - Costruzione di una nuova variabile: la tipologia edilizia

La tabella 5.12 illustra la costruzione di una nuova variabile sintetica, la ***tipologia edilizia***, otttenuta dall'esame dell'incrocio tra

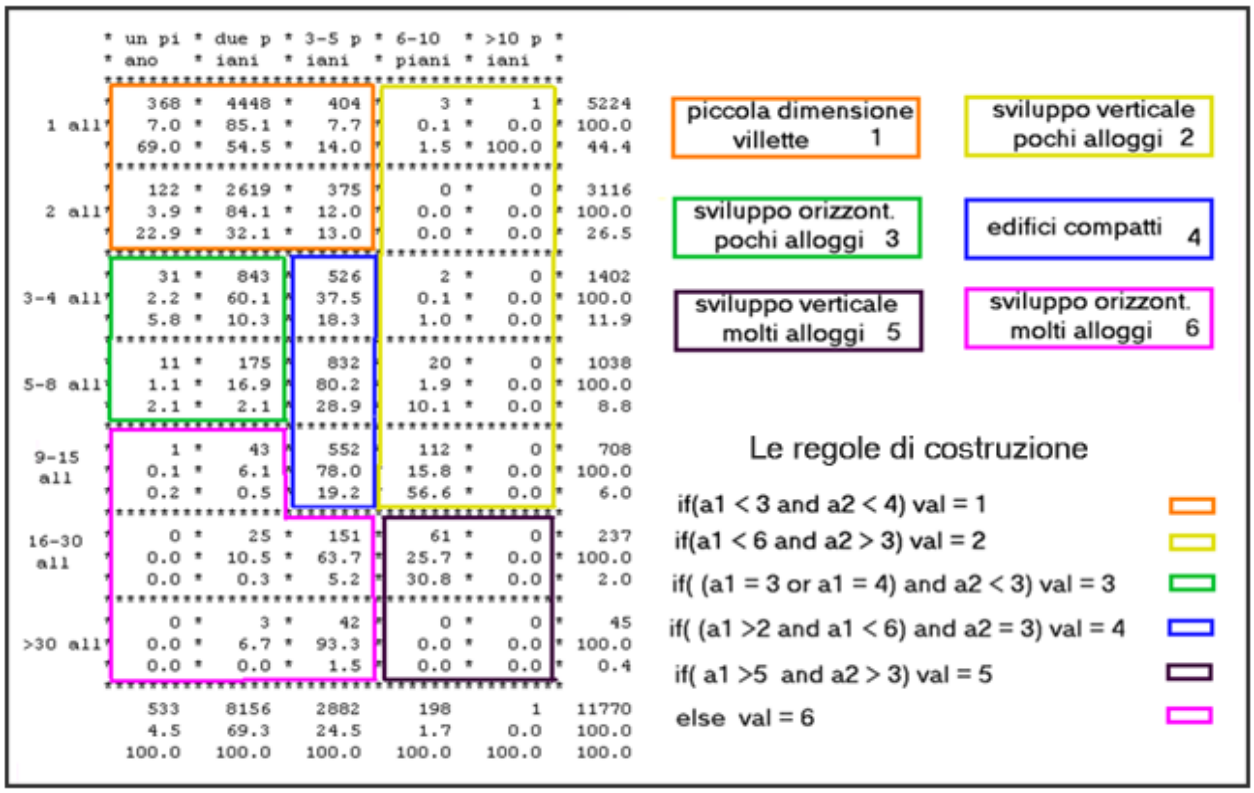
il ***numero di alloggi*** nell'edificio (sette categorie di riga) qui indicato con **a1**;

il ***numero di piani*** dell'edificio (cinque categorie di colonna), qui indicato con **a2**.

Ciascuno degli 11770 alloggi viene assegnato alla cella corrispondente alla combinazione delle caratteristiche dell'edificio in cui si trova.

Molte celle, corrispondenti agli edifici più alti o con un grande numero di alloggi, risultano vuote o quasi: nel Comune considerato gli edifici sono per lo più bassi, e mono- o bi-familiari.

Un esame della tabella consente di delimitare sei gruppi di celle definendo una nuova variabile categoriale a sei categorie, che si può salvare ed utilizzare in analisi ulteriori. **Ovviamente, è l'analista che decide, a suo giudizio e per i suoi obiettivi, il modo in cui la tavola viene ritagliata**. Il significato della suddivisione adottata qui è illustrato nella figura.



	un pi ano	due piani	3-5 piani	6-10 piani	>10 piani	
1 all	368	4448	404	3	1	5224
	7.0	85.1	7.7	0.1	0.0	100.0
	69.0	54.5	14.0	1.5	100.0	44.4
2 all	122	2619	375	0	0	3116
	3.9	84.1	12.0	0.0	0.0	100.0
	22.9	32.1	13.0	0.0	0.0	26.5
3-4 all	31	843	526	2	0	1402
	2.2	60.1	37.5	0.1	0.0	100.0
	5.8	10.3	18.3	1.0	0.0	11.9
5-8 all	11	175	832	20	0	1038
	1.1	16.9	80.2	1.9	0.0	100.0
	2.1	2.1	28.9	10.1	0.0	8.8
9-15 all	1	43	552	112	0	708
	0.1	6.1	78.0	15.8	0.0	100.0
	0.2	0.5	19.2	56.6	0.0	6.0
16-30 all	0	25	151	61	0	237
	0.0	10.5	63.7	25.7	0.0	100.0
	0.0	0.3	5.2	30.8	0.0	2.0
>30 all	0	3	42	0	0	45
	0.0	6.7	93.3	0.0	0.0	100.0
	0.0	0.0	1.5	0.0	0.0	0.4
	533	8156	2882	198	1	11770
	4.5	69.3	24.5	1.7	0.0	100.0
	100.0	100.0	100.0	100.0	100.0	100.0

Tabella 5.12 – Esempio di costruzione di una variabile ***tipologia edilizia*** con sei categorie

Cap. 6. - Menu di Analisi: le Analisi Fattoriali

6.1 - Rappresentazione geometrica e linguaggio

Si consideri una tavola di dati **X***(n,p)* in cui le righe rappresentino un insieme I di n unità statistiche, e le colonne riportino i valori assunti da p variabili su quelle unità. ***X*** è una tavola di descrizione e per semplicità supporremo per il momento che tutte le variabili siano quantitative. Comunque, le definizioni ed i concetti che introdurremo si possono estendere ad ogni tipo di tavola.

Il comportamento di ogni unità-riga della tavola (si pensi ad un Comune) è rappresentato da un vettore di p numeri reali ordinati corrispondenti ai valori delle p variabili osservati su quell'unità. Le p componenti di tale vettore si possono interpretare come le coordinate di un punto in uno spazio vettoriale (geometrico) R^p a p dimensioni e l'unità si può identificare con quel punto.

L'insieme I delle n unità si può rappresentare con una **nuvola di punti dotati di peso** (si ricordi che c'è un peso associato a ciascuna unità). Si può pensare ad ogni altro punto di R^p come ad una unità **virtuale** (vale a dire ad una combinazione di valori delle p variabili descrittive) che sarebbe forse possibile incontrare in un altro campione o che può avere un particolare significato per la nuvola, come ad esempio il suo punto centrale (baricentro). La figura 6.1 mostra un esempio nel semplice caso di due variabili.

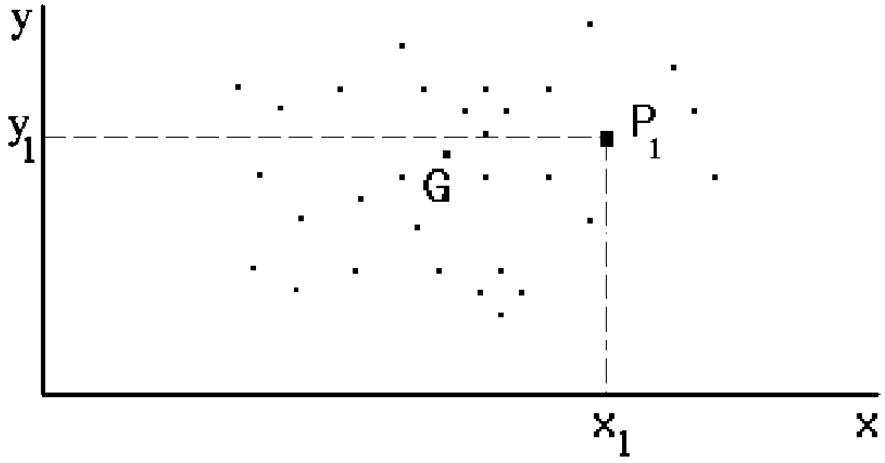


Figura 6.1 - Rappresentazione geometrica di un insieme di unità descritte da due variabili. **G** è il centro di gravità della nuvola.

Nella figura su ciascuno dei due assi ortogonali sono stati riportati i valori di una variabile. Ogni punto del piano risulta biunivocamente associato ad una coppia di valori (cioè ad una opportuna unità, reale o virtuale). Se le variabili fossero tre sarebbero necessari tre assi ortogonali per rappresentarle e la visualizzazione della rappresentazione geometrica sarebbe ancora possibile. Quando le variabili sono più di tre la nostra mente tridimensionale non riesce a visualizzare una

rappresentazione, ma la trattazione matematica utilizzata nel caso di due o tre variabili può venire generalizzata senza sforzo alcuno al caso di *p* dimensioni. Considereremo dunque come generale il caso di una nuvola di *n* punti-unità (o punti-oggetto) in R^p, ma si può continuare a pensare intuitivamente al caso bi-dimensionale senza perdita di generalità.

Possiamo anche guardare alla tavola **X** *per colonne*. Ogni colonna è un vettore di *n* numeri che rappresentano i valori assunti da una variabile sulle *n* unità. Esso si può identificare con un punto di uno spazio geometrico n-dimensionale R^n. In questo caso si ha una nuvola di *p* punti-variabile in R^n. La tavola ammette quindi *due rappresentazioni geometriche*: come nuvola di *n* punti-oggetto in R^p o rispettivamente come nuvola di *p* punti-variabile in R^n. Quanto al contenuto informativo, le due rappresentazioni geometriche risultano perfettamente equivalenti alla descrizione numerica offerta dalla tavola **X**. Sembra naturale centrare l'attenzione sulla nuvola degli *n* punti-oggetto in R^p per analizzare le differenze esistenti tra le unità rispetto alle variabili descrittive; tuttavia, anche l'altra rappresentazione può risultare utile.

I due spazi R^p ed R^n sono *duali*. In generale, risulta conveniente studiare in R^p le relazioni intercorrenti tra le unità (ad esempio, due unità globalmente simili sulle *p* variabili sono associate a due punti di R^p tra loro vicini, ecc.) centrando invece l'attenzione sulla nuvola in R^n per analizzare le relazioni tra le variabili (due variabili non correlate sono rappresentate da punti che giacciono in direzioni ortogonali rispetto all'origine, mentre due variabili altamente correlate giacciono in direzioni che formano tra loro un angolo piccolo, ecc.).

6.1.1 - La distanza

Come **indicatore globale** della ***dissimilarità*** tra due oggetti viene assunta la distanza tra i loro punti rappresentativi in R^p: tutte le variabili contribuiscono alla sua determinazione. Se si tratta di una tavola di descrizione quantitativa la *dissimilarità globale* tra le unità *i* e *k* si calcola come

$$d^2(i,k) = (x_{i1}-x_{k1})^2 + \ldots + (x_{ip}-x_{kp})^2$$

usando la distanza pitagorica tra i punti *i* e *k* di R^p (vedremo come si adotti una *metrica* diversa - cioè una diversa definizione di distanza - nel caso di una tavola di contingenza).

Poiché si debbono sommare i contributi alla distanza provenienti dalle diverse variabili, queste debbono essere espresse nella medesima unità di misura, oppure essere a-dimensionali. È anche conveniente equilibrare i diversi contributi, in modo da evitare la dominanza di qualche variabile solo in virtù dell'unità di misura in cui essa è espressa. Un modo per farlo è di **standardizzare** tutte le variabili prima dell'analisi, ed in ADDATI la cosa è realizzata automaticamente dalle routines di calcolo. Una volta standardizzata (cioè **centrata** sottraendo da ciascun valore assunto dalla variabile il suo valor medio e **ridotta** dividendo i valori così ottenuti per il suo scarto quadratico medio), ogni variabile viene ad avere media 0 e varianza 1.

6.1.2 - Il centro di gravità della nuvola

Il centro di gravità della nuvola di punti-oggetto in R^p è il punto **G** che ha come coordinate i valori medi delle *p* variabili. Esso può essere considerato come un oggetto *virtuale* che rappresenta i caratteri medi dell'intero sistema. Se le variabili sono centrate (cioè se hanno tutte media zero) il centro della nuvola ovviamente coincide con l'origine del sistema di riferimento (**G ≡ O**) e la nuvola stessa è detta **centrata**.

L'analisi che vogliamo eseguire s'interessa alle differenze di comportamento che esistono tra le *n* unità, e mira ad individuare le variabili cui tali differenze vanno ascritte. Da un punto di vista geometrico, si vuole osservare **di quanto** ed **in qual modo** ciascuna unità differisca dal comportamento medio dell'intero sistema, rappresentato dal centro **G** della nuvola (coincidente con l'origine **O** se la nuvola è centrata). È ragionevole assumere come indicatore di "*quanto*" la **distanza** di ciascun punto-oggetto da **G** ed associare "*in qual modo*" con la direzione di tale elongazione (cioè con le variabili che più contribuiscono a determinare tale distanza).

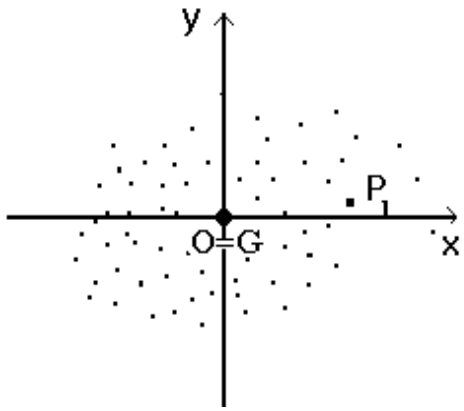


Figura 6.2 - La nuvola di figura 6.1 centrata.

6.1.3 - L'inerzia della nuvola

Supponiamo la nuvola centrata. Si dice **inerzia dell'unità** *i* rispetto al centro **G≡O** il prodotto della massa di *i* per il quadrato della sua distanza da **O**:

$$\text{Inerzia}(i) = m_i\, d^2(\mathbf{x_i},\mathbf{O}) = \Sigma_j\, m_i\, x_{ij}^{\,2}$$

Come misura della dispersione della nuvola si assume la sua **Inerzia totale** $\text{In}_{tot}(I)$**,** pari alla somma delle inerzie di tutti i suoi punti :

$$\text{In}_{tot}(I) = \Sigma_i\, m_i\, d^2(\mathbf{x_i},\mathbf{O})$$

L'inerzia della nuvola ha un'interpretazione semplice: essa nasce dalle differenze di comportamento tra le unità, cioè dal fatto che le variabili assumono in generale diversi valori in corrispondenza alle diverse unità ed hanno dunque varianza non nulla in *I*. In caso contrario la nuvola collasserebbe nel suo centro e l'inerzia sarebbe nulla.

E' facile verificare che **l'Inerzia totale è pari alla somma delle varianze delle *p* variabili**

$$\text{In}_{tot}(I) = \Sigma_i\, m_i\, d^2(\mathbf{x_i},\mathbf{O}) = \Sigma_i m_i(\Sigma_j x_{ij}^{\,2}) = \Sigma_j(\Sigma_i m_i x_{ij}^{\,2}) = \Sigma_j\, \mathbf{var}(j)$$

In particolare, quando le p variabili sono standardizzate il contributo di ciascuna di esse all'Inerzia vale 1 e l'Inerzia totale risulta dunque $\mathbf{In_{tot} = p}$.

6.1.4 - Interpretazione delle relazioni tra le variabili in R^n

In R^n ogni punto si può interpretare come una variabile, cioè come un vettore di *n* valori ordinati (le sue coordinate) misurate sulle *n* unità. Se le variabili sono centrate si può dimostrare che:

- la distanza di un punto-variabile *j* dall'origine è pari alla deviazione standard della variabile. Ne segue che se le variabili sono standardizzate tutti i loro punti rappresentativi giacciono sulla superficie di una sfera centrata sull'origine e di raggio 1;

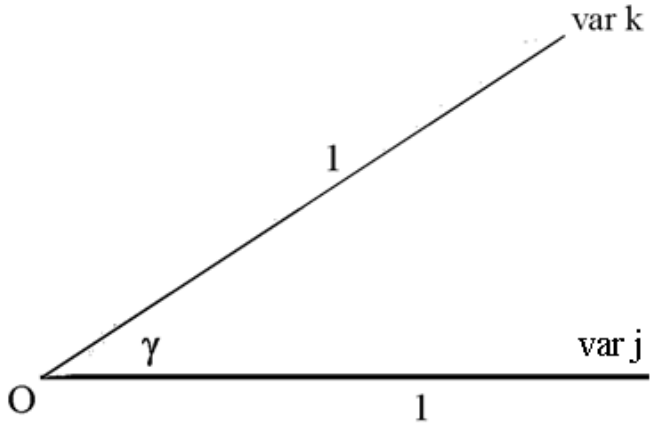


Figura 6.3 - Rappresentazione di punti-variabile in R^n. Si può dimostrare che **cos** γ **=** ***corr*****(var j, var k),** cioè che il coseno della distanza angolare tra i due punti-variabile *j* e *k* misura la correlazione tra le due variabili.

- la correlazione tra due variabili centrate *j* e *k* è pari al coseno dell'angolo formato dai due segmenti che congiungono i loro punti rappresentativi con l'origine. Va ricordato che la correlazione tra due variabili quantitative è una misura della forza della loro associazione sull'insieme *I*: essa può variare tra +1 (associazione positiva perfetta) e -1 (associazione perfetta negativa). Se le variabili sono standardizzate due variabili che abbiano correlazione +1 sono rappresentate da due punti coincidenti, mentre due variabili a correlazione -1 sono rappresentate da due punti opposti rispetto all'origine.

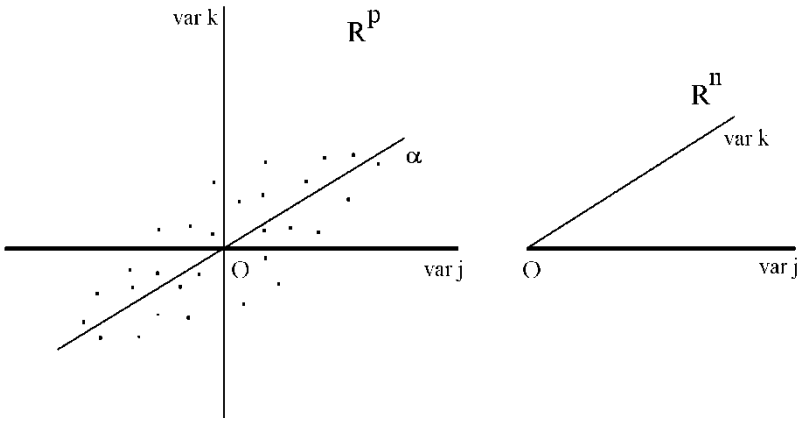


Figura 6.4 - La nuvola nei due spazi R^p e R^n. In R^p la nuvola è centrata e la sua proiezione su qualunque asse per O (ad esempio, l'asse α) risulta pure centrata. Le due variabili *j* e *k* sono altamente correlate. Il fatto si può vedere dalla forma allungata della nuvola in R^p e dalla piccola distanza angolare tra i due punti-variabile in R^n.

Le proprietà delle due nuvole (oggetti e variabili) sono dunque diverse: se le variabili sono centrate l'origine di R^p è il centro della nuvola di punti-oggetto, attorno al quale essi sono sparsi, con un

livello di dispersione misurato dall'inerzia della nuvola. Se si proietta la nuvola su di un asse qualunque per l'origine, la nuvola uni-dimensionale che si ottiene risulta anch'essa centrata. Nell'altro spazio, poiché la distanza angolare tra due punti è legata alla correlazione tra le variabili corrispondenti, la nuvola risulta in generale distribuita in modo non bilanciato attorno all'origine. Sse tutte le variabili hanno un'alta correlazione positiva la nuvola dei punti-variabile giace da una stessa parte rispetto ad **O**, senza alcuna simmetria. Questa differenza, che influenza l'interpretazione dei risultati analitici nei due spazi, è una conseguenza del diverso significato delle righe e delle colonne della tavola dei dati e del trattamento non simmetrico cui esse vengono sottoposte (la media è calcolata per le colonne e non per le righe; le colonne sono standardizzate, non le righe; ecc.).

6.2 – Introduzione alle analisi fattoriali: ACOMP ed ACORR

Non è questa la sede per una trattazione approfondita della teoria che sta a fondamento delle due Analisi Fattoriali incluse in ADDATI: l'Analisi in Componenti Principali e l'Analisi delle Corrispondenze. Si tratta certamente di un capitolo oltremodo interessante ed utile della Statistica, ma che va oltre i limiti di una Guida all'Uso. Comunque, l'utente che voglia padroneggiare ed utilizzare al meglio queste potenti tecniche statistiche come strumenti esplorativi e non semplicemente per ridurre la dimensionalità della descrizione in vista di una successiva classificazione, dovrebbe approfondirne la teoria ricorrendo a qualche testo specifico.

ACOMP ed **ACORR** sono molto simili. Entrambe accettano come input una tavola (anche molto grande) di dati ed esplorano le relazioni che intercorrono tra i suoi elementi (righe e colonne). Lo scopo è di semplificare la rappresentazione riconoscendo (cioè *costruendo* opportunamente) un **numero limitato di nuove variabili** sottogiacenti (dette **fattori**) sufficienti a riassumere gli aspetti più rilevanti della descrizione con una perdita di dettagli accettabile. Ciò si ottiene ruotando in un modo ottimale - rispetto alla nuvola - il sistema di riferimento nello spazio geometrico in cui il fenomeno è rappresentato (secondo quanto esposto nel paragrafo precedente, ogni riga ed ogni colonna della tavola si possono rappresentare come punti in uno spazio geometrico opportunamente definito).

La differenza tra le due analisi sta nella natura della tavola trattata:

- una **tavola di descrizione quantitativa** o di **variabili binarie** nel caso di **ACOMP**;
- un **tavola di contingenza** o una **tavola di variabili categoriali** nel caso di **ACORR.**

Entrambe le tecniche operano una trasformazione preliminare della tavola in input, diversa nei due casi.

Descriveremo più avanti in dettaglio i parametri di cui **ACOMP** abbisogna per un corretto funzionamento ed il modo di introdurli. Poiché **ACORR** pone all'incirca le medesime domande, ci limiteremo in quel caso ad una descrizione più succinta.

6.3 - L'Analisi in Componenti Principali (ACOMP)

Funzione Analizza una tavola di descrizione costituita da variabili **quantitative** o **binarie** (categoriali con esattamente due categorie).
Le variabili binarie incluse nella tavola da analizzare vengono ricodificate *al volo*: ciascuna di esse viene convertita in una nuova variabile con valore 1 per le unità che **assumono la prima categoria**, 0 per le altre; questo qualunque siano i codici originari. Una tale variabile può essere trattata come quantitativa ed elaborata con ACOMP.
Tutte le variabili vengono standardizzate dal programma: ognuna avrà media 0, varianza 1, e la stessa importanza nell'analisi.

L'utente può disattivare quest'operazione di standardizzazione e lavorare a partire dalla tavola delle varianze-covarianze invece della tavola delle correlazioni (quest'ultima è l'opzione più frequente). Tuttavia tale decisione dovrebbe essere presa solo da utenti esperti ed in circostanze particolari: meglio, in generale, attenersi all'opzione di default ed usare la matrice delle correlazioni.

Limiti La tavola da diagonalizzare viene preparata via via che vengono letti i casi (records) presenti nel file dei dati. Il tempo di calcolo aumenta proporzionalmente al numero dei casi, mentre la richiesta di memoria ne dipende solo per il numero degli indicatori da memorizzare, ed aumenta invece con il numero delle variabili trattate (le colonne della tavola), sia attive che supplementari.

Di fronte ad una diagnostica di memoria insufficiente è giocoforza ridurre il numero delle variabili complessive. Ma con tavole di dimensioni 'normali' il problema non dovrebbe porsi: si pensi che ADDATI è stato utilizzato per analizzare una tavola di circa cinque milioni di unità statistiche, ed una ventina di variabili! Comunque, poiché il programma è in grado di trattare un numero elevato di variabili (dell'ordine delle centinaia), ridurne il numero è consigliabile più per la necessità di ottenere risultati interpretabili con un minimo di chiarezza che per i limiti di memoria.

Consigli Le variabili da sottoporre all'analisi vanno scelte in modo avveduto, evitando di vanificare la potenza esplorativa del metodo con variabili scelte alla rinfusa. Giova ribadire che l'abilità dell'analista si manifesta soprattutto nella costruzione di ***tavole essenziali***, determinate anche a seguito di più tentativi.

Nota *Il programma esegue la standardizzazione automaticamente. I valori di ogni variabile vengono traslati in modo tale che la loro media sia nulla, cioè ogni valore è sostituito dalla sua differenza rispetto alla media corrente della variabile. Inoltre, per ogni variabile la scala di misura viene opportunamente cambiata dividendo tutti i suoi valori per la deviazione standard, così che tutte vengano ad avere varianza unitaria ed assumano la medesima rilevanza nell'analisi che segue.*

Si consideri la figura 6.5 che mostra un caso semplice, con solo due variabili (standardizzate). Si può rappresentare ogni unità statistica come un punto di uno spazio a due dimensioni R^2. La forma della nuvola è un ovale allungato, e ciò significa che le variabili sono fortemente correlate: il valore

di una variabile (ad es., la *y*) può essere desunto con buona approssimazione quando sia noto il valore dell'altra, e viceversa. La seconda variabile ripete dunque - almeno in parte - l'informazione già apportata dalla prima: solo una piccola parte della sua informazione è effettivamente originale (cioè indipendente dall'altra, non ripetitiva).

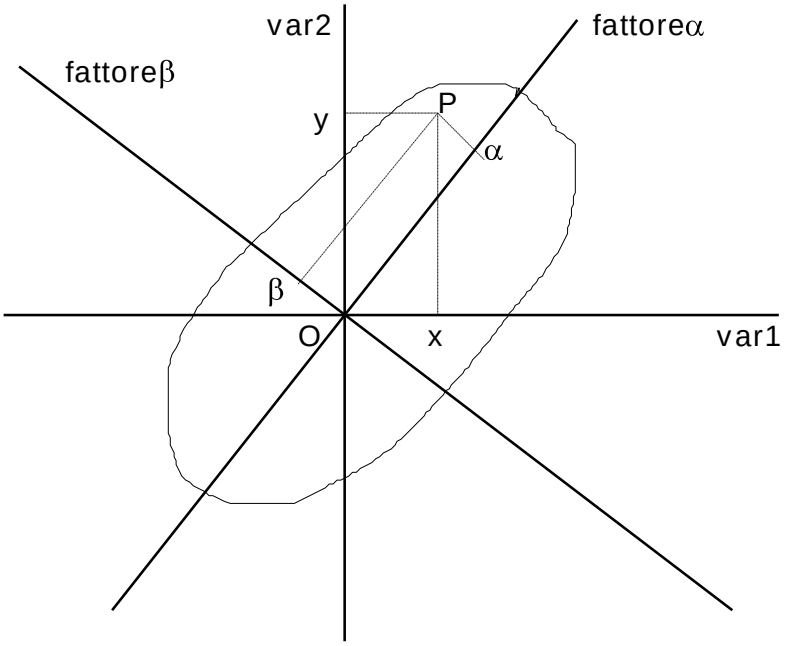


Figura 6.5 - Un generico punto P è rappresentato dalla coppia di coordinate (x, y) relativamente alle variabili originali, dalla coppia di coordinate (α, β) nel nuovo sistema di riferimento ruotato.

Consideriamo ora il fascio di tutte le possibili linee rette per l'origine O. Se si proiettano tutti i punti-oggetto su una qualsiasi di esse si ottiene una nuvola uni-dimensionale dispersa attorno ad **O**; a misura della sua dispersione si assume la sua **Inerzia**, definita nel paragrafo 6.1.

In particolare, l'inerzia che la nuvola mantiene quando venga proiettata sui due assi coordinati *x* e *y* vale 1. Essa è cioè pari alla varianza della corrispondente variabile, che è appunto 1 in questo caso per via della standardizzazione cui le variabili di partenza sono state sottoposte. **Su qualunque altro asse per O l'inerzia della nuvola proiettata è in generale diversa da 1**.

Esiste un asse - indicato in figura con α - sul quale la nuvola si proietta mantenendo la massima inerzia possibile, cioè conservando al meglio le distanze tra i suoi punti. E' questo il **primo asse fattoriale** e la distanza da **O** (presa con segno) della proiezione di un punto è la prima **coordinata fattoriale** di quel punto.

La nuvola si proietta sull'asse β normale ad α con un'inerzia molto minore. Questo completa la descrizione nel caso di due dimensioni. Si può facilmente dimostrare che la somma delle due inerzie sugli assi α e β **vale esattamente 2**, cioè è pari all'inerzia totale della nuvola.

Ogni coppia di assi ortogonali per **O** conduce ad una particolare **decomposizione dell'inerzia**. Il vantaggio della coppia (α, β) rispetto ad (*x*, *y*) sta nel fatto che la piccola frazione di inerzia "*spiegata*" dall'asse β può essere ignorata senza grande perdita d'informazione. Ciò porta ad una semplificazione uni-dimensionale accettabile dell'originaria descrizione bi-dimensionale.

Queste semplici considerazioni si possono estendere facilmente al caso di *p* variabili: l'insieme delle unità statistiche è rappresentato da una nuvola di *n* punti in uno spazio a *p* dimensioni. Il valore dell'inerzia totale, una volta standardizzate le variabili, è p. È sempre *possibile* - e *conveniente* quando almeno alcune delle variabili siano tra loro correlate - determinare un asse (detto *primo asse principale d'inerzia*) sul quale la nuvola si proietta conservando la massima inerzia possibile: il valore di tale inerzia è noto come l'**autovalore** associato all'asse. Si determina poi un secondo asse, ortogonale al primo, il quale spiega la massima frazione dell'inerzia residua e così via, finché la descrizione sia completa.

L'insieme dei nuovi assi costituisce un nuovo sistema di riferimento, alternativo a quello iniziale. Il fenomeno rappresentato è sempre il medesimo, ma è mutato il punto di vista da cui lo si osserva e ciò consente di focalizzare gli aspetti più rilevanti, espressi dai primi fattori. Poiché i fattori vengono ordinati in modo decrescente secondo l'inerzia spiegata da ciascuno (cioè secondo valori decrescenti dell'autovalore associato), il fatto di ignorare gli ultimi fattori può portare ad una riduzione nella dimensionalità della descrizione al costo di una perdita d'informazione non eccessiva.

6.3.1 - Un esempio

Illustreremo l'uso di **ACOMP** e l'interpretazione dei risultati su di un semplice esempio didattico relativo al Centro Storico di Venezia. I dati per l'elaborazione sono tratti dal Censimento della Popolazione e delle Abitazioni del 1981.

Il Censimento non definisce (né misura direttamente) il "**disagio abitativo**". Si tratta di un concetto complesso, che cercheremo di costruire a partire dalla simultanea considerazione dei valori di alcune variabili direttamente rilevate o facilmente calcolabili. Per ciascun alloggio, si determina se esso utilizzi o meno il piano terra o quello rialzato per scopi abitativi (per Venezia, si tratta di un indicatore di insalubrità per via della risalita d'umidità); la sua dotazione di servizi interni (bagno e riscaldamento) secondo tre categorie; l'affollamento pure in tre categorie.

Il **livello globale di disagio abitativo** in una sezione censuaria risulta implicitamente definito *dalla distribuzione dei caratteri di tutti i suoi alloggi*. Dopo l'analisi esplorativa con **ACOMP**, sezioni aventi profili distributivi simili verranno aggregate in una medesima classe; si otterranno cinque classi a diverso livello di disagio, come si vedrà quando passeremo ad illustrare il funzionamento di **NONGER**.

Il file di input VENEZIA.DAT, documentato in VENEZIA.TXT, ha 148 records. Esso è incluso tra i Dataset di esempio disponibili dalla pagina di ADDAWIN (cidoc.iuav.it/addawin.html).

Il file viene caricato in ADDAWIN (***File→Apri un Data Set***): al solito, il primo passo è il calcolo di alcune statistiche elementari, correlazioni, ecc., prima di cimentarsi con l'Analisi in Componenti Principali e la Classificazione.

Ogni record riguarda una sezione censuaria e contiene nell'ordine i seguenti campi (separati da spazi):

1. identificatore della sezione (è semplicemente il numero della sezione);
2. numero di alloggi occupati nella sezione (da usare come **peso** – vedi più avanti).

Variabili attive:

1. % di alloggi che utilizzano esclusivamente i *piani alti* (a scopo residenziale)*;*
2. % di alloggi che utilizzano (anche parzialmente) il *piano terra* o il piano rialzato a fini residenziali;

3. % di alloggi con *dotazione completa* di servizi interni (bagno, riscaldamento);

4. % di alloggi con *dotazione carente*;

5. % di alloggi con *dotazione scadente*;

6. % di alloggi *sovraffollati*;

7. % di alloggi in condizioni di affollamento *standard*;

8. % di alloggi *sottoutilizzati*.

Variabili supplementari :

9. % di alloggi con capofamiglia di *status elevato* (dirigenti, liberi professionisti, imprenditori, impiegati con titolo di studio superiore);

10. % di alloggi con capofamiglia *operaio;*

11. % di alloggi con capofamiglia di *altro status*.

I primi due campi contengono l'indicatore della sezione (dichiarato di tipo ID) ed il suo peso nell'analisi. Gli altri undici contengono i valori di altrettante variabili: le prime otto descrivono sommariamente le condizioni del patrimonio abitativo nella sezione censuaria e saranno assunte come **attive** nell'analisi; le altre tre descrivono la distribuzione dei capifamiglia su tre livelli di status e saranno assunte come **supplementari**.

Nota *Si può pensare che il Data Set sia stato costruito a partire da un altro che* ***contava*** *le abitazioni nelle diverse condizioni, passando da valori assoluti a percentuali mediante l'opzione* ***Utilità→ Crea nuove Variabili/DataSet****. Nel primo le variabili erano di tipo COUNT, nel secondo di tipo QUANTITATIVE.*

L'obiettivo dell'analisi è rispondere alle due seguenti questioni:

- il disagio abitativo, definito dalle tre variabili "uso del piano terra" (due categorie), "livello dei servizi interni all'alloggio" (tre categorie) e "livello di affollamento" (tre categorie) è distribuito in modo omogeneo in tutta la città, oppure esistono delle differenze spazialmente ben definite, legate al "pregio" di mercato della zona, che stimola interventi di risanamento distribuiti in modo non uniforme? In altri termini, si possono individuare zone con un livello di disagio superiore a quello medio cittadino, contrapposte ad altre con disagio abitativo inferiore?
- se la città non è omogenea dal punto di vista del disagio, fino a che punto emerge evidente una segregazione nell'uso del patrimonio abitativo? Fino a che punto cioè si manifesta evidente una concentrazione degli strati sociali meno abbienti nelle zone a maggior disagio residenziale?

Gli assi fattoriali (ed i gruppi prodotti dalla successiva classificazione) verranno costruiti solo sulla base delle relazioni che intercorrono tra le otto variabili che descrivono il patrimonio abitativo, ma la posizione delle tre variabili supplementari nello spazio fattoriale ci permetterà di rispondere al secondo quesito.

Si noti che le undici variabili **non sono indipendenti**. Ad esempio, le prime due (uso del piano terra o dei piani alti) sono legate dalla ovvia relazione analitica :

(quota di alloggi ai piani alti) = 100 - (quota di alloggi ai piani terra)

e la loro correlazione vale esattamente -1. Analogamente, anche le tre variabili relative alla dotazione dei servizi sono vincolate ad avere somma 100, così come le tre variabili che descrivono gli stati di affollamento. Ciò significa che lo spazio in cui la nuvola dei punti-sezione è immersa non ha dimensione pari ad 8 (numero delle variabili attive) ma dimensione intrinsecamente minore. La tavola delle variabili attive risulta dunque **ridondante** dal punto di vista informativo e sarebbe

possibile eliminare (o trattare come supplementari) alcune colonne attive senza cambiare l'esito dell'analisi.

La dipendenza tra le variabili si traduce nella nullità di alcuni autovalori (il che corrisponde ad una riduzione delle dimensionalità effettiva del fenomeno).

Per riconoscerle immediatamente, marcheremo le parti concernenti l'esempio con il simbolo "🗁".

6.3.2 - L'inserimento dei parametri di controllo dell'analisi

Si lanci ACOMP (*Analisi→Analisi in Componenti Principali*) dal Menu. Appare il dialogo mostrato nella figura 6.6, dove vanno inseriti i parametri di controllo dell'analisi.

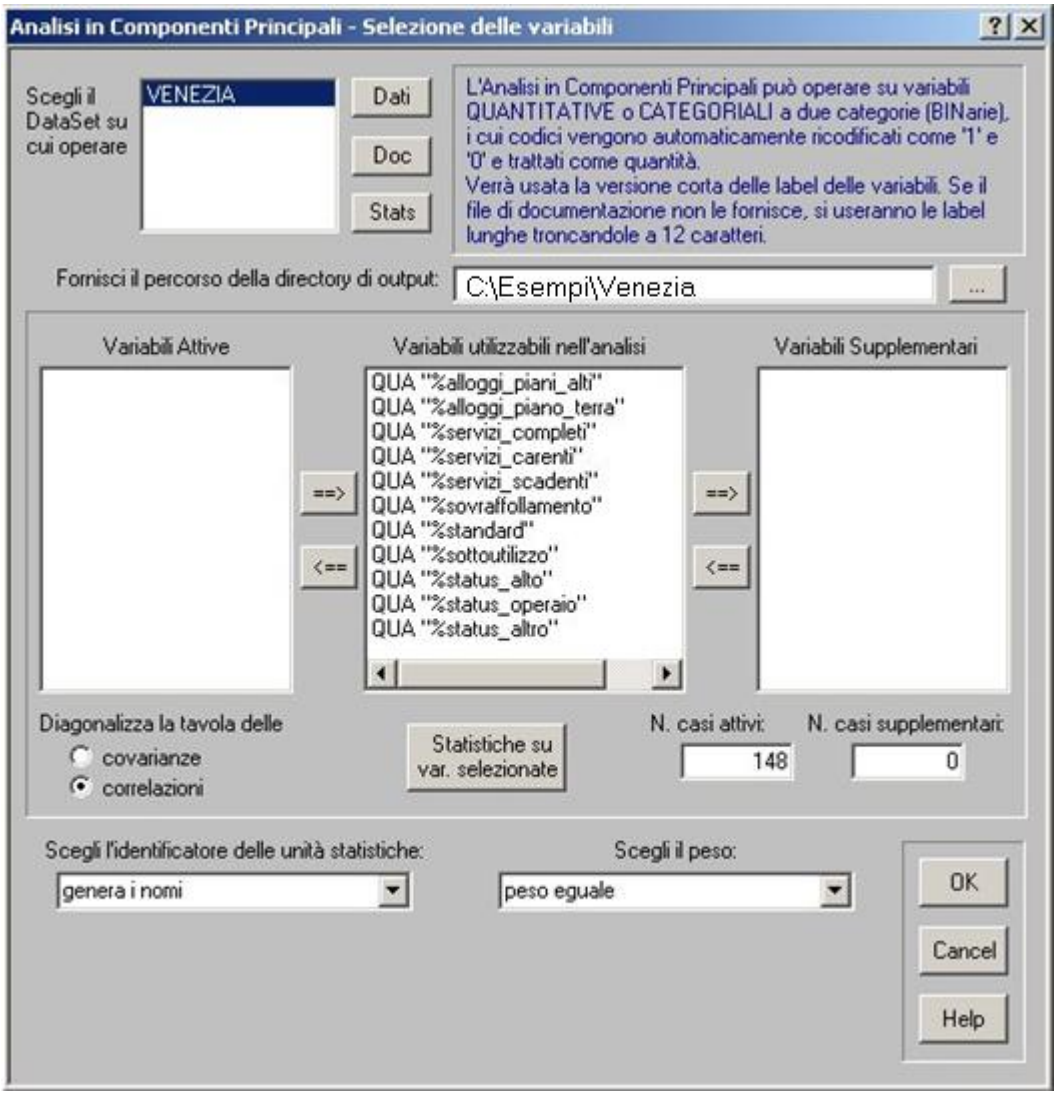


Figure 6.6 – ACOMP: il dialogo principale

Innanzitutto, va selezionato dalla lista in alto a sinistra il DataSet che contiene le variabili da trattare. Nel nostro caso è stato caricato solo il DS Venezia, ma non è sempre così: i DataSet caricati o prodotti nel corso della stessa sessione di lavoro possono essere numerosi.

Al centro del dialogo appaiono tre liste: quella di mezzo mostra tutte le variabili nel DS che si possono analizzare con ACOMP. La cosa dipende dalla loro scala di misura, come spiega la nota in alto a destra. L'analista seleziona le variabili da usare come attive o supplementari spostandole nella lista appropriata. Quelle lasciate nella lista centrale **verranno ignorate** nell'analisi.

- Va inserito il percorso della directory di output, dove verranno scritti tutti i file prodotti dall'analisi. In dettaglio, essi sono:

  1. un file denominato ACOMP-nn.OUT dove vengono scritte le informazioni numeriche dettagliate da usare nell'interpretazione dei risultati ('nn' è un numero progressivo per evitare di sovrascrivere output di analisi precedenti esistenti nella stessa cartella);
  2. un file con nome uguale alla label del Dataset analizzato ed estensione FPL (VENEZIA.FPL nel nostro caso); esso registra i dati che l'utilità FACPLAN userà per mostrare le proiezioni di variabili ed unità statistiche sui principali piani fattoriali;
  3. un file con nome uguale alla label associata al DS analizzato ed estensione .PCS (VENEZIA.PCS nel nostro caso), dove 'PCS' sta per Principal ComponentS; esso registra le coordinate fattoriali che l'analista decide di passare alla procedura di Classificazione (vedi più avanti);
  4. un file con nome uguale alla label associata al DS analizzato ed estensione .TMP (VENEZIA.TMP nel nostro caso), salvato insieme al file PCS, dove sono registrati i valori delle variabili analizzate.

- Si indichi un **identificatore** (cioè un campo di tipo ID), esistente nel record, per identificare le unità. In alternativa, si può chiedere al programma di generare automaticamente una lista di identificatori progressivi. Se le unità statistiche sono aree amministrative per le quali è disponibile un file cartografico (ad es., uno shapefile ESRI) l'id naturale da usare è il codice cartografico associato ad ogni unità. Esso consentirà all'analista di produrre una mappa della classificazione ottenuta.

- Si selezioni una variabile, tra quelle elencate nel Combo in basso a destra, da assegnare come **peso** alle unità statistiche. Si può usare come peso ogni variabile di tipo QUANTITATIVE o COUNT (la popolazione, di solito dichiarata come variabile di tipo COUNT, è una scelta comune quando si analizzano aree amministrative). In alternativa, quando è appropriato, si può assegnare il medesimo peso a tutte le unità. Si tratta della scelta tipica quando si lavori con unità al massimo livello di disaggregazione (famiglie, individui, o simili). Va però ricordato che, se si stanno analizzando dei dati d'inchiesta e si vogliono risultati riportati all'universo, si deve usare come peso il fattore di campionamento (*raising factor*) definito nella preparazione del campione.

- Oltre alle unità **attive**, sulle cui relazioni con le variabili verranno costruiti i fattori, si possono includere nell'analisi delle unità **supplementari**, che non contribuiscono alla costruzione dei fattori ma possono tuttavia essere proiettate sui piani fattoriali, e per le quali si possono calcolare i **contributi relativi** (quelli assoluti sono ovviamente nulli, come sarà evidente tra breve). Lo scopo è di ottenere maggiori informazioni esaminando come si collocano le unità supplementari rispetto a quelle attive ed alle variabili.
  Di norma, tutte le unità del DS vengono trattate come attive. Si può cambiare il numero delle unità attive ed aggiungerne di supplementari, ma si tenga presente che le prime unità nel DS, nel numero richiesto, vengono assunte come attive e le unità supplementari, se ve ne sono, sono quelle che seguono immediatamente, nel numero specificato. Tale condizione, alquanto restrittiva, può rendere necessario un riordino preliminare dei record del DataSet; essa potrebbe venire rilassata nelle future versioni del pacchetto.

Esempio *Se le righe attive descrivono un gruppo di Comuni ad una certa data, le righe supplementari possomno descrivere gli stessi Comuni ad una data diversa, permettendo una visualizzazione qualitativa ed un'interpretazione dei cambiamenti avvenuti.*

- Di norma, l'analisi elabora la tavola delle correlazioni tra le variabili. Tale scelta non va cambiata, a meno che non si sia particolarmente esperti e si abbiano buone ragioni per farlo.

La figura 6.7 mostra il dialogo riempito con i parametri di controllo dell'analisi su Venezia.

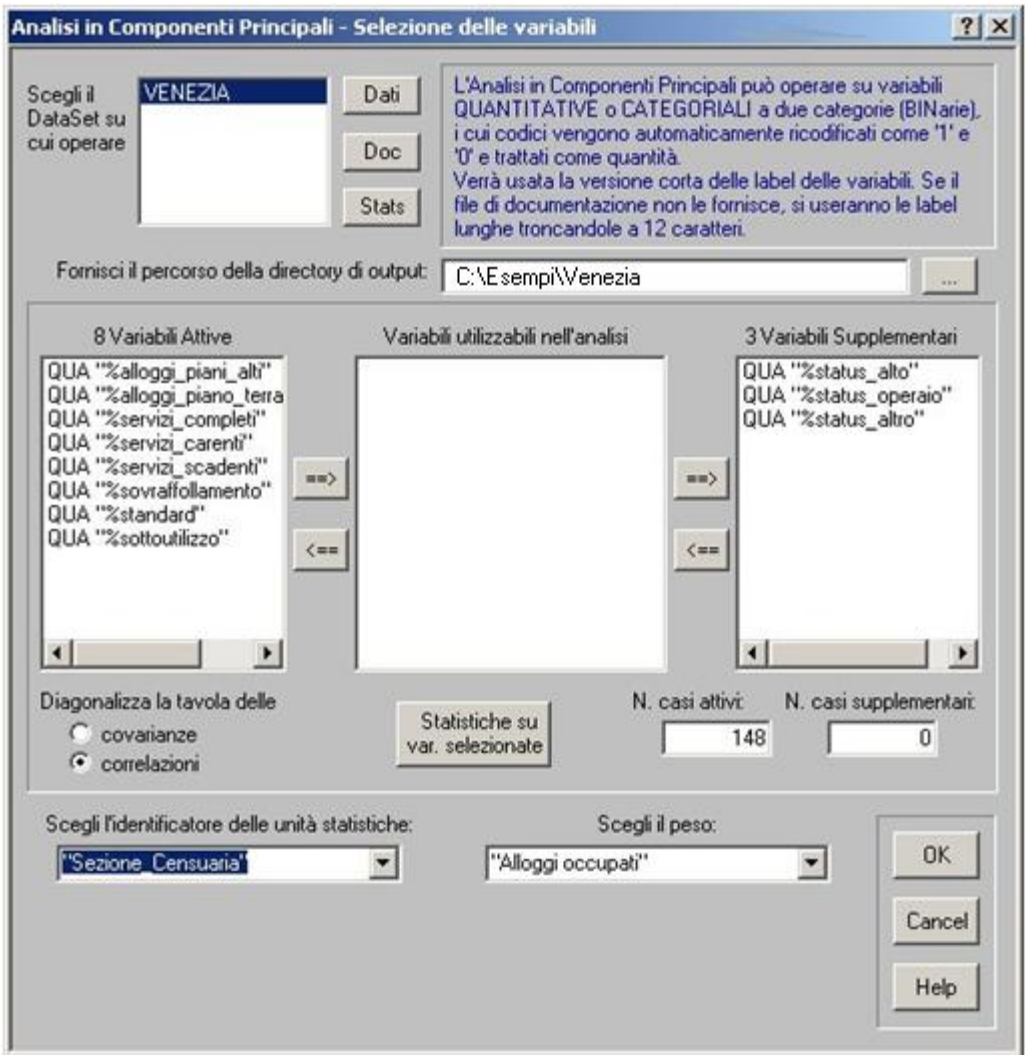


Figura 6.7 – Il dialogo di figura 6.6, riempito opportunamente.

Come si è detto, le variabili **attive** contribuiscono alla costruzione dei fattori, mentre quelle **supplementari** sono usate solo per scopi descrittivi. La varianza totale della tavola, distribuita tra i fattori, proviene solo dalle variabili attive.

In generale, solo parte della varianza delle variabili supplementari è "spiegata" nello spazio fattoriale descritto da quelle attive; in altri termini, una variabile supplementare non può in generale venire espressa esattamente come una combinazione lineare di quelle attive.

Il nostro esempio ha solo tre variabili supplementari, quelle relative allo status del CapoFamiglia.

Riempito il dialogo, si prema OK. Appare una finestra di editing del tipo di quella in figura 6.8.

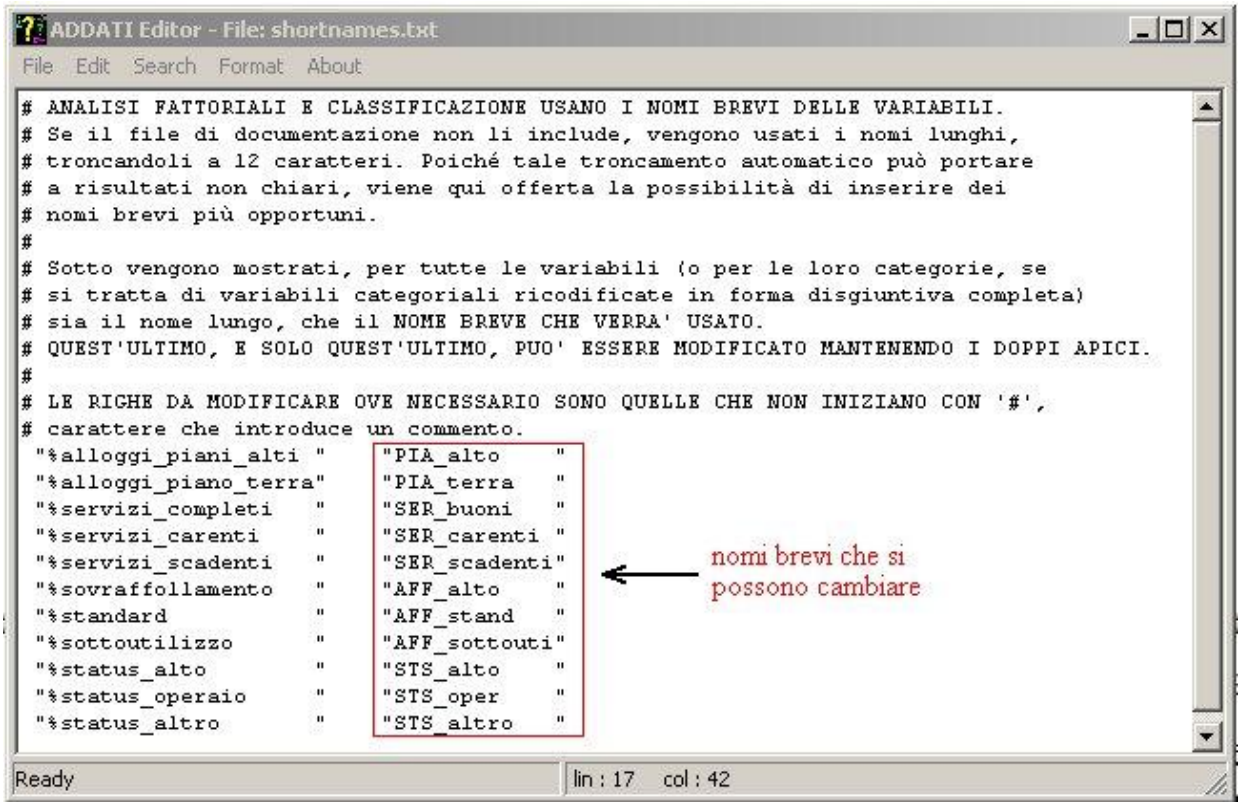


Figura 6.8 – La finestra di editing per confermare o modificare i nomi brevi delle variabili.

La finestra mostra alcune righe di commento, che iniziano con '#', seguite da altre che contengono, tra doppi apici, I **nomi lunghi e brevi** delle variabili da trattare.

Nota Per ogni variabile si può fornire nel file di documentazione del DataSet sia un nome ***lungo*** che uno ***corto***. I nomi corti sono usati dalle analisi multivariate e nella classificazione. Se non vengono forniti nomi brevi, il programma li genera prendendo i **primi dodici caratteri** del nome lungo. Poiché non è sicuro che ne risulti un nome chiaro, i nomi (lunghi e corti, forniti o creati) vengono mostrati nell'editor. Si possono modificare i nomi corti, mantenendo la limitazione di dodici caratteri ed i doppi apici.

In realtà il ruolo della finestra di editing è più complesso e dipende dalla scala di misura delle variabili (se siano QUANTITATIVE o binarie, e si usi una ACOMP; oppure se siano di tipo COUNT o CATEGORIAL, e richiedano un'Analisi delle Corrispondenze). Fino a quando non diventi abbastanza esperto **l'utente dovrebbe leggere attentamente le linee di istruzione**, **che variano con la scala delle variabili**.

Salvare il file nel caso si sia effettuato qualche cambiamento, e chiudere la finestra. I nomi brevi modificati vengono salvati nel file di documentazione: ricordarsi di salvare il DataSet uscendo da ADDAWIN se si vogliono mantenere le modifiche per uso futuro.

A questo punto le variabili vengono standardizzate, viene calcolata la tavola delle correlazioni, scritta nel file d'uscita, convenzionalmente chiamato ACOMPnn.OUT. A partire dalle correlazioni vengono determinati gli assi fattoriali e calcolati gli autovalori ad essi associati.

Appaiono poi simultaneamente due finestre, che offrono all'analista la possibilità di decidere quante Componenti Principali salvare per i diversi scopi.

La prima, riportata nella tabella 6.1, mostra il contenuto del file di uscita ACOMPnn.OUT (almeno, la parte di esso scritta fino a questo punto). La visualizzazione si localizza sulla tabella degli autovalori, che misurano il potere esplicativo delle diverse Componenti Principali (cioè la loro Inerzia). Sulla base della distribuzione dell'inerzia tra le Componenti Principali l'analista riempie l'altra, cioè il dialogo di figura 6.9, decidendo quante CP usare per la Classificazione, per l'interpretazione o per la visualizzazione dei piani fattoriali.

🗁 Il valore dell'inerzia ripartito tra le componenti principali è 8, corrispondente al numero delle colonne attive della tavola. La nuvola conserva un'inerzia di 5.01 (pari al 62.6% del totale) quando venga proiettata sul primo asse fattoriale: la relazione tra le variabili è dunque così forte che la prima Componente Principale basta da sola a sintetizzare il 62% dell'informazione complessivamente apportata dalla tavola! Si noti anche che cinque fattori esauriscono praticamente il 100 per cento dell'inerzia, a riprova del fatto che tra le otto variabili attive esistono delle relazioni analitiche legate al modo in cui sono state definite; quattro fattori sono sufficienti a spiegare il 96.3% dell'inerzia totale. Limitarsi dunque a quattro fattori nella successiva procedura di classificazione sembra costituire una semplificazione del tutto accettabile.

```
VIENE DIAGONALIZZATA LA TAVOLA DELLE CORRELAZIONI
DETERMINATI 6 FATTORI SIGNIFICATIVI - INERZIA SPIEGATA:
INERZIA TOTALE  = 8.000000
|    |           | INERZIA|INERZIA |
|  N | AUTOVALORE|SPIEGATA|CUMULATA|
|    |           |   (%)  |  (%)   |
|----|-----------|-----------------|
|  1 | 5.0113457 | 62.642 | 62.642 |******************************************
|  2 | 1.1802435 | 14.753 | 77.395 |**********
|  3 | 0.9061948 | 11.327 | 88.722 |********
|  4 | 0.6101296 |  7.627 | 96.349 |*****
|  5 | 0.2914368 |  3.643 | 99.992 |***
|  6 | 0.0006495 |  0.008 |100.000 |
```

Tabella 6.1 - Gli autovalori associati alle Componenti Principali sono una misura del loro potere esplicativo.

Giusto prima della tavola degli autovalori il file di uscita riporta la matrice delle correlazioni tra le variabili, riportata nella tabella 6.2

CORRELAZIONI (*1000)

	PIA_alto	PIA_terra	SER_buoni	SER_carenti	SER_scadenti	AFF_alto	AFF_stand	AFF_sottouti	STS_alto	STS_oper	STS_altro
PIA_alto	1000										
PIA_terra	-1000	1000									
SER_buoni	599	-599	1000								
SER_carenti	-502	502	-863	1000							
SER_scadenti	-495	495	-796	382	1000						
AFF_alto	-634	634	-795	641	691	1000					
AFF_stand	-81	81	-116	145	35	-206	1000				
AFF_sottouti	643	-643	817	-688	-670	-818	-394	1000			
STS_alto	578	-578	881	-734	-731	-745	-265	857	1000		
STS_oper	-651	651	-835	700	690	825	82	-824	-852	1000	
STS_altro	68	-68	-171	136	151	-60	346	-145	-365	-174	1000

Tabella 6.2 – La matrice delle correlazioni come appare nel file salvato da ACOMP.

La correlazione più elevata è quella tra *p_alto* e *p_terra*, pari a -1 (nella tabella i valori delle correlazioni sono moltiplicati per mille per comodità di lettura): ciò è dovuto al fatto che i due valori

sono legati da una relazione analitica, come è stato già rilevato. A parte ciò, risultano avere un'alta correlazione positiva le variabili *dotazione completa*, *sotto-utilizzo* e *status elevato*; queste stesse tre variabili risultano poi essere tutte negativamente correlate con *dotazione carente*, *dotazione scadente*, *sovraffollamento* e *status operaio*. Le variabili di quest'ultimo gruppo risultano poi tra loro tutte positivamente correlate.

La variabile *altro status* non ha correlazioni di rilievo con nessuna delle altre, a riprova del fatto che raccoglie status di risulta, non ben caratterizzati dal punto di vista del fenomeno che si sta considerando.

Esistono dunque gruppi di variabili fortemente correlate: ciò significa che la nuvola è marcatamente allungata nello spazio di rappresentazione e che il passaggio ai fattori può effettivamente comportare un risparmio dimensionale nella rappresentazione.

Nota *Non sempre le cose vanno in modo così conveniente e suggestivo: qui succede perché è molto forte la struttura delle relazioni tra le variabili attive, cioè perché sezioni caratterizzate da un'alta quota di "dotazione completa di servizi" presentano* ***simultaneamente*** *un'alta quota di "sottoutilizzo" (e di "status elevato") ed una bassa quota di "affollamento" e di dotazione carente (e di "status operaio"), e viceversa. Quando invece la struttura delle relazioni è molto più debole le variabili iniziali risultano pressoché indipendenti. La nuvola è allora quasi sferica ed un cambio di sistema di riferimento non porta vantaggi significativi: la dimensione del fenomeno non è riducibile senza una perdita significativa d'informazione.*

6.3.3 – Quante Componenti Principali registrare?

L'utente decide quante coordinate fattoriali facciano al suo caso, dopodiché queste vengono calcolate e, su richiesta, possono venire sia stampate per l'interpretazione che registrate su file per una eventuale successiva operazione di classificazione. Il contenuto delle tavole registrate nel file di uscita ACOMP-nn.OUT per la successiva interpretazione è descritto più avanti.

Secondo i valori inseriti nel dialogo di figura 6.9, la procedura di Classificazione automatica che seguirà assumerà come input n Componenti Principali, mentre le prime tre, che riassumono il xxx% dell'Inerzia totale, verranno usate per interpretare le caratteristiche delle Sezioni Censuarie e delle variabili, e per le proiezioni sui piani fattoriali. Sei è il massimo numero di CP utilizzabili per queste ultime operazioni.

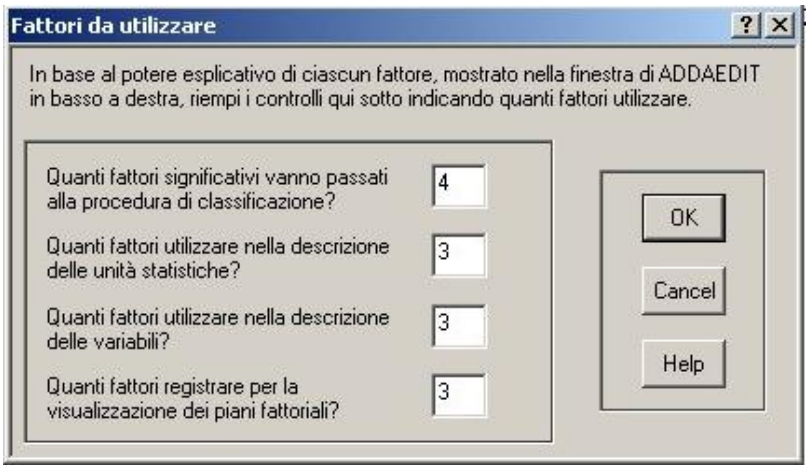


Figura 6.9 – Il dialogo che fissa il numero di CP da utilizzare

In generale:

Componenti Principali da utilizzare nella classificazione delle unità statistiche

È necessario registrare su file le coordinate fattoriali prima di procedere ad una Classificazione non-gerarchica. Nel caso di ACOMP, è in genere conveniente salvare un numero di fattori che riassumano tra l'80 ed il 90% dell'inerzia totale. Invece, quando si operi su variabili categoriali (con l'Analisi delle Corrispondenze) basta di solito una percentuale minore.

🗁 Esempio su Venezia: si è già deciso di usare per la classificazione i primi quattro fattori, che riassumono il 96.3% della varianza totale della tavola.

Descrizione delle unità statistiche

Conviene registrare queste informazioni sul file di uscita solo quando si abbia interesse a controllare il comportamento di (almeno) alcune unità. Meglio non richiederlo (inserendo '0') se le unità sono molto numerose, o se non si è particolarmente interessati al loro comportamento individuale.

Per ogni unità statistica, ed ogni fattore richiesto, vengono registratate le informazioni seguenti:

- la **coordinata fattoriale** dell'unità statistica sull'asse;
- il **contributo relativo** (quota dell'inerzia dell'unità spiegata da quel fattore);
- il **contributo assoluto** (quota dell'inerzia totale del fattore proveniente dall'unità);

Il loro significato verrà illustrato più avanti.

🗁 Poiché le sezioni censuarie sono 148, può non valere la pena di esaminarne il comportamento singolarmente attraverso la tavola dei contributi. Potrebbe tuttavia essere interessante individuare quelle caratterizzate da una prima coordinata fattoriale altamente positiva o negativa (come vedremo, il primo fattore ordina le sezioni secondo il livello del disagio abitativo). Chiederemo dunque la stampa dei contributi sui primi tre fattori.

Descrizione delle variabili

Va sempre richiesta almeno per i fattori più rilevanti, in quanto il significato dei fattori, in sede di interpretazione, si desume proprio dall'esame dei contributi delle variabili.

Per ciascuna variabile e ciascun fattore richiesto verranno salvate sul file di output ACOMP-nn.OUT le seguenti informazioni:

- la **coordinata fattoriale** della variabile sull'asse;
- il **contributo relativo** (quota dell'inerzia della variabile spiegata da quel fattore);
- il **contributo assoluto** (quota dell'inerzia totale del fattore proveniente dalla variabile);

🗁 Per Venezia C.S. richiederemo la stampa dei contributi su tre fattori, sufficienti a spiegare l'88.7% della variabilità totale della tavola. Useremo questi contributi per interpretare i fattori.

Proiezioni sui piani fattoriali

Possono facilitare l'interpretazione quando il piano considerato spiega un quota elevata della varianza globale. Ci si deve comunque limitare a considerare solo i punti **ben rappresentati**, poiché gli altri potrebbero risultare fuorvianti.
FACPLAN offre un'opzione specifica per selezionare i punti ben rappresentati.
E' opportuno insistere che gli spunti interpretativi offerti dai piani fattoriali **vanno sempre verificati sulle tavole dei contributi assoluti e relativi**.

🗁 Esempio su Venezia: la registrazione delle prime tre coordinate fattoriali sembra sufficiente per la visualizzazione dei piani fattoriali più interessanti.

Nota *E' possibile visualizzare sullo schermo le proiezioni della nuvola su uno o più piani fattoriali, mostrando qualunque combinazione si desideri di oggetti e variabili, sia attivi che supplementari. Si può ingrandire una parte del piano a piacere, editare graficamente l'immagine, salvarla su file e stamparla (si veda la spiegazione relativa a FACPLAN).*

6.3.4 –Lettura delle tavole dei contributi

Le informazioni elencate qui sotto vengono registrate nel file ACOMP-nn.OUT su richiesta dell'utente, separatamente per le variabili e per le unità attive e supplementari.

La tabella 6.3 mostra i contributi delle variabili, tratti da ACOMP-nn.OUT. Il significato è descritto in dettaglio nella tabella 6.4 per la variabile *p_alto* e per i primi due fattori.

QLT (qualità della rappresentazione): è la quota dell'inerzia della variabile spiegata globalmente da **tutti** i fattori dei quali è stata richiesta la registrazione (tre nel nostro caso). Rappresenta la somma dei contributi alla variabile da parte dei fattori in questione. Qui i primi tre fattori spiegano i 996/1000 della varianza della variabile *PIA_alto*.

INR (inerzia totale della variabile): poiché tutte le variabile sono standardizzate, esse contribuiscono in pari misura all'inerzia della nuvola, che vale esattamente 8 (ci sono otto variabili attive, tutte con varianza unitaria). INR è qui espresso come una frazione dell'inerzia totale (125/1000, corrispondente a 1/8).
Invece, nel caso di un'Analisi delle Corrispondenze le variabili - cioè le colonne della tavola - hanno in generale valori diversi di INR, che dipendono dal peso del punto e dalla sua distanza dal centro della nuvola.

n.	VAR ATT	QLT	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS	FAT 3	CON REL	CON ASS
1	PIA_alto	996	1	125	1000	-822	675	135	63	4	3	563	317	350
2	PIA_terra	996	1	125	1000	822	675	135	-63	4	3	-563	317	350
3	SER_buoni	956	1	125	1000	-927	859	171	-13	0	0	-311	97	107
4	SER_carenti	687	1	125	1000	779	606	121	114	13	11	261	68	75
5	SER_scadenti	664	1	125	1000	765	585	117	-117	14	12	256	65	72
6	AFF_alto	903	1	125	1000	877	770	154	-328	108	91	159	25	28
7	AFF_stand	988	1	125	1000	139	19	4	983	966	818	-56	3	3
8	AFF_sottouti	907	1	125	1000	-906	821	164	-269	72	61	-117	14	15

n.	VAR SUP	QLT	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS	FAT 3	CON REL	CON ASS
9	STS_alto	834	1	125	1000	-871	758	0	-156	24	0	-227	52	0
10	STS_oper	786	1	125	1000	878	771	0	-34	1	0	120	14	0
11	STS_altro	171	1	125	1000	79	6	0	345	119	0	213	45	0

Tabella 6.3 - I contributi delle variabili sui primi tre fattori.

	n.	VAR ATT	QLT	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS
num. d'ordine →	1	PIA_alto	996	1	125	1000	-822	675	135	63	4	3
della variabile		↑					↑	↑	↑	↑	↑	↑
		indicatore alfanumerico					informazioni sul fattore n. 1			informazioni sul fattore n. 2		

Tabella 6.4 - I contributi della variabile *p_alto* sui primi due fattori.

PESO Importanza della variabile nell'analisi. Poiché le variabili sono standardizzate PESO ha convenzionalmente lo stesso valore per tutte.

DIS è il quadrato della distanza del punto-variabile dall'origine (il valore nelle stampe è moltiplicato per 1000). Essa rappresenta la varianza della variabile, che qui vale 1 poiché tutte le variabili sono standardizzate: tutti i punti-variabile giacciono dunque sulla superficie di una iper-sfera di raggio unitario centrata sull'origine.

FAT1 è la coordinata del punto-variabile sul primo asse fattoriale (qui vale -0.822). Poiché la distanza di ogni punto-variabile dall'origine è esattamente 1, FAT1 risulta uguale al coseno dell'angolo formato dal segmento che congiunge il punto all'origine con il primo asse fattoriale (si veda la figura 6.3). Si può provare che la coordinata fattoriale misura la correlazione tra la variabile ed il primo fattore (considerato come quella nuova variabile, costruita come combinazione lineare delle variabili originali, che presenta la varianza massima).

CON REL **contributo relativo** (del fattore alla variabile): è la frazione (∗1000) dell'inerzia della variabile spiegata dal fattore. Qui il primo fattore spiega il 67.5% della varianza della variabile *PIA_alto* sull'insieme delle sezioni censuarie.

Si può dimostrare facilmente che per un'Analisi in Componenti Principali il contributo relativo è il quadrato della coordinata fattoriale corrispondente (FAT1); esso è dunque pari al quadrato della correlazione tra variabile e componente principale.

CON ASS **contributo assoluto** (della variabile alla varianza del fattore): è la quota (∗1000) della varianza del fattore che proviene dalla variabile. Qui il 13.5% della varianza del primo fattore proviene dalla variabile *PIA_alto*.

La tabella 6.5 mostra le informazioni sulle unità come appaiono in ACOMPnnn.TXT, mentre la tabella 6.6 mostra nei dettagli il significato per la sezione censuaria n.5 ed i primi due fattori.

n.	OGG ATT	QLT	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS	FAT 3	CON REL	CON ASS
1	1	992	5	6	8842	-2752	857	8	394	18	1	1018	117	6
2	2	977	4	4	6453	-1682	438	2	603	56	1	1763	482	15
3	3	959	5	4	6403	-2075	673	4	69	1	0	1353	286	9
4	4	949	4	6	11337	-3074	834	8	-1113	109	5	274	7	0
5	5	986	8	13	13101	-3305	834	17	-1317	132	12	-503	19	2
		…….												
46	46	958	9	18	15900	3810	913	26	225	3	0	813	42	6
47	47	977	6	16	21361	4349	886	23	-1076	54	6	890	37	5

Tabella 6.5 - I contributi delle singole unità sui primi tre fattori.

	n.	OGG ATT	QLT	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS
num. d'ordine →	5	5	986	8	13	13101	-3305	834	17	-1317	132	12
dell'unità		↑					↑	↑	↑	↑	↑	↑
		indicatore alfanumerico					informazioni sul fattore n. 1			informazioni sul fattore n. 2		

Tabella 6.6 - I contributi sui primi due fattori dell'unità numero 5.

Il significato è il seguente:

QLT (**qualità della rappresentazione**): è la quota dell'inerzia dell'oggetto spiegata globalmente da **tutti** i fattori dei quali è stata richiesta la stampa (tre fattori nel caso del nostro esempio). Rappresenta la somma dei contributi (relativi) dei fattori in questione all'oggetto considerato. La tabella mostra che i primi tre fattori spiegano 986/1000 dell'inerzia della sezione n.5.

INR (**inerzia totale dell'unità**): è la frazione (∗1000) dell'inerzia totale proveniente dal punto-unità:

INR = (inerzia dell'unità) / (inerzia totale)

dove l'inerzia totale è la somma degli autovalori. L'inerzia dell'unità rispetto all'origine (coincidente con l'origine degli assi) è definita come il prodotto della sua massa (PESO) per il quadrato della sua distanza dall'origine (DIS).

PESO **Peso dell'unità** nell'analisi: rappresenta l'importanza relativa dell'unità (i pesi sono scalati proporzionalmente così che la loro somma valga 1000).

Qui la sezione n.5 include l'8 per mille di tutti gli alloggi del Centro Storico di Venezia, mentre la sua inerzia è i 13/1000 di quella totale.

In generale, quanto più INR è grande rispetto a PESO tanto più peculiare è il comportamento dell'unità. Infatti, se tutti i punti fossero collocati alla medesima distanza dall'origine la loro inerzia sarebbe esattamente proporzionale al loro peso. Poiché qui sia INR che PESO sono scalati a 1000, essi sarebbero esattamente uguali.

Un valore di INR maggiore di PESO (come nel caso della sezione n.5) significa che il punto ha una distanza dall'origine maggiore della distanza media e dunque che il suo comportamento è piuttosto particolare (va ricordato che l'origine rappresenta il comportamento medio del sistema). E' necessario un attento esame per identificare le variabili cui va ascritto tale comportamento, cioè in che senso l'oggetto sia peculiare. Ci si può arrivare dai risultati dell'Analisi in Componenti Principali ma è di gran lunga più semplice considerare il profilo della classe alla quale l'oggetto è assegnato nella successiva classificazione: se si trascura la variabilità interna alla classe si può pensare che tale profilo rappresenti il comportamento di tutte le unità assegnate alla classe.

Va tenuto sempre presente che la distanza tra il punto-oggetto e l'origine è una misura della differenza globale tra il comportamento dell'oggetto ed il comportamento medio dell'intero sistema (cioè dell'insieme delle unità considerate). Un'Analisi in Componenti Principali ed un'Analisi delle Corrispondenze assumono sempre come riferimento tale comportamento medio ed analizzano in qual modo e di quanto le diverse unità differiscano da esso (e dunque anche l'una dall'altra). Nella maggior parte dei casi la cosa ammette un'interpretazione evidente.

DIS è il **quadrato della distanza** del punto-unità dall'origine. Maggiore è DIS più il profilo dell'unità differisce globalmente dal comportamento medio.

FAT1 è la coordinata dell'unità sul primo asse fattoriale (nel caso della sezione censuaria n.5 il valore va letto come -3.305).

CON REL **contributo relativo** (del fattore all'unità): è la frazione (∗1000) dell'inerzia dell'unità spiegata dal fattore. Nel nostro esempio il primo fattore è sufficiente a spiegare l'83.4% dell'inerzia della sezione n.5.

CON ASS **contributo assoluto** (dell'unità alla varianza del fattore): è la frazione (∗1000) dell'inerzia del fattore proveniente dall'unità considerata. Il 17 per mille della varianza del primo fattore proviene dalla sezione n.5.

6.3.5 - Interpretazione dei fattori

Il significato dei fattori, considerati come nuove variabili costruite, va derivato dalla loro correlazione con le variabili di partenza, individuando quali siano le variabili che presentano il maggior contributo assoluto sui vari fattori.

Fattore 1 Consideriamo la tabella 6.3: i contributi (assoluti) più elevati al fattore provengono dalle variabili *piano alto*, *servizi buoni* e *sottoutilizzo* che si proiettano sul semiasse negativo e da *piano terra*, *servizi carenti* o *scadenti* e *sovraffollamento* sul lato positivo dell'asse. Si tratta dell'insieme più cospicuo di relazioni riscontrabile nella matrice delle correlazioni. Si è già notato in precedenza come le variabili del primo gruppo risultino tra loro tutte positivamente correlate, così come quelle del secondo gruppo (sempre tra di loro); tutte le variabili del primo gruppo risultano invece negativamente correlate con quelle del secondo. Ove si ricordi che la coordinata fattoriale di una variabile misura la sua correlazione con il fattore, l'interpretazione del significato del primo fattore è immediata: esso offre una forte semplificazione della rappresentazione ordinando le sezioni censuarie secondo un livello di disagio abitativo che va crescendo con il valore della coordinata fattoriale.

Il primo fattore coglie dunque la ragione di maggior variabilità esistente nel sistema (almeno per quanto si può desumere dalla descrizione offerta dalle variabili utilizzate): l'**opposizione** esistente tra sezioni a basso disagio abitativo (cioè con quote d'uso dei piani alti, servizi interni buoni e sotto-utilizzo **sopra la media** della città e **simultaneamente** quote di piani terra, servizi carenti o scadenti e sovraffollamento **sotto la media**) e sezioni caratterizzate dal comportamento opposto.

Un rapido esame della tabella 6.5 evidenzia alcune sezioni coinvolte in questa opposizione: basta individuare quelle che ricevono contributi elevati (CON REL) dal primo fattore. Ad esempio, le sezioni n. 1, 4 e 5, con alto contributo relativo sul primo fattore e coordinata negativa (e dunque a basso disagio) da un lato, le sezioni 46 e 47, anch'esse caratterizzate da elevato contributo relativo ma con coordinata fattoriale positiva (e dunque ad elevato disagio) dall'altro.

Si noti che anche la variabile supplementare *status alto* è negativamente correlata con il fattore (e quindi con le condizioni di basso disagio), mentre lo *status operaio* si proietta sul semiasse positivo e quindi si lega alle condizione di disagio elevato.

La variabile *standard* ha una correlazione molto bassa con il primo fattore. Ciò significa che gli alloggi in condizione standard di affollamento si trovano equamente distribuiti attorno a condizioni di disagio medie.

Fattore 2 La tabella 6.3 mostra che ben l'81.8 per cento dell'inerzia del secondo fattore proviene dalla variabile *affollamento standard*, la cui varianza è a sua volta spiegata - per ben il 96.6% - dal fattore. Praticamente, il secondo fattore quasi si identifica con la distribuzione dell'*affollamento standard* sulle sezioni (la tabella 6.1 mostra che l'autovalore associato al secondo fattore vale 1.18, poco più del contributo offerto dalla variabile in questione). La ragione sta nel fatto che - come si può osservare dalla tabella 6.2 delle correlazioni - l'*affollamento standard* risulta pochissimo correlato con tutte le altre variabili e necessita quindi di un fattore esplicativo per conto suo; appunto, il secondo.

L'interpretazione potrebbe continuare per gli altri fattori, che diventano tuttavia sempre meno rilevanti.

La figura 6.10, ottenuta mediante il programma di utilità **FACPLAN**, (***Utilità→Mostra Piani Fattoriali***) presenta la proiezione delle variabili sul piano individuato dai primi due fattori, dove le considerazioni sin qui fatte sono puntualmente verificabili. Si noti che le scale dei due assi vengono fissate da **FACPLAN** *in modo indipendente*, con l'obiettivo di ottenere una proiezione a tutto schermo. Ne consegue che in figura la varianza del secondo fattore (pari a 1.18) è sovra-rappresentata rispetto a quella del primo fattore, pari a 5.01. La nuvola è cioè nella realtà molto più appiattita lungo il primo asse.

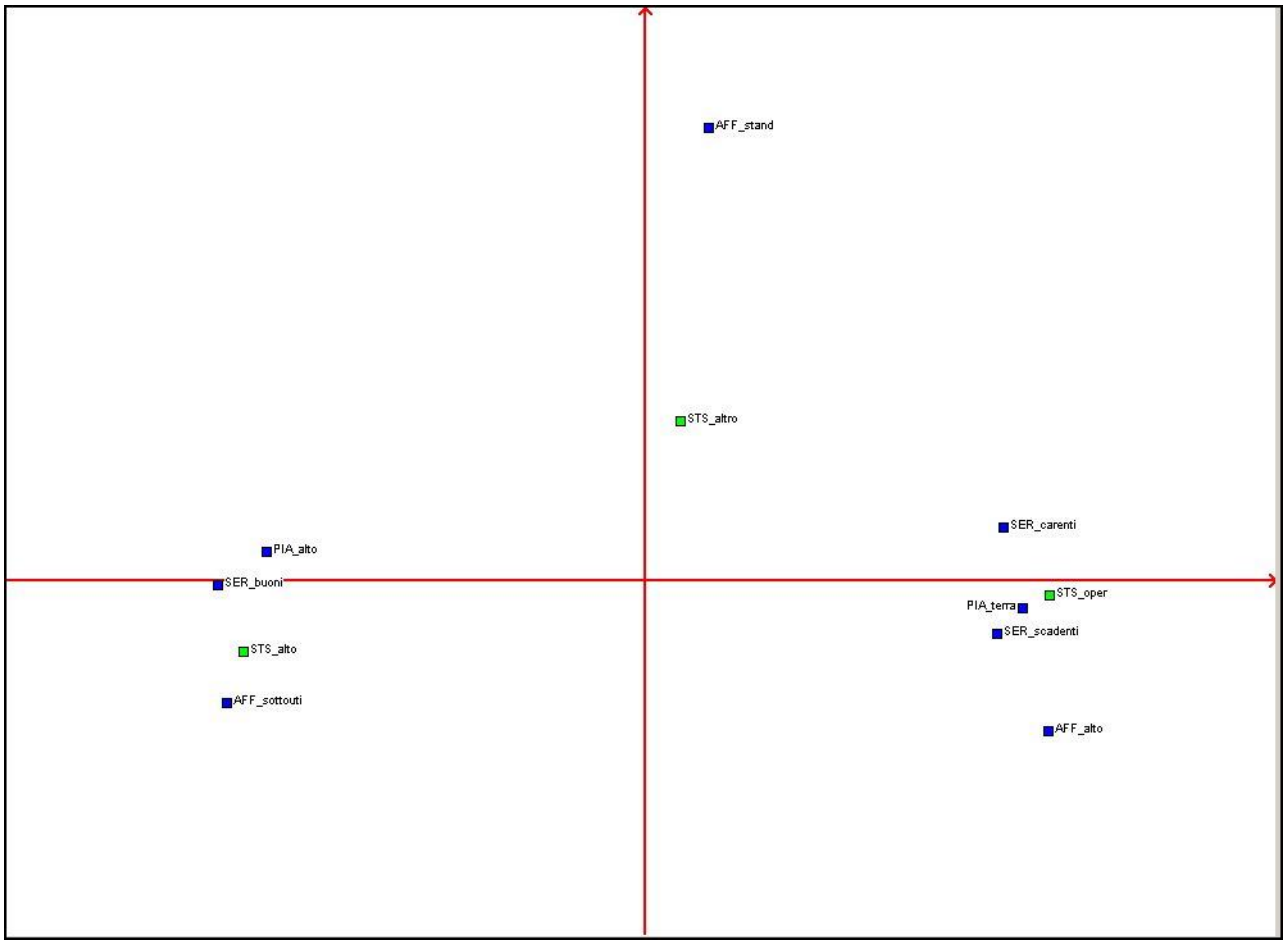


Figura 6.10 - Venezia: proiezione dei punti-variabile sul primo piano fattoriale.

La figura 6.11 proietta sul medesimo piano anche le unità Statistiche (le Sezioni Censuarie): tuttavia, per rendere meglio visibili le cose viene mostrata ingrandita solo la parte centrale della proiezione. Ove si tenga presente la discussione sul significato dei fattori, la posizione di ciascuna Sezione, purché ben rappresentata sul piano, dà un'immediata percezione delle sue caratteristiche in termini di disagio abitativo. Ad esempio, Sezioni come la 123 o la 55 hanno condizioni peggiori della media, mentre la 82 presenta condizioni migliori. Altre Sezioni, dalle caratteristiche più estreme, non sono visibili nella figura.

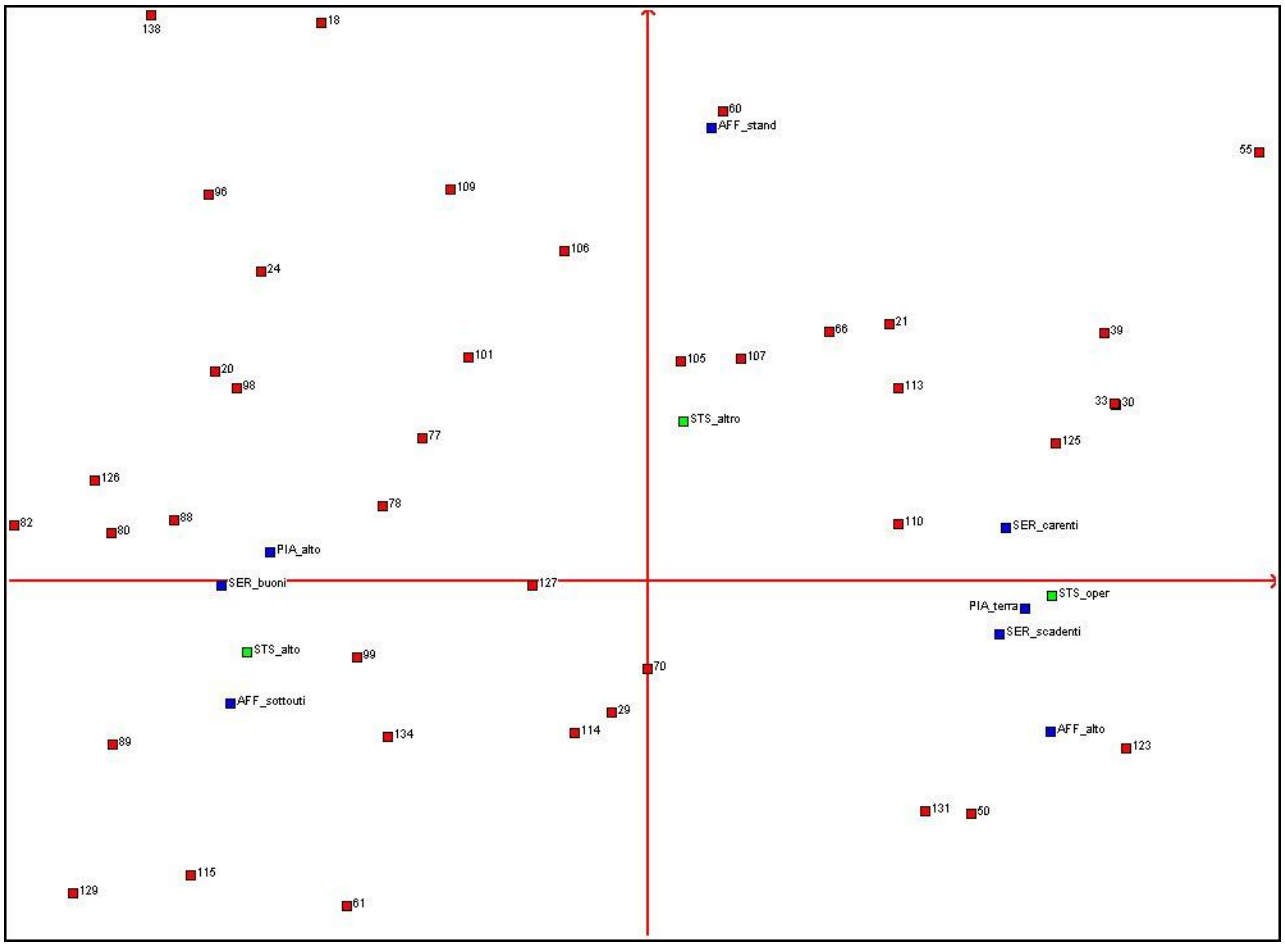


Figura 6.11 - Venezia: proiezione sovrapposta di variabili e Sezioni sul primo piano fattoriale.

I files registrati scritti da ACOMP

- **ACOMPnnn.TXT** è il file di testo che contiene le informazioni su cui si basa l'interpretazione. '**nnn**' è un numero progressivo volto ad impedire la sovrascrittura di file già esistenti
- **Label.PCS** (scritto su richiesta dell'utente) è un file binario che contiene le coordinate fattoriali delle unità ed altre informazioni necessarie a NONGER per operare la classificazione.
- **Label.TMP** (che accompagna il file .PCS) è un file binario che registra i valori originali delle variabili analizzate. È usato dal programma di Classificazione per calcolare i valori medi delle variabili nelle diverse classi.
- **Label.FPL** (scritto anch'esso su richiesta) è il file di testo che contiene le informazioni passate a **FACPLAN** per visualizzare le proiezioni sui piani fattoriali.

Nei nomi dei file elencati qui sopra '**Label**' sta per il nome del DataSet analizzato.

6.4 - L'Analisi delle Corrispondenze (ACORR)

Funzione, Limiti e Consigli

Vale esattamente quanto detto per **ACOMP**.

La tavola standard analizzata mediante un'Analisi delle Corrispondenze è una **tavola di contingenza,** ottenuta incrociando due variabili categoriali.

Se le due variabili incrociate hanno rispettivamente *n* e *p* categorie la tavola di contingenza ottenuta viene ad avere *n* righe e *p* colonne; la generica cella (i,j) conta le unità che prendono *simultaneamente* la categoria i della prima variabile **e** la categoria j della seconda.

Altre tavole ***solo apparentemente di struttura diversa*** possono essere pensate come tavole di contingenza e si possono trattare con un'Analisi delle Corrispondenze:

- tavole ottenute accostando fianco a fianco più tavole di contingenza che contano le stesse unità;
- tavole binarie ottenute a partire da tavole di descrizione qualitativa convertendo le variabili categoriali di descrizione in forma ***disgiuntiva completa*** (binaria).

Il secondo caso è meno intuitivo, ma risulta molto interessante per le possibili applicazioni allo spoglio di inchieste.

Esempio

*Si pensi ad una tavola le cui n righe rappresentino n unità descritte da p variabili categoriali, alcune delle quali potrebbero risultare da una ricodifica in classi (operata scegliendo dal Menu l'opzione "**Utilità→Crea nuove Variabili/Data Set**") di variabili direttamente osservate al livello quantitativo. Quando le p variabili vengono ricodificate in forma disgiuntiva completa ciascuna di esse dà origine ad una tavola di tipo 0-1 **che ha tante colonne quante sono le categorie della variabile in questione**.*

*Le p tavole binarie che ne risultano possono essere viste come tavole di contingenza, ciascuna delle quali incrocia la variabile "unità statistica" con una delle variabili descrittive. In ciascuna riga (unità) una cella che corrisponde ad una categoria **non assunta** da quell'unità contiene il valore 0 (cioè non include alcuna unità), mentre una cella che corrisponde ad una categoria **assunta** contiene 1 (cioè **conta esattamente una unità**). La struttura è piuttosto banale, ma si può certamente vederla come una tavola di contingenza multipla. Poiché le righe rappresentano un insieme di unità (persone, famiglie, Comuni, Sezioni Censuarie, ecc.) la struttura delle loro somiglianze può essere analizzata per mezzo di **ACORR**, seguita in caso da una classificazione.*

In una tavola di contingenza **righe e colonne hanno un ruolo simile** e sono trattate allo stesso modo da **ACORR**. Lo scopo del metodo è di analizzare la somiglianza tra le righe (rispetto alle colonne), quella tra le colonne (rispetto alle righe) e le relazioni che intercorrono tra righe e colonne.

Vista la simmetria della tavola, la trattazione analitica può focalizzarsi sia sulle righe che sulle colonne. La tabella 6.7 mostra un piccolo esempio didattico: un sistema consiste di tre unità geografiche, ciascuna descritta dall'area della superficie dedicata a certi tipi di coltivazione (sono considerati esplicitamente solo tre tipi, mentre il quarto conta tutte le coltivazioni residue). **ACORR** converte la tavola iniziale 6.7a) nella tavola dei **profili di riga** 6-7b) od in quella dei **profili di colonna** 6-7c). È sufficiente analizzare solo una di esse (il programma determina automaticamente

quale sia la più conveniente dal punto di vista del calcolo); ed i risultati relativi all'altra vengono poi ricavati mediante semplici trasformazioni.

	teff	mais	sorgo	altro		
area 1	40	60	50	100	250	
area 2	100	100	200	200	600	**a)**
area 3	80	120	100	200	500	
	220	280	350	500	1350	

	teff	mais	sorgo	altro		
area 1	.16	.24	.20	.40	250	
area 2	.17	.17	.33	.33	600	**b)**
area 3	.16	.24	.20	.40	500	
	.16	.21	.26	.37	1350	

	teff	mais	sorgo	altro		
area 1	.18	.21	.14	.20	.19	
area 2	.45	.36	.57	.40	.44	**c)**
area 3	.36	.43	.29	.40	.37	
	220	280	350	500	1350	

Tabella 6.7

a) Un esempio di tavola di contingenza: ogni cella conta l'area (in migliaia di ettari) per unità amministrativa e tipo di coltivazione.
Vengono anche mostrati i totali di riga e colonna (detti ***marginali*** della tavola).

b) I ***profili di riga*** calcolati a partire dalla tavola a). Ogni riga mostra, per ciascuna unità amministrativa, come l'area coltivata si ripartisca percentualmente tra i diversi tipi di coltivazione. L'ultima riga fornisce la medesima informazione riferita però all'intero sistema e rappresenta l'incidenza relativa globale delle diverse colture. La (eventuale) specializzazione relativa di un'area viene evidenziata confrontando il suo profilo con quello globale.
In **ACORR** il peso di ciascuna unità è proporzionale al suo marginale di riga (cioè all'area totale coltivata in quell'unità) ed è calcolato dal programma stesso in base ai dati.

c) I ***profili di colonna*** calcolati a partire dalla tavola a). Ogni colonna mostra come il corrispondente tipo di coltura si ripartisca percentualmente tra le aree geografiche. L'ultima colonna (marginale) dà la stessa informazione riferita all'intero sistema e rappresenta la distribuzione globale dei tipi di coltivazione sulle unità geografiche. La concentrazione relativa di un tipo di coltura viene determinata confrontando il suo profilo distributivo sulle aree con quello globale.
In **ACORR** il peso di ciascuna colonna è proporzionale al suo valore marginale (cioè alla superficie totale per quel tipo di coltivazione) ed è calcolato dal programma stesso.

Secondo la tavola 6.7b) ogni unità è rappresentata da un punto in uno spazio a 4 dimensioni (ci sono 4 variabili descrittive), le cui coordinate sono le componenti del profilo; al punto è assegnata una

massa proporzionale alla superficie coltivata in ciascuna unità geografica (è il marginale di riga, cioè la somma degli elementi di riga). In tutto, ci sono dunque tre punti-profilo (dotati di massa) in uno spazio a 4 dimensioni.

Simmetricamente, secondo la tabella 6.7c) ciascun tipo di coltivazione è rappresentato da un punto in uno spazio a tre dimensioni (ci sono infatti tre unità geografiche), le cui coordinate sono le componenti della corrispondente colonna-profilo. Al punto viene assegnata una massa proporzionale alla superficie complessivamente dedicata a quella coltivazione. In questo caso, ci sono quattro punti-profilo dotati di massa in uno spazio tridimensionale.

Si consideri la tabella 6.7b). I profili della prima e della terza unità sono identici: la ragione sta nel fatto che le due corrispondenti righe della tavola 6.7a) sono proporzionali. Le due unità presentano una diversa superficie coltivata (e dunque hanno una diversa rilevanza nell'analisi) ma **un'identica distribuzione percentuale** di tale superficie sui diversi tipi di coltivazione: i due punti-unità vengono a coincidere nello spazio di rappresentazione.

Se tutte le righe della tabella 6-7a) fossero proporzionali, tutti i punti-unità risulterebbero coincidenti: la nuvola collasserebbe in un punto e non vi sarebbero differenze di comportamento da analizzare. Invece, quando le unità hanno comportamenti differenti i loro punti rappresentativi sono dispersi attorno al centro della nuvola, che rappresenta il comportamento medio dell'intero sistema (vale a dire la combinazione percentuale media di coltivazioni, data dalla riga marginale della tavola 6.7b).

Considerazioni simili si possono fare sulla tavola 6.7c).

La distanza tra due punti-profilo in R^p viene calcolata secondo una modificazione dell'usuale formula pitagorica nota come ***distanza del chi-quadro***. Per la sua definizione si rinvia ad un testo di analisi statistica multivariata.

ACORR tratta la tavola dei profili in un modo molto simile a quello già spiegato per **ACOMP** Vengono determinati gli *assi fattoriali* ed i corrispondenti *autovalori*, sui quali si basa l'interpretazione. Vanno comunque tenute ben presenti le seguenti differenze:

- in **ACORR**, diversamente da quanto succede per **ACOMP**, le righe e le colonne giocano un ruolo *totalmente simmetrico*; le tavole dei contributi di riga e di colonna, scritti su ACORR.OUT, vengono interpretate esattamente allo stesso modo. Poiché **ACORR** non standardizza le colonne, non viene stampata alcuna tavola di correlazione e si preferisce parlare di *forte o debole associazione* tra due date linee (righe o colonne) rispetto alla totalità delle linee dell'altro insieme.
- in **ACORR il primo autovalore** (detto *triviale* o *banale*) **vale sempre 1**. Esso non riveste interesse alcuno poiché è una semplice conseguenza della trasformazione compiuta sulla tavola di partenza per passare ai profili; viene dunque ignorato. **Tutti gli altri autovalori (significativi) sono compresi tra 1 e 0**.
- Il numero degli autovalori non nulli è in generale diverso da quello che ci si aspetterebbe dalle dimensioni della tavola dei dati. La dimensionalità della rappresentazione, determinata dal numero delle colonne, è solo apparente (e ridondante): infatti, in ciascuna riga di ogni singola tavola di contingenza i valori delle celle assommano allo stesso totale, cioè al numero delle unità contate. Ne consegue che le colonne non sono linearmente indipendenti, e ciò riduce l'effettva dimensionaltà dello spazio di rappresentazione.
- L'Inerzia totale della nuova dipende in modo molto semplice dal numero delle variabili categoriali e da quello delle loro categorie:

$$Inerzia = \frac{n.\,di\ categorie}{n.\,di\ \text{var}iabili} - 1$$

Ad esempio, se la tavola da trattare consiste di cinque tavole di contingenza affiancate (cioè c'erano in origine cinque variabili categoriali) con 4, 4, 3, 4, 4 categorie rispettivamente (19 categorie complessive) l'Inerzia totale della nuvola risulta essere 2.8 = (19/5) –1.

- Se la tavola dei dati originale consisteva di variabili categoriali, ricodificate automaticamente in forma binaria da **ACORR**, la capacità esplicativa dei primi fattori appare generalmente più bassa che nel caso di **ACOMP** (o anche rispetto ad un'Analisi delle Corrispondenze su tavole di contingenza affiancate). È questo un effetto della ricodifica in forma disgiuntiva completa, che aumenta il numero delle colonne della tavola passata ad **ACORR**, introducendo dell'inerzia fittizia. Anche se la frazione d'Inerzia assegnata ai primi fattori sembra bassa, la loro importanza nell'interpretazione dei risultati è ugualmente rilevante.

6.4.1 - L'inserimento dei parametri di controllo dell'analisi

ACORR ed **ACOMP** necessitano all'incirca degli stessi parametri per controllare l'esecuzione. Vanno dunque riempiti dialoghi simili. Ci limiteremo qui ad illustrare solo le poche domande specifiche di **ACORR,** rinviando per le altre l'utente alla descrizione completa già fornita per **ACOMP**.

Si ricordi che un'**Analisi delle Corrispondenze** assume abitualmente come input una tavola formata da tavole di contingenza affiancate che contano le medesime unità elementari (famiglie, alloggi, individui, ecc.).

Tavole di descrizione qualitativa

Operando su dati urbani o regionali (specialmente dati derivanti dal Censimento o da un'inchiesta), è frequente incontrare tabelle nelle quali ogni riga descrive un'unità statistica elementare (ad un'impresa, o un nucleo familiare) mediante un insieme di variabili ***qualitative*** (o ***categoriali***). In questo caso **ACORR converte automaticamente le variabili qualitative** in forma **binaria (**o **disgiuntiva completa**): una colonna per ogni categoria, con valore 1 se l'unità in questione assume quella categoria, 0 altrimenti. La tavola binaria che si ottiene si può sottoporre ad un'Analisi delle Corrispondenze, nota in tal caso come **Analisi delle Corrispondenze Multiple**.

ACORR, come **ACOMP**, usa nomi brevi per variabili e categorie: viene presentata all'utente una pagina di editing, che propone nomi fino a 12 caratteri per le categorie. **Si leggano attentamente le istruzioni** sul modo in cui vanno modificate, perché tali istruzioni variano a seconda dei casi.

Allo scopo di facilitare l'interpretazione delle proiezioni sui piani fattoriali conviene usare **un prefisso comune** per le categorie di una stessa variabile qualitativa. Ad esempio: 'COM_1', 'COM_2', 'COM_3', 'COM_4' e 'COM_5+' per il numero dei componenti la famiglia; 'STA_1-2', 'STA_3', 'STA_4', 'STA_5' e 'STA6+' per il numero delle stanze; 'TIT_sup', 'TIT_medio', 'TIT_elem', 'TIT_alfab' e 'TIT_analf' per il livello di istruzione raggiunto; 'TV_si' e 'TV_no' per la presenza di un apparecchio TV; e così via. Succede spesso che le categorie di variabili diverse usino etichette simili, e questo modo di assegnarle evita confusioni.

Tavole di contingenza affiancate

Le tavole di dati che decrivono unità amministrative consistono spesso di variabili di tipo COUNT: per ogni unità amministrativa, si contano le frequenze di qualche tipo di unità elementare

sottogiacente (famiglie, individui, edifici...) nelle categorie delle variabili categoriali che descrivono gli aspetti d'interesse. Ad esempio, in ciascuna unità amministrativa:

- si può suddividere la popolazione in un certo numero di classi d'età;
- si possono contare gli abitanti per titolo di studio conseguito;
- si possono contare gli alloggi a seconda del livello di qualche tipo di servizio interno...

Ne risulta un certo numero di tavole di conteggio affiancate, ciascuna delle quali descrive il modo in cui le unità elementari si suddividono rispetto ad un particolare aspetto descrittivo (età, titolo di studio, ecc.)

Per poter calcolare correttamente i profili delle classi di una partizione bisogna che ADDAWIN conosca con esattezza di quante tavole di conteggio affiancate è costituita la tavola dei dati, e quali esse siano.

L'utente deve fornire quest'informazione separando gli i blocchi di variabili che fanno parte di una medesima tavola di contingenza. Ciò va fatto inserendo tra i blocchi, nella stessa pagina in cui si editano i nomi brevi delle variabili, una linea che contenga soltanto un asterisco '*'.

Si leggano dunque con attenzione le istruzioni che introducono la pagina di editing.

6.4.2 - La tabella dei contributi e la loro interpretazione

Le informazioni descritte qui sotto sono registrate (su richiesta dell'utente) nel file di output ACORR-nn.OUT, separatamente per righe e colonne attive e supplementari. L'interpretazione dei risultati si conduce allo stesso modo per righe e colonne, data la simmetria del loro ruolo.

QLT (**qualità della rappresentazione**): percentuale dell'inerzia del punto spiegata complessivamente dai fattori richiesti.

INR (**inerzia totale del punto**): contributo percentuale (∗1000) del punto all'inerzia globale della nuvola :
INR = (inerzia del punto) / (inerzia totale) dove l'inerzia totale è la somma degli autovalori. L'inerzia del punto rispetto all'origine (coincidente con il centro della nuvola) è il prodotto della massa del punto (PESO) per il quadrato della sua distanza (del chi-2) dall'origine (DIS).

PESO **Peso** del punto nell'analisi (normalizzato in modo che il peso totale di tutti i punti di un insieme – righe o colonne – assommi a 1000). Rappresenta l'importanza relativa del punto.

In generale, maggiore è INR rispetto a WEIG, più caratteristico è il comportamento del punto. Infatti, se tutti i punti si localizzassero alla medesima distanza dall'origine, la loro inerzia risulterebbe proporzionale al loro peso. In questo caso, poiché sia INR che PESO sono normalizzati a 1000, essi avrebbero esattamente lo stesso valore.

Un valore di INR maggiore di PESO significa che il punto ha una distanza dall'origine maggiore della distanza media, e che il suo comportamento è dunque piuttosto peculiare (si ricordi che l'origine rappresenta il comportamento medio del sistema). È invece necessario un esame più approfondito per identificare quali componenti del profilo siano responsabili del fatto, cioè sotto quali aspetti il punto sia 'diverso'. Va ricordato che la distanza tra un punto e l'origine degli assi è una misura della differenza globale tra il comportamento del punto e quello medio globale del sistema (ad esempio, se ogni riga descrive un distretto amministrativo di un Paese, il centro di gravità della nuvola rappresenta il comportamento medio del Paese).

Un'Analisi delle Corrispondenze assume come riferimento il comportamento medio del sistema complessivo, ed analizza quanto ed in che modo le diverse unità ne differiscano.

DIS è il **quadrato della distanza** del punto dall'origine. Maggiore è DIS, più il profilo del punto diverge dal comportamento medio globale, rappresentato dal centro della nuvola.

FAC1 è la **coordinata** del punto sul primo asse fattoriale.

REL CON **contributo relativo** (del fattore al punto): è la frazione (x 1000) dell'inerzia del punto spiegata dal fattore.

ABS CON **contributo assoluto** (del punto alla varianza del fattore): è la frazione (x 1000) dell'inerzia del fattore proveniente dal punto.

I file registrati da ACORR

ACORR-nn.OUT è il file di uscita, da interpretare.

Label.PCS (scritto su richiesta dell'utente) contiene le coordinate fattoriali passate a NONGER ed utilizzate per la classificazione delle unità.

Label.TMP (accompagna il file .PCS) è un file binario che registra i valori originali delle variabili analizzate, necessari dopo la procedura di Classificazione per calcolare i profili medi delle classi.

Label.FPL (scritto su richiesta dell'utente) registra l'informazione passata a FACPLAN per la visualizzazione delle proiezioni sui piani fattoriali.

Nei nomi dei file **'Label'** sta per il nome del DataSet le cui variabili vengono analizzate.

Cap. 7. – La Classificazione non gerarchica

7.1 - Alcune note sulla classificazione numerica

Lo scopo di una classificazione numerica è di raggruppare unità a comportamento simile in un numero limitato di **gruppi** (chiamati anche **classi** o **clusters**). La *similarità* tra due unità può venire osservata direttamente (ad esempio ponendo domande specifiche nel corso di un'inchiesta) o può venire **definita** e calcolata a partire da un insieme di variabili osservate che offrano una opportuna descrizione degli oggetti analizzati.

Si considerino ad esempio le province di un Paese, descritte dalla serie del loro reddito medio pro capite durante un certo numero di anni. Quali province hanno evoluzione simile? Non c'è una risposta assoluta: i risultati dipendono dal metodo utilizzato e sono almeno in parte soggettivi. Ad esempio, potremmo fare tutti i possibili confronti a coppie tra le province, ordinando poi le coppie secondo un livello decrescente di similarità percepita.

La similarità dipende dalle variabili prese in considerazione e quindi dalla particolare descrizione adottata per gli oggetti dell'analisi: due comuni possono avere popolazioni molto simili dal punto di vista della struttura demografica, ma presentare invece differenze sostanziali per quanto concerne il livello di scolarizzazione o l'occupazione.

Ci sono molti modi possibili per definire il livello di similarità di due oggetti.

Coerentemente con la rappresentazione geometrica adottata in ADDATI, dove ciascuna unità statistica è vista come un punto in uno spazio che ha tante dimensioni quante sono le variabili attive (si veda il paragrafo 6.1), si assumerà per la classificazione la stessa nozione di **distanza** già introdotta per le analisi fattoriali: una **distanza euclidea** (dopo la standardizzazione) per le variabili quantitative (trattate con **ACOMP**), una **distanza del chi-quadro** nel caso di variabili qualitative (trattate con **ACORR**). La distanza è un indicatore complesso, che si forma attraverso i contributi di tutte le variabili. La assumiamo convenzionalmente come un indicatore di dissimilarità e consideriamo due unità più simili tra loro di altre due quando i loro punti rappresentativi giacciono più vicini (nello spazio di rappresentazione) di quelli che rappresentano le altre due unità. Questa sembra una buona assunzione, sulla quale può esservi consenso.

Anche concordando sulla definizione di similarità, permangono alcuni problemi operativi:

- come misurare il livello di ottimalità di una partizione e come confrontare partizioni con lo stesso numero di classi e decidere quale sia la migliore?
- quante classi costruire? Come possiamo essere sicuri che tale numero si accordi con la struttura dell'insieme da classificare?
- quale algoritmo di classificazione conviene adottare?

Si possono identificare due grandi insiemi di metodi di classificazione, quelli **gerarchici** e quelli **non-gerarchici**. Entrambi lavorano in modo iterativo: essi ripetono una sequenza di operazioni prestabilita - che dipende dall'algoritmo scelto - fino a raggiungere una opportuna configurazione

finale. Entrambi presentano vantaggi e svantaggi.

7.1.1 – I metodi gerarchici

I ***metodi gerarchici ascendenti*** (o aggregativi) eseguono iterativamente le seguenti operazioni su un insieme di *n* unità elementari o di gruppi costituiti in precedenza:

- calcolano la similarità di ogni coppia di unità;
- aggregano le due unità più simili, riducendo così il loro numero ad n-1.

All'inizio del processo di aggregazione si hanno tanti gruppi quante sono le unità elementari, ciascuno consistente esattamente di una unità; alla fine, dopo n-1 passi di aggregazione, tutte le unità sono raggruppate in un solo cluster. Una partizione accettabile - con le unità elementari suddivise in un numero di gruppi abbastanza ridotto da garantire una buona sintesi ma che salvi al contempo una quota consistente di informazione - sta a qualche livello intermedio tra questi due estremi.

Il processo viene di solito rappresentato graficamente mediante un **albero di aggregazione**.

Le unità elementari sono mostrate a sinistra, alla base dell'albero (figura 7.1). Man mano che ci si muove verso destra le unità vengono aggregate, ad una distanza proporzionale alla loro dissimilarità. Se si taglia verticalmente l'albero ad un livello intermedio si ottiene una partizione. Più a destra si seziona l'albero, minore è il numero delle classi risultanti, ma anche minore è l'omogeneità interna delle classi ottenute). Va stabilito un criterio di compromesso per decidere come sia più conveniente sezionare l'albero.

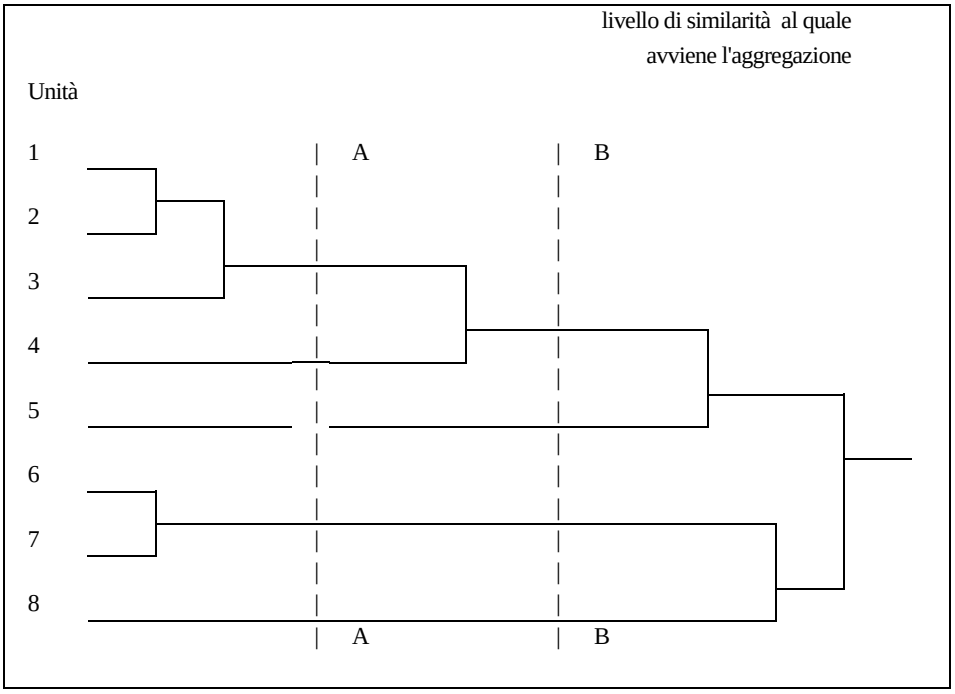


Figura 7.1 - Un albero di aggregazione gerarchica - Vengono aggregate otto unità: il numero dei gruppi diminuisce da sinistra verso destra, mentre la loro dissimilarità aumenta. La sezione AA dà una partizione in 5 classi, la sezione BB ne fornisce invece 4.

Una procedura gerarchica è consigliabile quando il numero degli oggetti non superi alcune decine.

Se sono di più - diciamo, sul centinaio - l'albero diventa difficile da leggere. Oltre a ciò, bisogna pensare che ad ogni passo vanno calcolati ***n(n-1)/2*** valori di similarità, se *n* è il numero dei gruppi esistenti a quel punto: il tempo richiesto cresce con il quadrato di *n*.

Un altro severo svantaggio di questi metodi sta nell'irreversibilità della scelta fatta ad ogni passo: quando due oggetti vengono aggregati non si torna più indietro. Tuttavia, l'algoritmo sceglie la coppia da aggregare solo in base a considerazioni locali, senza alcuna valutazione di tipo globale: lo si può paragonare ad un giocatore di scacchi che scelga sempre la mossa che gli procura il massimo vantaggio **immediato**, senza alcuna considerazione per le mosse successive, vale a dire senza alcuna strategia.

Nel caso dei metodi gerarchici succede spesso che il cammino complessivo di aggregazione potrebbe evolvere in maniera più soddisfacente scegliendo a qualche passo intermedio un'aggregazione diversa da quella localmente ottima. Come conseguenza, una partizione ottenuta tagliando l'albero ad un qualche livello intermedio risulta spesso tutt'altro che ottima.

7.1.2 - I metodi non gerarchici

Viene determinata in qualche modo una partizione iniziale con il numero di classi desiderato; la sua qualità viene poi migliorata mediante opportune riattribuzioni delle unità prossime ai confini tra le classi, quando ciò porti ad un aumento nel valore della **funzione-obiettivo**, che misura la bontà della partizione.

Il processo di riallocazione continua fino a raggiungere una configurazione finale che non è più ulteriormente migliorabile mediante piccoli spostamenti locali.

La partizione che si ottiene costituisce un **ottimo locale**: essa dipende dalla configurazione assunta inizialmente e dal numero delle classi richieste. Non si può escludere che possano esistere altre partizioni anche molto migliori con lo stesso numero di classi: esse non sono tuttavia raggiungibili a partire dalla partizione corrente operando solo riassegnazioni locali.

Alcune definizioni

Quando rappresentiamo un insieme di oggetti come punti in uno spazio multidimensionale assumiamo l'Inerzia (definita nel par. 6.1), alla quale contribuiscono tutte le unità, come misura della variabilità complessiva della tavola dei dati (o del suo contenuto informativo). Agiremo qui in coerenza con tale assunzione.

Sia $In_{tot} = \Sigma_i \, m_i * d_i^2$ l'Inerzia totale della nuvola (rispetto al suo centro), e si consideri una partizione generica della nuvola in *k* gruppi. Ogni unità appartiene ad uno ed un solo gruppo. $\mathbf{G}_j$ denota il centro della *j-esima* classe: le sue coordinate sono i valori medi delle *p* variabili, calcolate tenendo conto delle sole unità appartenenti alla classe.

La classe generica j della partizione ha un'***Inerzia Interna*** definita come

$$\mathbf{In_{int}}\,(j) = \Sigma_{i \in I_j} \, m_i * d^2(i, \mathbf{G}_j)$$

dove la somma è estesa alle sole unità appartenenti alla classe *j* e le distanze sono relative al centro di classe $\mathbf{G}_j$.

L'inerzia interna di una classe misura la dispersione dei suoi elementi attorno al centro di classe. Una buona partizione dovrebbe consistere di gruppi il più possibile omogenei, cioè con una bassa inerzia interna.

L'***Inerzia intraclasse*** (o ***interna***, o ***within-classes***) **di una partizione** è la somma delle inerzie

interne delle sue classi. Il suo valore dovrebbe essere il più basso possibile, e quindi globalmente le classi dovrebbero essere, ciascuna al proprio interno, il più possibile omogenee. I caratteri medi delle unità appartenenti ad una classe j sono rappresentate dalle coordinate del suo centro $\mathbf{G}_j$.

Lo scopo della classificazione è di offrire una rappresentazione semplificata dei fenomeni, nella quale tutte le unità nella stessa classe si identificano con il centro di classe e le differenze tra i loro comportamenti individuali vengono considerate irrilevanti. La nuvola iniziale si riduce così ad una nuova nuvola costituita dai k centri di classe, distribuiti intorno al centro globale. La sua inerzia è l'***Inerzia interclasse*** (o ***esterna***, o ***between classes***) della partizione:

$$\mathbf{In_{ext}} = \sum_j M(j) * d^2(\mathbf{G}_j,\mathbf{G})$$

dove M(j) è la massa della j-esima classe, pari alla somma delle masse di tutte le unità che ad essa afferiscono, e $d^2(\mathbf{G}_j,\mathbf{G})$ è il quadrato della distanza di $\mathbf{G}_j$ da **G**, centro globale della nuvola.

Immaginiamo che la nuvola sia già stata suddivisa in k gruppi, con centri $\mathbf{G}_1$, $\mathbf{G}_2$...$\mathbf{G}_k$ rispettivamente: si prova (teorema di Huyghens) che l'inerzia totale si può decomporre come segue:

$$\mathbf{In_{tot}} = \mathbf{In_{ext}} + \mathbf{In_{int}}$$

Dove $\mathbf{In_{tot}}$ è l'Inerzia Totale della nuvola di punti, $\mathbf{In_{int}}$ è l'Inerzia interna definita in precedenza e $\mathbf{In_{ext}}$ è l'Inerzia esterna.

In ADDATI la funzione-obiettivo per la classificazione non-gerarchica è

$$\mathbf{max\ (In_{ext} / In_{tot})} \qquad \text{che equivale a} \qquad \mathbf{min\ (In_{int} / In_{tot})}$$

e corrisponde ad un insieme di gruppi il più possibile compatti. Il valore della funzione obiettivo varia fra 0 e 1 (a parità di numero delle classi, il valore migliore è quello più alto).

In un'aggregazione gerarchica si hanno inizialmente tanti gruppi quante sono le unità ($In_{ext} = In_{tot}$; $In_{int} = 0$). Quando le unità vengono progressivamente aggregate, ad ogni passo l'inerzia interna aumenta mentre l'inerzia esterna diminuisce di una eguale quantità. Alla fine del processo, quando tutte le unità sono aggregate in una sola classe, si ha $In_{int} = In_{tot}$ e $In_{ext} = 0$. Ad ogni passo vengono aggregate le due unità che comportano il minimo incremento dell'inerzia interna.

Il metodo delle nubi dinamiche

La strategia di classificazione proposta in ADDATI è piuttosto articolata e verrà illustrata per fasi. Un ruolo importante assume in essa il metodo delle nubi dinamiche proposto da E.Diday (1971).

Il metodo di Diday richiede che l'utente decida il numero delle classi da costruire (in via orientativa, pari al numero dei gruppi che si desidererebbe ottenere alla fine del processo) e fornisca un numero equivalente di punti $\{S_1, S_2,...S_k\}$ da assumere come centri iniziali di aggregazione (sono anche noti come ***semi***).

Vengono iterati i due passi mostrati nello schema di figura 7.2. Viene calcolata la distanza di ogni unità da classificare da tutti i k semi e la unità viene assegnata alla classe associata al seme più vicino.

Viene così generata una partizione provvisoria con k classi (ogni unità appartiene ad una ed una sola classe). Vengono poi calcolati i centri delle classi, che assumono il ruolo dei centri iniziali, e la procedura di assegnazione viene ripetuta, ricalcolando poi ancora i centri. Si prosegue così: ad ogni iterazione qualche unità può cambiare di classe, finché si raggiunga una configurazione stabile.

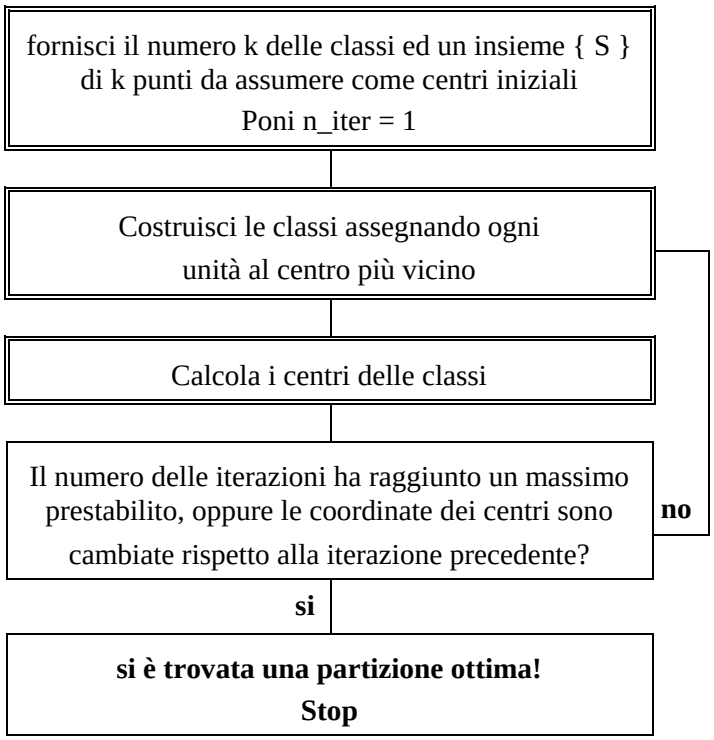


Figura 7.2 - Schema dell'algoritmo di classificazione non gerarchica di Diday.

Si può dimostrare che ad ogni iterazione ***l'Inerzia interna della partizione*** (che misura la dispersione interna delle sue classi) ***non può aumentare***. Di fatto essa diminuisce, oppure viene raggiunto un minimo e la procedura si arresta. Ciò significa che man mano che si procede i gruppi che si ottengono risultano sempre più compatti.

La partizione finale corrisponde ad un minimo dell'Inerzia Interna $\mathbf{In_{int}}$ oppure, per via del teorema di Huyghens, ad un massimo dell'Inerzia Esterna $\mathbf{In_{ext}}$. Si tratta comunque solo di un **ottimo locale**: la partizione non può più venire ulteriormente migliorata con piccoli cambiamenti, ma potrebbe esserlo con una ri-attribuzione più radicale. **Non vi è mai la certezza di aver trovato l'ottimo globale**, vale a dire la migliore partizione in assoluto con quel numero di classi: a causa della dimensione del problema, il raggiungimento di tale sicurezza richiederebbe un tempo di calcolo enorme.

Una volta scelto il numero k delle classi, la partizione finale dipende solo dall'insieme dei semi iniziali $\{S_1,...,S_k\}$, in quanto l'algoritmo è totalmente deterministico.

7.2 - La sequenza di classificazione in ADDATI

7.2.1 – Il metodo di Classificazione non gerarchica

L'input della routine di Classificazione è una tavola di coordinate fattoriali registrate da **ACORR** o **ACOMP** dopo aver analizzato una tavola di variabili quantitative o categoriali, o un insieme di tavole di contingenza affiancate.

Si è visto come l'algoritmo di Diday, che riassegna iterativamente le unità alle classi fino al raggiungimento di una partizione localmente ottima, richieda una decisione a priori sul numero dei

gruppi. Quando tale numero non si attaglia alla struttura di similarità dell'insieme da segmentare viene forzata una partizione che può a volte risultare fuorviante.

Inoltre, la partizione ottenuta - che rappresenta un ottimo locale e non globale - dipende dalla scelta dei centri intorno ai quali le unità sono aggregate secondo un criterio di minima distanza. Si possono concepire varie strategie per la scelta dei centri, ma in generale cambiando i centri iniziali cambia il risultato.

Allo scopo di superare almeno parzialmente tale problema ADDATI propone ed implementa una strategia di classificazione che usa in modo integrato sia procedure non gerarchiche che gerarchiche. Essa è il risultato evolutivo di un'altra strategia usata in passato, che indichiamo qui sotto come Metodo 1. Ci limiteremo qui a descrivere con qualche dettaglio i due metodi (quello usato nelle versioni precedenti del pacchetto e quello utilizzato attualmente) e le loro relazioni.

Riteniamo invece che il metodo di classificazione gerarchica ascendente, usato internamente dalla complessa procedura non gerarchica, sia semplice e diretto e non abbisogni di spiegazioni dettagliate. Esso era anche implementato come opzione direttamente utilizzabile (**CGA**: Classificazione Gerarchica Ascendente) nelle precedenti versioni del pacchetto, ma poiché l'interpretazione diventa quasi impossibile già non appena le unità statistiche da aggregare superino il centinaio (cioè quasi sempre), si è deciso di eliminarlo.

L'input per la routine di clustering è la tavola di coordinate fattoriali registrate da **ACORR** o **ACOMP** dopo aver elaborato una tavola di variabili osservate quantitative o categoriali, o un inseme di tavole di contingenza affiancate.

7.2.2 – Classificazione non gerarchica: l'algoritmo implementato in ADDATI

Allo scopo di produrre una partizione soddisfacente nelle versioni precedenti del pacchetto venivano nettamente distinte due fasi (*Metodo 1*): una **esplorativa,** che cercava di determinare il numero di classi più opportuno e suggeriva una buona scelta dei centri iniziali; la successiva **fase di ottimizzazione** generava poi la partizione finale ottima.

La sequenza è stata resa più agile a partire dalla versione 4.0, sotto il nome di *Metodo 2*. Per un certo tempo i *metodi 1 e 2* hanno convissuto nel pacchetto, e l'utente poteva scegliere quale usare. L'uso comparato, protratto abbastanza a lungo, ha portato alla decisione di elinare il *metodo 1* a partire dalla versione 5.2.

Ci limiteremo dunque qui a descrivere il solo *metodo 2*, distinguendolo ancora per comodità in una fase esplorativa ed una di ottimizzazione, anche se ora le due fasi sono implementate come un processo continuo, senza soluzioni di continuità che richiedano interventi dell'utente.

La fase esplorativa

Invece che una sola, si calcolano ***molte partizioni di base*** (diciamo, qualche decina). In linea di massima, per ogni partizione si consiglia di chiedere un numero di classi pari a quello su cui ci si vorrebbe attestare nella partizione finale. I centri iniziali sono per lo più scelti in modo casuale, ma ADDATI offre alcune alternative.

Vengono incrociate le due o tre partizioni che presentano il valore più elevato della funzione-obiettivo (che misura l'omogeneità interna delle classi prodotte), cioè le migliori in senso statistico. La ***partizione-prodotto*** ha un numero di classi a priori indeterminato: per costruzione, gli elementi di una classe sono stati classificati congiuntamente (cioè sono stati assegnati ad uno stesso gruppo) in tutte le partizioni di base incrociate, e sussiste dunque una ragionevole convinzione sulla fondatezza della loro somiglianza. Proprio per tale motivo le classi della partizione-prodotto sono

note come classi stabili o forme forti. Anche se spesso sono in numero eccessivo per gli scopi della ricerca, esse offrono una descrizione dettagliata e spesso esaustiva dei principali comportamenti ravvisabili nel contesto d'analisi dato.

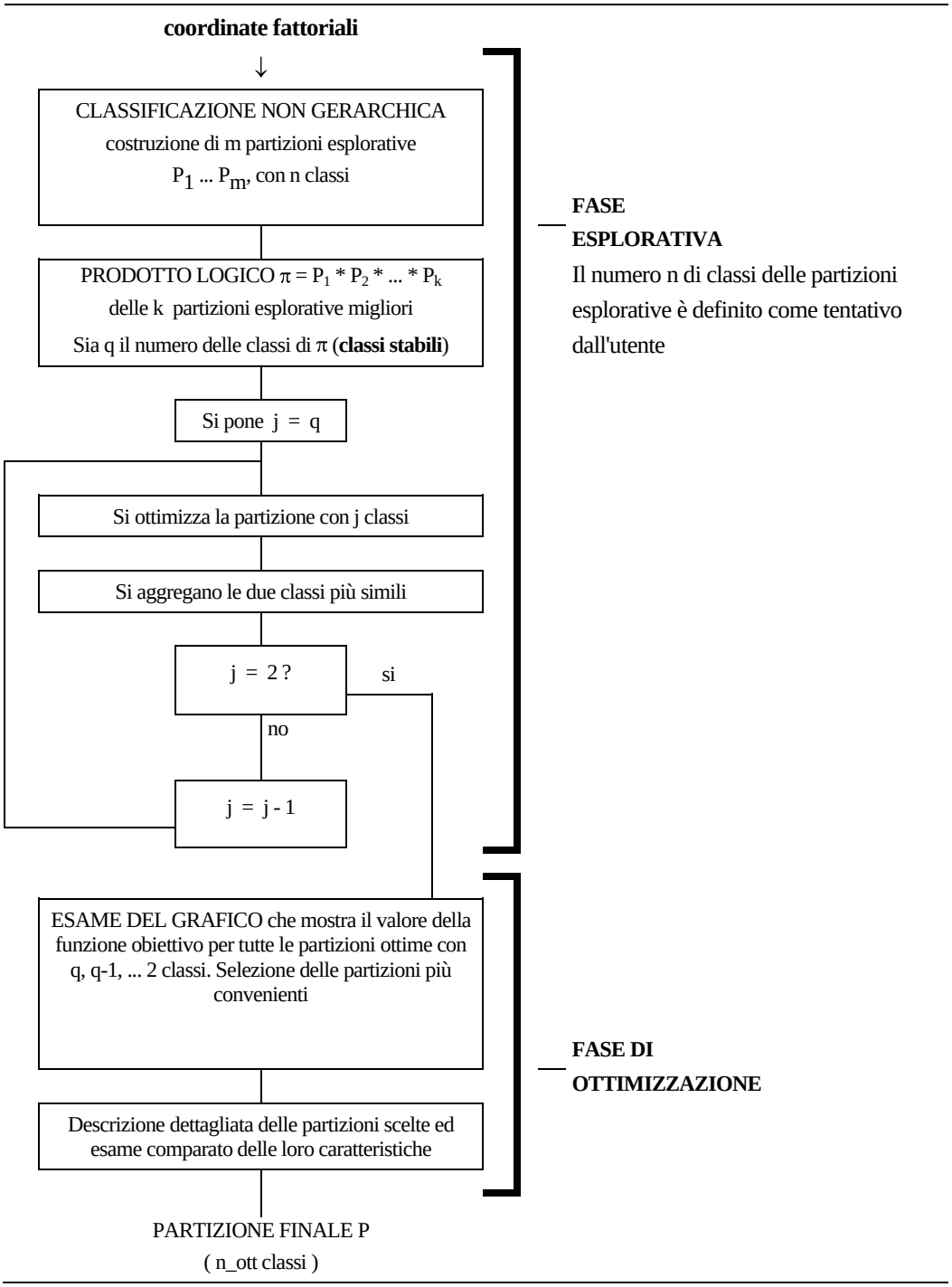


Figura 7.3 – La strategia di classificazione implementata in ADDATI (schema).

La fase di ottimizzazione

La partizione-prodotto così ottenuta, che consiste quasi sempre di un numero di classi troppo elevato per gli scopi della ricerca, viene assunta come la configurazione iniziale da ottimizzare. Essa è stata costruita in modo tale che i suoi gruppi dovrebbero rappresentare in dettaglio i diversi comportamenti emergenti nell'insieme I da classificare.

Sia *q* il numero delle ***classi stabili*** (le classi della partizione-prodotto) A questo punto vengono chiamate iterativamente due routines: la prima ottimizza la partizione corrente e ne salva su file una descrizione essenziale, sufficiente a ricostruirla con facilità; la seconda riduce di uno il numero dei gruppi aggregando i due più simili. Si ottiene così una partizione *non-ottima* con *q-1* classi.

Questa processo di ottimizzazione/aggregazione viene ripetuto, ottenendo una partizione non-ottima con *q-2* classi. La cosa prosegue iterativamente fino all'aggregazione totale dei gruppi.

Appare a questo punto sullo schermo il dialogo di figura 7.4.

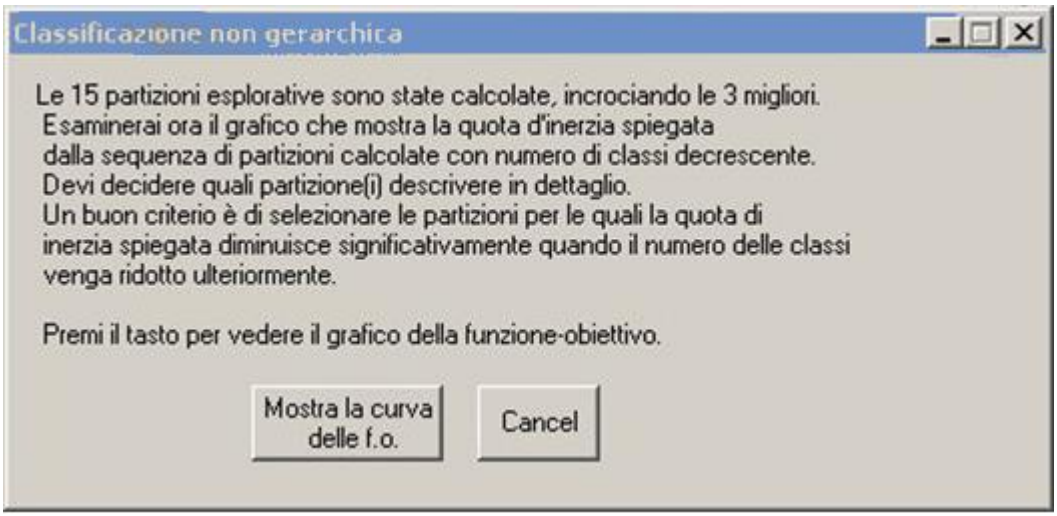


Figura 7.4 – Il dialogo per la scelta delle partizioni da descrivere

Premendo il bottone di sinistra appare un grafico come quello di figura 7.5. Esso mostra come varia il valore della funzione-obiettivo con il numero delle classi, per la sequenza di partizioni ottime calcolate (con numero di classi via via decrescente). Si possono prendere in considerazione le più promettenti, cioè quelle per le quali la f.o. diminuisce significativamente quando il numero delle classi venga diminuito di uno. L'utente può richiedere una descrizione dettagliata di tutte le partizioni cui è interessato: il confronto delle loro caratteristiche lo orienterà nella scelta finale.

Va sottolineato che il numero di classi inizialmente suggerito dall'analista è usato solo per costruire la partizione-prodotto. Essa è il punto di partenza per il passo successivo, che determina una sequenza di partizioni ottime con numero di classi progressivamente decrescente.

Il numero delle classi della partizione finale – scelta dopo l'esame del grafico della funzione-obiettivo e dei caratteri delle partizioni candidate nel caso ve ne sia più d'una - dovrebbe dunque emergere come una proprietà intrinseca dell'insieme da classificare.

Ovviamente, non potremo mai essere certi di aver trovato la partizione migliore in assoluto con quel numero di classi (il cosiddetto **optimum optimorum**). Il metodo è euristico e fornisce una partizione di buona qualità, non l'ottimo assoluto. Ma ciò è ben noto, e bisogna accettarlo.

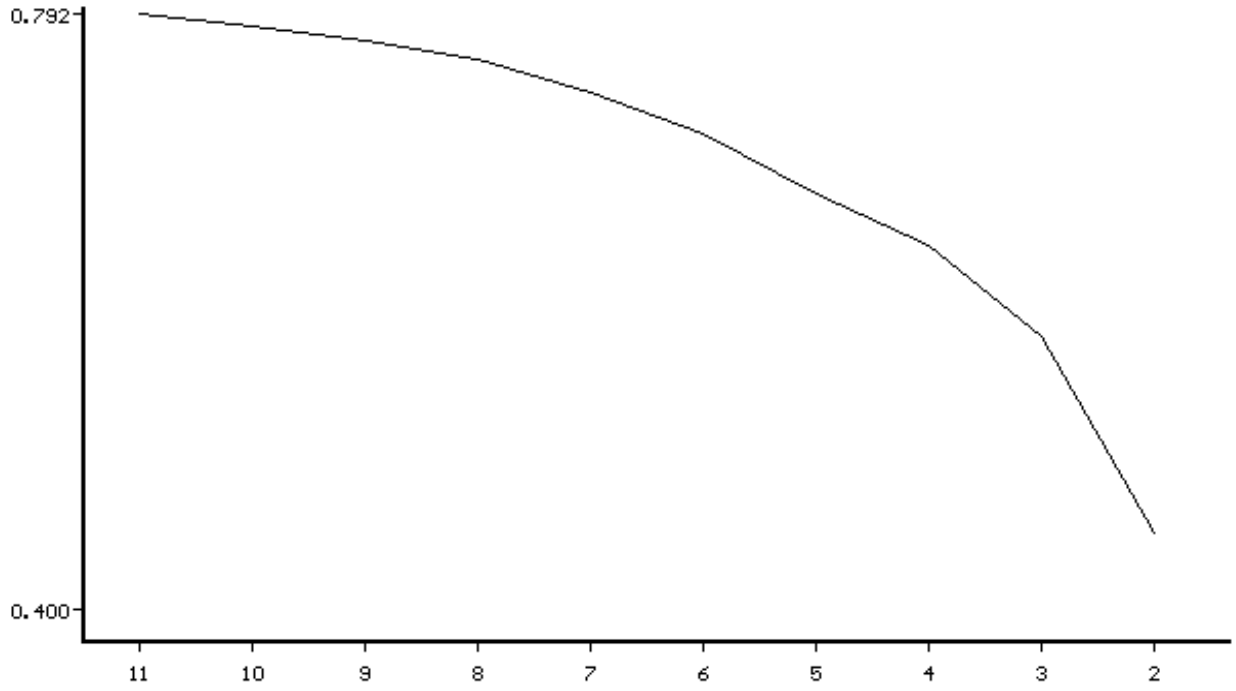


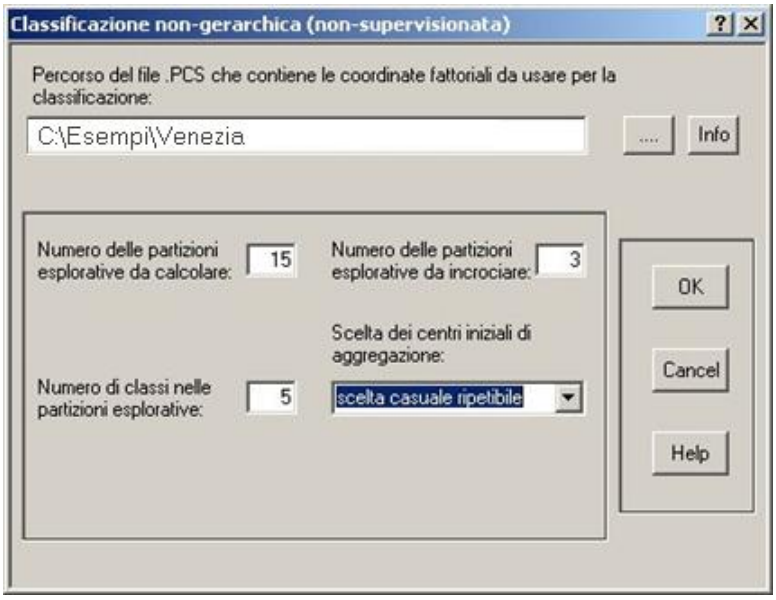
Figura 7.5 - Il grafico è un esempio di come il valore della funzione-obiettivo vada diminuendo man mano che il numero delle classi si riduce attraverso successive aggregazioni ed ottimizzazioni.

7.3 - I dialoghi della procedura di Classificazione NONGER

Funzione	Si calcola una partizione ottima su di un insieme di unità, usando un metodo non gerarchico.
Limiti	Per velocizzare il calcolo la tavola delle coordinate fattoriali viene caricata in memoria centrale. Si può usare l'intera memoria disponibile, ma se l'insieme delle unità da classificare è molto numeroso Windows dovrà ricorrere all'uso di memoria virtuale su disco, e l'esecuzione potrebbe diventare piuttosto lenta. Di solito, tuttavia, è abbastanza veloce. Viene permesso di incrociare un massimo di **tre partizioni esplorative**. La ragione di questo limite è di impedire ad un utente superficiale di produrre un numero elevatissimi di classi stabili, cosa possibile. Ottimizzare iterativamente una partizione con 200 classi, poi una di 199, e così via, può essere molto lento! E' bene che l'utente se ne convinca, e **per quanto possibile cerchi di non esagerare neppure con il numero delle classi richieste**.

7.3.1 - Parametri per la fase esplorativa

Vengono inseriti nel dialogo seguente (i valori mostrati si riferiscono all'esempio su Venezia).



Se la classificazione segue un'analisi fattoriale eseguita nella medesima sessione di lavoro, il percorso del file dove sono state salvate le coordinate fattoriali appare automaticamente. Se invece si intende elaborare coordinate fattoriali registrate in una sessione di lavoro precedente, l'analista deve inserire il nome del file, o sfogliare per puntare ad esso.

Dal file .PCS verranno anche letti

- il nome del file .TMP che registra il valore delle variabili sottoposte all'analisi fattoriale (ACOMP o ACORR). Tali valori servono al calcolo dei profili delle classi risultanti dalla classificazione.
- il file .FPL con le informazioni necessarie alla visualizzazione dei piani fattoriali. NONGER modificherà tale file aggiungendo i punti che rappresentano i centri delle classi. Inoltre, i quadratini che danno la posizione delle singole unità statistiche nel piano vengono sostituiti dal numero della classe cui ciascuna unità è stata assegnata.

Numero delle partizioni esplorative da calcolare

Indicativamente, 10-20. Il calcolo di un numero maggiore di ***partizioni di base*** (o ***esplorative***) aumenta la probabilità di ottenere alcune buone partizioni da incrociare per il calcolo della partizione-prodotto, ma richiede un tempo di calcolo proporzionalmente più lungo. Comunque, questo non è un problema con i computer moderni (lo era però in passato), a meno che il DataSet non sia davvero molto numeroso. L'occupazione di memoria non ne risente.

Ovviamente, succede che quando le unità da classificare sono relativamente poche non è necessario calcolare molte partizioni, mentre la cosa è opportuna, nella speranza di trovare qualche partizione buona, proprio quando esse sono numerose. Sfortunatamente, ciò aumenta i tempi di calcolo...

Numero delle partizioni esplorative da incrociare

Le migliori partizioni di base, calcolate in via esplorativa, vengono incrociate per determinare i gruppi più compatti emergenti dall'analisi (le cosiddette ***classi stabili*** o ***forme forti***).

La scelte su quante partizioni incrociare dipende dal livello di dettaglio desiderato per le classi stabili. Ad esempio, se vengono richieste partizioni in sette classi, e se ne incrociano due, in via teorica si potrebbero ottenere fino a 49 classi stabili (cioè, 49 combinazioni diverse delle due classi alle quali ciascuna unità viene assegnata nelle due partizioni).

In generale, più fortemente strutturato è l'insieme da classificare, più simili ci si può attenedere risultino le due partizioni incrociate, ed il numero delle classi stabili si ridurrà. Esse sarebbero esattamente sette se le due partizioni incrociate fossero identiche.

Allo scopo di evitare una frammentazione eccessiva della partizione-prodotto, che renderebbe più difficile l'interpretazione, ADDATI accetta di incrociare un massimo di tre partizioni esplorative.

Numero di classi nelle partizioni esplorative

Il numero è fornito in via tentativa, e dovrebbe rappresentare il numero ideale di gruppi che l'utente, per i suoi scopi operativi o di conoscenza, desidererebbe ottenere alla fine del processo.

Il numero delle classi della partizione finale viene effettivamente deciso dopo un attento esame del grafico che mostra come la f.o. diminuisca al diminuire del numero delle classi.

Scelta dei centri iniziali di aggregazione:

1. **scelta casuale ripetibile**
2. **scelta casuale non ripetibile**

Ulteriori opzioni sulla scelta dei centri inizali, già presenti nella versione 5.2, verranno probabilmente aggiunte in futuro. Comunque, una scelta casuale dei centri di aggregazione è una buona soluzione. Le due alternative attuali hanno il significato seguente.

Scelta casuale ripetibile

Se si debbono classificare n unità statistiche, vengono generati per ciascuna partizione tanti numeri interi tra 1 ed n quante sono le classi da costruire, e le unità con quei numeri d'ordine vengono assunte come centri iniziali di aggregazione. La generazione casuale parte da un **seme fisso** (cioè un valore inizialmente fissato, che determina la sequenza di numeri generati casualmente). Se si ripete l'analisi mantenendo il seme di partenza, vengono generati gli stessi centri di aggregazione e si ripete la sequenza di partizioni già trovata.

Scelta casuale non ripetibile

In questo caso il seme che determina la sequenza generata di numeri casuali non è fisso, ma viene derivato dall'orologio interna della macchina e cambia dunque di volta in volta. Ripetendo l'analisi si generano sequenze diverse di numeri casuali, e dunque anche sequenze di partizioni diverse.

7.4 - NONGER - Fase di ottimizzazione e descrizione delle partizioni

Supponiamo che la partizione-prodotto generata nella fase esplorativa consista di *q classi stabili*: **NONGER** comincia ad ottimizzare questa partizione riallocando opportunamente alcune unità (metodo di Diday). Poi vengono riuniti i due gruppi più simili, e la partizione ottenuta, con *q-1* classi, viene anch'essa ottimizzata. L'operazione viene ripetuta, generando una partizione in *q-2* classi, e così via fino ad ottenere una partizione con due sole classi.

A questo punto **NONGER** chiama una programma di utilità interno di ADDATI che visualizza il grafico della funzione-obiettivo vs. il numero delle classi, per la sequenza delle partizioni ottime appena calcolate: il valore della f.o. ovviamente diminuisce al diminuire del numero delle classi. La figura 7.6 si riferisce all'esempio di Venezia C.S., nel quale si sono determinate 11 *classi stabili*.

Esaminando il grafico l'utente può centrare l'attenzione su una o più *partizioni candidate*, con un numero di classi nel range desiderato ed un valore della f.o. sufficientemente alto.

Quando sceglie la partizione finale l'utente dovrebbe valutare il **trade-off** tra il livello di sintesi ottenibile (meno classi sono più convenienti) ed il valore della funzione-obiettivo, che rappresenta la quota d'informazione conservata. Il numero delle classi va ridotto per quanto possibile, ma non al prezzo d'impastare tutto: il valore della f.o. non deve diminuire troppo. E' dunque opportuno spingere il livello di aggregazione (passare cioè alla successiva partizione a destra lungo il grafico) se questo non costa troppo in termini di diminuzione della f.o.

Va dunque selezionata una partizione tale che un'aggregazione ulteriore comporti una diminuzione di qualità (valore della f.o.) che non si è disposti a pagare. Ciò succede nelle zone della curva nelle quali la pendenza cambi bruscamente quando ci si muova di un passo verso destra.

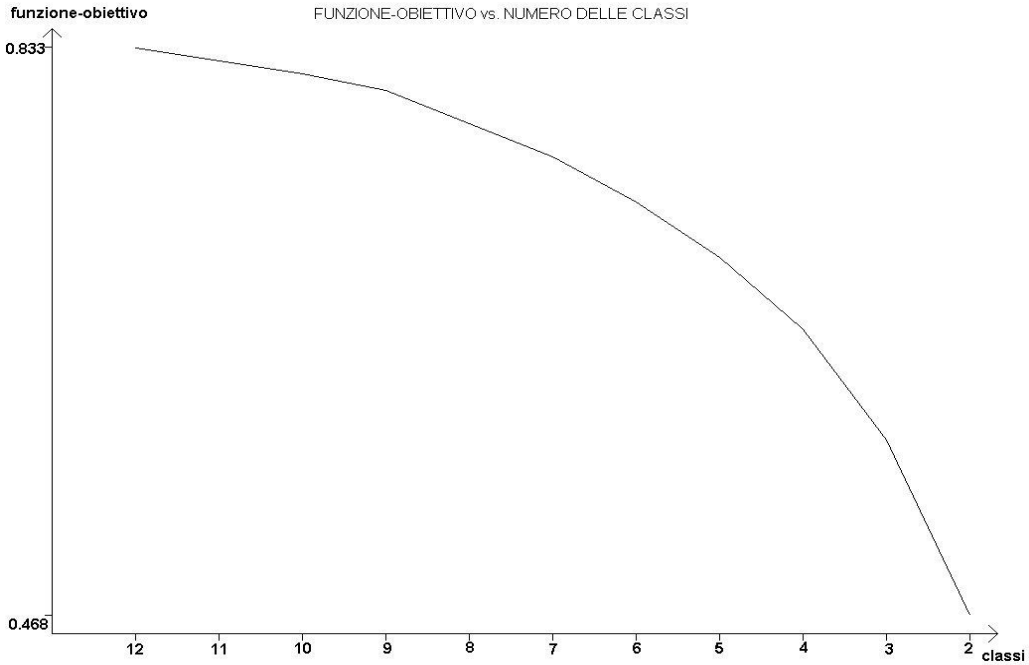
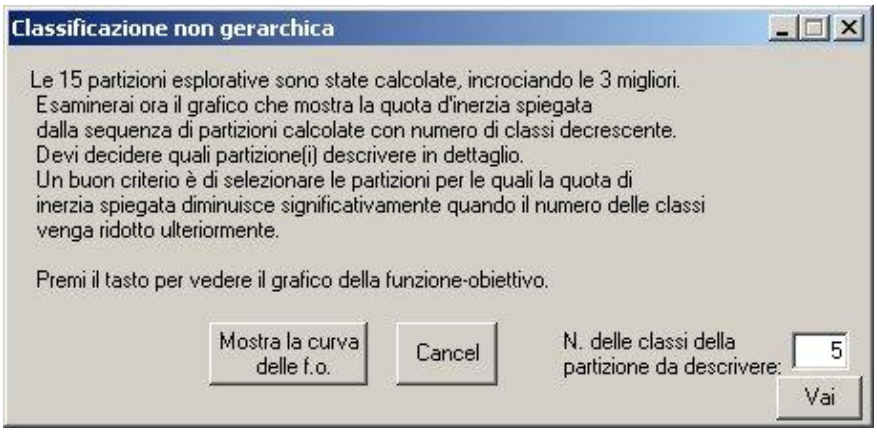


Figura 7.6 - Il grafico è un esempio di come il valore della funzione-obiettivo vada diminuendo man mano che il numero delle classi si riduce attraverso successive aggregazioni ed ottimizzazioni.

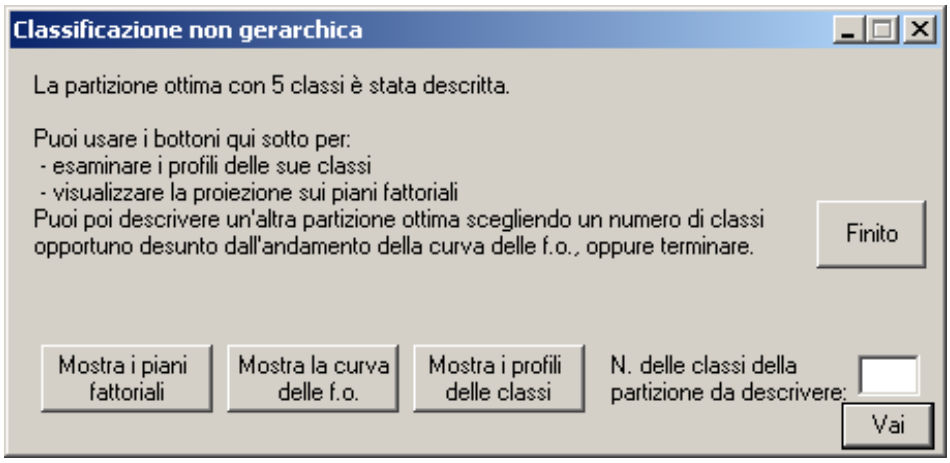
La decisione dovrebbe tener conto dei seguenti tre aspetti.

- Il numero dei gruppi che l'utente considera più adatto per i suoi scopi dovrebbe orientarlo verso la parte del grafico da considerare con attenzione.
- La diminuzione del valore della funzione-obiettivo conseguente ad ogni passo di aggregazione dovrebbe limitare il suo interesse ad una o a poche partizioni candidate, da descrivere in dettaglio.
- La decisione finale dovrebbe essere assunta dopo l'esame dei caratteri di queste partizioni candidate - specialmente dei profili delle classi emergenti - scegliendo quella che più risponde agli scopi dell'analisi.

Dopo aver deciso quali partizioni candidate meriti di esaminare a fondo, l'utente deve chiudere il grafico scegliendo dal Menu l'opzione *File→Exit*. Appare il dialogo mostrato qui sotto, nel quale si chiede all'utente di inserire il numero delle classi della partizione ottima da descrivere. La richiesta viene ripetuta fino a che non siano state descritte tutte le partizioni desiderate, le cui caratteristiche vanno confrontate per raggiungere la decisione finale.



Qui si è richiesto di descrivere la partizione ottima con cinque classi. Immediatamente dopo, appare il dialogo seguente.



È ora possibile esaminare in dettaglio le caratteristiche della partizione (dedicando un'attenzione particolare ai profili delle classi), visualizzare le proiezioni sui principali piani fattoriali, descrivere un'altra partizione, rivedere il grafico della f.o. e, una volta deciso su quale partizione finale attestarsi, terminare.

7.4.1 - L'esame dei profili di classe

Il profilo di ciascuna classe va confrontato con quello globale (cioè con i valori medi assunti dalle variabili sull'insieme delle unità statistiche) allo scopo di accertare quali variabili lo caratterizzino di più in quanto presentano nella classe un valore medio significativamente maggiore o minore di quello medio globale. Per facilitare la cosa, sotto i valori numerici delle variabili vengono scritti dei segni particolari il cui scopo è di attrarre velocemente l'attenzione dell'analista verso i valori da considerare.

Sotto ogni valore del profilo appare una delle stringhe seguenti:

"----", "--", "~~~", "++", "++++"

La stringa scritta è determinata dal **rapporto** tra la componente considerata del profilo di classe e la sua omologa sul profilo globale. Tale rapporto viene calcolato per ogni componente del profilo di ciascuna classe, e confrontato con un insieme di valori di soglia pre-definiti. La stringa è scelta secondo gli schemi riportati nelle tavole 7.1 e 7.2 qui sotto. Quelle riportati sono le soglie usate normalmente, adatte nella maggior parte dei casi.,

Quando le variabili sono QUANTITATIVE, i segni '+' e '-' usati per facilitare l'esame dei profili di classe hanno il significato seguente. Sia

xm(j,i) la **media** della variabile j nella classe i

xg(j) la **media globale** della stessa variabile j.

La variabile j è rilevante per il cluster i quando la differenza xm(j,i) - xg(j) è lontana dallo zero. Per valutare la sua significatività, tale differenza viene confrontata con la deviazione standard **σ(j)** della variabile. Il valore del rapporto

$R = [xm(j,i) - xmg(j)] / \sigma(j)$

viene confrontato con **quattro soglie s1,..., s4 scelte opportunamente**, i cui valori correnti sono riportati qui sotto. La stringa di aiuto è poi stabilita secondo lo schema seguente:

```
      ----       --       ~~~       ++      ++++    valore
   ---------->|<---------->|<---------->|<---------->|<---------  del
        -1.00       -0.20        0.20       1.00          rapporto R
```

Tabella 7.1 - I valori di soglia assunti come default per aiutare l'interpretazione dei profili di classe nel caso di variabili quantitative.

I simboli + e - adottati per descrivere il profilo di classe vanno interpretati come segue.

Il riferimento e` al RAPPORTO tra la frequenza di ciascuna variabile nella classe e la sua frequenza globale.

Con le soglie attuali, tale rapporto risulta cosi rappresentato:

```
  ----          - -          ~ ~          + +          ++++  valore
----------->|<---------->|<---------->|<---------->|<-------- del
         -1.00          -0.20          0.20           1.00          rapporto R
```

Tabella 7.2 - I valori di soglia assunti come default nel caso di analisi di una o più tavole di contingenza.

Nota: *Se la tavola analizzata consiste di variabili quantitative,i valori di soglia vengono automaticamente determinati dal programma e dipendono, **per ciascuna variabile**, dalla sua deviazione standard. Mentre le classi vengono determinate in base alle coordinate fattoriali, i profili vengono calcolati a partire dai valori delle variabili originali ed usando la loro unità di misura iniziale. I valori sono più intuitivi, e l'interpretazione più facile.*

Componenti del profilo

Le componenti del profilo hanno un diverso significato nei due casi.

- **Variabili quantitative** (tavole di misura): per ogni gruppo, una componente del profilo rappresenta il valore medio di una variabile nella classe. La corrispondente componente del profilo globale è la media globale della variabile (cioè, calcolata sull'insieme di tutte le unità).
- **Tavole di contingenza**: una componente del profilo rappresenta la frequenza di una variabile (o di una categoria) nella classe. La componente corrispondente del profilo globale è la frequenza globale di quella variabile o categoria.

Per quanto riguarda le variabili quantitative le indicazioni sui valori di soglia fornite dal programma sono ragionevoli e l'utente non dovrebbe cambiarle. Quanto alle tavole di contingenza, a volte può succedere che le unità statistiche, e dunque anche le classi, siano poco diverse: le soglie di default potrebbero allora non essere appropriate: la cosa dipende dalla definizione della variabile e dal suo significato. ADDATI dovrebbe allora consentire all'utente di cambiare le soglie. Tale possibilità, tuttavia, non è ancora implementata.

NONGER - L'interpretazione dei risultati

🗁 Illustreremo l'uscita della procedura di classificazione non-gerarchica con riferimento all'esempio sul Centro Storico di Venezia considerato per **ACOMP** (cfr. sezione 6.2). Supponiamo di aver registrato come input per la classificazione le prime quattro componenti principali (che riassumono il 96% dell'Inerzia totale) e di inserire i seguenti parametri per pilotare la fase esplorativa:

- partizioni esplorative da calcolare : 15
- numero di migliori partizioni da incrociare : 2
- numero di classi in ciascuna partizione esplorativa : 5
- scelta casuale ripetibile dei centri iniziali di aggregazione

Incrociando le due partizioni migliori come richiesto, la fase esplorativa genera 12 classi stabili. Esaminando il grafico della funzione-obiettivo, riportato nella figura 7.6, vediamo che la pendenza della curva s'impenna un po' di più in corrispondenza delle partizioni con 9 e con 4 classi. Stando alle considerazioni fatte in precedenza, queste sono le partizioni delle quali dovremmo chiedere la descrizione. Ma poiché 9 classi sono troppe e 4 sono poche rispetto a quello che volevamo, optiamo per la partizione ottima con 5 classi (potremmo comunque chiedere la descrizione di tutte, inclusa quella più dettagliata, con 12 classi...).

Per ogni partizione della quale si sia richiesta la descrizione, il file di output NG.OUT riporta una tabella che elenca il numero delle unità in ciascuno dei gruppi prodotti, ed il suo peso.

CLASSE	1	2	3	4	5	TOT
UNITÀ	29	36	40	31	12	148
PESO (%)	22.3	23.1	25.5	20.4	8.7	100.0

Tabella 7.3 - Numerosità e peso delle classi della partizione nel caso di Venezia.

🗁 La tabella 7.3 elenca le classi stabili ottenute nel caso di Venezia (si ricordi che si è assunto come peso di ciascuna sezione il numero dei suoi alloggi occupati).

Segue una descrizione dettagliata delle classi. La tabella 7.4 mostra come esempio l'informazione relativa alla classe 1 della partizione con 5 classi.

```
***************
*  CLASSE  1  *
***************

UNITA': 29 - PESO:  22.26%

UNITA' ASSEGNATE ALLA CLASSE: 21  28  48  55  56  57  59  60  66
67  70  77  86  105 106 107 108 109 110 113 114 123 125 127 131 132
134 137 148

UNITA' PIU' VICINA AL CENTRO DELLA CLASSE (d2 =  0.3319) : 106

UNITA' PIU' LONTANA DAL CENTRO DELLA CLASSE (d2 =  7.6026) : 48

RAGGIO DELLA CLASSE : 1.46562

DISTANZA DEL CENTRO DI CLASSE DAL CENTRO GLOBALE :  0.97449
```

Tabella 7.4 - Descrizione dettagliata della classe 1 (esempio di Venezia).

Vengono elencate le unità in ciascuna classe, unitamente al valore dei seguenti indicatori:

- il **raggio di classe**, che è un'indicatore della compattezza della classe. Sia $In_{int}(j)$ l'inerzia interna della j-esima classe, ottenuta sommando le inerzie rispetto al centro di classe $\mathbf{G}_j$ di tutte le unità assegnate alla classe. Tale valore si può scrivere come

$$In_{int}(j) = \sum_i m_i * d^2(i,\mathbf{G}_j) = M_j * d_j^2$$

  dove Mj è il peso di classe e d_j rappresenta il suo raggio medio, cioè la distanza da **G**j alla quale tutta la massa di classe dovrebbe distribuirsi per dar luogo ad un'inerzia pari a $In_{int}(j)$. Se ne può dedurre che

$$d_j = [In_{int}(j) / M_i]^{1/2}$$

- la **distanza del centro di classe** dal centro globale della nuvola, come indicatore della peculiarità dei caratteri della classe: maggiore è la distanza, più i caratteri medi della classe si differenziano dal comportamento medio generale rappresentato dal centro globale della nuvola.

🗁 La distanza del centro della classe 1 dal centro globale vale 0.97449 mentre, per confronto, l'analogo valore è 1.68473 per la classe 2. La prima risulta dunque più baricentrica, mentre la seconda esibisce caratteri più particolari. L'esame dei profili di classe (si veda più avanti) aiuta ad individuare esplicitamente di che tipo di peculiarità si tratti.

Una volta descritti i gruppi, vengono forniti alcuni parametri generali che caratterizzano la partizione:

- l'inerzia interna totale In_{int} (2.32453 per la partizione in 5 classi su Venezia);
- l'inerzia esterna totale In_{ext} (5.38338 per Venezia);
- l'inerzia totale della nuvola, pari alla somma degli autovalori relativi alle Componenti Principali utilizzate per la classificazione. Nel caso di Venezia il suo valore è 7.70791: siamo infatti partiti con otto variabili attive standardizzate, e dunque da un'Inerzia totale pari ad 8, utilizzando per la Classificazione le prime 4 CP, che riassumevano il 96.35 % dell'Inerzia totale;

il valore della funzione-obiettivo, che misura la qualità della partizione: il valore massimo è 1, raggiungibile solo quando si abbiano tante classi quante sono le unità effettivamente diverse (nel caso dell'esempio su Venezia il valore della funzione-obiettivo è 0.69842 = 5.38338 / 7.70791).

Seguono i profili delle classi della partizione. Per il significato degli aiuti all'interpretazione ("++", "--", ecc.) si rimanda alla spiegazione precedente.

🗁 Esempio su Venezia: la tabella 7.5 mostra i profili della partizione ottimizzata in 5 classi.

Le descrizioni di tutte le partizioni richieste vengono registrate in sequenza nel file **NGnnn.TXT**. Per ciascuna di esse l'informazione sulla classe di assegnazione delle diverse unità statistiche viene registrata su di un **file di testo** denominato '**NGCLASnn.TXT**', dove 'nn' sta per il numero delle classi richieste. Il file può essere direttamente caricato da **ARC/VIEW**, programma GIS che si può utilizzare per disegnare la mappa che rappresenta la classificazione quando si tratti di unità geografiche (Comuni, Sezioni censuarie e simili) per le quali si possiedano i file cartografici.

I profili delle classi della partizione scelta vengono salvati anche nel file **NGCLASnn.CSV**. Tutti i campi sono separati da virgole, ed il file può venire direttamente caricato in EXCEL come file di testo.

CLASS	NUM	PIA_ alto	PIA_t erra	SER_b uoni	SER_ca renti	SER_sc adenti	AFF_ alto	AFF_s tand	AFF_so ttouti
1	29	83.045	16.955	51.615	30.060	18.382	16.623	64.286	19.147
		--	++	~~~	~~~	~~~	~~~	++	--
2	36	94.857	5.143	55.719	29.321	15.014	13.892	60.877	25.288
		++++	----	++	--	--	--	~~~	++
3	40	90.024	9.976	66.981	22.235	10.909	11.426	54.384	34.317
		++	--	++++	--	--	--	--	++++
4	31	80.220	19.780	34.089	42.720	23.239	24.827	63.224	11.997
		--	++	----	++++	++	++	++	--
5	12	71.935	28.065	24.088	44.897	31.088	41.731	52.965	5.377
		----	++++	----	++++	++++	++++	----	----
PROFILO GLOBALE	148	86.012	13.988	50.504	31.772	17.797	18.529	59.771	21.774

CLASS	NUM	STS_ alto	STS_ oper	STS_a ltro
1	29	23.112	17.212	59.718
		~~	~~~	++
2	36	26.563	13.795	59.698
		++	--	++
3	40	33.801	10.292	56.021
		++++	--	--
4	31	15.188	25.062	59.763
		----	++	++
5	12	10.683	33.297	56.057
		----	++++	--
PROFILO GLOBALE	148	23.934	17.662	58.462

Tabella 7.5 - I profili della partizione ottimizzata in cinque classi.

I file scritti da NONGER

NONGER scrive i seguenti file che riportano in varia forma i risultati dell'analisi.

- **NGnnn.TXT**, dove '**nnn**' è un numero progressivo fissato in modo da evitare di sovrascrivere file esistenti prodotti da altre analisi. Il file elenca l'attribuzione delle unità alle classi, descrive i profili di classe ed offre una serie di informazioni utili all'interpretazione. Contiene in sequenza la descrizione di tutte le partizioni richieste.
- **NGCLASnn.TXT**, registra la classe di assegnazione di ogni unità statistica. '**nn**' è il numero delle classi della partizione cui il file si riferisce: viene salvato un file diverso per ogni partizione descritta.

  Questi file hanno tanti record quante sono le unità classificate, più uno di intestazione. Il loro formato li rende direttamente caricabili da **ArcView** come file di attributi in formato testo. Ogni record riporta l'identificatore dell'unità geografica **sia come stringa che come numero**, seguiti dal numero della classe cui l'unità è stata assegnata. L'informazione permette di integrare l'archivio originale oppure - nel caso si siano classificate unità geografiche - di visualizzare la mappa della classificazione.

- **NGCLASnn.CSV** (Comma-Separated Values) registra i profili delle classi in un formato immediatamente accettato da **EXCEL** come file di testo per ulteriori elaborazioni. Viene usata come separatore di campo la virgola. '**nn**' è il numero delle classi nella partizione cui il file si riferisce: anche qui, viene salvato un file di questo tipo per ciascuna partizione della quale si sia richiesta la descrizione.

- **NGnn.FPL** è un file che **NONGER** scrive per **FACPLAN** solo quando nell'analisi fattoriale che precede (**ACORR** o **ACOMP**) sia stato salvato il file necessario per visualizzare le proiezioni sui piani fattoriali. Se tale file esiste, **NONGER** ne legge il contenuto e lo arricchisce, riscrivendolo poi con il nome **NGnn.FPL**. Quando **FACPLAN** ne visualizza il contenuto, mostra la posizione delle unità non più con un quadratino ma mediante il numero corrispondente alla classe di assegnazione; sono inoltre visualizzati anche i centri delle classi. Si tratta di un altro modo di rappresentare come le classi (visibili come sub-nuvole contrassegnate dai numeri '1', '2', ecc.) si collochino rispetto alle variabili.