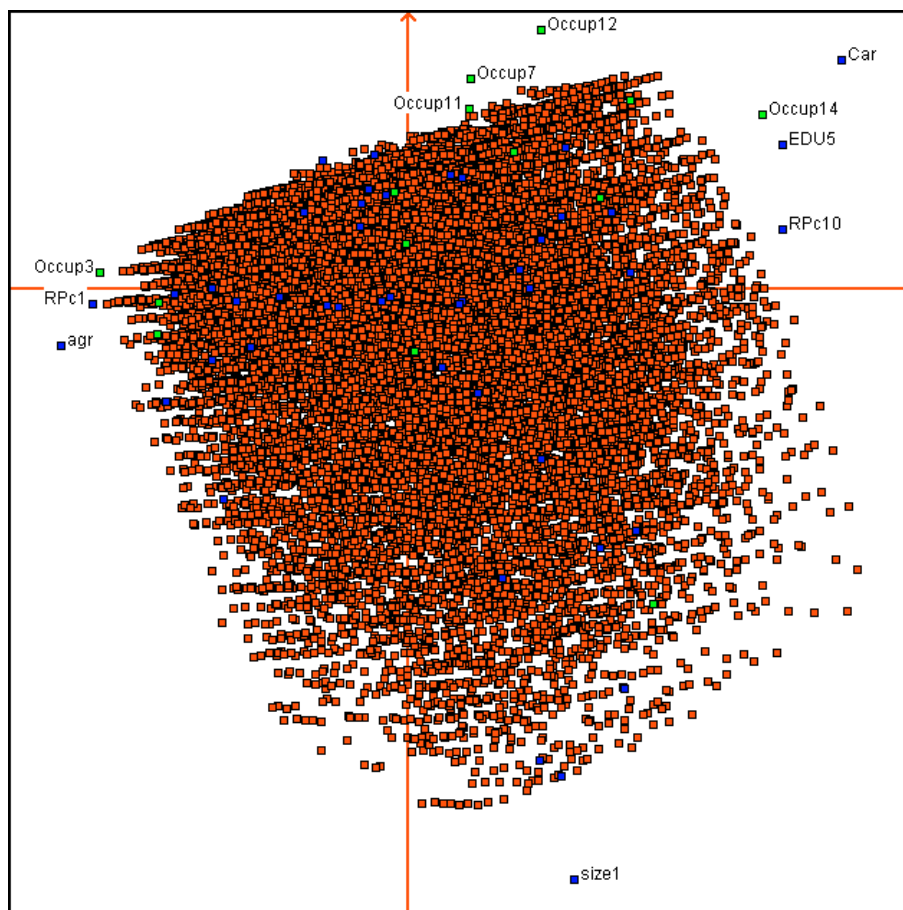


*Silvio Griguolo*

# ADDATI

Un pacchetto per l'analisi esplorativa dei dati  
(versione 5.2a – Luglio 2003)

## Guida all'uso



Istituto Universitario di Architettura di Venezia - Dipartimento di Pianificazione

*S. GRIGUOLO (\*)*

# **ADDATI**

**Un pacchetto per l'analisi esplorativa dei dati**

## **Guida all'uso**

---

(ver. 5.2a – Luglio 2003)

(\*) Questa edizione del Manuale costruisce sulla precedente, che illustrava la versione 4.3 di ADDATI, scritta con la collaborazione di Massimo Mazzanti.

<b>Istituto Universitario di Architettura di Venezia - Dipartimento di Pianificazione</b>
---

<b>INDICE .....</b>	<b>3</b>
<b>ADDATI V. 5.2A – LUGLIO 2003 - PRESENTAZIONE .....</b>	<b>1</b>
<i>Dove trovare i dati per le analisi?</i> .....	2
UNA BREVE DESCRIZIONE DELLA VERSIONE 5.2A .....	3
<i>Un po' di storia...</i> .....	3
<i>Attribuzioni e riconoscimenti</i> .....	5
<i>L'impostazione di questa Guida all'uso</i> .....	5
<b>CAP 1. - INSTALLAZIONE E GENERALITÀ.....</b>	<b>1</b>
1.1 – INSTALLAZIONE.....	1
1.2 – IL PROGRAMMA DI GESTIONE ADDATL.EXE.....	3
Il Menu FILE .....	4
<b>CAP 2. - IL MENU DI UTILITÀ.....</b>	<b>1</b>
Il formato di registrazione dei file di dati.....	1
I programmi di utilità.....	2
2.1 - MERGCHAR .....	4
2.2 - MERGFIELD .....	6
Possibili errori di esecuzione .....	8
2.3 - FIXFORM .....	11
2.4 - MISSVAL .....	12
Come funziona <b>MISSVAL</b> .....	12
Raw imputation.....	12
Un esempio .....	15
La struttura del file MISSVAL.PAR.....	15
2.5 - SELECT .....	19
L'operatore " / " .....	21
2.6 - SHOWREC .....	22
2.7 - RECODE.....	24
Le operazioni di ricodifica .....	26
Tasti per il controllo delle funzioni.....	27
2.8 - FACPLAN .....	29
2.9 - INTEGRA .....	31
<b>CAP. 3 - MENU DI ANALISI: L'INTERFACCIA UTENTE.....</b>	<b>1</b>
3-1 ALCUNI ASPETTI RILEVANTI IN ADDATI.....	1
3-2 L'INTERFACCIA UTENTE (ALCUNE NOTE GENERALI) .....	2
Cambiare le risposte già fornite .....	2
Caricamento di un file di parametri .....	3
Il caso di indicatori alfanumerici multipli .....	3
L'help in linea .....	3
Il formato di lettura dei dati .....	4
<b>CAP 4. - FONDAMENTI DI TEORIA E LINGUAGGIO .....</b>	<b>1</b>
4.1 - LE SCALE DI MISURA .....	1
Variabili quantitative (o continue) .....	2
Variabili categoriali (o qualitative) .....	2
Variabili categoriali ordinali .....	2
Variabili categoriali nominali .....	3
4-2 STANDARDIZZAZIONE DI VARIABILI CONTINUE.....	4
Esempio 1 - Indicatori di tendenza centrale e dispersione .....	6
Esempio 2 - Variabili continue: normalizzazione .....	7
Esempio 3 - Misura dell'associazione tra variabili continue.....	9
4-3 TIPI DI TAVOLE.....	12
Tavola di descrizione .....	12

Tavola di contingenza (o di conteggio).....	12
4-4 LA RAPPRESENTAZIONE GEOMETRICA.....	13
La distanza.....	14
Il centro di gravità della nuvola.....	15
L'inerzia della nuvola.....	15
Interpretazione delle relazioni tra le variabili in $\mathbb{R}^n$ .....	16
<b>CAP. 5 - MENU DI ANALISI: DISTRIBUZIONI ED INCROCI.....</b>	<b>1</b>
5.1 - DISTRIB.....	2
La struttura del file DISTRIB.PAR.....	2
5.2 - CROSSTAB.....	6
Un esempio, giusto per introdurre qualche considerazione teorica.....	8
Costruzione della tavola attesa nell'ipotesi di indipendenza.....	10
Un indicatore dello scostamento globale tra tavola osservata ed attesa.....	10
La struttura del file CROSS.PAR.....	12
<b>CAP. 6. - MENU DI ANALISI: LE ANALISI FATTORIALI.....</b>	<b>1</b>
6.1 - TYPOLOG.....	1
I parametri richiesti per l'esecuzione.....	3
6-2 LE ANALISI FATTORIALI: ACOMP ED ACORR.....	10
6-2.1 L'ANALISI IN COMPONENTI PRINCIPALI (ACOMP).....	11
Un esempio.....	13
L'inserimento dei parametri di controllo dell'analisi.....	14
A. Parametri che controllano il caricamento dei dati dal file di input.....	14
B. Parametri che controllano l'output.....	20
La tavola dei contributi e la loro interpretazione.....	22
Interpretazione dei fattori.....	25
6-2.2. - L'ANALISI DELLE CORRISPONDENZE.....	28
L'inserimento dei parametri di controllo dell'analisi.....	31
Le tavole dei contributi e la loro interpretazione.....	32
<b>CAP. 7. - MENU DI ANALISI: LE CLASSIFICAZIONI.....</b>	<b>1</b>
7-1 ALCUNE NOTE SULLA CLASSIFICAZIONE NUMERICA.....	1
I metodi gerarchici.....	2
I metodi non gerarchici.....	3
Alcune definizioni.....	3
Il metodo delle nubi dinamiche.....	4
7-2 - I METODI DI CLASSIFICAZIONE IN ADDATI.....	6
7-2.1 - Alcune note generali.....	6
Il metodo di classificazione non gerarchica.....	6
7-2.2 - Classificazione non gerarchica: il Metodo 1 (non più implementato in ADDATI).....	6
Metodo 1 - La fase esplorativa.....	7
La fase di Ottimizzazione.....	8
7-2.3 - Classificazione non gerarchica: il Metodo 2 attualmente utilizzato.....	8
7-3 - IL PROGRAMMA NONGER.....	11
Fase esplorativa.....	11
A. Domande relative alla lettura del file dei dati.....	11
<b>Blocco A1 (per una classificazione eseguita direttamente sulle variabili descrittive)</b> .....	12
<b>Blocco A2</b> (solo per una classificazione che segua un'analisi fattoriale).....	17
<b>Blocco B.</b> - Parametri per la classificazione.....	17
NONGER - Fase di ottimizzazione e descrizione delle partizioni ottenute.....	19
NONGER - L'interpretazione dei risultati.....	23
<b>APPENDICE: GLI ESEMPLI.....</b>	<b>1</b>
VENEZIA.....	1
DEGRADO.....	2

La precedente versione 3.1 di ADDATI era stata distribuita a Dipartimenti Universitari, Istituti di Ricerca o Pubbliche Amministrazioni dal Centro Interdipartimentale di Documentazione e Calcolo (CIDOC) della nostra Università.

Quella versione era stata scritta quando ancora imperavano DOS e Windows 3.1, e la CPU più veloce usata nei Personal Computer era il 40486. Molto è cambiato da allora: la normale configurazione in commercio usa oramai un Pentium 4 a 2GHz, e come Sistema Operativo Windows 2000 o XP (presto, immagino, le cose si evolveranno ulteriormente...sempre costretti a rincorrere Bill Gates ed i suoi tentativi di far soldi cambiando cose che funzionano benissimo come sono...). Comunque, modi di lavorare impensabili anche solo pochi anni fa, quando usciva la versione 3.1, sono ora abituali.

Con la versione 4.2, del 1997, non è sembrato più sufficiente rendere il pacchetto disponibile in rete sui server dell'Università, come si faceva prima. I PC cominciavano ormai ad essere molto diffusi, e sempre più numerosi erano gli studenti che chiedevano di potere disporre del pacchetto nel computer di casa, per svolgere in modo più comodo ed efficiente le esercitazioni loro richieste.

Più che nel pacchetto, in realtà il problema stava nella Guida all'Uso che lo accompagnava: circa 200 pagine che offrivano, seppur in forma volutamente semplificata, una quantità di nozioni teoriche e di consigli metodologici utili - forse indispensabili - per svolgere delle analisi corrette.

Si era allora deciso di rivedere il Manuale e di ristamparlo con le Edizioni Libreria Progetto di Padova, allegando ad esso il dischetto con i programmi. Poi varie altre modifiche sono state inserite nel pacchetto, che nelle sue successive versioni andava diventando sempre più diverso da quello che il Manuale raccontava. È sembrata allora opportuna un'opera di revisione radicale.

L'ultima versione cui questo Manuale si riferisce - la 5.2a del Luglio 2003 - è una versione mista Win32/DOS, e si sta andando ormai velocemente verso una pura versione Windows 32.

Nelle sue applicazioni più frequenti la 5.2a è molto sperimentata e dovrebbe essere esente da bachi maggiori. Comunque, saremo particolarmente riconoscenti a coloro che vorranno segnalarci errori di funzionamento con i dettagli necessari perché possiamo riprodurli, inviando un e-mail con il file dei dati zippato ed i valori dei parametri di controllo che hanno causato il problema all'indirizzo [silvio@cidoc.iuav.it](mailto:silvio@cidoc.iuav.it).

La versione aggiornata del pacchetto è sempre disponibile alla pagina Web <http://cidoc.iuav.it/~addati/addati.html>

Venezia, Luglio 2003

S. Griguolo

## **Dove trovare i dati per le analisi?**

---

Per quanto riguarda i dati censuari c'è **SINTESI** (Sistema INTerattivo per l'EStrazione delle Informazioni).

SINTESI (messo a punto da M. Mazzanti) è un'interfaccia grafica che consente un facile accesso a banche dati numeriche da parte degli utenti con accesso a Internet. Con poche e semplici operazioni è possibile selezionare i dati di interesse, salvarli su file in formato ASCII ed effettuare elaborazioni ed analisi con i programmi preferiti (i dati sono estratti in un formato immediatamente elaborabile da ADDATI).

E' anche possibile costruire interattivamente carte tematiche.

Sono attualmente disponibili i seguenti dati di fonte ISTAT:

- Censimento Popolazione e Abitazioni 1991
- Censimento Popolazione e Abitazioni 1981
- Censimento Agricoltura 1990
- Censimento Industria e Servizi 1991 (addetti Unità Locali)
- Censimento Industria e Servizi 1991 (Unità Locali)
- Censimento Industria e Servizi 1991 (addetti delle imprese)
- Censimento Industria e Servizi 1991 (imprese)
- Censimento Industria e Servizi 1991 (addetti delle Unità Locali artigiane)
- Censimento Industria e Servizi 1991 (Unità Locali artigiane)
- Censimento Industria e Servizi 1981 (addetti delle Unità Locali)
- Censimento Industria e Servizi 1981 (Unità Locali)
- Censimento Industria e Servizi 1981 (addetti delle imprese)
- Censimento Industria e Servizi 1981 (imprese)

I dati relativi ai Censimenti 2001 verranno aggiunti appena disponibili. I dati sono su base comunale, aggregati per regione (20 Banche Dati per ciascun Censimento).

L'indirizzo Internet di SINTESI è il seguente:

<http://cidoc.iuav.it/sintesi/index.html>

## Una breve descrizione della versione 5.2a

---

Si tratta ancora di una versione provvisoria Win32/DOS, ma un altro passo è stato compiuto verso un pacchetto puramente Windows 32: Il programma FACPLAN, che visualizza le proiezioni delle unità statistiche e delle variabili sui piani fattoriali, è stata riscritta come applicazione Win32.

Prima FACPLAN era un programma grafico DOS, dunque suscettibile di creare problemi girando in ambiente Windows, con chissà quale scheda video.

Gli altri programmi sono per lo più DOS a 32 bit, cavalli da tiro computazionali con uscita testuale, compatibili con tutte le versioni di Windows 32, da 95 a XP. Sono veloci e capaci di usare appieno la memoria centrale disponibile: non dovrebbero creare alcun problema.

Per installare il pacchetto, estrarre in una cartella temporanea gli archivi contenuti nel file di installazione <ADDA52aINST.ZIP>, poi lanciare **SETUP**.

**SETUP** è un'applicazione Win32 che installa ADDATI in Inglese o in Italiano, a scelta dell'utente. Il pacchetto è strutturato in modo da adattarsi facilmente ad altre lingue: la prossima potrebbe essere lo Spagnolo.

Installato il pacchetto, la configurazione e la lingua si possono cambiare in qualsiasi momento usando l'opzione "**configura**" dal menu FILE della routine di gestione ADDATI.EXE, che è anch'essa un'applicazione Windows32.

La versione 5.2a usa come DOS Extender WDOSX di Michael Tippach, compatibile sia con i Service Packs di Win2000 che con XP.

ADDATI dovrebbe dunque risultare pienamente compatibile con Windows 95/98/NT/2000. Qualche problema potrebbe presentarsi con XP, specialmente sui portatili, causato dalla visualizzazione del grafo della funzione obiettivo nel programma di Classificazione non gerarchica NONGER.

Il prossimo passo, molto presto, sarà la piena conversione a Windows 32.

Altre opzioni minori sono state aggiunte recentemente; ad esempio, una opzione nel Menu FILE volta a convertire file di dati che usano la tabulazione come separatore, salvati in EXCEL, in altri che usano come separatore lo spazio, come richiesto da ADDATI. La conversione è realizzata in modo controllato, in modo da permettere all'utente di eliminare facilmente tutti gli elementi superflui dal file dei dati, le cui variabili vengono poi incolonnate.

## Un po' di storia...

---

ADDATI ha alle spalle una storia di quindici anni, quando ci si riferisca ai primi esercizi di analisi territoriale dai quali il pacchetto ha iniziato a prendere forma.

Lo si può far risalire alla collaborazione di Silvio Griguolo con Piercarlo Palermo verso la fine degli anni '70 ed i primi anni '80 presso il DAEST (il Dipartimento di Analisi Economica e Sociale del Territorio dell'Istituto di Architettura di Venezia). L'apporto di Palermo è stato determinante oltre che nella formulazione della proposta metodologica anche nell'ispirazione alla stesura dei primi programmi di calcolo dai quali è iniziata la lunga evoluzione che ha portato alla versione attuale.

L'idea era di utilizzare le capacità esplorative e costruttive dell'Analyse des Données francese per affrontare alcuni temi di Analisi Territoriale nei quali risulta importante l'individuazione dei comportamenti più rilevanti che contraddistinguono, in un particolare

contesto, un insieme di unità statistiche. Si tratta spesso di analisi condotte al massimo livello di disaggregazione: ne è un esempio tipico l'analisi della condizione abitativa condotta sui fogli di famiglia del Censimento.

Si cominciò utilizzando i pacchetti francesi ADDAD e SPAD dapprima adattandoli, poi riscrivendone alcune routines, per arrivare infine a stendere dei programmi e formulare delle sequenze di analisi originali. A questa prima fase ha contribuito Luciano Vettoreto. Il linguaggio di programmazione era ancora il FORTRAN, ottimo per il calcolo scientifico ma piuttosto pesante nell'uso e con livelli primitivi di interattività (almeno allora): programmi di quei tempi, atti a girare in batch su macchine remote ed utilizzabili da altri con difficoltà (ma ci si adattava...).

Un pesante lavoro di sistemazione ed arricchimento ne trasse la prima versione del pacchetto (allora chiamato *ADDAEST*) che venne utilizzato, oltre che all'IUAV (ricordo *L'Atlante Tematico dei Comuni d'Italia* di Paolo Santacroce) da alcuni Enti Locali o di Ricerca.

A questo punto cominciarono ad arrivare sul mercato personal computers sempre più veloci e potenti e si affermarono linguaggi di programmazione che consentivano un controllo della macchina a livello più fine.

L'ADDATI attuale - scritto in C - sfrutta queste possibilità. Molto amichevole e di facile uso, può essere utilizzato in ogni ambito disciplinare sia per esercitazioni didattiche che per la ricerca.

E' interattivo, ed è stato arricchito con una serie di programmi originali di utilità, di grande aiuto per ricodificare variabili esistenti o per costruirne di nuove, per estrarre dagli archivi di partenza le variabili necessarie alla preparazione della tavola dei dati da analizzare ed anche per altre cose.

ADDATI non è certo un pacchetto statistico onnicomprensivo: implementa un numero limitato di cose ma lo fa - almeno così ci sembra - in modo molto amichevole ed efficiente.

Risulta particolarmente adatto per la trattazione di grandi basi di dati, anche a livello di massima disaggregazione. Basti pensare che ADDATI è stato utilizzato (da S.Griguolo, insieme a F.Gosen e D.Patassini) su di un PC 486 per analizzare la condizione abitativa di Addis Abeba a partire dai circa 238.000 fogli di famiglia del Censimento trattati *in forma disaggregata*.

Per capire con sufficiente profondità le tecniche statistiche usate, così come la metodologia di analisi ed i risultati delle ricerche empiriche condotte con ADDATI è consigliabile far ricorso a qualche testo specifico.

Sembra opportuno un particolare riferimento al volume

S. Griguolo, P.C. Palermo (a cura): *Nuovi Problemi e nuovi Metodi di Analisi Territoriale*, Angeli, Milano, 1984

ed alla estesa bibliografia ivi inclusa. Qualche testo specifico di Analisi dei Dati :

L.Lébart, A.Morineau, N.Tabard, *Téchniques de la description statistique*, Paris, Dunod, 1977

M.Jambu, *Classification automatique: méthodes et algorithmes*, Paris, Dunod, 1978

M.Jambu, *Exploration informatique et statistique des données*, Paris, Dunod, 1989



## Attribuzioni e riconoscimenti

---

- S.Griguolo ha concepito e progettato il pacchetto, ha scritto le routines per le finestre di testo ed è l'autore dei diversi programmi, con l'eccezione di **FACPLAN** e **RECODE**.
- Ciavarella (con contributi da parte di S.Griguolo) è l'autore della versione Win32 di **FACPLAN** (proiezioni su piani fattoriali).
- M.Mazzanti era il principale autore della precedente versione DOS del programma.
- **RECODE** è stato concepito congiuntamente da S. Griguolo e M. Mazzanti, e scritto quasi interamente da M.Mazzanti.
- Il presente manuale, scritto con Office 97, elabora quello preparato nel 1997 insieme a M. Mazzanti inserendo le necessarie informazioni sulle funzionalità nel frattempo aggiunte al pacchetto. Viene mantenuta l'impostazione tipografica, che deriva da quella di un precedente Manuale scritto da S.Griguolo per la versione inglese di ADDATI utilizzata dal progetto FAO/GCPS/RAF/256/ITA (IGADD Early Warning and Food Information System for Food Security, Djibouti). Tale impostazione tipografica, realizzata allora con Ventura 2.0, era opera di P.Santacroce ed A.Conte, che non manchiamo di ringraziare.

## L'impostazione di questa Guida all'uso

---

Il Capitolo 1 descrive il procedimento di installazione del pacchetto e le opzioni incluse nel *Menu FILE* di ADDATI.

Il Capitolo 2 descrive i programmi inclusi nel *Menu di Utilità*, da utilizzare nella preparazione della tavola da analizzare, o per facilitare l'interpretazione dei risultati.

Il Capitolo 3 descrive gli aspetti più rilevanti dell'interfaccia utente comune a tutti i programmi inclusi nel *Menu di Analisi*.

Il Capitolo 4, dal titolo "*Fondamenti di teoria e linguaggio*" fornisce brevemente alcuni elementi necessari per comprendere come operino i programmi di analisi e per interpretare i risultati. Se ne consiglia la lettura agli utenti ancora inesperti di Analisi dei Dati, e la consultazione per tutti quando sia necessario.

Il Capitolo 5, aggiunto con la versione 4.0 ed aggiornato in seguito, illustra i programmi **DISTRIB** (Distribuzioni) e **CROSSTAB** (Incroci). Vi vengono anche introdotte, in forma elementare, alcune nozioni teoriche necessarie per condurre una corretta interpretazione dei risultati.

Il Capitolo 6 entra nel vivo delle procedure di Analisi e riguarda la costruzione di Tipologie e le Analisi Fattoriali. Nell'espone l'Analisi in Componenti Principali si coglie l'occasione per fornire alcuni rudimenti di teoria, che vanno tuttavia puntualmente approfonditi su testi specifici. Viene sviluppata come esempio un'analisi sulle sezioni censuarie del Centro Storico di Venezia che l'utente può ripetere, introducendo le varianti che gli sembreranno opportune, utilizzando i files inclusi nella subdirectory <ESEMPI> inclusa nella cartella di installazione di ADDATI.

Il Capitolo 7 introduce sommariamente le tecniche di Classificazione numerica, descrivendo i due percorsi integrati di classificazione non gerarchica offerti da ADDATI e l'uso dei programmi relativi.

### L'autore è raggiungibile al seguente indirizzo:

Silvio Griguolo - Dipartimento di Pianificazione, S.Croce 1957, 30135 VENEZIA

telefono: 041-2572110 - fax : 041 52 08 945

e-mail : [silvio@cidoc.iuav.it](mailto:silvio@cidoc.iuav.it)

# Cap 1. - Installazione e Generalità

## 1.1 – Installazione

Estrarre in una cartella vuota i file contenuti nel file di installazione ADDA52aINST.ZIP. Si ottengono cinque file, tre dei quali necessari per installare il pacchetto.

- **SETUP.EXE**, è usato per controllare il processo d'installazione;
- **ADDATI.ZIP**, è il file zippato che contiene i moduli di ADDATI. Non va espanso direttamente, non ne risulterebbe un pacchetto funzionante. Si lasci che il lavoro venga fatto da SETUP.
- **UNZIP32.EXE**, è un'utilità di decompressione di pubblico dominio.

Durante l'installazione le versioni del Manuale nelle diverse lingue nelle quali è possibile l'installazione (attualmente solo Inglese ed Italiano) vengono copiate nella subdirectory <DOC> della cartella dove ADDATI viene installato. La versione in italiano installa anche alcuni esempi, contenuti nel file <EXMP\_ITA.ZIP>.

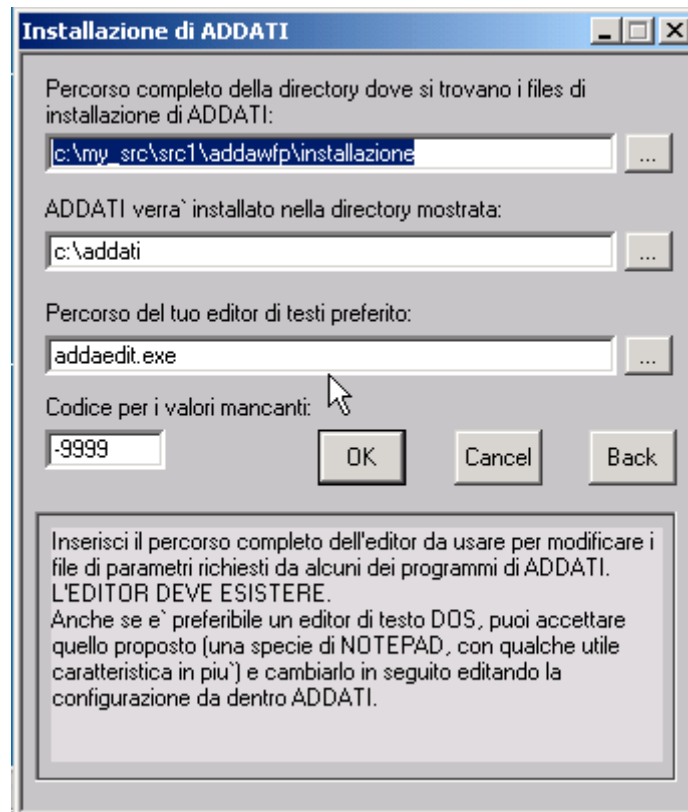
Lanciando **SETUP** appare il dialogo di fig. 1-1. La sua struttura è simile a quella degli altri dialoghi che verranno presentati in ADDATI: nella piccola finestra in basso viene mostrato un aiuto contestuale, che appare quando il mouse viene passato sopra un controllo (un bottone, un combo, un edit box, ecc.). Scegli la lingua di installazione e clicca su OK.

*Nota: La lingua può venire cambiata in qualunque momento durante l'uso del pacchetto: tutte le stringhe alfanumeriche, gli help, ecc. vengono immediatamente presentati nella nuova lingua.*



**Figura 1-1** Installazione – Il dialogo relativo alla scelta della lingua

Appare poi il dialogo di fig. 1-2. Il primo edit box mostra il nome della directory dalla quale **SETUP** è stato lanciato, ed ha solo uno scopo informativo.



**Figura 1-2** Installazione: il dialogo per la scelta dei valori di inizializzazione

Il secondo edit box propone di installare ADDATI in C:\ADDATI. La directory destinazione può essere cambiata a piacere, ma bisogna ricordare che i nomi delle directory dovrebbero rispettare il formato 8.3 del DOS, altrimenti alcuni programmi DOS di ADDATI potrebbero avere problemi. **Dunque, per favore, niente nomi lunghi né con spazi all'interno.**

Il terzo controllo richiede l'inserimento del percorso del programma di editing che ADDATI lancerà quando si dovrà editare un file di parametri, o in genere qualche file di testo. Si possono scegliere NOTEPAD o WORDPAD, ma ADDATI propone "addaedit.exe", il suo editor interno creato apposta per questo scopo, e con il quale ADDATI comunica in modo particolarmente utile. Si consiglia assolutamente di accettare l'offerta.

Il quarto controllo fissa il codice che indica un valore mancante. Il valore proposto può essere sempre modificato quando risulti necessario.

È importante che ADDATI possa distinguere un valore mancante da uno valido. Il codice, **che deve essere numerico**, viene utilizzato da alcuni programmi di utilità (ad esempio MERGFIELD o RECODE), oppure nel calcolo di Distribuzioni o Incroci.

**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 e tratterebbero dunque il codice di valore mancante come un normale dato numerico con risultati indesiderati. Il controllo sui valori mancanti verrà reso più severo nella versione Win32 in preparazione.*

I valori inseriti vengono registrati nel file di inizializzazione ADDATI.INI, scritto nella cartella di ADDATI. Essi possono venire modificati in qualsiasi momento editando ADDATI.INI, che è un file di testo, o scegliendo l'opzione “**configure**” dal Menu FILE di ADDATI.

Terminata l'installazione si possono cancellare i file estratti da ADDA52AINST.ZIP. Conviene poi creare sul desktop un collegamento al programma di gestione ADDATI.EXE.

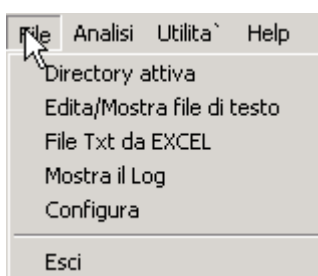
**Nota:** *L'icona è un po' strana ed ha una storia curiosa. Ho detto scherzando a mia figlia, che ha 10 anni: “Perché non mi fai un'icona per ADDATI?”. E lei di rimando: “Ma cosa fa ADDATI?”. Già, cosa fa ADDATI?*

*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...*

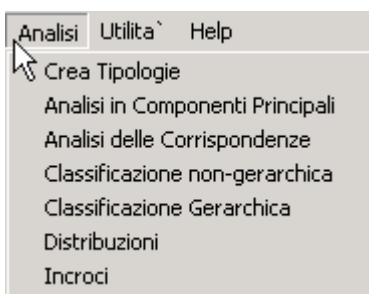
*Buon lavoro!!*

## 1.2 – Il programma di gestione ADDATI.EXE

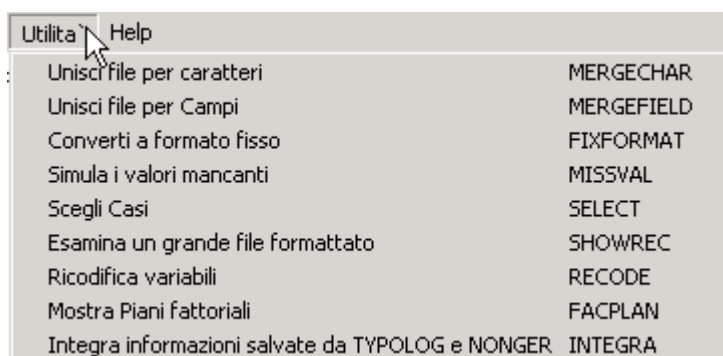
Il Menu di ADDATI offre i tre sottomenu mostrati nella figura 1-3.



a) Il menu FILE



b) Il menu di Analisi



c) il menu di Utilità

**Figura 1-3** I tre Menu del programma di gestione ADDATI.EXE

Qui verrà descritto brevemente il menu *FILE*. Il Cap. 2 si occupa del menu di Utilità ed i capitoli seguenti sono dedicati alle opzioni incluse nel menu di Analisi.

## 1. Directory attiva

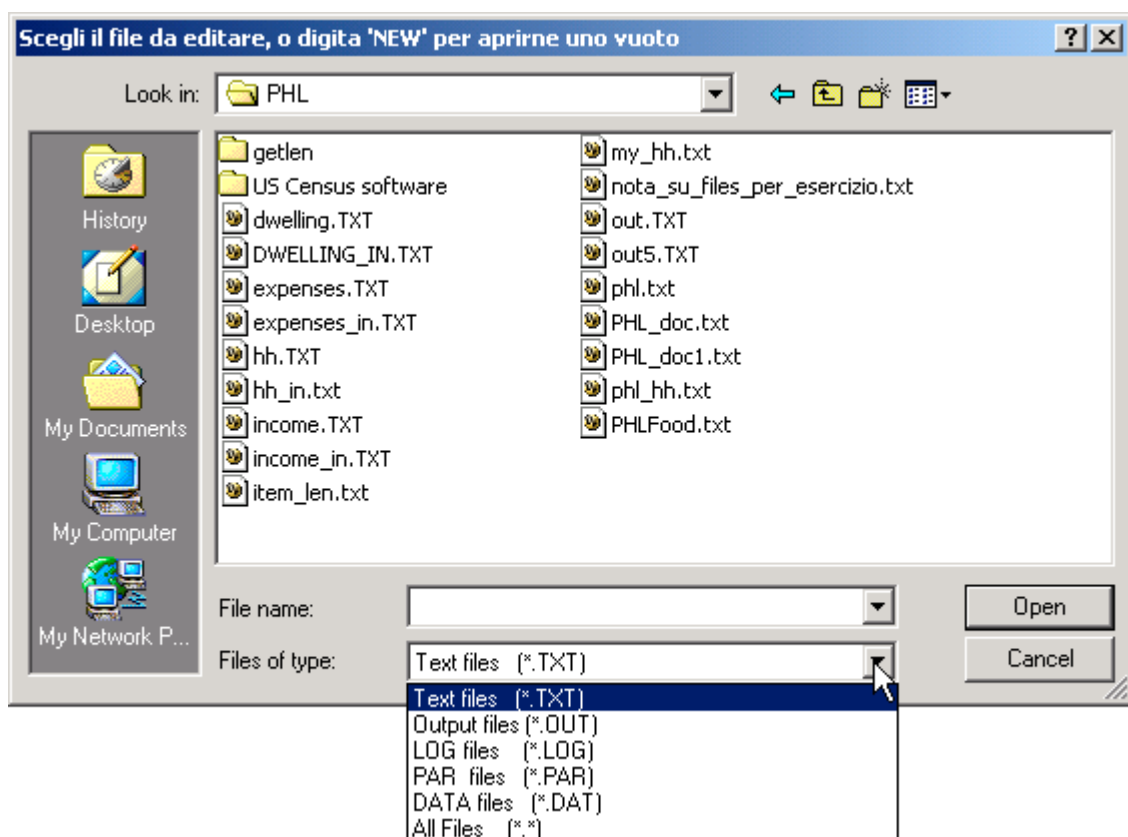
Scegliere la directory attiva è la prima operazione da fare immediatamente dopo aver lanciato ADDATI. Quando si parte, la directory corrente è quella di installazione del pacchetto: è più conveniente scegliere come attiva quella che contiene i dati sui quali si deve lavorare.

Questa opzione apre un dialogo che invita a scegliere la directory di lavoro. La scelta, una volta operata, viene confermata da un messaggio che appare nella finestra principale di ADDATI. Se ora viene lanciata una delle applicazioni DOS di ADDATI, essa eredita automaticamente la directory di lavoro.

Se un'applicazione DOS non trova il file che si vuole aprire, probabilmente questo passo è stato omesso, e la directory di lavoro non è quella voluta. È possibile accertarsene dall'interno dell'applicazione DOS premendo F10 per aprire una nuova shell: la directory attiva è specificata nel prompt.

## 2. Edita/Mostra file di testo

Questa voce evoca un dialogo per la scelta del file di testo da editare (fig. 1-4).



**Figura 1-4** Scelta del file di testo da editare

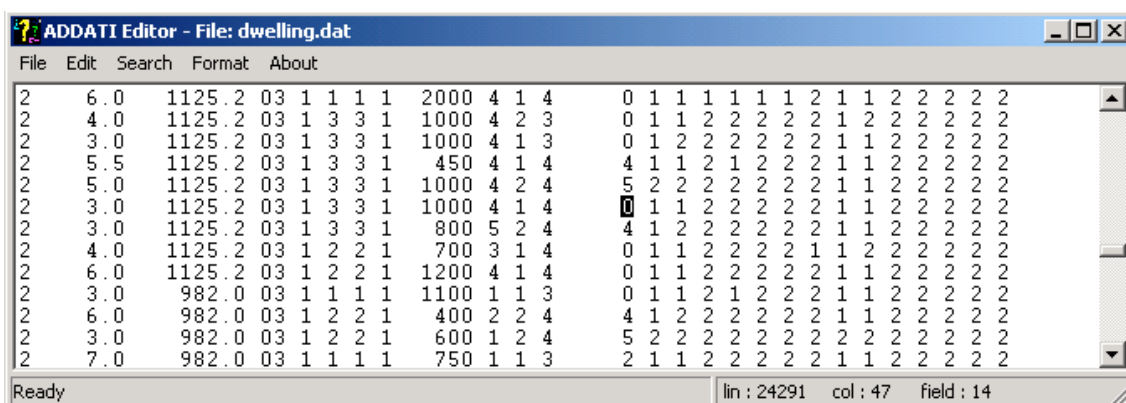
I file di testo solitamente usati da ADDATI hanno le estensioni mostrate in figura sotto "Files of type": **TXT** (file di testo generici, o documentazioni di file di dati), **OUT** (file prodotti dai programmi di analisi) **LOG** (file LOG di ADDATI), **PAR** (file di parametri, che l'utente riempie appropriatamente per controllare l'esecuzione di alcuni programmi);

**DAT** (estensione usuale per i file di dati). Selezionare l'estensione per visualizzare tutti i file del tipo di quello che si vuole aprire, 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 (a meno che non sia stato indicato un editor diverso durante l'installazione del pacchetto). **ADDAEDIT** ha alcune caratteristiche che ne rendono vantaggioso l'uso: 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* (o variabile: si parlerà dei *campi* più avanti) sopra il quale sta il cursore. I file di dati vengono automaticamente riconosciuti come tali se hanno l'estensione '.DAT'. Se ciò non è, si può forzare la visualizzazione del numero del campo scegliendo la voce "Show Field #" nel menu FILE dell'editor.



**Figura 1-5** La finestra dell'editor interno **ADDAEDIT**.

La figura 1-5 mostra la finestra dell'editor: il cursore è localizzato al record 24291, colonna 47, sul valore della variabile n.14, che è 0. Si può aprire un'altra copia (o *istanza*) dell'editor e visualizzare la documentazione (o *dizionario*) che accompagna il file dei dati, ispezionandone così il contenuto conoscendo il significato dei valori mostrati.

Si possono aprire simultaneamente tante istanze dell'editor quante siano necessarie. Si tratta di applicazioni indipendenti, ed ADDATI può essere simultaneamente usato senza restrizioni..

**ADDAEDIT** offre alcuni altri vantaggi, legati alla sua seppur limitata capacità di dialogare con ADDATI. Ma ne parleremo più avanti.

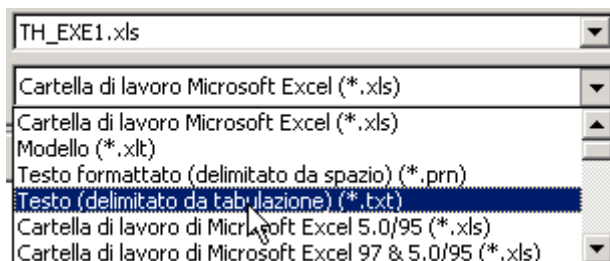
### 3. File txt da EXCEL

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

ADDATI elabora file di dati in formato testo, che devono contenere esclusivamente dati. Per tale ragione, i file di EXCEL devono essere convertiti in formato testo, eliminando ogni cosa diversa dai dati (ad esempio, intestazioni di colonna od altro).

Il foglio elettronico di EXCEL che contiene i dati va salvato come un **file di testo che usa la tabulazione come separatore**. Si tratta di una delle opzioni offerte da EXCEL quando si sceglie **"Save as"** (vedi la fig. 1-6). EXCEL offre anche la possibilità di esportare file di testo che usano come separatore lo spazio (l'opzione precedente nella fig. 1-6), il che è in effetti 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, l'inconveniente non sussiste.

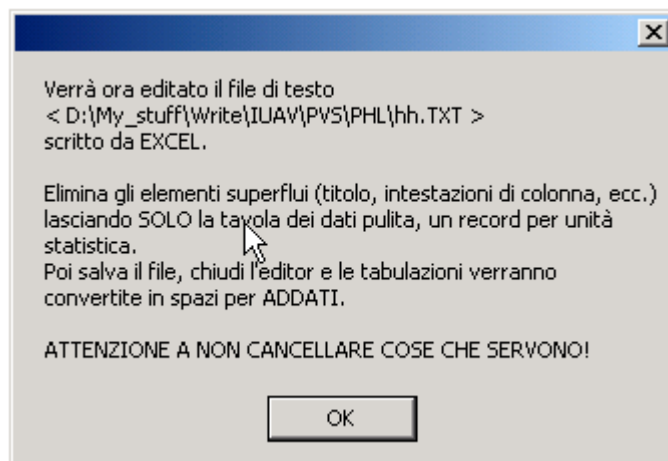


**Figura 1-6** L'opzione EXCEL per salvare file di testo con il tab come separatore di campo.

**Una volta esportati in questa forma i dati da EXCEL** (ad esempio con estensione .TXT, anche se la cosa non è strettamente necessaria), si scelga l'opzione "**File di testo da EXCEL**" dal menu FILE di ADDATI. Si apre il dialogo di fig. 1-7: si inserisca nell'edit box superiore il nome del file salvato da EXCEL, o si sfogli per puntarlo. ADDATI propone allora come output un file con lo stesso nome, ma con l'estensione cambiata in .DAT, nel quale le tabulazioni saranno convertite in spazi.



**Figura 1-7** Importazione di dati da EXCEL, il dialogo iniziale  
Premendo OK appare il messaggio seguente:



Si preme OK. ADDATI carica il file scritto da EXCEL nell'editor per permettere all'utente di eliminare il superfluo (documentazione, etichette, intestazioni, ecc.) **lasciando solamente i dati**. Si salva poi il file e si chiude 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 gli spazi che trova ***all'interno di un campo*** con underscore, poi ordina i valori per colonne per permetterne un facile esame. Tutti gli errori trovati durante l'operazione (variabili che mancano, record con un numero di campi diverso da quello che ci aspetta, ecc.) vengono segnalati.

Prestare 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 esso venga effettivamente terminato con la coppia di caratteri (ASCII 13 e 10) che rappresentano il fine record in DOS. Attenzione anche a non inserire qualche spazio nella nuova linea, altrimenti essa verrà interpretata da ADDATI come un record senza campi, il che può generare un errore.

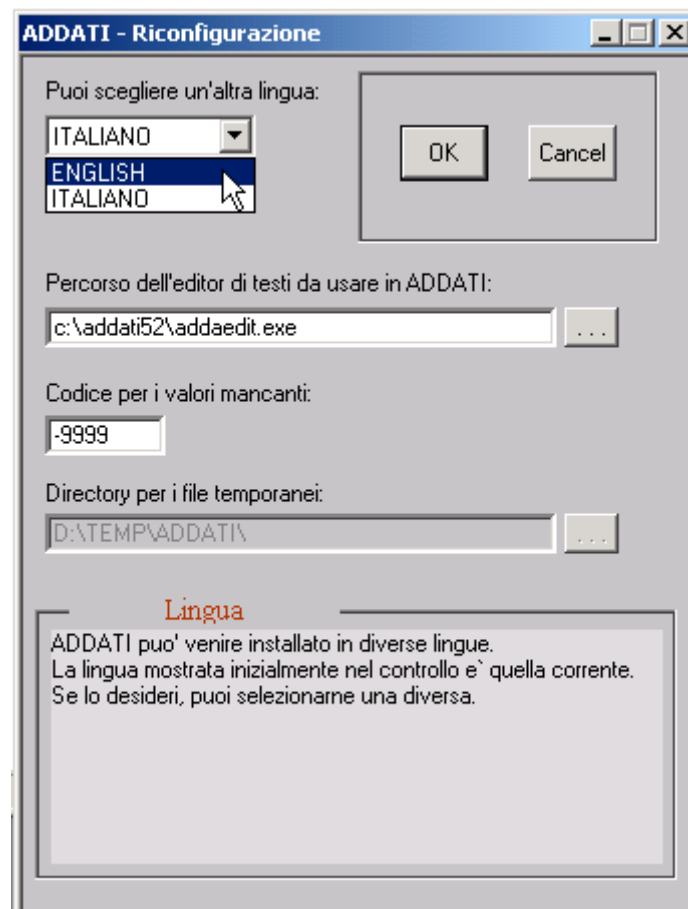
#### 4. Mostra il LOG

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

#### 5. Configura

L'opzione è utilizzata per cambiare i settaggi di inizializzazione. Viene mostrato il dialogo di fig.1-8: l'utente può cambiare la lingua, l'editor preferito ed il codice per i valori mancanti. Il cambiamento ha effetto immediato.





**Figura 1-8** Il dialogo di riconfigurazione.

## Cap 2. - Il Menu di Utilità

---

Questo menu include alcuni programmi di utilità di interesse generale che possono aiutare chi lavora con ADDATI a preparare la tavola da analizzare e ad interpretare i risultati. Tutti i programmi (per fondere, formattare, esaminare file o ricodificare variabili) operano su **file di testo**. Se i dati sono registrati in formato binario (ad esempio da applicazioni come DBASE o EXCEL) essi vanno "*esportati*" come file di testo prima che ADDATI li possa trattare.

Il Menu FILE offre un'opzione ("File txt da EXCEL") che facilita l'importazione di file di dati da EXCEL.

***Nota:** MERGCHAR, MERGFLD, FIXFORM, SELECT e RECODE vengono utilizzati in ADDATI per la manipolazione di files di dati allo scopo di preparare la tavola da sottoporre ad analisi, ma la loro utilità va oltre quest'uso.*

### Il formato di registrazione dei file di dati

Un file di dati (**di tipo testo**) può essere registrato secondo diversi tipi di **formato**. Intendiamo con ciò indicare il modo in cui sono accostati i valori (variabili) che costituiscono i suoi records. Poiché un programma deve conoscere il tipo di formato per poter caricare correttamente i dati, è importante precisare la convenzione di linguaggio che seguiremo in ADDATI.

Un record contiene la sequenza dei valori relativi alle variabili considerate. Per evidenti motivi l'ordine delle variabili deve essere il medesimo in tutti i records. Il file dei dati andrebbe sempre accompagnato da un *tracciato* o *documentazione* che informa l'utente sulla posizione occupata dalle diverse variabili.

In un record i valori possono essere registrati in uno dei due modi seguenti.

1. Senza spazi di separazione: il record si presenta nella forma "01212341234...." e solo la posizione permette di riconoscere il significato delle cifre lette (cioè, a quale variabile si riferiscano). Pertanto, tutti i record debbono avere la medesima lunghezza (lunghezza fissa) ed una variabile occuperà in tutti i record le medesime colonne. Si tratta del formato usato spesso quando il numero dei casi e delle variabili sia molto elevato, allo scopo di contenere la dimensione del file risultante (è il caso tipico dei dati censuari disaggregati o dei dati d'inchiesta).
2. Separandoli con degli spazi per una loro riconoscibilità immediata. Chiameremo nel seguito "campo" un gruppo di caratteri alfanumerici separato da altri gruppi simili mediante uno o più spazi. Un record strutturato in campi può avere o meno una lunghezza fissa. Se ce l'ha, pur mantenendo gli spazi di separazione viene assegnato ad ogni campo (variabile) un dato numero di colonne (caratteri), cioè una lunghezza di campo, in genere diversa da campo a campo. Le variabili risultano allora incolonnate e l'ispezione del file mediante visualizzazione risulta agevole.

La tabella 2-2 mostra le combinazioni possibili. Diremo convenzionalmente che un file è in **formato libero** quando i suoi record sono strutturati per campi (i record possono avere o meno lunghezza costante); lo diremo **formattato** (o anche in **formato fisso**) quando i

suoi record hanno *lunghezza* costante e ciascuna variabile occupa le medesime colonne in tutti i record.

Così definite, le due dizioni *non sono mutuamente esclusive*: la prima si riferisce alla struttura per campi, la seconda alla costanza della lunghezza del record. Un file in formato libero può essere *anche* formattato ed allora tutte le variabili che contiene risultano incolonnate; viceversa, un file *non strutturato per campi*, cioè non in formato libero, deve *necessariamente* avere i record di lunghezza fissa.

		SPAZIO COME SEPARATORE	
		Si	No
LUNGHEZZA DEL RECORD	Fissa	Formato libero formattato	Formato non libero, formattato
	Variabile	Formato libero, non formattato	Non prevista

**Tabella 2-1** Schema della terminologia convenzionale adottata per il formato.

### *I programmi di utilità*

Per ragioni di chiarezza e maneggevolezza conviene spesso registrare in file diversi variabili che descrivano aspetti diversi dell'insieme delle unità considerate (aree geografiche, famiglie, ecc.). Ad esempio, se si tratta dei comuni di una regione un file può contenere alcune variabili socioeconomiche, un altro descrivere i caratteri del patrimonio abitativo, un altro ancora gli aspetti demografici e così via. Naturalmente, tutti i file debbono contenere lo stesso numero di record, pari al numero dei comuni considerati nel medesimo ordine.

Sorge sovente la necessità di accostare informazioni tratte da file diversi. E' questa la funzione di **MERGCHAR** e **MERGFIELD**, che lavorano rispettivamente *per caratteri* (cioè su file *formattati*) e *per campi* (cioè su file *in formato libero*). **MERGCHAR** può inserire degli spazi tra i caratteri scelti, scrivendo così un file d'uscita *in formato libero*; **MERGFIELD** permette anche la costruzione e la registrazione di nuove variabili.

**FIXFORM** opera su file *in formato libero non formattati* (cioè con lunghezza di record variabile) ed incolonna i campi riportando tutti i record ad una lunghezza fissa. I campi sono ancora separati da spazi.

**MISSVAL** simula dei valori plausibili da sostituire ai valori mancanti di variabili quantitative.

**SELECT** consente la selezione di tutte e sole quelle unità che possiedono i requisiti specificati dall'utente mediante una opportuna condizione logica.

**SHOWREC** consente l'esame di un file *formattato di tipo testo* (tipicamente, un file di dati) quando la sua lunghezza eccessiva - anche parecchi Megabytes - renda difficile o impossibile caricarlo con un editor.

**RECODE** consente di ricodificare i valori delle variabili contenute in un file scritto in formato fisso. Sono previste quattro operazioni di ricodifica, che operano su variabili quantitative e categoriali.

**FACPLAN** permette la visualizzazione dei piani fattoriali dopo un'Analisi Fattoriale o dopo una Classificazione susseguente ad un'Analisi Fattoriale. In questa versione **FACPLAN** è stato convertito in un'applicazione Win32 che offre una quantità di opzioni.

**INTEGRA** viene utilizzata come passo finale di una sequenza di classificazione iniziata con la determinazione di tipologie; scrive, per ciascuna unità elementare, la classe cui è stata assegnata dalla procedura di classificazione.

I programmi di utilità sono descritti nel seguito nello stesso ordine in cui appaiono nel menu.

## 2.1 - MERGCHAR

### Fusione (merging) di file per CARATTERI

<b>Funzione</b>	consente all'utente di leggere, unità per unità, le colonne di caratteri desiderate da un massimo di <b>nove file di input</b> e di scriverle sul file d'uscita nell'ordine voluto. E' possibile inserire degli spazi dove sia necessario e scrivere la medesima colonna più di una volta.
<b>Requisiti</b>	I file in input debbono essere <b>file di testo formattati</b> , cioè in ciascuno di essi ogni record deve contenere lo stesso numero di caratteri. Tutti i file considerati debbono avere lo stesso numero di record, ciascuno dei quali descrive un'unità statistica (aree geografiche, famiglie, alloggi, ecc.). Le unità debbono stare nello stesso ordine in tutti i file.
<b>Limiti</b>	Max. lunghezza di un record (sia in ingresso che in uscita): <b>5120 caratteri</b> .

Si pensi ad un file formattato come ad una tabella rettangolare le cui righe sono i record. Una "**colonna**" è formata da tutti i caratteri (come 'a', o ';' o uno spazio) che occupano **una data posizione** nei diversi record del file.

Si suppone che i file da fondere forniscano diverse descrizioni delle medesime unità statistiche, descrizioni che vengono archiviate separatamente per ragioni di chiarezza e maneggevolezza.

Viene chiesto all'utente di fornire i nomi dei file dai quali vanno tratte le colonne da combinare, separandoli con degli spazi. Il nome deve includere il percorso completo se il file non sta nella directory di lavoro. L'ordine in cui vengono forniti i nomi dei file d'ingresso è l'ordine in cui essi verranno letti; esso è importante in quanto determina la particolare lettera dell'alfabeto che contrassegna le colonne di ciascun file.

Ad esempio, una risposta come "**file1 file2**" è accettabile. Il programma controlla che i file esistano, poi chiede il nome da dare al file d'uscita.

**Nota:** *Se esiste già un file con lo stesso nome del file d'uscita, viene chiesto all'utente se voglia sovrascriverlo oppure fornire un nome diverso.*

Il programma legge il primo record di ciascuno dei file d'ingresso e per ciascuno di essi lista sullo schermo la lunghezza del record e la lettera associata.

L'utente deve scrivere in una finestra scrollabile una stringa che identifica le colonne da copiare. Tale stringa consiste di gruppi alfanumerici separati da spazi; ciascun gruppo, **che non deve includere spazi al proprio interno**, comanda la scrittura sul file d'uscita di un carattere o di più caratteri consecutivi, ovvero l'inserzione di uno spazio. Un gruppo accettabile ha una delle seguenti forme:

- un **identificatore di carattere** (come **A3** o **C17**), che consiste di una lettera seguita da un numero intero; la lettera denota il file dal quale il carattere è tratto (**A** per il primo file, **B** per il secondo, ecc.) mentre l'intero specifica la posizione del carattere nel record. Un gruppo di questo tipo comanda la registrazione nel file d'uscita del corrispondente carattere per ogni caso incontrato nei file di input;
- una **sequenza di caratteri** (tratti dal medesimo file) espressa in forma compatta per mezzo dell'operatore **"/**". Ad esempio, "**c3/7 a1/20**" copia in uscita, nell'ordine,

tutti i caratteri dal terzo al settimo (inclusi) dal terzo file ed i primi 20 caratteri del primo file.

- **la lettera "b"** , che comanda la scrittura di uno spazio sul file di uscita.

Il programma assume che tutti i record di un file abbiano il medesimo numero di caratteri (contengano cioè lo stesso tipo di informazione per le diverse unità, formattata in modo simile). Se ciò non è, il programma avverte della condizione d'errore durante l'esecuzione.

Una risposta come **"b1/9 b a14 a15 a9 b16 b17"**

significa che si vuol copiare sul file d'uscita, *per tutti i casi* contenuti nei file di input, le seguenti colonne (cioè caratteri) tratte da due file (indichiamoli convenzionalmente con file1 e file2): le prime 9 colonne di file2, poi le colonne 14, 15 e 9 di file1 seguite dalle colonne 16 e 17 di file2. Inoltre, per ciascun record del file d'uscita viene lasciato uno spazio dopo i primi nove caratteri.

**Nota:** *Tutti i file in input debbono avere lo stesso numero di record (debbono cioè contenere informazioni che riguardano le medesime unità, assunte nel medesimo ordine) pena una condizione di errore segnalata dal programma.*

## 2.2 - MERGFIELD

### Fusione (merging) di file per campi

<b>Funzione</b>	Come <b>MERGCHAR</b> , permette di unire informazioni tratte da file diversi ma lavora per campi invece che per caratteri. Consente anche di definire nuove variabili costruendone il valore a partire da quelle contenute nei file in input.
<b>File</b>	Tratta file di testo in formato libero.
<b>Requisiti</b>	Tutti i file da fondere debbono avere lo stesso numero di record; tutti i record di un file debbono avere il medesimo numero di campi.
<b>Limiti</b>	fino a nove file in input; max. 5120 caratteri/600 campi per record (sia nei file in entrata che in quello di uscita).
<b>Avvertenza</b>	<b>MERGFIELD</b> riconosce il codice per i valori mancanti (registrato nel file di configurazione ADDATI.INI). Operazioni impossibili, o che coinvolgano valori mancanti danno come risultato un valore mancante, registrato su output. <b>Se il programma mostrasse comportamenti anomali</b> conviene controllare se il codice corrente di dato mancante sia quello desiderato. <b>Se necessario, si può cambiarlo usando l'opzione <i>configura</i> dal Menu FILE.</b>

Si tratta di un altro programma di utilità per fondere file (merging), però orientato ai campi e non ai caratteri come **MERGCHAR**.

Si intende qui per **campo** un **gruppo di caratteri alfanumerici** delimitato da spazi.

**Esempio** *Il record " 12 ax2 32.45 " consiste di tre campi.*

Non è necessario che i file in lettura siano formattati come nel caso di **MERGCHAR** che richiedeva che tutti i record di un file avessero la stessa lunghezza. Qui devono però avere lo stesso numero di campi. E' possibile registrare sul file d'uscita lo stesso campo più di una volta.

Si assume che i file da unire contengano descrizioni diverse dello stesso insieme di oggetti ed abbiano dunque lo stesso numero di record. Ogni record descrive una unità statistica (individuo, abitazione, ecc.) e le unità debbono essere nel medesimo ordine in tutti i file.

Per ciascun caso presente nei file di entrata **MERGFLD** consente di:

- copiare i campi desiderati da un massimo di nove file, scrivendoli sul file d'uscita nell'ordine che si desidera;
- calcolare e registrare sul file d'uscita il valore di un numero qualsiasi di nuove variabili, definite dall'utente a partire dai valori delle variabili in ingresso;
- definire fino a cinque variabili ausiliarie i cui valori possono venire utilizzati nei calcoli e salvati in uscita a richiesta.

Come nel caso di **MERGCHAR** viene richiesto all'utente di digitare, separandoli con degli spazi, i nomi dei file dai quali vanno estratti i campi da unire. Va fornito anche un percorso per i file che non si trovino nella directory di lavoro. L'ordine in cui vengono forniti i nomi dei file è importante, perché determina la lettera di identificazione associata a ciascuno di essi.

Viene letto il primo record da ciascuno dei file in entrata; vengono *elencati sullo schermo i nomi dei file e, per ciascuno di essi, il numero dei campi e la lettera associata.*

In una opportuna finestra scrollabile l'utente digita una stringa alfanumerica atta a specificare - seguendo alcune regole convenzionali - i campi da copiare sul file d'uscita e le nuove variabili da creare. Essa consiste di un certo numero di *gruppi* separati da spazi.

Ciascun gruppo, **che non deve contenere spazi al proprio interno**, comanda la scrittura in uscita di uno o più termini, ovvero definisce una variabile ausiliaria. Un gruppo viene costruito combinando i seguenti elementi:

- **numeri** (1, -3.2, 300...), caricati come valori interi o reali;
- **identificatori di campo** (come **A3** o **C17**), consistenti di una lettera seguita da un intero; la lettera contrassegna il file dal quale il campo viene caricato (A per il primo file, B per il secondo, ecc.) mentre l'intero specifica la posizione del campo nel record;
- **i simboli v1 ... v5** che rappresentano - quando sia necessario - fino a cinque variabili ausiliarie;
- **gli operatori** " + - \* : ^ " che debbono trovarsi tra i numeri, gli identificatori di campo o le variabili ausiliarie sui quali operano. Si noti che per indicare una divisione viene usato ":" e non "/", che ha un diverso significato;
- **il segno** "=", che si deve usare **solo per definire una variabile ausiliaria**; ad esempio, "**v2=(2\*a1+3\*a2):5**" definisce una variabile ausiliaria v2 (da usare nel calcolo di altre variabili) come media pesata dei primi due campi del primo file (*a1* ed *a2*), rispettivamente con pesi 2 e 3 (5 è il peso totale). Si noti l'uso delle parentesi per stabilire le priorità da rispettare nel calcolo. Variabili ausiliarie già definite possono venire utilizzate nel calcolo di altre variabili ausiliarie definite successivamente;
- **i simboli di funzione** "ln log exp sqrt", che rappresentano nell'ordine il **logaritmo naturale**, il **logaritmo in base 10**, la **funzione esponenziale** e la **radice quadrata**.

I gruppi digitati dall'utente vengono interpretati nell'ordine di inserzione. Un gruppo può consistere

- **di un identificatore di campo, o un numero**; esso comanda la scrittura di quell'elemento sul file di uscita. Ad esempio, "**b7**" copia in uscita il settimo campo del secondo file, mentre "**100**" ordina la registrazione in uscita del numero 100.
- **di una sequenza di campi consecutivi espressa in forma compatta** per mezzo del segno "/". Ad esempio, "**c3/7 a1/20**" copia in uscita, nell'ordine, i campi da 3 a 7 del terzo file ed i primi 20 campi del primo file.
- **della definizione di una variabile ausiliaria**, il cui valore, calcolato caso per caso, viene memorizzato per essere utilizzato in altri calcoli relativi alla stessa unità ma non viene di regola scritto in uscita.
- **dell'indicatore di una variabile ausiliaria**: ad esempio, "**v3**" ordina esplicitamente che venga scritto in uscita il valore della variabile ausiliaria v3, calcolato in precedenza.
- **di una sequenza di operazioni su numeri, identificatori di campo o variabili ausiliarie**. Ad esempio, "**sqrt((a1+a2):2)**" comanda il calcolo e la registrazione in uscita della radice quadrata della media aritmetica dei primi due campi del primo file.



- **di una sequenza di operazioni che vanno compiute su di un insieme di campi consecutivi**; i risultati vengono scritti in uscita. Ad esempio, si supponga che un file d'entrata contenga la popolazione di 200 comuni suddivisa in 10 classi d'età: quel file avrà 200 record (uno per comune) ciascuno di 10 campi. Il seguente comando permette all'utente di convertire l'informazione da valori assoluti a percentuali:

$$v1=a1+a2+a3+a4+a5+a6+a7+a8+a9+a10 \quad a1/10:v1*100$$

Il primo gruppo definisce la variabile ausiliaria v1 - calcolata record per record, cioè per ciascun comune - che rappresenta la popolazione totale del comune: il secondo comanda la scrittura sul file d'uscita dei valori percentuali ottenuti dividendo ciascuno dei campi a1...a10 per il loro totale v1 e moltiplicando il risultato per 100. Il termine a1/10 viene espanso nei 10 campi che rappresenta e l'operazione richiesta viene eseguita per ciascuno dei campi. In questo caso l'uso delle parentesi tonde non è strettamente necessario, poiché ":" e "\*" hanno la medesima priorità e la valutazione procede da sinistra a destra, portando al risultato voluto. Si noti che il gruppo "(a1/10:v1)\*100" è del tutto equivalente.

**Ricorda:** una sequenza di campi come a1/10 deve stare sempre all'inizio di un gruppo.

Le parentesi tonde vengono usate per aumentare la priorità. Ad un dato livello di parentesi una espressione viene valutata da sinistra a destra rispettando le seguenti regole di precedenza: vengono calcolate prima le funzioni **ln**, **log**, **exp** e **sqrt**, poi l'**elevazione a potenza** "^", poi le operazioni "\*" e ":" nel loro ordine, infine le operazioni "+" e "-".

**Nota:** Quando viene richiesto il calcolo di qualche nuova variabile il programma determina se i campi da usare per il calcolo contengano valori interi o reali (cioè se includano o meno dei decimali). In generale il risultato viene calcolato e salvato come reale (includendo cioè alcune cifre decimali) quando almeno uno dei termini utilizzati per il calcolo sia tale, oppure si faccia uso delle funzioni **ln**, **log**, **exp**, **sqrt**, dell'**elevazione a potenza** o della divisione. In tutti gli altri casi il risultato è registrato come intero.

Se almeno una variabile va calcolata in forma reale viene chiesto all'utente di specificare il numero di cifre decimali che desidera. La risposta dev'essere un numero intero, oppure zero se si vuol registrare solo la parte intera del risultato.

**Tutti i valori in virgola mobile (cioè reali) calcolati dal programma vengono registrati con il numero di cifre decimali indicato dall'utente.**

#### Possibili errori di esecuzione

- Qualcuno dei file di input termina prima degli altri: il programma segnala il nome del file che ha meno record di quanto ci si aspetti;
- Qualche record ha meno campi di quanto ci si attenda: anche questo errore viene segnalato.

**Nota:** In quest'ultimo caso si consiglia di sottomettere il file a **FIXFORM**, che controlla tutti i suoi record e segnala il numero d'ordine di quelli che hanno un diverso numero di campi. Questo problema non si presenta quasi mai per i file scritti da programma ma è piuttosto comune se i dati sono stati digitati da tastiera. **FIXFORM** offre una forma di controllo.

- se viene eseguita **un'operazione non valida** (una divisione per 0, la radice quadrata o il logaritmo di un numero negativo) viene scritto un messaggio di errore in un file denominato "**errors**". Conviene esaminarne sempre il contenuto ad esecuzione conclusa. In corrispondenza ad un'operazione non valida viene registrato in uscita il codice che rappresenta un valore mancante (definito nel file di configurazione ADDATI.INI). Se possibile, l'esecuzione andrebbe ripetuta dopo aver corretto i valori in input responsabili del problema: si ricordi che se si sta preparando una tavola da sottoporre ad una procedura di analisi fattoriale o di classificazione, questa **non deve presentare valori mancanti**.

Quando il programma termina regolarmente il file di uscita viene formattato prima di chiuderlo: tutti i suoi campi sono ordinati in colonne (mediante una chiamata automatica a **FIXFORM**) in modo da facilitare l'esame del suo contenuto.

**Esempio**

**file1** (582 records)

```
COM_01 32 25 14
COM_02 13 114 13
.....
```

**file2** (582 records)

```
COM_01 25.12 36.02 15.00 0.05
COM_02 124.35 181.23 27.65 0.24
.....
```

*I due file da unire contengono diverse variabili che descrivono 582 comuni. Ogni record consiste di 4 campi nel primo file, di 5 campi nel secondo. Il primo campo è il contrassegno (nome) del comune.*

*Non è necessario che i due file siano formattati, ma tutti i record di uno stesso file debbono avere il medesimo numero di campi.*

*Il comando: "**a1 a3 a4 a3+a4 (a4:a3)\*100 b2/4**" produce il seguente file di uscita (formattato):*

```
COM_01 25 14 39 56.00 25.12 36.02 15.00
COM_02 114 13 127 11.40 124.35 181.23 27.65
```

*dove il quarto campo è calcolato sommando il terzo ed il quarto campo del record corrispondente del primo file; il quinto campo è ottenuto dividendo il quarto campo per il terzo e moltiplicando il risultato per 100 (si ipotizza che siano state richieste due cifre decimali).*

**Nota:** Si osservi il trattino "\_" tra "COM" e "01": il suo scopo è di far sì che il programma accetti "COM\_01" come un campo: i nomi non possono consistere di due o tre parole separate da spazi, poiché esse verrebbero trattate come campi distinti con risultati indesiderati.

**Esempio:**

*la stringa "**c1/5 a1^a3 a5 v1=b1+b2 v2=sqrt(v1) b1:v2 b2:v2 a1/10:v2**"*

*riguarda campi tratti da due file. Vengono definite due variabili ausiliarie v1 e v2, rispettivamente come somma dei primi due campi del secondo file e come radice quadrata di tale somma. Esse sono utilizzate nei calcoli ma non ne è richiesta la registrazione in uscita.*

*Record per record vengono scritti in uscita i seguenti campi:*

- i primi 5 campi del terzo file;
- il valore ottenuto innalzando a potenza il primo campo del primo file; viene usato come esponente il valore del terzo campo del medesimo file ( $a1^{a3}$ );
- il quinto campo del primo file;

- *il rapporto tra il primo campo del secondo file e v2;*
- *il rapporto tra il secondo campo del secondo file e v2;*
- *i dieci rapporti tra il contenuto dei primi 10 campi del primo file e la variabile ausiliaria v2.*

*Vengono registrati in uscita 19 campi. Il sesto e gli ultimi 12 sono in forma reale, con il numero di decimali richiesti dall'utente.*

## 2.3 - FIXFORM

---

### Conversione di un file a formato fisso

<b>Funzione</b>	Converte un file da formato libero a formato fisso.
<b>File</b>	Accetta come input un file di testo in formato libero.
<b>Requisiti</b>	Tutti i record debbono avere lo stesso numero di campi.
<b>Limiti</b>	Max. <b>5120</b> caratteri e <b>600</b> campi per record.

Anche questo programma di utilità assume i *campi* (definiti nel caso di **MERGFLD**) come le unità costitutive di un record. Esso converte file **da formato libero a formato fisso**. Si ricordi che quando il formato è libero i record debbono avere il medesimo numero di campi, ma possono avere lunghezza diversa.

La conversione avviene aggiungendo o rimuovendo degli spazi in modo opportuno tra i campi in modo da dare a tutti i record il medesimo formato (ogni record viene ad avere la stessa lunghezza ed un dato campo occupa le medesime colonne in tutti i record). Il file originale viene cancellato e rimpiazzato dal nuovo file formattato, il cui contenuto può essere esaminato con l'editor di ADDATI, oppure ricorrendo al tasto-funzione **F3** o all'utilità **SHOWREC** offerte da ADDATI (vedi più avanti).

Il programma chiede il nome del file da convertire e lo legge due volte: una prima volta per determinare la massima lunghezza di ogni campo e poi per riscriverlo in base alle lunghezze così determinate.

E' possibile attribuire a **tutti** i campi la medesima lunghezza, ovvero assegnare ad ogni campo la minima lunghezza sufficiente a garantire l'incolonnamento (generalmente diversa da campo a campo), in modo da produrre un file più compatto.

## 2.4 - MISSVAL

### Simulazione dei valori mancanti di variabili quantitative

<b>Uso</b>	<b>MISSVAL</b> è usato per simulare valori mancanti di variabili continue. Si tratta di una versione ancora sperimentale: non è sempre garantito un buon risultato, per le ragioni spiegate nelle Note 1 e 2 più sotto.
<b>File</b>	<b>MISSVAL</b> è controllato mediante il file di parametri <b>MISSVAL.PAR</b> , come <b>DISTRIB</b> e <b>CROSSTAB</b> (si veda il cap. 5). Scegliendo l'opzione <b>MISSVAL</b> dal Menu di Utilità, il file dei parametri viene editato su richiesta dell'utente.

### Come funziona **MISSVAL**

1. Nel computo della media e della varianza di ciascuna variabile i valori mancanti vengono ignorati; nel calcolo delle covarianze un'unità statistica viene ignorata quando manca il valore di almeno una delle due variabili su cui si sta lavorando. Questo significa che il numero dei casi validi è diverso da variabile a variabile, e per ciascuna coppia di variabili quando si calcolino covarianze e correlazioni.
2. A seconda della strategia definita dall'utente nel file di comandi **MISSVAL.PAR** (vedi più avanti), ogni valore mancante è sostituito dalla media non condizionata della variabile (mean substitution) o da un valore plausibile, calcolato a partire dalle correlazioni esistenti tra le variabili e dall'insieme dei valori validi presenti nell'unità statistica che presenta il valore mancante da simulare (raw imputation method).

### Raw imputation

Limitiamoci per semplicità a considerare dapprima solo due variabili  $x$  e  $y$ , con media  $\bar{x}$  e  $\bar{y}$ , deviazione standard  $\sigma_x$  e  $\sigma_y$  e correlazione  $corr(x,y)$ . Se il valore di  $x$  è ignoto, ma la sua correlazione con  $y$  non è nulla, l'osservazione del valore di  $y$  dà qualche informazione sul valore di  $x$ . La distribuzione di  $x$  ne risulta modificata: il suo valore atteso  $ev(x)$  (cioè la media della sua distribuzione condizionale al valore di  $y$ ) è generalmente diverso da  $\bar{x}$ , ed anche la sua dispersione può cambiare. Ci si può attendere che:

- se  $y = \bar{y}$  o se la correlazione è nulla, allora  $ev(x) = \bar{x}$ ;
- la dispersione condizionale di  $x$  dipende dalla correlazione: se  $corr(x,y) = 0$ , la deviazione standard della distribuzione condizionale è  $\sigma_x$  come prima; se  $corr(x,y) = 1$  il valore di  $x$  è completamente determinato da quello di  $y$ , e la sua dispersione è 0. Così, la d.s. della distribuzione condizionale in qualche modo diminuisce quando la correlazione aumenta. Comunque non è a questo che siamo interessati, bensì solo al valore centrale della distribuzione condizionale.

Poiché è evidente che il valore atteso di  $x$  in corrispondenza all'unità statistica  $i$  dipende sia dalla correlazione tra  $x$  e  $y$  che dal valore di  $y$ , sembra ragionevole assumere che

$$z_i(x) = corr(x,y) * z_i(y) \quad (1)$$

dove  $z_i(x)$  e  $z_i(y)$  sono rispettivamente **lo z-score atteso** di  $x$  e **lo z-score osservato** di  $y$  in  $i$ .

La (1) è coerente con le condizioni elencate sopra: se  $y$  è una d.s. sopra la media, e la correlazione vale 1, allora anche  $x$  sarà una d.s. sopra la sua media. Se  $y$  è 0.5 d.s. sotto la

sua media, e la correlazione è  $\text{corr}(x,y) = 0.4$ , assumiamo che la distribuzione di  $x$  sia spostata in modo tale che il valore più probabile sia  $0.5*0.4$  d.s. sotto  $\bar{x}$ , e così via (Si tratta di una congettura che va approfondita: la formula usata nel software potrebbe venire modificata).

La direzione dello spostamento del valore atteso è comunque abbastanza evidente, e la (1) fornisce una buona approssimazione, certo migliore rispetto all'uso di  $\bar{x}$  (sostituzione con la media).

Accettata la (1), possiamo azzardarci ad estenderla come segue al caso in cui vengano osservate insieme ad  $x$  altre  $p$  variabili valide:

$$z_i(x) = \frac{1}{p} \sum_j \text{corr}(x, y_j) * z_i(y_j) \quad (2)$$

dove  $z_i(y_j)$  misura la deviazione dalla sua media, nell'unità  $i$ , della  $j$ -esima variabile il cui valore sia noto, assumendo come unità di misura la sua d.s.. L'indice  $j$  varia da 1 a  $p$ , include cioè tutte le variabili valide misurate simultaneamente al valore mancante da ricostruire (nella pratica, **MISSVAL** usa solo le variabili *migliori*).

Il risultato dipende com'è ovvio dall'insieme di variabili esplicative utilizzate per predire il valore di  $x$ , e non può essere altrimenti. Sfortunatamente, è facile rendersi conto che si può alterare il risultato aggiungendo una quantità di variabili scarsamente correlate con  $x$ : esse non contribuirebbero alla somma, per via del basso valore della loro correlazione con  $x$ , ma semplicemente perché ci sono, la divisione per  $p$  mortificherebbe la capacità predittiva delle poche variabili fortemente correlate con  $x$ , che sarebbero invece sufficienti per ricostruirne in modo accettabile il valore.

D'altro canto, anche una sola variabile con correlazione prossima ad 1 sarebbe sufficiente per simulare il valore  $x$ , ignorando tutte le altre. La formula dovrebbe dunque assegnare una importanza maggiore alle correlazioni più alte.

**MISSVAL** usa il valore assoluto delle correlazioni per pesare i valori più probabili generati individualmente da ciascuna variabile presa in considerazione, espressi dalla (1). La cosa funziona allora come segue:

$$z_i(x) = \frac{\sum_j |\text{corr}(x, y_j)| (\text{corr}(x, y_j) * z_i(y_j))}{\sum_j |\text{corr}(x, y_j)|} \quad (3)$$

La (3) mostra che per ogni unità statistica con un valore mancante da ricostruire l'indice  $j$  varia su un insieme di variabili definite dell'utente: quelle con la correlazione più elevata con la variabile il cui valore va simulato. Esprimendo nella (3) gli z-score in termini di media e deviazione standard, si ottiene per il caso  $i$

$$x_i = \bar{x} + \sigma_x \frac{\sum_j |\text{corr}(x, y_j)| (\text{corr}(x, y_j) * \frac{y_{i,j} - \bar{y}_j}{\sigma(y_j)})}{\sum_j |\text{corr}(x, y_j)|} \quad (4)$$

Supponiamo che l'analista abbia deciso di usare nella simulazione  $p$  variabili. Calcolate le medie, le deviazioni standard e le correlazioni (ignorando i valori mancanti) il file dei dati viene riletto, e quando si incontra un valore mancante:

1. tutte le variabili con valori validi (in quell'unità statistica) vengono ordinate in modo decrescente rispetto al valore assoluto della loro correlazione con la variabile da simulare;
2. vengono scelte le prime  $p$ ;
3. si calcola con la (4) un valore plausibile da sostituire a quello mancante.

**Nota** *La procedura descritta è appropriata quando i valori mancanti siano distribuiti casualmente (Missing At Random). Non sempre è così: succede spesso che il valore di una particolare variabile abbia una probabilità più elevata di mancare per le unità statistiche di un dato sottoinsieme (famiglie con simile condizione sociale, unità geografiche spazialmente contigue, ecc.). In tal caso, è necessario comportarsi con cautela.*

*Si ricordi comunque che si tratta di una simulazione che mira ad evitare di dover eliminare completamente i dati relativi ad un'unità statistica solo perché manca un valore: **non si ottiene il valore vero, solo uno plausibile**, di solito migliore della media della variabile, e certamente molto meglio di niente.*

*Se l'analisi riguarda un insieme di unità geografiche, può accadere che la correlazione di due variabili, piuttosto bassa sopra l'universo, risulti molto più maggiore sopra un particolare sottoinsieme (ad esempio, un insieme di province che costituiscano una regione omogenea). In tal caso può essere conveniente estrarre tali unità (con SELECT), eseguire una ricostruzione separata, poi ricostituire il database. Se è disponibile una classificazione attendibile delle unità, tale operazione può essere ripetuta separatamente per ciascuna classe.*

*In conclusione, spetta all'analista decidere come applicare l'algoritmo di simulazione in base alla sua conoscenza dello specifico contesto territoriale dell'analisi*

**Nota** *Meglio diffidare delle variabili con troppi valori mancanti: sono inaffidabili ed andrebbero eliminate. Se vengono mantenute, e si cerca di ricostruirne i valori mancanti, possono succedere cose strane. Ad esempio, le tavole delle correlazioni calcolate da MISSVAL potrebbero contenere dei valori maggiori di +1, o minori di -1.*

*Ciò dipende dal fatto che la media e la deviazione standard delle due variabili coinvolte, nonché la loro correlazione, **non sono calcolate sopra lo stesso insieme di unità statistiche**.*

*Media e deviazione standard riguardano ciascuna variabile singolarmente: esse sono calcolate tenendo conto di tutte le unità statistiche nelle quali la variabile assuma un valore valido. La correlazione invece è calcolata sulle unità statistiche per le quali il valore di **entrambe le variabili** sia valido, e questo fa parecchia differenza quando almeno una delle variabili abbia un alto numero di valori mancanti. Le statistiche sono allora calcolate su insiemi diversi di unità statistiche, e l'intera procedura diventa inaffidabile.*

*Ad esempio, in un'analisi condotta sulle unità amministrative della Cina 1732 valori di una particolare variabile, su 2114, mancavano: la correlazione di quella variabile con un'altra, utilizzabile per la ricostruzione, risultava valere 1.240! Chiaramente, **i valori mancanti non erano distribuiti casualmente!***

***La conclusione** da ricordare è che il procedimento funziona se i valori mancanti non sono troppi, e sono distribuiti casualmente. **Attenzione ai casi estremi!***

### Un esempio

Vengono mostrati qui sotto alcune righe estratte da un file che contiene 148 record (le unità statistiche descritte sono le Sezioni Censuarie del Centro Storico di Venezia). Il primo campo è l'identificatore dell'unità, il secondo il suo peso (pari al numero di alloggi occupati nella Sezione). Seguono poi 11 variabili, il cui significato non è qui rilevante.

7	233	86.70	13.30	62.23	-9999	18.03	13.30	51.50	35.19	38.63	9.44	51.93
9	214	92.99	7.01	77.10	-9999	7.48	10.75	54.67	35.05	40.19	7.94	52.33
108	277	79.42	-9999	52.35	28.88	18.77	-9999	68.23	-9999	28.15	11.19	-9999
109	250	83.20	16.80	58.40	26.40	-9999	13.20	64.80	22.00	-9999	12.40	60.00
<b>valori simulati da MISSVAL</b>												
7	233	86.70	13.30	62.23	22.198	18.03	13.30	51.50	35.19	38.63	9.44	51.93
9	214	92.99	7.01	77.10	18.477	7.48	10.75	54.67	35.05	40.19	7.94	52.33
109	250	83.20	16.80	58.40	26.40	15.442	13.20	64.80	22.00	26.480	12.40	60.00
<b>valori effettivamente osservati</b>												
7	233	86.70	13.30	62.23	19.74	18.03	13.30	51.50	35.19	38.63	9.44	51.93
9	214	92.99	7.01	77.10	15.89	7.48	10.75	54.67	35.05	40.19	7.94	52.33
109	250	83.20	16.80	58.40	26.40	15.20	13.20	64.80	22.00	27.60	12.40	60.00
<b>medie</b>												
		86.01	13.93	50.50	31.97	17.82	18.57	59.77	21.81	23.90	17.66	58.44

Nel primo blocco di quattro righe i valori di alcune variabili sono stati sostituiti con -9999, il codice per i valori mancanti, mostrati in rosso.

Il secondo blocco mostra i valori generate da **MISSVAL**. Poiché nel file MISSVAL.PAR è stato fissato un numero massimo di tre valori mancanti per record, il record #108, che ha quattro valori mancanti, viene escluso.

Il terzo blocco mostra i valori effettivamente osservati, mentre l'ultima riga riporta i valori medi delle diverse variabili.

In tutti i casi i valori sembrano costituire una scelta migliore del semplice uso della media.

### La struttura del file MISSVAL.PAR

MISSVAL.PAR è auto-illustrato. Il file include alcune righe di commento che cominciano con un punto interrogativo "?", ed altre righe *attive* nelle quali vanno inseriti i valori dei parametri di controllo.

Le righe di commento spiegano come vadano compilate le alter per adattare il file alla prova. **MISSVAL ignora le linee di commento** e legge da quelle attive le informazioni necessarie per l'esecuzione.

All'inizio è piuttosto facile commettere degli errori, che il programma segnala con dei messaggi chiari per aiutare a correggerli. Dunque niente fretta, ed attenzione nel preparare il file.



### Un esempio di file MISSVAL.PAR

- ? Questo è il file di controllo per **MISSVAL**.
- ? Tutti i record che cominciano con '?' sono **commenti**, ignorati dal programma.
- ? Il loro scopo è solo di spiegare come vadano riempiti quelli che non iniziano con '?',
- ? che sono quelli effettivamente caricati dal programma.
- ? Ogni riga attiva inizia con una **parola-chiave** (come **INPUT\_DATA\_FILENAME**, ecc).
- ? **Le parole-chiave non vanno modificate**, vanno solo cambiati i valori che le seguono in
- ? modo da adattarli alla particolare prova da svolgere.
- ? **1. Nome del file dei dati da controllare. Inserire il percorso complete se necessario.**
- ? **NOTA!**
- ? ADDATI si attende che il file sia in **formato libero** (valori separati da spazi). Se
- ? necessario, usare le utilità fornite da ADDATI (MERGCHAR, MERGFIELD e
- ? FIXFORM) per convertirlo.
- INPUT\_DATA\_FILENAME**                      **venezia.dat**
- ? **2. Nome del file di output. Se necessario inserire il percorso completo.**
- OUTPUT\_DATA\_FILENAME**                      **venezia1.dat**
- ? **3. Elencare i numeri d'ordine** (a partire da 1) delle **variabili continue** da
- ? controllare e, in caso, simulare. Usare lo spazio come separatore, e tanti record
- ? **'ORD\_#\_OF\_ITEMS\_TO\_CHECK'** quanti ne servano.
- ? Attenzione a non includere la variabile da usare come peso, se esiste, identificata dal
- ? valore assegnato alla parola-chiave **'WEIGHT'**.
- ? Le variabili indicate verranno controllate, ed i valori mancanti simulati.
- ? **Attenzione!**
- ? Tutti gli altri campi che non vanno trattati (variabili categoriali, etichette
- ? alfanumeriche, ecc.) verranno copiati immutati e conserveranno la medesima
- ? posizione nel file di output.
- ORD\_#\_OF\_ITEMS\_TO\_CHECK**   **3 4 5 6 7 8 9 10 11 12 13**
- ? **4. Peso delle unità statistiche**
- ? Nel calcolo delle statistiche relative alle variabili si può assegnare ad ogni caso un
- ? peso opportuno, oppure lo stesso peso a tutti i casi. Va usata come peso una delle
- ? variabili incluse nel file dei dati.
- ? **Il peso deve avere valore valido per tutte le unità statistiche.**
- ? **MISSVAL** controlla il valore dei pesi, ma meglio usare **DISTRIB** prima di **MISSVAL**
- ? per assicurarsi che ogni unità abbia un peso valido, individuare i valori errati e
- ? correggerli.
- ? Inserire **'0'** dopo la parola-chiave **'WEIGHT'** per indicare che viene assegnato a
- ? tutte le unità il medesimo peso, altrimenti inserire **il numero d'ordine, a partire**
- ? **da 1, del campo da assumere come peso.**

? Il valore qui di seguito significa che il secondo valore in ciascun record è assunto  
? come peso.

## **WEIGHT 2**

### ? **5. Etichette delle variabili**

? **Inserire dopo la parola-chiave 'LABELS' la lista delle etichette alfanumeriche,**  
? separate da spazi, che identificano nell'ordine le variabili (escludendo il peso).  
? Usare tante righe 'LABELS' quanto è necessario. Max. 12 caratteri per etichetta.  
? Qui vengono inserite alcune etichette fittizie.

**LABELS** var1 var2 var3 var4 var5 var6 var7 var8 var9 var10 var11

**LABELS** var12 var13

### ? **6. Numero Massimo di valori mancanti per record**

? Maggiore è il numero di valori mancanti per un'unità statistica, meno affidabile è  
? la simulazione.  
? Fornisci il **massimo numero tollerabile di valori mancanti** per unità: tutti i casi  
? che ne hanno di più verranno esclusi.  
? (se questo valore è messo a '0', verranno esclusi tutti i casi con un numero qualsiasi di  
? valori mancanti).

## **MAX\_MISSING\_VALUES 3**

### ? **7. Strategia di sostituzione**

? Se il parametro precedente non vale 0, si intende sostituire i valori mancanti con degli  
? altri valori ragionevoli. Ciò può essere fatto in più di un modo.

? La parola-chiave **STRATEGY** può essere seguita da

- ? • 'MEAN' per sostituire ad un valore mancante la media della variabile
- ? • 'CONDITIONAL' per sostituire ad esso un valore calcolato a partire dalle correlazioni
- ? tra la variabile di cui manca il valore e le variabili valide ad essa più correlate.
- ? Se **MAX\_MISSING\_VALUES** è messo a '0', il valore di 'STRATEGY' è ignorato.

## **STRATEGY CONDITIONAL**

### ? **8. Variabili usate per simulare i valori mancanti**

? Se il valore di 'STRATEGY' è 'CONDITIONAL', inserisci i numeri d'ordine  
? delle variabili tra le quali vanno scelte quelle da utilizzare nella simulazione.  
? Dovrebbero essere indicate solo le variabili più affidabili.  
? Dopo la parola-chiave, inserisci una lista di numeri d'ordine separato da spazi. Tutte  
? le variabili devono essere già state indicate nell'istruzione precedente che inizia  
? con la parola-chiave '**ORD\_#\_OF\_ITEMS\_TO\_CHECK**', e nessuna di esse  
? deve coincidere con il peso, se c'è.  
? Se 'STRATEGY' non vale 'CONDITIONAL', l'istruzione viene ignorata.

## **VARIABLES\_TO\_BE\_USED 3 4 5 6 7 8 9 10**

? **9. Scelta delle migliori variabili da usare**

? (usato solo se il valore di '**STRATEGY**' è '**CONDITIONAL**')  
 ? Specificare **quante variabili** vadano usate nella simulazione.  
 ? Tra le variabili candidate già specificate dalla parola-chiave  
 ? '**VARIABLES\_TO\_BE\_USED**', per ciascun valore mancante verranno scelte quelle  
 ? che hanno correlazioni più elevate con la variabile da simulare.

**SELECT\_BEST**            **3**

? **Esempio**            Le due righe seguenti:

? '**VARIABLES\_TO\_BE\_USED** 1 2 3 4 5 10 11 12 13 14 15'

? '**SELECT\_BEST** 3'

? significano che le variabili con numero d'ordine da 1 a 5 e da 10 a 15 devono essere  
 ? usate per simulare I valori mancanti delle alter variabili, scegliendo per ogni unità  
 ? statistica le tre che hanno la correlazione più alta con la variabile da simulare.

? **10. Formato di lettura dei dati**

? Si tratta di una stringa, inclusa tra doppi apici, che viene usata per specificare quali  
 ? variabili, tra quelle incluse nel record, vadano caricate, e quali ignorate.  
 ? Il formato segue la convenzione usuale in ADDATI (si veda il cap.3, oppure l'help  
 ? disponibile in linea per I programmi di Analisi Multivariata).  
 ? Se viene usato un **formato libero**, I record in input vanno preparati in modo che essi  
 ? consistano, nell'ordine, di **tutte e sole le variabili da leggere**, separate da spazi.  
 ? In tal caso basta inserire un **asterisco** '\*' invece di un formato dettagliato.

**FORMAT**    "\*"

## 2.5 - SELECT

### Un programma di utilità per selezionare unità statistiche

<b>Uso</b>	Permette di selezionare dal file d'ingresso solo quelle unità statistiche per le quali alcune variabili-chiave presentino delle combinazioni di valori definite dall'utente. Solo i record relativi alle unità scelte vengono trascritti sul file di uscita.
<b>File</b>	<p>Il programma legge:</p> <ul style="list-style-type: none"><li>• un file di ingresso, che contiene i record (uno per unità) da selezionare;</li><li>• un numero variabile di file di selezione (o di filtro), i quali contengono le informazioni necessarie per operare la selezione.</li></ul>
<b>Requisiti</b>	<p>I file di input e di selezione debbono contenere informazioni relative alle medesime unità, considerate nel medesimo ordine.</p> <p>Il file d'ingresso (dal quale vanno scelti i record da trascrivere) può essere in formato qualsiasi. I record dei file di selezione debbono invece essere organizzati <b>per campi</b>, ciascuno corrispondente al valore di una variabile, <b>separati da spazi</b>.</p> <p><b>I file debbono contenere informazioni relative alle medesime unità, considerate nel medesimo ordine.</b></p>
<b>Consiglio</b>	<p>Se i file di selezione non consistono di campi separati da spazi <b>SELECT</b> non riesce a caricare i valori delle variabili di filtro e non può funzionare. Conviene allora usare <b>MERGCHAR</b> per trascrivere su di un altro file le sole variabili da utilizzare per operare la selezione, separandole con degli spazi. Il file così scritto - in formato libero - può essere usato come file di selezione.</p>

**SELECT** permette di selezionare da un file solo quelle unità statistiche per le quali alcune variabili-chiave presentino delle combinazioni di valori definite dall'utente. E' così possibile selezionare per una successiva analisi (o escludere da essa) tutte le unità che presentano un opportuno insieme di caratteri (comuni localizzati in una regione voluta e con popolazione al di sopra di una data soglia, ecc.).

**Esempio** *Se si lavora sui 300 comuni di una regione data, tutti i file debbono avere esattamente 300 record (uno per comune) ed in ogni file il record i si riferisce al comune i-esimo. La ragione per cui può essere opportuno distribuire su diversi file la descrizione delle stesse unità sta nella convenienza di mantenere separate informazioni di natura diversa: la serie della popolazione comunale anno per anno, variabili relative al patrimonio edilizio, variabili di tipo socio-economico possono ben stare in file distinti.*

**Nota:** *Le variabili utilizzate per operare la selezione possono essere contenute nello stesso file di ingresso; è cioè lecito fornire per uno dei file di selezione un nome coincidente con quello del file d'ingresso.*

L'utente deve fornire la **condizione logica** (che può essere composta da più condizioni elementari) cui un'unità deve soddisfare per venire selezionata. Tale condizione logica specifica la combinazione accettabile dei valori delle variabili di selezione. La condizione logica globale è formata combinando i seguenti elementi:

- **numeri** (1, -3.2, 300...), che sono caricati come valori in virgola mobile;
- **identificatori di campo**, che consistono di una lettera seguita da un numero intero: la lettera (a partire da 'A') denota il file di selezione dal quale il campo va caricato; l'intero è il numero d'ordine del campo nel record;
- **gli operatori aritmetici** " +, -, \*, : ", che debbono trovarsi tra i numeri o gli identificatori di campo sui quali operano; si noti che per indicare una divisione è usato " : " invece di " / ", che ha un significato diverso;
- **gli operatori relazionali** " <, >, =, >=, <=, <> " (l'ultimo significa "diverso da"); essi debbono stare tra i due termini sui quali operano: numeri, identificatori di campo o gruppi composti che debbono essere pre-valutati ed includono almeno un operatore aritmetico;
- **gli operatori logici "AND, OR, XOR, NOT"**.

Si possono usare parentesi rotonde per incrementare la priorità. Ad uno stesso livello di parentesi un'espressione è valutata da sinistra a destra, rispettando le seguenti regole di precedenza:

1. vengono valutati prima gli operatori aritmetici, poi quelli relazionali, infine quelli logici;
2. " \* " e " : " sono valutati con precedenza su " + " e " - ";
3. "NOT" è valutato con precedenza sugli altri operatori logici.

Gli operatori aritmetici agiscono su quantità in virgola mobile (numeri, indicatori di campo o risultati di operazioni aritmetiche già calcolate) e producono come risultato un valore in virgola mobile.

Gli operatori relazionali confrontano valori in virgola mobile (numeri forniti dall'utente, valori di campi letti da file, risultati di operazioni aritmetiche); il risultato è un valore di verità (VERO se la condizione è soddisfatta, FALSO altrimenti).

Gli operatori logici agiscono su due valori di verità e forniscono un valore di verità come risultato. Fa eccezione l'operatore "NOT", che agisce su un solo termine, che segue convenzionalmente l'operatore.

**Nota:** *"AND" restituisce VERO solo se entrambi i suoi termini hanno valore VERO;*  
*"OR" restituisce VERO quando almeno uno dei suoi termini è VERO;*  
*"XOR" (OR eXclusivo) restituisce VERO se uno ed uno solo dei suoi termini è vero;*  
*"NOT" restituisce VERO se il termine su cui opera è FALSO, e viceversa.*

La condizione globale digitata dall'utente è convertita in maiuscolo (si possono dunque usare indifferentemente caratteri maiuscoli o minuscoli); essa viene analizzata sia per quanto riguarda la sintassi che l'accettabilità degli identificatori di campo. Viene poi letto un record sia dal file di ingresso che dai file di selezione; i campi indicati nella condizione vengono convertiti a valori in virgola mobile ed assegnati alle espressioni dove fungono da termini. Viene calcolato il valore di verità globale: se il risultato è VERO, il record è copiato su quello d'uscita.

L'operazione viene ripetuta per tutti gli altri record del file d'ingresso.

## L'operatore "/"

Questo operatore permette di specificare in modo conveniente che una stessa condizione dev'essere soddisfatta da un insieme di campi consecutivi.

**Esempio** Invece di: " **$A4 > .05 \text{ AND } A5 > .05 \text{ AND } A6 > .05 \text{ AND } A7 > .05$** "  
si può scrivere:  
 **$A4/7 > .05$**

Ciò significa che nel primo file di selezione tutti i campi dal quarto all'ottavo incluso debbono soddisfare la condizione. E' anche lecita una condizione come " $b3/6 \leq a1$ ", che richiede che il valore di tutti i campi dal terzo al sesto nel secondo file non superino il valore del primo campo del primo file.

Una condizione che usi l'operatore "/" viene espansa prima di ogni valutazione; a tutte le condizioni logiche che ne risultano viene assegnato il medesimo livello di priorità.

**Ricorda:** L'operatore '/' dev'essere incluso nel primo termine di una condizione; è illegale includerlo nel secondo o in entrambi i termini: " $a4/8 < b3/7$ " non è accettabile.

**Esempio** La condizione: " **$a1/10 > .05 \text{ AND } (b3 < 5 \text{ or } b3 > 10)$** "  
viene interpretata come segue.

La prima condizione, scritta in forma compatta, viene espansa nella

**$a1 > .05 \text{ AND } a2 > .05 \dots \text{ AND } a10 > .05$**

che richiede che i primi dieci valori del primo file siano tutti maggiori di .05. La valutazione parte dalla condizione in parentesi: è VERA quando la variabile corrispondente al terzo campo del secondo file di selezione è minore di 5 o maggiore di 10. Si immagini ad esempio che il campo contenga il mese d'inizio del ciclo vegetativo in un certo distretto: la condizione in parentesi è VERA solo quando il ciclo vegetativo inizi tra Novembre ed Aprile (inclusi). La condizione globale specifica che oltre a ciò, un distretto deve anche avere i primi 10 valori del primo file di selezione tutti maggiori di .05. In generale, gli spazi possono venire omessi quando non ne risulti compromesso il significato. La condizione precedente si potrebbe anche scrivere come:

**$a1/10 > .05 \text{ and } (b3 < 5 \text{ or } b3 > 10)$**

e venire ancora interpretata correttamente. Si noti comunque lo spazio tra 'or' e 'b3': senza di esso, 'orb3' verrebbe assunto come un identificatore di campo e rifiutato. Conviene esser chiari e scrivere una condizione senza lesinare in spazi e parentesi.

**Esempio** La condizione " **$a3 \neq b1 \text{ and } a4:a5 = .5$** "

seleziona tutte le unità per cui il campo 3 del file 1 è diverso dal campo 1 del secondo file, ed inoltre il rapporto tra i campi 4 e 5 del primo file vale esattamente .5. Le formulazioni:

**$\text{NOT } (a3 = b1) \text{ AND } a5 = 2 * a4$**  o

**$(a3 < b1 \text{ OR } a3 > b1) \text{ AND } a5 : a4 = 2$**

sono altri modi equivalenti di esprimere la medesima condizione.

## 2.6 - SHOWREC

### Visualizzazione e modifica per accesso diretto di un grande file di dati formattato

<b>Uso</b>	Consente l'esame di un file di dati <b>di tipo testo</b> quando la sua lunghezza eccessiva renda difficile o impossibile caricarlo con un editor.  È possibile muoversi molto rapidamente lungo un file di qualunque dimensione, <b>sovrascrivendo</b> il suo contenuto laddove siano richieste piccole modifiche.
<b>Requisiti</b>	Per un funzionamento regolare è <b>assolutamente richiesto che il file abbia un formato fisso</b> , cioè tutti i suoi record abbiano – <i>almeno logicamente</i> - la medesima lunghezza. L'effettiva presenza di un EOR (caratteri di fine record) alla fine di ciascun record non è strettamente necessaria: basta che il contenuto del file possa essere suddiviso in porzioni di eguale lunghezza che si possano considerare come <i>record logici</i> . Se non incontra dei caratteri di fine-record, <b>SHOWREC</b> chiede una lunghezza logica e la usa per mostrare il contenuto del file suddiviso in modo opportuno.
<b>Limiti</b>	Nessun limite nella dimensione del file esaminato.

Lo schermo può mostrare simultaneamente sino a 18 record. Per muovere il cursore spostandosi lungo il file vanno usati i **tasti freccia**.

I tasti CTRL-↑ ↓ → ←, PgUp e PgDn permettono spostamenti più rapidi.

I tasti **F5** ed **F6** portano il cursore all'inizio e rispettivamente alla fine del record corrente. La posizione nel file (riga e colonna) è mostrata e continuamente aggiornata nella parte bassa dello schermo, dove sono anche elencati i comandi disponibili.

#### Esempio

"**+1000**" porta avanti di 1000 record (o, nel caso, alla fine del file);

"**-5000,10**" porta indietro di 5000 record (oppure all'inizio del file) e mostra 10 record;

"**2000,12**" porta al record n.2000 e mostra 12 record.

Il tasto **F1** abilita/disabilita la condizione di **sovrascrittura**. In tale condizione si può scrivere sullo schermo cambiando il contenuto della pagina mostrata. I cambiamenti sono evidenziati in un colore diverso e vengono salvati a richiesta.

**Nota:** **SHOWREC** non è un editor: solo la parte del file mostrata sullo schermo è caricata in memoria centrale. Ciò permette di esaminare rapidamente parti del file tra loro anche molto lontane, ma consente solo di sovrascrivere il contenuto attuale. Il file non può cambiare la sua dimensione e niente può venire inserito o aggiunto.

**Nota:** **SHOWREC** può rivelarsi molto utile per apportare correzioni limitate a grandi file di dati (si pensi ad un file che contenga dati censuari al livello di massima disaggregazione, oppure che registri dei dati d'inchiesta). Conviene:

- passare il file a **FIXFORM** nel caso sia in formato libero, per dare la stessa lunghezza a tutti i record;
- usare qualcuna delle routine offerte da **ADDATI** che, pur finalizzate ad altri obiettivi, offrono un controllo della correttezza di codifica delle diverse variabili (ad esempio, **TYPOL**, o **RECODE**) e registrano su file gli errori riscontrati;
- usare tali informazioni per mirare immediatamente - all'interno di **SHOWREC** - ai record da modificare.



## 2.7 - RECODE

### Ricodifica di variabili quantitative o categoriali

<b>Funzione</b>	Ricodifica variabili sia quantitative che categoriali secondo diverse opzioni. Il file di uscita ha esattamente il medesimo formato di quello in input: i valori ricodificati sostituiscono quelli originali, mentre le variabili non ricodificate vengono trascritte inalterate.
<b>Requisiti</b>	Il file di input dev'essere <b>un file di testo in formato fisso</b> .
<b>Limiti</b>	La tavola di lavoro consente di ricodificare fino a <b>15</b> variabili per volta. Il limite è tuttavia solo apparente, perché una nuova sessione di lavoro può essere iniziata sullo stesso file non appena terminata la ricodifica precedente, senza abbandonare il programma. In tal caso il file appena scritto (contenente le variabili ricodificate) viene assunto come input.  I record del file di input non devono superare i <b>5120</b> caratteri.

Nella preparazione del file dei dati da sottoporre ad analisi è spesso necessario ridefinire il valore di una o più variabili. Il programma **RECODE** consente di effettuare alcune operazioni di ricodifica su variabili di tipo quantitativo o categoriale.

Il file dei dati **deve avere un formato fisso** (se necessario, si utilizzi **FIXFORM** per un pre-trattamento del file): ciascuna variabile è riconosciuta dalla posizione occupata all'interno del record e tale posizione si mantiene costante per tutti i record. Pertanto, la lettura dei dati *non avviene per campi* ma per colonne di caratteri. Ciò richiede non solo una particolare cura nella preparazione del file dei dati e nella stesura del tracciato record, ma anche una procedura un po' più laboriosa per definire le variabili che si vogliono ricodificare.

La scelta di operare su file organizzati per colonne presenta comunque un vantaggio fondamentale: si possono trattare in modo semplice e naturale tutti i casi nei quali l'informazione (il valore della variabile) non è presente. In effetti, nel caso di record organizzati per campi l'individuazione dei dati mancanti è possibile solo attraverso un codice specifico che rappresenta, appunto, un dato mancante. Se invece il record è in formato fisso non è assolutamente necessario introdurre alcuna codifica che segnali la mancanza del dato: è il programma stesso che, trovando solo spazi vuoti nelle posizioni assegnate ad una determinata variabile, interpreta tale situazione come mancanza di informazione.

Con **RECODE** è possibile effettuare più sessioni di ricodifica utilizzando sempre lo stesso file di input. E' altresì possibile cambiare i file di lavoro specificandone direttamente i nomi, oppure caricando un apposito file di parametri.

Durante una sessione si possono ricodificare fino a quindici variabili contemporaneamente. Le variabili ricodificate vengono scritte sul file di output **mantenendo esattamente la stessa posizione che avevano nel file originale**. Tutti gli altri dati vengono copiati senza variazioni.

VARIABILI DA RICODIFICARE					SOGLIE INFERIORI O MODALITA`	
var. n.	prima colon.	lunghezza Campo	tipo di operaz	n.nuove classi		
01						01
02						02
03						03
04						04
05						05
06						06
07						07
08						08
09						09
10						10
11						11
12						12
13						13
14						14
15						15
E' LA COLONNA DALLA QUALE INIZIA LA VARIABILE						

**Tabella 2-2** Il foglio di lavoro visualizzato da RECODE.

Il programma visualizza la schermata riportata in tabella 2.3, nella quale l'utente inserisce i dati relativi alle variabili che intende ricodificare. Ogni riga si riferisce ad una variabile ed è destinata a contenere le informazioni seguenti.

- Colonna 1** contiene la posizione (in caratteri dall'inizio del record) dalla quale inizia la lettura del valore della variabile. Tale numero deve appartenere all'intervallo 1...5120. Nella fase iniziale di ricodifica, il programma controlla che il valore inserito sia coerente con la lunghezza del record e con i parametri relativi ad eventuali altre variabili. Ad esempio, se al campo *PRIMA COLONNA* viene attribuito il valore 12, ciò significa che la lettura del valore della variabile in questione inizierà a partire dal dodicesimo carattere di ciascun record del file di input.
- Colonna 2** contiene la lunghezza (numero di caratteri) della variabile. Il numero inserito può andare da 1 a 10. Nella fase iniziale di ricodifica il programma controlla che il valore inserito sia coerente con il valore del campo *PRIMA COLONNA* relativo alla stessa variabile. Ad esempio, se per una variabile viene definita una *LUNGHEZZA CAMPO* pari a 4 mentre *PRIMA COLONNA* vale 12, la lettura della variabile in questione avverrà a partire dal dodicesimo carattere e terminerà al quindicesimo.
- Colonna 3** è destinata a contenere un numero da 1 a 4 che identifica l'operazione di ricodifica che si intende effettuare sulla variabile (vedi le note successive).
- Colonna 4** qui va inserito il numero delle classi (categorie o modalità) della variabile categoriale risultante dall'operazione di ricodifica. Il numero delle nuove classi richieste va da un minimo di uno ad un massimo di quindici se le operazioni di ricodifica scelte sono la 1 o la 2, mentre per le operazioni 3 e 4 si possono richiedere fino a 25 classi.
- Colonna 5** questo campo si attiva solamente se l'utente ha richiesto una operazione di ricodifica di tipo 1 o 2. In tal caso l'utente dovrà digitare i valori di soglia delle varie classi (nel caso di operazione 1) ovvero le modalità della variabile

che intende accorpate in ciascuna classe (operazione 2). Nel primo caso, per ciascuna nuova classe dovrà essere inserito il valore corrispondente alla soglia inferiore della stessa (l'indicazione della soglia è obbligatoria per tutte le classi previste); **i valori inferiori alla soglia minima non vengono ricodificati**. Nel secondo caso, per ciascuna nuova classe dovranno essere indicate, separate da uno o più spazi, tutte le modalità della variabile di partenza che si intendono aggregare (massimo 10); **le modalità non specificate verranno attribuite automaticamente all'ultima classe prevista**.

**Nota:** Eventuali errori nei parametri inseriti dall'utente, vengono segnalati all'inizio della fase di ricodifica.

*Tutto ciò che l'utente digita viene visualizzato nella cella attiva del tabellone, riconoscibile dal diverso colore rispetto a tutte le altre. Si dà conferma del valore inserito premendo ↵ o una freccetta che consenta lo spostamento nella direzione scelta.*

### Le operazioni di ricodifica

Sono possibili quattro tipi di operazioni di ricodifica. Esse operano esclusivamente su dati numerici: qualsiasi valore non numerico rilevato all'interno del file viene interpretato come un errore di codifica. Teoricamente, ogni ricodifica può essere effettuata indifferente su variabili sia quantitative che categoriali; è l'utente che deve valutare quale operazione scegliere in relazione alle proprie esigenze e tenendo conto delle caratteristiche proprie di ciascun tipo di ricodifica.

I nuovi valori delle variabili ricodificate vanno da 1 al numero di nuove classi definito dall'utente.

Di seguito sono riportate le caratteristiche peculiari dei **quattro tipi di ricodifica**. Per ciascuna viene segnalato il tipo delle variabili sulle quali *normalmente* dovrebbe operare.

1. *(per variabili quantitative e qualitative di tipo ordinale)* I valori della variabile sono ricodificati in base alle soglie specificate in colonna 5. Il valore della variabile ricodificata può variare da 1 (quando il valore originario è compreso tra la soglia inferiore della classe 1 e la soglia inferiore della classe 2 esclusa) ad  $n$  (numero di nuove classi) quando il valore iniziale è maggiore o uguale alla soglia inferiore dell'ultima classe.

**Esempio** *I valori di una variabile quantitativa devono essere suddivisi in cinque classi aventi i seguenti intervalli: (-15, -5), (-4, 6), (7, 14), (15, 20), (21 e oltre). I valori da impostare per ciascuna soglia dovranno essere:*

*soglia inferiore classe 1: -15 (o un valore inferiore)*

*soglia inferiore classe 2: -4*

*soglia inferiore classe 3: 7*

*soglia inferiore classe 4: 15*

*soglia inferiore classe 5: 21*

2. *(per variabili categoriali di tipo non ordinale)* Le nuove classi vengono determinate in base alle indicazioni fornite dall'utente nella colonna 5. La variabile assumerà il valore 1 se il suo valore iniziale è uno di quelli specificati dall'utente nella

riga corrispondente alla classe 1, e così via per le classi successive fino alla penultima.

**Esempio** Le otto categorie di una variabile qualitativa devono essere raggruppate in tre classi secondo il seguente schema: le modalità 1 e 3 confluiranno nella classe 1, le modalità 4, 6 e 7 nella 2 e le rimanenti modalità 2, 5 e 8 nella 3.

L'utente dovrà fornire i valori secondo questo schema:

modalità aggregate nella classe 1: 1 3

modalità aggregate nella classe 2 : 4 6 7

modalità aggregate nella classe 3: 2 5 8

**Nota:** Non è consentito inserire la stessa modalità in più di una classe. L'ultima classe raccoglie non solo le categorie esplicitamente assegnate ad essa, ma anche tutti i rimanenti valori non indicati ma presenti nel file dei dati.

3. (per variabili quantitative) I valori della variabile sono raggruppati in modo automatico in **intervalli aventi la medesima ampiezza**, calcolata dividendo la differenza tra massimo e minimo assoluti per il numero delle classi richieste.
4. (per variabili quantitative) I valori iniziali sono suddivisi in gruppi (in generale corrispondenti ad intervalli di diversa ampiezza), in modo tale che **ogni nuova classe contenga all'incirca lo stesso numero di casi**. Si ottengono in tal modo delle categorie **ben bilanciate**, che permettono di evitare indesiderati effetti di dominanza statistica quando la variabile è utilizzata in un'Analisi delle Corrispondenze.

### *Tasti per il controllo delle funzioni*

---

Di seguito sono riportate le *chiavi* per il completo controllo delle funzioni previste dal programma:

**ESC** termina il programma

**F1** inizia la fase di ricodifica

**F10** uscita temporanea al DOS; si ritorna a **RECODE** digitando EXIT

**'?'** aiuto contestuale

**ALT-F5** cancella tutti i parametri inseriti nel tabellone

**ALT-I** imposta un nuovo nome per il file di input

**ALT-O** imposta un nuovo nome per il file di output

**ALT-H** note generali sul programma

**ALT-P** carica il file dei parametri

Le **FRECCETTE** ed il tasto **ENTER** permettono di spostarsi all'interno del tabellone nelle direzioni consentite e, contemporaneamente, di confermare il valore eventualmente inserito.

**Nota:** Oltre a scrivere il file ricodificato, **RECODE** salva tre altri file che contengono utili informazioni:

- un file che registra le operazioni effettuate ed i risultati della ricodifica, denominato "REC\_????.DAT";
- un file che registra gli errori di lettura incontrati nel file originale; è denominato "REC\_????.ERR" e, in alcuni casi, potrebbe assumere dimensioni molto grandi (ad esempio a causa di errori nell'indicazione dei parametri di lettura, o di errori nel file dei dati quali presenza di caratteri non numerici o la mancanza di dati);
- un file di parametri che registra tutti i dati inseriti dall'utente relativamente alla ricodifica appena effettuata (inclusi i nomi dei files). Viene denominato "REC\_????.PAR" in modo automatico dal programma. In tal modo l'utente potrà ripetere esattamente la stessa prova, o modificare alcuni dei parametri immessi in precedenza, semplicemente caricando (con ALT-P) il file dei parametri salvato in precedenza, senza dover digitare completamente tutti i dati.

L'estensione "????" sta per un numero progressivo che parte da "0000", determinato automaticamente dal programma allo scopo di evitare la sovrascrittura non voluta di precedenti file di report presenti nella stessa directory di lavoro.

## 2.8 - FACPLAN

### Visualizzazione delle proiezioni delle unità/variabili sui piani fattoriali

<b>Funzione</b>	Permette di visualizzare proiezioni su piani fattoriali utilizzando le informazioni salvate da <b>ACORR</b> , <b>ACOMP</b> e <b>NONGER</b> su degli specifici file con estensione <b>.FPL</b> . E' possibile editare graficamente le immagini così ottenute, registrarle su file in formato PCX o stamparle. E' anche possibile ricaricare un file così salvato, o qualunque altro file di tipo PCX.
<b>File</b>	Richiede una particolare struttura del file in lettura e può dunque operare solo sui file con estensione <b>.FPL</b> scritti da <b>ACORR</b> , <b>ACOMP</b> o <b>NONGER</b> o su file strutturati in modo simile.
<b>Limiti</b>	Il file di input può contenere i valori relativi ad un massimo di <b>6</b> coordinate fattoriali.

L'interpretazione dei risultati di un'analisi fattoriale - in particolare, il riconoscimento del significato dei fattori - passa di norma attraverso l'esame dei contributi assoluti e relativi di ciascun punto. Risulta comunque spesso utile, allo scopo di cogliere rapidamente le relazioni più significative che intercorrono tra oggetti e/o variabili, esaminare il modo in cui i loro punti rappresentativi si dispongono sui principali piani fattoriali.

Se l'utente lo richiede, i programmi di Analisi Fattoriale (**ACOMP** ed **ACORR**) salvano le informazioni richieste da **FACPLAN** su dei file con estensione **.FPL**.

Quando viene eseguito, il programma **NONGER** (Classificazione non gerarchica) rilegge automaticamente quello che **ACOMP** o **ACORR** hanno scritto e lo modifica, sostituendo i quadratini che indicano la posizione delle unità statistiche sui piani fattoriali con il numero della loro classe di assegnazione. **NONGER** aggiunge anche i centri delle classi e salva l'uscita su di un file denominato NGnn.FPL, dove 'nn' è il numero delle classi della partizione.

**FACPLAN** è un'applicazione *Windows 32* che visualizza le proiezioni delle variabili e/o delle unità statistiche sui piani fattoriali più esplicativi.

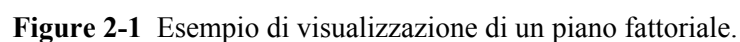
Prima di entrare in **FACPLAN**, viene richiesto all'utente il nome del file **.FPL** da utilizzare (va specificato il percorso completo se il file non si trova nella directory di lavoro).

Il file può contenere informazioni su più fattori (minimo 2, massimo 6): una qualsiasi coppia di essi determina il piano sul quale vengono proiettati i punti (oggetti/variabili). Al momento del caricamento il programma visualizza la proiezione sul piano individuato dai primi due fattori (quelli più importanti).

La Figura 2-1 mostra una schermata di **FACPLAN** che visualizza il piano 1-2 relativo ad un'analisi su 39615 famiglie di un'inchiesta (FIES2000) condotta nelle Filippine nell'anno 2000. Il numero delle unità appare in grigio nella casella 'Oggetti attivi', e si può leggere chiaramente solo ingrandendo la figura. I cerchietti gialli mostrano la posizione dei centri delle classi.

Le unità (famiglie) non sono visualizzate perché sono troppe. Esse vengono mostrate automaticamente al caricamento del programma solo se sono meno di 2000, tuttavia l'utente può forzarne la visualizzazione spuntando la casella relativa. Le etichette delle unità statistiche non vengono mai mostrate quando esse sono più di 2000: la cosa sarebbe troppo lenta, e la visualizzazione confusa.

- visualizzare ogni combinazione a piacere di variabili/unità attive/supplementari;
- visualizzare le proiezioni sul piano fattoriale determinato da qualunque coppia di fattori tra quelli dei quali si è richiesta la registrazione (fino a sei);
- limitare la rappresentazione alle sole classi indicate dall'utente (se le unità sono classificate), allo scopo di facilitare l'interpretazione dei caratteri delle classi rispetto alle variabili descrittive;
- visualizzare delle etichette identificative (nomi) per le variabili o le unità;
- consentire all'utente di tracciare linee che congiungano alcuni oggetti visualizzati (tipicamente, punti che rappresentano le categorie di una variabile categoriale);
- ...fare molte altre cose, che si possono scoprire usando il programma.



## 2.9 - INTEGRA

---

### Aggiornamento dell'archivio dei dati dopo una classificazione su unità elementari (a partire da **TYPOLÓG**)

**INTEGRA** è un semplicissimo programma che unisce alcune informazioni salvate da **TYPOLÓG** e **NONGER** dopo una sequenza di classificazione iniziata con **TYPOLÓG**, cioè aggregando un elevato numero di unità elementari (ad esempio, famiglie, alloggi, ecc.) in tipologie, sulla base dei valori di alcune variabili *attive*.

- **TYPOLÓG** salva sul file TYPCLAS l'informazione sulla tipologia alla quale ciascuna unità elementare appartiene;
- **NONGER** salva sul file NGCLASnn.TXT, dove 'nn' rappresenta il numero delle classi, la classe alla quale ciascuna tipologia è stata assegnata.

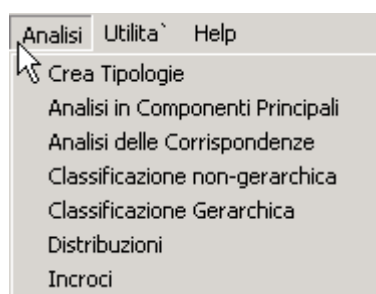
**INTEGRA** legge questi due file e registra ordinatamente su di un altro, chiamato CLASSI.CLS, la classe di assegnazione delle unità elementari. La colonna di questi valori, che rappresenta una nuova variabile sintetica, può essere accostata al file dei dati iniziali usando i programmi di utilità **MERGFIELD** o **MERGCHAR**.



## Cap. 3 - Menu di Analisi: l'interfaccia utente

---

Questo menu (vedi tabella 3.1) offre alcune analisi statistiche utili ad esprimere in forma sintetica l'informazione contenuta in una grande tavola di dati. Una quota minore dell'informazione viene perduta ma la parte più significativa viene restituita in forme facilmente interpretabili, permettendo così all'utente di focalizzare gli aspetti più rilevanti del caso studiato senza venire confuso o fuorviato da fatti relativamente poco importanti, o dalla grande quantità di dati.



**Tabella 3.1** - Il Menu di Analisi

Si tratta di tecniche multivariate, da applicare quando si abbia un insieme di unità statistiche (aree geografiche, famiglie, imprese, ecc.) descritte da molte variabili e si voglia esplorare la struttura delle relazioni che intercorrono tra le variabili o la somiglianza tra le unità. L'utilità in Analisi Territoriale è evidente: problemi di classificazione, analisi di dati censuari, spoglio di inchieste, trovano in ADDATI uno strumento appositamente concepito. Comunque, il pacchetto trova facile utilizzo in ogni disciplina che debba affrontare problemi di natura analoga.

I dati disponibili vanno organizzati in archivi. Le opzioni offerte dal Menu di Utilità consentono di effettuare l'estrazione delle variabili necessarie per una particolare analisi e di manipolarle opportunamente, preparando la tavola dei dati da sottoporre alle tecniche esplorative comprese in questo menu.

### 3-1 Alcuni aspetti rilevanti in ADDATI

---

I programmi di Analisi e di Utilità inclusi in ADDATI sono, con qualche eccezione, programmi DOS. Quelli soggetti alle richieste più pesanti sono a 32 bit e girano sotto un DOS Extender che li rende capaci di indirizzare, se necessario, tutta la memoria centrale disponibile. Sono scritti in C, un potente linguaggio di programmazione di alto livello che consente allo stesso tempo un controllo pressoché completo sulle funzioni di basso livello.

Purtroppo non si tratta di applicazioni Windows, e ad un utente abituato a lavorare con Windows l'interfaccia può apparire piuttosto rozza, anche se svolge bene il suo compito. Comunque ancora per poco, una versione pienamente Win32 è ad un buon livello di avanzamento.

Per quanto riguarda l'interfaccia, ADDATI mostra le seguenti caratteristiche:

- l'immissione dei parametri richiesti per un'analisi (numero degli oggetti o variabili attivi e/o supplementari, nomi degli oggetti e delle variabili, ecc.) avviene attraverso apposite finestre ed è sempre possibile correggere i parametri già inseriti. E' anche

possibile caricare un file di parametri relativo ad un'analisi condotta in precedenza (salvato automaticamente), adattando con un minimo sforzo tali parametri alla prova corrente;

- *l'uso della memoria è dinamico*: viene richiesto l'esatto ammontare necessario per allocare una tavola ogniqualvolta serve, per rilasciarla subito quando non serve più. Ciò garantisce un uso razionale della memoria centrale disponibile;
- il pacchetto è auto-illustrato. Esso offre un *help in linea*: il tasto '?' permette di ottenere una o più pagine di spiegazione specifica;
- i dati in input vengono accettati in *formato sia libero che fisso*: nel secondo caso, l'utente deve fornire un formato che segue alcune regole semplici ed è interpretato in modo abbastanza flessibile dalla routine che lo analizza. Esso controlla la lettura dei dati da file. Un formato di tipo FORTRAN è generalmente accettato, nel caso lo si preferisca;
- la visualizzazione delle proiezioni sui piani fattoriali è realizzata da una nuova applicazione WIN32 molto potente e di facile uso.

### 3-2 L'interfaccia utente (alcune note generali)

---

Alcuni aspetti sono comuni a tutti i programmi di Analisi Multivariata (**TPOLOG**, **ACOMP**, **ACORR**, **NONGER** e **AHC**). Li descriviamo con qualche dettaglio prima di passare a trattare i singoli programmi.

#### *Cambiare le risposte già fornite*

---

Quando si lancia un'analisi viene chiesto all'utente di fornire le informazioni necessarie per un'esecuzione corretta (numero delle unità statistiche, numero delle variabili e loro etichette, ecc.). La sequenza delle domande poste non è fissa, ma dipende dalle risposte che vengono fornite man mano.

E' sempre possibile in questa fase rivedere le risposte già date e correggerle se ci si accorge di aver fatto un errore, o se si è cambiata idea (ad esempio, in merito alle variabili da dichiarare come supplementari). Ci sono due modi per entrare in **condizione di correzione** (il che significa che si possono rivedere i valori già inseriti, non che si debba necessariamente modificarne qualcuno):

**F1** per rivedere le risposte fornite a cominciare dalla prima;

**PgUp** per rivedere le risposte a partire dall'ultima

***Nota:** La posizione corrente nella sequenza delle domande viene memorizzata. Le risposte fornite si possono scorrere con PgUp e PgDn.*

**F2** permette di tornare all'ultima domanda uscendo dalla fase di correzione.

Una modifica a qualche parametro già fornito può cambiare la sequenza delle domande previste. Le risposte a domande non più necessarie vengono abbandonate, liberando la memoria occupata; ma può anche sorgere la necessità di nuove domande, non poste prima. Ogni volta che l'utente esce dalla fase di correzione (con **F2**) il programma controlla che tutti i parametri siano accettabili e tra loro coerenti e pone le domande integrative eventualmente necessarie.

Durante la fase di inserimento parametri il tasto

- ‘?’ comanda la presentazione delle pagine di aiuto specifico eventualmente disponibili
- F10** permette un'uscita temporanea a DOS. Compiute le operazione desiderate, basta digitare **"EXIT"** per riprendere il programma esattamente dal punto in cui lo si è lasciato.  
Questa opzione è obsoleta lavorando sotto Windows: si può sempre rimpicciolire la finestra DOS con ALT-ENTER, ed usare Windows per qualsiasi operazione. F10 può ancora essere utile per verificare quale sia la directory attiva quando si stia lavorando sotto DOS.
- F3** chiama un programma di utilità interno che consente di vedere il contenuto di un file di testo. Anche questa opzione è ora obsoleta: si può sempre usare l'editor Win32 di ADDATI per visualizzare qualsiasi file, con tutti i vantaggi elencati nel capitolo 1.

---

### *Caricamento di un file di parametri*

Tutte le analisi che leggono un file di input esterno ad ADDATI (**TYPOLG**, **ACORR**, **ACOMP**, **NONGER**, **CAH**) necessitano di alcuni parametri digitati da tastiera. L'operazione di immissione dei parametri è rapida quando l'utente sia già pratico del pacchetto; può risultare un po' noiosa per un principiante, specialmente quando si debbano elencare i nomi di numerose variabili. È certamente consigliabile leggere con attenzione le schermate di aiuto per capire correttamente le domande. Comunque, una volta forniti tutti i parametri relativi ad un'analisi, essi vengono salvati su di un file di testo che ha lo stesso nome del programma ed estensione **".PAR"** (ACOMP.PAR, NONGER.PAR, ecc.). Se si vuol ripetere un'analisi cambiando solo qualche parametro basta caricare il file di parametri relativo ad essa (il programma pone una domanda specifica appena lanciato) e cambiare quello che si desidera.

---

### *Il caso di indicatori alfanumerici multipli*

Viene talvolta richiesta l'immissione simultanea di più indicatori: i nomi che contraddistinguono le unità statistiche o le variabili, i numeri d'ordine delle variabili supplementari (se ve ne sono), ecc. Va ricordato che i programmi accettano una forma compatta: **"variab1/50"** viene automaticamente espanso nei 50 indicatori **"variab01"...****"variab50"**; **"1/60"** è sviluppato nei numeri da 1 a 60. Si possono anche fornire sia indicatori singoli che altri in forma compatta, digitando parecchi gruppi alfanumerici separati da spazi. Il programma controlla il numero e l'accettabilità degli indicatori forniti.

---

### *L'help in linea*

Una o più schermate di aiuto, concernenti la domanda corrente, vengono mostrate quando si preme ‘?’ (questo è il tasto generalmente utilizzato per chiedere aiuto in ADDATI. Non tutte le domande e non tutti i programmi prevedono un help, ma la maggior parte sì).

Il sistema di help è piuttosto corposo ed offre una buona idea di quello che ADDATI fa; esso include anche alcuni suggerimenti di carattere metodologico. L'utente principiante dovrebbe far ricorso all'help con frequenza e leggerlo con attenzione.

Tutti i programmi compresi nel menu di analisi che debbono leggere dati da un file esterno (cioè non scritto da un altro programma di ADDATI nell'ambito di una analisi concatenata) richiedono l'immissione di un **formato di lettura**. Esso ha lo scopo di informare il programma sulla posizione occupata nel record dai vari elementi che debbono essere caricati (indicatore alfanumerico, peso del caso, valori delle variabili). L'utente può fornire due tipi di risposta: premere '\*' se il file di input è in **formato libero** (vedi cap. 2), oppure fornire un formato valido.

### Formato libero

Ogni record nel file di input deve contenere, **separati da spazi**, - tutti e soli gli elementi che il programma si aspetta di leggere come conseguenza delle risposte precedenti. Essi devono avere **esattamente** l'ordine seguente:

- un indicatore alfanumerico (cioè un nome) - fine a 12 caratteri - che contrassegna il caso al quale il record si riferisce. Questo se si è dichiarato che i nomi degli oggetti vanno letti da file (l'alternativa è di fornirli da tastiera).
- Il peso da assegnare al caso nell'analisi. Questo nei casi in cui l'analisi usa un peso e quando si sia dichiarato che ogni caso abbia generalmente un peso diverso (l'alternativa è di assegnare il medesimo peso a tutti i casi). I pesi vanno forniti come valori interi o reali.
- Tutte e sole le variabili da caricare, nello stesso ordine in cui sono state dichiarate.

Ovviamente, un file che contenga esattamente tanti record quanti sono i casi e dove ciascun record abbia esattamente la struttura sopra specificata deve essere stato preparato per mezzo di qualche programma esterno ad ADDATI (DBase o Excel), ovvero utilizzando le possibilità offerte dal Menù di Utilità. Questo metodo è rapido e consigliabile quando l'utente sappia a priori esattamente quali variabili vadano incluse nell'analisi, come attive o supplementari (si veda il cap. 6).

Se poi si cambia idea, anche a seguito del risultato di una prima prova, e si decide di ripetere l'analisi escludendo qualche variabile o aggiungendone qualcuna di nuova, bisogna ricreare un file opportuno.

E' allora spesso più conveniente progettare un file che includa tutte le variabili **potenzialmente utilizzabili**, usandolo come input per parecchie analisi specificando ogni volta quali variabili vadano lette e quali ignorate in quella specifica prova. Per far questo l'utente deve fornire al programma un **formato di lettura**.

### La sintassi del formato di lettura

Essa segue alcune semplici regole (si tratta di una versione semplificata del ben noto formato FORTRAN, che la routine di interpretazione è comunque quasi sempre in grado di comprendere correttamente). Vengono riconosciuti i seguenti tipi (l'uso di caratteri maiuscoli o minuscoli è irrilevante):

- **Tipo 'A'** per un **indicatore alfanumerico** (ad esempio, 'a6', 'A6', '6A' o '6a' indicano tutti un indicatore alfanumerico di 6 caratteri);
- **Tipo 'P'** per rappresentare un **peso** (ad es., 'P7' o '7P' per una variabile che occupa un campo da 7 caratteri, da usare come peso);
- **Tipo 'X'** per rappresentare il **salto** di alcuni caratteri (ad es., '5X' o 'x5' per ignorare 5 caratteri);

- **Un numero** indica una variabile con quella lunghezza di campo (così, '4' sta per una variabile su 4 colonne, ecc.). Va ricordato che il file è in **formato fisso** (si veda il capitolo 2) e che ad una variabile è pertanto riservata esattamente la medesima posizione (cioè un certo numero fisso di caratteri) in tutti i record. E' ciò che si intende qui per *lunghezza di campo* della variabile: lo spazio nel record ad essa riservato. Parte di tale spazio può occasionalmente essere vuoto (cioè contenere degli spazi bianchi), quando il valore effettivo sia tale da non occuparlo tutto.

Per più variabili consecutive aventi la medesima lunghezza di campo si può usare la forma compatta indicata nell'esempio seguente.

### Esempio

*'5\*4' indica la lettura di 5 variabili consecutive di 4 caratteri ciascuna; '4\*10' indica 4 variabili di 10 caratteri ciascuna, ecc.*

*Non occorre specificare l'eventuale presenza di decimali: è compito del programma rilevarli ed interpretarli correttamente. Così, '5\*6' è perfettamente equivalente alle notazioni FORTRAN '5F6.3' o '5I6' (che vengono peraltro correttamente interpretate);*

- se un caso consiste di più record, '/' **porta a capo record**; '2/' o '/2' portano due volte a capo record, facendo così saltare un record, ecc.

Un **formato di lettura** è costituito da un numero opportuno di questi gruppi, racchiusi opzionalmente tra parentesi tonde e separati da virgole o spazi. Consideriamo qualche esempio.

### Esempio 1

*Il formato "a6, p5, 2x, 5\*6 5x 6\*7"*

*Specifica che da ciascun record vanno letti nell'ordine un nome di 6 caratteri ed una variabile su 5 colonne da usare come peso, vanno ignorati due caratteri per leggere poi 5 variabili consecutive di 6 caratteri ciascuna; vanno ancora ignorati 5 caratteri e lette 6 variabili di 7 caratteri.*

*Naturalmente, il formato deve essere coerente con quanto dichiarato in precedenza: poiché il formato viene interpretato subito, le eventuali discrepanze vengono rilevate e ne è permessa la correzione. Qui il formato implica che indicatore e peso dei casi siano letti da file e che le variabili siano 11 in tutto.*

### Esempio 2

*Il formato "(3, 5X, 4, 4, / 4A 10x, 6\*3)"*

*Legge una prima variabile di 3 caratteri, ignora 5 caratteri leggendo poi 2 altre variabili di 4 caratteri ciascuna; passa al record successivo e legge un indicatore alfanumerico di 4 caratteri; ignora 10 caratteri e legge 6 variabili consecutive su 3 colonne ciascuna. Non viene letto alcun peso. Si noti che qui l'indicatore alfanumerico non è il primo elemento: deve esserlo solo se il formato è libero, altrimenti la sua posizione può essere qualsiasi ed è specificata dal formato stesso.*

Gli esempi precedenti sono assolutamente generali: le medesime regole valgono per tutti i programmi di analisi che richiedano la digitazione di un formato. Si noti comunque che **TYOLOG** non usa nomi per le unità e non si aspetta di leggerli; **ACORR** invece usa pesi determinati internamente e non pone dunque sui pesi alcuna domanda.

## Cap 4. - Fondamenti di teoria e linguaggio

---

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. In particolare, quelli già ricordati nell'Introduzione.

Questo capitolo si limita ad offrire un sommario molto semplificato dei concetti statistici sui quali si basano le procedure multivariate incluse in ADDATI e della terminologia utilizzata sia dai programmi che nel manuale.

### 4.1 - Le scale di misura

---

Quando si voglia eseguire 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 di caratteri demografici). In generale, vanno incluse nella tavola solo le variabili che rappresentano, nel modo più appropriato e completo possibile, i caratteri delle unità statistiche che sono ritenuti 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 qualche variabile scarsamente pertinente (a meno di non considerarla come supplementare) può distorcere il risultato in modo imprevedibile ed indesiderato. 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 possono 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** (che chiameremo 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. Il percorso di analisi esplora questa rete di relazioni e riduce la multidimensionalità del fenomeno, lasciando cadere in modo ottimale solo una piccola parte dell'informazione; le unità 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 varie scale; va poi scelto un percorso di analisi appropriato per la particolare scala adottata. In particolare, se la scala di misura non è la medesima per tutte le variabili, esse vanno prima sottoposte ad una opportuna **ricodifica** che le renda omogenee, vale a dire le riporti ad una stessa scala (utilità **RECODE**). Ciò non è ancora sufficiente per una impostazione corretta dell'analisi: se si tratta di variabili categoriali, il numero delle loro categorie e la frequenza di ciascuna di esse vanno per quanto è possibile bilanciati, allo scopo di evitare la dominanza statistica di una variabile sulle altre (ed una 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 4-1** Le scale di misura per le variabili.

La tabella 4-1 elenca i tipi di scala. È importante riconoscere quella usata per ciascuna variabile, in modo da impostare correttamente l'analisi.

#### *Variabili quantitative (o continue)*

Esse assumono valori numerici espressi secondo una opportuna unità di misura, oppure a-dimensionali: ha senso compiere su di essi 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. Una 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 *pesati* 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* (pesata) di ciascuna variabile quantitativa (calcolata sull'insieme dei casi) e la sua *deviazione standard* (o *scarto quadratico medio*), definite più avanti.

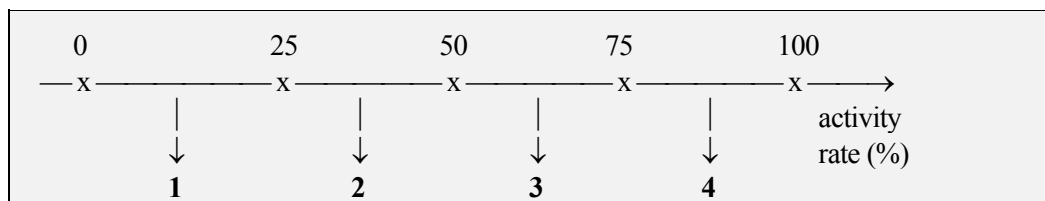
#### *Variabili categoriali (o qualitative)*

Assumono un insieme limitato di valori, rappresentati da codici opportuni. Tali codici possono essere dei numeri, anche se il loro significato non è numerico e nessuna operazione aritmetica su di essi ha significato. Si possono ulteriormente distinguere le variabili categoriali in **ordinali** e **nominali**.

#### *Variabili categoriali ordinali*

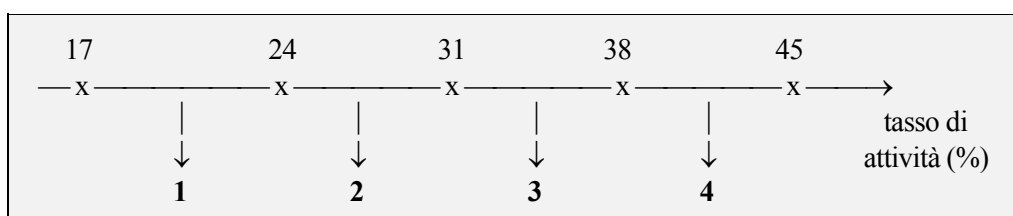
Sono di solito ottenute ricodificando in forma qualitativa delle variabili quantitative, allo scopo di renderle omogenee ad altre variabili categoriali contenute nella medesima tabella. Questa operazione causa una perdita (che bisogna cercare di minimizzare) d'informazione, ma non cambia molto il comportamento qualitativo delle unità statistiche e la struttura delle loro somiglianze. Come esempio, si consideri il tasso di attività dell'insieme dei comuni di una regione: esso può assumere teoricamente un qualsiasi valore tra 0 e 100% ed è dunque una variabile continua. Se si deve 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 4-2 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 stessa classe sono contrassegnati dal medesimo codice, e dunque le loro differenze vanno perse dopo la conversione alla scala categoriale.



**Tabella 4-2** Esempio di ricodifica di una variabile quantitativa in una variabile qualitativa a quattro categorie.

La ricodifica precedente risulta chiaramente inappropriata in molti casi: se ad esempio i valori del tasso considerato 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. Questo ci porta a decidere la nostra regola di ricodifica **dopo aver determinato** i valori massimo e minimo effettivamente assunti dalla variabile in questione. Se si vogliono ancora quattro classi corrispondenti ad intervalli di eguale ampiezza, si arriva alla ricodifica di tabella 4-3.



**Tabella 4-3** L'esempio precedente, con un intervallo di valori 17%-45%.

Si assuma ora che la maggior parte dei comuni abbiano 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 risulta indesiderabile perché *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 dimostrare che **la perdita di informazione è minima quando i valori di soglia che delimitano gli intervalli vengono determinati in modo da ottenere classi all'incirca equi-numerose**.

Tale operazione viene fatta automaticamente dalla routine **RECODE**, richiamabile dal Menù di Utilità di ADDATI.

Anche se vengono usati codici numerici, le etichette 1...4 non hanno in questo caso un significato numerico: 4 non è il doppio di 2 ed i codici A...D andrebbero altrettanto bene.

Quando una variabile quantitativa viene ricodificata l'ordine sottogiacente viene salvato: 2 non è il doppio di 1, ma rappresenta certamente un tasso di attività superiore. Per tale ragione, la variabile qualitativa prodotta dalla ricodifica è detta **ordinale**.

### Variabili categoriali nominali

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

Si pensi ad esempio di contrassegnare con un codice 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 assume per ogni unità statistica il valore corrispondente al numero della classe cui quell'unità è stata assegnata da una procedura di classificazione.

E' chiaro che i codici usati per le variabili categoriali non possono essere sottoposti a trattamento numerico: 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 4-4 mostra i valori ottenuti ricodificando il tasso d'attività della tabella 4.3: poiché i casi sono raggruppati in quattro classi (categorie), sono necessarie quattro variabili binarie.

17	24	31	38	45
— x —	x —	x —	x —	x —
↓	↓	↓	↓	↓
<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	tasso di attività (%)
↓	↓	↓	↓	
<b>1000</b>	<b>0100</b>	<b>0010</b>	<b>0001</b>	

**Tabella 4-4** L'esempio di tabella 4.3, seguito da una ricodifica in forma binaria. Vengono create quattro variabili binarie e quella corrispondente alla categoria assunta vale 1.

**Ricorda:** L'Analisi in Componenti Principali (**ACOMP**) va usata esclusivamente su variabili quantitative, mentre l'Analisi delle Corrispondenze (**ACORR**) va chiamata per analizzare tavole di contingenza (vedi più avanti), oppure tavole ottenute ricodificando in forma binaria variabili categoriali.

## 4-2 Standardizzazione 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, non sarebbe di alcun interesse e si potrebbe (si dovrebbe!) rimuoverla dall'analisi. 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 degli individui (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. Sia  $m_i$  il peso dell'unità  $i$ . I

pesi vengono **normalizzati** dalle analisi incluse in ADDATI mediante la seguente trasformazione:

$$m_i \leftarrow m_i / M$$

dove  $M = m_1 + \dots + m_n$  rappresenta la somma dei pesi. Una volta normalizzati, i pesi assommano ad 1 e ciascuno di essi misura in termini percentuali l'importanza dell'unità associata, cioè la quota-parte dell'intero sistema che essa rappresenta.

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

$$\bar{x} = \text{media}(x) = m_1 * x_1 + m_2 * x_2 + \dots + m_n * x_n = \sum_i m_i * x_i \quad (4.1)$$

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:

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

e la (4.1) diventa la **media semplice**:

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

La media è una misura di *tendenza centrale*: tutti i valori della variabile sono distribuiti intorno ad essa. La **varianza** della variabile è una misura della sua *dispersione*. Essa è data dalla

$$\text{varianza}(x) = \sigma^2(x) = m_1 * d_1^2 + \dots + m_n * d_n^2$$

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, pesando 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.

$$\text{stdev}(x) = \sqrt{\sigma^2(x)}$$

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:

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

I valori assoluti della variabile sono sostituiti dalle differenze rispetto alla media; 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.

I programmi **ACOMP** (Analisi in Componenti Principali) e **NONGER** (Classificazione non gerarchica) inclusi in ADDATI standardizzano automaticamente le variabili continue in entrata, così da attribuire a ciascuna di esse la medesima importanza nell'analisi.

### 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 4.5.

	Popolazione $pop_i$	Attivi $att_i$	tasso attività $t_i = pop_i/att_i$	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 4-5** Dati esemplificativi su tre comuni.

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

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

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.

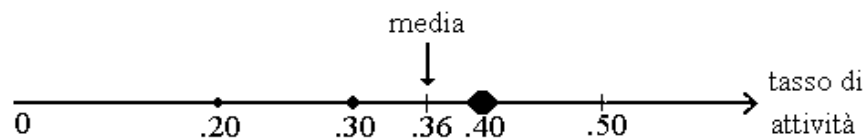
$$m_i = \text{peso del Comune } i = \text{pop}_i / \text{popolazione totale} = \text{pop}_i / (\text{pop}_1 + \text{pop}_2 + \text{pop}_3)$$

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:

$$\begin{aligned} \text{media (ponderata): } \bar{t} &= \sum_i m_i * t_i = m_1 * t_1 + m_2 * t_2 + m_3 * t_3 = \\ &= 0.27 * 0.30 + 0.667 * 0.40 + 0.067 * 0.20 = \mathbf{0.36} \end{aligned}$$

Anche nel calcolo di tutte le altre statistiche (scarto medio, varianza) va tenuto conto del peso.

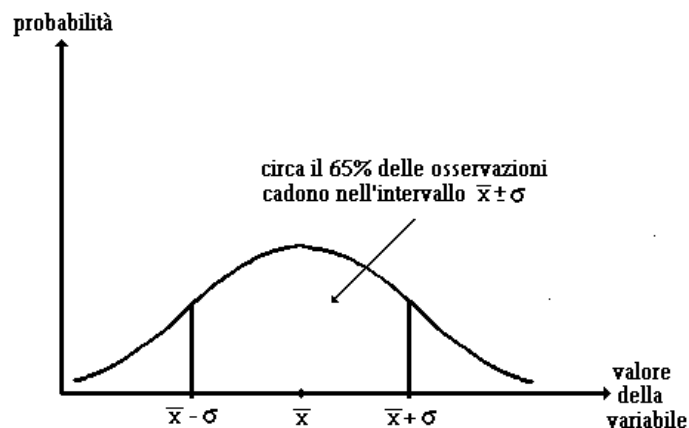


**Figura 4-1** I valori del tasso d'attività dei tre comuni rappresentati su di un asse. La grandezza di ogni punto è proporzionale alla popolazione. La media si colloca nel baricentro del sistema dei tre punti pesati.

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

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

$$\text{deviazione standard} = \sigma = \sqrt{0.0072} = .085.$$



**Figura 4-2** La distribuzione *normale* (o *gaussiana*) ed il significato di  $\sigma$ .

La **deviazione standard** ( o **scarto quadratico medio**)  $\sigma$  (sigma) misura la dispersione della distribuzione. La figura seguente mostra la distribuzione di una variabile normale ed il significato dello scarto quadratico medio.

#### *Esempio 2 - Variabili continue: normalizzazione*

La tabella 4-6 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**: percentuale di imprenditori, lib.professionisti, dirigenti ed impiegati sulla popolazione attiva

**dipl**: percentuale di laureati o diplomati sulla popolazione

**pov**: percentuale di iscritti all'elenco dei poveri sulla popolazione

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

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

**Tabella 4-6** Indicatori di benessere in cinque quartieri

È immediato osservare che:

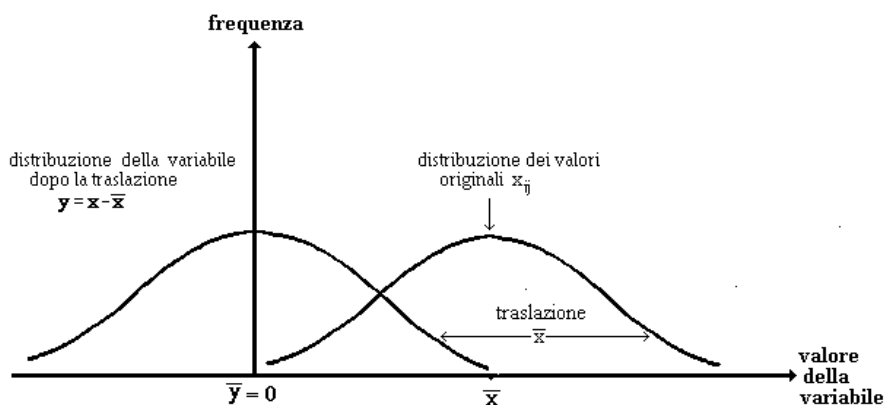
- Q4 appare il quartiere migliore, con i livelli più alti di *status* e *dipl*, e più bassi di *pov* e *affol*.
- E' difficile valutare i valori assoluti delle variabili nei vari quartieri; ha invece senso confrontare ciascuno di essi con la media globale, per capire quali quartieri assumano valori inferiori o superiori a quello medio relativo 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 seguente

$$y_{ij} \leftarrow x_{ij} - \bar{x}_j \quad (4.2)$$

dove  $\bar{x}_j$  è la media della variabile  $j$  ed  $y_{ij}$  misura lo scarto tra il valore della variabile nel quartiere  $i$  e la sua media.

- La (4.2) assegna a  $y_{ij}$  valori negativi in tutti i quartieri con valore  $x_{ij}$  **sotto la media**  $\bar{x}_j$ , valori positivi se  $x_{ij}$  è sopra la media.  
In particolare,  $y_{ij} = 0$  se  $x_{ij}$  è eguale alla media.



**Figura 4-3** Distribuzione dei valori di  $x$  prima della traslazione e distribuzione della corrispondente variabile centrata  $y$ .

- 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.

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 le diverse variabili hanno varianza diversa.

Ad esempio, la tabella 4-6 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  $\sigma$ . In tal modo, più alta è la dispersione (cioè il  $\sigma$ ) della variabile, più l'entità degli scarti sulla media viene ridimensionata. In altri termini, si usa la  $\sigma$  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  $z_j(0,1)$ .

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

La  $z_j$  si dice normalizzata o standardizzata.

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

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

	pop <sub>i</sub>	p <sub>i</sub>	status <sub>i</sub>	dipl <sub>i</sub>	Pov <sub>i</sub>	affoll <sub>i</sub>
<b>Q1</b>	28125	.148	-0.35	-0.17	0.15	-0.17
<b>Q2</b>	35853	.188	-0.79	-0.71	1.88	1.21
<b>Q3</b>	36169	.190	-0.71	-0.57	-0.48	0.86
<b>Q4</b>	30329	.159	2.19	2.26	-1.29	-1.90
<b>Q5</b>	60028	.315	-0.04	-0.29	-0.40	-0.17
<b>media</b>			0	0	0	0
<b>σ</b>			1	1	1	1

**Tabella 4-7** La tabella 4-6 espressa in termini di *z-scores*

### *Esempio 3 - Misura dell'associazione tra variabili continue*

Dall'esame della tabella 4-7 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, *pov* e *affoll* sopra la media;

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

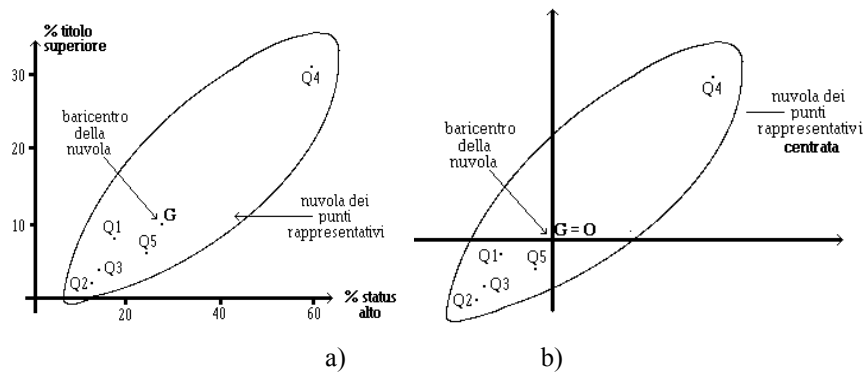
Riportiamo i valori delle due variabili *status* e *affoll* sui due assi di un piano cartesiano: poiché ogni quartiere è descritto da una coppia di tali valori (vedi la tabella 4.6), esso è univocamente rappresentato da un punto nel piano (*status*, *affoll*). Viceversa, ogni punto del piano individua 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 4-4a mostra la localizzazione dei punti rappresentativi dei cinque quartieri: si parla di **nuvola di punti**, la cui posizione sul piano è rappresentata schematicamente dall'ellissoide in figura.

Il termine *nuvola di punti* è forse esagerato quando i punti sono così pochi, ma nelle analisi reali essi sono molti di più e la nuvola è spesso molto fitta.

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

La figura 4-4b 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**. Tuttavia, la forma della nuvola, la sua dispersione, le distanze tra i punti quartiere rimangono identici nei due casi.



**Figura 4-4** La nuvola di punti (4-4a), e la stessa nuvola centrata (4-4b)

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:

$$\text{cov}(x, y) = \sum_i m_i (x_i - \bar{x})(y_i - \bar{y}) \quad (4.3)$$

Ogni caso  $i$  dà un contributo *positivo* alla costruzione della covarianza se entrambe le variabili assumono nell'unità  $i$  scarti dello stesso segno rispetto alla media (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*).

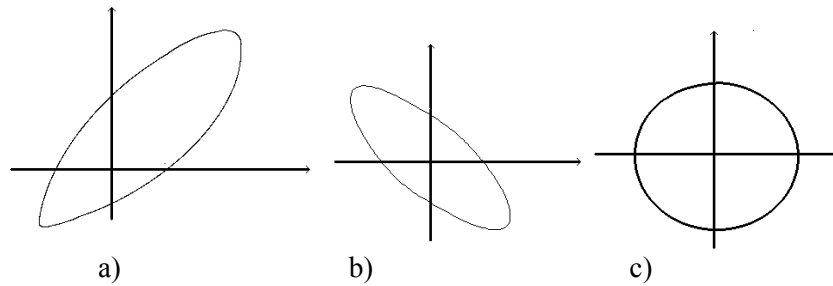
Si noti che il contributo di ciascun caso  $i$  va pesato con il suo peso  $m_i$ .

Se le due variabili sono **centrate** (cioè se  $\bar{x} = 0$  e  $\bar{y} = 0$ ) la (4.3) diventa

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

Si possono dare i seguenti casi:

- **covarianza elevata e positiva**: le due variabili tendono ad assumere *insieme* valori sopra la media, oppure sotto. La nuvola ha la forma mostrata in fig. 4-5a.
- **covarianza elevata e negativa**: le due variabili tendono a presentare *scarti di segno opposto* rispetto alla media (fig. 4-5b).
- **covarianza trascurabile (prossima a zero)**: scarti positivi di una variabile sono associati a scarti sia positivi che negativi dell'altra (fig. 4-5c).



**Figura 4-5** Diversi tipi di associazione tra variabili continue

In particolare:

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

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

- Il valore della covarianza tra due variabili **non è interpretabile in modo immediato**. Infatti, esso dipende dalla dispersione degli scarti, che è misurata dalla deviazione standard  $\sigma$  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  $\sigma_x$  e  $\sigma_y$ , poiché la varianza di entrambe le variabili è 1 dopo la normalizzazione. Tale valore è detto **correlazione** tra le variabili:

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

E' facile mostrare che  $-1 \leq \text{corr} \leq 1$ .

- Se la correlazione è prossima a 1 le due variabili assumono **insieme** valori molto alti o rispettivamente molto bassi (rispetto alla media).
- Se la correlazione è prossima a -1 le due variabili si muovono in modo opposto: quando l'una assume valori molto alti, l'altra assume valori molto bassi, 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 4-8 mostra le **correlazioni** (\*1000) tra le quattro variabili di tabella 4-6.

- La matrice è simmetrica, cioè per ciascuna coppia di variabili  $(i, j)$  risulta  $\text{corr}(i, j) = \text{corr}(j, i)$ : la correlazione è una proprietà *della coppia* di variabili, misurate su un dato insieme di unità statistiche (il "contesto", costituito nel nostro caso dai cinque quartieri), *indipendente dall'ordine in cui variabili o unità vengono considerate*.
- La correlazione di una variabile con se stessa è ovviamente 1 (indicato come 1000 nella tabella 4-8).
- *Status* e *dipl* presentano una forte correlazione positiva (= 0.984) e sono entrambe correlate negativamente con *affoll* (-0.948 e -0.915 rispettivamente). *Affoll* e *pov* sono 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 di esse è *quasi ripetuta* dalle altre due. I valori di correlazione non altissimi di *pov* con le altre variabili mostrano la parziale originalità dell'informazione apportata da questa variabile.

	<i>Status</i>	<i>Dipl</i>	<i>Pov</i>	<i>Affoll</i>
<i>Status</i>	1000	984	-671	-948
<i>Dipl</i>	984	1000	-643	-915
<i>Pov</i>	-671	-643	1000	748
<i>Affoll</i>	-948	-915	748	1000

**Tabella 4-8** Le correlazioni (\*1000) tra le quattro variabili.

### 4-3 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.

#### *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 file 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 incluse nella tavola da analizzare è delicata. Inoltre, le variabili si riferiscono in generale a diversi aspetti e sono misurate su scale diverse; vanno 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*.

Se le variabili sono tutte quantitative, esse vengono **standardizzate** e poi sottoposte ad un'**Analisi in Componenti Principali (ACOMP)**. Se sono categoriali, esse vanno convertite in forma binaria e la tavola che ne risulta va sottoposta ad un'**Analisi delle Corrispondenze (ACORR)**.

#### *Tavola di contingenza (o di conteggio)*

E' ottenuta contando unità statistiche di un medesimo 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, la tabella 4-9 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 (Censimento '81).

Le unità elementari sono le famiglie ed ogni famiglia è assegnata ad una cella a seconda della classe d'età del suo capofamiglia e del numero dei suoi componenti. Si tratta di due variabili categoriali, ottenute ricodificando due variabili quantitative sottogiacenti. 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. **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 4-9** Un esempio di tavola di contingenza. Vengono incrociate la classe di età del capofamiglia e la dimensione del nucleo familiare per tutte le famiglie del Centro Storico di Venezia (Censimento 1981).

Le tavole di contingenza vengono analizzate per mezzo dell'*Analisi delle Corrispondenze*.

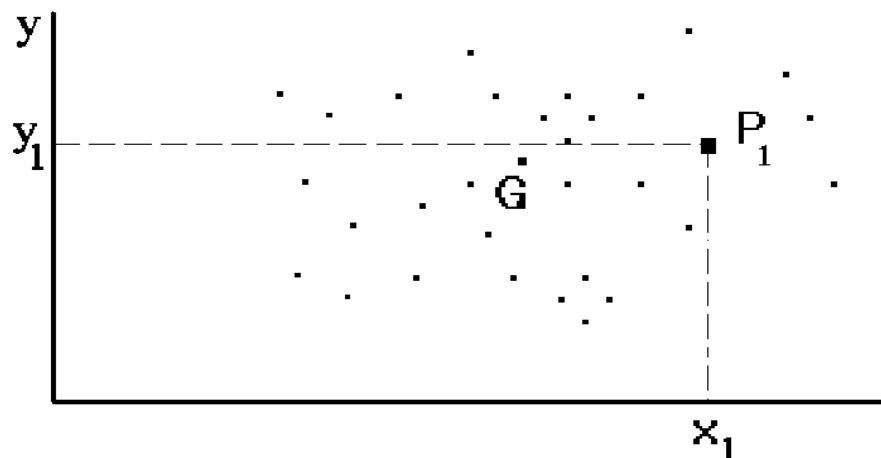
#### 4-4 La rappresentazione geometrica

Si consideri una tavola di dati  $X(n,p)$  in cui le righe rappresentino un insieme  $I$  di  $n$  unità 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 (si pensi ad un comune) è rappresentato da un vettore di  $p$  numeri reali ordinati corrispondenti ai valori delle  $p$  variabili. Le  $p$  componenti di tale vettore si possono interpretare come le coordinate di un punto in uno spazio vettoriale (geometrico)  $\mathbb{R}^p$  a  $p$  dimensioni e l'unità si può identificare con quel punto.

L'insieme  $I$  delle  $n$  unità si può rappresentare mediante una **nuvola di punti dotati di massa** (si ricordi che c'è un peso associato a ciascuna unità). Si può pensare ad ogni altro punto di  $\mathbb{R}^p$  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 4-6 mostra un esempio in un semplice caso con due variabili.

Nella figura ciascuno dei due assi ortogonali porta 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-oggetto in  $\mathbb{R}^p$ , ma si può continuare a pensare intuitivamente al caso bi-dimensionale senza perdita di generalità.



**Figura 4-6** Rappresentazione geometrica di un insieme di unità descritte da due variabili. **G** è il centro di gravità della nuvola.

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.).

### 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

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

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).

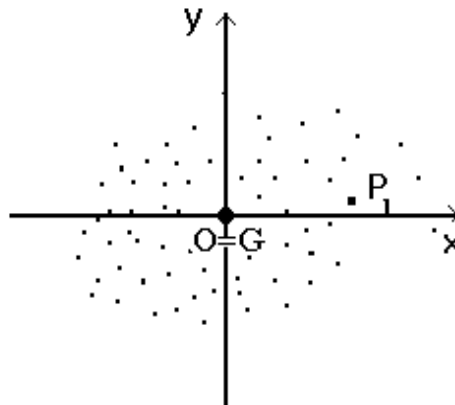
Poiché si debbono sommare i contributi alla distanza provenienti dalle diverse variabili, queste debbono essere espresse nella medesima unità di misura, oppure essere *a-dimensionali*. E' anche conveniente equilibrare i diversi contributi, in modo da evitare la dominanza di qualche variabile solo in virtù dell'unità di misura in cui essa è espressa. Un

modo per farlo è di **standardizzare** tutte le variabili prima dell'analisi: la cosa è realizzata automaticamente dai programmi 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.

### *Il centro di gravità della nuvola*

Il centro di gravità della nuvola di punti-oggetto in  $\mathbb{R}^p$  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 ( $\mathbf{G} \equiv \mathbf{O}$ ) e la nuvola stessa è detta **centrata**.

L'analisi che vogliamo eseguire s'interessa alle differenze di comportamento che esistono tra le  $n$  unità ed alle variabili cui tali differenze vanno ascritte: da un punto di vista geometrico si vuol 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). E' 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 4-7** La nuvola di figura 4-6 in forma centrata.

### *L'inerzia della nuvola*

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

$$\text{Inerzia}(i) = m_i d^2(\mathbf{x}_i, \mathbf{O}) = \sum_j m_i x_{ij}^2$$

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

$$\text{In}_{\text{tot}}(I) = \sum_i m_i d^2(\mathbf{x}_i, \mathbf{O})$$

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 collapserebbe nel suo centro e l'inerzia sarebbe nulla.

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

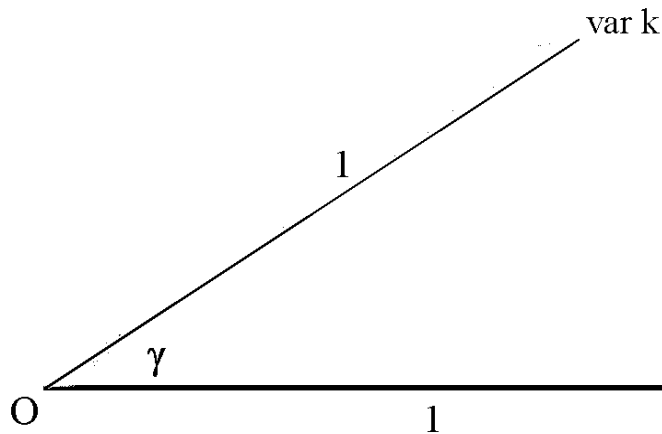
$$In_{tot}(I) = \sum_i m_i d^2(\mathbf{x}_i, \mathbf{O}) = \sum_i m_i (\sum_j x_{ij}^2) = \sum_j (\sum_i m_i x_{ij}^2) = \sum_j \text{var}(j)$$

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

### 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;

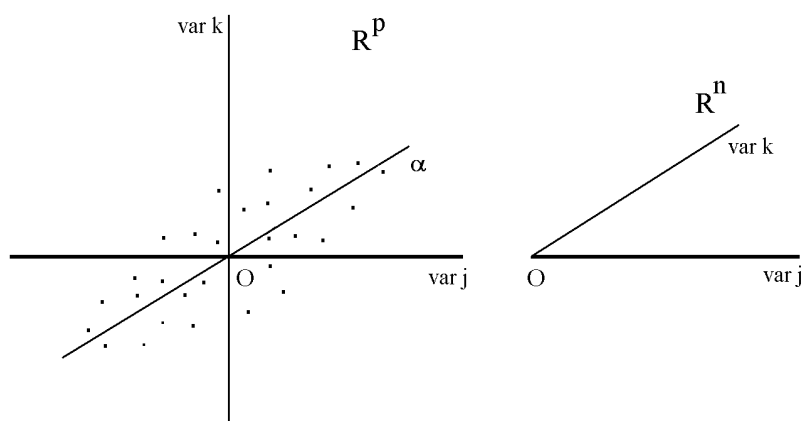


**Figura 4-8** Rappresentazione di punti-variabile in  $R^n$ . Si può dimostrare che

$$\cos \gamma = \text{corr}(\text{var } j, \text{var } k)$$

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 4-9** 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  $\alpha$ ) 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 (si veda la figura 4-9). Nell'altro spazio, poiché che 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: se 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, ecc.).

## Cap. 5 - Menu di Analisi: Distribuzioni ed Incroci

---

I programmi **DISTRIB** e **CROSSTAB**, aggiunti in questa versione, leggono i loro parametri di controllo (cioè le istruzioni che controllano l'esecuzione) dai file DISTRIB.PAR e CROSS.PAR, i cui prototipi si trovano nella directory di installazione di ADDATI.

L'utente deve copiare il file .PAR da utilizzare nella directory di lavoro, poi usare l'editor di ADDATI per adattarne il contenuto alla propria analisi, seguendo le istruzioni incluse nel file medesimo.

Quando si seleziona dal Menu di Analisi DISTRIB (o CROSSTAB) ADDATI controlla se un file denominato DISTRIB.PAR esista già nella cartella di lavoro. Se esiste, lo apre (su richiesta dell'utente) per permettere di modificarlo. Se invece non esiste, copia nella cartella di lavoro il prototipo presente nella directory di ADDATI, poi lo carica nell'editor.

In particolare:

- **DISTRIB** lavora in modo batch: dopo che l'utente ha opportunamente modificato il file DISTRIB.PAR, viene lanciato **DISTRIB** che carica il file dei parametri e segnala eventuali errori, nel qual caso il file DISTRIB.PAR viene editato automaticamente evidenziando la riga dove sta l'errore. L'utente corregge l'errore, salva il file e DISTRIB viene automaticamente rilanciato. Quando non ci sono più errori il programma calcola le distribuzioni richieste e le salva in un file denominato DISTRnn.OUT, dove 'nn' è un numero progressivo di due cifre, scelto dal programma per evitare di sovrascrivere file di uscita relativi a prove precedenti. Il file può essere esaminato con l'opzione '*Edita/Mostra file di testo*' dal menu FILE.
- Anche **CROSSTAB**, legge un file di parametri (CROSS.PAR), che specifica quale sia il file dei dati, quali variabili vadano caricate, le loro caratteristiche, ecc. Una volta caricate le variabili, **CROSSTAB** continua in modo interattivo chiedendo all'utente quali variabili vadano incrociate, i filtri da utilizzare, ecc. Gli incroci prodotti vengono registrati su di un file denominato CROSSnn.OUT.
- Il **formato di lettura** da fornire a **DISTRIB** e **CROSSTAB** segue la medesima sintassi degli altri programmi di ADDATI, illustrata nel capitolo 3.

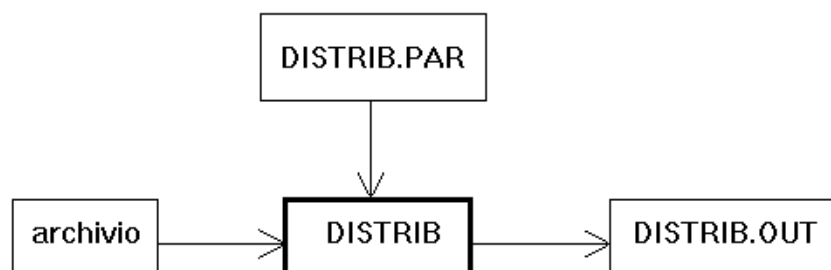
## 5.1 - DISTRIB

### Distribuzioni di variabili categoriali o continue

**Funzione** Calcola la distribuzione di un insieme di variabili categoriali o continue. Per ogni variabile categoriale viene fornita la frequenza di ciascuna modalità. Per le variabili continue vengono calcolate la media, la deviazione standard, il minimo ed il massimo. Oltre a ciò, le unità statistiche possono essere suddivisi in classi in modo automatico oppure in base a soglie fornite dall'utente.

**Limiti** La massima lunghezza di un record in input è 5120 caratteri.

Il file DISTRIB.PAR, dal quale **DISTRIB** legge i parametri di controllo dell'esecuzione, **deve risiedere nella directory corrente**.



**Figura 5.1** - I file letti e scritti dal programma **DISTRIB**.

Durante l'esecuzione **DISTRIB** compie le seguenti operazioni:

- calcola la distribuzione degli effettivi su tutte le modalità delle variabili categoriali delle quali si è comandata la lettura dal file dei dati;
- calcola media, deviazione standard, massimo e minimo assoluto delle variabili continue;
- permette di suddividere l'intervallo di variabilità di ciascuna variabile continua in un numero desiderato di segmenti di eguale ampiezza e di calcolare il numero degli effettivi che cadono in ciascun segmento (classe), tracciando l'istogramma corrispondente.

Quest'ultima possibilità è particolarmente utile quando si voglia ricodificare una variabile **continua** in un numero opportuno di classi per trattarla in un'analisi unitamente ad altre variabili categoriali.

Se gli effettivi sono distribuiti in modo pressoché uniforme su tutto l'intervallo di variabilità può essere sufficiente richiedere, in vista di una eventuale ricodifica da effettuare con il programma **RECODE** incluso nel Menù di Utilità, la suddivisione in un numero limitato di classi.

Se invece una variabile ha valori molto concentrati, chiedendo poche classi può succedere che la maggior parte delle unità statistiche si concentri in una sola di esse, mentre le altre risultano pressoché vuote. In tal caso l'operazione di ricodifica comporterebbe la perdita di una quantità rilevante di informazione. Per le variabili in queste condizioni si può allora chiedere in via esplorativa una suddivisione molto più fine (cioè in un numero di classi elevato), decidendo poi in base alla distribuzione ottenuta quali siano le soglie più opportune.

#### *La struttura del file DISTRIB.PAR*

DISTRIB.PAR è auto-illustrato. Il file include alcune linee di commento - distinguibili perché iniziano con "?" - e altre linee da riempire con i parametri richiesti. Le linee di commento



spiegano come vadano riempite le altre per adattare il file ad una prova particolare. Quando legge il file **DISTRIB** ignora le righe di commento e memorizza le altre le informazioni necessarie all'esecuzione.

Almeno all'inizio, è facile commettere degli errori nel preparare il file dei parametri. Il programma li segnala abbastanza chiaramente, permettendo di correggerli. Comunque, niente fretta e molta riflessione.

#### Listato-tipo del file di parametri DISTRIB.PAR

- ? Questo e' il file di controllo per DISTRIB.
- ? Tutti i records che iniziano con '?' sono di **commento**, e vengono ignorati dal programma.
- ? Essi spiegano come preparare gli altri records, che non iniziano con '?' e che contengono i
- ? parametri relativi alla prova che si sta conducendo.
- ? Ogni riga attiva comincia con una **parola-chiave** (come "**CASI**", "**DATA\_FILE**",
- ? "**TITOLO**", ecc.).
- ? **Le parole-chiave non vanno modificate.**
- ? Vanno invece adattati alla particolare prova i parametri che seguono le parole-chiave.
- ? **1. Numero dei casi.**
- ? Il numero dei casi specifica quanti casi vanno letti dal file di input. E' utile quando non si
- ? voglia leggerli tutti.
- ? Se si vogliono leggere tutti, si pu• sostituire il numero con la parola "**TUTTI**"
- CASI**            **TUTTI**
- ? **2. Numero delle variabili.**
- ? Il numero delle variabili specifica quante variabili vadano **effettivamente caricate** dal file
- ? ed elaborate. Esso non include né quelle che vengono saltate in lettura, specificando un
- ? formato opportuno (vedi più avanti), né l'eventuale peso attribuito al caso.
- VARIABILI**    **3**
- ? **3. Peso da assegnare a ciascun caso.**
- ? La parola-chiave **PESO** è seguita da '0' quando si voglia attribuire il medesimo peso
- ? a tutti i casi, da '1' altrimenti.
- ? In quest'ultimo caso, il peso da assegnare a ciascuna unità statistica deve essere il
- ? primo campo nel record se la lettura avviene in formato libero; può stare in qualunque
- ? posizione se viene specificato un formato di lettura.
- PESO**            **0**
- ? **4. Titolo della prova**, su di una riga, delimitato da doppi apici.
- TITOLO** "Distribuzione di alcune variabili censuarie a LUGO DI ROMAGNA"
- ? **5. Nome del file che contiene i dati** (fornire il percorso completo se sta in un'altra cartella).
- DATA\_FILE**    **LUGO.DAT**
- ? **6. (e seguenti)**    **Per ciascuna variabile**, nell'ordine di lettura dal file, fornire le seguenti
- ? informazioni:
- ? Se la variabile è categoriale (vale a dire se è rappresentata da codici interi consecutivi
- ? a partire da 1):
- ? • un record introdotto dalla parola-chiave "**VARIABILE**", seguito dal **nome della**
- ? **variabile** (max. 35 caratteri) e dal **numero delle sue categorie**. Usare lo spazio
- ? come separatore.

? • tanti records quanti sono necessari, introdotti dalla parola-chiave "**CATEGORIE**",  
 ? in cui si elencano le labels (nomi) delle categorie (max. 20 caratteri ciascuna).  
 ? Separare le labels con degli spazi. Andare a capo ed iniziare un altro record  
 ? "**VARIABILE**" quando le labels di tutte le categorie sono state elencate.

? Se la variabile è continua:

? • un record, ancora introdotto dalla parola-chiave "**VARIABILE**" seguita dal **nome**  
 ? **della variabile continua** (35 caratteri utili) e da uno '0', (**o da nulla**) per indicare  
 ? che si tratta di una variabile continua.

? • un record che controlla la descrizione della distribuzione della variabile (essendo  
 ? continua, non ci sono categorie).  
 ? Vi sono due modi per farlo: chiedere che l'intervallo di variabilità venga suddiviso  
 ? in un certo numero di classi e lasciare fare al programma, oppure definire in  
 ? dettaglio la sequenza dei valori di soglia desiderati.

? Un record come: "**CLASSI** 6"  
 ? impone al programma, una volta determinati i valori minimo e massimo assunti dalla  
 ? variabile, di dividere l'intervallo di variabilità nel numero richiesto di segmenti di  
 ? **eguale ampiezza**, e di contare la frequenza in ciascuno di essi.

? In alternativa, un record del tipo: "**SOGLIE** s1 s2 s3 s4"  
 ? impone al programma di suddividere il campo di variabilità utilizzando i valori di  
 ? soglia forniti. Nel caso mostrato, i quattro valori di soglia specificati determinano  
 ? cinque classi, che contengono rispettivamente:

? • i casi che assumono un valore della variabile <s1;  
 ? • i casi con valore della variabile >=s1 e < s2;  
 ? • i casi con valore della variabile >=s2 e < s3;  
 ? • i casi con valore della variabile >=s3 e < s4;  
 ? • i casi con valore della variabile >=s4.

? Si noti che ogni valore di soglia è incluso nella classe superiore.

? **ATTENZIONE!!!**

? Se il nome di una variabile o categoria include qualche spazio interno, racchiudere  
 ? il nome tra doppi apici per farlo interpretare correttamente dalla routine di lettura  
 ? (si vedano gli esempi qui sotto).

**VARIABILE** UBICAZIONE 3  
**CATEGORIE** centro frazione sparso  
**VARIABILE** "NUMERO COMPONENTI" 0  
**SOGLIE** 1 2 3 4 5 9  
**VARIABILE** ETA\_CF 0  
**CLASSI** 6

? **7. Formato di lettura dei dati.**

- ? E' una stringa delimitata da doppi apici, usata per specificare quali fra le variabili  
? incluse nel file di input vadano effettivamente caricate, quali ignorate, e quale  
? posizione occupi il peso, se presente.
- ? Il formato segue le regole abituali in ADDATI (vedi la [sintassi del formato](#) in questo  
? manuale, o l'help in linea per i programmi di Analisi Multivariata).
- ? Se la lettura avviene in [formato libero](#), ogni record in input deve essere stato preparato in  
? modo da contenere nell'ordine:
- ? • il peso del caso (se si è specificato "**PESO** 1");
  - ? • tutte e sole le variabili da leggere, separate da spazi;
- ? In tal caso, basta indicare un asterisco '\*' al posto del formato.
- ? La seguente istruzione di formato specifica che, per ciascun record:
- ? • vanno ignorati i primi tre caratteri;
  - ? • va letta la prima variabile (l'ubicazione) su due caratteri;
  - ? • vanno ignorati i 41 caratteri che seguono;
  - ? • va letta la seconda variabile (numero componenti) su tre caratteri;
  - ? • vanno ancora saltati due caratteri;
  - ? • va infine caricata la terza variabile (l'età del CapoFamiglia) su tre caratteri.

[FORMATO](#) "(3x, 2, 41x, 3, 2x, 3)"

## 5.2 - CROSSTAB

### Incroci tra variabili categoriali

**Funzione** Permette di incrociare coppie di variabili categoriali, calcolando la frequenza con la quale si presenta ciascuna combinazione di categorie delle due variabili.

Registra su file le tabelle che riportano gli incroci eseguiti, fornendo per ogni cella anche le percentuali calcolate sui totali di riga e di colonna. Viene calcolato il chi-quadro come indicatore della forza dell'associazione tra le variabili.

Sono possibili operazioni di filtro, che portano a selezionare segmenti di popolazione caratterizzati da particolari combinazioni dei valori di alcune delle variabili considerate, limitando poi gli incroci solo a tali segmenti.

Le informazioni necessarie per l'esecuzione (nome del file dei dati, nomi delle variabili, ecc.) non vengono fornite interattivamente, bensì lette dal file **<CROSS.PAR>** che viene cercato nella directory di lavoro. Tale file va preparato preventivamente dall'utente modificando opportunamente una copia del prototipo fornito nella directory di installazione di ADDATI.

**Limiti** Massima lunghezza di un record in input: 5120 caratteri

**N.B.** Si assume sempre che se una variabile ha  $n$  categorie (modalità), queste siano rappresentate dai codici 1... $n$ .

L'uso di **CROSSTAB** (in inglese gli *incroci* si chiamano *cross-tabulations*) è abbastanza semplice. Si prepara mediante l'editor di ADDATI il file CROSS.PAR che contiene i parametri di controllo della prova. Quando si lancia **CROSSTAB** dal menu di ADDATI, se l'utente lo richiede viene aperta l'eventuale copia di CROSS.PAR presente nella directory di lavoro; se non ne esiste alcuna, viene caricata la copia-prototipo distribuita con il pacchetto.

Eventuali errori presenti in CROSS.PAR vengono segnalati e corretti nel modo già descritto per **DISTRIB**. Poi il programma richiede interattivamente (la figura 5-3 mostra la schermata di esecuzione) quali incroci si vogliano calcolare. Il nome del file dei dati, che include le variabili da incrociare (due o più), viene specificato nel file CROSS.PAR. I risultati vengono scritti sul file CROSSnn.OUT, che può essere esaminato con l'editor **addaedit** per condurre l'interpretazione.

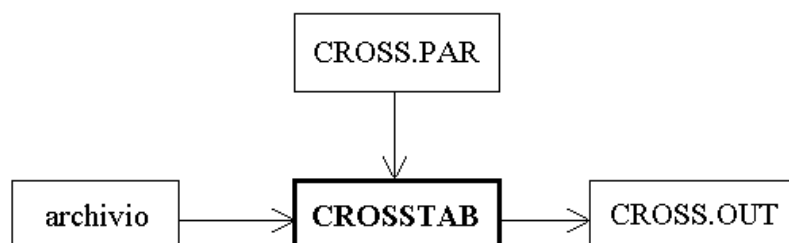


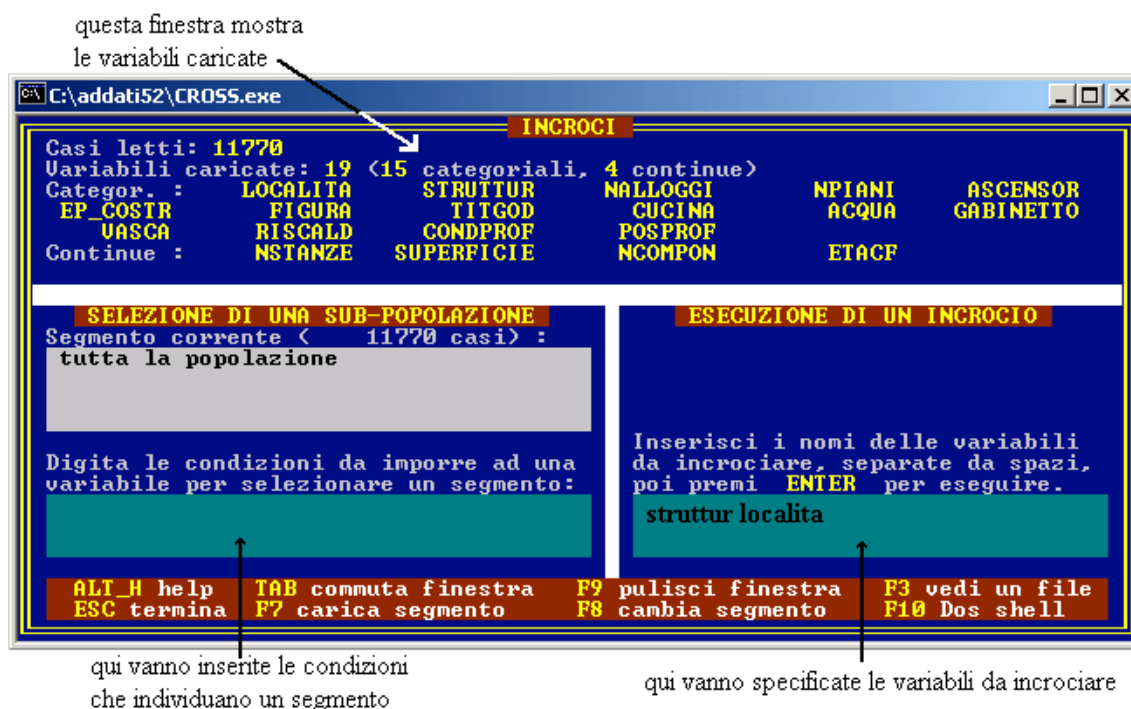
Figura 5.2 - I files letti e scritti dal programma **CROSSTAB**.

Oltre ad eseguire gli incroci richiesti, **CROSSTAB** fa quanto segue.

- Controlla che le variabili lette non abbiano valori fuori codifica. Se in qualche record una delle variabili da incrociare, con  $n$  categorie, presenta il valore 0 (che spesso indica un dato

mancante) o un valore maggiore di n (ad esempio, per un errore di digitazione dei dati) il record viene escluso dal calcolo dell'incrocio. I numeri d'ordine dei records esclusi vengono registrati in CROSSnn.OUT, il che permette di controllare gli errori (ad esempio, con **SHOWREC**) ed eventualmente di correggerli.

- Permette di selezionare tutti gli individui (una *sub-popolazione* o *segmento*) che assumono valori opportuni di una o più variabili (dette *variabili di filtro*): su ciascun segmento così selezionato si possono poi calcolare gli incroci desiderati. Per selezionare i casi voluti si possono combinare **quattro tipi di filtro**, che consentono rispettivamente di scegliere:
  1. tutti e soli i casi in cui la variabile di filtro assume un valore **superiore** ad un valore indicato: ad esempio, "> 3" porta ad accettare tutti i valori maggiori di 3;
  2. tutti e soli i casi in cui la variabile di filtro assume un valore **inferiore** ad un valore indicato: ad esempio, "< 3" porta ad accettare tutti i valori minori di 3;
  3. tutti i casi in cui la variabile di filtro assume un valore indicato: ad esempio, "4" seleziona solo i casi in cui la variabile di filtro ha valore 4;
  4. tutti i casi nei quali la variabile di filtro assume un valore **compreso** tra due valori indicati.



**Figura 5-3** La schermata di esecuzione di **CROSSTAB**.

È anche possibile leggere delle variabili continue, *che possono però essere usate esclusivamente come variabili di filtro*: non essendo categoriali non è possibile incrociarle, a meno che non siano prima state ricodificate in classi con **RECODE**. Ad esempio, si possono scegliere tutti i capifamiglia di età superiore a 45 anni, tutti gli alloggi con superficie inferiore a 100 mq., ecc.

E' possibile filtrare rispetto a più variabili (fino a quattro), *calcolando poi sul segmento così ottenuto un numero di incroci a piacere*.

Ad esempio, si può avere interesse ad analizzare la relazione tra titolo di godimento dell'alloggio e stato dei servizi interni, per valutare se sia mediamente vero che nuclei in proprietà ed in affitto fruiscono di condizioni comparabili: basterà un semplice incrocio tra titolo e servizi. Si può tuttavia avere qualche ragione di ritenere che la relazione da analizzare dipenda dall'ubicazione dell'alloggio, sia cioè diversa in zona centrale urbana o in zona agricola. Si selezioneranno allora gli alloggi di ubicazione centrale e **solo per questi** si farà l'incrocio tra

titolo e servizi; si procederà poi in modo analogo per gli alloggi in zona agricola, e si **confronteranno poi i due incroci ottenuti**.

Una ricerca, anche se poggia su obiettivi ed ipotesi formulati sin dall'inizio, segue di solito un percorso operativo determinato **dall'accumulo di passi intermedi**, ciascuno dei quali orienta i successivi.

Avendo presente questo, è consigliabile evitare di incrociare in modo esaustivo tutte le variabili a disposizione, producendo un insieme molto numeroso di tavole tra le quali è poi difficile raccapazzarsi.

È invece opportuno esercitare la propria intelligenza formulando, in base alle ipotesi ed agli obiettivi, un **piano di incroci** appropriato, volto ad indagare gli aspetti ritenuti più rilevanti, le associazioni più consistenti tra le variabili considerate, ecc. Analizzando i risultati ottenuti tale piano può essere poi approfondito od integrato con dei filtri opportuni.

Nel formulare un piano d'incroci vanno tenute presenti le distribuzioni delle variabili su cui si opera. Una volta individuati dei gruppi di variabili sensibilmente associate, è naturale approfondire l'indagine impostando un'analisi multivariata, nella quale vengono simultaneamente elaborati i valori di più variabili, non di due sole. Si vedano in proposito i capitoli 6 e 7.

La sequenza naturale è dunque: **DISTRIB - CROSS- Analisi Multivariata**.

*Un esempio, giusto per introdurre qualche considerazione teorica*

La tabella 5.1 mostra l'incrocio tra la dimensione del nucleo (variabile categoriale con quattro modalità) e l'età del capofamiglia per l'insieme delle famiglie residenti nel Centro Storico di Venezia alla data del Censimento 1981.

**Dimensione del Nucleo** (DMN1-4): 1, 2, 3, 4 o più componenti.

**Età del capofamiglia** (ECF1-3): < 35 anni, 35-55 anni, > 55 anni.

	ECF1 < 35 anni	ECF2 35-55 anni	ECF3 > 55 anni	totale
DMN1 (1 comp.)	1341	2852	9689	13882
DMN2 (2 comp.)	1680	2924	3521	8125
DMN3 (3 comp.)	1325	3693	1749	6767
DMN4 (>3 comp.)	1066	2526	1628	5220
totale	5412	11995	16587	33994

**Tabella 5.1** – Tutte le famiglie nel C.S. di Venezia, Censimento 1981. Incrocio tra la dimensione del nucleo e l'età del capofamiglia

La tabella 5.1 è detta *tavola degli effettivi*: essa conta quante famiglie assumono una certa combinazione di valori delle due variabili incrociate. Ad esempio,

- la cella (1,1) mostra che vi sono 1341 famiglie di un solo componente (ovviamente il capofamiglia) di età inferiore ai 35 anni.
- la cella (4,3) mostra che vi sono 1628 famiglie con quattro o più componenti e CF di età superiore a 55 anni.

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

Gli effettivi corrispondenti ai marginali possono essere distribuiti in molti modi sulle celle. Il numero di celle il cui valore può essere fissato arbitrariamente, rispettando i vincoli posti dai marginali, rappresenta i **gradi di libertà** della tabella. Una tavola con  $m$  righe ed  $n$  colonne ha  $(m-1)(n-1)$  gradi di libertà.

La **tavola di conteggio** viene trasformata in una **tavola di frequenze percentuali** dividendo tutti i suoi valori per il totale degli effettivi (che nel nostro caso sono 33994).

	ECF1 < 35 anni	ECF2 35-55 anni	ECF3 > 55 anni	$f_i$
DMN1 (1 comp.)	0.039	0.084	0.285	0.408
DMN2 (2 comp.)	0.049	0.086	0.104	0.239
DMN3 (3 comp.)	0.039	0.109	0.051	0.199
DMN4 (>3 comp.)	0.031	0.74	0.048	0.154
$f_j$	0.159	0.353	0.488	1.0

**Tabella 5.2** - Le frequenze percentuali osservate.

Ogni cella riporta la **percentuale** delle famiglie nelle quali si riscontra la particolare combinazione di caratteri che caratterizza la cella.

**Esempio** Le famiglie con un componente e CF < 35 anni sono il 3.9% del totale.

Ogni valore marginale rappresenta la frequenza della corrispondente categoria di una variabile, **a prescindere dal valore assunto dall'altra**.

**Esempio** Le famiglie che hanno un solo componente, a prescindere dall'età del CF, sono il 40.8% di tutte le famiglie residenti; le famiglie con CF minore di 35 anni sono il 15.9% di tutte le famiglie.

Ci proponiamo di verificare se esistano spiccate **associazioni** tra alcune categorie delle due variabili.

- Il fatto di osservare che una famiglia ha un solo componente, **aumenta la probabilità** che il CF risulti essere anziano?
- Il fatto di sapere che una famiglia ha quattro o più componenti **rende meno probabile** il fatto che il suo CF sia anziano?

Ovviamente, il raffronto va fatto con le frequenze che **ci aspetteremmo per la seconda variabile se non avessimo osservato il valore della prima**.

In altri termini: la conoscenza del valore assunto da una delle due variabili **apporta informazione** sulla distribuzione dell'altra?

Per rispondere dobbiamo calcolare i valori delle frequenze di cella che ci aspetteremmo **se le due variabili incrociate fossero indipendenti**.

Per **indipendenti** intendiamo che la frequenza sulle categorie di una variabile **rimane all'incirca la stessa** per qualunque valore dell'altra variabile. Ad esempio, se le variabili fossero indipendenti la percentuale di CF anziani dovrebbe essere circa il 48.8% e quella di CF giovani circa il 15.9% **qualunque sia il numero dei componenti la famiglia**. Si tratta delle frequenze che si osservano sull'intera popolazione. 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**, cioè da quella che ci attenderemmo nell'ipotesi di indipendenza.

Dobbiamo dunque costruire la tavola che ci aspetteremmo se le variabili fossero indipendenti e confrontarla con quella osservata, dopo aver definito un opportuno **indicatore dello scostamento globale tra le due tavole**.

#### Costruzione della tavola attesa nell'ipotesi di indipendenza

Prendiamo come esempio la cella (1,1).

La frequenza delle famiglie di un componente è  $f_{1.} = 0.408$  (cioè il 40.8% di tutte le famiglie). Tra queste, quante avranno il CF con meno di 35 anni? **Se le due variabili sono indipendenti** la probabilità di avere CF giovane **deve essere la stessa** che caratterizza la popolazione intera, cioè  $f_{.1} = 0.159$  (il 15.9%). Dovrebbe dunque avere un CF giovane il 15.9% del 40.8% di tutte le famiglie. Detta  $f_{11}^e$  la frequenza attesa nella prima cella (l'apice 'e' sta per '**expected**') si ha

$$f_{11}^e = f_{1.} * f_{.1} = 0.408 * 0.159 = 0.068$$

Tale valore risulta sensibilmente maggiore della frequenza effettivamente osservata che è 0.039, come mostra la Tabella 5.2.

**Date le distribuzioni marginali delle variabili**, le famiglie con un componente e CF giovane dovrebbero essere il 6.8% nell'ipotesi di indipendenza; sono invece solo il 3.9%. Dunque le due modalità (1 comp., < 35 anni) risultano **meno associate** di quanto ci si attenderebbe, e le due variabili non sembrano indipendenti. Ma se è vero che esse presentano un'associazione, quanto forte è?

In generale, la frequenza **attesa** nella cella ( $i,j$ ) è  $f_{ij}^e = f_{i.} * f_{.j}$ , pari al prodotto delle corrispondenti frequenze marginali. Essa va confrontata con la frequenza empiricamente osservata  $f_{ij}^o$ .

	ECF1	ECF2	ECF3	$f_{i.}$
	< 35 anni	35-55 anni	> 55 anni	
DMN1 (1 comp.)	0.068	0.144	0.199	0.408
DMN2 (2 comp.)	0.038	0.084	0.117	0.239
DMN3 (3 comp.)	0.032	0.070	0.097	0.199
DMN4 (>3 comp.)	0.024	0.054	0.075	0.154
$f_{.j}$	0.159	0.353	0.488	1.0

**Tabella 5.3** - Tavola calcolata delle frequenze **attese**  $f_{ij}^e = f_{i.} * f_{.j}$

#### Un indicatore dello scostamento globale tra tavola osservata ed attesa

Un tale indicatore

- deve tenere conto di tutte le celle;
- deve tenere 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 è più significativa 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, moltiplicata per il totale N degli effettivi:

$$\chi^2 = N * \sum_{ij} \frac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e} \quad (\text{la somma è estesa a tutti i valori di } i \text{ e } j).$$

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**, cioè quando per tutte le celle risulta  $f_{ij}^o = f_{ij}^e$ .

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

	ECF1 ( < 35 anni)	ECF2 (35-55 anni)	ECF3 ( > 55 anni)	
<b>DMN1</b>	<b>1341</b>	<b>2852</b>	<b>9689</b>	<b>13882</b>
(1 comp.)	9.7	20.5	69.8	100.0
	24.8	23.8	56.4	40.8
<b>DMN2</b>	<b>1680</b>	<b>2924</b>	<b>3521</b>	<b>8125</b>
(2 comp.)	20.7	36.0	43.3	100.0
	31.0	24.4	21.2	23.9
<b>DMN3</b>	<b>1325</b>	<b>3693</b>	<b>1749</b>	<b>6767</b>
(3 comp.)	19.6	54.6	25.8	100.0
	24.5	30.8	10.5	19.9
<b>DMN4</b>	<b>1066</b>	<b>2526</b>	<b>1628</b>	<b>5220</b>
(>3 comp.)	20.4	48.4	31.2	100.0
	19.7	21.1	9.8	15.4
	<b>5412</b>	<b>11995</b>	<b>16587</b>	<b>33994</b>
	15.9	35.3	48.8	100.0
	100.0	100.0	100.0	100.0

**Tabella 5.4** - La tavola d'incrocio presentata da CROSSTAB. In ogni cella il primo valore rappresenta il numero degli effettivi assoluti, il secondo è la percentuale sul totale di riga, il terzo la percentuale sul totale di colonna.

**Ricorda:** fatto un incrocio, si esamina la probabilità di indipendenza tra le variabili calcolata dal programma. Solo se l'associazione risulta abbastanza forte vale la pena di interpretare l'incrocio.

La tabella 5.4 mostra l'incrocio tra le due variabili come lo presenta **CROSSTAB**. Ogni cella (i,j) contiene tre valori:

- il **numero assoluto**  $x_{ij}$  dei casi che cadono nella cella;
- la **percentuale sul totale di riga**  $x_{ij}/x_{i.}$ , dove  $x_{i.}$  è il totale della riga i;
- la **percentuale sul totale di colonna**  $x_{ij}/x_{.j}$ , dove  $x_{.j}$  è il totale della colonna j;

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

**Nota:** Campione e universo- Il significato del  $\chi^2$

*Quando si ha a che fare con un campione, il  $\chi^2$  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 coincidono 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  $\chi^2$  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.*

### La struttura del file CROSS.PAR

Si tratta di un file di parametri simile a DISTRIB.PAR. Esso specifica le variabili che si vogliono far caricare da **CROSSTAB**, la loro posizione nel record, le etichette alfanumeriche delle categorie, ecc. Tuttavia, a differenza di **DISTRIB** che scriveva le distribuzioni su DISTRnn.OUT e terminava una volta letti ed organizzati opportunamente i dati, **CROSSTAB** continua l'esecuzione interattivamente, chiedendo all'utente quali incroci voglia fare, con quali filtri, ecc. E' possibile esaminare i risultati intermedi **senza uscire dal programma**, decidendo il modo più opportuno di proseguire.

### Listato-tipo del file di parametri CROSS.PAR

- ? Questo e' il file di controllo per CROSSTAB.
- ? Tutti i records che iniziano con '?' sono di commento, e vengono ignorati dal programma.
- ? Essi spiegano come preparare gli altri records, che non iniziano con '?' e che contengono
- ? i parametri relativi alla prova che si sta conducendo.
- ? Ogni riga attiva comincia con una **parola-chiave** (come **DATA\_FILE**, **TITOLO**, ecc.).
- ? **Le parole-chiave non vanno modificate.** vanno invece adattati alla particolare prova
- ? i parametri che seguono le parole-chiave.

- ? **1. Nome del file che contiene i dati**, incluso l'eventuale percorso:

**DATA\_FILE**            g:\analisi\lugo.dat

- ? **2. Titolo della prova**, in una riga, **delimitato da doppi apici**.

? Se non si vuole dare alcun titolo, non inserire nulla dopo la parola-chiave "**TITOLO**":  
**TITOLO** "Incroci su alcune variabili censuarie '91 a Lugo di Romagna"

- ? **3. Numero delle variabili da caricare.**

? Il programma prevede la lettura di un certo numero di variabili, da incrociare e/o da usare come filtri.

? Le variabili da incrociare debbono essere **categoriali**, mentre quelle da utilizzare per filtrare (cioè per selezionare una sub-popolazione o segmento) possono anche essere quantitative.

? **Attenzione!**

? Il programma richiede che una variabile **categoriale** ad n categorie assuma come codici **valori interi da 1 a n**, mentre lo '0' o blank viene assunto come dato mancante.  
 ? Fornire il **numero complessivo** delle variabili da leggere:

**N\_VARIABILI**            7

- ? **4. Nomi delle variabili e loro categorie.**

? Vanno forniti ora tanti gruppi di righe quante sono le variabili dichiarate, mantenendo l'ordine che le variabili presentano nel file. Per ogni variabile vanno fornite:

- ? • una riga intestata dalla parola-chiave "**VARIABILE**" seguita dal **nome della variabile** (fino a 12 caratteri) e, se si tratta di variabile categoriale, dal **numero delle sue categorie**.

? Se la variabile è **continua**, come numero delle sue categorie si specifichi "0" oppure nulla.

- ? • **solo se la variabile è categoriale**, una o più righe che inizino ciascuna con la parola-chiave "**CATEGORIE**", seguita dalle etichette (nomi) ordinate delle categorie. Usare le righe necessarie per fornire tutte le label di categoria, poi andare a capo per iniziare il gruppo relativo ad una nuova variabile.

? **Se la variabile è continua**, va inserito la sola riga "**VARIABILE**".

- ? Ubicazione dell'alloggio, con 3 categorie

**VARIABILE**            UBICAZIONE            3  
**CATEGORIE** centro frazione sparso

- ? Numero dei piani dell'edificio, con 5 categorie

**VARIABILE**            NPIANI            5  
**CATEGORIE** uno due 3-5 6-10 oltre10

? L'edificio è dotato di ascensore?

**VARIABILE** ASCENSORE 2

**CATEGORIE** si no

? Epoca di costruzione dell'alloggio, 6 categorie

**VARIABILE** "EPOCA COSTRUZIONE" 6

**CATEGORIE** ante19 19-45 46-60 61-71 72-81 post81

? Titolo di godimento, 3 categorie

**VARIABILE** TITGOD 3

**CATEGORIE** proprietà affitto altro

? Superficie dell'alloggio in mq, continua. Viene indicato '0' come numero delle categorie

**VARIABILE** SUPERFICIE 0

? Et  del CapoFamiglia: per indicare che   continua, non viene fornito il numero delle

? categorie.

**VARIABILE** ETACF

? **5. Formato di lettura dei dati.**

? E' una stringa delimitata da doppi apici, usata per specificare quali fra le variabili incluse

? nel file di input vadano effettivamente caricate, quali invece ignorate.

? Il formato segue le regole abituali in ADDATI (vedi la [sintassi del formato](#) in questo

? manuale, o l'help in linea per i programmi di Analisi Multivariata).

? Se la lettura avviene in [formato libero](#), ogni record in input deve essere stato preparato in

? modo da contenere, nell'ordine, tutte e sole le variabili da leggere, separate da spazi.

? In tal caso, basta indicare un asterisco '\*' al posto del formato.

? La seguente istruzione di formato specifica che, per ciascun record:

? • vanno ignorati i primi tre caratteri;

? • vanno letti due valori, di due caratteri ciascuno, assegnati alle variabili "ubicazione" e  
? numero di piani";

? • vanno ignorati i cinque caratteri successivi;

? • ...e cos  via.

**FORMATO** "(3x,2\*2,5x,2,4x,2,2,10x,4,15x,3)"

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

---

### 6.1 - TYPOLOG

---

#### Conversione di variabili categoriali in forma binaria e determinazione di tipi

**Funzione** Consente di **ricodificare in forma binaria** una tavola di variabili categoriali (si veda il par. 4.1), preparandola così per l'analisi delle Corrispondenze.

Oltre ad operare la semplice conversione, lasciando inalterato il numero dei casi, è anche possibile **aggregare questi in tipologie**: tutte le unità che presentano gli stessi valori nelle variabili scelte dall'utente come **attive** vengono assegnate alla stessa tipologia. Quando si opera su insiemi molto numerosi di unità elementari il riconoscimento di tipologie consente di ridurre fortemente il numero dei casi sottoposti all'Analisi delle Corrispondenze (le tipologie appunto, e non più le unità).

**Limiti** Max. lunghezza di un record in input: **5120 caratteri**.

**Consigli** Teoricamente non esistono limiti al numero dei casi elementari, né al numero delle tipologie che si possono costruire. Tuttavia, va tenuto presente che un numero di variabili attive troppo elevato - e con un alto numero di categorie ciascuna - può dar luogo ad un numero eccessivo di tipologie, pesante o proibitivo per **ACORR** o **NONGER**.

La tentazione - cui spesso indulge il principiante - di inserire come attive o supplementari tutte le variabili a disposizione può dar luogo a prove pesantissime, i cui risultati sono spesso confusi e di difficile interpretazione. Si tenga presente che il numero delle variabili effettivamente sottoposte ad **ACORR** - e che compariranno nei profili scritti da **NONGER** - è pari *alla somma delle categorie* delle variabili sottoposte a **TYPOLÓG**. Lo scopo di un'analisi esplorativa, che è di estrarre informazione utilizzabile da un insieme di dati tanto numeroso da non poterlo esaminare direttamente, non va vanificato costruendo tavole enormi: l'abilità dell'analista si manifesta soprattutto nella costruzione di **tavole essenziali**, determinate anche a seguito di più tentativi.

**TYPOLÓG** è il primo programma da utilizzare quando si vogliono analizzare tavole di variabili di tipo categoriale (qualitativo); la sequenza prevede poi l'esecuzione dell'Analisi delle Corrispondenze (**ACORR**, che elabora la tavola binaria scritta da **TYPOLÓG**), seguita da una procedura di classificazione.

L'input standard per l'Analisi delle Corrispondenze è una tavola ottenuta affiancando una o più tavole di contingenza, che contano le stesse unità elementari (famiglie, alloggi, imprese, ecc.).

Quando si lavora con dati disaggregati, spesso si ha a che fare con tavole in cui una riga descrive un'unità elementare mediante variabili **qualitative**. In tal caso le variabili vanno ricodificate in forma disgiuntiva completa (una colonna per ciascuna modalità: il valore è 1 se l'unità ricade in quella categoria, 0 altrimenti). La tavola binaria che ne risulta può

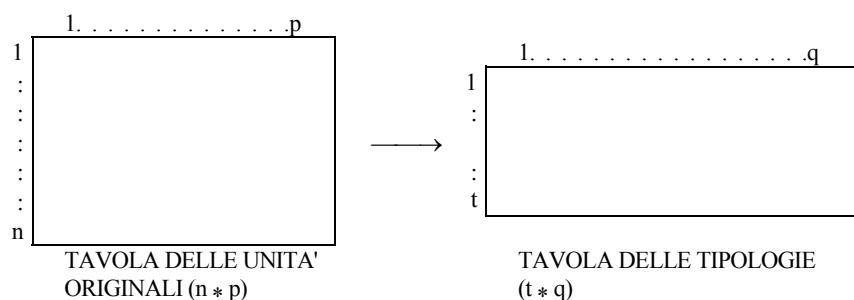
venire sottoposta ad un'Analisi delle Corrispondenze, nota in questo caso come "Analisi delle Corrispondenze Multiple".

**Nota:** Qui il termine binario si riferisce al tipo di codifica e significa solo che la tavola consiste di valori '0' e '1'. Il file è un file di testo.

La conversione in forma binaria è necessaria in quanto **ACORR** non può elaborare direttamente variabili qualitative, i cui valori non hanno generalmente un significato numerico in senso stretto. E' invece possibile trattare numericamente tavole binarie, nelle quali 0 rappresenta l'assenza di un determinato carattere mentre 1 ne definisce la presenza.

Le variabili in input possono essere **attive** o **supplementari**: le tipologie sono definite come combinazioni delle sole variabili attive.

**TPOLOG** calcola le tipologie determinate da tutte le possibili combinazioni dei valori delle variabili **attive** ed assegna ogni unità alla tipologia corrispondente. Viene registrata la tavola da sottoporre ad **ACORR**: ogni sua riga rappresenta una tipologia **pnderata con il numero delle unità ad essa assegnate**. Tutte le unità assegnate ad una tipologia assumono per definizione la stessa modalità di ogni variabile attiva, mentre si possono distribuire diversamente sulle categorie delle variabili supplementari (se ve ne sono).



**Figura 6-1** **TPOLOG** converte la tavola di partenza  $n \times p$  che descrive  $n$  oggetti mediante  $p$  variabili categoriali in una tavola  $t \times q$  che descrive le  $t$  tipologie incontrate ( $q$  è pari alla somma delle categorie delle  $p$  variabili descrittive).

La trasformazione è schematizzata nella figura 6-1. Si può provare che un'Analisi delle Corrispondenze condotta sulla tavola delle tipologie dà *risultati equivalenti* ad una condotta sulla tavola delle unità elementari (in forma binaria). La prima, tuttavia, offre un vantaggio computazionale spesso notevole, specialmente nella fase di classificazione.

Tutte le variabili nella tavola di input **devono essere categoriali**. Le eventuali variabili quantitative vanno ricodificate in forma categoriale (con il programma **RECODE**) prima di utilizzare **TPOLOG** per eseguire la codifica binaria. Per maggiori dettagli sul programma **RECODE** si veda il paragrafo 2-7.

**TPOLOG** può essere utilizzato con due finalità:

- solo per codificare in forma binaria una tavola di variabili categoriali;
- per assegnare inoltre ciascuna unità alla tipologia appropriata: la successiva procedura di classificazione potrà così elaborare le tipologie (aggregazioni di unità con caratteristiche uguali) e non le singole unità statistiche, con evidenti benefici dal punto di vista computazionale.

Questo secondo utilizzo del programma è particolarmente indicato quando si abbia a che fare con tavole di dati formate da moltissime unità (dati censuari disaggregati, dati raccolti tramite indagini campionarie, ecc.).

**Nota:** Partendo da *TYPOLÓG* sono stati elaborati (con un PC 486) i dati disaggregati dell'ultimo Censimento di Addis Abeba (237.917 famiglie). Dalla tavola iniziale sono state ricavate 749 tipologie sulla base dei valori delle variabili scelte per descrivere gli alloggi e 3946 tipologie definite in rapporto alle condizioni socio-demografiche delle famiglie.

**Esempio** Supponiamo di dover codificare tre variabili (tutte attive) che abbiano nell'ordine 4, 3 e 4 modalità. Il massimo numero di tipologie ottenibili combinando in tutti i modi possibili i valori delle variabili è  $4 \cdot 3 \cdot 4 = 48$ . Tutte le unità che presentano gli stessi valori delle variabili attive sono considerate **equivalenti** e vengono assegnate ad una medesima tipologia (le eventuali variabili supplementari possono assumere valori diversi ma hanno una funzione puramente descrittiva e non contribuiscono alla definizione delle tipologie). Solo le variabili attive sono realmente importanti nella valutazione dei diversi "comportamenti" delle unità indagate. La loro scelta rappresenta certamente la più importante e delicata ipotesi di lavoro nell'ambito di un'analisi.

**Nota:** Quando alcune variabili sono di tipo categoriale l'intera tavola deve essere riportata alla medesima scala di misura: la conversione in forma binaria è necessaria ed utile. Si tenga presente che quando il numero delle unità statistiche rimane nell'ordine di qualche centinaio la costruzione delle tipologie non comporta particolari vantaggi (in termini di tempi di elaborazione) nella successiva fase di classificazione. Conviene in questo caso eseguire una semplice ricodifica binaria, senza ridurre il numero dei casi passando alle tipologie.

### I parametri richiesti per l'esecuzione

Le domande che seguono vengono proposte nell'ordine dal programma (alcune possono non apparire, a seconda delle risposte date dall'utente alle domande precedenti).

Nell'illustrare il significato delle domande e le risposte da fornire si coglierà l'occasione per sviluppare una esercitazione a partire da uno dei file forniti come esempio (vedi l'Appendice): si tratta di *DEGRADO.DAT*. Le parti di testo relative a tale esercitazione saranno contraddistinte dal simbolo "📁".

📁 *DEGRADO.DAT* ha 200 records relativi ad altrettanti edifici. Ogni record descrive, mediante punteggio, lo stato di conservazione delle seguenti cinque componenti edilizie:

1. il tetto
2. la struttura portante
3. gli intonaci
4. gli infissi
5. i pavimenti delle parti comuni

Eccezion fatta per la struttura portante, il punteggio va da 1 a 4 denotando uno stato di degrado crescente. Il significato è il seguente:

1. condizione buona
2. la componente necessita di manutenzione ordinaria
3. è necessario un intervento di manutenzione straordinaria
4. la componente andrebbe sostituita

La condizione della struttura è invece codificata da 1 a 3, con il significato seguente:

1. condizione buona
2. è necessario un intervento di consolidamento
3. del tutto fatiscente, da demolire

Obiettivo dell'analisi è di raggruppare i 200 edifici in gruppi che possano definire implicitamente altrettante *classi di degrado riscontrabili in quel contesto*. Le categorie delle variabili descrittive potrebbero teoricamente dar luogo a  $4*3*4*4*4 = 768$  tipologie di degrado, ma è probabile che di fatto se ne incontrino parecchie di meno dato l'ovvio legame esistente tra i caratteri rilevati (ed anche perché le unità sono solo 200!). Ad esempio, è praticamente impossibile che in un edificio la struttura sia fatiscente e tutte le altre componenti in buone condizioni, o che alcune componenti siano ottime ed altre pessime, dato che gli interventi di risanamento tendono ad interessarle globalmente. Una volta costruita, la tavola delle tipologie può essere sottoposta all'Analisi delle Corrispondenze e ad una Classificazione che aggregi tipologie simili in un numero di classi contenuto.

**I parametri per questa analisi verranno letti :**

1. da tastiera
2. dal file TYP.PAR dove sono stati automaticamente salvati i parametri relativi ad un'analisi precedente, che possono ora venire modificati

**Nota:** L'uso dei files \*.PAR è descritto in dettaglio nel capitolo 3.

**Un titolo o commento per l'analisi :**

Si può scrivere un titolo o un commento (fino a 2000 caratteri) che verrà posto come intestazione del file di uscita (premere solo ↵ se non si desidera registrare alcun titolo).

**Nome del file dei dati :**

Inserire il nome del file contenente la tavola dei dati in input. Se si fornisce un nome errato si può correggerlo. Se il file non si trova nella directory corrente va fornito il path completo per rintracciarlo.



Il nome da fornire è "DEGRADO.DAT".

**Numero totale dei CASI (unità elementari) :**

Inserire il numero dei casi che debbono essere letti dal file di input ("TUTTI" comanda la lettura di tutti i casi presenti nel file dei dati). È possibile caricare ed elaborare solo una parte delle unità presenti nel file di input.



Si può digitare "200" oppure "TUTTI".

**Numero totale delle VARIABILI :**

Inserire il numero **complessivo** delle variabili categoriali - **sia attive che supplementari** - che dovranno essere lette ed utilizzate nelle analisi.




Nel file di input le variabili possono essere *già codificate in forma binaria* (detta anche *disgiuntiva completa*) oppure essere in forma categoriale. Nel primo caso lo scopo sarà quello di raggruppare opportunamente le unità in tipologie; nel secondo l'utente potrà anche solo richiedere una ricodifica binaria.


In ogni caso, **il numero che va fornito qui è quello delle variabili considerate in forma categoriale**: ogni variabile già codificata in forma binaria conta come una, anche se di fatto essa dà luogo a tante colonne della tavola quante sono le sue categorie.

**Nel file di input le variabili sono codificate :**

- 1. in forma DISGIUNTIVA COMPLETA;**
- 2. in forma CATEGORIALE** (cioè per ciascuna il codice parte da 1 e procede per interi successivi);
- 3. in forma LIBERA** (cioè le diverse categorie sono contraddistinte da codici interi non consecutivi o da codici alfanumerici).

 Nel nostro esempio sul degrado le variabili sono 5.

Ribadiamo che le variabili debbono essere di tipo categoriale (qualitative).

 Tutte le variabili sono in forma categoriale, con codici interi che vanno da 1 a 3 per la struttura e da 1 a 4 per le altre componenti.

#### **Forma disgiuntiva completa (o binaria)**

Ogni variabile viene suddivisa in tante variabili binarie (con valore '0' o '1') quante sono le sue categorie. Per ciascuna variabile ogni unità prende il valore 1 in corrispondenza ad una categoria assunta, 0 per le altre.

#### **Forma categoriale**

Ogni variabile assume solo valori interi. Se ha n categorie, i codici validi vanno da 1 a n.

#### **Codifica libera**

Una categoria può essere rappresentata da un codice qualsiasi, numerico o meno. L'utente dovrà fornire la sequenza dei codici adottati per ogni variabile, così che il programma possa riconoscerli.

**Ricorda:** Non sono consentite codifiche miste. Tutte le variabili devono essere in uno solo dei tre formati visti in precedenza. Se anche una sola variabile (ad es. il sesso) è codificata in forma libera ("M" e "F"), e tutte le altre assumono codici interi consecutivi (forma categoriale), è necessario dichiarare "codifica libera" e fornire i codici per tutte le variabili (nel caso di n modalità in forma categoriale, basta digitare "1/n").


**Lo scopo è di :**

- 1 preparare una tavola da analizzare con ACORR;**
- 2 operare solo una ricodifica binaria dei dati, registrando poi il risultato su file.**

Nel primo caso la tavola delle tipologie, determinata sulla base delle variabili attive presenti nel file di input (dichiarate successivamente), viene registrata su di un file denominato convenzionalmente ACORINP.LV, che verrà letto da **ACORR**. Il file comprende anche alcuni parametri necessari ad **ACORR** (numero dei casi e delle variabili, numero delle tipologie con i relativi indicatori, ecc.). Un altro file, - denominato TYPCLAS - registra la tipologia di attribuzione di ciascuna unità presente nel file di input.

La successiva procedura di classificazione opererà sulle tipologie: i dati presenti nel file TYPCLAS - unitamente alle informazioni salvate sul file NGCLASS dal programma **NONGER** (vedi par. 7.3) - consentiranno la determinazione della classe di attribuzione di ogni unità elementare.

Nel secondo caso non vengono costruite tipologie e quindi non si dà la distinzione tra variabili attive e supplementari. Tutte le variabili sono convertite in forma binaria e la tavola che ne risulta viene scritta su di un file denominato BINRECOD. Molte tra le domande che seguono non verranno presentate, in quanto non necessarie.

 Poiché vogliamo effettuare, se possibile, una riduzione dei casi e passare ad ACORR una tavola di dimensioni minori, si risponda '1'.


---

**Numero delle CATEGORIE di tutte le variabili :**

---

Nello stesso ordine in cui le variabili vengono lette va fornito il numero delle loro categorie. Si usino come separatori le virgole o gli spazi.

Il riferimento è alle variabili **categoriali** che descrivono le unità elementari. Se le variabili in input sono in realtà codificate in forma binaria, il numero totale delle loro categorie (vale a dire la *somma dei valori* che qui vanno forniti) deve eguagliare il numero delle colonne della tavola.

 Si risponda digitando la stringa "4 3 4 4 4".

---

**Digita nell'ordine gli indicatori per le categorie di ogni variabile :**


---

Gli indicatori (nomi delle categorie) servono ad **ACORR** che li leggerà dal file di lavoro ACORINP.LV. Ogni indicatore può essere al più di 12 caratteri e non può includere virgole o spazi, che hanno funzione di separatori. Al fine di evitare proiezioni grafiche troppo "dense" (vedi il paragrafo su **ACORR**) si consiglia di usare indicatori brevi.

E' consentita una forma compatta:

**Esempio** *Per definire 4 classi di reddito e 5 classi di età, invece di:*  
**reddito1 reddito2 reddito3 reddito4 età1 età2 età3 età4 età5**  
*si può scrivere:*  
**reddito1/4 età1/5**

**Nota:** *Vengono segnalati eventuali errori (troppi indicatori, o troppo pochi o troppo lunghi o comunque inaccettabili), in modo tale che l'utente possa provvedere alla correzione.*

 Conviene assegnare alle categorie dei nomi che consentano di distinguerle agevolmente nelle stampe successive. Poiché le categorie sono solo 19, non è necessario usare dei nomi particolarmente brevi: si può digitare ad esempio

*tetto1/4 strutt1/3 intonaci1/4 infissi1/4 pavim1/4*

o qualcosa di simile.

#### Quante variabili SUPPLEMENTARI?

Fornisci il numero delle variabili da trattare come **supplementari** (se ce ne sono; '0' = tutte le variabili sono **attive**).

Ancora una volta si fa riferimento alle variabili **categoriali** che descrivono le unità. Una variabile codificata in forma binaria è in ogni caso contata come **una**, indipendentemente dal numero delle sue categorie.

Nella successiva Analisi delle Corrispondenze (**ACORR**) le variabili attive contribuiscono alla costruzione dei fattori, mentre quelle **supplementari** sono usate solo per arricchire la descrizione. La varianza totale della tavola, che viene ripartita tra i fattori, proviene esclusivamente dalle variabili attive.

In generale, solo una parte della varianza delle variabili supplementari è "spiegata" nello spazio fattoriale generato dalle variabili attive.

Le variabili attive e supplementari possono trovarsi in qualsiasi ordine all'interno del record letto. Per distinguerle e trattarle convenientemente l'utente dovrà fornire i numeri d'ordine di quelle supplementari.



Nel nostro esempio useremo tutte le variabili come attive (e dunque la risposta da fornire è '0'). L'utente può ripetere la prova caricando i parametri salvati nel file TYP.PAR e definire qualche variabile come supplementare.

⇒ *La domanda seguente viene posta solo se vi sono variabili supplementari.*

#### Fornisci i NUMERI D'ORDINE delle variabili SUPPLEMENTARI :

Vengono ora richiesti i numeri d'ordine delle variabili supplementari, allo scopo di distinguerle dalle altre. Al solito, si fa riferimento alla forma categoriale, anche se le variabili sono di fatto in forma binaria: ogni variabile conta per una, a prescindere dal numero delle sue categorie. E' possibile una forma compatta.

*Esempio*    “ **2/5 12,13, 20** “ *significa che tra le variabili lette quelle supplementari sono la 2<sup>a</sup>, 3<sup>a</sup>, 4<sup>a</sup>, 5<sup>a</sup>, 12<sup>a</sup>, 13<sup>a</sup> e 20<sup>a</sup>. Ovviamente, bisognerà in questo caso aver prima dichiarato almeno 20 variabili totali ed esattamente 7 variabili supplementari.*

⇒ *Solo se si è dichiarata "codifica libera" in quanto almeno una variabile è codificata in forma libera, la seguente domanda viene proposta tante volte quante sono le variabili:*

#### Variabile n..... Fornisci i codici per le categorie :

L'utente deve specificare i codici adottati per ciascuna variabile nell'ordine nel quale sono state dichiarate.

I codici (alfanumerici o numerici) devono essere separati da **spazi** o **virgole**.

**Esempio** I codici per una ipotetica variabile "sesso", con modalità "M" e "F" (maschio e femmina) vanno forniti come "M F" o "M,F"; per un variabile con tre modalità rappresentate dai codici "0", "1" e "7" l'utente dovrà digitare "0 1 7" oppure "0,1,7".

---

**Ad ogni caso va associato :**

- 1** lo stesso peso
  - 2** un peso letto dal file dei dati.
- 

**Nota:** Se il peso va caricato dal file dei dati e si usa un formato di lettura libero, il contenuto del primo campo di ogni record è automaticamente assunto come il peso del caso.

Poiché ogni caso descrive un'unità elementare, la prima opzione (pesi eguali) è la più comune. Tuttavia, quando si analizzano dati raccolti in una inchiesta campionaria è talvolta necessario attribuire un peso diverso ad ogni caso, a seconda della strategia di campionamento adottata.

I pesi vanno forniti come **reali** o **interi**, non necessariamente normalizzati.



Nel nostro esempio useremo il medesimo peso per tutti i casi (opzione 1). Si tratta infatti di un piccolo insieme di edifici rilevati esaustivamente e non campionati. Ogni unità dunque rappresenta solo se stessa.

---

**Fornisci un formato per leggere il file dei dati :**

---

Il formato di lettura serve a specificare la posizione all'interno del file di input delle variabili e dell'eventuale peso.

Digita solo "\*" per leggere in **formato libero**: in tal caso ogni record in ingresso deve contenere nell'ordine, **separati da spazi**:

- il peso dell'unità (se letto dal file)
- tutte e solo le variabili da caricare.

Per una descrizione dettagliata della sintassi del formato di lettura si rinvia al capitolo 3.



Il file DEGRADO.DAT è registrato in formato libero. Si premerà quindi il tasto "\*".

**Nota:** **TYPOLÓG** non chiede di caricare gli indicatori alfanumerici (nomi) delle unità come fanno altri programmi in **ADDATI**: **TYPOLÓG** medesimo genera per le tipologie dei nomi convenzionali che passa ad **ACORR**.



Se tutti i parametri sono stati inseriti correttamente, **TYPOLÓG** dovrebbe riscontrare 86 tipologie di degrado sui 200 casi considerati. Tornati al Menu di Analisi, si preme F3 per visualizzare (tanto per farsene un'idea) il contenuto del file ACORINP.LV, scritto per **ACORR**.

Ovviamente, il numero delle tipologie cambia ripetendo la prova con una diversa definizione delle variabili attive.

### **I files scritti da TYPOLOG**

I seguenti files vengono salvati da **TYPOLOG** quando venga richiesta la preparazione di una tavola di tipologie per **ACORR**:

- **ACORINP.LV**, che registra la tavola delle tipologie rinvenute in base alle variabili attive dichiarate dall'utente. Ogni record descrive una tipologia ed ha in testa un indicatore alfanumerico. Il primo record contiene alcune informazioni generali passate ad **ACORR**.
- **TYPCLAS**, che consiste di tanti records quante sono le unità elementari. Ogni record registra il numero d'ordine della tipologia alla quale l'unità corrispondente è stata assegnata. Usato insieme a NGCLASS (scritto da **NONGER** dopo la classificazione delle tipologie), che registra la classe alla quale la tipologia è stata assegnata, permette l'attribuzione di ogni unità elementare ad una classe, consentendo di aggiungere tale informazione al file iniziale. Questa integrazione delle informazioni contenute in TYPCLAS e NGCLASS è operata dal programma **INTEGRA** del Menu di Utilità.
- **TIP.OUT**, che riporta i valori dei parametri forniti dall'utente ed elenca gli eventuali errori incontrati durante l'esecuzione.

Se è stata richiesta solo una ricodifica in forma binaria senza determinazione di tipologie la tavola ricodificata viene salvata su di un file denominato convenzionalmente **BINRECOD**.

## 6-2 Le 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: entrambi 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 minima. Ciò si ottiene ruotando in un modo ottimale - rispetto alla nuvola - il sistema di riferimento nello spazio geometrico in cui il fenomeno è rappresentato (si veda il cap. 4: ogni riga ed ogni colonna della tavola sono rappresentati come un punto in uno spazio geometrico opportunamente definito).

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

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

Entrambe le tecniche operano una trasformazione preliminare della tavola in entrata, 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.

## 6-2.1 L'Analisi in Componenti Principali (ACOMP)

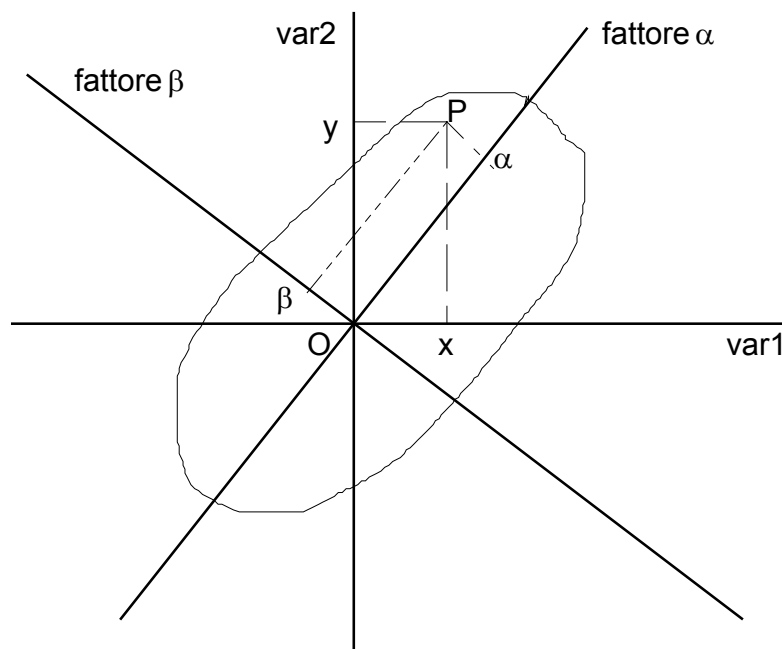
<b>Funzione</b>	Esegue un'Analisi in Componenti Principali di una tavola di descrizione quantitativa. Le variabili vengono standardizzate dal programma, così che ciascuna viene ad avere la medesima importanza nell'analisi.
<b>Limiti</b>	<p>La tavola da diagonalizzare viene preparata via via che vengono 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 è indipendente ed aumenta invece con il numero delle variabili trattate (cioè delle colonne della tavola), sia attive che supplementari.</p> <p>Di fronte ad una diagnostica di memoria insufficiente è giocoforza ridurre il numero delle variabili complessive. Comunque, poiché il programma è in grado di trattare un numero elevato di variabili (dell'ordine delle centinaia), tale riduzione è consigliata più dalla necessità di ottenere risultati interpretabili con un minimo di chiarezza che non dai limiti di memoria.</p>
<b>Consigli</b>	Anche qui come già nel caso di <b>TYPOLÓG</b> è fondamentale scegliere oculatamente le variabili da sottoporre all'analisi 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 <b>tavole essenziali</b> , determinate anche a seguito di più tentativi.

La tavola in input descrive un insieme di  $n$  unità (le righe) per mezzo di  $p$  variabili **quantitative** (le colonne). Per prima cosa il programma standardizza le variabili nel modo spiegato nel par. 4.2.

**Nota:** La standardizzazione è operata dal programma medesimo. 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 i suoi valori per la deviazione standard, così che tutte vengano ad avere varianza unitaria ed assumano nell'analisi la medesima rilevanza.

Si consideri la figura 6-2 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 allungata 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.

Consideriamo ora il fascio di tutte le possibili linee rette per l'origine. 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 cap. 4. In particolare, l'inerzia che la nuvola mantiene quando venga proiettata sui due assi coordinati  $x$  e  $y$  vale 1: è cioè pari alla varianza della corrispondente variabile, che è appunto 1 in questo caso per via della standardizzazione cui le variabili originarie sono state sottoposte. **Su qualunque altro asse l'inerzia della nuvola è in generale diversa.**



**Figura 6-2** Il punto P è rappresentato dalla coppia di coordinate  $(x, y)$  relativamente alle variabili originali, dalla coppia di coordinate  $(\alpha, \beta)$  nel nuovo sistema di riferimento.

Esiste un asse - indicato in figura con  $\alpha$  - 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  $\beta$  normale ad  $\alpha$  con un'inerzia molto minore e questo completa la descrizione nel caso di due dimensioni. Si può facilmente dimostrare che la somma delle due inerzie sugli assi  $\alpha$  e  $\beta$  **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  $(\alpha, \beta)$  rispetto ad  $(x, y)$  sta nel fatto che la piccola frazione di inerzia "*spiegata*" dall'asse  $\beta$  può essere ignorata senza una grossa perdita d'informazione; questo 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, dopo aver standardizzato, è  $p$ . E' 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: l'ammontare 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.

### *Un esempio*

---

Illustreremo l'uso di **ACOMP** e l'interpretazione dei risultati su di un esempio 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 dell'umidità risalente); 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 ha 148 records, ciascuno dei quali 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**).

#### **Variabili attive:**

1. % di alloggi che utilizzano esclusivamente i *piani alti*;
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 ed il suo peso nell'analisi. Gli altri 11 contengono i valori di altrettante variabili: le prime otto descrivono 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**.

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, ovvero 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 elevati 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 11 variabili **non sono indipendenti**. Ad esempio, le prime due (uso del piano terra o dei piani alti) sono legate dalla ovvia relazione analitica :

$$(\text{quota di alloggi ai piani alti}) = 100 - (\text{quota di alloggi ai piani terra})$$

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 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 "☞".

### *L'inserimento dei parametri di controllo dell'analisi*

Quando si lancia un'analisi bisogna fornire i parametri necessari a controllare l'esecuzione. Lo si fa rispondendo alle seguenti specifiche domande poste dal programma.

#### *A. Parametri che controllano il caricamento dei dati dal file di input*

**I parametri per questa analisi verranno letti :**

**1. da tastiera**

**2. dal file ACOMP.PAR dove sono stati automaticamente salvati i parametri relativi ad un'analisi precedente, che possono ora venire modificati**

Il significato di questa domanda è stato già illustrato in dettaglio nel cap. 3.

**Un titolo per questa analisi :**

Il titolo (che può essere un vero e proprio commento dettagliato all'analisi, lungo fino a 2000 caratteri) viene riportato in testa al file di uscita.

Si preme ↵ se non si desidera alcun titolo.

---

**Nome del file dei dati :**

---

Va fornito il nome del file che contiene la tavola da analizzare. Esso dev'essere organizzato nel modo seguente: ciascun record (o gruppo di records consecutivi) si deve riferire ad un'unità e deve contenere i valori delle variabili descrittive (**quantitative** nel caso di **ACOMP**) in un ordine fisso. Se l'analisi deve ignorare qualche variabile presente nel file l'utente deve fornire un **formato di lettura** appropriato (si veda più avanti).

***Nota:** Se il file dei dati non si trova nella directory di lavoro va fornito il path completo che consenta al programma di rintracciarlo.*

 Nel nostro esempio la risposta è "VENEZIA.DAT".


---

**Numero totale delle unità :**

---

Si tratta del numero totale dei casi (righe) inclusi nella tavola da analizzare. Specificare il numero dei casi da caricare consente di leggere e trattare - se lo si desidera - solo una parte di quelli presenti nel file.

**Non è qui possibile rispondere semplicemente "TUTTI" come per TYPOLOG:** infatti **ACOMP** (come tutti gli altri programmi di analisi) ha bisogno di conoscere - **prima di cominciare a leggere** il file dei dati - il valore di almeno due dei tre parametri "*numero totale dei casi*", "*numero dei casi attivi*" e "*numero dei casi supplementari*". Questo è necessario per determinare quando cominci la lettura di quelli supplementari che vanno trattati diversamente (sempre che ve ne siano).


 Nel caso di Venezia le unità (sezioni censuarie) sono 148.

---

**Numero totale delle variabili :**

---

Si digiti il numero di **tutte le variabili quantitative** che vanno effettivamente lette ed analizzate, siano esse **attive** o **supplementari**.

 Nel nostro esempio le variabili sono 11 (8 attive + 3 supplementari). Verranno anche caricati per ciascun caso un nome ed un peso, che non vanno però contati tra le variabili.

---

**Numero dei casi (unità, righe della tavola) supplementari :**  
**(0 = nessun caso supplementare)**

**ATTENZIONE! Nel file di input i casi supplementari - se ve ne sono -devono seguire quelli attivi.**


---

Oltre agli oggetti **attivi**, sulle cui relazioni con le variabili si costruiscono i fattori, è anche possibile includere nell'analisi un insieme di oggetti **supplementari**. Essi non hanno parte alcuna nella determinazione degli assi fattoriali, ma vengono però collocati nello spazio di rappresentazione in base ai valori che assumono per essi le variabili descrittive e possono venire proiettati sui piani fattoriali. Per ciascuno di essi il programma calcola i **contributi relativi**, mentre i **contributi assoluti** sono ovviamente nulli (si veda il cap. 4). Si può pensare che gli elementi supplementari vengano trattati assegnando loro nei calcoli un peso nullo: hanno un loro comportamento specifico, determinato dai valori delle variabili, ma non influenzano il risultato dell'analisi che è determinato dai soli oggetti attivi.

L'obiettivo è di raccogliere ulteriori informazioni in base al modo in cui questi punti supplementari si collocano rispetto a quelli attivi.

**Esempio** *Se le righe attive descrivono il comportamento dell'insieme dei comuni di una regione in un dato anno, quelle supplementari possono descrivere i medesimi comuni ad un anno diverso, consentendo una visualizzazione qualitativa delle variazioni intervenute, che può suggerire degli elementi di interpretazione.*

**Nota:** *Naturalmente, oggetti attivi e supplementari debbono essere descritti dalle stesse variabili (le colonne della tavola).*

 Nel nostro esempio tutte le sezioni censuarie sono attive e la risposta è dunque "0".

#### **Quante variabili supplementari?**

**Digita il numero delle variabili da trattare come SUPPLEMENTARI (se ve ne sono; 0 = tutte le variabili vanno considerate attive)**

Le variabili **attive** contribuiscono alla costruzione dei fattori, mentre quelle **supplementari** sono usate a solo scopo descrittivo. Alla varianza totale della tavola (inerzia della nuvola), ripartita tra i fattori, contribuiscono solo le variabili attive.

In generale solo una parte della varianza di una variabile supplementare è spiegata nello spazio fattoriale generato dalle variabili attive. In altri termini, una variabile supplementare non è in generale esprimibile come combinazione lineare di quelle attive.

Variabili attive e supplementari possono stare in qualunque posizione nel record in lettura. L'utente dovrà fornire i numeri d'ordine delle variabili supplementari per mettere il programma in grado di trattarle opportunamente.

 Venezia: ci sono 3 variabili supplementari, quelle relative allo status del capofamiglia.

#### **Vuoi la stampa della tavola in entrata? ( 1 = sì; 2 = no )**

Una copia della tavola dei dati letti può venire stampata su ACOMP.OUT nel caso ciò faciliti l'interpretazione dei risultati. Conviene non richiederlo quando la tavola sia molto grande o quando non ci sia un effettivo interesse all'esame dei dati in input.

#### **Gli indicatori alfanumerici (nomi) di riga**


- 1. verranno digitati da tastiera**
- 2. saranno letti dal file di input (se viene usato il formato LIBERO essi debbono trovarsi all'inizio del record corrispondente).**

La domanda riguarda gli indicatori alfanumerici che identificheranno le unità (righe) in ACOMP.OUT e nelle proiezioni sui piani fattoriali, se richieste.

Se i nomi degli oggetti vengono letti dal file di input il formato di lettura - fornito più avanti - deve specificare la loro posizione nel record.

La lunghezza massima di un indicatore è di 12 caratteri; esso **non deve includere virgole o spazi**, che vanno usati invece come separatori. Per un uso oculato della memoria disponibile e per evitare proiezioni grafiche troppo confuse è consigliabile scegliere nomi corti, compatibilmente con la necessità di chiarezza nell'interpretazione.

**Ricorda:** se viene usato un **formato libero** ogni nome deve stare in testa al record corrispondente.


 Nell'esempio di Venezia gli indicatori delle sezioni censuarie verranno letti dal file dei dati: la risposta da fornire è pertanto "2".

**Fornisci gli indicatori alfanumerici delle variabili:**

Va fornito un nome specifico per ciascuna variabile, rispettando l'ordine che esse hanno nel record. I nomi servono a distinguere le variabili nel file d'uscita e nelle proiezioni grafiche.

Vale tutto quanto è stato prima specificato per i nomi degli oggetti: Anche in questo caso gli indicatori possono avere al più 12 caratteri e non debbono includere spazi o virgole che vanno usati per separarli. E' accettata una risposta in forma compatta (del tipo "variab1/11") che risulta però qui poco conveniente perché confonde l'interpretazione: conviene senza dubbio che ogni variabile abbia un nome ben chiaro.

Eventuali errori (troppi o troppo pochi indicatori, o comunque inaccettabili per qualunque ragione) vengono rilevati ed è consentito correggerli.

 Nel caso di Venezia si può fornire la stringa seguente:  
*p\_alto p\_terra completa carente scadente sovraff standard sottout st\_alto st\_oper st\_altro*

**Ricorda:** L'ordine degli indicatori deve riflettere la posizione delle variabili nel record.

⇒ *La domanda seguente viene posta solo se è stato prima specificato che i nomi degli oggetti vengono forniti da tastiera. Nel caso del nostro esempio la richiesta non appare, dato che gli indicatori sono letti da file.*

**Fornisci gli indicatori per gli oggetti (righe) :**

E' consentita una risposta in forma compatta:

**Esempio** *Per indicare 150 oggetti attivi e 50 supplementari senza elencarli uno ad uno si può digitare*

*attivo1/150 suppl1/50*

*e la stringa verrà automaticamente espansa nei 200 nomi*

*attivo001 ... attivo150 suppl01 ... suppl50*

Eventuali errori vengono rimarcati dal programma ed è possibile correggerli.

⇒ *La domanda seguente appare solo se si è dichiarata la presenza di variabili supplementari.*

**Fornisci i numeri d'ordine delle variabili supplementari :**

Il programma deve poter identificare le variabili supplementari tra quelle lette per poterle trattare in modo appropriato. Anche qui è ammessa una risposta in forma compatta. Ad esempio, "2/5 12,13, 20" significa che tra le variabili da caricare la seconda, terza, quarta, quinta, dodicesima, tredicesima e ventesima sono supplementari. Ovviamente, bisognerà in questo caso aver dichiarato esattamente 7 variabili supplementari ed almeno 20 variabili in tutto.



Nel caso dell'esempio su Venezia le variabili supplementari sono le tre di "status" e la risposta da fornire è "9/11".

**Ad ogni caso va assegnato :**

- 1. lo stesso peso**
- 2. un peso particolare letto dal file dei dati**

Nel caso di una tavola di descrizione quantitativa trattata con **ACOMP** l'utente può fissare il peso di ogni unità, vale a dire la sua importanza nell'analisi.

Ad esempio, nel caso di una tavola di tassi è necessario restituire ad ogni unità la sua importanza assoluta assegnandole un peso opportuno. Le medie e le varianze pesate che vengono così calcolate sono corrette ed i valori che si ottengono hanno significato e rappresentano le effettive caratteristiche globali del sistema considerato. Ad esempio, se ogni riga rappresenta una unità geografica, descritta da variabili calcolate come tassi sulla popolazione, è conveniente usare la popolazione di ciascuna unità come peso. In tal modo i valori medi delle variabili rappresentano correttamente il comportamento medio del sistema.

L'opzione "*pesi eguali*" va usata con oggetti elementari (famiglie, abitazioni, ecc.). Se si usano pesi diversi, la posizione del peso nel record va dichiarata nel formato di lettura. Se si legge in formato libero il peso è convenzionalmente il secondo campo nel record se è presente l'indicatore alfanumerico del caso (che sta al primo posto); altrimenti il peso è il primo.

**Nota:** *Il peso va fornito come intero o reale, non necessariamente normalizzato (ci penserà il programma a trasformare i pesi in modo che abbiano somma unitaria). Esso non va contato tra le variabili.*



Esempio su Venezia: poiché i valori delle variabili sono calcolati come quote sul totale degli alloggi della sezione, si assumerà come peso tale totale, che costituisce il secondo campo di ciascun record letto. La risposta è dunque "2".

**Fornisci un FORMATO per leggere i dati dal file di input :**

Il formato serve a specificare la posizione dell'indicatore del caso e del peso (se letti da file), delle variabili da caricare e di quelle da ignorare.

Digita solo "\*" se la lettura avviene in **formato libero**: in tal caso ogni record deve contenere nell'ordine, separati da spazi:

- l'indicatore del caso (se viene letto da file)
- il peso del caso (se letto da file)
- le variabili

La sintassi del formato è descritta in dettaglio nel capitolo 3 al quale si rinvia.



Il file di input VENEZIA.DAT è stato preparato in modo da poter essere letto in formato libero; va risposto quindi con un asterisco.

A questo punto il programma carica i dati, standardizza le variabili, calcola la matrice delle correlazioni e la stampa sul file d'uscita ACOMP.OUT. A partire da tale matrice vengono determinati gli assi fattoriali e gli autovalori associati. Viene posta all'utente l'alternativa seguente:


**Digita :**

**1. per esaminare le quote di inerzia spiegate dai fattori e decidere quante coordinate fattoriali registrare o stampare;**

**2. per continuare.**

Scegliendo la prima opzione compaiono sullo schermo i risultati ottenuti fino a quel momento e registrati sul file di uscita ACOMP.OUT. L'utente può esaminare un sommario dei parametri da lui stesso forniti, seguito dalla tavola dei dati in input (se ne è stata richiesta la stampa), ovvero semplicemente dai valori medi delle diverse variabili. Seguono la matrice delle correlazioni e gli autovalori (si vedano come esempio le tabelle 6.1 e 6.2). In base alla sequenza delle capacità esplicative dei diversi fattori l'utente può decidere quante coordinate fattoriali gli convenga stampare o salvare su file.

La seconda opzione, scelta di solito **dopo** aver ispezionato il file di uscita, continua l'esecuzione.

 La tabella 6-1 riporta la matrice delle correlazioni per il nostro esempio su Venezia Centro Storico.

CORRELAZIONI (* 1000)											
	p_al to	p_ter ra	comp leta	care nte	scad ente	sovr aff	stan dard	sott out	st_a lto	st_o per	st_a ltro
p_alto	1000										
p_terra	-1000	1000									
completa	599	-599	1000								
carente	-502	502	-863	1000							
scadente	-495	495	-796	382	1000						
sovraff	-634	634	-795	641	691	1000					
standard	-81	81	-116	145	35	-206	1000				
sottout	643	-643	817	-688	-670	-818	-394	1000			
st_alto	578	-578	881	-734	-731	-745	-265	857	1000		
st_oper	-651	651	-835	700	690	825	82	-824	-852	1000	
st_altro	68	-68	-171	136	151	-60	346	-145	-365	-174	1000

**Tabella 6-1** 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.

La tabella 6-2 mostra la sequenza degli autovalori per l'esempio su Venezia. Essi misurano, nell'ordine, il potere esplicativo dei fattori.

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 una sola nuova variabile costruita basta 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.

INERZIA TOTALE = 8.000009				
n.	AUTOVALORE	INERZIA SPIEGATA (%)	INERZIA CUMULATA (%)	
1	5.0113397	62.642	62.642	*****
2	1.1802406	14.753	77.395	*****
3	0.9061959	11.327	88.722	*****
4	0.6101317	7.627	96.349	*****
5	0.2914360	3.643	99.992	**
6	0.0006524	0.008	100.000	
7	0.0000121	0.000	100.000	

**Tabella 6-2** Gli autovalori associati alle componenti principali; sono una misura della capacità esplicativa delle componenti.

**Nota:** *Non sempre le cose vanno in modo così conveniente: 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.*

## B. Parametri che controllano l'output

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 di stampa è descritto più avanti.

Il programma pone alcune domande che hanno lo scopo di controllarne l'uscita. Le coordinate fattoriali calcolate per ciascuna unità vengono scritte insieme ai contributi su ACOMP.OUT (per essere stampate ed interpretate); su ACOMP.FPL (a richiesta) per il programma **FACPLAN** che mostra graficamente le proiezioni della nuvola sui piani fattoriali; su COORRIG.LV o COORCOL.LV (a richiesta) per la fase di classificazione successiva.

**Per quanti fattori vuoi stampare i contributi per le colonne (variabili)?**

**Digita il numero di fattori richiesto:**

Questa stampa andrebbe sempre richiesta perché il significato dei fattori si desume dall'esame dei contributi delle variabili; la decisione su quanti fattori valga la pena considerare dipende invece ovviamente dal loro potere esplicativo, mostrato dalla tavola degli autovalori.

Per ciascuna variabile e ciascun asse richiesto verranno stampate 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.

**Per quanti fattori vuoi stampare il contributo per gli oggetti?**

**Digita il numero di fattori richiesto :**

E' opportuno richiedere questa stampa **solo** quando si intendano interpretare i risultati facendo uso anche dei contributi relativi agli oggetti e non solo di quelli sulle variabili. Conviene invece tralasciarla quando gli oggetti siano troppo numerosi o quando non si sia interessati al comportamento di qualche particolare unità. Si eviterà così uno spreco di carta, poiché la stampa dei contributi relativi a 50 oggetti o variabili richiede un modulo.

**Nota:** *La stampa della tavola in entrata è per lo più inutile quando si tratti di una tavola di tipologie scritta da **TYPOLÓG**.*

Per ogni unità e per ciascun asse fattoriale richiesto vengono stampate le seguenti informazioni:

- la **coordinata fattoriale** dell'unità sull'asse;
- il **contributo relativo** (frazione dell'inerzia dell'unità spiegata da quel fattore);
- il **contributo assoluto** (frazione della varianza totale del fattore - misurata dall'autovalore associato - che proviene da quell'unità).



Poiché le sezioni censuarie sono ben 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.

**VARIABILI ( COLONNE ):**

**Quante coordinate fattoriali vuoi registrare su file?**

**( 0 = nessuna )**

Le coordinate fattoriali delle colonne (variabili) vanno registrate per poterle usare per una successiva classificazione che aggregi variabili a comportamento simile (cioè sufficientemente correlate sull'insieme delle unità statistiche). Comunque, per uno studio di questo tipo si possono anche utilizzare altri metodi statistici.




Poiché non intendiamo classificare le variabili la risposta è "0".

**OGGETTI ( righe ):**

**Quante coordinate fattoriali vuoi registrare su file?**

E' necessario salvare su file le coordinate fattoriali se si vuol operare una successiva classificazione degli oggetti con **NONGER** (metodi 1 o 2) o **CGA**.


Se il numero delle unità è molto elevato (alcune centinaia o migliaia) conviene limitare il numero dei fattori passati alla procedura di classificazione, che è piuttosto pesante quanto a tempo di calcolo (le cose vanno ovviamente meglio su macchine dotate di co-processore). Comunque, per non perdere troppa informazione è conveniente salvare un numero di fattori sufficiente a spiegare globalmente almeno l'80 o il 90 per cento dell'inerzia.

 Venezia: abbiamo già deciso di utilizzare per la classificazione delle sezioni quattro fattori, che spiegano il 96.3% dell'inerzia. La risposta è dunque "4".

**Quanti fattori vuoi registrare per le proiezioni grafiche sui piani fattoriali? ( 0 = non si vogliono proiezioni )**

**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).*

**Ricorda:** La proiezione su di un piano fattoriale può facilitare l'interpretazione quando quel piano spiega una 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 selezionarli. 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.

### La tavola dei contributi e la loro interpretazione

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

La tabella 6-3 mostra i contributi delle variabili, tratti da ACOMP.OUT. Il significato è descritto in dettaglio nella 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 *p\_alto*.

**INR** (**inerzia totale della variabile**): poiché tutte le variabili 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.

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

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	p_alto	996	1	125	1000	-822	675	135	63	4	3	563	317	350
2	p_terra	996	1	125	1000	822	675	135	-63	4	3	-563	317	350
3	buoni	956	1	125	1000	-927	859	171	-13	0	0	-311	97	107
4	carente	687	1	125	1000	779	606	121	114	13	11	261	68	75
5	scadenti	664	1	125	1000	765	585	117	-117	14	12	256	65	72
6	sovraff	903	1	125	1000	877	770	154	-328	108	91	159	25	28
7	standard	988	1	125	1000	139	19	4	983	966	818	-56	3	3
8	sottout	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	st_alto	834	1	125	1000	-871	758	0	-156	24	0	-227	52	0
10	st_oper	786	1	125	1000	878	771	0	-34	1	0	120	14	0
11	st_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 → della variabile	1	p_alto	996	1	125	1000	-822	675	135	63	4	3
		↑					↑	↑	↑	↑	↑	↑
		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.

**DIS** è il **quadrato della distanza** del punto-variabile dall'origine (il valore nelle stampe è moltiplicato per 1000). Si può provare che essa rappresenta la varianza della variabile e vale 1 per tutte, dato che 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 (va letta come -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 4-8). 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 *p\_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 *p\_alto*.

La tabella 6-5 mostra le informazioni concernenti le unità come appaiono in ACOMP.OUT, mentre la tabella 6-6 mostra nei dettagli il significato per la sezione censuaria n.5 ed i primi due fattori.

n.	ATT OGG	QTL	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.	VAR ATT	QTL	PESO	INR	DIS	FAT 1	CON REL	CON ASS	FAT 2	CON REL	CON ASS
num. d'ordine dell'unità →	5	5	986	8	13	13101	-3305	834	17	-1317	132	12
		↑					↑	↑	↑	↑	↑	↑
		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:

**QTL (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:

$$\text{INR} = (\text{inerzia dell'unità}) / (\text{inerzia totale})$$

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.

### *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 loro); tutte le variabili del primo gruppo risultano invece negativamente correlate a 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 cattura 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 porta ad evidenziare 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.

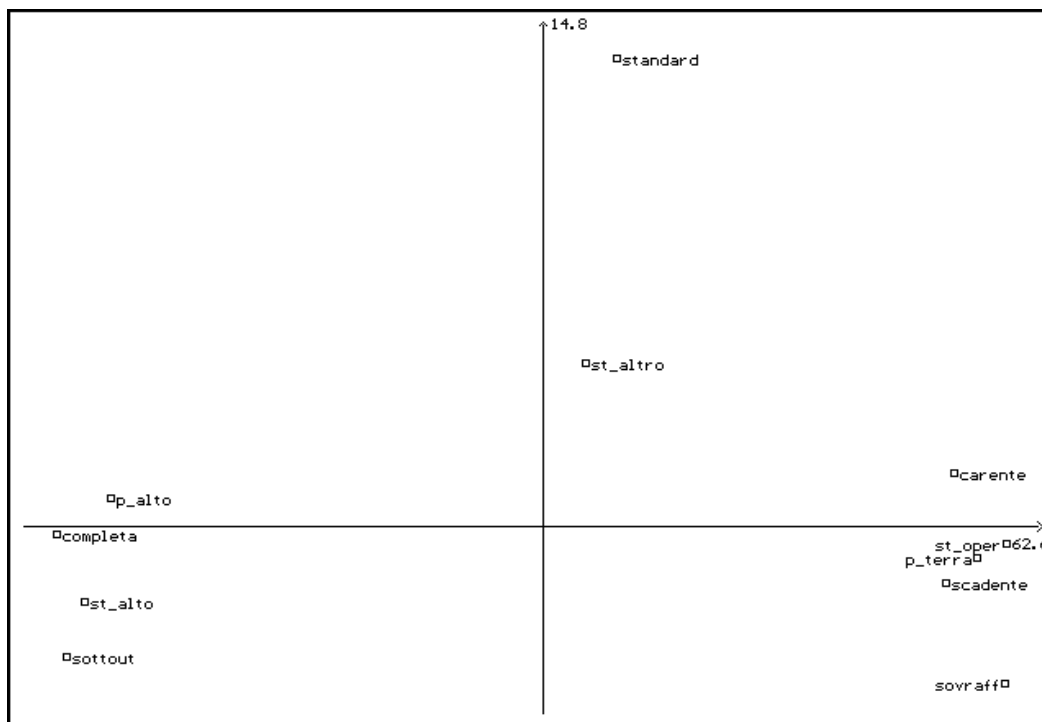
Si noti che anche la variabile supplementare *status alto* è negativamente correlata con il fattore (e quindi con le condizioni di basso disagio), mentre lo *status operaio* si proietta sul semiasse positivo e quindi si lega alle condizioni di disagio elevato.

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 6-2 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 6.1 delle correlazioni - l'affollamento standard risulta pochissimo correlato con tutte le altre variabili e necessita quindi di un fattore esplicativo per suo conto; appunto, il secondo.

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

La Figura 6-3 mostra 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 il solo obiettivo di ottenere una proiezione a tutto schermo. Come conseguenza, la varianza del secondo fattore (pari a 1.18) risulta sovrarappresentata in figura rispetto a quella del primo fattore, pari a 5.01. La nuvola è cioè molto più appiattita lungo il primo asse.



**Figura 6-3** Venezia: proiezione dei punti-variabile sul primo piano fattoriale.

#### **I files scritti da ACOMP**

- **ACOMP.OUT** è il file da stampare, che contiene le informazioni su cui si basa l'interpretazione.
- **COORRIG.LV** (scritto su richiesta dell'utente) contiene le coordinate fattoriali delle unità ed altre informazioni necessarie a NONGER per una classificazione delle unità.
- **COORCOL.LV** (scritto su richiesta dell'utente) contiene le coordinate fattoriali delle variabili ed altre informazioni necessarie a NONGER per una classificazione sulle variabili.
- **ACOMP.FPL** (scritto anch'esso su richiesta) contiene le informazioni passate a **FACPLAN** per mostrare le proiezioni sui piani fattoriali.

## 6-2.2. - L'Analisi delle Corrispondenze

### 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 trattate con un'Analisi delle Corrispondenze:

- tavole ottenute accostando fianco a fianco parecchie 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); si veda in proposito il paragrafo 4.1.

Il secondo caso è meno intuitivo, ma risulta molto interessante per le possibili applicazioni nello 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 previa ricodifica in classi (mediante **RECODE**) 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" 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 (comuni, famiglie, 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.



	Teff	mais	sorgo	altro		
area 1	40	60	50	100	250	<b>a)</b>
area 2	100	100	200	200	600	
area 3	80	120	100	200	500	
	220	280	350	500	1350	
	teff	mais	sorgo	altro		
area 1	.16	.24	.20	.40	250	<b>b)</b>
area 2	.17	.17	.33	.33	600	
area 3	.16	.24	.20	.40	500	
	.16	.21	.26	.37	1350	
	teff	mais	sorgo	altro		
area 1	.18	.21	.14	.20	.19	<b>c)</b>
area 2	.45	.36	.57	.40	.44	
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 il "mix" globale tra le coltivazioni. 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.

Vista la simmetria della tavola, la trattazione analitica può basarsi 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'ammontare 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). E' sufficiente analizzare solo una

di esse (il programma determina automaticamente quale sia la più conveniente dal punto di vista del calcolo); i risultati relativi all'altra vengono poi ricavati mediante semplici trasformazioni.

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 collapserebbe 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 sono compresi tra 1 e 0.**

**ACORR** ed **ACOMP** necessitano all'incirca degli stessi parametri per controllare l'esecuzione e pongono dunque all'utente all'incirca le stesse domande. 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 tratta di un'esecuzione concatenata a TYPOLOG?**

- 1. no - è un'analisi indipendente**
- 2. si - l'input è una tavola di tipologie prodotta da TYPOLOG**

Un'Analisi delle Corrispondenze assume come input una o più tavole di contingenza affiancate che contano le stesse unità (famiglie, alloggi, ecc.) secondo caratteri diversi.

Quando si lavora con dati urbani o regionali (specialmente dati tratti dal Censimento o da inchieste mediante questionario) si ha spesso occasione di analizzare tavole nelle quali ogni riga rappresenta un'unità elementare (un'impresa, una famiglia, ecc.), descritta da un insieme di **variabili qualitative**. In tal caso le variabili vanno ricodificate 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 può venire trattata con un'Analisi delle Corrispondenze (si parla in questo caso di **Analisi delle Corrispondenze multiple**).

Il programma **TYPOLÓG** legge il file dei dati elementari, riconosce tutte le unità che risultano identiche rispetto ai valori delle variabili attive, le aggrega in **tipologie opportunamente pesate** e scrive una tavola binaria usata come input da **ACORR**.

Lo scopo di questa domanda è dunque di informare **ACORR** sulla natura della tavola in input. Se **ACORR** non opera su di una tavola di tipologie tutti i parametri necessari per una corretta esecuzione vanno forniti da tastiera rispondendo a specifiche domande.

La risposta da fornire è dunque "2" se si intende analizzare una tavola di tipologie scritta da **TYPOLÓG** su di un file convenzionalmente denominato ACORINP.LV. Ovviamente tale file deve esistere, cioè deve essere stato prima mandato in esecuzione **TYPOLÓG** per creare la tavola delle tipologie sulla base di un numero opportuno di variabili attive (categoriali).

**Nota:** se **TYPOLÓG** è stato utilizzato solo per ricodificare in forma binaria una tavola di descrizione qualitativa senza determinare tipologie la risposta corretta da fornire è "1". Il file BINRECOD scritto in questo caso da **TYPOLÓG** non contiene alcuna informazione aggiunta in merito alle dimensioni della tavola, ai nomi delle categorie, ecc.; tali informazioni sono invece presenti in ACORINP.LV. Bisogna dunque digitare "BINRECOD" come nome del file di input e fornire da tastiera tutti i parametri richiesti.



Il lettore usi **ACORR** per elaborare la tavola delle 85 tipologie determinate da **TYPOLÓG** a partire dai dati sullo stato di conservazione di 200 edifici (file DEGRADO.DAT; si veda il par. 6.1). In tal caso va risposto "2" al quesito mostrato sopra.

INERZIA TOTALE = 2.800000

AUTOVALORE BANALE (0) = 1.000000

n.	AUTOVALORE	INERZIA SPIEGATA (%)	INERZIA CUMULATA (%)	
1	0.6722702	24.010	24.010	*****
2	0.4772470	17.045	41.054	*****
3	0.3629881	12.964	54.018	*****
4	0.2111022	7.539	61.557	*****
5	0.1811097	6.468	68.026	*****
6	0.1597108	5.704	73.730	*****
7	0.1330668	4.752	78.482	*****
8	0.1209372	4.319	82.801	*****
9	0.1105108	3.947	86.748	*****
10	0.0960970	3.432	90.180	*****
11	0.0879850	3.142	93.322	*****
12	0.0745672	2.663	95.985	*****
13	0.0664623	2.374	98.359	****
14	0.0459456	1.641	100.000	***

**Tabella 6-8** Gli autovalori associati alle componenti principali; sono una misura della capacità esplicativa delle componenti.

La tabella 6.8 mostra la sequenza degli autovalori. Va notato quanto segue.

- Il primo autovalore (pari ad 1) è banale conseguenza del passaggio ai profili operato dal programma e non ha alcuna rilevanza esplicativa.
- Vi sono solo 14 autovalori non nulli, anche se le variabili descrittive passate ad ACORR erano 19. Ciò significa che la nuvola è di fatto contenuta in uno spazio a 14 dimensioni. La dimensionalità iniziale (19) è solo apparente (e ridondante) poiché le variabili iniziali, codificate in forma disgiuntiva completa, sono legate da relazioni che diminuiscono i gradi di libertà (per ogni tipologia, i valori corrispondenti alle categorie di ciascuna delle cinque variabili categoriali iniziali sono vincolati ad assommare allo stesso valore, pari al numero degli edifici assegnati a quella tipologia).
- L'inerzia totale (pari a 2.8) è computabile in modo semplice a partire dal numero delle variabili iniziali e delle loro categorie (si veda in proposito Griguolo e Palermo, 1984, cit.).
- La capacità esplicativa dei primi fattori è in generale più contenuta di quella vista per ACOMP o di quella che si otterrebbe operando con ACORR direttamente su di una tavola di contingenza. Anche questo è un effetto della ricodifica in forma disgiuntiva completa, che ha moltiplicato il numero delle colonne della tavola passata ad ACORR introducendo dell'inerzia fittizia. Anche se la quota di inerzia spiegata dai primi fattori appare bassa, la loro rilevanza ai fini interpretativi è egualmente notevole.

Potremmo decidere qui di salvare per la classificazione 10 coordinate fattoriali (sufficienti a riassumere il 90% dell'inerzia).

### *Le tavole dei contributi e la loro interpretazione*

Le informazioni qui sotto elencate vengono scritte (su richiesta) sul file ACORR.OUT, separatamente per le righe e le colonne attive/supplementari. Poiché l'interpretazione è condotta esattamente allo stesso modo sia per le righe che per le colonne si parlerà genericamente di *punto* (riga o colonna).

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	tetto1	562	63	49	2175	-1036	493	101	-323	48	14	-210	20	8
2	tetto2	254	22	64	8091	2	0	0	462	26	10	1358	228	112
3	tetto3	391	74	45	1703	307	55	10	741	322	85	-154	14	5
4	tetto4	587	41	57	3878	1037	277	66	-1088	305	102	-128	4	2
5	strutt1	633	121	28	653	-631	609	72	-90	12	2	87	12	3
6	strutt2	587	73	45	1740	894	459	87	424	103	28	-209	25	9
7	strutt3	470	6	69	32333	1846	105	30	-3344	346	141	782	19	10
8	intonaci1	638	74	45	1703	-942	521	98	-255	38	10	-367	79	27
9	intonaci2	348	31	60	5452	-391	28	7	451	37	13	1240	282	131
10	intonaci3	625	74	45	1703	772	350	66	608	217	57	-312	57	20
11	intonaci4	628	21	64	8524	1176	162	43	-1910	428	161	563	37	18
12	infissi1	730	77	44	1597	-984	607	111	-299	56	14	-329	68	23
13	infissi2	584	36	59	4556	-33	0	0	738	120	41	1454	464	210
14	infissi3	502	46	55	3348	830	206	47	652	127	41	-753	169	72
15	infissi4	413	41	57	3878	947	231	55	-819	173	58	186	9	4
16	pavim1	664	86	41	1326	-861	560	95	-262	52	12	-266	53	17
17	pavim2	626	27	62	6407	-151	4	1	766	92	33	1844	531	253
18	pavim3	590	56	51	2571	710	196	42	757	223	67	-663	171	68
19	pavim4	615	31	60	5452	1238	281	71	-1309	314	111	329	20	9

**Tabella 6-9** I contributi delle variabili sui primi tre fattori.

**QLT** (**qualità della rappresentazione**): è la quota dell'inerzia del punto spiegata globalmente da tutti i fattori dei quali è stata richiesta la stampa. La qualità rappresenta la somma dei contributi (relativi) dei fattori in questione al punto considerato. La tabella 6-9 mostra che i primi 3 fattori spiegano i 562/1000 dell'inerzia della variabile tetto1.

**INR** (**inerzia totale del punto**): è la frazione (\*1000) dell'inerzia totale proveniente dal punto:


$$\text{INR} = (\text{inerzia del punto}) / (\text{inerzia totale})$$

dove l'inerzia totale è la somma degli autovalori (escluso quello banale). L'inerzia del punto rispetto all'origine (coincidente con il centro della nuvola) è definita come il prodotto della sua massa (PESO) per il quadrato della sua distanza (del chi-quadro) dall'origine (DIS).

**PESO** **massa del punto** nell'analisi (le masse sono scalate proporzionalmente così che la loro somma valga 1000). Rappresenta l'importanza relativa del punto considerato.

Vale quanto si è già osservato nel caso di ACOMP: in generale, quanto più INR è grande rispetto a PESO tanto più peculiare è il comportamento del punto considerato. 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, i due valori risulterebbero esattamente uguali.

Un valore di INR maggiore di PESO (come ad esempio nel caso di strutt3) significa che il punto ha una distanza dall'origine maggiore della distanza media e dunque che il suo comportamento è piuttosto peculiare (va ricordato che l'origine rappresenta il comportamento medio del sistema).

 In questo caso, ciò significa che la categoria strutt3 (struttura irrecuperabile) caratterizza un numero molto basso di edifici (come mostra il suo peso) e dunque che la sua distribuzione sui 200 edifici è necessariamente molto diversa da quella delle altre categorie.

Va sempre tenuto ben presente che la distanza tra un punto e l'origine è una misura della differenza globale che esiste tra il comportamento del punto ed il comportamento medio

dell'intero sistema (ad esempio, se le righe rappresentano i comuni di una regione il centro della nuvola rappresenta i caratteri medi dell'intera regione).

Un'Analisi delle Corrispondenze assume sempre come riferimento il comportamento medio dell'intero sistema ed analizza in qual modo e di quanto le diverse unità differiscano da esso (e l'una dall'altra).

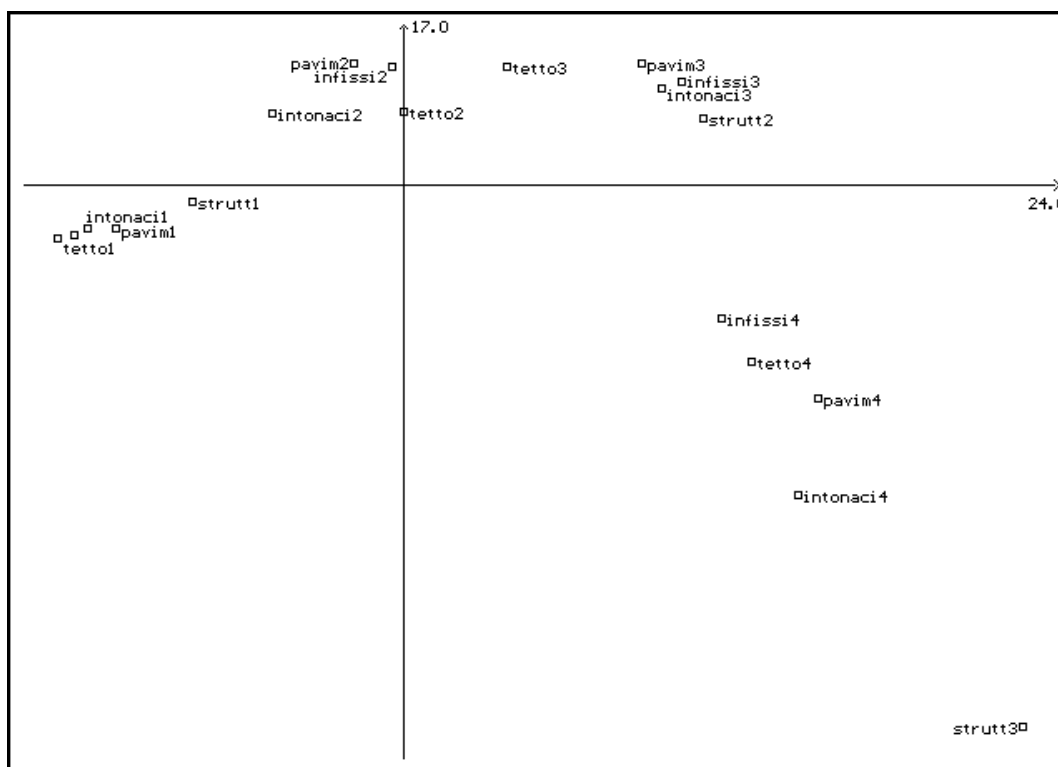
**DIS** è il quadrato della distanza del punto dall'origine. Maggiore è DIS più il profilo del punto differisce globalmente dal comportamento medio, rappresentato dal centro della nuvola.

**FAT1** è la coordinata del punto in questione sul primo asse fattoriale.

**CON REL** contributo relativo (del fattore al punto): è la frazione (\*1000) dell'inerzia del punto spiegata dal fattore. Nel caso di strutt1, ad esempio, il primo fattore spiega il 60.9% dell'inerzia.

**CON ASS** contributo assoluto (del punto all'inerzia del fattore): è la frazione (\*1000) dell'inerzia del fattore proveniente dal punto considerato. Nella tavola 6.9 il 7.2% dell'inerzia del primo fattore proviene da strutt1.

La figura 6-4 mostra la proiezione della nuvola sul piano determinato dai primi due assi fattoriali. In questo caso la struttura della nuvola è molto caratteristica, poiché le diverse categorie di ciascuna delle variabili iniziali presentano di fatto un ordinamento che denota un peggioramento dello stato di conservazione della corrispondente componente edilizia nel passare da 1 a 4.



**Figura 6-4** Analisi dello stato di conservazione: proiezione dei punti-variabile sul primo piano fattoriale.

### **I files scritti da ACORR**

- **ACORR.OUT** è il file di uscita, da stampare ed interpretare.
- **COORRIG.LV** (scritto su richiesta): contiene le coordinate fattoriali delle righe ed alcune altre informazioni che vanno passate a **NONGER** o a **CGA** per eseguire una classificazione delle righe.
- **COORCOL.LV** (scritto su richiesta): contiene le coordinate fattoriali delle colonne ed alcune altre informazioni passate a **NONGER** o a **CGA** per eseguire una classificazione delle colonne.
- **ACORR.FPL** (scritto su richiesta): contiene le informazioni passate a **FACPLAN** per visualizzare le proiezioni grafiche sui piani fattoriali.

### 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 dei 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 (vedi par. 4.4), 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.



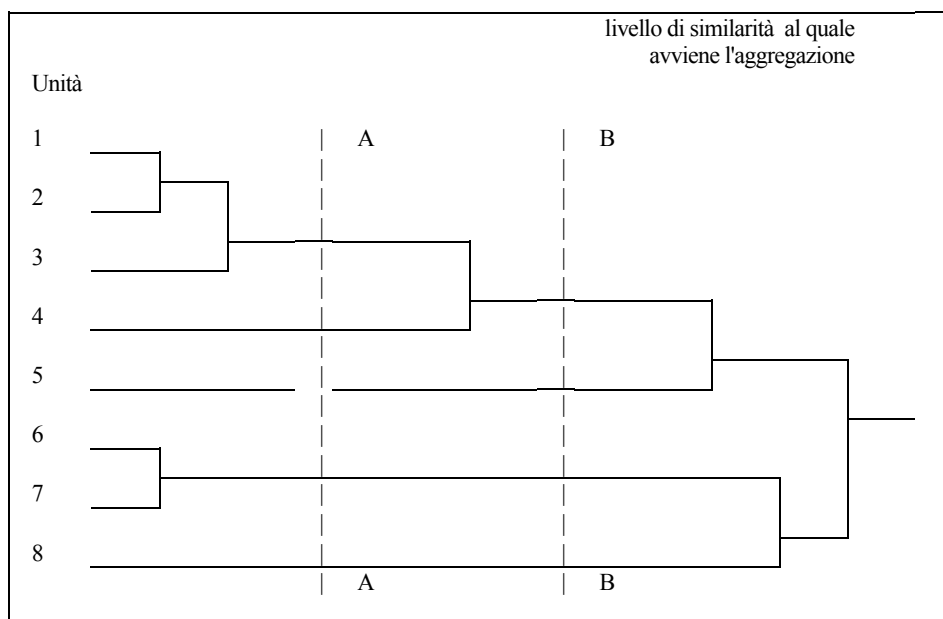
I **metodi gerarchici ascendenti** (o aggregativi) eseguono iterativamente le seguenti operazioni su un insieme di  $n$  unità elementari o di gruppi costituiti in precedenza:

- 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). Vanno stabiliti alcuni criteri per sezionare l'albero in modo conveniente.



**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.

### *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**. Non è escluso 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.

La partizione che si ottiene alla fine del processo dipende dalla configurazione assunta inizialmente e dal numero delle classi richieste.

### *Alcune definizioni*

---

Quando rappresentiamo un insieme di oggetti come punti in uno spazio multidimensionale assumiamo l'**Inerzia** (definita nel par. 4.4), 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  $In_{tot} = \sum_i m_i * d_i^2$  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.  $G_j$  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

$$In_{int}(j) = \sum_{i \in I_j} m_i * d^2(i, G_j)$$

dove la somma è estesa alle sole unità appartenenti alla classe  $j$  e le distanze sono relative al centro di classe  $G_j$ .

**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 intraclassa** (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  $G_j$ .

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 interclassa** (o **esterna**, o **between classes**) della partizione:

$$In_{ext} = \sum_j M(j) * d^2(G_j, G)$$

dove  $M(j)$  è la massa della  $j$ -esima classe, pari alla somma delle masse di tutte le unità che ad essa afferiscono, e  $d^2(G_j, G)$  è il quadrato della distanza di  $G_j$  da  $G$ , centro globale della nuvola.

Immaginiamo che la nuvola sia già stata suddivisa in  $k$  gruppi, con centri  $G_1, G_2 \dots G_k$  rispettivamente: si può provare (teorema di Huyghens) che l'inerzia totale si può decomporre come segue:

$$In_{tot} = In_{ext} + In_{int}$$

Dove **In<sub>tot</sub>** è l'Inerzia Totale della nuvola di punti, **In<sub>int</sub>** è l'Inerzia interna definita in precedenza e **In<sub>ext</sub>** è l'Inerzia esterna.

In ADDATI la funzione-obiettivo per la classificazione non-gerarchica è

$$\max (In_{ext} / In_{tot}) \quad \text{che equivale a} \quad \min (In_{int} / In_{tot})$$

e corrisponde ad un insieme di gruppi il più possibile compatti. Il valore della funzione obiettivo varia fra 0 e 1 (a parità di classi, il valore migliore è quello più alto).

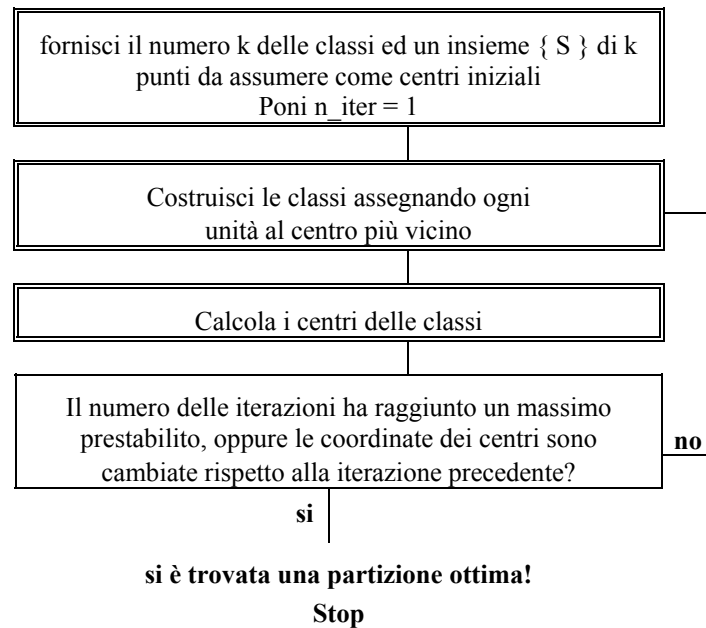
In un'aggregazione gerarchica si hanno inizialmente tanti gruppi quante sono le unità ( $In_{ext} = In_{tot}$ ;  $In_{int} = 0$ ). Quando le unità vengono progressivamente aggregate, ad ogni passo l'inerzia interna aumenta mentre l'inerzia esterna diminuisce di una eguale quantità. Alla fine del processo, 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  $\{S_1, S_2, \dots, S_k\}$  da assumere come centri iniziali di aggregazione (sono anche noti come **semi**).



**Figura 7-2** Schema dell'algoritmo di classificazione non gerarchica di Diday.

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 calcolati i centri delle classi, che assumono il ruolo dei centri iniziali, poi viene ripetuta la procedura di assegnazione e vengono ricalcolati i centri. Ad ogni iterazione qualche unità può cambiare di classe, finché si raggiunga una configurazione stabile.

Si può dimostrare che ad ogni iterazione *'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.

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.

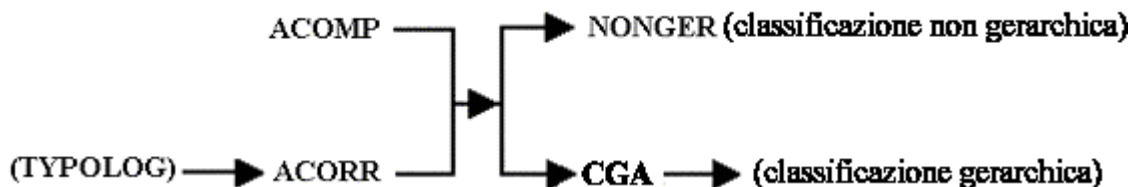
Una volta scelto il numero  $k$  delle classi, la partizione finale dipende solo dall'insieme dei semi iniziali  $\{S_1, \dots, S_k\}$ , in quanto l'algoritmo è totalmente deterministico.

## 7-2 - I metodi di classificazione in ADDATI

---

### 7-2.1 - Alcune note generali

---



ADDATI include un percorso di classificazione non gerarchica ed uno di classificazione gerarchica. Essi sono basati sulla teoria illustrata brevemente nei paragrafi precedenti e sono implementati rispettivamente dai programmi **NONGER** e **CGA**. Mostriamo qui sotto i possibili schemi analitici.

### *Il metodo di classificazione non gerarchica*

---

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 di partenza 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, implementato nel programma **CGA**, sia piuttosto semplice e diretto e non abbisogni di spiegazioni dettagliate.

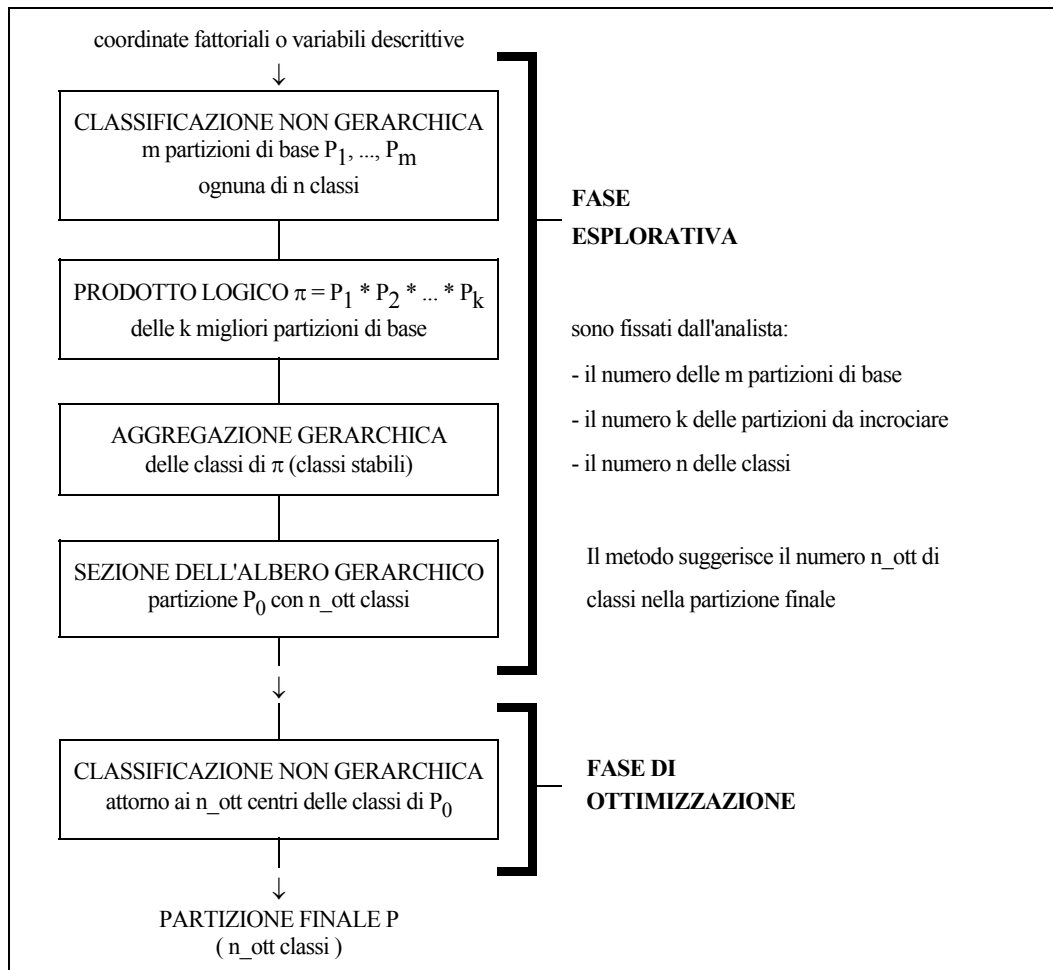
L'input può essere una tavola di descrizione quantitativa, una tavola di contingenza oppure una tavola di coordinate fattoriali registrate da **ACORR** o **ACOMP**.

### 7-2.2 - Classificazione non gerarchica: il Metodo 1 (non più implementato in ADDATI)

---

Allo scopo di produrre una partizione soddisfacente si procedeva in due tappe: una **fase esplorativa** forniva informazioni sul numero più conveniente di classi e suggeriva una scelta dei centri d'aggregazione iniziali; una successiva **fase di ottimizzazione** produceva la partizione ottima finale.

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; comunque, ADDATI offre alcune alternative.



**Figura 7-3** La strategia di classificazione mista già implementata in ADDATI (schema).

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.

Dopo avere esaminato la composizione e le caratteristiche delle classi stabili (elencando le unità assegnate a ciascuna di esse, e calcolando in ciascuna il valore medio delle variabili descrittive), **NONGER** chiama una routine che le aggrega gerarchicamente

secondo la loro somiglianza globale: poiché il loro numero non supera di solito qualche decina, un'aggregazione di tipo gerarchico risulta accettabile.

Veniva poi mostrato sullo schermo l'albero gerarchico e l'utente poteva decidere il numero di classi della partizione finale. A tale scopo, **ADDATI** offriva una routine che consentiva di descrivere ogni partizione ottenibile sezionando l'albero in modo da ottenere un numero di classi conveniente (almeno in via ipotetica): l'utente poteva confrontare i caratteri delle *partizioni candidate* (valore della funzione-obiettivo e significato delle classi in ciascuna di esse).

La fase esplorativa consentiva di fermare l'attenzione su una partizione soddisfacente (non ottima), il cui numero di classi - che non coincideva in generale con quello fissato all'inizio - rappresentava una utile *informazione emergente*.

### *La fase di Ottimizzazione*

---

La partizione sulla quale si decideva di attestarsi (registrata su file) non risultava in generale ottima. Essa veniva allora migliorata controllando l'allocatione alle classi degli elementi di confine ed eventualmente riassegnandoli in modo da aumentare il valore della funzione-obiettivo. Si tratta di cambiamenti spesso piccoli, a volte invece cospicui: la cosa dipende da quanto l'insieme da classificare è strutturato in modo netto, o da come aveva lavorato il metodo nella fase esplorativa.

### *7-2.3 - Classificazione non gerarchica: il Metodo 2 attualmente utilizzato*

---

L'uso prolungato del pacchetto ha messo in evidenza quanto segue.

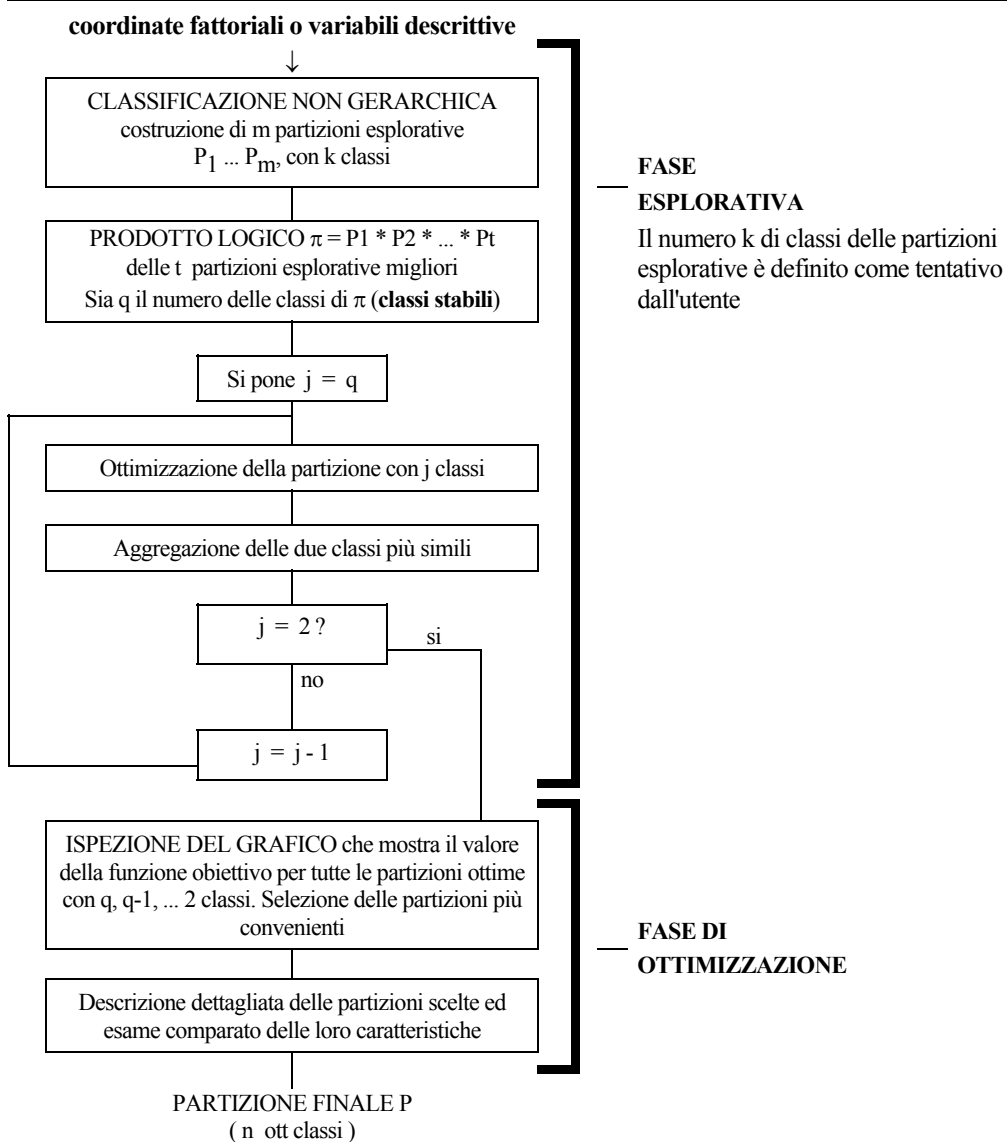
- La sequenza di operazioni piuttosto articolata presentata come Metodo 1 attenua effettivamente la dipendenza della partizione ottima dalla scelta dei centri iniziali. Tuttavia, il risultato è ancora condizionato dal numero di classi scelto per costruire le partizioni esplorative. Il programma è costretto ad usare quel numero di classi anche quando esso non si attagli alla struttura propria dell'insieme da classificare, e ciò può influenzare negativamente la qualità dei risultati.
- Le partizioni ottenute tagliando l'albero che rappresenta l'aggregazione gerarchica delle classi stabili possono risultare tutt'altro che ottime. E' ben vero che la partizione selezionata viene poi ottimizzata attraverso un'altra chiamata a **NONGER**, ma comunque l'albero da cui essa è desunta non rappresenta una sequenza di partizioni ottime e non consente dunque di determinare con certezza quale sia il numero delle classi in corrispondenza al quale conviene arrestare la procedura di aggregazione. Si tratta del numero di classi cui corrisponde, in caso di ulteriore aggregazione, la massima caduta del valore della funzione-obiettivo.

La strategia illustrata dalla figura 7-4 è almeno in parte diversa. Viene ancora determinato un numero piuttosto elevato di partizioni esplorative, con un numero di classi definito dall'utente come tentativo, e le migliori partizioni vengono ancora incrociate per costruire le classi stabili.

La partizione-prodotto (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.

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; l'altra riduce il numero dei gruppi aggregando i due più simili. Si continua così finché tutti i gruppi non siano aggregati.



**Figura 7-4** Strategia di classificazione implementata in ADDATI (Metodo 2, schema).

Appare sullo schermo un grafico che mostra come il valore della funzione-obiettivo diminuisce con il diminuire del numero delle classi (si tenga a mente che in questo caso si ha a che fare con una **sequenza di partizioni ottime**. E' così possibile focalizzare le più promettenti (vale a dire, quelle cui corrisponde un'alta perdita d'inerzia quando il numero dei gruppi venga ulteriormente ridotto di una unità). A richiesta, le partizioni così selezionate vengono descritte in modo esaustivo e la loro comparazione porta alla scelta definitiva.

L'intera sequenza prevista dal metodo 2 è tutta calcolata da **NONGER**; il che risulta operativamente più conveniente rispetto al metodo precedente, che richiedeva il lancio consecutivo di tre programmi (**NONGER** fase esplorativa - **PROFIL** - **NONGER** fase di ottimizzazione). Si tratta tuttavia di un aspetto non essenziale: anche il metodo 1 poteva con facilità venir compattato in un solo programma.



La differenza sostanziale sta nel fatto che il metodo 2 usa il numero delle classi iniziali - definito dall'utente - *solo per costruire la partizione incrociata*, la quale è il punto di partenza di una seconda fase che vede l'ottimizzazione sequenziale di un insieme di partizioni con un numero di classi via via decrescente. Il numero di classi della partizione selezionata alla fine del processo - deciso in base al grafico che mostra l'andamento della funzione-obiettivo ed ai caratteri delle partizioni candidate - dovrebbe rappresentare realmente un carattere intrinseco dell'insieme da classificare.

Ovviamente non si può mai esser certi di aver raggiunto la partizione globalmente ottima con un dato numero di classi (il cosiddetto *optimum optimorum*), e probabilmente è proprio così. Entrambi i metodi sono di tipo euristico e producono una partizione di buona qualità, non la migliore in assoluto. Ma ciò è ben noto per ragioni combinatorie, e va accettato.

Può anche accadere che il metodo 1 porti - a volte - ad una partizione finale migliore, cioè con un valore più elevato della funzione-obiettivo. Non si può mai dire. Ma se ci si propone di avere un'idea del numero di classi che meglio si attaglia alla struttura interna dell'insieme, la seconda procedura - che produce **una sequenza di partizioni ottime** - appare più affidabile.

## 7-3 – Il programma NONGER

---

<b>Funzione</b>	Produce una partizione ottimizzata dell'insieme da classificare seguendo un metodo non gerarchico.
<b>Limiti</b>	Per rendere più spedito il calcolo, piuttosto pesante, la tavola dei dati viene registrata in memoria centrale. NONGER è un programma a 32 bit, che gira sotto un DOS Extender, capace di indirizzare tutta la memoria disponibile.
<b>Consigli</b>	Nella fase esplorativa si chiede l'incrocio di <b>non più di tre partizioni di base</b> . Si cerchi di limitare, per quanto possibile, anche il numero delle classi.

### *Fase esplorativa*

---

Poiché la sequenza delle domande effettivamente poste dal programma dipende dalle risposte già fornite, porremo tra linee orizzontali e marcheremo con una freccia le domande condizionate da risposte precedenti.

Si possono suddividere le domande in due blocchi:

- domande relative al caricamento dei dati
- domande poste dal programma dopo il caricamento dei dati e finalizzate all'immissione dei parametri di controllo per la classificazione.

### *A. Domande relative alla lettura del file dei dati*

---

**Questa classificazione è concatenata ad un'Analisi Fattoriale?**

**no - analisi autonoma**

**si - l'input è una tavola di coordinate fattoriali scritta da ACORR o ACOMP**

Lo scopo è di informare il programma sul formato ed il contenuto del file dei dati in lettura. Se si tratta di un file di coordinate registrato da una precedente Analisi Fattoriale, esso include già alcune informazioni che è altrimenti necessario specificare da tastiera.

I profili dei gruppi sono sempre espressi utilizzando i valori delle variabili descrittive di partenza.

#### **Analisi autonoma**

Le unità statistiche (oggetti, righe della tavola analizzata) vengono classificate direttamente in base ai valori letti nel file di descrizione iniziale. **Non è possibile classificare le variabili (cioè le colonne della tavola).**

In questo caso l'utente deve fornire tutte le informazioni necessarie per l'esecuzione (numero delle unità e delle variabili, nomi, ecc.) rispondendo a specifiche domande poste dal programma.

### **Esecuzione concatenata ad un'Analisi Fattoriale**

La similarità tra gli oggetti è calcolata a partire dalle coordinate fattoriali registrate su file. Ignorando le ultime, meno esplicative, si ottiene un'esecuzione più veloce.

La maggior parte dei parametri viene letta da un file di lavoro.

In questo caso è possibile anche raggruppare le variabili secondo la loro somiglianza sulle unità statistiche (cioè la loro correlazione).

La risposta data a questa domanda provoca una biforcazione nella sequenza delle domande successive. Per una classificazione su coordinate fattoriali si ignori il blocco di domande A1 (appropriato per un'analisi autonoma) e si vada direttamente ad A2.

### ***Blocco A1 (per una classificazione eseguita direttamente sulle variabili descrittive)***

#### **I parametri per quest'analisi verranno**

##### **1. forniti da tastiera**

##### **2. letti dal file NG.PAR dove sono stati salvati quelli relativi ad un'analisi precedente, che possono venire ora opportunamente modificati**

Ogni volta che si esegue un'analisi autonoma la sequenza delle risposte fornite viene registrata su di un file denominato NG.PAR e può essere ricaricata a piacimento. Ciò permette di evitare la noiosa operazione di reinserimento dei parametri allorché, per qualsiasi ragione, si debba ripetere un'analisi con parametri uguali o simili (ad esempio, quando nella fase di ottimizzazione viene letto lo stesso file utilizzato nella fase esplorativa). L'utente può poi utilizzare le chiavi **F1** o **↑** per rivedere ed eventualmente modificare i parametri ricaricati. Ogni volta che viene modificato un parametro il programma controlla la coerenza del nuovo valore con tutti quelli già inseriti e pone, in caso, le necessarie domande integrative.

Si esce dalla fase di modifica premendo **F2**.

I parametri eventualmente letti dal file NG.PAR includono tutte le risposte alle domande del blocco A1.

#### **Nome del file dei dati :**

Il file deve essere organizzato in modo che ogni record (o uno stesso numero di records) si riferisca ad un'unità e contenga le variabili che la descrivono, precedute eventualmente dall'identificatore dell'unità e/o dal suo peso.

Il file può trovarsi in una directory diversa da quella di lavoro: in tal caso va specificato il path completo.

#### **Numero totale delle unità da leggere dal file dei dati :**

Vanno incluse anche le unità supplementari se ve ne sono (cfr. più avanti). Nel file di input ogni unità deve essere descritta da uno stesso numero di records (solitamente uno). Eventuali records in più verranno ignorati.

#### **Numero totale delle variabili :**

E' il numero complessivo delle variabili, **sia attive che supplementari**, che verranno lette dal file dei dati. Se il file contiene più variabili, quelle da caricare verranno specificate fornendo un formato di lettura (si veda più sotto). **Il nome ed il peso dell'unità non vanno contati come variabili.**

#### **Numero delle unità (righe) supplementari :**

I records supplementari -se ve ne sono- **devono seguire** quelli attivi nel file di input. Digita "0" se non vi sono oggetti supplementari.

Oltre alle righe **attive**, sulla cui similarità verranno costruite le partizioni, è possibile introdurre nell'analisi un insieme di unità dette **supplementari**, che non partecipano alla costruzione delle partizioni, ma che vengono assegnate ai gruppi con profilo medio più simile al loro. Lo scopo è di ricavare ulteriori informazioni osservando come gli oggetti supplementari si collocano nelle classi determinate da quelli attivi.

Ad esempio, se le righe attive descrivono il comportamento di un insieme di comuni ad un anno dato, quelle supplementari possono descrivere i medesimi comuni ad uno o più anni diversi, consentendo così di verificare ed interpretare qualitativamente le variazioni avvenute nel sistema.

**Ricorda:** Gli oggetti attivi e quelli supplementari devono essere descritti dalle stesse variabili (colonne della tavola).

#### **Quante variabili supplementari?**

Va inserito il numero delle variabili supplementari (se ve ne sono; 0 = tutte le variabili sono **attive**). Esse possono stare in posizione qualsiasi all'interno del record: allo scopo di consentire al programma di trattarle opportunamente, l'utente dovrà specificare per ciascuna il numero d'ordine che ne identifica la posizione nel record.

Le **variabili supplementari** (se ve ne sono) descrivono gli oggetti ma non contribuiscono alla costruzione delle classi che sono determinate esclusivamente dalla struttura delle similarità tra le unità rispetto alle sole variabili **attive**.

#### **Gli indicatori alfanumerici (nomi) di riga**

**1. verranno digitati da tastiera**

**2. saranno letti dal file di input (se viene usato il formato libero essi debbono trovarsi all'inizio del record corrispondente).**

Il programma utilizza un'identificatore (lungo fino a 12 caratteri e non contenente spazi o virgole) per identificare ciascuna unità. La domanda viene posta per indicare al programma se i nomi verranno inseriti da tastiera o dovranno essere letti dal file di input. In quest'ultimo caso, il formato di lettura dovrà specificare la posizione dell'indicatore in ciascun record; in caso di **formato libero** (si veda più avanti) esso deve trovarsi all'inizio del record corrispondente.

#### **Fornisci gli indicatori alfanumerici delle variabili :**

Ogni variabile (attiva o supplementare) è identificata da un'etichetta alfanumerica lunga al massimo 12 caratteri e che non deve includere né spazi né virgole.

I nomi delle variabili vanno inseriti da tastiera nello stesso ordine in cui le variabili sono caricate dai records in lettura. Essi vanno separati mediante virgole o spazi. E' consentita una scrittura compatta.

**Esempio:** *Per denotare 4 classi di reddito e 5 classi di età, invece di*  
**reddito1 reddito2 reddito3 reddito4 eta1 eta2 eta3 eta4 eta5**  
*è accettata e viene espansa automaticamente la risposta seguente:*  
**reddito1/4 eta1/5**

Qualsiasi errore (troppi nomi, o troppo pochi o inaccettabili) viene segnalato ed è possibile correggerlo.

---

⇒ *La domanda seguente viene posta solo se è stato prima specificato che i nomi delle unità vengono letti da tastiera:*

**Fornisci gli indicatori alfanumerici degli oggetti :**

Essi seguono le medesime regole dei nomi delle variabili. Anche in questo caso è consentita una forma compatta.

**Esempio:** *Per definire un'etichetta per 150 oggetti attivi e 50 supplementari senza dover inserire i nomi uno ad uno, l'utente può digitare:*

**attivo1/150 supplem1/50**  
e la stringa sarà sviluppata automaticamente nelle 200 etichette  
**attivo001....attivo150 supplem01....supplem50**

⇒ *La domanda seguente viene posta solo se è stata dichiarata in precedenza la presenza di variabili supplementari.*

**Fornisci i NUMERI D'ORDINE che identificano le variabili SUPPLEMENTARI tra tutte quelle caricate :**

Vanno specificati i numeri d'ordine delle variabili supplementari per distinguerle dalle altre variabili lette dal file dei dati. La sequenza dei numeri d'ordine *deve tener conto delle sole variabili effettivamente caricate*; altre variabili eventualmente presenti nel file di input, ma ignorate in lettura, non vanno prese in considerazione. Anche in questo caso è possibile una scrittura in forma compatta.

**Esempio:** *la stringa di risposta "2/5 12, 13, 20" specifica che fra tutte le variabili lette quelle supplementari sono la 2<sup>a</sup>, 3<sup>a</sup>, 4<sup>a</sup>, 5<sup>a</sup>, 12<sup>a</sup>, 13<sup>a</sup> e 20<sup>a</sup>. Naturalmente, in questo caso l'utente deve avere dichiarato esattamente 7 variabili supplementari ed almeno 20 variabili complessivamente.*

---

**La tavola sottoposta all'analisi è**

- 1. una tavola di misure quantitative**
- 2. una tavola ottenuta accostando una o più tavola di contingenza.**

Per una tavola di misura (variabili **quantitative**), la distanza (euclidea) tra gli oggetti è calcolata applicando la ben nota *formula pitagorica* dopo aver standardizzato le variabili. Per le tavole di contingenza viene invece usata la *metrica del chi-quadro*.

Questa informazione è necessaria anche per il corretto calcolo dei profili dei gruppi risultanti dalla classificazione: nel caso di variabili **quantitative** il profilo di una classe consiste nelle medie che le diverse variabili presentano in quella classe; tali valori vanno confrontati con il profilo medio globale. Se si tratta di tavole di contingenza, il profilo indica la frequenza delle categorie di ogni variabile nella classe.

⇒ *La domanda seguente viene posta solo se si tratta di una tavola di misura.*

**Ad ogni unità statistica va assegnato :**

- 1. lo stesso peso**
- 2. un peso particolare letto dal file dei dati.**

È noto che ogni unità è rappresentata come un punto dotato di massa in un opportuno spazio geometrico; essa va dunque opportunamente pesata.

Per una tavola di contingenza (si veda **ACORR**) non è necessario fissare i pesi esogenamente; essi vengono calcolati dal programma stesso in base alla tavola dei dati. Per una tavola di descrizione quantitativa il peso è invece fissato dall'analista.

***Nota** Se si opera su una tavola di tassi, è necessario restituire ad ogni oggetto la sua importanza assoluta attribuendogli un peso opportuno.*

Medie, varianze e correlazioni vengono calcolate tenendo conto dei pesi delle unità ed i valori risultanti rappresentano delle proprietà effettive del sistema. Il centro di una classe definisce il comportamento medio di quella classe.

***Esempio:** Se ogni riga rappresenta un comune, descritto da variabili calcolate come valori percentuali sulla popolazione, conviene usare la popolazione del comune come peso.*

La posizione del peso nel record in lettura verrà specificata nel **formato di lettura**. Se si utilizza un **formato libero**, il peso deve trovarsi immediatamente **dopo** il nome dell'oggetto (se letto da file) e **prima** delle variabili.

Si userà l'opzione "*pesi uguali*" quando si tratti di oggetti elementari (individui, alloggi, ecc.). Se invece il peso è letto da file, esso va fornito come numero **intero** o **reale**, non necessariamente normalizzato (ci pensa il programma). **Il peso non deve essere contato tra le variabili.**

⇒ *La domanda seguente viene posta solo se si sta analizzando una tavola di contingenza.*

**Di quante tavole di contingenza affiancate consiste la tavola da analizzare?**

Va fornito **un intero positivo** (c'è **almeno una** tavola di contingenza). Ogni sub-tavola di contingenza rappresenta il modo in cui le **medesime unità elementari di conteggio** si distribuiscono sulle categorie di una diversa variabile qualitativa.

**Esempio:** *Si pensi ad una tavola le cui righe rappresentino i comuni di una regione e le colonne diano, per ogni comune, la distribuzione della popolazione su 4 classi di età e 4 livelli di istruzione. In tal caso le unità di conteggio sono gli abitanti: si può pensare di aver costruito la tavola da analizzare incrociando, per la popolazione della regione, il comune di appartenenza con la classe d'età e con il titolo di studio, affiancando poi le due tabelle d'incrocio ottenute.*

Le tavole di contingenza **debbono contenere lo stesso numero di effettivi**. In caso contrario la classificazione viene validamente eseguita ma i profili delle classi ottenute risultano interpretabili con qualche difficoltà. In questo caso è **opportuno modificare** la tavola dei dati variando **proporzionalmente** gli effettivi di ciascuna sub-tavola in modo da riportare tutte le tavole di contingenza al medesimo totale.

**Esempio:** *Se alcune tavole contano gli abitanti ed altre gli alloggi, i due totali risultano generalmente diversi. Per un corretto calcolo dei profili è necessario scegliere un totale di riferimento (ad es. la popolazione) e riportare tutte le tavole al medesimo valore moltiplicando gli effettivi di ciascuna riga per una costante opportuna. Così, se un comune ha 20000 abitanti distribuiti per condizione professionale e 7000 alloggi suddivisi per epoca di costruzione, gli alloggi effettivi in ciascuna categoria dell'epoca di costruzione andranno moltiplicati per 20000/7000, in modo da portare il loro totale a 20000, senza alterarne la distribuzione percentuale.*

**Nota:** *In generale, se la tavola analizzata ricodifica in forma disgiuntiva completa un insieme di variabili categoriali va qui fornito il numero di tali variabili, ognuna delle quali è stata espansa in una tavola di contingenza.*

---

#### **Fornisci un FORMATO per leggere il file dei dati :**

È richiesto un formato di lettura che specifichi la posizione e la lunghezza dell'indicatore di riga e del peso (se presenti), nonché delle variabili da leggere, ignorando quelle che non si vogliono caricare.

Per una spiegazione dettagliata sul formato di lettura si veda il capitolo 3. Qui ci limitiamo a ricordare che per dichiarare un **formato libero** va premuto il tasto "\*". In tal caso ogni record del file dei dati deve contenere, **nell'ordine e separati da spazi**:

- il nome dell'unità statistica(se letto da file)
- il peso dell'unità statistica (solo per una tavola di misura e solo se si è dichiarato in precedenza che ad ogni unità va assegnato un peso letto da file)
- le variabili

<b>Ricorda:</b> Se il record in lettura non è in formato libero ovvero se l'ordine degli elementi non è quello appena descritto <b>deve essere fornito esplicitamente un formato</b> .
--

A questo punto, il file di input viene aperto e letto. I parametri vengono salvati sul file NG.PAR e possono essere ricaricati se si esegue poi un'altra prova.

### ***Blocco A2 (solo per una classificazione che segua un'analisi fattoriale)***

---

In questo caso quasi tutti i parametri vengono letti dal file di lavoro registrato da **ACORR** o **ACOMP** e vengono poste solo altre due domande:

**Digita un commento (fino a due pagine) da usare come titolo per l'analisi. Esso verrà scritto in testa al file d'uscita.**

**Se non desideri alcun titolo, premi ↵.**

**Si vogliono classificare :**

- 1. le righe della tavola sottoposta all'analisi fattoriale**
- 2. le colonne della medesima tavola.**

L'informazione è necessaria affinché il programma possa aprire il file appropriato che contiene le coordinate fattoriali salvate in precedenza (COORRIG.LV per le righe, COORCOL.LV per le colonne). Si tenga presente che in una classificazione delle colonne (variabili) non vengono calcolati i profili delle classi.

Questo è quanto per il caso **A2**. Il file dei dati (COORRIG.LV o COORCOL.LV) viene aperto e letto. Segue poi un secondo gruppo di domande che mirano al controllo della procedura di classificazione.

### ***Blocco B. - Parametri per la classificazione***

---

**Quante partizioni esplorative?**

Indicativamente, qualche decina. Un numero maggiore di *partizioni di base* (o *esplorative*) aumenta la probabilità di ottenere delle partizioni di buona qualità da incrociare. E' richiesto un tempo di calcolo proporzionalmente maggiore, ma questo non è più un problema. Il numero delle partizioni esplorative non influenza la richiesta di memoria centrale.

Ovviamente, se l'insieme da classificare è poco numeroso (qualche decina di unità) è inutile richiedere molte partizioni esplorative; se le unità sono alcune centinaia o migliaia è utile aumentare il numero delle partizioni.

**Quante partizioni esplorative (le migliori) si vogliono incrociare per generare le classi stabili?**

Tra le partizioni di base, calcolate in via esplorativa, le migliori (cioè quelle con il valore più alto della funzione-obiettivo, in numero deciso dall'analista) vengono memorizzate ed incrociate per evidenziare i gruppi omogenei emergenti dall'analisi (le cosiddette **classi stabili** o **forme forti**).

La scelta del numero delle partizioni da incrociare dipende dal livello di dettaglio con cui si vogliono costruire le classi stabili. Ad esempio, se si richiedono partizioni di 7 classi, l'incrocio di due di esse potrebbe in teoria produrre fino a 49 classi stabili.



Tuttavia, quanto meglio strutturato è l'insieme tanto più coerenti sono i risultati delle due partizioni e tanto minore risulta il numero delle classi stabili (si ridurrebbero a 7 se le due partizioni fossero identiche).

Per non frammentare eccessivamente la *partizione-prodotto*, il che ne renderebbe difficile la lettura, conviene di solito non incrociare più di 2 o 3 partizioni di base. Si tratta comunque di una scelta guidata dall'esperienza dell'utente e dal buon senso, legata anche al numero delle classi richieste per ciascuna partizione.

#### Quante classi in ogni partizione?

L'utente deve specificare il numero delle classi da calcolare per ciascuna partizione. Si dovrebbe indicare, come tentativo, un numero di gruppi all'incirca pari a quello che si vorrebbe ottenere nella partizione finale.

Il numero effettivo di classi della partizione finale viene invece deciso dopo aver attentamente ispezionato il diagramma che mostra come la funzione-obiettivo diminuisca con il numero delle classi.

#### Opzioni per la scelta dei centri iniziali di aggregazione :

1. **scelta casuale ripetibile**
2. **scelta casuale non ripetibile**
3. **centri iniziali forniti da tastiera**

E' conveniente utilizzare una **scelta casuale** (è il programma stesso ad eseguirla), ma l'utente può optare per una qualsiasi delle scelte sopra indicate.

#### Scelta casuale ripetibile

Se si debbono aggregare  $n$  unità il programma genera tanti numeri interi tra 1 ed  $n$  (inclusi) quante sono le classi e gli oggetti con quei numeri d'ordine vengono assunti come centri iniziali di aggregazione. La generazione dei numeri casuali parte da un **seme fisso** (un valore che ne determina la sequenza): come conseguenza, ripetendo l'analisi viene generata la medesima sequenza di partizioni.

#### Scelta casuale non ripetibile

In questo caso il seme varia (esso viene determinato di volta in volta in base all'orologio interno della macchina). Una ripetizione dell'analisi dà luogo a centri iniziali diversi e produce dunque risultati non identici.

#### Centri di aggregazione forniti da tastiera

Talvolta può essere utile cercare di *guidare* l'aggregazione assumendo come centri iniziali alcune unità opportunamente fissate dall'esterno. Questa opzione può venire utilizzata quando vi sia ragione di pensare che ciò porti ad una partizione meglio interpretabile ed utilizzabile.

Se la scelta è questa, viene chiesto all'utente di fornire, per ogni partizione esplorativa, i numeri d'ordine delle unità da assumere di volta in volta come centri iniziali.

---

⇒ *La domanda che segue appare solo se si è seguito un percorso d'analisi che usi ACORR. In ogni altro caso l'informazione viene letta dal file di lavoro.*

**Di quante tavole di contingenza affiancate consiste la tavola da analizzare?**

Il significato della domanda è già stato illustrato per la fase esplorativa.

⇒ *La domanda seguente viene posta solo se vi sono unità supplementari.*

**L'analisi include alcune unità supplementari. Puoi scegliere :**

- 1. di classificare solo le unità attive**
- 2. di costruire le partizioni solo in base alle relazioni di similarità esistenti tra le unità attive, assegnando poi ogni unità supplementare alla classe più simile (in questo caso il profilo delle classi sarà computato solo sulle unità attive assegnate a ciascuna di esse).**

A questo punto vengono calcolate le partizioni esplorative richieste. Durante l'esecuzione, il programma informa l'utente sulle operazioni in corso. Le migliori partizioni esplorative, nel numero richiesto dall'analista, vengono poi incrociate per determinare le **classi stabili** (o **forme forti**).

---

#### ***NONGER - Fase di ottimizzazione e descrizione delle partizioni ottenute***

---

Sia  $q$  il numero delle **classi stabili** ottenute nella fase esplorativa. **NONGER** procede ora aggregando i due gruppi più simili e cercando poi di ottimizzare la partizione ottenuta riallocando opportunamente alcuni elementi. Ne risulta una partizione ottima con  $q-1$  classi, sulla quale viene ripetuta la medesima operazione, che produce ora una partizione con  $q-2$  classi. Si procede così iterativamente fino ad ottenere alla fine una partizione ottima con due classi.

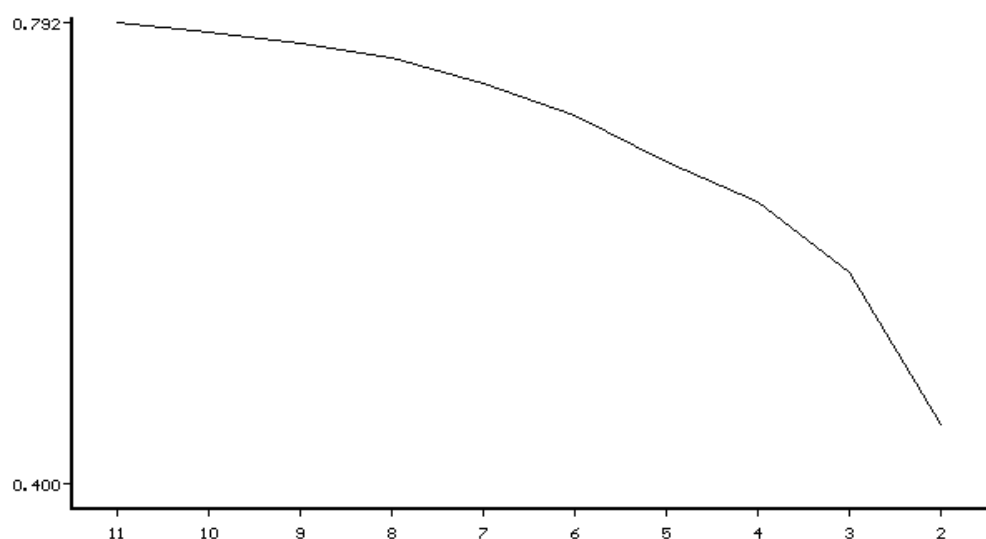
Le partizioni ottenute, con un numero di classi variabile da  $q$  a 2, costituiscono degli **ottimi locali**.

**NONGER** chiama un programma di utilità che disegna sullo schermo il grafico che mostra come il valore della funzione-obiettivo della sequenza di partizioni ottime vada diminuendo al decrescere del numero delle classi (la figura 7-5 si riferisce al caso di Venezia: vengono determinate 11 classi stabili).

Dall'esame del grafico l'utente può scegliere una o più partizioni *promettenti*, con un numero di classi prossimo a quello che si vorrebbe idealmente ottenere ed un valore della funzione-obiettivo ancora sufficientemente elevato.

Il criterio è la **compensazione** (trade-off) tra il livello di sintesi raggiunto (il numero delle classi dev'essere contenuto) ed il valore della funzione-obiettivo, che rappresenta in un certo senso la quota d'informazione mantenuta. E' conveniente ridurre il numero dei gruppi per conseguire una maggior sintesi, a patto che in corrispondenza il valore della funzione-obiettivo non diminuisca eccessivamente. Ciò succede quando il valore della funzione-obiettivo della partizione immediatamente più aggregata (la successiva verso destra nel grafico) si mantenga ancora abbastanza elevato. In caso contrario, cioè se la f.o. diminuisce troppo, il prezzo da pagare in termini di perdita d'informazione può risultare inaccettabile. Conviene attestarsi su una delle partizioni la cui funzione-obiettivo

diminuisca sensibilmente (cioè la pendenza della linea spezzata aumenti visibilmente) quando il numero delle classi venga ridotto di una unità.



**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.

La decisione dovrebbe tenere conto dei seguenti tre elementi.

- 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 loro profili - scegliendo quella che più risponde agli scopi dell'analisi.

Una volta scelte le possibili partizioni candidate, si esce dall'utility che mostra il grafico scegliendo l'opzione '**ESCI**' dal menu '**FILE**', e **NONGER** riassume il controllo. Viene allora chiesto all'analista il numero delle classi della partizione ottima che vuole descrivere. La richiesta viene poi ripetuta in modo da permettere la descrizione di tutte le partizioni desiderate pervenendo, attraverso il confrontando dei loro caratteri, alla scelta finale.

Dopo aver scritto sul file di output **NG.OUT** la partizione richiesta, il programma mostra una schermata come quella di tabella 7-1 se sono state analizzate variabili quantitative, o come quella di tabella 7-2 nel caso di variabili qualitative.

Nel caso di VARIABILI QUANTITATIVE, i simboli '+' e '-' usati come aiuti all'interpretazione dei profili hanno il significato seguente. Sia					
$xm(j,i)$ il valore medio della variabile $j$ nella classe $i$ , e					
$xg(j)$ il valore medio globale della medesima variabile $j$ .					
La variabile $j$ caratterizza fortemente il comportamento della classe $i$ quando lo scarto $xm(j,i) - xg(j)$ si scosta notevolmente da 0. Per valutarne la significatività, tale scarto va confrontato con la deviazione standard $\sigma(j)$ della variabile. Il valore del rapporto					
$R = [xm(j,i) - xmg(j)] / \sigma(j)$					
Viene confrontato con 4 soglie $s1, \dots, s4$ <i>prefissate opportunamente</i> ed i cui valori correnti sono riportati qui sotto. Il valore dell'aiuto viene determinato secondo lo schema seguente:					
---	--	~~	++	++++	valore
-----> <-----> <-----> <-----> <----->	del				
-1.00	-0.20	0.20	1.00	rapporto R	

**Tabella 7-1** Schermata relativi agli aiuti all'interpretazione dei profili nel caso di variabili quantitative.

I simboli + e - adottati per descrivere il profilo di classe vanno interpretati come segue. Il riferimento è al RAPPORTO tra la frequenza di ciascuna variabile nella classe e la sua frequenza globale. Con le soglie attuali, tale rapporto risulta così rappresentato:					
---	--	~~	++	++++	valore
-----> <-----> <-----> <-----> <----->	del				
-1.00	-0.20	0.20	1.00	rapporto R	

**Tabella 7-2** Schermata d'aiuto all'interpretazione dei profili nel caso di variabili qualitative.

I profili di classe vanno confrontati con il profilo globale allo scopo di verificare quali variabili caratterizzino maggiormente i diversi gruppi. E' possibile interpretare convenientemente i risultati focalizzando l'attenzione sui caratteri peculiari di ciascun gruppo, segnalati da valori di qualche variabile significativamente più alti o più bassi rispetto a quello assunto dalla stessa variabile nel profilo medio globale. Per aiutare l'utente ad individuarli a colpo d'occhio, sotto il valore di ogni componente del profilo viene trascritto uno dei seguenti simboli:

"---", "--", "~~", "++", "++++"

Il simbolo che viene scritto è determinato dal **rapporto** tra il valore della componente di profilo in una particolare classe e nel profilo globale. Tale rapporto **viene calcolato per ogni componente** e confrontato poi con alcuni valori di soglia. La stringa simbolica è scelta sulla base dello schema mostrato nelle tabelle 7.1 e 7.2. I valori di soglia mostrati sono quelli di default proposti dal programma; di solito abbastanza appropriati.

**Nota:** Quando si lavora con variabili quantitative le soglie sono calcolati dal programma e dipendono, per ciascuna variabile, dalla sua deviazione standard. Ogni variabile ha le sue soglie, che dipendono dalla sua specifica dispersione. Mentre le classi sono calcolate a partire dalle coordinate fattoriali o - se si lavora direttamente sulla tavola dei dati senza passare per ACOMP - a partire dalle variabili standardizzate (al fine di equilibrarne l'importanza), i profili sono calcolati sui valori originali e fanno uso delle unità di misura di partenza. Ciò semplifica l'interpretazione.

### **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 componente omologa del profilo globale è la media della variabile sull'insieme di tutte le unità.
- **Tavole di contingenza** (incluso il caso di analisi su descrizioni qualitative che partano con **TPOLOG**): una componente del profilo rappresenta la frequenza di una variabile (o di una categoria) nella classe. La componente omologa del profilo globale è la frequenza globale di quella variabile o categoria.

In linea di massima, per quanto riguarda le variabili quantitative le indicazioni sui valori di soglia fornite dal programma sono ragionevoli e non dovrebbero essere cambiate dall'utente. Nel caso di tavole di contingenza può capitare a volte che le diverse unità (e quindi anche le diverse classi) siano poco differenziate: in tal caso l'insieme delle soglie di default può risultare poco appropriato (ad esempio, se in quasi nessun gruppo la differenza tra le frequenze della variabili e le corrispondenti componenti del profilo globale supera il 20%). In questo caso l'utente può utilmente cambiare, i valori di soglia. Tuttavia, bisogna rilevare che in tal caso i valori di soglia sono i medesimi per tutte le variabili (non dipendono cioè dalla variabilità propria di ciascuna variabile, come nel caso delle variabili quantitative): a volte è quindi necessario ripiegare su valori di compromesso.

Seguendo le indicazioni visualizzate sullo schermo, l'utente deve rispondere alla seguente domanda:

**Digita :**

- 1. per mantenere questi valori di soglia**
- 2. per cambiarli opportunamente**

Se l'utente sceglie di cambiarli, gli vengono richiesti quattro nuovi valori di soglia, **reali** e in **ordine ascendente**, separati da spazi.

Viene effettuato un controllo sulla correttezza dei valori inseriti. Se vengono accettati, i simboli di aiuto vengono modificati di conseguenza. Una volta registrati i profili sul file di output, appare la domanda seguente:

**I profili sono stati descritti. Puoi ora:**

- 1. continuare**
- 2. esaminare i profili ottenuti e decidere se cambiare le soglie**
- 3. cambiare le soglie per gli aiuti**

L'utente può esaminare (dall'interno del programma) il file di output e decidere se le soglie correnti siano adeguate, cioè se esse consentano di ben rappresentare le differenze esistenti tra i profili delle diverse classi. Se così non fosse, l'utente può modificare le soglie, ed allora la parte del file di output che contiene la descrizione dei profili viene sovrascritta. Questa operazione può essere ripetuta fino ad ottenere una descrizione soddisfacente.

Illustreremo l'uscita di **NONGER** con riferimento all'esempio sul Centro Storico di Venezia considerato per **ACOMP** (cfr. sezione 6-2). Supponiamo di aver salvato per la classificazione le prime cinque componenti principali (che spiegano praticamente il 100 per cento dell'inerzia) e di inserire i seguenti parametri per pilotare la fase esplorativa:

- partizioni esplorative da calcolare : 8
- 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 10 classi stabili. Esaminato il grafico della funzione-obiettivo, decidiamo di chiedere la descrizione della partizione ottima con 5 classi.

Dopo un sommario del valore dei parametri forniti dall'utente, il file d'uscita **NG.OUT** riporta il numero di iterazioni eseguite per ciascuna partizione richiesta ed il valore finale della funzione-obiettivo.

Per ogni partizione della quale si sia richiesta la descrizione, viene scritta una tavola che elenca il numero delle unità in ciascuno dei gruppi prodotti ed il suo peso.

10 CLASSI STABILI NELLA PARTIZIONE INCROCIATA.											
CLASSE	1	2	3	4	5	6	7	8	9	10	TOT
UNITÀ	28	22	20	20	20	12	10	8	5	3	148
PESO (%)	17.5	13.9	13.4	12.8	15.4	9.4	7.6	5.2	3.2	1.6	100.0

**Tabella 7-3** Numerosità e peso delle 10 classi stabili ottenute 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 10 classi.

*****
* CLASSE 1 *
*****
UNITA' : 28 PESO : 17.54%
UNITA' ASSEGNATE ALLA CLASSE:
2 18 20 24 27 29 50 77 78 80 82 87 88 89 90 91 92 93 94 96
98 99 101 102 103 112 126 138
UNITA' PIU' VICINA AL CENTRO DI CLASSE (d2 = 0.2184) : 82
UNITA' PIU' LONTANA DAL CENTRO DI CLASSE (d2 = 27.1010) : 50
RAGGIO DI CLASSE : 1.21268
DISTANZA DEL CENTRO DI CLASSE DAL CENTRO GLOBALE : 1.44331

**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  $G_j$  di tutte le unità assegnate alla classe. Tale valore si può scrivere come

$$In_{int}(j) = \sum_i m_i * d^2(i, G_j) = M_j * d_j^2$$

dove  $M_j$  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

$$d_j = [In_{int}(j) / M_j]^{1/2}$$

- 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 1.4433 mentre, per confronto, l'analogo valore è 2.9074 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  $In_{int}$  (1.960 per le 10 classi stabili su Venezia);
- l'inerzia esterna totale  $In_{ext}$  (6.039 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 8.0, poiché siamo partiti da otto variabili attive standardizzate e tutte le Componenti Principali significative sono state tenute in conto nella classificazione);
- 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  $0.755 = 6.039 / 8.0$ ).

Seguono i profili di classe delle partizioni delle quali viene richiesta la descrizione (si vedano la spiegazione che segue, e la tabella 7.5 come esempio). 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 **NG.OUT**. 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' rappresenta il numero delle classi richieste. Questi file hanno un formato compatibile con **ARC/VIEW**, programma GIS che si può utilizzare quando si siano classificate unità geografiche e si voglia disegnare sullo schermo la mappa che rappresenta il risultato della classificazione.

CLASSE	NUM	p_alto	p_terra	buoni	carente	scadenti	sovraff	standard	sottout
1	42	90.338 ++	9.661 --	66.942 ++++	22.303 --	10.874 --	11.283 --	54.639 --	34.198 ++++
2	36	94.580 ++	5.420 --	55.082 ++	29.784 ~~	15.189 --	14.498 --	60.912 ~~	24.648 ++
3	33	83.005 --	16.995 ++	49.993 ~~	31.250 ~~	18.807 ~~	17.086 ~~	64.326 ++	18.639 --
4	32	78.024 ~~	21.976 ++	31.647 --	44.347 ++++	24.063 ++	28.320 ++	61.574 ++	10.164 --
5	5	68.764 --	31.236 ++++	18.357 --	39.655 ++	42.091 ++++	49.491 ++++	46.551 --	4.056 --
PROFILO GLOBALE	148	86.012	13.988	50.504	31.772	17.797	18.529	59.771	21.774
CLASSE	NUM	st_alto	st_oper	st_altro					
1	42	33.659 ++++	10.251 --	56.199 --					
2	36	26.335 ++	14.159 --	59.543 ++					
3	33	22.056 --	18.076 ~~	59.880 ++					
4	32	14.392 --	26.881 ++++	58.767 ~~					
5	5	7.609 --	35.904 ++++	56.592 --					
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.

- **NG.OUT** che elenca l'attribuzione delle unità alle classi, descrive i profili di classe ed offre una serie di informazioni utili all'interpretazione.
- **NGCLASnn.TXT**, dove 'nn' è il numero delle classi della partizione cui il file si riferisce (dunque, un file per ciascuna partizione della quale si è richiesta la descrizione)

Questi file hanno tanti record quante sono le unità classificate, più uno di intestazione. Il loro formato li rende direttamente caricabili, come file di testo, da ARC/VIEW. 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.

- **NG.FPL** è un file che **NONGER** scrive per **FACPLAN** solo quando il percorso di classificazione abbia incluso **ACORR** o **ACOMP**, e sia stato salvato il file necessario per visualizzare le proiezioni sui piani fattoriali. Se quel 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 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.

Il programma di installazione crea all'interno di \ADDATI la directory \ADDATI\ESEMPI, che contiene a sua volta due sottodirettorie denominate VENEZIA e DEGRADO. In esse si trovano alcuni files di dati sui quali l'utente può esercitarsi seguendo le brevi istruzioni qui riportate.

### VENEZIA

---

Vengono forniti due file di dati:

- **VENEZIA.DAT** contiene i dati sulle 148 sezioni censuarie del Centro Storico di Venezia utilizzati per l'esempio sviluppato e commentato nel par. 5.2 relativo ad ACOMP. L'utente può leggere quel paragrafo eseguendo simultaneamente l'esercitazione. La sequenza dei programmi da applicare è

**ACOMP ⇒ NONGER**

- **VE\_ASS.DAT** contiene le medesime informazioni, con la differenza che il numero di alloggi occupati per sezione e categorie delle variabili descrittive è espresso come valore assoluto invece che nella forma percentuale usata per VENEZIA.DAT. Poiché si tratta in questo caso di 4 tavole di contingenza accostate la sequenza da utilizzare è

**ACORR ⇒ NONGER**

Il contenuto dei due file è specificato in dettaglio nei file di documentazione (quelli con estensione **.DOC**). Vengono anche forniti i due file di parametri ACORR.PAR ed ACOMP.PAR da caricare per le due prove, rivedendo poi i parametri (ed eventualmente cambiandoli) dall'interno di **ACOMP** o **ACORR**.

Come esercizio l'utente può utilizzare **MERGFIELD** per costruire VENEZIA.DAT a partire da VE\_ASS.DAT (magari fornendo un nome diverso da VENEZIA.DAT per il file d'uscita per evitare di sovrascriverlo).

### *Visualizzazione della mappa della classificazione*

---

Ottenuta la partizione finale (ad esempio in cinque classi, secondo l'esempio sviluppato nel par. 6.3), si esaminino con attenzione i profili delle classi. La forte struttura di relazione esistente tra le variabili attive fa sì che il primo fattore - che si può interpretare come un indicatore del disagio abitativo - riassume una quota molto elevata dell'inerzia della nuvola. Tuttavia, la scelta casuale dei centri iniziali porta ovviamente a classi il cui ordinamento da 1 a 5 è casuale e non corrisponde ad un ordinamento evidente del livello del disagio.

Per una rappresentazione di più immediata interpretazione, se le classi non risultano già per caso ordinate è opportuno riordinarle a partire dall'esame dei loro profili: ad esempio, in modo da ridenominare come nuova classe 1 quella a minor disagio e come classe 5 quella a maggior disagio. La cosa si può fare operando con **RECODE** (Menu di Utilità) sul file NGCLASnn.TXT che riporta la classe di attribuzione delle sezioni e che costituisce l'input per la visualizzazione della mappa geografica. Se le classi sono ordinate

secondo livelli crescenti di disagio si può scegliere di rappresentarle con una sequenza cromatica opportuna (ad esempio mediante sfumature via via più intense di giallo-rosso o di verde-blu).

**L'analisi va condotta all'interno della cartella <VENEZIA>** che contiene i dati ed include anche tutti i files necessari per la mappatura: CROMA.VID, GRAF\_TAB e MAPNAME, oltre alla subdirectory MAP dove stanno le coordinate geografiche delle sezioni censuarie di Venezia. ***Questi files non vanno toccati.*** Viene anche fornito il programma **MAPPING.EXE**, che è una semplificazione ad hoc di un programma cartografico un tempo distribuito dal CIDOC (Centro di Calcolo dell'Istituto Universitario di Architettura di Venezia), ed ora abbandonato (d'altra parte, neppure il CIDOC esiste più...).

Per disegnare sullo schermo la mappa si digiti "**MAPPING**" e si fornisca su richiesta il nome del file che contiene la classe di attribuzione delle sezioni (NGCLASnn.TXT, dove 'nn' è il numero delle classi) e si scelga una delle sequenze cromatiche proposte

**Il file NGCLASnn.TXT può essere caricato direttamente in ARC/VIEW** (beninteso, bisogna avere lo shapefile delle Sezioni Censuarie di Venezia Centro Storico...). Come esempio, comunque, si può usare **MAPPING**, che dispone del vettoriale necessario.

## DEGRADO

---

Si tratta dei dati relativi allo stato di conservazione di 200 edifici di un Centro Storico cui si è fatto cenno nel par.6.1 (**TPOLOG**) e poi nella parte del par. 6-2.2 relativa ad **ACORR**. Viene fornito il file TYP.PAR che contiene i parametri da far caricare a **TPOLOG**. L'utente può attenersi alla sequenza

**TPOLOG ⇒ ACORR ⇒ NONGER**