I modelli di analisi statistica multidimensionale dei dati:

L'analisi delle corrispondenze multiple

Obiettivi dell'unità didattica

- Comprendere la procedura dell'Analisi delle Corrispondenze Multiple
- Effettuare un'ACM utilizzando Tanagra

4.4 L'analisi delle corrispondenze multiple

- Il ruolo delle variabili nell'ACM
- L'inerzia spiegata e la correzione di Benzecrì
- Esempio: gli utenti di Ebay

L'interdipendenza tra le variabili originarie

E' un'analisi di tipo fattoriale che ha come scopo quello di individuare dimensioni soggiacenti alla struttura dei dati, dimensioni intese a riassumere l'intreccio di relazioni di "interdipendenza" tra le variabili originarie.

L'analisi delle corrispondenze (ACM) trasforma una tabella di contingenza in una rappresentazione grafica al fine di facilitare l'interpretazione dell'informazione contenuta nella tabella stessa.

Punti salienti dell'analisi

Caratteristiche: è un'estensione dell'AC allo studio simultaneo di più caratteri. Il punto di partenza è una tabella di contingenza particolare, nota come matrice di Burt (vedi Unit1.3), in cui le modalità riga coincidono con le modalità colonna.

Obiettivo: individuare le associazioni che possono sussistere tra più variabili qualitative

Ambiti di applicazione: analisi di dati provenienti da indagini effettuate mediante questionario (per esempio per analizzare un segmento di mercato o per lanciare un nuovo prodotto).

Le tipologie di variabili nell'ACM

Nell'ACM possiamo distinguere due tipologie di variabili:

- *attive*, cioè variabili che entrano direttamente nell'analisi concorrendo alla formazione degli assi fattoriali;
- *supplementari o illustrative*, cioè variabili di tipo "passivo" che sono escluse dalla fase di estrazione dei fattori, ma si utilizzano successivamente considerando la loro posizione sugli assi fattoriali come ausilio per la loro interpretazione

La valutazione della dispersione dei profili

Per valutare la dispersione dei profili, riga e colonna, rispetto al loro "centro di gravità" viene utilizzata la metrica del Chi^2.

L'inerzia totale è proporzionale al Chi^2 ma non è esattamente uguale al rapporto tra il Chi^2 e il numero di osservazioni. Essa è funzione del numero di modalità (q) e del numero di variabili attive (p):

Inerzia=(q-p)/p

I fattori estraibili

Vengono estratti degli assi fattoriali, ortogonali tra loro, che spiegano ciascuno, in ordine decrescente, il massimo della variabilità della matrice dei dati (inerzia).

Il numero massimo di fattori estraibili è pari al numero di modalità (q) meno il numero di variabili attive (p).

I fattori da considerare si determinano in base alla loro rilevanza, ovvero alla quota di inerzia totale che spiegano.

La correzione di Benzecrì

Benzecrì indica un limite, pari a 1 diviso il numero di variabili attive (=1/p), al di sotto del quale l'autovalore, e, dunque, il fattore ad esso associato, diviene insignificante (*correzione di Benzecrì*).

Per interpretare il significato degli assi fattoriali, assieme alle coordinate fattoriali, si utilizzano i seguenti indicatori:

- la *massa,* ovvero la frequenza relativa della modalità rapportata al numero di variabili attive
- il *contributo assoluto* e *il contributo relativo (coseno quadrato)*

Inerzia spiegata dai primi fattori

Tipicamente nell'ACM le quote di inerzia spiegate dai primi fattori non sono molto elevate a causa del grande numero di modalità, e di conseguenza di variabilità, presente nella matrice dei dati.

Talora può risultare interessante ripetere l'ACM dopo aver ripulito la matrice dei dati compattando le categorie che nella prima analisi non presentano contributi assoluti abbastanza elevati sui fattori considerati; in questo modo infatti si riduce l'inerzia globale (*q* diminuisce mentre *p rimane fisso).*

I passi dell'ACM

I passi principali dell'interpretazione dei risultati dell'ACM sono:

- Esaminare le entrate della matrice di Burt (quali coppie di modalità si presentano con maggior frequenza e quali mai)
- Le variabili latenti vengono interpretate in ordine crescente d'importanza analizzando le modalità che contribuiscono maggiormente alla spiegazione dell'inerzia totale
- Si esaminano graficamente le proiezioni dei punti riga nel piano formato dalle dimensioni latenti (le prime due e così via)

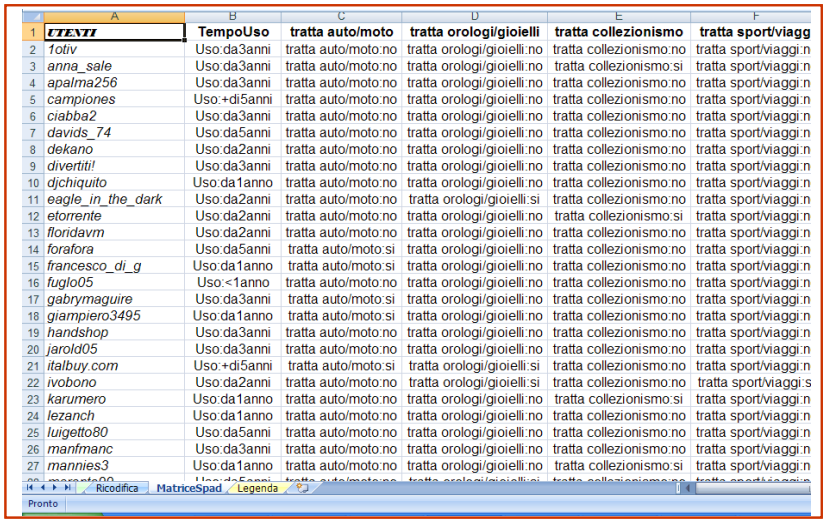
Esempio: gli utenti di Ebay

Nel file "Indagine_utenti_Ebay", sono presenti le risposte degli utenti (acquirenti e/o venditori) di Ebay (la più grande piattaforma al mondo di aste on-line) ad una serie di domande che avevano lo scopo di dare una spiegazione ad un semplice quesito:

Come mai 135 milioni di persone si fidano di un perfetto sconosciuto?

Il campione (70 individui), per semplicità di calcolo, non è rappresentativo dell'intera popolazione ed è composto esclusivamente da utenti italiani.

Esempio: la matrice originaria



UTENTI	TempoUso	tratta auto/moto	tratta orologi/gioielli	tratta collezionismo	tratta sport/viagg
1otiv	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
anna_sale	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:si	tratta sport/viaggi:n
apalma256	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
campiones	Uso:+di5anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
ciabba2	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
davids_74	Uso:da5anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
dekano	Uso:da2anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
divertiti!	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
djchiquito	Uso:da1anno	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
eagle_in_the_dark	Uso:da2anni	tratta auto/moto:no	tratta orologi/gioielli:si	tratta collezionismo:no	tratta sport/viaggi:n
etorrente	Uso:da2anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:si	tratta sport/viaggi:n
floridavm	Uso:da2anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
forafora	Uso:da5anni	tratta auto/moto:si	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
francesco_di_g	Uso:da1anno	tratta auto/moto:si	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
fuglo05	Uso:<1anno	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
gabrymaguire	Uso:da3anni	tratta auto/moto:si	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
giampiero3495	Uso:da1anno	tratta auto/moto:si	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
handshop	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
jarold05	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
italbuy.com	Uso:+di5anni	tratta auto/moto:si	tratta orologi/gioielli:si	tratta collezionismo:no	tratta sport/viaggi:n
ivobono	Uso:da2anni	tratta auto/moto:no	tratta orologi/gioielli:si	tratta collezionismo:no	tratta sport/viaggi:s
karumero	Uso:da1anno	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:si	tratta sport/viaggi:n
lezanch	Uso:da1anno	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
luigetto80	Uso:da5anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
manfmanc	Uso:da3anni	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:no	tratta sport/viaggi:n
mannies3	Uso:da1anno	tratta auto/moto:no	tratta orologi/gioielli:no	tratta collezionismo:si	tratta sport/viaggi:n

Ricodifica | MatriceSpad | Legenda

Pronto

Esempio: ACM con Tanagra

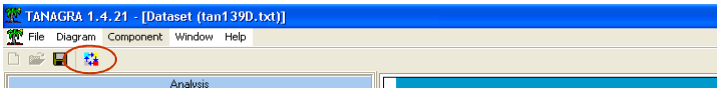
Come per l'ACP, poiché il dataset è in formato Excel, è necessario aprire in C→Programmi→Tanagra il *componente aggiuntivo di Microsoft Office Excel,* successivamente selezionare le celle d'interesse per l'analisi (tutto il dataset con le etichette relative) ed, infine, premere il comando Execute Tanagra che si trova sotto la voce Componenti Aggiuntivi di Excel.

In questo modo si aprirà automaticamente Tanagra e si potrà iniziare l'analisi.

Esempio: un primo sguardo al dataset

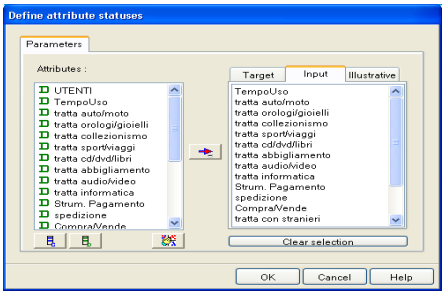
Sotto la voce Dataset description si vede che le variabili oggetto dell'analisi sono 21 (il programma riporta la voce 22 attributes perché elabora come variabile anche l'elenco degli individui "UTENTI") e 70 individui.

A questo punto si clicca sul comando Define Status per iniziare l'analisi.

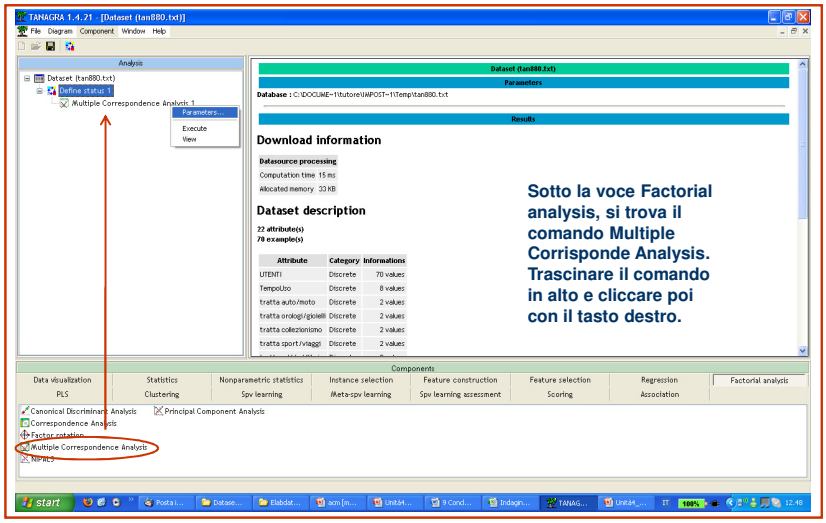


Esempio: la selezioni delle variabili

Si aprirà automaticamente una finestra e si dovranno inserire le variabili oggetto dell'analisi sotto la voce Input, mentre l'elenco degli individui (UTENTI) e le variabili socio-demografiche (che andranno in supplementare) dovranno essere spostate sotto la voce Illustrative.



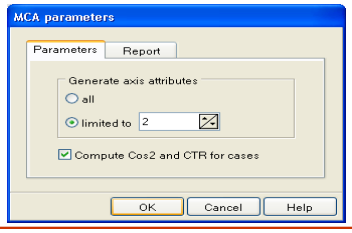
Esempio: il comando ACM



Esempio: la scelta degli assi

Si aprirà a questo punto una finestra in cui bisogna scegliere il numero di assi da generare. La scelta è a discrezione dell'analista. In questo caso, per semplicità di calcolo, scegliamo di propendere per 2 assi fattoriali. Si deve spuntare anche la voce Compute Cos2 per avere il contributo relativo di ciascuna modalità alla costruzione degli assi.

Sotto la voce Report della stessa finestra indicare il parametro 3,00, invece del 4,00 che si trova di default.



Esempio: l'output dell'ACM

Cliccando nuovamente con il tasto destro sulla voce Multiple Corrispondence Analysis e selezionando la voce View appariranno tre tabelle:

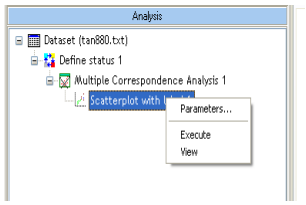
- nella prima è presente l'elenco degli autovalori e l'inerzia spiegata da ciascuno di loro.
- nella seconda tabella si trovano le coordinate di ciascuna modalità
- nella terza è indicato il coseno al quadrato di ciascuna modalità

N.B. La correzione di Benzecrì non è implementata in Tanagra, quindi si avranno percentuali di inerzia spiegata molto basse.

Esempio: come ottenere il piano fattoriale

Per ottenere la rappresentazione grafica degli assi fattoriali bisogna trascinare, come per l'ACP, la voce *Scatterplot with label*, che si trova in Components sotto la voce Data Visualization, nella voce Multiple Corrispondence Analysis che abbiamo costruito precedentemente.

Si deve poi cliccare sulla voce appena spostata con il tasto destro e selezionare la voce View.

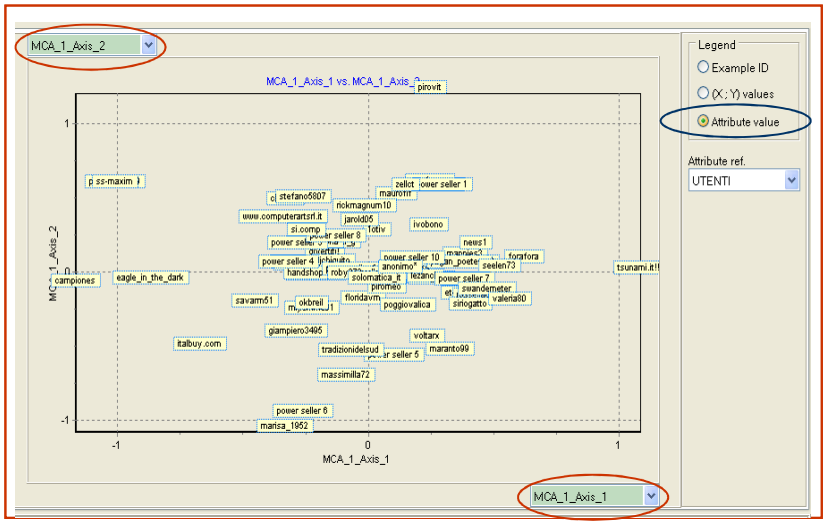


Esempio: il piano fattoriale

Il grafico si apre con il primo asse sia sull'ascisse che sull'ordinata. Sarà necessario, quindi, cambiare uno dei due assi (cerchiati nella figura successiva in rosso) e selezionare il secondo asse.

Inoltre, inizialmente appariranno gli individui con i rispettivi id: per visualizzare i nomi è necessario spuntare nel quadrato Legend che si trova in alto a destra la voce Attribute values (cerchiato in blu nella figura successiva).

Esempio: gli assi fattoriali



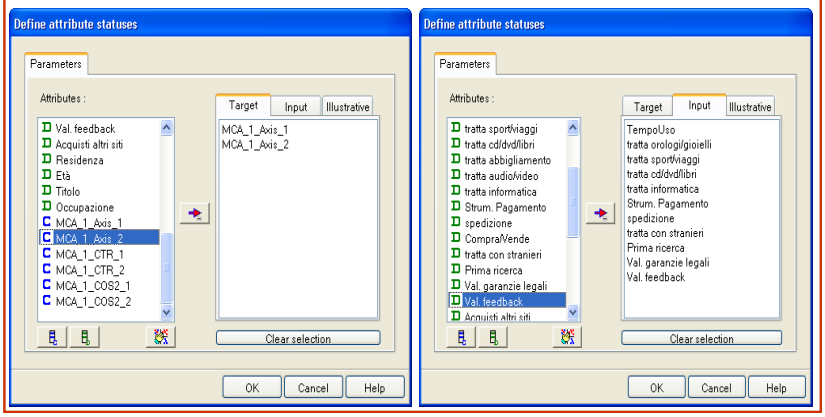
Esempio: come interpretare gli assi

Per una corretta interpretazione degli assi è necessario aggiungere un nuovo componente al diagramma precedentemente formato.

Si aggiunga un nuovo Define Status a questo punto si indicheranno come Target i primi due assi e come Input le variabili che più hanno caratterizzato la costruzione dei fattori. Sono quelle evidenziate in rosa nella seconda tabella dell'output dell'ACM in cui sono indicate le coordinate di ciascuna modalità (Factors characterization – Coordinates and Test values).

Esempio: selezione Target e Input

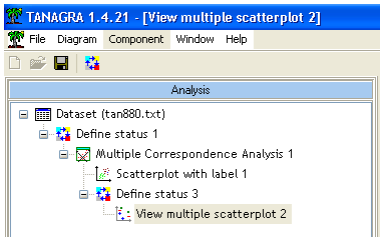
Per una maggiore comprensione è qui indicata la composizione dei Target e degli Input



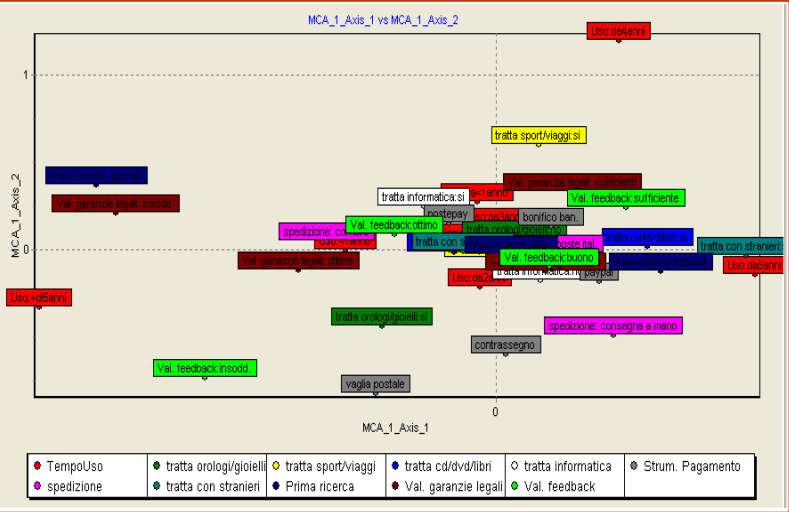
Esempio: il diagramma finale

Si aggiunga, infine, il componente *View Multiple Scatterplot* che si trova in Data visualization e cliccando con il tasto destro del mouse sul nuovo componente si selezioni View per visualizzare il piano fattoriale.

Il diagramma sulla sinistra dovrebbe essere così composto:



Esempio: le variabili sul piano fattoriale



Esempio: l'interpretazione del primo asse

- Appare evidente che per valori positivi del primo asse abbiamo utenti molto "globalizzati" sia perché trattano con altri utenti stranieri sia perché si affidano molto nelle loro ricerche ad internet.
- Per valori negativi del primo asse si hanno utenti, pur di vecchia data, che sono insoddisfatti di alcune garanzie di Ebay (garanzie legali e feedback). Attenzione a non farsi ingannare dalla presenza della modalità "Valutazione Garanzie legali: ottime" che appare in contrasto con quanto detto in precedenza. Tale modalità, infatti, ha un valore Test inferiore al limite 3 fissato nel report in precedenza.

Esempio: l'interpretazione del secondo asse

- Per valori negativi del secondo asse appaiono utenti molto legati a tipologie di pagamento "antiquate" (vaglia postale e contrassegno) e come spedizione preferiscono la consegna a mano. Inoltre, per una conferma è sufficiente aggiungere come Input le variabili età e titolo di studio per vedere che in tale zona si concentrano coloro che hanno oltre 50 anni di età e con licenza media.