CABRAS

annali

università degli studi di cagliari

nuova serie - vol. XXI- anno accademico 2004/2005

ANNALI **DELLA FACOLTÀ DI ECONOMIA DI CAGLIARI**

direttore: **prof. roberto malavasi**

redattore: **prof. gianfranco sabattini**

francoangeli

UNIVERSITÀ DEGLI STUDI DI CAGLIARI

ANNALI
DELLA
FACOLTÀ DI ECONOMIA

DIRETTORE
Prof. Roberto Malavasi

REDATTORE
Prof. Gianfranco Sabattini

Nuova Serie, vol. XXI - Anno Accademico 2004-2005

FrancoAngeli

INDICE

PARTE I

Analisi esplorativa grafica di dati economici. Breve guida didattica all'uso di R.

di *Stefano Cabras*[1] e *Maria Eugenia Castellanos*[2]

Riassunto. L'analisi esplorativa dei dati, oltre agli usuali scopi descrittivi, offre notevoli potenzialità interpretative quando si utilizzano le rappresentazioni grafiche che i programmi statistici di elaborazione automatica contengono in quantità sempre maggiore e con rese più accurate. Certe suggestioni grafiche possono fornire essenziali contributi per la costruzione dei modelli statistici ed indicare proprietà di sintesi necessarie alla loro validazione, basandosi sul contenuto informativo dei soli dati, prima ancora cioè della formulazione delle teorie sottostanti al fenomeno in studio. In particolare il lavoro illustra brevemente ed in forma didattica l'ambiente statistico R con specifico riguardo all'analisi in Componenti Principali, all'analisi delle Corrispondenze ed all'analisi dei gruppi e mostra, attraverso l'applicazione ad uno studio reale di dati socio-economici tratto dalla letteratura, come l'interpretazione di opportune rese grafiche dell'analisi esplorativa possa condurre a risultati molto prossimi a quelli ottenuti con l'uso di modelli statistici sostenuti da teorie economiche.

Parole chiave. Ambiente statistico R; Analisi in componenti principali; Analisi delle corrispondenze multiple; Analisi dei gruppi; Modelli lineari generalizzati; Modelli probit; Stime di densità.

1. Introduzione

Scopo di questo lavoro è far emergere ed illustrare in maniera sintetica le potenzialità che l'analisi esplorativa può offrire, con particolare riguardo agli aspetti grafici, nella costruzione di modelli statistici in generale e specificamente nell'ambito di fenomeni di natura economica.

L'analisi esplorativa, prima fase dell'analisi dei dati, è in grado non solo di supportare le ipotesi di lavoro di partenza, ma anche di orientare lo studio verso la formulazione di nuove ipotesi, di verificare qualitativamente l'esistenza di opportune assunzioni sulle variabili in gioco e di suggerire eventuali mo-

1. Ricercatore di Statistica.
2. Department of Statistics, Rey Juan Carlos University, Madrid (Spain).

delli statistici parametrici, quali ad esempio i modelli lineare generalizzati (Dobson, 2001) da utilizzare a fini inferenziali.

I supporti grafici, di cui sono usualmente dotati i programmi di analisi dei dati, facilitano la lettura dell'informazione statistica, rendendola maggiormente incisiva e rapida soprattutto in presenza di grandi basi di dati.

In questa breve trattazione faremo ricorso all'ambiente di programmazione R, specificamente dedicato alla statistica e basato su un software *open source* che lo rende scaricabile gratuitamente dalla rete e continuamente aggiornabile e di cui sono facilmente reperibili manuali dedicati ad una didattica generale della statistica a vari livelli di approfondimento (si veda, ad esempio, Bortot, *et al.* 2000 oppure Iacus e Masarotto, 2003).

Questa esposizione delle routine dell'analisi esplorativa grafica sarà orientata, oltre che allo scopo didattico, a mostrare come il loro uso ed un'interpretazione appropriata delle analisi possano condurre a risultati che nella fase descrittiva anticipano le risposte fornite dai modelli statistico-probabilistici, giustificandone l'assunzione e semplificando la loro costruzione. L'utilità di tale modo di procedere appare particolarmente efficace nello studio longitudinale di dati *panel* di natura economica e sociale (Wooldridge, 2002). Considereremo la base di dati trattata da Vella e Verbeek in un loro lavoro del 1998. In esso gli autori analizzano la carriera lavorativa di un campione di 554 giovani statunitensi, seguiti nel corso di 8 anni, in base all'affiliazione al sindacato. Mediante modelli di regressione lineare ed un modello *probit* ad effetti misti (McCulloch e Searle, 2001) studiano le caratteristiche socio-economiche degli affiliati al sindacato e l'effetto che l'affiliazione ha in media sull'entità del salario. L'uso di alcune delle tecniche grafiche offerte dall'analisi esplorativa in ambiente R conduce a conclusioni simili a quelle ottenute da Vella e Verbeek, senza che sia necessario ricorrere ad una teoria economica, ma lasciandosi guidare dal contenuto informativo dei dati.

In questo modo viene messa in luce la potenza interpretativa dell'analisi esplorativa grafica che, tuttavia, non può sostituirsi alla modellizzazione né surrogarne il contenuto teorico.

Nella sezione 2 viene introdotto sinteticamente il programma R, e le routine dell'analisi esplorativa grafica. In particolare la sez. 2.1. contiene alcune notizie generali sul programma, come scaricarlo dalla rete, l'accesso all'*help* in linea ed i principali comandi che saranno commentati soltanto nelle parti essenziali, rinviando il lettore alle pagine d'aiuto in linea descritte in seguito. La sez. 2.2. è dedicata ai comandi specifici per l'analisi esplorativa e all'illustrazione delle rese grafiche tramite un esempio elementare tratto dall'*Economic Research Department of the Union Bank of Switzerland* (http://www.ubs.com) relativo allo studio del benessere economico di alcune capitali nel 1991. Si noti che l'uso dell'analisi multivariata in questo esempio ha soltanto valenza didattica, giacché l'esigua dimensione dello spazio la renderebbe superflua. Nella sezione 3 viene proposta e commentata l'analisi

esplorativa grafica per lo studio del caso di Vella e Verbeek, ed è sottolineata la concordanza dei risultati con quelli da essi ottenuti e che sono richiamati, insieme ai loro modelli, nell'Appendice.

Alcune brevi considerazioni conclusive si trovano nella sezione 4.

Le sequenze di comandi R utilizzati nel testo sono disponibili all'indirizzo: http://www.unica.it/~s.cabras/annali/city-workers.R.

2. L'ambiente statistico R

2.1. Alcune indicazioni generali

R è un ambiente di programmazione per l'analisi statistica distribuito gratuitamente in rete (http://www.r-project.org) sotto licenza GPL (*General Public License*), sviluppato da un team di statistici e informatici (*R Core Team*) ed aperto al contributo di tutti i ricercatori. È un software *open source*, essendo il suo codice sorgente distribuito liberamente e quindi aperto a chiunque voglia aumentarne le funzionalità. Una comunità attiva di utenti di R contribuisce, fin dalla sua creazione, alla produzione di nuovi *packages* (pacchetti) che sono censiti in rete presso il *Comprehensive R Archive Network* (*CRAN*, http://cran.r-project.org). Questa rete di raccolta permette di aggiornare continuamente il programma, scaricando i nuovi pacchetti e le nuove versioni dei pacchetti già installati attraverso un procedimento completamente visibile all'utente finale. Il buon utilizzo di R richiede dunque che il computer in uso sia connesso in rete.

Essendo R un ambiente di programmazione più che un programma propriamente detto, per ogni problema d'analisi di dati possono esistere diverse soluzioni, tutte valide in linea di principio. Da questo punto di vista, le tecniche utilizzate in questo lavoro, sebbene ampiamente consolidate, risentono comunque dell'esperienza e delle preferenze degli autori e potrebbero essere sostituite con altre ugualmente o diversamente efficaci: la flessibilità di R è la sua principale caratteristica.

I comandi di R sono digitati su una riga di comando della finestra principale denominata *consolle*, visibile appena aperto il programma. La riga inizia con il simbolo ">" chiamato *prompt* dei comandi. Ad esempio per creare un vettore con i valori 10, 20 e 30, si deve digitare:

```
>x=c(10,20,30)
```

In alternativa si può utilizzare

```
>x<-c(10,20,30)
```

La funzione `c()` indica la concatenazione di elementi, in questo caso gli scalari 10, 20, 30.

Per visualizzare il vettore appena creato è sufficiente richiamarlo con il suo simbolo, (*Nota*: nel seguito alcuni commenti saranno inseriti direttamente nel codice, così come accade quando si programma, avendo l'avvertenza di anteporre al commento il simbolo #):

```
>#  R differenzia maiuscole con minuscole
>X
Errore: oggetto "X" non trovato
>x
[1] 10 20 30
```

R è *orientato ad oggetti*. Tra gli oggetti principali richiamiamo: i vettori, le matrici, i dataframe (`data.frame()`) ed i grafici. Gli oggetti (o classi di oggetti) sono entità astratte, tecnicamente definiti da *proprietà* (es.: numero di righe di un vettore o titolo di un grafico) e *metodi* (es.: estrazione di un elemento da un vettore oppure aggiunta di un punto in un grafico).

L'elemento `x` appena creato è un'*istanza* dell'oggetto vettore dal quale possiamo estrarre alcuni elementi, ad es. il primo e l'ultimo:

```
>x[c(1,3)]
[1] 10 30
```

In un ambiente orientato ad oggetti sono fondamentali le *gerarchie* create tramite l'*ereditarietà* di metodi e di proprietà tra gli oggetti. Per esempio, l'oggetto *matrice* eredita i metodi e le proprietà dell'oggetto vettore, creando una gerarchia tra matrici e vettori. Conoscere R significa conoscere le gerarchie tra gli oggetti con i loro metodi e proprietà.

Una classe di elementi fondamentali in R sono le *funzioni*. Esse interagiscono con i metodi e le proprietà degli oggetti. Per esempio, possiamo creare una matrice 3×3 tramite la funzione `array()`

```
>X=array(x,dim=c(3,3))
>X
     [,1] [,2] [,3]
[1,]   10   10   10
[2,]   20   20   20
[3,]   30   30   30
```

La matrice `X` eredita quindi il metodo di estrazione di un elemento come mostra il seguente esempio:

```
>#  Estrazione dell'elemento in riga 1, colonna 3
>X[1,3]
[1] 10
```

Oggetti e funzioni sono contenute in librerie (o *packages*). Gli oggetti e le funzioni di base sono contenute nella libreria `base` che viene caricata in me-

moria automaticamente. I pacchetti si installano fisicamente sul computer scaricandoli dalla rete tramite il comando `install.packages("nome pacchetto")` e si richiamano nella memoria di R con il comando `library("nome pacchetto")`.

È importante saper ricorrere alle pagine di aiuto sulle funzioni, in particolare quelle in formato html accessibili tramite un motore di ricerca installato con R (menu Aiuto/Guida Html/Search Engine & Keywords).

Le pagine di aiuto hanno tutte le stessa struttura con le seguenti sezioni:

- *Description*: descrizione della funzione;
- *Usage*: sintassi di utilizzo della funzione con impostazione dei parametri di default;
- *Arguments*: descrizione dei parametri (o argomenti) della funzione;
- *Details*: descrizione dettagliata sul funzionamento della funzione;
- *Value*: valori che restituisce la funzione. In questa sezione sono elencati tutti i risultati che possono essere ottenuti dalla funzione;
- *Note*: ulteriori informazioni sulla funzione;
- *References*: riferimenti bibliografici sulla spiegazione del metodo statistico implementato nella funzione;
- *See Also*: altre funzioni relazionate con la funzione in esame;
- *Examples*: esempio dimostrativo della funzione. L'esempio è di solito autocontenuto, pertanto è possibile copiarlo e incollarlo nella consolle di R per eseguire la funzione.

Le pagine di aiuto su tutte le funzioni e pacchetti censiti sono disponibili sulla pagina web del *CRAN* (su collegamento "Search"). Per preparare l'ambiente di lavoro è utile cancellare la memoria di R e indicare la directory di lavoro nella quale è presente la base dati:

```
>rm(list=ls(all=TRUE))      # Svuota la memoria
>setwd("C:/miacartella")    # Directory di lavoro
```

2.2. *L'analisi esplorativa grafica*

Per introdurre le potenzialità di R, limitatamente all'analisi esplorativa con metodi grafici, utilizziamo una semplice base di dati relativa al benessere economico di alcune capitali nel 1991 (Fonte: *Economic Research Department of the Union Bank of Switzerland*). I dati si possono scaricare con R, seguendo la sequenza:

```
>download.file("http://www.unica.it/~s.cabras/annali/city.csv" ,dest="city.csv")
```

La base di dati è in formato CSV (facilmente ottenibile con Excel) dove i campi sono separati da un ";" (`sep=";"`); la prima riga contiene i nomi del-

le variabili (header=TRUE), mentre le etichette delle unità statistiche sono i nomi delle città (row.names="citta"):

```
>dati=read.table(file="city.csv",sep=";",header=TRUE,r
ow.names="citta")
```

La base dati, contenuta nell'oggetto dati, è costituita da 46 osservazioni e 4 variabili.

```
>dim(dati)
[1] 46  4
```

In particolare ogni città è caratterizzata dalle variabili

```
>colnames(dati)
[1] "continente" "ore" "prezzi" "salario"
```

Le variabili quantitative "ore" "prezzi" e "salario" indicano, rispettivamente, le ore annuali medie lavorate, l'indice dei prezzi al consumo e un indice del salario medio orario. La variabile qualitativa "continente" (tipo factor in R) indica il continente in cui è situata la città. Per rendere "visibili" in R le colonne della base dati come vettori si utilizza il comando attach()

```
>attach(dati) # Per un sommario dei dati utilizzare
summary(dati)
```

Iniziamo l'analisi studiando separatamente le singole variabili, ad esempio rappresentando la distribuzione di frequenza dell'indice dei prezzi. Possiamo iniziare con un semplice istogramma (Figura 1) dove nelle ordinate sono riportate le frequenze relative delle singole classi (prob=TRUE):

```
>hist(prezzi,xlim=c(3,150),xlab="Prezzi",ylab="Frequenze
Relative/Densità",main="Distribuzione marginale dei
Prezzi",prob=TRUE)
```

L'istogramma è un'approssimazione "grossolana" della distribuzione, tanto più quando le numerosità campionarie non sono elevate come in questo caso. Esistono altri metodi possibili di rappresentare la distribuzione della variabile a partire dai dati osservati: ipotizzando una distribuzione parametrica, come ad esempio la distribuzione normale. Per disegnare la densità normale sull'istogramma si utilizza il comando lines(x,y):

```
># Il vettore dei primi n interi si può ottenere con 1:n
># x e y sono vettori di punti che definiscono la linea
># dnorm(x, mean, sd) restituisce la densità della di-
stribuzione
># normale con media=mean e scarto=sd.
>lines(x=0:140,y=dnorm(0:140,mean=mean(prezzi),sd=sd(prez
zi)),lty=3)
```

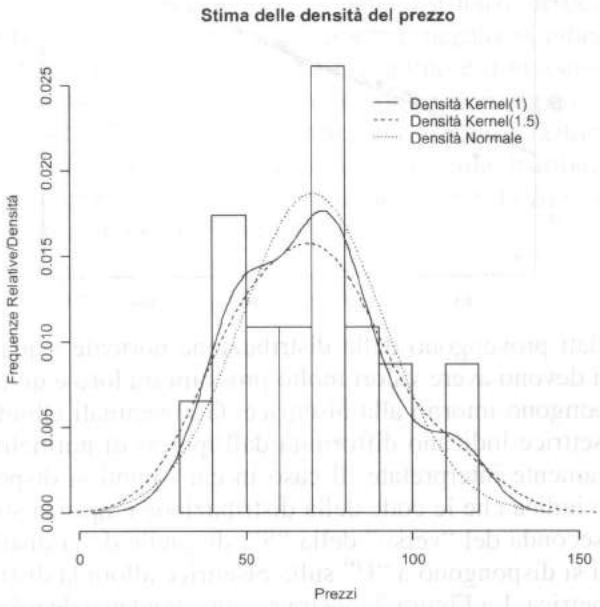
Un altro metodo, (metodo di stime kernel di densità (Silverman, 1986)), consiste nello stimare la densità (`density()`) ricorrendo alla convoluzione di funzioni denominate *kernel*, per esempio densità normali. La stima della densità dipende da un parametro di "lisciatura" (`adjust`). Qui sono riportati due esempi con valori diversi di adjust (Figura 1):

```
>lines(density(prezzi,adjust = 1)) # Meno liscia
>lines(density(prezzi,adjust = 1.5),lty=2) # Più liscia
```

Una legenda (`legend()`) rende più agevole la lettura del grafico:

```
>legend(85,0.025,c("Densità Kernel(1)","Densità
Kernel(1.5)","Densità Normale"),lty=1:3,bty="n")
```

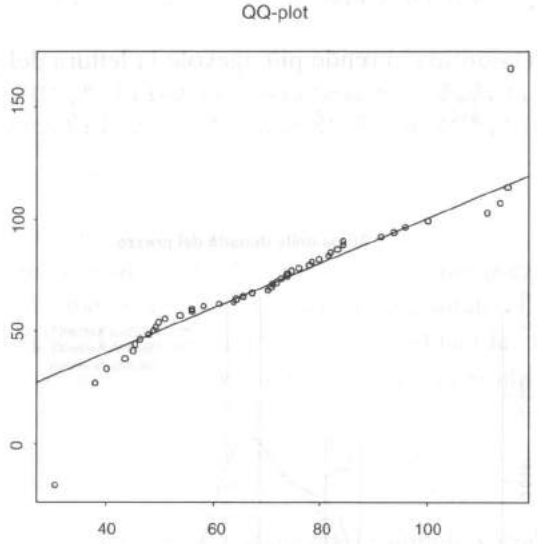
Figura 1



La Figura 1 mostra le differenti stime della distribuzione dei prezzi. Esse si differenziano per la diversa probabilità assegnata nelle code della distribuzione. Con l'istogramma, la variabilità stimata dei prezzi coincide con il rango osservato. Assumendo invece il modello normale si ottengono, in questo caso, code meno "pesanti" rispetto a quelle stimate con la densità kernel. L'assunzione di normalità per l'indice dei prezzi può essere verificata, oltre che con test non-parametrici, come il test di Kolmogorov-Smirnov (Stephens, 1986), anche con un diagramma del tipo *QQ-plot* (`qqplot()`) dove in ascisse e ordinate sono comparati, rispettivamente, i quantili delle distribuzioni empirica e teorica (normale in questo caso):

```
>qqplot(prezzi,rnorm(10000,mean=mean(prezzi),sd=sd(pre
zzi)),ylab ="Quantili teorici",xlab="Quantili empirici",
main="QQ-plot")
>#Aggiungiamo la bisettrice del piano
>abline(0,1)
```

Figura 2



Quando i dati provengono dalla distribuzione normale i quantili teorici e quelli empirici devono avere valori molto prossimi tra loro e quindi i punti del grafico si dispongono intorno alla bisettrice. Gli eventuali allontanamenti dei punti dalla bisettrice indicano difformità dall'ipotesi di normalità che possono essere variamente interpretate. Il caso in cui i punti si dispongono a "S" sulla bisettrice indica che le code della distribuzione empirica sono più o meno pesanti, a seconda del "verso" della "S", di quelle della distribuzione teorica. Se i punti si dispongono a "U" sulla bisettrice allora la distribuzione empirica è asimmetrica. La Figura 2 mostra sia una, tendenziale asimmetria (a sinistra o negativa), sia una diversa pesantezza delle code della distribuzione empirica rispetto alla normale.

Un'altra rappresentazione molto efficace dei dati osservati è offerta dai BoxPlot (`boxplot()`):

```
>mioboxplot=boxplot(prezzi,main="BoxPlot
Prezzi",ylab="Prezzi", xlim=c(20,130))
>#  Questo testo chiarisce la rappresentazione boxplot
>#  e i valori restituiti dalla funzione boxplot (mio-
boxplot$stats)
>text(x=1.3,y=mioboxplot$stats[,1],labels=c("min","I
quartile", "mediana","III quartile","max"))
```

an(prezzi),sd=sd(pre
'Quantili empirici",
ano

100

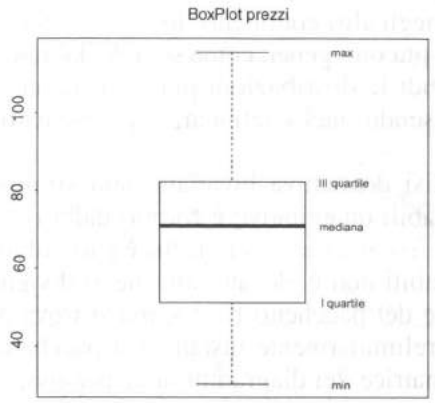
normale i quantili teorici e
tra loro e quindi i punti del
ventuali allontanamenti dei
si di normalità che posso-
punti si dispongono a "S"
e empirica sono più o me-
lle della distribuzione teo-
allora la distribuzione em-
ndenziale asimmetria (a si-
e code della distribuzione

lati osservati è offerta dai

,main="BoxPlot
esentazione boxplot
zione boxplot (mio-

,labels=c("min","I
nax"))

Figura 3

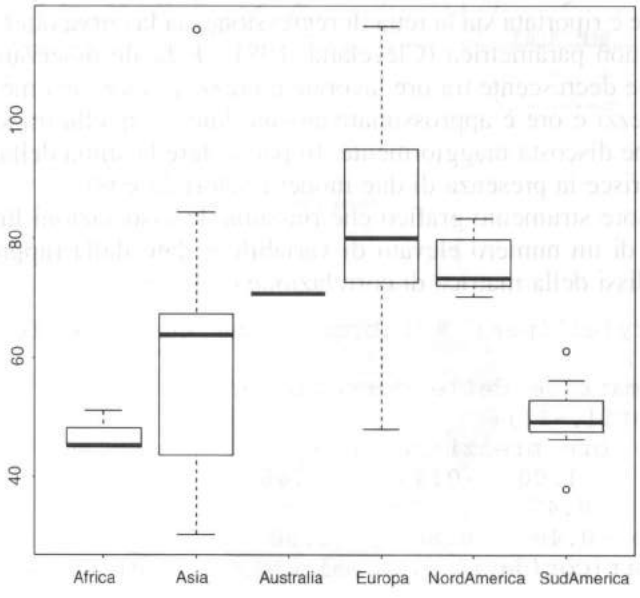


In Figura 3 è rappresentato il BoxPlot della distribuzione dei prezzi. Anche in questa rappresentazione appare la asimmetria (negativa), infatti la mediana è più vicina al terzo quartile che al primo. Il minimo e il massimo possono non comparire nei segmenti estremi quando esistono valori atipici (si veda l'aiuto `>?boxplot`). Si noti che il BoxPlot fornisce una rappresentazione sintetica della distribuzione empirica senza alcuna assunzione sulla distribuzione teorica.

Con il BoxPlot possiamo anche visualizzare le distribuzioni dei prezzi condizionatamente ai continenti (Figura 4):

```
>boxplot(prezzi~continente,main="BoxPlot Prezzi rispet-
to ai continenti",ylab="Prezzi",cex.axis=0.8)
```

Figura 4



La Figura 4 suggerisce che in Europa e nord America i prezzi al consumo sono più elevati che negli altri continenti. In Africa e Sud America i prezzi sono inferiori ed anche più omogenei, come si vede dal fatto che le "scatole" sono più piccole e quindi le distribuzioni più concentrate. Il BoxPlot dell'Australia è degenere essendo quel continente rappresentato da un solo dato: la città di Sidney.

Passando all'analisi descrittiva bivariata, uno strumento molto utile per rappresentare le variabili quantitative è fornito dalla matrice di diagrammi a dispersione (`scatterplot.matrix()`), che è particolarmente efficace quando il numero di variabili non è elevato, altrimenti diviene illeggibile. Questa funzione non fa parte del pacchetto `base`, ma si trova nel pacchetto `car`. È dunque necessario preliminarmente installare il pacchetto `car` e successivamente disegnare la matrice dei diagrammi a dispersione:

```
>#(Indicare il server dal quale scaricare il pacchetto)
>install.packages("car")
>#Carichiamo il pacchetto in memoria
>library(car)
>#Disegniamo la matrice dei diagrammi a dispersione,
considerando l'intera base dati tranne la colonna conti-
nente (dati[,-1])
>scatterplot.matrix(dati[,-1],groups=continente, col=rep
("black",7),cex=1.2) # Simboli 20% più grandi (cex=1.2)
```

La Figura 5 mostra i diagrammi a dispersione delle tre variabili quantitative prese a due a due. La distribuzione marginale stimata con metodo kernel è rappresentata nei grafici sulla diagonale della matrice. Nei singoli diagrammi a dispersione è riportata sia la retta di regressione sia la curva ottenuta con una regressione non parametrica (Cleveland, 1981). È facile osservare che esiste una relazione decrescente tra ore lavorate e prezzi e salari, ma mentre la relazione tra prezzi e ore è approssimativamente lineare, quella tra salari ed ore lavorate se ne discosta maggiormente. In particolare la stima della densità dei salari suggerisce la presenza di due mode: i valori 25 e 60.

Un ulteriore strumento grafico che riassume le associazioni lineari, anche in presenza di un numero elevato di variabili, è dato dalla rappresentazione mediante ellissi della matrice di correlazione (`cor()`):

```
>library(ellipse) # Libreria che contiene la funzione
plotcorr()
>#  La matrice delle correlazioni
>cor(dati[,-1])
            ore prezzi salario
ore        1.00  -0.45   -0.46
prezzi    -0.45   1.00    0.80
salario   -0.46   0.80    1.00
>plotcorr(cor(dati[,-1]),main="Matrice di Correlazione")
```

merica i prezzi al consumo
e Sud America i prezzi so-
lal fatto che le "scatole" so-
ntrate. Il BoxPlot dell'Au-
sentato da un solo dato: la

strumento molto utile per
lla matrice di diagrammi a
rticolarmente efficace quan-
diviene illeggibile. Questa
trova nel pacchetto `car`. È
acchetto `car` e successiva-
rsione:

```
ricare il pacchetto)
a
ammi a dispersione,
e la colonna conti-
=continente, col=rep
ù grandi (cex=1.2)
```

delle tre variabili quantitati-
timata con metodo kernel è
rice. Nei singoli diagrammi
ia la curva ottenuta con una
facile osservare che esiste
e salari, ma mentre la rela-
are, quella tra salari ed ore
re la stima della densità dei
25 e 60.
associazioni lineari, anche
dato dalla rappresentazione
r( )):

```
contiene la funzione
```

```
e di Correlazione")
```

Figura 5

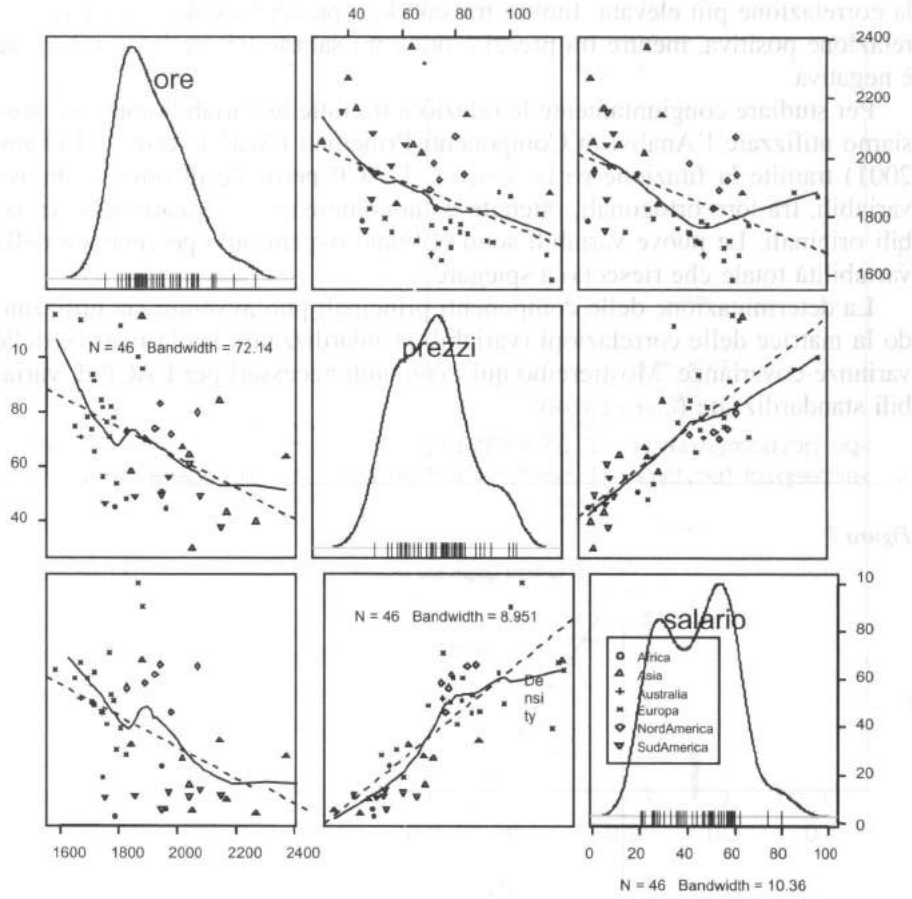
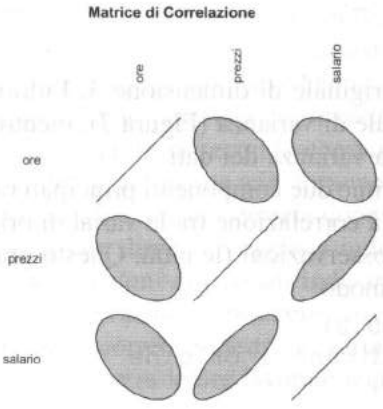


Figura 6



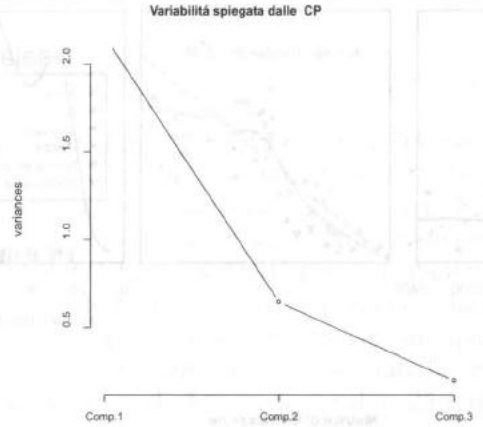
Nella Figura 6 l'associazione lineare tra prezzi e salari è positiva e presenta la correlazione più elevata. Inoltre tra salario e prezzi è evidenziata una correlazione positiva, mentre tra prezzi e ore e tra salario ed ore la correlazione è negativa.

Per studiare congiuntamente le relazioni tra tutte le variabili continue possiamo utilizzare l'Analisi in Componenti Principali (ACP, Everitt and Dunn, 2001) tramite la funzione `princomp()`. L'ACP permette di ottenere nuove variabili, tra loro ortogonali, ottenute dalla combinazione lineare delle variabili originali. Le nuove variabili sono ordinate rispetto alla percentuale della variabilità totale che riescono a spiegare.

La determinazione delle componenti principali può avvenire sia utilizzando la matrice delle correlazioni (variabili standardizzate), sia la matrice delle varianze-covarianze. Mostreremo qui i comandi necessari per l'ACP di variabili standardizzate (`cor=TRUE`):

```
>pc=princomp(dati[,-1],cor=TRUE)
>screeplot(pc,type="lines",main="Variabilità spiegata dalle CP")
```

Figura 7

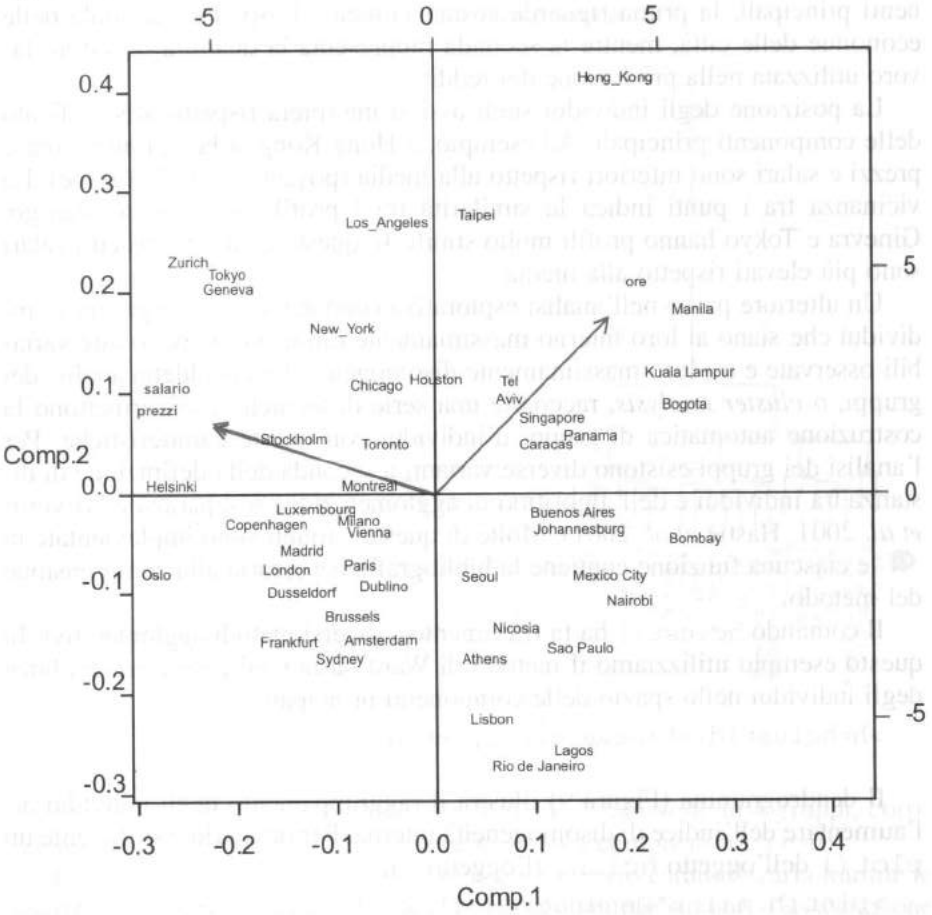


Essendo lo spazio originale di dimensione 3, l'ultima componente "spiega" una bassa percentuale di varianza (Figura 7), mentre le prime due spiegano la quasi totalità della varianza dei dati.

Per interpretare le prime due componenti principali rappresentiamo sul piano da esse individuato la correlazione tra le variabili originali e le componenti congiuntamente alle osservazioni (le città. Questo grafico, chiamato *biplot*, si ottiene nel seguente modo:

```
>biplot(pc,cex=0.8)
># Le linee indicano l'origine
>abline(v=0,h=0)
```

Figura 8



Nel biplot (Figura 8) sono rappresentate due scale: la prima (gradazioni a sinistra e in basso) si riferisce alle correlazioni, indicate con le frecce, tra variabili e componenti; la seconda fornisce le coordinate dei punti sulle componenti principali (gradazione a destra e in alto). Dal grafico si evince che la prima componente è negativamente correlata con i prezzi ed i salari e positivamente con le ore lavorate; mentre la seconda componente è correlata positivamente con tutte e tre le variabili, ma più fortemente con le ore lavorate. Frecce con la stessa direzione indicano correlazione positiva tra variabili (si veda prezzi-salari); frecce con direzione opposta indicano correlazione negativa. Infine, le frecce con direzioni ortogonali indicano assenza di correlazione. Sulla base di queste osservazioni è possibile, avere una rappresentazione sintetica della struttura della correlazione tra le variabili: i prezzi e salari sono tra loro molto correlati; mentre le ore lavorate tendono ad essere pochissi-

mo (e negativamente) correlate con le altre due. Interpretando le due componenti principali, la prima riguarda sostanzialmente l'aspetto reddituale delle economie delle città, mentre la seconda rappresenta la quantità di fattore lavoro utilizzata nella produzione del reddito.

La posizione degli individui sugli assi si interpreta rispetto al significato delle componenti principali. Ad esempio, a Hong Kong si lavora molte ore e prezzi e salari sono inferiori rispetto alla media (posizionata nell'origine). La vicinanza tra i punti indica la similarità tra i profili: ad esempio Zurigo, Ginevra e Tokyo hanno profili molto simili. In queste città i prezzi ed i salari sono più elevati rispetto alla media.

Un ulteriore passo nell'analisi esplorativa conduce a formare gruppi di individui che siano al loro interno massimamente omogenei rispetto alle variabili osservate e tra loro massimamente disomogenei. La cosiddetta analisi dei gruppi, o *cluster analysis*, raccoglie una serie di tecniche che permettono la costruzione automatica di gruppi d'individui con queste caratteristiche. Per l'analisi dei gruppi esistono diverse varianti a seconda della definizione di distanza tra individui e dell'algoritmo di agglomerazione o separazione (Everitt *et al.* 2001, Hastie *et al.* 2001). Molte di queste varianti sono implementate in R e ciascuna funzione contiene la bibliografia necessaria alla comprensione del metodo.

Il comando `hclust()` ha fa riferimento a diversi metodi agglomerativi. In questo esempio utilizziamo il metodo di Ward basato sulle distanze euclidee degli individui nello spazio delle componenti principali.

```
>h=hclust(dist(pc$scores),method="ward")
```

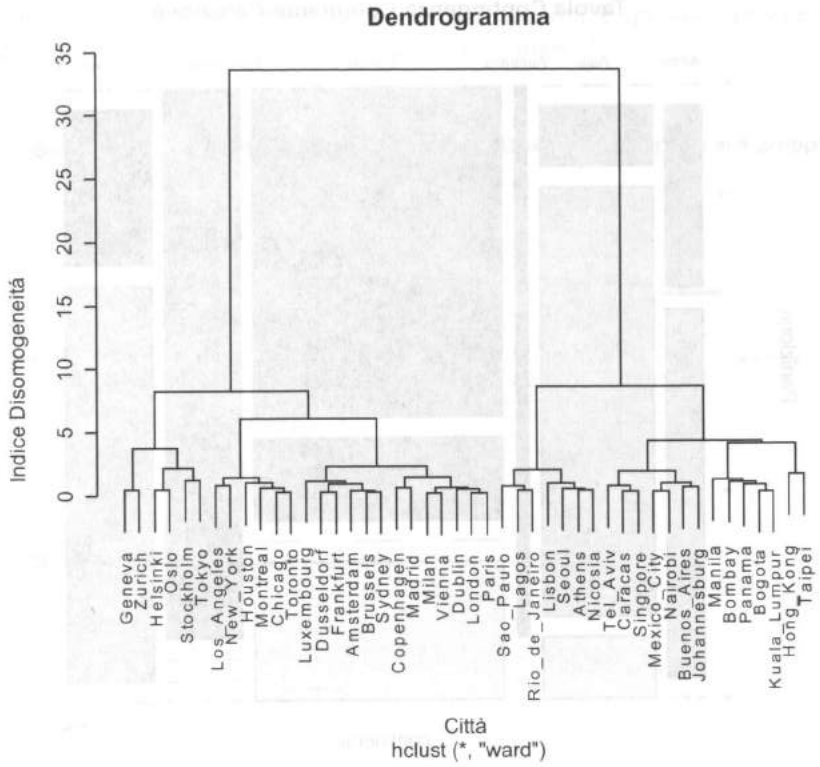
Il dendrogramma (Figura 9) illustra il raggruppamento degli individui all'aumentare dell'indice di disomogeneità interna. Per ottenerlo è sufficiente un `plot()` dell'oggetto `hclust` (l'oggetto `h`).

```
>plot(h,main="Dendrogramma",xlab="Città",ylab="Indice
Disomogeneità")
```

La Figura 9 illustra il procedimento agglomerativo delle città. Possiamo osservare, ad esempio, che Ginevra e Zurigo sono simili tra loro e formano un gruppo per valori molto bassi di disomogeneità. Il gruppo formato da Ginevra e Zurigo è dissimile da quello formato da Taipei e Hong Kong, infatti, questi due gruppi si aggregano solo per elevati livelli di disomogeneità.

Il dendrogramma può essere di aiuto per la scelta del numero di gruppi da analizzare. Infatti la maggiore diminuzione dell'indice di disomogeneità si ottiene dividendo l'intero campione in due gruppi, mentre le successive partizioni comportano una diminuzione minore della disomogeneità interna. Questo implica che un numero elevato di gruppi (al massimo 46) implica una maggiore omogeneità interna. La scelta del numero di gruppi è inoltre legata ad esigenze di sintesi del fenomeno analizzato e alla interpretabilità dei grup-

Figura 9



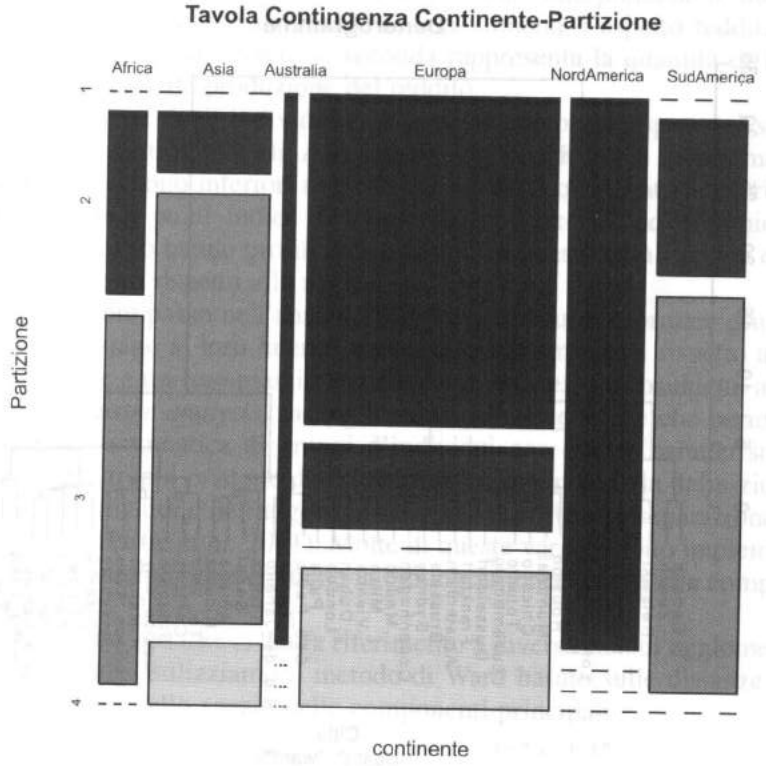
pi. In questo caso appare ragionevole dividere il campione in 4 gruppi, corrispondenti ad un indice di disomogeneità interna compreso tra 5 e 10.

Per poter interpretare la suddivisione è necessario caratterizzarla tramite le variabili presenti nella base dei dati. In particolare per studiare l'associazione tra gruppi e continenti è utile rappresentare la corrispondente tavola di contingenza (`table()`) tramite un diagramma denominato *mosaico* (`plot(table())`).

```
>#  Costruiamo il vettore che definisce la partizione in
4 gruppi
>partizione=cutree(h,k=4)
>plot(table(continente,partizione),col=grey(0:4/4),mai
n="Tavola Contingenza Continente-Partizione")
```

La Figura 10 mostra il mosaico che rappresenta la tavola di contingenza tra gruppo (righe) e continente (colonne). La dimensione della casella è proporzionale alla frequenza congiunta. La dimensione orizzontale della casella è proporzionale al profilo riga, mentre quella verticale al profilo colonna. Per agevolarne la lettura, la colorazione delle caselle è per gruppi (righe).

Figura 10

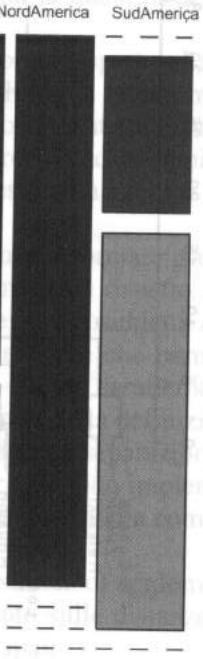


Analizzando la Figura 10, si osserva che i continenti Africa, Asia e Sud America sono essenzialmente caratterizzanti del gruppo 3, infatti esso contiene solo città situate in questi tre continenti e la maggior parte delle città africane, asiatiche e sud americane sono contenute in esso. Considerazioni analoghe possono essere fatte su Australia, Europa e Nord America per il gruppo 1.

Per descrivere i 4 gruppi rispetto alle tre variabili quantitative utilizziamo nuovamente i BoxPlot:

```
>par(mfrow=c(2,2))
>boxplot(prezzi~partizione,main="BoxPlot prezzi rispetto ai 4 gruppi",ylab="Prezzi",cex.axis=0.8)
>boxplot(ore~partizione,main="BoxPlot ore rispetto ai 4 gruppi", ylab="Ore",cex.axis=0.8)
>boxplot(salario~partizione,main="BoxPlot salario rispetto ai 4 gruppi",ylab="Salario",cex.axis=0.8)
```

La Figura 11 indica che i salari ed i prezzi al consumo più elevati sono caratterizzanti del gruppo 4 dove il numero d'ore lavorate tende ad essere più basso rispetto agli altri gruppi. Il terzo gruppo è caratterizzato da un numero



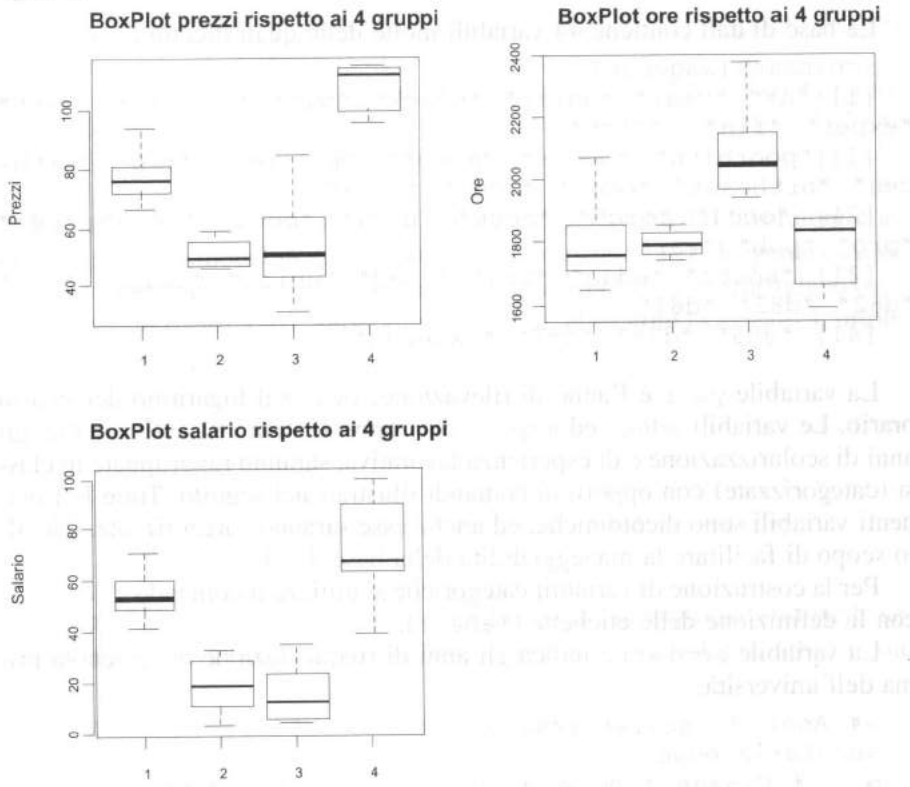
tinenti Africa, Asia e Sud
ppo 3, infatti esso contie-
ggior parte delle città afri-
sso. Considerazioni analo-
l America per il gruppo 1.
li quantitative utilizziamo

```
Plot prezzi rispet-
=0.8)
t ore rispetto ai 4
xPlot salario ri-
.axis=0.8)
```

sumo più elevati sono ca-
orate tende ad essere più
atterizzato da un numero

d'ore lavorate superiore rispetto agli altri gruppi. I gruppi 1 e 2 si differenziano tra loro, nonostante il numero di ore lavorate sia simile, perché prezzi e salari del gruppo 1 sono più elevati di quelli del gruppo 2.

Figura 11



3. Studio di un caso reale

Allo scopo di mostrare le potenzialità interpretative dell'analisi esplorativa grafica anche ai fini della costruzione di modelli, con particolare riguardo allo studio di un panel data di natura socio-economica, consideriamo la base dati relativa a 545 giovani statunitensi occupati, seguiti nella loro carriera lavorativa nei primi 8 anni di lavoro dal 1980 al 1987. La base dati fornita dal *National Longitudinal Survey* (http://www.bls.gov/nls/home.htm) può essere caricata in R direttamente dalla pagina web:

```
>#  Scarica  il  file  dalla  rete
>download.file("http://www.stata.com/data/jwooldridge/eacsap/wagepan.dta"  ,dest="wagepan.dta",mode="wb")
>library(foreign)  #  Libreria  per  la  conversione  files
```

```
>#  Leggi  il  file  in  formato  STATA
>wagepan=read.dta(file="wagepan.dta",  convert.dates  =
TRUE,  tz  =  NULL,  convert.factors  =  TRUE,  missing.type  =
FALSE,  convert.underscore=TRUE,  warn.missing.labels=TRUE)
>attach(wagepan)
```

La base di dati contiene 44 variabili molte delle quali dicotomiche.

```
>colnames(wagepan)
 [1]  "nr"  "year"  "agric"  "black"  "bus"  "construc"  "ent"
"exper"  "fin"  "hisp"
 [11]"poorhlth"  "hours"  "manuf"  "married"  "min"  "nrth-
cen"  "nrtheast"  "occ1"  "occ2"  "occ3"
 [21]  "occ4"  "occ5"  "occ6"  "occ7"  "occ8"  "occ9"  "per"
"pro"  "pub"  "rur"
 [31]  "south"  "educ"  "tra"  "trad"  "union"  "lwage"  "d81"
"d82"  "d83"  "d84"
 [41]  "d85"  "d86"  "d87"  "expersq"
```

La variabile `year` è l'anno di rilevazione, `lwage` il logaritmo del salario orario. Le variabili `educ` ed `exper` che rappresentano, rispettivamente, gli anni di scolarizzazione e di esperienza lavorativa, saranno raggruppate in classi (categorizzate) con opportuni comandi illustrati nel seguito. Tutte le rimanenti variabili sono dicotomiche, ed anche esse saranno categorizzate. Ciò allo scopo di facilitare la maneggiabilità della base dei dati.

Per la costruzione di variabili categoriche si utilizza il comando `factor()` con la definizione delle etichette (`labels`).

La variabile `scolariz` indica gli anni di frequentazione della scuola prima dell'università.

```
>#  Anni  di  scolarizzazione
>scolariz=educ
>    #  Classe  fino  a  9  anni  di  frequentazione
>    scolariz[scolariz<=9]=9
>scolariz=factor(scolariz,labels=c("<=9","10","11","12
","13","14","15","16"))
```

La variabile `esperlav` indica gli anni di esperienza lavorativa.

```
>#  Anni  di  esperienza  lavorativa
>esperlav=exper
>    #  Definizione  delle  Classi  estreme
>    esperlav[esperlav<=5]=5
>    esperlav[esperlav>=12]=12
>esperlav=factor(esperlav,labels=c("<=5","6","7","8","
9","10","11","12"))
```

La variabile `settore` indica il settore di attività del lavoratore e assume le seguenti modalità: agricoltura (`agric`), settore minerario (`miner`), setto-

re delle costruzioni (const), commercio (comme), trasporti (trasp), finanziario (finan), servizi alle imprese (servimp), servizi alle persone (servper), attività ricreative (ricrea), manifattura (manif), attività professionali (profe), pubblica amministrazione (pammin).

```
>#  Settore di occupazione dei lavoratori
>settore=agric+2*min+3*construc+4*trad+5*tra+6*fin+7*b
us+8*per+9*ent+10*manuf+11*pro+12*pub
>settore=factor(settore,labels=c("agric","miner","con-
st","comme","trasp","finan","servimp","servper","ri-
crea","manif","profe","pammin"))
```

La variabile attivita indica il tipo di attività svolta dal lavoratore e assume le seguenti modalità: lavoratore in proprio (profess), dirigente (manager), commerciante (commer), segretario/segretaria (segret), artigiano (artig), operaio (operaio), agricoltore (agric), allevatore (allevat), impiegato in genere (impiegat).

```
>#  Attivita' lavorativa svolta
>attivita=occ1+2*occ2+3*occ3+4*occ4+5*occ5+6*occ6+7*oc
c7+8*occ8+9*occ9
>attivita=factor(attivita,labels=c("profess","mana-
ger","commerc","segret","artig","operaio","agricol","al-
levat","impiegat"))
```

La variabile zona indica la zona di residenza del lavoratore e assume le seguenti modalità: ovest (ovest), centro (centro), nordest (nordest), sud (sud).

```
>#  Zona di residenza del lavoratore
>zona=nrthcen+2*nrtheast+3*south
>zona=factor(zona,labels=c("ovest","centro","norde-
st","sud"))
```

La variabile etnia indica l'etnia del lavoratore con le seguenti modalità: bianco (bianco), africano (african), ispanico (ispanico).

```
>#  Etnia
>etnia=black+2*hisp
>etnia=factor(etnia,labels=c("bianco","african","ispa-
nico"))
```

La variabile salute indica lo stato di salute del lavoratore al momento dell'intervista.

```
>#  Salute
>salute=factor(poorhlth,labels=c("sano","malato"))
```

La variabile `statocivile` indica lo stato civile del lavoratore al momento dell'intervista.

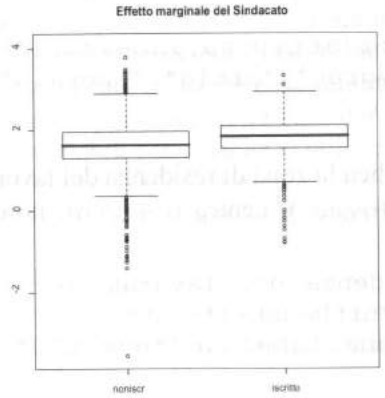
```
>#  Stato Civile
>statocivile=factor(married,labels=c("nonsposato","spo
sato"))
>#  Iscrizione al sindacato
>sindacato=factor(union,labels=c("noniscr","iscritto"))
```

L'analisi esplorativa sarà effettuata con specifica attenzione allo scopo del già citato lavoro di Vella and Verbeek, mettendo quindi in luce l'effetto dell'affiliazione al sindacato per poter fare i confronti con i risultati da loro ottenuti, (vedi Appendice).

Il BoxPlot in Figura 12 mostra l'*effetto sindacato* marginalmente alle differenti condizioni socio-economiche dei lavoratori: i lavoratori iscritti tendono ad avere salari più elevati.

```
>boxplot(lwage~sindacato,ylab="Log Salario",main="Effetto
marginale del Sindacato")
```

Figura 12



L'effetto sindacato, condizionatamente ad alcune condizioni socio-economiche e la tendenza negli anni 1980-1987, è evidenziato dalla serie di BoxPlot riportati in Figura 13.

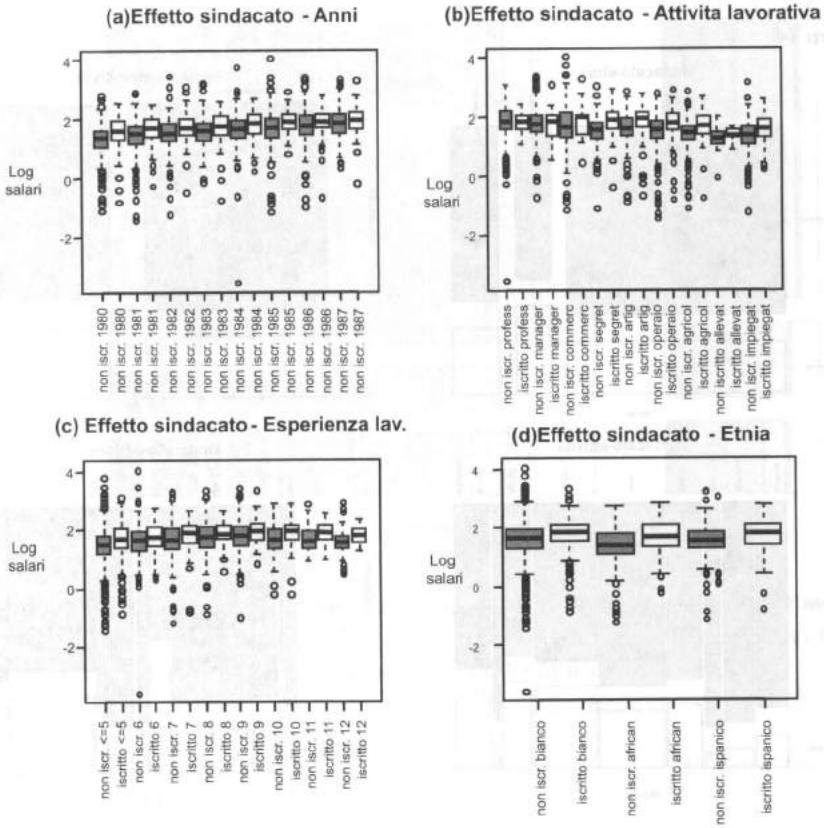
```
>par(mfrow=c(2,2))
>boxplot(lwage~sindacato*year,col=grey(1:2/2),las=2,ce
x.axis=0.8,main="(a)Effetto sindacato - Anni",ylab="Log
salari")
>boxplot(lwage~sindacato*attivita,las=2,col=grey(1:2/2
),cex.axis=0.8,main="(b)Effetto sindacato - Attivita la-
vorativa",ylab="Log salari")
>boxplot(lwage~sindacato*esperlav,las=2,col=grey(1:2/2),cex.
axis=0.8,main="(c)Effetto sindacato - Esperienza lav.",ylab="Log
salari")
```

```
>boxplot(lwage~sindacato*etnia,las=2,col=grey(1:2/2),c
ex.axis=0.7,main="(d)Effetto sindacato - Etnia",ylab="Log
salari")
```

Dalla Figura 13 si evince che l'effetto sindacato è più elevato per i salari bassi ed inoltre si osservano facilmente alcune altre caratteristiche di tale effetto che pure emergono dall'analisi di Vella and Verbeek. In particolare (Figura 13 (a)) vi è una tendenza generale all'incremento dei salari, con un effetto sindacato maggiore nei primi anni. Possono essere date varie interpretazioni alle interazioni tra l'effetto sindacato e le altre variabili:

- l'interazione con gli anni di esperienza lavorativa rivela che l'effetto sindacato è maggiore per i lavoratori con più anni di esperienza (Figura 13 (c)):
- l'interazione con l'etnia rivela che l'effetto sindacato è più elevato per i lavoratori di colore (Figura 13 (d));
- l'interazione con il tipo di lavoro (Figura 13 (b)) indica che l'effetto sindacato è più elevato per operai, artigiani, agricoltori e impiegati. L'effetto appare invece negativo per i manager.

Figura 13

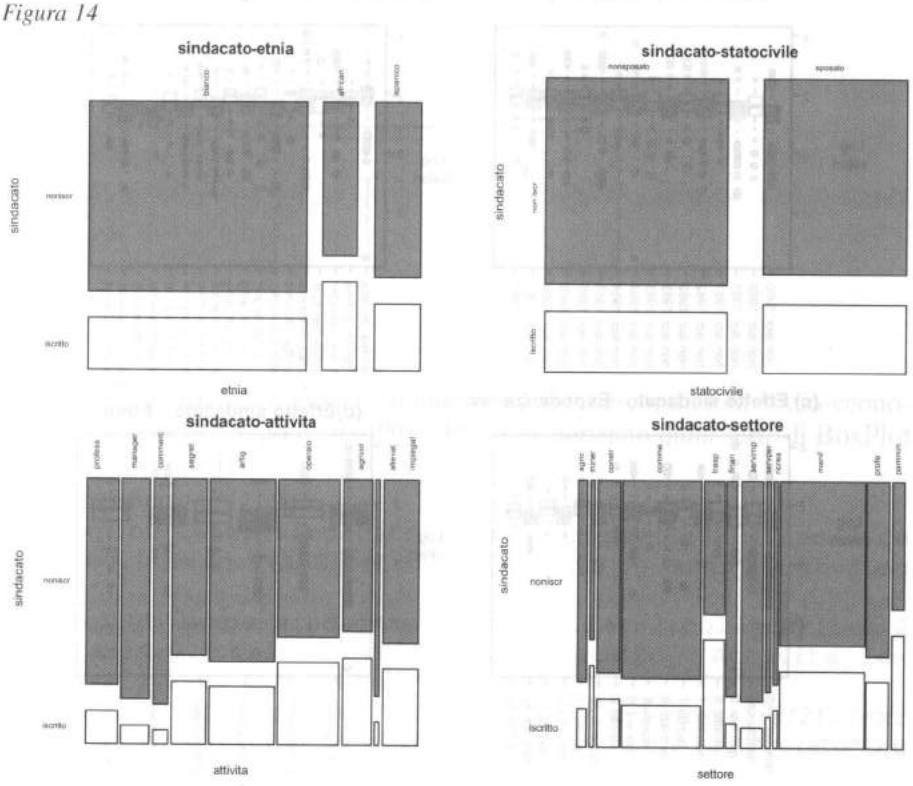


Attraverso l'analisi dei dati si vogliono anche caratterizzare, rispetto a certe variabili socio-economiche, i lavoratori affiliati al sindacato. La seguente serie di mosaici (Figura 14) descrive l'appartenenza al sindacato rispetto alle variabili: etnia, stato civile, tipo di attività lavorativa e settore economico. Si noti che queste variabili sono anche quelle che risultano più significative nel modello probit stimato da Vella and Verbeek (1998).

```
>par(mfrow=c(2,2),mar=c(2.1, 4, 2, 2)) # mar=c(bottom, left,
top, right)
>plot(table(etnia,sindacato),col=grey(1:2/2),las=2,main="sin
dacato-etnia")
>plot(table(statocivile,sindacato),col=grey(1:2/2),main="sin
dacato-statocivile")
>plot(table(attivita,sindacato),col=grey(1:2/2),las=2,main="
sindacato-attivita")
>plot(table(settore,sindacato),col=grey(1:2/2),cex.axis=0.8,
las=2,main="sindacato-settore")
```

Dalla Figura 14 si osserva inoltre che l'affiliazione al sindacato è più diffusa tra gli africani, gli sposati, agricoltori ed operai e coloro che lavorano nel settore dei trasporti e della pubblica amministrazione.

Figura 14



aratterizzare, rispetto a cer-
i al sindacato. La seguente
za al sindacato rispetto alle
iva e settore economico. Si
sultano più significative nel
8).

```
 # mar=c(bottom, left,
1:2/2),las=2,main="sin
=grey(1:2/2),main="sin
ey(1:2/2),las=2,main="
y(1:2/2),cex.axis=0.8,
```

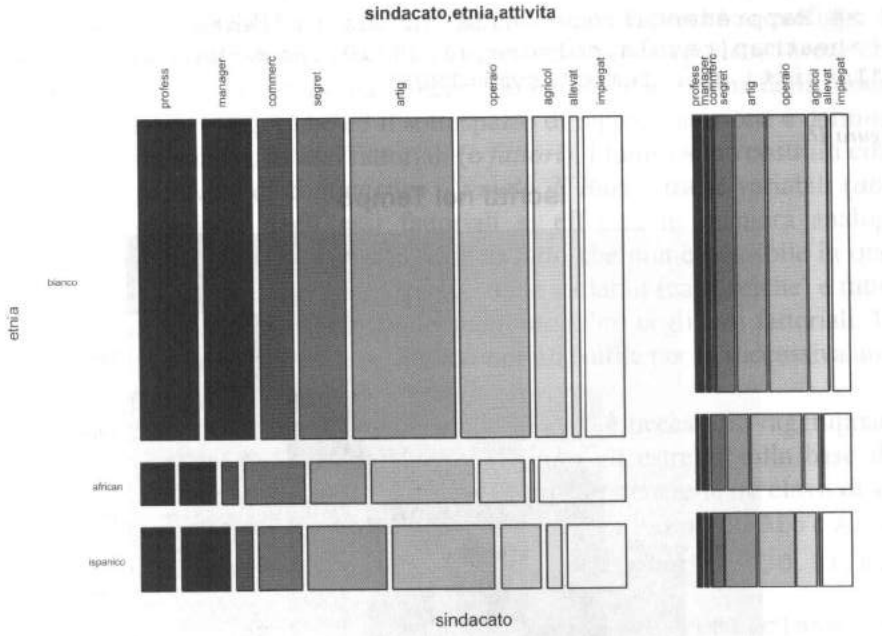
ione al sindacato è più dif-
ai e coloro che lavorano nel
one.

Il mosaico permette di rappresentare la relazione congiunta tra più variabili categoriche. Consideriamo ad esempio la relazione tra affiliazione al sindacato, etnia e tipo di attività. Questa può essere visualizzata con il seguente mosaico a tre dimensioni:

```
>plot(table(sindacato,etnia,attivita),col=grey(1:9/9),
las=2,main="sindacato,etnia,attivita",cex.axis=0.7)
```

La Figura 15 mostra il mosaico della tavola di contingenza a tre vie. A sinistra sono rappresentate le frequenze congiunte tra le variabili `etnia` e `attivita` lavorative per i non iscritti; mentre a destra quelle per gli iscritti al sindacato. La colorazione delle caselle è per attività. Si può quindi notare che il gruppo più numeroso è quello degli artigiani bianchi non iscritti. Mentre tra gli iscritti, il gruppo più numeroso è costituito da operai bianchi o africani.

Figura 15



È interessante anche valutare l'affiliazione dei lavoratori al sindacato rispetto agli anni di rilevazione. Con i seguenti comandi si costruisce una tavola di migrazione *da e verso* il sindacato rispetto agli anni:

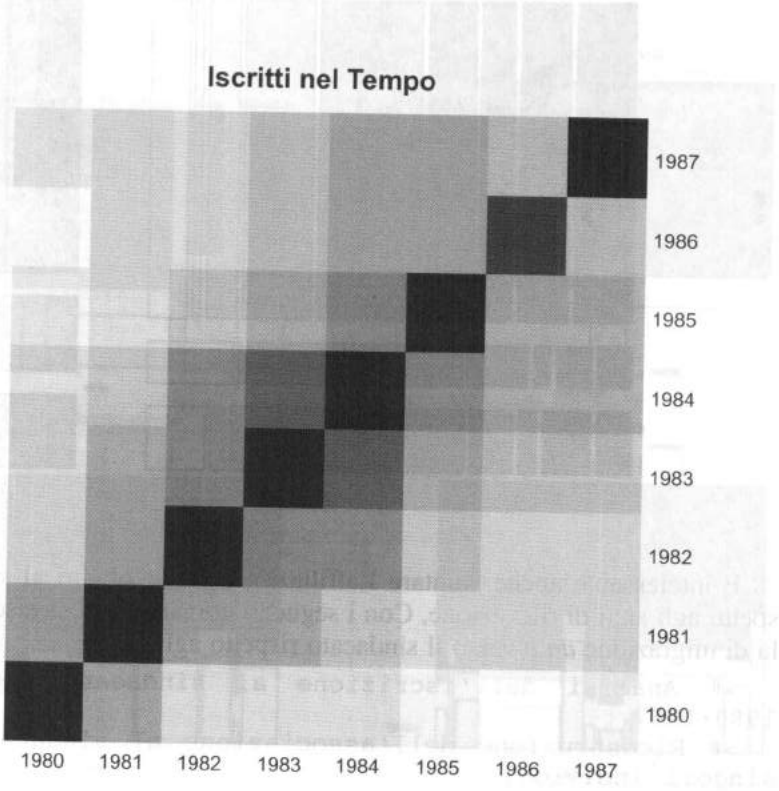
```
>#  Analisi  dell'iscrizione  al  sindacato  negli  anni
1980:1987
>#  Ricostruzione  dell'associazione  al  sindacato  per  i
singoli  individui
```

```
>percorsi=apply(as.matrix(unique(nr)),1,function(i)
sindacato[which(nr==i)])-1
>rownames(percorsi)=1980:1987
># Tavola delle migrazioni
>tavola=percorsi%*%t(percorsi)
```

La matrice `tavola` di dimensioni 8x8 è simmetrica (righe e colonne indicano gli anni 1980-1987). La diagonale indica il numero di lavoratori iscritti in un dato anno al sindacato, mentre l'elemento, $(i,j,j>i)$ indica il numero di lavoratori iscritti nell'anno i che continuano ad essere iscritti nell'anno j. A differenza di Vella e Verbeek che nel loro lavoro considerano solamente la migrazione a distanza di un anno (j-i=1), qui l'uso dell'intera matrice fornisce ulteriori informazioni sul fenomeno di abbandono del sindacato. La rappresentazione della matrice (`tavola`) può essere ottenuta mediante la funzione `heatmap()`, specificando che la matrice è simmetrica (`symm=TRUE`):

```
># Rappresentazione della Tavola mediante un heat map
>heatmap(tavola,col=grey(10:0/10),Rowv=NA,Colv=NA,main
="Iscritti nel Tempo",symm=TRUE)
```

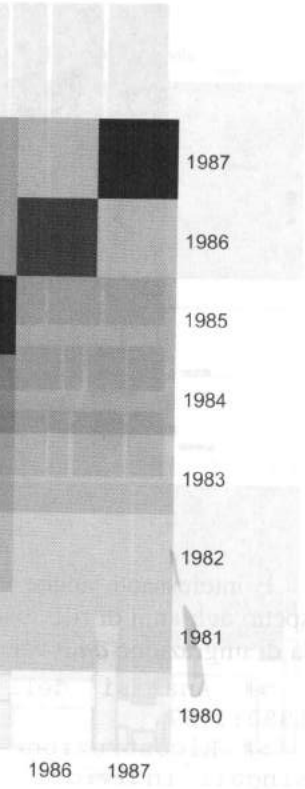
Figura 16



(nr)),1,function(i)

metrica (righe e colonne in-
numero di lavoratori iscrit-
(*i,j,j>i*) indica il numero di
essere iscritti nell'anno *j*. A
onsiderano solamente la mi-
dell'intera matrice fornisce
o del sindacato. La rappre-
enuta mediante la funzione
netrica (symm=TRUE):

ediante un heat map
Rowv=NA,Colv=NA,main



In Figura 16 sono presenti differenti tonalità di grigio: la più scura indica la frequenza maggiore. Il percorso di affiliazione al sindacato inizia sulla diagonale e prosegue a destra (o verso l'alto, data la simmetria della matrice). Si vede facilmente che il tasso di abbandono degli iscritti è più elevato nei primi anni (1980-1982) e nell'ultimo (1986-1987), come mostrano le più evidenti differenze di colore tra la diagonale e l'anno successivo. Tuttavia, il tasso di abbandono in intervalli di tempo più lunghi sembra appiattirsi in quanto le differenze di colore tra elementi lontani dalla diagonale sono meno intense rispetto a quelli ad essa più prossimi.

Una visione congiunta del fenomeno di affiliazione al sindacato ed il suo effetto sul salario è ottenibile incrociando tutte le variabili presenti nella base dei dati. Per fare ciò è necessario introdurre una tecnica statistica più raffinata: l'*Analisi delle COrrispondenze MUltiple* (ACOMU, Greenacre 1993, tra i molti riferimenti recenti che si potrebbero dare). L'ACOMU è implementata nella funzione `mca()` della libreria `MASS` (Venables e Ripley, 1999). L'obiettivo dell'ACOMU è similare a quello dell'ACP poiché si prefigge di trovare un sottospazio di rappresentazione per variabili categoriche invece che continue. Lo spazio originario ha dimensione uguale alla somma delle modalità delle variabili categoriche ed il sottospazio di rappresentazione è definito, anche in questa analisi, da assi fattoriali (o *fattori*). I fattori sono costruiti considerando l'ipertavola di contingenza – *tavola di Burt* – tra le variabili (*attive*). L'interpretazione degli assi fattoriali si effettua in maniera analoga all'ACP, con la differenza teorica dovuta al fatto che non è possibile in questo caso ricorrere alle correlazioni tra assi delle variabili (categoriche) e fattori, ma si considerano le coordinate dei *punti modalità* negli assi fattoriali. La proiezione dei *punti individuo* sugli assi fattoriali è utile per la successiva analisi dei gruppi.

Per poter introdurre anche i salari nell'ACOMU è necessario raggrupparli in classi (`findInterval()`) dopo aver definito gli estremi sulla base dei quantili (33% e 66%). Questa sequenza permette di ottenere le tre classi di salario che abbiamo scelto, indicandole con le etichette: `Basso`, `Medio` e `Alto`:

```
>estremi.classi=round(quantile(lwage,prob=c(0,0.33,0.6
6,1)),2)
>logsalarioclassi=findInterval(lwage,estremi.classi,al
l.inside=TRUE)
>logsalarioclassi=factor(logsalario.classi,labels=c("B
asso","Medio","Alto"))
```

In definitiva, la base di dati (`dat`) che contiene le variabili categoriche da utilizzare nella costruzione della Tavola di Burt può essere definita:

```
>dat=data.frame(scolariz,settore,attivita,etnia,salu-
te,statocivile,sindacato,zona,logsalarioclassi,esperlav)
```

e la seguente funzione calcola le coordinate nei punti individuo e dei punti modalità sui primi due assi fattoriali (nf=5)

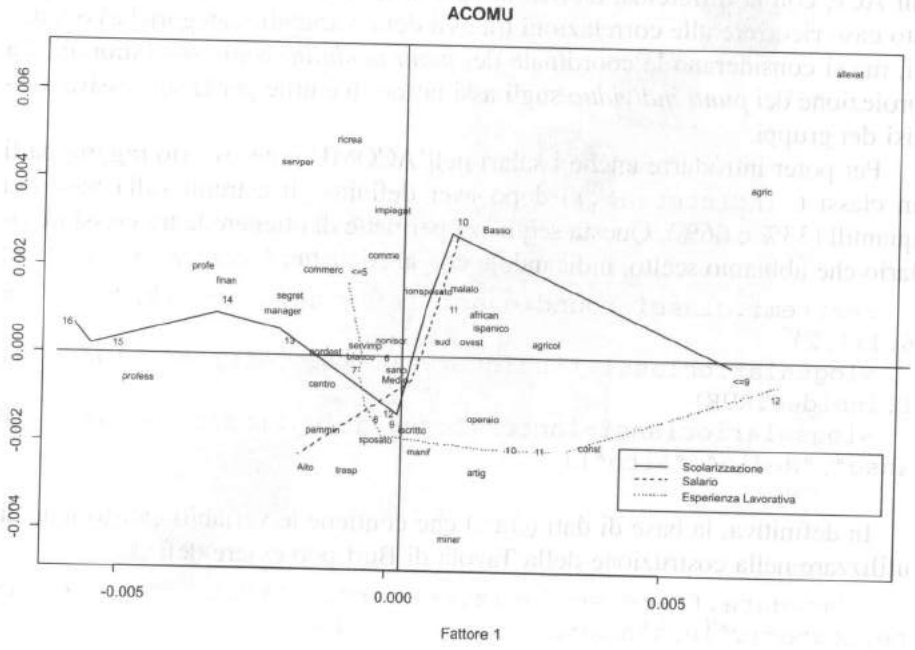
```
>library(MASS)
>acomu=mca(dat, nf = 5,abbrev=TRUE)
```

coordinate dei punti modalità sui primi due fattori sono riportati nella matrice acomu$cs, mentre le etichette delle modalità sono i nomi delle righe della matrice acomu$cs (rownames(acomu$cs)):

```
>plot(acomu$cs,type="n",main="ACOMU",xlab="Fattore
1",ylab="Fattore 2")
>text(acomu$cs,rownames(acomu$cs),cex=0.7)
>#  Congiungiamo con una spezzata i punti-modalità delle va-
riabili ordinabili
>lines(acomu$cs[1:8,1],acomu$cs[1:8,2])
>lines(acomu$cs[43:45,1],acomu$cs[43:45,2],lty=2)
>lines(acomu$cs[46:53,1],acomu$cs[46:53,2],lty=3)
>abline(v=0,h=0)
>legend(0.004,-0.002,c("Scolarizzazione","Salario","Esperienza
Lavorativa"),lty=1:3,cex=0.8)
```

In Figura 17 sono rappresentati i punti modalità sul piano dei primi due fattori.

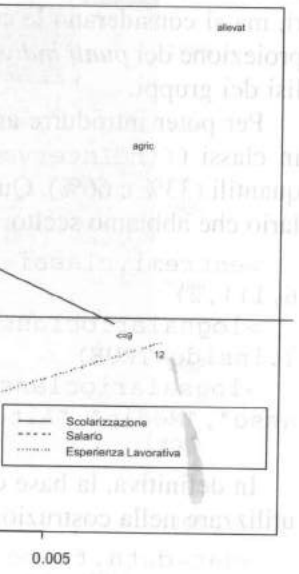
Figura 17



unti individuo e dei punti
sono riportati nella matri-
ono i nomi delle righe del-

```
OMU",xlab="Fattore
.7)
ti-modalità delle va-
2],lty=2)
2],lty=3)
,"Salario","Esperienza
```

ul piano dei primi due fat-

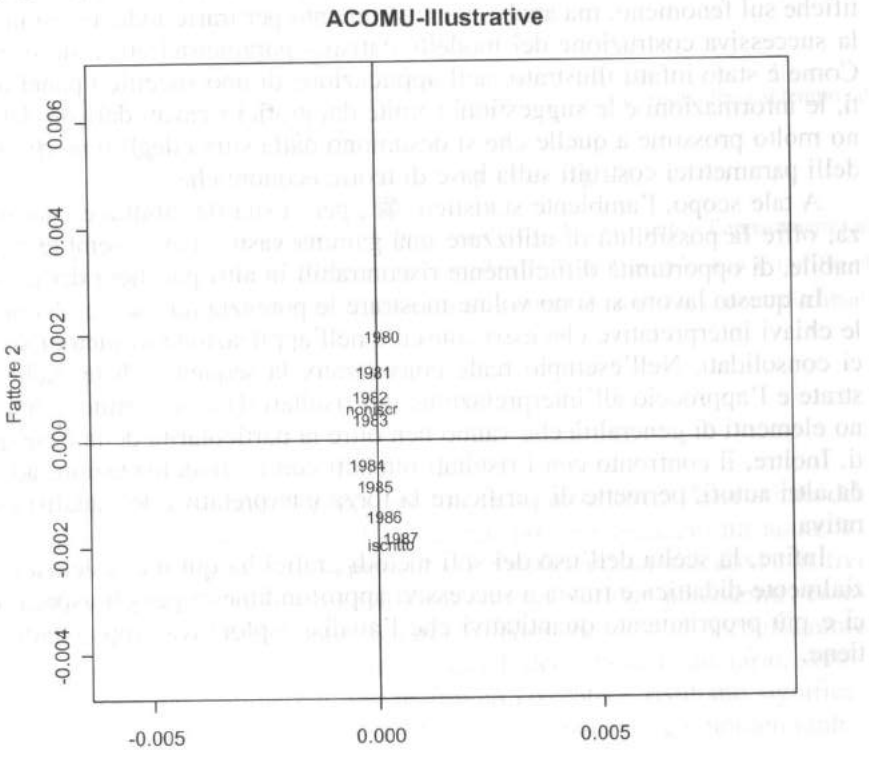


È facile intuire che il primo fattore rappresenta l'associazione tra scolarizzazione, esperienza lavorativa e settore d'occupazione. Infatti a destra si trovano principalmente i settori primari (agricoltura, allevamento e costruzioni), mentre a sinistra quelli dei servizi (finanziario e commercio). La scolarizzazione e l'esperienza lavorativa si dispongono prevalentemente lungo questo asse nei versi opposti: la bassa scolarizzazione coincide con lunghe esperienze di lavoro.

Il secondo fattore prende significato dal salario associato all'iscrizione al sindacato e da alcuni settori economici. Ad esempio, gli impiegati nel settore dei servizi alle persone o nel settore ricreativo hanno bassi salari e tendono a non iscriversi al sindacato. Nella parte inferiore del grafico si situano invece gli iscritti caratterizzati prevalentemente dall'essere sposati, lavorare nei settori dei trasporti, manifattura, minerario e pubblica amministrazione.

Per caratterizzare meglio gli assi fattoriali è anche possibile utilizzare altre variabili (*illustrative*) che non sono state inserite nella costruzione della tavola di Burt, ma possono essere proiettate sugli assi fattoriali già determinati. Studiamo, ad esempio, la caratterizzazione degli assi fattoriali rispetto agli anni.

Figura 18



```
>plot(acomu$cs,type="n",main="ACOMU-Illustrative",xlab="Fattore
1",ylab="Fattore 2")
>text(coord.ill,rownames(coord.ill),cex=0.8)
>text(acomu$cs[37:38,1:2],names(acomu$cs[37:38,1]),cex=0.8)
>abline(v=0,h=0)
```

La Figura 18 mostra che la posizione degli anni nel piano fattoriale è essenzialmente associata all'esperienza lavorativa. La prossimità tra l'anno 1987 ed il numero d'iscritti suggerisce che nel 1987 il numero d'iscritti al sindacato è stato in assoluto il maggiore.

4. Considerazioni conclusive

L'analisi esplorativa grafica dei dati può essere vista non solo come una prima fase dello studio di un fenomeno – in particolare socio-economico – necessaria per la descrizione dei dati e per una prima verifica delle ipotesi scientifiche sul fenomeno, ma anche come strumento per trarre indicazioni utili alla successiva costruzione dei modelli statistici parametrici atti a descriverlo. Come è stato infatti illustrato, nell'applicazione di uno specifico panel di dati, le informazioni e le suggestioni fornite dai grafici ricavati dai soli dati, sono molto prossime a quelle che si desumono dalla stima degli opportuni modelli parametrici costruiti sulla base di teorie economiche.

A tale scopo, l'ambiente statistico R, per la sua flessibilità e completezza, offre la possibilità di utilizzare una gamma vastissima, e sempre aggiornabile, di opportunità difficilmente riscontrabili in altri pacchetti dedicati.

In questo lavoro si sono volute mostrare le potenzialità R, e alcune delle chiavi interpretative che esso consente nell'applicazione di metodi statistici consolidati. Nell'esempio reale considerato, la sequenza di tecniche illustrate e l'approccio all'interpretazione dei risultati da esse fornite, contengono elementi di generalità che vanno ben oltre la particolarità della base di dati. Inoltre, il confronto con i risultati ottenuti con le modellizzazioni adottate da altri autori, permette di verificare la forza interpretativa dell'analisi esplorativa.

Infine, la scelta dell'uso dei soli metodi grafici ha qui una valenza essenzialmente didattica e rinvia a successivi approfondimenti per gli aspetti teorici e più propriamente quantitativi che l'analisi esplorativa ampiamente contiene.

Appendice

I modelli stimati in Vella and Verbeek (1998) sono essenzialmente due: uno per predire l'affiliazione al sindacato e l'altro per quantificare l'effetto sindacato sul salario.

Il modello utilizzato per spiegare l'appartenenza al sindacato è il seguente:

$$\begin{cases} U_{it}^{*} = \gamma_1' Z_{it} + \gamma_2 U_{i,t-1} + \theta_i + \eta_{it}, t = 1,\ldots,T; i = 1,\ldots N, \\ U_{it} = I\left(U^{*}_{it} > 0\right), t = 1,\ldots,T; i = 1,\ldots N, \end{cases}$$

dove U^{*}_{it} è una variabile latente che rappresenta il beneficio che l'individuo i trae dall'appartenenza al sindacato al tempo t. U^{*}_{it} dipende da:

- Z_{it} che rappresenta un trend deterministico e le condizioni socioeconomiche dell'individuo i al tempo t;
- $U_{i,t-1}$ che rappresenta l'appartenenza al sindacato nell'anno precedente;
- θ_i rappresenta l'effetto aleatorio individuale dovuto alla propensione dell'individuo i ad iscriversi al sindacato. θ_i è distribuito con legge normale con media zero e varianza σ_θ^2 stimata con il metodo della massima verosimiglianza;
- η_{it} rappresenta l'effetto casuale individuale al tempo t. Per questo effetto si assume una legge normale multivariata (di dimensione 7 con media zero e matrice di covarianza non diagonale).

La stima dell'effetto sindacato avviene sul logaritmo del salario che per l'individuo i al tempo t è indicato con w_{it}. Il modello stimato è il seguente

$$w_{it} = \beta\, X_{it} + \delta U_{it} + U_{it}(\alpha_{1,i} + \varepsilon_{1,it}) + (1 - U_{it})(\alpha_{0,i} + \varepsilon_{0,it}), t = 1,\ldots,T; i = 1,\ldots N,$$

dove β rappresenta l'effetto delle condizioni socio-economiche, X_{it}. U_{it} indica l'appartenenza al sindacato, mentre $\alpha_{k,i}$ e $\varepsilon_{k,it}$, sono gli effetti aleatori dell'individuo quando non è iscritto al sindacato, k=0 e quando lo è k=1. Anche in questo caso gli effetti aleatori si assumono distribuiti con legge normale.

Risultati

La Tabella II in Vella e Verbeek mostra la significatività delle stime dei coefficienti per il primo modello. Coefficienti positivi indicano un aumento della probabilità di iscriversi al sindacato. I coefficienti significativi positivi sono quelli relativi a: appartenenza al sindacato nel periodo precedente, essere sposati ed africani. Per quanto riguarda i settori economici, i coefficiente negativi significativi sono principalmente quelli del settore finanziario, servizi alle imprese e del commercio. Per le attività lavorative, risultano significativamente negativi i coefficienti per professionisti, manager e commercianti.

La Tabella V mostra invece i risultati del modello per lo studio dei salari rispetto all'appartenenza al sindacato. Gli incrementi salariali per gli iscritti al sindacato sono risultati maggiori per coloro che hanno avuto un periodo di scolarizzazione più lungo e maggiore esperienza lavorativa, per gli africani, gli agricoltori e le persone sposate.

Bibliografia

[1] Bortot, P., Ventura, L., Salvan, A. (2000). *Inferenza Statistica: Applicazioni con S-PLUS e R*, Cedam, Padova.

[2] Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, **35**, 54

[3] Dobson, A.J. (2001). *An Introduction to Generalized Linear Models*, Second Edition. Chapman and Hall/CRC, London.

[4] Everitt, B.S., Dunn, G. (2001). *Applied multivariate data analysis*. Arnold, London.

[5] Everitt, B.S., Landau, S., Leese, M. (2001). *Cluster Analysis*, (4th Ed.). Arnold.

[6] Greenacre, M.J. (1993). *Correspondence analysis in practice*. Academic Press.

[7] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

[8] Iacus, S. M., Masarotto, G. (2003). *Laboratorio di statistica con R*, McGraw-Hill Italia.

[9] McCulloch, C., Searle, S. (2001). *Generalized, linear and mixed models*. Wiley, New York.

[10] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[11] Struyf, A., Hubert, M., Rousseeuw, P.J. (1997). Integrating Robust Clustering Techniques in S-PLUS. *Computational Statistics and Data Analysis*, **26**, 17–37.

[12] Stephens, M.A. (1986). Tests Based on EDF Statistics. Chap. 4 in *Goodness-of-fit Techniques*, R.B. D'Agostino and M.A. Stephens Eds., Marcel Dekker, New York.

[13] Vella, F., Verbeek, M. (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Journal of Applied Econometrics*, **13**, 163-183.

[14] Venables, W.N., Ripley, B.D. (1999). *Modern Applied Statistics with S-plus* (3rd Ed.). Springer.

[15] Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.