SDE 2017

Marco Riani
mriani@unipr.it
http://www.riani.it

RAPPRESENTAZIONI GRAFICHE

Nella comunicazione una figura vale più di cento numeri e di mille parole!

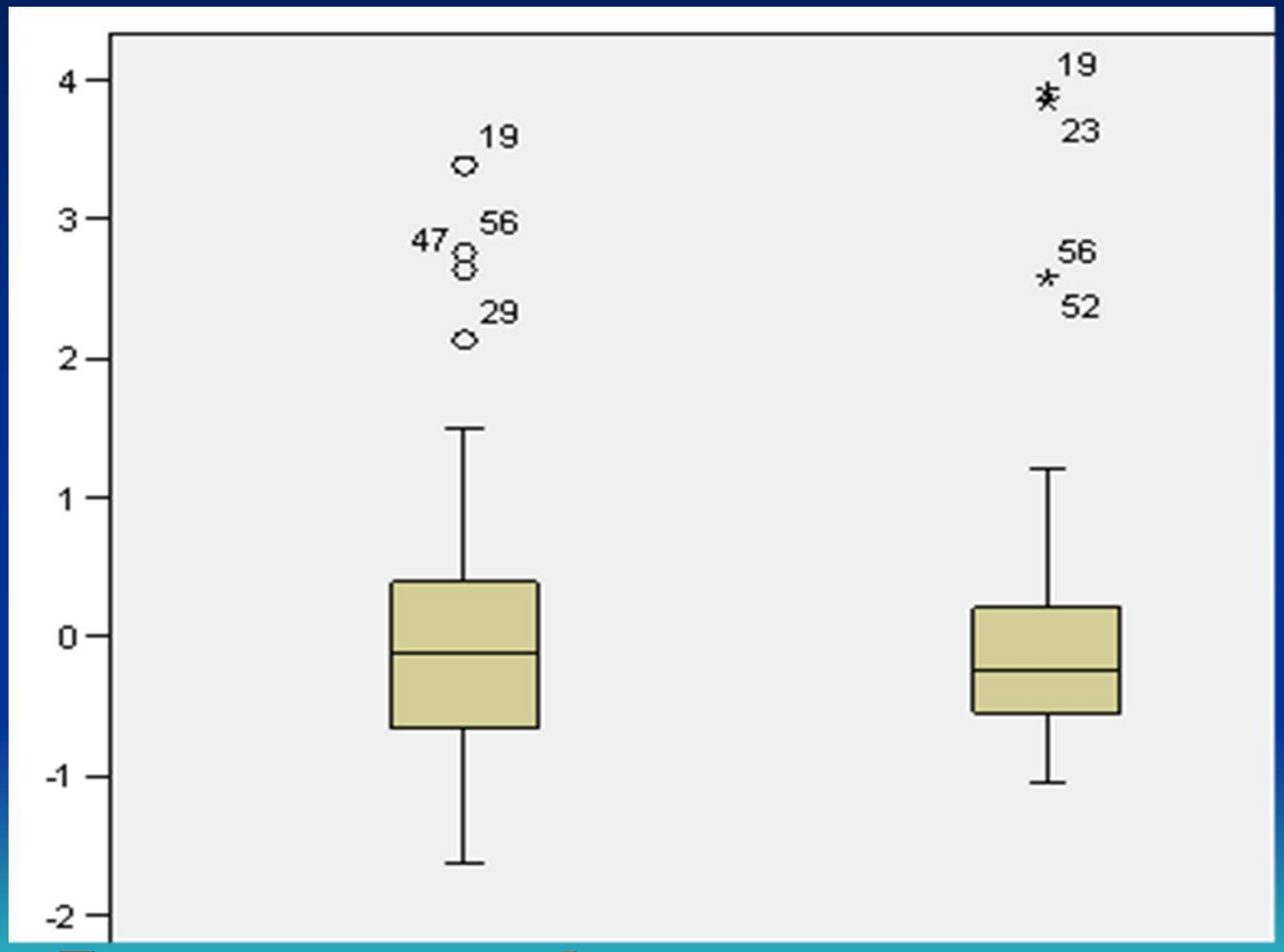
SCOPI DELLE RAPPRESENTAZIONI GRAFICHE

NELLE ANALISI PRELIMINARI:

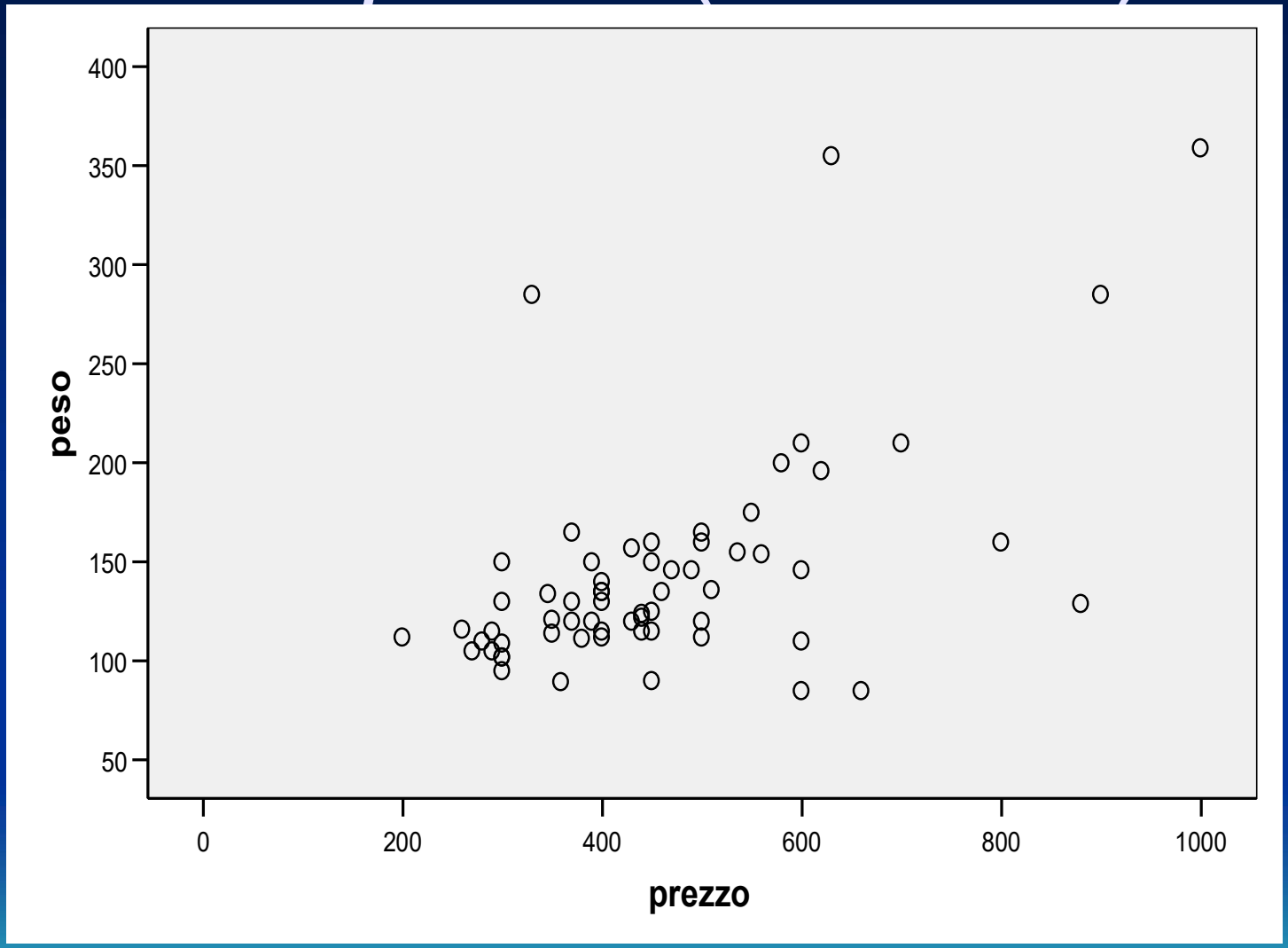
- individuazione di valori anomali
- ricerca di strutture (*pattern*) nei dati
  - relazioni, anche non lineari, tra variabili
  - somiglianze tra unità e gruppi omogenei

- NELLA PRESENTAZIONE DEI RISULTATI:
  comunicazione ai “non addetti ai lavori”

Smartphone



Smartphone (continua).

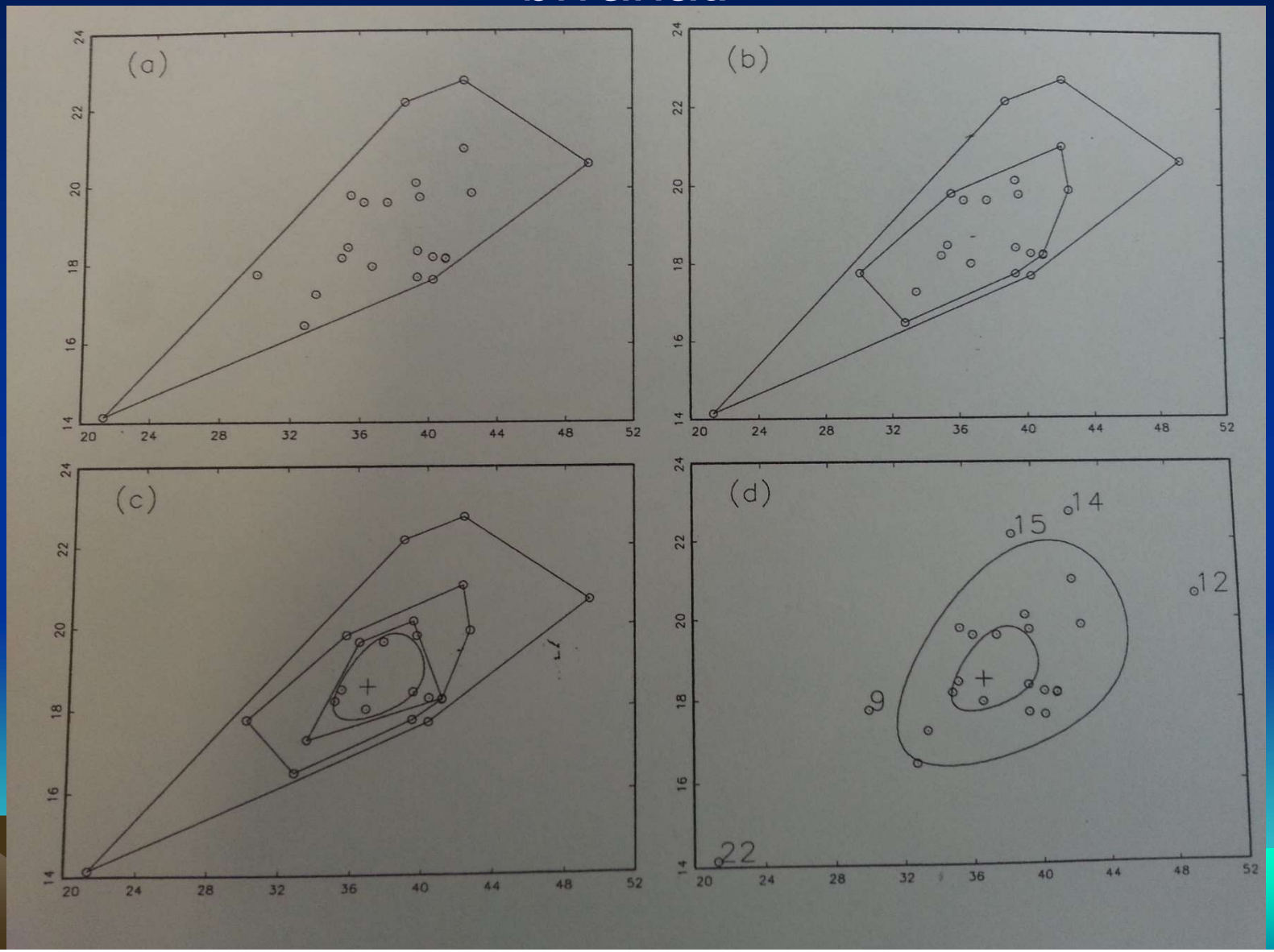


Necessità di avere un contorno robusto bivariato

Il *convex hull* nel diagramma di dispersione

- **In un diagramma di dispersione unisco tra loro i punti più esterni: ottengo un "guscio convesso"**
- **Ripeto l'operazione con riferimento ai punti interni rimanenti, etc.**
- **Ottengo una successione di poligoni convessi, che forniscono una stima delle "curve di livello"**
- **Informazioni sugli outliers e sulla forma di distribuzione bivariata**
- **I gusci più interni non risentono degli outliers**
- **Il "contorno interno" (*core*) che contiene una percentuale non maggiore del 50% ha un significato analogo alla scatola del boxplot**

Scatter della performance (asse x) e volatilità (asse y) di 23 fondi di investimento con convex hull e boxplot bivariati



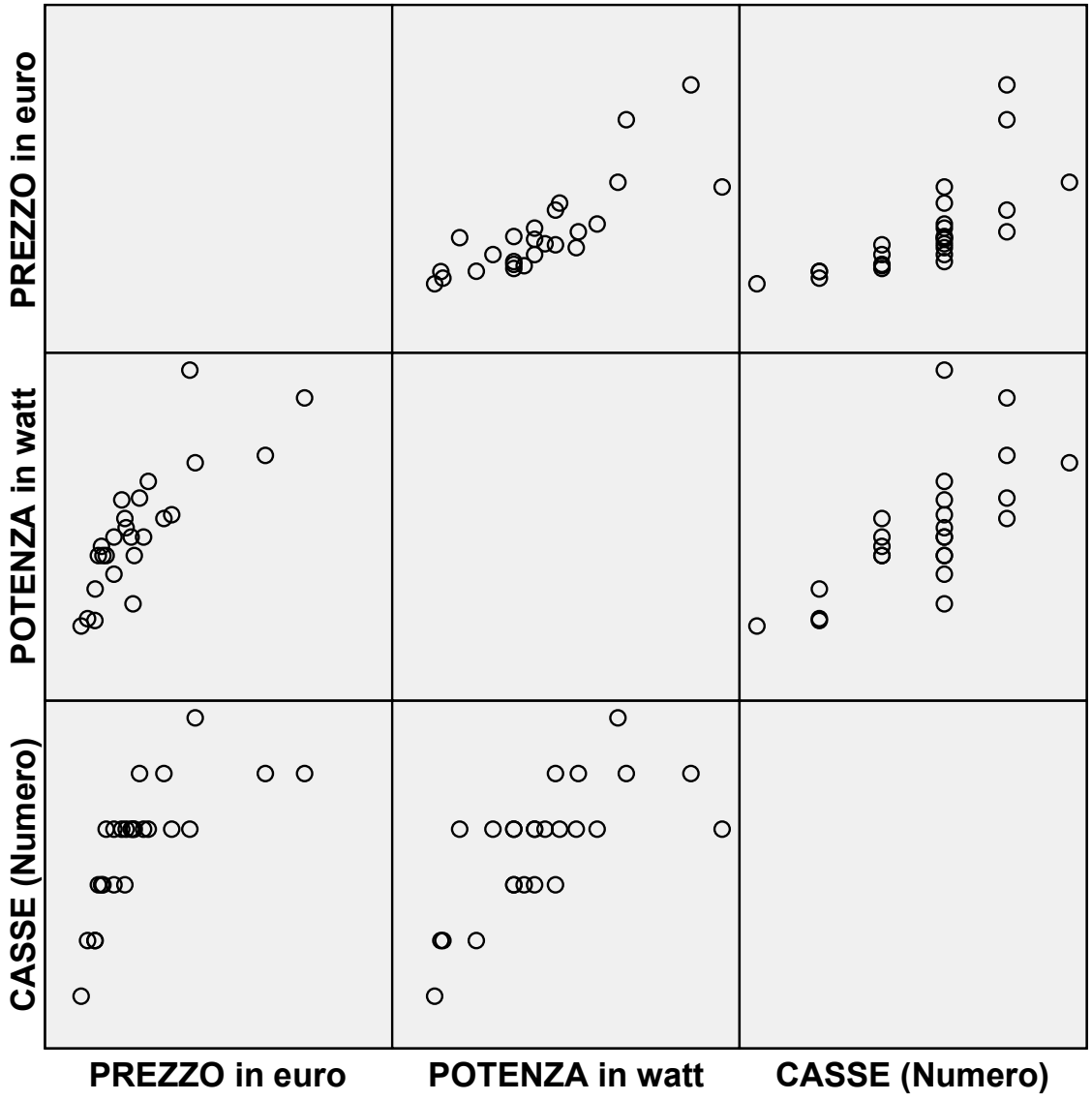
RAPPRESENTAZIONE DI 3 VARIABILI

- Diagramma di dispersione a tre dimensioni (in R^3) :

  - di difficile lettura, sovrapposizione di punti, mancanza di prospettiva. Esempio: Benessere regionale: reddito medio, mediano, % famiglie povere

- MATRICE DEI DIAGRAMMI DI DISPERSIONE (*scatterplot matrix*): informazioni: outliers univariati e bivariati, relazioni tra coppie di variabili, gruppi omogenei di unità, estensione a più di 3 variabili

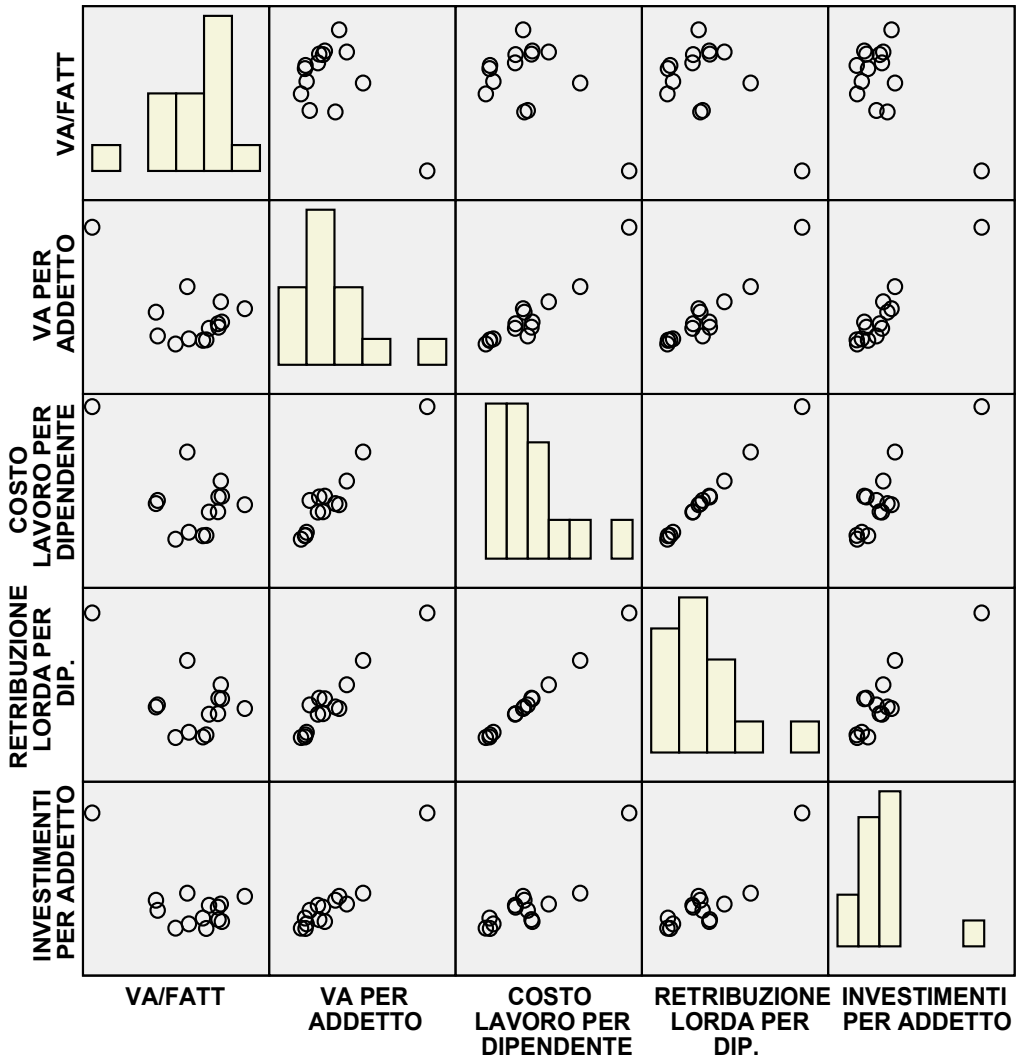
Esempio di scatter plot matrix



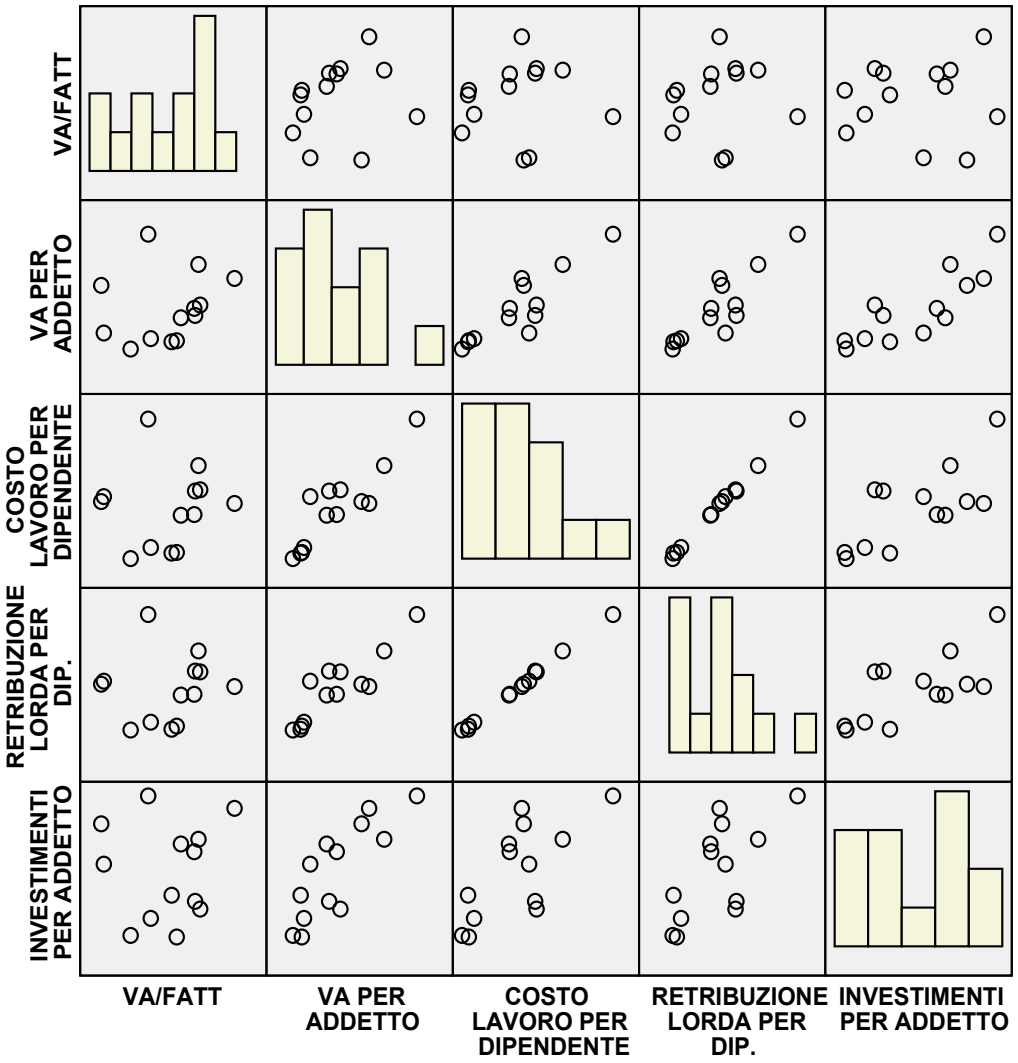
RAPPRESENTAZIONE DI PIU' DI 3 VARIABILI

- **MATRICE DEI DIAGRAMMI DI DISPERSIONE**
  limitazione: se il numero di variabili è troppo grande i singoli riquadri diventano illeggibili
- **ESEMPIO: 14 settori industriali**

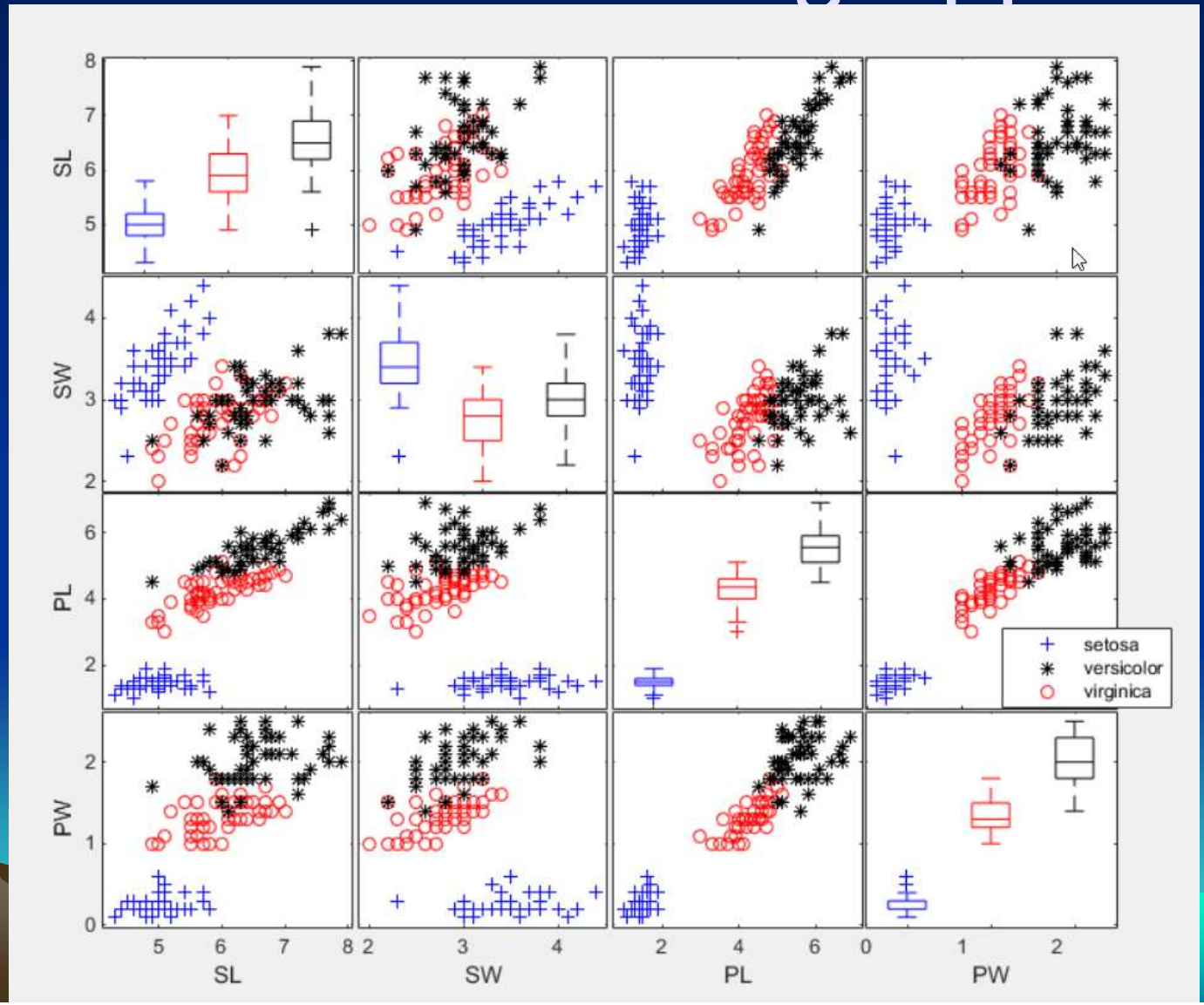
SCATTERPLOT MATRIX (14 settori, con raffinerie)



13 settori (senza raffinerie)



Scatter plot matrix con indicazione dei gruppi



Esercizio

Si disegni una matrice dei diagrammi di dispersione riferita a 10 unità ed a 3 variabili X, Y, Z che presenti tutte le seguenti caratteristiche:

- **Correlazione tra X ed Y diretta e molto elevata.**
- **Correlazione inversa tra X e Z di discreta entità.**
- **Un valore anomalo, eccezionalmente grande, per Z.**

COORDINATE PARALLELE

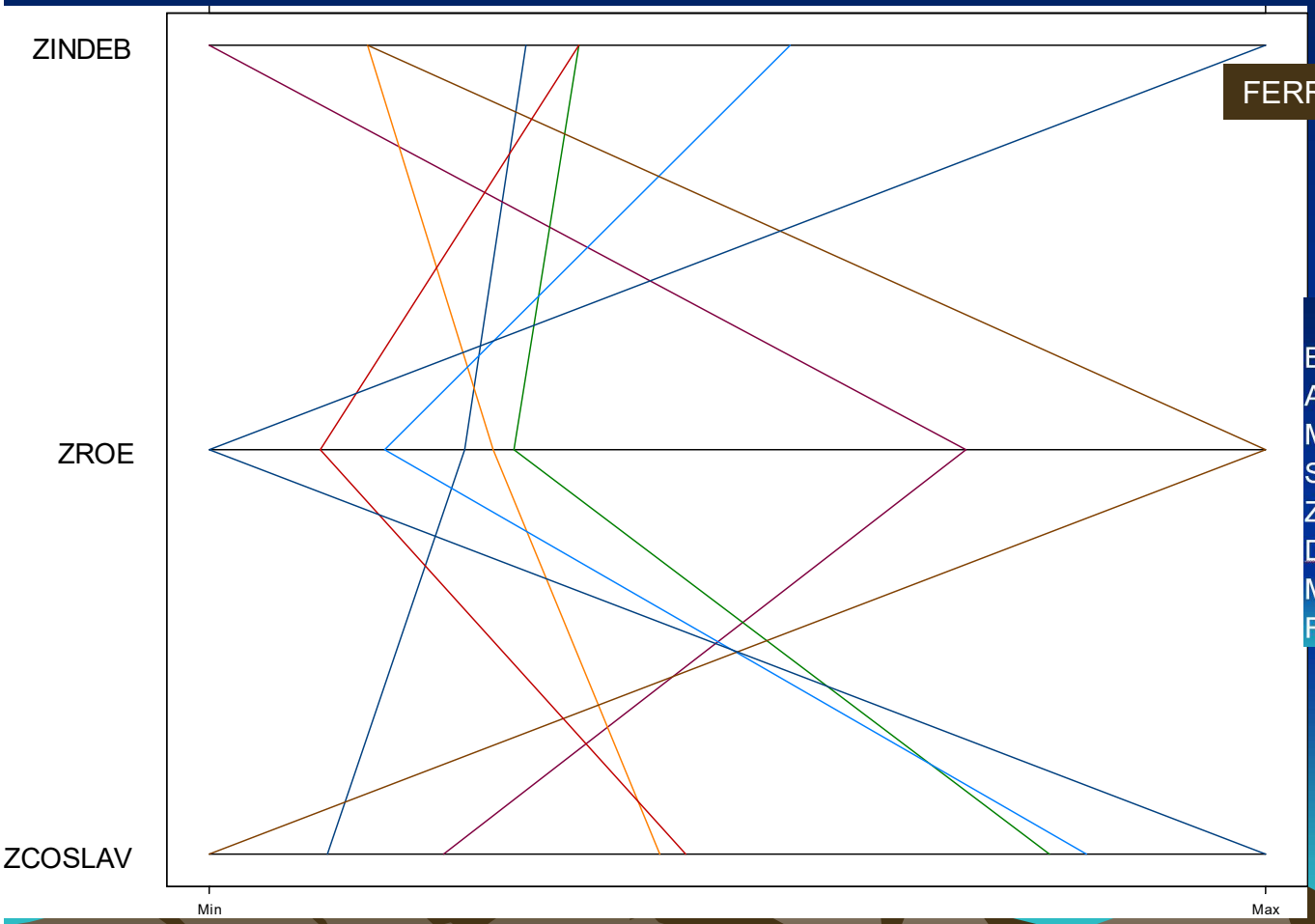
***n* unità statistiche e *p* variabili**

- Si tracciano p parallele equidistanti
- Su ogni parallela si indica la scala di misurazione
- Abitualmente si considerano le variabili in termini di scostamenti standardizzati
- Per ogni unità si individuano i rispettivi valori (standardizzati) su ciascuna parallela e si uniscono tra loro, ottenendo una **spezzata** rappresentativa di tale unità

FASHION:	costlav	roe	indeb
• BENETTON	31,30	9,90	,70
• ARMANI	37,60	29,30	,10
• MARZOTTO	70,40	11,80	,80
• STEFANEL	49,30	11,00	,40
• ZEGNA	72,40	6,80	1,20
• DOLCEeG.	24,90	40,90	,40
• MISSONI	50,70	4,30	,80
• FERRE'	82,10	,00	2,10

	Zcostlav	Zroe	Zindeb
• BENETTON	-1,01	-,32	,18
• ARMANI	-,71	1,09	-1,15
• MARZOTTO	,87	-,18	-,02
• STEFANEL	-,15	-,24	-,67
• ZEGNA	,96	-,54	,63
• DOLCEeG.	-1,32	1,93	-,67
• MISSONI	-,08	-,72	-,02
• FERRE'	1,43	-1,03	2,09

FASHION



FERRE’

	Zcostlav	Zroe	Zindeb
BENETTON	-1,01	-,32	,18
ARMANI	-,71	1,09	-1,15
MARZOTTO	,87	-,18	-,02
STEFANEL	-,15	-,24	-,67
ZEGNA	,96	-,54	,63
DOLCEeG.	-1,32	1,93	-,67
MISSONI	-,08	-,72	-,02
FERRE'	1,43	-1,03	2,09

D. e G.

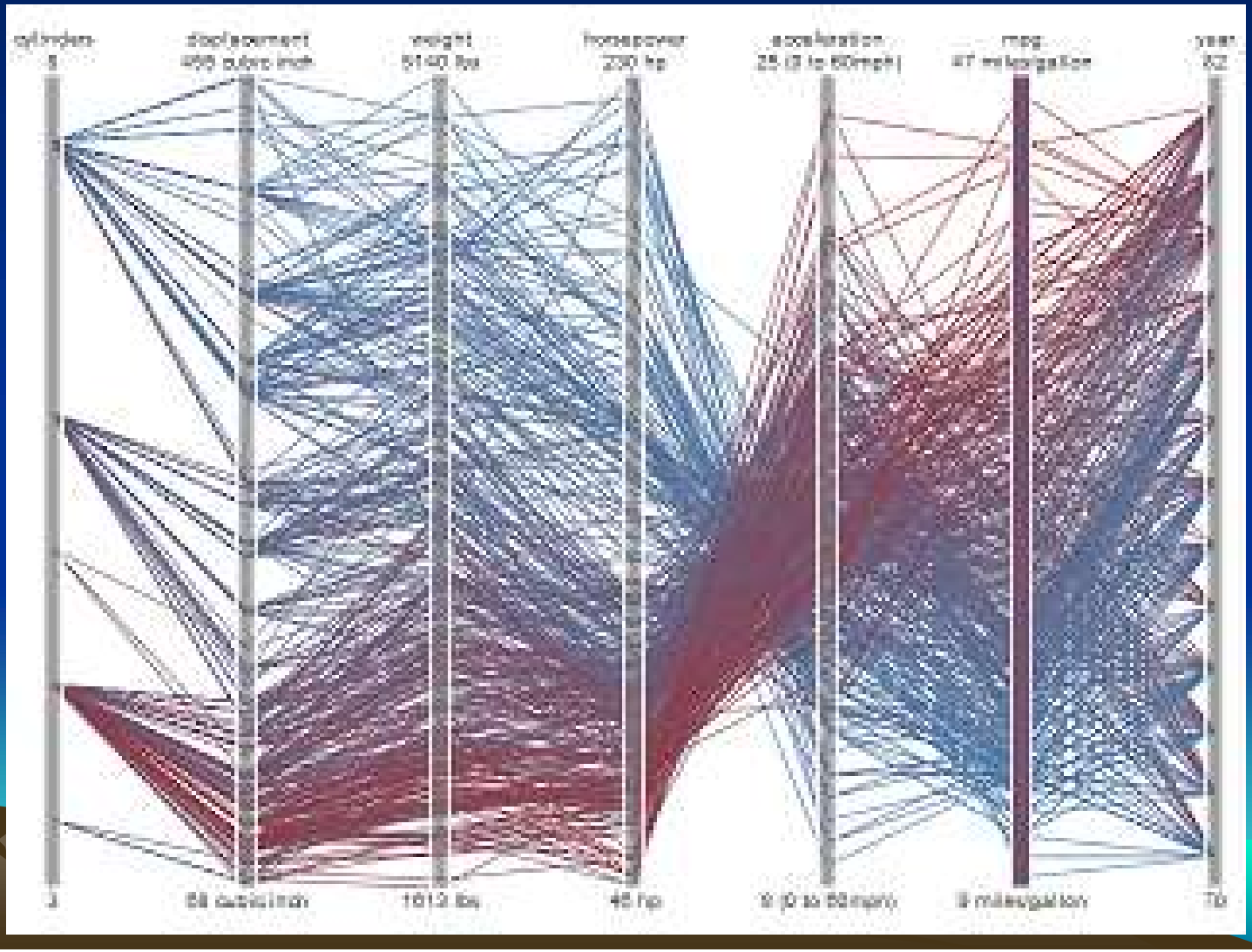
INFORMAZIONI FORNITE DALLE COORDINATE PARALLELE

- **Ogni spezzata pone in evidenza i valori della corrispondente unità per tutte le variabili**
- **Segnalano la relazione tra coppie di variabili consecutive nel grafico**
- **Mostrano gli eventuali gruppi omogenei di unità (fasci di spezzate con andamento simile)**
- **Segnalano gli outliers multivariati (spezzate con andamento molto diverso dalle altre)**

DIFETTO:

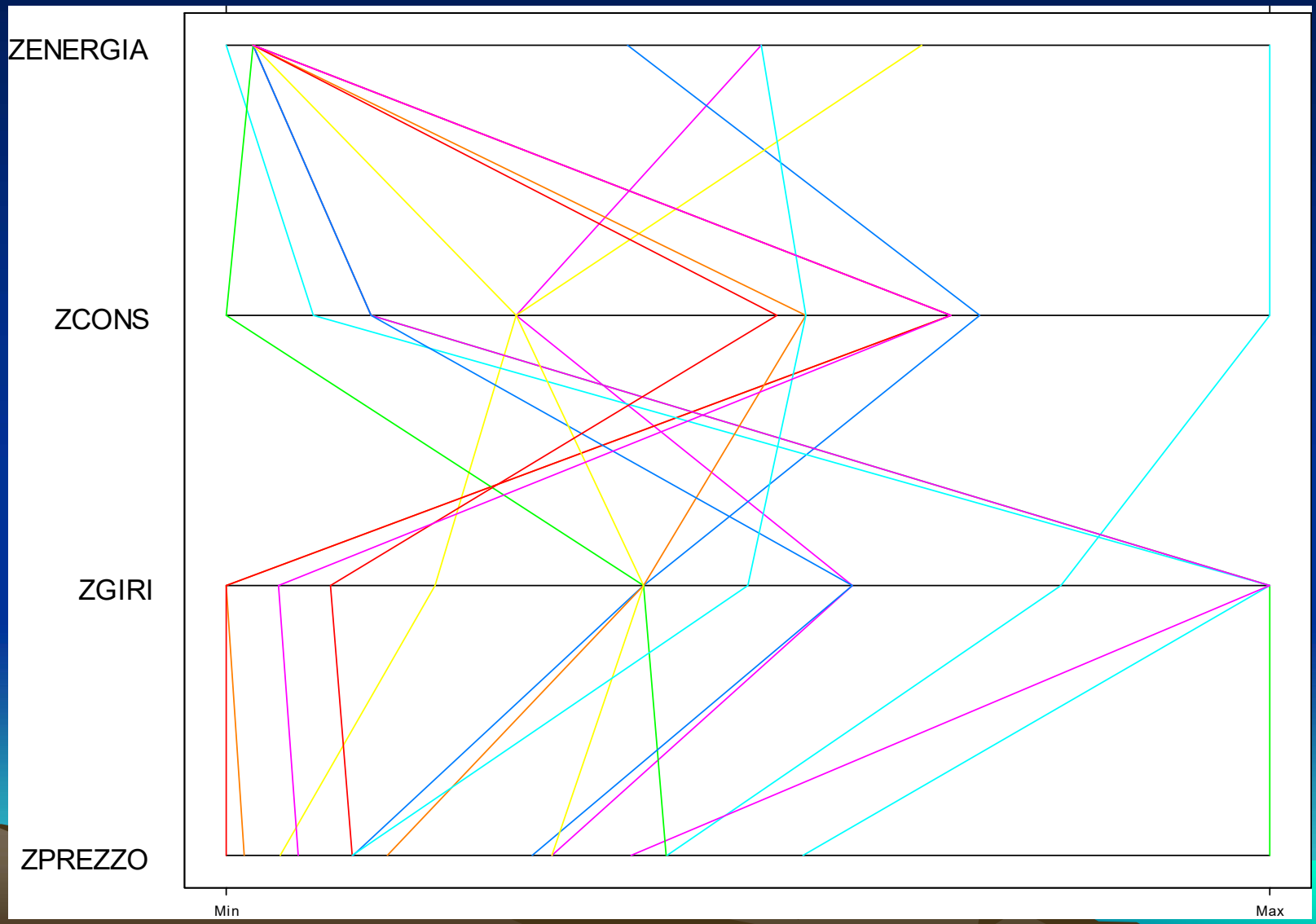
"Lettura" meno immediata rispetto alla *scatterplot matrix*.

Esempio di rappresentazione in coordinate parallele



LAVATRICI

	prezzo	giri	cons	energia
prezzo	1	0.822022	-0.52604	-0.09361
giri	0.822022	1	-0.45639	0.123789
cons	-0.52604	-0.45639	1	0.42507
energia	-0.09361	0.123789	0.42507	1



Esercizio

- Supponiamo di essere in presenza di 10 unità e 5 variabili di cui le prime 3 correlate in maniera forte e diretta e le ultime due correlate in maniera forte ma indiretta con le prime tre. Disegnare il diagramma in coordinate parallele che ci si attende

RAPPRESENTAZIONI ICONICHE

n unità statistiche e *p* variabili

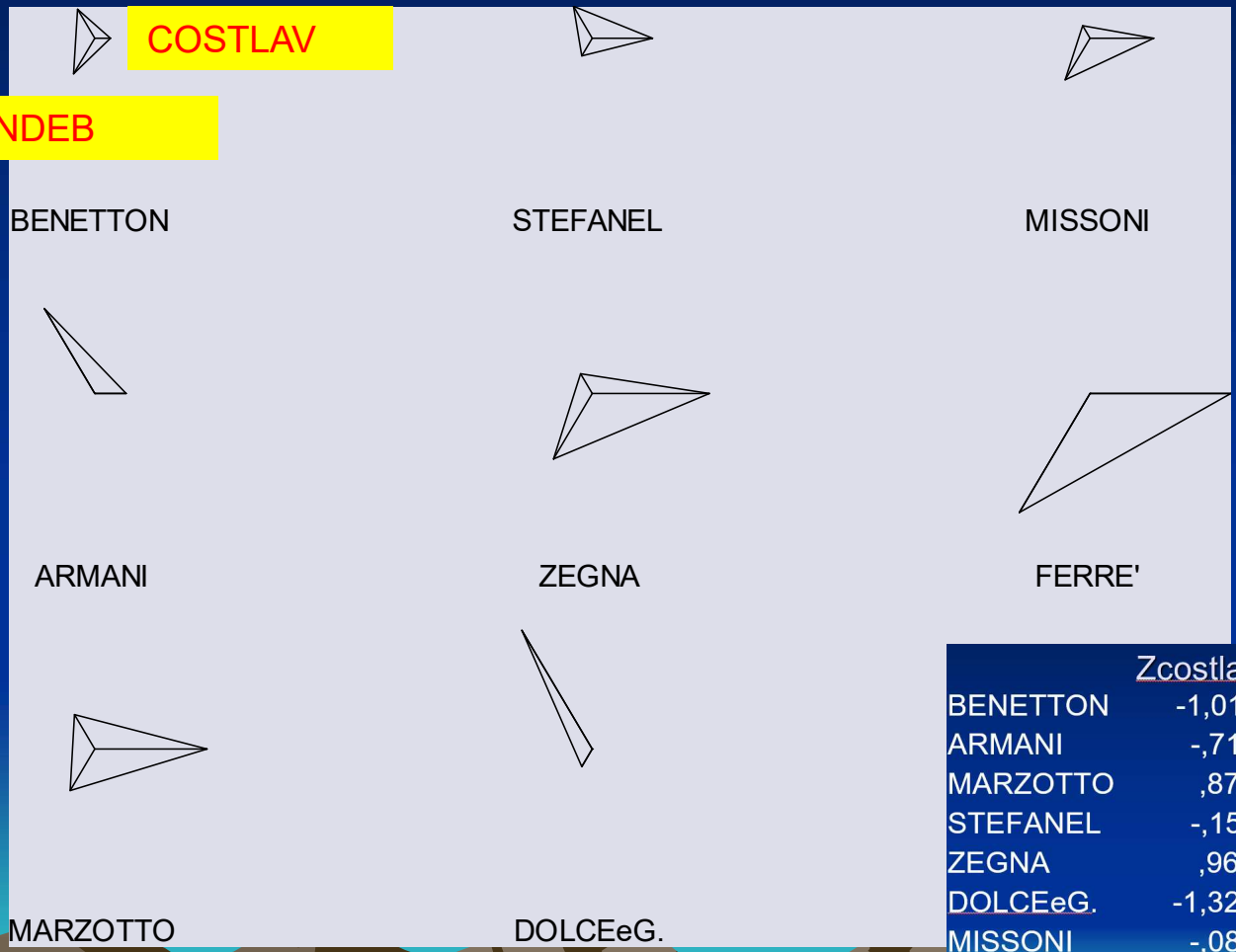
- Si disegna per ogni unità statistica una figura con p caratteristiche
- Le figure più utili sono le cosiddette “**stelle**”:
- Si divide l’angolo giro in p parti uguali
- Si associa ad ogni semiretta uscente dal centro una variabile
- Si scegli un’opportuna unità di misura per ogni semiretta
- Per l’unità statistica considerata s’individuano i rispettivi valori delle p variabili sulle p semirette e uniscono tra loro
- Si ottiene un poligono irregolare chiamato “stella”
- Analogamente si traccia una stella per ogni unità statistica

STELLE FASHION

ROE

COSTLAV

INDEB

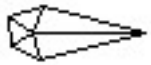


	Zcostlav	Zroe	Zindeb
BENETTON	-1,01	-,32	,18
ARMANI	-,71	1,09	-1,15
MARZOTTO	,87	-,18	-,02
STEFANEL	-,15	-,24	-,67
ZEGNA	,96	-,54	,63
DOLCEeG.	-1,32	1,93	-,67
MISSONI	-,08	-,72	-,02
FERRE'	1,43	-1,03	2,09

14 settori e 5 variabili

Alimentari e bevande

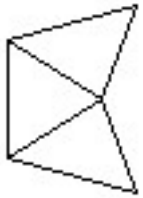
Carta e stampa



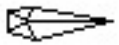
Minerali non metallici

Mezzi di trasporto

Tessili e abbigliamento



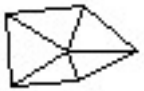
raffinerie



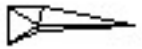
Prodotti in metallo

Altre industrie

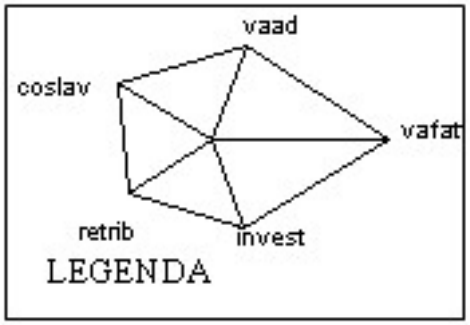
Cuoio



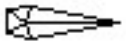
Chimica



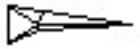
Meccanica



Legno



Gomma e plastica



Apparecchi elettrici

INFORMAZIONI FORNITE DALLE STELLE

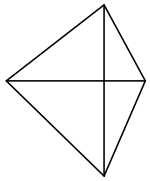
- **Evidenziano l'ordine di grandezza di ogni unità**
- **Mostrano le somiglianze tra le unità (stelle di forma analoga)**
- **Segnalano gli outliers multivariati (stelle di forma molto diversa dalle altre)**
- **In una stella, una punta molto acuta segnala un valore eccezionalmente grande di quella variabile**

DIFETTI

Non evidenziano le relazioni tra le variabili

16 LAVATRICI, 4 VARIABILI

(prezzo, giri, consumo acqua, energia)



LGELECTRONICS

PHILCO

ELECTROLUX

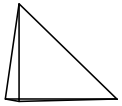
BOSCH

WHIRLPOOL

SAMSUNG

SIEMENS

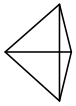
CANDY



MIELE

IGNIS

ARISTON



SANGIORGIO

INDESIT

AEG

REX

ZOPPAS

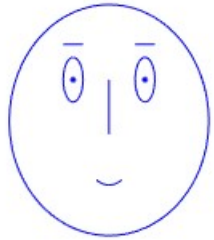
Le facce di Chernoff

Alimentari e bevande

Tessili e abbigliamento

Cuoio

Legno

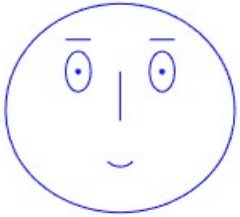


Carta e stampa

raffinerie

Chimica

Gomma e plastica



Minerali non metallici

Prodotti in metallo

Meccanica

Apparecchi elettrici

Mezzi di trasporto

Altre industrie

TRASFORMAZIONI DEI DATI

- Trasformazioni lineari:
- Scostamenti dalla media: $x_i - M$
- Scostamenti standardizzati:

$$z_i = (x_i - M) / \sigma = x_i/\sigma - M/\sigma$$

- Visualizzazione nel diagramma di dispersione: *origine degli assi nel punto di coordinate* (M_X, M_Y) = **centroide**
- Esempio: tablets (*peso, prezzo*)

TRASFORMAZIONI NON LINEARI

- Trasformazione logaritmica, quadratica, cubica, ecc. (casi particolari della trasformazione di Box and Cox, p. 205)

MOTIVAZIONI:

- Ricondurre la distribuzione alla simmetria

Esempio: distribuzione con forte asimmetria positiva (coda destra più lunga) - Trasformazione logaritmica:

$x = 10$; $\text{Log } x = 1$

$x = 100$; $\text{Log } x = 2$

- Ricondurre alla linearità la relazione tra due (o più) variabili

Esempio: relazione tra PIL pro capite e speranza di vita in 109 Paesi (pp. 207-212)