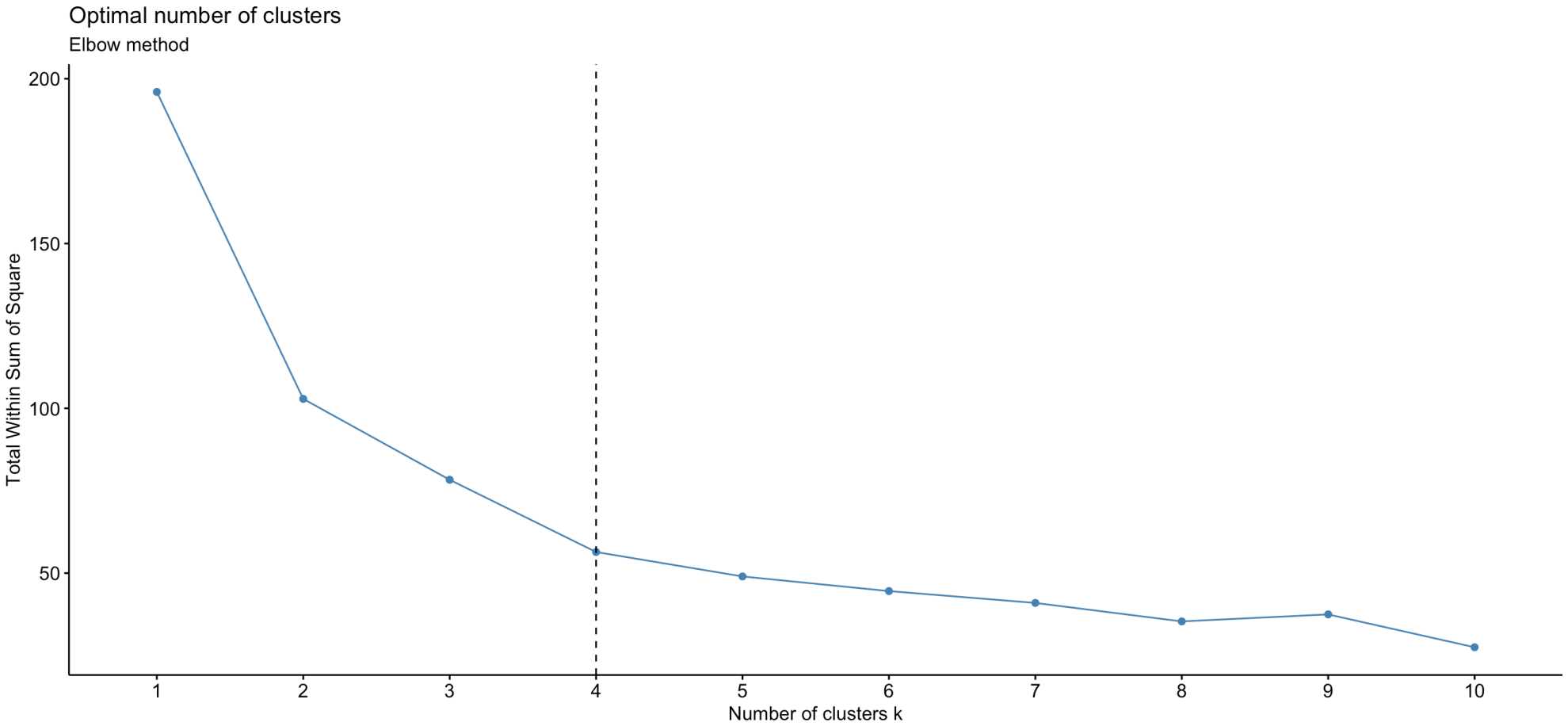
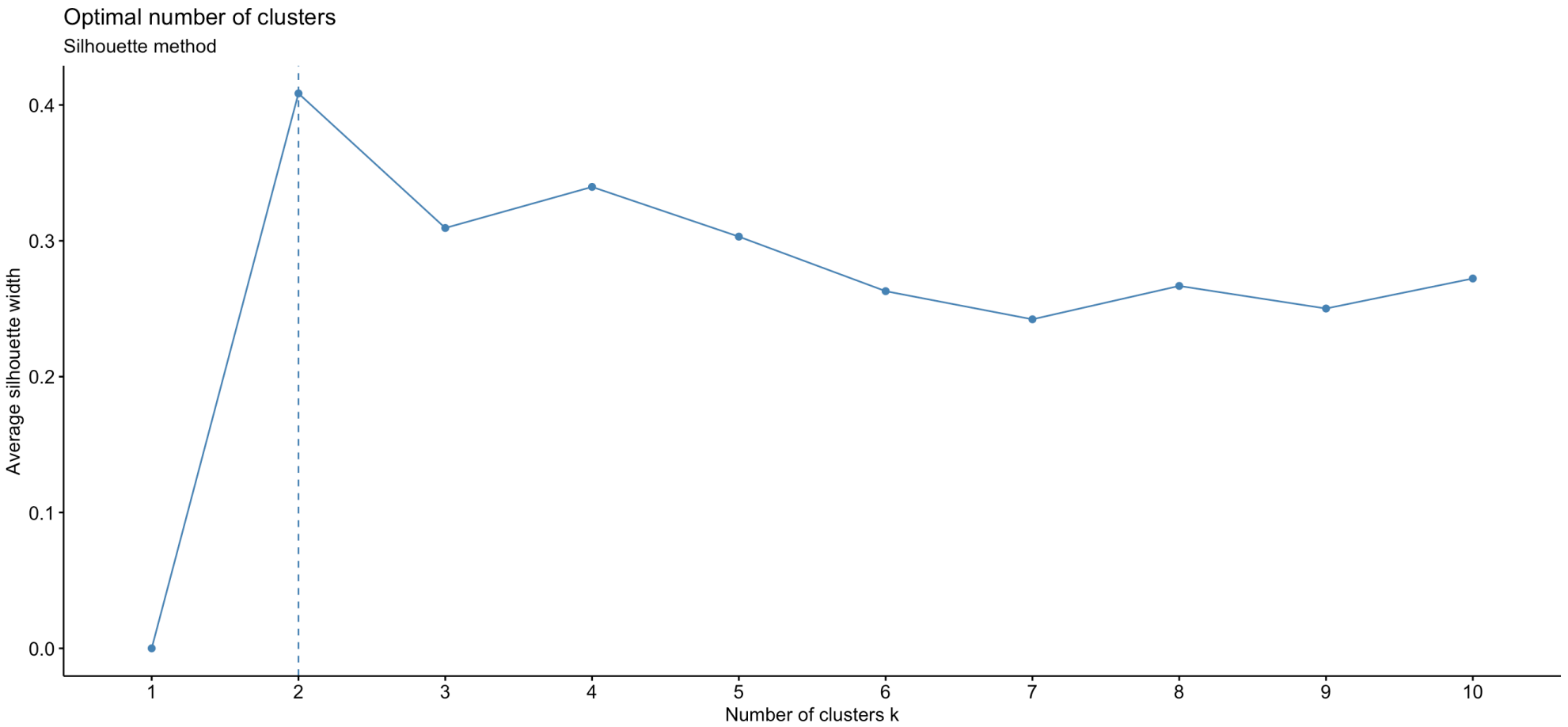
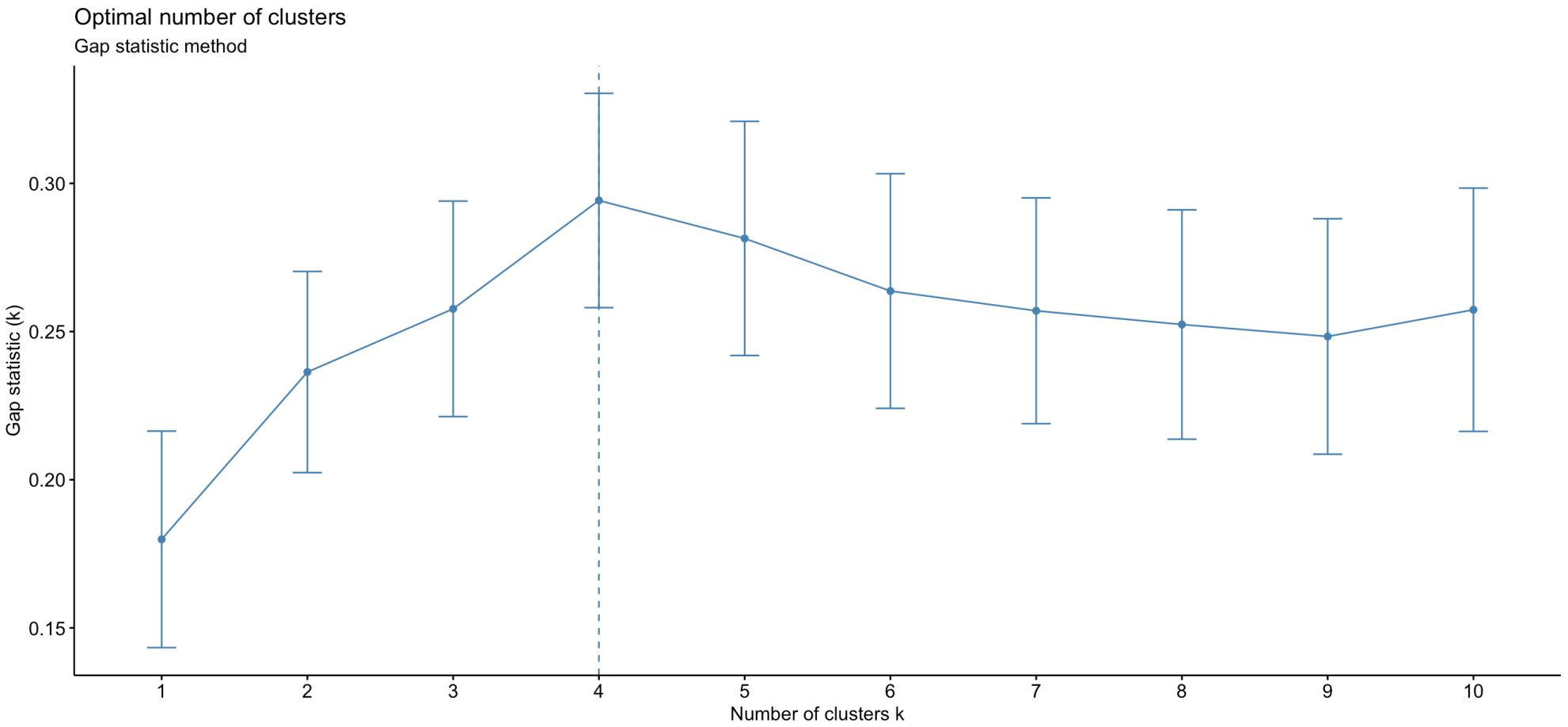
Criminalità stati USA



Criminalità stati USA



Criminalità stati USA



Criminalità stati USA

```
nbus <- NbClust(data = UScale, distance = "euclidean", min.nc = 2, max.nc
= 15, method = "kmeans")

*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 7 proposed 4 as the best number of clusters
* 2 proposed 10 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 3 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

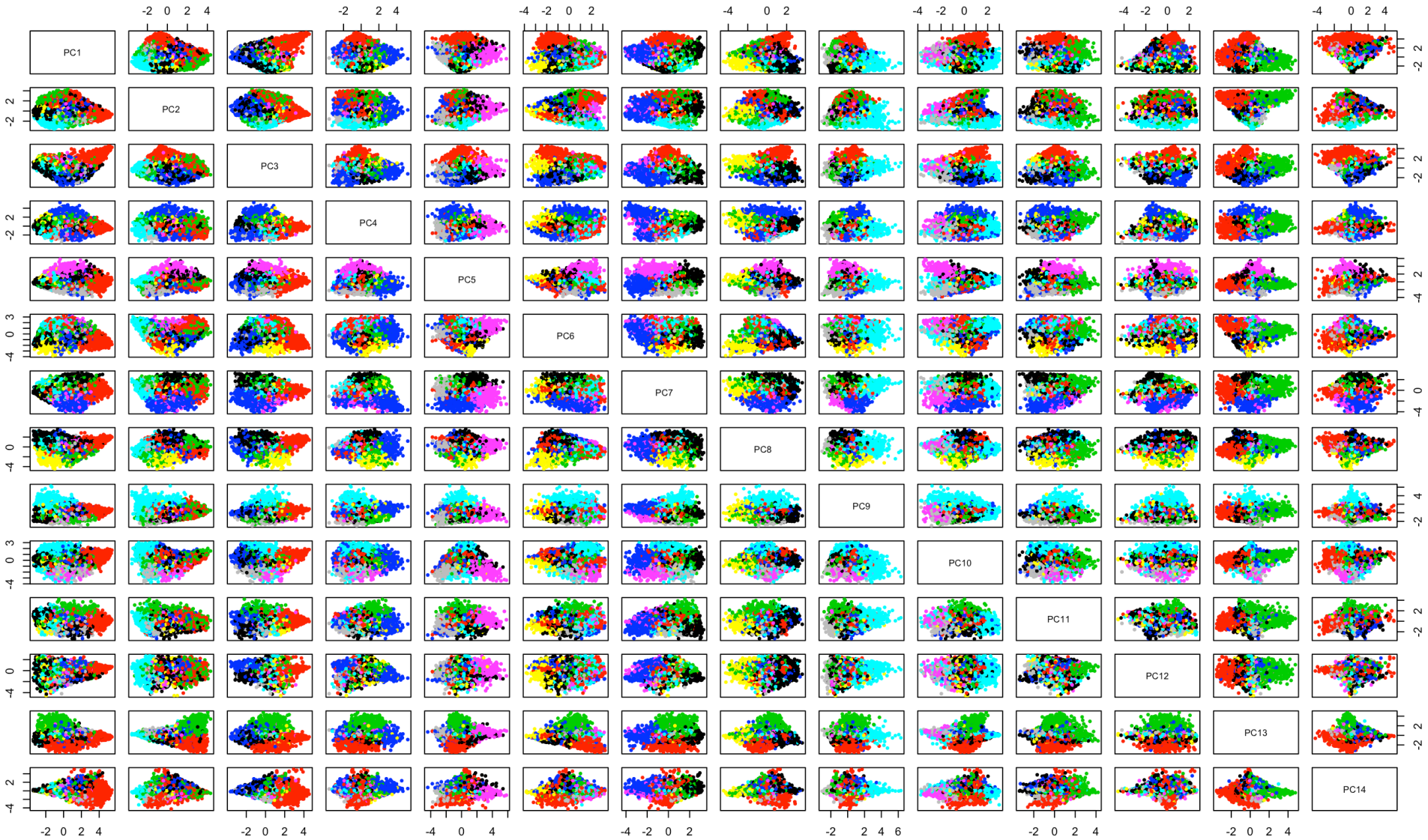
Misurare la qualità del clustering ottenuto

- Due metodi: estrinseco vs intrinseco
- Estrinseco: supervisionato, cioè i gruppi sono noti
  - Confrontare un clustering con la verità del terreno usando una certa misura di qualità del clustering
  - Ad esempio uso di metriche di precisione
- Intrinseca: senza supervisione, cioè i gruppi non sono noti
  - Valutare la bontà di un clustering considerando la bontà dei gruppi e la loro compattezza
  - Ad esempio coefficiente Silhouette

Misurare la qualità del clustering ottenuto

- Misura della qualità del clustering: $Q(C, C_g)$, per un clustering C dati i gruppi noti Cg.
- Q è buono se soddisfa i seguenti 4 criteri essenziali:
  - Omogeneità del gruppi: più puro, meglio è
  - Completezza del gruppi: dovrebbe assegnare unità appartenenti alla stessa categoria nei gruppi noti allo stesso gruppo
  - Sacco di stracci (*rag bag*): mettere una unità eterogeneo in un gruppo puro dovrebbe essere penalizzato più che metterlo in un “sacco di stracci” (ad es. Categoria "miscellanea" o "altro")
  - Conservazione di piccoli gruppi: dividere in pezzi una piccola categoria è più dannoso che dividere in pezzi una grande categoria

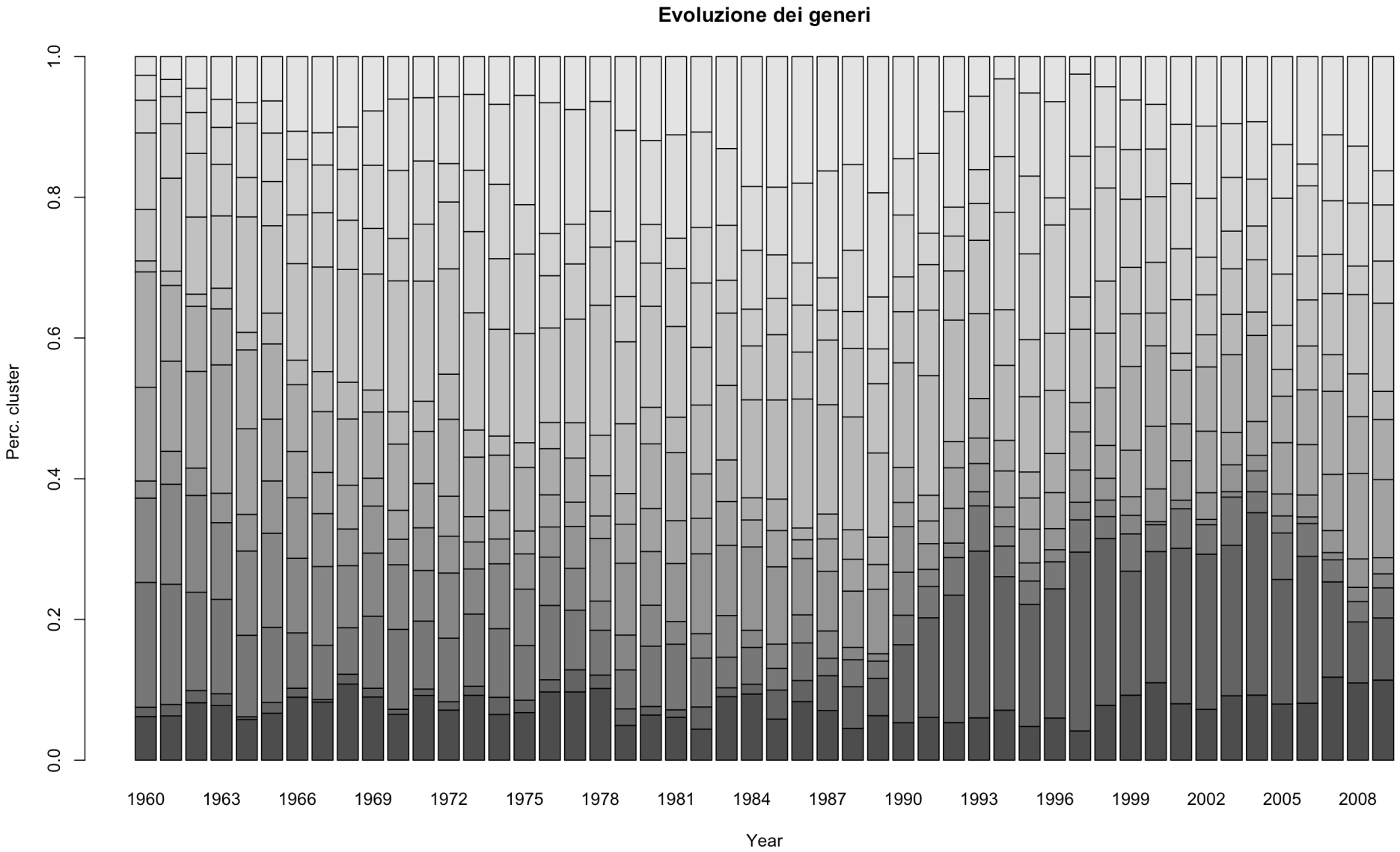
Esempio: definire il genere di musica



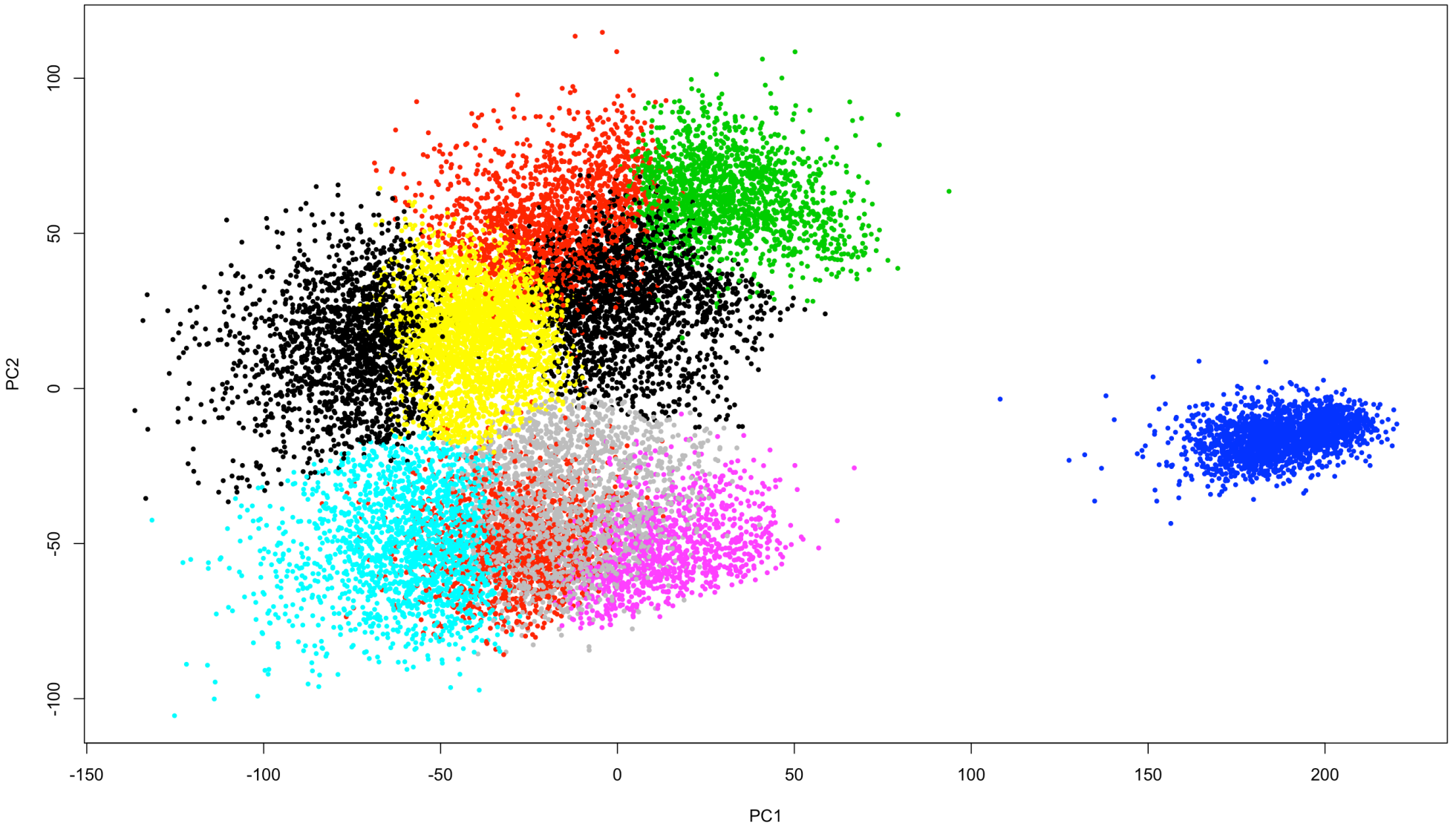
Esempio: interpr. il genere di musica

1	Chicago	The Miracles	Bobby Vee	Cher	Diana Ross	Elton John	Four Tops	Gladys Knight & The Pips	James Brown	Mary Wells
Freq	9	7	6	6	6	6	6	6	6	6
2	LL Cool J	James Brown	Jay-Z	Dickie Goodman	Eminem	Nas	Busta Rhymes	Kris Kross	50 Cent	OutKast
Freq	13	11	11	10	10	9	7	7	6	6
3	Elvis Presley	Neil Diamond	Barry Manilow	Elton John	Jim Reeves	Brook Benton	Eddy Arnold	Jerry Butler	Johnny Mathis	Bobby Darin
Freq	32	13	10	10	10	9	9	9	9	8
4	James Brown	Ray Charles	B.B. King	Elvis Presley	Bobby Darin	Marvin Gaye	Stevie Wonder	Aretha Franklin	Bobby Bland	Ramsey Lewis
Freq	30	13	12	12	10	10	10	9	9	9
5	Eric Clapton	The Rolling Stones	Neil Diamond	Jay & The Americans	John Cougar Mellencamp	R.E.M.	Styx	The Cars	Brad Paisley	Cher
Freq	7	7	6	5	5	5	5	5	4	4
6	Connie Francis	Brenda Lee	The Supremes	Lesley Gore	Martha & The Vandellas	Aretha Franklin	Etta James	Barbra Streisand	Dionne Warwick	The Shirelles
Freq	29	19	16	14	14	13	12	11	11	11
7	Toby Keith	Johnny Cash	Kenny Chesney	John Denver	Alan Jackson	George Strait	Johnny Tillotson	Bobby Vinton	Brooks & Dunn	Duane Eddy
Freq	14	13	13	12	11	11	11	9	9	9
8	Madonna	David Bowie	Janet Jackson	Otis Redding	Pet Shop Boys	Queen	The Cars	Duran Duran	Melissa Etheridge	Rod Stewart
Freq	10	7	6	6	6	6	6	5	5	5
9	Elton John	Neil Diamond	Rascal Flatts	The Rolling Stones	Fats Domino	Rod Stewart	Sam Cooke	Spinners	The Beatles	Three Dog Night
Freq	16	16	13	12	10	9	9	9	9	9
10	Elvis Presley	Barbra Streisand	Brook Benton	Al Green	Dionne Warwick	Mariah Carey	Smokey Robinson	Andy Williams	Boyz II Men	Kenny Rogers
Freq	24	11	10	8	8	8	8	7	7	7
11	Elvis Presley	Bobby Bland	Ike & Tina Turner	James Brown	Johnnie Taylor	Booker T. & The MG's	Sly & The Family Stone	The Temptations	Aerosmith	Ray Charles
Freq	10	9	8	8	8	7	7	7	6	6
12	The Impressions	Earth Wind & Fire	Aretha Franklin	Barry White	Commodores	Four Tops	James Brown	Rod Stewart	Steely Dan	The 5th Dimensio
Freq	11	9	8	8	8	8	8	8	8	8
13	Aerosmith	The Rolling Stones	Bon Jovi	Kiss	Hannah Montana	Bryan Adams	REO Speedwagon	The Beatles	Van Halen	The Supremes
Freq	16	15	14	14	13	11	11	11	11	10

Esempio: evoluzione genere di musica



Esempio: caratteristiche giocatori



Esempio: caratteristiche giocatori

	1	2	3	4	5	6	7	8	9	10
CAM	23	3	1	0	337	24	79	368	97	26
CB	138	516	1036	0	0	0	55	1	32	0
CDM	211	186	15	0	3	0	328	2	203	0
CF	1	0	0	0	26	16	0	17	2	12
CM	304	24	1	0	49	1	519	193	303	0
GK	0	0	0	2025	0	0	0	0	0	0
LAM	0	0	0	0	17	0	0	2	0	2
LB	499	67	36	0	3	1	505	1	210	0
LCB	33	297	243	0	0	0	43	0	32	0
LCM	23	17	0	0	47	3	128	20	155	2
LDM	18	37	4	0	7	0	84	5	88	0
LF	0	0	0	0	8	0	0	3	3	1
LM	47	1	0	0	322	42	114	416	77	76
LS	0	0	0	0	57	28	1	15	0	106
LW	5	0	0	0	122	40	19	156	10	29
LWB	25	4	2	0	1	0	28	1	17	0
RAM	0	0	0	0	16	0	1	2	2	0
RB	515	67	34	0	0	1	501	4	169	0
RCB	24	307	273	0	0	0	32	0	25	1
RCM	17	17	1	0	32	1	144	17	159	3
RDM	16	48	0	0	6	1	75	5	97	0
RF	0	0	0	0	11	0	0	0	1	4
RM	44	2	0	0	324	65	134	427	71	57
RS	2	0	0	0	46	25	3	17	2	108
RW	1	0	0	0	123	38	22	155	8	23
RWB	36	4	1	0	0	0	35	1	10	0
ST	5	4	0	0	249	791	14	101	8	980

clus	Clausola	Stipendio	Valore
1	624808.5	1998.492	339529.9
2	6471702.7	13986.900	3351122.9
3	1223322.5	3458.738	676814.3
4	2997590.0	6800.592	1585032.1
5	13427200.2	23409.192	6952018.3
6	664438.5	2052.778	352810.2
7	2216212.5	5621.858	1212493.0
8	1169646.4	2995.342	639930.1
9	13337728.2	28052.779	6910033.7
10	4545180.0	10404.895	2481549.0

Esempio: caratteristiche giocatori

```
cargioc <- fifa19[rowSums(is.na(fifa19[,55:88])) == 0,]
fifa19cl <- kmeans(cargioc[,55:88],centers = 10)
clusplot(cargioc[,55:88], fifa19cl$cluster)
fifa19pr <- prcomp(cargioc[,55:88])$x[,1:2]
par(mar=c(4.5,4.5,1,1))
plot(fifa19pr,cex=0.5,pch=19,col=fifa19cl$cluster)
table(cargioc$Position,fifa19cl$cluster)
aggregate(cargioc[,90:92],by=list(clus=fifa19cl$cluster),m
ean,na.rm=T)
```

Esempio: immagini telerilevate

Con telerilevamento, o remote sensing, si definisce il metodo di osservazione della superficie terrestre e del suo ambiente dall'atmosfera o dallo spazio tramite l'utilizzo della radiazione elettromagnetica nel campo ottico o delle microonde. La Radiazione ElettroMagnetica (REM) E' il piu' utile campo di forza per il telerilevamento dal momento che stabilisce una connessione diretta e ad elevata velocita' (simile a quella della luce) fra il sensore, o l'osservatore, e l'oggetto di osservazione. Le variazioni della REM nella quantità e nelle sue proprietà divengono una preziosa fonte di dati per interpretare le proprietà del mezzo con cui essa ha interagito. La validità dell'uso della radiazione elettromagnetica come vettore di informazioni sugli oggetti che ci circondano è confermata dal fatto che l'uomo, come tutti gli animali, ne fa un largo uso attraverso il suo sensore, l'occhio, e il suo eleaboratore d'immagini, il cervello con la corteccia associativa. I sensori che registrano le variazioni del campo elettromagnetico al suo interagire con l'oggetto osservato possono essere "attivi" se sono accoppiati a un trasmettitore che emette la radiazione (ad esempio un radar), o "passivi", se invece sfruttano come fonte di radiazione il sole o la terra stessa. Quest'ultimi sono i sensori piu' utilizzati nel telerilevamento. I satelliti per il "remote sensing" sono essenzialmente di 2 tipi:

- satelliti orbitali: hanno orbite con altezza inferiore ai 2000 km e, dal momento che periodo orbitale e altezza dell'orbita sono correlati da una relazione precisa, con un periodo inferiore alle 2 ore. Eliosincroni son detti quei satelliti che osservano la stessa area alla stessa ora, in quanto l'angolo formato dal piano orbitale con la direzione terra-sole rimane costante.
- satelliti geostazionari: sono tali quei satelliti in cui la velocita' angolare e' uguale alla rotazione della terra. In tal caso il satellite appare fisso ad una altezza di 36.000 km sopra l'equatore.

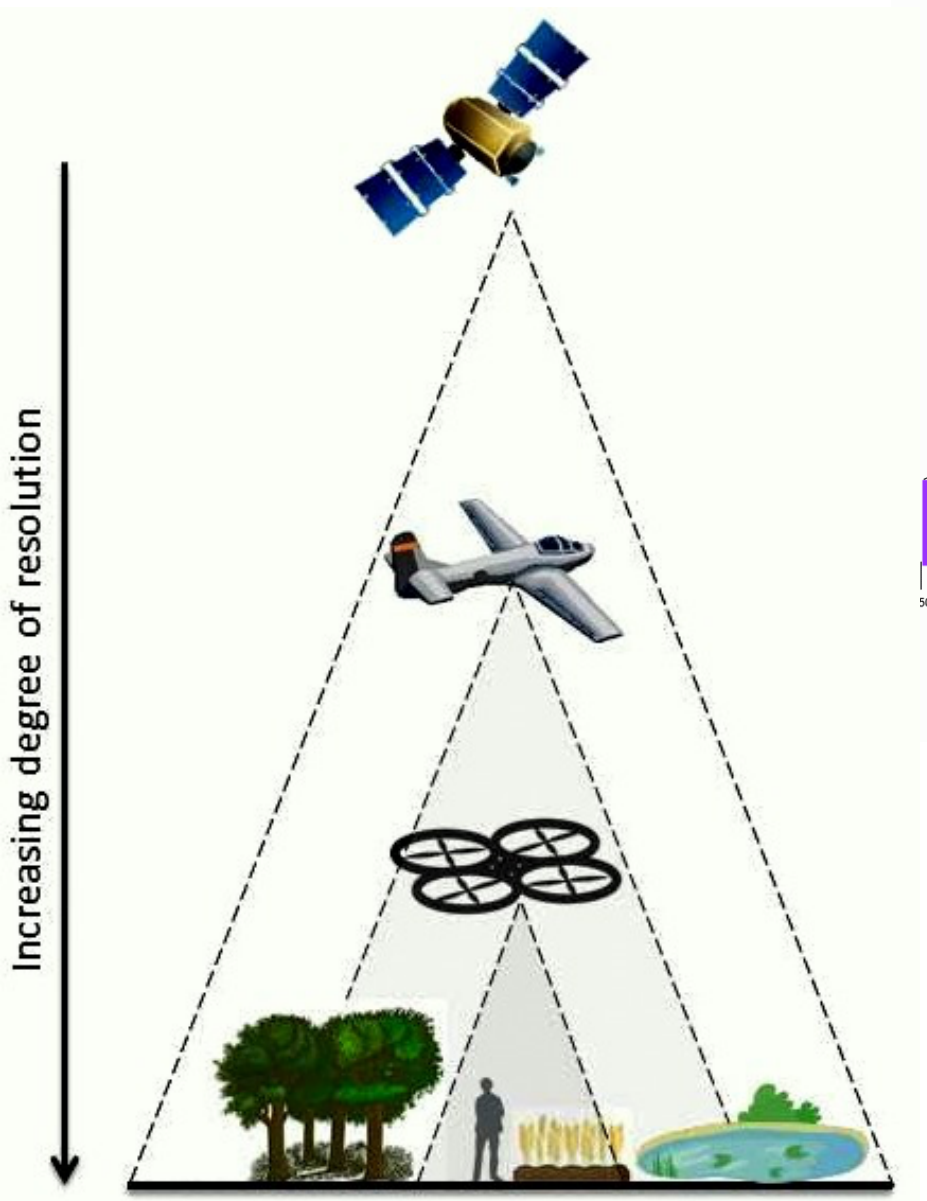
Esempio: immagini telerilevate

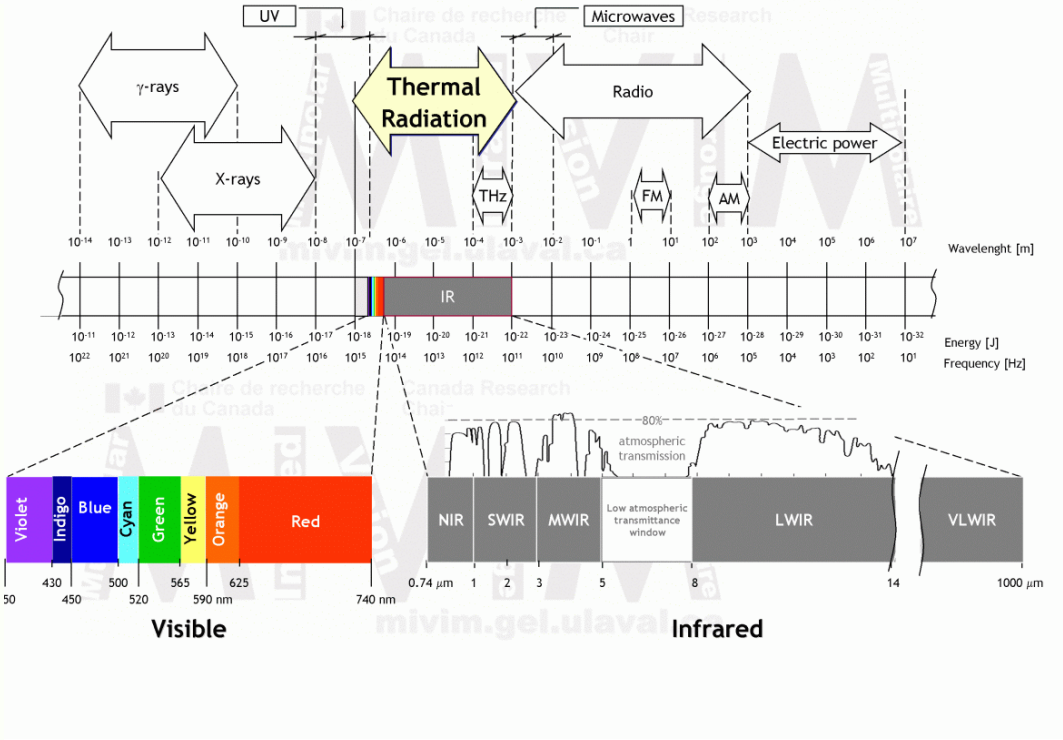
Sono geostazionari alcuni satelliti meteorologici (Meteosat, GOES, ecc.) e quelli per le telecomunicazioni (Eutelsat, Intelsat, Telecom, ecc.). Per l'Agricoltura sono usati i satelliti americani della serie LANDSAT di prima e seconda generazione (dal LANDSAT 1, lanciato nel 1972 al LANDSAT 5 lanciato nel 1984) con una ripetitività di 16 gg (ovvero passano sulla stessa area ogni 16 giorni) e un periodo orbitale di 98,9 minuti con un'altezza di 705 km (LANDSAT 5).

Questi satelliti hanno l'importante caratteristica di portare a bordo un sensore (TM=Thematic Mapper) con ben 7 bande di osservazione (4 nel visibile e nel vicino infrarosso, 2 nell'infrarosso medio ed una nel termico) e una risoluzione spaziale di 30 m al suolo.

Tuttavia la sua ripetitività di 16 giorni sebbene lo renda molto utile ai fini del rilevamento dell'area investita nelle varie colture, risulta poco utile per il rilevamento delle variabili agroclimatiche che influenzano le rese o di altri cambiamenti di stato della vegetazione che sono spesso caratterizzati da periodi di variazione più brevi (soprattutto se si tiene conto che la presenza di nubi, spesso riduce la disponibilita' di buone immagini ad 1 ogni 40-60 giorni o più).

Esempio: immagini telerilevate



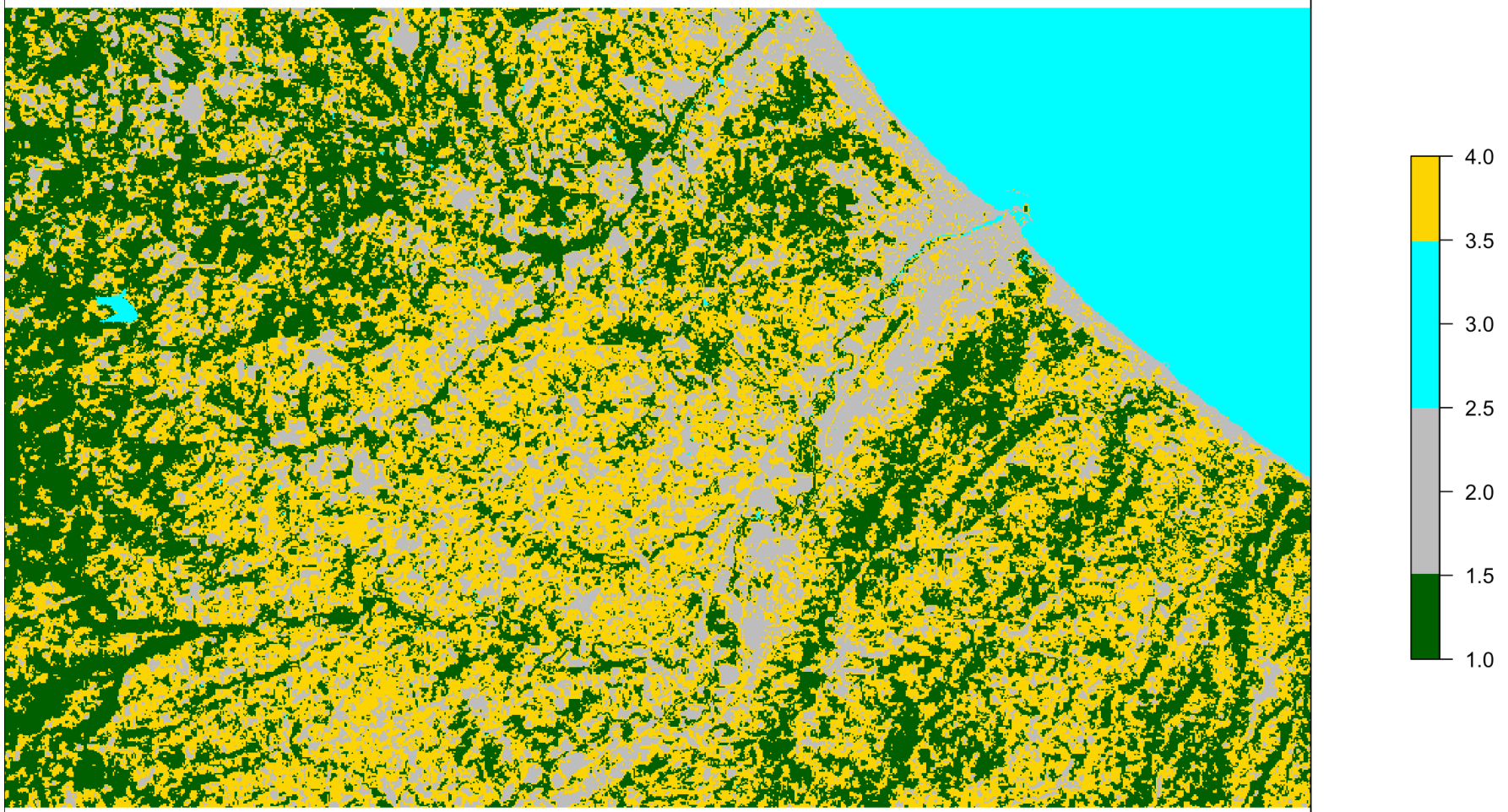


Esempio: immagini telerilevate

Esempio: immagini telerilevate

Esempio: immagini telerilevate

Esempio: immagini telerilevate



Esempio: immagini telerilevate

```
library(raster)
ls8 <- stack('agric.png')[[-4]]
names(ls8) <- c('blue', 'green', 'red')
plotRGB(ls8)
nr <- getValues(ls8)
kmls8 <- kmeans(na.omit(nr), centers = 10, iter.max = 500, nstart = 5,
algorithm="Lloyd")
table(kmls8$cluster)
sclu <- ls8[[1]]
knr <- setValues(sclu, kmls8$cluster)
plot(knr, main = 'Unsupervised classification', col = terrain.colors(10))
sent3 <- stack('pescara sentinel.png')[[-4]]
names(sent3) <- c('blue', 'green', 'red')
plotRGB(sent3)
nr <- getValues(sent3)
kmsent3 <- kmeans(na.omit(nr), centers = 4, iter.max = 500, nstart = 5,
algorithm="Lloyd”)
table(kmsent3$cluster)
sclu <- sent3[[1]]
knr <- setValues(sclu, kmsent3$cluster)
plot(knr, main = 'Unsupervised classification', col =
c("darkgreen","gray","cyan","gold"))
```

Riassunto

- La Cluster Analysis raggruppa le unità in base alla loro somiglianza ed ha ampie applicazioni
- La misura della similarità può essere calcolata per vari tipi di dati
- Gli algoritmi di clustering possono essere suddivisi in metodi di partizionamento, metodi gerarchici, metodi basati sulla densità, metodi basati sulla griglia e metodi basati su modelli
- Gli algoritmi K-means e K-medoids sono algoritmi di clustering basati sul partizionamento
- Birch e Chameleon sono interessanti algoritmi di clustering gerarchico, e ci sono anche algoritmi di clustering gerarchici probabilistici
- DBSCAN, OPTICS e DENCLU sono interessanti algoritmi basati sulla densità
- STING e CLIQUE sono metodi basati sulla griglia, in cui CLIQUE è anche un algoritmo di clustering subspaziale
- La qualità dei risultati del clustering può essere valutata in vari modi
- Clustering basato sul modello di probabilità
- Raggruppamento sfocato
- Clustering model-based della distribuzione di probabilità
  - L'algoritmo EM
- Grafici di clustering per grafi e dati di rete
  - Raggruppamento di grafi: taglio min-cut vs. taglio più corto
  - Metodi di clustering ad alta dimensionalità
  - Metodi di clustering specifici del grafico, ad es. SCAN

Bibliografia

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD’ 99.
- Beil F., Ester M., Xu X.: "Frequent Term-Based Text Clustering", KDD'02
- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. VLDB’ 98.
- V. Ganti, J. Gehrke, R. Ramakrishan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.

Bibliografia

- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB’98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In *ICDE'99*, pp. 512-521, Sydney, Australia, March 1999.
- A. Hinneburg, D.I A. Keim: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD’98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
- G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *COMPUTER*, 32(8): 68-75, 1999.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB’98.

Bibliografia

- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- L. Parsons, E. Haque and H. Liu, Subspace Clustering for High Dimensional Data: A Review, SIGKDD Explorations, 6(1), June 2004
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB' 98.
- A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-Based Clustering in Large Databases, ICDT'01.
- A. K. H. Tung, J. Hou, and J. Han. Spatial Clustering in the Presence of Obstacles, ICDE'01
- H. Wang, W. Wang, J. Yang, and P.S. Yu. Clustering by pattern similarity in large data sets, SIGMOD' 02
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB' 97
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : An efficient data clustering method for very large databases. SIGMOD'96
- X. Yin, J. Han, and P. S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", VLDB'06

Bibliografia (Metodi di clustering più sofisticati)

- C. C. Aggarwal, C. Procopiuc, J. Wolf, P. S. Yu, and J.-S. Park. Fast algorithms for projected clustering. *SIGMOD´99*
- S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56:5:1–5:37, 2009.
- J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.
- K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ”nearest neighbor” meaningful? *ICDT´99*
- Y. Cheng and G. Church. Biclustering of expression data. *ISMB´00*
- I. Davidson and S. S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. *SDM´05*
- I. Davidson, K. L. Wagstaff, and S. Basu. Measuring constraint-set utility for partitional clustering algorithms. *PKDD´06*
- C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. American Stat. Assoc.*, 97:611–631, 2002.
- F. H¨oppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Wiley, 1999.
- G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. *KDD´02*
- H.-P. Kriegel, P. Kroeger, and A. Zimek. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowledge Discovery from Data (TKDD)*, 3, 2009.
- U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007

Bibliografia (Metodi di clustering più sofisticati)

- G. J. McLachlan and K. E. Bkasford. *Mixture Models: Inference and Applications to Clustering*. John Wiley & Sons, 1988.
- B. Mirkin. Mathematical classification and clustering. *J. of Global Optimization*, 12:105–108, 1998.
- S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1, 2004.
- A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. NIPS'01
- J. Pei, X. Zhang, M. Cho, H. Wang, and P. S. Yu. Maple: A fast algorithm for maximal pattern-based clustering. *ICDM'03*
- M. Radovanovi´c, A. Nanopoulos, and M. Ivanovi´c. Nearest neighbors in high-dimensional data: the emergence and influence of hubs. *ICML'09*
- S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1:27–64, 2007.
- A. K. H. Tung, J. Hou, and J. Han. Spatial clustering in the presence of obstacles. *ICDE'01*
- A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-based clustering in large databases. *ICDT'01*
- A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. In *Handbook of Computational Molecular Biology, Chapman & Hall*, 2004.
- K. Wagstaff, C. Cardie, S. Rogers, and S. Schr¨odl. Constrained k-means clustering with background knowledge. *ICML'01*
- H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. *SIGMOD'02*
- X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: A structural clustering algorithm for networks. *KDD'07*
- X. Yin, J. Han, and P.S. Yu, "Cross-Relational Clustering with User's Guidance", KDD'05

Conclusione

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Libera traduzione
"La validazione dei gruppi ottenuti è la parte più difficile e frustrante della Cluster Analysis.
Senza un forte sforzo in questa direzione, la Cluster Analysis rimarrà un'arte oscura accessibile solo a quei veri credenti che hanno esperienza e grande coraggio."