



Università degli Studi di Salerno
Dipartimento di Informatica

Corso di Metodi e tecniche per l'analisi dei dati

Dati ambientali nella città

Indicatori sulla raccolta differenziata per le regioni

Docente

Prof.ssa Amelia Giuseppina Nobile

Studente

Maria Domenica Guida
Matr. 0522500236

Anno Accademico 2014/15

Indice

1	Introduzione	3
1.1	Organizzazione del lavoro	3
2	Strumenti e dati utilizzati	4
2.1	Strumenti	4
2.2	Fonte Dati	4
3	Analisi dei dati	5
3.1	Inserimento dati	5
3.2	Ripartizione dati	6
3.3	Dati in percentuale	7
4	Analisi Grafica	9
4.1	Rappresentazione grafica	9
4.1.1	Carta e Cartone	11
4.1.2	Vetro	12
4.1.3	Plastica	13
4.1.4	Metalli	14
4.1.5	Raccolta Selettiva	15
4.1.6	Rifiuti verdi, organici e legno	16
4.1.7	Altro	17
4.2	Rappresentazione dei valori numerici	17
4.2.1	Carta e Cartone	19
4.2.2	Vetro	20
4.2.3	Plastica	21
4.2.4	Metalli	22
4.2.5	Raccolta Selettiva	23
4.2.6	Rifiuti verdi, organici e legno	24
4.2.7	Altro	25
4.3	Confronto tra Nord, Centro e Sud	26
4.3.1	Carta e Cartone	26
4.3.2	Vetro	27
4.3.3	Plastica	27
4.3.4	Metalli	28
4.3.5	Raccolta Selettiva	28
4.3.6	Rifiuti verdi, organici e legno	29
4.3.7	Altro	29
4.4	Variabili a confronto	30
5	Statistica descrittiva con R	31
5.1	Indici di sintesi: posizione e dispersione	31
5.1.1	Indice di posizione	31
5.1.2	Calcolo degli indici di posizione con R	32
5.1.3	Indici di dispersione	32
5.1.4	Calcolo degli indici di dispersione con R	33

5.1.5	Boxplot	34
5.2	Correlazioni tra le variabili	35
5.2.1	Covarianza campionaria	36
5.2.2	Coefficiente di correlazione campionario	36
5.2.3	Scatterplot	37
6	Clustering	38
6.1	Cluster gerarchico	39
6.1.1	Metodo del legame singolo	40
6.1.2	Metodo del legame completo	41
6.1.3	Metodo del legame medio	43
6.1.4	Metodo del centroide	44
6.1.5	Metodo della mediana	45
6.2	Metodi non gerarchici	47
6.2.1	K-Means	47

1 Introduzione

L'obiettivo del progetto è quello di analizzare una mole di dati, realizzando alcuni grafici con il software R.

R è un potente sistema che utilizza un linguaggio di programmazione, costituito da una varietà di strumenti per l'analisi statistica dei dati e per la loro visualizzazione. R è quindi un ambiente integrato che permette di elaborare i dati, eseguire i calcoli e produrre grafici.

Sulla fonte dei dati scelti, effettueremo l'analisi delle frequenze assolute e relative generando grafici specifici (grafici a barre, a torta, ecc.), esamineremo gli indici di posizione e di dispersione, la correlazione tra le variabili e, infine, effettueremo l'analisi dei cluster.

Per effettuare l'analisi richiesta sono stati scelti dei dati dal sito <http://www.istat.it>. L'ISTAT è l'Istituto Nazionale di Statistica ed è il principale produttore di statistica ufficiale a supporto dei cittadini.

1.1 Organizzazione del lavoro

Il seguente lavoro è stato così suddiviso:

- Nel capitolo 2 verranno descritti i dati utilizzati e gli strumenti

2 Strumenti e dati utilizzati

In questa sezione verranno descritti gli strumenti utilizzati per l'analisi e successivamente i dati presi in considerazione.

2.1 Strumenti

L'analisi dei dati è stata effettuata utilizzando il software RStudio che utilizza come motore il linguaggio R versione 3.1.1 su piattaforma Apple.

2.2 Fonte Dati

Come annunciato precedentemente, i dati utilizzati per eseguire l'analisi sono stati estratti dal sito dell'ISTAT. Questi ultimi, si riferiscono all'indagine che raccoglie informazioni ambientali relative alle regioni italiane.

In particolare sono stati prelevati i dati relativi all'anno 2012 riguardanti i chili pro capite di raccolta differenziata per tipo di rifiuto per le regioni italiane.

Nella seguente tabella vengono mostrati i dati utilizzati espressi in chili pro capite:

	Carta e cartone	Vetro	materie plastiche	metalli	Raccolta selettiva	Organico e legno	altro	totale
Piemonte	80.79	37.42	29.37	7.41	0.47	121.32	21.62	298.40
Valle d'Aosta	71.02	41.30	22.37	8.84	0.87	90.52	56.54	291.45
Liguria	41.98	17.89	9.24	2.14	0.25	28.19	20.72	120.40
Lombardia	73.08	42.18	14.91	4.14	0.61	89.16	18.59	242.68
Prov.Aut. Bolzano	75.68	34.05	5.22	2.40	2.14	108.79	19.16	247.43
Prov.Aut. Trento	86.46	33.99	22.63	6.69	1.58	149.41	36.53	337.29
Veneto	75.58	42.02	21.06	7.77	0.82	133.44	26.41	307.10
Friuli-Venezia Giulia	76.81	33.97	16.14	4.49	0.82	121.84	36.39	290.45
Emilia-Romagna	96.52	34.28	25.44	10.54	0.58	156.82	38.41	362.59
Toscana	76.75	24.26	14.92	3.54	0.49	110.57	20.87	251.40
Umbria	83.82	31.36	22.09	6.26	0.36	120.98	26.99	291.87
Marche	73.69	28.53	20.67	4.32	0.45	121.71	30.37	279.73
Lazio	39.42	18.06	6.12	4.83	0.20	42.90	12.55	124.07
Abruzzo	64.85	27.08	19.13	4.06	0.25	95.94	15.68	227.00
Molise	14.81	9.82	2.53	3.26	0.17	7.67	4.31	42.57
Campania	49.51	21.32	24.53	6.89	0.52	112.13	23.96	238.85
Puglia	34.53	9.78	7.75	0.72	0.11	31.00	10.23	94.12
Basilicata	54.63	8.82	3.01	3.97	0.05	28.57	5.49	104.54
Calabria	21.61	2.72	2.16	4.92	0.00	16.49	11.09	57.78
Sicilia	20.88	4.74	3.52	2.87	0.06	17.60	8.85	58.20
Sardegna	50.12	29.28	14.20	3.73	0.21	111.42	22.28	231.19.

Sono stati scelti questi tipi di dati per verificare quali sono le regioni che differenziano di più in Italia. I dati e l'informazione statistica, derivati dall'indagine e annualmente diffusi, rappresentano un essenziale strumento informativo a supporto del monitoraggio dello stato dell'ambiente urbano e delle attività poste in essere dalle amministrazioni per assicurare la buona qualità dell'ambiente nelle città.

3 Analisi dei dati

3.1 Inserimento dati

L'operazione d'inserimento dati prevede di caricare la tabella da analizzare "raccolta differenziata dei rifiuti urbani per le regioni - chili" all'interno del software R per svolgere l'operazione di analisi. I seguenti comandi sono stati utilizzati per importare i dati in R:

Listing 1: Inserimento della tabella in R.

```
> Piemonte <- c(80.79, 37.42, 29.37, 7.41, 0.47, 121.32, 21.62, 298.40)
> Valle_d_Aosta <- c(71.02, 41.30, 22.37, 8.84, 0.87, 90.52, 56.54, 291.45)
> Liguria <- c(41.98, 17.89, 9.24, 2.14, 0.25, 28.19, 20.72, 120.40)
> Lombardia <- c(73.08, 42.18, 14.91, 4.14, 0.61, 89.16, 18.59, 242.68)
> Prov_Aut_Bolzano <- c(75.68, 34.05, 5.22, 2.40, 2.14, 108.79, 19.16, 247.43)
> Prov_Aut_Trento <- c(86.46, 33.99, 22.63, 6.69, 1.58, 149.41, 36.53, 337.29)
> Veneto <- c(75.58, 42.02, 21.06, 7.77, 0.82, 133.44, 26.41, 307.10)
> Friuli_Venezia_Giulia <- c(76.81, 33.97, 16.14, 4.49, 0.82, 121.84, 36.39, 290.45)
> Emilia_Romagna <- c(96.52, 34.28, 25.44, 10.54, 0.58, 156.82, 38.41, 362.59)
> Toscana <- c(76.75, 24.26, 14.92, 3.54, 0.49, 110.57, 20.87, 251.40)
> Umbria <- c(83.82, 31.36, 22.09, 6.26, 0.36, 120.98, 26.99, 291.87)
> Marche <- c(73.69, 28.53, 20.67, 4.32, 0.45, 121.71, 30.37, 279.73)
> Lazio <- c(39.42, 18.06, 6.12, 4.83, 0.20, 42.90, 12.55, 124.07)
> Abruzzo <- c(64.85, 27.08, 19.13, 4.06, 0.25, 95.94, 15.68, 227.00)
> Molise <- c(14.81, 9.82, 2.53, 3.26, 0.17, 7.67, 4.31, 42.57)
> Campania <- c(49.51, 21.32, 24.53, 6.89, 0.52, 112.13, 23.96, 238.85)
> Puglia <- c(34.53, 9.78, 7.75, 0.72, 0.11, 31.00, 10.23, 94.12)
> Basilicata <- c(54.63, 8.82, 3.01, 3.97, 0.05, 28.57, 5.49, 104.54)
> Calabria <- c(21.61, 2.72, 2.16, 4.92, 0.00, 16.49, 11.09, 57.78)
> Sicilia <- c(20.88, 4.74, 3.52, 2.87, 0.06, 17.60, 8.85, 58.20)
> Sardegna <- c(50.12, 29.28, 14.20, 3.73, 0.21, 111.42, 22.28, 231.19)
```

Di conseguenza è stata creata la matrice su cui eseguire l'analisi. I comandi utilizzati sono i seguenti:

Listing 2: creazione della matrice

```
> matriceAnalisi<-matrix(c(Piemonte,Valle_d_Aosta, Lombardia,Prov_Aut_Bolzano,
Prov_Aut_Trento, Veneto,Friuli_Venezia_Giulia, Liguria,Emilia_Romagna,Toscana,
Umbria,Marche,Lazio,Abruzzo,Molise,Campania,Puglia,Basilicata,Calabria,Sicilia,
Sardegna), nrow=21,ncol=8,byrow=TRUE)
```

Per associare ogni riga ad una diversa regione sono state create delle etichette.

Il comando utilizzato per creare le etichette delle regioni è il seguente:

```
> labelRegioni <-c("Piemonte","Valle d'Aosta", "Lombardia",
"Provincia Autonoma Bolzano", "Provincia Autonoma Trento",
"Veneto","Friuli Venezia Giulia", "Liguria", "Emilia Romagna","Toscana",
"Umbria","Marche","Lazio","Abruzzo","Molise","Campania","Puglia",
"Basilicata","Calabria","Sicilia","Sardegna")
```

Per associare ogni colonna al rifiuto relativo sono state create delle etichette, per ragioni di comodità alcuni nomi sono stati abbreviati.

Il comando utilizzato per creare le etichette dei rifiuti è il seguente:

```
>labelRifiuti <-c("carta/cartone","vetro","plastica","metalli","selettiva",
"verde/organici/legno","altro","totale");
```

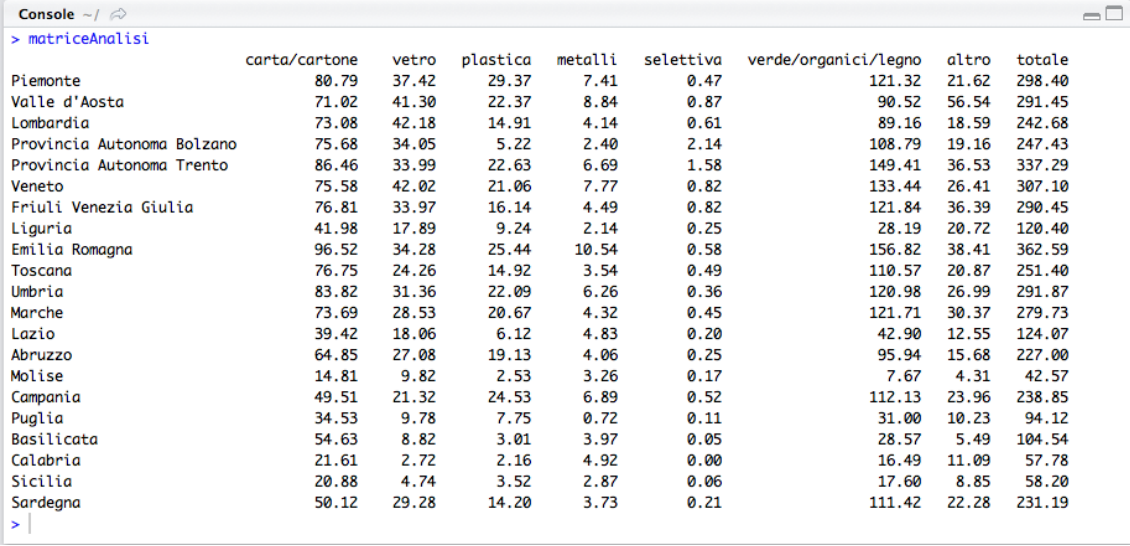
Per associare le etichette create alla matrice è stato utilizzato il seguente comando:

```
> rownames(matriceAnalisi)<-labelRegioni
> colnames(matriceAnalisi)<-labelRifiuti
```

Per visualizzare la matrice creata, bisognerà eseguire il comando:

```
> matriceAnalisi
```

In questo modo potremo visualizzare il seguente risultato:



	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro	totale
Piemonte	80.79	37.42	29.37	7.41	0.47	121.32	21.62	298.40
Valle d'Aosta	71.02	41.30	22.37	8.84	0.87	90.52	56.54	291.45
Lombardia	73.08	42.18	14.91	4.14	0.61	89.16	18.59	242.68
Provincia Autonoma Bolzano	75.68	34.05	5.22	2.40	2.14	108.79	19.16	247.43
Provincia Autonoma Trento	86.46	33.99	22.63	6.69	1.58	149.41	36.53	337.29
Veneto	75.58	42.02	21.06	7.77	0.82	133.44	26.41	307.10
Friuli Venezia Giulia	76.81	33.97	16.14	4.49	0.82	121.84	36.39	290.45
Liguria	41.98	17.89	9.24	2.14	0.25	28.19	20.72	120.40
Emilia Romagna	96.52	34.28	25.44	10.54	0.58	156.82	38.41	362.59
Toscana	76.75	24.26	14.92	3.54	0.49	110.57	20.87	251.40
Umbria	83.82	31.36	22.09	6.26	0.36	120.98	26.99	291.87
Marche	73.69	28.53	20.67	4.32	0.45	121.71	30.37	279.73
Lazio	39.42	18.06	6.12	4.83	0.20	42.90	12.55	124.07
Abruzzo	64.85	27.08	19.13	4.06	0.25	95.94	15.68	227.00
Molise	14.81	9.82	2.53	3.26	0.17	7.67	4.31	42.57
Campania	49.51	21.32	24.53	6.89	0.52	112.13	23.96	238.85
Puglia	34.53	9.78	7.75	0.72	0.11	31.00	10.23	94.12
Basilicata	54.63	8.82	3.01	3.97	0.05	28.57	5.49	104.54
Calabria	21.61	2.72	2.16	4.92	0.00	16.49	11.09	57.78
Sicilia	20.88	4.74	3.52	2.87	0.06	17.60	8.85	58.20
Sardegna	50.12	29.28	14.20	3.73	0.21	111.42	22.28	231.19

3.2 Ripartizione dati

Successivamente sono state raggruppate le regioni nelle tre aree geografiche di riferimento, Nord, Centro e Sud, e creata la matrice con le ripartizioni con le dovute etichette. I comandi eseguiti per effettuare questa ripartizione sono i seguenti:

```
> Nord<-round((Piemonte + Valle_d_Aosta + Liguria + Lombardia +
Prov_Aut_Bolzano + Prov_Aut_Trento + Veneto + Friuli_Venezia_Giulia +
Emilia_Romagna)/9,2)
> Centro<-round((Marche+Toscana+Umbria+Lazio)/4,2)
> Sud<-round((Abruzzo+Molise+Campania+Puglia+Basilicata+Calabria+Sicilia+
Sardegna)/8,2)

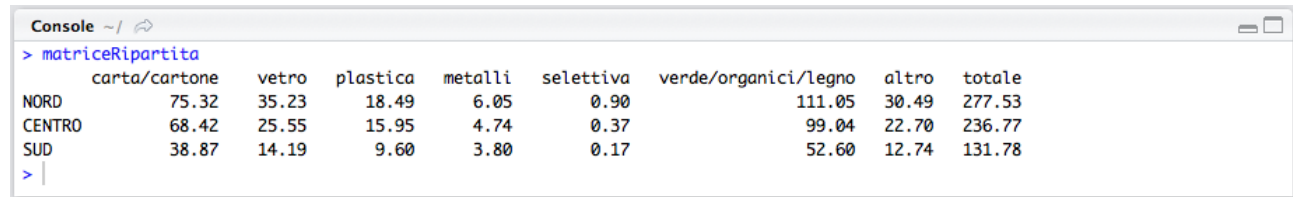
#creo la matrice ripartita
> matriceRipartita <- matrix(c(Nord, Centro, Sud),nrow=3,ncol=8,byrow=TRUE)

#Etichetta area geografica
> labelArea <-c("NORD", "CENTRO", "SUD");
```

#aggiunte le etichette dell'area geografica e dei rifiuti alle righe e alle colonne della matrice ripartita:

```
> rownames(matriceRipartita)<-labelArea
> colnames(matriceRipartita)<-labelRifiuti
```

La matrice visualizzata sarà la seguente:



	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro	totale
NORD	75.32	35.23	18.49	6.05	0.90	111.05	30.49	277.53
CENTRO	68.42	25.55	15.95	4.74	0.37	99.04	22.70	236.77
SUD	38.87	14.19	9.60	3.80	0.17	52.60	12.74	131.78

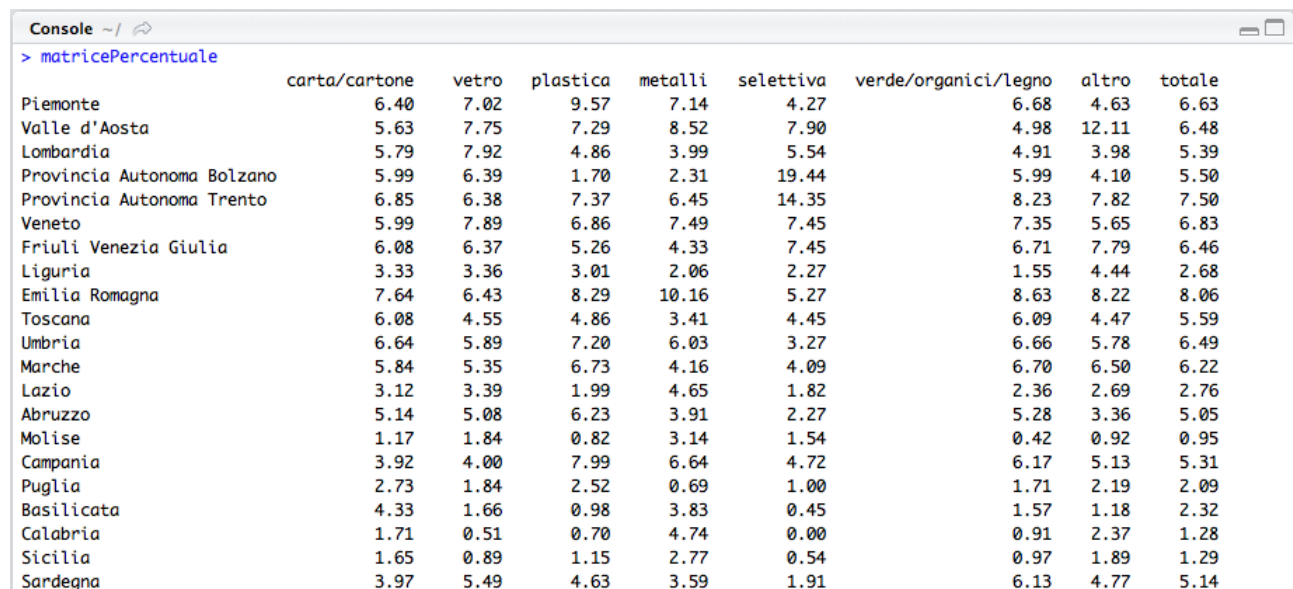
3.3 Dati in percentuale

Il passo successivo è stato quello di creare una tabella che mostra le percentuali dei rifiuti riciclati dagli abitanti di ogni singola regione rispetto al totale riciclato in Italia.

Per fare ciò è stata calcolata la distribuzione delle frequenze relative dei rifiuti riciclati da ogni singolo abitante in ogni regione rispetto al totale in Italia, il tutto moltiplicato per 100. I comandi eseguiti per effettuare questa operazione sono i seguenti:

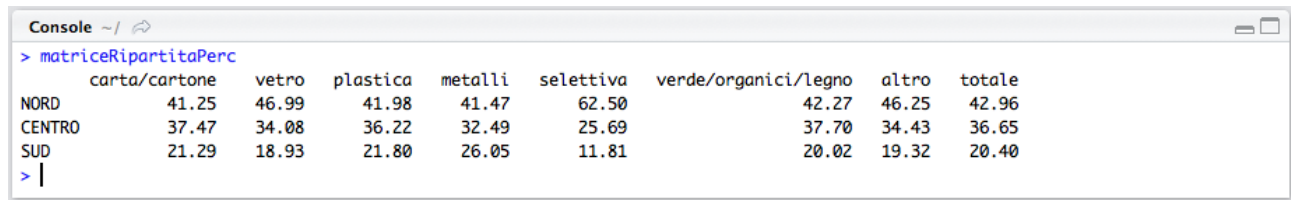
```
> matricePercentuale<-round((prop.table(matriceAnalisi,2)*100),2)
> matriceRipartitaPerc<-round((prop.table(matriceRipartita,2)*100),2)
```

Con i seguenti risultati:



	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro	totale
Piemonte	6.40	7.02	9.57	7.14	4.27	6.68	4.63	6.63
Valle d'Aosta	5.63	7.75	7.29	8.52	7.90	4.98	12.11	6.48
Lombardia	5.79	7.92	4.86	3.99	5.54	4.91	3.98	5.39
Provincia Autonoma Bolzano	5.99	6.39	1.70	2.31	19.44	5.99	4.10	5.50
Provincia Autonoma Trento	6.85	6.38	7.37	6.45	14.35	8.23	7.82	7.50
Veneto	5.99	7.89	6.86	7.49	7.45	7.35	5.65	6.83
Friuli Venezia Giulia	6.08	6.37	5.26	4.33	7.45	6.71	7.79	6.46
Liguria	3.33	3.36	3.01	2.06	2.27	1.55	4.44	2.68
Emilia Romagna	7.64	6.43	8.29	10.16	5.27	8.63	8.22	8.06
Toscana	6.08	4.55	4.86	3.41	4.45	6.09	4.47	5.59
Umbria	6.64	5.89	7.20	6.03	3.27	6.66	5.78	6.49
Marche	5.84	5.35	6.73	4.16	4.09	6.70	6.50	6.22
Lazio	3.12	3.39	1.99	4.65	1.82	2.36	2.69	2.76
Abruzzo	5.14	5.08	6.23	3.91	2.27	5.28	3.36	5.05
Molise	1.17	1.84	0.82	3.14	1.54	0.42	0.92	0.95
Campania	3.92	4.00	7.99	6.64	4.72	6.17	5.13	5.31
Puglia	2.73	1.84	2.52	0.69	1.00	1.71	2.19	2.09
Basilicata	4.33	1.66	0.98	3.83	0.45	1.57	1.18	2.32
Calabria	1.71	0.51	0.70	4.74	0.00	0.91	2.37	1.28
Sicilia	1.65	0.89	1.15	2.77	0.54	0.97	1.89	1.29
Sardegna	3.97	5.49	4.63	3.59	1.91	6.13	4.77	5.14

Figura 1: Matrice dei dati con percentuale




```
Console ~/   
> matriceRipartitaPerc  
      carta/cartone  vetro  plastica  metalli  selettiva  verde/organici/legno  altro  totale  
NORD      41.25    46.99    41.98    41.47    62.50           42.27    46.25    42.96  
CENTRO     37.47    34.08    36.22    32.49    25.69           37.70    34.43    36.65  
SUD        21.29    18.93    21.80    26.05    11.81           20.02    19.32    20.40  
> |
```

Figura 2: Matrice dei dati ripartita con percentuale

4 Analisi Grafica

In questo capitolo verranno rappresentati graficamente in R i dati discussi in precedenza. La prima analisi riguarderà i dati percentuali, mentre tutte le altre riguarderanno quelli numerici e per ogni sezione sarà presente sia il codice utilizzato sia conclusioni eseguite.

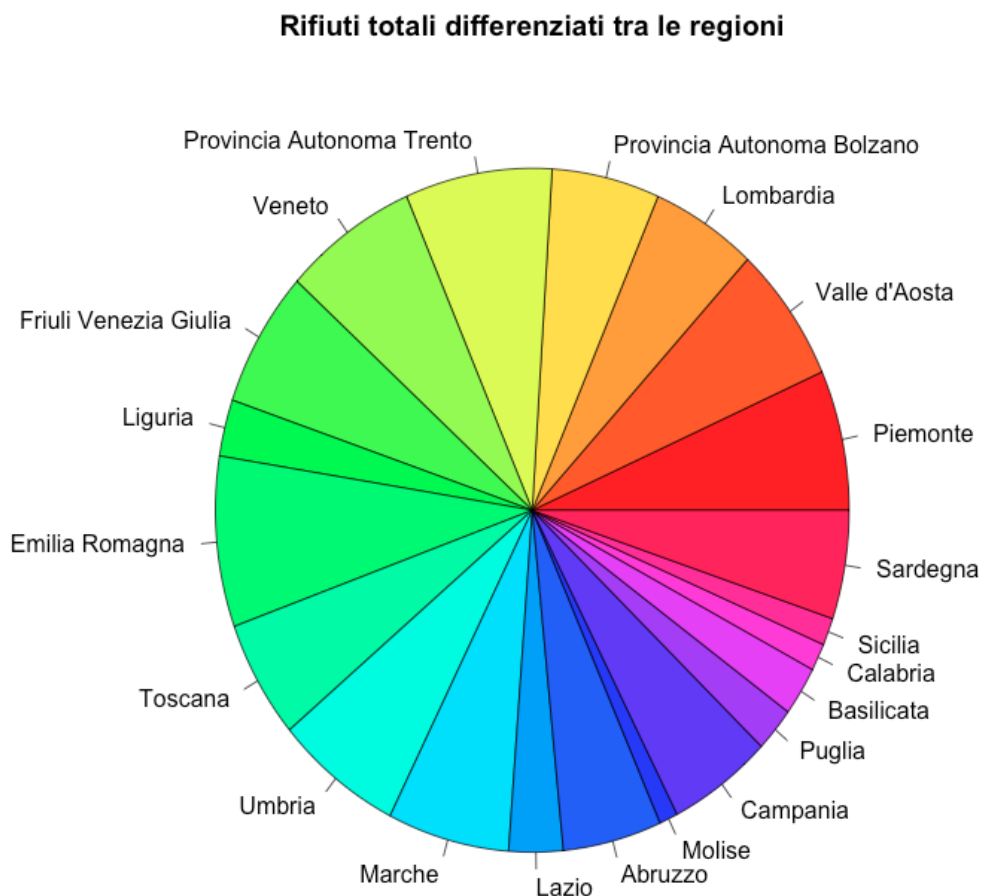
4.1 Rappresentazione grafica

Di seguito sono stati creati vari grafici per la rappresentazione dei dati.

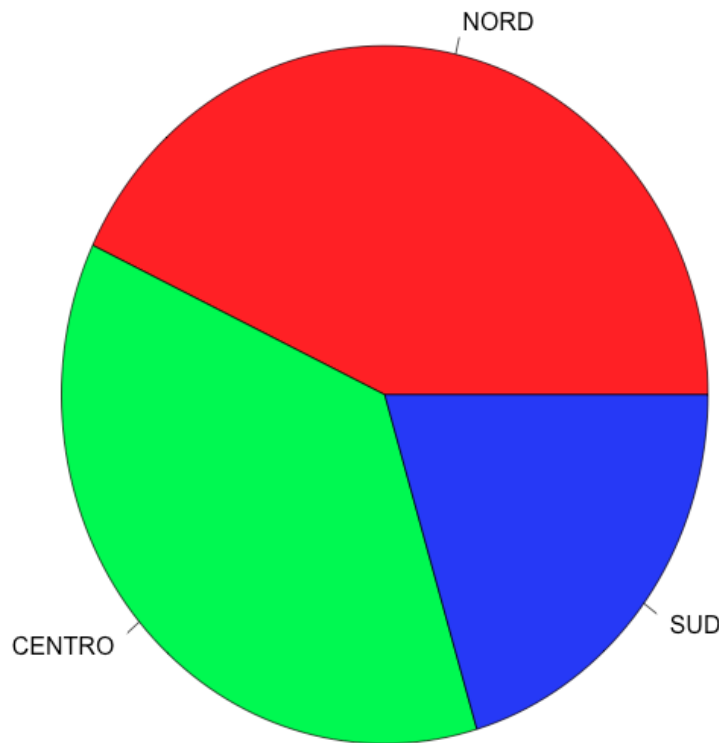
Prima di iniziare con la rappresentazione dettagliata della raccolta differenziata di ogni singolo rifiuto, viene presentato un diagramma a torta che darà un'idea del risultato finale, e che rappresenta la frequenza assoluta del totale dei rifiuti riciclati nell'anno 2012 dalle regioni italiane e il diagramma per le regioni ripartire in Nord, Centro e sud. Il codice utilizzato è il seguente:

```
> pie(matriceAnalisi[,8], main = "Rifiuti totali differenziati tra le regioni",  
      col=rainbow(21,s=0.7))  
> pie(matriceRipartita[,8], main = "Rifiuti totali differenziati tra Nord, Centro  
e Sud",col=rainbow(3,s=0.7))
```

Con il seguente risultato



Rifiuti totali differenziati tra Nord, Centro e Sud



Da questi due grafici si evince quali sono le regioni in cui gli abitanti differenziano di più in Italia. Successivamente è stato costruito in R, per ogni tipologia di rifiuto, il corrispettivo grafico a barre utilizzando il comando `barplot` ed il comando `text` per inserire il testo.

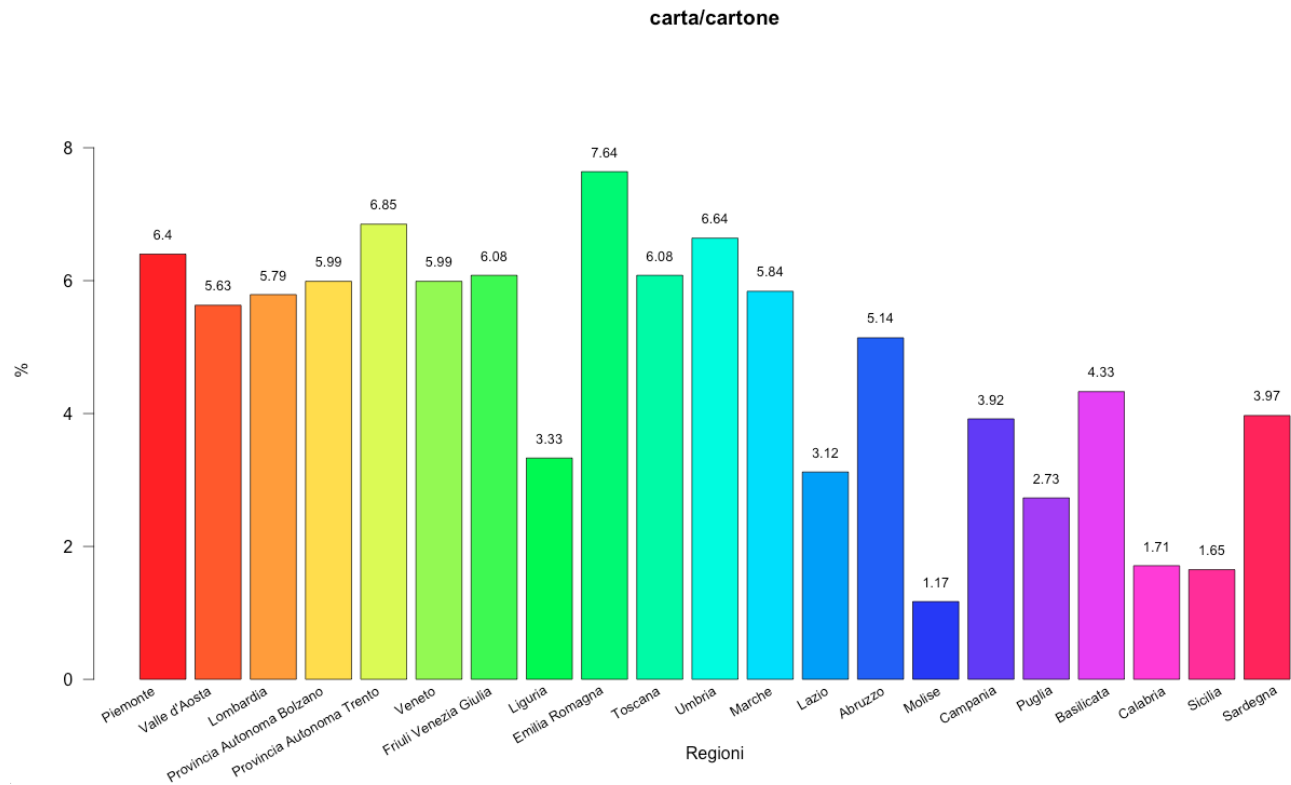
Con `barplot`, infatti, non è possibile inserire le etichette delle variabili in obliquo e ciò si può ottenere soltanto utilizzando opportunamente il parametro `srt` della funzione `text`.

Il codice utilizzato è il seguente:

```
#grafico a barre sulle percentuali
> for(i in 1:7){
>   valPerc<-max(matricePercentuale[,i])
>   livelloPercentuale<-valPerc+(valPerc/4.5)
>   bpt17 <- barplot(matricePercentuale[,i],main=labelRifiuti[i], xlab="Regioni",
>     ylab="%", names.arg="", ylim=c(0,livelloPercentuale),las=2,col=rainbow(21,s=0.9))
>   text(seq(0.5,25,by=1.22), par("usr")[3]-0.15, srt = 30, adj = 1, xpd = TRUE,
>     labels = paste(rownames(matriceAnalisi)),cex=0.8)
> }
```

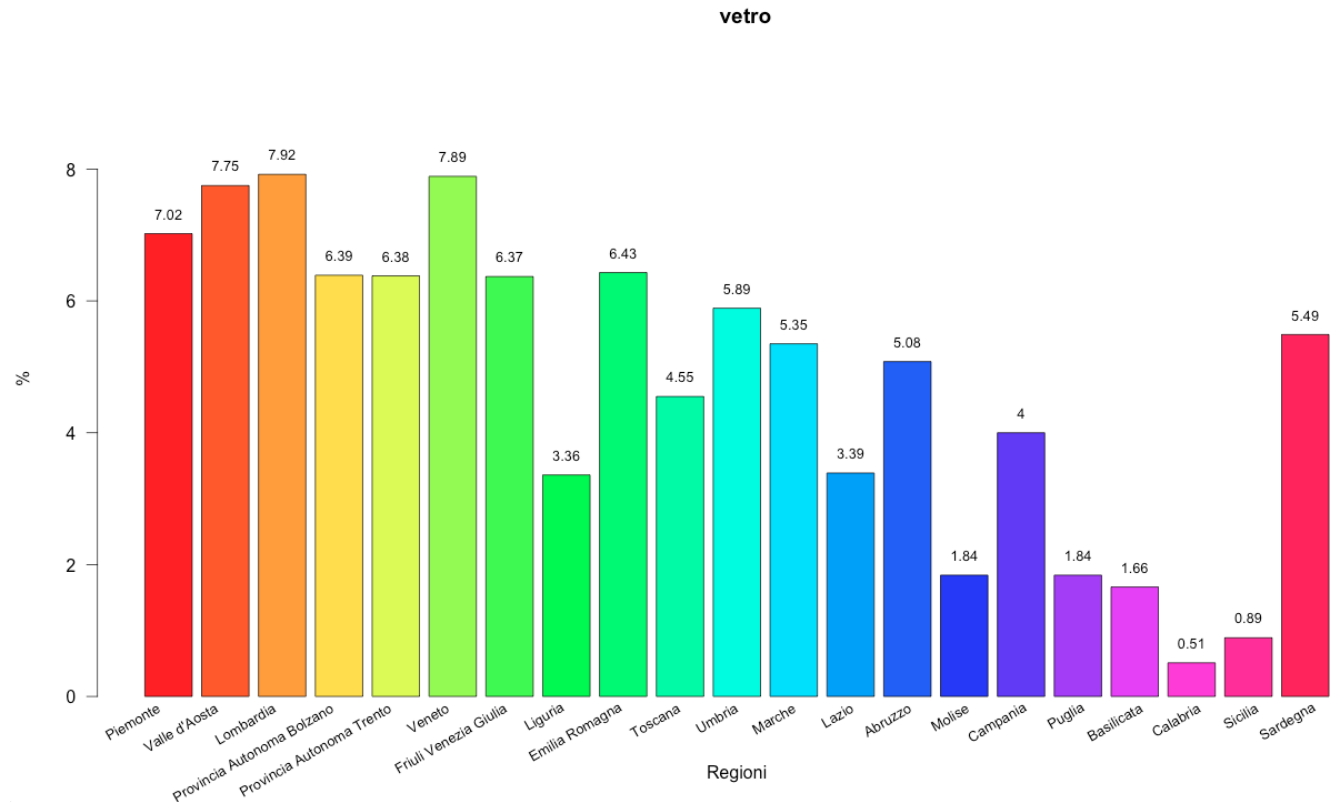
Essendo che, per ogni tipo di rifiuto, era da applicare la stessa funzione, il tutto è stato inserito all'interno di un ciclo `for`, modificando opportunamente le variabili, in modo da non essere ripetitivi. Di seguito sono riportati i 7 grafici dati restituiti dal codice precedente.

4.1.1 Carta e Cartone



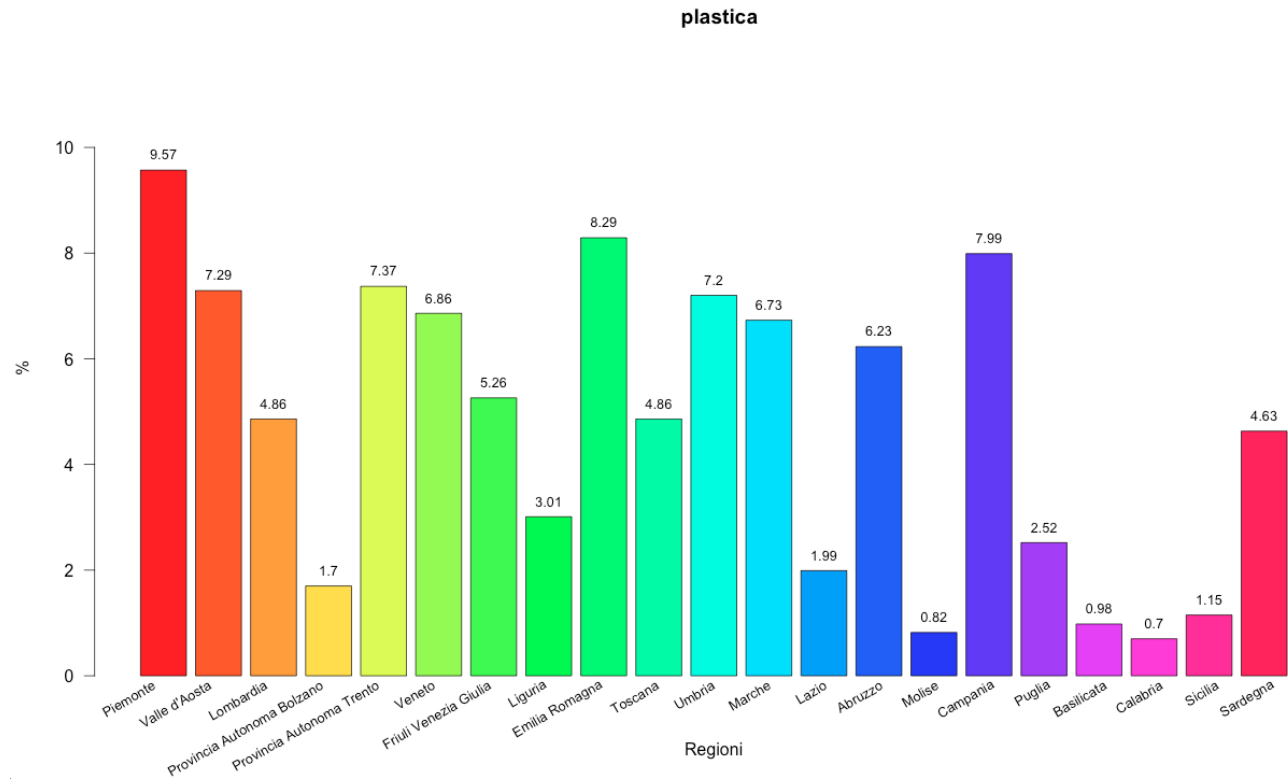
Dal grafico risulta che l'Emilia Romagna è la regione in cui gli abitanti riciclano la maggior quantità di carta e cartone, con il 7.84%, seguita dalla Provincia Autonoma del Trento, con il 6.85%, e dall'Umbria con il 6.64%. Mentre nelle ultime posizioni abbiamo il Molise con l'1.17%, la Sicilia con l'1.65% e la Calabria con l'1.71%.

4.1.2 Vetro



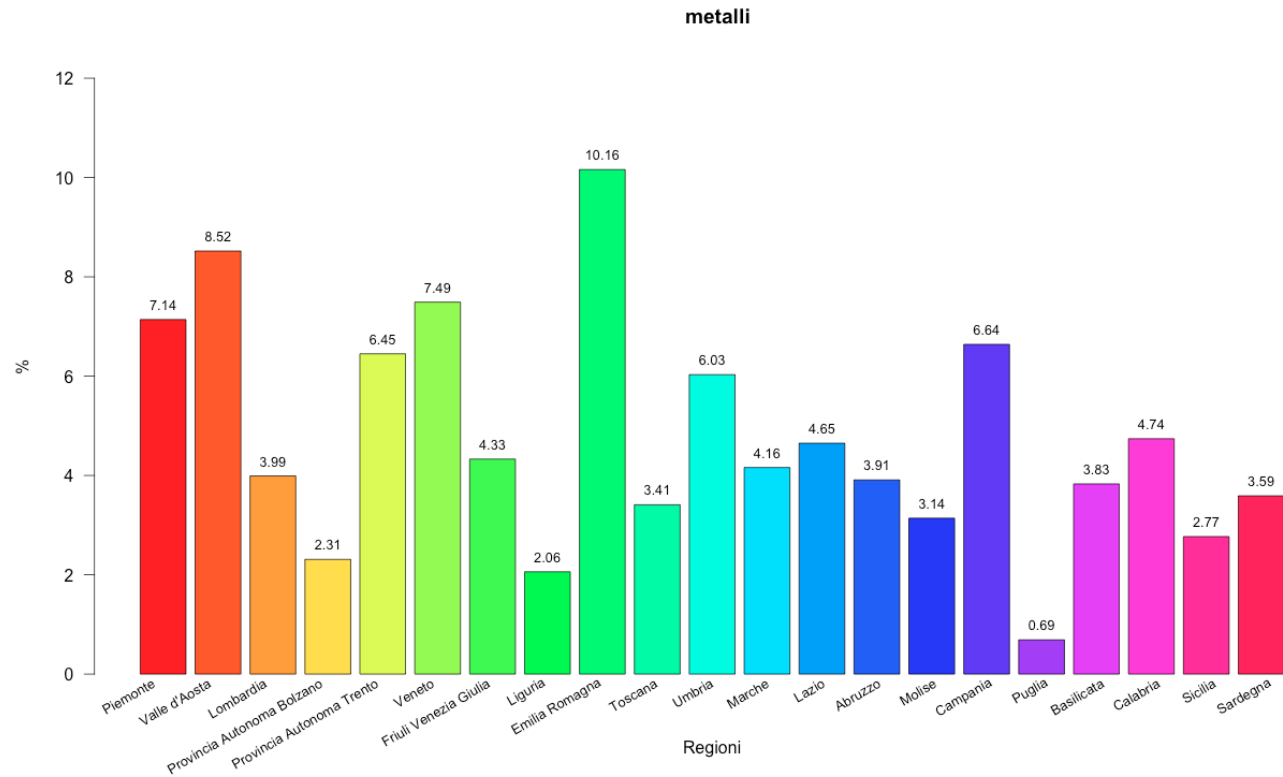
Dal grafico risulta che la Lombardia è la regione in cui gli abitanti riciclano la maggior quantità di vetro, con il 7.92%, subito seguita dal Veneto con il 7.89% e dalla Valle D'Aosta con il 7.75%. Mentre nelle ultime posizioni abbiamo la Calabria con lo 0.51% e dalla Sicilia con lo 0.89%.

4.1.3 Plastica



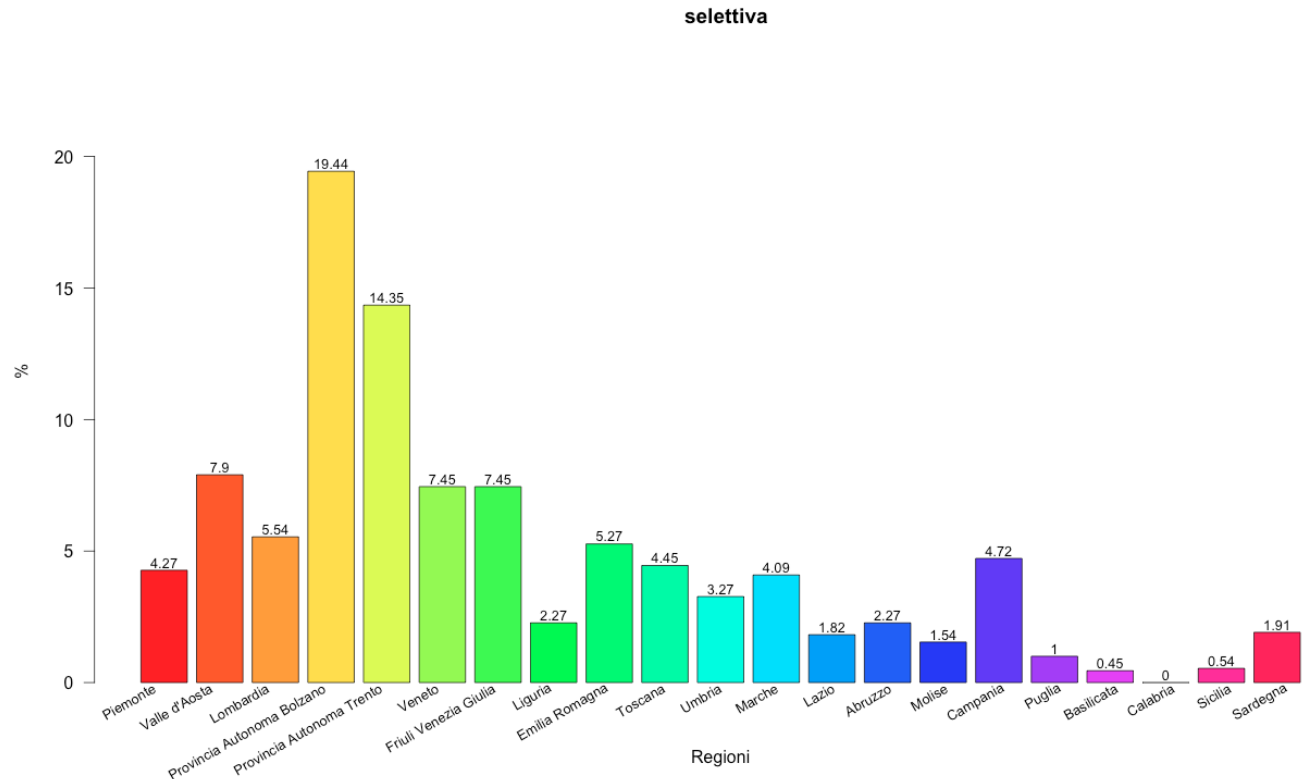
Dal grafico risulta che il Piemonte è la regione in cui gli abitanti differenziano la maggior quantità materiale plastico, con il 9.57%, seguita dall'Emilia Romagna, con l'8.29%, e la Campania con il 7.99%. Mentre nelle ultime posizioni abbiamo la Calabria con lo 0.70%, la Basilicata con lo 0.98% e la Calabria con l'1.15%.

4.1.4 Metalli



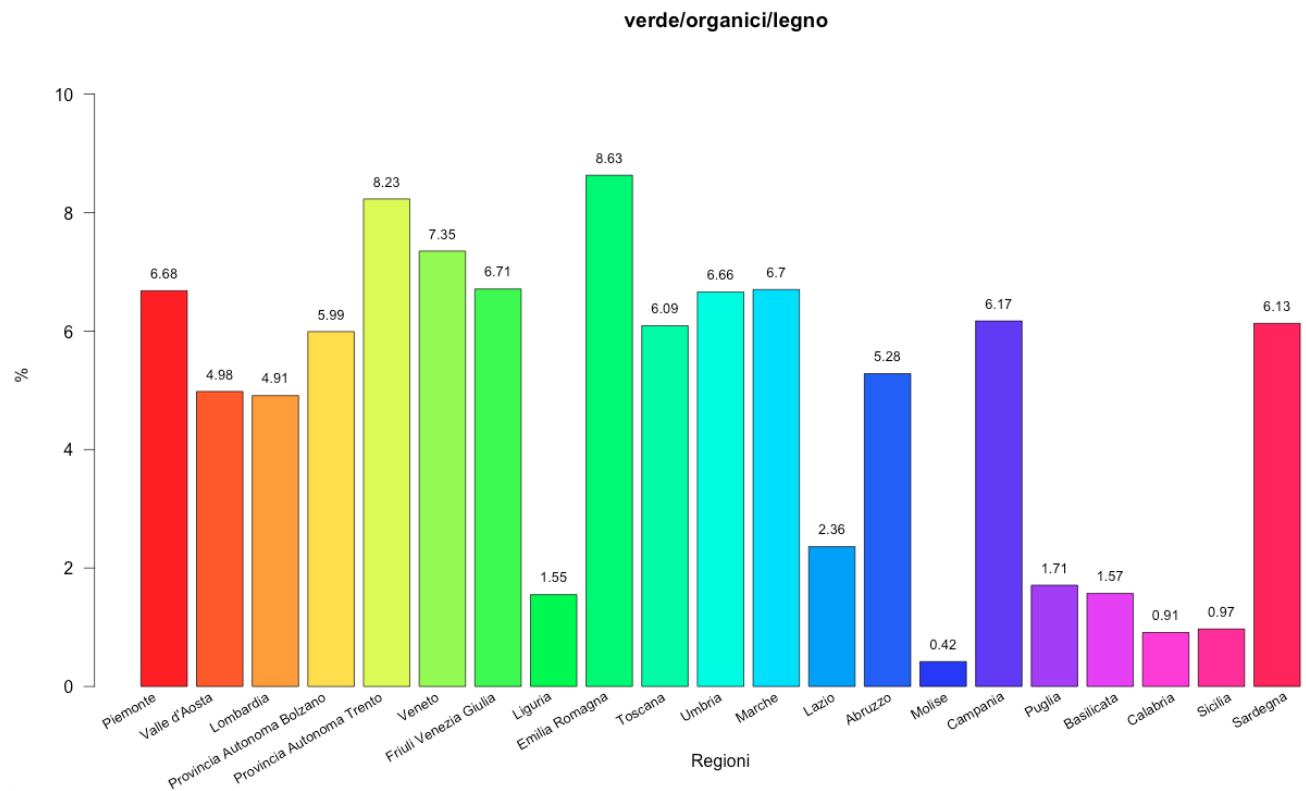
Dal grafico risulta che l'Emilia Romagna è la regione in cui gli abitanti differenziano la maggior quantità di metalli, con il 10.16%, seguita dalla Valle D'Aosta, con l'8.52%, e il Veneto con il 7.49%. Mentre nelle ultime posizioni abbiamo la Puglia con lo 0.69%, la Liguria con il 2.06% e la Provincia Autonoma di Bolzano con il 2.31%.

4.1.5 Raccolta Selettiva



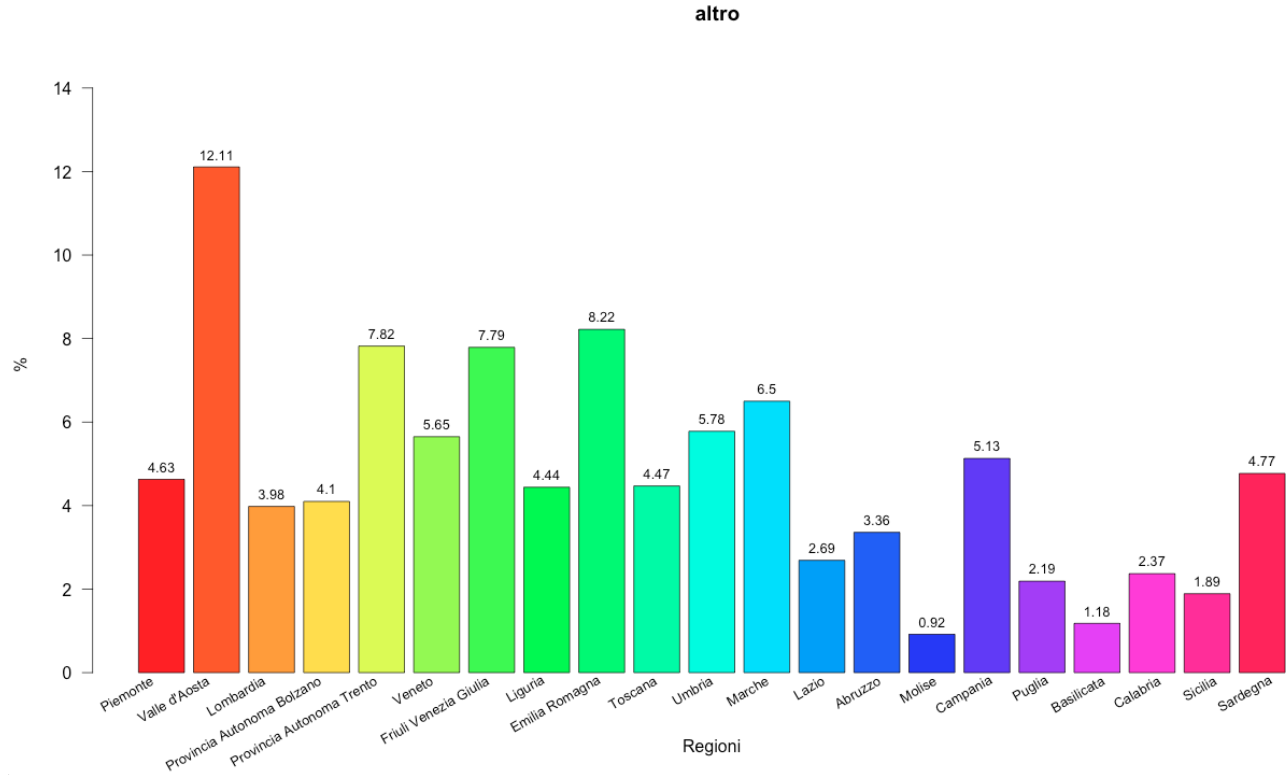
Dal grafico risulta che la Provincia Autonoma di Bolzano è la regione in cui gli abitanti differenziano la maggior quantità di raccolta selettiva, con il 19.44%, seguita dalla Provincia Autonoma del Trento, con il 14.35%, e la Valle D'Aosta con il 7.90%. Mentre nelle ultime posizioni abbiamo la Calabria con lo 0%, probabilmente perchè non viene effettuato questo tipo di riciclaggio, la Basilicata con lo 0.45% e la Sicilia con l'0.54%.

4.1.6 Rifiuti verdi, organici e legno



Dal grafico risulta che l'Emilia Romagna è la regione in cui gli abitanti riciclano la maggior quantità di rifiuti verdi, organici e legno, con il 7.84%, seguita dalla Provincia Autonoma del Trento, con l'8.23%, e dal Veneto con il 7.35%. Mentre nelle ultime posizioni abbiamo il Molise con lo 0.42%, la Calabria con lo 0.91% e la Sicilia con lo 0.97%.

4.1.7 Altro



Dal grafico risulta che la Valle D'Aosta è la regione in cui gli abitanti riciclano la maggior quantità di altri rifiuti non indicati, con il 12.11%, seguita dall'Emilia Romagna, con l'8.22%, e dalla Provincia Autonoma del Trento con il 7.82%. Mentre nelle ultime posizioni abbiamo il Molise con lo 0.92%, la Basilicata con l'1.18% e la Sicilia con l'1.89%.

4.2 Rappresentazione dei valori numerici

Oltre alla rappresentazione dei valori percentuali è stata effettuata anche un'analisi sui corrispondenti valori numerici sia in base alla frequenza relativa che a quella assoluta.

Consideriamo una variabile X ed indichiamo con x_1, x_2, \dots, x_k le modalità distinte da essa assunte, consideriamo poi un campione di n osservazioni di X . Se indichiamo con n_i il numero di volte in cui ciascuna modalità x_i è presente nel campione, ossia, la frequenza assoluta con la quale appare nel campione, l'insieme $x_i, n_i, i = 1, 2, \dots, k$ si chiama distribuzione di frequenza.

Se non esistono dati mancanti, la somma delle frequenze assolute è sempre uguale alla numerosità del campione, ovvero $n = n_1 + n_2 + \dots + n_k$. La frequenza relativa, invece, risulta essere il rapporto tra la frequenza assoluta e la numerosità del campione, ovvero $f_i = n_i / n$.

Se non esistono dati mancanti la somma delle frequenze relative è sempre unitaria, ovvero $f_1 + f_2 + \dots + f_n = 1$.

Nelle successive sezioni saranno presentati i grafici relativi alla distribuzione di frequenza assoluta e relativa per ogni indicatore, per la prima sono stati utilizzati grafici a barre creati tramite il comando `barplot`, per la seconda grafici a bastoncino creati tramite il comando `plot`.

Il codice utilizzato è il seguente:

```

#grafici a barra sui chili pro capite
> for(i in 1:7){
>   valPerc<-max(matriceAnalisi[,i])
>   livelloPercentuale<-valPerc+(valPerc/4.5)
>   livelloLabel<-(valPerc/50)
>   bptl7 <- barplot(matriceAnalisi[,i],main=paste(labelRifiuti[i],
(frequenza assoluta)",... = ""), xlab="Regioni",ylab="Kg pro capite",
  names.arg="", ylim=c(0,livelloPercentuale),las=2,col=rainbow(21,s=0.9))
>   text(seq(0.5,25,by=1.22), par("usr")[3], srt = 30, adj = 1, xpd = TRUE,
  labels = paste(rownames(matriceAnalisi)),cex=0.8)
>   text(x=bptl7, y=matriceAnalisi[,i]+livelloLabel, labels=matriceAnalisi[,i],
  xpd=TRUE, cex=0.8)
> }

#grafici a bastoncini su frequenze relative di ogni singolo rifiuto
> for(i in 1:7){
>   valPerc<-max(matriceRelativa[,i])
>   livelloPercentuale<-valPerc+(valPerc/4.5)
>   livelloLabel<-(valPerc/50)
>   frequenzeRelative<-round(prop.table(matriceAnalisi[,i]),2)
>   plot(as.table(frequenzeRelative),main=paste(labelRifiuti[i],
(frequenza relativa)",... = ""),xlab="Regioni",ylab="",xaxt="n",
  ylim=c(0,livelloPercentuale),las=2,col=rainbow(21,s=0.9))
>   axis(1,at=1:22,labels=FALSE)
>   text(seq(1,21,by=1), par("usr")[3]-0.0028, srt = 30, adj = 1, xpd = TRUE,
  labels = paste(rownames(matriceRelativa)),cex=0.8)

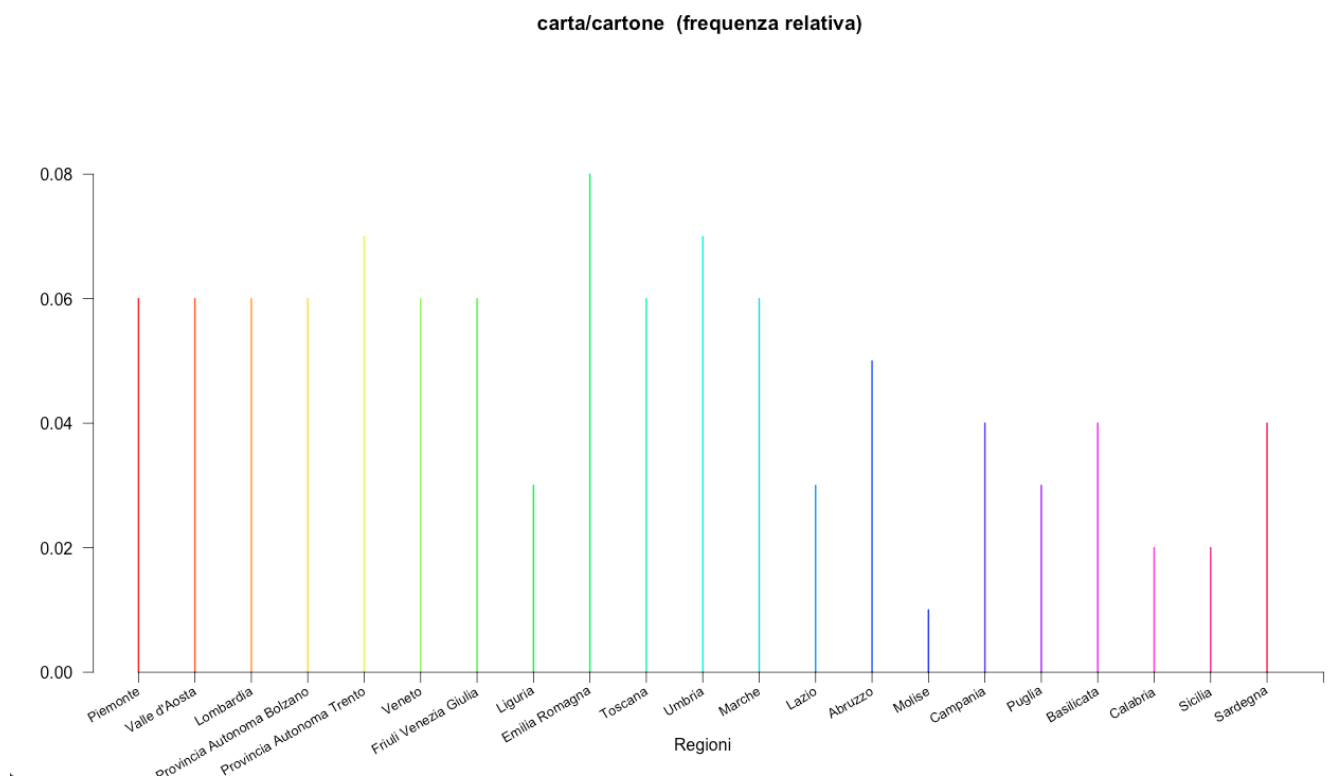
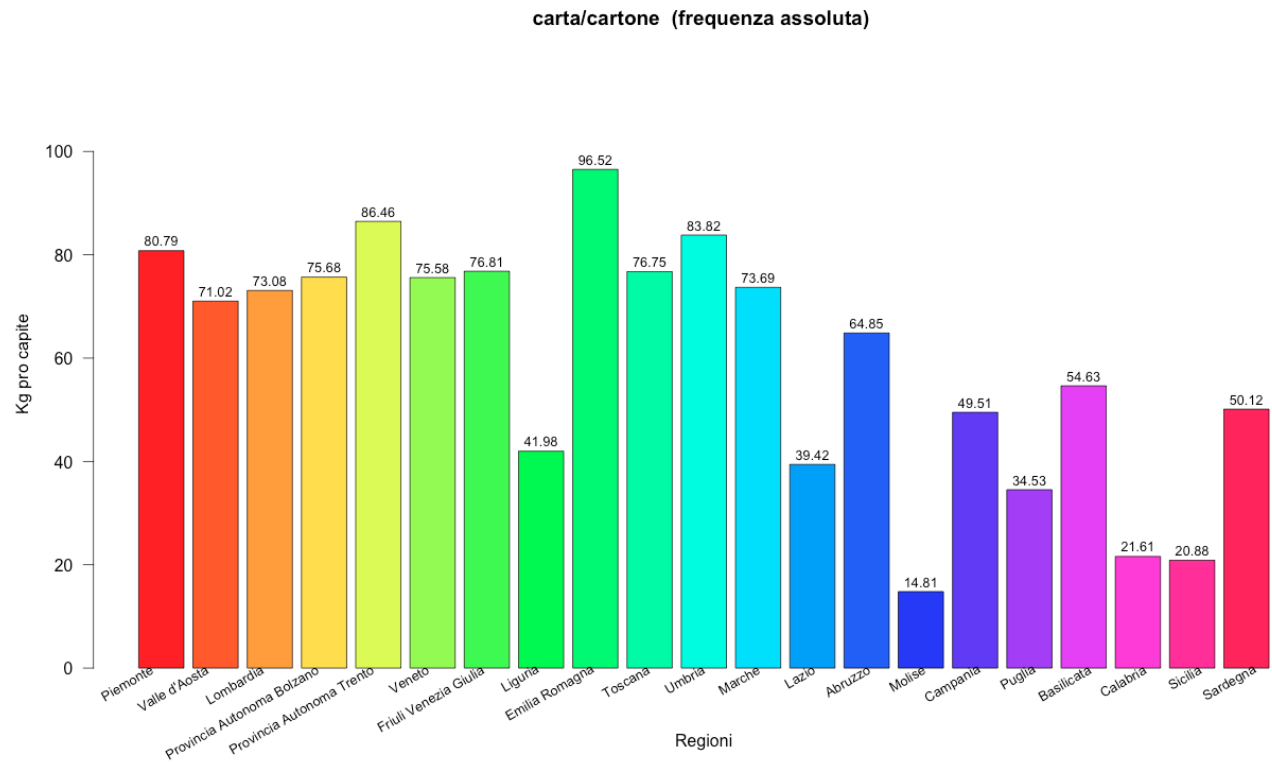
```

Sia la funzione che crea il grafico a barre sulle frequenze assolute che la funzione che crea il grafico a bastoncino per le frequenze relative, sono state inserite all'interno di un ciclo for, poichè per tutti i tipi di rifiuti erano da applicare entrambe le funzioni.

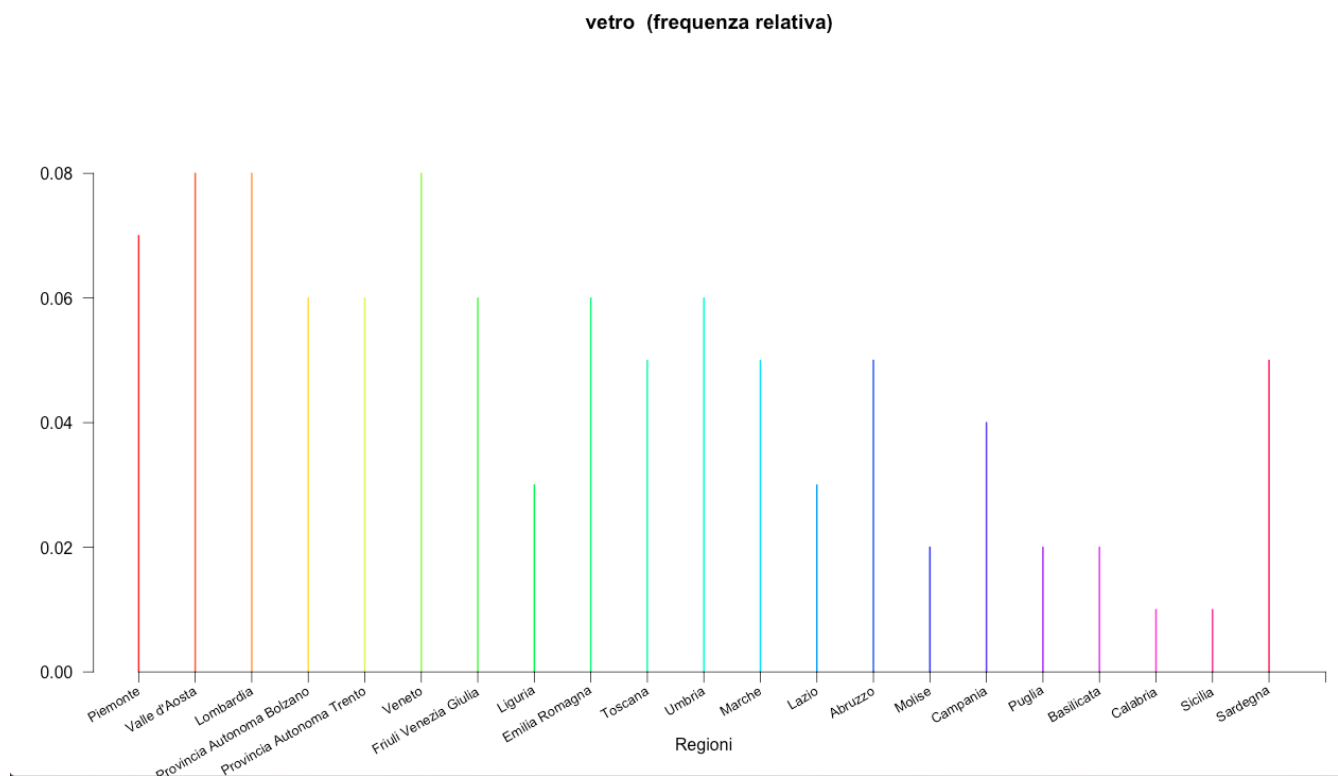
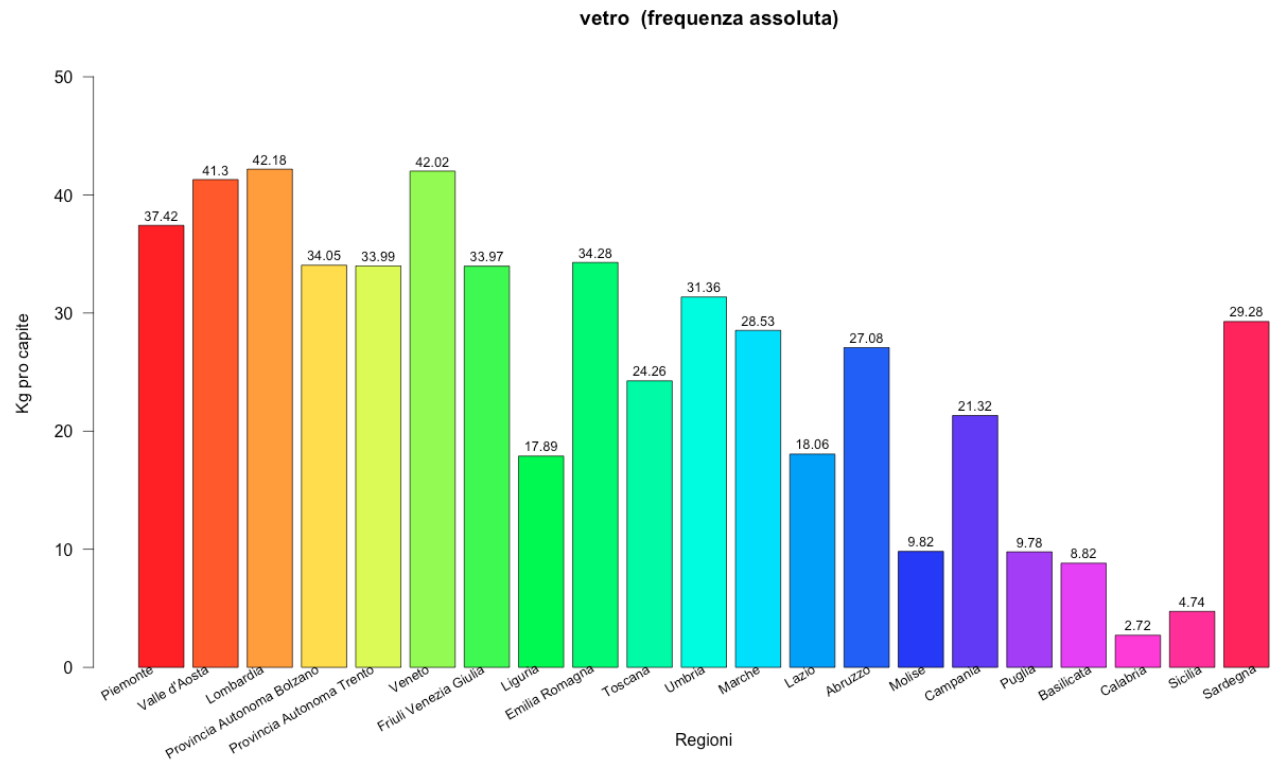
Sia nella funzione barplot che nella funzione plot, sono state modificate le variabili in modo da poter essere compatibili con qualsiasi grafico dei rifiuti. Ad esempio in entrambe le funzioni è stata usata una variabile chiamata livelloPercentuale per settare in modo opportuno, il valore di ylim, che rappresenta la lunghezza dell'asse y. Per fare in modo che questo valore fosse adatto per tutti i valori relativi ai vari tipi di rifiuti, sono state effettuate numerose prove, finchè non è stata trovata una funzione che era adatta a tutti i valori.

Di seguito vengono riportati i risultati ottenuti.

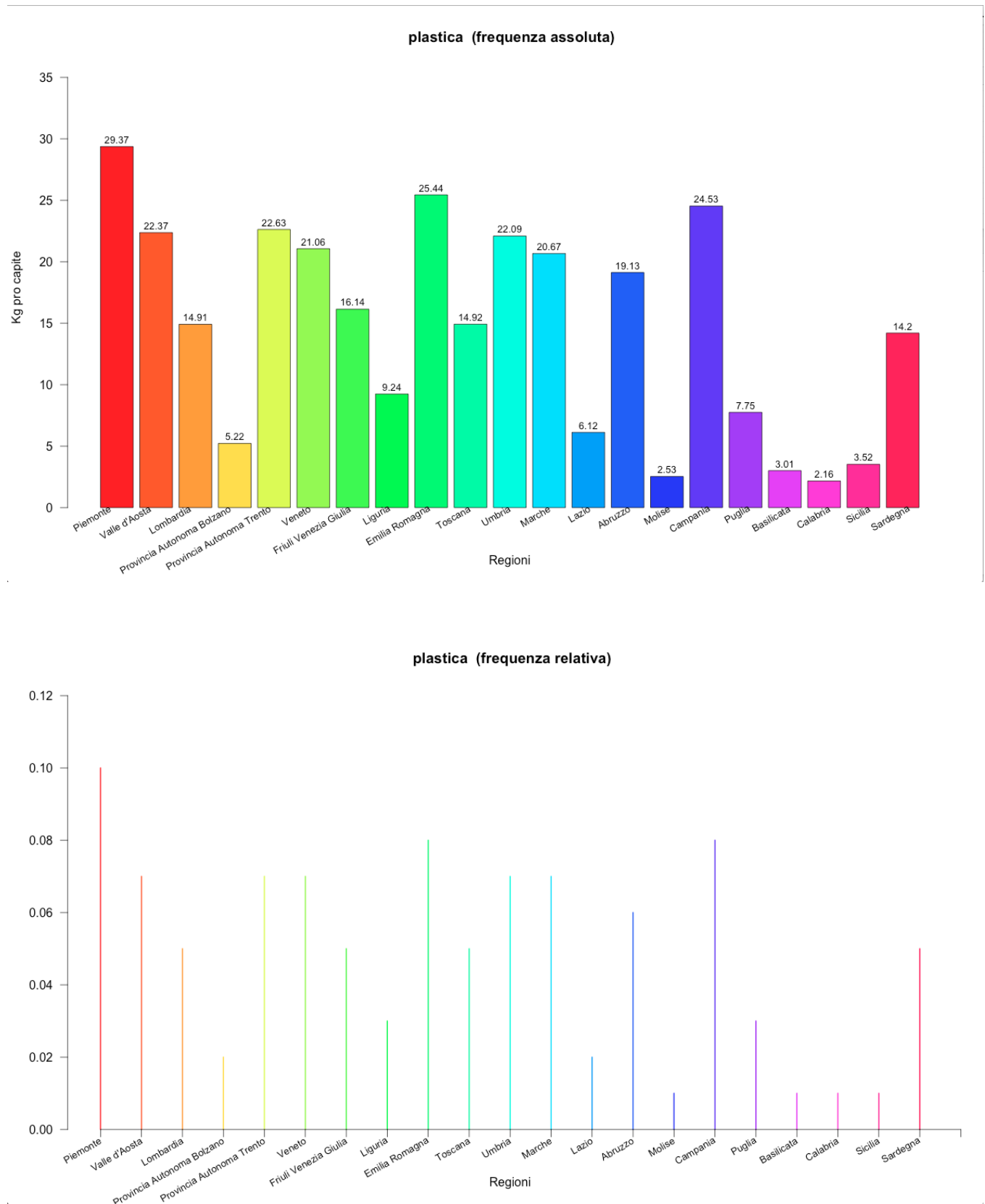
4.2.1 Carta e Cartone



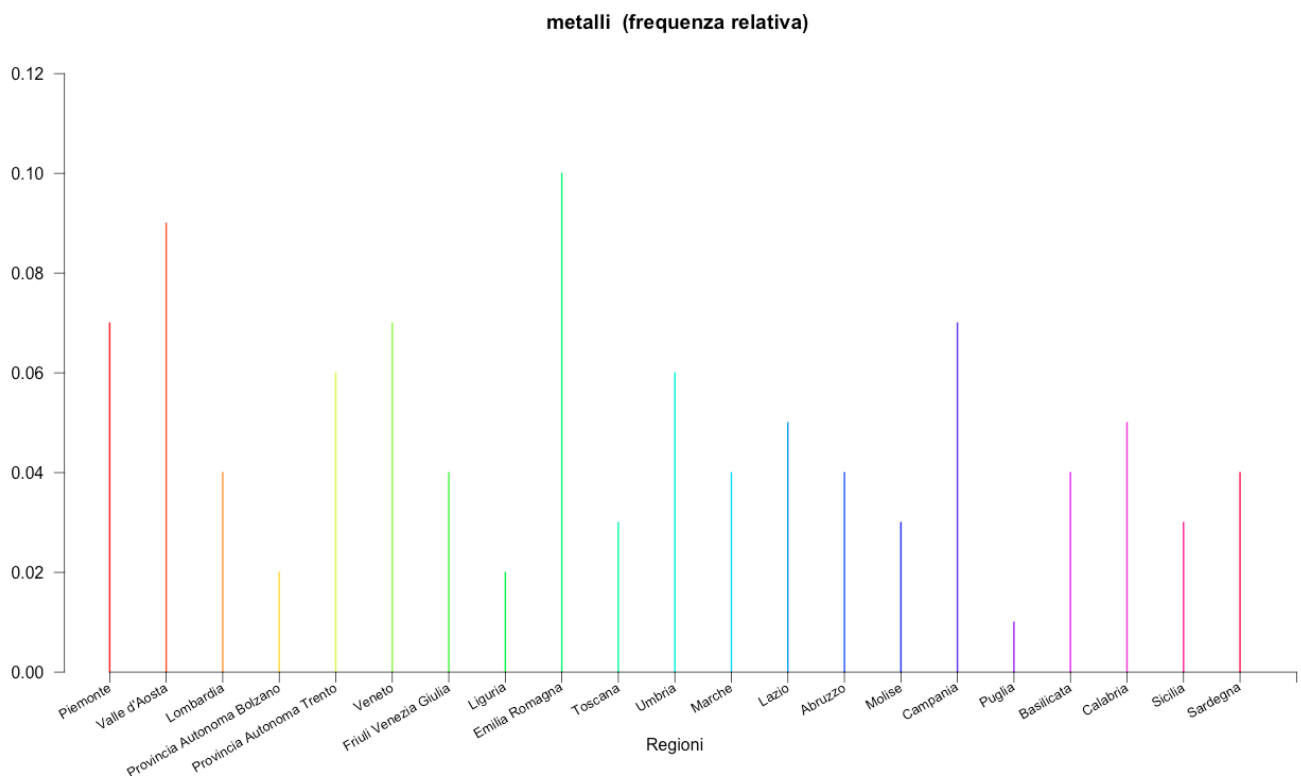
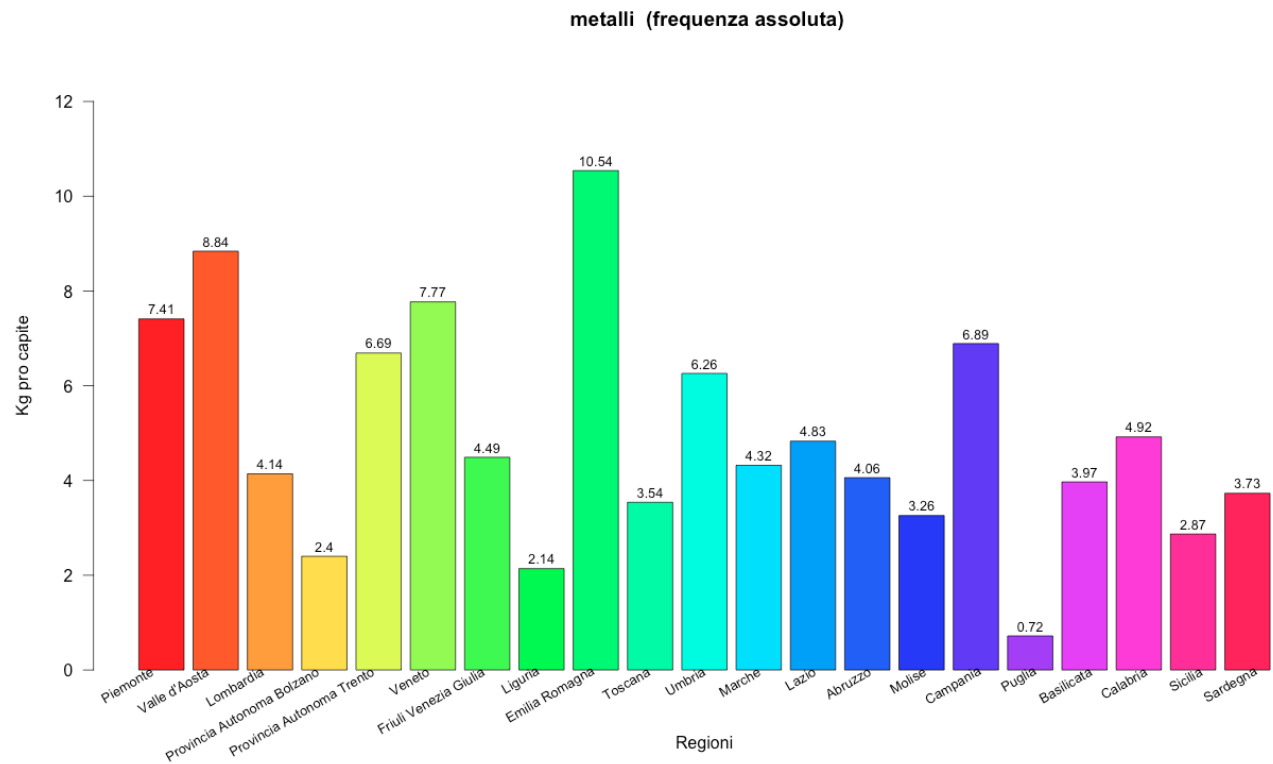
4.2.2 Vetro



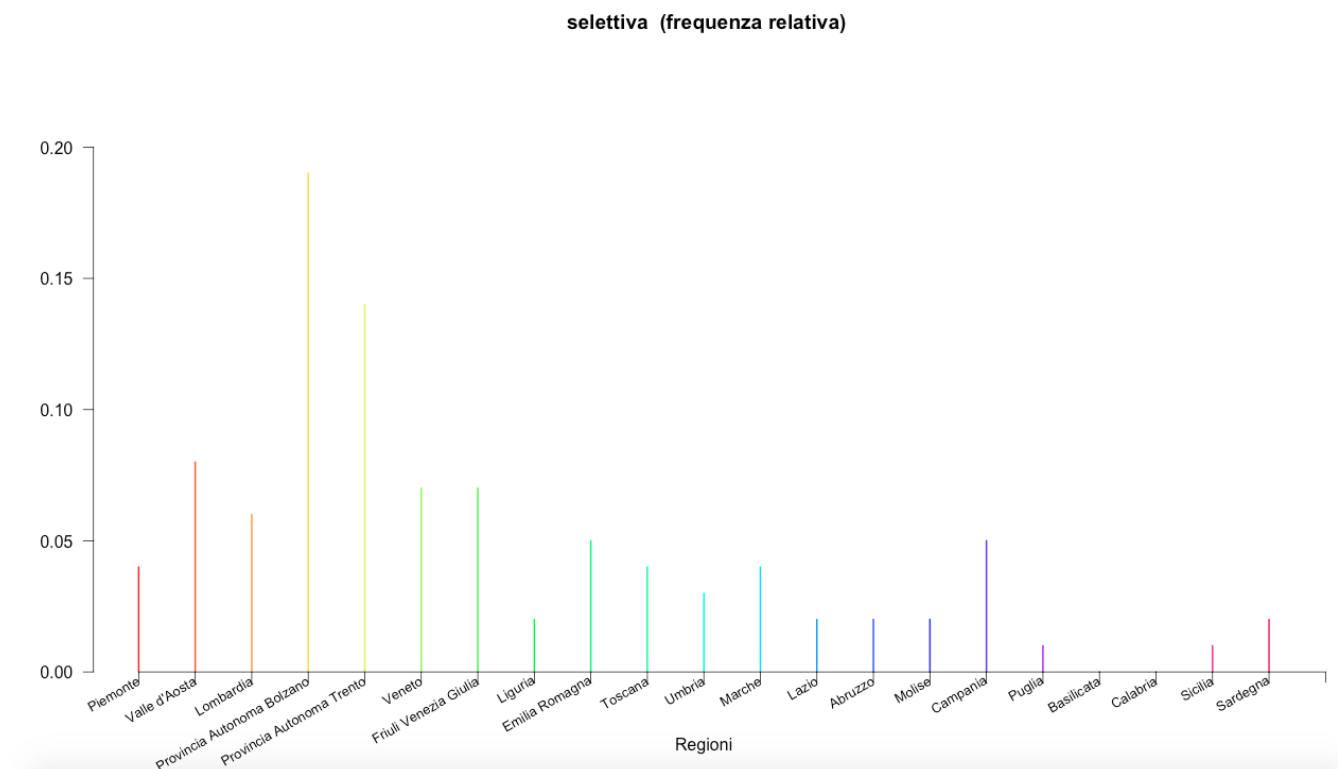
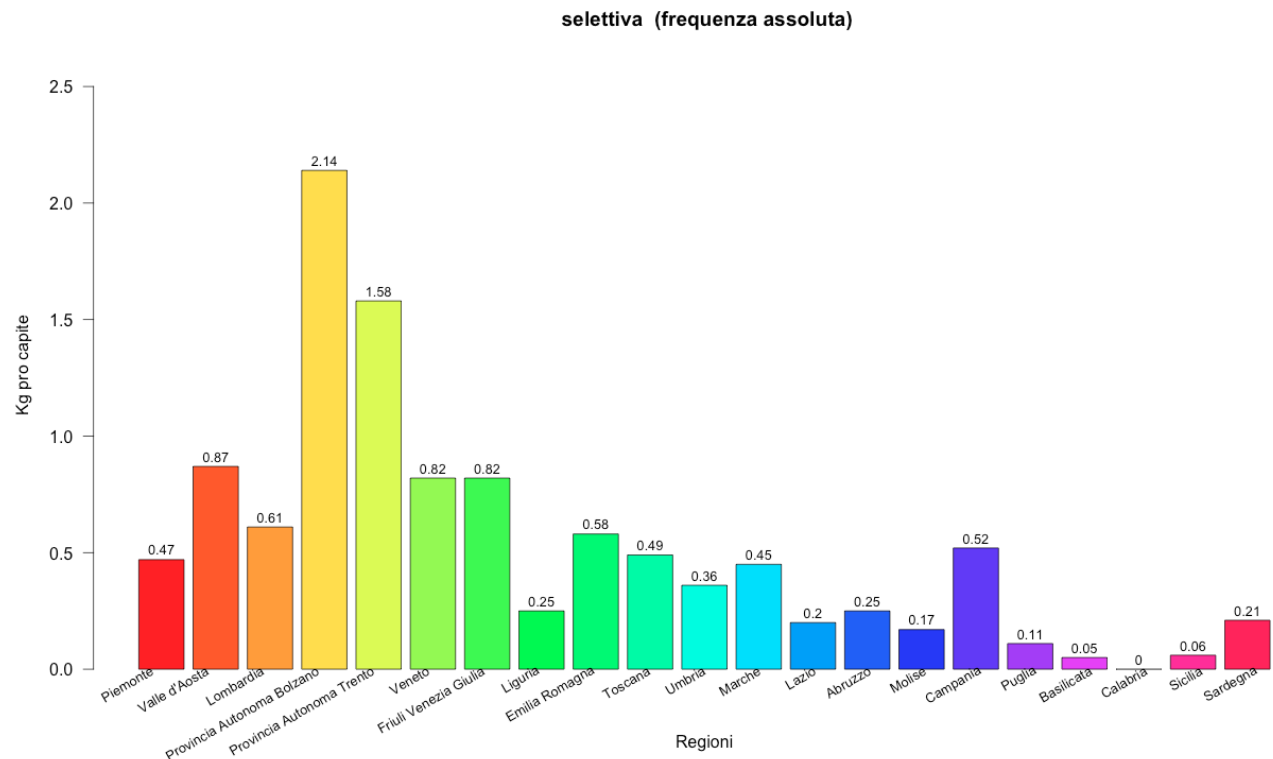
4.2.3 Plastica



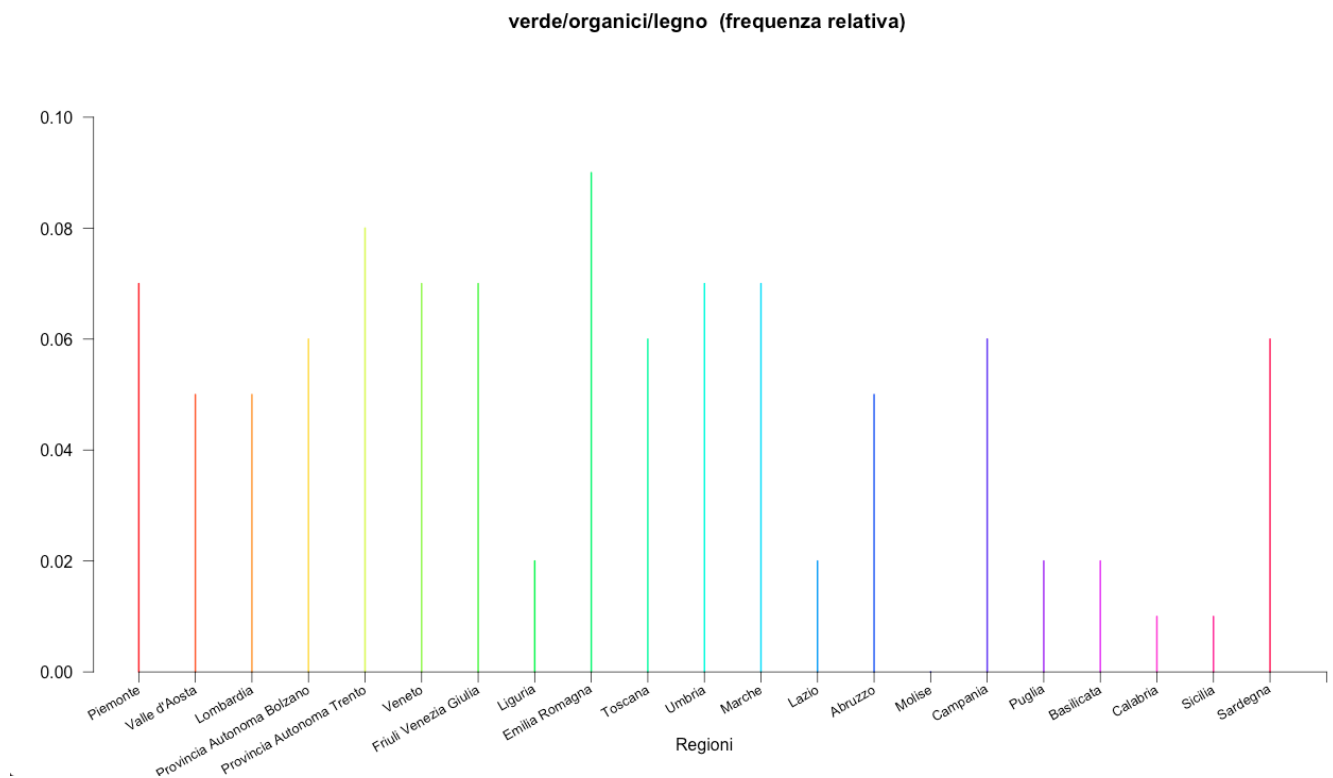
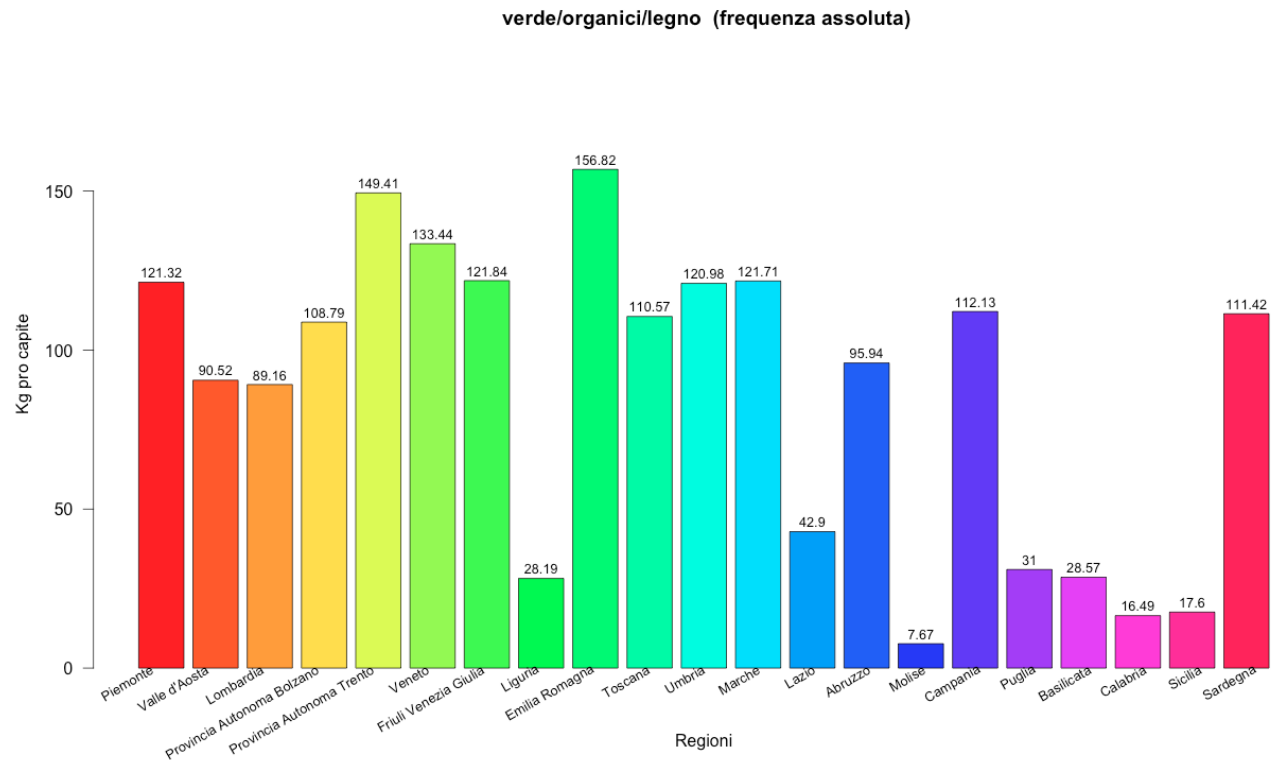
4.2.4 Metalli



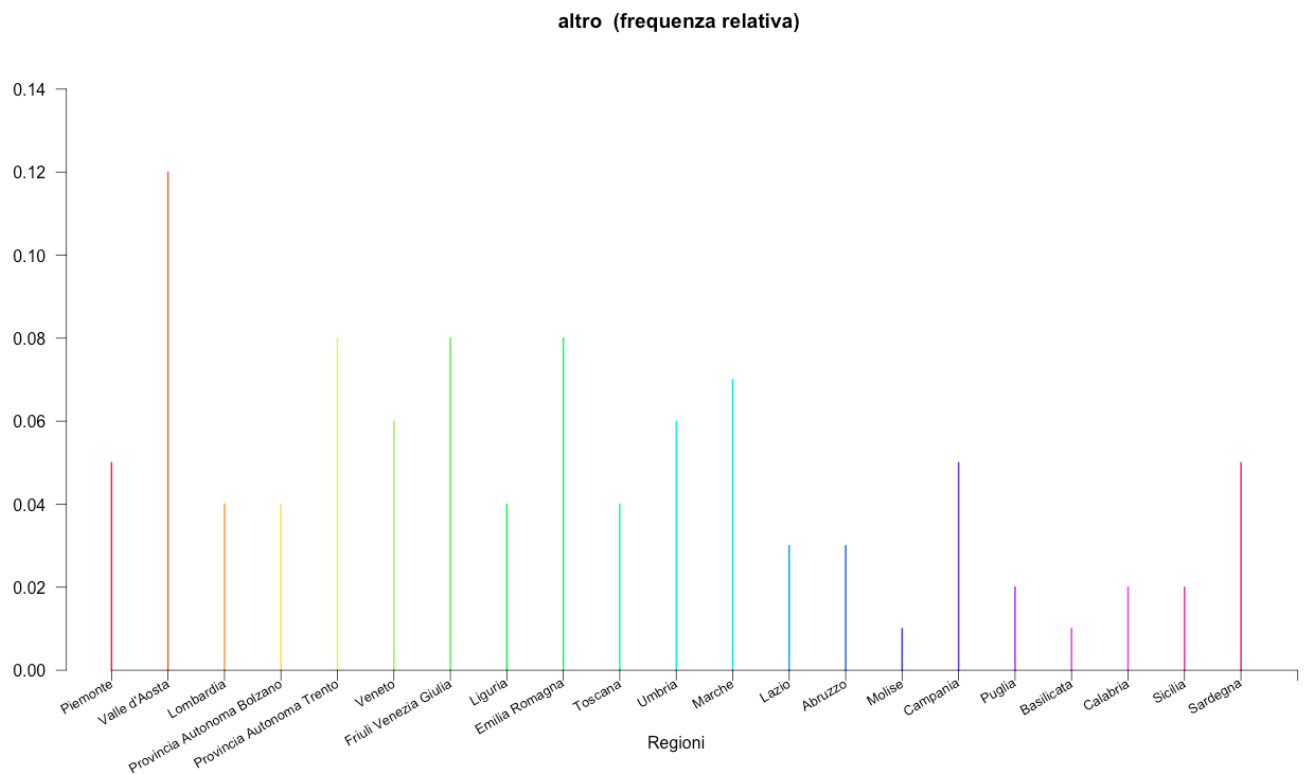
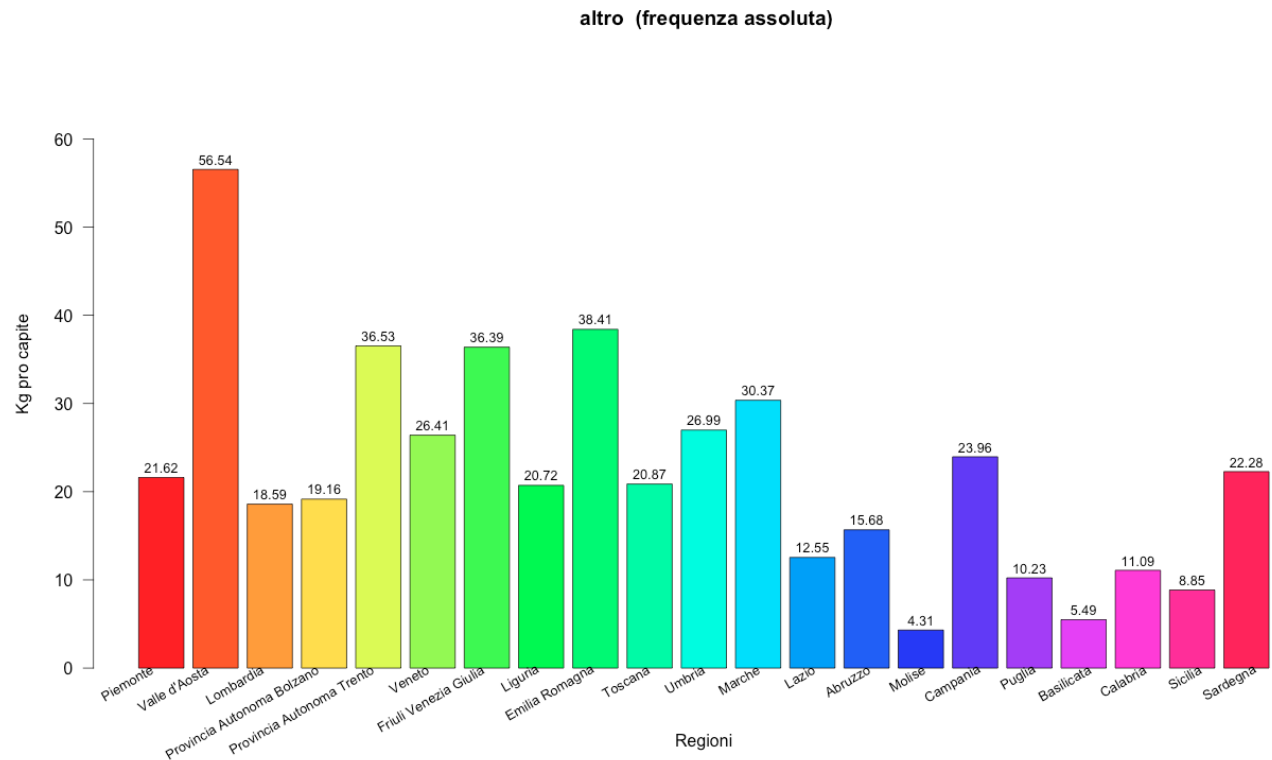
4.2.5 Raccolta Selettiva



4.2.6 Rifiuti verdi, organici e legno



4.2.7 Altro



4.3 Confronto tra Nord, Centro e Sud

In precedenza sono stati raggruppati i dati relativi alle regioni in modo da suddividere il tutto in aree geografiche, quali Nord, Centro e Sud in modo da ottenere una visione generale della situazione in Italia.

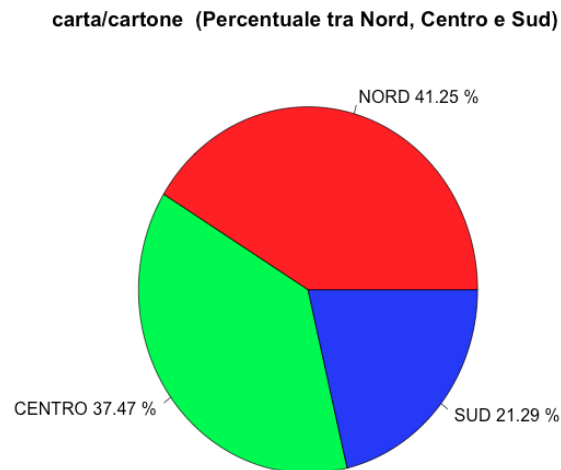
Dal momento che le categorie da analizzare sono poche, si è scelto di rappresentare tali dati utilizzando dei grafici a torta.

Il codice utilizzato per generare i 7 grafici è il seguente:

```
#grafici a torta sulle precentuali
> for(i in 1:7){
>   pie(matriceRipartitaPerc[,i], labels=paste(rownames(matriceRipartitaPerc),
matriceRipartitaPerc[,1],"%"),
main = paste(labelRifiuti[i], " (Percentuale tra Nord, Centro e Sud)",... = ""),
col=rainbow(3,s=0.9))
> }
```

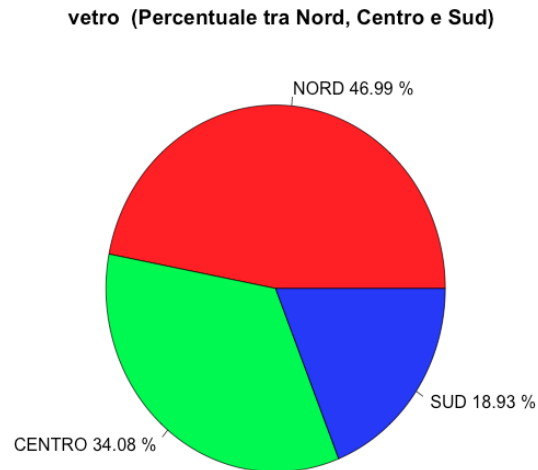
I risultati ottenuti da questo codice vengono riportati di seguito.

4.3.1 Carta e Cartone



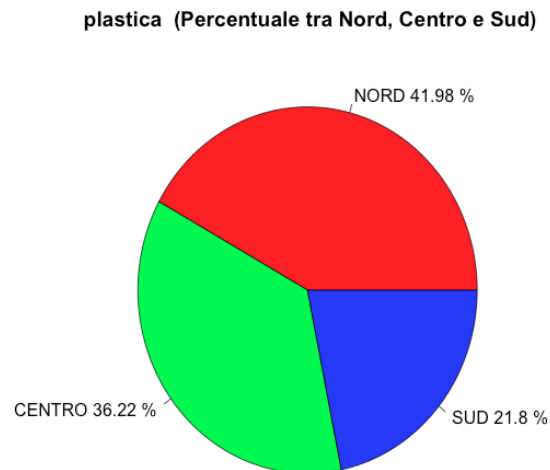
Dal grafico si può vedere in modo marcato che al Sud si differenzia di meno la carta e il cartone, mentre tra Centro e Nord la differenza è minima.

4.3.2 Vetro



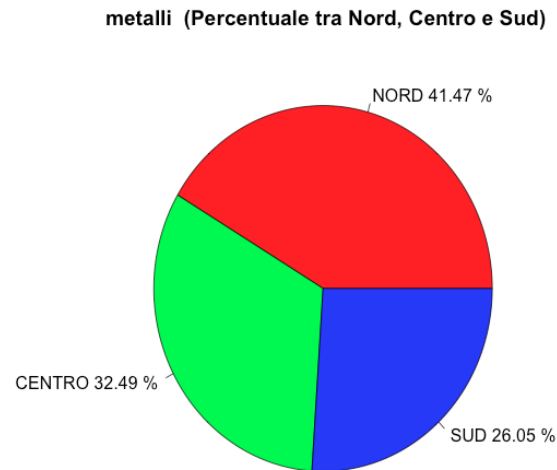
Dal grafico risulta che la quantità maggiore di vetro viene differenziata al Nord, mentre a Sud la quantità minore.

4.3.3 Plastica



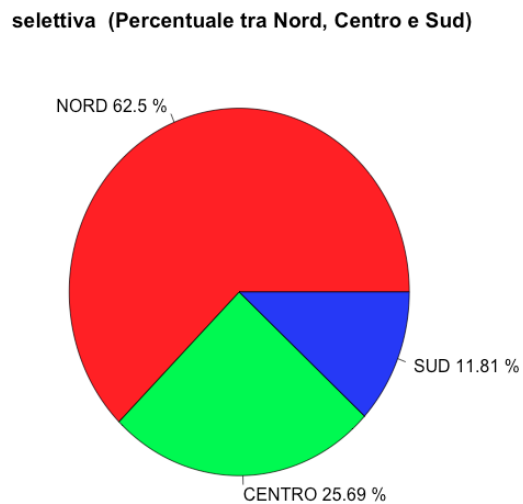
Dal grafico si nota che al Sud, come nei casi precedenti, si differenzia la minor quantità di plastica, mentre al Nord la quantità maggiore.

4.3.4 Metalli



Dal grafico si nota, che il Sud differenzia la quantità minore di metalli, ed il Nord la quantità maggiore, ma a differenza degli altri rifiuti visti in precedenza, abbiamo dei valori ravvicinati.

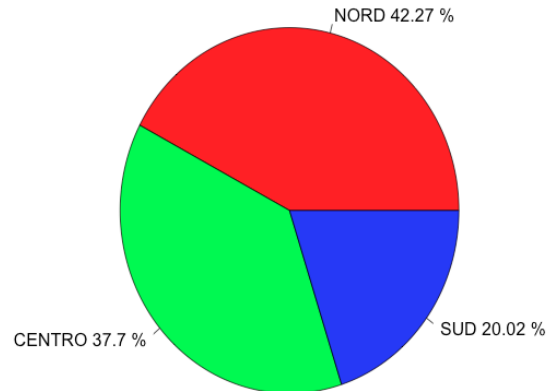
4.3.5 Raccolta Selettiva



Dal grafico si nota in maniera molto marcata, che il Nord effettua la maggior quantità di raccolta selettiva, con il 62.5%, mentre il sud la quantità minore, con solo l'11.82%.

4.3.6 Rifiuti verdi, organici e legno

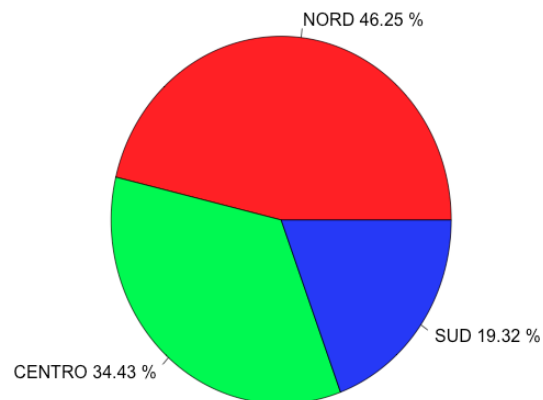
verde/organici/legno (Percentuale tra Nord, Centro e Sud)



Dal grafico risulta che la quantità maggiore di rifiuti verdi, organici e legno viene differenziata al Nord, mentre a Sud la quantità minore.

4.3.7 Altro

altro (Percentuale tra Nord, Centro e Sud)



Dal grafico risulta che la quantità maggiore di altri rifiuti non riportati viene differenziata al Nord, mentre a Sud la quantità minore.

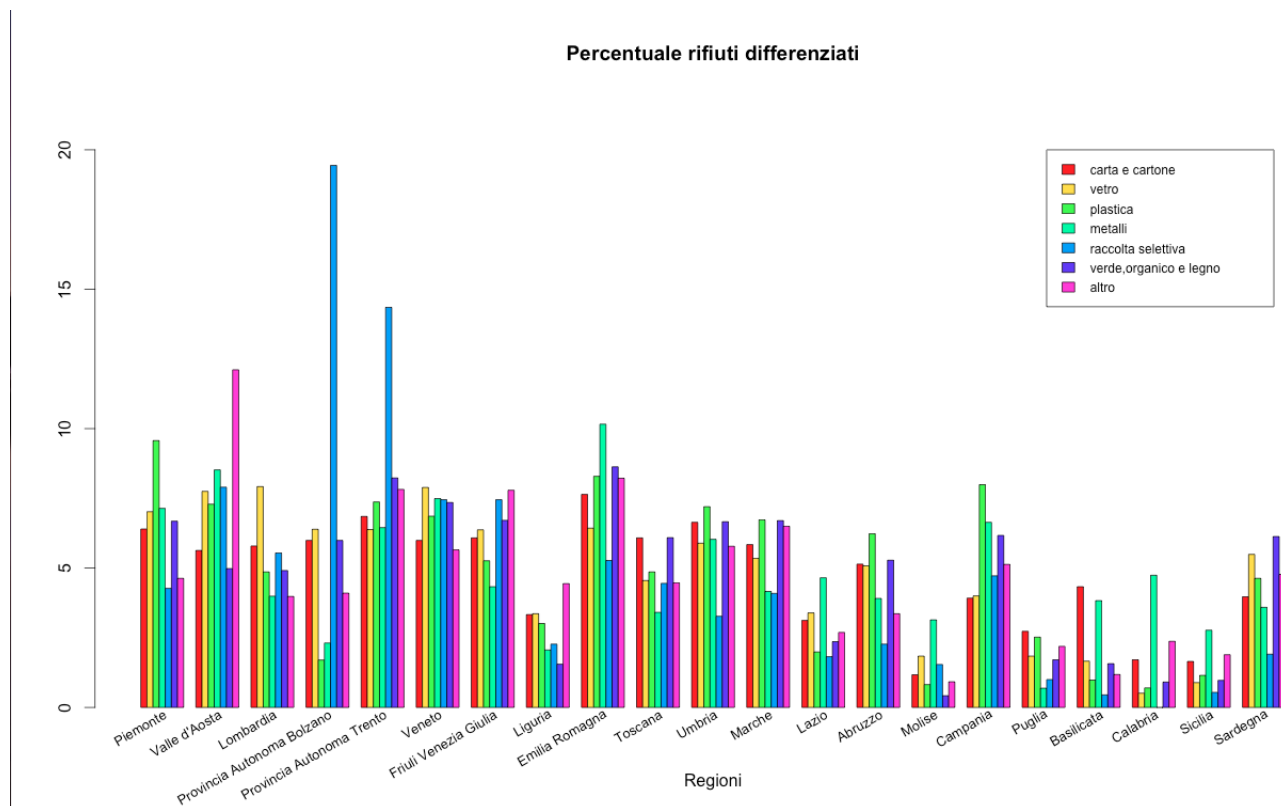
4.4 Variabili a confronto

In questa sezione è stato fatto un confronto tra le variabili al fine di capire quali sono i rifiuti maggiormente riciclati da ogni regione.

Per fare ciò è stato compilato il seguente codice:

```
> barplot(t(matricePercentuale[,1:7]), beside=TRUE, space = c(0, 2),
  main="Percentuale rifiuti differenziati", xlab = "Regioni", ylim=c(0,22),
  xaxt = "n", col=rainbow(7,s=0.9))
> legend(x=150,y=20, legend = c(labelRifiuti2[1:7]), fill = rainbow(7,s=0.9),
  cex = 0.75)
> text(seq(6,190,by=9), par("usr")[3]-0.2, srt = 30, adj = 1, xpd = TRUE,
  labels = rownames(matricePercentuale), cex=0.8)
```

Con il seguente risultato:



5 Statistica descrittiva con R

5.1 Indici di sintesi: posizione e dispersione

Gli indici di sintesi sono utili per descrivere dati numerici, in particolare ne prenderemo in considerazione due:

- **Indici di posizione:** rappresentano misure di centralità, ovvero sono utili per individuare attorno a quali valori è centrato l'insieme dei dati, in particolare utilizzeremo media, mediana, moda e quartili.
- **Indici di dispersione:** servono per quantificare la dispersione dei dati, in particolare utilizzeremo varianza, deviazione standard e coefficiente di variazione.

5.1.1 Indice di posizione

Definiamo di seguito con maggiore dettaglio gli indici di posizione prima introdotti.

Media e mediana campionaria

Supponiamo di avere un insieme x_1, x_2, \dots, x_n di n valori statistici detto campione o ampiezza di numerosità n . La media campionaria è la media aritmetica di questi valori e si denota con \bar{x} :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Prendiamo l'intervallo di ampiezza n considerato in precedenza ed ordiniamolo dal minore al maggiore. Se n è dispari si definisce mediana campionaria il valore in posizione $\frac{(n+1)}{2}$ mentre se n è pari la mediana campionaria è invece definita come la media aritmetica dei valori che occupano le posizioni $\frac{n}{2}$ e $(\frac{n}{2} + 1)$. Questa definizione della mediana campionaria assicura che lo stesso numero dei valori cada sia a sinistra che a destra della mediana stessa.

La media e la mediana sono due statistiche utili per descrivere misure di centralità dei dati, presentano però delle differenze e vanno usate opportunamente a seconda dei casi. La media campionaria, infatti, utilizza tutti i dati ed è influenzata in maniera sensibile da valori molto alti o bassi mentre la mediana dipende solo da uno o due valori centrali senza risentire dei dati che si trovano agli estremi. Inoltre, l'uso della mediana presenta lo svantaggio di dover riordinare i dati, il che non è richiesto nella media ma è da preferire a quest'ultima quando si desidera eliminare gli effetti di valori estremi molto diversi tra di loro.

Moda campionaria

La moda campionaria di un insieme di dati, se esiste, è il valore a cui corrisponde la massima frequenza. Se esistono più valori con frequenza massima ciascuno di essi è detto valore modale.

La moda campionaria è maggiormente utilizzata quando si trattano dati di tipo qualitativo per i quali non è possibile calcolare media e mediana. La moda campionaria può non essere utile quando i dati sono numerosi e per la maggior parte diversi tra di loro; in tali casi la moda può non esistere o essere lontana dal centro dell'insieme dei dati.

Quartili

Si definiscono quartili e si indicano con Q1, Q2 e Q3, i tre valori che dividono l'insieme dei dati ordinati in quattro parti uguali:

- **Q1** è il minimo valore osservato tale che almeno il 25% (1/4) dei dati è minore o uguale a questo;
- **Q2** è il minimo valore osservato tale che almeno il 50% (1/2) dei dati è minore o uguale a questo e coincide con la mediana;
- **Q3** è il minimo valore osservato tale che almeno il 75% (3/4) dei dati è minore o uguale a questo;

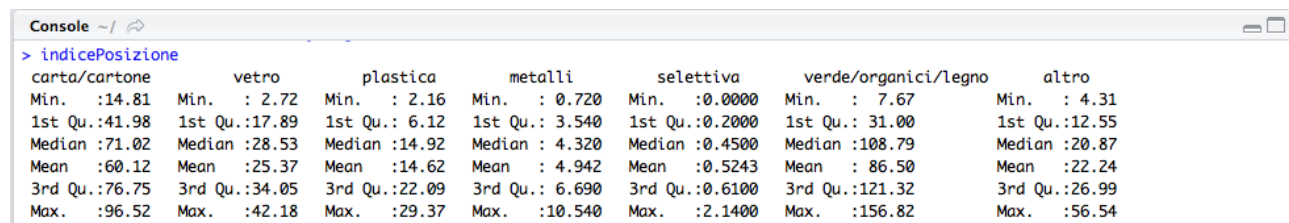
5.1.2 Calcolo degli indici di posizione con R

In R è possibile calcolare tutti gli indici di posizione (tranne la moda) utilizzando la funzione `summary()`.

Per fare ciò è stato eseguito il seguente codice:

```
> indicePosizione<-summary(matriceAnalisi)
```

Con il seguente risultato:



```

> indicePosizione
carta/cartone      vetro      plastica      metalli      selettiva      verde/organici/legno      altro
Min.   :14.81  Min.   : 2.72  Min.   : 2.16  Min.   : 0.720  Min.   :0.0000  Min.   : 7.67  Min.   : 4.31
1st Qu.:41.98  1st Qu.:17.89  1st Qu.: 6.12  1st Qu.: 3.540  1st Qu.:0.2000  1st Qu.: 31.00  1st Qu.:12.55
Median :71.02  Median :28.53  Median :14.92  Median : 4.320  Median :0.4500  Median :108.79  Median :20.87
Mean   :60.12  Mean   :25.37  Mean   :14.62  Mean   : 4.942  Mean   :0.5243  Mean   : 86.50  Mean   :22.24
3rd Qu.:76.75  3rd Qu.:34.05  3rd Qu.:22.09  3rd Qu.: 6.690  3rd Qu.:0.6100  3rd Qu.:121.32  3rd Qu.:26.99
Max.   :96.52  Max.   :42.18  Max.   :29.37  Max.   :10.540  Max.   :2.1400  Max.   :156.82  Max.   :56.54

```

Confrontando la media campionaria e la mediana campionaria è possibile conoscere la forma della distribuzione. Infatti, se queste misure sono uguali la distribuzione di frequenza è simmetrica; se la media campionaria è sensibilmente maggiore della mediana campionaria la distribuzione è più sbilanciata verso destra, se accade il contrario è sbilanciata verso sinistra.

5.1.3 Indici di dispersione

Per quanto utili, gli indici di posizione purtroppo non tengono conto della variabilità esistente tra i dati. Infatti esistono distribuzioni di frequenza che, pur avendo la stessa media campionaria, sono molto diverse tra loro.

Per sopperire a tale mancanza considerano una seconda categoria d'indici di sintesi ovvero gli indici di dispersione, che misurano la dispersione dei dati intorno alla media.

Varianza campionaria

Dato un insieme di dati numerici x_1, x_2, \dots, x_n si definisce varianza campionaria, e si denota con s^2 , la quantità:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

dove \bar{x} denota la media campionaria dei dati. In R è possibile calcolare la varianza campionaria di un vettore numerico v attraverso il comando `var(v)`.

Si noti che la varianza campionaria è una stima della varianza, più è grande e più i dati si discostano

dalla media.

Deviazione standard campionaria

Si definisce deviazione standard campionaria la radice quadrata della varianza campionaria, ovvero:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

In R è possibile calcolare la deviazione standard di un vettore numero v attraverso il comando `sd(v)`.

Coefficiente di variazione

Il coefficiente di variazione è utile per calcolare variazioni esistenti tra diversi campioni dei dati. Dato un insieme di dati numerici x_1, x_2, \dots, x_n si definisce coefficiente di variazione il rapporto tra la deviazione standard campionaria ed il modulo della media campionaria, ovvero:

$$CV = \frac{s}{|\bar{x}|}$$

In R non esiste una funzione specifica per il calcolo del coefficiente di variazione, ma può essere facilmente definita nel seguente modo:

```
> cv <- function(x) { sd(x) / abs(mean(x)) }
```

Utilizzeremo di seguito questa nuova funzione `cv` per il calcolo del coefficiente di variazione.

5.1.4 Calcolo degli indici di dispersione con R

di seguito sono stati calcolati tutti gli indici di dispersione relativi ad ogni rifiuto e per ognuno di questo è stato creato un vettore.

Di seguito viene riportato il codice utilizzato per effettuare questi calcoli:

```
#creo il vettore delle varianze
> vettoreVarianza<-c()
> for(i in 1:7)
> vettoreVarianza[i]<-var(matriceAnalisi[,i])

#creo il vettore della deviazione standard
> vettoreDS<-c()
> for(i in 1:7)
> vettoreDS[i]<-sd(matriceAnalisi[,i])

#creo il vettore dei coefficienti di variazione
> vettoreCV<-c()
> for(i in 1:7)
> vettoreCV[i]<-cv(matriceAnalisi[,i])
```

I dati raccolti, insieme alla media, sono stati inseriti all'interno di una matrice.

```
> matriceIndiceDispersione<-matrix(c(vettoreMedia, vettoreVarianza, vettoreDS,
  vettoreCV), nrow=7, ncol=4, byrow=FALSE)
> labelIndici <- c(" Media", " Varianza", " Deviazione Standard", " Coefficiente di v
> colnames(matriceIndiceDispersione)<-labelIndici
> rownames(matriceIndiceDispersione)<-labelRifiuti2[1:7]
```

Con il seguente risultato:

	Media	Varianza	Deviazione Standard	Coefficiente di variazione
carta e cartone	60.1209524	563.4062890	23.7361810	0.3948071
vetro	25.3747619	157.5115962	12.5503624	0.4946002
plastica	14.6195238	75.5058348	8.6894093	0.5943702
metalli	4.9423810	5.7436590	2.3965932	0.4849066
raccolta selettiva	0.5242857	0.2716757	0.5212252	0.9941625
verde,organico e legno	86.4985714	2303.1144229	47.9907744	0.5548158
altro	22.2400000	158.2803100	12.5809503	0.5656902

5.1.5 Boxplot

Il boxplot è utile per rappresentare graficamente le informazioni ottenute grazie agli indici descrittivi, consiste nel disegno di una scatola i cui estremi sono Q1 e Q3, tagliata da una linea orizzontale in corrispondenza di Q2 (ossia della mediana). Dalla scatola, inoltre, escono due segmenti che terminano in corrispondenza dei valori minimo e massimo ed è possibile anche verificare l'eventuale presenza di outlier, ovvero di valori che si discostano notevolmente dalla media e che sono rappresentati da piccoli cerchi.

Di seguito viene riportato il codice R utilizzato per costruire il boxplot, i valori degli outlier sono stati individuati salvando l'output del comando boxplot in una variabile, esaminandolo e trovando le corrispettive regioni.

```
> indiciDescrittivi <- boxplot(matriceAnalisi[,1:7], main="Boxplot",
  ylab="Numero", xlab="Frequenza", col=rainbow(7, s=0.9))
```

Di seguito eseguendo il comando

```
> indiciDescrittivi$out
```

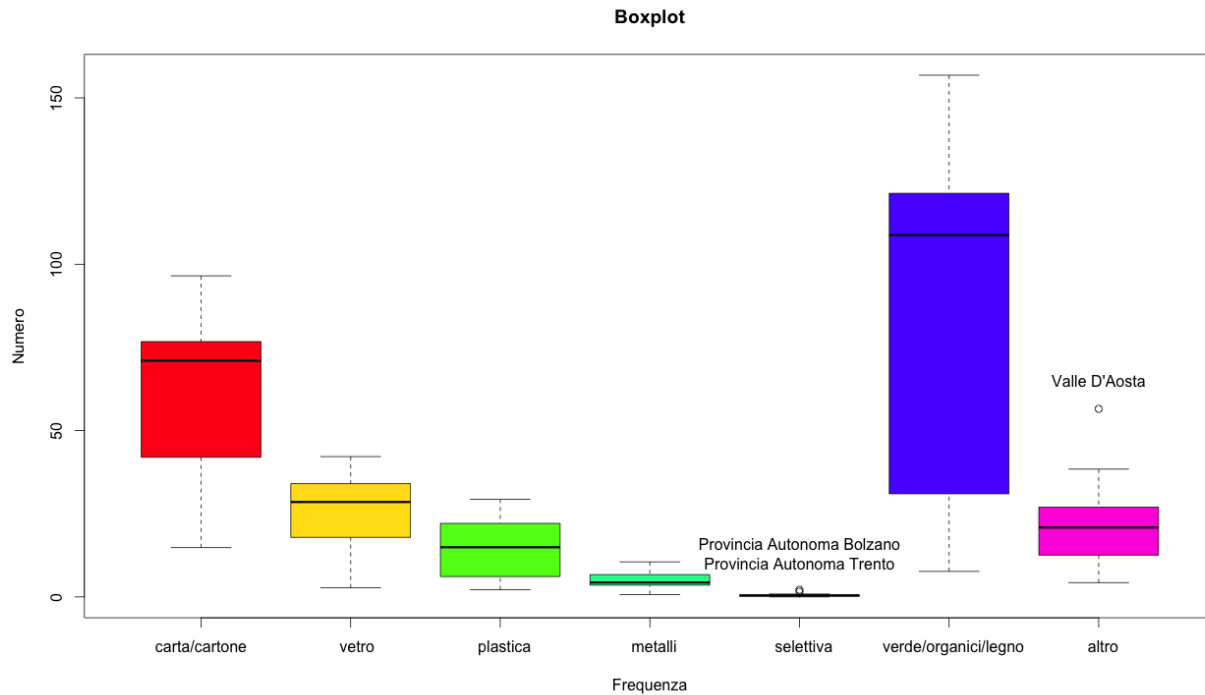
è possibile poter vedere i valori degli outlier che sono stati confrontanti con i dati in nostro possesso.

	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro
Piemonte	80.79	37.42	29.37	7.41	0.47	121.32	21.62
Valle d'Aosta	71.02	41.30	22.37	8.84	0.87	90.52	56.54
Lombardia	73.08	42.18	14.91	4.14	0.61	89.16	18.59
Provincia Autonoma Bolzano	75.68	34.05	5.22	2.40	2.14	108.79	19.16
Provincia Autonoma Trento	86.46	33.99	22.63	6.69	1.58	149.41	36.53
Veneto	75.58	42.02	21.06	7.77	0.82	133.44	26.41
Friuli Venezia Giulia	76.81	33.97	16.14	4.49	0.82	121.84	36.39
Liguria	41.98	17.89	9.24	2.14	0.25	28.19	20.72
Emilia Romagna	96.52	34.28	25.44	10.54	0.58	156.82	38.41
Toscana	76.75	24.26	14.92	3.54	0.49	110.57	20.87
Umbria	83.82	31.36	22.09	6.26	0.36	120.98	26.99
Marche	73.69	28.53	20.67	4.32	0.45	121.71	30.37
Lazio	39.42	18.06	6.12	4.83	0.20	42.90	12.55
Abruzzo	64.85	27.08	19.13	4.06	0.25	95.94	15.68
Molise	14.81	9.82	2.53	3.26	0.17	7.67	4.31
Campania	49.51	21.32	24.53	6.89	0.52	112.13	23.96
Puglia	34.53	9.78	7.75	0.72	0.11	31.00	10.23
Basilicata	54.63	8.82	3.01	3.97	0.05	28.57	5.49
Calabria	21.61	2.72	2.16	4.92	0.00	16.49	11.09
Sicilia	20.88	4.74	3.52	2.87	0.06	17.60	8.85
Sardegna	50.12	29.28	14.20	3.73	0.21	111.42	22.28

Una volta identificati gli outlier, sono state aggiunte le label con le corrispettive regioni all'interno del nostro grafico, utilizzando i seguenti comandi:

```
> text(5, 16.0, labels="Provincia Autonoma Bolzano")
> text(5, 10, labels="Provincia Autonoma Trento")
> text(7, 65, labels="Valle D'Aosta")
```

Il risultato finale è il seguente:



Come si può notare dal boxplot per quanto riguarda la raccolta selettiva i valori della Provincia Autonoma Del Trento e della Provincia Autonoma Di Bolzano si discostano dalla media, mentre per altri tipi di rifiuti sono i valori della Valle D'Aosta che si discostano dalla media. Ciò sta a significare che per la raccolta selettiva, le Province Autonome del Trento e di Bolzano, differenziano una quantità significativa in più rispetto alla media, mentre per altri tipi di rifiuto è la Valle D'Aosta.

5.2 Correlazioni tra le variabili

Spesso nelle indagini statistiche si osservano più variabili quantitative per uno stesso gruppo d'individui ed in tal caso è necessario vedere se esiste una correlazione tra esse. Date due variabili X e Y per verificare se esiste una dipendenza si può disegnare il diagramma di dispersione o scatterplot: si pongono in ascissa i dati relativi ad una delle due variabili (variabile indipendente) e in ordinata quelli relativi all'altra variabile (variabile dipendente).

Lo scatterplot serve ad evidenziare se i punti sono sparsi senza apparente regolarità oppure se ne esiste una (ossia le variabili sono connesse mediante una relazione lineare, quadratica, etc). Per ottenere una misura quantitativa della correlazione si possono inoltre usare:

- Covarianza campionaria;
- Coefficiente di correlazione campionario.

5.2.1 Covarianza campionaria

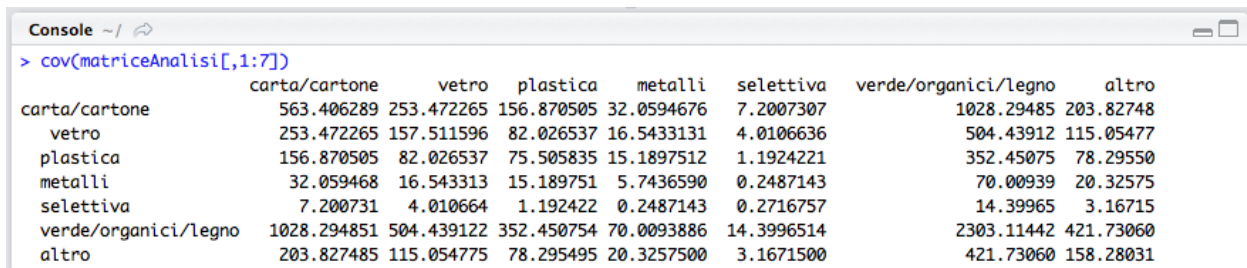
Per ottenere una misura quantitativa della correlazione tra le variabili si considera la covarianza campionaria. Assegnato un campione bivariato (x_i, y_i) per $i = 1, 2, \dots, n$ di una variabile quantitativa bidimensionale (X, Y) , siano \bar{x} e \bar{y} rispettivamente le medie campionarie di x_1, x_2, \dots, x_n e y_1, y_2, \dots, y_n . La covarianza campionaria tra le due variabili X e Y è così definita:

$$C_{xy} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}$$

La covarianza campionaria può avere segno positivo, negativo o nullo:

- Se $C_{xy} > 0$ si dice che le variabili sono correlate positivamente;
- Se $C_{xy} < 0$ si dice che le variabili sono correlate negativamente;
- Se $C_{xy} = 0$ si dice che le variabili sono non correlate.

In R la covarianza campionaria è calcolata attraverso il comando `cov()`.



```

> cov(matriceAnalisi[,1:7])

```

	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro
carta/cartone	563.406289	253.472265	156.870505	32.0594676	7.2007307	1028.29485	203.82748
vetro	253.472265	157.511596	82.026537	16.5433131	4.0106636	504.43912	115.05477
plastica	156.870505	82.026537	75.505835	15.1897512	1.1924221	352.45075	78.29550
metalli	32.059468	16.543313	15.189751	5.7436590	0.2487143	70.00939	20.32575
selettiva	7.200731	4.010664	1.192422	0.2487143	0.2716757	14.39965	3.16715
verde/organici/legno	1028.294851	504.439122	352.450754	70.0093886	14.3996514	2303.11442	421.73060
altro	203.827485	115.054775	78.295495	20.3257500	3.1671500	421.73060	158.28031

Come si può notare, le variabili dei rifiuti, sono tutte correlate positivamente.

5.2.2 Coefficiente di correlazione campionario

Date due variabili X e Y , il coefficiente di correlazione tra le due variabili è definito come:

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

Ovvero è il rapporto tra la covarianza di X e Y e il prodotto della deviazione standard di X indicata con s_x e la deviazione standard di Y indicata con s_y . Il coefficiente di correlazione campionario ha lo stesso segno della covarianza campionaria e può positivo, negativo o nullo:

- Se $r_{xy} > 0$ si dice che le variabili sono correlate positivamente;
- Se $r_{xy} < 0$ si dice che le variabili sono correlate negativamente;
- Se $r_{xy} = 0$ si dice che le variabili non sono correlate.

A differenza della covarianza campionaria però il suo valore è compreso nell'intervallo $[-1, 1]$.

In R può essere calcolato attraverso il comando `cor()`.

Console ~/ ↻							
> coeffCorrelazioneCamp<-cor(matriceAnalisi[,1:7])							
> coeffCorrelazioneCamp							
	carta/cartone	vetro	plastica	metalli	selettiva	verde/organici/legno	altro
carta/cartone	1.0000000	0.8508702	0.7605718	0.5635742	0.5820232	0.9027117	0.6825563
vetro	0.8508702	1.0000000	0.7521559	0.5500117	0.6131046	0.8375191	0.7286768
plastica	0.7605718	0.7521559	1.0000000	0.7294005	0.2632779	0.8451824	0.7161979
metalli	0.5635742	0.5500117	0.7294005	1.0000000	0.1991045	0.6087013	0.6741225
selettiva	0.5820232	0.6131046	0.2632779	0.1991045	1.0000000	0.5756637	0.4829807
verde/organici/legno	0.9027117	0.8375191	0.8451824	0.6087013	0.5756637	1.0000000	0.6984960
altro	0.6825563	0.7286768	0.7161979	0.6741225	0.4829807	0.6984960	1.0000000

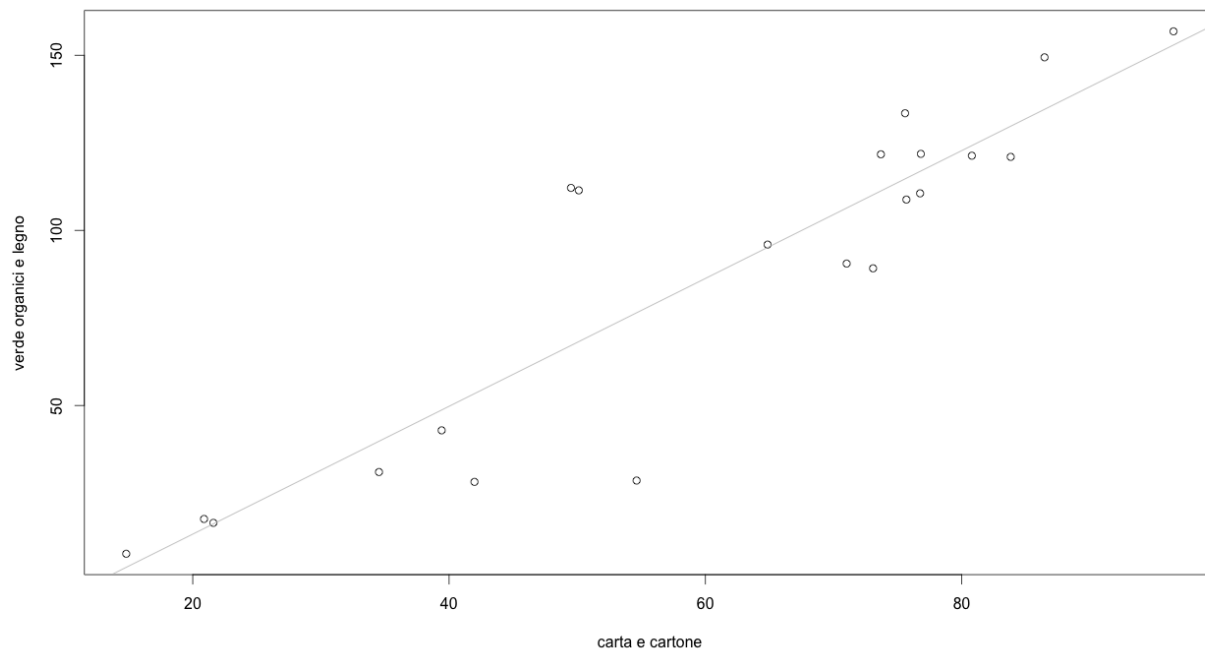
Come si può notare, le variabili dei rifiuti, sono tutte correlate positivamente. Le due correlazioni maggiori sono tra carta/cartone e verde/organico/legno.

5.2.3 Scatterplot

Di seguito viene riportato il diagramma scatterplot dei due indicatori che presentano il coefficiente di correlazione più alto, carta/cartone e verde/organico/legno, per evidenziare la relazione lineare tra i dati. Il codice R utilizzato è il seguente:

```
#scatterplot carta/cartone e verde/organico/legno
> plot(matriceAnalisi[,1], matriceAnalisi[,6], xlab="carta e cartone",
      ylab="verde organici e legno")
> abline(lm(matriceAnalisi[,6]~matriceAnalisi[,1]), col="gray")
```

Con il seguente risultato



6 Clustering

Lo scopo del clustering è quello di partizionare i dati in gruppi in modo tale da rendere minima la dissimilarità tra oggetti appartenenti ad uno stesso cluster e massima la dissimilarità tra oggetti appartenenti a cluster differenti.

In questo modo si otterrà quindi un'alta omogeneità all'interno dei gruppi ed un'alta eterogeneità tra gruppi distinti.

Il clustering può essere sfruttato per molti scopi tra i quali:

- Ridurre i dati in caso di un grande numero di osservazioni che risultano intrattabili a meno che non vengano classificati in gruppi che possono essere considerati come singole unità;
- Generare ipotesi sulla natura dei dati;
- Produrre gruppi che formano la base di uno schema di classificazione utile in studi successivi per scopi di previsioni di un qualche tipo.

Formalmente sia $I = I_1, I_2, \dots, I_n$ un insieme di n unità appartenenti ad una popolazione ideale.

Assumiamo che esista un insieme di caratteristiche (chiamate anche features) $C = C_1, C_2, \dots, C_p$ osservabili. Il problema dell'analisi dei cluster consiste nel determinare m sottoinsiemi (detti cluster) d'individui I , con m intero minore di n tale che I_i appartenga soltanto ad un unico sottoinsieme.

Gli individui che appartengono ad uno stesso cluster vengono detti simili, mentre quelli che appartengono a cluster differenti dissimili, risulta quindi cruciale per il problema del clustering definire in maniera precisa i termini di somiglianza e differenza.

La somiglianza può essere definita attraverso un *coefficiente di similarità* oppure mediante una *misura di distanza*, mentre i coefficienti di similarità assumono valori compresi tra 0 e 1 le misure di distanza possono assumere qualsiasi valore maggiore o uguale di zero.

Date queste considerazioni, il primo passo per effettuare un raggruppamento dei dati in input è quello di calcolare le distanze tra tutte le possibili coppie di unità e d'inserirle in una matrice denominata matrice delle distanze.

Per costruire la matrice delle distanze è possibile scegliere tra diverse metriche in base al problema in esame, le opzioni sono:

- Metrica euclidea;
- Metrica del valore assoluto o metrica di Manhattan;
- Metrica del massimo o metrica di Chebycey;
- Metrica di Minkowski;
- Distanza di Canberra;
- Distanza di Jaccar.

Tra queste la più nota è la metrica euclidea così definita:

$$d_2(X_i, X_j) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$

dove x_{ik} è il valore della k -esima caratteristica dell'individuo I . Sebbene la metrica euclidea non tenga conto della forma della distribuzione e sia fortemente influenzata dall'unità di misura in

base alla quale è valutata ciascuna delle p caratteristiche, è molto diffusa e di facile utilizzo e per questo motivo è stata scelta come metrica per costruire la matrice delle distanze per i dati presi in considerazione per questa tesina.

Una volta calcolata la matrice delle distanze, il passo successivo risulta essere quello di scegliere un algoritmo per il raggruppamento.

I metodi di raggruppamento si distinguono in tre tipi:

- Metodi di enumerazione completa;
- Metodi gerarchici;
- Metodi non gerarchici.

I metodi di enumerazione spesso sono computazionalmente onerosi perchè prevedono il calcolo della funzione obiettivo per ogni possibile partizione dell'insieme totale di n individui di m cluster, per questo motivo spesso si preferiscono i metodi gerarchici e non gerarchici che verranno trattati in dettaglio nelle sezioni successive.

6.1 Cluster gerarchico

I metodi di clustering gerarchico possono essere di due tipi:

- **Agglomerativi:** si parte da una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo per poi giungere alla fine ad una situazione in cui si ha un unico cluster che contiene tutti gli individui;
- **Divisivi:** si parte da una situazione in cui si ha un solo cluster che contiene tutti gli n individui, per poi giungere, ad una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo.

In entrambi metodi lo scopo è quello di ottenere una partizione dei dati che possa essere rappresentata su un particolare diagramma ad albero detto dendrogramma.

In seguito approfondiremo i metodi agglomerativi, tralasciando quelli divisivi.

Sebbene esistano diversi metodi gerarchici agglomerativi, essi possiedono una struttura di fondo comune che può essere esplicitata attraverso i seguenti passi:

- **Passo 1:** a partire dalla matrice X originaria dei dati o dalla matrice scalata si costruisce la matrice delle distanze;
- **Passo2:** Si individuano i due cluster meno distanti e si raggruppano in un unico cluster. Fatto questo si calcola la distanza tra il cluster originato dopo l'agglomerazione e tutti gli altri gruppi già esistenti;
- **Passo 3:** Si ricalcola la matrice delle distanze eliminando una riga ed una colonna;
- **Passo 4:** Si ripete il procedimento a partire dal passo 2 fino ad esaurire tutte le possibilità di raggruppamento;
- **Passo 5:** Si rappresentare il risultato ottenuto attraverso un dendrogramma dove ad ogni livello di distanza corrisponde una partizione.

Dopo aver scelto l'opportuna metrica al passo 1, ciò che differenzia sostanzialmente i vari algoritmi gerarchici di tipo agglomerativo è soltanto la definizione del concetto di distanza tra cluster al passo 2. Le tecniche che prenderemo in considerazione per calcolare la distanza saranno:

- Metodo del legame singolo
- Metodo del legame completo
- Metodo del legame medio
- Metodo del centroide
- Metodo della mediana

Prima di applicare qualsiasi di queste metodologie è stato necessario effettuare delle operazioni preliminari come la standardizzazione dei valori attraverso il comando `scale()` di R e la costruzione della matrice delle distanze. Il codice eseguito è il seguente:

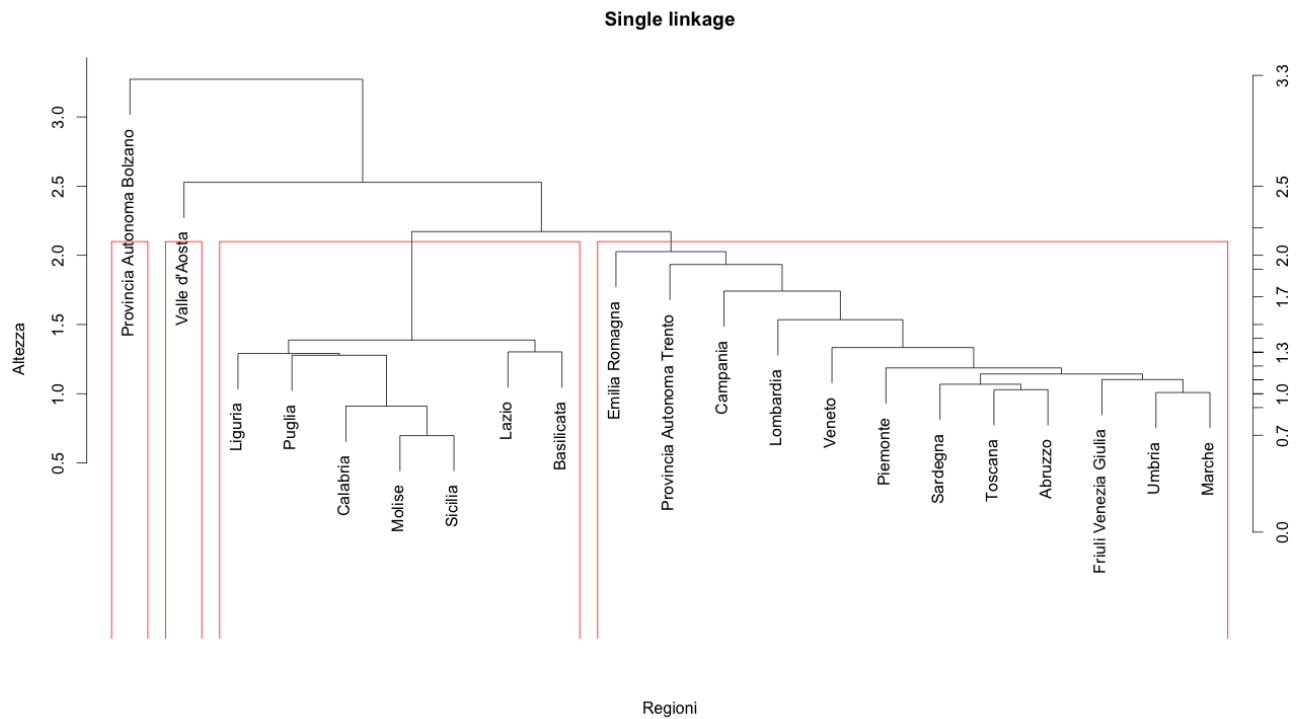
```
#dati scalati
> datiScalati <- scale(matriceAnalisi[,1:7])
#matrice delle distanze
> matriceDistanze <- dist(datiScalati,method="euclidean",diag=TRUE,upper=TRUE)
```

6.1.1 Metodo del legame singolo

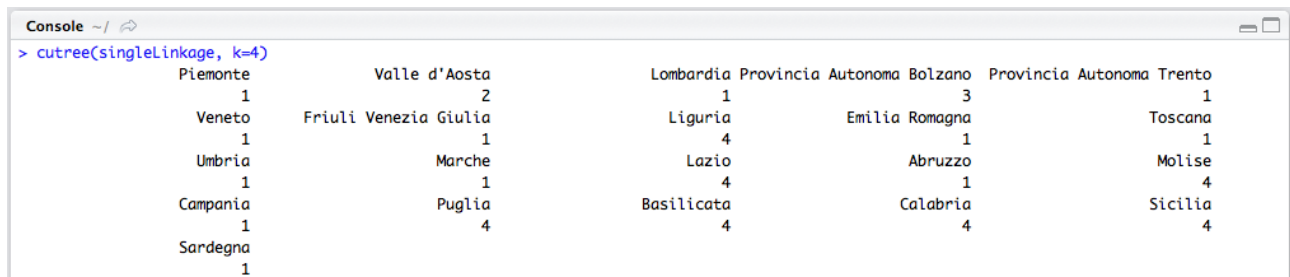
Nel metodo del legame singolo (chiamato anche single linkage) la distanza tra due cluster è definita come la distanza tra i due oggetti più vicini. Un vantaggio di tale metodo è quello di individuare gruppi di qualsiasi forma e di mettere in luce eventuali valori anomali. Uno svantaggio è che può dare origine a catene d'individui.

Di seguito viene riportato il codice R usato per generare il dendrogramma.

```
> singleLinkage <-hclust(matriceDistanze,method="single")
> plot(singleLinkage,main="Single linkage",xlab="Regioni",
      ylab="Altezza",sub="")
> axis(side=4, at=round(c(0, singleLinkage$height),1))
> rect.hclust(singleLinkage , k=4, border="red")
```



Grazie all'ausilio del comando `cutree` è possibile ottenere una matrice che indica a quale cluster appartiene ogni regione.



Ovvero:

- **C1:** Piemonte, Veneto, Umbria, Campania, Sardegna, Friuli Venezia Giulia, Marche, Lombardia, Emilia Romagna, Abruzzo, Provincia Autonoma Trento, Toscana;
- **C2:** Valle D'Aosta;
- **C3:** Provincia Autonoma Bolzano;
- **C4:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.

6.1.2 Metodo del legame completo

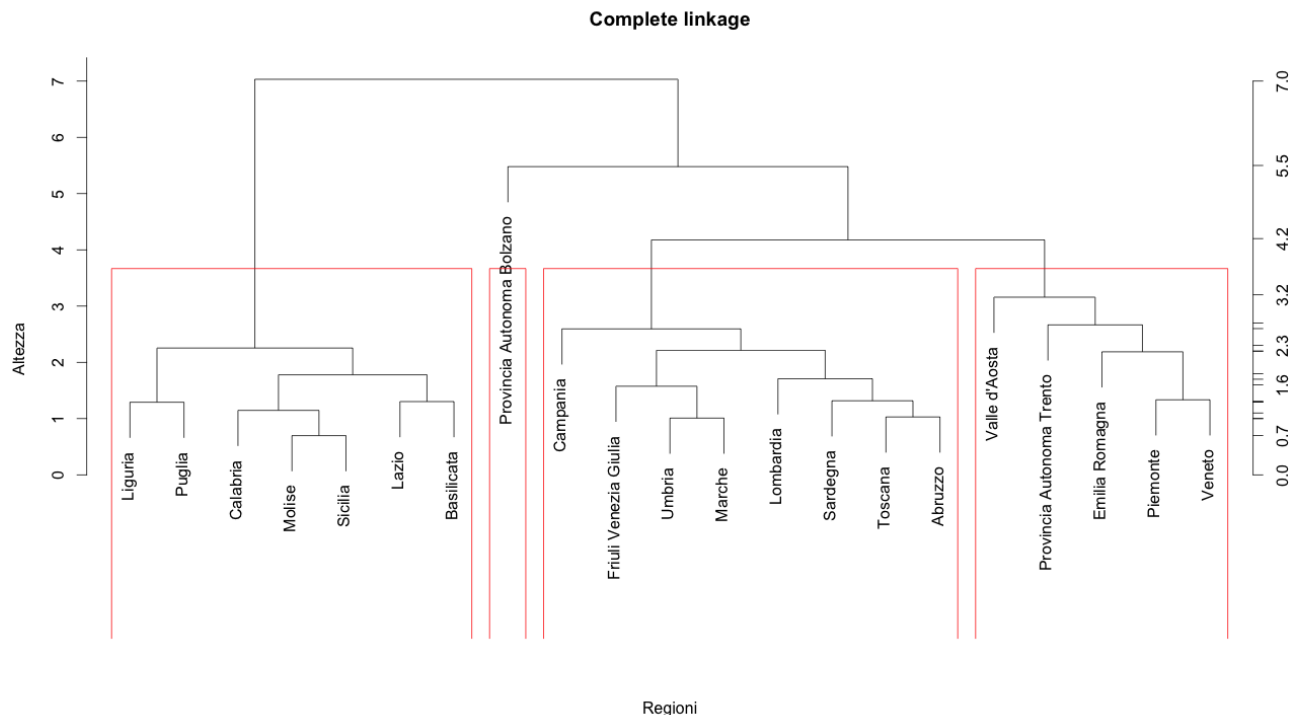
Nel metodo del legame completo (chiamato anche complete linkage) ad ogni passo per ogni coppia di cluster si prende in considerazione la distanza tra i due elementi più lontani.

La coppia di cluster che alla fine presenta il minimo di queste distanze viene scelta per l'unione.

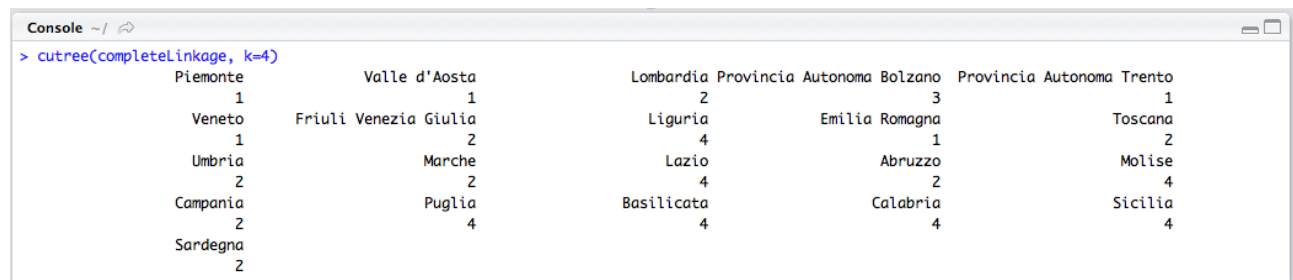
Questo metodo individua soprattutto i gruppi di forma ellissoidale, ossia una serie di punti che si addensano attorno ad un nucleo centrale. Il dendrogramma costruito con questo metodo ha i rami molto più lunghi rispetto al dendrogramma ottenuto con il metodo del legame singolo poichè i gruppi si formano a livelli di distanza maggiori.

Di seguito viene riportato il codice R per generare il dendrogramma.

```
#metodo legame completo
> completeLinkage <-hclust(matriceDistanze ,method="complete")
> plot(completeLinkage ,main="Complete linkage",xlab="Regioni",
      ylab="Altezza", sub="")
> axis(side=4, at=round(c(0, completeLinkage$height),1))
> rect.hclust(completeLinkage , k=4, border="red")
```



Come prima, grazie all'ausilio del comando `cutree` è possibile ottenere una matrice che indica a quale cluster appartiene ogni regione.



Ovvero:

- **C1:** Piemonte, Veneto, Valle D'Aosta, Emilia Romagna, Provincia Autonoma Trento;

- **C2:** Umbria, Campania, Sardegna, Friuli Venezia Giulia, Marche, Lombardia, Toscana;
- **C3:** Provincia Autonoma Bolzano;
- **C4:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.

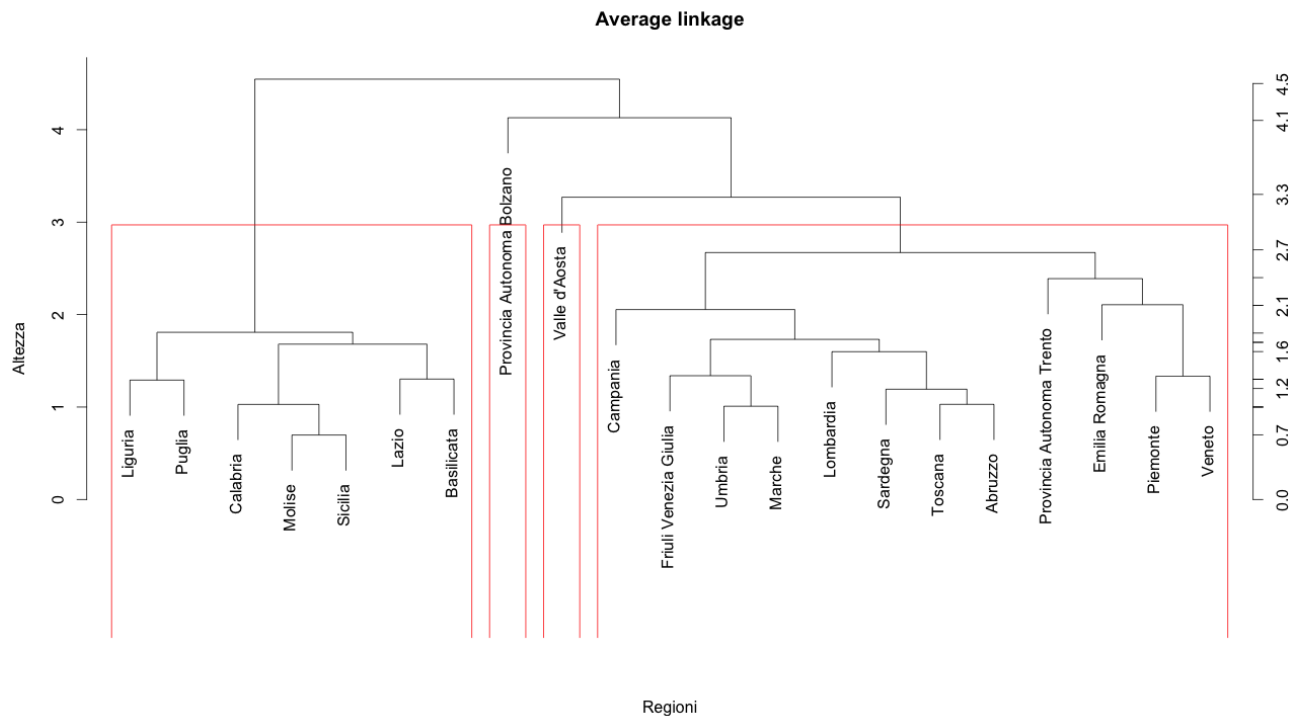
6.1.3 Metodo del legame medio

Nel metodo del legame medio (average linkage method) la distanza tra due cluster è rappresentata dalla media delle distanze tra tutte le coppie di elementi presenti nei due gruppi.

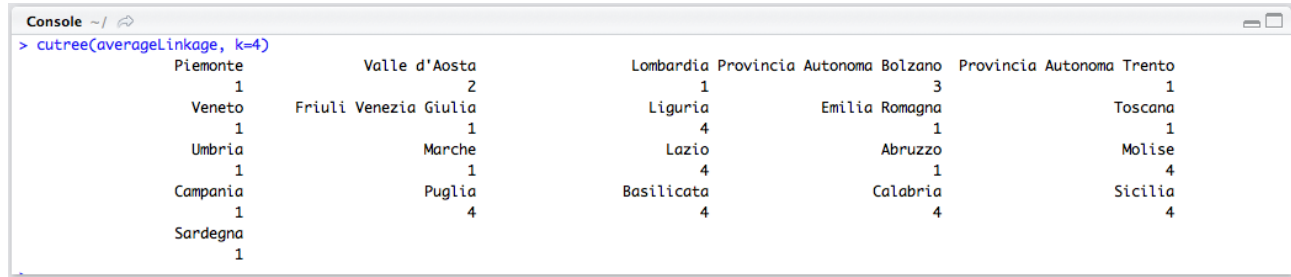
Uno svantaggio di questo metodo è che nel caso in cui si considerano cluster molto differenti, la distanza sarà molto vicina a quella del cluster più numeroso.

Di seguito viene riportato il codice R per generare il dendrogramma.

```
#metodo del legame medio
> averageLinkage <-hclust(matriceDistanze ,method="average")
> plot(averageLinkage ,main="Average linkage",xlab="Regioni",
      ylab="Altezza", sub="")
> axis(side=4, at=round(c(0, averageLinkage$height),1))
> rect.hclust(averageLinkage , k=4, border="red")
```



Come prima, grazie all'ausilio del comando *cutree* è possibile ottenere una matrice che indica a quale cluster appartiene ogni regione.



Ovvero:

- **C1:** Piemonte, Veneto, Umbria, Campania, Sardegna, Friuli Venezia Giulia, Marche, Lombardia, Emilia Romagna, Abruzzo, Provincia Autonoma Trento, Toscana;
- **C2:** Valle D'Aosta;
- **C3:** Provincia Autonoma Bolzano;
- **C4:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.

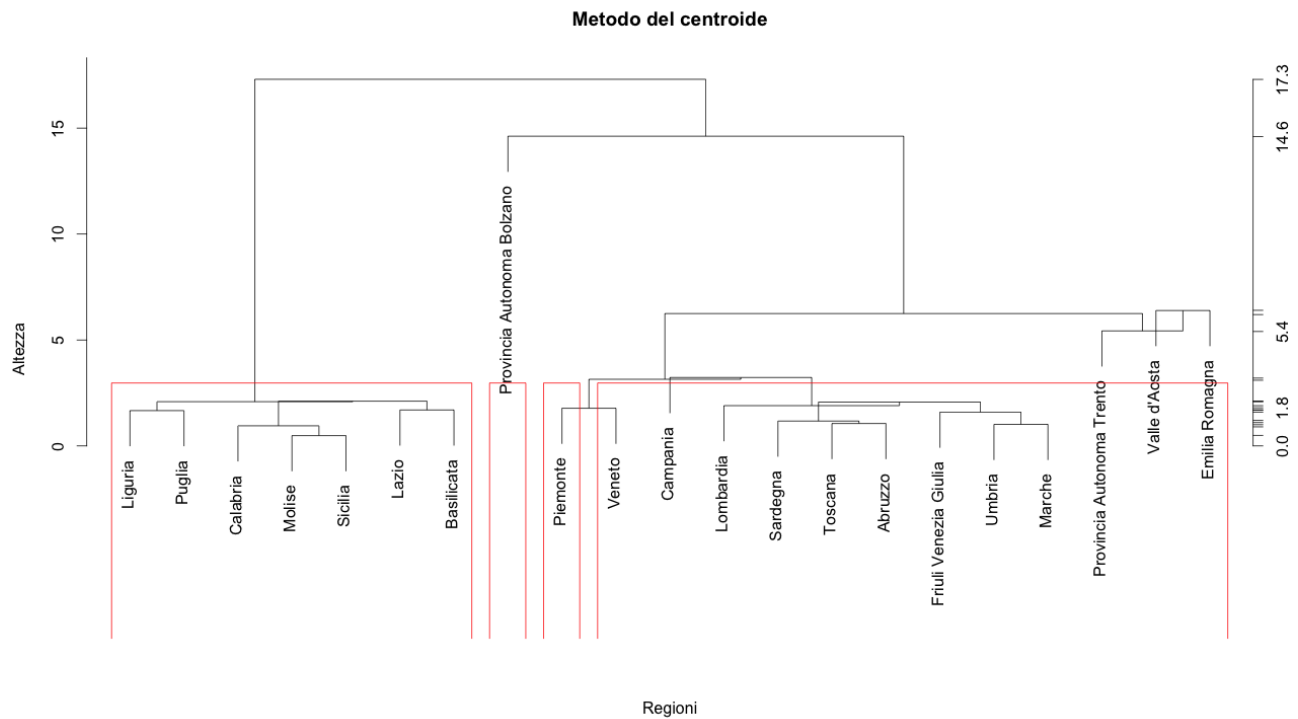
6.1.4 Metodo del centroide

Nel metodo del centroide la distanza tra due cluster è definita come la distanza tra i centroidi, ossia tra le medie campionarie calcolate sugli individui appartenenti ai due gruppi.

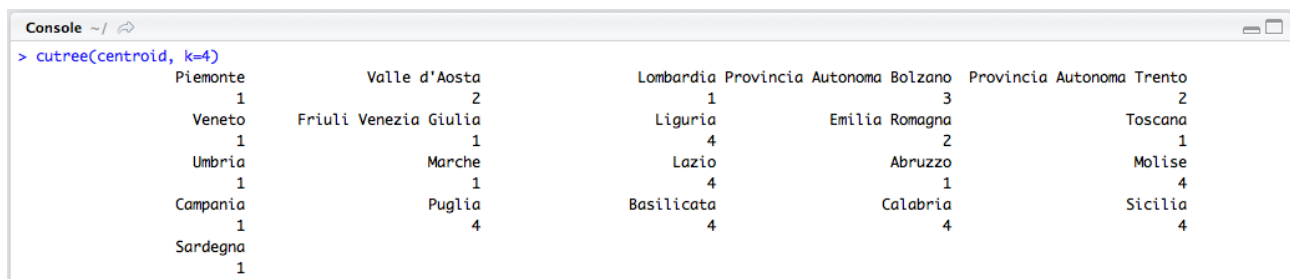
Il metodo del centroide può dare origine a fenomeni gravitazionali, per cui i gruppi grandi tendono ad attrarre al loro interno i piccoli gruppi. Inoltre le distanze in cui si verificano le successive agglomerazioni possono essere non crescenti. Uno svantaggio del metodo del centroide è che se le misure dei due cluster da unire sono molto differenti il centroide del nuovo cluster sarà molto vicino a quello del cluster più numeroso. A differenza dei casi precedenti, per il metodo del centroide bisogna elevare la matrice delle distanze al quadrato.

Di seguito viene riportato il codice R per generare il dendrogramma.

```
#metodo del legame medio
> averageLinkage <-hclust(matriceDistanze ,method="average")
> plot(averageLinkage ,main="Average linkage",xlab="Regioni",
ylab="Altezza",sub="")
> axis(side=4, at=round(c(0, averageLinkage$height),1))
> rect.hclust(averageLinkage , k=4, border="red")
```



Come prima, grazie all'ausilio del comando *cutree* è possibile ottenere una matrice che indica a quale cluster appartiene ogni regione.



Ovvero:

- **C1:** Piemonte, Veneto, Umbria, Campania, Sardegna, Friuli Venezia Giulia, Marche, Lombardia, Abruzzo, Toscana;
- **C2:** Valle D'Aosta, Emilia Romagna, Provincia Autonoma Trento;
- **C3:** Provincia Autonoma Bolzano;
- **C4:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.

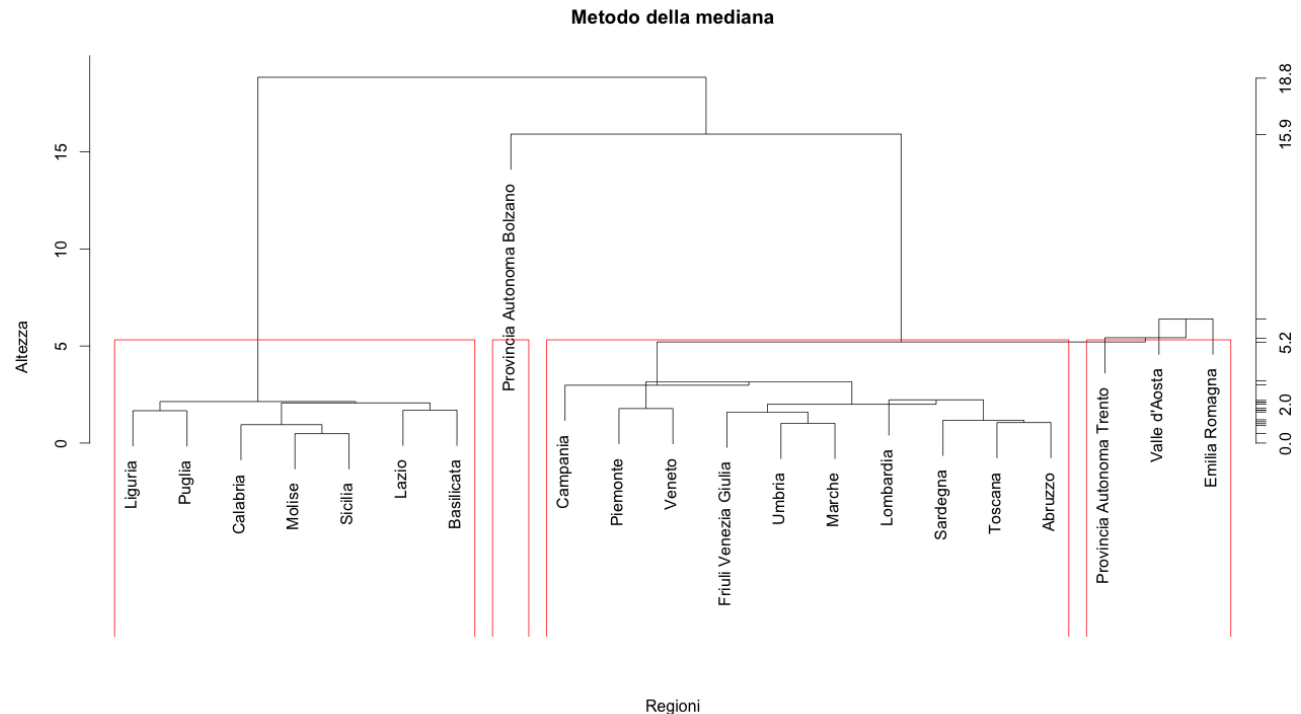
6.1.5 Metodo della mediana

Il metodo della mediana è simile a quello del centroide, con la differenza che la dimensione del cluster non influisce sul calcolo del centroide. Infatti, quando due cluster si aggregano, il nuovo centroide risulta essere la semisomma dei due centroidi precedenti. Come il metodo del legame singolo può

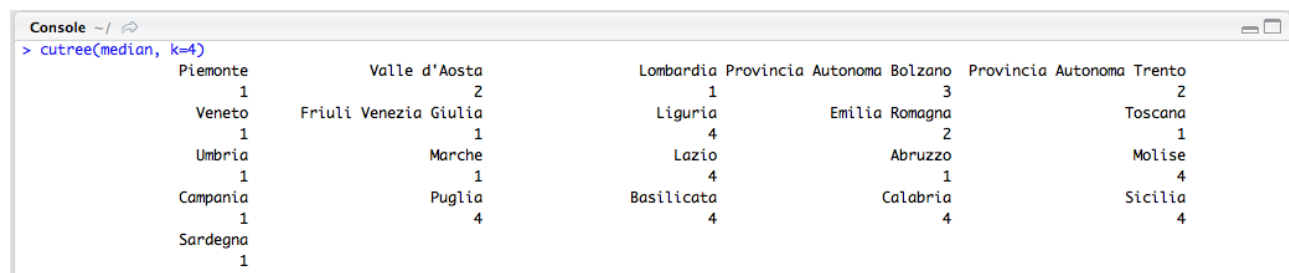
comportare la formazione di una catena tra gli individui.

Di seguito viene riportato il codice R usato per generare il dendrogramma.

```
#metodo del legame medio
> averageLinkage <-hclust(matriceDistanze ,method="average")
> plot(averageLinkage ,main="Average linkage",xlab="Regioni",
ylab="Altezza",sub="")
> axis(side=4, at=round(c(0, averageLinkage$height),1))
> rect.hclust(averageLinkage , k=4, border="red")
```



Come prima, grazie all'ausilio del comando *cutree* è possibile ottenere una matrice che indica a quale cluster appartiene ogni regione.



Ovvero:

- **C1:** Piemonte, Veneto, Umbria, Campania, Sardegna, Friuli Venezia Giulia, Marche, Lombardia, Abruzzo, Toscana;
- **C2:** Valle D'Aosta, Emilia Romagna, Provincia Autonoma Trento;

- **C3:** Provincia Autonoma Bolzano;
- **C4:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.

6.2 Metodi non gerarchici

Lo scopo dei metodi non gerarchici è quello di ottenere un'unica partizione degli n individui di partenza in cluster. A differenza dei metodi gerarchici in tali tecniche è consentito riallocare gli individui già classificati ad un livello precedente dell'analisi.

Esistono numerose tecniche di metodi non gerarchici ed è impossibile raggrupparli tutti in un unico tipo; in alcuni bisogna fissare a priori il numero di cluster da formare mentre in altri viene determinato durante l'analisi. In molti casi, inoltre, bisogna fissare un insieme di punti di riferimento. In generale gli algoritmi di tipo non gerarchico procedono, data una prima partizione, a riallocare gli individui nel gruppo con centroide più vicino, fino a che per nessun individuo si verifica che sia minima la distanza rispetto al centroide di un gruppo diverso da quello a cui esso appartiene.

6.2.1 K-Means

Il metodo non gerarchico più utilizzato è noto come k-means e consiste nei seguenti passi:

- **Passo 1:** fissare a priori il numero k di cluster da formare specificando k punti di riferimento iniziali;
- **Passo 2:** Considerare tutti gli individui e attribuire ciascuno di essi al cluster individuato dal punto di riferimento da cui ha distanza minore;
- **Passo 3:** Calcolare il centroide di ognuno dei k gruppi così ottenuti. Tali centroidi costituiscono i punti di riferimento per i nuovi cluster;
- **Passo 4:** Valutare la distanza di ogni unità da ogni centroide ottenuto al passo precedente. Se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora si procede a spostare l'individuo presso il cluster che ha il centroide più vicino;
- **Passo 5:** Ricalcolare i centroidi dei k gruppi così ottenuti;
- **Passo 6:** Ripetere il procedimento a partire dal punto (4) fino a che i centroidi non subiscono ulteriori modifiche rispetto all'iterazione precedente. Si procede così iterativamente a spostamenti successivi fino a raggiungere una configurazione stabile, ossia gli individui all'interno di ogni cluster non cambiano al ripetersi del procedimento.

Il k-means è un metodo veloce a livello computazionale ma presenta il problema dei minimi locali infatti, dal momento che i valori iniziali possono essere diversi, ogni specifica allocazione porterà ad un ottimo che è locale al problema ma non è necessariamente quello globale.

Inoltre, vi è il problema di dover determinare a priori il numero di cluster da utilizzare.

In R l'analisi del metodo del k-means viene fatta attraverso l'utilizzo della funzione `kmeans()`:

```
kmeans(X, centers , iter.max = N, nstart = M)
```

Dove X rappresenta l'input iniziale, `centers` il numero di cluster che vogliamo formare, `iter.max` il numero d'iterazioni che vogliamo effettuare (di default è 10) e `nstart` il numero di volte in cui vogliamo ripetere la procedura di scelta casuale dei punti di riferimento.

Scelta casuale dei punti di riferimento

Di seguito viene applicato il metodo del k-means scegliendo casualmente i punti di riferimento iniziali ed effettuando 10 iterazioni.

Il comando utilizzato è:

```
> km <- kmeans(matriceAnalisi, centers=3, iter.max=10, nstart=1)
```

Con il seguente risultato:

```

Console ~/
> km <- kmeans(matriceAnalisi , centers=3, iter.max=10, nstart=1)
> km
K-means clustering with 3 clusters of sizes 6, 7, 8

Cluster means:
  carta/cartone vetro  plastica  metalli  selettiva  verde/organici/legno  altro  totale
1   64.99833 29.69500 15.485000 4.126667 0.7033333 104.66833 20.09000 239.75833
2   32.55143 10.26143 4.904286 3.244286 0.1200000 24.63143 10.46286 85.95429
3   80.58625 35.35875 22.471250 7.040000 0.7437500 127.00500 34.15750 307.36000

Clustering vector:
      Piemonte      Valle d'Aosta      Lombardia Provincia Autonoma Bolzano Provincia Autonoma Trento
      3          3          1          1          3
      Veneto Friuli Venezia Giulia      Liguria      Emilia Romagna      Toscana
      3          3          2          3          1
      Umbria      Marche      Lazio      Abruzzo      Molise
      3          3          2          1          2
      Campania      Puglia      Basilicata      Calabria      Sicilia
      1          2          2          2          2
      Sardegna
      1

Within cluster sum of squares by cluster:
[1] 2229.465 8946.502 10030.907
(between_SS / total_SS = 92.0 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

```

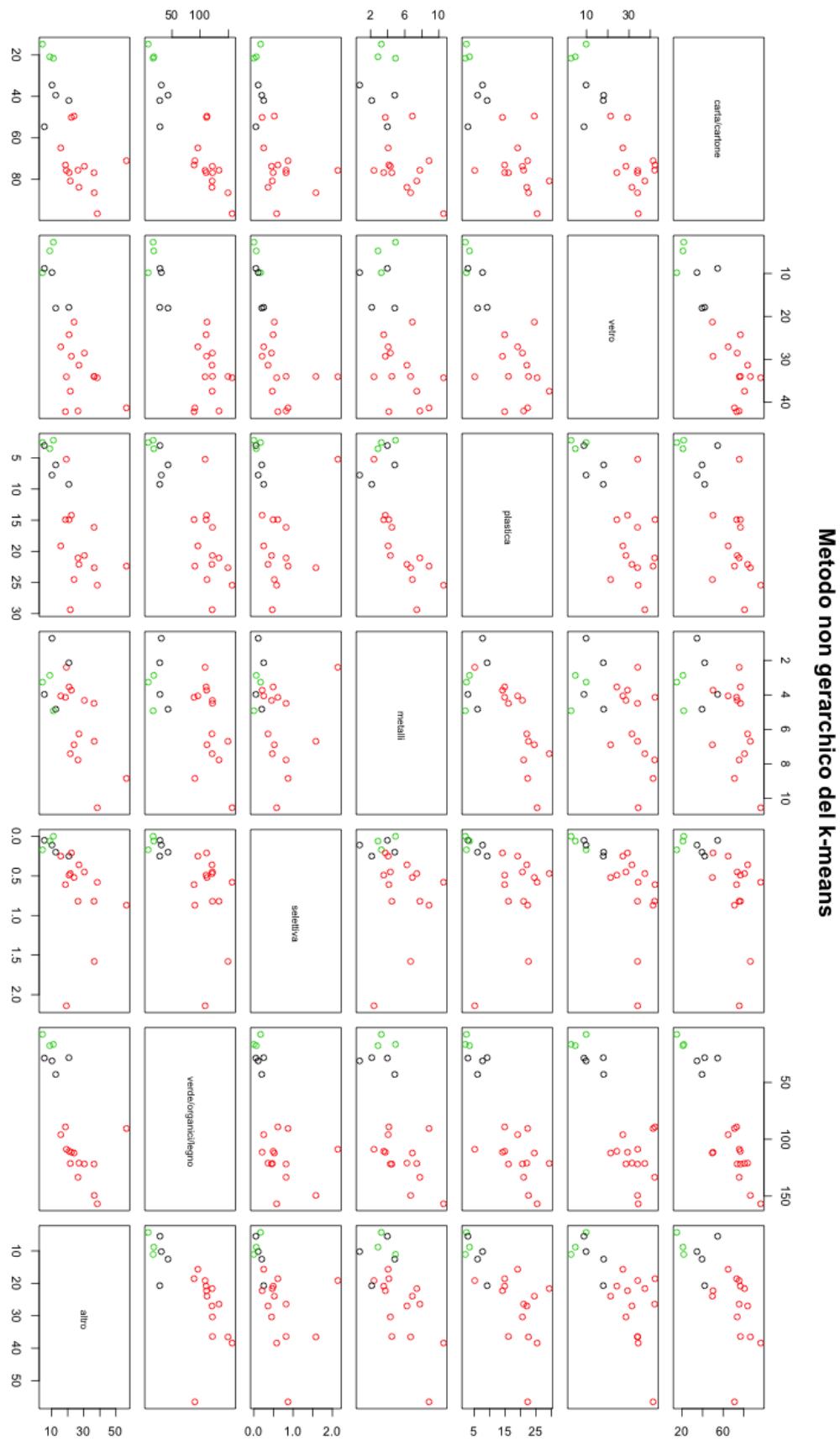
In questo caso il k-means ha individuato la seguente partizione:

- **C1:** Campania, Sardegna, Lombardia, Provincia Autonoma Bolzano, Toscana, Abruzzo.
- **C2:** Puglia, Liguria, Lazio, Basilicata, Calabria, Molise, Sicilia.
- **C3:** Piemonte, Veneto, Umbria, Valle D'Aosta, Friuli Venezia Giulia, Marche, Emilia Romagna, Provincia Autonoma Trento.

È possibile rappresentare graficamente i cluster:

```
> pairs(matriceAnalisi , col=km$cluster , main="Metodo non gerarchico del k-means")
```

con il seguente risultato:



Ripetizione della procedura di scelta casuale dei punti di riferimento

Dal momento che la partizione iniziale è scelta casualmente non è detto che la procedura del k-means conduca sempre allo stesso risultato con configurazioni iniziali differenti, per questo motivo in questa sezione si è deciso di porre il parametro $nstart = 10$ ed osservare i risultati.

```

Console ~/
> km <- kmeans(matriceAnalisi[,1:7] , centers=3, iter.max=10, nstart=10)
> km
K-means clustering with 3 clusters of sizes 7, 7, 7

Cluster means:
  carta/cartone vetro plastica metalli selettiva verde/organici/legno altro
1  65.85857 31.35286 16.468571 4.800000 0.7271429 102.64714 25.29714
2  32.55143 10.26143  4.904286 3.244286 0.1200000  24.63143 10.46286
3  81.95286 34.51000 22.485714 6.782857 0.7257143 132.21714 30.96000

Clustering vector:
      Piemonte      Valle d'Aosta      Lombardia Provincia Autonoma Bolzano Provincia Autonoma Trento
      3          1          1          1          1          3
      Veneto Friuli Venezia Giulia      Liguria      Emilia Romagna      Toscana
      3          3          2          3          1
      Umbria      Marche      Lazio      Abruzzo      Molise
      3          3          2          1          2
      Campania      Puglia      Basilicata      Calabria      Sicilia
      1          2          2          2          2
      Sardegna
      1

Within cluster sum of squares by cluster:
[1] 3306.124 2448.792 2219.442
(between_SS / total_SS = 87.8 %)

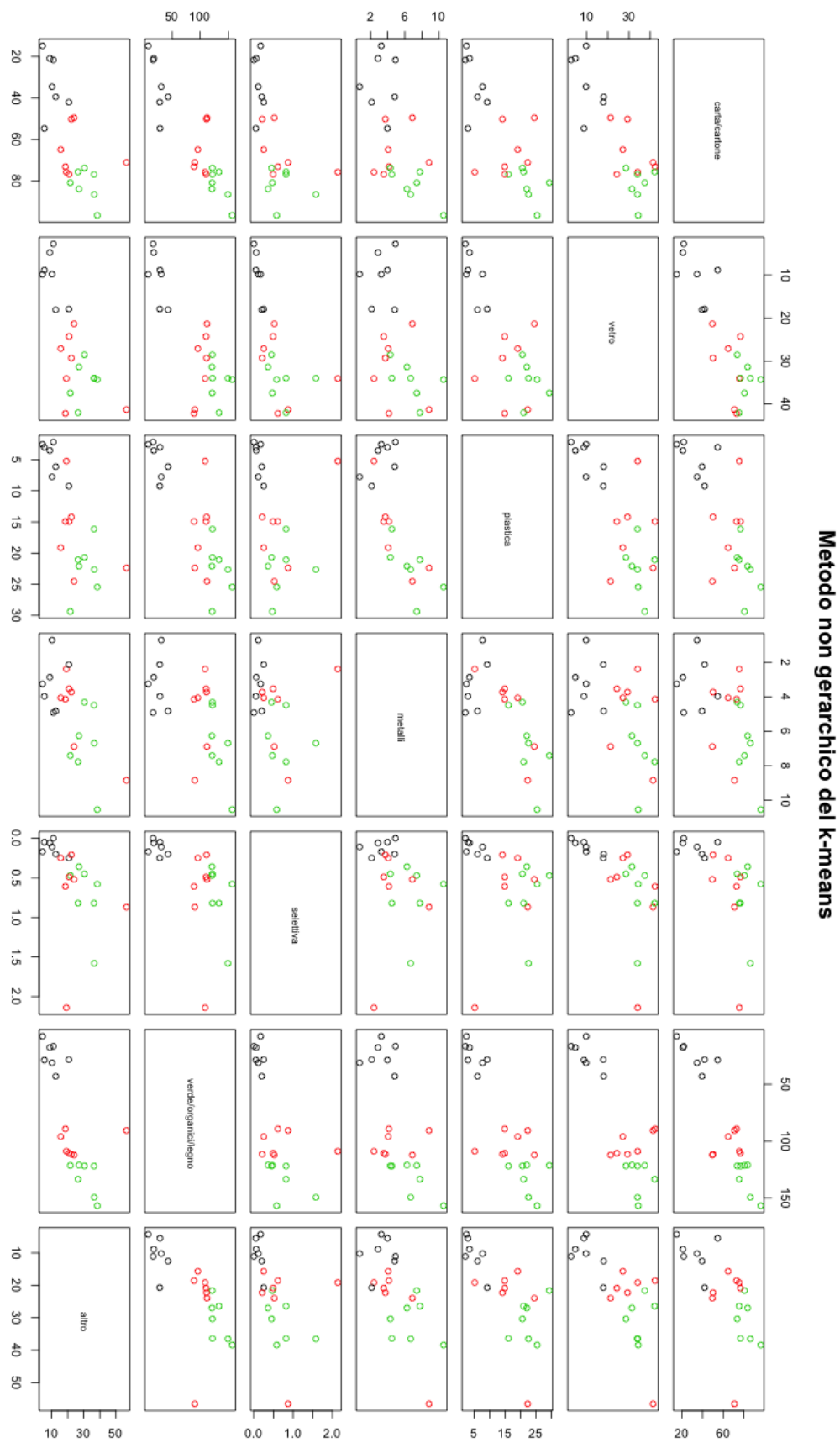
Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"

```

Dopo 10 iterazioni sono stati formati 3 cluster:

- **C1:** Campania, Sardegna, Valle D'Aosta, Lombardia, Provincia Autonoma Bolzano, Abruzzo, Toscana.
- **C2:** Molise, Sicilia, Puglia, Liguria, Lazio, Basilicata, Calabria.
- **C3:** Piemonte, Veneto, Umbria, Friuli Venezia Giulia, Marche, Emilia Romagna, Toscana.

Come si può notare il risultato è differente dal precedente raggruppamento.



Scelta dei centroidi

In alternativa alla scelta casuale dei punti di riferimento si possono utilizzare i centroidi dei cluster ottenuti con la tecnica del centroide utilizzando la funzione `aggregate()`.

```
#tecnica dei centroidi
tree<-hclust(matriceDistanze2 ,method="centroid")
> taglio<-cutree(tree ,k=3,h=NULL)
> tagliolist<-list(taggio)
> centroidiIniziali<-aggregate(matriceAnalisi[,1:7],tagliolist,mean) [, -1]
```

```
Console ~/
> centroidiIniziali
  carta/cartone vetro plastica metalli selettiva verde/organici/legno altro
1  73.76923 32.84538 20.573846 6.052308 0.6176923 118.09692 28.81846
2  75.68000 34.05000 5.220000 2.400000 2.1400000 108.79000 19.16000
3  32.55143 10.26143 4.904286 3.244286 0.1200000 24.63143 10.46286
```

```
> k <- kmeans(matriceAnalisi[,1:7],centers=centroidiIniziali,iter.max =10)
```

```
Console ~/
> k
K-means clustering with 3 clusters of sizes 7, 7, 7

Cluster means:
  carta/cartone vetro plastica metalli selettiva verde/organici/legno altro
1  81.95286 34.51000 22.485714 6.782857 0.7257143 132.21714 30.96000
2  65.85857 31.35286 16.468571 4.800000 0.7271429 102.64714 25.29714
3  32.55143 10.26143 4.904286 3.244286 0.1200000 24.63143 10.46286

Clustering vector:
      Piemonte          Valle d'Aosta          Lombardia Provincia Autonoma Bolzano
      1              2              2              2
Provincia Autonoma Trento          Veneto          Friuli Venezia Giulia          Liguria
      1              1              1              3
      Emilia Romagna          Toscana          Umbria          Marche
      1              2              1              1
      Lazio          Abruzzo          Molise          Campania
      3              2              3              2
      Puglia          Basilicata          Calabria          Sicilia
      3              3              3              3
      Sardegna
      2

Within cluster sum of squares by cluster:
[1] 2219.442 3306.124 2448.792
(between_SS / total_SS = 87.8 %)

Available components:
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"
[9] "ifault"
```

- **C1:** Piemonte, Provincia Autonoma Trento, Emilia Romagna, Veneto, Friuli Venezia Giulia, Umbria, Marche.
- **C2:** Sardegna, Valle D'Aosta, Toscana, Abruzzo, Lombardia, Campania, Provincia Autonoma Bolzano.
- **C3:** Lazio, Puglia, Basilicata, Calabria, Molise, Liguria, Sicilia.

