PROGETTO DI METODI E TECNICHE PER L'ANALISI DEI DATI

Ambiente utilizzato: **R**

Data base utilizzato: Principi attivi contenuti nei prodotti fitosanitari per regione 2015

Prof.ssa:
Amelia G. Nobile

Studentessa:
Carmen Bisogni

Anno 2016-2017

Prendiamo in esame il data base che raccoglie i dati relativi ai principi attivi contenuti nei prodotti fitosanitari, prodotti cioè utilizzati per proteggere i vegetali da organismi nocivi.
La situazione presa in analisi è relativa ai kg per ettaro, all'anno 2015 e suddivisa in regioni.
Le sostanze prese in esame sono: Fungicidi, Insetticidi-Acaricidi, Erbicidi e Vari.
Gli scopi di questa analisi sono:

- Individuare regioni che fanno particolare utilizzo di una specifica sostanza
- Individuare relazioni tra una o più sostanze utilizzate
- Individuare la distribuzione delle sostanze nelle varie regioni
- Individuare eventuali anomalie e valori fuori dal range accettabile
- Dividere le regioni in gruppi in base al quantitativo di sostanze utilizzato

Tabella utilizzata

Principi attivi contenuti nei prodotti fitosanitari per ettaro di superficie trattabile (in chilogrammi)) -Dettaglio per regioni(2015):

REGIONI	Fungicidi	Insetticidi e acaricidi	Erbicidi	Vari
Abruzzo	5,63	0,26	0,46	0,13
Basilicata	2,17	0,23	0,27	0,68
Calabria	1,58	1,10	0,37	0,28
Campania	4,68	1,05	0,77	5,15
Emilia-Romagna	5,97	1,27	1,44	0,64
Friuli-Venezia Giulia	31,92	3,45	6,41	10,07
Lazio	2,83	0,48	0,64	3,64
Liguria	66,42	3,83	15,72	1,60
Lombardia	1,80	0,34	1,53	0,78
Marche	1,82	0,13	0,65	0,06
Molise	0,76	0,09	0,28	0,12
Piemonte	5,39	0,55	1,77	0,27
Puglia	3,95	0,65	0,64	0,26
Sardegna	1,77	0,18	0,24	0,20
Sicilia	5,89	0,90	0,51	2,10
Toscana	4,12	0,25	0,58	0,21
Trentino-Alto Adige	0,55	0,36	0,02	0,63
Umbria	2,82	0,07	0,48	0,06
Valle d'Aosta/Vallée d'Aoste	18,14	1,96	2,58	0,22
Veneto	1,41	0,44	0,05	0,01

Una prima analisi potrebbe ad esempio essere la seguente:
Ci chiediamo quante regioni utilizzino quantitativi di particolari sostanze all'interno di un certo range.

Fungicidi:

```
>Fungicidi<-
c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82,0.76,5.39,3.95,1.77,5.89,4.12,0.55,2.82,18
.14,1.41)
>table ( cut ( Funghicidi, breaks =c (0,1 ,5 ,10 ,40 ,70) ))
```

```
 (0,1]  (1,5]  (5,10] (10,40] (40,70]
    2     11      4      2      1
```

Da questa analisi risulta che la maggior parte delle regioni utilizza non più di 5kg/ettaro di Fungicidi, solo 4 ne utilizzano un quantitativo tra i 5 e i 10 e solo 3 regioni superano i 10kg/ettaro.

Insetticidi e Acaricidi:

```
>InsAca<-
c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13,0.09,0.55,0.65,0.18,0.90,0.25,0.36,0.07,1.96,
0.44)
> table(cut(InsAca,breaks=c(0,0.5,1,1.5,3,4)))

(0,0.5] (0.5,1] (1,1.5] (1.5,3]   (3,4]
    11      3       3       1       2
```

Ovviamente nel caso di Insetticidi-Acaricidi abbiamo scelto range diversi poiché il quantitativo utilizzato è nettamente inferiore.
Abbiamo quindi che la maggior parte delle regioni utilizza meno di 1.5kg/ettaro, una sola utilizza tra 1.5 e 3kg/ettaro e due ne utilizzano tra 3 e 4kg/ettaro.

Erbicidi:

```
> Erbicidi<-
c(0.46,0.27,0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65,0.28,1.77,0.64,0.24,0.51,0.58,0.02,0.48,2.5
8,0.05)
> table(cut(Erbicidi,breaks=c(0,1,1.5,2,3,7,16)))

 (0,1] (1,1.5] (1.5,2]   (2,3]   (3,7]  (7,16]
    14      1       2       1       1       1
```

Per gli erbicidi complessivamente la maggior parte delle regioni utilizza meno di 1kg/ettaro, e in particolare solo una regione supera i 7kg/ettaro.

Vari:

```
> Vari<-
c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06,0.12,0.27,0.27,0.20,2.10,0.21,0.63,0.06,0.22,0.01)
> table(cut(Vari,breaks=c(0,0.1,0.5,1,2,3,10,11)))

 (0,0.1] (0.1,0.5]   (0.5,1]     (1,2]     (2,3]    (3,10]   (10,11]
       3         8         4         1         1         2         1
```

Anche per i vari si ripete la stessa situazione, seppur con range differenti, la maggior parte delle regioni utilizza al più 1kg/ettaro, due ne utilizzano tra 3 e 10kg/ettaro e solo una supera i 10kg/ettaro.

Ora per avere un quadro completo trattiamo la tabella completa dei dati quantitativi:

```
> m<-
cbind(c(5.63,0.26,0.46,0.13),c(2.17,0.23,0.27,0.68),c(1.58,1.10,0.37,0.28),c(4.68,1.05,0.77,5.15),c(5.79,1.27,1.44,0.64),c(31.92,3.45,6.41,10.07),c(2.83,0.48,0.64,3.64),c(66.42,3.83,15.72,1.60),c(1.80,0.34,1.53,0.78),c(1.82,0.13,0.65,0.06),c(0.76,0.09,0.28,0.12),c(5.39,0.55,1.77,0.27),c(3.95,0.65,0.64,0.26),c(1.77,0.18,0.24,0.20),c(5.89,0.90,0.51,2.10),c(4.12,0.25,0.58,0.21),c(0.55,0.36,0.02,0.63),c(2.82,0.07,0.48,0.06),c(18.14,1.96,2.58,0.22),c(1.41,0.44,0.05,0.01))
> colnames(m)<-c("Abruzzo","Basilicata","Calabria","Campania","Emilia-Romagna","Friuli-Venezia Giulia","Lazio","Liguria","Lombardia","Marche","Molise","Piemonte","Puglia","Sardegna","Sicilia","Toscana","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto")
> rownames(m)<-c("Fungicidi","Insetticidi e acaricidi","Erbicidi","Vari")
> m
```

	Abruzzo	Basilicata	Calabria	Campania	Emilia-Romagna
Fungicidi	5.63	2.17	1.58	4.68	5.79
Insetticidi e acaricidi	0.26	0.23	1.10	1.05	1.27
Erbicidi	0.46	0.27	0.37	0.77	1.44
Vari	0.13	0.68	0.28	5.15	0.64

	Friuli-Venezia Giulia	Lazio	Liguria	Lombardia
Fungicidi	31.92	2.83	66.42	1.80
Insetticidi e acaricidi	3.45	0.48	3.83	0.34
Erbicidi	6.41	0.64	15.72	1.53
Vari	10.07	3.64	1.60	0.78

	Marche	Molise	Piemonte	Puglia	Sardegna	Sicilia
Fungicidi	1.82	0.76	5.39	3.95	1.77	5.89
Insetticidi e acaricidi	0.13	0.09	0.55	0.65	0.18	0.90
Erbicidi	0.65	0.28	1.77	0.64	0.24	0.51
Vari	0.06	0.12	0.27	0.26	0.20	2.10

	Toscana	Trentino-Alto Aldige	Umbria	Valle d'Aosta
Fungicidi	4.12	0.55	2.82	18.14
Insetticidi e acaricidi	0.25	0.36	0.07	1.96
Erbicidi	0.58	0.02	0.48	2.58
Vari	0.21	0.63	0.06	0.22

	Veneto
Fungicidi	1.41
Insetticidi e acaricidi	0.44
Erbicidi	0.05
Vari	0.01

Osservazione: Per tutta la matrice di dati abbiamo la stessa unità di misura (kg/ettaro)

Ora calcoliamo le frequenze marginali sulle righe:

> margin.table(m,1)

Fungicidi	Insetticidi e acaricidi	Erbicidi	Vari
169.44	17.59	35.41	27.11

Queste frequenze marginali rappresentano in realtà il totale di kg/ettaro di sostanza utilizzata in tutta Italia, indipendentemente dalla regione.

(R ha in effetti sommato per riga)

Notiamo che la sostanza più utilizzata è il Fungicida, la meno utilizzata è l'Insetticida-Acaricida.

E sulle colonne:

>margin.table(m,2)

Abruzzo	Basilicata	Calabria	Campania
6.48	3.35	3.33	11.65

Emilia-Romagna	Friuli-Venezia Giulia	Lazio	Liguria
9.14	51.85	7.59	87.57

Lombardia	Marche	Molise	Piemonte
4.45	2.66	1.25	7.98

Puglia	Sardegna	Sicilia	Toscana
5.50	2.39	9.40	5.16

Trentino-Alto Aldige	Umbria	Valle d'Aosta	Veneto
1.56	3.43	22.90	1.91

Qui quindi stiamo in effetti sommando per colonna, cioè vedendo quanti kg/ettaro di prodotti fitosanitari utilizza ogni regione, indipendentemente dalla sostanza.

Possiamo ad esempio manualmente notare che la regione che utilizza più prodotti fitosanitari è la Liguria.

Calcoliamo le frequenze relative congiunte:

Osservazione: Non sono calcolate per riga o colonna, ma per tutta la matrice di dati, ciò significa

che così come sono scritte prescindono dalla sostanza utilizzata.

```
> mr<-prop.table(m)
> mr
> sum(mr)
[1] 1
```

	Abruzzo	Basilicata	Calabria	Campania	Emilia-Romagna
Fungicidi	0.0225606091	0.008695652	0.006331397	0.018753757	0.023201763
Insetticidi e acaricidi	0.0010418754	0.000921659	0.004407934	0.004207574	0.005089160
Erbicidi	0.0018433180	0.001081948	0.001482669	0.003085554	0.005770387
Vari	0.0005209377	0.002724905	0.001122020	0.020637147	0.002564616

	Friuli-Venezia Giulia	Lazio	Liguria	Lombardia
Fungicidi	0.12791024	0.011340413	0.266159086	0.007212983
Insetticidi e acaricidi	0.01382488	0.001923462	0.015347626	0.001362452
Erbicidi	0.02568624	0.002564616	0.062993388	0.006131036
Vari	0.04035263	0.014586255	0.006411541	0.003125626

	Marche	Molise	Piemonte	Puglia	Sardegna
Fungicidi	0.0072931276	0.0030454819	0.021598878	0.015828491	0.0070927670
Insetticidi e acaricidi	0.0005209377	0.0003606492	0.002203967	0.002604688	0.0007212983
Erbicidi	0.0026046884	0.0011220196	0.007092767	0.002564616	0.0009617311
Vari	0.0002404328	0.0004808656	0.001081948	0.001041875	0.0008014426

	Sicilia	Toscana	Trentino-Alto Aldige
Fungicidi	0.023602484	0.0165097175	2.203967e-03
Insetticidi e acaricidi	0.003606492	0.0010018032	1.442597e-03
Erbicidi	0.002043679	0.0023241835	8.014426e-05
Vari	0.008415147	0.0008415147	2.524544e-03

	Umbria	Valle d'Aosta	Veneto
Fungicidi	0.0113003406	0.0726908435	5.650170e-03
Insetticidi e acaricidi	0.0002805049	0.0078541374	1.763174e-03
Erbicidi	0.0019234622	0.0103386095	2.003606e-04
Vari	0.0002404328	0.0008815869	4.007213e-05

Da cui si ricavano le frequenze relative marginali sulle righe:

```
>margin.table(mr,1)
```

Fungicidi	Insetticidi e acaricidi	Erbicidi	Vari
0.67898217	0.07048688	0.14189541	0.10863554

e sulle colonne:

Abruzzo	Basilicata	Calabria	Campania	Emilia-Romagna
0.025966740	0.013424163	0.013344019	0.046684031	0.036625927

Friuli-Venezia Giulia	Lazio	Liguria	Lombardia
0.207773993	0.030414747	0.350911641	0.017832098

Marche	Molise	Piemonte	Puglia
0.010659187	0.005009016	0.031977560	0.022039671

Sardegna	Sicilia	Toscana	Trentino-Alto Aldige
0.009577239	0.037667802	0.020677219	0.006251252

Umbria	Valle d'Aosta	Veneto
0.013744741	0.091765177	0.007653777

Ancora una volta, ad uno sguardo attento, ricaviamo le stesse informazioni viste per le frequenze assolute.

Possiamo poi calcolare la distribuzione delle frequenze relative condizionate f(regione|sostanza):

>prop.table(mr,1):

	Abruzzo	Basilicata	Calabria	Campania
Fungicidi	0.033227101	0.012806893	0.009324835	0.02762040
Insetticidi e acaricidi	0.014781126	0.013075611	0.062535532	0.05969301
Erbicidi	0.012990681	0.007624965	0.010449026	0.02174527
Vari	0.004795278	0.025082995	0.010328292	0.18996680

	Emilia-Romagna	Friuli-Venezia Giulia	Lazio	Liguria
Fungicidi	0.03417139	0.1883853	0.01670208	0.39199717
Insetticidi e acaricidi	0.07220011	0.1961342	0.02728823	0.21773735
Erbicidi	0.04066648	0.1810223	0.01807399	0.44394239
Vari	0.02360752	0.3714496	0.13426780	0.05901881

	Lombardia	Marche	Molise	Piemonte	Puglia
Fungicidi	0.01062323	0.010741265	0.004485364	0.031810670	0.023312087
Insetticidi e acaricidi	0.01932916	0.007390563	0.005116543	0.031267766	0.036952814
Erbicidi	0.04320813	0.018356396	0.007907371	0.049985880	0.018073990
Vari	0.02877167	0.002213205	0.004426411	0.009959425	0.009590557

	Sardegna	Sicilia	Toscana	Trentino-Alto Aldige
Fungicidi	0.010446176	0.03476157	0.024315392	0.0032459868
Insetticidi e acaricidi	0.010233087	0.05116543	0.014212621	0.0204661740
Erbicidi	0.006777746	0.01440271	0.016379554	0.0005648122
Vari	0.007377352	0.07746219	0.007746219	0.0232386573

	Umbria	Valle d'Aosta	Veneto
Fungicidi	0.016643059	0.107058546	0.0083215297
Insetticidi e acaricidi	0.003979534	0.111426947	0.0250142126
Erbicidi	0.013555493	0.072860774	0.0014120305
Vari	0.002213205	0.008115087	0.0003688676

Questa tabella ci dà 1 se per ogni sostanza sommiamo i valori di tutte le regioni (somma unitaria sulle righe).

E la distribuzione delle frequenze relative condizionate f(sostanza|regione):

>prop.table(mr,2):

	Abruzzo	Basilicata	Calabria	Campania	Emilia-Romagna
Fungicidi	0.86882716	0.64776119	0.47447447	0.40171674	0.63347921
Insetticidi e acaricidi	0.04012346	0.06865672	0.33033033	0.09012876	0.13894967
Erbicidi	0.07098765	0.08059701	0.11111111	0.06609442	0.15754923
Vari	0.02006173	0.20298507	0.08408408	0.44206009	0.07002188

	Friuli-Venezia Giulia	Lazio	Liguria	Lombardia
Fungicidi	0.61562199	0.37285903	0.75847893	0.40449438
Insetticidi e acaricidi	0.06653809	0.06324111	0.04373644	0.07640449
Erbicidi	0.12362584	0.08432148	0.17951353	0.34382022
Vari	0.19421408	0.47957839	0.01827110	0.17528090

	Marche	Molise	Piemonte	Puglia	Sardegna
Fungicidi	0.68421053	0.608	0.67543860	0.71818182	0.74058577
Insetticidi e acaricidi	0.04887218	0.072	0.06892231	0.11818182	0.07531381
Erbicidi	0.24436090	0.224	0.22180451	0.11636364	0.10041841
Vari	0.02255639	0.096	0.03383459	0.04727273	0.08368201

	Sicilia	Toscana	Trentino-Alto Aldige	Umbria
Fungicidi	0.62659574	0.79844961	0.35256410	0.82215743
Insetticidi e acaricidi	0.09574468	0.04844961	0.23076923	0.02040816
Erbicidi	0.05425532	0.11240310	0.01282051	0.13994169
Vari	0.22340426	0.04069767	0.40384615	0.01749271

	Valle d'Aosta	Veneto
Fungicidi	0.792139738	0.738219895
Insetticidi e acaricidi	0.085589520	0.230366492
Erbicidi	0.112663755	0.026178010
Vari	0.009606987	0.005235602

Questa tabella ci dà 1 se per ogni regione sommiamo i valori di tutte le sostanze (somma unitaria sulle colonne).

GRAFICI DI VARIABILI QUANTITATIVE:

Tipi di grafico:

- Grafico a bastoncini
- Grafico a barre
- Grafico a torta

Nel nostro caso non ha senso considerare dei grafici a bastoncini per visualizzare le frequenze delle sostanze, poiché ovviamente siccome stiamo parlando di un quantitativo preciso in kg/ettaro nessuna regione avrà un numero esattamente uguale ad un'altra. Per questo non considereremo questo tipo di grafici e preferiremo invece quelli a barre e a torta.

Grafici a barre e a torta in R:

Inserimento matrice:

```
> m<-
cbind(c(5.63,0.26,0.46,0.13),c(2.17,0.23,0.27,0.68),c(1.58,1.10,0.37,0.28),c(4.68,1.05,0.77,5.15),
c(5.79,1.27,1.44,0.64),c(31.92,3.45,6.41,10.07),c(2.83,0.48,0.64,3.64),c(66.42,3.83,15.72,1.60),c
(1.80,0.34,1.53,0.78),c(1.82,0.13,0.65,0.06),c(0.76,0.09,0.28,0.12),c(5.39,0.55,1.77,0.27),c(3.95,
0.65,0.64,0.26),c(1.77,0.18,0.24,0.20),c(5.89,0.90,0.51,2.10),c(4.12,0.25,0.58,0.21),c(0.55,0.36,0.
02,0.63),c(2.82,0.07,0.48,0.06),c(18.14,1.96,2.58,0.22),c(1.41,0.44,0.05,0.01))
> colnames(m)<-
c("Abr","Bas","Cal","Cam","EmRo","FrVen","Laz","Lig","Lom","Mar","Mol","Pie","Pug","Sar","Sic","Tos","TreA
lt","Umb","ValAo","Ven")
```

Definizione delle righe:

```
> Fun<-m[1,]
> Ins<-m[2,]
> Erb<-m[3,]
> Va<-m[4,]
```

Grafici a barre (per sostanza):

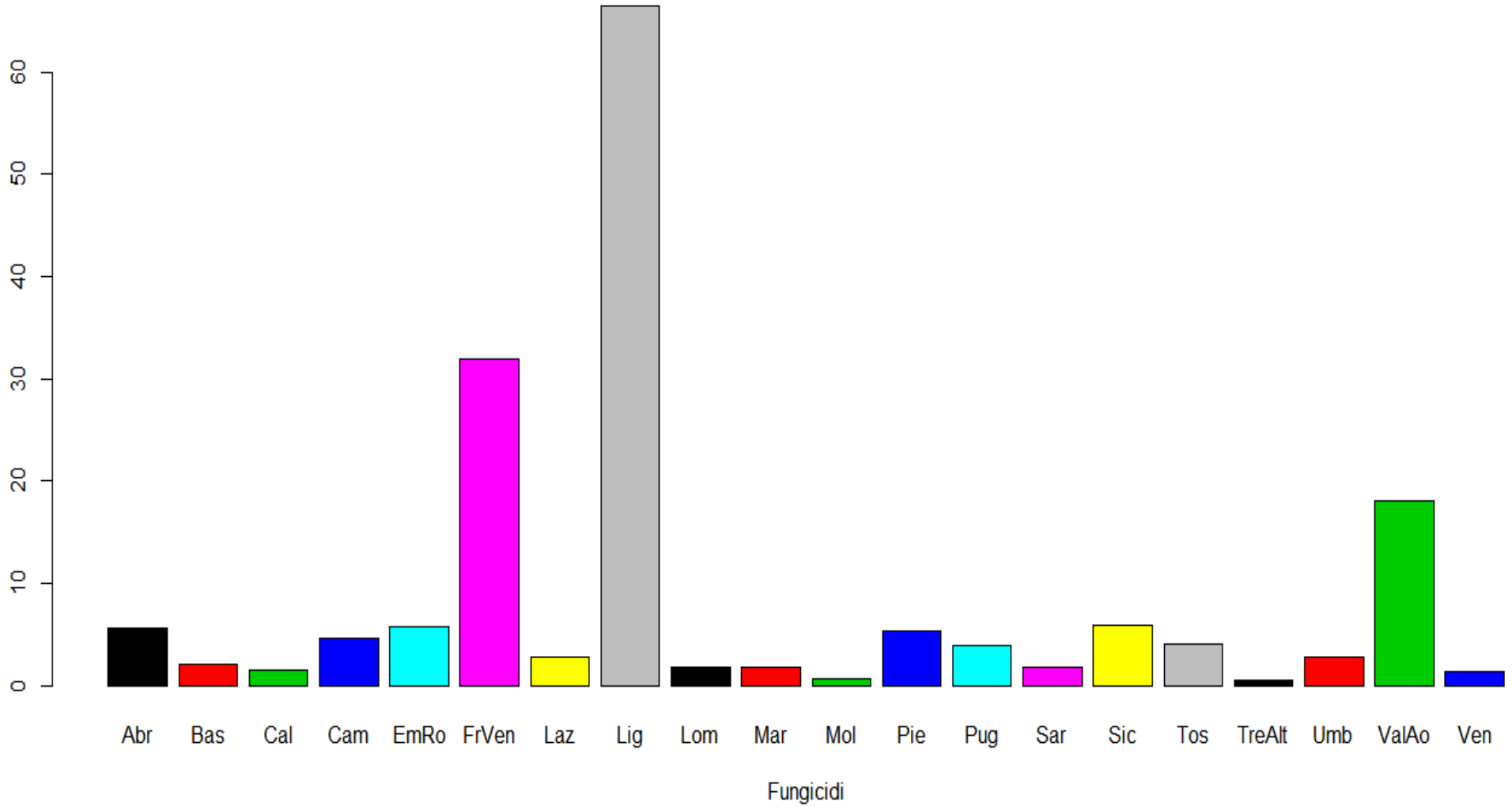
```
> barplot (Fun , xlab="Fungicidi", col =1:20)
> barplot (Ins , xlab="Insetticidi e Acaricidi", col =1:20)
> barplot (Erb , xlab="Erbicidi", col =1:20)
> barplot (Va , xlab="Vari", col =1:20)
```

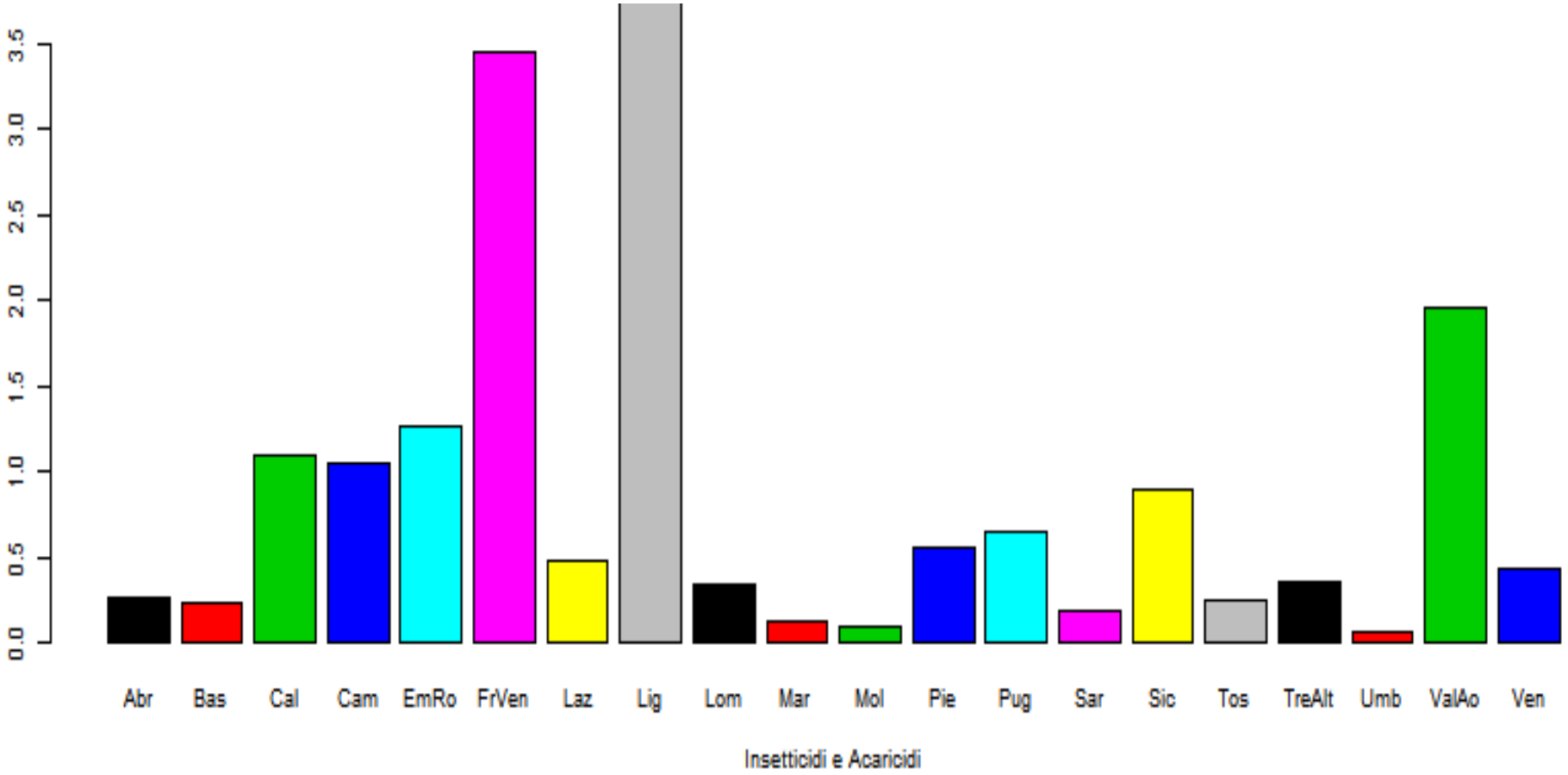
Sulle ordinate ci sono sempre i kg/ettaro.

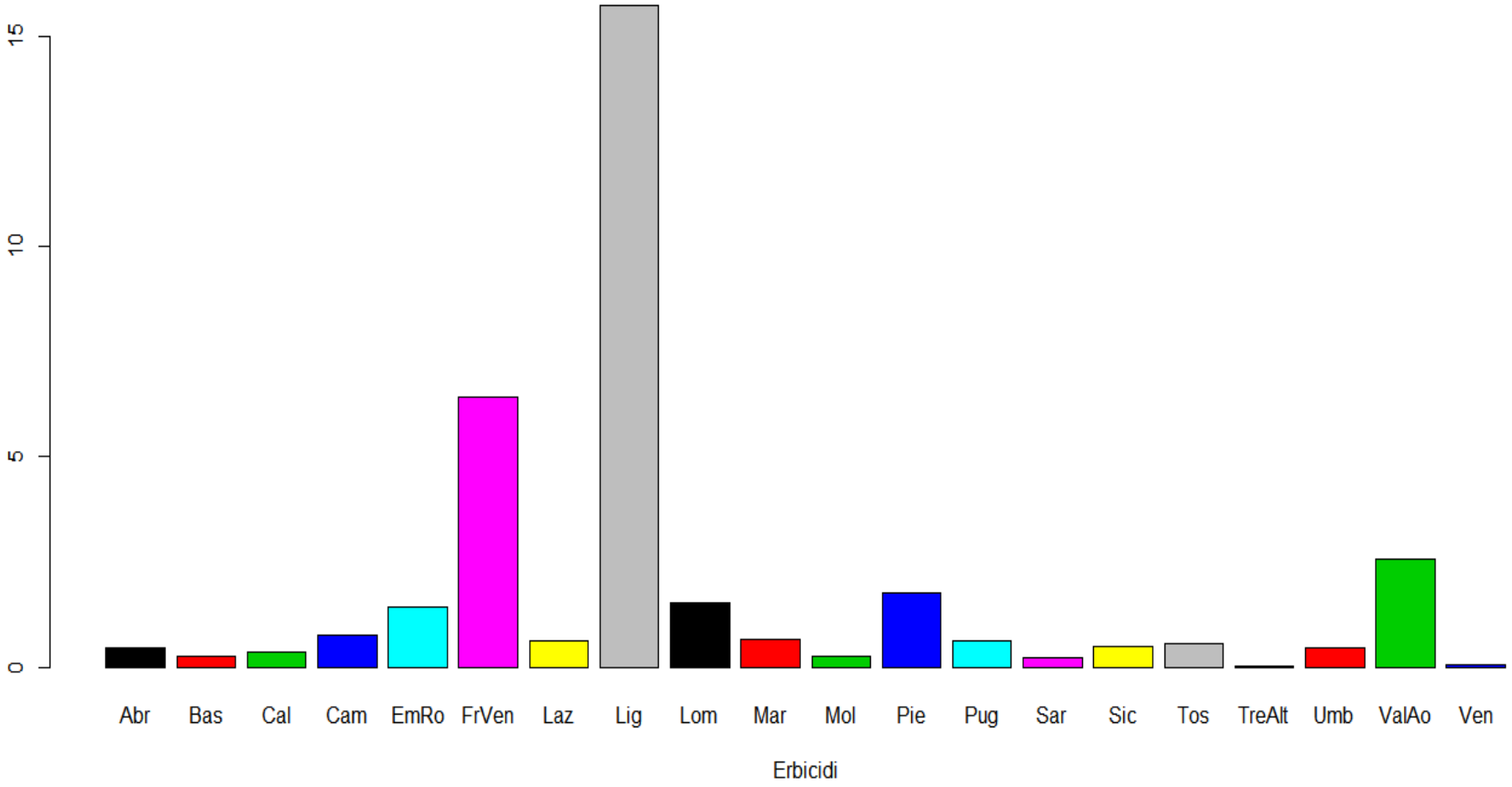
Grafici a torta (per sostanza):

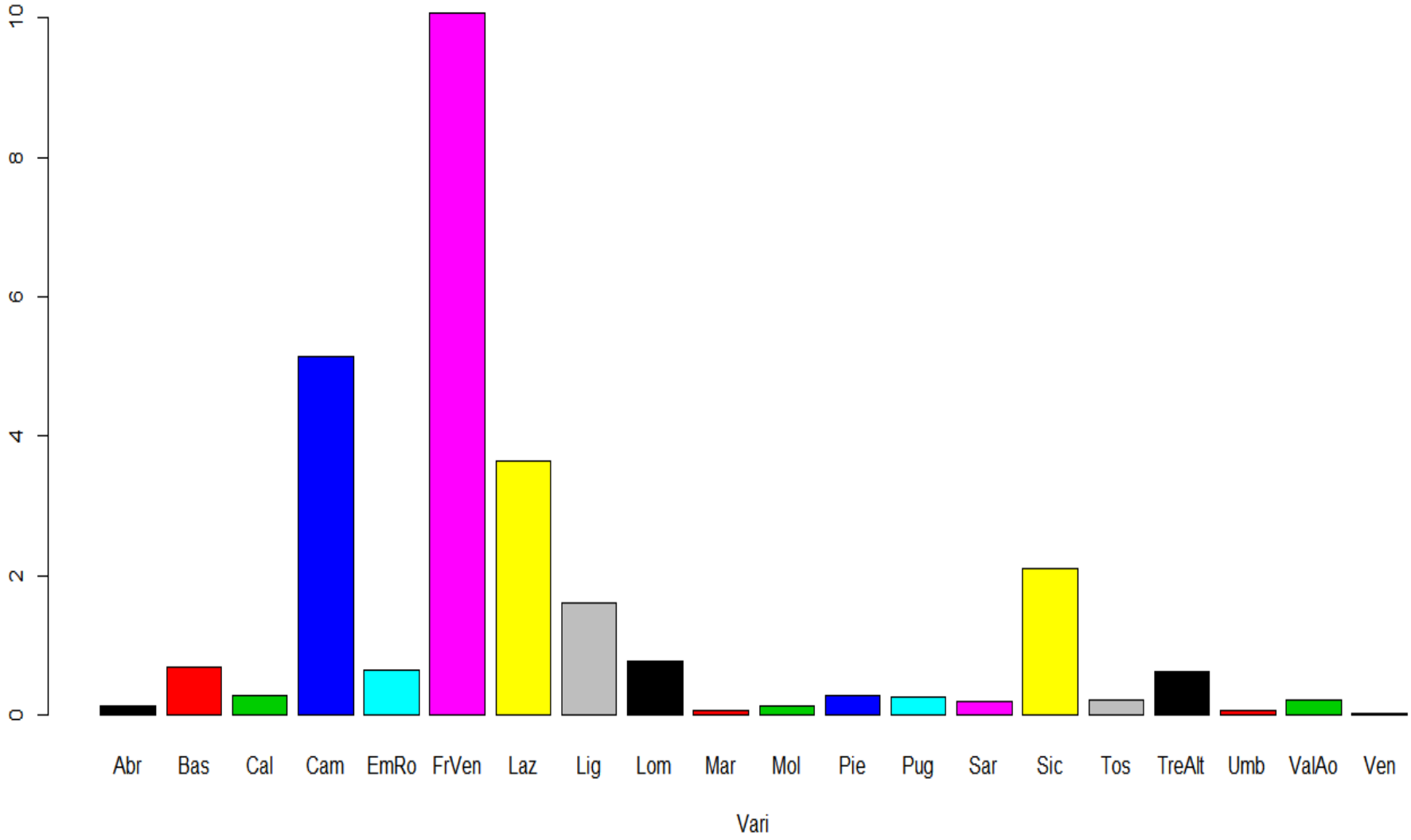
```
> pie (Fun , xlab="Fungicidi", col =1:20)
> pie(Ins , xlab="Insetticidi e Acaricidi", col =1:20)
> pie(Erb , xlab="Erbicidi", col =1:20)
> pie(Va , xlab="Vari", col =1:20)
```

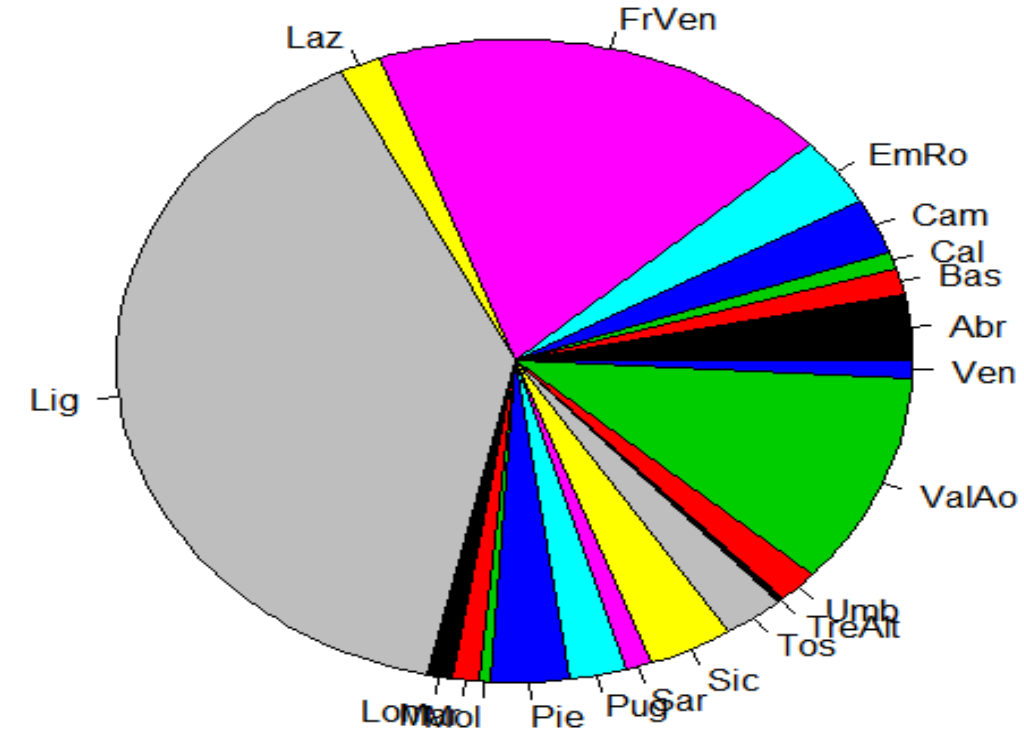
Possiamo osservare i risultati nelle pagine seguenti.


0
10
20
30
40
50
60
Abr
Bas
Cal
Cam
EmRo
FrVen
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAlt
Umb
ValAo
Ven
Fungicidi

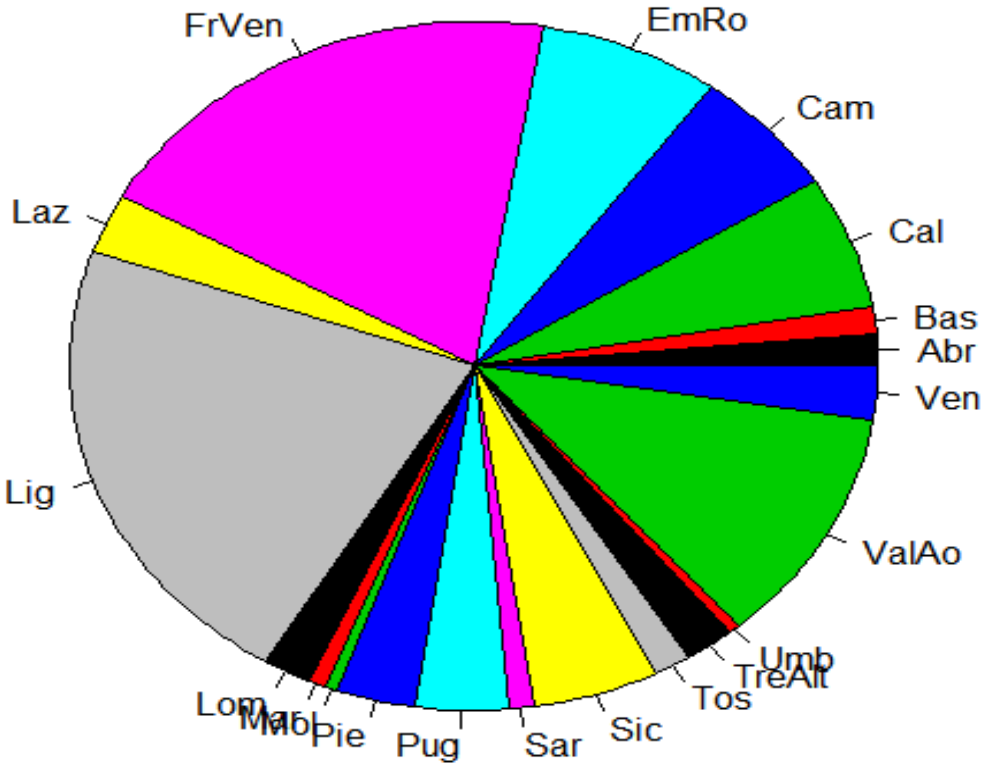

0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Abr
Bas
Cal
Cam
EmRo
FrVen
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAlt
Umb
ValAo
Ven
Insetticidi e Acaricidi


0
5
10
15
Abr
Bas
Cal
Cam
EmRo
FrVen
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAlt
Umb
ValAo
Ven
Erbicidi

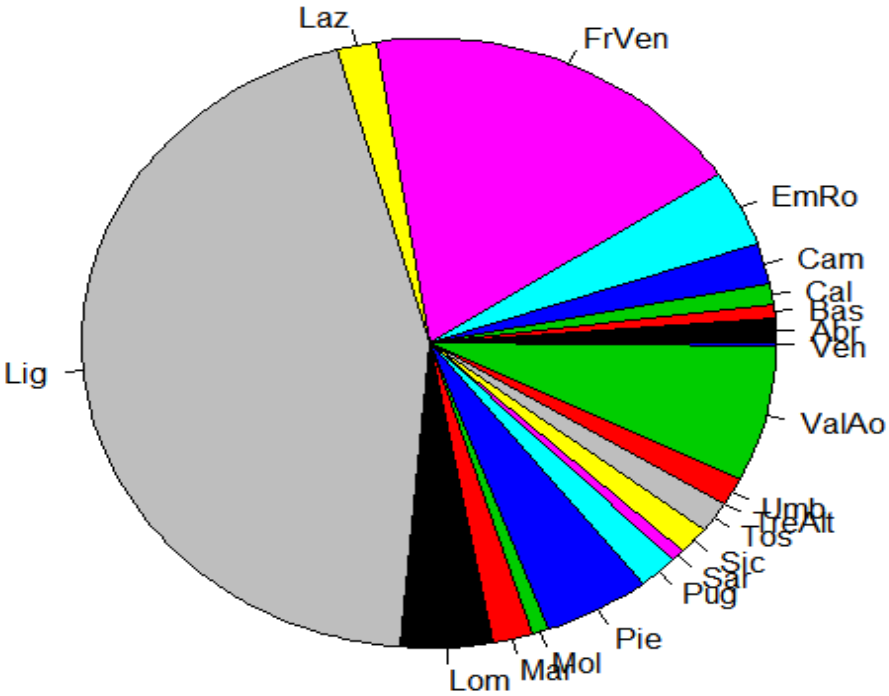

0
2
4
6
8
10
Abr
Bas
Cal
Cam
EmRo
FrVen
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAlt
Umb
ValAo
Ven
Vari



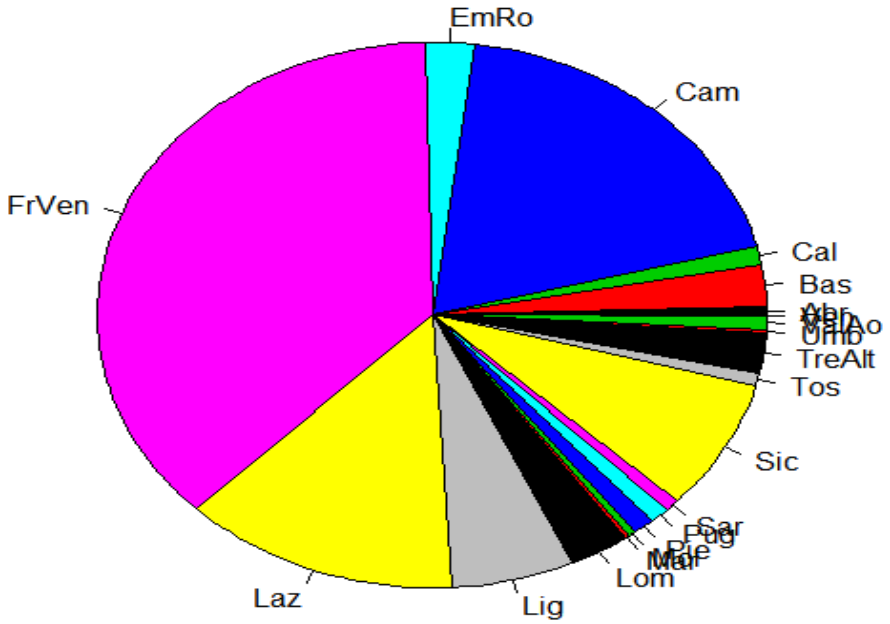
Fungicidi



Insetticidi e Acaricidi



Erbicidi



Vari

Analizzando i risultati ottenuti risulta, da entrambi i tipi di grafico che:

La regione che fa più uso di:

- Fungicidi: Liguria
- Insetticidi-Acaricidi: Liguria
- Erbicidi: Liguria
- Vari: Friuli-Venezia Giulia

La regione che fa meno uso di:

- Fungicidi: Trentino-Alto Adige; Molise (dai grafici non riusciamo a stabilire quale delle due)
- Insetticidi-Acaricidi: Umbria
- Erbicidi: Basilicata; Molise (dai grafici non riusciamo a stabilire quale delle due)
- Vari: Veneto

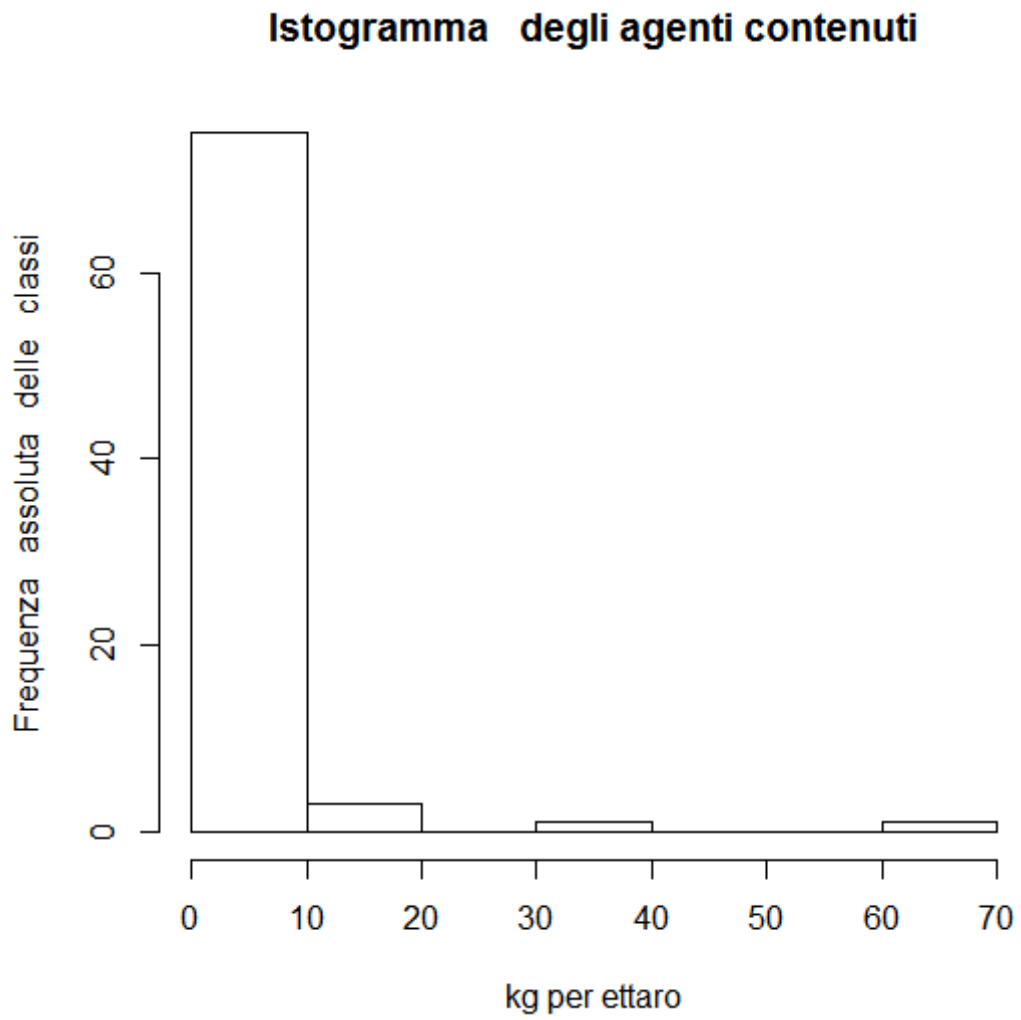
ISTOGRAMMI:

Nel calcolo di R per l'istogramma viene fatta una automatica divisione in classi.
Sulle ordinate avremo le frequenze (relative o assolute delle classi, a seconda della matrice scelta), sulle ascisse invece sempre l'unità di misura, nel nostro caso kg/ettaro. Fissate quindi le basi, le altezze devono essere tali che l'area di ogni rettangolo sia uguale alla frequenza (relativa o assoluta) della classe scelta.
Per cui applicandolo alla matrice m risulta:

```
> h<-hist(m, freq=TRUE ,main=" Istogramma  degli agenti contenuti ", ylab=" Frequenza  assoluta delle  classi ",xlab="kg per ettaro")
```

Volendo vedere quale sia stata la divisione in classi otteniamo:

```
> str(h)
List of 6
 $ breaks : num [1:8] 0 10 20 30 40 50 60 70
```



$ counts : int [1:7] 75 3 0 1 0 0 1
$ density : num [1:7] 0.09375 0.00375 0 0.00125 0 ...
$ mids : num [1:7] 5 15 25 35 45 55 65
$ xname : chr "m"
$ equidist: logi TRUE
– attr(*, "class")= chr "histogram"

Cioè fornisce i punti di suddivisione in classi (breaks), le frequenze assolute delle classi (counts), la densità delle classi (density) e i punti centrali delle classi (mids).

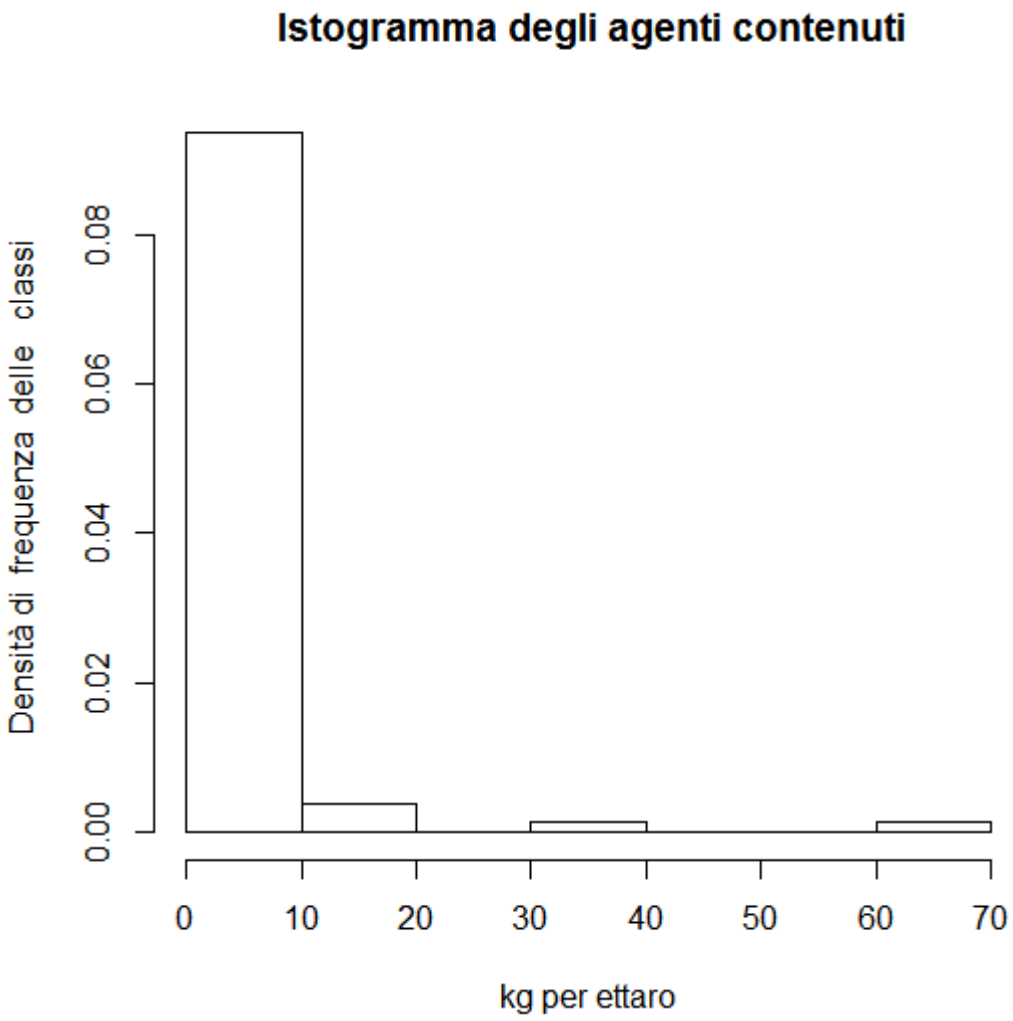
L'istogramma rappresenta le frequenze assolute delle classi. Per ottenere le relative possiamo moltiplicare gli elementi del vettore h$density per 10 (ampiezza effettiva di ogni classe).

```
> f <-10*h$ density
> f
[1] 0.9375 0.0375 0.0000 0.0125 0.0000 0.0000 0.0125
> sum(f)
[1] 1
```

la cui somma è infatti proprio unitaria, come ci aspettiamo che sia.

Si può poi realizzare l'istogramma delle frequenze relative in questo modo:

```
> hist(m ,freq =FALSE , main =" Istogramma degli agenti contenuti",xlab="kg per ettaro",
ylab="Densità di frequenza delle classi ")
```



Dagli istogrammi notiamo complessivamente che la maggior parte degli agenti contenuti nei prodotti fitosanitari non superano i 10kg/ettaro, alcuni arrivano fino ai 20kg/ettaro e solo alcuni hanno valori più elevati, compresi cioè tra 30 e 40kg/ettaro e i 60 e 70kg/ettaro.

BOXPLOT:

Per il boxplot, nello scopo di avere dati più significativi, divideremo lo studio per i 4 agenti contenuti nei prodotti fitosanitari, quindi Fungicidi, Insetticidi e Acaricidi, Erbicidi e Vari.

Tramite il comando quantile otteniamo:

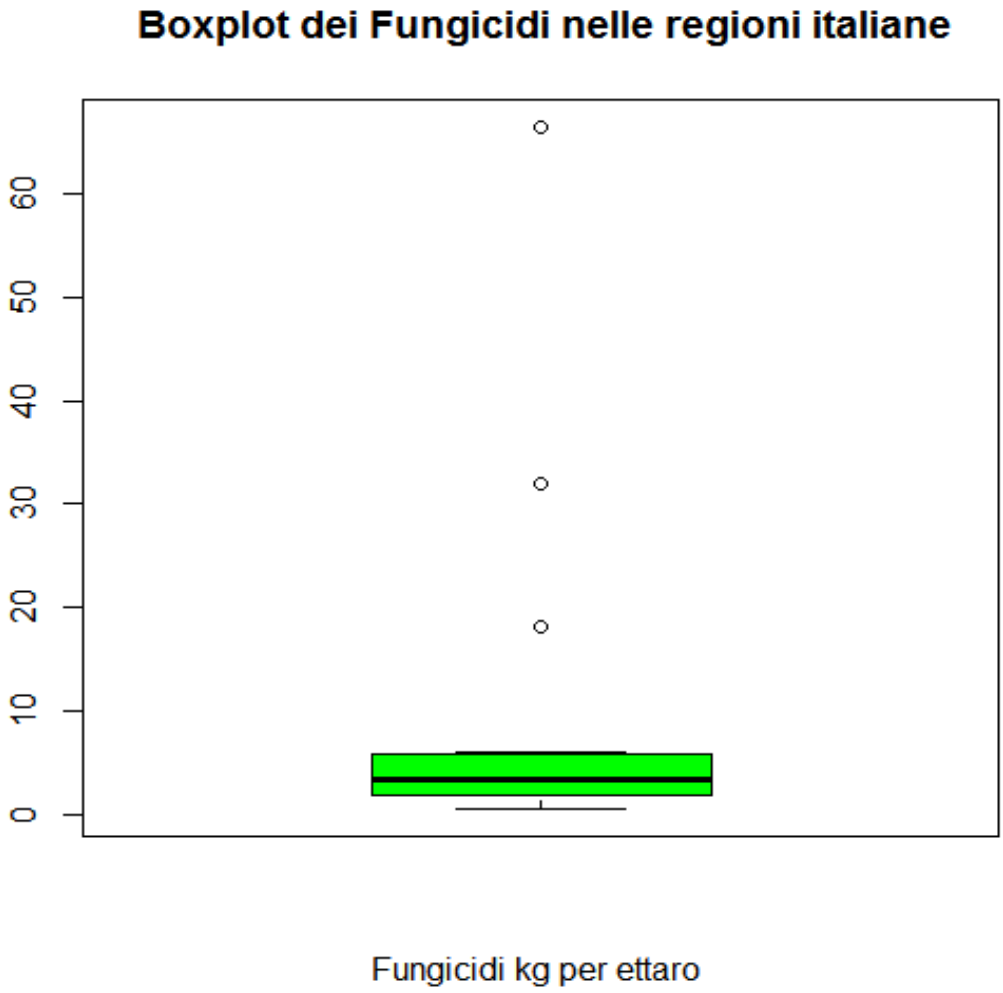
```
>Fungicidi<-
c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82,0.76,5.39,3.95,1.77,5.89,4.12,0.55,2.82,18
.14,1.41)
> quantile(Fungicidi)
    0%    25%    50%    75%   100%
 0.5500  1.7925  3.3900  5.6950 66.4200
```

Il primo e l'ultimo valore rappresentano il minimo e il massimo numero di kg per ettaro, cioè Q0 e Q4. I valori di 25%,50% e 75% rappresentano rispettivamente i valori per cui il 25%,50% e 75% dei dati sono alla loro sinistra e vengono chiamati rispettivamente primo, secondo e terzo quartile (Q1,Q2 e Q3). Notiamo che Q2 rappresenta proprio la mediana dei dati.
Infatti analizzando gli indici otteniamo:

```
> summary(Fungicidi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.550   1.792   3.390   8.481   5.695  66.420
```

Questo comando ci fornisce anche la media.

Ora disegnamo il boxplot e facciamo alcune considerazioni osservando il grafico:
>boxplot (Fungicidi ,xlab ="Fungicidi kg per ettaro ",main =" Boxplot dei Fungicidi nelle regioni italiane",col="green")



L'aspetto più interessante di questo boxplot sono i valori contrassegnati con dei cerchietti.
Essi rappresentano dei valori anomali.
Notiamo quindi che c'è un uso molto elevato di Fungicidi che va ben oltre quello delle altre regioni in 3 regioni in particolare. Possiamo scoprire quali sono osservando la tabella dei dati, anche vista a pagina 2:

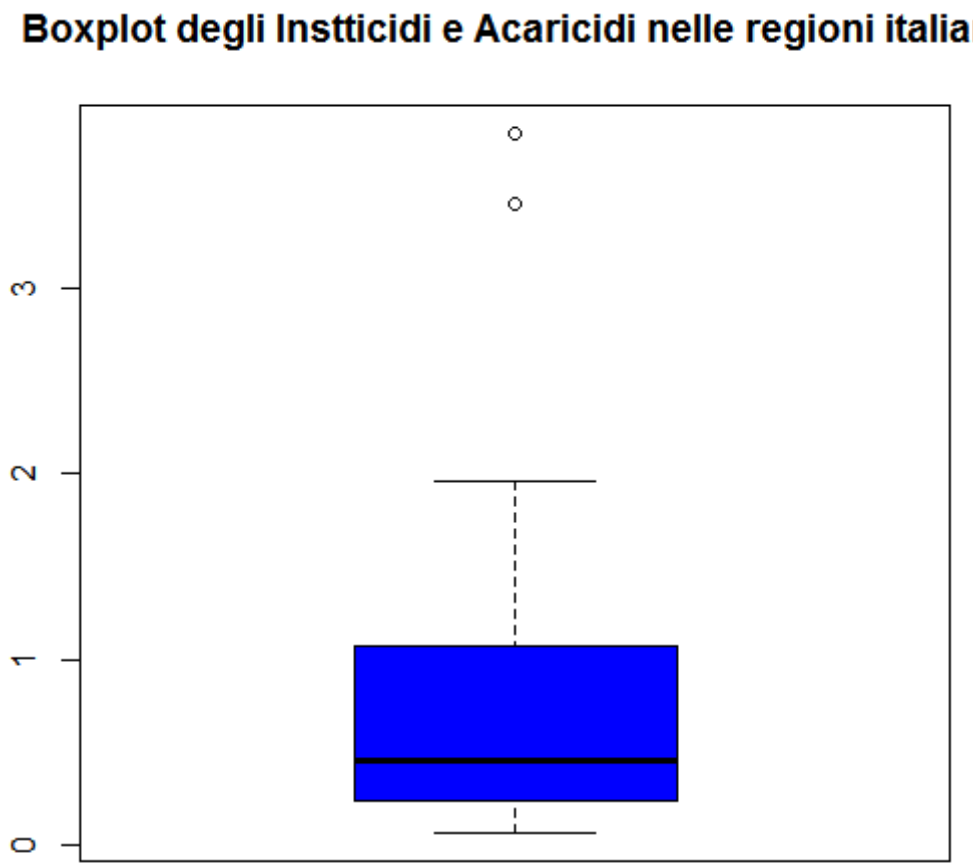
Valle D'Aosta: 18,14 kg per ettaro
Friuli-Venezia Giulia: 31,92 kg per ettaro
Liguria: 66,42 kg per ettaro

Queste "anomalie" potrebbero essere dovute ad una errato raccoglimento dei dati o, come si può facilmente supporre, evidenziano delle regioni con un significativo livello di utilizzo di Fungicidi.

Le aree colorate rappresentano le regioni con un utilizzo di kg di Fungicidi per ettaro fuori dalla norma, ciò ci porta a pensare che le piantagioni di queste zone siano particolarmente infestate da felci, muschi, licheni, epatiche e sfagni.

Passiamo ora al boxplot riguardante gli insetticidi e gli acaricidi:

```
>InsAca<-
c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13,0.09,0.55,0.65,0.18,0.90,0.25,0.36,0.07,1.96,
0.44)
> quantile(InsAca)
    0%    25%    50%    75%   100%
0.0700 0.2450 0.4600 1.0625 3.8300
> summary(InsAca)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0700  0.2450  0.4600  0.8795  1.0620  3.8300

>  boxplot (InsAca ,xlab ="Insetticidi e acaricidi kg per ettaro ",main =" Boxplot degli Instticidi e
Acaricidi nelle regioni italiane",col="blue")
```



La situazione per gli insetticidi e acaricidi è meglio distribuita, infatti sono ben evidenti sia il baffo superiore che quello inferiore.
Possiamo poi notare dai dati forniti dai comandi che la mediana è la metà della media, questo evidenzia che non c'è una buona *centralità* dei dati. Di conseguenza, osservando le distanze tra primo e terzo quartile dalla linea mediana capiamo che non c'è neppure una grande *simmetria.*
C'è anche abbastanza *dispersione* poiché la distanza tra il baffo inferiore e il primo quartile, e quello superiore e il terzo quartile non è proporzionata.
Infine, come nel caso precedente abbiamo delle anomalie, questa volta meno "gravi" delle precedenti e corrispondono alle regioni:

Friuli-Venezia Giulia: 3,45 kg per ettaro
Liguria: 3,83 kg per ettaro

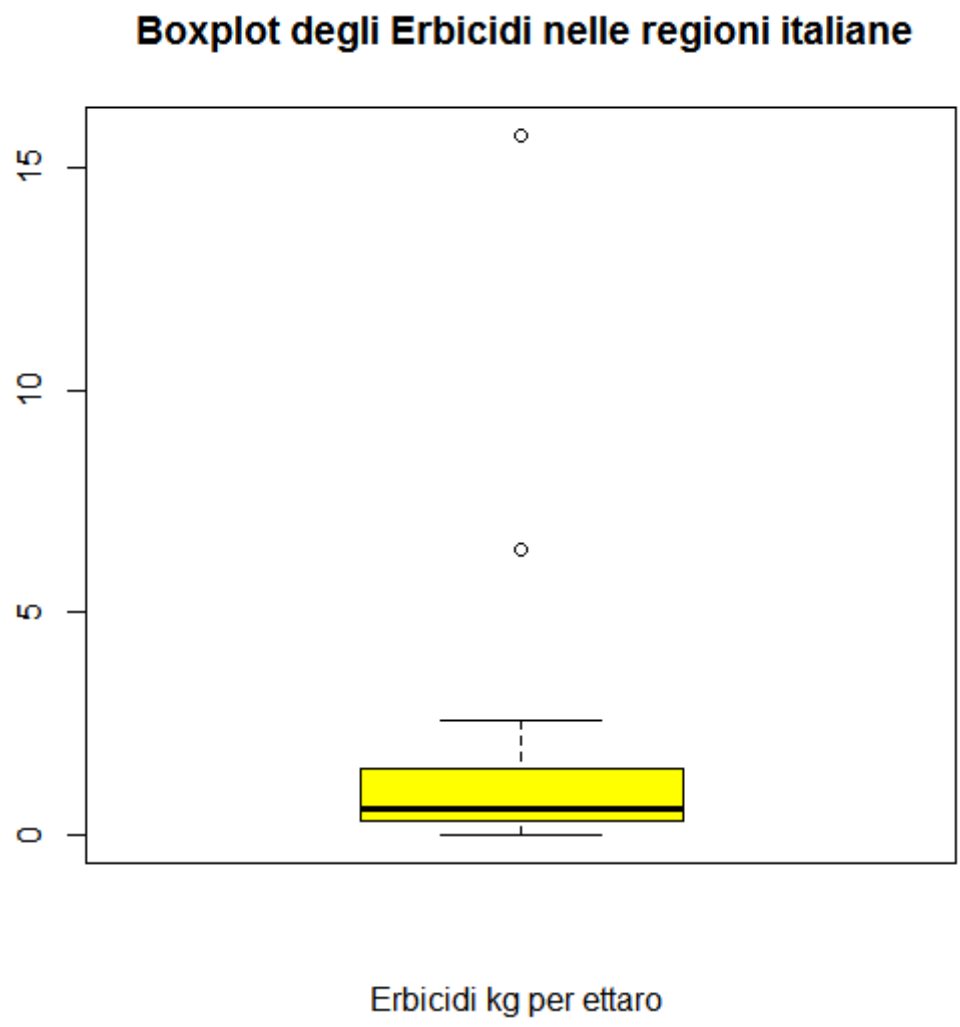
Sono anche le medesime regioni del caso dei Fungicidi precedente, ma questa volta la regione Valle D'Aosta rientra nei parametri.

Probabilmente, ancora una volta, possiamo dire che queste due regioni sono particolarmente soggette ad attacchi di afidi e aleurodidi (mosche bianche).

Boxplot per Erbicidi:

```
> Erbicidi<-
c(0.46,0.27,0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65,0.28,1.77,0.64,0.24,0.51,0.58,0.02,0.48,2.5
8,0.05)
> quantile(Erbicidi)
    0%    25%    50%    75%   100%
0.0200 0.3475 0.6100 1.4625 15.7200
> summary(Erbicidi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0200 0.3475 0.6100 1.7700 1.4630 15.7200
```

```
>  boxplot (Erbicidi,xlab ="Erbicidi kg per ettaro ",main =" Boxplot degli Erbicidi nelle regioni
italiane",col="yellow")
```



Anche in questo caso abbiamo una situazione analoga al caso precedente, dove però i valori anomali si discostano di più dall'estremità del baffo.
Le regioni coinvolte sono:

Friuli-Venezia Giulia: 6,41 kg per ettaro
Liguria: 15,72 kg per ettaro.

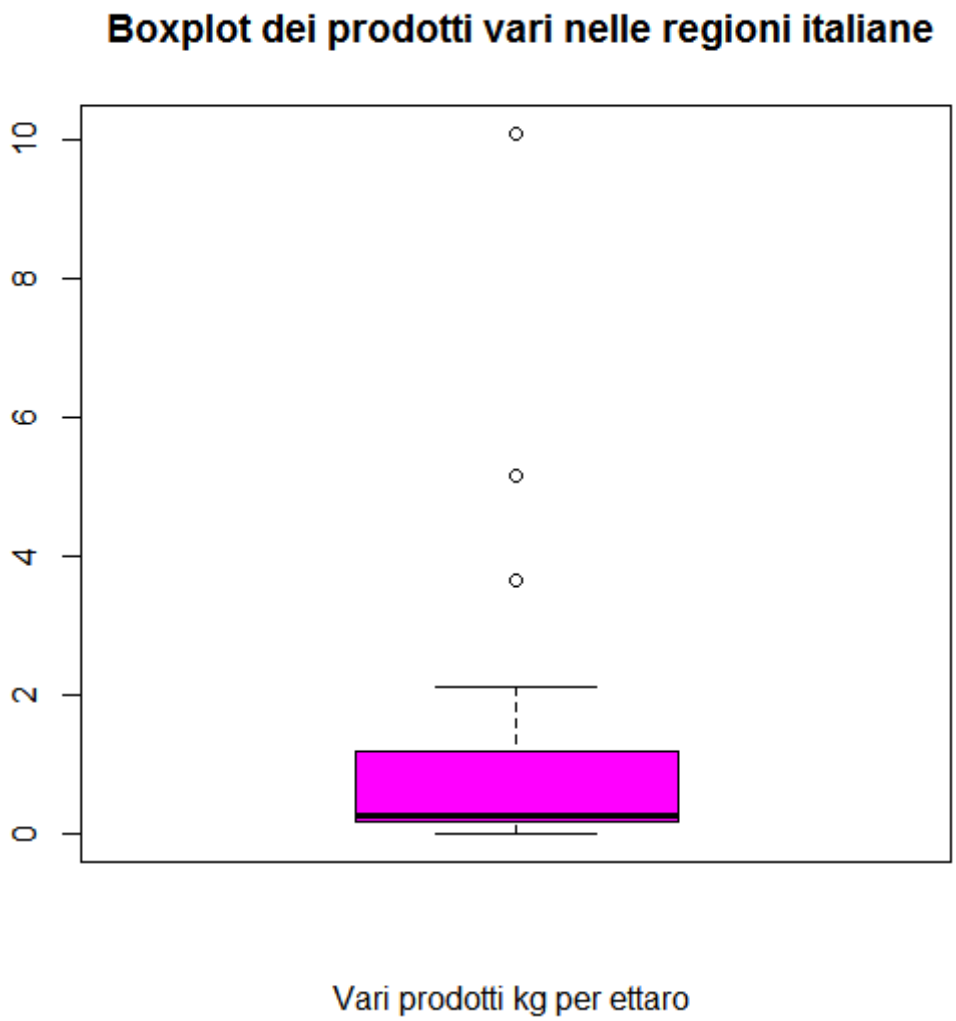
Dalle informazioni degli ultimi tre boxplot potremmo dedurre che:

I prodotti utilizzati dalle regioni Friuli-Venezia Giulia e dalla Liguria hanno gli stessi principi attivi
oppure
Le coltivazioni del Friuli-Venezia Giulia e della Liguria hanno climi/terreni/condizioni atmosferiche più sfavorevoli e che incorrono maggiormente in problemi per cui ci sia il bisogno di utilizzare questi tre prodotti
oppure
I dati sono stati prelevati in maniera troppo generica senza tenere conto di particolari altre sostanze utilizzate nelle altre regioni.

A tale proposito andiamo ad analizzare la situazione di tutti gli altri prodotti che cadono nella categoria "Vari".

```
> Vari<-
c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06,0.12,0.27,0.27,0.20,2.10,0.21,0.63,0.06,0.2
2,0.01)
> quantile(Vari)
    0%     25%     50%     75%    100%
 0.0100  0.1825  0.2750  0.9850 10.0700
> summary(Vari)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.1825  0.2750  1.3560  0.9850 10.0700
```

>boxplot (Vari,xlab ="Vari prodotti kg per ettaro ",main =" Boxplot dei prodotti vari nelle regioni italiane",col="magenta")



La situazione in questo caso presenta poca o quasi nessuna simmetria e di conseguenza una situazione completamente mal distribuita, sia per quanto riguarda la posizione della mediana, sia per quanto riguarda l'estensione di baffo superiore e inferiore.
Ciò poteva essere sospettato dalla forte differenza tra media e mediana fornita dal comando summary.
Notiamo inoltre 3 valori anomali che corrispondono alle regioni:

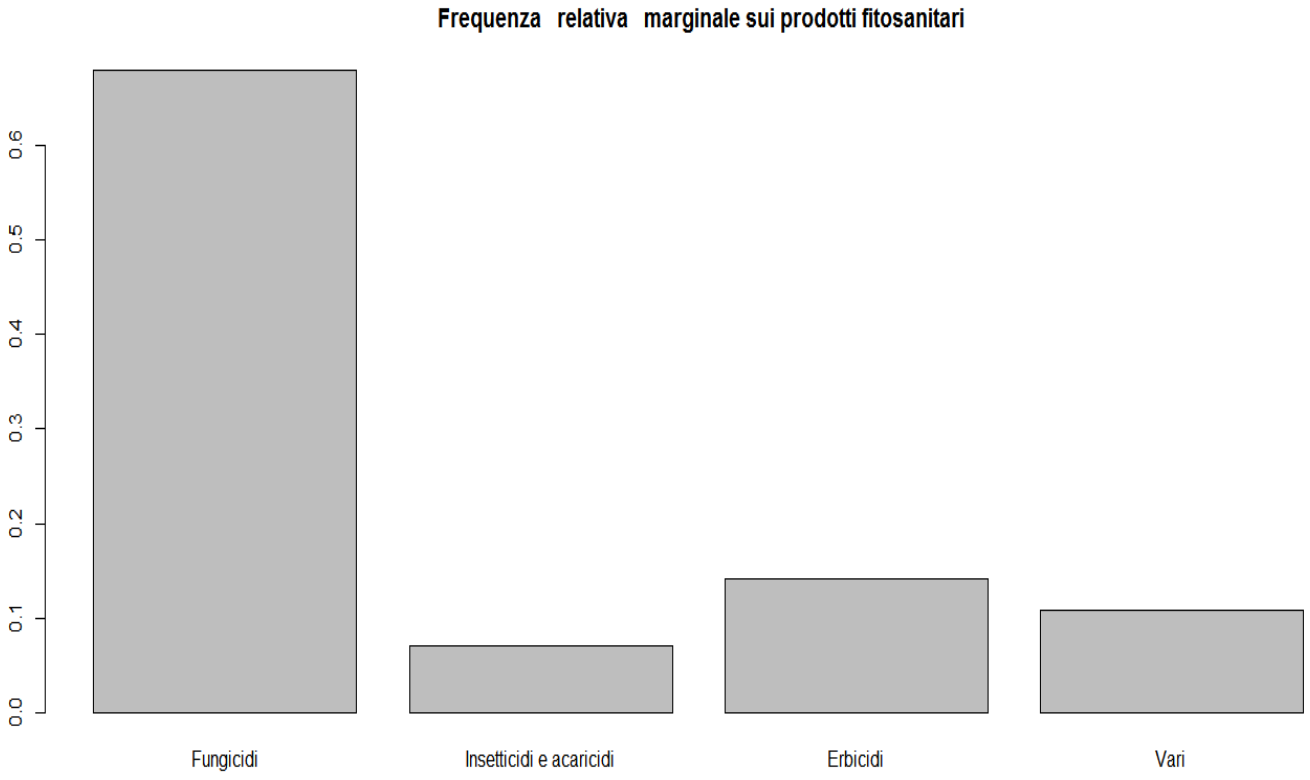
Lazio: 3,64 kg per ettaro
Campania: 5,15 kg per ettaro
Friuli-Venezia Giulia: 10,07 kg per ettaro

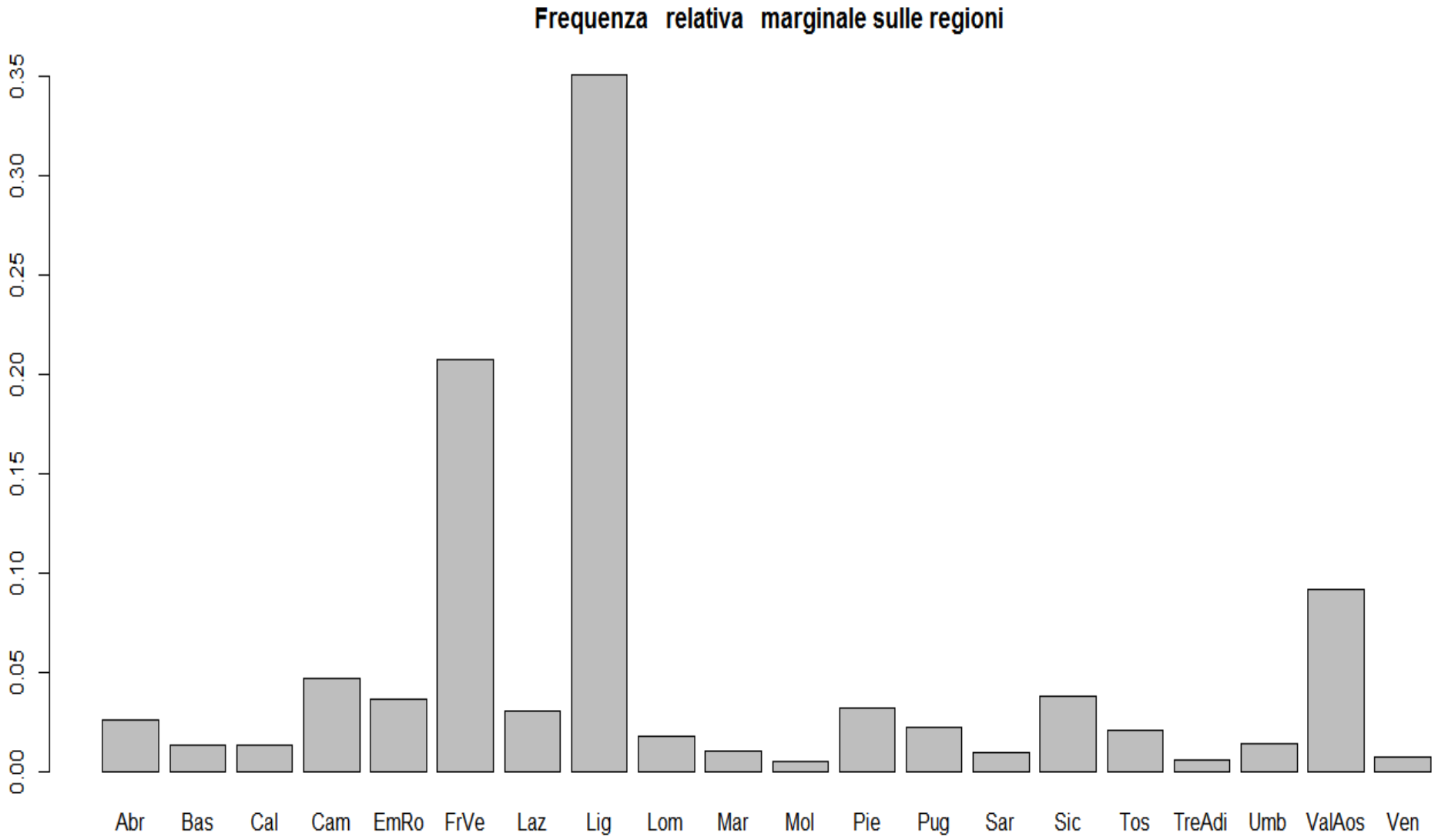
Nonostante in questo caso la regione Liguria, presente negli altri boxplot non figuri in questa categoria, si può notare facilmente che comunque i kg per ettaro utilizzati in questa ultima e nel Friuli-Venezia Giulia di prodotti fitosanitari supera la quantità considerata "normale".

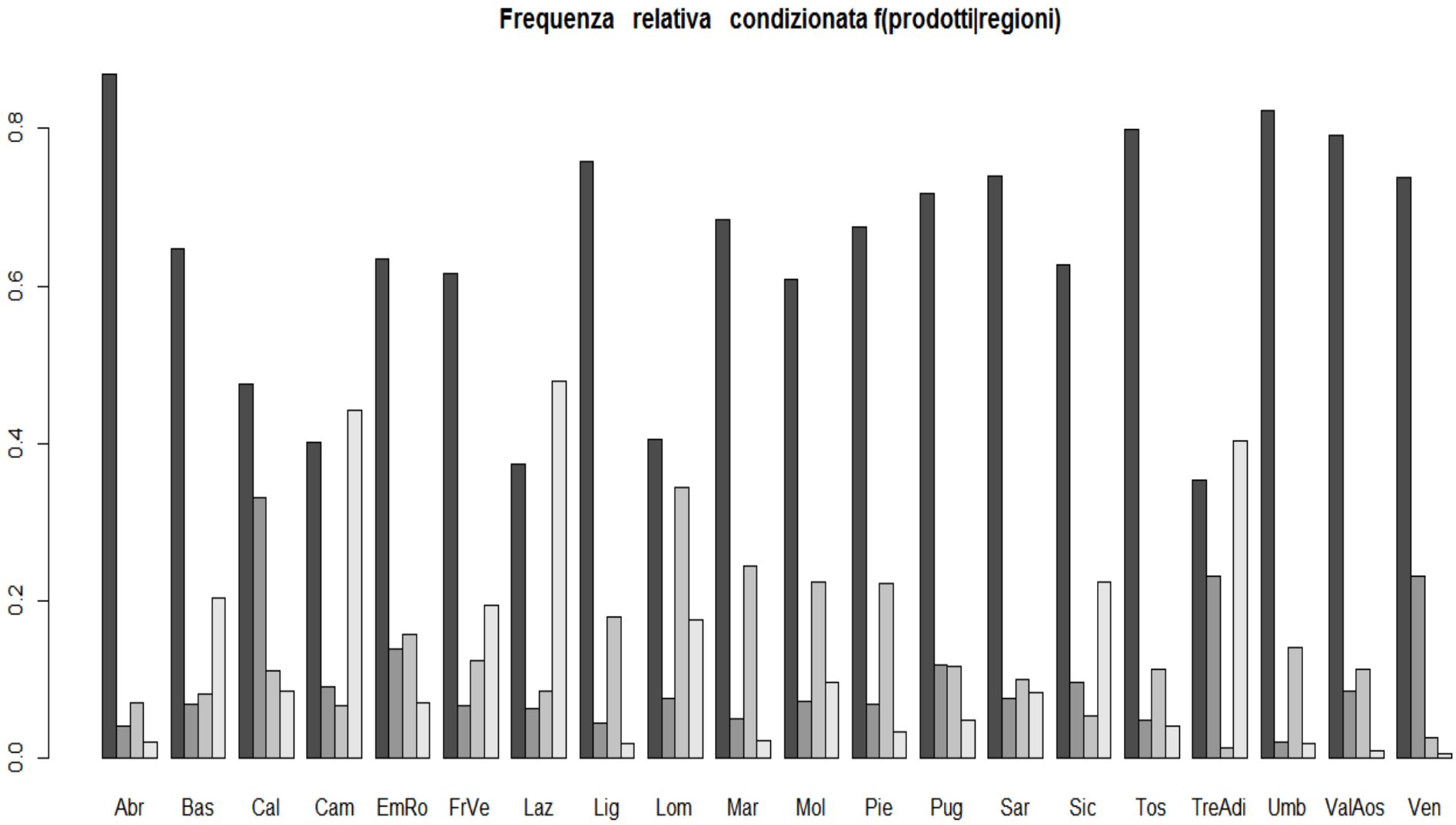
MATRICI DI DATI

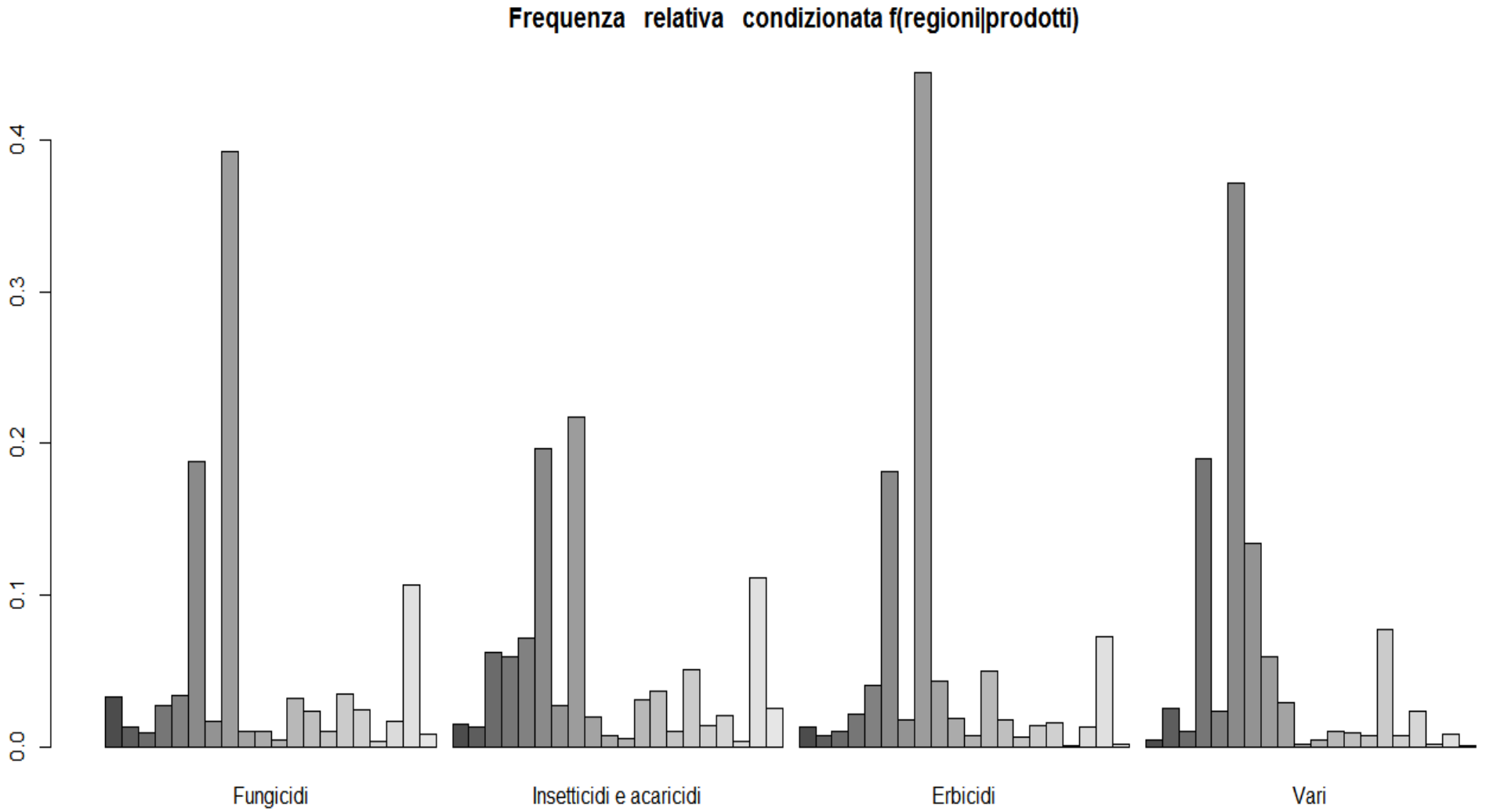
Ricollegandoci alla prima sezione trattata in cui abbiamo parlato di frequenze relative e marginali per la nostra tabella di dati numerici, andiamo a graficare le frequenze relative marginali e condizionate su righe e colonne.

```
> mr <-prop.table (m)
> barplot ( margin.table (mr ,1) ,  main=" Frequenza   relativa   marginale sui prodotti fitosanitari")
> barplot ( margin.table (mr ,2) , main=" Frequenza   relativa   marginale sulle regioni")
> barplot ( prop.table (m ,2) ,beside =TRUE , main=" Frequenza   relativa   condizionata f(prodotti|
regioni)")
>  barplot ( t(prop.table (m ,1)) ,beside =TRUE , main=" Frequenza   relativa   condizionata
f(regioni|prodotti)")
```



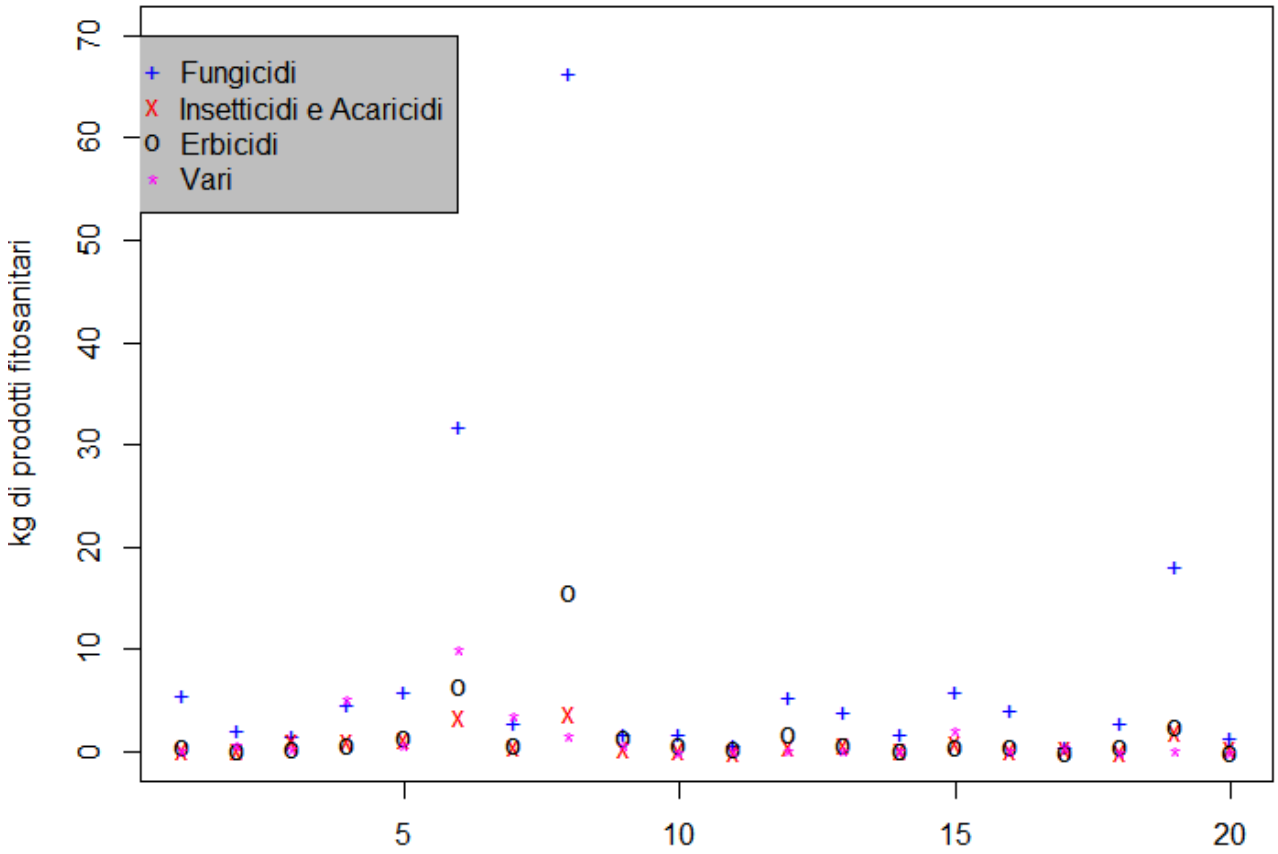

Frequenza relativa marginale sulle regioni
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Abr
Bas
Cal
Cam
EmRo
FrVe
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAdi
Umb
ValAos
Ven


Frequenza relativa condizionata f(prodotti|regioni)
0.0
0.2
0.4
0.6
0.8
Abr
Bas
Cal
Cam
EmRo
FrVe
Laz
Lig
Lom
Mar
Mol
Pie
Pug
Sar
Sic
Tos
TreAdi
Umb
ValAos
Ven


Frequenza relativa condizionata f(regioni|prodotti)
0.0
0.1
0.2
0.3
0.4
Fungicidi
Insetticidi e acaricidi
Erbicidi
Vari

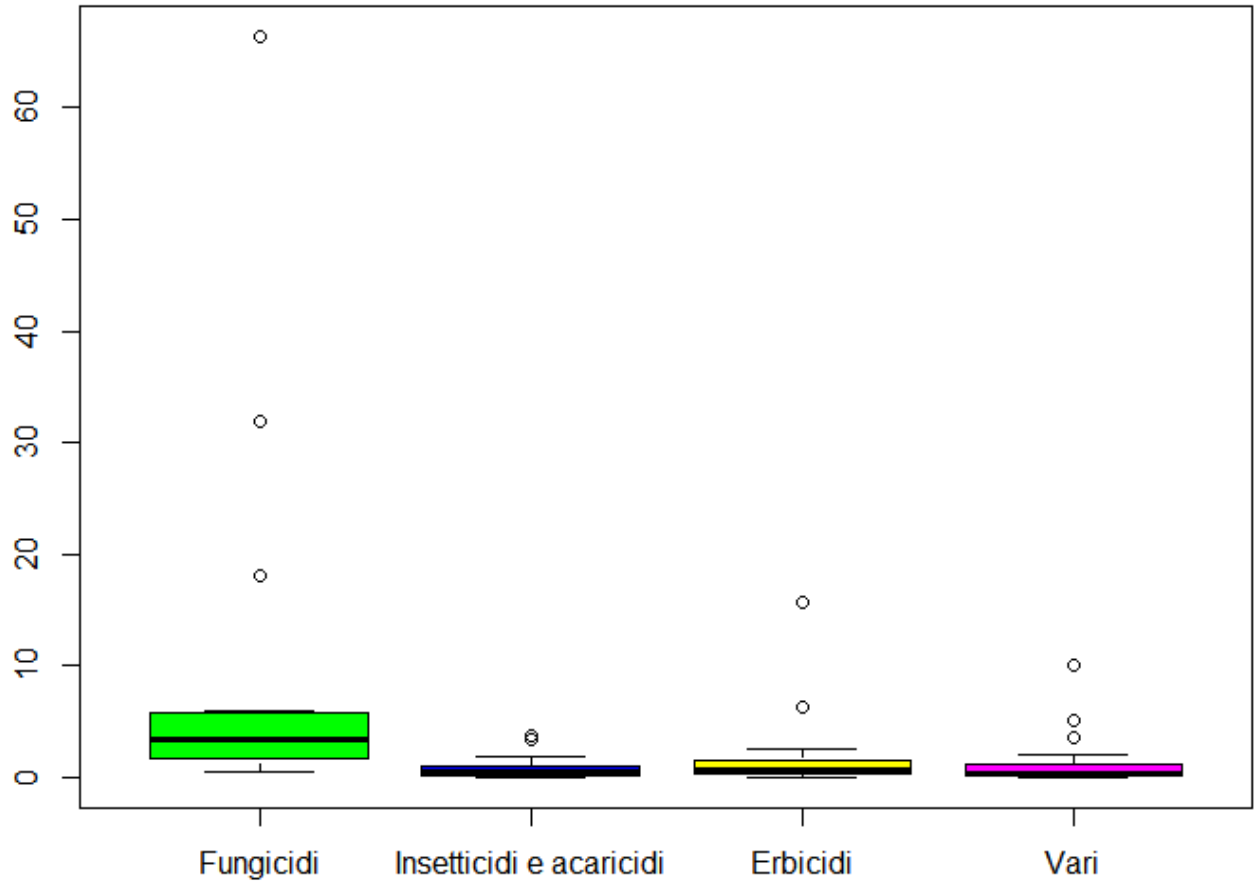
Ora facciamo un grafico congiunto con i dati a nostra disposizione:

```
> plot( Fungicidi ,pch ="+",ylim=c (0,70) ,ylab="kg di prodotti fitosanitari",col="blue")
>  points (InsAca,pch ="x",col ="red ")
>  points (Erbicidi,pch ="o",col ="black ")
>  points (Vari,pch ="*",col ="yellow")
>  points (Vari,pch ="*",col ="magenta")
>legend (0,70, c("Fungicidi","Insetticidi e Acaricidi","Erbicidi","Vari"), pch=c("+","x","o","*"),col =c("
blue","red ","black","magenta"),bg="gray ",cex =1)
```



Confrontiamo ora i dati confrontando i boxplot:

> boxplot(Fungicidi,InsAca,Erbicidi,Vari,names=c("Fungicidi","Insetticidi e acaricidi","Erbicidi","Vari"),col=c("green","blue","yellow","magenta"))



Purtroppo a causa dei valori anomali non può essere fatto un confronto su ciò che vediamo, tuttavia possiamo vedere cosa ci restituisce la funzione summary:

```
> summary(Fungicidi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.550   1.792   3.390   8.481   5.695  66.420
> summary(InsAca)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0700  0.2450  0.4600  0.8795  1.0620  3.8300
> summary(Erbicidi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0200  0.3475  0.6100  1.7700  1.4630 15.7200
> summary(Vari)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0100  0.1825  0.2750  1.3560  0.9850 10.0700
```

Dal terzo quartile di tutti e quattro notiamo che il 75% dei dati non supera i 6kg/ettaro di prodotti fitosanitari, questo vuol dire che laddove vi sono dei numeri particolarmente elevati, come nel caso dei Fungicidi, essi sono delle eccezioni, e non sono quindi rappresentativi della situazione complessiva dei prodotti fitosanitari in Italia.

SCATTERPLOT:

Gli scatterplot, o anche diagrammi di dispersione, mettono in evidenza le relazioni tra le variabili, in cui ogni coppia di osservazioni viene rappresentata sotto forma di un punto in un piano euclideo.

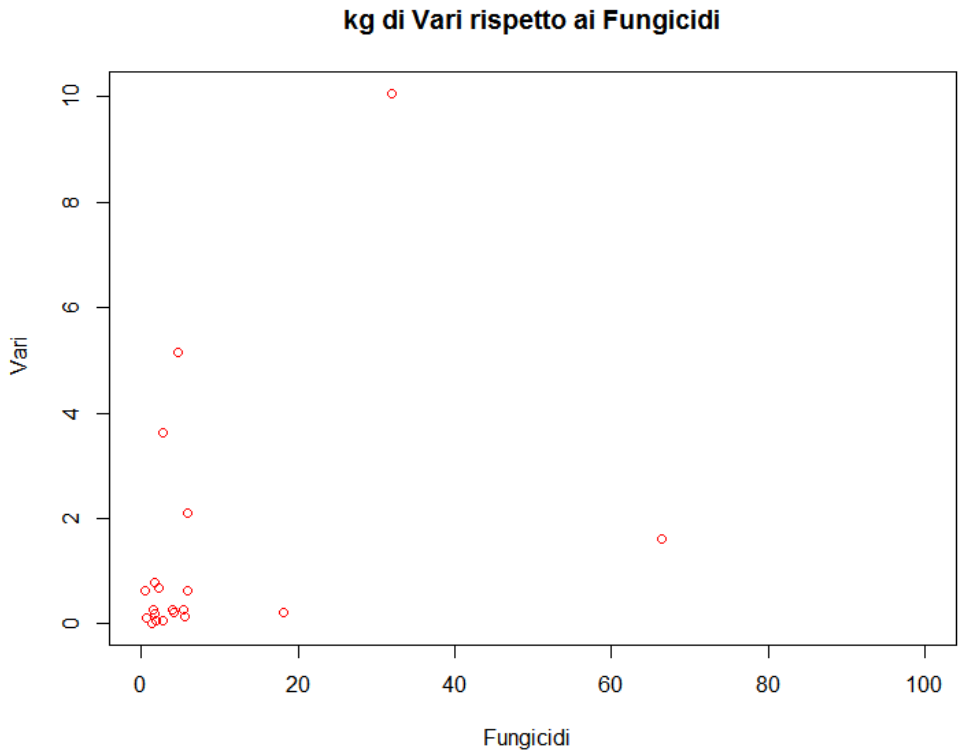
Sull'asse delle ascisse viene posta la variabile indipendente, sulle ordinate la dipendente, otteniamo così una nuvola di punti di cui siamo interessati a studiare, se esiste, una qualche regolarità.

```
>df<-
data.frame(Fungicidi=c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82,0.76,5.39,3.95,1.77,
5.89,4.12,0.55,2.82,18.14,1.41),
InsAca=c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13,0.09,0.55,0.65,0.18,0.90,0.25,0.36,0.
07,1.96,0.44),
Erbicidi=c(0.46,0.27,0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65,0.28,1.77,0.64,0.24,0.51,0.58,0.02,
0.48,2.58,0.05),
Vari=c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06,0.12,0.27,0.27,0.20,2.10,0.21,0.63,0.0
6,0.22,0.01))
> df
```

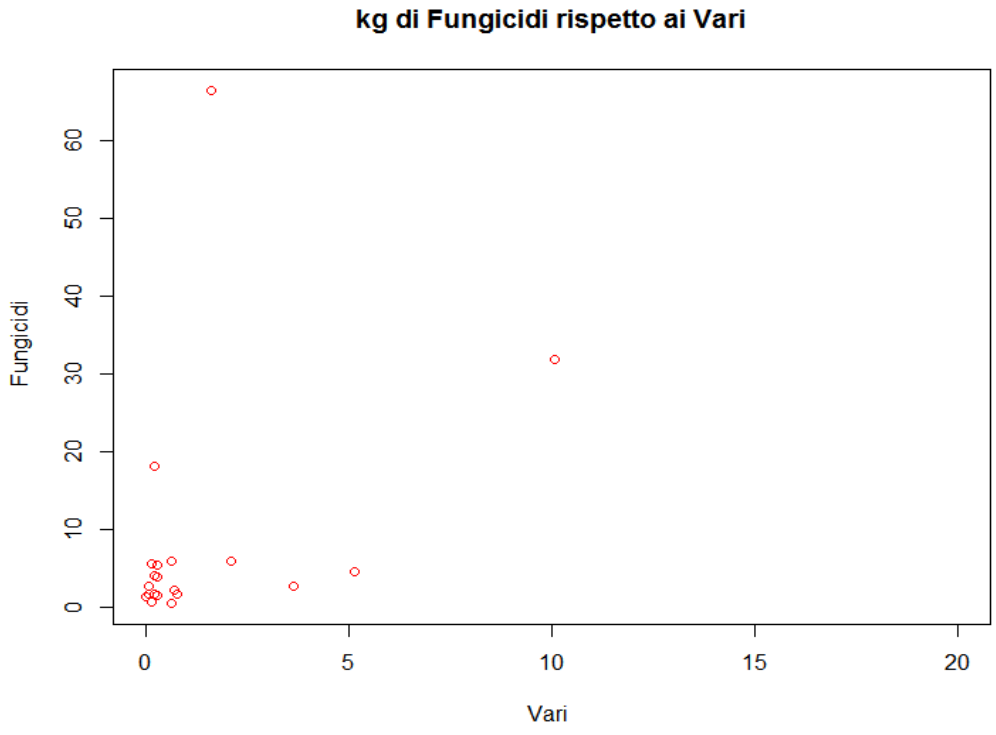
	Fungicidi	InsAca	Erbicidi	Vari
1	5.63	0.26	0.46	0.13
2	2.17	0.23	0.27	0.68
3	1.58	1.10	0.37	0.28
4	4.68	1.05	0.77	5.15
5	5.97	1.27	1.44	0.64
6	31.92	3.45	6.41	10.07
7	2.83	0.48	0.64	3.64
8	66.42	3.83	15.72	1.60
9	1.80	0.34	1.53	0.78
10	1.82	0.13	0.65	0.06
11	0.76	0.09	0.28	0.12
12	5.39	0.55	1.77	0.27
13	3.95	0.65	0.64	0.27
14	1.77	0.18	0.24	0.20
15	5.89	0.90	0.51	2.10
16	4.12	0.25	0.58	0.21
17	0.55	0.36	0.02	0.63
18	2.82	0.07	0.48	0.06
19	18.14	1.96	2.58	0.22
20	1.41	0.44	0.05	0.01

Potremmo a questo punto essere interessati ai kg di Vari componenti dati i Fungicidi,e viceversa, ad esempio:

> plot(df$Fungicidi ,df$Vari , main =" kg di Vari rispetto ai Fungicidi ", xlab="Fungicidi ",ylab=" Vari", xlim=c(0 ,100) ,col ="red ")

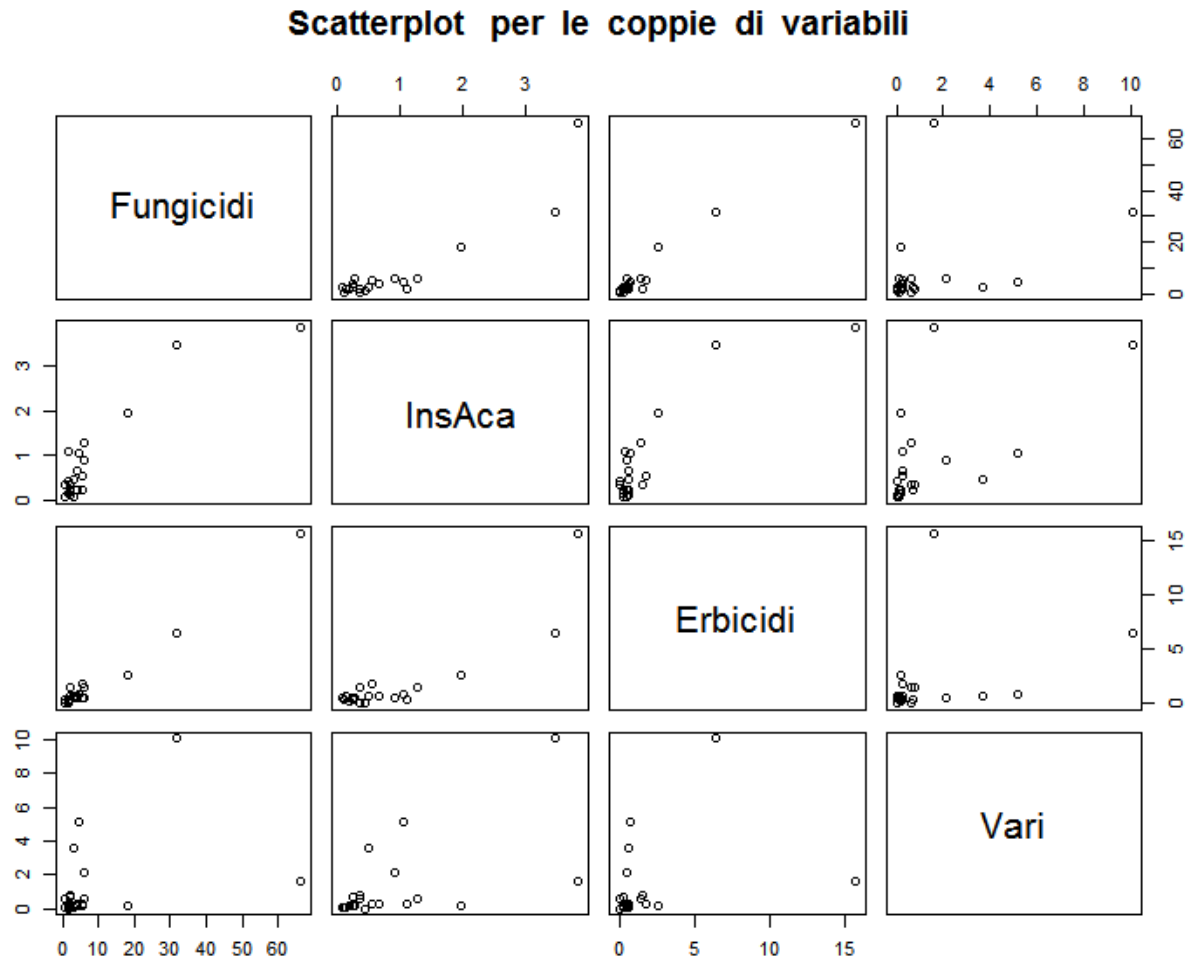


> plot(df$Vari ,df$Fungicidi , main =" kg di Fungicidi rispetto ai Vari ", xlab="Vari ",ylab=" Fungicidi", xlim=c(0 ,20) ,col ="red ")



Possiamo poi graficare tutte le possibili relazioni tra le sostanze contenute nei prodotti fitosanitari:

>pairs (df ,main =" Scatterplot per le coppie di variabili ")



Notiamo che una qualche regolarità è presente, nella maniera più interessante, solo nel caso Fungicidi-Erbicidi.

Studieremo la regressione per questa coppia di variabili quando tratteremo nelle prossime pagine la statistica bivariata.

STATISTICA DESCRITTIVA UNIVARIATA

La statistica descrittiva fa uso di metodi di natura logica e matematica per interpretare i fenomeni osservati.

Funzione di distribuzione empirica discreta

A partire dalle frequenze relative cumulate possiamo definire la funzione di distribuzione empirica discreta, supponendo i dati ordinati in maniera crescente, essa è definita mediante la formula:

$$F(x) = \frac{\#\{x_i \leq x, i = 1, 2, \ldots, n\}}{n} = \begin{cases} 0, & x < z_1 \\ F_1, & z_1 \leq x < z_2 \\ \ldots & \\ F_i, & z_i \leq x < z_{i+1} \\ \ldots & \\ 1, & x \geq z_k \end{cases}$$

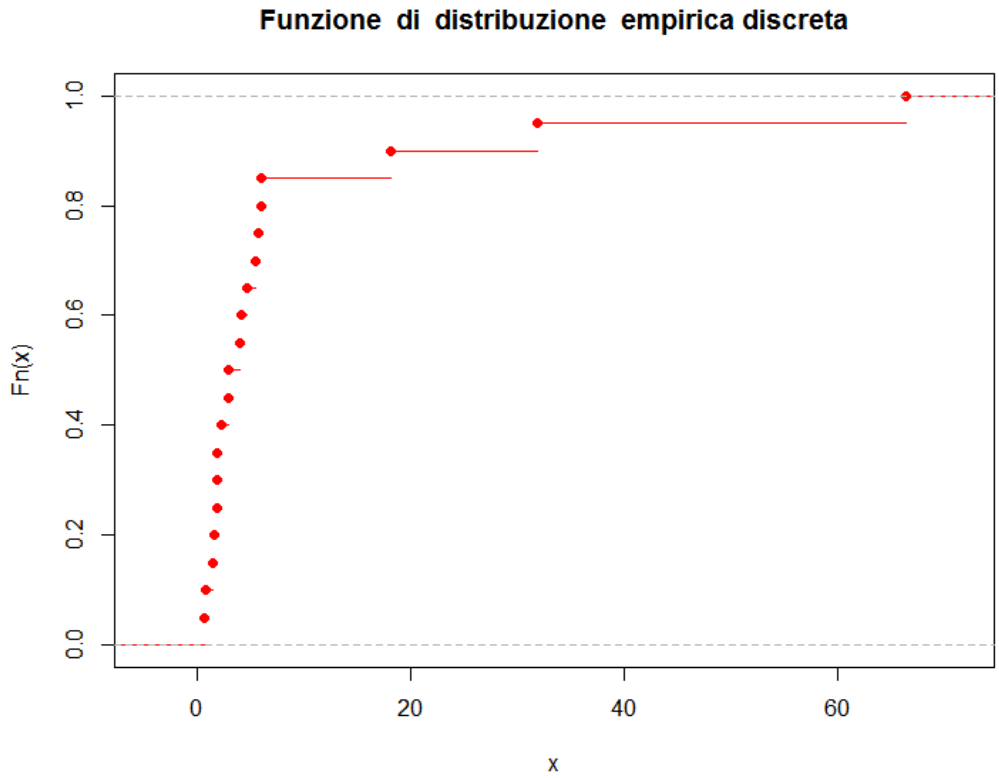
dove # è la cardinalità dell'insieme(noi abbiamo 20 regioni), i z_i sono possibili valori del campione, le x_i gli effettivi valori del campione e le F_i la proporzione dei dati del campione minori o uguali di z_i.

Ovviamente non si può fare ciò per tutta la tabella m, lo vediamo quindi a titolo di esempio (poiché non di particolare interesse ai fini della nostra trattazione) nel caso dei Fungicidi.

```
>round ( cumsum ( table (Fungicidi)/ length (Fungicidi)) ,3)
  0.55  0.76  1.41  1.58  1.77   1.8  1.82  2.17  2.82  2.83  3.95  4.12  4.68
  0.05  0.10  0.15  0.20  0.25  0.30  0.35  0.40  0.45  0.50  0.55  0.60  0.65
  5.39  5.63  5.89  5.97 18.14 31.92 66.42
  0.70  0.75  0.80  0.85  0.90  0.95  1.00
```

Queste sono le frequenze relative cumulate di Fungicidi arrotondate alla terza cifra decimale, poi grafichiamo:

```
>plot( ecdf( m),main=" Funzione  di  distribuzione  empirica discreta ",verticals =FALSE ,col ="red
")
```



Ad esempio questa funzione nel punto x=30 vale

> ecdf(Fungicidi) (30)

[1] 0.9

Funzione di distribuzione empirica continua

Per quella continua invece i dati vengono raccolti in classi distinte e la funzione è così definita:

$$F(x) = \begin{cases} 0, & x < z_1 \\ \dots & \\ F_i, & x = z_i \\ \dfrac{F_{i+1} - F_i}{z_{i+1} - z_i}\, x + \dfrac{z_{i+1} F_i - z_i F_{i+1}}{z_{i+1} - z_i}, & z_i < x < z_{i+1} \\ F_{i+1}, & x = z_{i+1} \\ \dots & \\ 1, & x \geq z_{k+1} \end{cases}$$

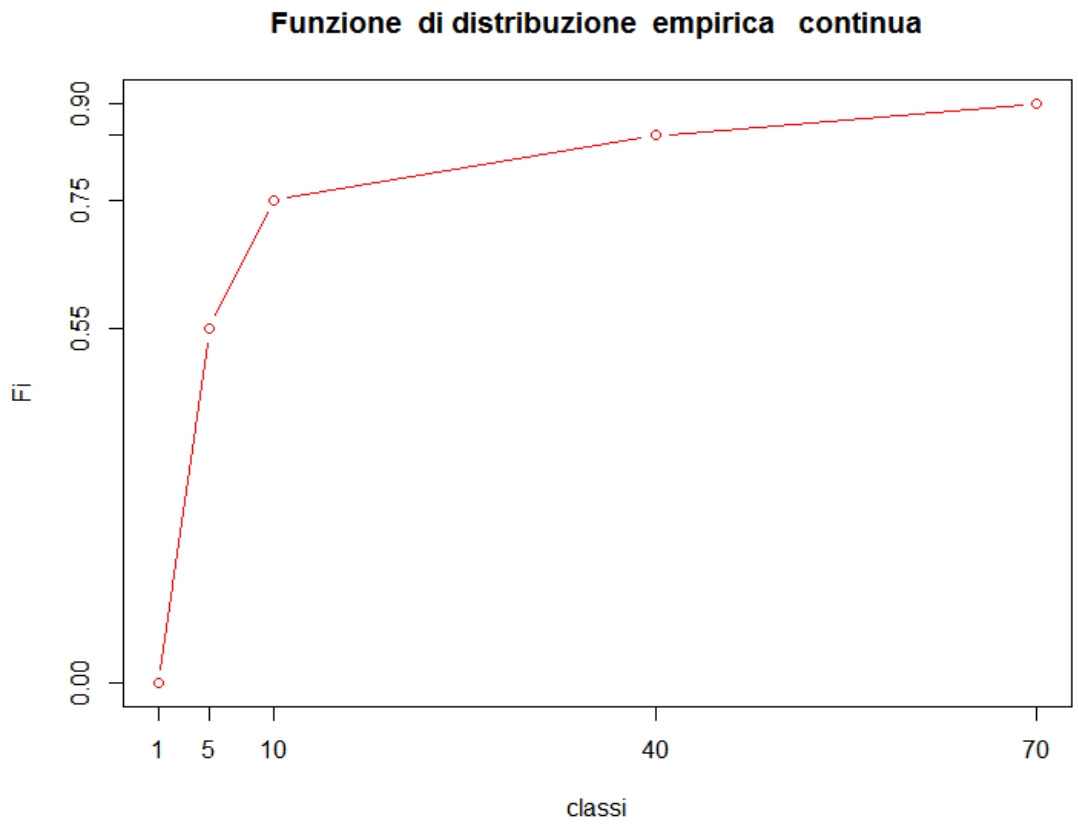
Stiamo aggiungendo cioè i segmenti passanti per i punti, o meglio le F_i delle classi.

Quindi in R:

```
>classi<-c (1 ,5 ,10 ,40 ,70)
> Fi <-cumsum (table (cut (Fungicidi , breaks =classi , right =FALSE )))/length (Fungicidi)
> Fi
 [1,5)  [5,10) [10,40) [40,70)
  0.55    0.75    0.85    0.90
```

Quelle così ottenute sono le frequenze relative cumulate delle classi.

```
>  Fi <-c(0, Fi)
> Fi
      [1,5)  [5,10) [10,40) [40,70)
  0.00   0.55   0.75   0.85   0.90
> plot( classi , Fi , type = "b", axes = FALSE , main = " Funzione  di distribuzione  empirica  continua
", col ="red ")
>  axis (1, classi )
>  axis (2, format (Fi , digits = 2))
```

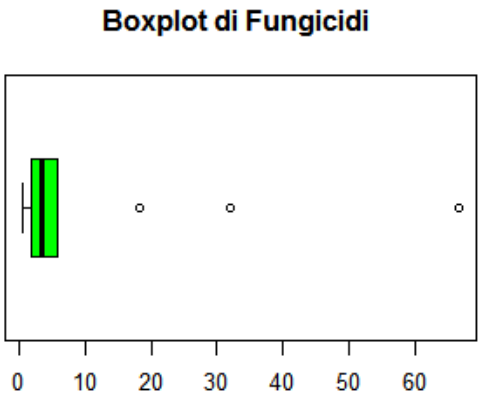


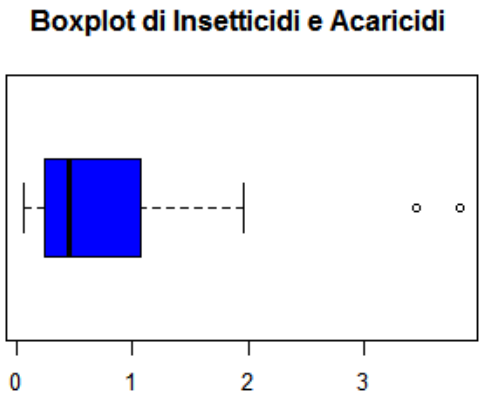
INDICI DI POSIZIONE E DI DISPERZIONE

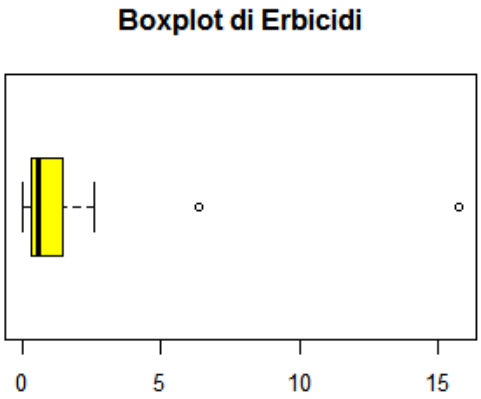
Introduciamo questi indici per poter confrontare al meglio i nostri dati. Per prima cosa confrontiamo in una unica finestra grafica i boxplot dei dati a nostra disposizione:

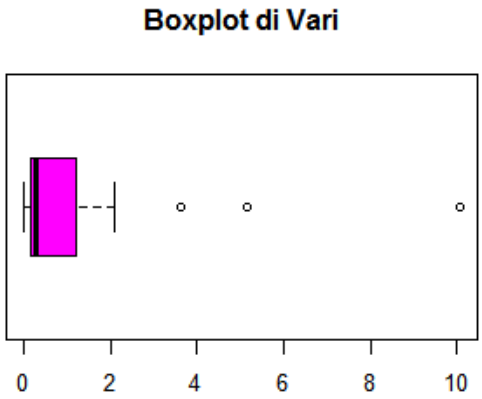
```
>  par ( mfrow =c(2 ,2) )
>  boxplot (Fungicidi , horizontal =TRUE ,col ="green ",main="Boxplot di Fungicidi ")
>  boxplot (InsAca , horizontal =TRUE ,col ="blue ",main="Boxplot di Insetticidi e Acaricidi ")
>  boxplot (Erbicidi , horizontal =TRUE ,col ="yellow ",main="Boxplot di Erbicidi")
>  boxplot (Vari , horizontal =TRUE ,col ="magenta ",main="Boxplot di Vari")
```

e otteniamo:









Calcoliamo ora la mediana:

> median (Fungicidi)
[1] 3.39
> median (InsAca)
[1] 0.46
> median (Erbicidi)
[1] 0.61
> median (Vari)
[1] 0.275

e la media:

> mean(Fungicidi)
[1] 8.481
> mean(InsAca)
[1] 0.8795
> mean(Erbicidi)
[1] 1.7705
> mean(Vari)
[1] 1.356

Esse sono misure di *centralità* dei dati.
Notiamo che, come visto già con altri strumenti, mediana e media sono molto distanti tra loro per tutte e quattro le categorie prese in analisi. Ciò, come già sottolineato più volte, evidenzia una non simmetria e una cattiva distribuzione dei dati.
Questo con molta probabilità è dovuto ai valori anomali che possiamo vedere nei boxplot.

La cattiva distribuzione dei dati può essere osservata nei grafici riguardanti gli istogrammi.

Calcolo dei quantili con percentuali scelte:

```
>  quantile (Fungicidi ,c(0 ,0.2 ,0.4 ,0.6 ,0.8 ,1) , type =2)
   0%    20%    40%    60%    80%   100%
 0.550  1.675  2.495  4.400  5.930 66.420
>  quantile (InsAca,c(0 ,0.2 ,0.4 ,0.6 ,0.8 ,1) , type =2)
  0%   20%   40%   60%   80%  100%
0.070 0.205 0.350 0.600 1.185 3.830
>  quantile (Erbicidi,c(0 ,0.2 ,0.4 ,0.6 ,0.8 ,1) , type =2)
   0%    20%    40%    60%    80%   100%
 0.020  0.275  0.495  0.645  1.650 15.720
>  quantile (Vari,c(0 ,0.2 ,0.4 ,0.6 ,0.8 ,1) , type =2)
   0%    20%    40%    60%    80%   100%
 0.010  0.125  0.245  0.635  1.850 10.070
```

Attraverso la funzione summary si possono confrontare i dati già trovati nelle pagine precedenti.
Tuttavia questi indici di posizione non tengono conto della variabilità tra i dati, infatti, dati che presentano stessa media e mediana, potrebbero però essere distribuiti in maniera completamente diversa.

Per questo motivo vengono introdotti degli *indici di dispersione* o di *variabilità*: varianza campionaria e deviazione standard.

Nel nostro caso i dati non presentano questo tipo di problema, ma per completezza ci occuperemo comunque degli indici di dispersione, in R vengono semplicemente implementati dai comandi sd e var.

```
> var(Fungicidi)
[1] 239.0682
> sd(Fungicidi)
[1] 15.46183
> var(InsAca)
[1] 1.121973
> sd(InsAca)
[1] 1.059232
```

```
> var(Erbicidi)
[1] 12.78625
> sd(Erbicidi)
[1] 3.575786
> var(Vari)
[1] 5.981309
> sd(Vari)
[1] 2.445672
```
Più i dati si discostano dalla media, più la varianza è elevata, come possiamo osservare dal valore dei Fungicidi (ricordiamo infatti che nei Fungicidi c'era anche il valore più alto che si discostava dal baffo superiore del boxplot).

Tuttavia, volendo fare un confronto tra i vari dati, siccome essi non hanno lo stesso range di variazione, calcoliamo il coefficiente di variazione:

```
>  sd(Fungicidi)/abs (mean (Fungicidi))
[1] 1.823114
>  sd(InsAca)/abs (mean (InsAca))
[1] 1.204358
>  sd(Erbicidi)/abs (mean (Erbicidi))
[1] 2.019648
>  sd(Vari)/abs (mean (Vari))
[1] 1.803593
```

Differentemente da quanto potevamo immaginare, nonostante il valore della varianza più alto fosse quello relativo ai Fungicidi, con questo confronto dei dati, notiamo che la dispersione massima è quella dei dati relativi agli Erbicidi.

Nell'ottica di valutare ancora la simmetria della distribuzione di frequenza dei dati introduciamo la skeweness campionaria, che ci permette di valutare la presenza di uno sbilanciamento eccessivo verso sinistra o verso destra.

In R non c'è un comando diretto per il calcolo della skewness, per questo costruiamo la funzione:

```
> skw<-function(x){
+ n<-length(x)
+ m2<-(n-1)*var(x)/n
+ m3<-(sum((x-mean(x))^3))/n
+ m3/(m2^1.5)
+ }
```

quindi:

```
> skw(Fungicidi)
[1] 2.991909
> skw(InsAca)
[1] 1.861369
> skw(Erbicidi)
[1] 3.280842
> skw(Vari)
[1] 2.615516
```

La skewness presenta una asimmetria positiva per tutti e quattro gli insiemi di dati, questo ci porta a dire che la distribuzione di frequenza ha, in tutti e quattro i casi, la coda di destra più allungata.

Siamo poi interessati a valutare in che modo queste distribuzioni di frequenza si discostano da una normale, cioè se sono più piatte (platicurtiche) o più piccate (leptocurtiche).

Così come per la skewness, per la curtosi dobbiamo costruire una apposita funzione:

```
>  curt <- function (x){
+  n<-length (x)
+  m2 <-(n -1) *var (x)/n
+  m4 <- (sum ( (x- mean(x))^4) )/n
+  m4/(m2 ^2)
+ }
```

Ottenendo quindi:

```
> curt(Fungicidi)
[1] 11.24697
> curt(InsAca)
[1] 5.409923
> curt(Erbicidi)
[1] 13.00663
> curt(Vari)
[1] 9.3291
```

Ci interessa confrontare tali risultati con il valore 3, poiché la normale ha curtosi uguale a 3. Poichè tutte le curtosi risultano essere maggiori di 3, possiamo dire che le corrispondenti distribuzioni di frequenza sono più piccate di una normale, in particolare quella relativa agli Erbicidi, esse sono quindi leptocurtiche.

STATISTICA DESCRITTIVA BIVARIATA

Abbiamo già calcolato indici di dispersione e posizione del nostro data frame nelle precedenti sezioni. In particolare ci interessano: mediana campionaria, media campionaria e deviazione standard, li riassumiamo nella seguente tabella:

	Fungicidi	Insetticidi e Acaricidi	Erbicidi	Vari
Mediana campionaria	3.39	0.46	0.61	0.275
Media campionaria	8.481	0.8795	1.7705	1.356
Deviazione standard	15.46183	1.059232	3.575786	2.445672

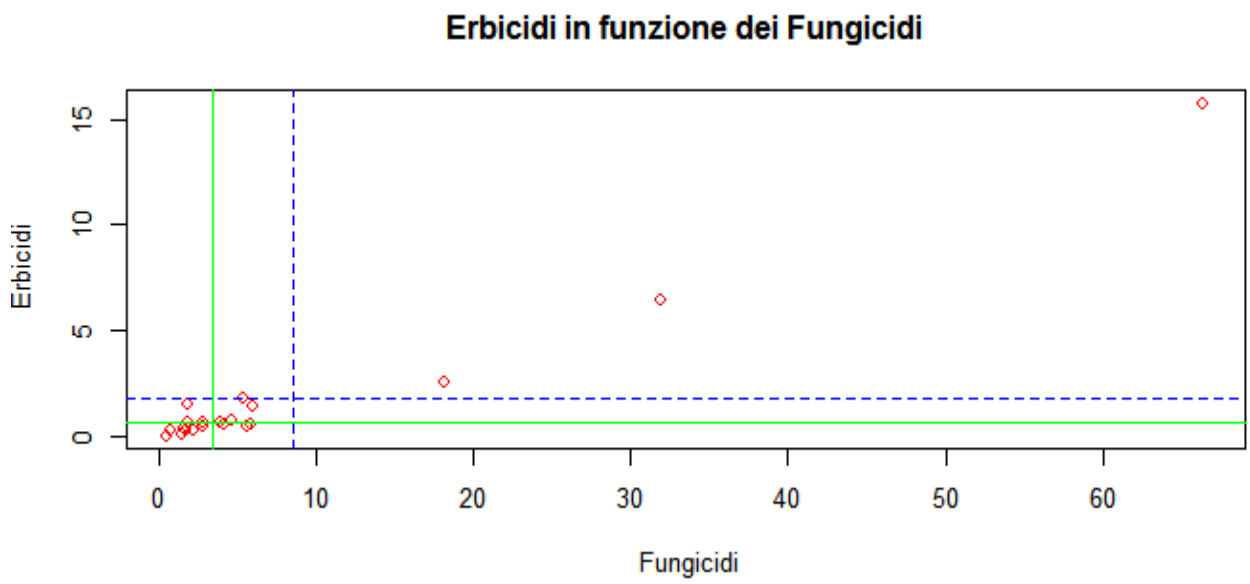
Possiamo poi disegnare lo scatterplot. Il seguente a titolo di esempio rappresenta gli Erbicidi in funzione dei Fungicidi:

```
> plot(Fungicidi,Erbicidi, main=" Erbicidi in funzione dei Fungicidi ", xlab=" Fungicidi",ylab =" Erbicidi", col ="red ")
```

disegnamo poi anche le linee della media e della mediana:

```
> abline (v= median (Fungicidi ),lty =1, col =" green ")
> abline (v=mean(Fungicidi),lty =2, col =" blue")
> abline (h=median (Erbicidi ),lty =1, col =" green ")
> abline (h=mean(Erbicidi),lty =2, col =" blue")
> legend (18 ,30 , c(" Mediana "," Media "),pch =0, col =c(" green","blue "), cex =0.8)
```

Otteniamo:



Siccome stiamo considerando più variabili quantitative (Fungicidi,Insetticidi e Acaricidi, Erbicidi, Vari) per ogni regione, ha senso chiedersi se esiste una correlazione tra le variabili. La covarianza campionaria ci fornisce una misura di questa correlazione.

L'equazione che la rappresenta è $$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}).$$

Nel nostro caso n=20 numero delle regioni, $\overline{x}$ e $\overline{y}$ sono le medie campionarie delle due variabili scelte e le x e le y sono i valori delle variabili.

Le variabili che scegliamo sono Fungicidi ed Erbicidi.

```
> cov(Fungicidi, Erbicidi)
[1] 54.55063
```

Invece il coefficiente di correlazione viene calcolato nel modo seguente: $$r_{xy} = \frac{C_{xy}}{s_x\, s_y}$$

dove al numeratore abbiamo la covarianza campionaria, al denominatore il prodotto delle deviazioni standard campionarie.

In R per le medesime variabili abbiamo:

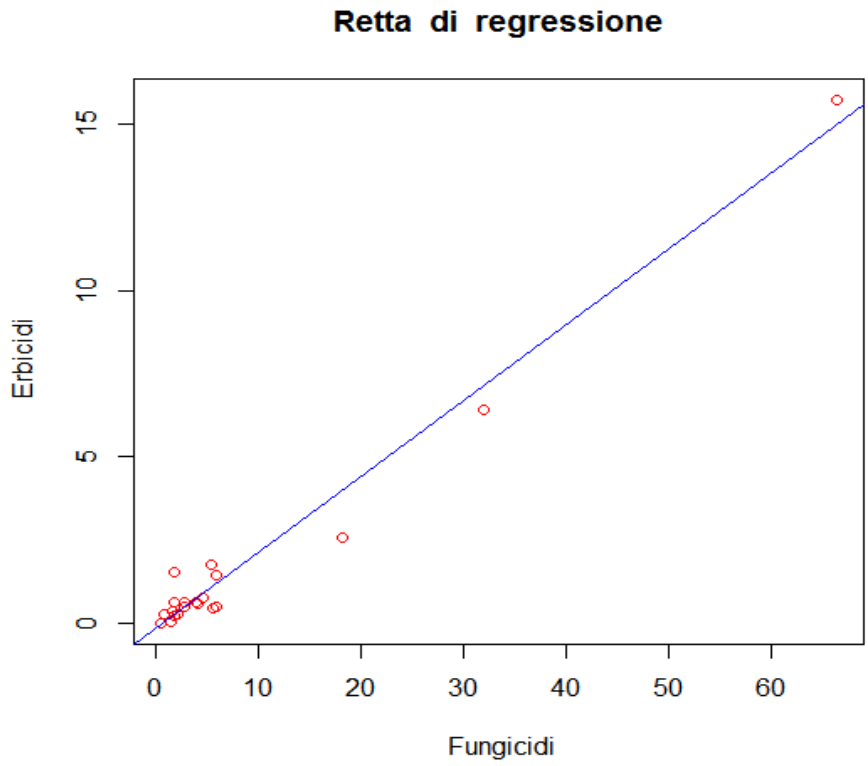
```
> cor(Fungicidi, Erbicidi)
[1] 0.9866595
```

Deduciamo che, siccome la covarianza campionaria è positiva, le variabili sono correlate positivamente e in maniera forte.
Una ulteriore conferma ci è data dal coefficiente di correlazione campionario, poiché esso è prossimo all'unità.

Tramite il seguente codice grafichiamo lo scatter plot e la retta di regressione stimata:

```
> plot(Fungicidi ,Erbicidi ,main =" Retta  di  regressione ", xlab=" Fungicidi",ylab=" Erbicidi", col ="red ")
> abline (lm(Erbicidi ~ Fungicidi), col =" blue")
```

Ottenendo:



lm è un comando di R, che crea automaticamente la retta di regressione, volendo costruirla utilizzando il metodo dei minimi quadrati abbiamo:

Equazione della retta: $Y = \alpha + \beta X$

Dobbiamo determinare α e β, lo facciamo mediante il metodo dei minimi quadrati, che consiste nel trovare quelli che minimizzano la somma:

$$Q = \sum_{i=1}^{n} \left[y_i - (\alpha + \beta x_i) \right]^2$$

otteniamo quindi:

$$\beta = \frac{s_y}{s_x} r_{xy}, \qquad \alpha = \overline{y} - \beta \overline{x}.$$

Possiamo implementare in R ottenendo il codice:

```
> beta <-(sd(Erbicidi)/sd(Fungicidi))*cor (Fungicidi ,Erbicidi )
> alpha <-mean (Erbicidi )-beta*mean (Fungicidi )
>  c(alpha , beta)
[1] -0.1646960  0.2281802
```

e quindi i coefficienti della retta di regressione.

Siamo poi interessati a sapere di quanto la retta si discosta dai dati osservati, per questo individuiamo i residui cioè gli scostamenti tra dati osservati e stimati.

$$E_i = y_i - (\alpha + \beta\, x_i)$$

La cui media campionaria è 0.

Calcoliamo prima il vettore dei valori stimati:

```
>  stime <-fitted (lm(Erbicidi ~Fungicidi ))
> stime
```

```
          1           2           3           4           5           6
1.119958345 0.330454969 0.195828671 0.903187187 1.197539602 7.118814923
          7           8           9          10          11          12
0.481053879 14.991030668 0.246028307 0.250591911 0.008720934 1.065195105
         13          14          15          16          17          18
0.736615666 0.239182902 1.179285189 0.775406294 -0.039196901 0.478772077
         19          20
3.974492228 0.157038043
```

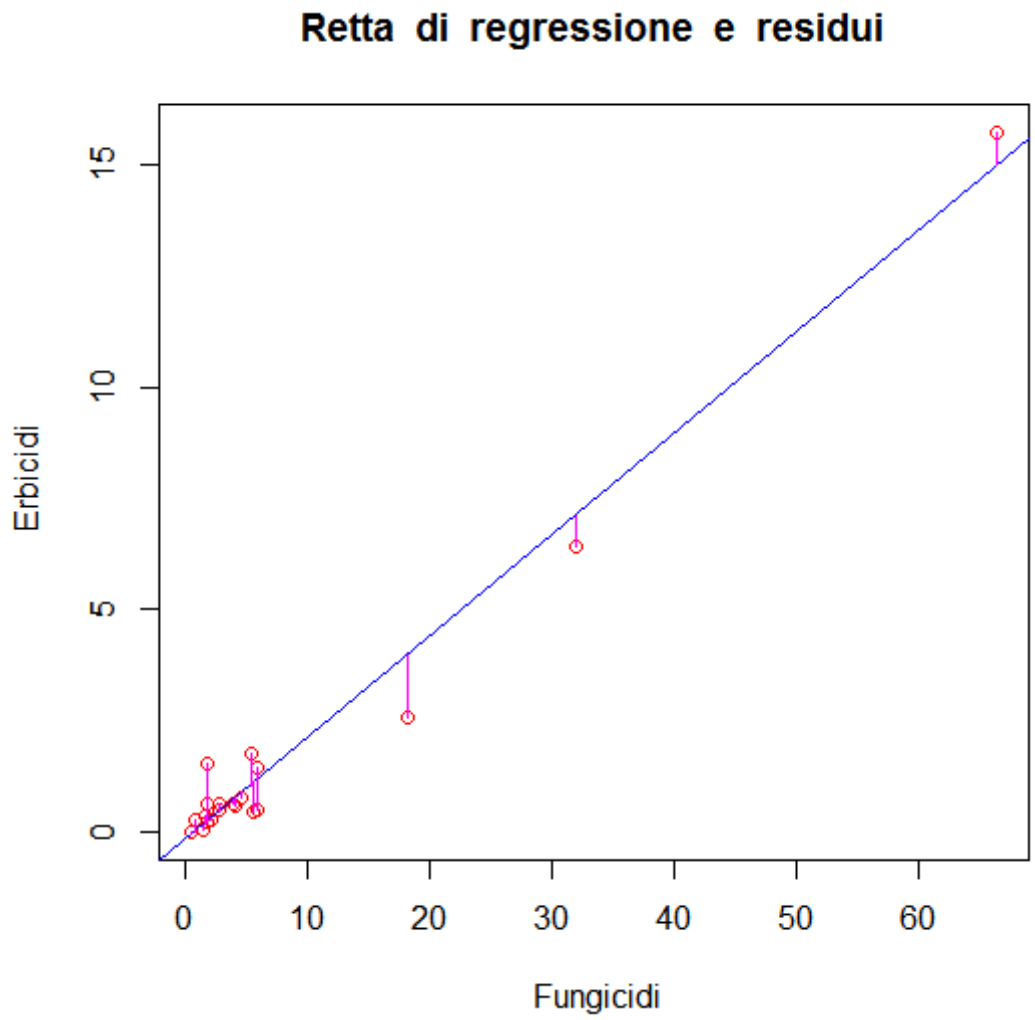
E poi quello dei residui:

```
> residui <-resid (lm(Erbicidi~Fungicidi))
> residui
```

```
            1           2           3           4           5
-0.6599583453 -0.0604549691  0.1741713291 -0.1331871871  0.2424603981
            6           7           8           9          10
-0.7088149230  0.1589461210  0.7289693321  1.2839716925  0.3994080892
           11          12          13          14          15
 0.2712790657  0.7048048947 -0.0966156655  0.0008170975 -0.6692851886
           16          17          18          19          20
-0.1954062938  0.0591969007  0.0012279227 -1.3944922284 -0.1070380425
```

Rappresentiamoli sul grafico:

```
>plot(Fungicidi ,Erbicidi ,main =" Retta  di  regressione  e  residui ", xlab=" Fungicidi",ylab="
Erbicidi", col ="red ")
>  abline (lm(Erbicidi~Fungicidi), col =" blue")
>  stime <-fitted (lm(Erbicidi~Fungicidi))
> segments (Fungicidi ,stime ,Fungicidi ,Erbicidi ,col="magenta")
```

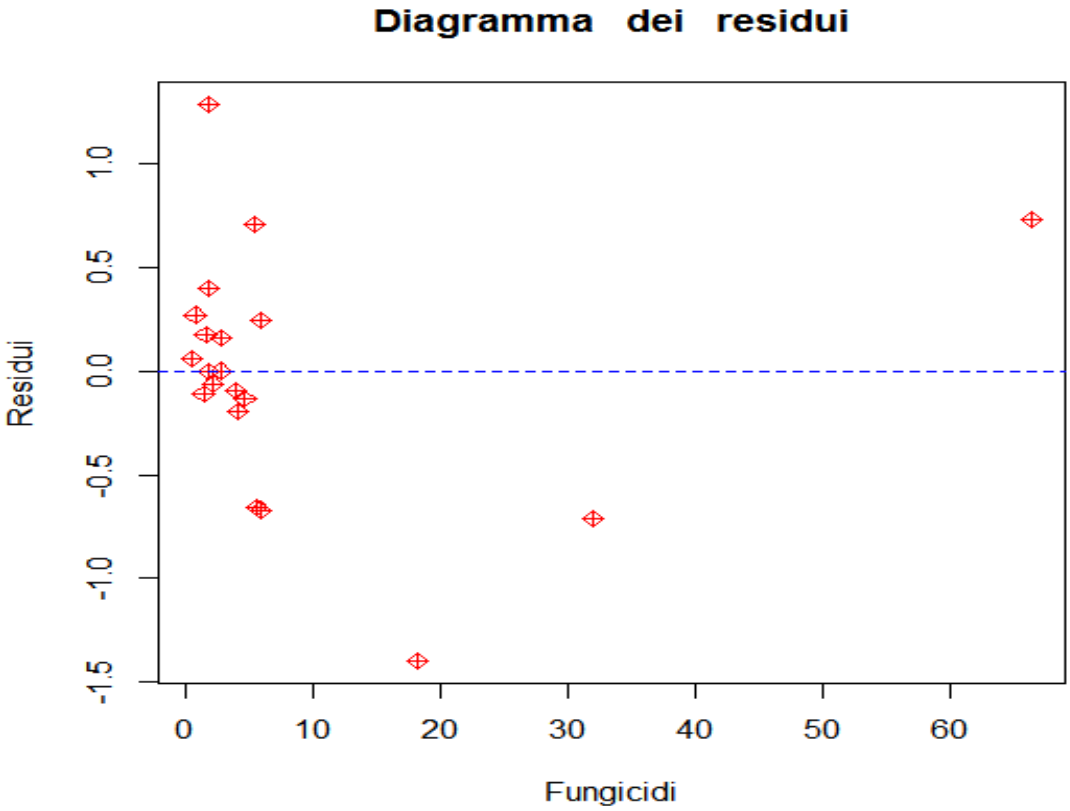


Studiamo poi il diagramma dei residui:

> residui <-resid (lm(Erbicidi ~Fungicidi))

> plot(Fungicidi ,residui , main =" Diagramma dei residui ", xlab="Fungicidi",ylab =" Residui ",pch =9, col ="red ")

> abline (h=0, col =" blue",lty =2)



I punti sono disposti in maniera piuttosto casuale rispetto alla retta. Tuttavia vi è un accumulo di punti nella parte iniziale, questo indica che la maggior parte dei valori non superano il 10.

Ora invece ci occuperemo della regressione lineare multipla, cioè sceglieremo una come variabile dipendente e le altre tre saranno indipendenti.

Ricostruiamo il dataframe:

	Fungicidi	InsAca	Erbicidi	Vari
1	5.63	0.26	0.46	0.13
2	2.17	0.23	0.27	0.68
3	1.58	1.10	0.37	0.28
4	4.68	1.05	0.77	5.15
5	5.97	1.27	1.44	0.64
6	31.92	3.45	6.41	10.07
7	2.83	0.48	0.64	3.64
8	66.42	3.83	15.72	1.60
9	1.80	0.34	1.53	0.78
10	1.82	0.13	0.65	0.06
11	0.76	0.09	0.28	0.12
12	5.39	0.55	1.77	0.27
13	3.95	0.65	0.64	0.27
14	1.77	0.18	0.24	0.20
15	5.89	0.90	0.51	2.10
16	4.12	0.25	0.58	0.21
17	0.55	0.36	0.02	0.63
18	2.82	0.07	0.48	0.06
19	18.14	1.96	2.58	0.22
20	1.41	0.44	0.05	0.01

Calcoliamo covarianza e correlazione:

> cov(df)

	Fungicidi	InsAca	Erbicidi	Vari
Fungicidi	239.06824	14.871358	54.550631	13.726873
InsAca	14.87136	1.121973	3.289390	1.512424
Erbicidi	54.55063	3.289390	12.786247	2.790286
Vari	13.72687	1.512424	2.790286	5.981309

> cor(df)

	Fungicidi	InsAca	Erbicidi	Vari
Fungicidi	1.0000000	0.9080263	0.9866595	0.3630049
InsAca	0.9080263	1.0000000	0.8684654	0.5838270
Erbicidi	0.9866595	0.8684654	1.0000000	0.3190649
Vari	0.3630049	0.5838270	0.3190649	1.0000000

La correlazione è forte tra Fungicidi e Insetticidi-Acaricidi, tra Fungicidi ed Erbicidi, tra Erbicidi e Insetticidi-Acaricidi(meno forte delle precedenti).
Il relativo grafico scatterplot per coppia di variabili è già stato trattato in sezioni precedenti.

I coefficienti della regressione multipla vengono trovati nel modo seguente:
(stiamo considerando gli Erbicidi come variabile dipendente)

>lm(df $Erbicidi ~df $Fungicidi +df $ InsAca+df$Vari)

```
Call:
lm(formula = df$Erbicidi ~ df$Fungicidi + df$InsAca + df$Vari)

Coefficients:
 (Intercept)  df$Fungicidi     df$InsAca       df$Vari
    0.021722      0.261316     -0.534477      0.001937
```

I segni dei regressori di Fungicidi e Vari sono positivi, questo implica che essi (seppur in minima parte) hanno un effetto positivo sull'aumento degli Erbicidi.
Calcoliamo ora i residui con una formula analoga a quella lineare:

$$E_i = y_i - \widehat{y}_i = y_i - (\alpha + \beta_1\, x_{i,1} + \beta_2\, x_{i,2} + \ldots \beta_p\, x_{i,p}) \qquad (i = 1, 2, \ldots, n)$$

Nel nostro caso n=20, y sono i valori osservati, x gli stimati.

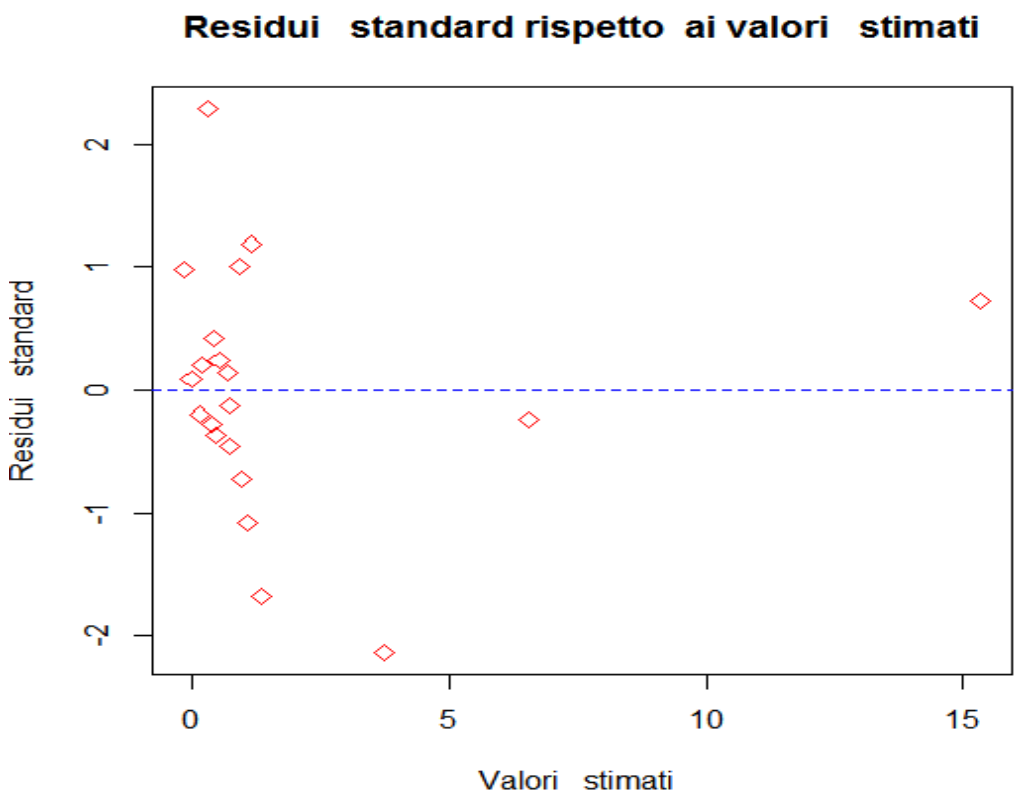
Valori stimati:
>stimemult <-fitted (lm(df $Erbicidi ~df $Fungicidi +df $ InsAca+df$Vari))
> stimemult

```
          1           2           3           4           5           6
1.35422085  0.46716593 -0.15278073  0.69345734  0.90423432  6.53850011
          7           8           9          10          11          12
0.51174912 15.33440644  0.31188008  0.42795194  0.17245191  1.13677771
         13          14          15          16          17          18
0.70703443  0.38843345  1.08391369  0.96513287 -0.02574548  0.72133694
         19          20
3.71485163  0.15502744
```

Residui:
>residuimult <-resid (lm(df $Erbicidi ~df $Fungicidi +df $ InsAca+df$Vari))
> residuimult

```
          1           2           3           4           5           6
-0.89422085 -0.19716593  0.52278073  0.07654266  0.53576568 -0.12850011
          7           8           9          10          11          12
 0.12825088  0.38559356  1.21811992  0.22204806  0.10754809  0.63322229
         13          14          15          16          17          18
-0.06703443 -0.14843345 -0.57391369 -0.38513287  0.04574548 -0.24133694
         19          20
-1.13485163 -0.10502744
```

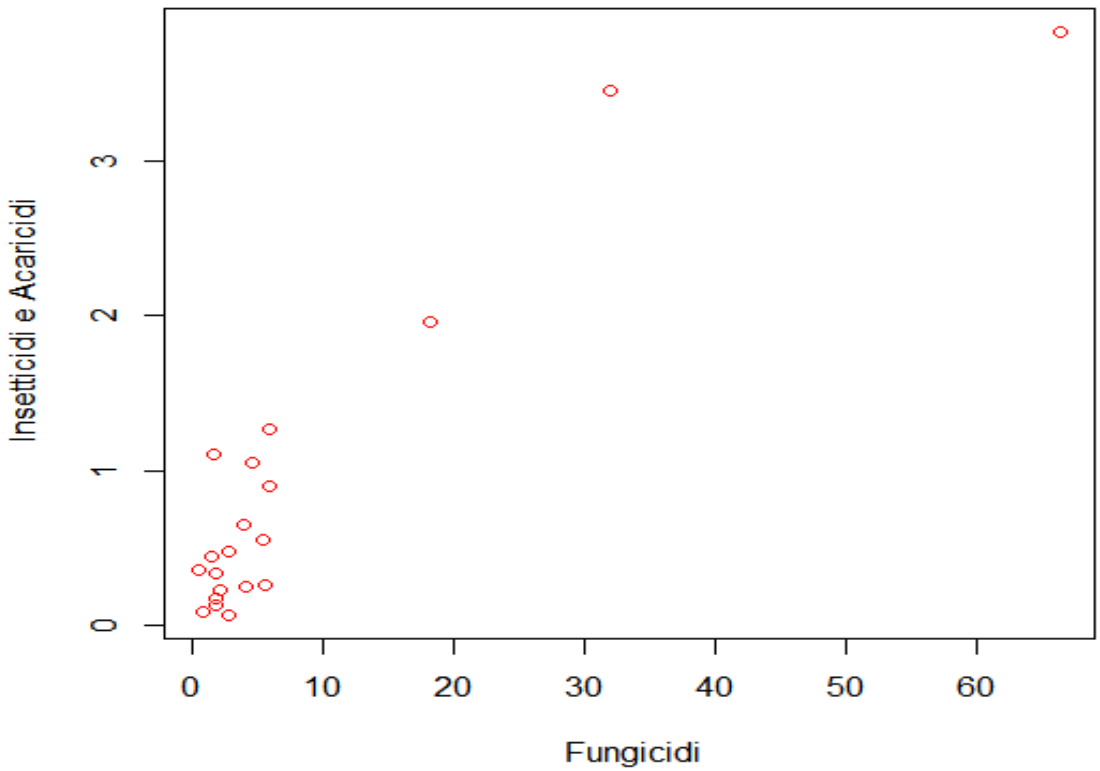
Il grafico viene poi realizzato nel seguente modo(con i residui standardizzati):
>residuimultstandard <- residuimult /sd(residuimult)
> plot(stimemult , residuimultstandard , main=" Residui standard rispetto ai valori stimati ", xlab="Valori stimati ",ylab =" Residui standard ",pch =5, col ="red ")
> abline (h=0, col ="blue ",lty =2)



Nel caso di Fungicidi-Erbicidi la regressione lineare funziona bene, tuttavia ci sono casi in cui la regressione lineare, seppur semplice non è il metodo migliore per approssimare i dati.

Ad esempio abbiamo visto che c'è una correlazione ache tra Fungicidi e Insetticidi-acaricidi. Tuttavia graficando otteniamo:

> plot(df$Fungicidi ,df$InsAca , xlab=" Fungicidi",ylab=" Insetticidi e Acaricidi", col ="red ")



Che non è adatto approssimare con una regressione lineare.

Calcoliamo i coefficienti nel modo lineare:

> lex <-lm(I(log (df$InsAca))~df$Fungicidi)

> lex

```
Call:
lm(formula = I(log(df$InsAca)) ~ df$Fungicidi)

Coefficients:
 (Intercept)  df$Fungicidi
    -1.11801       0.04824
```

Intercept rappresenta la α e df$Fungicidi la β

Tuttavia i modelli studiati non producono un buon risultato su questi dati, ci limitiamo ad osservare che con un apposito modello non lineare questi dati potrebbero essere ben approssimati.

ANALISI DEI CLUSTER:

Lo scopo dell'analisi dei cluster è raggruppare i dati in maniera tale che gli elementi di uno stesso gruppo siano tra loro il più possibile simili e gli elementi appartenenti a gruppi distinti siano tra loro il più possibile diversi.

Più nello specifico vogliamo che ogni individuo (regione) con le sue caratteristiche appartenga ad uno solo degli insiemi.

Occorre quindi introdurre un coefficiente di somiglianza o *similarità,* oppure una misura di distanza tra due individui.

I coefficienti di similarità assumono valori compresi tra 0 e 1, le misure di distanza possono assumere qualsiasi valore reale maggiore o uguale a zero.

Utilizzando il comando *dist* in R possiamo calcolare la distanza calcolata utilizzando le misure di distanza tra le righe della matrice dei dati.

Ad esempio con la metrica Euclidea:

```
>m<-
data.frame(Fungicidi=c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82,0.76,5.39,3.95,1.77,
5.89,4.12,0.55,2.82,18.14,1.41),
InsetticidiAcaricidi=c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13,0.09,0.55,0.65,0.18,0.90,
0.25,0.36,0.07,1.96,0.44), Erbicidi=c(0.46,0.27,
0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65,0.28,1.77,0.64,0.24,0.51,0.58,0.02,0.48,2.58,0.05),
Vari=c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06,0.12,0.27,0.26,0.20,2.10,0.21,0.63,0.0
6,0.22,0.01))
>rownames(m)<-c("Abruzzo","Basilicata","Calabria","Campania","Emilia-Romagna","Friuli-Venezia
Giulia","Lazio","Liguria","Lombardia","Marche","Molise","Piemonte","Puglia","Sardegna","Sicilia","Tosca
na","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto")
>  dist(m, method ="euclidean",diag=TRUE ,upper =TRUE )
```

Ottenendo:

	Abruzzo	Basilicata	Calabria	Campania
Abruzzo	0.0000000	3.5087177	4.1398913	5.1791022
Basilicata	3.5087177	0.0000000	1.1291590	5.2156879
Calabria	4.1398913	1.1291590	0.0000000	5.7870027
Campania	5.1791022	5.2156879	5.7870027	0.0000000
Emilia-Romagna	1.5349919	4.1100000	4.5360225	4.7435746
Friuli-Venezia Giulia	28.9058178	31.9578253	32.5325037	28.3512539
Lazio	4.4989888	3.0653874	3.6482050	2.4585361
Liguria	62.7948843	66.1858852	66.7011499	63.6840718
Lombardia	4.0302233	1.3215900	1.4905033	5.3360097
Marche	3.8175909	0.8132035	1.0608016	5.9117256
Molise	4.8762998	1.5236141	1.3138493	6.4675343
Piemonte	1.3701825	3.5901114	4.0961811	5.0565304
Puglia	1.7389077	1.9126160	2.4274884	4.9620459
Sardegna	3.8677254	0.6275349	0.9517353	5.8316721
Sicilia	2.0882050	4.0449104	4.6848799	3.2949507
Toscana	1.5169047	2.0297537	2.6875826	5.0391765
Trentino-Alto Aldige	5.1244512	1.6450836	1.3614331	6.2069236
Umbria	2.8173569	0.9362692	1.6306440	5.5147257
Valle d'Aosta	12.8020545	16.2351933	16.7290436	14.4769023
Veneto	4.2453857	1.0578280	0.7998750	6.1646573

	Emilia-Romagna	Friuli-Venezia Giulia	Lazio
Abruzzo	1.5349919	28.9058178	4.4989888
Basilicata	4.1100000	31.9578253	3.0653874
Calabria	4.5360225	32.5325037	3.6482050
Campania	4.7435746	28.3512539	2.4585361
Emilia-Romagna	0.0000000	28.1385980	4.4859447
Friuli-Venezia Giulia	28.1385980	0.0000000	30.4907658
Lazio	4.4859447	30.4907658	0.0000000
Liguria	62.1739182	36.7261678	65.4712044
Lombardia	4.2756871	32.0469187	3.1705205
Marche	4.4139098	32.4100309	3.7361879
Molise	5.4911292	33.4486861	4.1178878
Piemonte	1.0490948	28.8066051	4.3809017
Puglia	2.2911133	30.3263895	3.5647861
Sardegna	4.5234611	32.4838606	3.6341712
Sicilia	1.7719481	27.9713121	3.4537661
Toscana	2.3210773	30.2372039	3.6722609
Trentino-Alto Aldige	5.6763545	33.5197061	3.8284853
Umbria	3.5525343	31.5214118	3.6069655
Valle d'Aosta	12.2499388	17.4298566	15.8759724
Veneto	4.8797029	32.8872833	3.9424612

	Liguria	Lombardia	Marche	Molise	Piemonte
Abruzzo	62.7948843	4.0302233	3.8175909	4.8762998	1.3701825
Basilicata	66.1858852	1.3215900	0.8132035	1.5236141	3.5901114
Calabria	66.7011499	1.4905033	1.0608016	1.3138493	4.0961811
Campania	63.6840718	5.3360097	5.9117256	6.4675343	5.0565304
Emilia-Romagna	62.1739182	4.2756871	4.4139098	5.4911292	1.0490948
Friuli-Venezia Giulia	36.7261678	32.0469187	32.4100309	33.4486861	28.8066051
Lazio	65.4712044	3.1705205	3.7361879	4.1178878	4.3809017
Liguria	0.0000000	66.2567204	66.4554475	67.5707570	62.7039927
Lombardia	66.2567204	0.0000000	1.1564169	1.7726252	3.6400412
Marche	66.4554475	1.1564169	0.0000000	1.1250333	3.7709150
Molise	67.5707570	1.7726252	1.1250333	0.0000000	4.8878523
Piemonte	62.7039927	3.6400412	3.7709150	4.8878523	0.0000000
Puglia	64.3569367	2.4043918	2.2016812	3.2617327	1.8331939
Sardegna	66.5923074	1.4237275	0.4389761	1.0179391	3.9480502
Sicilia	62.4824767	4.4524712	4.6194155	5.5629399	2.3041267
Toscana	64.2281722	2.5725279	2.3090691	3.3783576	1.7670880
Trentino-Alto Aldige	67.8109777	1.9660875	1.5451861	0.6668583	5.1627318
Umbria	65.5265351	1.6535417	1.0161201	2.0706521	2.9229266
Valle d'Aosta	50.0901118	16.4631741	16.5360757	17.6312592	12.8533731
Veneto	66.9766317	1.7162168	0.7916439	0.7810250	4.3449396

	Puglia	Sardegna	Sicilia	Toscana
Abruzzo	1.7389077	3.8677254	2.0882050	1.5169047
Basilicata	1.9126160	0.6275349	4.0449104	2.0297537
Calabria	2.4274884	0.9517353	4.6848799	2.6875826
Campania	4.9620459	5.8316721	3.2949507	5.0391765
Emilia-Romagna	2.2911133	4.5234611	1.7719481	2.3210773
Friuli-Venezia Giulia	30.3263895	32.4838606	27.9713121	30.2372039
Lazio	3.5647861	3.6341712	3.4537661	3.6722609
Liguria	64.3569367	66.5923074	62.4824767	64.2281722
Lombardia	2.4043918	1.4237275	4.4524712	2.5725279
Marche	2.2016812	0.4389761	4.6194155	2.3090691
Molise	3.2617327	1.0179391	5.5629399	3.3783576
Piemonte	1.8331939	3.9480502	2.3041267	1.7670880
Puglia	0.0000000	2.2664730	2.6886056	0.4415880
Sardegna	2.2664730	0.0000000	4.6017062	2.3755210
Sicilia	2.6886056	4.6017062	0.0000000	2.6706554
Toscana	0.4415880	2.3755210	2.6706554	0.0000000
Trentino-Alto Aldige	3.4878933	1.3244244	5.5864300	3.6396428
Umbria	1.2957237	1.0916959	3.7783991	1.3247264
Valle d'Aosta	14.3818427	16.6319361	12.6097343	14.2648028
Veneto	2.6279840	0.5190376	4.9861508	2.7750856

	Trentino-Alto Aldige	Umbria	Valle d'Aosta	Veneto
Abruzzo	5.1244512	2.8173569	12.8020545	4.2453857
Basilicata	1.6450836	0.9362692	16.2351933	1.0578280
Calabria	1.3614331	1.6306440	16.7290436	0.7998750
Campania	6.2069236	5.5147257	14.4769023	6.1646573
Emilia-Romagna	5.6763545	3.5525343	12.2499388	4.8797029
Friuli-Venezia Giulia	33.5197061	31.5214118	17.4298566	32.8872833
Lazio	3.8284853	3.6069655	15.8759724	3.9424612
Liguria	67.8109777	65.5265351	50.0901118	66.9766317
Lombardia	1.9660875	1.6535417	16.4631741	1.7162168
Marche	1.5451861	1.0161201	16.5360757	0.7916439
Molise	0.6668583	2.0706521	17.6312592	0.7810250
Piemonte	5.1627318	2.9229266	12.8533731	4.3449396
Puglia	3.4878933	1.2957237	14.3818427	2.6279840
Sardegna	1.3244244	1.0916959	16.6319361	0.5190376
Sicilia	5.5864300	3.7783991	12.6097343	4.9861508
Toscana	3.6396428	1.3247264	14.2648028	2.7750856
Trentino-Alto Aldige	0.0000000	2.4028109	17.8518851	1.0636259
Umbria	2.4028109	0.0000000	15.5791559	1.5206578
Valle d'Aosta	17.8518851	15.5791559	0.0000000	16.9896527
Veneto	1.0636259	1.5206578	16.9896527	0.0000000

Scalando e standardizzando le variabili possiamo avere dei valori delle distanze che siano indipendenti dalle misure, nel nostro caso ci si riferisce sempre a kg per ettaro quindi non c'è questa necessità.

Volendone comunque fornire un esempio utilizziamo la funzione di R *scale*:

>scale(m)

```
                      Fungicidi InsetticidiAcaricidi      Erbicidi
Abruzzo              -0.1843895          -0.58485745 -0.3664928267
Basilicata           -0.4081664          -0.61317984 -0.4196279942
Calabria             -0.4463249           0.20816960 -0.3916621165
Campania             -0.2458312           0.16096561 -0.2797986059
Emilia-Romagna       -0.1623999           0.36866317 -0.0924272256
Friuli-Venezia Giulia 1.5159265           2.42675718  1.2974768937
Lazio                -0.3654806          -0.37715989 -0.3161542469
Liguria               3.7472274           2.78550752  3.9011001033
Lombardia            -0.4320963          -0.50933106 -0.0672579358
Marche               -0.4308028          -0.70758783 -0.3133576591
Molise               -0.4993587          -0.74535102 -0.4168314064
Piemonte             -0.1999116          -0.31107430 -0.0001398294
Puglia               -0.2930442          -0.21666632 -0.3161542469
Sardegna             -0.4340365          -0.66038384 -0.4280177575
Sicilia              -0.1675739           0.01935364 -0.3525098878
Toscana              -0.2820494          -0.59429825 -0.3329337735
Trentino-Alto Aldige -0.5129405          -0.49044947 -0.4895426883
Umbria               -0.3661274          -0.76423262 -0.3608996511
Valle d'Aosta         0.6246996           1.02007825  0.2263837796
Veneto               -0.4573197          -0.41492308 -0.4811529250

                            Vari
Abruzzo              -0.50104127
Basilicata           -0.27617575
Calabria             -0.43971431
Campania              1.55136769
Emilia-Romagna       -0.29252960
Friuli-Venezia Giulia 3.56289200
Lazio                 0.93400961
Liguria               0.09996295
Lombardia            -0.23529111
Marche               -0.52966052
Molise               -0.50512973
Piemonte             -0.44380277
Puglia               -0.44789124
Sardegna             -0.47242202
Sicilia               0.30438615
Toscana              -0.46833356
Trentino-Alto Aldige -0.29661807
Umbria               -0.52966052
Valle d'Aosta        -0.46424509
Veneto               -0.55010284

attr(,"scaled:center")
           Fungicidi InsetticidiAcaricidi             Erbicidi
              8.4810               0.8795               1.7705
                Vari
              1.3555
attr(,"scaled:scale")
           Fungicidi InsetticidiAcaricidi             Erbicidi
           15.461832            1.059232             3.575786
                Vari
            2.445906
```

Dove, negli attributi troviamo le medie campionarie e le deviazioni standard campionarie. Infatti otteniamo gli stessi valori utilizzando i comandi appositi:

> apply (m,2, mean)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
8.4810	0.8795	1.7705	1.3555

> apply (m,2, sd)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
15.461832	1.059232	3.575786	2.445906

Notiamo che sono gli stessi di quella scalata.

Possiamo poi quindi calcolare la distanza:

>z<-scale(m)

>dist(z, method ="euclidean",diag=TRUE , upper = TRUE)

	Abruzzo	Basilicata	Calabria	Campania	Emilia-Romagna
Abruzzo	0.00000000	0.32290260	0.83779264	2.18630408	1.01403855
Basilicata	0.32290260	0.00000000	0.83880745	1.99627683	1.06383488
Calabria	0.83779264	0.83880745	0.00000000	2.00483089	0.46645088
Campania	2.18630408	1.99627683	2.00483089	0.00000000	1.86685948
Emilia-Romagna	1.01403855	1.06383488	0.46645088	1.86685948	0.00000000
Friuli-Venezia Giulia	5.58974473	5.53445940	5.25800428	3.84338776	4.88350690
Lazio	1.46213449	1.23805610	1.49731936	0.82846109	1.46695128
Liguria	6.73724925	6.90134372	6.55344265	6.51300444	6.10152714
Lombardia	0.47668389	0.37039630	0.81365630	1.92907115	0.92060836
Marche	0.28182377	0.29150161	0.92362033	2.26283101	1.15559406
Molise	0.35709164	0.27966588	0.95756296	2.26575499	1.22678675
Piemonte	0.46118208	0.58198535	0.69544240	2.06974503	0.70345625
Puglia	0.39080614	0.45898706	0.45798343	2.03548346	0.65868776
Sardegna	0.26950369	0.20366755	0.87001586	2.19720957	1.13035680
Sicilia	1.00710608	0.89417017	0.81766249	1.25953485	0.73891497
Toscana	0.10873168	0.24637883	0.82171112	2.15725560	1.01506705
Trentino-Alto Aldige	0.41687975	0.17704891	0.72288546	1.98865104	1.00929215
Umbria	0.25701058	0.30378901	0.98032360	2.28204266	1.20551494
Valle d'Aosta	1.89296076	2.04622712	1.47948354	2.41139250	1.08397135
Veneto	0.34485146	0.34719326	0.63918668	2.19843060	0.95835105

	Friuli-Venezia Giulia	Lazio	Liguria	Lombardia
Abruzzo	5.58974473	1.46213449	6.73724925	0.47668389
Basilicata	5.53445940	1.23805610	6.90134372	0.37039630
Calabria	5.25800428	1.49731936	6.55344265	0.81365630
Campania	3.84338776	0.82846109	6.51300444	1.92907115
Emilia-Romagna	4.88350690	1.46695128	6.10152714	0.92060836
Friuli-Venezia Giulia	0.00000000	4.57345293	4.88652609	5.35762101
Lazio	4.57345293	0.00000000	6.73778159	1.20462459
Liguria	4.88652609	6.73778159	0.00000000	6.64702706
Lombardia	5.35762101	1.20462459	6.64702706	0.00000000
Marche	5.74087515	1.50192799	6.91489200	0.43188638
Molise	5.79752521	1.49490669	7.03641424	0.50521693
Piemonte	5.30822323	1.42478507	6.37840731	0.37576244
Puglia	5.38054368	1.39307401	6.58955393	0.46058259
Sardegna	5.70910558	1.44065232	6.95888077	0.45738324
Sicilia	4.68722543	0.77080314	6.41188457	0.84976456
Toscana	5.59182542	1.42160394	6.77555624	0.39322504
Trentino-Alto Aldige	5.54217041	1.25661728	6.95096725	0.43471567
Umbria	5.76400395	1.51464785	6.93407603	0.49214378
Valle d'Aosta	4.48754996	2.27644350	5.16611956	1.89593137
Veneto	5.66119225	1.49655431	6.89548400	0.52911679

	Marche	Molise	Piemonte	Puglia	Sardegna
Abruzzo	0.28182377	0.35709164	0.46118208	0.39080614	0.26950369
Basilicata	0.29150161	0.27966588	0.58198535	0.45898706	0.20366755
Calabria	0.92362033	0.95756296	0.69544240	0.45798343	0.87001586
Campania	2.26283101	2.26575499	2.06974503	2.03548346	2.19720957
Emilia-Romagna	1.15559406	1.22678675	0.70345625	0.65868776	1.13035680
Friuli-Venezia Giulia	5.74087515	5.79752521	5.30822323	5.38054368	5.70910558
Lazio	1.50192799	1.49490669	1.42478507	1.39307401	1.44065232
Liguria	6.91489200	7.03641424	6.37840731	6.58955393	6.95888077
Lombardia	0.43188638	0.50521693	0.37576244	0.46058259	0.45738324
Marche	0.00000000	0.13203995	0.56214825	0.51640623	0.13660841
Molise	0.13203995	0.00000000	0.67502420	0.57921095	0.11261143
Piemonte	0.56214825	0.67502420	0.00000000	0.34273659	0.60060820
Puglia	0.51640623	0.57921095	0.34273659	0.00000000	0.47945728
Sardegna	0.13660841	0.11261143	0.60060820	0.47945728	0.00000000
Sicilia	1.13793675	1.16374720	0.89116759	0.79918158	1.06872715
Toscana	0.19775319	0.28005901	0.44532683	0.37871652	0.19111537
Trentino-Alto Aldige	0.37315630	0.33752528	0.62557256	0.41983179	0.26418909
Umbria	0.09824347	0.14777420	0.60868657	0.56023064	0.15224096
Valle d'Aosta	2.09630993	2.18989608	1.58230263	1.63291327	2.09121574
Veneto	0.33901171	0.34221300	0.56543503	0.32243442	0.26391427

	Sicilia	Toscana	Trentino-Alto Aldige
Abruzzo	1.00710608	0.10873168	0.41687975
Basilicata	0.89417017	0.24637883	0.17704891
Calabria	0.81766249	0.82171112	0.72288546
Campania	1.25953485	2.15725560	1.98865104
Emilia-Romagna	0.73891497	1.01506705	1.00929215
Friuli-Venezia Giulia	4.68722543	5.59182542	5.54217041
Lazio	0.77080314	1.42160394	1.25661728
Liguria	6.41188457	6.77555624	6.95096725
Lombardia	0.84976456	0.39322504	0.43471567
Marche	1.13793675	0.19775319	0.37315630
Molise	1.16374720	0.28005901	0.33752528
Piemonte	0.89116759	0.44532683	0.62557256
Puglia	0.79918158	0.37871652	0.41983179
Sardegna	1.06872715	0.19111537	0.26418909
Sicilia	0.00000000	0.99355535	0.87129865
Toscana	0.99355535	0.00000000	0.34366823
Trentino-Alto Aldige	0.87129865	0.34366823	0.00000000
Umbria	1.16152276	0.20122099	0.40910772
Valle d'Aosta	1.59845518	1.93423318	2.02893286
Veneto	1.00957879	0.30257274	0.27041233

	Umbria	Valle d'Aosta	Veneto
Abruzzo	0.25701058	1.89296076	0.34485146
Basilicata	0.30378901	2.04622712	0.34719326
Calabria	0.98032360	1.47948354	0.63918668
Campania	2.28204266	2.41139250	2.19843060
Emilia-Romagna	1.20551494	1.08397135	0.95835105
Friuli-Venezia Giulia	5.76400395	4.48754996	5.66119225
Lazio	1.51464785	2.27644350	1.49655431
Liguria	6.93407603	5.16611956	6.89548400
Lombardia	0.49214378	1.89593137	0.52911679
Marche	0.09824347	2.09630993	0.33901171
Molise	0.14777420	2.18989608	0.34221300
Piemonte	0.60868657	1.58230263	0.56543503
Puglia	0.56023064	1.63291327	0.32243442
Sardegna	0.15224096	2.09121574	0.26391427
Sicilia	1.16152276	1.59845518	1.00957879
Toscana	0.20122099	1.93423318	0.30257274
Trentino-Alto Aldige	0.40910772	2.02893286	0.27041233
Umbria	0.00000000	2.12477866	0.38106682
Valle d'Aosta	2.12477866	0.00000000	1.93338415
Veneto	0.38106682	1.93338415	0.00000000

Si possono poi definire anche altre metriche, ad esempio quella del massimo,di Manhattan, di Minkowsi, di Canberra, ne faremo una prova sulle prime 5 regioni:

```
>m5<-data.frame(Fungicidi=c(5.63,2.17,1.58,4.68,5.97),
InsetticidiAcaricidi=c(0.26,0.23,1.10,1.05,1.27), Erbicidi=c(0.46,0.27, 0.37,0.77,1.44),
Vari=c(0.13,0.68,0.28,5.15,0.64))
> rownames(m5)<-c("Abruzzo","Basilicata","Calabria","Campania","EmiliaRomagna")
>m5
```

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Abruzzo	5.63	0.26	0.46	0.13
Basilicata	2.17	0.23	0.27	0.68
Calabria	1.58	1.10	0.37	0.28
Campania	4.68	1.05	0.77	5.15
EmiliaRomagna	5.97	1.27	1.44	0.64

```
> dist(m5, method ="manhattan",diag=TRUE ,upper =TRUE )
```

	Abruzzo	Basilicata	Calabria	Campania	EmiliaRomagna
Abruzzo	0.00	4.23	5.13	7.07	2.84
Basilicata	4.23	0.00	1.96	8.30	6.05
Calabria	5.13	1.96	0.00	8.42	5.99
Campania	7.07	8.30	8.42	0.00	6.69
EmiliaRomagna	2.84	6.05	5.99	6.69	0.00

```
> dist(m5, method ="maximum",diag=TRUE , upper = TRUE)
```

	Abruzzo	Basilicata	Calabria	Campania	EmiliaRomagna
Abruzzo	0.00	3.46	4.05	5.02	1.01
Basilicata	3.46	0.00	0.87	4.47	3.80
Calabria	4.05	0.87	0.00	4.87	4.39
Campania	5.02	4.47	4.87	0.00	4.51
EmiliaRomagna	1.01	3.80	4.39	4.51	0.00

```
> dist(m5, method ="minkowski" ,4,diag =TRUE , upper =TRUE)
```

	Abruzzo	Basilicata	Calabria	Campania	EmiliaRomagna
Abruzzo	0.0000000	3.4605600	4.0518745	5.0223959	1.1956927
Basilicata	3.4605600	0.0000000	0.9210827	4.5785235	3.8137922
Calabria	4.0518745	0.9210827	0.0000000	5.0586964	4.3939201
Campania	5.0223959	4.5785235	5.0586964	0.0000000	4.5180807
EmiliaRomagna	1.1956927	3.8137922	4.3939201	4.5180807	0.0000000

>dist(m5, method ="canberra",diag=TRUE ,upper =TRUE)

	Abruzzo	Basilicata	Calabria	Campania	EmiliaRomagna
Abruzzo	0.000000	1.444101	1.653654	1.897987	1.867568
Basilicata	1.444101	0.000000	1.384385	2.254541	1.874677
Calabria	1.653654	1.384385	0.000000	1.766210	1.635651
Campania	1.897987	2.254541	1.766210	0.000000	1.298051
EmiliaRomagna	1.867568	1.874677	1.635651	1.298051	0.000000

Notiamo che nella metrica di Manhattan, del massimo e di Minkowski l'Abruzzo assume valori più vicini a quelli dell'Emilia-Romagna.

Nella metrica di Canberra questo non accade, ciò potrebbe essere spiegato dal fatto che quest'ultima è poco sensibile all'asimmetria delle distribuzioni delle variabili (caratteristiche) e alla presenza di eventuali valori anomali (outlier).

Misure di similarità

Così come la matrice delle distanze, possiamo introdurre la matrice di similarità. Essa è significativa poiché, come abbiamo accennato assume solo valori tra 0 e 1.
E' possibile e facile passare da una misura di distanza ad una di similarità, il viceversa è invece molto difficile.

Andiamo ora a studiare le misure di non omogeneità totale e tra cluster.

Andiamo a calcolare, nel nostro data frame iniziale m, la media(già vista), la varianza e la matrice delle covarianze:

>apply (m,2, mean)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
8.4810	0.8795	1.7705	1.3555

>apply (m,2, var)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
239.068241	1.121973	12.786247	5.982458

> w<-cov (m)

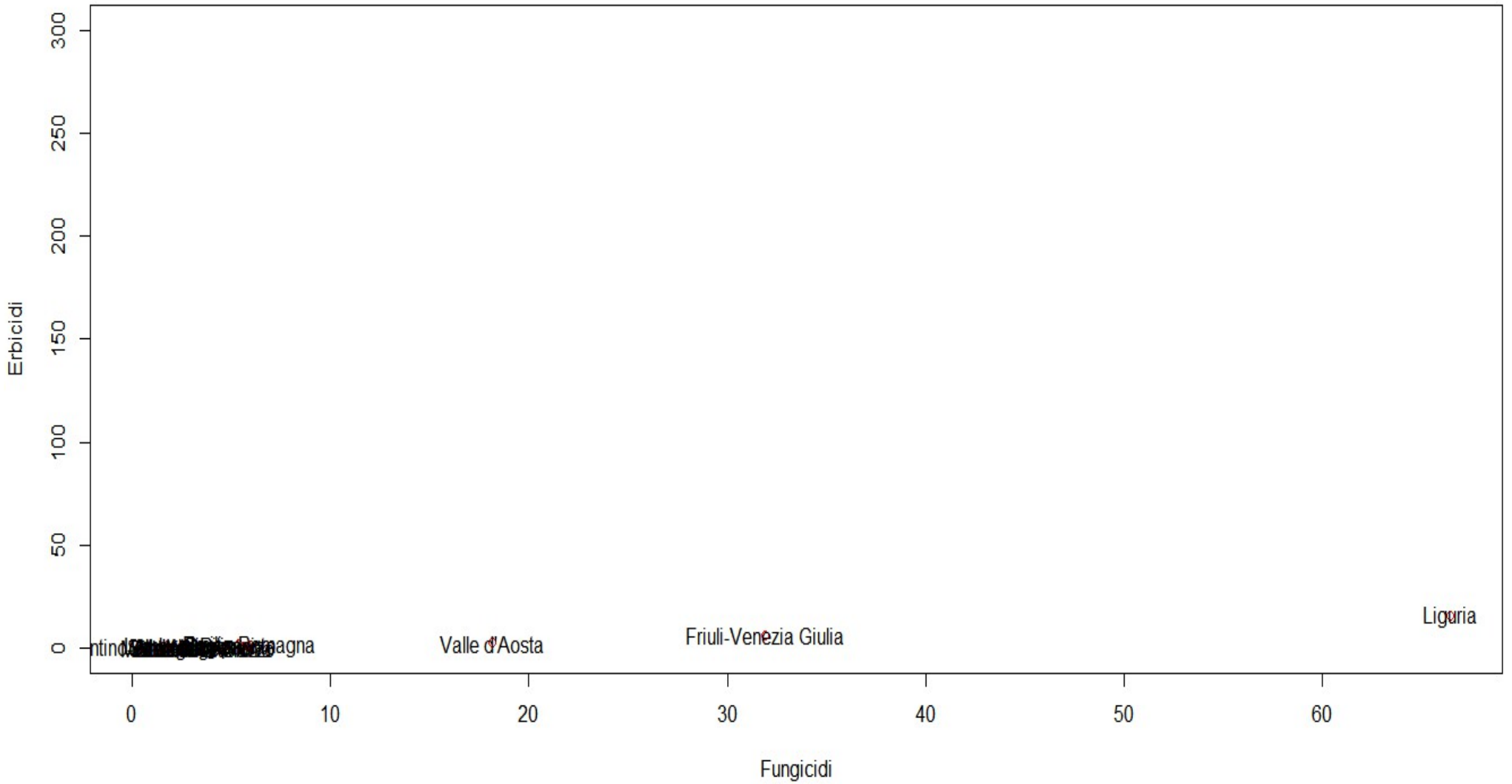
>w

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	239.06824	14.871358	54.550631	13.729257
InsetticidiAcaricidi	14.87136	1.121973	3.289390	1.512545
Erbicidi	54.55063	3.289390	12.786247	2.790881
Vari	13.72926	1.512545	2.790881	5.982458

Le caratteristiche sono tutte correlate positivamente.
Possiamo poi rappresentare i 20 punti relativi alle regioni tramite uno scatterplot, si scelgono ad esempio due(Fungicidi ed Erbicidi)

```
>plot(m$Fungicidi , m$Erbicidi ,col ="red ",xlab="Fungicidi",
ylab="Erbicidi",ylim=c(0 ,300) )
>text( m$Fungicidi, m$Erbicidi +0.1 , c("Abruzzo","Basilicata","Calabria","Campania","Emilia-
Romagna","Friuli-Venezia
Giulia","Lazio","Liguria","Lombardia","Marche","Molise","Piemonte","Puglia","Sardegna","Sicilia","Tosca
na","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto"))
abline (lm(m$Erbicidi~m$Fungicidi),lty =2, col ="blue ")
```

A causa dei forti valori anomali presenti nei Fungicidi il grafico che otteniamo è quello nella pagina seguente.


Erbicidi
300
250
200
150
100
50
0
Valle d'Aosta
Friuli-Venezia Giulia
Liguria
0
10
20
30
40
50
60
Fungicidi

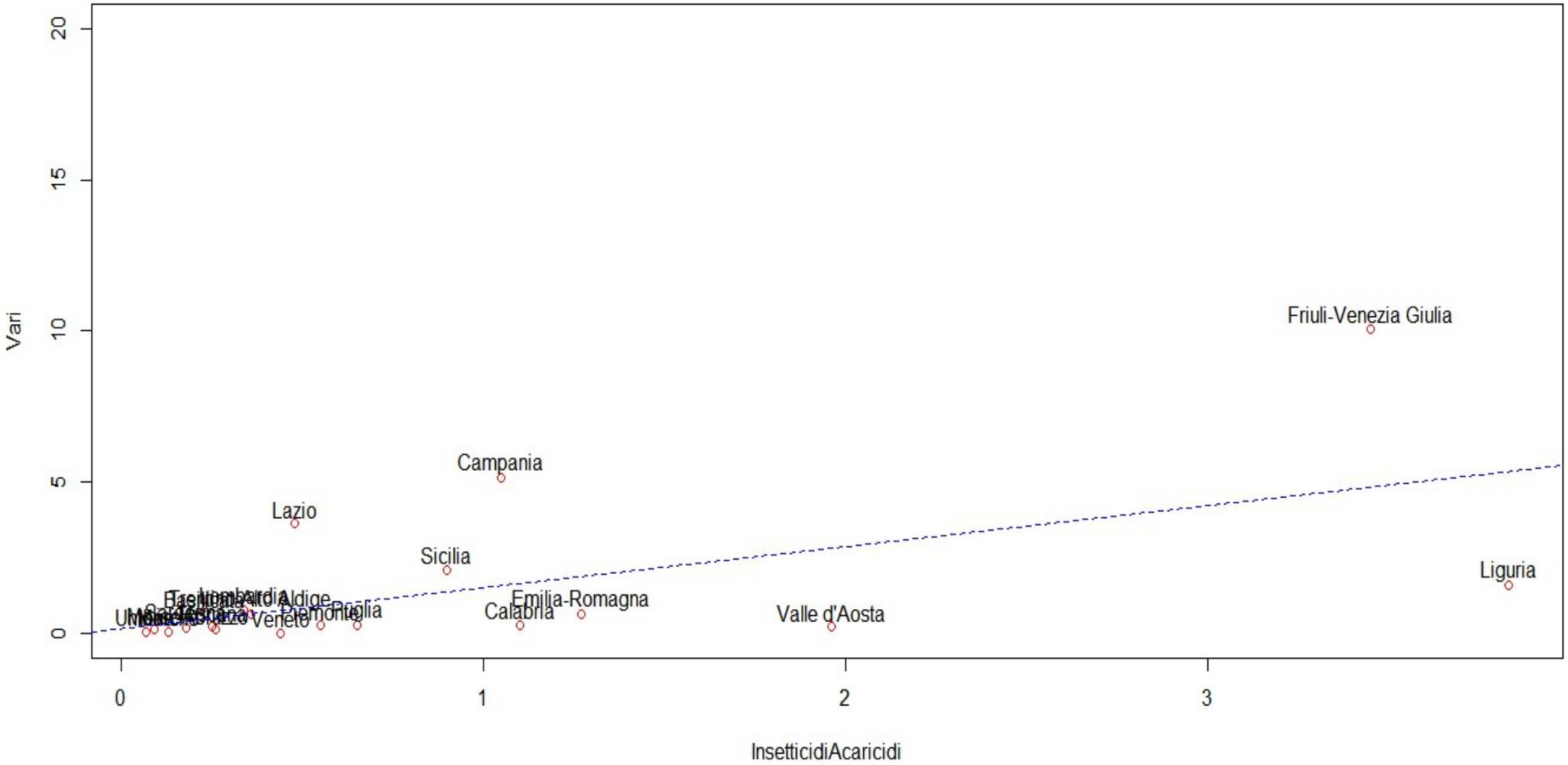
Possiamo quindi pensare di ripetere l'operazione ad esempio con Insetticidi-Acaricidi e Vari che presentano valori più piccoli:

```
>plot(m$InsetticidiAcaricidi , m$Vari ,col ="red ",xlab="InsetticidiAcaricidi",
ylab="Vari",ylim=c(0 ,20) )
>text( m$InsetticidiAcaricidi, m$Vari+0.5 , c("Abruzzo","Basilicata","Calabria","Campania","Emilia-Romagna","Friuli-Venezia
Giulia","Lazio","Liguria","Lombardia","Marche","Molise","Piemonte","Puglia","Sardegna","Sicilia","Toscana","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto"))
```

Aggiungiamo anche la retta di regressione:

```
>abline (lm(m$Vari~m$InsetticidiAcaricidi),lty =2, col ="blue ")
```

Pagina seguente.


Vari
InsetticidiAcaricidi
0
5
10
15
20
0
1
2
3
Friuli-Venezia Giulia
Campania
Lazio
Sicilia
Liguria
Emilia-Romagna
Calabria
Valle d'Aosta
Veneto

Vogliamo ora calcolare la matrice di non omogeneità totale, essa dipende dalla numerosità del campione e dalla covarianza.
Calcoliamo quindi per prima cosa la matrice non omogeneità statistica:

```
> n <-nrow (m)
>w<-cov(m)
> h <-(n -1) *w
>h
```

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	4542.2966	282.55581	1036.46199	260.85589
InsetticidiAcaricidi	282.5558	21.31750	62.49841	28.73836
Erbicidi	1036.4620	62.49841	242.93870	53.02675
Vari	260.8559	28.73836	53.02675	113.66670

Calcoliamo poi la misura di non omogeneità statistica, essa è la traccia della matrice h:

```
>trh<-sum ( diag(h))
>trh
[1] 4920.219
```

La misura di non omogeneità statistica trh dipende sia dall'omogeneità interna sia dalla numerosità del gruppo.

Ci sono anche altri modi per calcolare la misura di non omogeneità ad esempio:

```
>d <-dist (m, method ="euclidean",diag=FALSE , upper =FALSE )
>tr <-sum (d^2) /n
> tr
[1] 4920.219
```

O ancora, in un unico passaggio:

```
>trh<-(n -1) *sum ( apply (m ,2, var ))
> trh
[1] 4920.219
```

Occupiamoci ora della più interessante misura di non omogeneità tra cluster:
Questa misura è importante perché al termine della classificazione in cluster individui appartenenti allo stesso cluster dovrebbero essere il più possibile omogenei tra loro, quelli appartenenti a cluster distinti da questi, il più possibile differenti.
Cioè vogliamo individuare i cluster in maniera tale da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between).

Riconsideriamo quindi la matrice dei dati *m*.
Consideriamo come due cluster ad esempio le prime 10 e le seconde 10 regioni.
Quindi nel primo cluster abbiamo: m1={Abruzzo, Basilicata, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Lazio, Liguria, Lombardia, Marche}
Nel secondo cluster: m2={Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-Alto Aldige, Umbria, Valle d'Aosta, Veneto}

Per m1:

```
>m1<-data.frame(Fungicidi=c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82),
InsetticidiAcaricidi=c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13), Erbicidi=c(0.46,0.27,
0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65),
Vari=c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06))
>row.names(m1)<-c("Abruzzo","Basilicata","Calabria","Campania","Emilia-Romagna","Friuli-
Venezia Giulia","Lazio","Liguria","Lombardia","Marche")
>m1
```

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Abruzzo	5.63	0.26	0.46	0.13
Basilicata	2.17	0.23	0.27	0.68
Calabria	1.58	1.10	0.37	0.28
Campania	4.68	1.05	0.77	5.15
Emilia-Romagna	5.97	1.27	1.44	0.64
Friuli-Venezia Giulia	31.92	3.45	6.41	10.07
Lazio	2.83	0.48	0.64	3.64
Liguria	66.42	3.83	15.72	1.60
Lombardia	1.80	0.34	1.53	0.78
Marche	1.82	0.13	0.65	0.06

```
> apply (m1 ,2, mean)
```

Fungicidi	InsetticidiAcaricidi	Erbicidi
12.482	1.214	2.826

Vari
2.303

```
> apply (m1 ,2, var )
```

Fungicidi	InsetticidiAcaricidi	Erbicidi
442.611818	1.804693	23.844960

Vari
10.253134

```
>w1<-cov (m1)
>w1
```

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	442.61182	25.709436	101.966353	20.292560
InsetticidiAcaricidi	25.70944	1.804693	5.794707	2.405531
Erbicidi	101.96635	5.794707	23.844960	3.712669
Vari	20.29256	2.405531	3.712669	10.253134

Matrice di non omogeneità statistica:

> n1 <-nrow(m1)

> h1 <-(n1 -1) *w1

>h1

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	3983.5064	231.38492	917.69718	182.63304
InsetticidiAcaricidi	231.3849	16.24224	52.15236	21.64978
Erbicidi	917.6972	52.15236	214.60464	33.41402
Vari	182.6330	21.64978	33.41402	92.27821

Misura di non omogeneità statistica:

> tr1 <-sum (diag (h1))

>tr1

[1] 4306.631

Questa misura è grande quindi già possiamo pensare che suddividere nei primi 10 e nei secondi 10 non è la scelta ottimale, andiamo avanti per il secondo gruppo:

Per m2:

>m2<-data.frame(Fungicidi=c(0.76,5.39,3.95,1.77,5.89,4.12,0.55,2.82,18.14,1.41),
InsetticidiAcaricidi=c(0.09,0.55,0.65,0.18,0.90,0.25,0.36,0.07,1.96,0.44),
Erbicidi=c(0.28,1.77,0.64,0.24,0.51,0.58,0.02,0.48,2.58,0.05),
Vari=c(0.12,0.27,0.26,0.20,2.10,0.21,0.63,0.06,0.22,0.01))

>row.names(m2)<-c("Molise","Piemonte","Puglia","Sardegna","Sicilia","Toscana","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto")

>m2

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Molise	0.76	0.09	0.28	0.12
Piemonte	5.39	0.55	1.77	0.27
Puglia	3.95	0.65	0.64	0.26
Sardegna	1.77	0.18	0.24	0.20
Sicilia	5.89	0.90	0.51	2.10
Toscana	4.12	0.25	0.58	0.21
Trentino-Alto Aldige	0.55	0.36	0.02	0.63
Umbria	2.82	0.07	0.48	0.06
Valle d'Aosta	18.14	1.96	2.58	0.22
Veneto	1.41	0.44	0.05	0.01

> apply (m2 ,2, mean)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
4.480	0.545	0.715	0.408

> apply (m2 ,2, var)

Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
26.5144667	0.3152722	0.6724944	0.3814844

> w2 <-cov (m2)

>w2

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	26.514467	2.71157778	3.81152222	0.26710000
InsetticidiAcaricidi	2.711578	0.31527222	0.36497222	0.08331111
Erbicidi	3.811522	0.36497222	0.67249444	-0.04322222
Vari	0.267100	0.08331111	-0.04322222	0.38148444

Notiamo che in questo caso c'è qualche correlazione negativa.

Matrice di non omogeneità statistica:

> n2 <-nrow (m2)

> h2 <-(n2 -1) *w2

> h2

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
Fungicidi	238.6302	24.40420	34.30370	2.40390
InsetticidiAcaricidi	24.4042	2.83745	3.28475	0.74980
Erbicidi	34.3037	3.28475	6.05245	-0.38900
Vari	2.4039	0.74980	-0.38900	3.43336

Misura di non omogeneità statistica:

>tr2 <-sum (diag(h2))

>tr2

[1] 250.9535

Questa volta la non omogeneità interna è più piccola ma è comunque troppo grande.

Vediamo quella tra cluster:

> meanm1 <-apply (m1 ,2, mean)

>meanm2 <-apply (m2 ,2, mean)

Traccia tra i cluster:

> trBetween <-((n1*n2)/(n1+n2))*sum ((meanm1 - meanm2)^2)

>trBetween

[1] 362.6346

Misura di non omogeneità totale:

> trUnion <-tr1 +tr2 + trBetween

>trUnion

[1] 4920.219

Matrice di non omogeneità statistica tra i due cluster:

> d<-apply (m1 ,2, mean)-apply (m2 ,2, mean)

> hBetween <-(n1*n2/(n1+n2))*d%*%t(d)

> hBetween

```
     Fungicidi InsetticidiAcaricidi  Erbicidi      Vari
[1,] 320.16002           26.766690 84.461110 75.818950
[2,]  26.76669            2.237805  7.061295  6.338775
[3,]  84.46111            7.061295 22.281605 20.001725
[4,]  75.81895            6.338775 20.001725 17.955125
```

Misura di non omogeneità tra i cluster:

```
> trBetween <-sum (diag( hBetween ))
> trBetween
[1] 362.6346
```

Notiamo che la misura di non omogeneità statistica dei due cluster(m1= 4306.631 ed m2=250.9535) è minore della misura di non omogeneità ottenuta unendoli (4920.219).
Inoltre la misura di non omogeneità interna (tr1+tr2=4557.5845) è superiore alla misura di non omogeneità tra i cluster (362.6346). Questo vuol dire che la divisione dei cluster non è stata fatta nel modo giusto.

Vediamo quindi come possiamo ottimizzare la scelta dei cluster:

Studiamo metodi gerarchici e non gerarchici.

Gerarchici:

I metodi gerarchici di tipo *agglomerativo* partono da una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo per giungere, attraverso successive unioni dei cluster meno distanti tra loro, ad una situazione in cui si ha un solo cluster che contiene tutti gli n individui.
Otterremo una sequenza di partizioni detto dendrogramma, sulle ordinate avremo le distanze, sulle ascisse gli individui.
Vari metodi possono essere utilizzati.

- **Metodo del legame singolo:** la distanza tra i gruppi è la minima tra tutte le distanze che si possono calcolare tra gli individui dei gruppi. Questa tecnica mette meglio in risalto i valori anomali, tuttavia possono ad un certo punto, proprio a causa del legame singolo, trovarsi nello stesso cluster individui non simili.
  Utilizziamo la matrice m non scalata e calcoliamo:

  ```
  >d <-dist (m, method ="euclidean",diag=TRUE , upper = TRUE)
  >hls <-hclust (d,method ="single")
  >str (hls )
  ```

```
List of 7
 $ merge      : int [1:19, 1:2] -10 -13 -20 -2 -11 4 -3 -18 -5 -9 ...
 $ height     : num [1:19] 0.439 0.442 0.519 0.628 0.667 ...
 $ order      : int [1:20] 8 6 19 15 13 16 9 18 3 2 ...
 $ labels     : chr [1:20] "Abruzzo" "Basilicata" "Calabria" "Campania" ...
 $ method     : chr "single"
 $ call       : language hclust(d = d, method = "single")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

$merge contiene la sequenza del processo di agglomerazione

$height indica il livello di distanza a cui è avvenuta la fusione di due cluster

$order è una permutazione delle regioni per costruire il dendrogramma

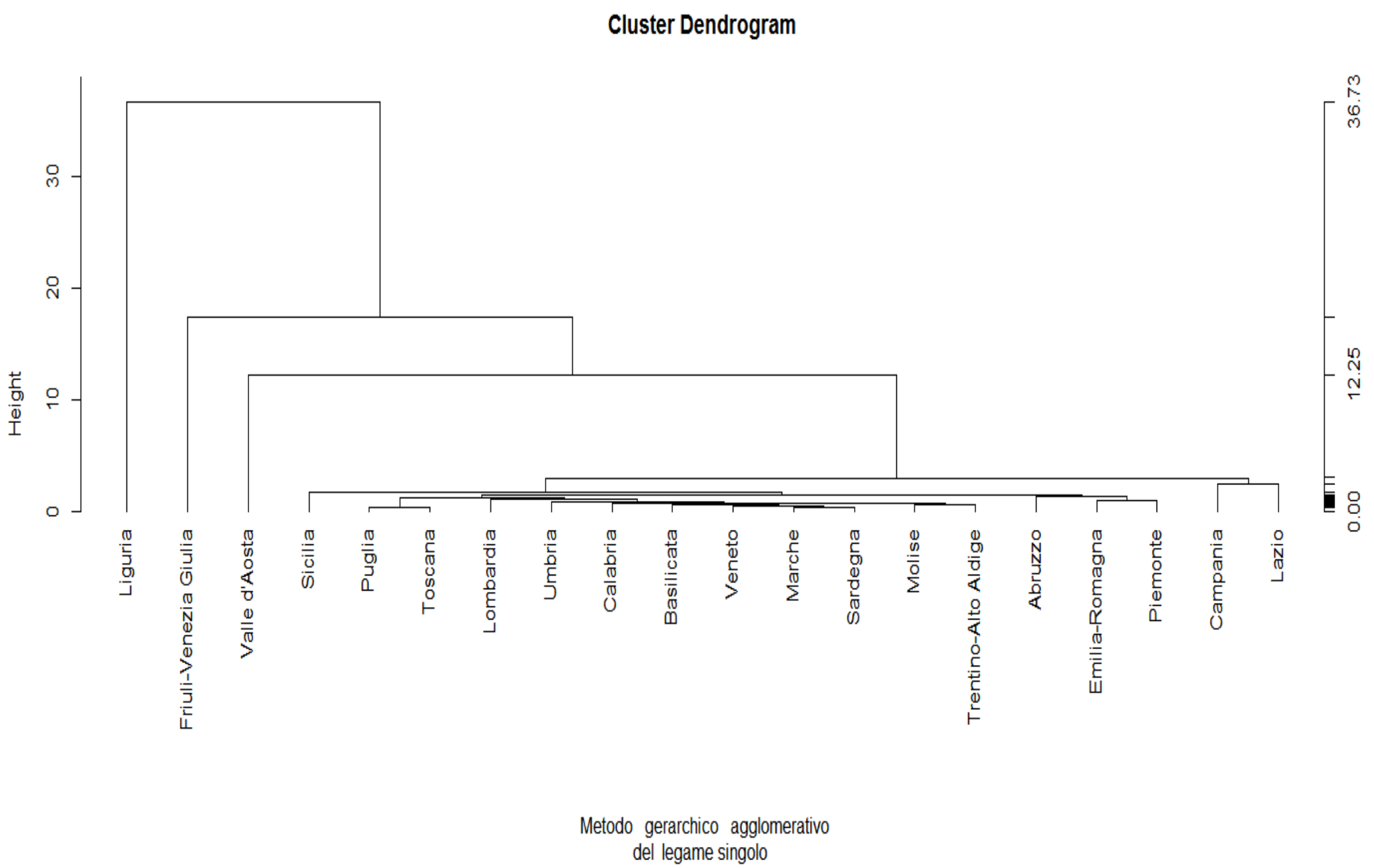
$labels sono le etichette

Dendrogramma:

> plot(hls ,hang =-1, xlab=" Metodo gerarchico agglomerativo", sub ="del legame singolo ")

> axis(side =4, at=round (c(0, hls $ height) ,2))

Nella pagina seguente il dendrogramma.


Cluster Dendrogram
Height
0
10
20
30
Liguria
Friuli-Venezia Giulia
Valle d'Aosta
Sicilia
Puglia
Toscana
Lombardia
Umbria
Calabria
Basilicata
Veneto
Marche
Sardegna
Molise
Trentino-Alto Aldige
Abruzzo
Emilia-Romagna
Piemonte
Campania
Lazio
0.00
12.25
36.73
Metodo gerarchico agglomerativo
del legame singolo

- **Metodo del legame completo:** In questo metodo la distanza tra gruppi viene presa come la massima distanza tra gli individui, essa rappresenta il diametro della sfera che contiene i punti.

  > hlc <-hclust (d, method ="complete")

  >str(hlc)

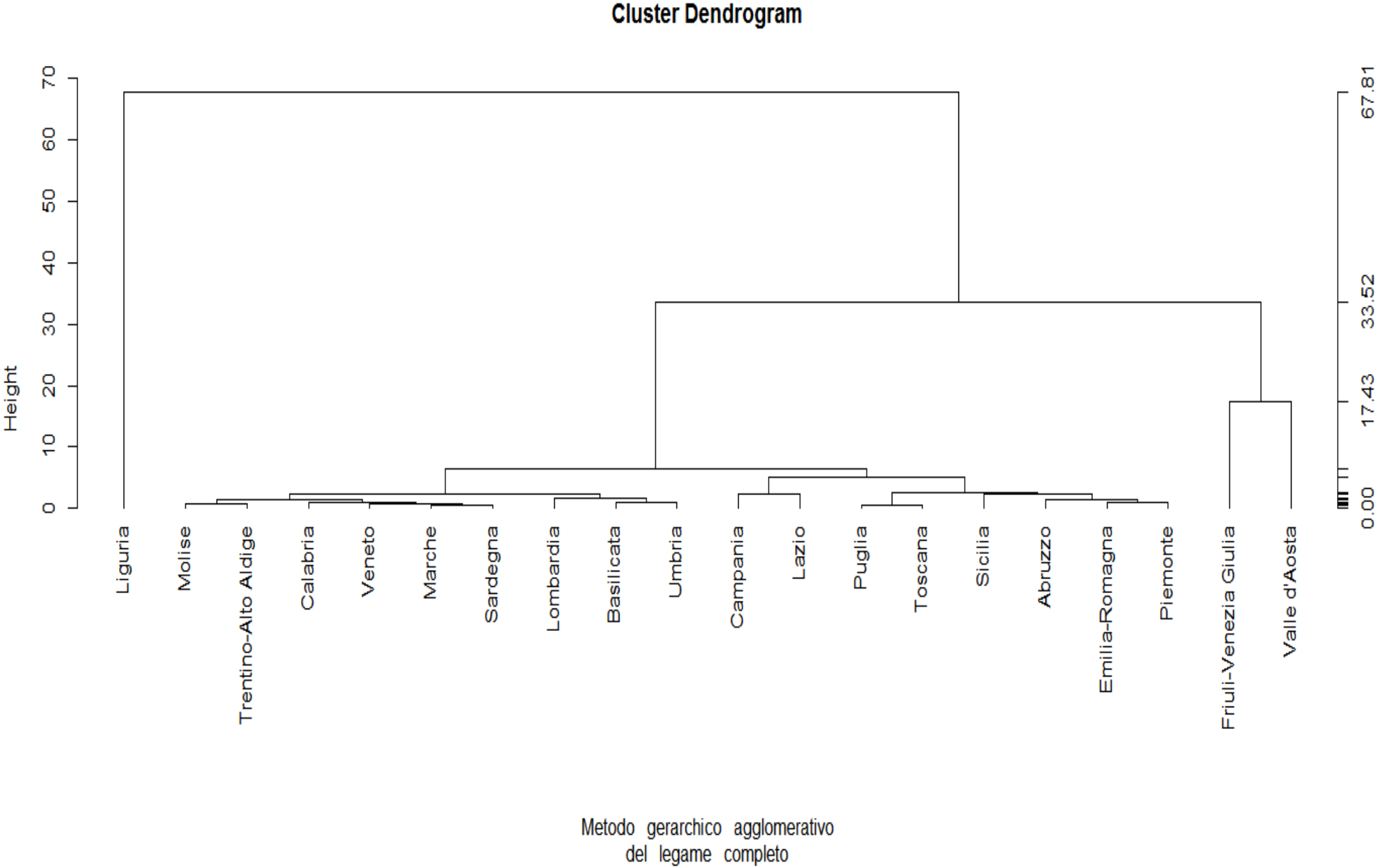
```
List of 7
 $ merge      : int [1:19, 1:2] -10 -13 -11 -20 -2 -5 -3 -1 3 -9 ...
 $ height     : num [1:19] 0.439 0.442 0.667 0.792 0.936 ...
 $ order      : int [1:20] 8 11 17 3 20 10 14 9 2 18 ...
 $ labels     : chr [1:20] "Abruzzo" "Basilicata" "Calabria" "Campania" ...
 $ method     : chr "complete"
 $ call       : language hclust(d = d, method = "complete")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

  Dendrogramma:

  > plot(hlc ,hang =-1, xlab="Metodo gerarchico agglomerativo", sub =" del legame completo ")

  >axis(side =4, at= round (c(0, hlc $ height) ,2))

Nella pagina seguente il dendrogramma.


Cluster Dendrogram
Height
0
10
20
30
40
50
60
70
Liguria
Molise
Trentino-Alto Aldige
Calabria
Veneto
Marche
Sardegna
Lombardia
Basilicata
Umbria
Campania
Lazio
Puglia
Toscana
Sicilia
Abruzzo
Emilia-Romagna
Piemonte
Friuli-Venezia Giulia
Valle d'Aosta
0.00
17.43
33.52
67.81
Metodo gerarchico agglomerativo
del legame completo

Notiamo che la suddivisione in cluster risulta leggermente diversa nel secondo caso, nel secondo caso il salto dal primo al secondo livello è molto più elevato.
Possiamo spiegare questo fenomeno dicendo che il dendrogramma con il metodo del legame completo privilegia l'omogeneità tra gli elementi del gruppo a scapito della differenziazione tra gruppi, quindi abbiamo rami più lunghi rispetto al quelli del dendrogramma precedente.

- **Metodo del legame medio**
  Nel metodo del legame medio la distanza tra i gruppi è la media aritmetica delle distanze tra tutte le coppie che compongono i due gruppi. Il problema del procedere in questo modo è che se le misure sono troppo distanti (come nel nostro caso), la distanza media tra gli elementi del cluster sarà molto più vicina a quella del cluster più numeroso.

  > hlm <-hclust (d, method ="average")
  >str(hlm)

```
List of 7
 $ merge      : int [1:19, 1:2] -10 -13 -20 -11 -2 -3 -5 -18 4 -1 ...
 $ height     : num [1:19] 0.439 0.442 0.655 0.667 0.833 ...
 $ order      : int [1:20] 8 6 19 4 7 9 11 17 18 3 ...
 $ labels     : chr [1:20] "Abruzzo" "Basilicata" "Calabria" "Campania" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

  Dendrogramma:
  > plot(hlm ,hang =-1, xlab=" Metodo gerarchico agglomerativo", sub ="del legame medio ")
  > axis(side =4, at=round (c(0, hlm $ height) ,2))

Nella pagina seguente il dendrogramma.

Cluster Dendrogram
Height
0
10
20
30
40
50
60
Liguria
Friuli-Venezia Giulia
Valle d'Aosta
Campania
Lazio
Lombardia
Molise
Trentino-Alto Adige
Umbria
Calabria
Basilicata
Veneto
Marche
Sardegna
Sicilia
Puglia
Toscana
Abruzzo
Emilia-Romagna
Piemonte
0.00
15.30
30.19
62.88
Metodo gerarchico agglomerativo
del legame medio

Quest'ultimo è molto simile al dendrogramma del legame singolo per quanto riguarda la suddivisione in cluster, tuttavia i salti di questo risultano avere valori intermedi tra quelli del legame singolo e del completo.

- **Metodo del centroide**: la distanza tra i gruppi è la distanza tra le medie campionarie sugli individui appartenenti a due gruppi, dette appunto centroidi.
  Questa procedura può dare origine ad una sorta di fenomeno gravitazionale per cui i gruppi piccoli tendono ad essere attratti da quelli più grandi. Le distanze in cui avvengono poi le successive agglomerazioni quindi possono essere non crescenti. Anche qui come svantaggio, se le misure di due cluster da unire sono molto diverse otterremo che il nuovo centroide sarà molto vicino a quello del cluster più numeroso.

  Dobbiamo calcolare la matrice dei quadrati delle distanze euclidee poiché sono questi ultimi ad essere utilizzati in questo metodo:

  >d2<-d^2

  >hc <-hclust (d2,method ="centroid")
  >str(hc)

```
List of 7
 $ merge      : int [1:19, 1:2] -10 -13 -20 -11 -2 -3 -5 4 -18 -1 ...
 $ height     : num [1:19] 0.193 0.195 0.4 0.445 0.604 ...
 $ order      : int [1:20] 8 6 19 4 7 15 1 5 12 18 ...
 $ labels     : chr [1:20] "Abruzzo" "Basilicata" "Calabria" "Campania" ...
 $ method     : chr "centroid"
 $ call       : language hclust(d = d2, method = "centroid")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

  Dendrogramma:
  > plot(hc , hang=-1, xlab =" Metodo gerarchico agglomerativo", sub ="del centroide ")
  > axis(side =4, at=round (c(0, hc$ height) ,2))

Nella pagina seguente il dendrogramma.

Cluster Dendrogram
Metodo gerarchico agglomerativo
del centroide
Height
0
1000
2000
3000
4000
Liguria
Friuli-Venezia Giulia
Valle d'Aosta
Campania
Lazio
Sicilia
Abruzzo
Emilia-Romagna
Piemonte
Umbria
Puglia
Toscana
Lombardia
Molise
Trentino-Alto Adige
Calabria
Basilicata
Veneto
Marche
Sardegna
0.00
907.08
3944.91

- **Metodo della mediana:** questo metodo è simile a quello del centroide, ma è indipendente dalla numerosità dei cluster. Infatti il centroide risultante dall'aggregazione di due gruppi è la semisomma dei centroidi precedenti.
  Anche questo metodo, come quello del legame singolo può dare origine a delle catene.

  >hmed <-hclust (d2 , method ="median")
  > str(hmed)

```
List of 7
 $ merge      : int [1:19, 1:2] -10 -13 -20 -11 -2 -3 -5 4 -18 -1 ...
 $ height     : num [1:19] 0.193 0.195 0.4 0.445 0.699 ...
 $ order      : int [1:20] 8 6 19 4 7 15 1 5 12 18 ...
 $ labels     : chr [1:20] "Abruzzo" "Basilicata" "Calabria" "Campania" ...
 $ method     : chr "median"
 $ call       : language hclust(d = d2, method = "median")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```

  Dendrogramma
  > plot(hmed , hang=-1, xlab =" Metodo gerarchico agglomerativo", sub =" della mediana ")
  > axis(side =4, at= round (c(0, hmed$ height) ,2))

Cluster Dendrogram
Metodo gerarchico agglomerativo
della mediana
Height
0
500
1000
1500
2000
Liguria
Friuli-Venezia Giulia
Valle d'Aosta
Campania
Lazio
Sicilia
Abruzzo
Emilia-Romagna
Piemonte
Umbria
Puglia
Toscana
Lombardia
Molise
Trentino-Alto Adige
Calabria
Basilicata
Veneto
Marche
Sardegna
0.00
538.41
2175.54

Questo dendrogramma è molto simile al precedente, se non per l'ultimo livello che risulta avere un salto differente.

- **Metodo di Lance e Williams:** questo metodo è ricorsivo, infatti il calcolo della matrice dei quadrati delle distanze dipende unicamente dalla medesima matrice al livello precedente, esso include tutti i metodi precedentemente visti.
  Nella sua formula infatti compaiono dei coefficienti che variano a seconda del metodo che si vuole utilizzare.

In conclusione ogni metodo ha un suo scopo preciso ed ha vantaggi e svantaggi. Nel nostro caso non abbiamo informazioni sulla struttura dell'insieme che stiamo analizzando, siamo solo a conoscenza di valori anomali in Liguria, Friuli-Venezia Giulia e Valle d'Aosta. Infatti in tutti i metodi queste regioni vengono isolate sin dai primi passaggi, possiamo quindi considerare validi tutti i metodi visti, in caso di indecisione comunque si preferisce il metodo del legame singolo.

Screeplot

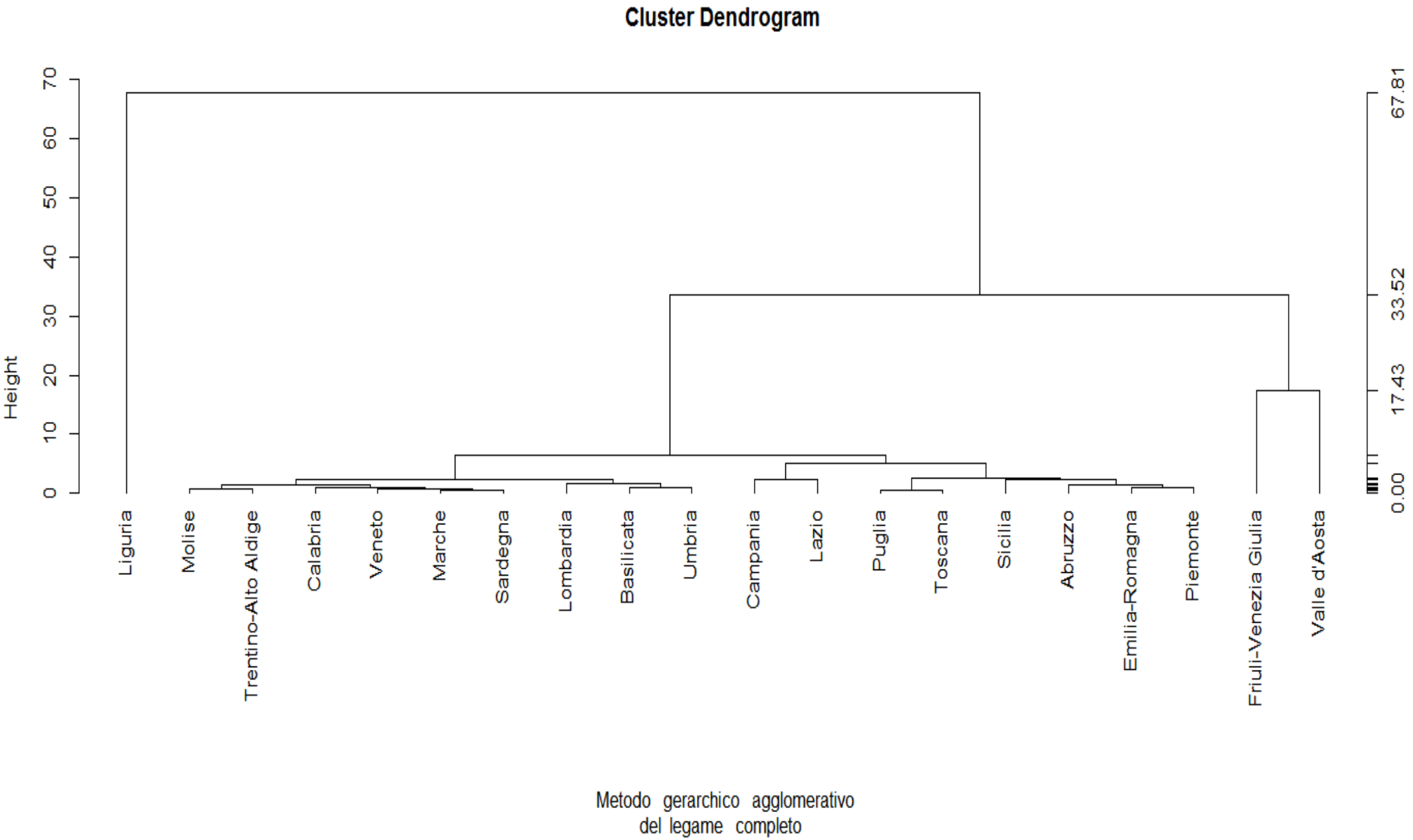
Lo screeplot ci aiuta a scegliere una buona partizione del dendrogramma. In esso poniamo sull'asse delle ordinate i gruppi di numeri ottenuti con un metodo gerarchico (legame singolo) e su quello delle ascisse le distanze a cui avvengono le successive aggregazioni tra gruppi. Lo scopo è: se passando da più gruppi a meno gruppi otteniamo un forte incremento della distanza, tagliamo il dendrogramma al passo precedente. Possiamo quindi avere una visione complessiva delle altezze a cui sono avvenute le agglomerazioni.

```
> d <-dist (m, method ="euclidean",diag=TRUE , upper = TRUE)
>hlc <-hclust (d,method ="complete")
```

```
 [1]  0.4389761  0.4415880  0.6668583  0.7916439  0.9362692  1.0490948
 [7]  1.0608016  1.5349919  1.5451861  1.6535417  2.3041267  2.4028109
[13]  2.4585361  2.6886056  5.1791022  6.4675343 17.4298566 33.5197061
[19] 67.8109777
```

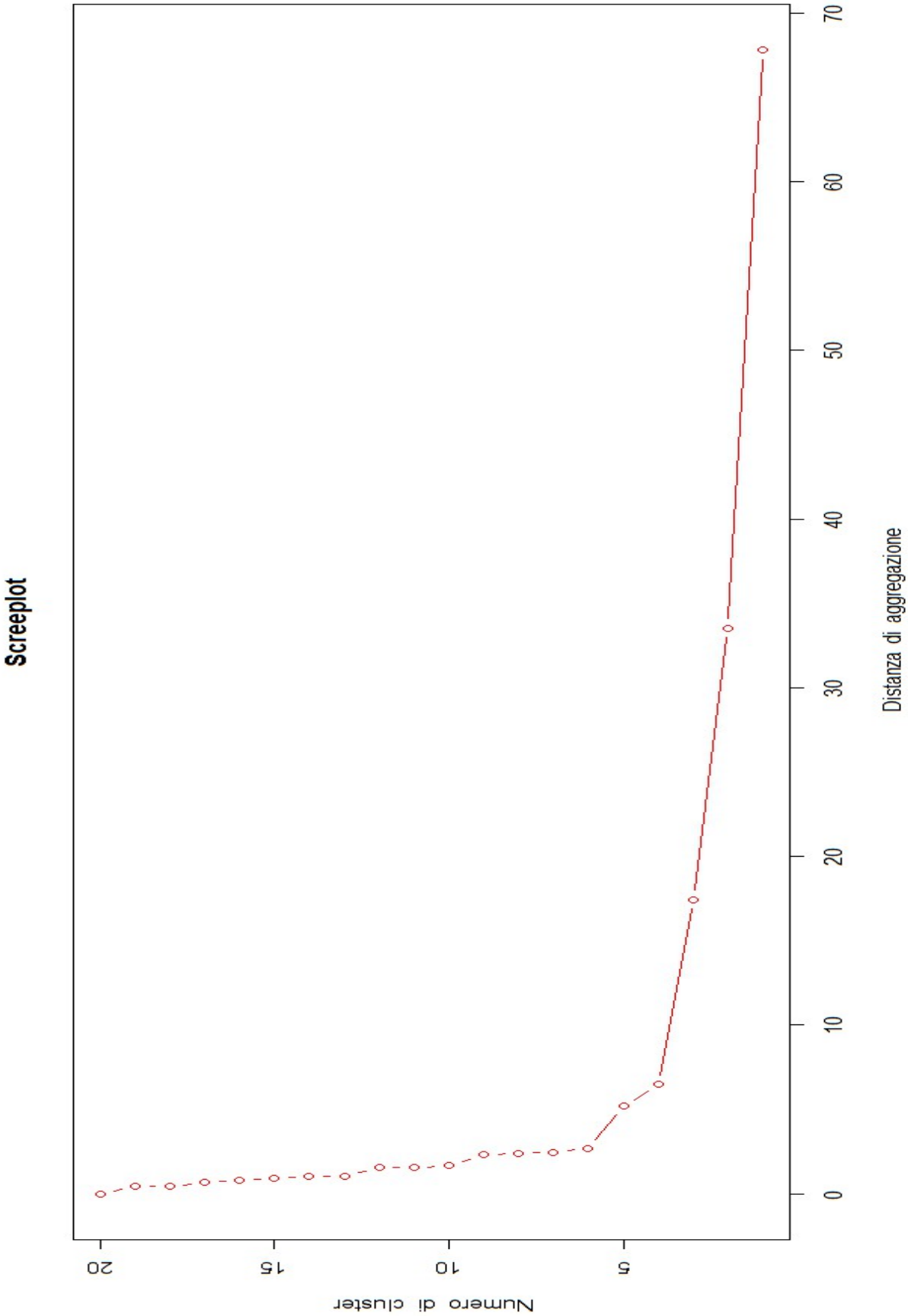
```
> plot(hlc ,hang =-1, xlab=" Metodo  gerarchico  agglomerativo", sub ="del  legame  completo ")
> axis( side =4, at=round (c(0, hlc $ height ) ,2))
```

Grafico in pagine seguenti.


Cluster Dendrogram
Height
0
10
20
30
40
50
60
70
Liguria
Molise
Trentino-Alto Aldige
Calabria
Veneto
Marche
Sardegna
Lombardia
Basilicata
Umbria
Campania
Lazio
Puglia
Toscana
Sicilia
Abruzzo
Emilia-Romagna
Piemonte
Friuli-Venezia Giulia
Valle d'Aosta
0.00
17.43
33.52
67.81
Metodo gerarchico agglomerativo
del legame completo

Occupiamoci ora del grafico screeplot:

> plot(rev (c(0, hlc$height)),seq (1 ,20) ,type="b", main=" Screeplot ",xlab=" Distanza di aggregazione", ylab="Numero di cluster ", col ="red ")



Il grafico suggerisce una suddivisione in 4 gruppi, poiché:

67,8109777-33,5197061=34,2912716
33,5197061-17,4298566=16,0898495
17,4298566-6,4675343=10,9623223

Le altre distanze sono tutte inferiori a 5.

Analisi dettagliata del dendrogramma.

Visti tutti i metodi, e in particolare notato che per i nostri dati essi si comportano in maniera simile, ne scegliamo uno e lo analizziamo nel dettaglio.
Utilizziamo ad esempio il metodo del centroide.

Quindi abbiamo le medesime linee di codice viste precedentemente:

```
>d <-dist (m, method ="euclidean",diag=TRUE , upper = TRUE)
>d2<-d^2
>hc <-hclust (d2,method ="centroid")
> plot(hc , hang=-1, xlab =" Metodo   gerarchico agglomerativo", sub ="del   centroide ")
> axis( side =4, at=round (c(0, hc$height ) ,2))
```

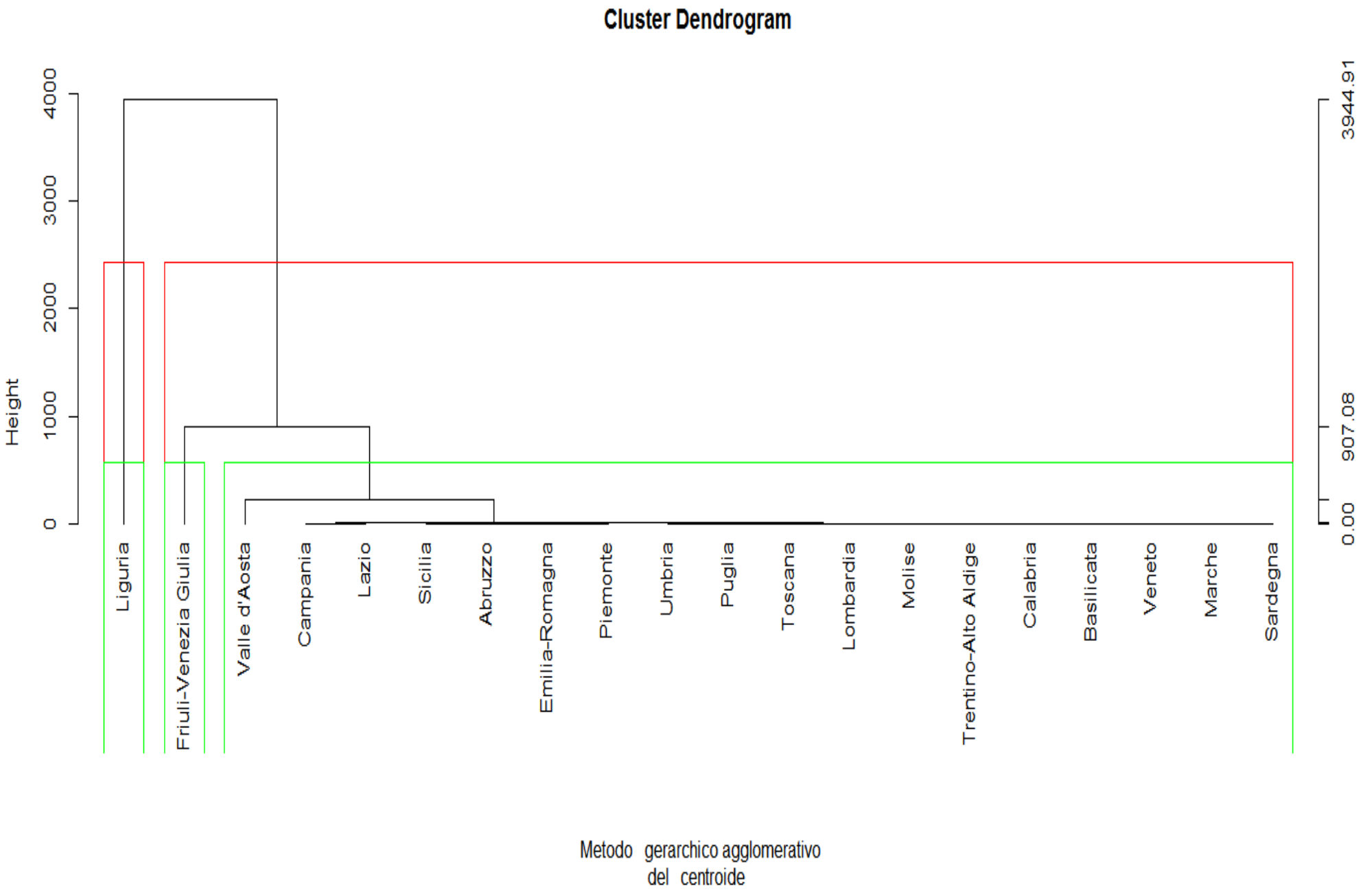
Vogliamo ad esempio evidenziare le prime due partizioni che si ottengono con il metodo:

```
> rect.hclust (hc , k = 2, border = "red ")
```

e magari anche quelle al passo successivo:

```
> rect.hclust (hc , k = 3, border = " green ")
```

Otteniamo quindi:
(grafico pagina seguente)


Cluster Dendrogram
Height
0
1000
2000
3000
4000
3944.91
907.08
0.00
Liguria
Friuli-Venezia Giulia
Valle d'Aosta
Campania
Lazio
Sicilia
Abruzzo
Emilia-Romagna
Piemonte
Umbria
Puglia
Toscana
Lombardia
Molise
Trentino-Alto Aldige
Calabria
Basilicata
Veneto
Marche
Sardegna
Metodo gerarchico agglomerativo
del centroide

Vediamo ora la suddivisione delle regioni in cluster utilizzando il comando *cutree*:

>cutree (hc, k = 2, h = NULL)

Abruzzo	Basilicata	Calabria
1	1	1
Campania	Emilia-Romagna	Friuli-Venezia Giulia
1	1	1
Lazio	Liguria	Lombardia
1	2	1
Marche	Molise	Piemonte
1	1	1
Puglia	Sardegna	Sicilia
1	1	1
Toscana	Trentino-Alto Aldige	Umbria
1	1	1
Valle d'Aosta	Veneto	
1	1	

Come ci aspettavamo, nel primo passaggio e quindi la divisione in due cluster solo la Liguria è stata isolata, poiché ci ricordiamo che essa presenta i valori anomali più consistenti.

Proviamo per k=3

>cutree (hc, k = 3, h = NULL)

Abruzzo	Basilicata	Calabria
1	1	1
Campania	Emilia-Romagna	Friuli-Venezia Giulia
1	1	2
Lazio	Liguria	Lombardia
1	3	1
Marche	Molise	Piemonte
1	1	1
Puglia	Sardegna	Sicilia
1	1	1
Toscana	Trentino-Alto Aldige	Umbria
1	1	1
Valle d'Aosta	Veneto	
1	1	

Vediamo che ora i cluster sono 3 poiché ricordiamo che anche il Friuli-Venezia Giulia presentava dei valori anomali.

Ci aspettiamo quindi che per k=4 i cluster saranno 4 e che venga isolata la Valle d'Aosta, l'altra regione che presentava valori molto al di sopra della norma, isolati quindi anche Friuli-Venezia Giulia e Liguria, e le altre regioni cadranno nello stesso cluster, infatti:

>cutree (hc, k = 4, h = NULL)

Abruzzo	Basilicata	Calabria
1	1	1
Campania	Emilia-Romagna	Friuli-Venezia Giulia
1	1	2
Lazio	Liguria	Lombardia
1	3	1
Marche	Molise	Piemonte
1	1	1
Puglia	Sardegna	Sicilia
1	1	1
Toscana	Trentino-Alto Aldige	Umbria
1	1	1
Valle d'Aosta	Veneto	
4	1	

Tre sono le regioni con valori anomali, per questo il numero minimo e indispensabile di cluster da creare sono 4 (creati proprio al passo 4) per isolare le anomalie.
La situazione diventa poi da questo punto in poi più dettagliata.
Ad esempio, per k=10 abbiamo:

>cutree (hc, k = 10, h = NULL)

Abruzzo	Basilicata	Calabria
1	2	2
Campania	Emilia-Romagna	Friuli-Venezia Giulia
3	1	4
Lazio	Liguria	Lombardia
5	6	7
Marche	Molise	Piemonte
2	2	1
Puglia	Sardegna	Sicilia
8	2	9
Toscana	Trentino-Alto Aldige	Umbria
8	2	8
Valle d'Aosta	Veneto	
10	2	

Che ci fornisce una suddivisione già molto più interessante.

Possiamo poi dare uno sguardo alla classificazione degli individui all'aumentare del numero dei cluster:

>cutree (hc , k = 1:10)

	1	2	3	4	5	6	7	8	9	10
Abruzzo	1	1	1	1	1	1	1	1	1	1
Basilicata	1	1	1	1	1	2	2	2	2	2
Calabria	1	1	1	1	1	2	2	2	2	2
Campania	1	1	1	1	2	3	3	3	3	3
Emilia-Romagna	1	1	1	1	1	1	1	1	1	1
Friuli-Venezia Giulia	1	1	2	2	3	4	4	4	4	4
Lazio	1	1	1	1	2	3	5	5	5	5
Liguria	1	2	3	3	4	5	6	6	6	6
Lombardia	1	1	1	1	1	2	2	2	2	7
Marche	1	1	1	1	1	2	2	2	2	2
Molise	1	1	1	1	1	2	2	2	2	2
Piemonte	1	1	1	1	1	1	1	1	1	1
Puglia	1	1	1	1	1	2	2	7	7	8
Sardegna	1	1	1	1	1	2	2	2	2	2
Sicilia	1	1	1	1	1	1	1	1	8	9
Toscana	1	1	1	1	1	2	2	7	7	8
Trentino-Alto Aldige	1	1	1	1	1	2	2	2	2	2
Umbria	1	1	1	1	1	2	2	7	7	8
Valle d'Aosta	1	1	1	4	5	6	7	8	9	10
Veneto	1	1	1	1	1	2	2	2	2	2

Saremo poi naturalmente interessati a conoscere media, varianza e deviazione standard campionaria di questi nuovi gruppi ottenuti. Facciamo questo con il comando *aggregate*.

```
>taglio <-cutree (hc , k =6, h = NULL)
> tagliolist <-list( taglio )
> aggregate (m, tagliolist , mean )
```

	Group.1	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
1	1	5.720000	0.7450000	1.0450000	0.7850000
2	2	2.068182	0.3490909	0.4645455	0.2990909
3	3	3.755000	0.7650000	0.7050000	4.3950000
4	4	31.920000	3.4500000	6.4100000	10.0700000
5	5	66.420000	3.8300000	15.7200000	1.6000000
6	6	18.140000	1.9600000	2.5800000	0.2200000

```
> aggregate (m, tagliolist , var )
```

	Group.1	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
1	1	0.06946667	0.19096667	0.4367000	0.81483333
2	2	1.32369636	0.09084909	0.1714673	0.07354909
3	3	1.71125000	0.16245000	0.0084500	1.14005000
4	4	NA	NA	NA	NA
5	5	NA	NA	NA	NA
6	6	NA	NA	NA	NA

```
> aggregate (m, tagliolist , sd )
```

	Group.1	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
1	1	0.2635653	0.4369973	0.66083281	0.9026812
2	2	1.1505200	0.3014118	0.41408607	0.2711994
3	3	1.3081475	0.4030509	0.09192388	1.0677312
4	4	NA	NA	NA	NA
5	5	NA	NA	NA	NA
6	6	NA	NA	NA	NA

Ovviamente gli ultimi tre valori non sono pervenuti, e comunque non di interesse, perché in quei cluster risulta esserci un solo elemento.

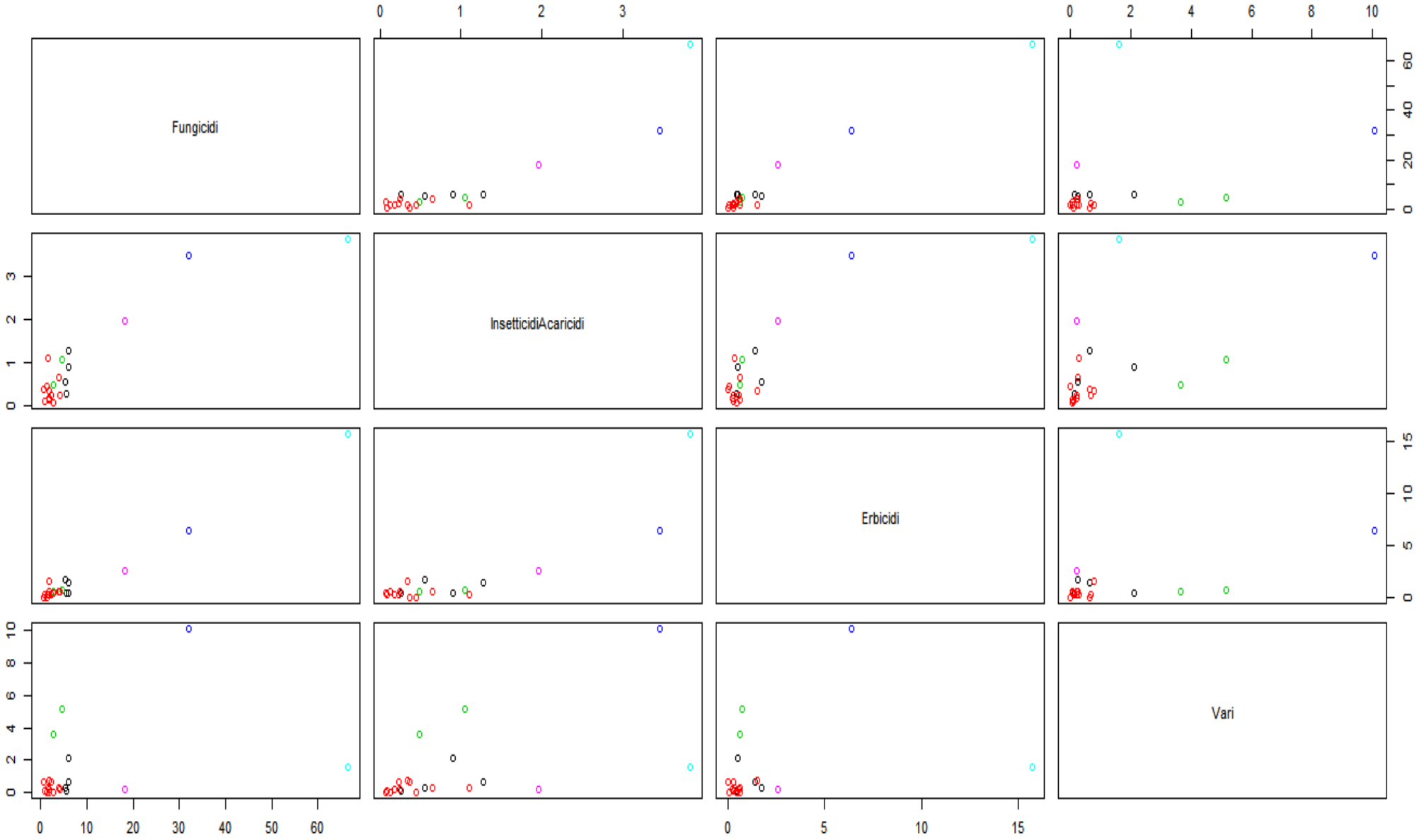
Vediamo che già con la sola esclusione dei valori anomali, questi indici sono molto più rappresentativi della realtà.

Ora visualizziamo a due a due la divisione in cluster a seconda della coppia di sostanze scelte:

```
> agmean <- aggregate (m, tagliolist , mean)[, -1]
> plot(m, col = taglio , main = " Metodo  gerarchico  del  centroide ")
> points (agmean ,col = 1:10, pch =20, cex =1)
```

Grafico a pagina seguente.

Metodo gerarchico del centroide



Ora che la divisione in cluster è avvenuta (scegliamo k=6 da questo punto in poi) possiamo studiare le misure di non omogeneità statistiche totali, quindi la somma della misura di non omogeneità tra cluster(trS), e per le singole regioni di uno stesso cluster(trB).

$$\text{tr}\, T = \text{tr}\, S + \text{tr}\, B$$

Sappiamo che siccome i dati sono fissati la trT è sempre fissata. Dobbiamo quindi cercare di minimizzare trB e massimizzare trS.

Calcolo della misura di non omogeneità totale:

```
> n<-nrow(m)
> trHI <-(n -1) *sum ( apply (m,2, var ))
> trHI
[1] 4920.219
```

Richiamiamo i comandi:

```
> d <-dist (m, method ="euclidean",diag=TRUE , upper = TRUE)
> d2 <-d^2
> hc <- hclust (d2 , method = "centroid")
> taglio <-cutree (hc , k =6, h = NULL)
> num <-table (taglio )
> tagliolist <-list (taglio )
> agvar <- aggregate (m, tagliolist , var )[, -1]
```

Misure di non omogeneità dei gruppi:

1)
```
> trH1 <-(num [[1]] -1) * sum ( agvar [1, ])
> trH1
[1] 4.5359
```
2)
```
> trH2 <-(num [[2]] -1) * sum ( agvar [2, ])
> trH2
[1] 16.59562
```
3)
```
> trH3 <-(num [[3]] -1) * sum ( agvar [3, ])
> trH3
[1] 3.0222
```

Per il 4, 5 e 6 non li calcoliamo poiché sappiamo che hanno un singolo elemento e quindi è necessario isolarli.
Notiamo che la somma di queste non omogeneità è molto più piccola della non omogeneità totale, questo ci dice quindi che è grande la non omogeneità tra cluster, proprio come volevamo, per questo la divisione in cluster per k=6 è soddisfacente.

METODI NON GERARCHICI

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.

Il metodo più utilizzato è il k-means, esso richiede a priori il numero di cluster e dei valori di riferimento iniziali (regioni), attribuisce ad ogni regione il cluster contenete il riferimento a distanza minima, calcola il centroide dei gruppi ed esso diviene il nuovo riferimento. In questo modo ad ogni passo sposta le regioni nel cluster a cui sono più vicine e ripete il procedimento fino a raggiungere una configurazione stabile, cioè le regioni restano negli stessi cluster al passo successivo.
Purtroppo con questo metodo il risultato finale può dipendere dai punti scelti in partenza.

Utilizziamo quindi il nostro data.frame m:

```
>m<-
data.frame(Fungicidi=c(5.63,2.17,1.58,4.68,5.97,31.92,2.83,66.42,1.80,1.82,0.76,5.39,3.95,1.77,
5.89,4.12,0.55,2.82,18.14,1.41),
InsetticidiAcaricidi=c(0.26,0.23,1.10,1.05,1.27,3.45,0.48,3.83,0.34,0.13,0.09,0.55,0.65,0.18,0.90,
0.25,0.36,0.07,1.96,0.44), Erbicidi=c(0.46,0.27,
0.37,0.77,1.44,6.41,0.64,15.72,1.53,0.65,0.28,1.77,0.64,0.24,0.51,0.58,0.02,0.48,2.58,0.05),
Vari=c(0.13,0.68,0.28,5.15,0.64,10.07,3.64,1.60,0.78,0.06,0.12,0.27,0.26,0.20,2.10,0.21,0.63,0.0
6,0.22,0.01))
>rownames(m)<-c("Abruzzo","Basilicata","Calabria","Campania","Emilia-Romagna","Friuli-Venezia
Giulia","Lazio","Liguria","Lombardia","Marche","Molise","Piemonte","Puglia","Sardegna","Sicilia","Tosca
na","Trentino-Alto Aldige","Umbria","Valle d'Aosta","Veneto")
```

Scegliamo la divisione in 6 cluster e il numero massimo di iterazioni (10) e il numero di volte in cui ripetere la procedura (1):

```
>km <-kmeans (m, center =6, iter.max =10, nstart =1)
K-means clustering with 6 clusters of sizes 2, 6, 1, 1, 9, 1

Cluster means:
  Fungicidi InsetticidiAcaricidi   Erbicidi       Vari
1  3.755000            0.7650000  0.7050000  4.3950000
2  5.158333            0.6466667  0.9000000  0.6016667
3 66.420000            3.8300000 15.7200000  1.6000000
4 18.140000            1.9600000  2.5800000  0.2200000
5  1.631111            0.3266667  0.4322222  0.3133333
6 31.920000            3.4500000  6.4100000 10.0700000

Clustering vector:
          Abruzzo       Basilicata         Calabria         Campania   Emilia-Romagna
                2                5                5                1                2
```

```
Friuli-Venezia Giulia    Lazio    Liguria     Lombardia      Marche
                    6        1          3             5          5

                  Molise   Piemonte   Puglia  Sardegna    Sicilia
                       5          2        2         5          2

Toscana  Trentino-Alto Aldige              Umbria        Valle d'Aosta               Veneto
      2                    5                   5                    4                    5

Within cluster sum of squares by cluster:
[1] 3.022200 9.190700 0.000000 0.000000 6.955444 0.000000
 (between_SS / total_SS =  99.6 %)

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
             "size"         "iter"         "ifault"
```

La divisione in cluster è diversa da quella ottenuta con i metodi gerarchici visti, ripetendo la procedura per 8 volte il risultato sarà lo stesso.

Tuttavia, come abbiamo detto, siamo partiti da punti di riferimento casuali, per questo ad una diversa scelta otterremo risultatati diversi.
Potremmo invece scegliere come punti di riferimento i centroidi ottenuti con la medesima tecnica gerarchica, quindi:

```
> d <-dist (m, method ="euclidean",diag=TRUE , upper = TRUE)
> d2 <-d^2
> hl <- hclust (d2 , method = "centroid")
> taglio <-cutree (hl , k =6, h =NULL )
> tagliolist <-list( taglio )
> centroidiIniziali <-aggregate (m, tagliolist , mean)[,-1]
> centroidiIniziali
```

	Fungicidi	InsetticidiAcaricidi	Erbicidi	Vari
1	5.720000	0.7450000	1.0450000	0.7850000
2	2.068182	0.3490909	0.4645455	0.2990909
3	3.755000	0.7650000	0.7050000	4.3950000
4	31.920000	3.4500000	6.4100000	10.0700000
5	66.420000	3.8300000	15.7200000	1.6000000
6	18.140000	1.9600000	2.5800000	0.2200000

Ora applichiamo il k-mean:

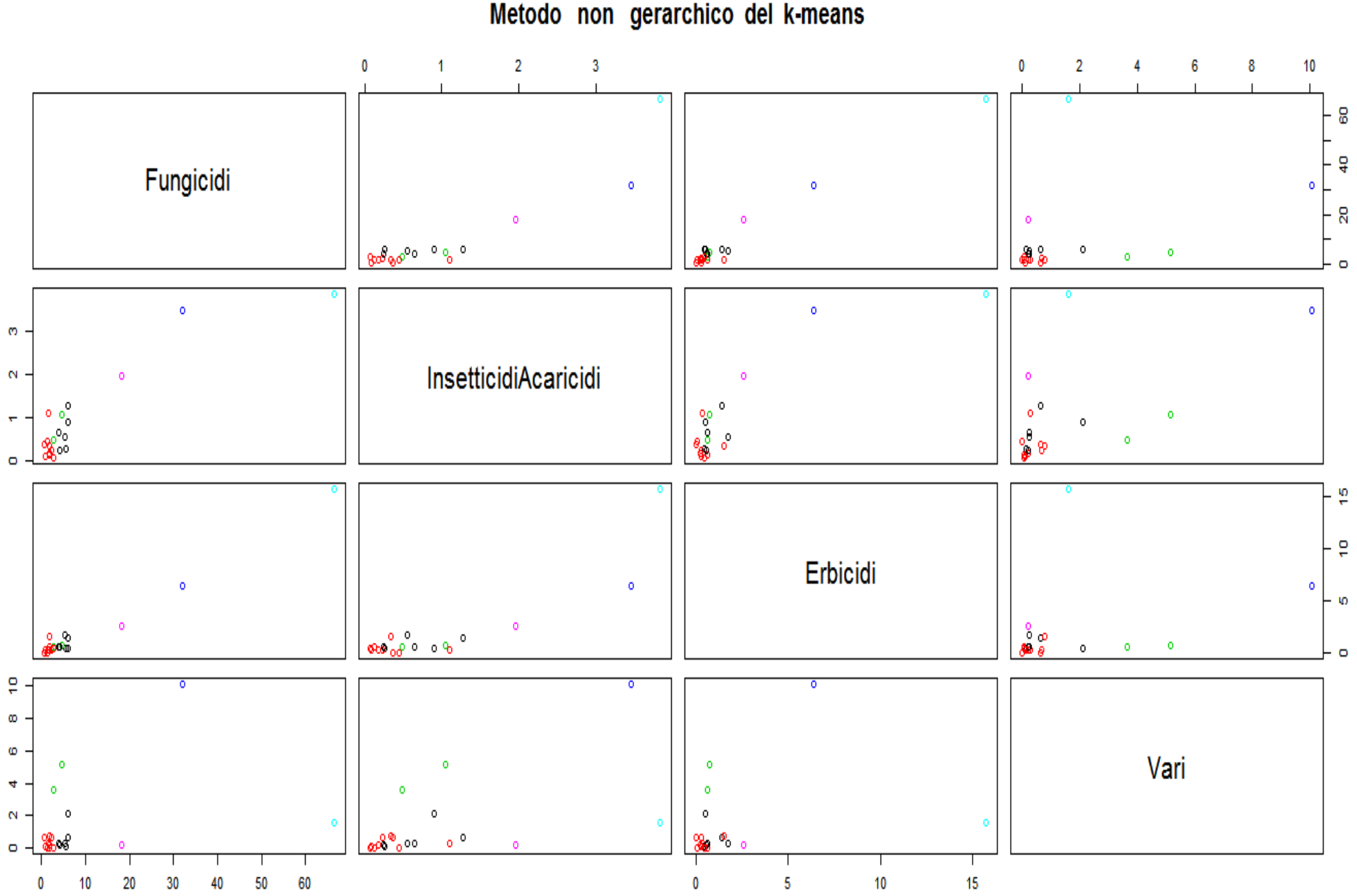
```
> km <-kmeans (m, centers = centroidiIniziali , iter.max = 10)
> km
```

```
K-means clustering with 6 clusters of sizes 6, 9, 2, 1, 1, 1

Cluster means:
  Fungicidi Insetticidi Acaricidi   Erbicidi       Vari
1  5.158333            0.6466667  0.9000000  0.6016667
2  1.631111            0.3266667  0.4322222  0.3133333
3  3.755000            0.7650000  0.7050000  4.3950000
4 31.920000            3.4500000  6.4100000 10.0700000
5 66.420000            3.8300000 15.7200000  1.6000000
6 18.140000            1.9600000  2.5800000  0.2200000


Clustering vector:
      Abruzzo     Basilicata      Calabria      Campania Emilia-Romagna Friuli-Venezia Giulia
            1              2             2             3              1                     4
        Lazio       Liguria     Lombardia       Marche       Molise     Piemonte       Puglia
            3             5             2            2            2            1            1
     Sardegna       Sicilia       Toscana Trentino-Alto Aldige       Umbria Valle d'Aosta
            2             1             1                    2            2              6
       Veneto
            2


Within cluster sum of squares by cluster:
[1] 9.190700 6.955444 3.022200 0.000000 0.000000 0.000000
 (between_SS / total_SS =  99.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
          "betweenss"    "size"         "iter"         "ifault"
```

Rappresentiamoli graficamente:

> plot(m, col = km$cluster , main = " Metodo non gerarchico del k-means ")

> points (km$center , col = 1:2, pch = 8, cex =1)


Metodo non gerarchico del k-means
Fungicidi
InsetticidiAcaricidi
Erbicidi
Vari

In conclusione possiamo dire che:

Con i dati a nostra disposizione hanno avuto particolare importanza i risultati ottenuti con:

- boxplot: abbiamo individuato subito i valori che sono troppo lontani dagli altri, e le regioni quindi per cui sarebbe necessario effettuare un controllo per limitare l'utilizzo di prodotti fitosanitari o considerare prodotti più efficienti per i problemi che affliggono le piantagioni di queste regioni in particolare.
- Indici di dispersione: Nonostante i Fungicidi siano la sostanza più utilizzata considerando i kg/ettaro, i valori per cui l'utilizzo nelle regioni presenta una maggiore variazione da regione a regione interessa gli Erbicidi. Sarebbe quindi interessante fare un'analisi sui tipi di terreno/coltivazioni/condizioni atmosferiche che portano ad una così forte variazione tra regione e regione.
- Analisi dei cluster: sarebbe interessante studiare da un punto di vista biovegetale cosa accomuna regioni come Calabria e Lombardia oppure come Piemonte e Puglia, fisicamente distanti, da farle cadere nello stesso cluster.

Il risultato di questa analisi fornisce quindi, come sopra descritto, degli interessanti spunti per ulteriori lavori di analisi su altri fattori riguardanti le coltivazioni delle regioni.