ClusterAnalisys.R

pr43569

Mon Nov 23 15:32:26 2015

Obiettivi della ricerca

L'obiettivo del documento è trovare un insieme di regioni europee che si ritengono essere strutturalmente simili alla Provincia Autonoma di Trento. La metodologia proposta prevede un analisi di cluster su alcuni indicatori ritenuti importanti. Gli indicatori che si è pensato rappresentativi della situazione demografica e economica regionale sono: * Demografia: - Popolazione media - Indice di vecchiaia - Età media - Superficie * Economia: - Addetti nell'industria - Addetti nei servizi - Addetti totali - Valore aggiunto totale - Percentuale di valore aggiunto dall'agricoltura - Percentuale di valore aggiunto dall'industria - Percentuale di valore aggiunto dai servizi - Occuapti totali - Percentuale di occupati nell'agricoltura - Percentuale di occupati nell'industria - Percentuale di occupati nei servizi * Istruzione: - NEET - Percentuale di persone 25-64 con almeno diploma superiore * Ricerca e innovazione - Percentuale di occupati in settori ad alto contenuto tecnologico o ad alta intensità di conoscenza

Preparazione dataset

Leggo i dati preparati da Paolo nel file 'LASTVALUE.csv'

```
indOrig <-
    read.csv2(
        'LASTVALUE.csv', sep = ',', skip = 21, header = T, stringsAsFactors = F
    )
```

Il Dataset è composto da 272 osservazioni su 18 variabili

Poiché la procedura di clustering non può lavorare con i valori NA sono costretto a eliminare tutte le righe in cui compare un NA. Da notare che si escludono parecchie regioni e interi stati, particolarmente REGNO UNITO, GERMANIA e BELGIO. Delle 272 regioni europee ne rimangono utilizzabili solo 136.

```
indOrig <- na.omit(indOrig)
```

Poiché i valori fanno riferimento a fenomeni diversi tra loro standardizzo gli indicatori.

```
ind <- scale(indOrig[5:22])
row.names(ind) <- indOrig$GEO
```

K-Medie

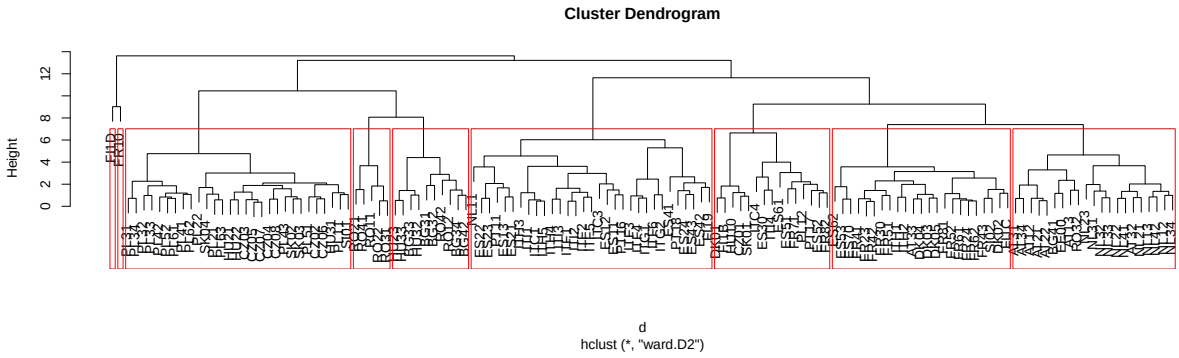
Il metodo di clustering delle K-Medie è il più utilizzato. Richiede che l'analista specifici il numero di cluster da estrarre. Un grafico della somma dei quadrati all'interno dei gruppi per il numero di cluster estratti può aiutare a scegliere il numero più appropriato di cluster. L'analista decide la numerosità di cluster

individuando il punto in cui l'aggiunta di un cluster (asse orizzontale) non comporta un significativo aumento dell'informazione (asse verticale). Tuttavia dopo approfondita analisi il metodo delle K-Medie non risulta adeguato nel caso specifico, in quanto i punti sono troppo ravvicinati tra loro, e non si conformano a dei cluster chiaramente separati, e quindi la procedura delle K-Medie non riesce a determinare i cluster in modo deterministico. Il grafico di seguito mostra la rappresentazione dei punti su uno spazio tridemnsionale utlizzando le prime tre componenti principali.

Metodo gerarchico

Si decide perciò di utilizzare un metodo di clustering gerarchico, che accoppia via via i punti più vicini nello spazio. Si rende però necessario decidere il tipo di distanza da utilizzare, il metodo di clusterizzazione e infine il numero di cluster. Si sceglie di utilizzare la distanza massima per minimizzare le differenze; si sceglie cioè come distanza tra due punti il massimo tra le distanze di ogni dimensione

```
# Matrice delle distanze
d <- dist(ind, method = "maximum")
# Crea l'albero
fit <- hclust(d, method="ward.D2")
# Disegna il grafico
plot(fit)
# Divide l'albero all'altezza di 9 cluster
group <- cutree(fit, k=9)
# Disegna i rettangoli rossi attorno ai 9 cluster
rect.hclust(fit, k=9, border="red")
```



Analisi dei cluster

Trento (ITH2) risulta essere all'interno del cluster numero 2

Il cluster 2 risulta essere così composto:

NUTS2

Denominazione

Stato

AT33

Tirol

AUSTRIA

DK02

Sjælland

DANIMARCA

DK03

Syddanmark

DANIMARCA

DK04

Midtjylland

DANIMARCA

DK05

Nordjylland

DANIMARCA

ES53

Illes Balears

SPAGNA

ES62

Región de Murcia

SPAGNA

ES70

Canarias (ES)

SPAGNA

FI1C

Etelä-Suomi

FINLANDIA

FR23

Haute-Normandie

FRANCIA

FR24

Centre (FR)

FRANCIA

FR30

Nord - Pas-de-Calais

FRANCIA

FR41

Lorraine

FRANCIA

FR42

Alsace

FRANCIA

FR43

Franche-Comté

FRANCIA

FR51

Pays de la Loire

FRANCIA

FR52

Bretagne

FRANCIA

FR61

Aquitaine

FRANCIA

FR62

Midi-Pyrénées

FRANCIA

FR81

Languedoc-Roussillon

FRANCIA

ITH1

Provincia Autonoma di Bolzano/Bozen

ITALIA

ITH2

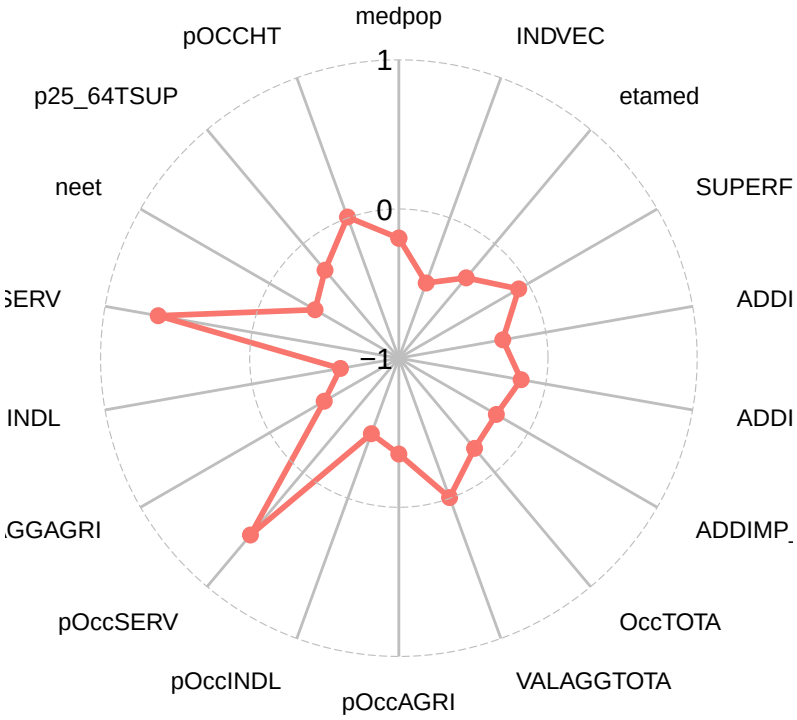
Provincia Autonoma di Trento

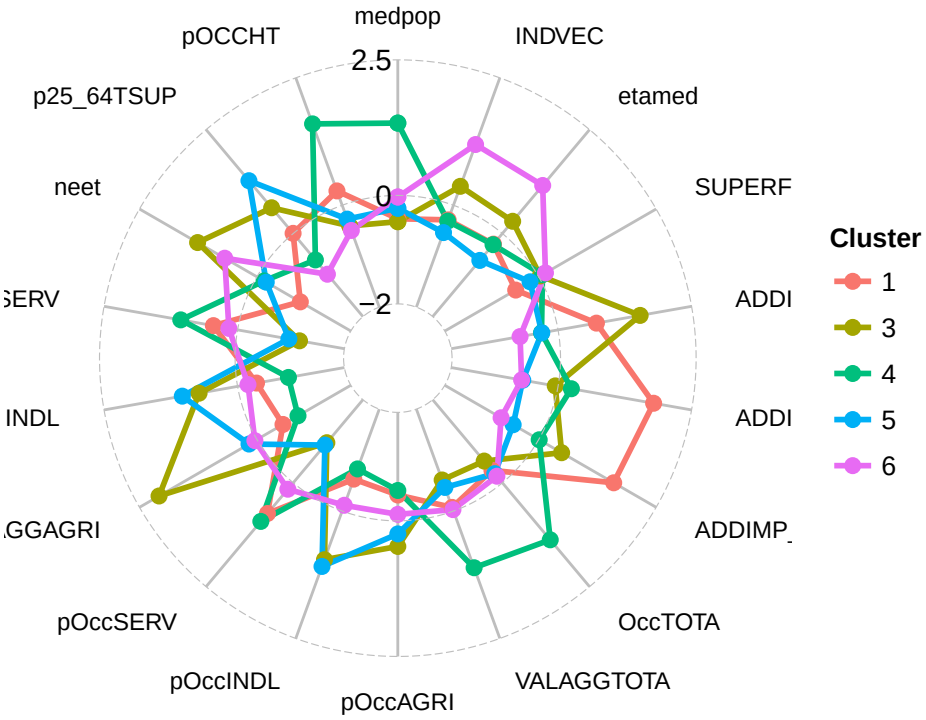
ITALIA

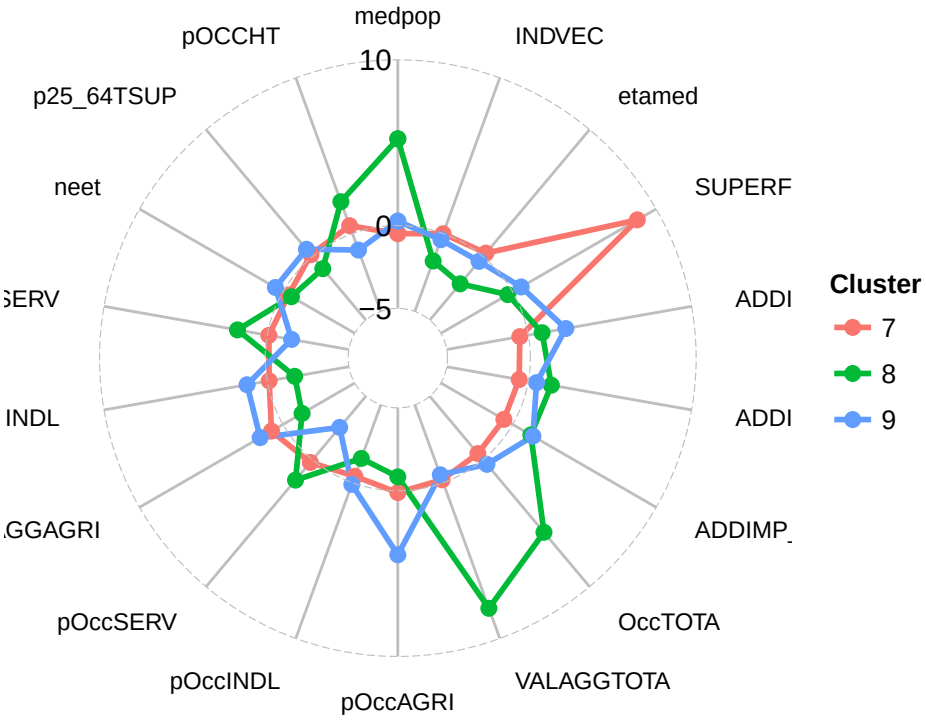
SI02

Zahodna Slovenija (NUTS 2010)

SLOVENIA


medpop
INDVEC
etamed
SUPERF
ADDI
ADDI
ADDIMP_
OccTOTA
VALAGGTOTA
pOccAGRI
pOccINDL
pOccSERV
GGAGRI
INDL
SERV
neet
p25_64TSUP
pOCCHT
1
0
−1


medpop
INDVEC
etamed
SUPERF
ADDI
ADDI
ADDIMP_
OccTOTA
VALAGGTOTA
pOccAGRI
pOccINDL
pOccSERV
.GGAGRI
INDL
SERV
neet
p25_64TSUP
pOCCHT
2.5
0
−2
Cluster
1
3
4
5
6


medpop
INDVEC
etamed
SUPERF
ADDI
ADDI
ADDIMP_
OccTOTA
VALAGGTOTA
pOccAGRI
pOccINDL
pOccSERV
GGAGRI
INDL
SERV
neet
p25_64TSUP
pOCCHT
10
0
−5
Cluster
7
8
9