TDE 10/02/21

Problema 1.

1. Supponiamo che i dati siano normali e li verifichiamo attraverso un test basato su un approssimazione della verifica di normalità di tutte le combinazioni lineari delle componenti. A questo punto basiamo il nostro intervallo di confidenza su un ellisse del tipo:

Dove:

In particolare l’ellisse è centrata in = [211.12500, 20.79167, 27.04167]. Definite come le coppie di autovettori e autovalori di S: le direzioni principali degli assi sono dati dagli autovettori di S: = [0.98529663 0.05642585 0.1612658

0.09957569 -0.95665950 -0.2736554

0.13883524 0.28568993 -0.9482120 ]

e le lunghezze dei semi assi da:

i = 1 15.693925

i = 2 5.842251

i = 3 3.965390

1. Costruiamo 4 intervalli simultanei per combinazioni lineari della media, gli intervalli di confidenza T2 sono del tipo:

In particolare avremo per il primo intervallo

195.6451, 211.125, 226.6049]

Per il secondo

[14.88768, 20.79167, 26.69566]

Per il terzo

[22.38644, 27.04167, 31.69689]

Per il quarto

[ 238.9733, 258.9583, 278.9434]

1. Dobbiamo testare se

H0: vs H1:

Per fare il test e trovarne il p-value fruttiamo:

Dove

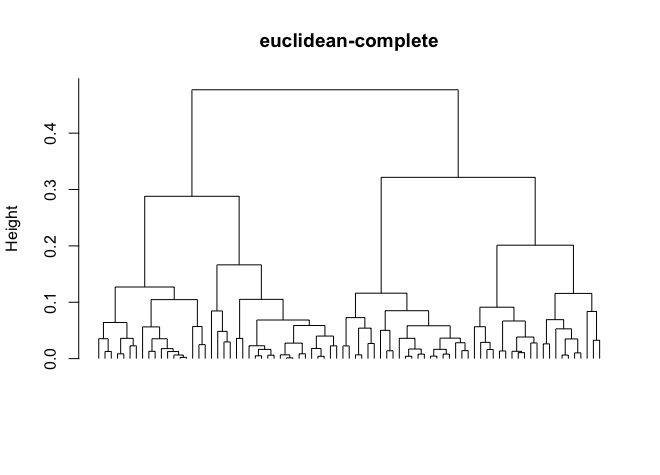
La nostra regione di rifiuto sarà quindi:

Rifiutiamo l’ipotesi al 95% and il p-value è poco più grande del 0.5% in particolare 0.005304853

Problema 2.

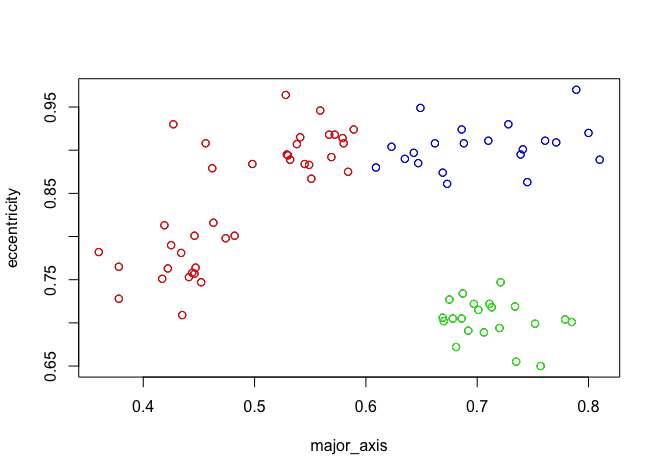
1. Vogliamo utilizzare un algoritmo di clustering per raggruppare i nostri dati basandoci sulla loro distanza. In partcioare utiizziamo un algoritmo di custering aglomeritavo dove le unità inizialemnte e poi i clustering più vicini vengono aggregati.

Utilizzando in particolare come distanza tra le unità una distanza gaussiana e tra i clustering un complete linkage trivaimo il seguente dendogramma

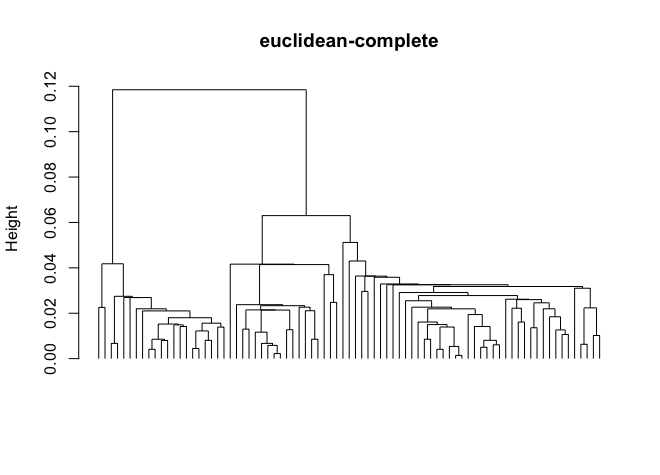


Tagliano in particolare questo dendogramma quando siamo in presenza di tre cluster troviamo dei gruppi di numerisità 39, 21 e 21 con media

Possiamo rappresentare quindi i tre cluster

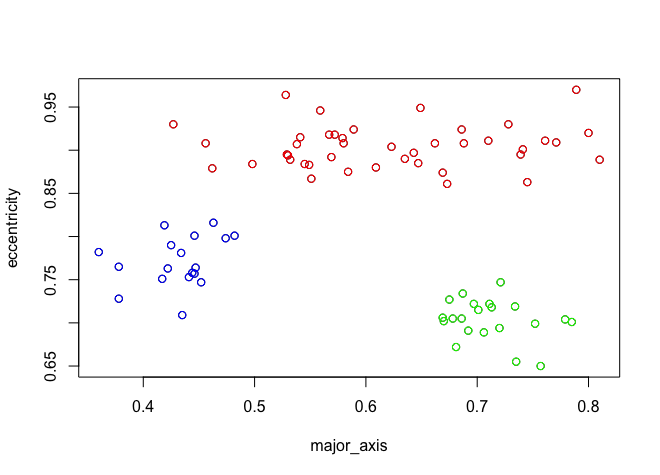


1. Valutimao il clustering in base al cophenetic coefficient che guarda la correlazione tra la distanza in cui deu. Unità sono state aggregate nello stesso clustering e quella originale : la correlazione è pari a 0.7823651. Notiamo che il custering rosso sembra unificare due clustering diversi. Essendo le strutture dei clustering separate e non ellisoidali un buon metodo di linkage potremme essere il sigle linkage. Ripetendo l’analisi di prima con il nuovo linkage abbiamo:



La numerosità e ora di 42,18 e 21 e le medie diventano

Facendo infine un grafico dei cluster otteniamo



E’ da notare però che il cophenetic coeffiecint si abbassa in questo caso.

1. Vogliamo or andare a fare inferenza sulle tre diverse popolazioni. Dobbiamo suppore che esse siamo normali e lo possiamo controllare attraverso un test approssimato. Poi il modello di intervallo di confidenza per le componenti di ognuno dei tre gruppi sarà:

Dove per la prima componente e per la seconda. k in questo caso è 4.

Per il primo cluster troviamo

major\_axis [ 0.5848970, 0.6205476, 0.6561983]

eccentricity [0.8951629, 0.9041190, 0.9130752]

Per il secondo

major\_axis [ 0.4124099, 0.4312778, 0.4501457]

eccentricity [0.7540783, 0.7709444, 0.7878106]

Per il terzo

major\_axis [0.6936225, 0.7118571, 0.7300918]

eccentricity [0.6910732, 0.7036667, 0.7162601]

Problema 3.

1. Il modello volendo formulare una regressione lineare è

average downslope displacement rate= **β0 + β1\*** annual precipitations + **β2\***hardness of subsurface rocks

**+ β3\*** quantity of coarse debris **+ β4\***the quantity of fine debris +

Possiamo poi trovare i coefficienti grazie al Least Square methods infatti definita Z= [1, annual precipitations, hardness of subsurface rocks , quantity of coarse debris, the quantity of fine debris] inteso come la matrice disegno di tutte le misurazioni delle quantità possiamo vedere che **β** = (Z’Z)-1Z’y Dove con y si intendono le misurazioni di average downslope displacement rate sul training set.

The coefficient for the specific problem are

**β0 β1** **β2** **β3** **β4**

20.514647296 0.006115268 0.011423423 0.385216816 0.195448455

The model assumptions are verified throw a shapiro test on the residuals on their visualization to assert independence and zero mean.

1. We base in general on the test:

H0: **βi= 0** vs H1: **βi != 0**

As a pivotal quantity we use t0 = βI / S\*sqrt(diagi(Z’Z)-1 ) ~ t(n-r+1)

We will reject H0 if t0 > t1-alpha/2(n-r+1)

In particolare non rifiutiamo H0 solo per hardness of subsurface rocks

Possiamo quindi proporre un modello ridotto del tipo

average downslope displacement rate= **β0 + β1\*** annual precipitations  **+ β2\*** quantity of coarse debris **+ β3\***the quantity of fine debris +

**β0 β1** **β2** **β3**

20.560257374 0.006131644 0.386020255 0.194141815

c) Dobbiamo testare ora una combinazione delle beta in particolare vogliamo eseguire il test

H0: **β2-2\*β3= 0** vs H1: **β2-2\*β3!= 0**

Definiamo a’= (0,0,1,-2) Possiamo sfruttare come quantità pivotale

As a pivotal quantity we use t0 = a’\*β / S\*sqrt(a’ (Z’Z)-1 a) ~ t(n-r+1)

We will reject H0 if t0 > t1-alpha/2(n-r+1)

Il p-value del test è di 0.9836 quindi non possiamo rifiutare l’ipotesi nulla di conseguenza un modello ridotto può essere del tipo:

average downslope displacement rate= **β0 + β1\*** annual precipitations  **+ β2\*(** the quantity of fine debris + 2 quantity of coarse debris) +

I nuovi parametri stimati saranno:

**β0 β1** **β2**

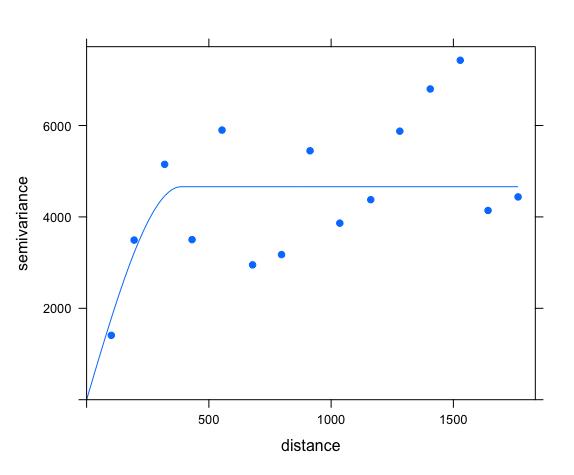
20.562007919 0.006131581 0.193288986

d) Dobbiamo adesso passare alla parte di predizione sulla media abbiamo in particolare che la stima puntuale sulla media di una vuova osservazione dove z0 = (700, 5, 10 , 8) E’ data semplimente da z0\***β**= 30.26621

L’intervallo invece è valutato come z0\***β** +- S sqrt(z0\*(Z’Z)^(-1)z0’\* (r+1)F(1-apha)(r+1,n-r+1))= [ 30.04873 30.48368]

Problema 4.

1. Cominciamo suppendo il nostro random field abbia la proprietà di second order stationarity. Per prima cosa dobbiamo stimare e capire come è fatta la struttura di varainza dei nostri dati. In particolare partiamo dai singoli dati e da essi stimiamo attraverso un binned variogram il variogramma dei dati e dato il binned variogram ci focalizziamo sul fittarlo con un modello valido. Una volta fittato il modello possiamo trovare la media desiderata predicendo uno dei punto del nostro dataset in media trovaimo in particolare che a0 = 263.5469 per fare questo in particolare usiamo Ordinary Kriging che sfrutta i Generalized Least square per trovare i parametri. Il modello per δ(si) è un modello a media nulla e varianza ∑ il cui variogramma è modellizato come



Possiamo notare che le assunzioni sulla stazionarietà e sull isotropia non sembrano adatte a questo modelli infatti vediamo che il variogramma non tende a una costate ma sale e che nelle diverse direzioni abbiamo variogrammi diversi dovremmo quindi cambiare tipo di modello per la media

1. Assumiamo quindi per la media un modello del tipo y(si) = a0,g + a1,g · d(si) + δ(si), abbiamo in particolare che come prima possiamo usufruire di un Universal Krigging per valutare le stime dei parametri di regressione della media utilizzando i regressori in modo corretto. Allora:

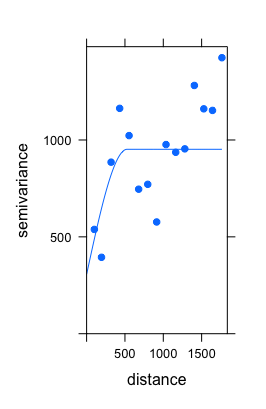
a01 = 447.1602

a02 = 667.8809

a11 = -0.06926014

a12 = -447.1696

Il nuovo modello per i residui invece + basato su un variogramma del tipo



Anche in questo caso le assunzioni di stazionarietà sembrano essere violate come l’isotropia ma il modello sembra descrivere meglio il fenomeno.

1. Sceglierei il secondo modello perché ci dà la possibilità di generare due modelli diversi in base al periodo dell’anno, cosa che possiamo vedere essere presente anche del primo variogramma entrambi non sono particolarmente conformi alle assunzioni
2. Supponiamo di sfruttare il modello scelto ovvero il secondo e quindi di sfruttare un approccio di predizione dato da Universal Kriging, le assunzioni sui residui sono sempre quelli di stazionarietà e di isotropia. In particolare andiamo a fornire un intervallo basato sulla media da che supponiamo gli anni siamo indipendenti. Il prezzo al giorno suggerito è 429.9638.