Laboratorio 4

Nicola Agostini, Roberto Cedolin, Lisa Parma

22 Maggio 2019

Domanda 1

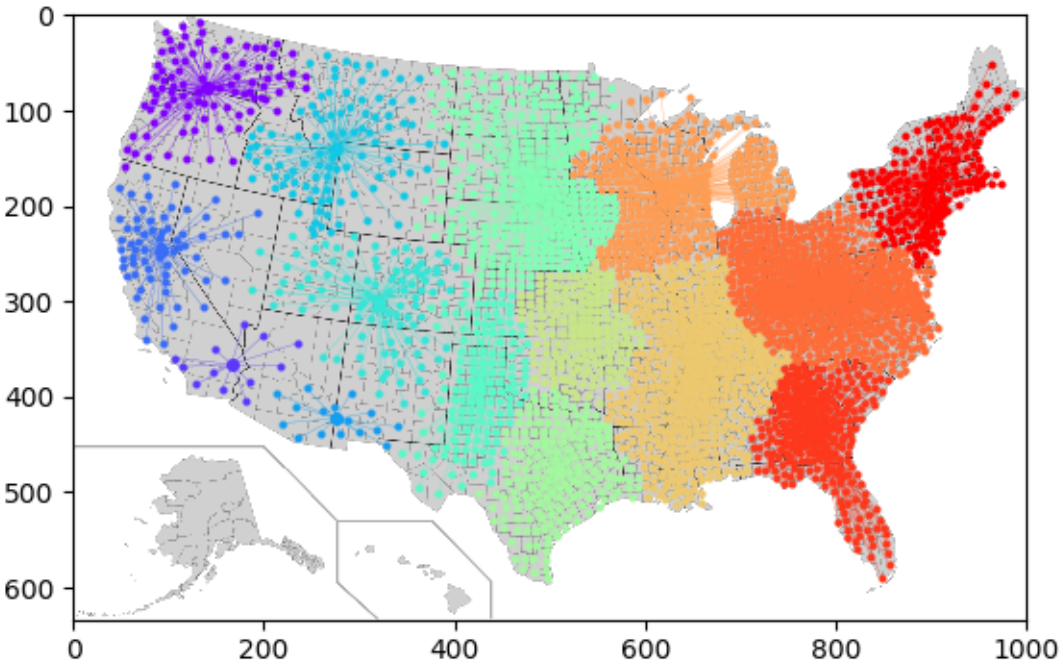

Figura 1: Clustering gerarchico con 15 cluster

Domanda 2

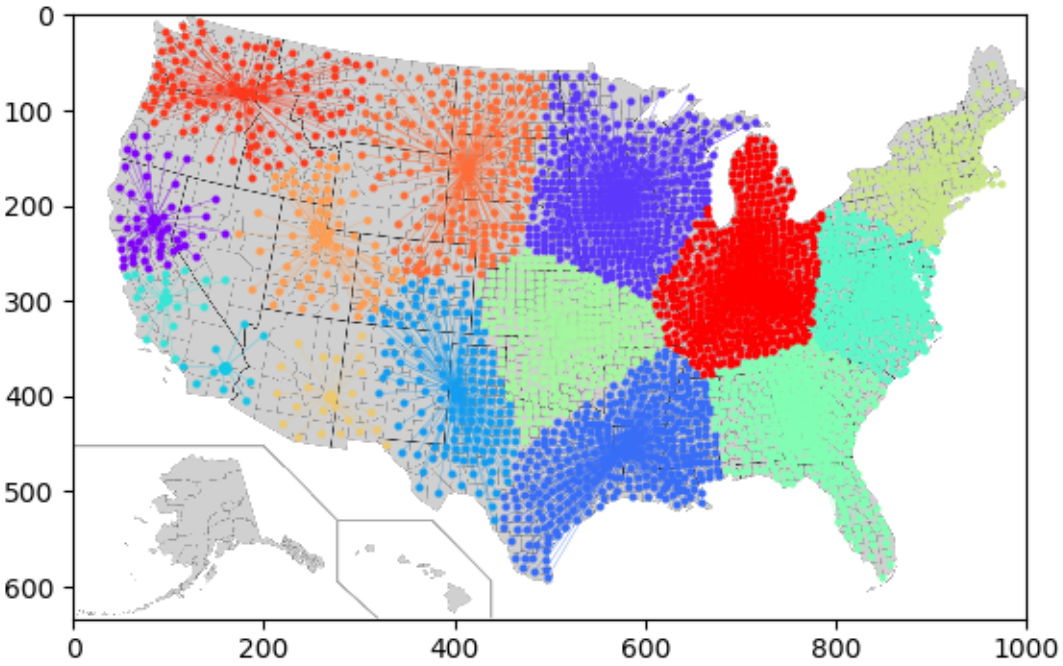

Figura 2: Clustering K-Means con 15 cluster e 5 iterazioni

Domanda 3

Quando il numero di cluster che si vuole in output è basso risulta essere più efficiente l'algoritmo K-Means con un numero basso di iterazioni perchè esegue un controllo per ciascun cluster dei punti che sono a distana minima e calcola di conseguenza i cluster di appartenenza. Quando il numero di cluster è basso

K-Means è asintoticamente migliore in termini di complessità temporale rispetto all'algoritmo di clustering gerarchico. Quest'ultimo, infatti, crea prima un cluster per ciascun punto del dataset e poi unisce i cluster più vicini fino ad ottenerne il numero desiderato in output.

Clustering gerarchico complessità: $O(n^2 log(n))$

Clustering K-Means complessità: $O(q * (n + k))$
dove q=numero di iterazioni richieste e
k=numero di cluster richiesti in output

Domanda 4

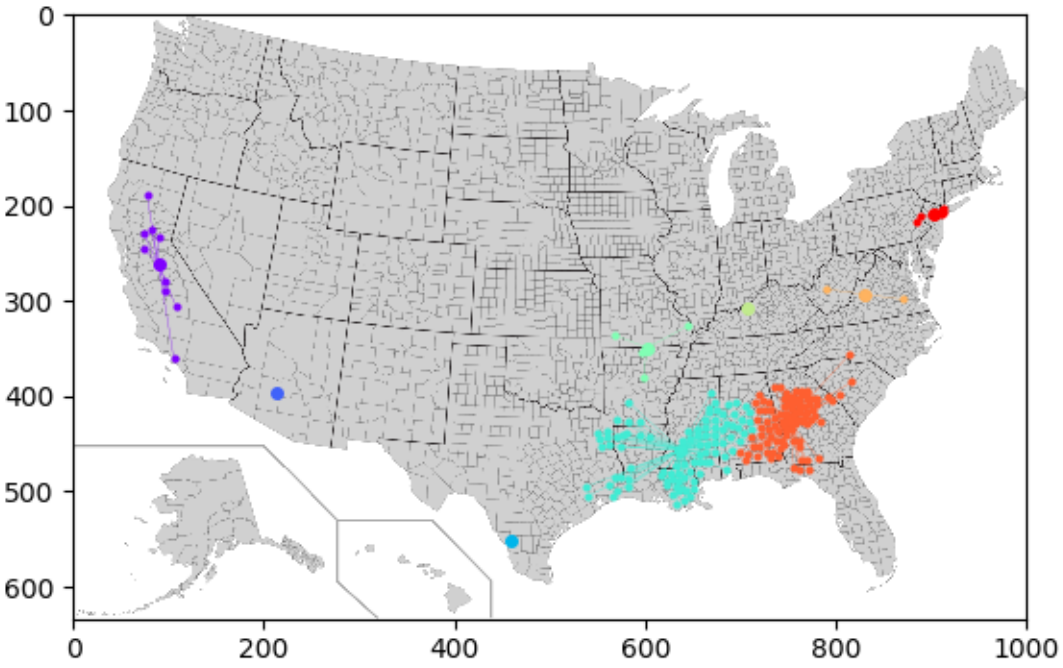

Figura 3: Clustering gerarchico con 212 nodi e 9 cluster

Domanda 5

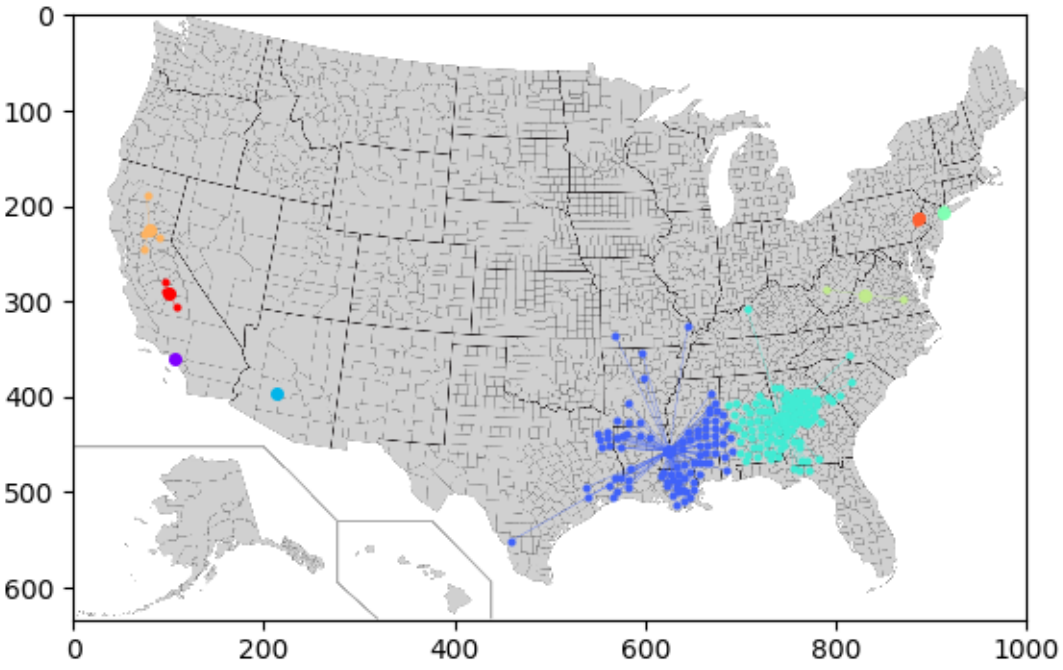

Figura 4: Clustering K-Means con 212 nodi, 9 cluster e 5 iterazioni

Domanda 6

Clustering gerarchico con 212 nodi e 9 cluster:
Distorsione $= 1.968 * 10^{11}$

Clustering K-Means con 212 nodi, 9 cluster e 5 iterazioni:
Distorsione $= 9.538 * 10^{10}$

Domanda 7

L'algoritmo K-Means inizializza i centroidi nelle contee aventi popolazione maggiore: di conseguenza vengono generati vari cluster nella costa occidentale americana ciascuno avente centroide distinto. Nel caso del clustering gerarchico, invece, vengono unite le varie contee in un numero minore di cluster in quanto esso non tiene conto della popolazione ma solo della distanza fra le contee. La distorsione viene calcolata utilizzando la distanza dei punti del cluster pesata rispetto alla popolazione della contea. Per questo motivo, la distorsione ottenuta utilizzando l'algoritmo gerarchico risulta essere più alta rispetto all'utilizzo dell'algoritmo K-Means.

Domanda 8

Il metodo di clustering che richiede minore supervisione umana per generare cluster a bassa distorsione è quello gerarchico in quanto nel clustering K-Means è necessario stabilire un valore ottimale dell'iperparametro q per mantenere una distorsione bassa.

Domanda 9

Di seguito sono indicate le distorsioni calcolate per ciascun metodo di clustering con numero di cluster variabile da 6 a 20.

212 Nodi

Distorsione clustering gerarchico:

6. 355133019395.9377
7. 211984221669.04886
8. 207994590136.3608
9. 196752213374.95956
10. 145885753905.5376
11. 73339588185.56026
12. 72986951812.24315
13. 62925657807.43205
14. 37490382899.7963

15. 30754683330.12072
16. 30522710097.315456
17. 19653705327.8246
18. 18992753329.124424
19. 18908757016.55021
20. 14493513775.288986

Distorsione clustering K-Means:

6. 194204933067.88162
7. 121658767347.90419
8. 120818040272.97252
9. 95382765365.33682
10. 76125162275.26419
11. 69404408805.67369
12. 56845918429.6827
13. 51057083117.0554
14. 45060501518.22067
15. 41440317826.9728
16. 36589222769.01075
17. 36454273396.203926
18. 18494050364.1729
19. 17545382190.64733
20. 16127651929.373045

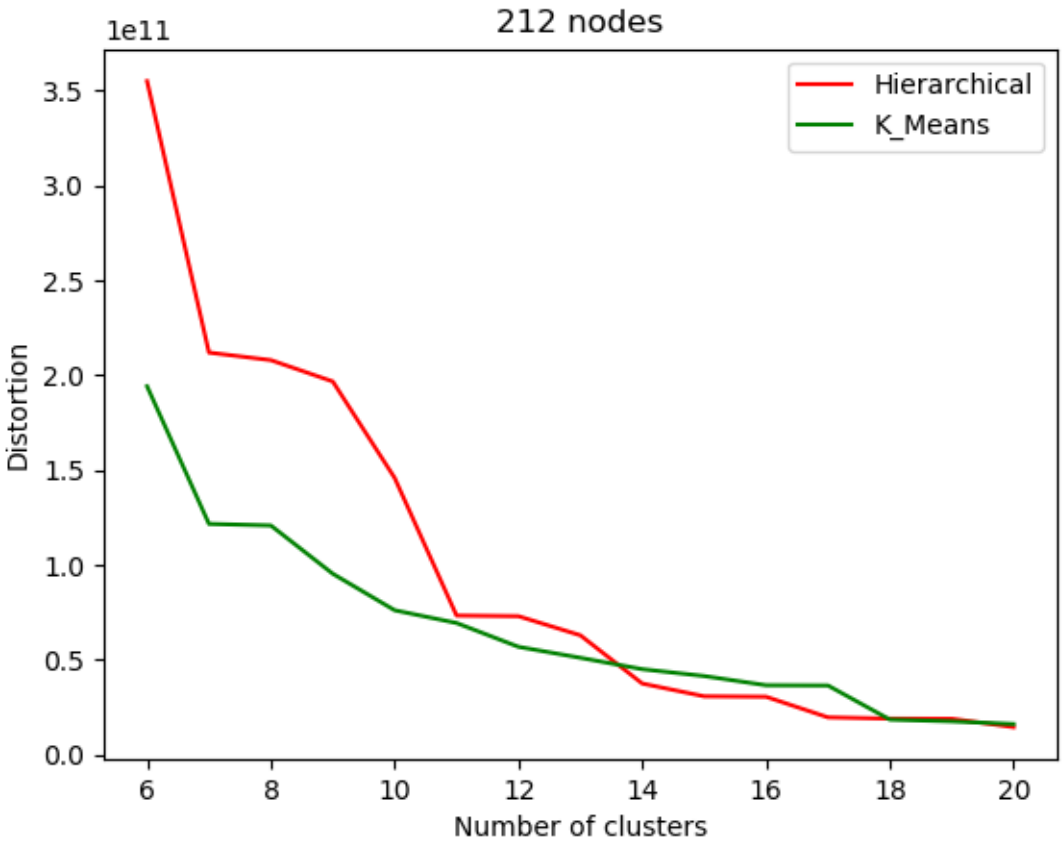

Figura 5: Grafico della distorsione al variare del numero di cluster

562 Nodi

Distorsione clustering gerarchico:

6. 1341235435450.2866
7. 1335053122466.4336
8. 1018233474132.4292
9. 895932791169.6853
10. 475194045058.0631
11. 415736597128.8702

12. 392219684447.10583
13. 306004502344.7203
14. 286056690543.1298
15. 267364824249.12415
16. 226459885742.40872
17. 205523712689.7545
18. 148414234692.82202
19. 139791163888.83295
20. 130939811152.133

Distorsione clustering K-Means:

6. 1488108061889.3071
7. 888829728600.4204
8. 806775313087.8262
9. 667410677315.3434
10. 564496280838.2029
11. 547508227071.0067
12. 414026729326.09894
13. 283500925210.1014
14. 251957008358.28006
15. 438051796942.9473
16. 386286714904.1847
17. 382637712753.42505
18. 305324658619.52454
19. 204602109047.81152
20. 203734980028.14786

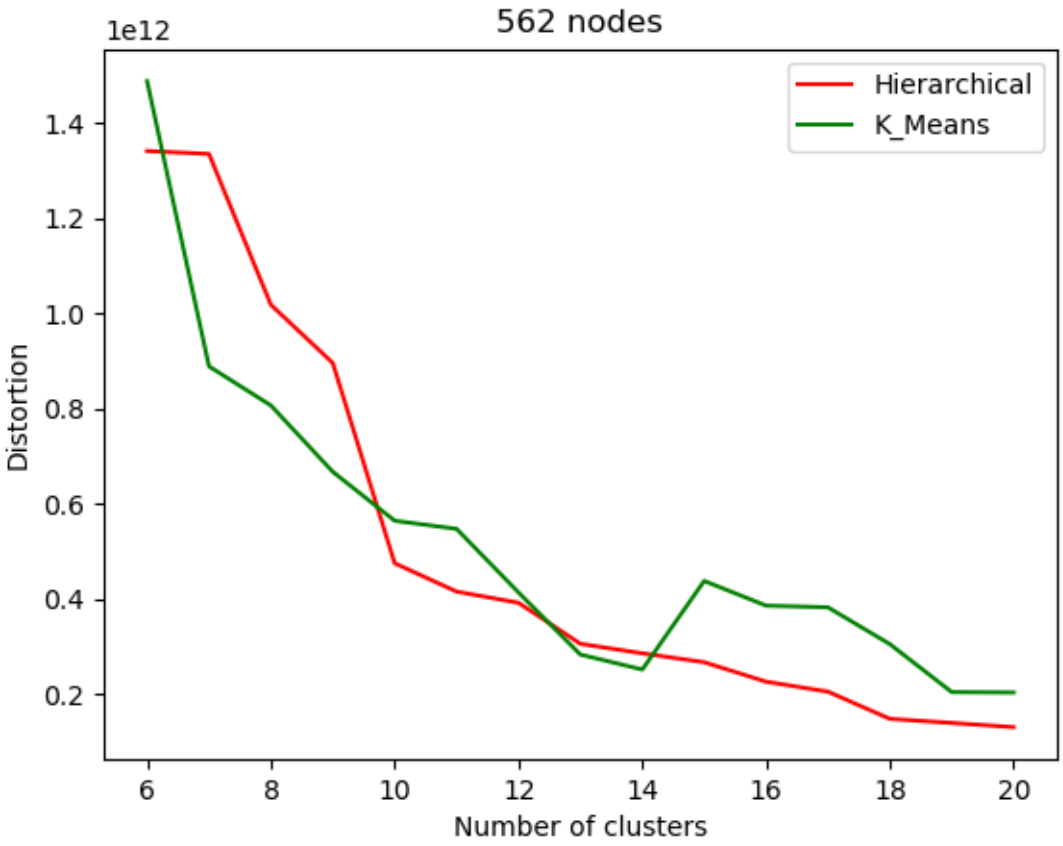

Figura 6: Grafico della distorsione al variare del numero di cluster

1041 Nodi

Distorsione clustering gerarchico:

6. 2497247032903.0283
7. 2222619830374.2134
8. 1375667596348.1187
9. 1140417892328.6836
10. 868707383504.1082
11. 846162895364.3564

12. 840742774899.233
13. 830979648143.1234
14. 619593702887.9216
15. 575365007811.4012
16. 501056283628.2639
17. 467062645124.6728
18. 421625589616.8628
19. 370370804442.0912
20. 366663659842.13776

Distorsione clustering K-Means:

6. 2524928272641.2495
7. 1568258715796.538
8. 1180744256177.8633
9. 1089884917987.2756
10. 846902771250.8722
11. 775488768350.1039
12. 774276048773.2698
13. 772758756794.6791
14. 711957846651.9125
15. 643895092233.2773
16. 624335242151.4619
17. 458568955530.61066
18. 425410848826.7105
19. 420450065218.2312
20. 383420553007.08875

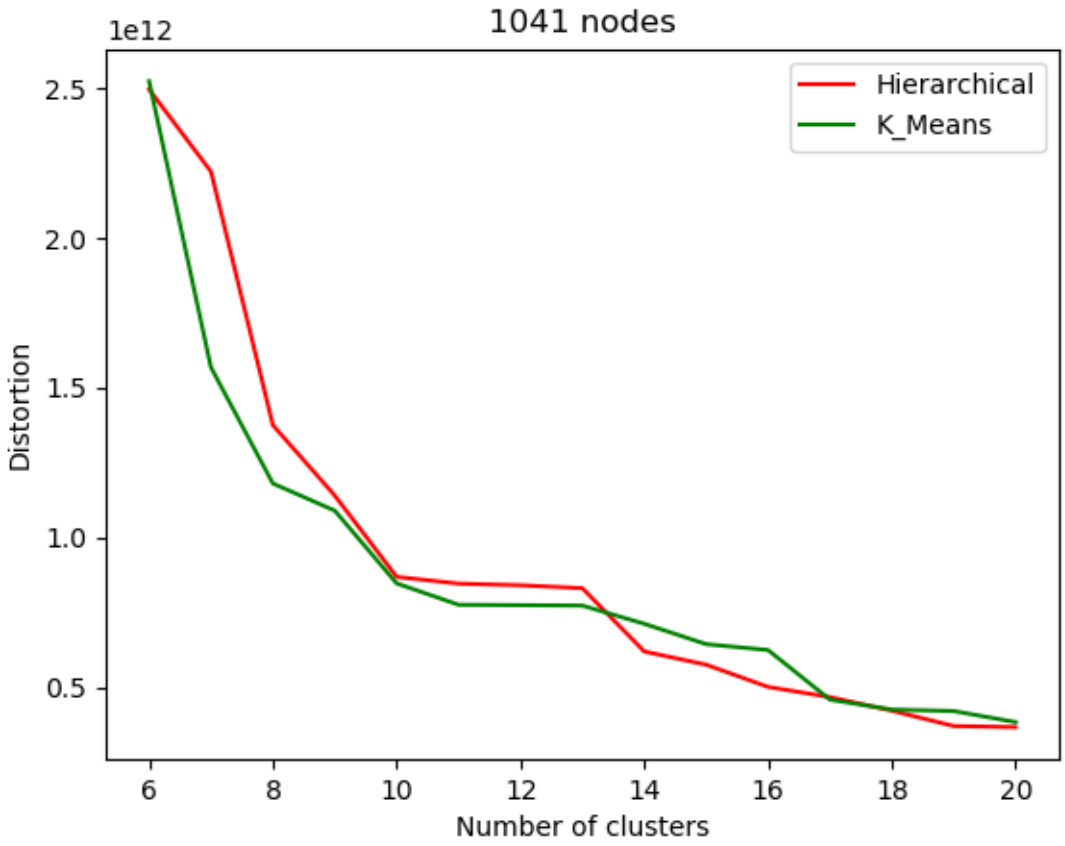

Figura 7: Grafico della distorsione al variare del numero di cluster

Domanda 10

No, nessuno fra gli algoritmi di clustering implementati risulta essere sempre migliore dell'altro in quanto, per ciascun set di dati, per un numero di cluster variabile fra 6 e 10-12 la distorsione risulta essere maggiore nel clustering gerarchico. Per numero di cluster maggiori risulta essere variabile: vi è una generale tendenza del clustering K-Means ad avere distorsione maggiore.