clasificarea nesupervizata a ratei crimelor violente in statele unite ale americii

Tema 4. Analiza de cluster (Metode de recunoaștere nesupervizată a formelor)

Proiect realizat de:

Andreea-Maria Ionescu

Grupa 1069, Seria B

Cuprins

[**1. Despre Analiza de Clusterizare 2**](#_Toc442919470)

[**2. Semnificația Modelului 3**](#_Toc442919471)

[**3. Matricea distanțelor 5**](#_Toc442919472)

[**4. Evoluția Clasificării (Tabelul Ierarhie) 5**](#_Toc442919473)

[**5. Graficul Dendrogramă 8**](#_Toc442919474)

[**6. Bibliografie 11**](#_Toc442919475)

# Despre Analiza de Clusterizare

Analiza de clusterizare, cunoscută și sub denumirea de clasificare nesupervizată reprezintă un ansamblu de algoritmi ce sunt utilizați pentru a determina grupări naturale ale datelor sau pentru a furniza o împărțire convenabilă a datelor în grupuri.

Diferența dintre analiza discriminantă (clasificare supervizată) și analiza de cluster (clasificare nesupervizată) este aceea că nu cunoaștem în prealabil informații cu privire la grupele în care se va face clasificarea în cazul clusterizării. Așadar, analiza de cluster este o metodă exploratorie întrucât numărul grupelor este subordonat scopului analizei. Practic, atunci cand se realizează împărțirea colectivităților în grupe, nu se ține cont de o clasificare subiectivă, avută din start, ci se iau în considerare doar variabilele cauzale. Odată ce se realizează împărțirea folosind algoritmii de clusterizare, se poate observa specificitatea grupelor.

Datele care sunt analizate reprezintă valorile relațiilor dintre indivizii și variabilele aflate în studiu luate două câte două, adică distanțele sau disimiliaritățile. Aceste valori sunt înregistrate sub forma Matricei de Distanță. Astfel, valorile supuse analizezi reprezintă deja rezultatul unor calcule ce implică masurarea distanței dintre instanțe.

În urma prelucrării, grupele formate prin clasificare sunt formate din indivizi asemănători între ei, cu o disimilaritate/distanță foarte mică. Având în vedere că în analiza de cluster se urmărește formarea de grupe omogene, aceasta constituie o metodă de sinteză informațională, aspect ce amintește de analiza în componente principale, atâta tot că se referă în principal la instanțe, ci nu la variabile. Uneori, în diverse analize, o grupă omogenă poate fi reprezentată de un singur individ: centrul grupei.

Printre motivele principale pentru care o grupare a datelor este necesară amintim:

* Identificarea trăsăturilor fundamentale ale datelor (unde se studiază relațiile semnificative existente între date)
* Obținerea unor reprezentări avantajoase în efectuarea analizelor
* Stocarea și regăsirea rapidă a informației (se îmbunătățește viteza de acces prin intermediul unor strategii de rutare pentru informația stocată. Practic, eficiența unei grupe se raportează la eficiența timpului în care anumite date sunt regăsite)

# Semnificația Modelului

În cele ce urmează, se va prezenta setul de date ales, cât și semnificația fiecărei variabile în parte. Datele au fost preluate de la următoarele surse:

* World Almanac and Book of facts 1975 (Rata Criminalității).
* Statistical Abstracts of the United States 1975 (Rate Urbanistice).

Setul de date conține statistici care fac referire la arestarile efectuate pe un eșantion de 100,000 de rezidenți. Motivele aresturilor sunt asalt, crimă, viol în fiecare dintre cele 50 de state din Statele Unite ale Americii în anul 1973. De asemenea, o altă variabilă oferită este procentul de populație care locuiește în zona urbană.

Descrierea semnificației variabilelor care se regăsesc în tabelul de date este:

| **Denumirea variabilei** | **Descriere** |
| --- | --- |
| *State* | Denumirea fiecărui stat din SUA |
| *Murder* | Numărul de arestări efectuat în urma crimelor (la 100,000 locuitori) |
| *Assault* | Numărul de arestări efectuat în urma asalturilor (la 100,000 locuitori) |
| *UrbanPop* | Procentul de populație care locuiește în zona urbană |
| *Rape* | Numărul de arestări efectuat în urma unui viol (la 100,000 locuitori) |

# Matricea distanțelor

**Matricea de disimilaritate** este matricea distanţelor între cazuri (variabile). Este o matrice simetrică. Aceasta este calculată și reprezentată în fișierul *Hclust\_USArrests\_Dist.csv.* Așadar, putem obesrva că distanța dintre variabila 1 și 2, care în cazul de față sunt Alabama și Alaska este aproximativ 33.45.

După cum se poate observa și în fișierul .csv rezultat, distanțele dintre două variabile luate două câte două diferă între ele. Astfel, se pot indentifica atât distanțe mai mici, cât și distanțe mai mari, adică valori extreme. Acest lucru poate fi justificat astfel:

* Cea mai mică distanță este între variabila 15 (statul Iowa) și variabila 29 (statul New Hampshire) în valoare de aproximativ 1.637
* Cea mai mare distanță este între variabila 9 (statul Florida) și variabila 33 (statul North Carolina) în valoare de aproximativ 40.57

# Evoluția Clasificării (Tabelul Ierarhie)

Pentru a putea observa evoluția clasificării, se folosesc algoritmii ierarhici de clasificare. Pașii necesari pentru construirea ierarhiei sunt:

**Date de intrare:** un set de n(n-1)/2 distanțe (în cazul de față 50(50-1)/2 = 1225 distanțe)

**Pasul 1:** se selectează obiectele care vor fi agregate: *i* și *k* (adică statele în cazul de față)

**Pasul 2:** se adună termenii *i* și *k* și se înlocuiesc cu un nou obiect *,* actualizându-se și distanțele pentru toți termenii  și

**Pasul 3:** atâta timp cât rămân cel puțin doi termeni, se revine la pasul 1

În cazul setului de date prezentat în cadrul acestui proiect, ierarhia cuprinde 50 de partiții: prima partiție cuprinde grupele formate din câte un singur obiect, iar ultima partiție conține o singură grupă formată din toate cele 50 de obiecte, adică toate cele 50 de state. Acestea se pot observa și în fișierul *Hclust\_USArrests\_Hierarchy.csv.*

|  |  |  |
| --- | --- | --- |
| Clustere | Distanta | Structura cluster |
| Cluster 1 | 1.637071 | {15,29} |
| Cluster 2 | 2.930301 | {14,16} |
| Cluster 3 | 4.425683 | {17,26} |
| Cluster 4 | 4.653314 | {13,28} |
| Cluster 5 | 6.781347 | {32,{13,28}} |
| Cluster 6 | 6.932051 | {6,43} |
| Cluster 7 | 7.013321 | {46,50} |
| Cluster 8 | 7.466369 | {35,44} |
| Cluster 9 | 7.990828 | {19,41} |
| Cluster 10 | 8.164966 | {7,38} |
| Cluster 11 | 8.687923 | {37,{46,50}} |
| Cluster 12 | 9.505086 | {36,47} |
| Cluster 13 | 9.606248 | {2,24} |
| Cluster 14 | 11.79604 | {4,42} |
| Cluster 15 | 12.1595 | {18,22} |
| Cluster 16 | 12.4499 | {49,{15,29}} |
| Cluster 17 | 12.90994 | {21,30} |
| Cluster 18 | 14.37405 | {34,45} |
| Cluster 19 | 14.64969 | {48,{19,41}} |
| Cluster 20 | 14.97598 | {{14,16},{7,38}} |
| Cluster 21 | 15.29314 | {12,{17,26}} |
| Cluster 22 | 15.99542 | {3,31} |
| Cluster 23 | 17.6639 | {20,{3,31}} |
| Cluster 24 | 18.37716 | {1,8} |
| Cluster 25 | 19.66384 | {27,{{14,16},{7,38}}} |
| Cluster 26 | 21.17735 | {25,39} |
| Cluster 27 | 21.93961 | {23,{49,{15,29}}} |
| Cluster 28 | 23.63303 | {40,{2,24}} |
| Cluster 29 | 24.02332 | {{37,{46,50}},{36,47}} |
| Cluster 30 | 24.31707 | {{32,{13,28}},{18,22}} |
| Cluster 31 | 26.38409 | {10,{6,43}} |
| Cluster 32 | 29.42131 | {{35,44},{27,{{14,16},{7,38}}}} |
| Cluster 33 | 33.57301 | {{21,30},{{37,{46,50}},{36,47}}} |
| Cluster 34 | 35.53871 | {{12,{17,26}},{{35,44},{27,{{14,16},{7,38}}}}} |
| Cluster 35 | 35.99556 | {11,{23,{49,{15,29}}}} |
| Cluster 36 | 37.17006 | {{4,42},{10,{6,43}}} |
| Cluster 37 | 38.51251 | {{1,8},{{32,{13,28}},{18,22}}} |
| Cluster 38 | 38.69539 | {{25,39},{{21,30},{{37,{46,50}},{36,47}}}} |
| Cluster 39 | 39.28172 | {5,{20,{3,31}}} |
| Cluster 40 | 40.5752 | {9,33} |
| Cluster 41 | 47.48333 | {{34,45},{48,{19,41}}} |
| Cluster 42 | 55.72851 | {{40,{2,24}},{{1,8},{{32,{13,28}},{18,22}}}} |
| Cluster 43 | 64.92211 | {{11,{23,{49,{15,29}}}},{{34,45},{48,{19,41}}}} |
| Cluster 44 | 74.66021 | {{5,{20,{3,31}}},{{40,{2,24}},{{1,8},{{32,{13,28}},{18,22}}}}} |
| Cluster 45 | 79.20067 | {{{4,42},{10,{6,43}}},{{25,39},{{21,30},{{37,{46,50}},{36,47}}}}} |
| Cluster 46 | 99.92664 | {{{12,{17,26}},{{35,44},{27,{{14,16},{7,38}}}}},{{11,{23,{49,{15,29}}}},{{34,45},{48,{19,41}}}}} |
| Cluster 47 | 118.7738 | {{9,33},{{5,{20,{3,31}}},{{40,{2,24}},{{1,8},{{32,{13,28}},{18,22}}}}}} |
| Cluster 48 | 193.5202 | {{{{4,42},{10,{6,43}}},{{25,39},{{21,30},{{37,{46,50}},{36,47}}}}},{{{12,{17,26}},{{35,44},  {27,{{14,16},{7,38}}}}},{{11,{23,{49,{15,29}}}},{{34,45},{48,{19,41}}}}}} |
| Cluster 49 | 338.9873 | {{{9,33},{{5,{20,{3,31}}},{{40,{2,24}},{{1,8},{{32,{13,28}},{18,22}}}}}},{{{{4,42},{10,{6,43}}},  {{25,39},{{21,30},{{37,{46,50}},{36,47}}}}},{{{12,{17,26}},{{35,44},{27,{{14,16},{7,38}}}}},  {{11,{23,{49,{15,29}}}},{{34,45},{48,{19,41}}}}}}} |

Partițiile pot fi evidențiate și în graficul dendrogramă prin secționări paralele cu axa distanțelor. Numărul de clustere dintr-o partiție este egal cu numărul de brațe intersectate de secțiune.

Un algoritm general de clusterizare poate fi scris în pseudocod în felul următor:

**Procedure Ierarhie(D,n)**

**List[] P**

**for i=1,n**

**P[i] = {i}**

**endfor**

**for i=1,n-1**

**call Select(D;k,j)**

**P[k] = P[k] + P[j]**

**P[j] = {}**

**call Actualizare(D,k,j)**

**call Afisare(P)**

**endfor**

**return**

**end**

Unde:

**D** – matricea de distanțe

**P** – vector de liste, utilizat pentru memorarea unei partiții, adică memorarea unei liste pentru fiecare grupă. În această listă apar valorile index ale instanțelor din grupă

**Procedura Select()** – individualizează fiecare implementare și implementează principiul de grupare

**Procedura Actualizare()** – la fel ca procedura anterioara, individualizează fiecare implementare, însă aceasta are rolul de a recalcula distanțele dintre clusteri după joncțiune

Algoritmul se aplică în toate cazurile, indiferent de metoda de grupare ierarhică. Pentru realizarea cluster-urilor din fișierul *Hclust\_USArrests\_Hierarchy.csv,* s-a aplicat metoda *hclust* din pachetul DiscriMiner. Această metoda grupeaza clusterele și realizează joncțiunile prin **legătură completă** (complete linkage). Astfel, selecția clusterelor se face prin distanță minimă. Actulizarea distanțelor după joncționare se face după relația:

= Maxim

Distanța dinte două clustere este dată de cei mai îndepărtați doi indivizi.

# Graficul Dendrogramă

Outputul grafic al agoritmilor ierarhici este **Graficul Dendrogramă.**

**Axa OX** reprezintă axa variabilelor, în cazul de față fiind axa unde se găsesc toate statele analizate.

**Axa OY** – Height, reprezintă axa cu distanțele dintre variabile (determinată în urma aplicării algoritmului ierarhic).

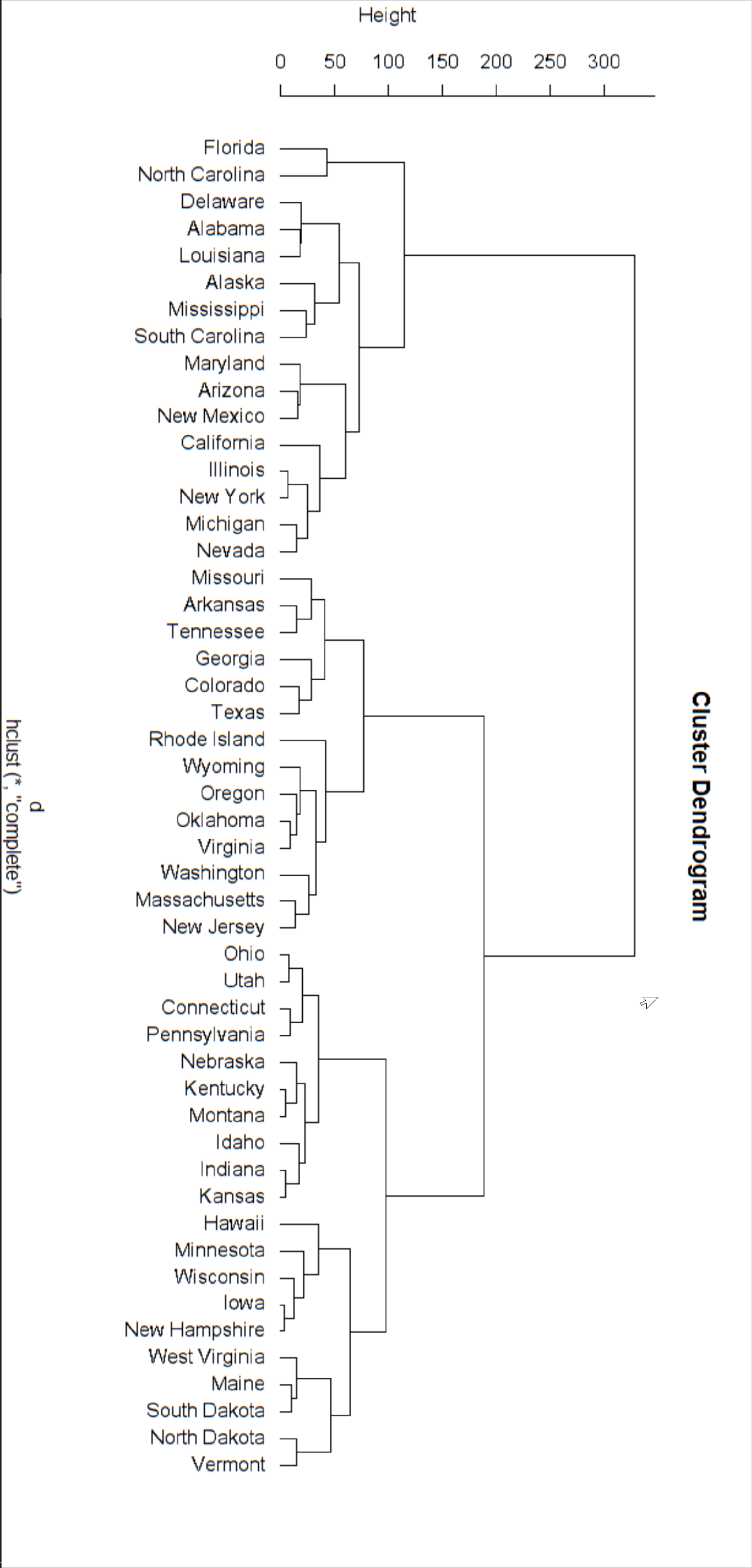
Așadar, graficul evidențiează distanțele de agregare din ierarhie.

**Interpretarea graficului**

Statele joncționează între ele și se formează clustere din ce în ce mai mari. Citind de la stânga la dreapta, se poate observa din imaginea de mai jos a graficului faptul că ies în evidență 3 clustere principale care cuprind cea mai mare parte din state în urma joncțiunilor. Acestea au următoarea componență:

1. **Clusterul principal 1**
   * Este format din statele Florida, North Carolina, Delaware, Alabama, Louisiana, Alaska, Mississippi, South Carolina, Maryland, Arizona, New Mexico, California, Illionis, New York, Michigan, Nevada
   * Se găsește la o distanță de aproximativ 118.77 față de celalte două clustere principale
   * Apar cu preponderență state care se regăsesc ca zonă geografică în partea de Sud-Est a Statelor Unite ale Americii
2. **Clusterul principal 2**
   * Este format din statele Missouri, Arkansas, Tennessee, Georgia, Colorado, Texas, Rhode Island, Wyoming, Oregano, Oklahoma, Virginia, Washington, Massachusetts, New Jersey
   * Se găsește la o distanță de aproximativ 79.2 față de celelalte două clustere principale
   * Apar cu preponderență state care se regăsesc ca zonă geografică în partea Est-Centrală a Statelor Unite ale Americii
3. **Clusterul principal 3**
   * Este format la rândul lui din două clustere mai mici, unul având statele Ohio, Utah, Connecticut, Pennsylvania, Nebraska, Kentuchy, Montana, Idaho, Indiana, Kansas, iar celălalt statele Hawaii, Minnesota, Wisconsin, Iowa, New Hampshire, West Virginia, Maine, South Dakota, North Dakota, Vermot care au joncționat pentru a forma cluterul principal 3
   * Se găsește la o distanță de aproximativ 99.92 față de celelalte două clustere principale
   * Apar cu preponderență state care se regăsesc ca zonă geografică în partea de Nord-Vest a Statelor Unite ale Americii

În final, clusterele principale 2 și 3 ajung să joncționeze între ele la o distanță de aproximativ 193.52, iar acestea două să ajungă în final să se unească cu cel dintâi cluster principal, acest lucru realizându-se la o distanță de aproximativ 339.



# Bibliografie

<http://ip.ase.ro>

<https://vincentarelbundock.github.io/Rdatasets/doc/datasets/USArrests.html>

<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

<http://www.r-tutor.com/gpu-computing/clustering/hierarchical-cluster-analysis>

<https://rpubs.com/gaston/dendrograms>

<https://www.youtube.com/watch?v=48cZ2cMBpio>

<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html>