Killian Blain 2022-12-14 Analyse en composantes principales 1. d <- read.table("Eaux.data", header = TRUE)</pre> d Pays Bicarbonates Sulfates Chlorures Calcium Magnesium ## Nom ## 1 Aix-les-Bains F-73 341 27 3 84 23 ## 2 Arvie F-63 2195 31 387 170 92 F-42 85 ## 3 Badoit 1300 40 40 190 F-68 ## 4 Beckerish 263 23 9 91 5 ## 5 Cayranne F-84 287 24 Chambon F-45 96 ## 6 298 9 23 6 ## 7 Contrex F-88 403 1187 9 486 84 Cristal-Roc F-72 200 70 ## 8 15 Saint-Cyr F-45 250 71 ## 9 5 20 6 ## 10 Evian F-74 357 10 4 78 24 Ferita 73 18 ## 11 Italie 14 18 ## 12 Hepar F-88 403 1479 10 555 110 ## 13 Saint-Hyppolite F-37 256 6 23 86 3 Laurier ## 14 F-77 186 10 16 64 ## 15 0geu F-64 183 48 11 16 44 35 ## 16 **Ondine** F-95 398 218 15 157 Plancoet F-22 17 ## 17 207 38 54 ## 18 Perrier F-30 348 140 51 31 4 ## 19 Ribes Espagne 168 24 8 55 253 11 ## 20 La_Salvetat F-34 820 25 Sanpellegrino 430 174 51 ## 21 Italie 245 52 ## 22 Spa Belgique 110 65 5 4 1 ## 23 F-74 332 8 103 16 Thonon 14 ## 24 Valvert F-08 204 18 68 ## 25 Veri Espagne 196 18 6 58 6 ## 26 Verniere F-54 1170 158 18 190 72 ## 27 Vichy_Celestin F-03 2989 138 235 103 10 Vichy_St-Yorre ## 28 F-03 4368 174 322 90 11 ## 29 Viladreau 59 7 6 16 2 Espagne Vittel 15 36 ## 30 F-88 402 306 202 Volvic F-63 65 6 8 10 ## 31 6 Sodium ## ## 1 ## 2 650 ## 3 150 3 ## 4 ## 5 23 ## 6 11 ## 7 9 ## 8 4 ## 9 11 5 ## 10 ## 11 13 ## 12 14 18 ## 13 ## 14 ## 15 ## 16 8 ## 17 ## 18 ## 19 9 ## 20 7 ## 21 33 ## 22 3 ## 23 ## 24 2 ## 25 13 ## 26 ## 27 1172 ## 28 1708 ## 29 ## 30 3 ## 31 boxplot(d\$Bicarbonates) 0 0 0 On remarque une concentration autour d'environ 300 mais il y a également quelques valeurs extrêmes allant jusqu'à plus de 4000. boxplot(d\$Sulfates) 0 0 500 0 On observe encore une fois une concentration de valeurs autour de 50 avec quelques valeurs extrêmes. boxplot(d\$Chlorures) 0 0 300 100 0 Cette fois-ci, on a moins de variables extrêmes et des valeurs groupées autour de 25. boxplot(d\$Calcium) 0 200 100 0 On a ici des valeurs autour de 100 et presque aucune valeur extrême. boxplot(d\$Magnesium) 0 100 9 40 20 0 On a des valeurs centrées autour de 20 et quelques valeurs extrêmes. boxplot(d\$Sodium) 0 0 0 Le graphe est dur à lire, on a l'air d'avoir majoritairement des petites valeurs (en dessous de 50) mais plusieurs valeurs très extrêmes sont présentes (plus de 1000). 3. cov(d[,-c(1,2)]) Bicarbonates Sulfates Chlorures Calcium Magnesium ## Bicarbonates 881787.632 4057.470 75563.23441 9564.32473 5539.8591 ## Sulfates 4057.470 111192.824 -1292.56989 36652.92581 7065.0559 75563.234 -1292.570 8604.89892 48.17527 641.2075 ## Chlorures 9564.325 36652.926 ## Calcium 48.17527 14784.66452 2954.3398 5539.859 7065.056 641.20753 2954.33978 956.1806 ## Magnesium ## Sodium 342397.995 -1595.884 30225.03871 -1291.97634 561.1957 Sodium ## Bicarbonates 342397.9946 ## Sulfates -1595.8839 ## Chlorures 30225.0387 -1291.9763 ## Calcium ## Magnesium 561.1957 140504.8065 ## Sodium Sans grande surprise, les valeurs sur lesquelles on a remarqué des valeurs extrêmes possèdent des variances très élevées. 4. cor(d[,-c(1,2)]) Chlorures Bicarbonates Sulfates Calcium Magnesium ## Bicarbonates 1.00000000 0.01295791 0.867472370 0.083765768 0.1907861 ## Sulfates 0.01295791 1.00000000 -0.041787152 0.903991883 0.6851841 0.86747237 -0.04178715 1.000000000 0.004271157 0.2235403 ## Chlorures ## Calcium 0.08376577 0.90399188 0.004271157 1.000000000 0.7857506 ## Magnesium 0.19078611 0.68518406 0.223540260 0.785750619 1.0000000 0.97275533 -0.01276784 0.869257384 -0.028346747 0.0484171 ## Sodium ## Sodium ## Bicarbonates 0.97275533 ## Sulfates -0.01276784 ## Chlorures 0.86925738 ## Calcium -0.02834675 ## Magnesium 0.04841710 ## Sodium 1.00000000 Il semble exister une corrélation entre : - Bicarbonates et sodium - Bicarbonates et chlorures - Sulfates et calcium - Chlorures et sodium - Calcium et magnesium - et peut-être magnesium et sulfates 5. $pr \leftarrow prcomp(d[,-c(1,2)], center = TRUE, scale. = TRUE)$ ## Standard deviations (1, .., p=6): ## [1] 1.70350342 1.58644233 0.60574244 0.36312279 0.28612339 0.02506336 ## Rotation $(n \times k) = (6 \times 6)$: PC6 PC1 PC2 PC3 PC4 PC5 ## Bicarbonates 0.5337300 -0.2307045 -0.1144369 0.5101045 -0.01589122 0.62317959 ## Sulfates 0.2161086 0.5416699 -0.4940136 -0.4096336 0.41805399 0.27068923 ## Chlorures 0.5070475 -0.2427741 0.3038885 -0.6917086 -0.32423425 0.08959175 ## Calcium 0.2460346 0.5511043 -0.1874260 0.2079009 -0.71047216 -0.22941014 ## Magnesium 0.3055515 0.4562073 0.7185556 0.1949564 0.36307454 -0.11117519 ## Sodium 0.5073830 -0.2873828 -0.3147315 0.1109149 0.28854232 -0.68217221 summary(pr) ## Importance of components: PC1 PC2 PC3 PC4 ## Standard deviation 1.7035 1.5864 0.60574 0.36312 0.28612 0.02506 ## Proportion of Variance 0.4837 0.4195 0.06115 0.02198 0.01364 0.00010 ## Cumulative Proportion 0.4837 0.9031 0.96427 0.98625 0.99990 1.00000 Il semble qu'il y ait deux composantes principales qui nous permettent de récupérer 90% de l'information. Le paramètre center permet de décaler les valeurs afin que leurs moyennes soient 0. Le paramètre scale permet de diviser toutes les valeurs par l'écart-type afin d'avoir des valeurs sur une plus petit échelle. 6. $pca \leftarrow prcomp(d[,-c(1,2)], center = TRUE, scale. = TRUE)$ df <- scale(d[,-c(1,2)], pca\$center, pca\$scale) %*% pca\$rotationplot(df) text(df, label=d\$Nom) Hepar **Contrex** 4 2 PC2 Les eaux qui sortent du lot sont 0 Ar⊽ie -2 Vichy_©elestin Vichy_St-0 3 5 PC1 Hepar, Contrex, Arvie, Vichy Celestin et Vichy St-Yorre. On pourrait aussi dire que Sanpellegrino, Vittel, Ondine, Verniere, Badoit et Salveta sortent 7. $barplot(pr\$sdev \land 2, names.arg = c('PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6'))$ par(new = TRUE) $s = cumsum((pr\$sdev \land 2) * 100 / sum(pr\$sdev \land 2))$ plot(s, type = 'b', axes = FALSE) axis(4, col = 'red') 100 2.5 2.0 80 1.5 On retient deux composantes car 1.0 9 0.5 20 0.0 PC1 PC2 PC3 PC5 PC6 PC4 Index elles représentent environ 90% de l'information. 8. biplot(pr) -2 0 2 SCHARRER 9.0 7 Magnesium 0.4 7 0.2 Le magnésium, les sulfates et le 0.0 0 -0.2 -0.2 0.6 0.0 0.2 PC1 calcium semblent être corrélés entre eux, tout comme le bicarbonatem, le sodium et le chlorures. 9. biplot(pr, choices = 2:3) -2 0 2 2 9.0 0.4 3 Magnesium 26 0.2 Chloru<u>re</u>s On perd toute corrélation en 0.0 0 Calciu odium -0.2 7 Sulfat 27 -0.2 0.0 0.2 0.4 0.6 PC2 choisissant les composantes principales 2 et 3. biplot(pr, choices = c(1,3))-2 9.0 0.4 3 Magnesium 7 26 0.2 PC3 Chlorur 0.0 Sodiur 7 PC1 Même réponse en prenant les composantes principales 1 et 3. 10. square <- pr\$rotation ^ 2 rowSums(square) Calcium ## Bicarbonates Sulfates Chlorures Magnesium Sodium quality <- rowSums(square[,1:2])</pre> quality Magnesium ## Bicarbonates Sulfates Chlorures Calcium Sodium 0.3160365 0.3380922 0.3401092 0.3642490 0.3014868 0.3400264 Aucune variable n'a l'air mieux represéntée qu'une autre sur le premier plan factoriel, les valeurs sont environ toutes les mêmes. ##Partitionnement 1. d <- read.table("temperatures.csv", header=TRUE, sep = ";", row.names = 1)</pre> January February March April May June July August September 5.7 8.2 12.5 14.8 17.1 17.1 14.5 ## Amsterdam ## Anvers 3.1 6.2 8.9 12.9 15.5 17.9 17.6 14.7 9.7 11.7 15.4 20.1 24.5 27.4 27.2 9.1 23.8 ## Athena ## Barcelone 9.1 21.7 10.3 11.8 14.1 17.4 21.2 24.2 24.1 ## Berlin -0.2 0.1 4.4 8.2 13.8 16.0 18.3 14.4 9.0 11.9 15.0 18.3 20.4 ## Bordeaux 5.6 6.7 17.6 ## Brest 6.9 8.5 9.7 12.6 15.0 17.0 17.0 15.4 6.8 3.3 6.7 8.9 12.8 15.6 17.8 15.0 ## Bruxelles 3.3 ## Budapest -1.1 0.8 5.5 11.6 17.0 20.2 22.0 16.9 ## Copenhague -0.4 -0.4 1.3 5.8 11.1 15.4 17.1 16.6 13.3 ## Cracovie -3.7 -2.0 1.9 7.9 13.2 16.9 18.4 17.6 13.7 ## Dublin 4.8 5.0 5.9 7.8 10.4 13.3 15.0 12.7 ## Edimbourg 2.9 3.6 4.7 7.1 9.9 13.0 14.7 12.1 ## Francfort 0.2 1.8 5.4 9.7 14.3 17.5 19.0 18.3 14.8 8.7 ## Genes 8.7 11.4 13.8 17.5 21.0 24.5 21.8 ## Geneva 0.1 5.1 9.4 13.8 17.3 19.4 15.0 3.1 10.2 14.0 17.2 9.7 ## Helsinki -5.8 -6.2 -2.7 -5.0 -0.3 ## Kiev -5.9 7.4 14.3 17.8 19.4 18.5 13.7 10.5 ## Lisbonne 11.3 12.8 14.5 16.7 19.4 21.5 20.4 ## Londres 3.4 8.3 11.9 15.1 16.9 14.0 4.2 5.5 16.5 6.6 9.4 12.2 16.0 20.8 24.7 ## Madrid 5.0 24.3 19.8 1.1 ## Milan 8.0 12.6 17.3 21.3 23.8 18.9 ## Minsk -6.9 -6.2 -1.9 5.4 12.4 15.9 17.4 16.3 11.6 -7.6 -2.0 -9.3 6.0 13.0 16.6 18.3 16.7 11.2 ## Moscou 6.1 ## Nantes 8.9 11.0 14.6 17.7 19.6 19.6 17.0 ## Oslo -4.3 -3.8 -0.6 4.4 10.3 14.9 16.9 15.4 11.1 10.5 ## Palerme 11.5 13.3 16.9 20.9 23.8 24.5 22.3 3.7 7.3 9.7 13.7 16.5 19.0 18.7 ## Paris 3.7 16.1 -1.3 ## Prague 0.2 3.6 8.8 14.3 17.6 19.3 18.7 14.9 ## Reykjavik -0.3 0.8 2.9 6.5 9.3 11.1 8.2 10.5 13.7 17.8 21.7 24.4 ## Rome 7.1 20.9 ## Sarajevo 9.3 13.8 17.0 18.9 18.7 -1.4 0.8 4.9 15.2 14.1 16.1 19.7 23.4 26.7 ## Seville 10.7 11.8 26.7 24.3 9.7 14.3 17.7 20.0 15.8 ## Sofia -1.7 0.2 4.3 19.5 ## St Petersburg -8.2 -3.7 3.2 10.0 15.4 18.4 16.9 -7.9 11.5 ## Stockholm -3.5 -3.5 -1.3 3.5 9.2 14.6 17.2 16.0 11.7 ## Zurich -0.7 0.7 4.3 8.5 12.9 16.2 18.0 17.2 14.1 October November December Mean Amplitude Latitude Longitude ## Amsterdam 11.4 7.0 4.4 9.9 14.6 52.20 4.50 ## Anvers 11.5 6.8 4.7 10.3 15.0 51.10 4.20 ## Athena 19.2 14.6 11.0 17.8 18.3 37.60 23.50 10.0 16.2 ## Barcelone 17.5 13.1 15.1 41.20 2.20 10.0 52.30 ## Berlin 4.2 1.2 9.1 18.5 13.20 ## Bordeaux 13.5 8.5 6.1 12.7 14.8 44.50 -0.30 ## Brest 7.4 11.6 12.7 9.5 10.2 48.20 -4.30 4.4 10.3 50.50 ## Bruxelles 11.1 6.7 14.4 4.20 ## Budapest 11.3 5.1 0.7 10.9 23.1 47.30 19.00 ## Copenhague 8.8 4.1 1.3 7.8 17.5 55.40 12.30 ## Cracovie 8.6 -1.7 7.7 22.1 50.00 19.60 2.6 5.4 9.3 ## Dublin 9.7 53.20 -6.10 10.2 ## Edimbourg 8.7 5.3 3.7 8.3 11.8 55.00 -3.00 ## Francfort 9.8 4.9 1.7 9.8 18.8 50.10 8.40 ## Genes 17.8 12.2 10.0 16.1 15.9 44.30 9.40 ## Geneva 9.8 4.9 1.4 9.7 19.3 46.10 6.10 ## Helsinki 5.2 0.1 -2.3 4.8 23.4 60.10 25.00 -3.6 7.1 ## Kiev 7.5 1.2 25.3 50.30 30.30 17.4 ## Lisbonne 13.7 11.1 15.9 11.4 38.40 -9.10 ## Londres 10.2 6.3 4.4 9.7 13.5 51.40 0.00 ## Madrid 5.4 13.9 13.9 8.7 19.7 40.20 -3.40 13.1 2.6 12.6 ## Milan 6.9 22.7 45.30 9.20 ## Minsk 5.8 0.1 -4.2 5.5 24.3 53.50 27.30 -6.0 5.1 ## Moscou 5.1 -1.1 27.6 55.45 37.37 13.5 6.4 12.5 ## Nantes 9.1 13.5 47.10 -1.30 ## Oslo 5.7 -2.9 5.6 21.2 59.50 10.50 0.5 ## Palerme 18.4 14.9 12.0 16.6 14.0 38.10 13.10 ## Paris 12.5 7.3 5.2 11.2 15.3 48.50 2.20 ## Prague 9.4 3.8 0.3 9.2 20.6 50.00 14.20 ## Reykjavik 4.5 1.7 0.2 4.6 11.4 64.10 -21.60 ## Rome 16.5 11.7 8.3 15.4 17.3 41.50 12.30 ## Sarajevo 10.5 5.1 0.8 9.4 20.3 43.50 18.30 19.4 ## Seville 14.5 11.2 18.2 16.0 37.20 -5.60 ## Sofia 10.7 5.0 0.6 9.6 21.7 42.40 23.20 ## St Petersburg 5.2 -5.3 4.5 26.6 59.60 30.20 -0.4 ## Stockholm 6.5 -1.6 5.8 20.7 59.20 18.00 1.7 0.3 8.7 ## Zurich 8.9 3.9 18.7 47.20 8.30 Sunshine Precipitation Elevation 1662.0 838.2 2 ## Amsterdam ## Anvers 12 1654.0 785.0 ## Athena 2847.0 414.1 170 ## Barcelone 2591.0 588.0 9 ## Berlin 1437.0 570.0 34 ## Bordeaux 2035.0 944.0 21 ## Brest 1529.0 1210.0 34 ## Bruxelles 1546.0 852.0 13 117 ## Budapest 1988.0 532.0 ## Copenhague 1539.0 613.0 5 219 ## Cracovie 1469.0 631.0 ## Dublin 1453.0 714.0 10 47 ## Edimbourg 1426.0 704.0 ## Francfort 113 1662.0 629.0 ## Genes 2222.0 1072.0 2 1828.0 1005.0 373 ## Geneva ## Helsinki 1858.0 655.0 51 ## Kiev 1825.0 621.0 179 ## Lisbonne 2800.0 774.0 ## Londres 1632.0 601.0 35 ## Madrid 2769.0 421.0 667 1017.0 ## Milan 1915.0 120 1754.0 690.0 220 ## Minsk ## Moscou 1731.0 707.0 156 788.0 1954.0 27 ## Nantes 1668.0 763.0 ## Oslo 23 ## Palerme 2646.0 610.0 14 ## Paris 1661.0 649.6 79 ## Prague 1667.0 525.0 288 2 ## Reykjavik 1268.0 798.0 ## Rome 2472.0 804.0 21 ## Sarajevo 1769.0 928.0 518 7 ## Seville 2917.4 538.8 ## Sofia 2177.0 581.0 650 ## St Petersburg 1551.0 633.0 3 ## Stockholm 1821.0 539.0 1 ## Zurich 1531.0 1031.0 408 distances <- dist(d[1:12])</pre> cluster <- hclust(distances)</pre> plot(cluster, main = "Cluster Dendrogram") **Cluster Dendrogram** 50 40 30 20 Height 10 distances hclust (*, "complete") (dendrogram <- as.dendrogram(cluster))</pre> ## 'dendrogram' with 2 branches and 37 members total, at height 49.10509 (cuted <- cut(dendrogram, 21))</pre> ## 'dendrogram' with 2 branches and 4 members total, at height 49.10509 ## ## \$lower ## \$lower[[1]] ## 'dendrogram' with 2 branches and 7 members total, at height 11.07384 ## \$lower[[2]] ## 'dendrogram' with 2 branches and 9 members total, at height 20.01375 ## 'dendrogram' with 2 branches and 10 members total, at height 13.84124 ## \$lower[[4]] ## 'dendrogram' with 2 branches and 11 members total, at height 16.46086 La hauteur représente la différence de températures entre les groupes. On peut couper à la hauteur 21 pour former quatre groupes différents.

Analyse de données - Rapport