Groundwater spring classification in Southern Bandung area: PCA dan cluster analysis application

Arif Susanto and Dasapta Erwin Irawan

December 20, 2015

Contents

Abstract

This paper discusses groundwater spring classification based on geological observation supported by multivariate statistics. Seven groundwater spring sites located in the volcanic area of Southern Bandung area have been observed and sampled to test the hydrochemistry contents. We have measured 30 variables for each sample including: physical properties (turbidity, TDS, EC, etc), major elements (Ca, Na, Mg, etc), and trace elements (SiO_2, B, As, etc). R statistical packages were used to fit the principal component analysis (PCA) and cluster analysis (CA). We find three clusters on the CA: cluster 1 (Situ Kince, Bedil), cluster 2 Ciseupan, Ciblegblegan), and cluster 3 (Citawa, Cigoong, Cikoleberes). The PCA shows SiO2 is the strongest variable to control cluster 1; nitrate, EC, TDS, hardness to control cluster 2; and no string variables signal for cluster 3.

Introduction

This paper discusses groundwater spring classification based on geological observation supported by multivariate statistics.

Materials and methods

The data set discussed in this paper came from Comunity Empowerment Research 2015 entitled ".....". We managed to set up a hydrochemical data set with a dimension of seven rows and 33 columns.

We observed 15 spring sites but only seven of them were analysed in the laboratory.

Field mapping

The field geological mapping was conducted in April-July 2015.

Laboratory testing

We tested water samples at the Water Quality Lab, Dept of Environmental Engineering ITB.

Statistical analysis

In this paper we will use free and open source R statistical software. R is a system for statistical computation and graphics, consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files. R can be freely downloaded has a home page at http://www.R-project.org/.

Both methods, PCA and CA, have used extensively to classify hydrochemical data. A few examples were taken from Irawan et.al (2009), ..., ..., ..., In all of those examples, PCA were used to reduce the dimensionality of the data set. In this case we need to reduce the number of measured variables into groups of variables that significantly contribute to the hydrochemical outputs. The PCA methods will simultaneously transform the multidimensional axis of from the original data set to make a new set of Principal Components (PC's). The original variables were then projected on to the new created PC, based on its loading values. Each sample then were also plotted against those PC's to see the most controlling PC for each sample.

Package yang diperlukan:

1. PCA: `princomp()` atau `prcomp()`, gunanya untuk mengekstrak variabel (component) berpengaruh dalam suatu data set dengan jumlah variabel yang sangat banyak. Fungsi ini akan mengelompokkan variabel menjadi lebih ringkas, misal: bila semua kita punya 33 variabel, maka nantinya akan dapat menjadi dua atau tiga kelompok variabel yang disebut PC (principal component)
2. Cluster: `kmeans()` dan `hclust()`, gunanya untuk menguji kemiripan sampel berdasarkan perhitungan Euclidean distance dan mengelompokkannya dalam sebuah dendogram.

Namun demikian dalam kesempatan ini saya akan menggunakan package:

1. `pcamethods` yang ditulis oleh Wolfram Stacklies, Henning Redestig, dan Kevin Wright. link
2. `cluster` yang ditulis oleh Friedrich Leisch dan Bettina Gruen link 3. `vegan` ditulis oleh Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner.

Tahapannya akan saya jelaskan lebih rinci besok ya per blok kode. Data set juga akan segera tersedia setelah publikasi diterbitkan.

1. Instalasi dan load library

Dalam analisis ini kami menggunakan beberapa package: `pcaMethods`, `cluster`, `vegan` dan aplikasi pembuka format xls `readxl`.

Package `pcamethods` tersedia di server repo Bioconductor, sehingga cara pengunduhan dan instalasinya pun berbeda. Untuk membuka file data dengan format xls ada beberapa package lainnya, misalnya 'readr'. Kami membuka file langsung dari ormat xls karena ditemui masalah saat membuka file dengan fungsi `read.csv` standar. Semantara ini kami menduga masalah ada di konversi `unicode utf-xxx`.

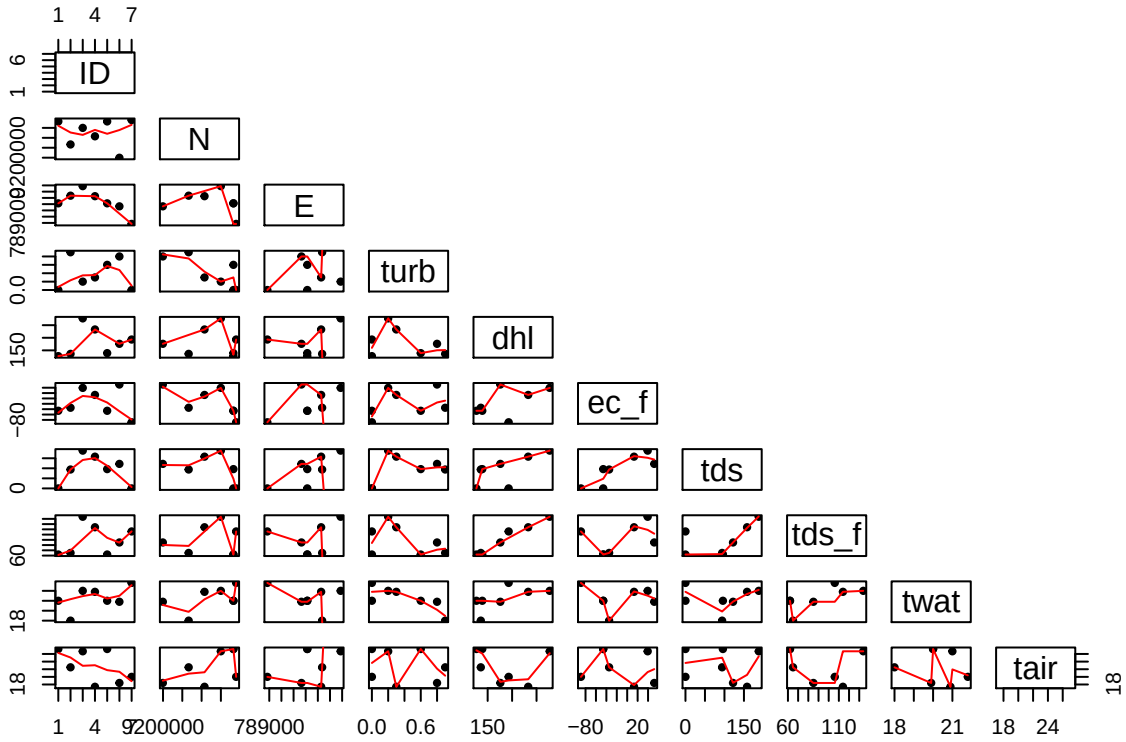
Regional hydrogeological setting

The hydrogeological setting of the area is fairly complex due to various volcanic systems.

Results and discussions

PCA

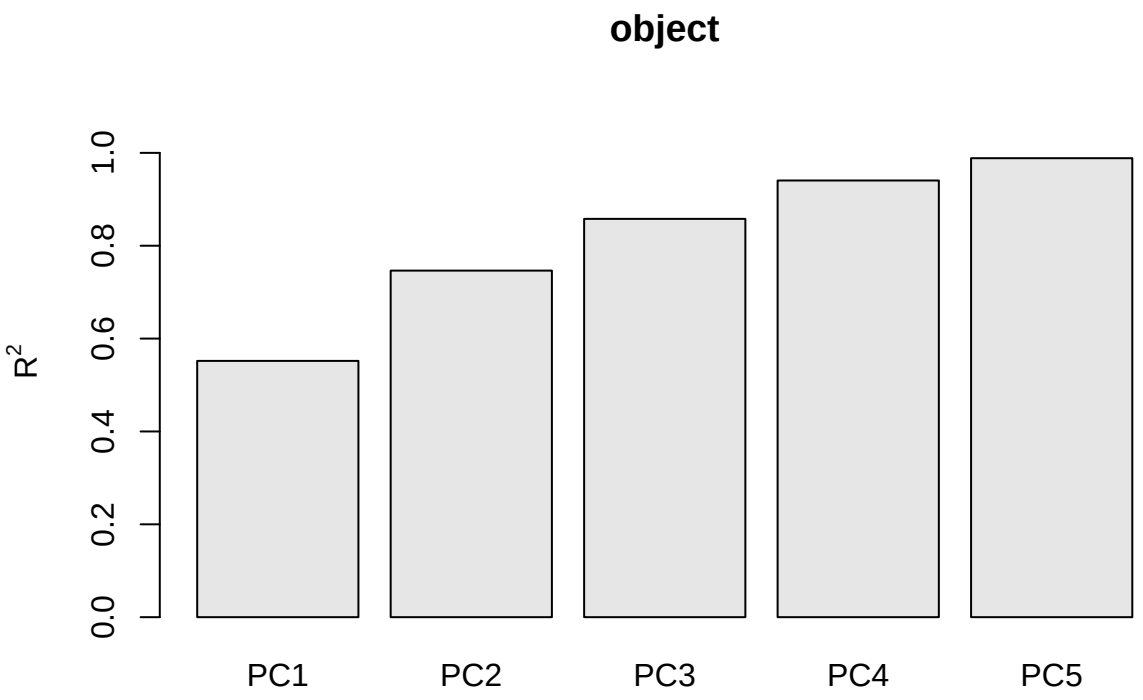
```
library(pcaMethods)
df <- read.csv("data_cisanti.csv")
df <- na.omit(df) # omitting NA if any

# Exploratory using pairs() function
# Assesing data patterns
pairs(df[1:10],
      lower.panel=panel.smooth,
      upper.panel=NULL,
      pch=20)
```



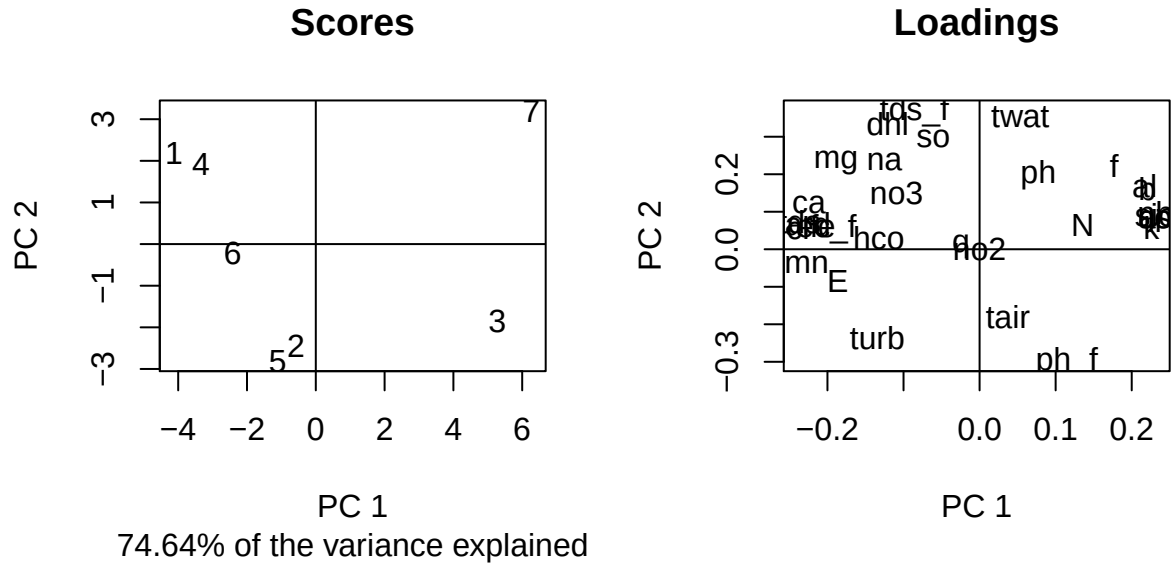

```
# Run PCA (using pcamethods package)
## svdImpute = standard pca, with imputation, standardised, method univariate (uv)
pca <- pca(df,
           method = "svdImpute",
           scale = "uv",
           center = T,
           nPcs = 5,
           evalPcs = 1:5)
summary(pca)
```

```
## svdImpute calculated PCA
## Importance of component(s):
##                   PC1    PC2    PC3     PC4     PC5
## R2            0.552 0.1945 0.1115 0.08264 0.04804
## Cumulative R2 0.552 0.7464 0.8579 0.94051 0.98854
```

```
## Evaluating results
plot(pca, type="lines")
```




```
slplot(pca) # default function in pcamethods but not big enough
```




```
loadings(pca) # loadings of each variables
```

```
##               PC1          PC2           PC3          PC4          PC5
## N       0.13561565  0.067235936 -3.090428e-01 -0.229998892 -0.355399811
## E      -0.18640077 -0.083279176 -2.649212e-01 -0.082744946  0.119390668
## turb   -0.13425438 -0.236789387  2.508494e-01  0.218689809 -0.095461654
## dhl    -0.12072414  0.335801575 -1.400544e-01 -0.004297532  0.008575054
## ec_f   -0.20127317  0.058908938 -1.959546e-02  0.067843227  0.401111075
## tds    -0.23635845  0.072375018 -5.727448e-02 -0.015768366 -0.019529285
## tds_f  -0.08556510  0.369935409 -1.083089e-01  0.019145799  0.005026424
```

```
## twat   0.05347677  0.352526452 -7.933662e-02 -0.140094401 -0.062848379
## tair   0.03728633 -0.180764247 -4.570466e-01  0.004021935 -0.152790488
## ph     0.07722381  0.198496875  1.998978e-01  0.078242033 -0.578448219
## ph_f   0.11470078 -0.298283118 -8.056462e-02 -0.225256174 -0.226319510
## q     -0.02431153  0.005555798 -6.294331e-05 -0.621359727 -0.006258281
## hard  -0.23748527  0.068167047 -3.931357e-02 -0.020611066 -0.013232936
## ca    -0.22408322  0.118082088 -9.945594e-02  0.088623704 -0.015810934
## mg    -0.18869450  0.229700582 -1.614958e-02 -0.140818323  0.011509006
## fe    -0.20605342  0.063494660 -2.341299e-01  0.135874705 -0.091662356
## mn    -0.22725322 -0.040150254 -2.147860e-02 -0.057719265 -0.239293010
## k      0.22611527  0.057318843 -8.131144e-02 -0.049121645  0.225377968
## na    -0.12483226  0.232400206  1.413893e-01 -0.358042492  0.057003358
## li     0.23054588  0.077713515 -7.037698e-02 -0.029413722  0.150825498
## nh     0.23112724  0.097792247 -5.166800e-02 -0.010657859  0.109381897
## co     0.23054588  0.077713515 -7.037698e-02 -0.029413722  0.150825498
## hco   -0.13245252  0.031152331  3.180080e-01 -0.363704139  0.040312808
## cl    -0.23863522  0.058576577 -3.521203e-02 -0.003917488 -0.015628316
## so    -0.06084941  0.294658301  2.896670e-01  0.217901488  0.030981286
## no2    0.00000000  0.000000000  0.000000e+00  0.000000000  0.000000000
## no3   -0.10899650  0.148584300 -4.036184e-01  0.191619363 -0.014894643
## sio    0.23113506  0.092737423 -5.647441e-02 -0.015448539  0.120025512
## b      0.22056864  0.160973426  1.495117e-02  0.053895873 -0.037915730
## al     0.21714222  0.169866372  2.554158e-02  0.063848962 -0.061293453
## as     0.23054588  0.077713515 -7.037698e-02 -0.029413722  0.150825498
## f      0.17720808  0.223949610  1.018918e-01  0.132912041 -0.229500309
```

```
scores(pca) # scores of each samples respectively to each variables
```

```
##          PC1        PC2        PC3        PC4         PC5
## 1 -4.1198522  2.1974572 -3.1864665  1.0198745 -0.01921167
## 2 -0.5730037 -2.4534493 -0.3409084 -0.8818771 -1.81878975
## 3  5.2817359 -1.8378178 -1.6676760 -1.0930760  1.58169782
## 4 -3.3355644  1.9197928  1.4123501 -2.7482052  0.23549515
## 5 -1.1038752 -2.8120111  0.7906250  1.2769001 -0.46800339
## 6 -2.4167773 -0.2233124  2.0363043  1.5539281  1.41293504
## 7  6.2673369  3.2093406  0.9557715  0.8724556 -0.92412320
```

```
row.names(scores(pca))
```

```
## [1] "1" "2" "3" "4" "5" "6" "7"
```

Cluster

Conclusions

Acknowledgements

References

http://www2.stat.unibo.it/montanari/Didattica/Multivariate/CA_lab.pdf

http://cc.oulu.fi/~jarioksa/opetus/metodi/sessio3.pdf

http://www2.stat.unibo.it/montanari/Didattica/Multivariate/PCA_lab1.pdf

http://bioconductor.wustl.edu/bioc/vignettes/pcaMethods

https://cran.r-project.org/web/packages/vegan/vignettes/intro-vegan.pdf

http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf