Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

Hydrochemical assessment of Semarang area using multivariate statistics: A sample based dataset

Irawan Dasapta Erwin[1] and Putranto Thomas Triadi[2]

[1]Faculty of Earth Sciences and Technology, Institut Teknologi Bandung, Jalan Ganesa No. 10, Bandung - 40132, Indonesia
[2]Faculty of Engineering, Universitas Diponegoro, Jalan Prof. H. Soedarto, SH, Tembalang, Kota Semarang - 50275, Indonesia

Correspondence to: Dasapta Erwin Irawan (dasaptaerwin@outlook.co.id)

Abstract. The following paper describes in brief the data set related to our project "Hydrochemical assessment of Semarang Groundwater Quality". All of 58 samples were taken in 1992, 1993, 2003, 2006, and 2007 using well point data from several reports from Ministry of Energy and Mineral Resources and independent consultants. We provided 20 parameters in each samples (sample id, coord X, coord Y, well depth, water level, water elevation, TDS, pH, EC, K, Ca, Na, Mg, Cl, SO4, HCO3, year, ion balance, screen location, and chemical facies). The chemical composition were tested in the Water Quality Laboratory, Universitas Diponegoro using mas spectrofotometer method.

The statistical treatment for the dataset (available on Zenodo doi:10.5281/zenodo.57293) were described as follows: (1) data preparation in to csv file format, load it in to R environment; (2) data treatment, including: correlation matrix, cluster analysis using kmeans and hierarchical cluster analysis, and principal component analysis. For analysis and visualizations, We used the following R packages: `ggplot2`, `dplyr`, `factomineR`, `factoExtra`, `cluster`, `ggcorrplot`, and `ape`.

1 Introduction

The following paper describes in brief the data set related to our project "Hydrochemical assessment of Semarang Groundwater Quality". The aim of this project is to understand the water quality classification and distribution in Semarang area and to explain the underlying processes. This analysis is very important with the vast development of infrastructure (Putranto and Rüde (2016)) and urban settlement in coastal area and the rate of salinity encroachment (Rahmawati and Marfai (2013)). The location of the study is Semarang area, Indonesia.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

2 General description of the dataset

2.1 Samples

All of 58 taken in 1992, 1993, 2003, 2006, and 2007 in 1992, 1993, 2003, 2006, and 2007 using well point data from several reports from Ministry of Energy and Mineral Resources and independent consultant. We provided 20 parameters in each samples: `sample id`, `coord X`, `coord Y`, `well depth`, `water level`, `water elevation`, `TDS`, `pH`, `EC`, `K`, `Ca`, `Na`, `Mg`, `Cl`, `SO4`, `HCO3`, `year`, `ion balance`, `screen location`, and `chemical facies`. The chemical composition were tested in the Water Quality Laboratory, Universitas Diponegoro using mass spectrofotometer method. The laboratory procedures followed the SNI (Indonesia National Standard) for water quality measurement (BSN (2012)), which is comply to the US-EPA standards. The original dataset is available on Zenodo (Irawan and Putranto (2016)).

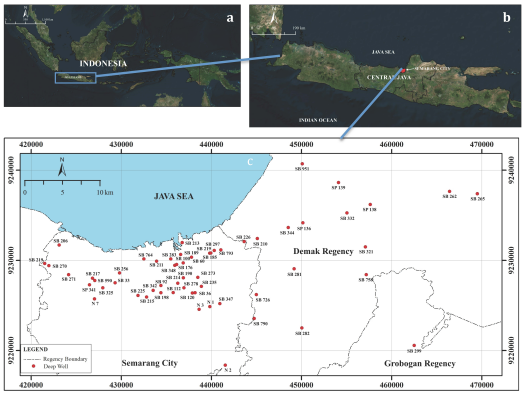

Figure 1. The location of well point and the Stiff diagram

2.2 The value of dataset

The following list describes the value of the dataset:

- It provides the current setting of water quality as the baseline of environmental monitoring of the area and serves as a source of groundwater quality indicator for the regional planning of the area,
- It promotes the importance of open government dataset and enriches the library of water quality dataset of the area,
- It sets an example of data re-use and re-analysis in hydrogeological research landscape.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

3 Geographical coverage

The sampling area is Semarang area, the capital of Mid Java Province, Java, Indonesia. The sampling points were distributed from the southern volcanic highland to the coastal area. The coordinate of the area is (420000, 9240000) and (470000, 9220000). We plotted the data points using UTM-WGS84-48S projection system.

4 Statistical design

The hierarchical cluster analysis (HCA) and principal component analysis (PCA) are both widely used in the hydrochemical analysis (Adams et al. (2001); King et al. (2014); Ayenew et al. (2009); Deon et al. (2015); Wilkinson (2014); Maechler et al. (2016)). We have applied the two approaches on groundwater in volcanic area on various locations (Irawan et al. (2009); Herdianita et al. (2010)). The R implementation was based on Coghlan (2009).

4.1 Data preparation

The dataset was formatted in the csv (comma separated value) before parsed in to R program (R Core Team (2016)) for analysis using the following R packages: ggplot2 (Wickham (2009)), dplyr (Wickham and Francois (2016)), factomineR (Lê et al. (2008)), factoExtra (Kassambara and Mundt (2016)), cluster (Maechler et al. (2016)), ggcorrplot (Kassambara (2016)), and ape (Paradis et al. (2004)).

```
df <- as.data.frame(read.csv("data_smg.csv")) # loading as data frame
head(df)              # checking header
is.na(df)             # checking NAs in df
df2 <- df[c(2,5:18)] # subsetting df, exclude var with NAs
head(df2)
is.na(df2)            # checking NAs in df2
str(df2)              # checking data type in df2
is.numeric(df2)       # checking data type in df2
rownames(df2) <- df2$location # setting col location as row names
str(df2)              # checking data type in df2
```

4.2 Data treatment

The dataset was treated using the following method: correlation matrix, HCA, and PCA. the steps and R code can be described below.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

4.2.1 Correlation matrix

Here we used `PerformanceAnalytics` and `ggcorrplot` packages to build a correlation matrix. The following is the code.

```
## using PerformanceAnalytics
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)
chart.Correlation(df2, histogram=TRUE, pch=19) # visual PA

## using ggcorrplot
install.packages("ggcorrplot")
library(ggcorrplot)
correl <- round(cor(df2), 1)     # rounding correl matrix
head(correl[, 1:14])             # view headers
p.mat <- cor_pmat(df2)           # compute p-values
head(p.mat[, 1:14])              # view headers
ggcorrplot(correl)               # making heatmap
```

4.2.2 Hierarchical cluster analysis (CA)

We build the CA using k-means and hierarchical clustering by implementing R base function and `factoextra` package, based on the following code.

```
install.packages("factoextra")
#install_github("kassambara/factoextra")
install.packages("cluster")
library(cluster)
library(factoextra)

### k means method
km2 <- kmeans(df2, 2, nstart = 25) # kmeans with 2 centers
km3 <- kmeans(df2, 3, nstart = 25) # kmeans with 3 centers
km2$cluster                        # extracting cluster number
km2$centers                        # extracting cluster means (or centers)
plotkm2 <- plot(df2,
                col = km2$cluster,
                pch = 19,
                frame = T,
                main = "K-means with k = 2") # notes: need longer x axis
```

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

```
105     points(km2$centers,
               col = 1:2,
               pch = 8, cex = 3)

        km3$cluster                     # extracting cluster number
110     km3$centers                     # extracting cluster means (or centers)
        plotkm3 <- plot(df2,
                        col = km3$cluster,
                        pch = 19,
                        frame = T,
115                     main = "K-means with k = 3")
        points(km3$centers,
               col = 1:2,
               pch = 8,
               cex = 3)
120
        ### evaluating cluster
        df2 <- scale(df2)
        head(df2)
        fviz_nbclust(df2,
125                  kmeans, method = "wss") +
                     geom_vline(xintercept = 3,
                                linetype = 2)   # determining optimal no cluster
        km3.res <- kmeans(df2, 3, nstart = 25)  # running kmeans with 4 cluster
        print(km3.res)                          # print output
130     fviz_cluster(km3.res, data = df2)       # vis output

        pam.res <- pam(scale(df2), 3)           # running pam cluster with 3 cluster
        pam.res$medoids                         # extract medoids
        clusplot(pam.res,
135              main = "Cluster plot, k = 3",
                 color = TRUE)
        plot(silhouette(pam.res),  col = 2:5)
        fviz_silhouette(silhouette(pam.res))
        clarax <- clara(df2, 3, samples = 5)    # using clara method
140     fviz_cluster(clarax,
                     stand = FALSE,
```

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

```
                    geom = "point",
                    label=T,
                    pointsize = 1)

### Creating dendogram
distdf2.res <- dist(df2,
                    method = "euclidean")
hcadf2 <- hclust(distdf2.res,
                 method = "complete")
plot(hcadf2,
     hang = -1)        # dendogram vis
rect.hclust(hcadf2,
            k = 3,
            border = 2:4) # dendogram vis with grouping

### using nbclust pack to evaluate no of cluster
install.packages("NbClust") # for more precise no of cluster
library("NbClust")
resdf2.nb <- NbClust(df2,
                     distance = "euclidean",
                     min.nc = 2, max.nc = 10,
                     method = "complete",
                     index ="gap")
resdf2.nb                  # print the results
resdf2.nb$All.index        # All gap statistic values
resdf2.nb$Best.nc          # Best number of clusters
resdf2.nb$Best.partition   # calculate best partition
nbdf2 <- NbClust(df2,
                 distance = "euclidean",
                 min.nc = 2,
                 max.nc = 10,
                 method = "complete",
                 index ="all")
nbdf2
fviz_nbclust(nbdf2) + theme_minimal()
dev.off()                  # delete the '#' sign whenever
                           # you want to clean the plot screen
```

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016


```
distdf2.res <- dist(df2,
                    method = "euclidean")
hcadf2 <- hclust(distdf2.res,
                 method = "complete")
plot(hcadf2,
     hang = -1)          # dendogram vis
rect.hclust(hcadf2,
            k = 3,
            border = 2:4) # dendogram vis with grouping

#### rotating the plot

#### using ape
# load package ape; remember to install it: install.packages('ape')
install.packages("ape")
library(ape)
plot(as.phylo(hcadf2),
     cex = 0.9,
     label.offset = 1,
     type = "unrooted")

plot(as.phylo(hcadf2),
     cex = 0.9,
     label.offset = 1)
```

4.2.3 Principal component analysis (PCA)

The PCA is applied using R base function and visualized using `factominer` and `factoextra` packages. The following is the code.

```
df <- as.data.frame(read.csv("data_smg.csv")) # loading as data frame
head(df)              # checking header
is.na(df)             # checking NAs in df
df2 <- df[c(2,5:18)] # subsetting df, exclude var with NAs
head(df2)
is.na(df2)            # checking NAs in df2
str(df2)              # checking data type in df2
is.numeric(df2)       # checking data type in df2
```

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

```
rownames(df2) <- df2$location # setting col location as row names
str(df2)              # checking data type in df2

install.packages("FactoMineR")
    library("FactoMineR")
    library(factoextra)
    res.pca <- PCA(df2, graph = FALSE)
    eigenvalues <- res.pca$eig
    head(eigenvalues[, 1:2])
    barplot(eigenvalues[, 2], names.arg=1:nrow(eigenvalues),
    main = "Variances",
    xlab = "Principal Components",
    ylab = "Percentage of variances",
    col ="steelblue")
    # Add connected line segments to the plot
    lines(x = 1:nrow(eigenvalues), eigenvalues[, 2],
  type="b", pch=19, col = "red")
    res.pca$var$contrib
    fviz_pca_var(res.pca)

    fviz_pca_var(res.pca, col.var="steelblue")+
     theme_minimal()

    res.pca$ind$contrib
    plot(res.pca, choix = "ind")

    fviz_pca_biplot(res.pca,  geom = "text")
```

5 Conclusions

The present study integrates geological, hydrogeological data, and statistical analysis to construct a hydrogeological model of the aquifer system in Semarang. The statistical treatment shows a consistent pattern of anomalous setting at well point 37 (University Sultan Agung 2/Unisula-2). The anomaly needs more in depth analysis to understand the underlying processes in the groundwater flow.

This paper is one of our preliminary example of data paper in Indonesia. Hopefully this can trigger more data papers to endorse open science in our country.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Acknowledgements. The authors are thankful to the Department of Energy and Resources of Central Java Province and Geological Agency in Bandung for providing hydrogeological data. Hopefully this paper will initiate a mass movement on open government data and data reuse in Indonesia.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Open Access Earth System Science Data Discussions

References

Adams, S., Titus, R., Pietersen, K., Tredoux, G., and Harris, C.: Hydrochemical characteristics of aquifers near Sutherland in the Western Karoo, South Africa, Journal of Hydrology, 241, 91–103, 2001.

Ayenew, T., Fikre, S., Wisotzky, F., Demlie, M., and Wohnlich, S.: Hierarchical cluster analysis of hydrochemical data as a tool for assessing the evolution and dynamics of groundwater across the Ethiopian rift, International journal of physical sciences, 4, 76–90, http://www.academicjournals.org/journal/IJPS/article-abstract/D64DFAE18634, 2009.

BSN: Standard tests for water sample, Tech. rep., National Board for Standards, http://sisni.bsn.go.id/index.php/sni_main/sni/detail_sni/7689, 2012.

Coghlan, A.: Little Book of R for Multivariate Analysis! — Multivariate Analysis 0.1 documentation, Wellcome Trust Sanger Institute, Cambridge, U.K., https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/, affiliation: Wellcome Trust Sanger Institute, Cambridge, U.K., 2009.

Deon, F., Förster, H.-J., Brehme, M., Wiegand, B., Scheytt, T., Moeck, I., Jaya, M., and Putriatni, D.: Geochemical/hydrochemical evaluation of the geothermal potential of the Lamongan volcanic field (Eastern Java, Indonesia), Geothermal Energy, 3, 1–21, doi:10.1186/s40517-015-0040-6, 2015.

Herdianita, N. R., Julinawati, T., and Amorita, I. E.: Hydrogeochemistry of Thermal Water from Surface Manifestation at Gunung Ciremai and Its Surrounding, Cirebon, West Java–Indonesia, in: Proceedings World Geothermal Congress 2010, http://www.geothermal-energy.org/pdf/IGAstandard/WGC/2010/1476.pdf, 2010.

Irawan, D. E. and Putranto, T. A.: Dataset: hydrochemical assessment of Semarang area, Indonesia, doi:10.5281/zenodo.57293, http://dx.doi.org/10.5281/zenodo.57293, 2016.

Irawan, D. E., Puradimaja, D. J., Notosiswoyo, S., and Soemintadiredja, P.: Hydrogeochemistry of volcanic hydrogeology based on cluster analysis of Mount Ciremai, West Java, Indonesia, Journal of hydrology, 376, 221–234, http://www.sciencedirect.com/science/article/pii/S002216940900434X, 2009.

Kassambara, A.: ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2', https://CRAN.R-project.org/package=ggcorrplot, r package version 0.1.1, 2016.

Kassambara, A. and Mundt, F.: factoextra: Extract and Visualize the Results of Multivariate Data Analyses, https://CRAN.R-project.org/package=factoextra, r package version 1.0.3, 2016.

King, A. C., Raiber, M., and Cox, M. E.: Multivariate statistical analysis of hydrochemical data to assess alluvial aquifer–stream connectivity during drought and flood: Cressbrook Creek, southeast Queensland, Australia, Hydrogeology Journal, 22, 481–500, doi:10.1007/s10040-013-1057-1, http://link.springer.com/10.1007/s10040-013-1057-1, 2014.

Lê, S., Josse, J., and Husson, F.: FactoMineR: A Package for Multivariate Analysis, Journal of Statistical Software, 25, 1–18, doi:10.18637/jss.v025.i01, 2008.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K.: cluster: Cluster Analysis Basics and Extensions, r package version 2.0.4 — For new features, see the 'Changelog' file (in the package source), 2016.

Paradis, E., Claude, J., and Strimmer, K.: APE: analyses of phylogenetics and evolution in R language, Bioinformatics, 20, 289–290, 2004.

Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-29, 2016
Manuscript under review for journal Earth Syst. Sci. Data
Published: 28 July 2016

Putranto, T. and Rüde, T.: Hydrogeological Model of an Urban City in a Coastal Area, Case study: Semarang, Indonesia, Indonesian Journal on Geoscience, 3, 17–27, https://ijog.geologi.esdm.go.id/index.php/IJOG/article/view/227, 2016.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2016.

Rahmawati, N. and Marfai, M.: Salinity Pattern in Semarang Coastal City: An Overview, Indonesian Journal of Geosciences, 8, doi:10.17014/ijog.3.1.17-27, https://ijog.geologi.esdm.go.id/index.php/IJOG/article/view/160, 2013.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, http://ggplot2.org, 2009.

Wickham, H. and Francois, R.: dplyr: A Grammar of Data Manipulation, https://CRAN.R-project.org/package=dplyr, r package version 0.5.0, 2016.

Wilkinson, D. J.: Multivariate Data Analysis using R: a course notes, Tech. rep., https://www.staff.ncl.ac.uk/d.j.wilkinson/teaching/mas8381/notes14.pdf, 2014.