Classifying hot water chemistry: Application of multivariate statistics

Prihadi Sumintadireja, Dasapta Erwin Irawan, Rina Herdianita, Yuano Rezky, Prana Ugiana Gio, Anggita Agustin

Faculty of Earth Sciences and Technology, Institut Teknologi Bandung
Ministry of Energy and Mineral Resources of Indonesia
Faculty of Math and Natural Sciences, Universitas Sumatera Utara
dasaptaerwin@outlook.co.id

Resumo

The following paper is a try out on the application of multivariate analysis (regression tree, principal component analysis, and cluster analysis) for classifying hot water chemistry. The number of sample analysed was 416 from all over Indonesia. Regression tree technique has failed to read the data structure due to multi-collinearity effect therefore PCA and cluster analysis were applied. We used open source R statistical packages to do the calculation. Such technique classifies hot water samples into three major clusters: cluster 1 (pure hot water), cluster 2 (mixing water), and cluster 3 (cold-meteoric water). Similar clustering were also detected in the PCA plot. The statistical is able to detect the close and open geothermal system based on data structure. This robust method should be applied to more geothermal system with larger dataset to see its performance.

1. Introduction

Indonesia ranked 3rd in the world in both geothermal electricity production and geothermal generating capacity in 2014. Our current geothermal capacity of 1.3 GW consists of plants clustered around Java, Bali, North Sumatra, and North Sulawesi (Bronto, 2006; Darmawan et al., 2015; Fauzi, 2015; Manalu, 1988; Smillie et al., n.d.). However, despite of the various techniques used in geothermal exploration, the application of multivariate analysis/statistics is still very low (Agustin et al., 2015). Such large quantity of measured parameters in each observation site needs to be deeply analysed using a robust multivariate technique. This paper is one of our methodological trial offering correlations between statistical and deterministic model.

2. Materials and method

2.1 Data

The data was gathered from various sources: EBTKE reports, researches from ITB and other universities. At least 30 variables were measured on each samples (Table 1).

Tabela 1: *The description of variables in the dataset*

No	Parameters	Parameters	Units
1	elv	elevation	masl
2	wtemp	water temperature	oC
3	atemp	air temperature	oC
4	ph	pH	-
5	ec	Electro Conductivity	mikroSiemens/cm
6	SiO2	Silica	mg/L
7	B	Boron	mg/L
8	Al	Aluminium	mg/L
9	Fe	Iron	mg/L
10	Ca	Calsium	mg/L
11	Mg	Magnesium	mg/L
12	Na	Sodium	mg/L
13	K	Potassium	mg/L
14	Li	Lithium	mg/L
15	As	Arsen	mg/L
16	NH4	Amonium	mg/L
17	F	Fluor	mg/L
18	Cl	Chlor	mg/L
19	SO4	Sulfate	mg/L
20	HCO3	Bicarbonate	mg/L

2.2 Software and code

To ensure the reproducibility of this work and to support open science principals, we made the code and dataset available on OSF server (https://osf.io/wbnz3/). We used R statistical program and R Studio. It is open source and can be downloaded from (https://cran.r-project.org and http://rstudio.com), along with some add on packages.

2.3 Regression tree method

Regression tree mainly uses the principal of linear model. It is one of the most frequently used machine-learning methods for building prediction models. As a result, the partitioning can be represented graphically as a decision tree (ath and Fabricius, 2000; Faraway, 2005; Gio, Prana Ugiana and Rosmaini, Elly, 2015; Gregor et al., 2002; Loh, 2011; Rothwell et al., 2008). Relying on the assumption of normal distribution and no collinearity within the dataset, we expect the method could split the samples into several classes (Kuhn, 2008; Lewis, 2000).

2.4 Cluster analysis and principal component analysis

Cluster analysis (CA) is an unsupervised classification method of a set of objects in groups. The HCA is useful for variables or observations, categorical data, or continuous variables. Principal component analysis (PCA) is another form of unsupervised classification method applying the rotation of covariance matrix to reduce the dimension of the dataset by retaining most of the information in the original set (Chen et al., 2007; Fitzpatrick et al., 2007; Irawan et al., 2009; King et al., 2014; Ritzi et al., 1993; Seyhan et al., 1985).

3. Results and discussions

3.1 Regression tree

Regression tree technique has failed to read the data structure due to collinearity effect. The so called "statistical hazards" is identified from the extremely high correlation coefficients (R^2): ec vs B, Na, Cl; Na vs K, Li; NH4 vs K, Li. Only three branches are extracted from the procedure, involving air temperature (atemp), elevation (elv), and B (Boron concentration), which is very difficult to explain (Figure 1).

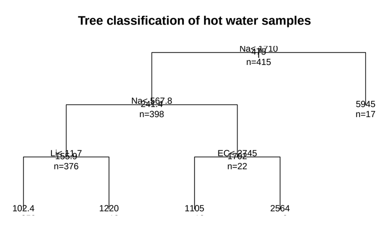

Figura 1: *Data classification based on regression tree analysis.*

3.2 Cluster analysis and principal component analysis

The PCA extracts five new sets of variable accounted for 88% of variation from the original set of 30 variables.(Figure 2), with anomalous data are shown by samples from East Flores and Karaha Bodas. The majority of volcanic samples is distributed from the center of the plot upwards, while the majority non-volcanic system from the center to the right (Figure 3),

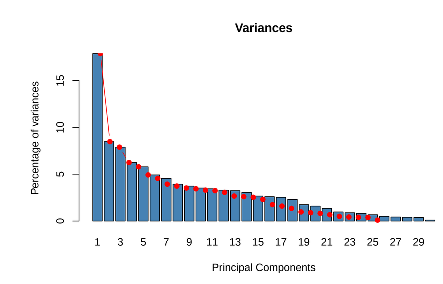

Figura 2: *Eigenvalues from PCA.*

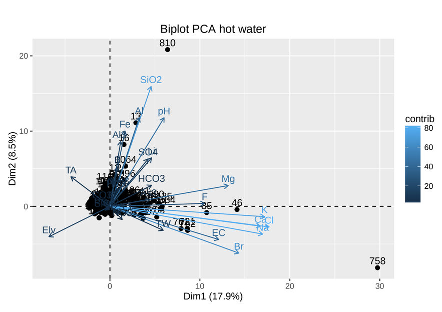

Figura 3: *Biplot from PCA.*

We identify three major clusters based on the dendogram: cluster 1 (pure hot water), which is divided into 1a volcanic system and 1b non-volcanic system, cluster 2 (mixing water), and cluster 3 (cold-meteoric water). Mixing processes occur in each system due to the high rainfall (Figure 4).

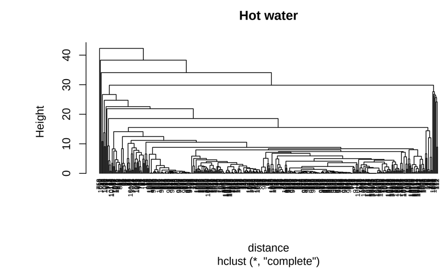

Figura 4: *Dendogram from cluster analysis.*

After evaluating the results we are not be able to draw a firm conclusion at this moment. We need to sort out the dataset based on the heterogeneity.

4. References

1. Agustin, A., Irawan, D.E., Susanto, A., Herdianita, N.R., 2015. Application of Geochemical Methods in Geothermal Exploration in Indonesia: a Literature Review (Part 1), PROCEEDINGS, 3rd IIGW ITB, Bandung, Indonesia.
2. ath, G. De', Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.
3. Bronto, S., 2006. Fasies gunung api dan aplikasinya. Indones. J. Geosci. 1, 59–71.
4. Chen, K., Jiao, J.J., Huang, J., Huang, R., 2007. Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China. Environ. Pollut. 147, 771–780.
5. Darmawan, I.G.B., Setijadji, L.D., Wintolo, D., 2015. Geology and Geothermal System in Rajabasa Volcano South Lampung Regency, Indonesia (Approach to Field Observations, Water Geochemistry and Magnetic Methods). Geology 19, 25.
6. Faraway, J.J., 2005. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press.
7. Fauzi, A., 2015. Geothermal resources and reserves in Indonesia: an updated revision. Geotherm. Energy Sci. 3, 1–6. doi:10.5194/gtes-3-1-2015
8. Fitzpatrick, M.L., Long, D.T., Pijanowski, B.C., 2007. Exploring the effects of urban and agricultural land use on surface water chemistry, across a regional watershed, using multivariate statistics. Appl. Geochem. 22, 1825–1840.
9. Gio, Prana Ugiana, Rosmaini, Elly, 2015. Belajar olah data dengan piranti lunak statistik. USU Press, Medan.
10. Gregor, J., Garrett, N., Gilpin, B., Randall, C., Saunders, D., 2002. Use of classification and regression tree (CART) analysis with chemical faecal indicators to determine sources of contamination. N. Z. J. Mar. Freshw. Res. 36, 387–398.
11. Irawan, D.E.., Puradimaja, D.J.., Notosiswoyo, S.., Soemintadiredja, P.., 2009. Hydrogeochemistry of volcanic hydrogeology based on cluster analysis of Mount Ciremai, West Java, Indonesia. J. Hydrol. 376, 221–234.
12. King, A.C., Raiber, M., Cox, M.E., 2014. Multivariate statistical analysis of hydrochemical data to assess alluvial aquifer–stream connectivity during drought and flood: Cressbrook Creek, southeast Queensland, Australia. Hydrogeol. J. 22, 481–500. doi:10.1007/s10040-013-1057-1
13. Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.
14. Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis, in: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California. pp. 1–14.
15. Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1, 14–23.
16. Manalu, P., 1988. Geothermal development in Indonesia. Geothermics 17, 415 – 420. doi:http://dx.doi.org/10.1016/0375-6505(88)90070-3
17. Ritzi, R.W., Wright, S.L., Mann, B., Chen, M., 1993. Analysis of temporal variability in hydrogeochemical data used for multivariate analyses. Groundwater 31, 221–229.
18. Rothwell, J.J., Futter, M.N., Dise, N.B., 2008. A classification and regression tree model of controls on dissolved inorganic nitrogen leaching from European forests. Environ. Pollut. 156, 544–552.
19. Seyhan, E., Griend, A.A.V.D., Engelen, G.B., 1985. Multivariate Analysis and Interpretation of the Hydrochemistry of a Dolomitic Reef Aquifer, Northern Italy. Water Resour. Res. 21, 1010–1024. doi:http://dx.doi.org/10.1029/wr021i007p01010
20. Smillie, A., Satar, S., Saptadji, N., Aminzadeh, F., Setianingsih, R., n.d. Capacity Building in the Geothermal Sector in Indonesia, a Unique Collaboration.