Application of Cluster Analysis and Principal Component Analysis for Assessment of Groundwater Quality—A Study in Semarang, Central Java, Indonesia

T R N Amanah¹, T T Putranto² and M Helmi³,⁴

¹Master Program of Environmental Sciences, School of Postgraduate Studies, Diponegoro University, Jl.Prof.H.Soedarto, S.H.Tembalang, Tembalang, Kota Semarang, Jawa Tengah 50275 - Indonesia
²Department of geological Engineering, Faculty of Engineering, Diponegoro University, Jl.Prof.H.Soedarto, S.H.Tembalang, Tembalang, Kota Semarang, Jawa Tengah 50275 - Indonesia
³Oceanography Department, Faculty of Fisheries and Marine Science, Diponegoro University, Jl.Prof.H.Soedarto, S.H.Tembalang, Tembalang, Kota Semarang, Jawa Tengah 50275 - Indonesia
⁴Center for Coastal Rehabilitation and Mitigation Studies (CoREM), Diponegoro University, Jl.Prof.H.Soedarto, S.H.Tembalang, Tembalang, Kota Semarang, Jawa Tengah 50275 – Indonesia

¹triretnaning.nuramanah@gmail.com
²putranto@ft.undip.ac.id

Abstract. The research study is located in Semarang with 10 samples of ground water gathered from dug wells. Cluster analysis and principal component analysis of ground water data is used to find water quality. The findings show lead concentration, cadmium concentration, and TDS in the water have surpassed the standard by the rules of Indonesian Health Ministry Year 2010. Those surpassing concentrations make water not suitable to consume because it will have impact on health, such as: inducting injuries through oxidative stress, epigenetic changes in DNA expression, central nerve, hematopoietic, incurable liver and kidney failures. Cluster analysis shows 7 locations are included into first cluster meanwhile three other locations are included into second, third, and fourth. From factorial analysis of four main factors, such as manganese, nitrate, zinc, and cadmium. Those factors are assumed to pollute because anthropogenic factors, such as: industrial, domestic, and agriculture waste causing lowering water quality.

Keywords: ground water quality, cluster analysis, principal component analysis.
1. Introduction

Water is a vital natural source for living creature existence because it plays an important role. Physical and chemical parameters of water are important to classify and assess water quality. To preserve water source is an obligation for each individual [1]. Utilizing waters for various purposes makes quantity and quality of water equally important [2, 3]. For recent years, the needs of domestic, industrial, and agriculture make water demand increasing in some regions. The use of pesticide fertilizer, pollution flow from highland to lower land, unplanned urbanization, quick growing population, and increasing industrialization make water quality lower and becomes a problem of water source [4]. For growing countries, village exodus makes demography of a quicker society growth in the city. Drainage system and changing topography due to unmanaged society growth and urbanization influence the quality and quantity of ground water [5].

This research is done in Semarang, Central Java, Indonesia. Ground water is very important in this studied area because it is used to drink, irrigate, and support industrial purposes. Hydro-chemical data of the system is used to determine main factors and the mechanism of ground water chemical control in the regions. As the time goes by with advancement, the changes of water quality can be understood by using ground water geo-chemical study. The contamination of other metal in water has impact on health so it makes the water unsuitable to consume, based on the rule of Indonesian Health Ministry number 2010. This research is done to identify hydro-geo chemical factors considered to take responsibilities in decreasing water quality using statistic multivariate concept through cluster analysis and principal component analysis.

2. Methods

Semarang, the studied era, is 373.70 km2, geographically located in 6°50’ – 7°10’ South Latitude and 109°35’ – 110°50’ East Longitude. Topographically, it is located above hill area, lower land area and beach area. The taken water samples were done in 2017 in the studied area by gathering from 10 dug wells. (Figure 1). This water sample is tested using twelve parameters: TDS, pH, Fe, CaCO3, Mn, NO3, NO2, Cd, Cr, Zn, SO4, and Pb. To determine the quality of the water and hydro-geo chemical in past occasion is done by using various statistical analysis method [6, 7, 8, 9]. Correlational analysis, hierarchical cluster analysis, and principal component analysis are multivariate statistical analysis done in this research using SPSS 22.

Figure 1. Map of the study area presenting the sampling location points.
Sample determination into different hydro-chemic group uses cluster analysis. Classifying water has been successfully done by some researches using this technique [10, 11, 12, 13, 14]. Grouping samples is done by observing the similarities of one to another and also the different parameter comparisons. Q-mode classification is a sample classification based on the parameters. Classifying samples into different hydro-chemic group in this research uses Q mode hierarchical cluster analysis (HCA). Similarity measurement is done by measuring the distance of straight lines between two spots in c-dimension space determined by c-variable or called as Euclidean distance. Ward method is done to find out group with similarities in each member compared to other members outside of the group [15].

The determination of water quality is done by observing the content of chemical compounds of the water by using correlation analysis as statistical stool [16, 17, 18, 19]. Correlational analysis is only a measurement of how well a certain variable predicting another variable [20].

Multivariate analysis used to get the factors explaining variants inside of the observed data in which subset variable is not correlated is called by factorial analysis [21, 22]. Factorial analysis is used to reduce greater data by gaining the factors explaining the data without reducing the information by determining variants among variables. The numbers of factors resulted show the total of possibility of variation source inside data in which later is ordered based on reliability. The highest Eigen vector score in this factor is the most important variation source in the data meanwhile factors having lowest Eigen vector skill is the most useless process contributing to chemical compound variables. PCA is used to identify many data patterns from some parameters. Factorial analysis using principal component analysis method is used to find out the causes of the changes of water quality by evaluating water quality using various tested parameters [23, 24, 25]. Principal component analysis and factorial analysis are done using Kaiser Normalization and Varimax Rotation [26].

### 3. Results and discussions

From table 1, it can be said that the concentration of TDS varies from 370 until 1680 with average 787.5. It shows the solids are from organic and inorganic compounds dissolved into water in some sample sport, surpassing the standard quality of drinking water with maximum limit 500. The highest TDS is from natural factors and waste, both domestic or industries. Score ground water pH is about 6.34 until 8.13, with average 7.3, showing neutral in nature. Iron concentration in studied area is about 0.0055 mg/L until 0.5909 mg/L, under the determined limitation.

![Table 1. Quantitative chemical analysis results](image-url)
Calcium carbonate concentration is between 89.17 mg/L until 108.75 mg/L, with average 102.62 mg/L. Manganese concentration in water is between 0.0032 mg/L until 1.5351 mg/L, with average 0.2749 mg/L. Nitrate concentration in water is between 0.1196 mg/L until 3.5535 mg/L, with average 0.5257 mg/L. Nitrite concentration is between 0.0014 mg/L until 1.0891 mg/L, with average 0.15761 mg/L. Cadmium (Cd) is poisonous metal, with lower dosage compared to other metals, this metal can disturb some biological systems [27, 28, 29]. The allowed maximum limitation based on the rule of Indonesian Health Ministry year 2010 is 0.003 mg/L. Meanwhile, the concentration of Cadmium is between 0.0001 mg/L until 1.0891 mg/L, with average 0.00442 mg/L, so it is said the concentration of Cadmium surpassing the limitation.

Cadmium inside of body will have impact on health, such as oxidative stress causing inducing injuries of the tissues [30, 31, 32], epigenetic changes in DNA expression [33, 34, 35], hindrance or increase of regulation of transport vessels, especially in S1 kidney tubules proximal segment [36]. Chromium concentration in water is 0.0001 mg/L. Zinc concentration is about 0.0031 mg/L until 0.0504 mg/L, with average 0.05187 mg/L. Sulfate concentration is about 12.460 mg/L until 94.584 mg/L, with average 43.9988 mg/L. Lead concentration in the water is 0.0376 mg/L until 0.29 mg/L, with average 0.0842 mg/L. The maximum limitation of lead in water is about 0.01 mg/L. Toxicity of lead in body is very dangerous with potency causing incurable health problem effects. It is known to disturb some body functions, especially central nerve, hematopoietic, liver and kidney systems [37].

**Table 2.** Correlation coefficient matrix

|        | TDS  | PH   | Fe    | CaCO3 | Mn    | NO3  | NO2  | Cd   | Zn   | So4  | Pb   |
|--------|------|------|-------|-------|-------|------|------|------|------|------|------|
| TDS    |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| 1.00 | -0.35| 0.53  | -0.27 | 0.35  | -0.40| -0.49| -0.12| -0.30| -0.19| 0.19 |
| Sig.   | 0.33 | 0.11 | 0.46  | 0.32  | 0.25  | 0.15 | 0.74 | 0.40 | 0.60 | 0.59 |      |
| PH     |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| -0.35| 1.00 | -0.04 | 0.32  | -0.05 | 0.34 | 0.34 | 0.33 | 0.42 | 0.04 | -0.17|
| Sig.   | 0.33 | 0.91 | 0.36  | 0.90  | 0.34  | 0.34 | 0.35 | 0.22 | 0.91 | 0.65 |      |
| Fe     |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| 0.53 | -0.04| 1.00  | -0.24 | 0.853*| -0.24| -0.26| 0.14 | 0.03 | 0.71 | 0.671*|
| Sig.   | 0.11 | 0.91 | 0.50  | 0.00  | 0.51  | 0.46 | 0.71 | 0.93 | 0.03 | 0.02 |      |
| CaCO3  |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| -0.27| 0.32 | -0.24 | 1.00  | 0.01  | 0.32 | 0.39 | 0.34 | -0.03| -0.23| -0.04|
| Sig.   | 0.46 | 0.36 | 0.50  | 0.97  | 0.36  | 0.26 | 0.34 | 0.92 | 0.53 | 0.91 |      |
| Mn     |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| 0.35 | -0.05 | 0.853*| 0.01  | 1.00  | -0.08| -0.09| 0.02 | 0.33 | -0.52| 0.906*|
| Sig.   | 0.32 | 0.90 | 0.00  | 0.97  | 0.83  | 0.80 | 0.96 | 0.35 | 0.13 | 0.00 |      |
| NO3    |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| -0.40| 0.34 | -0.24 | 0.32  | -0.08 | 1.00 | 0.979**| 0.25 | -0.27| 0.21 | 0.02 |
| Sig.   | 0.25 | 0.34 | 0.51  | 0.36  | 0.83  | 0.00 | 0.49 | 0.46 | 0.56 | 0.97 |      |
| NO2    |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| -0.49| 0.34 | -0.26 | 0.39  | -0.09 | 0.979**| 1.00 | -0.24| -0.18| 0.11 | -0.03|
| Sig.   | 0.15 | 0.34 | 0.46  | 0.26  | 0.80  | 0.00 | 0.50 | 0.61 | 0.76 | 0.93 |      |
| Cd     |      |      |       |       |       |      |      |      |      |      |      |
| Pearson| -0.12| -0.33| -0.14 | 0.34  | -0.02 | -0.25| -0.24| 1.00 | 0.06 | 0.04 | 0.22 |
| Sig.   | 0.74 | 0.35 | 0.71  | 0.34  | 0.96  | 0.49 | 0.50 | 0.88 | 0.92 | 0.54 |      |
From table 2, it can be said that the correlation: nitrate and nitrite, iron and manganese, iron and sulfate, iron and lead, and manganese and lead. There are four positive correlation: nitrate and nitrite, iron and manganese, iron and lead, manganese and lead. One negative correlation is iron and sulfate. The correlation among nitrate and nitrite, iron and manganese, and manganese and lead are strong correlation. The correlation among iron and lead, iron and sulfate, are strong correlation. Correlational analysis among the parameters can be seen in Figure 2.

![Figure 2. Correlation Analysis Graphs Presenting correlation](image-url)
Cluster is grouped into 4 to be more specific, gained from 7 locations (SG1, SG2, SG4, SG5, SG6, SG7, SG8) of first samples and 3 other locations included in second (SG10), third (SG9), and fourth (SG3) (Figure 3).

![Figure 3. Dendogram hierarchiesl cluster](image)

Varimax rotated factor loadings is presented in Table 3. Four factors having greater Eigen score than 1 and rotation is solved in six iteration. Figure 4 presents Scatter plot resulted based on Eigen scores of each component.

|                | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|----------------|----------|----------|----------|----------|
| TDS            | 0.373    | -0.504   | -0.328   | -0.342   |
| PH             | 0.000    | 0.492    | 0.671    | -0.160   |
| Fe             | 0.926    | -0.224   | 0.045    | -0.239   |
| CaCO3          | 0.071    | 0.474    | 0.206    | 0.668    |
| Mn             | 0.950    | 0.015    | -0.188   | -0.016   |
| NO3            | -0.074   | 0.954    | -0.126   | -0.069   |
| NO2            | -0.075   | 0.960    | -0.026   | -0.003   |
| Cd             | -0.027   | -0.271   | -0.160   | 0.895    |
| Zn             | -0.154   | -0.236   | 0.895    | 0.063    |
| So4            | -0.721   | 0.132    | -0.419   | -0.123   |
| Pb             | 0.841    | 0.049    | -0.343   | 0.154    |
| Eigen Value    | 3.61     | 2.42     | 1.68     | 1.48     |
| % Variance     | 32.78    | 21.99    | 15.32    | 13.42    |
| Cumulative % variance | 32.78 | 54.77 | 70.09 | 83.51 |
Factor 1 explains 32.78% of total variants, pH concentration and calcium carbonate with the lowest loading, iron and lead with strong positive loading. It shows the contamination source of the metal may be from domestic and industrial wastes. Factor 2 explains 21.99% from all total variant with manganese, sulfate, and lead with sufficiently lower loading so they do not contribute significantly. Calcium Carbonate pH have less strong positive loading but Nitrite and Nitrate have strong positive loading. It shows the source of contamination may be from agriculture and domestic wastes. Factor 3 explains 15.32% of total variants with iron and calcium carbonate concentrations with lower positive loading, slightly contributing. Zinc concentration and pH have strong positive loading with high contribution. Factor 4 explains 13.42% of total variants with cadmium concentration and Calcium Carbonate with high contribution so it can be assumed the pollutants are from anthropogenic such as from industrial wastes.

4. Conclusion

There are three strong positive correlation: nitrate and nitrite, iron and manganese, and manganese and lead. One strong positive correlation is iron and lead meanwhile the negative and strong one is iron and sulfate. Lead is dangerous for body because it will negatively and incurably affect health: disturbing kidney, liver, hematopoietic, and central nerve problems. The concentration of lead in this research surpasses the determined limitation, 0.0842 mg/L so the concentration of lead surpasses the limitation of drinking water. Besides that, TDS and cadmium also surpass the limitation, 787.5 and 0.00442 mg/L. Cadmium will influence health by inducting injuries through oxidative stress, epigenetic changes in DNA expression, and kidney problems. Other parameters are found to be in line with the limitation of drinking water. Cluster analysis groups 10 locations into 4 clusters. It is also gained 7 locations are included in first cluster meanwhile 3 other locations are grouped into second, third, and fourth. Factorial analysis result gains 4 main factors: manganese, nitrite, zinc, and cadmium. Those are assumed that anthropogenic factors, such as domestic, industrial, and agriculture wastes strongly contributing to polluted water as drinking water.
5. References

[1] Jothivenkatachalam K, Nithya A and Mohan S C 2010 Correlation analysis of drinking water quality in and around Perur block of Coimbatore district, Tamil Nadu, India Rasayan Journal 13. Chemistry 3(4) 649-654

[2] Schiavo M A, Havser S, Gusimano G and Gatto L 2006 Geochemical characterization of groundwater and sub-marine discharge in the southeastern Sicily Cont Shelf Res 26(7) 826–834

[3] Subramani T, Elango L and Dhamodarasamy SR 2005 Groundwater quality and its suitability for drinking and agricultural use in Chithar River Basin, Tamilnadu, India Environ Geo 47 1099–1110

[4] Joarder M A M, Raihan F, Alam J B and Hasanuzzaman S 2008 Int J Environ. Res 2(3) 291

[5] Vasanthavigar M, Srinivasamoorthy K, Vijayaragavan K, Rajiv G R, Chidambaram S, Sarama V S, Anandhan P, Manivannan R and Vasudevan S 2010 Application of water quality index for groundwater quality assessment: Thirumanimuttar Sub-Basin,Tamilnadu, India Environ Monit Assess 171(1–4) 595–609

[6] Nelson J D and Ward R D 1981 Statistical considerations and sampling techniques for ground—water quality monitoring Groundwater 19(6) 617–625

[7] Briz-Kishor BH, Murali G 1992 Factor analysis for revealing hydrochemical characteristics of a watershed Environ Geol Water Sci 19(1) 3–9

[8] Kumar A R and Riyazuddin P 2008 Application of chemometric techniques in the assessment of groundwater pollution in a suburban area of Chennai city, India Curr Sci 94(8) 235–242

[9] Routroy S, Harichandran R, Mohanty J K and Panda CR 2013 A statistical appraisal to hydrogeochemistry of fluoride contaminated ground water in Nayagarh District, Odisha J Geol Soc India 81 350–360

[10] Alther G A 1979 A simplified statistical sequence applied to routine water quality analysis: a case history Ground Water 17 556-561

[11] Williams R E 1982 Statistical identification of hydraulic connections between the surface of a mountain and internal mineralized sources Ground Water 20 466-478

[12] Farnham I M, Stetzenbach K J, Singh A K and Johannesson K H 2000 Deciphering groundwater flow systems in Oasis Valley, Nevada, Using trace element chemistry, multivariate statistics, and Geographical Information System Math Geol 32 943-968

[13] Alberto W D, Pilar D M D, Valeria A M, Fabiana P S, Cecilia H A and Angeles B M D L 2001 Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A Case Study: Suquía River Basin (Cordoba-Argentina) Water Res 35(12) 2881-2894

[14] Meng S X and Maynard J B 2001 Use of statistical analysis to formulate conceptual models of geochemical behavior: water chemical data from Butucatu aquifer in Sao Paulo State, Brazil J of Hydrology 250 78-97

[15] Güler C, Thyne G D, McCray J E and Turner A K 2002 Evaluation of graphical and multivariate statistical methods for classification of water chemistry data Hydrogeology Journal 10 455-474

[16] Box G F P, Hunter W G, Hunter J S 1978 Statistics for experiments.An introduction to design data analysis and model building (New York: Wiley) p 653

[17] Chapman D 1996 On the behalf of UNESCO, WHO. UXEP. Water quality assessments—a guide to use biota, sediments and water in environmental monitoring (London: F & F Spoil) Chapter 9

[18] Shrivastava V S and Patil P R 2002 Tapti river water pollution by industrial wastes: a statistical approach Nat Environ Pollut Technol 1(3) 279–283

[19] Zeng X and Rasmussen T C 2005 Multivariate statistical characterization of water quality in Lake Lanier, Georgia, USA J Environ Qual 34 1980–1991
[20] Kurumbein W C and Graybill F A 1965 *An introduction to statistical models in geology* (New York: McGraw-Hill)
[21] Anazawa K and Ohmori H 2005 The hydrochemistry of surface waters in Andesitic Volcanic area, Norikura volcano, central Japan *Chemosphere* **59** 605-615
[22] Brown C E 1998 *Applied Multivariate Statistics in Geohydrology and Related Sciences* (New York: Springer)
[23] Helena B, Pardo R, Vega M, Barrado E, Fernandez J M and Fernandez L 2000 Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis *Water Res* **34**(3) 807–816
[24] Reghunath R, Murthy T R S and Raghavan B R 2002 The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India *Water Res* **36** 2437–2442
[25] Yidana S M 2010 Groundwater classification using multivariate statistical methods: Southern Ghana *J Afr Earth Sci* **57** 455–469
[26] Kaiser H F 1958 The varimax criterion for analytic rotation in factor analysis *Psychometrika* **23** 187–200
[27] Nordberg G, Nogawa K, Nordberg M and Friberg L 2007 Cadmium. In: *Handbook on toxicology of metals* ed Nordberg G et al (New York: Academic Press) pp. 65-78
[28] Järup L, Berglund M, Elinder C G, Nordberg G and Vahter M 1998 Health effects of cadmium exposure - a review of the literature and a risk estimate *Scand J Work Environ Health* **24** (Suppl 1) I-51
[29] Bernard A 2004 Dysfunction induced by cadmium: Biomarkers of critical effects *Biometals* **17** 519-523
[30] Matović V, Buha A, Bulat Z, and Dukić-Cosić D 2011 Cadmium toxicity revisited: focus on oxidative stress induction and interactions with zinc and magnesium *Arhiv za Higijenu Rada i Toksikologiju* **62**(1) 65–76
[31] Patra R C, Rautray A K and Swarup D 2011 Oxidative stress in lead and cadmium toxicity and its melioration *Veterinary Medicine International* **2011** Article ID 457327
[32] Cuypers, Plusquin M, Remans T, Jozefczak M, Keunen E, Gielen H, Opdenakker K, Nair A M, Munters E, Artois T J, Vangronsveld J and Smeets K 2010 Cadmium stress: an oxidative challenge *BioMetals* **23**(5) 927–940
[33] Wang B, Shao C, Li Y, Tan Y and Cai L 2012 Cadmium and its epigenetic effects *Current Medicinal Chemistry* **19**(16) 2611–2620
[34] Martinez-Zamudio R and Ha H C 2011 Environmental epigenetics in metal exposure *Epigenetics* **6**(7) 820–827
[35] Luparello C, Sirchia R, and Longo A 2011 Cadmium as a transcriptional modulator in human cells *Critical Reviews in Toxicology* **41**(1) 75–82
[36] Vesey D A 2010 Transport pathways for cadmium in the intestine and kidney proximal tubule: focus on the interaction with essential metals *Toxicology Letters* **198**(1) 13–19
[37] Kalia K and Flora S J 2005 Strategies for safe and effective therapeutic measures for chronic arsenic and lead poisoning *J Occup Health* **47** 1–21

**Acknowledgements**

This article is presented at the International Conference on Smart City Innovation 2018 that supported by the United States Agency for International Development (USAID) through the Sustainable Higher Education Research Alliance (SHERA) Program for Universitas Indonesia’s Scientific Modeling, Application, Research and Training for City-centered Innovation and Technology (SMART CITY) Project, Grant #AID-497-A-1600004, Sub Grant #IE-00000078/UI-1