Data analysis on sea water quality data in Jakarta Bay using Principal Components Analysis (PCA) method during transitional monsoon 2012

A Martina and I M Radjawane
Oceanography research group, Bandung Institute of Technology
Jalan Ganesha 10, Bandung 40132, Indonesia
E-mail: annisa.martina@oceanography.itb.ac.id

Abstract. To get a conclusion from a data matrix consisting of 3 individuals and 2 variables is relatively easy. However, it is very difficult to understand the large amount of data. Therefore, it requires data analysis methods for an easier representation. Based on sea water quality data in Jakarta Bay from BPLHD DKI Jakarta (Jakarta Environmental Management Board), there are 24 biological, physical, and chemical parameters in 23 stations. Based on the quality and quantitative of data, we use only one set data on October 2012 as representative of the Second Transition monsoon. Analysing was conducted for 10 parameters namely turbidity, total suspended solid (TSS), temperature, pH, salinity, dissolved oxygen (DO), biological oxygen demand (BOD), methylene blue active substances, phenol, and zinc (Zn) at 23 stations. Consequently, in this paper, we get a conclusion from the data using principal component analysis (PCA) method for its application in data analysis. The method of PCA is used to analyse the data matrix from a similarity point of view between stations and correlation between parameters. The result of PCA is four principal components i.e. PC 1 (27.73% of the variance) is mainly related to TSS, temperature, salinity, and DO. PC 2 (16.33% of the variance) is mainly related to BOD. PC 3 (12.39% of variance) is mainly related to MBAS, phenol, and zinc. PC 4 explains 11.09% of variances related mainly to turbidity.

1. Introduction
Jakarta Bay is semi enclosed bay that located facing to capitol city of Indonesia, Jakarta Greater Area at the north part of Java Island. This bay is characterized by shallow water with 13 rivers system pour out to this bay and connected with the Java Sea [1]. Jakarta as one of the fast growing Asian mega-cities. The polluted water around this bay has long history evidences that has been investigated by many of researchers. Suwandana et al. summarized from several researchers the impact of rapid development and increase of population introduce some environmental issues such as urban waste, air and river water pollution, coastal pollution, and drink water shortage [2]. In recent years, the red tides, fish mortality and oxygen depleted water mass been frequently occured in Jakarta Bay [3]. The latest issue is related with the microplastic marine debris [4].

To determine the level of pollution in Jakarta Bay there are numerous water quality indicators i.e physical, chemical and biological parameters. Temperature, turbidity and total suspended sediment (TSS) can be classified as physical indicacors. Salinity, pH, Biological Oxygen Demand (BOD), Dissolved Oxygen and type of metal parameter can be classified as chemical indicators. The number of biological parameters is less and the common measure is Estherichia coli and fecal coliform.

The polluted coastal water and its process can be evaluated from several water quality method using observation [2,5], sattelite as well as numerical simulation [1,3]. In this research we will applied...
the principal component analysis (PCA) method as one of statistical technique to reduced a large set of parameters (variables) to a smaller new independent set data but still can get the information from large dataset. This method can be performed in data matrix from a similarity point of view between stations and parameters. So we can analyze the contribution of specific parameters in selected stations.

2. Materials and Method
In this paper, we use the water quality data in Jakarta Bay provided from BPLHD Jakarta Province (Regional Environmental Management Agency) [6] to analysis the data. Every year BPLHD conducts the observation and records sea water quality data in Jakarta Bay and estuary surrounding the bay. There are 2 or 3 times observations in a year to cover the dry, transition, and rainy season. For example, in 2012 there were 3 observations: July to represent dry season, October to represent the second transition season, and December to represent rainy season. We have to select the data based on the quality and quantity of the data. In this study, we found the best dataset in October 2012 that represents the second transition monsoon. The monitoring indicators include 3 physical, 19 chemical and 2 biological parameters respectively. Water sampling observation was taken in fixed 23 sampling stations at Jakarta Bay namely A1-A7, B1-B7, C2-C5, and D3-D6. All of the sampling stations locations can be seen on Figure 1 and the longitude and latitude position shown in Table 1.

![Figure 1. Study sites of with 23 sampling stations at Jakarta Bay (A1-A7, B1-B7, C2-C6, and D3-D6) (Source: BPLHD DKI Jakarta, 2011)]](image-url)
Table 1. Study sites of Water Quality at Jakarta Bay in longitude and latitude

| No | Station | Position |
|----|---------|----------|
| 1  | A1      | 106°42'20"E 05°59'40"S |
| 2  | A2      | 106°44'50"E 05°59'00"S |
| 3  | A3      | 106°47'20"E 05°58'20"S |
| 4  | A4      | 106°50'00"E 05°57'50"S |
| 5  | A5      | 106°52'40"E 05°57'10"S |
| 6  | A6      | 106°55'20"E 05°56'30"S |
| 7  | A7      | 106°58'00"E 05°56'00"S |
| 8  | B1      | 106°42'50"E 06°02'00"S |
| 9  | B2      | 106°45'30"E 06°01'30"S |
| 10 | B3      | 106°48'00"E 06°01'00"S |
| 11 | B4      | 106°50'40"E 06°00'20"S |
| 12 | B5      | 106°53'20"E 05°59'40"S |
| 13 | B6      | 106°56'00"E 05°59'00"S |
| 14 | B7      | 106°58'40"E 05°58'30"S |
| 15 | C2      | 106°46'10"E 06°04'10"S |
| 16 | C3      | 106°48'50"E 06°03'30"S |
| 17 | C4      | 106°51'20"E 06°02'50"S |
| 18 | C5      | 106°54'00"E 06°02'10"S |
| 19 | C6      | 106°56'40"E 06°01'40"S |
| 20 | D3      | 106°49'30"E 06°05'50"S |
| 21 | D4      | 106°52'00"E 06°05'20"S |
| 22 | D5      | 106°54'40"E 06°04'40"S |
| 23 | D6      | 106°57'20"E 06°04'00"S |

(Source: BPLHD DKI Jakarta, 2011)

Table 2. Parameters selected for using PCA method

| No | Parameter                               | Unit |
|----|-----------------------------------------|------|
| 1  | Turbidity (Turb)                         | NTU  |
| 2  | Total Suspended Solid (TSS)              | mg/L |
| 3  | Temperature (Temp)                       | °C   |
| 4  | pH                                      |      |
| 5  | Salinity (Sal)                           | °/oo |
| 6  | Dissolved Oxygen (DO)                    | mg/L |
| 7  | Biological Oxygen Demand (BOD)           | mg/L |
| 8  | Methylene blue active substances (MBAS)  | mg/L |
| 9  | Phenol (Phe)                             | mg/L |
| 10 | Zinc (Zn)                                | mg/L |
For the case of October 2012, from 23 parameters, only 3 physical parameters and 7 chemical parameters can be analyzed. We eliminated a lot of not available data that could not be analyzed. So, we selected total of 10 parameters collected from 23 sampling stations to be analyzed. The selected 10 parameters can be seen on Table 2.

The water sample will be analyzed by principal component analysis (PCA) method to reduce the variables (parameters) from 10 variables to 4 new variables (4 principal components). The new variables can represent the all original variable as water quality indicators, or we can say that is principal indicators. All the procedures were performed using MATLAB R2014a software for calculation implementation of PCA. In additional we compare the each PC component with water quality standards number 51/2004 published by the Minister of Environment.

3. Results and Discussion

The first step before doing the PCA method is standardizing data, because there are unit differences between one parameter with the another parameter, then calculates covariance from the standardized data, as seen in Table 3.

Table 3. Covariance of standardized data

|       | Turb  | TSS   | Temp  | pH    | Sal   | DO    | BOD   | MBAS  | Phe   | Zn    |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Turb  | 1.000 | 0.655 | 0.125 | 0.260 | -0.218| 0.126 | -0.047| 0.098 | 0.040 | -0.100|
| TSS   | 0.655 | 1.000 | 0.262 | 0.130 | -0.498| 0.360 | -0.101| 0.274 | -0.052| 0.107 |
| Temp  | 0.125 | 0.262 | 1.000 | 0.470 | -0.515| 0.330 | 0.352 | 0.133 | -0.101| -0.013|
| pH    | 0.260 | 0.130 | 0.470 | 1.000 | -0.118| 0.365 | 0.227 | -0.048| -0.240| 0.007 |
| Sal   | -0.218| -0.498| -0.515| -0.118| 1.000 | -0.444| 0.003 | -0.195| -0.033| -0.022|
| DO    | 0.126 | 0.360 | 0.330 | 0.365 | -0.444| 1.000 | 0.161 | 0.231 | -0.083| 0.058 |
| BOD   | -0.047| -0.101| 0.352 | 0.227 | 0.003 | 0.161 | 1.000 | -0.232| -0.105| -0.104|
| MBAS  | 0.098 | 0.274 | 0.133 | -0.048| -0.195| 0.231 | -0.232| 1.000 | -0.198| -0.049|
| Phe   | 0.040 | -0.052| -0.101| -0.240| -0.033| -0.083| -0.105| -0.198| 1.000 | 0.164 |
| Zn    | -0.100| 0.107 | -0.013| 0.007 | -0.022| 0.058 | -0.104| -0.049| 0.164 | 1.000 |

In Table 3 we can see that highest covariance between parameters is the turbidity and total suspended solid (TSS) i.e. 0.655. The interpretation of this value that’s the both parameters still have correlation. Therefore, PCA can produce the new variables which do not have any correlation between the parameters, but still described situation from the original data.

\[
D = \begin{bmatrix}
2.77 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.63 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.23 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.11 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0.93 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.66 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.65 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.24 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.18 \\
\end{bmatrix}
\]

**Figure 2.** The decompositions matrix (D) calculated from original data set
The original data can be transformed to an uncorrelated variable (independent variable) using the eigenvector of the correlation matrix. Eigenvector can be determined by finding the eigenvalue first. Eigenvalue can be found with Eigen-value Decomposition (EVD) or Singular-value Decomposition (SVD) [9]. In this research we use EVD, and Figure 2 is decomposition matrix (D) that has been sorted from the large to the small value.

Based on decomposition matrix (Figure 2), there are 4 principal components having significant eigenvalue > 1 that were performed in the first to fourth matrix columns. Eigenvalue for PC 1, PC 2, PC 3 and PC 4 is 2.77, 1.63, 1.23, and 1.11 respectively, then variance and cumulative variance for the principal component can be seen in Table 4 and Figure 3.

**Table 4. Variance dan cumulative variance of principal component.**

| Principal component | Proportion of variance | Cumulative proportion |
|---------------------|------------------------|-----------------------|
| PC 1                | 27.73%                 | 27.73%                |
| PC 2                | 16.33%                 | 44.07%                |
| PC 3                | 12.39%                 | 56.45%                |
| PC 4                | 11.09%                 | 67.54%                |
| PC 5                | 9.31%                  | 76.85%                |
| PC 6                | 6.60%                  | 83.45%                |
| PC 7                | 6.45%                  | 89.90%                |
| PC 8                | 5.84%                  | 95.74%                |
| PC 9                | 2.44%                  | 98.17%                |
| PC 10               | 1.83%                  | 100.00%               |

**Figure 3.** Proportion of variance from principal component.

Based on Table 4, PC1, PC2, PC3, and PC4 can explain the diversity data around 67.54%. The loadings of the four principal components with PCA are shown in Table 5. PC 1 (27.73% of the variance) is mainly related to TSS, temperature, salinity, and DO. PC 2 (16.33% of the variance) is mainly related to BOD. PC 3 (12.39% of variance) is mainly related to MBAS, phenol, and zinc. PC 4 explains 11.09% of variances related mainly to turbidity.
Table 5. Loadings of 10 parameters on 10 principal components.

| Parameter          | PC 1     | PC 2     | PC 3     | PC 4     | PC 5     | PC 6     | PC 7     | PC 8     | PC 9     | PC 10    |
|--------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Turb               | 0.3275   | -0.2805  | 0.0121   | 0.6549   | 0.1184   | -0.0309  | 0.0500   | 0.1735   | -0.2450  | -0.5267  |
| TSS                | 0.4393   | -0.3737  | 0.0412   | 0.2079   | 0.0784   | -0.2625  | 0.2154   | -0.0851  | 0.4252   | 0.5594   |
| Temp               | 0.4183   | 0.3109   | 0.1099   | -0.1444  | -0.1913  | -0.2416  | -0.4774  | 0.2051   | 0.4820   | -0.3132  |
| pH                 | 0.3268   | 0.3805   | 0.003    | 0.1691   | 0.4601   | 0.4078   | -0.3629  | 0.0758   | -0.2314  | 0.3884   |
| Sal                | -0.4295  | 0.1279   | -0.1637  | 0.2319   | 0.3970   | 0.1352   | 0.1715   | 0.4416   | 0.5601   | -0.0681  |
| DO                 | 0.4073   | 0.0768   | 0.0498   | -0.3000  | 0.0306   | 0.5643   | 0.5282   | -0.1484  | 0.1975   | -0.2762  |
| BOD                | 0.1071   | 0.5864   | 0.1406   | 0.1191   | -0.1923  | -0.3413  | 0.5215   | 0.3250   | -0.2390  | 0.1502   |
| MBAS               | 0.2075   | -0.3086  | -0.4902  | -0.3969  | -0.0439  | 0.0142   | -0.0013  | 0.6388   | -0.2143  | 0.0879   |
| Phe                | -0.1140  | -0.2364  | 0.652    | 0.0728   | -0.3735  | 0.3869   | -0.0886  | 0.4161   | 0.0184   | 0.1739   |
| Zn                 | 0.0026   | -0.1516  | 0.5213   | -0.3971  | 0.6282   | -0.3200  | 0.0569   | 0.1045   | -0.1314  | -0.1385  |

The scores matrix provided information about the distribution of patterns or sources of contamination among samples, and the matrix of loadings defined the contribution of the original variables to each one of these contamination patterns or sources [10]. We choose loading value > 0.4 which can describe the variables that affected to sea water quality in Jakarta Bay. Therefore we can resume the result of PCA method as seen in Table 6:

Table 6. Summary of Principal Component Analysis (PCA)

| Principal Component (PC) | Parameter                                    | Loading | Proportion of variance |
|--------------------------|----------------------------------------------|---------|------------------------|
| PC 1                     | Total Suspended Solid (TSS)                   | 0.4393  | 27.73%                 |
| PC 1                     | Temperature                                  | 0.4183  |                        |
| PC 1                     | Salinity                                     | -0.4295 |                        |
| PC 2                     | Dissolved Oxygen (DO)                         | 0.4073  |                        |
| PC 2                     | Biological Oxygen Demand (BOD)                | 0.5864  | 16.33%                 |
| PC 3                     | Methylen Blue Active Substances (MBAS)        | -0.4902 |                        |
| PC 4                     | Phenol                                       | 0.652   | 12.39%                 |
| PC 4                     | Zinc                                         | 0.5213  |                        |
| PC 4                     | Turbidity                                    | 0.6549  | 11.09%                 |

The principal components score plot can describe their different characteristics and find out the correlation between different parameters by the parameters lines [10], as seen in Figure 4. The red dots denote the station number and the blue lines denote the parameters. All ten parameters are represented by a vector, and they can describe for the contribution to the two principal components. For example, in Figure 4, the PC 1 has positive coefficients for eight parameters and has negative coefficient for two parameters (Salinity and Phenol). The highest positive coefficients in the PC 1 are TSS, temperature, and DO, and the highest negative coefficient is Salinity. This interpretation is corresponding to value in Table 5.
If two-parameter lines get closer together, the correlation between them will be strong. For example, in Figure 4, TSS and turbidity was overlap, consequently to see more clearly, we try to create score plot in three dimensions as seen in Figure 5.

![Figure 4](image1.png)

*Figure 4*. The loadings of variables and scores of the 23 stations for the PC 1, and PC 2, respectively.

We can see that TSS and turbidity lines in Figure 5 are not overlap, but they just close together. Therefore TSS and turbidity have the highest positive correlation coefficient (0.655). If the angle between two parameter lines is about 90°, then the interpretation is there is no correlation between them, for example temperature and salinity (-0.515).

![Figure 5](image2.png)

*Figure 5*. The loadings of variables and scores of the 23 stations for the PC 1, PC 2, and PC 3, respectively.
From the result above, then first, we can see the important of PC 1 (TSS, temperature, salinity and DO) affected the water quality of Jakarta Bay. Figure 6 shows the value of TSS, salinity, and DO at all stations locations still in accordance with quality standards for biota and marine tourism, but temperature at nine station locations are not available for coral and seagrass (temperature value > 30°C). The TSS value in general is less than 5 mg/L, except for B1, B2, C2 and D3 stations (Figure 6a). Salinity is below the quality standards for all stations (Figure 6c). The range of DO is between 6-8 mg/L as seen in Figure 6d.

Figure 6. a. TSS (top left), b. temperature (top right), c. salinity (bottom left), and d. DO (bottom right) distribution at 23 stations in Jakarta Bay in October 2012

The second, we can see the important of PC 2 (BOD) affected the water quality of Jakarta Bay. BOD at all station locations still in accordance with quality standards for marine biota, but BOD at C3 is not fulfill to marine tourism standard quality with the BOD value > 10 mg/L (Figure 7).

Figure 7. BOD distributions at 23 stations in Jakarta Bay (October 2012)
The third, we can see the important of PC 3 (MBAS, Phenol, and Zinc) affected the water quality of Jakarta Bay. In Figure 8 the MBAS value in all station locations are to marine biota standard quality, but not available to marine tourism, the MBAS value > 0.001 mg/L. The phenol value > 0 mg/L at all station locations, it means the phenol value at all station locations are not available to marine tourism and biota. The zinc value at all location > 0.005 mg/L, it means the zinc values at all locations are not available to marine biota, but only 9 stations from 23 stations are not available to marine tourism.

Figure 8. a. MBAS (top left), b. Phenol (top right), and c. Zinc (bottom) at 23 station in Jakarta Bay (October 2012)

The fourth, we can see the important of PC 4 (turbidity) affected the water quality of Jakarta Bay (Figure 9). Turbidity value at B7, C4, and D3 stations > 5 NTU, it means the turbidity value in three locations not meet to marine tourism and biota standard quality value.

Figure 9. Turbidity distribution at 23 stations in Jakarta Bay (October 2012)
4. Conclusions
Using principal component analysis (PCA) method, from 10 parameters (the original variables) in data set can be reduced to 4 new variables which affect the sea water quality in Jakarta Bay at October 2012 (the second transition season). The four new variables can explain 67.54% from the total variance; it means the sea water quality can be explained with 4 new variables. The four new variables consist of PC 1 (TSS, temperature, salinity and DO), PC 2 (BOD), PC 3 (MBAS, Phenol, and Zinc), and PC 4 (turbidity).

Acknowledgement
Part of this research was funded by the Oceanography Research Group P3MI Grant FY 2018. We would like to thank Mr. Arlanda Sofyan for helping us arranged the data and BPLDH DKI Jakarta that provided and allowed to use the water quality data.

References
[1] Radjawane I M and Riandini F 2010 Numerical simulation of cohesive sediment transport in Jakarta Bay. International Journal of Remote Sensing and Earth Sciences, Vol 6. Doi: 10.30536/j.ijijreses, 2009. v6.a1240.
[2] Suwandana, E, Kawamura K, Tanaka K, Sakuno Y and Rahardjo P 2011 Evaluating spatio-seasonal pattern of river- and groundwater quality in the city of Jakarta, Indonesia using pollution index. Journal of Japanese Agricultural Systems Society 27(3) 91-102
[3] Nurdjaman S, Radjawane I M and Jamelina S 2014 Numerical simulation of dissolved oxygen in Jakarta Bay. 2014. AIP Conference Proceedings 1589, 369, doi: 10.1063/1.4868820.
[4] Manalu A A, Haryadi S, Wardiatno Y 2017 Microplastics abundance in coastal sediments of Jakarta Bay Indonesia AACL Bioflux 10(5) 1164-1173.
[5] Damar, Ario. 2003. Effects of enrichment on nutrient dynamics, phytoplankton dynamics and productivity in Indonesian tropical waters: a comparison between Jakarta Bay, Lampung Bay and Semangka Bay. Germany: Forschungs-und Technologiezentrum Westküste
[6] BPLHD Jakarta Province 2012 Laporan Hasil Uji Kualitas Air Laut (Jakarta: BPLHD)
[7] Emery, W. J. and R. E. Thomson. 2001. Data Analysis Methods in Physical Oceanography. 2nd and revised edition (Amsterdam: Elsevier sciences) p 654
[8] Bengen, D. G. 2000. Teknik Pengambilan Contoh dan Analisis Data Biofisik Sumberdaya Pesisir (Bogor: Institut Pertanian Bogor)
[9] Anton, Howard and Chris Rorres. 2010. Elementary Linear Algebra. Applications Version. 10th edition (Denver: John Wiley & Sons, Inc.) p 592
[10] Wu M L, Wang Y S, Dong J D, Sun C C, Wang Y T, Sun F L and Cheng H 2011 Investigation of Spatial and Temporal Trends in Water Quality in Daya Bay, South China Sea. International Journal of Environmental Research and Public Health, Vol 8. Doi: 10.3390/ijerph8062352