International Journal of Engineering & Technology, 7 (3.14) (2018) 115-122

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Spatial Assessment and the Most Significant Parameters for Drinking Water Quality Using Chemometric Technique: A Case Study at Malaysia Water Treatment Plants

H. M. Zolkipli[1], H. Juahir[1,2*], G. Adiana[1*], N. Zainuddin[3], A. Ismail[1], A. B. H. M. Maliki[4], N.I. Hussain[1], M. K. A. Kamarudin[1,4], M. E. Toriman[5], M. Mokhtar[6]

[1]*East Coast Environmental Research Institute, Universiti Sultan ZainalAbidin, Gong BadakCampus, 21300 Kuala Nerus, Terengganu, Malaysia*

[2]*Faculty of Bioresource and Food Industry, Universiti Sultan ZainalAbidin, Besut Campus, 22200 Besut, Terengganu, Malaysia*

[3]*Ministry of Health, Engineering Services Division Department, Aras 3-7, Block E3, Kompleks E, Presint 1, Federal Government Administrative Centre, 62590 Putrajaya, Malaysia*

[4]*Faculty of Applied Society Science, Universiti Sultan ZainalAbidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia*

[5]*Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*

[6]*Institute for Environment and Development (LESTARI) & Former Deputy Vice-Chancellor, Research and Innovation Affairs, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.*

**Corresponding author E-mail: hafizanjuahir@unisza.edu.my, adiana.ghazali@gmail.com*

Abstract

The objectives of this study are to determine the most significant spatial variation of drinking water pollutant and to identify the most significant parameters in each group of physico- chemical parameters (PCPs), Inorganic parameters (IOPs), heavy metals and organic parameters (HMOPs) and pesticides parameters (PPs). The Discriminant Analysis (DA) and One- Way Analysis of variance (ANOVA) showed spatial variation on four station categories and the variance of four group parameter in water drinking quality while principle component analysis (PCA) was carried out to identify the most significant of each water quality parameters base on given group. DA and ANOVA successfully reduced the physico and inorganic pollutants concentration with significant value 98.63% and 96.90%. PCA revealed six most significant drinking water quality parameters for PCPs, nine significant parameters for IOPs, fourteen parameters on HMOPs and four significant of PPs with the p value less than 0.05 ($p < 0.05$). Therefore, this study proves that chemometric method is the alternative way to explain the characteristic of the drinking water quality and could reduce several parameters and sampling points in the future sampling strategy.

Keywords: *Discriminant analysis; One- way analysis of variance; Principal component analysis; Drinking water quality.*

1. Introduction

Water acts as a substantial role in keeping the human health and welfare. Fresh drinking water is instantly recognized as a fundamental need of human organisms. Roughly 780 million people do not possess access to clean and safe water and around 2.5 billion people do not have proper sanitation. As a result, roughly 6–8 million people drop dead each year due to water related diseases and calamities [1]. Consequently, water quality control is a top-priority policy agenda in many regions of the universe [2]. Water quality defines the physicochemical, biological and radiological physiognomies of water [3]. It usually measures the condition of water with respect to the provisions of living fellow and any human requirement or purpose [4].To make sufficient in drinking water quality for human being, the flow of the treated water was monitored by Ministry of Health (MOH) from the raw source (R) until the water distributed to the consumers called Auxiliary (A). This monitoring was involving in the programmed call National Drinking Water Quality Surveillance Programme (NDWQSP) which is implement on 1893 with the aim to enhance the banner of health of the people by ensuring the safety and acceptability of the drinking water supplied to the consumers that complies with the stipulated standards, thereby cutting down the incidence of water-borne diseases or intoxication associated with poor quality public water supplies.

Several water quality (WQ) parameters have been measured and monitored throughout the Malaysian water supply scheme. The selected WQ parameters were classified into four groups, namely Group 1: Physico-chemical parameters (total coliform, E-Coli, color, pH, residual chlorine, temperature and conductivity), Group 2: Inorganic parameter (total dissolved solids, chlorine, nitrogen-ammonia, nitrogen- nitrate, ferum, florine, hardness, aluminum, Mangan, chemical oxygen demand, biochemical oxygen demand and total organic carbon), Group 3: Heavy metals and organic parameter (mercury, cadmium, arsenic, cyanide, lead, chromium, zinc, natrium/ sodium, sulphate, selenium, argentum, magnesium,

mineral oil, chloroform, bromoform, dibromochloromethane and bromochloromethane) and Group 4: Pesticides (Aldrin/ dieldrin, DDT, H & He, methoxychlor, lindane, chlordane and endosulfan). All of the parameters were state in drinking water quality standard (DWQS) with their maximum limit. Under the NDWQSP, a national expert committee established the set of the National Drinking Water Quality Standards. The criteria were prepared after taking into consideration the situation in Malaysia and reviewing recommendations made by the World Health Organization (WHO) as easily as the practices in other nations. A number of guidelines were also produced under the NDWQSP to ensure its successful implementation and to achieve its objective effectively [5].

Since NDWQSP was fulfilled, millions of WQ data were collected and dealing with tones of the dataset is a challenging task towards a comprehensive understanding how to elaborate of information display. Hence, the application of chemometrics technique is the best exercise that can be used in order to identify the most important parameter in each group. Chemometrics is the chemical study that uses mathematical, statistical and other methods employing Formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical data by analyzing the data [6]. In other words, chemometrics solves the following projects in the area of chemistry how to get chemically relevant information out of measured chemical data, how to interpret and expose this information, and how to get such information into data [7]. Thus, this study aims to identify the most significant parameters in each group of PCPs, IOPs, HMOPs and PPs and to determine the most significant spatial variation of drinking water pollutant.

2. Methodology

2.1. Data Collection

Large data set of drinking water quality was obtained from Department of Engineering Services, Malaysian Ministry of Health (MOH). The five years data collection from 2012- 2016 base on four group parameters were produced from 460 operational water treatment plants and 548 of water courses in Malaysia. There is six's intake sampling station which are raw or river catchment (R) followed by treatment plant outlet (TPO), two of service reservoir outlets (SRO) and two of distribution system also called auxiliary (A). Out of 44 parameters from 90 variables state in DWQS use in this study were grouping into four group parameters such as Group 1: Physico-chemical parameters namely total coliform (cfu), E-Coli (mg/l), colour (TCU), pH, residual chlorine (mg/l), temperature (^{o} C) and conductivity (mg/l), Group 2: Inorganic Parameter likes total dissolved solids (mg/l), chlorine (mg/l), nitrogen-ammonia (mg/l), nitrogen- nitrate (mg/l), ferum (mg/l), Florine (mg/l), hardness (mg/l), aluminum (mg/l), mangan (mg/l), chemical oxygen demand (mg/l), biochemical oxygen demand (mg/l) and total organic carbon (mg/l), Group 3: Heavy metals and organic parameter such as mercury (mg/l), cadmium (mg/l), arsenic (mg/l), cyanide (mg/l), plumbum/ lead (mg/l), chromium (mg/l), zinc (mg/l), natrium/ sodium (mg/l), sulphate (mg/l), selenium (mg/l), argentum (mg/l), magnesium (mg/l), mineral oil (mg/l), chloroform (mg/l), bromoform (mg/l), dibromochloromethane (mg/l) and bromochloromethane (mg/l) and last but not least Group 4: Pesticides namely (aldrin/ dieldrin, DDT, H & He, methox, lindane, chlordane and endosulfan). The secondary data on four group parameters of the drinking water quality of the water treatment plants were further investigated.

2.2. Data Pre-Treatment

Regarding of 5 years data that we are collected from Ministry of Health (MOH), we conclude that all the data's need to clean up to make it smooth for running an analysis. All missing data's needs to remove because we only use analytical to run the data. Missing data excluded are the abjad's, symbols, data blanks and typing error.

2.3. Data Analysis

After treated the data, the Multivariate statistical analysis, such as PCA and DA- ANOVA were performed by utilizing the XLSTAT 2014.The purpose of PCA and DA- ANOVA in this study is to identified the spatial variation drinking water pollutant and to distinguish the 44 drinking Water parameters originating from the 460 water treatment plants (WTPs) in Malaysia. These multivariate methods predict the line of descent of the pollutants from the water sources in society to curb problems originating from WTPs.

2.5. Discriminant Analysis

Discriminant analysis (DA) gives statistical classification of samples distribution common properties and is implemented with prior knowledge of relationship of objects to a particular group. It builds up a discriminant function for each group operating on fresh data [11-15]. DA accepts the same discriminant ability to experimental data with and without calibration [11]. The discriminant function state as per below [16]

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + ... + d_{ip}Z_p \quad (1)$$

where z = the score on each predictor, and d_i = discriminant function coefficient. The discriminant function score for a case can be formed with raw scores and unstandardized discriminant function targets. The mean discriminant function coefficient can be computed for each group, these group means are called centroids, which are created in the scaled down space produced by the discriminant function, reduced from the initial predictor variables. Once the discriminant functions are defined groups are differentiated, the utility of these roles can be studied via their ability to correctly reform each data point to their a priori group.

DAs were implemented on data matrix by using the standard, forward stepwise and backward stepwise modes. In forward stepwise mode, variables are included step-by-step beginning with the more significant until no important modifications are found [17, 44]. Present study, DA use to discriminate the drinking water pollutants for determine the most significant spatial variation.

2.4. One-Way Analysis of Variance

One-way analysis of variance, one- way ANOVA Is a technique that can be applied to compare means of two or more samples (using the F distribution). This technique can be utilized only for numerical response data, the "Y", usually one variable, and usually categorical input data, the "X", always one variable, hence "one-way" [10]. The generally used normal linear models for a finally randomized experiment are [11]

$$Y_{i,j} = \mu_j + €_{ij} \text{ (the mean models)} \quad (2)$$

or

$$Y_{i,j} = \mu + t_j + €_{ij} \text{ (the effect models)} \quad (3)$$

where Y_{ij} are observations, μ_j is the observation for treatment group, μ is the grand mean of the observation, t_i is a treatment effect, $€_{ij}$ is an effect of random error.

The ANOVA tests the null hypothesis that samples in whole groups are drawn from populations with the same mean values. To perform this, two estimates are made of the population division. These ideas rely on several assumptions. The ANOVA produces an F-statistic, the proportion of the variance calculated among the means to the variable within the samples. If the group means are drawn from populations with the same mean values, the variation between the group means should be lower than the variance of the samples, following the central limit theorem [10, 46]. The hypothesis can be tested as follows:

Ho: $\mu_1 = \mu_2 = \mu_3 ..... \mu_k$ (4)

Ha: at least one of the means is different.
The F statistics calculate the ratio between mean square among and mean square within and the decision rule is rejection of Ho if F > Fcrit otherwise accept Ho.

$F = MSA/ MSW$ (5)

One-way ANOVA use to confirm the discriminant analysis with the significant variables ($p < 0.05$).

2.6. Principle Component Analysis

Principle component analysis (PCA) is a technique commonly applied for decreasing the dimensions of multivariate problems [18]. It cuts the dimensionality of data set by clarifying the correlation amongst a huge number of variables in terms of a smaller number of principal components or PCs without dropping much data [12, 19-22, 41-43]. The principles component PC have showed as [16]

$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \cdots + a_{im}x_{mj}$ (6)

where z is the component score, a is the component loading, x the measured value of variable, $_i$ is the component number, $_j$ is the sample number and $_m$ is the total number of variables.
Rotation of the PCs by PC varimax rotation to build a better correlation between the PCs and the original variables. Varimax rotation guarantees that each variable is greatly correlated with only one component and has a close zero relationship with the other components [23]. PCs with eigenvalue more than 1 was used for the rotation [24]. The factor loadings after rotation are significant because they reproduce how much the variable gives to that specific PC and to what extent one variable is similar to the other. Range of that factor loading is greater than 0.75 called stronger, 0.75- 0.50 as moderate and 0.49- 0.30 as weak [25]. In this study, PCA practice for identify the most significant parameters in each group of PCPs, IOPs, HMOPs and PPs.

3. Results and Discussion

3.1. The Most Significance Spatial Variation of Drinking Water Pollutant

After analyzed the data collection use multivariate technique, the finding on Figure 1 plotting the discriminant function for physico – chemical, inorganic, heavy metal and organic and pesticide pollutants detected from four station categories. Those observe from DA standard mode on raw data and treated data. Meanwhile Table 1 (physico – chemical parameter), Table 2 (inorganic parameter), Table 3 (heavy metal and organic parameters) and Table 4 (pesticide parameters) showed the determination of spatial variation for all pollutants between station categories by confusion matrix of DA. Combination with one- way ANOVA for the statistical significant difference for all the parameters sampled from four different station categories (R, A, TPO and SRO) ($p < 0.05$) presented in Table 5.

3.1.1. Physico - Chemical Parameter (PCP)

The confusion matrix summarized the reclassification of the observations, which is the ratio of the number (54.75%) of observations that have been not well classified over the total number of observations. Therefor to continue the analysis of DA using forward stepwise and backward stepwise mode can't be proceed. So, to continue the analysis of DA the correct classification needs to achieve the percentage performance at least up to 70% (Table 1 PCPs and Figure 1 first line) with eliminate all the misclassified data from the data set (treated data).

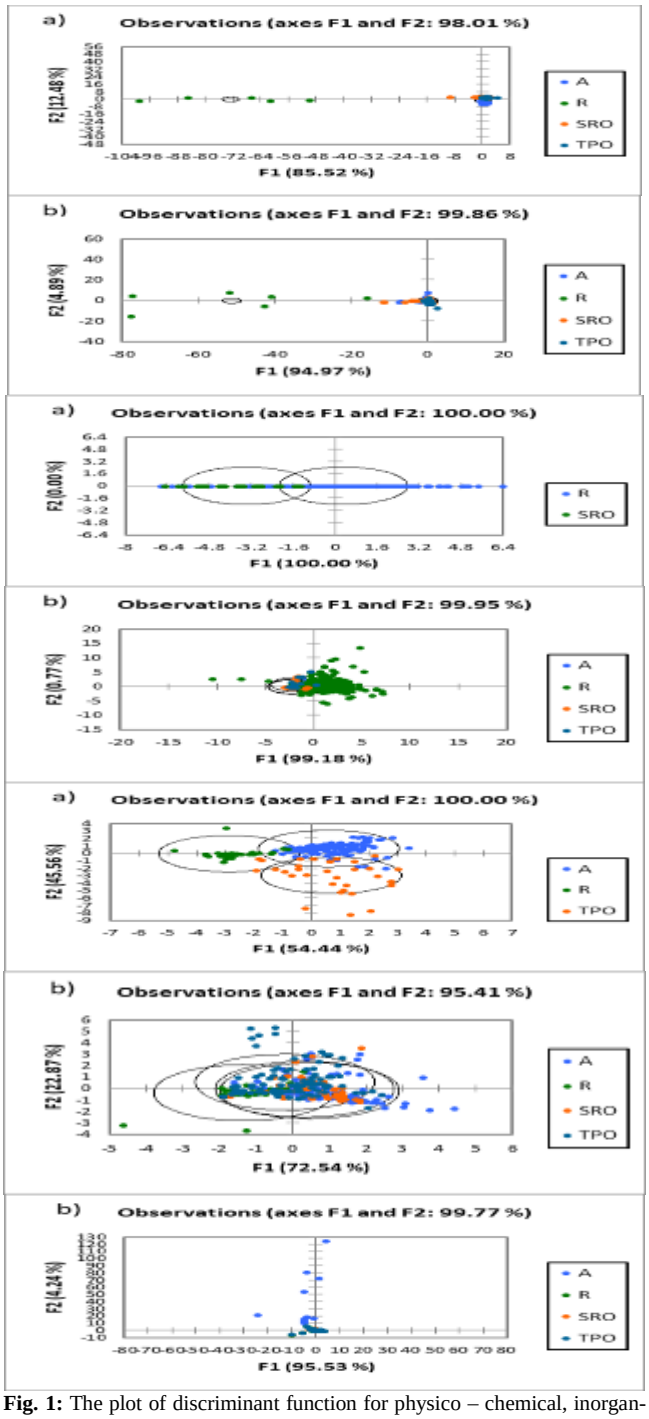

Fig. 1: The plot of discriminant function for physico – chemical, inorganic, heavy metal and organic and pesticide pollutants detected from four station categories a) standard DA mode (treated data) b) Standard DA mode (raw data).

The treated data contribute the accuracy of spatial classification using standard mode DA were 98.53% so further analysis of DA using forward stepwise and backward stepwise mode will be carried out in order to identify the most significant water quality parameters which plays an important role in discriminating the drinking water quality (Table 1). Table 5 showed the analysis of variance of four group parameters sampled from the various station point. All the station category is viewed after run the ANOVA which are station R, SRO, TPO and A. The variances of the station showed the significant differences ($p < 0.05$) in PCP parameters between the sampling station. There are seven of eight parameters was statistically significant except temperature ($p >$

0.05). On the other hand, for the future sampling strategy, only the most significant physico-chemical parameters (seven) should be taken into consideration for all the sampling station categories.

Table 1: Determination of spatial variation for physico chemical pollutants between station categories

St. Category	St. Category assigned by DA				Total	% correct
	A	R	SRO	TPO		
Standard DA mode (Raw data)						
A	1278	0	497	35	1810	70.61%
R	1	4	0	1	6	66.67%
SRO	746	0	680	32	1458	46.64%
TPO	56	0	295	50	401	12.47%
Total	2081	4	1472	118	3675	54.75%
Standard DA mode (Treated data)						
A	1197	0	4	3	1204	99.42%
R	1	2	1	1	5	40.00%
SRO	1	0	650	11	662	98.19%
TPO	0	0	6	26	32	81.25%
Total	1199	2	661	41	1903	98.53%
Stepwise forward DA mode (Treated data)						
A	1197	0	4	3	1204	99.42%
R	0	4	0	1	5	80.00%
SRO	1	0	650	11	662	98.19%
TPO	0	0	6	26	32	81.25%
Total	1198	4	660	41	1903	98.63%
Stepwise backward DA mode (Treated data)						
A	1197	0	4	3	1204	99.42%
R	0	4	0	1	5	80.00%
SRO	1	0	650	11	662	98.19%
TPO	0	0	6	26	32	81.25%
Total	1198	4	660	41	1903	98.63%

3.1.2. Inorganic Parameters (Iop)

The spatial variation of IOP which are make from four station categories for raw data (Table 2) shows only two station R and SRO was significantly discriminated (99.39% and 87.78%) compared to station A and TPO entertained very similar in pattern or same in their characteristic that are 0.00% in other words the P value more than 0.05 ($p > 0.05$) even though the percent of correction state the high value 95.04%. Therefore, two stations (R and SRO) significantly discriminate ($p < 0.05$), due to the different inorganic pollutant patterns obtained respectively. The different pollutant patterns between these two station categories (R and SRO) have expected due the difference in water quality, where for R station considered as the untreated water, while for SRO station considered as the treated water.

Table 2: Determination of spatial variation for Inorganic pollutants between station categories.

St. Category	St. Category assigned by DA				Total	% correct
	A	R	SRO	TPO		
Standard DA mode (Raw Data)						
A	0	4	41	0	45	0.00%
R	0	1779	11	0	1790	99.39%
SRO	0	22	158	0	180	87.78%
TPO	0	8	15	0	23	0.00%
Total	0	1813	225	0	2038	95.04%
Standard DA mode (Treated data)						
R	-	1747	33	-	1780	98.15%
SRO	-	28	130	-	158	82.28%
Total	-	1775	163	-	1938	96.85%
Stepwise forward DA mode (Treated data)						
R	-	1748	32	-	1780	98.20%
SRO	-	28	130	-	158	82.28%
Total	-	1776	162	-	1938	96.90%
Stepwise backward DA mode (Treated data)						
R	-	1748	32	-	1780	98.20%
SRO	-	28	130	-	158	82.28%
Total	-	1776	162	-	1938	96.90%

Hence, to determine the most significant inorganic pollutant due to spatial pattern or differences of station categories, all the misclassified data given by DA standard mode (raw data) was removed from the data set. The data called treated data was run again using the standard, stepwise forward and stepwise backward DA modes. So, the results show not much increase the value of correct percentage is 95.04% for raw data and 96.90% for treated data (Table 2) and illustrated into the plot discriminant function for IOP pollutant detected from four categories station (Figure 1 line 2). From the analysis, only two station R and SRO was under consideration due to the similarity of Inorganic pollutant pattern among to the others three station are A, TPO and SRO which is received treated water from the treatment plant. One- Way ANOVA illustrates only Chloride (Cl^-) is not significant compared to others eleven parameters which are dissolved solids, nitrogen-ammonia, nitrogen- nitrate, ferum, fluoride, hardness, aluminium, mangan, chemical oxygen demand, biochemical oxygen demand and total organic carbon ($p < 0.05$). The variation of the various station points which give the similar pattern showed no significant among them ($p > 0.05$). Therefore, the water treatment plant could be reduced the inorganic pollutant concentration successfully and produced drinking water with complying the permissible limit of inorganic pollutant for drinking water quality standard.

3.1.3. Heavy Metal and Organic Parameters (HMOP)

The plot of discriminant function for HMOP was very well discriminate after eliminated the misclassified data and re analyzed using DA standard mode compared run with the raw data (Figure 1, line 3). The DA show the percent of correction for heavy metal and organic pollutant is 53.46% in the moderate value which is need to be increase the value of percentage correction classification. The elimination of misclassified HMOP pattern (treated data) was employed in order to increase the correct classification of HMOP pattern. The highest correct classification was found to be 91.45% given by the standard, stepwise forward and stepwise backward DA mode. Three station categories (A, R and TPO) were taken into the consideration as independent variables after removal all the misclassified HMOP pattern was done (in this case, 100% data obtained from SRO station are misclassified and eliminated). Three of station categories show statistically significant among them which are station A give the value is 98.11% followed by station R is 88.10% and TPO is 63.64% the value. 10 out of 16 individual HMOP parameters shows significant difference due to spatial variation ($p < 0.05$), except Hg, Cd, As, Cr, Zn and Ag ($p > 0.05$) (Table 5). Based on this result, as for the future sampling strategy the HMOP parameter and sampling station could be reduced to 10 and 6 respectively.

Table 3: Determination of spatial variation for Heavy metal and organic pollutants between station categories

St. Category	St. Category assigned by DA				Total	% correct
	A	R	SRO	TPO		
Standard DA mode (Raw Data)						
A	158	11	0	17	186	84.95%
R	15	39	0	6	60	65.00%
SRO	48	3	0	7	58	0.00%
TPO	64	24	0	27	115	23.48%
Total	285	77	0	57	419	53.46%
Standard DA mode (Treated data)						
A	156	0	-	3	159	98.11%
R	5	37	-	0	42	88.10%
TPO	9	3	-	21	33	63.64%
Total	170	40	-	24	234	91.45%
Stepwise forward DA mode (Treated data)						
A	156	0	-	3	159	98.11%
R	5	37	-	0	42	88.10%
TPO	9	3	-	21	33	63.64%
Total	170	40	-	24	234	91.45%
Stepwise backward DA mode (Treated data)						
A	156	0	-	3	159	98.11%
R	5	37	-	0	42	88.10%
TPO	9	3	-	21	33	63.64%

Total	170	40	-	24	234	91.45%

3.1.4. Pesticide Parameters (PP)

Discriminant Analysis failed to show the significant of four station categories for PP (Table 4, Figure 1 line 4). The percentage of correction classification displayed only station A (99.98%) give the significant spatial variation among station R, SRO and TPO ($p < 0.05$). The output shows highly similar patterns ($p > 0.05$) with the data obtained from A station category. Failure of this standard DA mode analysis unable to proceed stepwise forward and stepwise backward DA mode. So, the analyzation was confirmed that the water quality patterns are similar to each other for treated water (R) and un- treated water (A, SRO and TPO).

Table 4: Determination of spatial variation for pesticide pollutants between station categories

St. Location	St. location assigned by DA				Total	% correct
	A	R	SRO	TPO		
Standard DA mode (Raw data)						
A	12944	2	1	0	12947	99.98%
R	6849	0	0	1	6850	0.00%
SRO	6004	0	0	0	6004	0.00%
TPO	7000	2	0	0	7002	0.00%
Total	32797	4	1	1	32803	39.46%

Out of two from seven PP showed the most significant parameters between four station categories that is H & He and chlorodyne ($P < 0.05$) and the rest of four PP statistically showed no significant (Table 5). For further sampling activities therefore, the sampling design could have reduced to only two sampling stations out of four sampling stations because the pesticide patterns show the similarity in patterns for all the stations. It is suggested to maintain the sampling station R and SRO for further analysis.

Table 5: One-way Analysis of variance of four parameters sampled from the various station point

Parameters PCP	R	SRO	TPO	A	Pr > F	Significant
TC	9824.20b	0.28a	0.00a	0.01a	0.00	Yes
E-coli	674.40b	0.00a	0.00a	0.00a	0.00	Yes
Turbidity	126.66b	1.63a	1.79a	1.54a	0.00	Yes
TCU	0.02a	0.17a	0.00a	3.39b	0.00	Yes
PH	6.73ab	7.44c	6.43a	7.24b	0.00	Yes
Residual chlorine	5.92d	1.27b	3.13c	1.06a	0.00	Yes
Temperature	29.76a	29.17 a	29.96a	29.54a	0.64	No
Conductivity	264.60bc	136.89b	412.60c	97.6a	0.00	Yes
IOP						
TDS	61.59b	0.00a			0.00	Yes
Cl	5.02a	6.54a			0.17	No
NH3-N	0.21b	0.08a			0.00	Yes
NO3-N	0.63b	0.04a			0.00	Yes
Fe	0.96b	0.03a			0.00	Yes
Fl	0.11a	0.367b			0.00	Yes
Hardness	21.97b	14.69a			0.03	Yes
Aluminum	0.37b	0.10a			0.02	Yes
Mangan	0.07b	0.01a			0.00	Yes
COD	7.39b	0.00a			0.00	Yes
BOD	1.69b	0.00a			0.00	Yes
TOC	2.69b	0.00a			0.00	Yes
HMOP						
Hg	0.00a		0.01a	0.00a	0.10	No
Cd	0.00a		0.00a	0.00a	0.24	No
As	0.04a		0.06a	0.00a	0.11	No
Pb	0.00ab		0.02b	0.00a	0.01	Yes
Cr	0.00a		0.00a	0.00a	0.19	No
Cu	0.00a		0.01b	0.00a	0.00	Yes
Zn	0.00a		0.09a	0.03a	0.12	No
Na	1.40a		19.91c	4.37b	0.00	Yes
SO4	2.05a		10.05b	9.49b	0.00	Yes
Se	0.00a		0.00ab	0.00b	0.05	Yes
Ag	0.00a		0.00a	0.00a	0.24	No
Mg	0.49a		0.69a	1.10b	0.00	Yes
Chloroform	0.00a		0.04b	0.03b	0.00	Yes
Bromoform	0.00a		0.00ab	0.00b	0.04	Yes
$CHBr_2Cl$	0.00a		0.00a	0.00b	0.00	Yes
$CHBrCl_2$	0.00a		0.00a	0.01b	0.00	Yes
PP						
Aldrin	0.00a	0.00a	0.00a	0.00a	0.72	No
DDT	0.02a	0.02a	0.02a	0.02a	0.51	No
H & He	0.00b	0.00a	0.00c	0.00c	0.00	Yes
Methoxychlor	0.00a	0.00a	0.01a	0.01a	0.69	No
Lindane	0.01a	0.01a	0.01a	0.01a	0.66	No
Chlordane	0.03b	0.04a	0.03c	0.03c	0.00	Yes
Endosulfan	0.01a	0.01a	0.01a	0.01a	0.15	No

3.2. The Most Significance Parameters in Each Group Using PCA.

3.2.1. Group 1: Physico- Chemical Parameters

After varimax rotation from eight PCs, only three VFs which represent 60.339% of the variance of data were selected due to the eigenvalues is greater than 1 (>1). Table 6 highlights that four out of eight parameters showed the most significant parameters. PC1 or D1 give the high percentage of variance value is 29.268% and highlight total coliform (TC) and E- coli as the most important variables and Illustrate of these two variables into the Figure 2 (A) for more understand. The coliform bacteria which can cause serious human illness and also considered as indicator organism which is their occurrence informs of the potential existence of disease causing organisms and should aware the person in authority for the water to take defensive action in quality of drinking water [29]. Second higher of variance (D2) is 16.167% was showed colour and residual chlorine for the strong (-0.736) and moderate (0.539) factor loading. The positive or negative sign just to show the direction of the factor was loaded as per Figure 1 (A). At time, water can have an unpleasant colour such as green or blue water, black or dark brown water, brown, red, orange or yellow water and milky white or cloudy water which are their own classification type. So, if this colour was happened, unsafe water for human being will appear [30, 47]. That is why colour is one of most significant parameters should under caution. Different of the residual chlorine which one needed in the safe drinking water. This variable become as a treatment method call chlorination used to extinguish or eliminate bacteria, viruses or other organisms in water [29]. In the meantime, D3 describe the moderate value of factor loading for pH is 0.684 and conductivity is 0.588 with the percentage of variance value is 14.904%. The permissible limit from DWQS give the value range of pH is 6.5- 9.0 and usually has no direct impact on water consumers but alert to pH control is required at all stages of water treatment to ensure acceptable water clarification and disinfection [31]. Conductivity is the measurement of a material to conduct electric current and strongly correlated with conductive ions from dissolved salts and inorganic materials [32].

Table 6: Factor loading and percentage of variance after varimax rotation for Physico- chemical parameter

Parameters	D1	D2	D3
TC	0.930	0.006	-0.051
E- coli	0.906	0.101	-0.150
Turbidity	0.635	-0.244	0.355
Color (TCU)	-0.043	-0.736	-0.179
pH	-0.044	-0.001	0.684
Residual chlorine	0.492	0.539	-0.440
Temperature	-0.060	0.380	-0.034
Conductivity	0.056	0.497	0.588
Variability (%)	29.268	16.167	14.904
Cumulative %	29.268	45.435	60.339

3.2.2. Group 2: Inorganic Parameters

PCA was describes nine of twelve variables most significant in group 2. These variables are TDS, Cl and Harness group in D1 followed by D2 including nitrogen- ammonia, nitrogen- nitrate, D3 consist Al and Mg and lastly COD and TOC in factor loading V4 (table 7) illustrate into Figure 2 (B) for more cleared. The VF1 contributes about 19.308% of the variation in the drinking water quality data under NDWQS from inorganic parameters. It has high loadings from three parameters, which are TDS (0.909), Cl (0.832) and Hardness (0.882). These factors can be interpreted as a mineral salt component which is total dissolved solids (TDS) is the term used to define the inorganic salts and minor sums of organic matter existing in result in water. The main elements are usually calcium, magnesium, sodium, and potassium cations and carbonate, hydrogen carbonate, chloride, sulphate, and nitrate anions [33]. Hardness is definite by Standard Methods as the amount of calcium and magnesium concentrations, both stated as calcium carbonate ($CaCO_3$), in milligrams per liter [34]. The VF2 demonstrate 14.072% of the variance in the data. It is displays high loading from NH3-N (0.868) and NO3-N (0.821). Nitrate and ammonia are the most common forms of nitrogen in water systems. Nitrogen can be a significant factor monitoring algal growth when other nutrients, such as phosphate, are rich. If phosphate is not plentiful it may limit algal growth rather than nitrogen [35]. Factor loading D3 displayed that two variables (aluminium and mangan) had the variation 13.717% of the drinking water quality data. Both of them in the strong and moderate loading which are mangan is 0.824 and aluminium is 0.669. As for the last PCs, D4 is accountable for 13.861% of the total variance with moderate positive loadings on chemical oxygen demand (COD) and total organic carbon (TOC). At large, COD is the amount of oxygen consumed by chemical reaction to reduce the organic compounds in the water column [36] whereas TOC describes the amount of carbon in the organic compound in certain material [37]. Therefore, the positive loadings on both parameters show that they are reliable to each other in the collected water samples.

Table 7: Factor loading and percentage of variance after varimax rotation for Inorganic parameter

Parameters	D1	D2	D3	D4
TDS	0.909	-0.010	0.031	0.093
Cl	0.832	0.135	-0.001	-0.039
NH3-N	0.042	0.868	0.140	0.040
NO3-N	0.030	0.821	0.002	0.142
Fe	-0.041	0.213	0.647	0.171
Fl	-0.011	0.264	-0.204	-0.535
Hardness	0.882	-0.034	0.003	0.029
Aluminum	0.003	-0.009	0.669	-0.020
Mangan	0.061	-0.020	0.824	0.042
COD	0.042	0.214	-0.018	0.748
BOD	0.098	0.281	0.190	0.513
TOC	-0.034	0.031	-0.058	0.700
Variability (%)	19.308	14.072	13.717	13.861
Cumulative %	19.308	33.380	47.097	60.959

3.2.3. Group 3: Heavy Metal and Organic Parameters

Figure 3 (C), explain the variances loading among all HMO parameters. Almost wholly the factor loadings located into positive loading. Similar to the inorganic parameter, all selected heavy metals and organic parameters show similar loadings on PC1 and PC2. There are five VFs was described after varimax rotation with eigenvalue more than 1 (> 1). Out of fourteen from sixteen variables showed the most significant variables on HMOP with the different value of variance (Table 8). D1 showed the high percentage of variance (33.018%) with six positive strong loading parameters namely mercury with value 0.968 tracked by cadmium (0.973), arsenic (0.976), chromium (0.887), selenium (0.782) and argentum (0.978). All the variables called as trace element in the water which is relatively in low concentration less than 0.1% (<0.1%) [38]. Magnesium (Mg) (0.701), dibromochloromethane ($CHBr_2Cl$) (0.862) and bromodichloromethane ($CHBrCl_2$) (0.906) are in the group VF2 distributes the moderate and high loading. The variation of these three variables is 14.753%. Mg is well known as one of the richest element in the upper central crust with the concentration of 13510 ppm [39, 45] whereas $CHBr_2Cl$ and $CHBrCl_2$ are decontamination consequences in the chlorinated water. Those compounds were resulting from the response of chlorine with natural organic matter and bromide ions [40]. Last three of the factor loadings of group three are D3 which load the variation is 11.491 followed by D4 is 10.195 and lastly D5 with the variation value is 6.918. Both of strong loading variables of D3

are Lead (0.919) and Copper (0.952) are come from the corrosion of galvanized iron pipe. The guideline required that action be taken to fight the corrosion of lead and the subsequent infection of drinking water [41]. Sodium (Na) and chloroform have a strong and moderate positive loading respectively which are 0.854 and 0.710. Bromoform is a disinfection by-product usually formed during the chlorination of water [42]. This variable described into factor loading D5 with the strong value is 0.830.

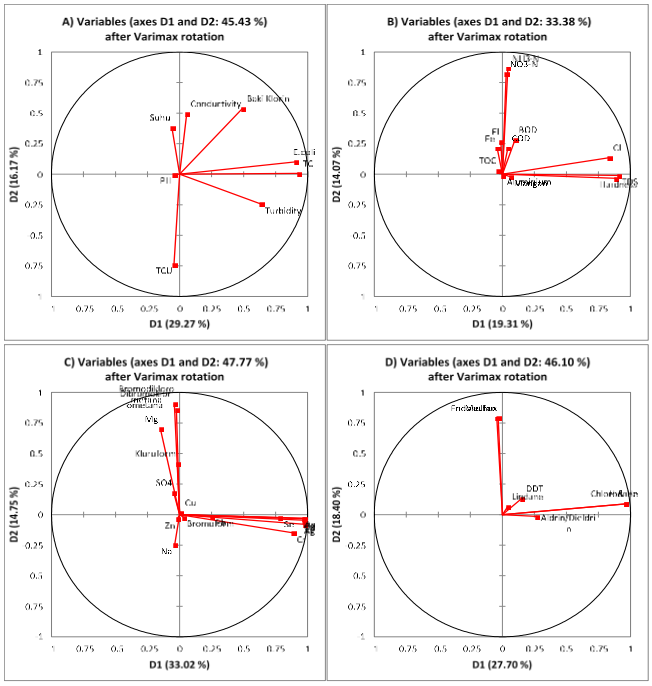

Fig. 3: Four parameters in the water samples was described into PCA plot A) Pyisico- chemical parameters, B) Inorganic parameters, C) Heavy metal and organic parameters and D) Pesticide parameters

Table 8: Factor loading and percentage of variance after varimax rotation for Heavy Metal and Organic parameter

Parameters	D1	D2	D3	D4	D5
Hg	0.968	-0.034	0.064	-0.054	-0.069
Cd	0.973	-0.046	0.059	-0.054	-0.052
As	0.976	-0.029	0.063	-0.056	-0.076
Pb	0.247	-0.013	0.919	-0.024	0.045
Cr	0.887	-0.149	0.069	0.116	0.033
Cu	0.009	0.019	0.952	-0.045	0.054
Zn	-0.016	-0.033	0.204	-0.151	0.409
Na	-0.039	-0.246	0.011	0.854	-0.125
SO4	-0.046	0.182	-0.120	0.563	0.419
Se	0.782	-0.025	0.033	-0.034	0.158
Ag	0.978	-0.079	0.080	-0.037	-0.017
Mg	-0.150	0.701	0.046	-0.051	0.075
Chloroform	-0.018	0.419	-0.035	0.710	-0.062
Bromoform	0.029	-0.022	-0.060	0.099	0.830
Dibromochloromethane	-0.025	0.862	-0.015	0.057	-0.034
Bromodichloromethane	-0.041	0.906	-0.037	0.118	-0.070
Variability (%)	33.018	14.753	11.491	10.195	6.918
Cumulative %	33.018	47.771	59.262	69.457	76.375

3.2.4. Group 4: Pesticide Parameters

Two (2) PCs were obtained for pesticide parameter with the total variance in the data set sums up about 46.102% (Table 9). All selected inorganic parameters show like loadings on PC1 and PC2 (Figure 2 D). D1 is described for 27.701% of the pesticide parameter with strong factor loadings on chlordane, H & He. D2 highlight 18.401% of the total variance with strong positive loadings on methoxychlor and endosulfan. The four presented pesticides are classified as organochlorine insecticides. However, Chlordane, H and He are in the form of viscous liquid whereas Methoxychlor and endosulfan are in the form of solid, namely crystal shape [Based on the Pesticide Action Network (PAN) Pesticide Database].

Table 9: Factor loading and percentage of variance after varimax rotation for Pesticide parameter

Parameters	D1	D2
Aldrin/Dieldrin	0.270	-0.015
DDT	0.152	0.132
H & He	0.959	0.087
Methoxychlor	-0.033	0.792
Lindane	0.041	0.065
Chlordane	0.959	0.088
Endosulfan	-0.043	0.790
Variability (%)	27.701	18.401
Cumulative %	27.701	46.102

4. Conclusion

Based on the prolonged explanations of the concentration of the physico chemicals, inorganic, heavy metal and organic and pesticide parameters in the water distributed to the water supply in Malaysia, the chemometric statistical technique assisted to deliver significant effort on the spatial variability and the most important of parameters of a huge and complicated drinking water quality data. The application of DA in this study has succeeded to discriminate or distinguish the spatial location among four station categories while One – way ANOVA has managed to discriminate the drinking water parameters. I) PCP; the variances of the station showed the significant differences ($p < 0.05$) in PCP parameters between the sampling station while ii) IOP; the variation of the various station points which give the similar pattern showed no significant among them ($p > 0.05$). Therefore, the water treatment plant could be reduced the inorganic pollutant concentration successfully and produced drinking water with complying the permissible limit of inorganic pollutant for drinking water quality standard followed by iii) HMOP; based on this result, as for the future sampling strategy the HMOP parameter and sampling station could be reduced to ten (10), and six (6), respectively lastly iv) PP; for further sampling activities therefore, the sampling design could have reduced to only two sampling stations out of four sampling stations because the pesticide patterns show the similarity in patterns for all the stations. It is suggested to maintain the sampling station R and SRO for further analysis.

The PCA achieved to expose the most significant parameters in each group where is Group 1; six variables from the eight respectively which are responsible to drinking water quality variations. There are Total coliform, E- coli, color, residual chlorine, pH and conductivity. Same with group 2, the variables that accountable to drinking water quality variations has 9 the most significant parameters out of twelve. While for group 3 resulted five VFs with each VFs has their strong and moderate variables which are out of fourteen from sixteen variables. Last group 4 is pesticide parameters that including four the most important parameter out of seven. Thus, for the future and effective management in sampling task recommended to select only significant parameters in each group to be collected and analyzed as it may reduce the cost of sample collection and analysis. Hence, the sampling design could be reduced the sampling point which are no variation each other or same pattern in their characteristic for example is only two sampling stations out of four sampling stations for pesticides parameter. It is suggested to maintain the sampling station R and SRO for further analysis. Moreover, this research discoveries may assist as a reference for other related studies carried out in the future.

Acknowledgement

First and foremost, the authors would like to thank the Malaysian Ministry of Health for providing us with the secondary data and valuable source of information.

References

[1] UN-Water, An increasing demand, facts ans figures, UN-Water, coordinated by UNESCO in collaboration with UNECE and UN-DESA, 2013, http://www.unwater.org/water-cooperation-2013/en/.

[2] World Health Organization (WHO), *Guidelines for Drinking-Water Quality*, WHO Press, Geneva, Switzerland, 4th edition, 2011.

[3] Diersing, N. (2009). Water Quality: Frequently Asked Questions. Florida Keys National Marine Sanctuary, Key West, FL.

[4] Johnson, D. L., Ambrose, S. H., Bassett, T. J., Bowen, M. L., Crummey, D. E., Isaacson, J. S., ... & Winter-Nelson, A. E. (1997). Meanings of environmental terms. *Journal of environmental quality*, *26*(3), 581-589.

[5] Laporan KKM 2016

[6] Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y. K. A. U. F. M. A. N., & Kaufman, L. (1988). Chemometrics: a textbook.

[7] Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 109-115.

[8] Ismail, Azimah, Mohd Ekhwan Toriman, Hafizan Juahir, Sharifuddin Md Zain, Nur Liyana Abdul Habir, Ananthy Retnam, Mohd Khairul Amri Kamaruddin, Roslan Umar, and Azman Azid. "Spatial assessment and source identification of heavy metals pollution in surface water using several chemometric techniques." *Marine pollution bulletin* 106, no. 1 (2016): 292-300.

[9] Kannel, P. R., Lee, S., Kanel, S. R., & Khan, S. P. (2007). Chemometric application in classification and assessment of monitoring locations of an urban river system. *Analytica Chimica Acta*, *582*(2), 390-399.

[10] Howell, David (2002). Statistical Methods for Psychology. Duxbury. pp. 324–325.

[11] K.P. Singh, A. Malik, S. Sinha, Water quality assessment and apportionment of pollution sources of Gomti River (India) using multivariate statistical techniques: a case study, Anal. Chim. Acta 538 (2005) 355

[12] D.A. Wunderlin, M. Diaz, M.M.V. Ame, S.F. Pesce, A.C. Hued, M. Bistoni, Patern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River basin (CordobaArtgentina), Water Res. 35 (2001) 2881.

[13] P.A. Rogerson, Statistical Methods for Geography, Sage Publications, London, 2001.

[14] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 3rd ed., Prentice Hall, Englewood Cliffs, NJ, 1992, 642 pp.

[15] K.P. Singh, A. Malik, D. Mohan, S. Sinha, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study, Water Res. 38 (2004) 3980.

[16] Poulsen, J., & French, A. (2008). Discriminant function analysis. *Retrieved from.*

[17] M. Vega, R. Pardo, E. Barrado, L. Deban, Assesment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water Res. 32 (1998) 3581.

[18] B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis, Water Res. 34 (2000) 807.

[19] J.E. Jackson, A Users Guide to Principal Components, Wiley, New York, 1991.

[20] R.R. Meglen, Examining large databases: A chemometric approach using principal component analysis, Mar. Chem. 39 (1992) 217.)

[21] Statheropoulos, M., Vassilliadis, N., Pappa, A., 1998. Principal component and canonical correlation analysis for examining air pollution and meteorological data. Atmospheric Environment 32, 1087e1095.

[22] Kim, J.O., Mueller, C.W., 1987. Introduction to Factor Analysis: What It Is and How to Do It. In: Quantitative Applications in the Social Sciences Series. Sage University Press, Newbury Park

[23] Liu, C.W., Lin, K.H., Kuo, Y.M., 2003. Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. Science of the Total Environment 313, 77e89.

[24] Simeonov, V., Einax, J. W., Stanimirova, I., & Kraft, J. (2002). Environmetric modeling and interpretation of river water monitoring data. Analytical and Bioanalytical Chemistry, 374, 898–905.

[25] Kim, J. O., & Mueller, C. W. (1987). Introduction to factor analysis: what it is and how to do it. Quantitative applications in the social science series. Newbury Park: Sage University Press.

[26] Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. Science of the Total Environment, 313, 77–89.

[27] http://www.env.gov.bc.ca/wsd/plan_protect_sustain/groundwater/library/ground_fact_sheets/pdfs/coliform(020715)_fin2.pdf

[28] https://www.doh.wa.gov/portals/1/Documents/pubs/331-286.pdf

[29] http://www.who.int/water_sanitation_health/dwq/chemicals/ph_revised_2007_clean_version.pdf

[30] Brown W S. 2016. Physical properties of seawater. In M. R. Dhanak, & N. I. Xiros (Eds.), *Springer Handbook of Ocean Engineering. Cham: Springer*, pp. 101-110

[31] http://www.who.int/water_sanitation_health/dwq/chemicals/tds.pdf

[32] http://www.fwspubs.org/doi/suppl/10.3996/052013-JFWM033/suppl_file/patnodereference+s8.pdf?code=ufws-site

[33] https://www.waterboards.ca.gov/water_issues/programs/swamp/docs/cwt/guidance/3310en.pdf

[34] *Clair N. S., McCarty, P. L. and Parkin, G. F. 2003. Chemistry for Environmental Engineering and Science (5th ed.). New York: McGraw-Hill.*

[35] Schumacher, B. A. 2002. Methods for the Determination of Total Organic Carbon (TOC) in Soils and Sediments. *Ecological Risk Assessment Support Center. US. Environmental Protection Agency* 23p.

[36] Pais, I., & Jones Jr, J. B. (1997). *The handbook of trace elements*. CRC Press.

[37] Wedepohl, K. H. 1995. The composition of the continental crust: Ingerson Lecture. *Geochimica et Cosmochimica Acta* 59(7):1217-1232.

[38] ATSDR. 2005. Bromoform and dibromochloromethane (CAS # 75-25-2 and 124-48-1). U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES, Public Health Service Agency for Toxic Parameter and Disease Registry. http://www.atsdr.cdc.gov/toxfaq.html.

[39] Schock, M. R., Cantor, A. F., Triantafyllidou, S., DeSantis, M. K., & Scheckel, K. G. (2014). Importance of pipe deposits to Lead and Copper Rule compliance. *Journal: American Water Works Association*, *106*(7).

[40] Ivahnenko, t., & zogorski, j. S. (2001). Sources and occurrence of chloroform and other trihalomethanes in drinking-water supply wells in the United States.

[41] Al-Odaini, N. A., Zakaria, M. P., Zali, M. A., Juahir, H., Yaziz, M. I., & Surif, S. (2012). Application of chemometrics in understanding the spatial distribution of human pharmaceuticals in surface water. *Environmental monitoring and assessment*, *184*(11), 6735-6748.

[42] Maliki, A. B. H. M., Abdullah, M. R., Juahir, H., Abdullah, F., Abdullah, N. A. S., Musa, R. M., ... & Nasir, N. A. M. (2018, April). A multilateral modelling of Youth Soccer Performance Index (YSPI). In *IOP Conference Series: Materials Science and Engineering* (Vol. 342, No. 1, p. 012057). IOP Publishing.

[43] Maliki, A. B. H. M., Abdullah, M. R., Juahir, H., Muhamad, W. S. A. W., Nasir, N. A. M., Musa, R. M., ... & Abdullah, N. A. S. (2018, April). The role of anthropometric, growth and maturity index (AGaMI) influencing youth soccer relative performance. In *IOP Conference Series: Materials Science and Engineering* (Vol. 342, No. 1, p. 012056). IOP Publishing.

[44] Al-Odaini, N. A., Zakaria, M. P., Zali, M. A., Juahir, H., Yaziz, M. I., & Surif, S. (2012). Application of chemometrics in understanding the spatial distribution of human pharmaceuticals in surface water. *Environmental monitoring and assessment*, *184*(11), 6735-6748.

[45] Aris, A. Z., Abdullah, M. H., Praveena, S. M., Yusoff, M. K., & Juahir, H. (2010). Extenuation of saline solutes in shallow aquifer of a small tropical island: a case study of Manukan Island, North Borneo. *Environment Asia*, *3*, 84-92.

[46] Juahir, H., Zain, S. M., Aris, A. Z., Yusof, M. K., Samah, M. A. A., & Mokhtar, M. (2010). Hydrological trend analysis due to land use changes at Langat River Basin. *Environment Asia*, *3*(2010), 20-31.

[47] Kamarudin, M. K. A., Toriman, M. E., Rosli, M. H., Juahir, H., Aziz, N. A. A., Azid, A., ... & Sulaiman, W. N. A. (2015). Analysis of meander evolution studies on effect from land use and climate change at the upstream reach of the Pahang River, Malaysia. *Mitigation and Adaptation Strategies for Global Change*, *20*(8), 1319-1334.