Int. J. Aquat. Biol. (2022) 10(1): 25-31
ISSN: 2322-5270; P-ISSN: 2383-0956
Journal homepage: www.ij-aquaticbiology.com

Short Communication

The effect of non-prepared ecological data on the decision of multivariate analysis: A case study of water quality data of the Shatt Al-Arab River, Iraq

Mujtaba A.T. Ankush*

College of Agriculture, University of Basrah, Basrah, Iraq.

Abstract: Many ecological studies deal with a multivariate approach in the statistical analysis of the data. Aquatic ecologists advise applying the multivariate techniques to interpret environmental data information through various techniques such as biplots plotting in PCA, RDA, and CCA to achieve this goal. This work examines the importance of the application of the preparation of raw data before performing the statistical techniques in terms of scaling and transforming data and its important impact on the results of the PCA analysis of the water quality. The results showed the effect of data processing on the outputs of the analysis. However, raw data must be presented to evaluate appropriate methods of data processing before applying analysis techniques. The results of the study showed that some data did not show a clear change when converting their raw data, while other variables had a significant and clear effect on improving the normal distribution of values after the conversion of raw data, and this was evident through the Shapiro–Wilk test which was conducted for the variables where the values increased significantly.

Article history:
Received 14 December 2021
Accepted 1 February 2022
Available online 25 February 2022

Keywords:
Multivariate analysis
Shapiro–Wilk
Standardize
Transformation
Box whisker

Introduction

Data analysis is one of the most important stages that the researcher should perform for data mining knowledge discovery in databases (KDD), which aims to discover useful information from large sets of data (Mannila, 1996). The data is growing exponentially over the past few decades and it has become a challenging task for scientists to handle the data themselves; hence, data science and artificial intelligence were born. Because of these techniques, the data was handled correctly (Arora and Malik, 2015). Statistical mistakes have grown frequent in the scientific literature; statisticians documented that around half of all published articles include at least one statistical error (Ewuzie et al., 2021). It is significant to run essential data preparation, and every "preparatory step" can save you time, rework, and wrong conclusions (Rokach and Maimon, 2007).

The variables in the data matrix can be on the same scale, same range, or not mixed range. In the first case, all the data are numbers e.g. morphometric measurements in millimeters. We often have a data matrix with variables on two or more scales e.g. composition of stomach contents, expressed as a set of percentages, along with morphometric measurements in millimeters and categorical classifications such as gender (Greenacre and Primicerio, 2013).

Ensuring the normal distribution of the data, changing the weights of different types of variables, and removing the effect of units of measurement are the most important reasons for performing data transformation and standardization before launching a multivariate analysis (Jongman et al., 1995; Greenacre and Primicerio, 2013). Different magnitudes of changes in variables do not necessarily have the same or similar ecological significance. Hence, this study aimed to investigate the importance of transforming raw data before proceeding with multivariate applications using the Shatt Al-Arab River water quality data.

Materials and Methods

The data set of this study was obtained from three sampling stations in the Shatt Al-Arab River near

*Correspondence: Mujtaba A.T. Ankush
E-mail: mujtaba.tahir@uobasrah.edu.iq

Table 1. Averages and ranges of 9 major water quality variables during the study period in Abu Alkhaseeb (site 1), Ashaar (sites 2), and Garmat Ali (site 3) of Shatt Al-Arab River.

Parameter	Unit	Site1		Site2		Site3	
		Mean	Min-Max	Mean	Min-Max	Mean	Min-Max
WT	°C	24.47	13.4-31.8	14.00	14-32.8	25.14	16.9-33
pH		8.14	7.3-8.94	7.20	7.2-8.9	8.38	7.1-9.2
EC	µS/cm	4173.67	3260-5350	2958.00	2958-4920	3357.79	2530-4135
DO	mg/l	5.37	4.4-5.96	4.20	4.2-5.7	5.66	4.1-6.5
BOD	mg/l	2.16	1.82-2.9	1.26	1.26-3	2.04	1.24-2.72
Turbidity	NTU	14.49	6.9-28.5	6.00	6-36	27.33	12-100
TH	mg/l	1460.83	920-2000	880.00	880-2400	1370.56	780-2800
NO3	mg/l	1.60	0.74-2.5	0.66	0.66-2.74	1.57	0.54-2.72
PO4	mg/l	2.55	1.02-3.88	1.00	1-4.02	2.62	1.4-3.84

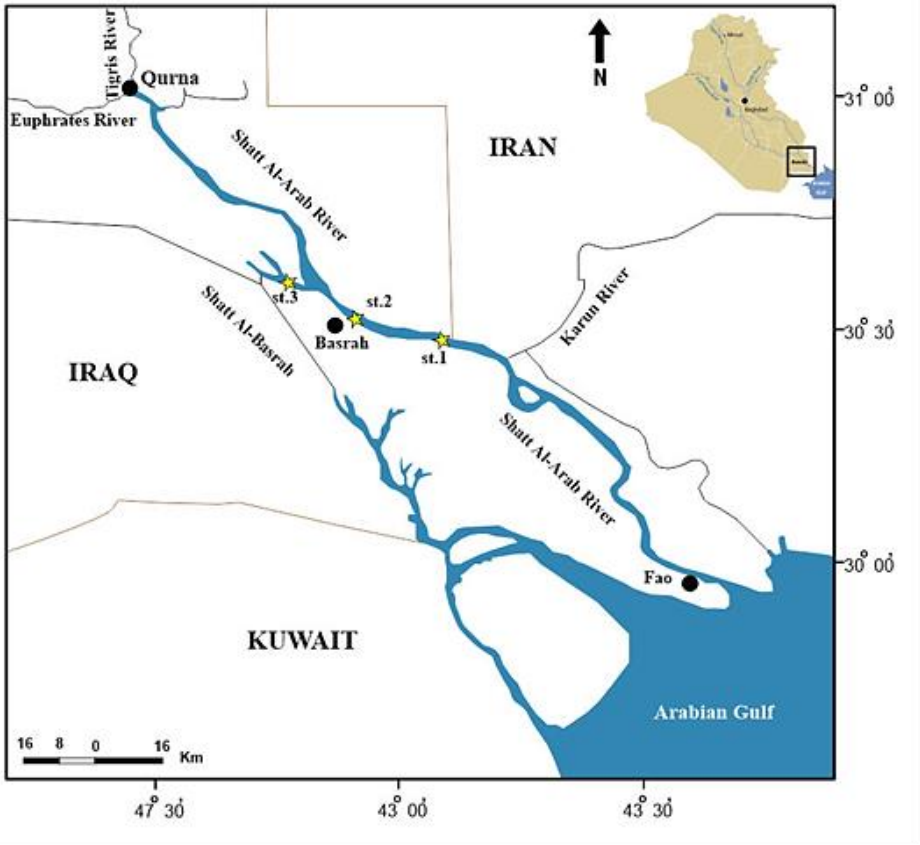

Figure 1. Map of Shatt Al-Arab River showing sampling sites,

Basrah, southern Iraq (Fig. 1). The total annual freshwater discharge of the Shatt Al-Arab is about 105.70 km³/year. This river depends on the flow rate of freshwater flowing from the Euphrates-Tigris system from Iraq, and Karkheh, Karun from Iran (Al-Asadi and Alhello, 2019). Several studies have been performed to evaluate the water quality of this river using multivariate techniques (Al-Ankush, 2013; Moyel and Hussain, 2015; Lateef et al., 2020). The data set of the present study covers a period from January to December 2020 including nine water quality variables. The means and ranges (minimum and maximum) of these variables are shown in Table 1.

There are many methods to standardize variables (van Tongeren, 1995; Gower and Warrens, 2017). Categorical variables or a combination of categorical and numeric variables are not standardized. All natural

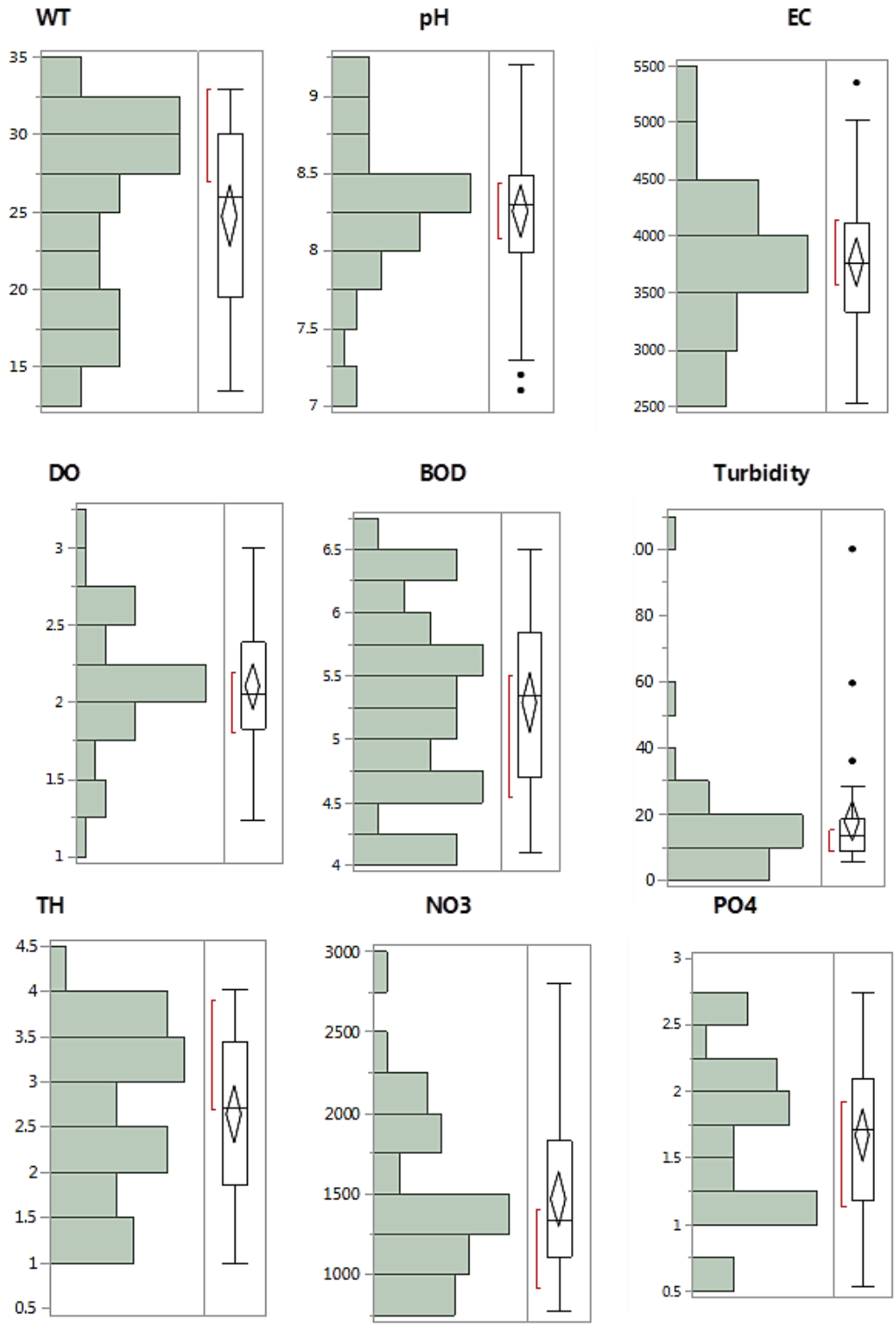

Figure 2. Box-and-whisker plots of water quality variables before transformation (WT, pH, EC, DO, BOD, Turbidity, TH, NO_3, PO_4.

and social science studies usually assume that the normally distributed standard variable is a mean of zero and a variance of one (Bhandari, 2020). When doing univariate or multivariate studies based on variable frequency distributions, selecting the right analytic technique is crucial. Despite the fact that linear statistical procedures are founded on the assumption that variables are normally distributed, most biological and physical measurements have a log-normal distribution, which means they are

Table 2. Factor scores of non-transformed water variables

Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
WT	**-0.89612**	0.15160	0.21736	-0.18784	0.11226	0.00968	0.08946	-0.02665	**0.26466**
pH	**-0.37803**	**0.31667**	**0.76967**	0.20686	**0.28523**	0.02449	0.10016	0.11054	-0.13205
EC	**0.73212**	0.14407	-0.27285	0.07326	**0.43610**	-0.17989	**0.35619**	0.10097	0.06172
DO	0.31449	**0.68948**	-0.12701	**0.30545**	0.00818	**0.55726**	0.00119	-0.05929	0.04644
BOD	0.01971	**-0.78733**	0.05734	**0.53074**	-0.09034	0.14547	-0.03142	**0.23660**	0.09199
Turbidity	0.31010	**0.60198**	0.19826	0.23904	**-0.58604**	-0.29291	0.08749	0.06762	0.05950
TH	**0.76349**	0.01652	**0.34981**	0.18994	**0.28069**	-0.18166	**-0.34398**	-0.13384	0.10165
NO3	**0.53381**	**-0.55734**	**0.44429**	-0.10603	-0.19022	0.15596	**0.29089**	-0.22472	0.01234
PO4	**0.64625**	0.07272	0.24539	**-0.63882**	-0.06604	0.18226	-0.09638	**0.24624**	0.03558

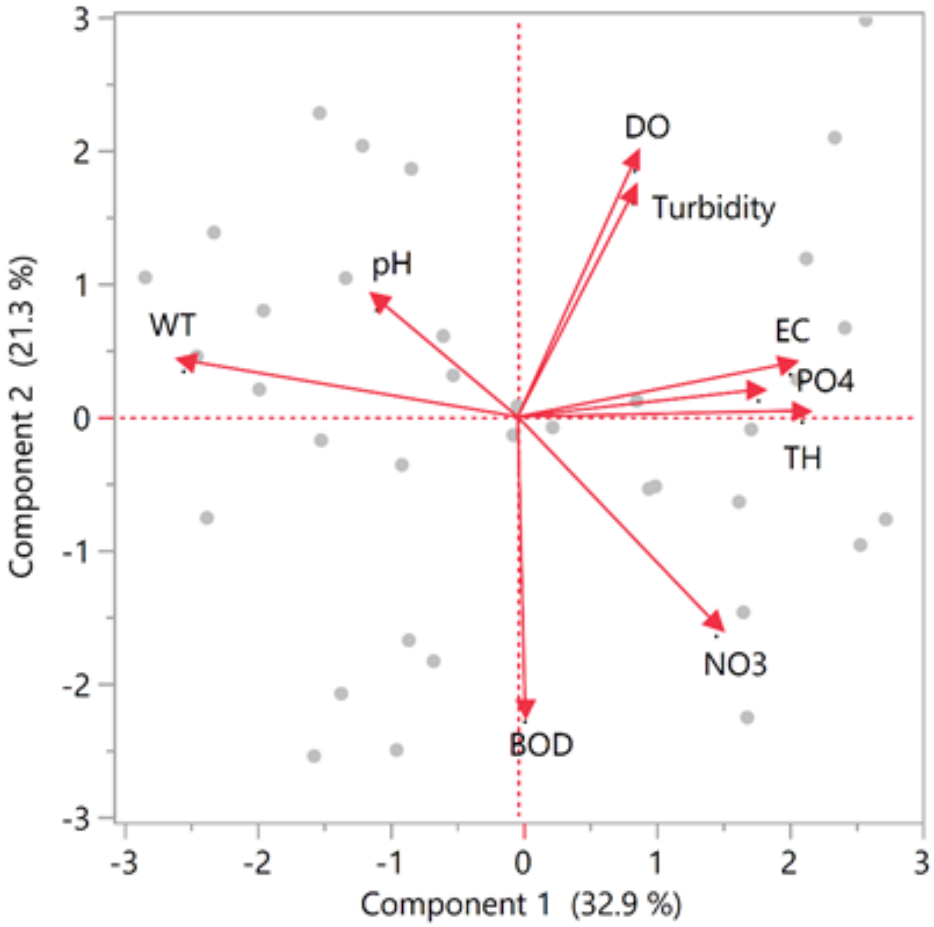

Figure 3. PCA biplot of water quality variables for 36 samples from three sites sampling using non-transformed values.

positively skewed (Limpert et al., 2001).

Results and Discussions

The normality test was carried out using the box plot technique (Fig. 2), for the water variables data before transformation, and some data variables were not distributed normally. Outliers were identified in variables such as pH, EC, and turbidity. According to the normality test, data transformed (Fig. 4) were normally distributed and no outliers were observed in the box plots of data variables.

A principal component analysis (PCA) biplot was done based on a correlation matrix using all nine selected variables, non-transformed (Fig. 3). The first

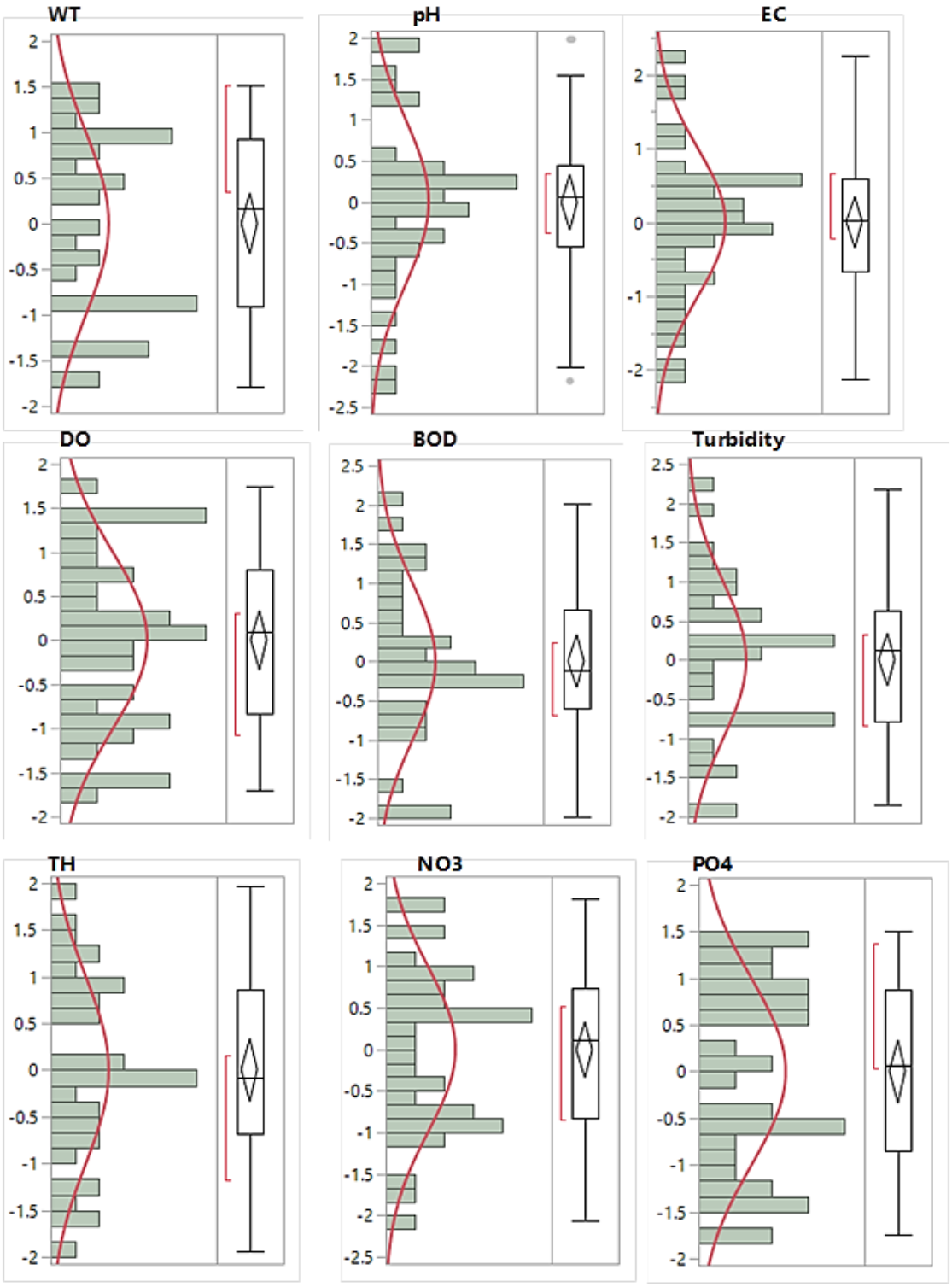

Figure 4. Box-and-whisker plots of studied environmental variables after transformation (WT, pH, EC, DO, BOD, Turbidity, TH, NO_3, PO_4).

and second components explained 54.2% of the total variations, the first PCA axis variation was accounted for 32.9% highly dominated by WT, TH, EC, PO_4, and NO_3 with weights of -0.896, 0.763, 0.732, 0.646, and 0.533, respectively (Table 2), whereas BOD, DO, turbidity and NO3 were highly linked with the second axis with variation 21.3 with weights of -0.787, 0.689, 0.601, and -0.557, respectively (Table 3).

Shapiro-Wilk test was applied to tests of normality. If the $P>0.05$, the null hypothesis, which assumes the

Table 3. Factor scores of non-transformed water variables.

Variable	Before transformation	After transformation
WT	0.0267	0.0374
pH	0.1838	0.2488
EC	0.7338	0.8816
DO	0.187	0.187
BOD	0.3572	0.3512
Turbidity	0.0001	0.7793
TH	0.0462	0.6852
NO_3	0.444	0.4429
PO_4	0.0294	0.03

Table 4. *P*-values for the Shapiro–Wilk test for all variables dataset before and after transformation.

Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Std WT	-0.90701	0.01783	0.10440	0.23376	0.16852	0.08562	0.11574	0.03572	0.24712
Std pH	-0.45433	0.26018	0.77007	0.08724	0.27229	0.14204	0.01865	0.08511	-0.15285
Std EC	0.71776	0.26478	-0.22065	-0.13878	0.47312	0.00789	0.31990	0.14291	0.01036
Std DO	0.20290	0.74736	-0.05139	-0.28826	-0.15952	0.52782	-0.07760	-0.04322	0.05117
Std BOD	0.13363	-0.73155	0.27187	-0.50247	-0.18669	0.10477	-0.01298	0.26686	0.05815
Std Turbidity	-0.06640	0.73301	0.26422	-0.15574	-0.45397	-0.33950	0.19517	0.06325	0.02803
Std TH	0.71862	0.15041	0.47719	-0.16954	0.27098	-0.21327	-0.22966	-0.12086	0.13508
Std NO3	0.60579	-0.46154	0.38399	0.26347	-0.22158	0.18744	0.27578	-0.20591	0.02120
Std PO4	0.63137	0.17522	0.03015	0.67915	-0.15262	0.05164	-0.13897	0.24928	0.03335

data are normally distributed, is accepted; otherwise, If $P<0.05$, then the null hypothesis is rejected. Table 4 shows the Shapiro–Wilk test of normality P-values for all variables. The datasets for pH, EC, turbidity, and TH are deemed normally distributed after transformation compared with these variables' non-transformed datasets.

The results showed the effect of data processing on the outputs of the analysis. The results showed that some data did not show a clear change when converting their raw data, while other variables had a significant and clear effect on improving the normal distribution of values after the conversion of raw data, and this was evident through the Shapiro–Wilk test which was conducted for the variables where the values increased significantly.

References

Al-Ankush M.A.T. (2013). Monitoring of Shatt Al-Arab river using water quality environmental modeling and benthic diatoms indices. Basrah, college of Agriculture. https://scholar.google.com/citations?view_op=view_ citation&hl=en&user=280m164AAA AJ&citation_for_ view=280m164AAAAJ:GnPB-g6to BAC

Arora D., Malik P. (2015). Analytics: Key to Go from Generating Big Data to Deriving Business Value. 2015 IEEE First International Conference on Big Data Computing Service and Applications. pp: 446–452.

Asadi S.A.R.A., Alhello A.A. (2019). General assessment of Shatt Al-Arab River, Iraq. International Journal of Water, 13(4): 360.

Bhandari P. (2020). Understanding normal distributions. Scribbr. https://www.scribbr.com/statistics/normal-distribution.

Ewuzie U., Aku N.O., Nwankpa S.U. (2021). An appraisal of data collection, analysis, and reporting adopted for water quality assessment: A case of Nigeria water quality research. Heliyon, 7(9): e07950.

Gower J.C., Warrens M.J. (2017). Similarity, Dissimilarity, and Distance, Measures of. In: Wiley StatsRef: Statistics Reference Online. John Wiley and Sons, Ltd. pp: 1-11.

Greenacre M.J., Primicerio R. (2013). Multivariate analysis of ecological data. Fundación BBVA. 336 p.

Jongman R.H.G., Braak C.J.F.T., van Tongeren O.F.R.

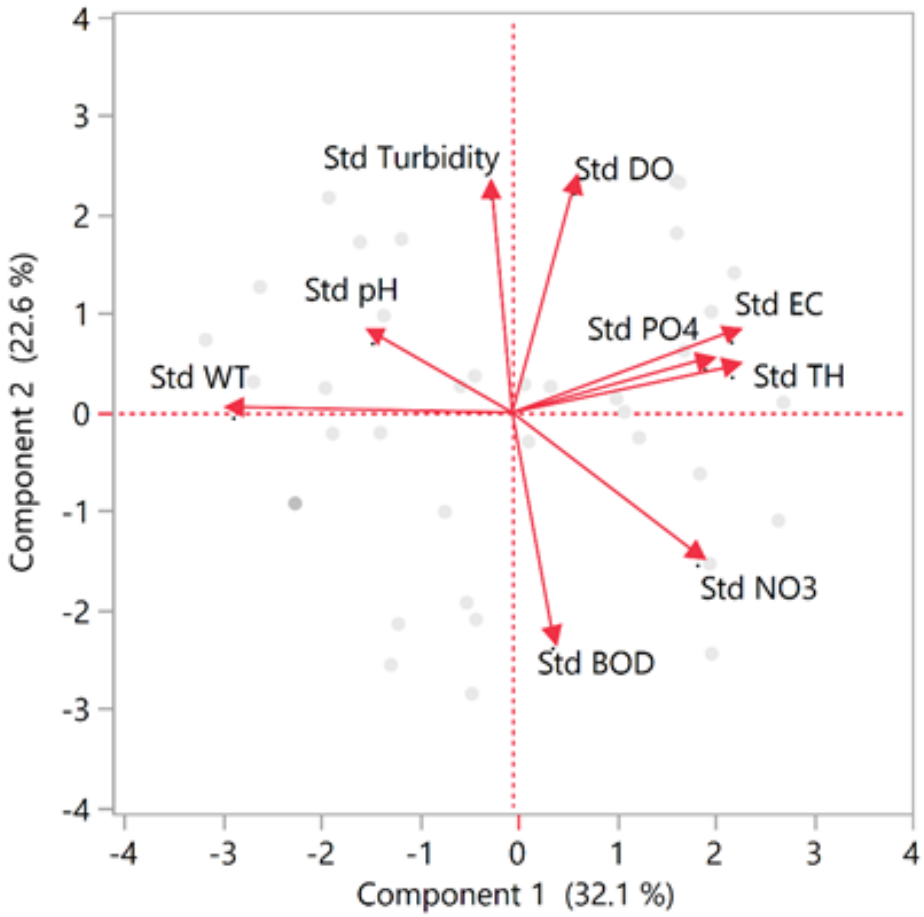

Figure 5. PCA biplot related to water quality variables for 36 samples from three sites sampling using transformed values.

(1995). Data analysis in community and landscape ecology. Cambridge University Press. 322 p.

Lateef Z.Q., Al-Madhhachi, A.-S. T., Sachit, D. E. (2020). Evaluation of water quality parameters in Shatt AL-Arab, southern Iraq, using spatial analysis. Hydrology, 7(4): 79.

Limpert E., Stahel W.A., Abbt M. (2001). Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. BioScience, 51(5): 341-352.

Mannila H. (1996). Data mining: Machine learning, statistics, and databases. Proceedings of 8th International Conference on Scientific and Statistical Data Base Management, 2-9.

Moyel M., Hussain N. (2015). Water quality assessment of the Shatt al-Arab River, Southern Iraq. *Jou*rnal of Coastal Life Medicine, 3(6): 1.

Rokach L., Maimon O.Z. (2007). Data mining with decision trees: theory and applications. World Scientific. 264 p.

van Tongeren O.F.R. (1995). Cluster analysis. In: C.J.F.T. Braak, O.F.R. van Tongeren, R.H.G. Jongman (Eds.), Data Analysis in Community and Landscape Ecology. Cambridge University Press. pp: 174-212.