



Identification of temporal and spatial patterns of river water quality parameters using NLPKA and multivariate statistical techniques

M. Rezaali¹ · A. Karimi¹ · N. Moghadam Yekta^{2,3} · R. Fouladi Fard⁴

Received: 12 July 2019 / Revised: 10 October 2019 / Accepted: 16 October 2019 / Published online: 31 October 2019
© Islamic Azad University (IAU) 2019

Abstract

River water quality is assessed by collecting samples from rivers. During this process, a significant amount of data is generated, which often results in challenges in interpreting the dataset. In this study, 14 water quality parameters of the Gadarchay River basin in Iran, collected monthly, were analyzed to identify pollution sources and patterns. Nonlinear principal component analysis was compared with frequently used multivariate statistical techniques. Results suggested that spatial and temporal nonlinear principal component analysis outperformed the other multivariate techniques by explaining 80.34% and 80.78% of the total variances, respectively. Cluster analysis categorized 20 sampling stations into three groups: less polluted, moderately polluted and highly polluted. Discriminant analysis was carried out both spatially and temporally for each of the three groups. The performance of the spatial discriminant analysis for less polluted, moderately polluted, highly polluted and overall was observed to be 95.83%, 70.14%, 64.58% and 76.85%, respectively. Temporal discriminant analysis was also done for each season to find the most significant variables. The performance of temporal discriminant analysis for summer, winter, autumn and spring was 85%, 85%, 40% and 61.67%, respectively. For source identification, principal component analysis was implemented on raw data. The results of spatial and temporal discriminant analysis were used to better interpret the results of principal component analysis for the less polluted, moderately polluted and highly polluted groups; five principal components covered 76% of the variance, four principal components covered 75% of the variance, and four principal components covered 77% of the variance, respectively.

Keywords River water quality assessment · Nonlinear principal component analysis · Discriminant analysis · Principal component analysis · Auto-associative artificial neural network

Introduction

In Iran, rivers are the main drinking water source for most populated centers. As a result, surface water pollution in the region poses a serious threat to public health. Surface water pollution has a variety of potential sources. In this region,

pollution is transferred mostly from municipal and industrial sources such as wastewater and urban runoff (Khaledian et al. 2018; Shrestha and Kazama 2007). To protect cities from water pollution, a variety of water quality (WQ) monitoring programs, both constant and intermittent, are used by regional governments to estimate spatiotemporal variations in WQ parameters. Such WQ monitoring programs produce a substantial amount of data (Alberto et al. 2001) that must be studied and analyzed continuously in what has become an expensive and labor-intensive approach (Chapman 1996).

Many studies have applied multivariate techniques for the purpose of data reduction, i.e., the process of removing non-significant data from a big dataset, pollution source identification and locating significant parameters. Helena et al. (2000) used principal component analysis/factor analysis (PCA/FA) for the temporal evolution of groundwater composition in an alluvial aquifer in Spain. Box and bivariate plots were used to interpret the results. PCA/FA extracted five principal components (PCs) from 16 variables

Editorial responsibility: M. Abbaspour.

✉ A. Karimi
karimi@qut.ac.ir

¹ Department of Civil Engineering, Faculty of Engineering, Qom University of Technology (QUT), Qom, Iran

² Department of Environment Engineering, Faculty of Science and Technology, Islamic Azad University, North Tehran Branch, Tehran, Iran

³ Lar Consulting Engineers, Tehran, Iran

⁴ Research Center for Environmental Pollutants, Qom University of Medical Sciences, Qom, Iran



recorded from two surveys. These PCs explained 71.4% of total variance, and the source of pollution was found to be the mineralization processes in the aquifer. Other significant parameters, ranked from most to least significant, included lead, aluminum, iron, nitrate, cadmium, copper and zinc.

Traditionally, multivariate methods have been used for several purposes, such as feature extraction of Landsat images (Balázs et al. 2018), summarizing the high spatial variability (Peña-Gallardo et al. 2019), and extracting spatial and temporal variabilities of rainfall (Suhaila and Yusop 2017). In multivariate statistical methods, linear mapping is usually applied to achieve various goals, including feature extraction and image compression. More recently, the introduction of artificial intelligence (AI) approaches has stimulated the development of new methods based on multivariate analysis and AI approaches, such as nonlinear principal component analysis (NLPCA). The main difference between PCA, a well-known statistical multivariate technique, and NLPCA is nonlinear mapping between the original and the reduced data (Kramer 1991). This feature of NLPCA renders it as a good alternative for multivariate statistical analysis in water resources studies. The current research uses NLPCA for feature extraction and dimensionality reduction of WQ parameters of the Gadarchay River, West Azerbaijan Province, Iran, and assesses its performance with other common multivariate techniques including PCA/FA and DA. To the best of the authors' knowledge, this is the first time that NLPCA has been applied to the WQ assessment of rivers globally.

According to preliminary studies, currently, the river suffers from being exposed to several anthropogenic pollutions (Laar Consulting Engineers 2018). Considering the

fact that the river is the main drinking and irrigation water source of multiple population centers in the basin, constant WQ monitoring is needed. As mentioned, WQ monitoring creates a large amount of data which makes it hard for the decision-makers to manage the WQ of the river efficiently. The motivation behind this study lies on the importance of dimensionality reduction of these large matrixes of WQ to help managers analyze the river WQ more efficiently.

Materials and methods

Study area

The basin area of the Gadarchay River spans 875 km² in the province of West Azerbaijan in Iran. The annual cumulative precipitation in this province is 351.7 mm. The river is 110 km long. There are 14 rural districts and 168 villages in the basin with a total population of 119,815 (Laar Consulting Engineers 2018).

The study area is comprised of the Gadarchay River basin, which is surrounded by the Barandouzchay, Zaab and Mahabad watersheds. The majority of the Gadarchay River basin is located in the mountainous region of Dalamper Bozorg and Baadgoole. The Ghalazchay, Kaanirash, Sheykhanchay, Balaghchichay and Mohamad Shah tributaries flow into the Gadarchay River along the river's path to Lake Urmia, into which the Gadarchay River discharges after passing the Bahramlou Bridge. Figure 1 illustrates the study area. For ease of analysis, the study area is divided into eight main regions.

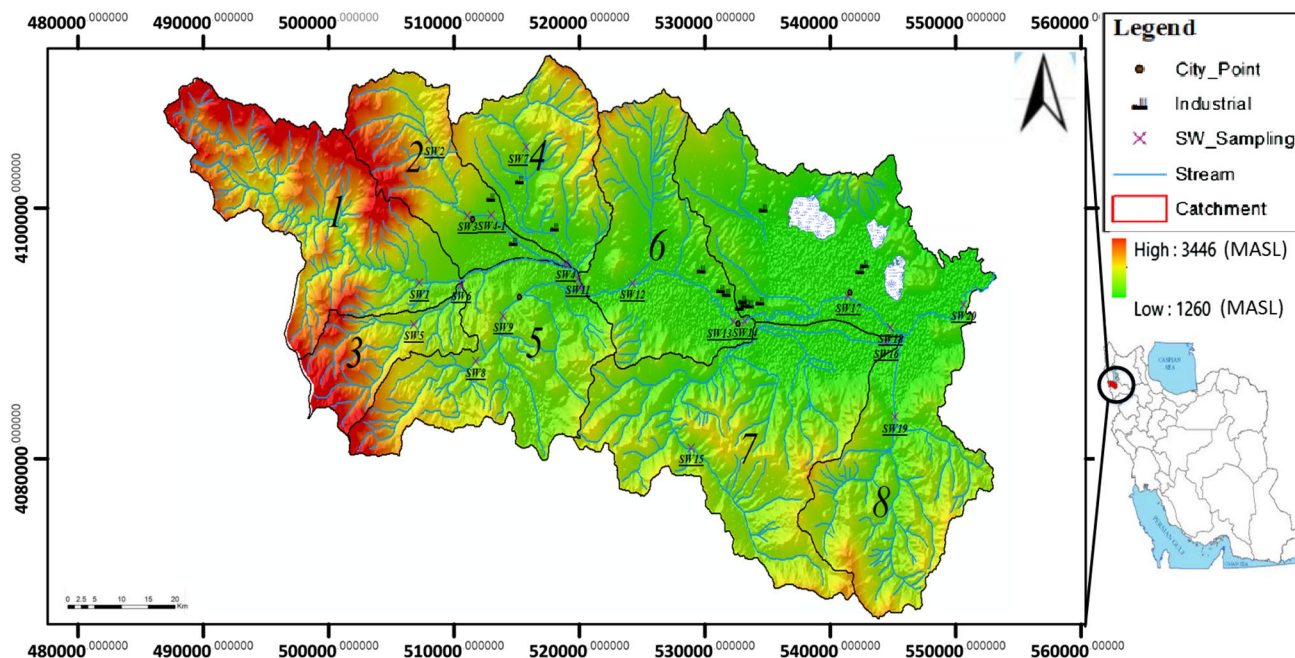


Fig. 1 The Gadarchay River and its tributaries



The first region encompasses the highest upstream point of the GR, which originates from the southern domains of Bikhul Mountain. Due to the region's mountainous terrain and its nearness to the western border of the country, no monitoring station was chosen upstream from this zone. The second region is situated mainly in the watershed of the river Ghalazchay, which after passing from the city of Ashnouye discharges into the GR. In this region, no industrial sources of pollution are reported, save two fish hatchery centers. The land use upstream of this city is mainly agricultural. The third region is the watershed of the Sheykhanchay River, which is considered to be a perennial river without any reported industrial pollution sources. The vegetation type in this region differs season to season, from grassland to tundra. The fifth region contains the Kanirash River watershed, which is one of the main permanent water sources of the GR. The upstream side of this region is covered with grass and other types of vegetation, and in the lower altitudes, dry farming is practiced. The city of Naghde in the sixth region is considered to be the main pollution source of the GR. Two sampling stations were located upstream and downstream of this city to carefully monitor WQ variations during the program. Other primary potential pollution sources are located in the eighth region, Mohamadyar City. Similar to Naghde, two sampling stations were chosen near the city to investigate the contribution of Mohamadyar City to pollution in the GR.

Monitored parameters

Fifty-four samples were collected through the Gadarchay River WQ monitoring program. In the current study, 14 parameters collected from 20 stations along the river from 10/22/2012 to 10/3/2013 were used based on the availability and continuity of data records during the study period. The parameters used in this study were water temperature (WT), turbidity (TR), total suspended solids (TSS), pH, electrical conductivity (EC), chemical oxygen demand (COD), 5-day biochemical oxygen demand (BOD), dissolved oxygen (DO),

nitrite (NO₂), nitrate (NO₃), phosphate (PO₄), total phosphorus (TP), total coliform (TC) and fecal coliform (FC). These parameters were chosen based on their potential risks to human health and the surrounding environment (e.g., FC, TC, etc.), or based on their impact on other parameters (e.g., WT, pH and DO, etc.), or less-studied parameters (e.g., TP, EC, etc.). Table 1 displays details about the measurement units and the analytical methods used to analyze the samples.

Data preprocessing

The distribution of all variables was analyzed using a Kolmogorov–Smirnov (K–S) test. Three different methods of dimensionality reduction were used, i.e., DA, PCA, PFA and NLPCA. PCA, PFA, CA and NLPCA were performed on standardized data with a zero mean and unique standard deviation, while DA was performed on raw data. Since the purpose of this paper is to assess the performance of each of the above methods, other statistical tests on the original dataset, which are popular mainly due to their suitability for use with the PCA method, were not performed.

CA

The primary purpose of this multivariate technique is to classify a specific group of objects based on their similarities (Moya et al. 2015; Rakotondrabe et al. 2018; Shrestha and Kazama 2007). Agglomerative hierarchical cluster (AHC) is one of the most applied techniques for the classification of objects based on different methods such as Euclidean distance, Dice coefficient, and Chi-square distance. The output of this approach is usually plotted on a dendrogram, which is an illustrative summary of the defined clusters (Moya et al. 2015; Rakotondrabe et al. 2018). Based on previous research, CA was performed on the normalized dataset in the current study using Ward's method in order to improve the comparative analysis (Alberto et al. 2001; Li et al. 2018; Shrestha and Kazama 2007). Ward's method analyzes the variance of the input data to calculate the distance between

Table 1 General information for the parameters measured in the current study

Parameter	Unit	Analytical method	Parameter	Unit	Analytical method
WT	C	Laboratory and field	NO ₂	mg/lit NO ₂ ⁻	Colorimetric
TR	NTU	Nephelometric	NO ₃	mg/lit NO ₃ ⁻	Ultraviolet spectrophotometric
TSS	mg/lit	Gravimetric	PO ₄	mg/lit P	Vanadomolybdophosphoric acid colorimetric
TP	mg/lit P	Vanadomolybdophosphoric acid colorimetric	pH	–	Electrometric
TC	MPN/100 ml	Multiple tube method	EC	µmhos/cm	Platinum electrode
FC	MPN/100 ml	Multiple tube method	COD	mg/lit O ₂	Closed reflux-titrimetric
DO	mg/lit O ₂	Membrane electrode	BOD	mg/lit O ₂	Respirometric

the clusters (Li et al. 2019; Peña-Gallardo et al. 2019). In the current study, this method was applied to the Euclidean distance, with an aim to minimize it. In an attempt to increase the within-group inertia as little as possible and to keep the clusters homogenous, this method aggregates two groups. It is noteworthy that Ward's criterion can only be used in classes with quadratic distance, i.e., Euclidean or Chi-square distance. Although this method has been widely used in the literature, it has two primary disadvantages: (1) Analysis may be slow for cases in which the datasets are large and (2) the dendrogram may be unreadable if too many variables are used. The AHC calculation process begins with the calculation of dissimilarity between predefined objects. The minimization of the agglomeration criterion is emphasized during the calculation of the first two main clusters. Then, the dissimilarity between the two clusters (or classes) and the next object is computed by the agglomeration criterion. This process continues until all of the objects (or variables) have been clustered (XLSTAT 2018a).

PCA/FA

PCA and FA are multivariate statistical tools designed to extract, from a larger group of data, the critical variables that contribute most of the variance. More specifically, PCA highlights variables that improve the description of the dataset relative to the other variables (Fouladi Fard et al. 2018). It also maximizes analysis simplification by giving the user the ability to eliminate other variables with a minimum loss of information (Gulgundi and Shetty 2018; Helena et al. 2000; Noshadi and Ghafourian 2016). The detailed mathematical basis of the PCA method is widely available in the literature, for example, in Shrestha and Kazama (2007). Mathematically, each principal component (PC) is a linear combination of the original dataset and orthogonal eigenvectors. This approach reduces information redundancy (Johnson and Wichern 1992).

FA is considered to be an extension of PCA (Johnson and Wichern 1992). The goal in FA is to further simplify PCA by reducing the contribution of less important variables through the application of varimax rotation, a process that generates varifactors (VFs). In the context of WQ assessment, there is a notable difference between PCA and FA. PC is a linear combination of WQ variables, while VF is able to incorporate unobservable, hypothetical, "latent" variables (Alberto et al. 2001; Helena et al. 2000; Vega et al. 1998). In the present study, based on the previous literature (Gulgundi and Shetty 2018; Li et al. 2018) PCs with eigenvalues less than one were not considered into further analysis, while PCs with eigenvalues greater than one were used to select the most suitable PCs and VFs.

DA

Introduced by Fisher (1936), DA has been slightly modified over the course of time but remains both explanatory and predictive. Although the current literature suggests better performance from DA than PCA (Alberto et al. 2001; Singh et al. 2005; Vega et al. 1998), in the sense that it uses linear combinations of variables, DA is considered to be similar to PCA and FA. Computationally, PCA calculates the vector(s) that has the largest variance among the original dataset, while DA explicitly models the difference between two classes using a vector that best discriminates between the classes (Martinez and Kak 2001). The mathematical equation that represents DA is presented in Eq. 1 (Alberto et al. 2001; Johnson and Wichern 1992; Shrestha and Kazama 2007; Singh et al. 2005).

$$f(G_i) = k_i + \sum_{j=1}^n w_{ij}p_{ij} \quad (1)$$

where i corresponds to the number of groups (G), k_i is the constant inherent to each group, n is the number of variables used to classify a set of data into a given group, and w_j is the weight coefficient, assigned by DA to a given selected variable p_j . To assess the performance of the DA, a confusion matrix was used to compare the predicted output against the real observation to calculate the percentage of well-classified observations.

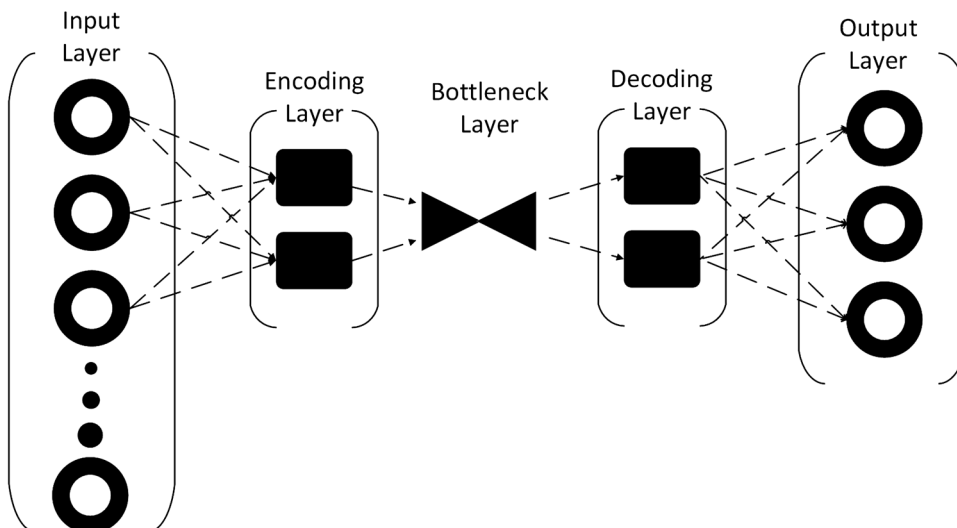
NLPCA

In PCA, a straight line is fitted through the middle of the data cluster. NLPCA differs in that a curved line is generated and then passed through the middle of the data cluster. The principal difference between the NLPCA method and traditional PCA is that PCA only employs linear mapping between the input data and the first PC, while NLPCA supports nonlinear mapping by training an auto-associative artificial neural network (AANN) (Hsieh 2004).

NLPCA trains an AANN using three hidden layers between the input and output layers. The output layer and three hidden layers contain four transfer functions (activation functions) (Hsieh 2004). Figure 2 provides a schematic of the NLPCA process.

As shown, the first layer (from left) is the input layer where the data are sorted as a matrix in a time series format. The second layer is the encoding layer, where a nonlinear function reduces the dimensions of the input data into single-dimension data. Data compression is achieved in the following layer, the bottleneck layer, by using a bottleneck neuron. The next layer, the decoding layer, recovers the lowered-dimension data to the

Fig. 2 Schematic network topology of the NLPCA process



original form by using inverse transform mapping. Similar to the linear mapping in PCA, NLPCA can be defined by Eq. 2:

$$T = G(Y) \tag{2}$$

where G is a nonlinear vector function composed of f individual nonlinear vector functions, Y is a row of an $(n \times m)$ data matrix, and T is a single row of $(n \times f)$ scores matrix. Consequently, Eq. 3 presents the definitive version of Eq. 2:

$$T_i = G_i(Y) \tag{3}$$

where G_i is the i th nonlinear factor of Y . The inverse transformation of Eq. 3, Y'_i which restores the original dimensionality of data using H_i as a second nonlinear function, is shown in Eq. 4:

$$Y'_i = H_i(T) \tag{4}$$

This process continues until the ANN minimizes the cost function. The following equation (Eq. 5) defines the cost function (Kramer 1991):

$$J = (\|Y - Y'\|)^2 \tag{5}$$

where J is the cost function, which is minimized during the training period. To this end, a function of the following form (Cybenko 1989) can fit any nonlinear function $\vartheta = f(u)$ to an arbitrary degree of accuracy (see Eq. 6):

$$\vartheta_k = \sum_{j=1}^{N_2} w_{jk_2} \sigma \left(\sum_{i=1}^{N_1} w_{ij_1} u_i + \theta_{j_1} \right) \tag{6}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is a sigmoidal transfer function implemented as a monotonically increasing function. Equation 6 is a feedforward ANN with N_1 input, a hidden layer comprised of N_2 node with a sigmoidal transfer function and a linear output node. w_{jk_2} is the weight on the connection node

i in layer k to node j in layer $k + 1$ and θ is bias (Kramer 1991).

The NLPCA model is able to utilize a pre-PCA to reduce the contribution of unimportant data. This method may improve the performance of the process. To complete this step, data must first be normalized. However, in the current study, to qualify the NLPCA performance, this option was not used, and therefore, the input data were applied in a raw format without any normalization (Scholz et al. 2008).

In standard PCA, the ranking of the variables is easily achieved by measuring the absolute value of the loading matrix. However, in NLPCA, since the components are curves, no global ranking is possible. In NLPCA, the rank order differs for each time step; in other words, the rank order in NLPCA is dependent on time. The tangent direction $dz = dx/dt$ at the curve of components value x at time t given by the first component for the sample point(s) defined in PC may be a reliable method to rank variables in each time step of the l_2 -normalized values of dz (Scholz et al. 2008). With the application of the bottleneck AANN, the training process undertaken using the mapped data was found to be more consistent compared to the same process under a regular multi-layer perceptron feedforward ANN. Other network parameters including the number of neurons (nonlinear components) in each layer, maximum iterations, type of NLPCA, i.e., hierarchical, circular, etc., and weight decay coefficient were optimized by a trial-and-error process.

Software

For the multivariate statistical methods, i.e., PCA/FA, DA and CA, XLSTAT software version 2016 was used (XLSTAT

2016). For the NLPCA method, MATLAB version 2017a was used (MATLAB 2017).

Results and discussion

By referring to Fig. 2 and the result of the K–S test, it was shown that the distribution of data did not follow a normal distribution at a 5% significance level. After identifying the data distribution, Spearman correlation analysis was used to study the spatial correlation between the stations (see Fig. 4). As a side note, since the multivariate techniques used were nonparametric, the distribution of dataset did not affect results and hence, was not of importance (Razmkhah et al. 2010). In Fig. 3, the mentioned parameters are displayed along with their basic statistical analysis results, including minimum value, maximum value, mean and standard deviation.

Spatial clustering

Since the 20 stations were located in different parts of the basin (e.g., upstream, downstream, tributaries and the main river), it was important to classify them based on their WQ parameters. To this end, CA was used. Figure 4 shows the dendrogram of the CA (right).

As shown in Fig. 4, all of the clusters that yielded a statistical significance of $D_{\text{link}}/D_{\text{max}} < 60\%$ were classified. Then, the clusters were divided into three main sub-clusters, i.e., less polluted (LP), moderately polluted (MP) and highly polluted (HP), based on the largest decrease in Shannon's entropy between a node and the next node (Shannon 1948). Figure 5 provides more explanation of how the stations were clustered into three major groups. Stations 5, 2, 1 and 3 in the LP cluster are located near the upstream portion of the basin. The primary pollution sources in this area are land use and erosion. The effects of anthropogenic pollution in the LP region were less significant than in the other clusters. Although an industry is active upstream of station 3, it is not especially water-dependent and so has no discernible impact on local WQ. In the MP cluster, however, the effects of anthropogenic pollution are more noticeable than in the LP cluster. Stations in this cluster are mainly located midstream in the Gadarchay River basin. Domestic and industrial wastewater (Gabris et al. 2018), fertilizers in agricultural runoff (AlKhader et al. 2019) and erosion are significant sources of pollution (Hunt et al. 2019) in this region. Certain stations in this group, such as 6, 7, 8, 19 and 15, are more significantly affected by agricultural land use than domestic wastewater. Agricultural jobs are dominant in the villages upstream of this region. Other stations in this group

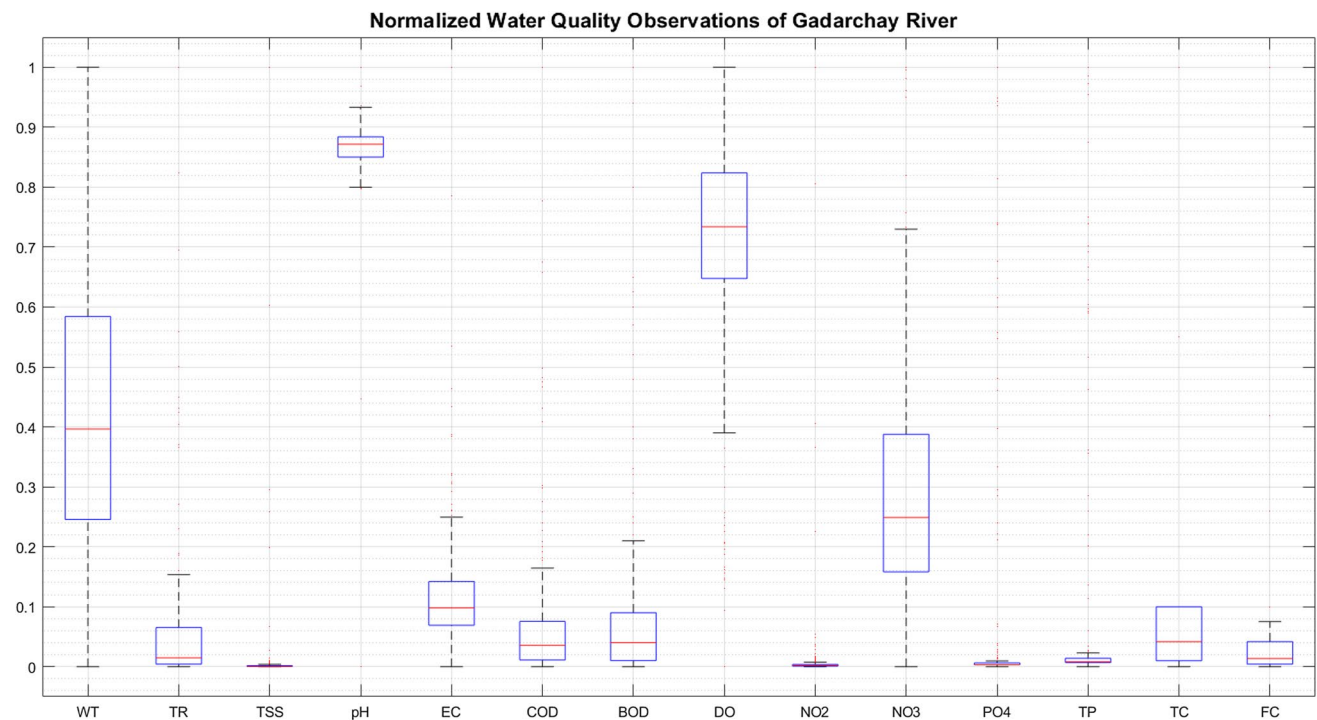


Fig. 3 Basic statistical analysis on the parameters used in the Gadarchay River study from the 20 monitoring stations from 10/22/2012 to 10/3/2013



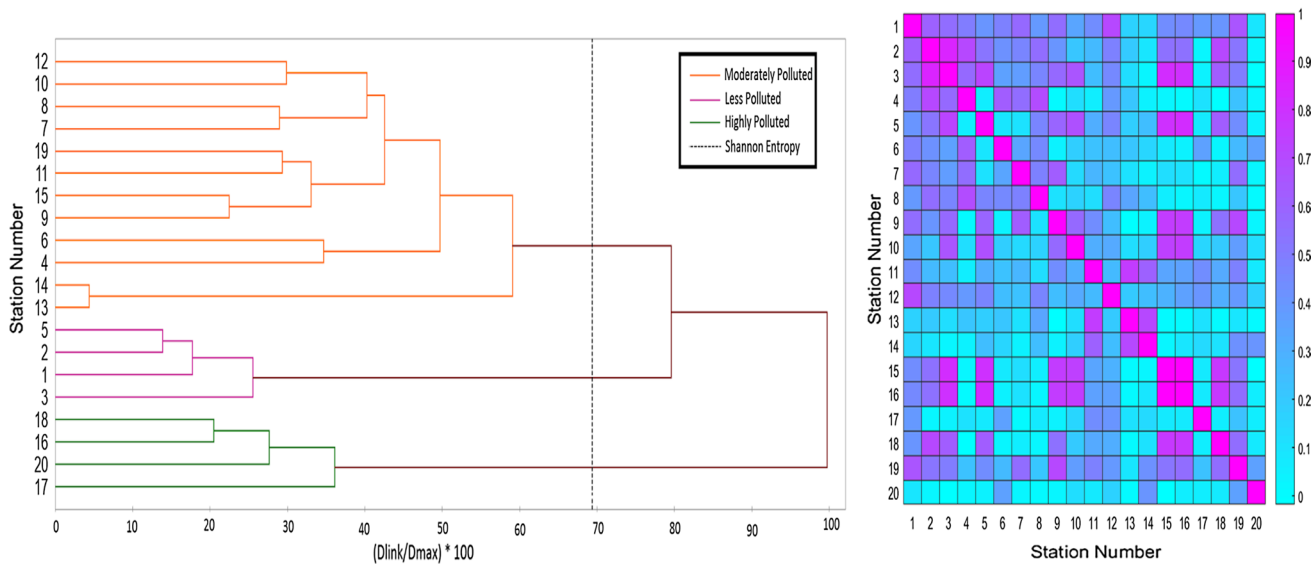


Fig. 4 Dendrogram of the clustered stations based on their WQ parameters (right) and correlation heatmap of each station (left)

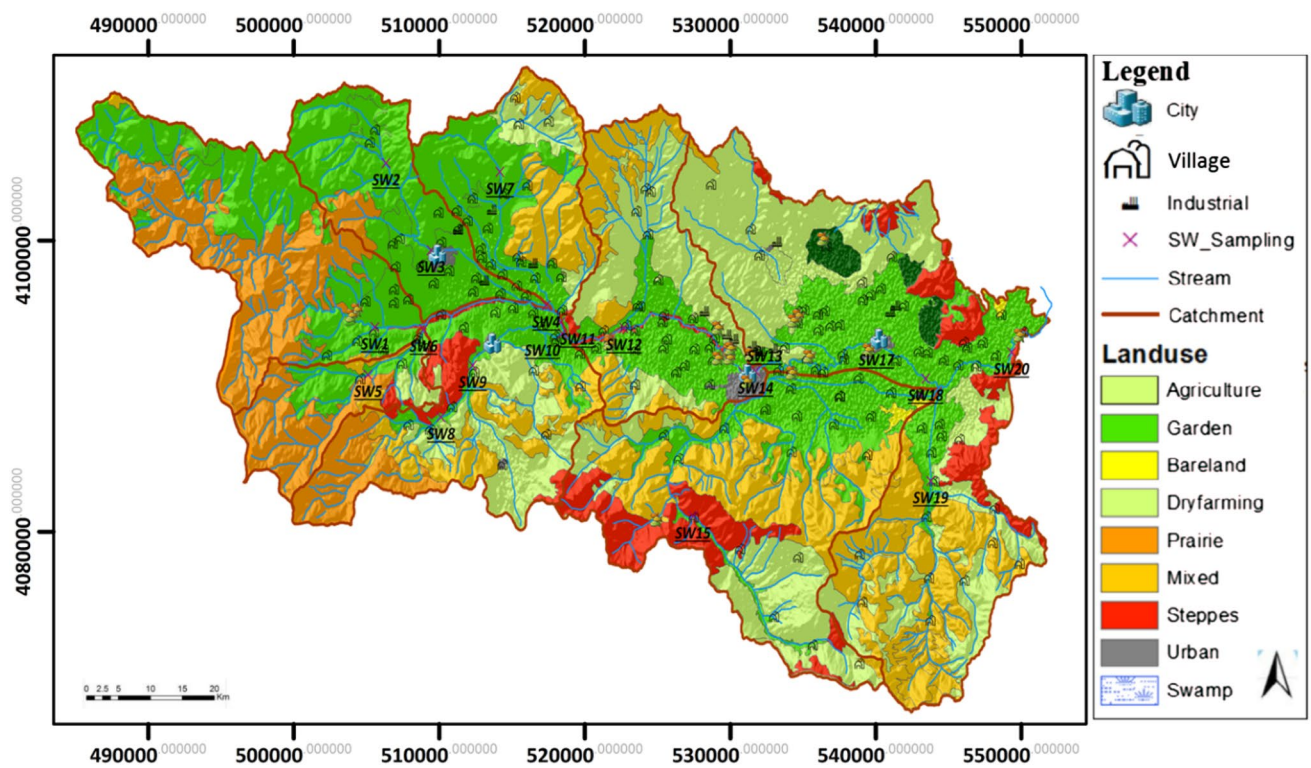


Fig. 5 Water quality-correlated parameters of Gadarchay Basin

(e.g., 13, 14, 10, 12, 4, 11 and 9) are primarily affected by pollution from agricultural sources, fish breeding centers, industrial practices and domestic wastewater. Stations in the HP category are generally located downstream of the Gadarchay River. The pollution sources in this area consist

mainly of domestic and industrial wastewater discharge, urban and agricultural runoff and fish breeding centers that use groundwater.

The dominant land use in the LP cluster is mainly garden and prairie, which likely contribute very little to the

pollution of the river. Considering the distance of the stations from the most upstream points in this cluster, natural attenuation and a low human population may support better WQ in this section. On the other hand, in the HP cluster, stations 20, 16, 17 and 18 are at the lowest elevation and are the most downstream points of the Gadarchay River. The higher population and the lack of wastewater treatment plants for most of the cities in this region are factors in the relatively poor WQ in this cluster. The results of this analysis and other studies suggest that CA can contribute considerably to the dimensionality reduction of stations (Alberto et al. 2001; Shrestha and Kazama 2007; Singh et al. 2005).

Temporal discrimination

The Spearman correlation test was used to assess whether to group temporal variations in seasonal form or wet/dry form. The correlation analysis results revealed that WQ parameters have a higher correlation with the seasonal form, i.e., from winter to autumn during a year. Among the considered parameters, seasonal variations were more closely correlated with WT, pH, EC, DO, PO₄, NO₃ and TP, with *p* values smaller than 0.05.

The raw data were grouped into four seasons and analyzed by the Box test (Chi-square and Fisher's F asymptotic approximation) to study the level of equality among the covariance of grouped input data. The results suggested that the within-class covariance matrix is not equal, with a significance level of $\alpha=0.01$. This is an essential step in using DA since the equality of the covariance matrix is a measure of whether the linear discriminant function (when

the within-class covariance matrix is equal) or quadratic discriminant function (when the within-class covariance matrix is not equal) is more appropriate for the model in question. Besides, the Box test was found to be oversensitive to sample size, suggesting that increasing the sample size may increase the bias from real results (Cohen 2008).

After grouping the raw data, DA was applied. Since the performance of the three versions of DA, i.e., standard, forward stepwise and backward stepwise, was similar according to the results of the confusion matrix, only the results of the backward DA were provided in the current study to avoid redundancy. Classification functions are often used to determine to which group each case most likely belongs. In Table 2, the classification functions of each variable in backward stepwise mode and their corresponding Wilks' Lambda and *p* value are provided.

Smaller Wilks' Lambda values suggest higher contributions to the model (Huberty 1994; IBM 2018). Contributing variables arranged from the highest to lowest Wilks' Lambda values are WT, TC, EC, NO₃, pH, FC, COD and NO₂.

Also, a confusion matrix was used to evaluate the performance of the DA. The confusion matrix counts the number of correct classifications versus misclassifications assigned by the DA. Table 3 shows the confusion matrix as a measure of DA performance. Note that standard DA outperformed the forward and backward stepwise modes.

The results indicate that the total performance of DA for discriminating between seasonal groups is about 68%. There are several possible explanations for the lower performance of DA in spring and autumn, for example the use

Table 2 Classification function obtained by Eq. 1 and temporal analysis of the Gadarchay River basin

Variables	Backward stepwise classification functions				Lambda	<i>p</i> value
	Autumn	Spring	Summer	Winter		
WT	0.887	1.336	1.834	0.696	0.347	<0.0001
TR ^a						
TSS						
pH	31.860	31.054	32.808	31.280	0.960	0.021
EC	0.008	0.006	0.007	0.007	0.946	0.004
COD	-0.011	-0.039	-0.056	-0.016	0.969	0.062
BOD						
DO						
NO ₂	-3.281	-3.063	-3.428	-3.158	0.972	0.080
NO ₃	-0.196	-0.394	-0.460	-0.147	0.957	0.016
PO ₄						
TP						
TC	0.000 ^b	0.000	0.000	0.000	0.943	0.003
FC	0.000	0.000	0.000	0.000	0.966	0.042
Constant	-130.864	-127.361	-149.408	-124.747		

^aValues in strikethrough indicate non-significant by backward stepwise DA

^bValues in 0.000 form are smaller than the tolerance value, i.e., 0.0001



Table 3 Confusion matrix of temporal standard DA

	Autumn	Spring	Summer	Winter	Total	% correct
Autumn	24	11	0	25	60	40.00
Spring	2	37	12	9	60	61.67
Summer	0	9	51	0	60	85.00
Winter	8	1	0	51	60	85.00
Total	34	58	63	85	240	67.92

of fertilizers, groundwater and agricultural pesticides, along with some macro-scale variables such as erosion. However, the main reason is suggested by the temperature box plots. Spring and summer are transitional seasons, as observable in Fig. 6. The first (the whiskers' upper bounds) and the third quartiles (the whiskers' lower bounds) of the autumn season cover almost all of the first and the third quartiles of the winter season. This is a possible reason for the misclassification of autumn as winter. On the other hand, the whiskers' spring season bounds are overlapped considerably by the

minimum bound of the summer season. This overlapping phenomenon occurs throughout almost all seasons for EC and DO, as seen in Fig. 6.

Spatial discrimination

The results of “Spatial clustering” section, spatial CA, were used to group the raw input data into three categories, i.e., LP, MP and HP. After grouping, they were used as the input data for spatial DA. The sites were used as dependent

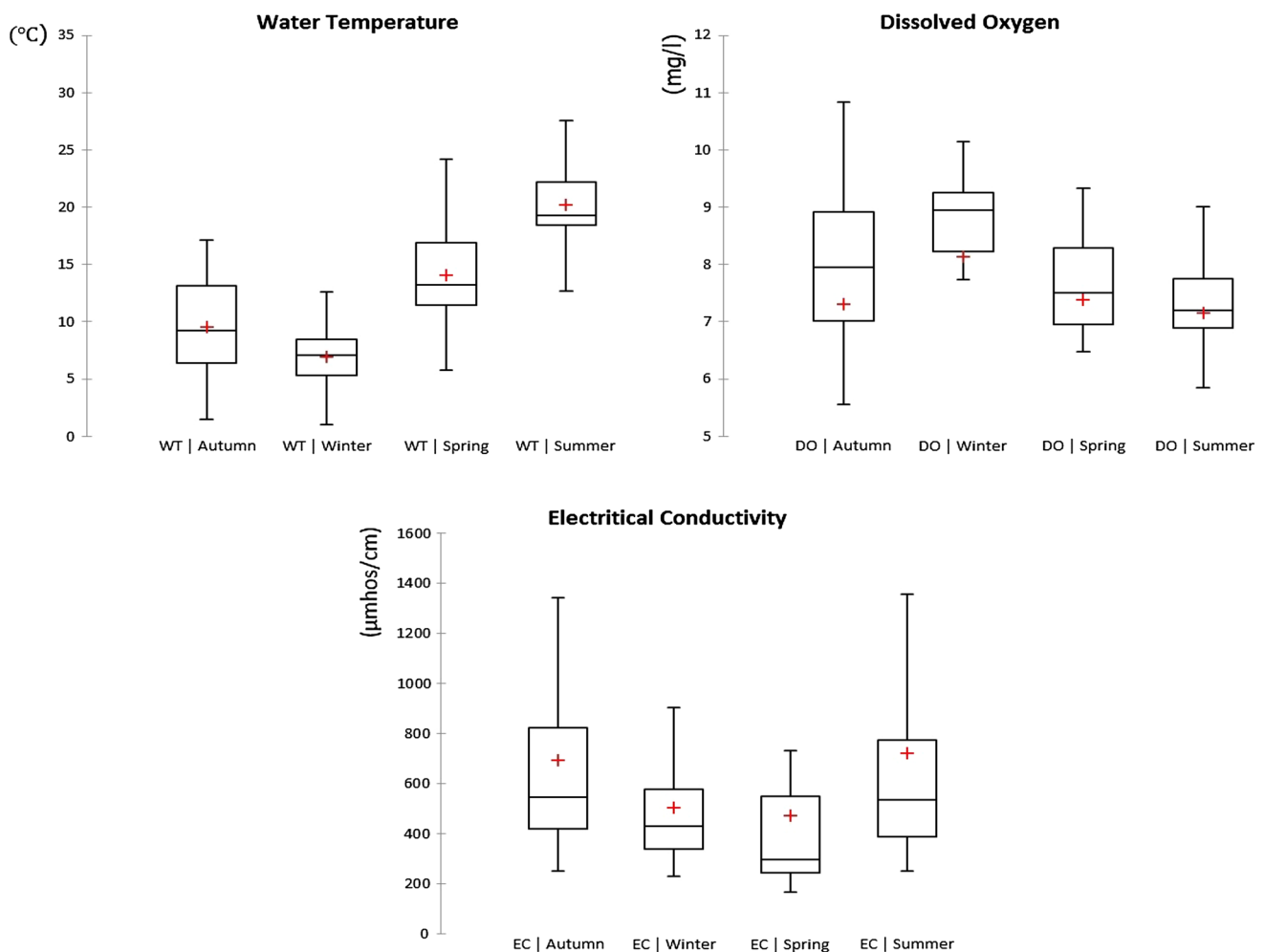


Fig. 6 Box plots of the most discriminating variables

variables, while the parameters were used as independent/explanatory variables. The classification functions and standardized canonical discriminant function coefficients are provided in Table 4.

As shown in Table 4, parameters arranged from most to least significant are: DO, TP, NO₃, EC, BOD, TR, WT, pH, TSS and TC. Further analysis suggests that TC contributes nothing but noise and is therefore insignificant since its *p* value is greater than 0.1, its Wilks' Lambda is the greatest among all parameters, and its univariate *F* value is lower than one at 0.561 (Huberty and Olejnik 2006).

Since the number of observations for the various groups of dependent variables differs, there is a risk of penalizing classes with a low number of observations in establishing the model (XLSTAT 2018b). To solve this, weight correction should be applied to the final results so that the performance

of each class is not overestimated or underestimated by the confusion matrix.

As shown in Table 5, although the overall performance of both versions did not vary significantly, the individual class performance of groups with lower members, i.e., LP and HP, was considerably overestimated. The HP and MP groups were penalized since they had the fewest members. This suggests a bias in non-weight-corrected results.

Figure 7 shows DO and NO₃, the two most significant variables, to help clarify the lower performance of spatial DA with respect to HP stations. This figure demonstrates the overlapping of the first and third quartiles of HP by MP stations for both DO and NO₃, which may contribute to the lower performance of the HP sites compared to the LP and MP groups.

PCA/FA results

Based on the literature and the CA outputs particular to the current study, PCA/FA was done on standardized data for the three regions, LP, MP and HP (Alberto et al. 2001; Singh et al. 2005). The input matrix was in [parameters × observations] form. The PCA results for the LP, MP and HP stations are provided in Table 6.

As Table 6 suggests, PCA results for the LP sites yielded five components explaining 76% of the total variance. Lower PCs extracted from the MP and HP sites accounted for 75% and 77% of the total variance, respectively. The relative importance of each PC is implied by its eigenvalue. Kim and Muller (1978) posit that eigenvalues greater than one are significant. Therefore, in the current study, only those PCs with eigenvalues greater than one undergo varimax rotation, as also suggested by Abdi and Williams (2010). Table 6 provides the results of varimax rotation for each spatial cluster, i.e., the LP, MP and HP stations.

The first five PCs in the LP group and the first four in the MP and HP groups were subjected to a varimax rotation based on the lowest eigenvalue, i.e., one, of each component. Since the results of varimax rotation due to the selection of multiple varifactors may not be one or two, squared cosine is used to avoid misinterpretation of PCs with lower squared cosine values due to projection effects. Squared cosine is also a measure of importance for each of the varifactors.

Table 4 Classification function obtained by Eq. 1 and spatial analysis of the Gadarchay River basin

Variables	Backward stepwise classification functions			Lambda	<i>p</i> value
	LP	MP	HP		
WT	0.013	0.122	0.099	0.939	0.001
TR ^a	-0.012	-0.004	-0.005	0.905	<0.0001
TSS	0.000	0.000	0.000	0.980	0.095
pH	35.710	36.660	35.338	0.959	0.007
EC	0.000	0.001	0.003	0.710	<0.0001
COD					
BOD	0.413	0.411	0.358	0.816	<0.0001
DO	1.543	1.112	0.514	0.559	<0.0001
NO ₂					
NO ₃	-0.163	0.269	-0.088	0.683	<0.0001
PO ₄					
TP	-9.118	-8.972	-7.814	0.683	<0.0001
TC	0.000	0.000	0.000	0.995	0.572
FC	0.000	0.000	0.000		
Constant	-142.881	-151.305	-136.266		

^aA strikethrough indicates an insignificant variable as determined by backward stepwise DA

^bValues in 0.000 form are smaller than the tolerance value, i.e., 0.0001

Table 5 Confusion matrix of spatial DA with and without weight correction

	Confusion matrix with weight correction					Confusion matrix without weight correction				
	LP	MP	HP	Total	Correct (%)	LP	MP	HP	Total	Correct (%)
LP	76.7	3.3	0.0	80	95.83	32	16	0	48	66.67
MP	19.4	56.1	4.4	80	70.14	13	128	3	144	88.89
HP	10.0	18.3	51.7	80	64.58	0	19	29	48	60.42
Total	106.1	77.8	56.1	240	76.85	45	163	32	240	78.75

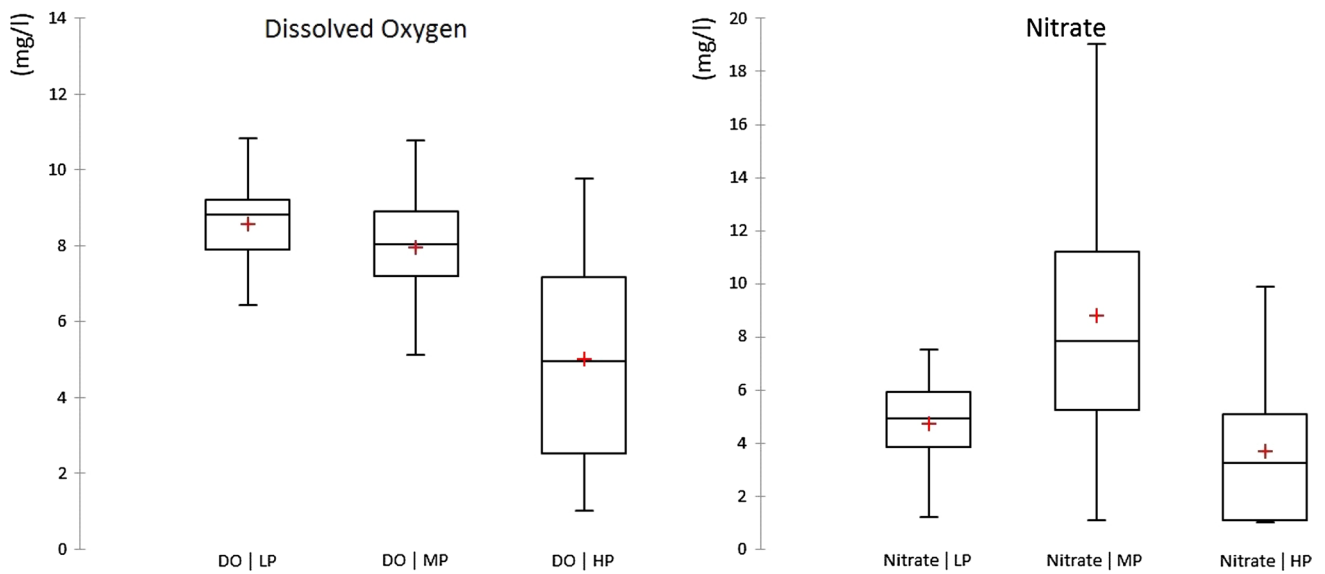


Fig. 7 Box and whisker plots of DO (left) and NO₃ (right)

Table 6 PCA results and factor loadings after varimax rotation of each group of the LP, MP and HP stations

	LP					MP				HP			
	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
PCA													
Eigenvalue	3.8	2.3	1.8	1.7	1.1	3.9	3.3	1.9	1.4	5.2	2.5	1.8	1.3
Variability (%)	27.3	16.5	12.7	12.1	7.7	27.5	23.6	13.6	10.3	37.1	18.1	12.8	9.2
Cumulative (%)	27.3	43.8	56.5	68.6	76.4	27.5	51.1	64.7	75.0	37.1	55.1	67.9	77.0
	VF1	VF2	VF3	VF4	VF5	VF1	VF2	VF3	VF4	VF1	VF2	VF3	VF4
FA													
WT	0.7^a	-0.2	0.1	0.4	0.4	0.1	0.1	0.6	0.4	-0.5	0.0	-0.1	0.6
TR	-0.1	0.3	0.9	0.0	0.0	-0.1	0.7	-0.4	0.0	-0.1	-0.1	0.9	-0.2
TSS	0.0	0.1	0.9	0.1	0.0	0.0	0.8	-0.1	0.0	-0.2	-0.1	0.9	0.1
pH	-0.1	0.1	-0.3	-0.2	0.4	0.9	-0.1	-0.2	0.1	0.2	0.1	0.3	0.5
EC	-0.1	0.1	-0.3	-0.1	0.6	0.0	-0.1	0.7	-0.3	0.2	0.2	-0.3	0.7
COD	0.0	0.9	0.2	0.0	-0.1	0.0	0.9	0.3	0.1	0.9	0.0	0.0	0.2
BOD	0.0	1.0	0.1	0.0	-0.1	0.0	0.8	0.4	0.1	0.9	0.0	0.0	0.2
DO	-0.7	0.1	-0.2	-0.4	-0.2	0.2	-0.3	-0.8	-0.2	-0.8	-0.1	0.2	-0.1
NO ₂	0.8	-0.2	0.1	-0.1	0.0	1.0	0.0	0.0	0.0	-0.1	0.8	-0.2	0.0
NO ₃	-0.2	0.3	-0.1	-0.1	-0.8	0.1	0.2	-0.2	-0.2	-0.7	0.0	0.1	0.4
PO ₄	0.9	0.2	-0.3	-0.1	0.0	1.0	0.0	0.0	0.0	0.9	0.2	-0.1	-0.2
TP	0.8	0.4	-0.2	-0.1	-0.1	1.0	0.0	0.1	0.0	0.9	0.2	-0.1	-0.1
TC	0.0	0.0	0.1	0.9	0.2	0.0	0.2	0.1	0.8	0.1	0.9	0.0	0.0
FC	0.0	0.0	-0.1	0.9	-0.2	0.0	0.0	0.0	0.9	0.1	0.9	0.0	0.2
Variance	21.9	16.3	15.3	14.0	10.5	27.2	21.0	14.1	12.7	35.2	17.0	14.5	10.3
CV ^b	21.9	38.2	53.5	67.5	78.0	27.2	48.2	62.3	75.0	35.2	52.2	66.7	77.0

^aIndicates values with the largest squared cosine and therefore higher importance

^bCumulative variance

Lower values indicate lower importance, and higher values indicate higher importance (Abdi and Williams 2010).

FA results of LP sites

This study used the results of both spatial and temporal DA for the first time to determine whether the loading of each VF was affected by spatial or temporal variations. Among the five VFs in the LP group, VF1 covers the greatest variance. As the cosine values of the VF1 were highest, the current study suggests that dissolved oxygen contributes the highest negative loading on WT, NO₂, PO₄ and TP. An increase in temperature can cause a decrease in oxygen solubility. Lower oxygen solubility can lead to a higher chance of eutrophication in phosphorus-rich aqueous environments. In addition, biochemical reactions are highly dependent on temperature: A 10 °C rise in temperature can cause reaction rates to double. Consequently, bacterial oxidation can lead to higher NO₃ concentrations (Ireland 2001). The concentration of NO₂ in aqueous solution is relatively lower than its reduced form (ammonia) or its oxidized form (NO₃). The contribution of wastewater discharge from upstream of the river in raising PO₄, TP, and NO₃ levels, and consequently, decreasing DO cannot be ignored. BOD and COD are indicators of the amount of organic pollution and the total amount of chemically oxidizable organic matter discharge into a river, respectively. Since bacteria are not capable of oxidizing all types of matter, COD is assumed to be higher than BOD in water bodies. Therefore, as VF2 indicates, BOD and COD have the highest positive loadings.

In VF3, TR was found to have the highest positive loading on TSS. A major source of TSS is the erosion of the upstream lands of LP sites. The source of the TR in LP sites, where wastewater discharge contribution is small, is the same as TSS. Consequently, it is expected that these variables have the highest positive loadings. In VF4, TC and FC were found to have positive loadings. As indicated by the relatively higher loading of TC compared to FC and the basic definition of TC, TC includes a wider range of bacteria than FC. This suggests that the primary source of the bacteria is environmental, not fecal. The lowest variance is covered by VF5. VF5 suggests the highest positive loadings on pH and EC and, conversely, negative loading on NO₂. Although pH and NO₃ may not have a direct influence on each other, a lower pH solution (more acidic) can change the kinetics of NO₃ to NO₂ reactions since nitrifying bacteria are very sensitive to pH (Holt et al. 1995; Skadsen and Sanford 1996; Watson et al. 1981).

FA results of MP sites

In MP sites, VF1 covers the largest amount (27.2%) of the variance of all the VFs. VF1 has strong positive loadings on

pH, NO₂, PO₄ and TP. When these loadings are compared to the same at LP sites, a more significant contribution of point source wastewater pollution is found in the MP areas. VF2 specifies loadings on TR, TSS, BOD, COD and NO₃. These loadings imply the existence of both wastewater and land-use pollution effects in this area. VF3 covers 14.1% of the variance, and WT, EC and DO have the greatest loadings. This VF illustrates the seasonal variations of WQ in this category. VF4 covers the lowest variance among VFs. Compared to VF4 in LP sites, the VF4 trend in MP sites is toward a relatively higher loading of FC than TC, which indicates a higher contribution of wastewater discharge at MP versus LP sites.

FA results of HP sites

VF1 of HP sites is dominated by domestic and industrial wastewater pollution. The highest loadings are observed in COD, BOD, DO, NO₃, PO₄ and TP. The impact of VF1 on these variables corresponds to their location in the downstream section of the Gadarchay River. VF2 covers 17% of the variance in HP sites and suggests that NO₂, TC and FC have the highest loadings. This may be a consequence of nitrification along the river. VF3 covers 14.5% of the variance and indicates that TSS and TR have the highest positive loading on each other. Since HP-suspended solids from agricultural and garden land use may be carried down the river, it is expected that the highest loadings will be between TSS and TR. VF4 covers 10.3% of total variance; WT, pH and EC have the highest loadings and indicate a seasonal variation in the Gadarchay River.

Temporal NLPCA

Since NLPCA is a data-driven method that demands a considerable amount of data for the training process, the input data were not divided into the three major clusters, i.e., LP, MP and HP, and were not fed into the AANN. Instead, the whole cluster was used to train the model, and the CA results were used to label the data. This does not mean that it is impossible to divide data and feed it into the AANN model. Despite the difficulties in estimating the true total variance during the reconstruction process, the PC's variance was not found to be overestimated. PCA preprocessing was done on the raw data, not for dimensionality reduction (dimensionality was still 14), but for rotating the space data by PCA. Weight initialization was selected as linear. Unlike its default value, which is random weight initialization, the optimization process by this method, i.e., linear weight initialization, was found to be more efficient, consistent and time-saving. Table 7 provides the three PCs and their corresponding variances as extracted from each of the spatial groups.

Table 7 The extracted components and the performance metrics of temporal NLPCA

	Variance	Cumulative	Training error (MSE)	Train <i>R</i> coefficient	Test error (MSE)	Test <i>R</i> coefficient
PC1	50.12	50.12	0.14	0.92	0.12	0.89
PC2	25.37	75.49				
PC3	5.29	80.78				

As described in Table 7, NLPCA extracted three PCs, covering more than 97% of the total variance. This suggests that the NLPCA method can be considered a good alternative to the PCA method, which, under optimal conditions, extracted 77% of the total variance with four PCs. Figure 8 gives an illustrative visualization of the NLPCA and its extracted PCs.

As described in “NLPCA” section, under NLPCA the extracted PCs are curves in data space. Therefore, one cannot describe a global ranking of the variables for the whole period. This is evident in the current study, given that WQ parameters along the river during a month or season can change considerably, as discussed in “Temporal discrimination” section. To this end, the tangents or the derivatives for the component values were calculated over the entire study period.

There are challenges in accurately discerning which variables rank highest in different seasons. To address this, the box plot “dz” over 1 year of the sampling period is provided in Fig. 9. This approach may answer the question as

to which variables can generally be considered significant according to the proposed NLPCA method. By performing a normality test, it was found that “dz” does not follow the normal distribution. Hence, unlike the median, the mean value of each parameter is not a good representative of the whole dataset.

Referring to Fig. 9 and the median value of each parameter, WT is the most significant variable and, in all cases, the median is skewed toward the third quartile. This agrees with the fact that WT can potentially affect all other variables in water bodies, especially DO, TR and NO₃. FC and TC in the summer and spring seasons have greater median values than in the winter and autumn seasons. This makes sense knowing that TC and FC populations are highly affected by temperature, especially in the winter and autumn seasons. In these seasons, TSS is more affected by the water flow rate and its density. These results are consistent with Gurjar and Tare (2019); Shrestha and Kazama (2007); Sun et al. (2019), and with the temporal DA results. For better readability, Table 8 shows the median of

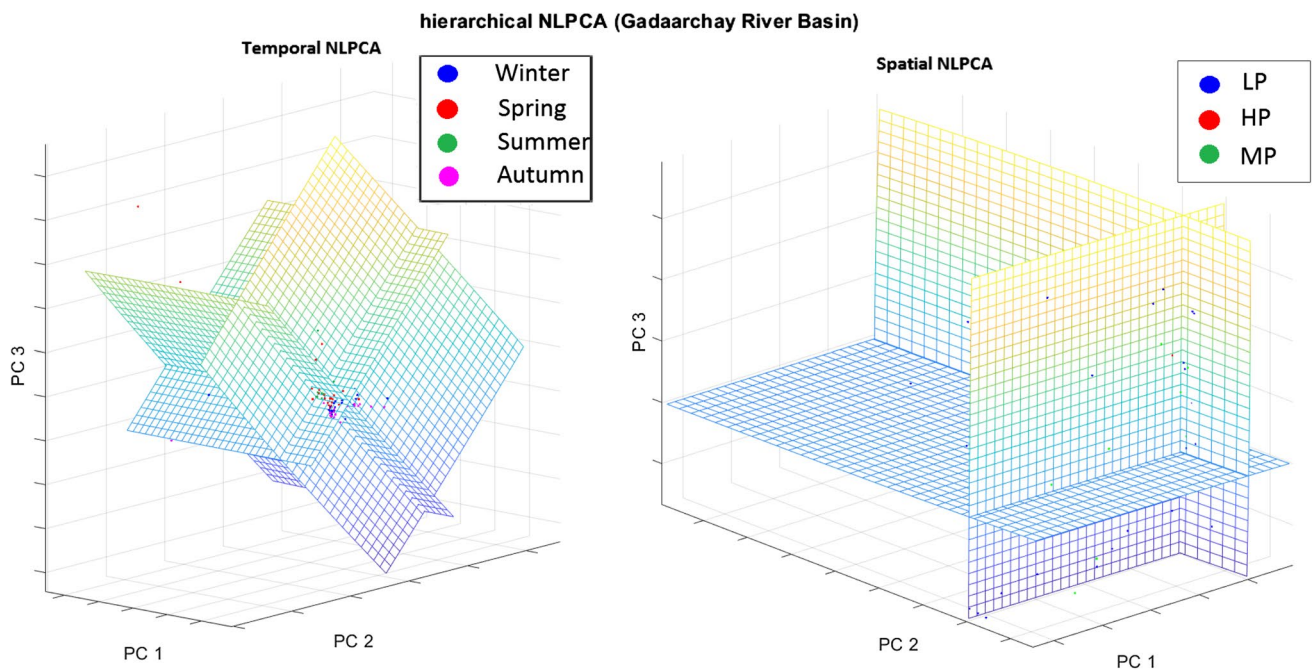


Fig. 8 Extracted components by the bottleneck NLPCA process with linear weight initialization captured at iteration 300

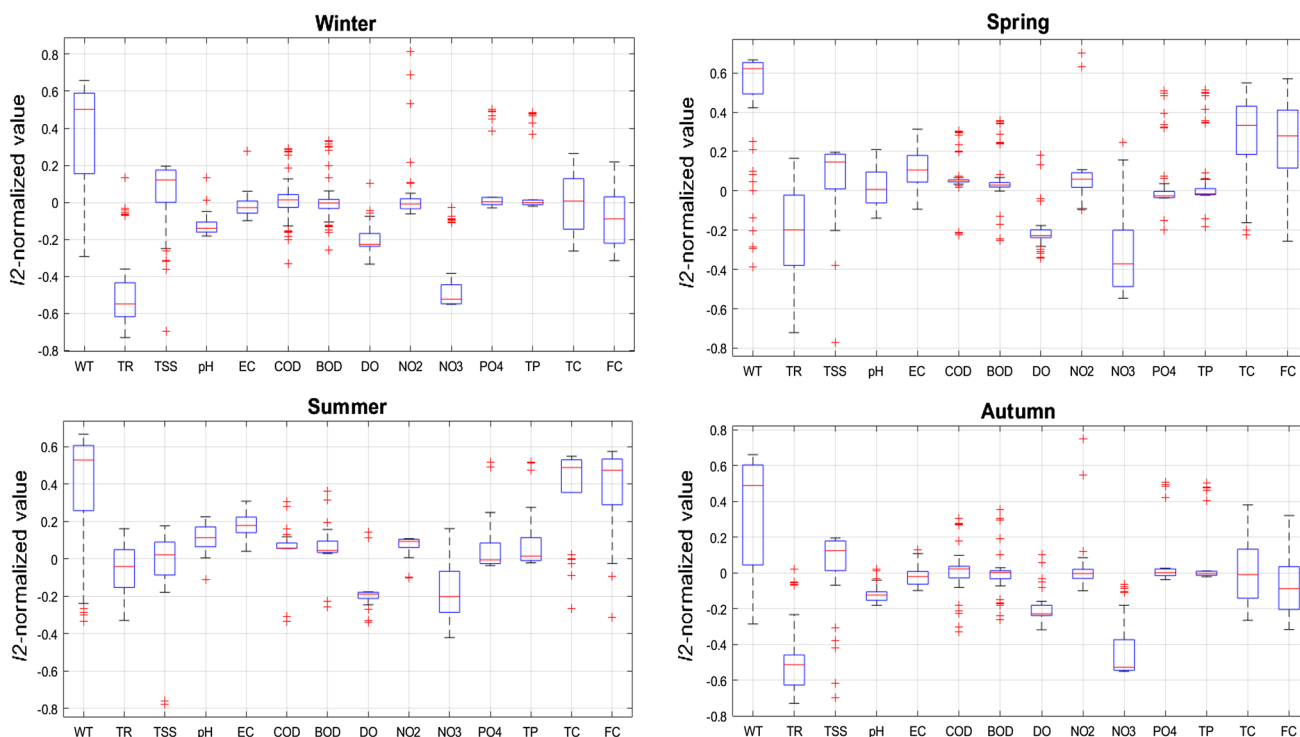


Fig. 9 Box and whisker plots of dz value calculated by the temporal NLPCA

Table 8 Temporal ranking of each variable based on the median value of l2-normalized dz

Parameters	Winter	Spring	Summer	Autumn
WT	2.18	2.16	1.64	2.16
TR	-1.81	-1.03	-0.6	-1.71
TSS	0.7	0.25	-0.45	0.66
pH	-0.3	-0.17	0.02	-0.3
EC	0.11	0.17	0.31	0.11
COD	0.24	0.04	-0.23	0.24
BOD	0.2	-0.05	-0.27	0.16
DO	-0.62	-1.04	-1.22	-0.66
NO ₂	0.21	0.05	-0.05	0.21
NO ₃	-1.78	-1.69	-1.39	-1.75
PO ₄	0.31	-0.26	-0.46	0.33
TP	0.29	-0.21	-0.37	0.31
TC	0.29	1.03	1.5	0.29
FC	-0.07	0.84	1.43	-0.04

the l2-normalized value of each parameter for each season. As a side note, this does not indicate that short median values are not significant since any of these variables can be the most significant at some points of time. These results are simply a general measure of significance.

Spatial NLPCA

Table 9 provides performance metrics of the spatial NLPCA performed on the dataset. According to this table, spatial NLPCA explains 80.34% of the variance by three components. These results challenge the results of spatial PCA and DA.

Figure 10 shows box and whisker plots of “dz” at each station. According to this figure, WT is generally the most significant variable at all stations. Although the box plot of each station is rather similar, they have a different distribution. For example, PO₄, TP, BOD and COD in HP stations have a broader interquartile range than at LP and MP stations. While the interquartile range of DO is more

Table 9 The extracted components and the performance metrics of spatial NLPCA

	Variance	Cumulative	Training error (MSE)	Train R coefficient	Test error (MSE)	Test R coefficient
PC1	61.24	61.24	0.13	0.92	0.1	0.92
PC2	14.19	75.43				
PC3	4.91	80.34				

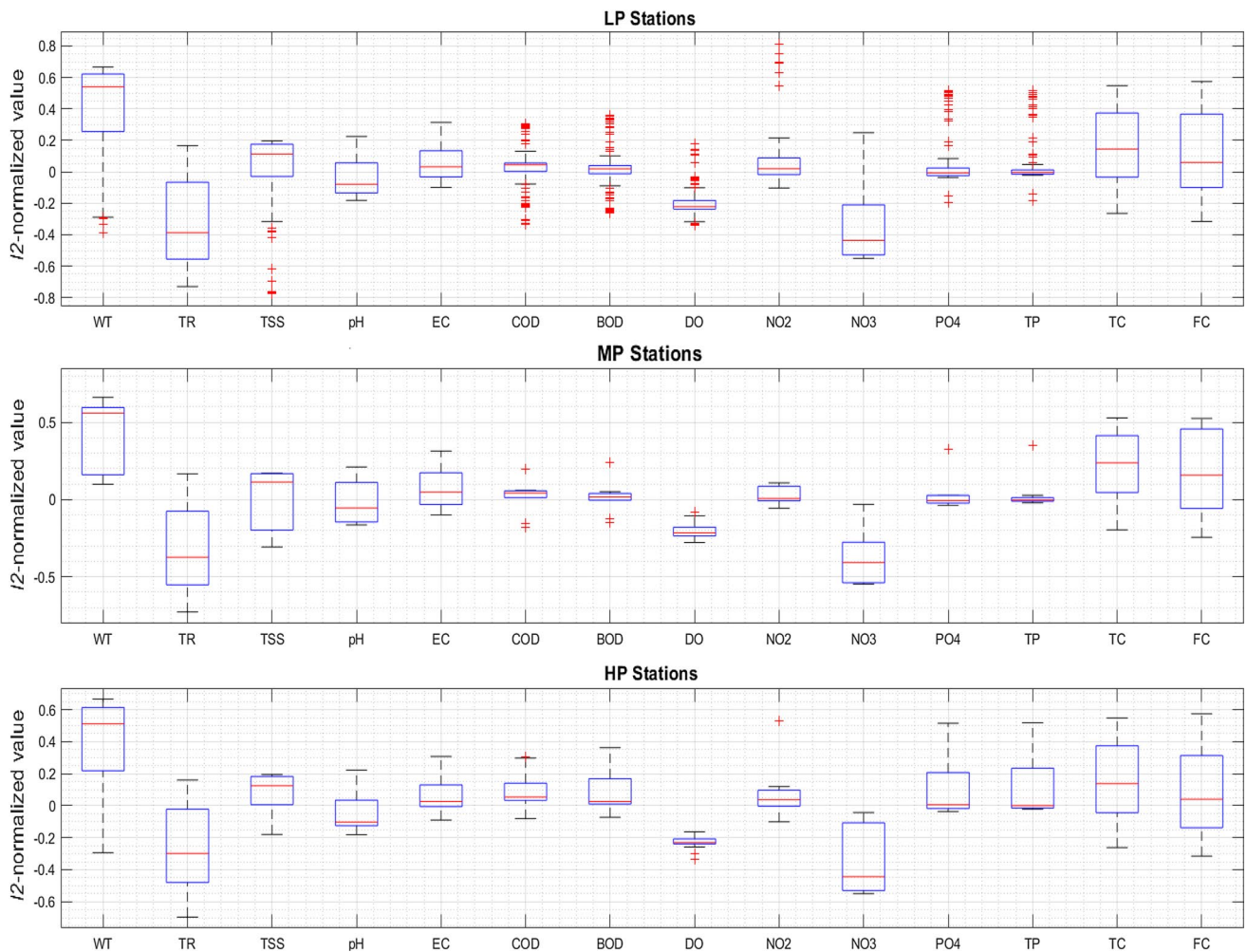


Fig. 10 Box and whisker plots of d_z values obtained by spatial NLPCA

limited than at LP and MP stations, in contrast, outliers at LP stations are more abundant than at MP and HP stations. This may be due to higher members in LP stations and hence, higher diversity, especially for TSS, COD, BOD, DO, NO_2 , PO_4 and TP. The distribution of outliers is of importance. For example, the outlier of TSS lies under the minimum, while for NO_2 it lies above the maximum.

Table 10 explains the median value of each variable in spatial NLPCA. Based on this table, TSS at LP and HP stations have larger median values. According to the PCA source identification results, the anthropogenic pollution effect may contribute more to HP than LP stations. The value for COD at HP stations is greater than at MP and LP stations. The value for BOD is greater at MP stations than LP and HP stations. TR in all three groups has significant negative value reflecting the impact of erosion in all LP, MP and HP stations. DO and NO_3 in all three clusters have negative median d_z value, while WT and TC have strong positive median d_z value. High values of NO_3 in all three clusters indicate

Table 10 Spatial ranking of each variable based on the median value of $1/2$ -normalized d_z

Parameters	LP	MP	HP
WT	2.1	2.05	2.13
TR	-1.44	-1.42	-1.25
TSS	0.32	0.09	0.34
pH	-0.25	-0.07	-0.3
EC	0.14	0.21	0.12
COD	0.15	0.02	0.21
BOD	0.1	0.15	0.13
DO	-0.81	-0.88	-0.97
NO_2	0.14	0.11	0.14
NO_3	-1.67	-1.62	-1.72
PO_4	0.06	0.11	0.04
TP	0.07	0.11	0.05
TC	0.59	0.82	0.59
FC	0.29	0.54	0.24

agricultural drainage (fertilizers and manure), a decrease in DO concentration, and an increase in TC concentration.

Conclusion

In the current study, several different multivariate analysis methods, CA, PCA/FA and DA, were compared with NLPCA, an AANN-based technique. The effects of temporal and spatial variation on the WQ parameters of the Gadar-chay River basin in Iran were evaluated with the mentioned techniques. The spatial grouping of 20 sampling stations was determined using CA on standardized data. AHC provided three homogenous groups of objects on the basis of their descriptions by a set of WQ parameters. The CA results then used spatially grouped variables as inputs for DA. A discussion of the most suitable way to interpret the results of DA when using different sampling sizes followed. CA divided the stations into three classes: LP, MP and HP.

Spatial DA was performed on raw data inputs that were divided into the three mentioned groups based on CA results. By analyzing the p values and applying Wilks' Lambda analysis to spatial DA, parameters were ranked from most to least significant as follows: DO, TP, NO_3 , EC, BOD, TR, WT, pH and TSS. With a performance of 95.83%, the best performance was observed for the LP stations as identified via the confusion matrix with the weight correction technique. For the MP and HP stations, the performance of spatial DA was observed to be 70.14% and 64.58%, respectively. The overall performance was 76.85%. By comparing the results of spatial DA with and without weight correction, the effect of each individual class size on the estimation of real performance of spatial DA was discovered.

Spearman correlation analysis was used to compare the dry/wet classification and the seasonal classification. The results were comparable; however, the seasonal form was observed to have a higher correlation with the data. Temporal DA was also performed on the raw data, which were grouped into four seasonal classes, i.e., autumn, winter, spring and summer. Ranked from most to least significant variables in seasonal DA form, the variables were WT, TC, EC, NO_3 , pH, FC, COD and NO_2 . Since the group size was equal, weight correction was not applied in temporal DA. The best performance was observed for both summer and winter with 85% correct classification. Superior classification in these groups was found to be due to improved discrimination ability achieved through using temperature maximums and minimums. For autumn and spring, the performance was 40% and 61.67%, respectively.

PCA/FA was also performed on the standardized spatially divided datasets. The results of both spatial and temporal DA were used to better interpret the results of PCA/FA. PCA extracted five PCs for the LP stations covering approximately 76% of the total variance, four PCs for the MP stations covering approximately 75% of the total variance, and four PCs for the HP stations covering approximately 77% of total variance—the best performance in this group. In addition, the FA results helped to identify the origin of pollution and suggested that LP stations were mainly affected by erosion, MP stations were more affected by anthropogenic pollution than erosion, and HP stations were primarily affected by anthropogenic pollution.

NLPCA is capable of processing nonlinearities with more accuracy than PCA/FA, DA and CA. For the entire dataset, temporal and spatial NLPCA extracted only three PCs defining approximately 80.78% and 80.34% of the total variance, respectively. This method differs from PCA/FA in that it extracts components dynamically, which can result in the identification of significant variables during the sampling period. NLPCA was capable of specifying the significance of each variable in each time step. NLPCA could discriminate each variable in different seasons and different locations separately.

Based on the results of this study, it can be concluded that NLPCA is a potentially reliable method for river WQ assessment. Also, this study contributed to certain practical details of the implementation of DA and estimation of its real performance, hence, avoiding overestimation due to sample size differences. It is recommended that further research using NLPCA be conducted to provide a better understanding of WQ interactions. The introduced methodology illustrates the usefulness of NLPCA, and its results can help decision-makers to analyze WQ parameters along the river both spatially and temporally more effectively. Also, precautionary measures based on pollution source identification can be undertaken to ensure the quality of drinking water.

Acknowledgements The authors would like to extend their appreciation to the Lar Consulting Engineers Company for providing introductory data and helping them during the sampling period (Lar Consulting Engineers 2018). Authors would like to appreciate anonymous reviewers who improved the quality of the manuscript.

References

Abdi H, Williams LJ (2010) Principal component analysis. *Wiley Interdiscip Rev Comput Stat* 2:433–459

- Alberto WD, del Pilar DM, Valeria AM, Fabiana PS, Cecilia HA, de los Angeles BM (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia River Basin (Córdoba–Argentina). *Water Res* 35:2881–2894
- AlKhader AM, Qaryouti MM, Okasheh TYM (2019) Effect of nitrogen on yield, quality, and irrigation water use efficiency of drip fertigated grafted watermelon (*Citrullus lanatus*) grown on a calcareous soil. *J Plant Nutr* 42:737–748
- Balázs B, Bíró T, Dyke G, Singh SK, Szabó S (2018) Extracting water-related features using reflectance data and principal component analysis of Landsat images. *Hydrol Sci J* 63:269–284
- Chapman DV (1996) Water quality assessments: a guide to the use of biota, sediments and water in environmental monitoring. World Health Organization, Geneva
- Cohen BH (2008) Explaining psychological statistics. Wiley, New York
- Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst* 2:303–314
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7:179–188
- Fard RF, Naddafi K, Hassanvand MS, Khazaei M, Rahmani F (2018) Trends of metals enrichment in deposited particulate matter at semi-arid area of Iran. *Environ Sci Pollut Res* 25:18737–18751
- Gabris MA, Jume BH, Rezaali M, Shahabuddin S, Nodeh HR, Saidur R (2018) Novel magnetic graphene oxide functionalized cyanopropyl nanocomposite as an adsorbent for the removal of Pb(II) ions from aqueous media: equilibrium and kinetic studies. *Environ Sci Pollut Res* 25:27122–27132
- Gulgundi MS, Shetty A (2018) Groundwater quality assessment of urban Bengaluru using multivariate statistical techniques. *Appl Water Sci* 8:43
- Gurjar SK, Tare V (2019) Spatial–temporal assessment of water quality and assimilative capacity of river Ramganga, a tributary of Ganga using multivariate analysis and QUEL2K. *J Clean Prod* 222:550–564
- Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Res* 34:807–816
- Holt D, Todd RD, Delanoue A, Colbourne JS (1995) A study of nitrite formation and control in chloraminated distribution systems. In: *Proceedings of the 1995 AWWA WQTC*, New Orleans
- Hsieh WW (2004) Nonlinear multivariate and time series analysis by neural network methods. *Rev Geophys*. <https://doi.org/10.1029/2002rg000112>
- Huberty CJ (1994) Applied discriminant analysis. vol 519.535 HUB. CIMMYT
- Huberty CJ, Olejnik S (2006) Applied MANOVA and discriminant analysis, vol 498. Wiley, New York
- Hunt ND, Hill JD, Liebman M (2019) Cropping system diversity effects on nutrient discharge, soil erosion, and agronomic performance. *Environ Sci Technol* 53:1344–1352
- IBM (2018) Tests of equality of group means. https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/tutorials/discrim_bankloan_groupmean.html. Accessed 30 June 2018
- Ireland E (2001) Parameters of water quality—interpretation and standards. Wexford, EPA, ISBN 133
- Johnson RA, Wichern D (1992) Applied multivariate statistical analysis. Prentice Hall, Englewood Cliffs
- Khaledian Y, Ebrahimi S, Natesan U, Basatnia N, Nejad BB, Bagmohammadi H, Zeraatpisheh M (2018) Assessment of water quality using multivariate statistical analysis in the Gharaso river, Northern Iran. In: Sarma AK, Singh VP, Bhattacharjya RK, Kartha SA (eds) *Urban ecology, water quality and climate change*. Springer, Cham, pp 227–253. https://doi.org/10.1007/978-3-319-74494-0_18
- Kim J-O, Muller C (1978) Introduction to factor analysis: what it is and how to do it, Series: Quantitative Applications in the Social Sciences. Sage, Beverly Hills
- Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. *AICHE J* 37:233–243
- Laar Consulting Engineers (2018) Gadarchay WQ Assessment Project. <http://www.lar-co.com/HomePage.aspx?TabID=4786&Site=lar-co&Lang=en-US>. Accessed 08 Aug 2018
- Li T, Li S, Liang C, Bush RT, Xiong L, Jiang Y (2018) A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. *J Clean Prod* 190:1–11
- Li P, Tian R, Liu R (2019) Solute geochemistry and multivariate analysis of water quality in the Guohua phosphorite mine, Guizhou Province, China. *Expo Health* 11:81–94
- Martinez AM, Kak AC (2001) PCA versus LDA. *IEEE Trans Pattern Anal Mach Intell* 23:228–233. <https://doi.org/10.1109/34.908974>
- MATLAB (2017) MATLAB and Statistics Toolbox Release 2017a. The MathWorks Inc, Natick
- Moya CE, Raiber M, Taulis M, Cox ME (2015) Hydrochemical evolution and groundwater flow processes in the Galilee and Eromanga basins, Great Artesian Basin, Australia: a multivariate statistical approach. *Sci Total Environ* 508:411–426. <https://doi.org/10.1016/j.scitotenv.2014.11.099>
- Noshadi M, Ghafourian A (2016) Groundwater quality analysis using multivariate statistical techniques (case study: Fars province, Iran). *Environ Monit Assess* 188:419
- Peña-Gallardo M et al (2019) Complex influences of meteorological drought time-scales on hydrological droughts in natural basins of the contiguous United States. *J Hydrol* 568:611–625. <https://doi.org/10.1016/j.jhydrol.2018.11.026>
- Rakotondrabe F, Ngoupayou JRN, Mfonka Z, Rasolomanana EH, Abolo AJN, Ako AA (2018) Water quality assessment in the Bétaré-Oya gold mining area (East-Cameroon): multivariate statistical analysis approach. *Sci Total Environ* 610:831–844
- Razmkhah H, Abrishamchi A, Torkian A (2010) Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: a case study on Jajrood River (Tehran, Iran). *J Environ Manag* 91:852–860
- Scholz M, Fraunholz M, Selbig J (2008) Nonlinear principal component analysis: neural network models and applications. In: Gorban AN, Kégl B, Wunsch DC, Zinovyev AY (eds) *Principal manifolds for data visualization and dimension reduction*. Springer, Berlin, Heidelberg, pp 44–67
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. *Environ Model Softw* 22:464–475
- Singh KP, Malik A, Sinha S (2005) Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study. *Anal Chim Acta* 538:355–374
- Skadsen J, Sanford L (1996) The effectiveness of high pH for control of nitrification and the impact of ozone on nitrification control. In: *Proceedings of the 1996 AWWA water quality technology conference*



- Suhaila J, Yusop Z (2017) Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoret Appl Climatol* 129:229–242
- Sun X, Zhang H, Zhong M, Wang Z, Liang X, Huang T, Huang H (2019) Analyses on the temporal and spatial characteristics of water quality in a seagoing river using multivariate statistical techniques: a case study in the Duliujian river, China. *Int J Environmental Res Public Health* 16:1020
- Vega M, Pardo R, Barrado E, Debán L (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Res* 32:3581–3592
- Watson SW, Valois FW, Waterbury JB (1981) The family nitrobacteraceae. In: Starr MP, Stolp H, Trüper HG, Balows A, Schlegel HG (eds) *The prokaryotes: a handbook on habitats, isolation, and identification of bacteria*. Springer, Berlin, Heidelberg, pp 1005–1022. https://doi.org/10.1007/978-3-662-13187-9_80
- XLSTAT (2016) XLSTAT. <https://www.xlstat.com/en/>. Accessed 30 June 2018
- XLSTAT (2018a) AHC. <https://www.xlstat.com/en/solutions/features/agglomerative-hierarchical-clustering-ahc>. Accessed 30 June 2018
- XLSTAT (2018b) Discriminant Analysis (DA). <https://www.xlstat.com/en/solutions/features/discriminant-analysis-da>. Accessed 30 June 2018

