Dynamic Multivariate Analysis for Pollution Profiling and Abatement Recommendation in La Buong Watershed of Vietnam

Luan Hong Pham (hongluanosgeo@gmail.com)  
Center of Water Management and Climate Change  https://orcid.org/0000-0003-3435-1338

Dung Duc Tran  
Center of Water Management and Climate Change

Hien Dieu Thi Le  
Thu Dau Mot University

Toan Quang Dinh  
Department of Natural Resources and Environment of Thanh Hoa, Vietnam

Quan Hong Nguyen  
Institute of Circular Economy Development

Khoi Nguyen Dao  
Faculty of Environment, University of Science, Ho Chi Minh National University of Vietnam

Hue Thanh Thi Nguyen  
Center of Water Management and Climate Change

Au Hai Nguyen  
Institute of Environment and Resources, National University of Ho Chi Minh city

Research Article

Keywords: Cluster analysis, Dynamic Factor Analysis, La Buong watershed, Water pollution, Water quality management

DOI: https://doi.org/10.21203/rs.3.rs-367101/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Analysis of temporal patterns of high-dimensional time-series water quality data is essential in informing better pollution management. In this study, Dynamic Factor Analysis (DFA) and Cluster Analysis (CA) were adopted to analyze time-series water quality data monitored at five stations SB1, SB2, SB3, SB4 and SB5 on La Buong river in the Southern Vietnam. Application of DFA identified two temporal patterns in SB1 and SB2 and three temporal patterns in SB3, SB4 and SB5. Analysis of factor loadings of water variables revealed run-off-driven patterns with the contribution of Total Suspended Solid (TSS), turbidity or Fe at all stations. The association of other variables like BOD$_5$, COD at SB1, SB2, SB4, and SB5 to this run-off pattern exposed their sharing of common driver. On the contrary, separation of variables like Phosphate (PO$_4^{3-}$) in SB3, SB4 and SB5 from run-off pattern suggested their local point-source origin. The derived factors from DFA were later used in time-point CA to explore temporal distribution of pollution intensities. Comparisons between clusters' value and two regulatory benchmarks A2 and B1 for drinking and irrigation water respectively suggested land-use approach for abating TSS, Fe and BOD$_5$, COD at most sites. The control of point sources of BOD$_5$ and COD pollutants is needed at SB3 along with PO$_4^{3-}$, Ammonium (NH$_4^+$) and *Escherichia coli* (E.coli) at SB1 and SB4.

1. Introduction

Long-term monitoring of water quality in a watershed is essential to understand hydro-chemical properties and water resource pollution (Vega et al. 1998; Felipe-Sotelo et al. 2007). Continuous surveying of water quality over time also helps keep track of the dynamics of water quality. Insights from analyzing water quality time series can unravel the link between anthropogenic and natural drivers and water quality (see Diamantini et al. 2018 for instances) and inform appropriate policies towards better managing quality of water resource. However, long-term monitoring program of water quality usually originate multi-dimensional time-series datasets which require sophisticated method for analysis (Dixon and Chiswell 1996).

Univariate statistical analysis is the approach used to describe each water quality variable's value distributions and, to its best, the extent they are correlated. However, shortcomings of univariate method appear in case comprehensive interpretation of dataset is needed because it is unable to capture the complexity of data structure (Le et al. 2017). Insights generated from univariate methods can hardly support integrated water quality management, for it usually targets specific types of pollution. In such context, composite indices like Water Quality Index (WQI) were formulated as a measure considering multipule water parameters. The index was proved to fit the objectives of communicating about the health status of water bodies rather than the structural composition of water datasets (Pesce and Wunderlin 2000; Wunderlin et al. 2001). Other information can be derived from datasets like sources of pollution would also be hidden in WQI after normalization and weighting of water parameters (Le et al. 2017).

Multivariate statistical analysis with two popular techniques of factor analysis (FA) and cluster analysis (CA), is an approach to handle multi-constituent dataset. The FA explores latent structures that explain the correlation between variables (Liu et al. 2003; Kowalkowski et al. 2006; Budaev 2010). The primary objectives of FA are to capture maximum variabilities in the original dataset with a minimum number of factors and therefore, minimize information redundancy (Kowalkowski et al. 2006) while CA is a multivariate technique of grouping observations into clusters possessing high internal homogeneity and high external heterogeneity between groups (Shrestha and Kazama 2007; Le et al. 2017). FA combined with CA has multiple applications like explaining temporal and spatial characteristics of water quality (Vega et al. 1998; Felipe-Sotelo et al. 2007; Juahir et al. 2011; Vialle et al. 2011; Magyar et al. 2013) and exploring sources driving water quality (Singh et al. 2005; Le et al. 2017). The dynamic factor analysis (DFA) was suggested as FA for time-series data for its ability in exploring temporal co-variabilities in dataset (Molenaar 1985). Understanding this dynamic relationships in time-series data had many applications especially in ecological and environmental researches, like noted works by Y. M. Kuo et al. (2014) and Y-M. Kuo and Chang (2010) in air and groundwater pollution, Zuur, Tuck, and Bailey (2003) in fishery, Kisekka et al. (2013) and Muñoz-Carpena, Ritter, and Li (2005) in water pollution.

La Buong river is a downstream branch of the Dong Nai river system, one of the largest and most important national river basins in Vietnam (Khoi et al. 2019). The La Buong watershed has been undergone rapid socio-economic development, imposing high pressures on water environment (Fig. 1). Drivers of such pressures include population increase, urbanization and agricultural intensification to name a few. Khoi et al. (2019) used a modelling approach with SWAT model to show that stream flow and water quality of the La Buong river would be impaired by climate change and land use and land covers (LULCs) changes. However, the application of SWAT model in assessing water quality has many uncertainties especially in area featured by diffused and heterogeneous sources of pollution (Glavan and Pinter 2012). The modelling approach also aimed at predicting water quality rather than understanding driving forces of pollutions. To address complexity of water pollution in the area, it is essential to adopt a data-driven multivariate time-series technique. Therefore, in this research, we applied dynamic multivariate technique for pollution characterization and later suggested policies for local authorities in water quality management. The paper's objective is twofold: i) to understand temporal patterns of water quality parameters and their drivers, and ii) to assess pollution intensities of pollution patterns and suggest abatement solutions.

2. Material And Method
2.1 Study area, sample collection and data preprocessing

La Buong river traverses a distance of 52km with a catchment of 478.5 km² (Fig. 1). The terrain of the catchment area is relatively flat, with an average elevation of 93m. The catchment is affected by a tropical monsoon climate, with the rainy season from May to October and the dry season from November to April (Khoi et al. 2019). The mean annual rainfall is 1999 mm with 87–93% in the rainy season and 7–13% in the dry season. The yearly average evaporation is about 1155 mm, and the average air temperature is 26°C. The La Buong catchment is characterized by high agricultural activities. Rhodic Ferralsols and Ferric Acrisols are the dominant soil in the catchment taking up to 75% of catchment area.

Five monitoring sites were distributed along La Buong river. Figure 1 shows the locations of these sites along with borders of sub-basins. Figure 2 presented land-use characteristic at each sub-basin. Four sites SB1, SB2, SB4 and SB5 are located near residential area. The most intensive level of built-up was observable in SB1 (Fig. 2a) and SB5 (Fig. 2e). Besides built-up, agriculture is the main land-use type in SB1, SB2 (Fig. 2b) and SB4 (Fig. 2d) sub-basin. Lastly, SB3 (Fig. 2c) is the site directly receives discharge from Giang Dien industrial park.

Water quality data were monitored monthly during 2009–2017 by the Department of Natural Resources and Environment of Dong Nai province. The monitoring program measured 21 water parameters. However, five parameters including Salinity, Zinc, Total grease, Aldrin and Endosulfan were excluded in this study due to their low variances (see Figure A1 in the Appendix). Hence, our study analyzed 16 parameters, including Temperature, pH, Conductivity, Turbidity, Total Suspended Solid (TSS), DO, COD, BOD₅, Ammonium (NH₄⁺), Nitrite (NO₂⁻), Nitrate (NO₃⁻), Phosphate (PO₄³⁻), Lead (Pb), Iron (Fe), Escherichia coli (E.coli) and Coliform.

Preprocessing of the original dataset included scanning for missing values and outliers, and imputation. Outliners are unique values distinctive to other observations that can distort the result of multivariate analysis. In this research, outliers were identified and trimmed using Interquartile Range (IQR), a method defines outliers as values above the upper whisker limit (third quartile + 1.5IQR) or below the lower whisker limit (first quartile – 1.5IQR) (Barbato et al. 2011). Imputation of missing values was conducted to maintain dataset dimension, taking into account that more than 50% of original information would be lost due to missing values and outlier removal. The imputed dataset was normalized (Eq. 1) to avoid scale differences (Kowalkowski et al. 2006).

\[
Z = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}
\]  

(1)

2.2 Correlation analysis

Correlation analysis was conducted to explore relationships among variables in dataset and filter irrelevant variables with no significant associations with others (Aguilera et al. 2018). The maximal information coefficient (MIC) proposed by Reshef et al. (2011) was applied to analyze the pair-wise association of time-series variables. The coefficient is more equitable compared to others measures such as mutual information estimation or squared Pearson/Spearman correlation R² (Reshef et al. 2011; Kinney and Atwal 2014; Albanese et al. 2018). MIC captures a wide range of correlation both linear and nonlinear, and provides a score for functional relationships. MIC score ranges from 0 for statistically independent variables to 1 for probability for noiseless functional relationships. Pairwise MICs were computed using python module minepy (Albanese et al. 2013).

2.3 Dynamic factor analysis (DFA) and cluster analysis (CA)

2.3.1 DFA

DFA was adopted for analyzing dynamic patterns of dataset. The DFA is a dimensionality reduction technique used for time-series data (Kuo et al. 2014). The method is useful for identifying latent temporal pattern in multivariate datasets by mining their lagged covariance. The full mathematical formula for the DFA was given in (2). In this study, the DFA was conducted using Python module statsmodels (Seabold and Perktold 2010).

\[
y_t = \Delta f_t + \beta x_t + u_t
\]  

(2)

in which:

\(y\): observed data
\(f\): unobserved factors
\(\Delta\): matrix of factor loadings
x: exogenous variables

B: matrix of regression parameters

u: error term

The aim of the DFA model is to find a set of M independent factors from a dataset of space n x N (n is number of observations, N is number of variables, N > M). Loading of M factors is a matrix of size N x M explaining the relative importance of factors to variables (Zuur et al. 2003a; Kuo et al. 2014). Canonical correlation coefficient is another quantity showing cross-correlation between M factors and N variables. In this research, N was 16 normalized water quality parameters and the DFA model was constructed for each monitoring site.

Exogenous (or explainable) variables (x) can be included to improve the model fitness (Kuo et al. 2014). The regression matrix B is of size i x N, with i is number of exogenous variables. Regression matrix helps identify exogenous variables that had significant influences on response variable (ibid). However, exogenous variables were not included for unraveling impacts of driver variables is out of scope of this research.

To find optimal value for number of factors, the Akaike's information criterion (AIC) (Akaike 1974) was suggested. The AIC judges performance of the DFA models on balancing between model fitting and number of underlying factors (Kuo et al. 2014). Models with low AICs are preferred for their simplicity but still ensuring less information loss (Kisekka et al. 2013). In this study, AIC was adopted as a metric for selection of number of factors of model. The developed model was later validated with observed values using Nash-Sutcliffe (NSE) coefficient of efficiency (Nash and Sutcliffe 1970)(3). The coefficients of determination (R^2), which quantifies correlations between variables and factors, were combined with graph validation to select valid patterns (factors) and their associated variables. The derived factors and factor loadings of variables were analyzed to generate insights about temporal structure of 16 water quality parameters at each site.

\[
NSE = 1 - \frac{\sum_{t=1}^{k} (Q_m - Q_m^o)^2}{\sum_{t=1}^{k} (Q_m - \bar{Q_m})^2}
\]

in which:

\(Q_m\): mean of observed variable

\(Q_m^o\): modeled value of variable

\(Q_m^t\): observed value of variable at time t

### 2.3.2 CA and cluster profiling

CA was conducted to find clusters of time points in a time-series dataset based on similarities of their values. In this study, the Python scikit-learn module’s hierarchical algorithm (Pedregosa et al. 2011) was utilized for clustering data. The Ward method was selected for defining similarities between observations. Number of clusters was identified by interpreting the Dendrogram and the validity of the clustering result was examined by silhouette score (Rousseeuw 1987). The independent factors derived from DFA were inputs for CA to address the concern about the existence of multi-collinearity in a dataset, which can impose implicit weighting and affect the analytical result (Kisekka et al. 2013; Kuo et al. 2013).

The derived clusters were profiled in terms of their temporal features and pollution intensities. To link pollution intensities with temporal patterns, only parameters that showed associations to derived dynamic factors were selected. Two Vietnamese regulatory standards of A2 and B1 for drinking and irrigation water respectively were adopted as benchmarks for assessing pollution intensities. The 95% confident intervals (CIs) of water quality parameters were calculated for each cluster using Bootstrap method (Efron 1979) to make statistical inference about level of pollution on comparing to the benchmark regulatory.

### 3. Results And Discussion

#### 3.1 Pairwise correlation of water quality variables

Figure 3 presents pair-wise MIC values at five monitoring sites. MIC values identified prevalent pair-wise associations of variables in five monitoring sites. The highest associations were among three variables TSS, Turbidity and Fe with MIC values ranging from 0.85 to 0.99. The high correlations between TSS, Turbidity and Fe mainly arriving from soil indicated their sharing of common source, probably driven by run-off in agricultural areas. Besides high correlation between BOD5 and COD which was intuitive, another noted type of correlation was among DO and nutrient parameters (reactive Nitrogen and PO4^3-). The highest correlation was between DO and NO2^- in SB1 and SB2 (Fig. 3-a,b), probably explained by denitrification process forming Nitrite in low DO condition (Le et al. 2017, 2019).
Cross-group association were important as they may indicate sharing of common drivers of pollution (Wunderlin et al. 2001). In SB1 and SB2 (Fig. 3-a,b), a significant correlation was observable between Fe and NO$_3^-$ (0.66 and 0.85 for SB1 and SB2 respectively). SB2 was also marked for its high correlation between run-off-related variables (like TSS, Fe), and organic parameters (BOD$_5$, COD). This indicated run-off source of NO$_2^-$ and organic pollution in sub-basin SB1 and SB2 respectively. The SB3 (Fig. 3-c), by contrast, had very low correlation between these run-off and organic groups (for instances, MIC value of Fe-COD pair was only 0.48). The noted relationships in SB3 was between organic (BOD$_5$ and COD) and nutrient parameters (PO$_4^{3-}$ and NO$_3^-$). In SB4 and SB5 (Fig. 3-d,e), such cross-group association was significant among Fe, E.coli and COD and among TSS, BOD$_5$ and PO$_4^{3-}$ respectively. Also from MIC analysis, individual variables associate with at least one other variable at all five stations, therefore we decided to sustain all 16 time-series variables subjected in the DFA.

### 3.2 Multivariate data analysis

#### 3.2.1 Model selection and validation

Analysis of AICs for the DFA models with different numbers of factors showed that in the SB1, three factors gave optimal result (AIC was 2365); for other four stations, the five-factor model gave the smallest AIC values (AIC values were 2235, 1594, 1170 and 2210 for SB2, SB3, SB4 and SB5 respectively).

Figure 4 shows the results of model validation using the Nash-Sutcliffe coefficient. Performance ranges for the coefficient was suggested in Chiew and McMahon (1993). The selected models for five sites performed well in almost all variables in that all NSE values were at an acceptable level (higher than 0.3). Selected models performed equally well on TSS, DO, COD, NO$_3^-$, and Fe, which had NSE values in satisfactory levels. On the other hand, NSE values of pH and E.coli were among the lowest, especially in SB1, SB2, and SB3 sites.

#### 3.2.2 Temporal patterns of variables and drivers of pollution

An analysis of determination coefficients of variables combined with graph validation showed that valid factors could capture temporal patterns of variables with coefficient higher than 0.4. This finding was similar to work by Y. M. Kuo, Chiu, and Yu (2014) suggesting that a correlation higher than 0.25 is moderate. Figure 5 presents derived valid factors projected onto space of scaled variables (for determination coefficients, see Table A2 of Appendix).

Table 1 shows loadings of variables in valid factors for five pollution sites. In SB1 site, two common patterns were recognized including factor 1 of Turbidity, TSS, COD, BOD$_5$ and Fe, and factor 2 of NH$_4^+$. Similarly, SB2 site owned two common patterns, first one captured fluctuation of Turbidity, TSS, COD, BOD$_5$, NO$_2^-$, PO$_4^{3-}$, and Fe and second one described trend of conductivity. In other three sites, there were three latent temporal patterns. In SB3 site, the first pattern included Turbidity, TSS, NO$_3^-$ and Fe; the second pattern included DO and COD and the last pattern described BOD$_5$ and PO$_4^{3-}$. The SB4 site possessed three patterns, the first one included turbidity, TSS, DO, COD, BOD$_5$, Fe, NO$_3^-$ and Coliform, the second included Temperature and E. coli, the third included PO$_4^{3-}$. Lastly, SB5 site' three patterns include one of TSS and COD, BOD$_5$, another one of DO and PO$_4^{3-}$ and last one of NO$_3^-$.

Most of the variables were positively correlated to their associated factors. SB2 was exceptional because of negative loadings of variables in the first factor, leading to inverse co-movement of factors and response variables. The signs of loading also revealed the dynamic relationships between variables. For example, the intuitive inverse correlation between DO and COD with loadings of 0.28 and −0.34 respectively in factor 2 of SB3. Similarly, it can be inferred that PO$_4^{3-}$ was the main driver of low DO in SB5 site on considering their inverse relationship (loadings of -0.32 and 0.38, respectively).

The first factor in each monitoring site explained the dominant pattern in the dataset. Analysis of this pattern’s components showed the contributions of at least one parameter related to run-off pollution, i.e., TSS, Turbidity, and Fe, making it pattern of run-off pollution. Similar inferences about sources of pollution can be found in Wang et al. (2013) and Razmkhah, Abrishamchi, and Torkian (2010). Association of other variables with run-off pollution revealed insightful information about drivers of these variables. In specific, association of COD and BOD$_5$ with run-off pollution in SB1, SB2, SB4 and SB5 suggested that organic pollution at these sites could be linked to area sources of organic pollutants. Similarly, NO$_2^-$ pattern in SB2, NO$_3^-$ patterns in SB3 and SB4 sites and total coliform pattern in SB4 could be driven by run-off from agricultural areas. In addition, loadings of variables in run-off factors also revealed features of run-off surface. In specific, the low loading to Fe in run-off factor of SB5 site (lying in the most urbanized sub-basin) compared to other sites indicated low influence of traditional natural surface's run-off. This insight was legitimate taking in account the impacts of quarry site upstream of SB5 and dominant urban land-use in its sub-basin. Similar findings about influence of local sources of pollution were suggested by Nguyen et al. (2019) and Felipe-Sotelo et al. (2007).

Contrarily, the independence of derived factors to each others at each site suggested distinctions of drivers of pollution. In the SB1, NH$_4^+$ owned a pattern that diverted from run-off components, pointing to sources other than run-off of NH$_4^+$. A very likely source of NH$_4^+$ in the SB1 was...
human excreta (Van Drecht et al. 2003) taking in account location of SB1 in residential area. Effluence from residential area was also likely the main source of PO₄³⁻ in the SB3, SB4 and SB5 as suggested by Riemersma et al. (2006). Similarly, in the SB3, separation of COD and BOD₅ from run-off factor revealed that organic pollution might be driven by point sources’ activities; such influence was very likely on considering that water bodies at SB3 site receive discharge from industrial zone nearby. Also, the isolation of COD from BOD pattern in SB3 proved the influence of industrial effluent. Same inference was also valid in case of E.coli in SB4 site. Unlike total Coliform which was run-off driven, trend of E.coli in SB4 was driven by discrete point sources of surrounding livestock farms. Lastly, pattern of PO₄³⁻ at the SB3, SB4 and SB5 were driven by point sources. 3.2.3 Pollution profile and policy implication

Application of Hierarchical clustering for factors showed that at SB1 and SB4, there were optimally two clusters, and at the three remaining sites SB2, SB3 and SB5, there were three clusters. Figure A4 of Appendix showed dendograms plotting distances between clusters at each site.

Figure 6 and 7 presented temporal and pollution characteristics of clusters respectively; for numeric presentation of cluster profiles, see Table A3 of Appendix. Water quality parameters associated with run-off factors realized in the DFA models had an apparently higher range of mean distributions in clusters occupied mainly by wet-month data points, i.e. May, June, July, August, September and October. For example, in SB1 site, cluster 1 composed approximately 71% by wet-season data points (Fig. 6-a) had TSS, Fe, Turbidity (CIs of 52.73–94.08, 4.83–8.38 and 36.80–73.86 respectively) significantly higher than these in cluster 2 featured by 63% of dry-season data points (CIs of 18.86–30.48, 1.72–2.83 and 13.21–21.78 respectively). Similarly, the BOD₅ and COD values of wet-period cluster in the SB1 had CIs of 5.69–7.64 and 18.21–27.99 respectively compared to 4.03–4.92 and 10.88–14.57 in dry-period cluster. Such inference was weakly adoptable in case of NO₃⁻ in SB3 and Coliform in SB4 where CIs of these two parameters in rainy-season clusters showed higher ranges of values but not strictly significant.

Cluster profiling of water quality variables detached from run-off pattern was complex for it was unlikely possible to describe seasonal characteristics of these clusters. The NH₄⁺ parameter in SB1 site for instance, had indistinguishable CIs on comparing between wet and dry-season cluster (CIs of 0.39–0.82 and 0.34–0.59 respectively). Similar remark was adoptable to, conductivity in SB1; DO, COD, BOD₅ and PO₄³⁻ in SB2; Temperature and PO₄³⁻ in SB4 and lastly DO, PO₄³⁻ and NO₃⁻ in SB5. E.coli parameter in SB4 was an exception as its CIs in wet season was significantly lower than in dry season (150.00-236.28 compared to 290-1085.70).

Profiling of clusters in term of their seasonal characteristics and varying pollution intensities supported pollution management on comparing to regulatory limit A2 (for drinking water) and B1 (for irrigation water) (Fig. 7). Run-off-induced pollution that breach the two standards was rather a year-round phenomenon especially for Fe which exceeded both the criteria of A2 and B1. The problem with TSS was less severe in dry season in SB1 (cluster 2), SB2 (cluster 1), SB5 (cluster 1) when pollution level only exceeded lower regulatory level B1. This also suggested that TSS and Fe pollutions of La Buong river is rather a natural phenomenon which are intensified in wet season by surface run-off. This phenomenon was also observed in another study in Dong Nai river basin (Quan and Meon 2015). Therefore, a proper abatement scheme for TSS and Fe pollution should consider controlling of pollution sources by land-use practices in wet season combined with water treatment solutions before using. Furthermore, for TSS pattern at SB5 site showed the influence of urban run-off and quarry site in wet season, measures like green infrastructure for abating run-off contamination and pollution control at quarry site at SB5 should be considered.

For organic parameters that could be linked to run-off drivers like BOD₅, COD in the SB1, SB2 SB4, and SB5, an obvious pattern of exceeding regulatory levels (mostly B1) could be observed in wet-season clusters. This finding was strongly policy-relevant as organic pollution abatement in SB1, SB2, and SB4 sites could be conducted through land-use management policy at watershed scale aiming at limiting the influence of run-off. For SB5, more concerns should be on urban run-off. A similar suggestion was also suitable for NO₂⁻ and PO₄³⁻ pollution in SB2. On the contrary, exceeding the regulatory limit of BOD₅ and COD at the SB3 site could be reduced at point sources with a particular focus on the industrial zone.

Eutrophication parameters excluded from run-off-driven groups like NH₄⁺ in SB1, PO₄³⁻ in SB3, SB4 and SB5, and NO₃⁻ in SB5, the NH₄⁺ and PO₄³⁻ showed significant nutrient pollution. The separation of these parameters from the run-off pattern pointed out that reduction for point sources is more efficient. Abating NH₄⁺ pollution at SB1 and PO₄³⁻ pollution at SB3, SB4 and SB5 should target domestic waste treatment from the residential area.

4. Conclusion

In this study, temporal patterns of 16 water quality parameters monitored from 2010 to 2017 at five La Buong watershed stations were analyzed using DFA and CA techniques. The DFA helped identify explicitly the pattern of run-off-driven variables and other patterns related to point sources. DFA results also helped unravel the influences of local sources on pollution patterns like effluence from the industrial park at SB3 affecting BOD₅ and COD wastewater from residential areas as sources of NH₄⁺ in the SB1 PO₄³⁻ in the SB4 and SB5 or discrete source of E.coli in the SB4. This finding suggested that DFA is a data-driven bottom-up method suitable for pollution management in the local context.
Time-point clustering of temporal patterns have segmented time-series data into clusters of distinctive temporal structure. The profiling of temporal clusters compared to regulatory values suggested that in order to address Fe and TSS, a combination of land-use measures and water treatment plan is necessary. The land-use approach was also suitable to control organic pollution observed in the SB1, SB2, SB3 and SB5. This finding supported the application of the dynamic multivariate approach in developing an integrated pollution management scheme. On the contrary, the point-source measures worked better for organic pollution in SB3 and E.coli problem in SB4 as well as NH$_4$+ in the SB1.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the co-author Nguyen Hong Quan, email nh.quan@iced.org.vn on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the Vietnam National University of Ho Chi Minh city under project C2020-24-09/HĐ – KHCN

Authors’ contributions

All authors contributed to the study in various aspects. LHP developed the concept, conducted data analysis, and wrote the first draft of the manuscript, while DDT contributed to revising the structure and editing. HDTL performed correlation analysis and, together with TQD writing various parts of the manuscript. QHN provided monitoring data and revised the manuscript. KND, HTTN, and AHN presented their comments on the methods and discussion

References

1. Aguilera R, Sabater S, Marcé R (2018) A methodological framework for characterizing the spatiotemporal variability of river water-quality patterns using dynamic factor analysis. J Environ Informatics 31:97–110. https://doi.org/10.3808/jei.201600333
2. Akaike H (1974) A New Look at the Statistical Model Identification. IEEE Trans Automat Contr 19:716–723. https://doi.org/10.1109/TAC.1974.1100705
3. Albanese D, Filosi M, Visintainer R et al (2013) Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics 29:407–408. https://doi.org/10.1093/bioinformatics/bts707
4. Albanese D, Riccadonna S, Franceschi P, Donati C (2018) A practical tool for maximal information coefficient analysis. Gigascience 1–8. https://doi.org/10.1093/gigascience/giy032
5. Barbato G, Barini EM, Genta G, Levi R (2011) Features and performance of some outlier detection methods. J Appl Stat 38:2133–2149. https://doi.org/10.1080/02664763.2010.545119
6. Budaev SV (2010) Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. Ethology 116:472–480. https://doi.org/10.1111/j.1439-0310.2010.01758.x
7. Chiew FHS, McMahon TA (1993) Assessing the adequacy of catchment streamflow yield estimates. Aust J Soil Res 31:665–680. https://doi.org/10.1071/SR930665
8. Diamantini E, Lutz SR, Mallucci S et al (2018) Driver detection of water quality trends in three large European river basins. Sci Total Environ 612:49–62. https://doi.org/10.1016/j.scitotenv.2017.08.172
9. Dixon W, Chiswell B (1996) Review of aquatic monitoring program design. Water Res 30:1935–1948. https://doi.org/10.1016/0043-1354(96)00087-5
10. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7:1–26. https://doi.org/10.1214/aos/1176348654
11. Felipe-Sotelo M, Andrade JM, Carlosena A, Tauler R (2007) Temporal characterisation of river waters in urban and semi-urban areas using physico-chemical parameters and chemometric methods. Anal Chim Acta 583:128–137. https://doi.org/10.1016/j.aca.2006.10.011

12. Glavan M, Pintar M (2012) Strengths, Weaknesses, Opportunities and Threats of Catchment Modelling with Soil and Water Assessment Tool (SWAT) Model. Water Resour Manag Model. https://doi.org/10.5772/34539

13. Juahir H, Zain SM, Kamil M et al (2011) Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. https://doi.org/10.1007/s10166-010-1411-x

14. Khoi DN, Nguyen VT, Sam TT, Nqi PTT (2019) Evaluation on effects of climate and land-use changes on streamflow and water quality in the La Buong River Basin, Southern Vietnam. Sustain 11: https://doi.org/10.3390/SU11247221

15. Kinney JB, Atwal GS (2014) Equitability, mutual information, and the maximal information coefficient. 2014:21–26. https://doi.org/10.1073/pnas.1309933111

16. Kisekka I, Migliaccio KW, Muñoz-Carpena R et al (2013) Dynamic factor analysis of surface water management impacts on soil and bedrock water contents in Southern Florida Lowlands. J Hydrol 488:55–72. https://doi.org/10.1016/j.jhydrol.2013.02.035

17. Kowalkowski T, Zbytniewski R, Szpejna J, Buszewski B (2006) Application of chemometrics in river water classification. Water Res 40:744–752. https://doi.org/10.1016/j.watres.2005.11.042

18. Kuo Y-M, Chang F-J (2010) Dynamic Factor Analysis for Estimating Ground Water Arsenic Trends. J Environ Qual 39:176–184. https://doi.org/10.2134/jeq2009.0098

19. Kuo YM, Chiu CH, Yu HL (2014) Influences of ambient air pollutants and meteorological conditions on ozone variations in Kaohsiung, Taiwan. Stoch Environ Res Risk Assess 29:1037–1050. https://doi.org/10.1007/s00477-014-0968-2

20. Kuo YM, Jang CS, Yu HL et al (2013) Identifying nearshore groundwater and river hydrochemical variables influencing water quality of Kaoping River Estuary using dynamic factor analysis model. J Hydrol 486:39–47. https://doi.org/10.1016/j.jhydrol.2013.01.027

21. Le TTH, Fettig J, Meon G (2019) Kinetics and simulation of nitrification at various pH values of a polluted river in the tropics. Hydrobiol 19:54–65. https://doi.org/10.1016/j.ecolhyd.2018.06.006

22. Le TTH, Zeunert S, Lorenz M, Meon G (2017) Multivariate statistical assessment of a polluted river under nitrification inhibition in the tropics. Environ Sci Pollut Res 24:13845–13862. https://doi.org/10.1007/s11356-017-0989-2

23. Liu CW, Lin KH, Kuo YM (2003) Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Sci Total Environ 313:77–89. https://doi.org/10.1016/S0048-9697(02)00683-6

24. Magyar N, Hatvani IG, Székely IK et al (2013) Application of multivariate statistical methods in determining spatial changes in water quality in the Austrian part of Neusiedler See. Ecol Eng 55:82–92. https://doi.org/10.1016/j.ecoleng.2013.02.005

25. Molenaar PCM (1985) A dynamic factor model for the analysis of multivariate time series. Psychometrika 50:181–202. https://doi.org/10.1007/BF02294246

26. Muñoz-Carpena R, Ritter A, Li YC (2005) Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to Everglades National Park. J Contam Hydrol 80:49–70. https://doi.org/10.1016/j.jconhyd.2005.07.003

27. Nash JE, Sutcliffe IV (1970) River ow forecasting through conceptual models. Part I - A discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1007/10.1080/15715124.2019.1700513

28. Nguyen HD, Hong Quan N, Quang NX et al (2019) Spatio-temporal pattern of water quality in the Saigon-Dong Nai river system due to waste water pollution sources. Int J River Basin Manag 0:1–34. https://doi.org/10.1080/00750770109555783

29. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830

30. Pesce SF, Wunderlin DA (2000) Use of Water Quality Indices To Verify the Córdoba City (Argentina) on Suquía River. Wat Res 34:2915–2926

31. Quan NH, Meon G (2015) Nutrient Dynamics During Flood Events in Tropical Catchments: A Case Study in Southern Vietnam. Clean - Soil Air Water 43:652–661.

32. 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 Manage 91:852–860.

33. Reshef D, Reshef Y, Finucane H et al (2011) Detecting Novel Associations in Large Data Sets. Sci Transl Med 334:1518–1524

34. Riemersma S, Little J, Ontkean G, Moskal-Hebert T (2006) Phosphorus sources and sinks in watersheds: a review. Alberta Soil Phosphorus Limits Proj 5:82

35. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

36. Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference

37. 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. https://doi.org/10.1016/j.envsoft.2006.02.001
38. 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. https://doi.org/10.1016/j.aca.2005.02.006
39. Van Drecht G, Bouwman AF, Knoop JM et al (2003) Global modeling of the fate of nitrogen from point and nonpoint sources in soils, groundwater, and surface water. Global Biogeochem Cycles 17:. https://doi.org/10.1029/2003gb002060
40. 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. https://doi.org/10.1016/S0043-1354(98)00138-9
41. Vialle C, Sablyrolles C, Lovera M et al (2011) Monitoring of water quality from roof runoff: Interpretation using multivariate analysis. Water Res 45:3765–3775. https://doi.org/10.1016/j.watres.2011.04.029
42. Wang Y, Wang P, Bai Y et al (2013) Assessment of surface water quality via multivariate statistical techniques: A case study of the Songhua River Harbin region, China. J Hydro-Environment Res 7:30 – 40. https://doi.org/10.1016/j.jher.2012.10.003
43. Wunderlin DA, Maria Del Pilar D, Maria Valeria A et al (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A Case Study: Suquía River basin (Córdoba-Argentina). Water Res 35:2881 – 2894. https://doi.org/10.1016/S0043-1354(00)00592-3
44. Zuur AF, Fryer RJ, Jolliffe IT et al (2003a) Estimating common trends in multivariate time series using dynamic factor analysis. Environmetrics 14:665 – 685. https://doi.org/10.1002/env.611
45. Zuur AF, Tuck ID, Bailey N (2003b) Dynamic factor analysis to estimate common trends in fisheries time series. Can J Fish Aquat Sci 60:542 – 552. https://doi.org/10.1139/F03-030

Tables

Table 1. Factor loadings of variables in factors for each pollution sites (bold underlined numbers present significant loadings)

| Parameters | SB1 | SB2 | SB3 | SB4 | SB5 |
|------------|-----|-----|-----|-----|-----|
| Temperature | -0.04 | -0.05 | -0.05 | -0.14 | -0.06 | -0.22 | -0.28 | -0.04 | **-0.32** | -0.01 | -0.20 | -0.14 | -0.05 |
| Conductivity | -0.09 | -0.24 | -0.14 | **-0.32** | 0.20 | -0.27 | -0.25 | 0.03 | -0.22 | 0.10 | 0.06 | -0.31 | 0.15 |
| pH | -0.11 | -0.05 | -0.23 | -0.07 | -0.07 | -0.08 | -0.16 | 0.24 | 0.17 | -0.03 | 0.25 | -0.14 | 0.22 |
| Turbidity | **0.37** | 0.26 | **-0.32** | -0.18 | **0.42** | 0.13 | -0.15 | **0.33** | -0.19 | 0.20 | 0.27 | 0.25 | 0.12 |
| TSS | **0.38** | 0.01 | **-0.29** | 0.18 | **0.42** | 0.16 | 0.09 | **0.34** | -0.15 | 0.16 | **0.35** | 0.17 | 0.05 |
| DO | -0.10 | 0.19 | 0.21 | 0.03 | 0.13 | **0.28** | -0.02 | **0.28** | 0.22 | 0.20 | -0.08 | **0.38** | 0.06 |
| COD | **0.38** | -0.17 | **-0.36** | 0.20 | 0.18 | **-0.34** | 0.12 | **0.32** | -0.09 | -0.22 | **0.36** | -0.06 | 0.01 |
| BOD$_5$ | **0.36** | -0.15 | **-0.29** | 0.27 | 0.10 | -0.24 | **0.34** | **0.27** | -0.22 | -0.21 | **0.31** | -0.08 | 0.07 |
| NH$_4^+$ | 0.12 | **-0.39** | -0.09 | 0.03 | -0.05 | -0.23 | 0.21 | -0.18 | -0.09 | -0.12 | 0.13 | -0.24 | 0.08 |
| NO$_2^-$ | 0.24 | -0.28 | **-0.27** | -0.18 | 0.19 | -0.27 | -0.15 | -0.11 | -0.28 | -0.04 | 0.27 | -0.22 | -0.18 |
| NO$_3^-$ | 0.25 | -0.09 | -0.23 | -0.16 | **0.33** | -0.17 | -0.06 | **0.26** | -0.27 | -0.04 | 0.26 | 0.03 | **-0.35** |
| PO$_4^{3-}$ | 0.11 | -0.29 | **-0.31** | 0.13 | 0.10 | -0.17 | **0.34** | 0.15 | 0.10 | **-0.29** | 0.28 | **-0.32** | 0.10 |
| Pb | 0.18 | 0.15 | -0.26 | -0.02 | **0.30** | 0.14 | 0.10 | 0.14 | -0.08 | -0.17 | 0.16 | 0.21 | 0.15 |
| Fe | **0.35** | 0.21 | **-0.32** | -0.08 | **0.41** | 0.19 | 0.02 | **0.38** | -0.02 | 0.11 | 0.30 | 0.29 | 0.09 |
| E.coli | 0.16 | 0.14 | -0.13 | 0.07 | -0.23 | 0.06 | 0.13 | 0.22 | **0.35** | 0.04 | 0.29 | 0.05 | -0.15 |
| Coliform | 0.25 | 0.20 | -0.24 | -0.02 | -0.25 | 0.13 | -0.04 | **0.30** | 0.13 | -0.16 | 0.17 | 0.11 | -0.32 |
Figure 1

Map of study area and locations of monitoring sites

(a) (b) (c) (d) (e)
Figure 2

Land-use characteristics at monitoring sites (a-SB1, b-SB2, c-SB3, d-SB4, e-SB5) (background image is false-color composite of Sentinel-2 data taken in March 2020)

(a)  
(b)  
(c)  
(d)  
(e)  

Figure 3

Pairwise Maximal Information Coefficient between variables in five stations (values less than 0.5 were masked out)

Figure 4

Nash-Sutcliffe coefficients for dynamic factor models at the five stations
Figure 5

Please see the Manuscript file for complete figure caption.

Figure 6

Temporal profiles of derived clusters (percent of months) at each monitoring site, blank cell presents zero percent
Figure 7
Cluster profiles for five monitoring sites (red and black dot lines showed regulatory benchmarks A2 and B1)

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARY.docx