STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, September 2022, pp 1204–1221.
Published online in International Academic Press (www.IAPress.org)

Spatial assessment of water river pollution using the stochastic block model: Application in different station in the Litani river, Lebanon

Alya Atoui [1,2], Abir El Haj [3], Yousri Slaoui [3], Ali Fadel[4], Kamal Slim[5], Samir Abbad Andaloussi[1], Régis Moilleron [1], Zaher Khraibani [2,6,*]

[1] *Leesu, Univ Paris Est Creteil,F-94010,France*
[2] *Rammal Rammal Laboratory, PhyToxE group, Lebanese University, Nabatieh, Lebanon*
[3] *Laboratoire de Mathématiques et Applications, Université de Poitiers, Poitiers, 86000, France*
[4] *Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS - Université de Tours, Tours, 37200, France*
[5] *Lebanese Atomic Energy Commission, National Council for Scientific Research (CNRS), Beirut, Lebanon*
[6] *Department of Applied Mathematics, Lebanese University, Faculty of Sciences, Hadat, Lebanon*

Abstract Water pollution is a major global environmental problem. In Lebanon, water pollution threatens public health and biological diversity. In this work, a non-classical classification method was used to assess water pollution in a Mediterranean River. A clustering proposal method based on the stochastic block model (SBM) was used as an application on physicochemical parameters in three stations of the Litani River to regroup these parameters in different clusters and identify the evolution of the physicochemical parameters between the stations. Results showed that the used method gave advanced findings on the distribution of parameters between inter and intra stations. This was achieved by calculating the estimated connection matrices between the obtained clusters and the probability vector of belonging of the physicochemical parameters to each cluster in the different stations. In each of the three stations, the same two clusters were obtained, the difference between them was in the estimated connection matrices and the estimated cluster membership vectors. The power of SBM proposed methods is demonstrated in simulation studies and a new real application to the sampling physicochemical parameters in Litani River. First, we compare the proposed method to the classical principal component analysis (PCA) method then to the Hierarchical and the K-means clustering methods. Results showed that these classical methods gave the same two clusters as the proposed method. However, unlike the proposed SBM method, classical approaches are not able to show the blocks structure of the three stations.

Keywords Litani River, Water quality, Physicochemical parameters, Stochastic Block Model, Clustering.

DOI: 10.19139/soic-2310-5070-1547

1. Introduction

Water is an essential natural resource for any ecosystem. Maintaining its quality is a major concern for society, especially with the increasing future water needs. Freshwater ecosystems are natural compartments necessary for the continuity of life. They are essential for various activities such as the production of drinking water, industrial and agriculture activities, hydropower generation [Ridzuan et al.(2020)], and recreational activities [Simpi et al.(2011)]. Unfortunately, they are among the most seriously threatened ecosystems due to anthropogenic activities over the past century [Dudgeon et al. (2006)]. However, the quality of groundwater and surface water in Lebanon, in most cases, does not meet international standards levels. The impact of human activities in the sense of simplification, modification, and alteration of natural ecosystems is becoming increasingly alarming

*Correspondence to: Zaher Khraibani (Email: zaher.khraibani@ul.edu.lb). Department of Applied Mathematics, Lebanese University, Hadat, Lebanon.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

[Fadel et al.(2014)], [Qiang et al.(2019)]. This impact affects both terrestrial and aquatic ecosystems and therefore the environment. Statistical modeling is an effective method to understand, analyze and provide recommendations for various problems in aquatic systems. In the present study, statistical methods were used to characterize the physicochemical parameters of the water in several stations on Litani River to assess the status of quality and the intensity of pollution of the water, to locate the sources of pollution, and to establish a disturbance gradient between the different three stations (Lake Qaraoun, Jeb-Jenin, and Ghzayel). Litani River is the longest freshwater resource in Lebanon. It starts from the west of Baalbek in the fertile Bekaa plain and empties into the Mediterranean Sea north of Tyre. It has a length of 172 km, a basin area of 2186 km^2, and a discharge of 360 million m^3/year. The Litani River Basin consists of two subbasins: the Upper and Lower Litani Basins, which are joined in the middle part by Lake Qaraoun (i.e. capacity of 220 million m^3). Over the last 20 years, the Litani River has experienced rapid population growth and development, resulting in large-scale land-use changes, rapid population growth, and industrial and agricultural activities. This has caused a deterioration in the water quality of the Litani River [Hayek et al.(2021)]. The main sources of pollution in the Litani River Bassin include the waste discharged by factories manufacturing agrifood products, pesticides, herbicides, and especially the wastewater of a hundred villages and urban agglomerations. There are conventional statistical methods and analyses in the literature that are directly applicable to studying water pollution in the Litani River. The statistical research in general and especially at the Litani River that has been carried out on the study of water quality is a comparative descriptive study based on classical statistical methods such as data analysis (Principal component analysis), descriptive statistics, inferential statistics, cluster analysis, [Fadel et al.(2021)], [Ali H. et al.(2020)], [Mark et al.(2012)], [Chaden et al. (2014)], trend analysis method, [Groppo et al.(2008)], [Esterby (1993)]. Many methods use global dimension reduction techniques to solve high dimensionality problems. Classical clustering methods are most often used in the literature. Among these methods, the most popular is the principal component analysis (PCA) which is often used in data mining and image analysis. However, PCA is a linear technique, that only considers linear dependencies between variables. In addition, among the most conventionally used techniques are k-means or hierarchical classification. These techniques are easy to implement, flexible in the level of granularity, applicable to any type of attribute, and can easily manipulate any form of similarity or distance. Their disadvantages compared to the SBM method are their; high cost, time complexity, and generally having trough results. Moreover, the levels of the nodes of the hierarchy are no longer defined except by the order in which they appear. In this paper, we adopt a new approach to classify the physicochemical parameters at the Litani River in Lebanon. This approach is based on the SBM method, and it aims to show the relationship between these parameters to have a good understanding of the water quality in three different stations of the river. An important aspect of the proposed method in this work is to determine elements with similar properties based on the observed and modeled relationships. Since there are many possible applications and many approaches that exist for the detection of these so-called clusters or groups. For a more general overview of clustering methods, we recommend that you see the Fortunato review [Fortunato, S.(2010)] for further details. In this work, we focus on the stochastic block model (SBM) because this model is different from others based on a generative formulation. The SBM method is not only a partition but also a description of the relationship between the deduced groups. In addition, SBM can create and describe a variety of different structures, whereas most approaches would only be able to identify one kind. The advances in network analysis are used by many people unwittingly in their lives. Many systems in our lives can be adequately represented as networks, such as social networks, citation, and co-citation networks. A network consists of several vertices connected by edges. Most real-world networks are weighted, where edges joining nodes are often associated with weights representing their strength, intensity, or capacity. Recently, several authors focused on the partition of network nodes into groups, clusters, or communities that have different connection behavior within groups than between groups. Indeed, the number of links joining nodes within a cluster is higher than between different clusters. Thus, since clusters can differ strongly in their characteristics, it is quite important to find these clusters. This paper aims to cluster the physicochemical parameters of the Litani river in various stations. The considered network is a weighted indirect network without a self-loop. It consists of nodes that represent the physicochemical parameters and an edge joining a pair of parameters is weighted by the Wasserstein distance between the values of these two parameters. We present here a model for the generation of networks with community structure, the so-called stochastic block model (SBM). This model is proposed by [Anderson et al.(1992)] and [Holland et al.(1983)] and

aims to produce classes, called blocks, or more generally clusters in networks. Several authors have proposed some generalization of this model [Mariadassou et al.(2010)] have treated the case of weighted networks by using the SBM model [Airoldi et al.(2008)], [Ng and Murphy (2021)] and then [Latouche et al.(2011)] have focused on the SBM model with overlapping clusters. More recently, Barbillon et al.[Barbillon et al.(2017)] have dealt with the case of multiplex networks, where several edges of different types can exist between a pair of nodes and [Zreik et al.(2017)] and [Matias and Miele (2017)] have extended the model to deal with dynamic networks. However, [El Haj et al.(2020)] have proposed a binomial SBM to deal with co-citation networks. We are interested here in estimating the parameters in the SBM model as well as in clustering the vertices of the considered network. Several authors have focused on estimation methods. First, [Snijders and Nowicki (1997)] has proposed a maximum likelihood inference based on the expectation-maximization (EM) algorithm to estimate the probabilities of connection between nodes and to predict the clusters in the SBM model having only two blocks. Then, [Nowicki and Snijders (2001)] have generalized the previous work by proposing a Bayesian approach based on Gibbs sampling to deal with the case of the SBM model with an arbitrary number of blocks. However, the EM approach is intractable in the case of the SBM model because of the dependency of the edges in the network. To tackle this issue, [Daudin et al.(2008)] has proposed a variational approach by introducing the variational expectation maximization (VEM) algorithm while [Latouche et al.(2012)] has proposed a variational Bayesian approach based on variational Bayes EM algorithm (VBEM) to estimate these parameters. To prove the validity of this method in the environmental field, we introduce some applications of the proposed approach using first simulated data, then using the physicochemical parameters. We develop and implement the method on simulated data sets (to validate the procedure) and to show the efficiency of our approach by calculating the root mean square error and the Adjusted Rank Index. Also, we propose to use the variational approach developed by [Daudin et al.(2008)] to estimate the parameters as well as to classify the nodes in a Gaussian SBM model. This approach is consistent under the SBM model according to Celisse et al.[Celisse et al. (2012)].

2. Environmental concept

2.1. Site Selection and Description

Three sites have been selected in the Litani basin for the sampling to achieve the SBM study, Jeb-Jenin station, Lake Qaraoun, and Ghzayel. The selection of these 3 stations in the Litani river is due to two main concepts; the first one is, the stationed distribution of zones, one at the upper (Jeb-Jenin, Ghzayel), of the river, the second in the middle (Lake Qaraoun). The second concept is due to the importance of the three zones, while the (Jeb-Jenin and Ghzayel) station in the Upper Litani River is located in the zone with complex industries and a huge population, the second is Lake Qaraoun, the only lake in Lebanon, located in the middle of the river and divided it into two parts (see Figure1).

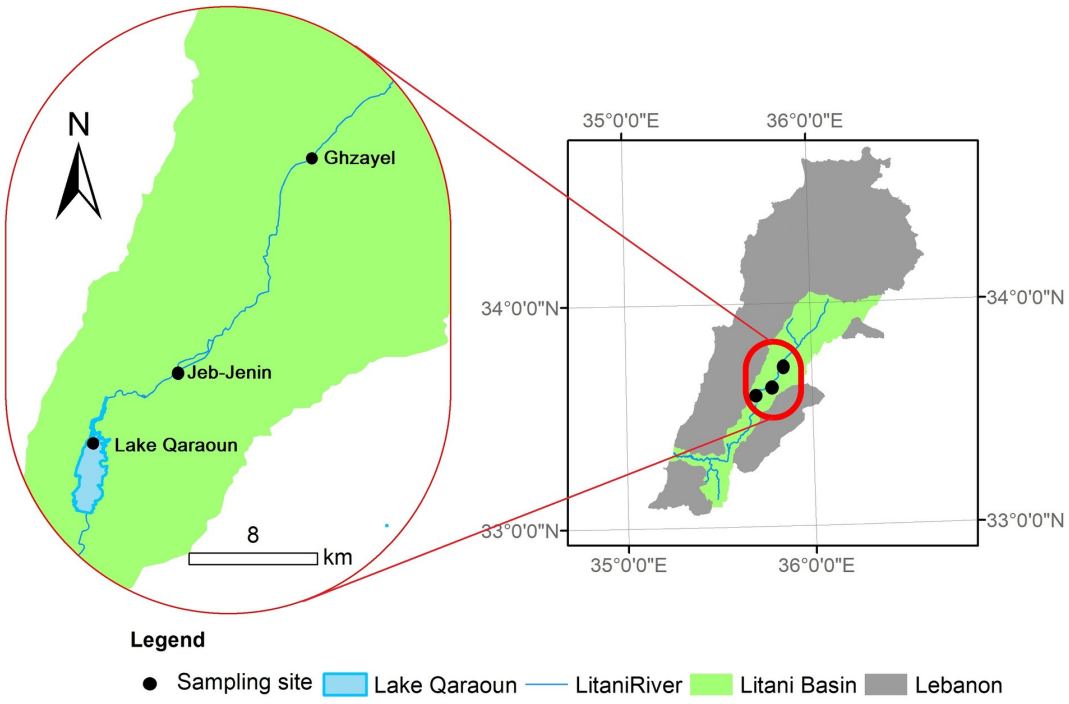

Figure 1. Litani river and located three stations.

The Litani River has 16 tributaries with Berdawni, Chtoura, and Ghzayel rivers being the most important ones. The average discharge of the Ghzayel stations is 2.8 (m^3/s) and 891 (m) as an altitude, with the coordinate: $N33°\ 43'\ 56''N$ (Latitude), $E35°\ 56'\ 52''E$ (Longitude). Noted that the description of the sites Jeb-Jenin (Upper Litani River), with the coordinate: $N33°\ 38'\ 331''\ E35°\ 46'\ 79''$ with elevation above the sea level equal to 920 (m). Jeb-Jenin is the largest and most populated town in its district. Lake Qaraoun (Middle of Litani River), with the coordinate: $N33°\ 46'\ 42''$ (Latitude) $E35°\ 53'\ 26''$ (Longitude), with elevation above the sea level equal to 864 (m). Qaraoun is the biggest lake in Lebanon with a total capacity of 220 Million cube meters.

2.2. Sampling test

Several parameters such as Temperature, DO, pH, conductivity, total dissolved solids (TDS), Salinity, Ammonia, Nitrite, Nitrate, SO4, PO4, were analyzed for each water sample collected during the 2008-2018 periods in three stations using spectrophotometry (Table 1) [Fadel et al.(2019)] [Fadel et al.(2021)]. All samplings were taken from the same points in the three stations. The same sampling method was used for all samples, 2 (m) from the side, and 20 (cm) deep from the sub-surface. Each sample consists of 5 trials, while each measurement was performed 5 times and the average is recorded. These samples were taken at the beginning of each month.

Table 1. Methods and quality control used to measure the physicochemical parameters.

Variable	Used instrumentation and methods/solutions	Accuracy (sensitivity)	Range test
Temp	LaMotte, Salt/TDS/Conductivity/Temp TRACER	±1.0 °C	0–65 °C
Salinity (Sal, mg/L)		±2 mg/L	0–9999 ppm (mg/L)
Total dissolved solids (TDS, mg/L)		±2 mg/L	0–9999 mg/L
Conductivity (EC, S/cm)		±2 S/cm	0–1999 S/cm
pH	LaMotte,—pH Meter	±0.01 pH	0–14 pH
Dissolved oxygen (DO, mg/L)	LaMotte, DO 6 Plus Dissolved oxygen meter	±0.3 mg/L	0–20 mg/L
Ammonia ($NH_3, mg/L$)	Ionic strength adjustor (ISA) for ammonium determinations by ion selective electrode (ISE) method	±0.05 mg/L $NH_3 - N$	0.1–10.0 mg/L $NH_3 - N$
Sulfate ($[SO_4], mg/L$)	Spectrophotometry using powder pillows Hach 8051	±0.5 mg/L SO4	2–70 mg/L SO_4
Phosphate ($PO_4, mg/L$)	Spectrophotometry using USEPA PhosVer 3 ascorbic acid method	±0.06 mg/L PO_4	0.06–5.00 mg/L PO_4
Nitrite ($NO_2, mg/L$)	Spectrophotometry using USEPA NitriVer 3 diazotization method	±0.002 mg/L $NO_2 - N$	0.002–0.300 mg/L $NO_2 - N$
Nitrate ($NO_3, mg/L$)	Spectrophotometry using NitraVer 5 cadmium reduction method	±0.2 mg/L $NO_3 - N$	0.1–10.0 mg/L $NO_3 - N$

3. Proposed statistical method

3.1. The model

In this section, we present the necessary tools to develop our methodology to clustering the physicochemical parameters. For this reason, we consider a weighted undirected network represented by $G : ([n], X)$, where $[n]$ is the set of weighted nodes $\{1, \ldots, n\}$ for all $n \geq 1$ and X is the symmetric weighted matrix of dimensions $n \times n$ encoding the of intensity of the observed interactions between nodes. In this context, the weighted nodes denotes the physicochemical parameters and the adjacency matrix X encodes the interaction between these parameters such as, for all $i, j \in \{1, \ldots, n\}$,

$$X_{ij} = \begin{cases} m_{ij} & \text{if the nodes } i \text{ and } j \text{ interact with an interaction weight } m_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

We denote by Q the number of clusters ($Q > 1$) and by Z the binary indicator matrix labeling the assignment of the physicochemical parameters into groups. We have for all $i \in \{1, \ldots, n\}$ and $q \in \{1, \ldots, Q\}$,

$$Z_{iq} = \begin{cases} 1 & \text{if node } i \text{ belongs to group } q \\ 0 & \text{otherwise.} \end{cases}$$

3.2. Mixture model with latent classes

We propose to generate the stochastic block model as follows:

- The (latent) vectors Z_i , for $i \in \{1, \ldots, n\}$, are independent and sampled from a multinomial distribution as follows
$$Z_i \sim \mathcal{M}(1, \alpha = (\alpha_1, \ldots, \alpha_Q)),$$
where $\alpha = (\alpha_1, \ldots, \alpha_Q)$ is the vector of class proportions of dimension $1 \times Q$ such as
$$\sum_{q=1}^{Q} \alpha_q = 1.$$
- The (observed) variables $\{X_{ij}, i, j \in [n], i < j\}$ are independent conditionally on $\{Z_i = q, Z_j = l\}$, and are sampled from a Gaussian distribution as follows
$$X_{ij} \mid Z_{iq} Z_{jl} = 1 \sim \mathcal{N}(\mu_{ql}, \sigma^2_{ql}),$$
where μ_{ql} and σ^2_{ql} denotes respectively the mean and the covariance parameters associated to the Gaussian distribution.

3.3. Inference

We are interested here in estimating the parameter $\theta = (\alpha, \mu, \Sigma)$ of the model in a weighted undirected network. However, we claim that all results obtained in this paper can be extended to directed networks.

Since the variable Z is latent, our model belongs to the class of incomplete data models. The log-likelihood of the incomplete data can be expressed as follows

$$\log P_\theta(X) = \log \sum_z \mathbb{P}_\theta(X, Z), \tag{1}$$

where $\mathbb{P}_\theta(X, Z)$ is the joint distribution such that

$$\mathbb{P}_\theta(X, Z) \quad = \quad \mathbb{P}_{\mu,\sigma}(X \mid Z)\mathbb{P}_\alpha(Z),$$

where

$$\begin{aligned}\mathbb{P}_{\mu,\sigma}(X \mid Z) &= \prod_{i<j}^{n}\prod_{q,l}^{Q}\mathbb{P}_{\mu_{ql},\sigma_{ql}}(X_{i,j} \mid Z_i = q, Z_j = l) \\ &= \prod_{i<j}^{n}\prod_{q,l}^{Q}\left(\frac{1}{(2\pi)^{1/2}\sigma_{ql}}e^{-\frac{1}{2}\frac{(X_{ij}-\mu_{ql})^2}{\sigma_{ql}^2}}\right)^{Z_{iq}Z_{jl}}\end{aligned}$$

and

$$P_\alpha(Z) \quad = \quad \prod_{i}^{n}\prod_{q}^{Q}\mathbb{P}_{\alpha_q}(Z_i) = \prod_{i}^{n}\prod_{q}^{Q}\alpha_q^{Z_{iq}}.$$

The equation (1) is intractable for large networks since it requires a summation over all the possible values of Z. Thus, we propose to use the expectation-maximization (EM) algorithm. It is an iterative method that consists in computing $P_\theta(Z \mid X)$. Hence, it is intractable in this context due to the dependency of the variables X_{ij}. Therefore, we use in the sequel the variational expectation maximization (VEM) algorithm. This method overcomes the issue by maximizing a lower bound of the log-likelihood based on an approximation of the true conditional distribution of the latent variable Z given the observed variable X.

We rely on a variational decomposition of the incomplete log-likelihood as follows

$$\log \mathbb{P}_\theta(X) = J_\theta(R_X(Z)) + \mathrm{KL}(R_X(Z) \parallel P_\theta(Z \mid X)), \tag{2}$$

where $P_\theta(Z \mid X)$ is the true conditional distribution of Z given Y, $R_X(Z)$ is an approximate distribution of $P_\theta(Z \mid X)$ and KL is the Kullback-Leibler divergence between $P_\theta(Z \mid X)$ and $R_X(Z)$ defined by

$$\mathrm{KL}(R(.) \parallel P_\theta(. \mid Z)) = -\sum_Z R_X(Z)\log\frac{P_\theta(Z \mid X)}{R_X(Z)},$$

and $J(.)$ is a lower bound of $\log \mathbb{P}_\theta(X)$ of the form

$$J_\theta(R_X(.)) \quad = \quad \sum_Z R_X(Z)\log\frac{P_\theta(X, Z)}{R_X(Z)}. \tag{3}$$

The log likelihood of the incomplete data $\log \mathbb{P}_\theta(X)$ does not depend on the distribution $R_X(Z)$, thus, maximizing the lower bound J_θ with respect to $R_X(Z)$ is equivalent to minimize the Kullback Leibler divergence KL.

According to [Blei et al. (2003)], the approximate distribution $R_X(Z)$ can be factorized over the latent variables Z_i as follows

$$R_X(Z) = \prod_{i=1}^{n} R_{X,i}(Z_i) = \prod_{i=1}^{n} h(Z_i; \tau_i), \tag{4}$$

where $\{\tau_i \in [0,1]^Q, i = 1, \ldots, n\}$ are the variational parameters associated with $\{Z_i, i = 1, \ldots, n\}$ such as $\sum_q \tau_{iq} = 1, \ \forall i \in \{1, \ldots, n\}$ and h is the multinomial distribution with parameters τ_i.

By combining equations (1), (3) and (4), we obtain

$$\begin{aligned} J_\theta(R_{X,A}(Z)) &= -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q \\ &\quad + \sum_{i<j} \sum_{q,l} \tau_{iq}\tau_{jl} \Big(-\log\left((2\pi)^{1/2}\sigma_{ql}\right) - \frac{1}{2}\frac{(X_{ij} - \mu_{ql})^2}{\sigma_{ql}^2} \Big). \end{aligned} \tag{5}$$

The VEM algorithm alternates between the optimization of τ and $\theta = (\alpha, \mu, \sigma)$ until the convergence of the lower bound. During the E-step, the parameter θ of the model is fixed. We maximize $J_\theta(R_X)$ with respect to τ. Under the condition $\sum_q \tau_{iq} = 1$, for $i \in \{1, \ldots, n\}$, we obtain $\hat{\tau}$ by a fixed point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_l \left(\frac{1}{(2\pi)^{1/2}\sigma_{ql}} e^{-\frac{1}{2}\frac{(X_{ij} - \mu_{ql})^2}{\sigma_{ql}^2}} \right)^{\hat{\tau}_{jl}}. \tag{6}$$

The estimation of τ is obtained from (6) by iterating a fixed-point algorithm until convergence.

During the M-step, the parameter τ is fixed. We maximize first $J_\theta(R_X)$ with respect to α. Under the condition $\sum_q \alpha_q = 1$, we obtain

$$\hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}.$$

Then, we maximize $J_\theta(R_X)$ with respect to μ and Σ respectively, we obtain

$$\hat{\mu}_{ql} = \frac{\sum_{i<j} \tau_{iq}\tau_{jl} X_{ij}}{\sum_{i<j} \tau_{iq}\tau_{jl}},$$

and

$$\hat{\sigma}_{ql}^2 = \frac{\sum_{i<j} \tau_{iq}\tau_{jl} (X_{ij} - \hat{\mu}_{ql})^2}{\sum_{i<j} \tau_{iq}\tau_{jl}}.$$

4. Choice of the number of groups

In practice, the number of groups is unknown and should be estimated. We use the integrated classification likelihood (ICL) criterion in order to perform the selection of the most adequate number of blocks $\hat{Q}$. Roughly, this criterion is proposed by [Daudin et al.(2008)] and is based on the complete data variational log-likelihood penalized by the number of parameters.

The (ICL) is of the form

$$\begin{aligned} ICL(Q) &= \sum_{i<j} \sum_{q,l} \hat{\tau}_{iq}\hat{\tau}_{jl} \left(-\log((2\pi)^{1/2}\hat{\sigma}_{ql}) - \frac{1}{2}\frac{(X_{ij} - \hat{\mu}_{ql})^2}{\hat{\sigma}_{ql}^2} \right) - \sum_i \sum_q \hat{\tau}_{iq} \log \hat{\tau}_{iq} \\ &\quad + \sum_i \sum_q \hat{\tau}_{iq} \log \hat{\alpha}_q - \frac{1}{2} \left(Q(Q+1) \log \frac{n(n-1)}{2} + (Q-1) \log n \right). \end{aligned}$$

The VEM algorithm is run for different values of Q then $\hat{Q}$ is chosen such that ICL is maximized

$$\hat{Q} = \text{argmax}_Q(ICL(Q)).$$

5. Simulated data

First, before applying the SBM on physicochemical parameters, we perform the stochastic block model using simulated data with a Gaussian output distribution. The graph has $n = 100$ vertices. We choose a number of clusters Q equal to three.
We use in the simulation the following parameters:

$$\bar{\alpha} = (0.5, 0.25, 0.25),$$

$$\bar{\mu} = \begin{bmatrix} 20 & 5 & 5 \\ 5 & 20 & 5 \\ 5 & 5 & 20 \end{bmatrix}$$

and

$$\bar{\sigma} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Now we sample $S = 100$ random graphs according to the same mixture model. Then we calculate in Table 2, Table 3 and Table 4, for each parameter, the estimated Root Mean Square Error (RMSE) defined by:

$$RMSE(\bar{\alpha}_q) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\alpha}_q^{(s)} - \bar{\alpha}_q)^2}, \quad RMSE(\bar{\mu}_{qr}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\mu}_{qr}^{(s)} - \bar{\mu}_{qr})^2}$$

$$\text{and} \quad RMSE(\bar{\sigma}_{qr}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\sigma}_{qr}^{(s)} - \bar{\sigma}_{qr})^2},$$

where the superscript s labels the estimates obtained in simulation s.

Table 2. Root Mean Square Error of the parameter $\bar{\alpha}_q$ for the simulated data.

RMSE($\bar{\alpha}_1$)	**RMSE**($\bar{\alpha}_2$)	**RMSE**($\bar{\alpha}_3$)
0.002	0.004	0.007

Table 3. Root Mean Square Error of the parameter $\bar{\mu}_{qr}$ for the simulated data.

RMSE	$\bar{\mu}_1$	$\bar{\mu}_2$	$\bar{\mu}_3$
$\bar{\mu}_1$	0.02	0.017	0.014
$\bar{\mu}_2$	0.017	0.031	0.01
$\bar{\mu}_3$	0.014	0.01	0.05

According to Table 2, Table 3 and Table 4, we can clearly show that the RMSE of the model's parameters are close to zero. This means that the obtained estimated parameters are close to the observed parameters.

Table 4. Root Mean Square Error of the parameter $\bar{\sigma}_{qr}$ for the simulated data.

RMSE	$\bar{\sigma}_1$	$\bar{\sigma}_2$	$\bar{\sigma}_3$
$\bar{\sigma}_1$	0.07	0.013	0.04
$\bar{\sigma}_2$	0.013	0.01	0.08
$\bar{\sigma}_3$	0.04	0.08	0.01

In order to compare the estimated clustering results to the simulated ones, we propose to calculate the Adjusted Rand Index (ARI) proposed by [Hubert et al.(1985)]. It is a measure of agreement between two data partitions. The ARI has a value between 0 and 1, with 0 indicating that the two data clustering do not agree on any pair of points and 1 indicating that the data clustering is exactly the same.

The average of the ARI between the simulated clustering results and the estimated clustering results obtained using the proposed method is equal to 0.88. This means a high agreement between the two partitions of the nodes.

6. Clustering in environmental network

The data has the form of a physicochemical parameter-by-river station matrix. Each cell represents the value of the physicochemical parameter for each river station. We apply the SBM with Gaussian distributed weight method to cluster the physicochemical parameter of the considered network. Thus, we transform the matrix data into a physicochemical parameter- by- physicochemical parameter matrix of dimension 11×11. The associated network has 11 vertices connected by weighted edges. A weight associated with a pair of physicochemical parameters represents the Wasserstein distance between the values of these two physicochemical parameters. The Wasserstein distance (also known as the earth mover's distance) is a measure of the distance between two probability distributions. It is available in the package "transport" of the software R under the name "Wasserstein".
In each station, the physicochemical parameters data are collected monthly between 2008 and 2018. It has the form of a matrix of dimension 95×11, where each cell $C_{i,j}$ of this matrix represents the value of the j^{th} physicochemical parameter at the i^{th} month. We transform this matrix into a weighted matrix X of dimension 11×11, where each cell $m_{i,j}$ represents the Wasserstein distance between the i^{th} and the j^{th} physicochemical parameters.
The network associated with this matrix is formed by 11 nodes, where each node represents one of the 11 physicochemical parameters, and an edge presented between two of these parameters is weighted by the computed Wasserstein distance between this pair of parameters.
In the following, we are interested in comparing the clustering of the networks associated with the three river stations: Qaraoun, Jeb-Jenin, and Ghzayel. By applying the Gaussian SBM, we show that it reveals two clusters for each Jeb-Jenin station as shown in figure 2 by using "Gephi" software with the layout algorithm "Force Atlas".

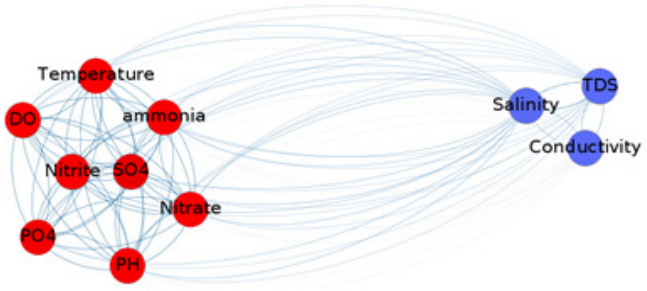

Figure 2. Grouping the physicochemical parameters of the Qaraoun station into clusters.

Several classical statistical methods are used in the literature to show the relationship between the physicochemical properties, among them we count the principal component analysis (PCA), the hierarchical classification, etc. We used in this article the SBM method, which is suitable for clustering big data to recover the community structure.
We noted that hierarchical clustering requires the measure of similarity and dissimilarity between clusters. So, it is required to determine the proximity matrix which contains the distance between each pair of physicochemical parameters using a distance function (i.e. Euclidean distance, Manhattan distance, etc.).
In the following tables, we give the weights matrix associated with the nodes.

Table 5. Weight matrix for the Jeb-Jenin station.

Jeb-Jenin	Temp.	pH	DO	Cond	TDS	Sal	Amo	Nitrite	Nitrate	SO4	PO4
Temp	0	11.67	15.34	667.46	466.12	339.12	12.13	18.45	10.8	15.22	16.27
pH		0	4.15	678.66	477.32	350.32	5.21	7.25	4.66	25.48	5.61
DO			0	682.81	481.47	354.47	4.74	3.15	5.49	29.63	2.35
Cond				0	201.34	328.34	678.56	685.91	677.48	653.17	683.043
TDS					0	127	477.22	484.57	476.14	451.83	481.7
Sal						0	350.22	357.57	349.14	324.83	354.7
Amo							0	7.35	1.75	25.38	4.48
Nitrite								0	8.43	32.73	2.88
Nitrate									0	24.3	5.6
SO4										0	29.86
PO4											0

Table 6. Weight matrix for the Qaraoun station.

Qaraoun	Temp.	pH	DO	Cond	TDS	Sal	Amo	Nitrite	Nitrate	SO4	PO4
Temp.	0	11.192	13.43	404.44	279.071	194.33	18.55	18.81	9.74	13.05	18.59
pH		0	2.32	415.63	290.26	205.52	7.36	7.62	4.91	24.13	7.66
DO			0	417.88	292.51	207.77	5.11	5.37	4.8	26.32	5.37
Cond				0	125.37	210.1	422.99	423.25	413.91	391.6	423.01
TDS					0	84.73	297.62	297.88	288.54	266.23	297.64
Sal						0	212.89	213.15	203.8	181.49	212.91
Amo							0	0.4	9.08	31.39	0.4
Nitrite								0	9.34	31.65	0.53
Nitrate									0	22.31	9.1
SO4										0	31.41
PO4											0

Based on the above tables we can conclude the following: In the three stations, two clusters are obtained after applying the SBM with Gaussian distributed weight method. TDS, salinity, and conductivity form the first cluster, and the rest of the parameters form the second one. Although the clusters contain the same parameters in all three stations, the weight matrix differs from one station to another. In other words, the relation between the parameter is more or less strong depending on the type of activity around each station.

First, we will explain the relationship between the parameters of each cluster, then we will demonstrate how can this relationship be affected by the surrounding environment.

In the first cluster, TDS, salinity, and conductivity are directly related because firstly, TDS is the dissolved solids in water, of which a high percentage is usually mineral salts, therefore, water salinity increases with high TDS

Table 7. Weight matrix for the Ghzayel station.

Ghzayel	Temp.	pH	DO	Cond	TDS	Sal	Amo	Nitrite	Nitrate	SO4	PO4
Temp	0	10.76	12.98	405.66	278.66	192.95	18.07	17.94	9.63	9.59	17.78
pH		0	2.24	416.42	289.42	203.67	7.31	7.26	3.50	5.79	7.027
DO			0	413.28	298.21	217.29	6.21	5.67	6.1	27.22	5.88
Cond				0	126.99	212.85	423.74	423.60	415.26	411.33	423.45
TDS					0	85.85	296.74	296.60	288.26	284.33	296.45
Sal						0	210.88	210.75	202.41	198.48	210.59
Amo							0	0.15	8.47	12.40	0.35
Nitrite								0	8.34	12.27	0.44
Nitrate									0	3.93	8.18
SO4										0	12.11
PO4											0

levels. Secondly, water salinity boosts its conductivity, the more saline the water, the higher its ability to pass an electrical flow.

In the second cluster, ammonia, nitrite, and nitrate are part of the nitrogen cycle. Ammonia goes through the nitrification process to produce nitrite, which will then transform into nitrate after oxidation using the dissolved oxygen in the water. This cycle plays an important role in the fluctuation of water's pH, because the presence of ammonia increases water's pH, and after nitrification, pH levels decrease again.

pH is an important parameter since it is responsible for the biodiversity in the water, which in turn affects the amount of dissolved oxygen (DO). Also, phosphate and nitrogen products initiate the eutrophication phenomenon responsible for the growth of plants and algae, and this is directly related to the decrease in DO levels, as shown by [Fadel et al.(2015)].

We can see a correlation between temperature and other parameters. Temperature affects water solubility in general, that is why we can see connections between temperature and all the other parameters, whether it is a positive or negative relation, the temperature has a direct effect on the presence of many elements in the water, and thus on its overall quality. We can also find an indirect relation, in fact, the change in some physicochemical parameters is mainly due to precipitations and springs during the wet season, where, by default, the temperature is lower than usual.

Regarding the weight matrix, the intensity of a relationship between a parameter and another is mainly affected by the type of activity around the station we are studying. For example, Qaraoun lake was initially created to generate hydropower, it is then surrounded by factories and engines. On the other hand, Jeb-Jenin is located in the Bekaa valley, where the Lebanese agricultural activity is mainly practiced, and the use of fertilizers and pesticides is uncontrolled. Contrarily, the Ghzayel river mainly contains springs and it is considered a touristic attraction, which makes it less subjected to chemicals and industrial wastes, but it is still subjected to pollution from wastewater. This difference in human activities resulted in different interactions between the physicochemical parameters. In Jeb-Jenin station, a strong relation between nitrite and conductivity is tracked, and this can be explained as follows: during the wet season, springs carry the dissolved solids into the river increasing water is conductivity, and since this station is surrounded by agricultural activity, springs will also carry phosphate, which is present in most fertilizers, from cultivated land to the river. The presence of phosphate at a high concentration will eventually lower the amount of DO as explained before, thus the oxidation of nitrite into nitrate will be inhibited resulting in higher concentrations of nitrite. In Qaraoun lake, we can see the same correlation as in Jeb-Jenin, because the type of activity and the sources of pollution around these two stations are similar. This relationship between nitrite and conductivity doesn't exist in the Ghzayel river. Instead, we can see a strong relation between ammonia and conductivity, which is a sign that the origin of pollution in the Ghzayel river comes from wastewater.

7. Estimation of the parameters

We compare here the estimated parameters obtained by applying the proposed method to the three different stations of the Litani river: Qaraoun, Jeb-Jenin and Ghzayel.
We compute now the error for the estimated parameters. Let $\hat{g}$ be the estimated community membership vectors,

Table 8. Comparison of the estimated parameters obtained by applying the proposed method for the three stations: Qaraoun, Jeb-Jenin and Ghzayel.

	Qaraoun	Jeb-Jenin	Ghzayel
$\hat{\alpha}$	(0.73, 0.27)	(0.73, 0.27)	(0.72, 0.28)
$\hat{\mu}$	$\begin{pmatrix} 12.51 & 302.23 \\ 302.23 & 140.12 \end{pmatrix}$	$\begin{pmatrix} 12.17 & 499.33 \\ 499.33 & 218.9 \end{pmatrix}$	$\begin{pmatrix} 7.23 & 265.79 \\ 265.79 & 141.95 \end{pmatrix}$
$\hat{\sigma}$	$\begin{pmatrix} 9.72 & 86.98 \\ 86.98 & 52.34 \end{pmatrix}$	$\begin{pmatrix} 9.9 & 135.65 \\ 135.65 & 83.34 \end{pmatrix}$	$\begin{pmatrix} 5.79 & 129.71 \\ 129.71 & 53.03 \end{pmatrix}$

$\hat{N}_k = \{i : 1 \leq i \leq n, \hat{g}_i = k\}$ the set of all the physicochemical parameters that belong to the cluster K and $\hat{n}_k = \mid \hat{N}_k \mid$ for all $1 \leq$ k $\leq$ Q. The plug-in estimator of μ and σ^2 are respectively

$$\tilde{\mu}_{ql} = \begin{cases} \frac{\sum_{i \in \hat{N}_q} \sum_{j \in \hat{N}_l} X_{ij}}{\hat{n}_q \hat{n}_l} & q \neq l \\ \frac{\sum_{i,j \in \hat{N}_q, i<j} X_{ij}}{\hat{n}_q(\hat{n}_q - 1)/2} & q = l \end{cases}$$

$$\tilde{\sigma}^2_{ql} = \begin{cases} \frac{\sum_{i \in \hat{N}_q} \sum_{j \in \hat{N}_l} (X_{ij} - \tilde{\mu}_{ql})^2}{\hat{n}_q \hat{n}_l} & q \neq l \\ \frac{\sum_{i,j \in \hat{N}_q, i<j} (X_{ij} - \tilde{\mu}_{ql})^2}{\hat{n}_q(\hat{n}_q - 1)/2} & q = l \end{cases}$$

We compute for the three river stations the mean error E_μ between the estimated mean connection matrix $\hat{\mu}_{ql}$ and the plug-in estimator $\tilde{\mu}_{ql}$ and the mean error E_σ between $\hat{\sigma}_{ql}$ and $\tilde{\sigma}_{ql}$ defined respectively as:

$$E_\mu = \frac{\sum_{q,l}^{Q} (\hat{\mu}_{ql} - \tilde{\mu}_{ql})^2}{Q^2}$$

$$E_\sigma = \frac{\sum_{q,l}^{Q} (\hat{\sigma}_{ql} - \tilde{\sigma}_{ql})^2}{Q^2}$$

We obtain $E_\mu = 8.75 \times 10^{-4}$, $E_\mu = 5 \times 10^{-5}$ and $E_\mu = 0.011$ for the Qaraoun, Jeb-Jenin and Ghzayel stations respectively. However for the standard deviation parameters, we obtain $E_\sigma = 0.0113$, $E_\sigma = 0.0822$ and $E_\sigma = 4.45 \times 10^{-3}$ for these stations respectively. Thus, E_μ and E_σ are close to 0 and then the results obtained using the proposed SBM method are satisfying.

8. Classical clustering method

We have already used the SBM method to classify the physicochemical parameters in the three stations of the Litani river in order to assess the pollution of the Litani river. The SBM method was successfully treated on a basis of data available over 11 years, thus it would be more interesting and useful if we apply the method of SBM on a bigdata. In order to give more of the SBM method's advantages compared to the classical classification methods which are used many times in the literature, we summarize below the different classification methods which have been used to classify the physicochemical parameters at the Quaraoun station as an example.

8.1. PCA method

Unfortunately, there is no specific method for deciding how many main axes are sufficient. In general, the choice of axes in the PCA will depend on the specific field of application and data set. In practice, we tend to look at the first major axes with the highest percentage of inertia in order to find interesting patterns in the data [Jolliffe (1986)]. For that, we choose the first two principal components which explain approximately 52% of the variation.

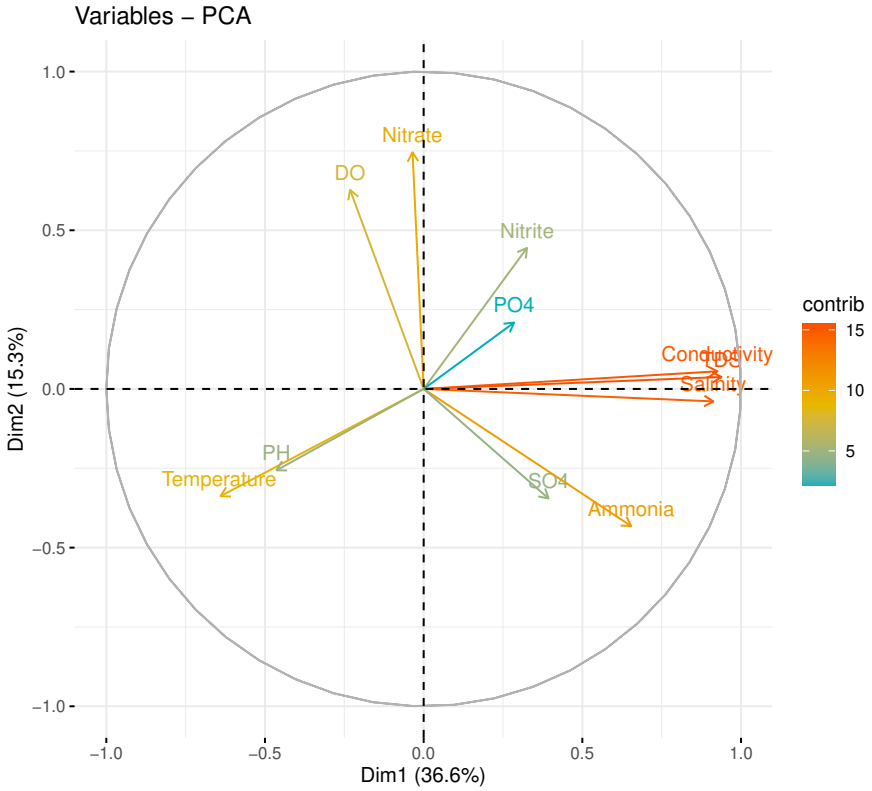

Figure 3. Correlation circle and contribution of the physicochemical parameters of the Qaraoun station.

Figure 3 shows the relationships between the physicochemical parameters. First, the positively correlated parameters are grouped together (TDS, Conductivity, Salinity). Then the negatively correlated parameters are positioned on opposite sides of the origin of the graph (opposite quadrants). The distance between the parameters and the origin measures the quality of the representation of the parameters. Parameters that are far from the origin are well represented by the PCA. We point out that the quality of representation of the parameters on the PCA graph is called cos^2 (cosine squared); the parameters with the moderate values of cos^2 will be colored in "blue" and the parameters with high values of cos^2 will be colored in "red". So, Conductivity, Salinity, and TDS are close to a circle which indicates a good representation of these parameters on the main axes. PO4 and SO4 have a low cos^2 (close to the center of the circle) which indicates that these two parameters are not perfectly represented by the first two main axes. So, we can see that the parameters are contributed on 5 axes in the below figure 4.

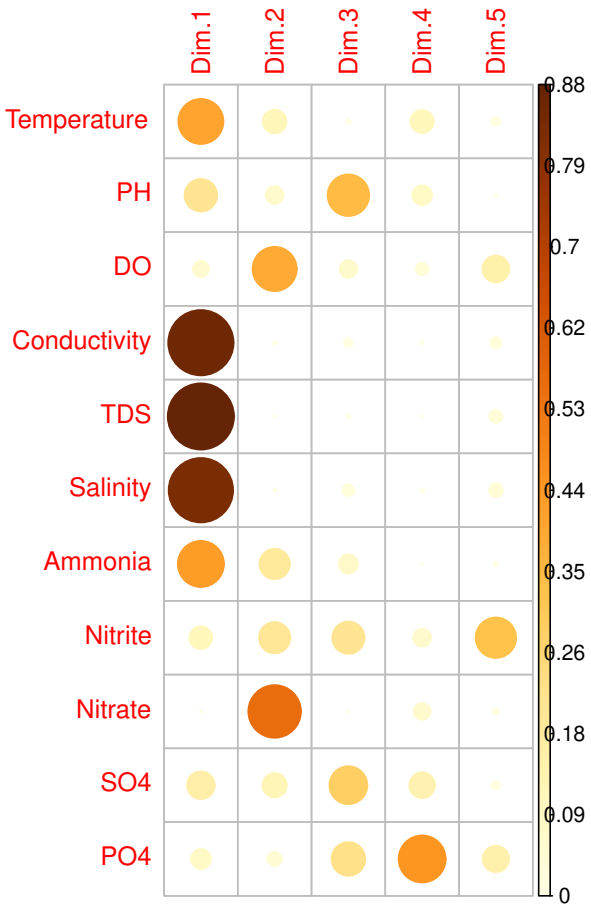

Figure 4. Contribution of the physicochemical parameters on different five axis.

We notice that TDS, Salinity, and Conductivity are related just between them by axis 1 (Dim1), while the other parameters are related to each other on different axes, we can classify the parameters according to two clusters (cluster 1: Conductivity, TDS, Salinity) and (cluster 2: Temperature, pH, DO, Ammonia, Nitrite, Nitrate, SO4, PO4).

8.2. Hierarchical Cluster

In this subsection, we present the classical hierarchical method to classify the physicochemical parameters of the Quaraoun station. Note first that the hierarchical classification does not require determining the number of classes previously, unlike the k-means method. Indeed, we can choose a number of classes by observing the dendrogram.

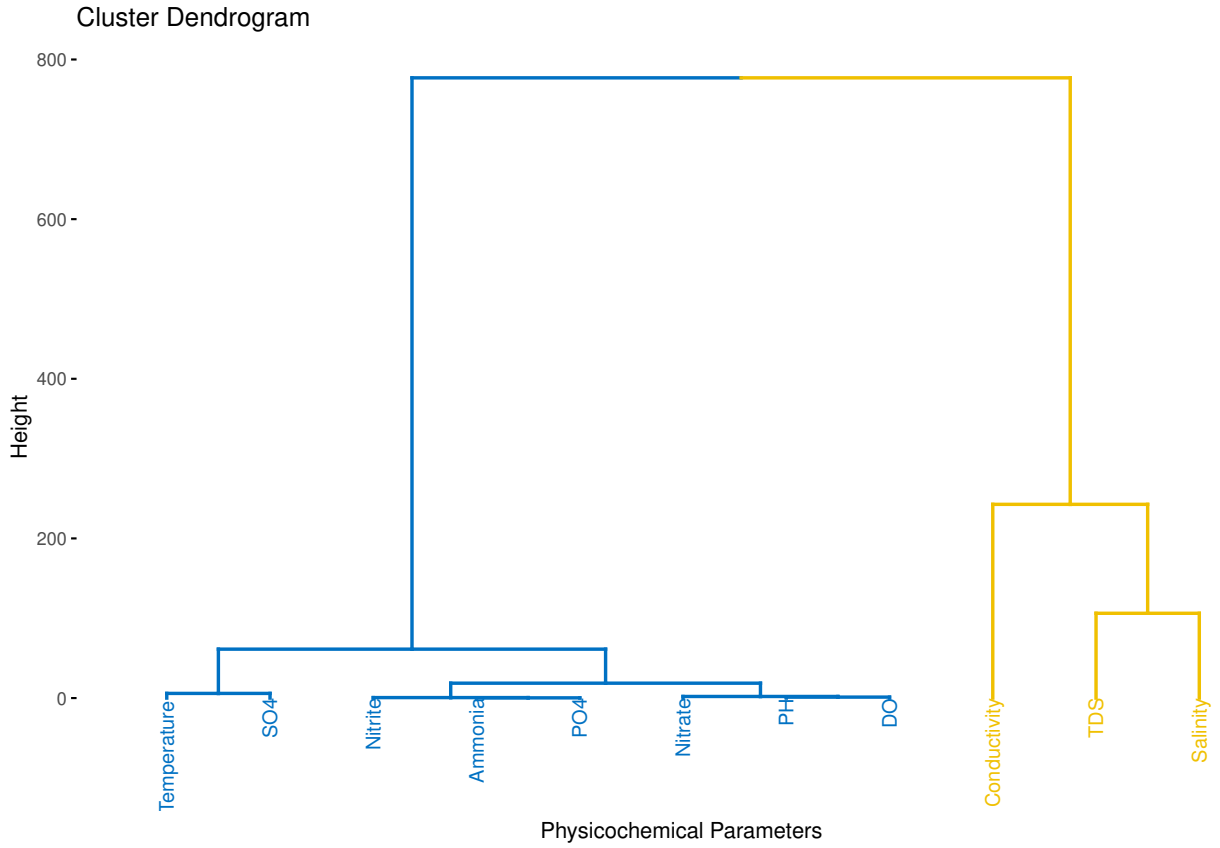

Figure 5. Dendogram of the physicochemical parameters.

Based on the figure 5, we classify the physicochemical parameters into two clusters. We also notice that the hierarchical classification gives the same distribution of the physicochemical parameters in two clusters but without knowing the intensity of connection between these parameters. In addition, the hierarchical classification has a strong algorithmic complexity in time and space. Hierarchical clustering is therefore more suitable for small samples.

8.3. K-means Cluster

The k-means method is generally simple, understandable, and applicable to large data. We note that the number of the classes must be fixed at the start by the method of k-means, in addition, it does not detect the noisy data, and the results depend on the initial drawing of the center's classes [Bradley et al. (1998)].
According to the figure 6, the k-means method classifies the physicochemical parameters into two clusters; the distances between the parameters of cluster 1 are close. Contrarily, in cluster 2 the distance between conductivity and the two parameters Salinity and TDS is far.

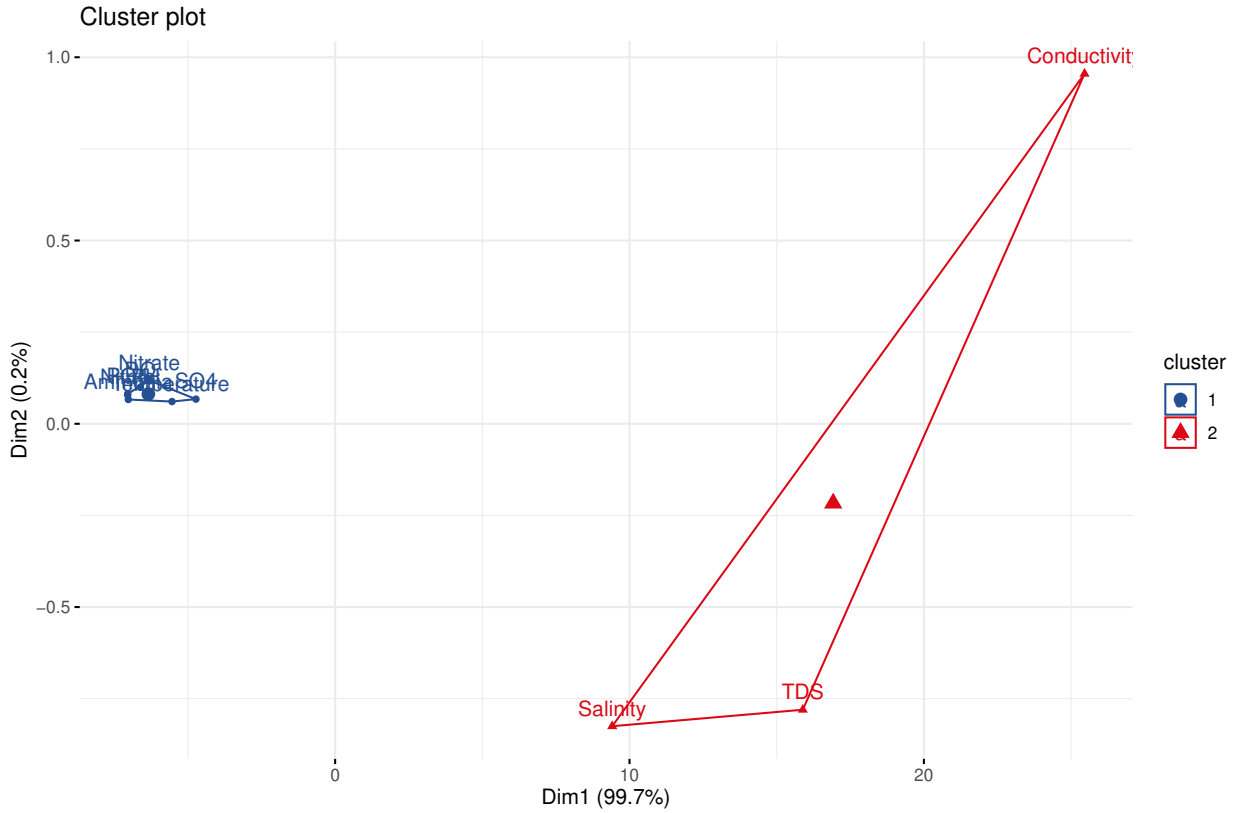

Figure 6. K-means cluster of the physicochemical parameters.

We notice that the two k-means and hierarchical methods classify the parameters into two groups similar to the SBM method, then the three classification methods are indeed applicable to the datasets of the physicochemical parameters of the Quaraoun over a period between 2008 and 2018. Note that each of these methods has different characteristics that are applicable according to the types of data.

9. Conclusion

In this article, we have introduced a new approach in the environmental field based on statistical models for the classifications of physicochemical parameters at the Litani River in Lebanon, in particular the SBM. This method differs from the other classical methods since it can show the block structure of the three station networks by estimating the connection matrices between the obtained clusters and the membership vectors belonging of the physicochemical parameters to each cluster of these three stations. Using the SBM, we were able to identify two clusters in each station. The clusters were similar in each station, however, the significance of the correlation between the parameters of each network was different. We noticed that the type of human activity around the stations was mainly responsible for this difference in correlations. We explained the different types of pollution caused by certain activities and how they can change the behavior of these parameters and how they react together. Thus, we found that to make a good plan to improve water quality, it is best to locate the station correctly, study the type of activity around it, and specify the type of pollution to which the station is subjected. After analyzing the results, we can deal with the parameters as groups of related parameters instead of treating each parameter separately, which is cost-effective and time-saving.

REFERENCES

[Ali H. et al.(2020)]. Hayek, A. , Tabaja, N. , Andaloussi, S. , Toufaily, J. , Garnie-Zarli, E. , Toufaili, A. and Hamieh, T. Evaluation of the Physicochemical Properties of the Waters on the Litani River Station Qaraoun, American Journal of Analytical Chemistry, 11 (2020) 90-103. doi: 10.4236/ajac.2020.112007.

[Airoldi et al.(2008)]. Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic block models. The Journal of Machine Learning Research, 9 (2008) 1981-2014.

[Anderson et al.(1992)]. Anderson, C. J., Wasserman, S., and Faust, K. Building stochastic block models, Social Networks, 14 (1992) 137-161.

[Barbillon et al.(2017)]. Barbillon, P., Donnet, S., Lazega, E., and BarHen, A. Stochastic block models for multiplex networks: an application to a multilevel network of researchers, Journal of the Royal Statistical Society: Series A (Statistics in Society), 180 (2017) 295-314.

[Blei et al. (2003)]. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation, The Journal of Machine Learning Research, 3 (2003) 993-1022.

[Bradley et al. (1998)]. Bradley, P.S., Fayyad, U.M., Reina, C. Scaling clustering algorithms to large databases. In: KDD'98 proceedings of the fourth international conference on knowledge discovery and data mining, (1998) 9–15.

[Celisse et al. (2012)]. Celisse, A., Daudin, J. J., and Pierre, L. Consistency of maximum-likelihood and variational estimators in the stochastic block model, Electronic Journal of Statistics, 6 (2012) 1847-1899.

[Chaden et al. (2014)]. Chaden M.H., Nada N, Sadek A, Bachar K., Mohamad F., Ali Y., Joumana T., Frederic V.,Tayssir H. Water Quality of the upper Litani River Basin, Lebanon, Physics Procedia, 55 (2014) 279-284.

[Daudin et al.(2008)]. Daudin, J., Picard, F., and Robin, S. A mixture model for random graph, Statistics and Computing, 18 (2008) 1-36.

[Dudgeon et al. (2006)]. Dudgeon D, Angela HA, Gessner MO, Kawabata ZI, Knowler DJ, Lévêque C, Naiman RJ, Prieur RAH, Soto D, Stiassny MLJ, Sullivan CA. Freshwater biodiversity: importance, threats, status, and conservation challenges. Biological Reviews of the Cambridge Philosophical Society, 81(2) (2006) 163-182 doi: 10.1017/s1464793105006950.

[El Haj et al.(2020)]. El Haj, A., Slaoui, Y., Louis, P-Y., and Khraibani, Z. Estimation in a Binomial Stochastic Block model for a Weighted Graph by a Variational Expectation Maximization Algorithm, Communication in Statistics Simulation and Computation, (2020) doi:10.1080/03610918.2020.1743858.

[Fadel et al.(2014)]. Fadel, A., Lemaire, B. J., Atoui, A., Vinçon-Leite, B., Amacha, N., Slim, K., Tassin, B. First assessment of the ecological status of Karaoun reservoir, Lebanon. Lakes Reservoirs, Research Management, 19(2) (2014) 142-157.

[Fadel et al.(2019)]. Fadel, A., Sharaf, N., Siblini, M., Slim, K., Kobaissi, A. A simple modeling approach to simulate the effect of different climate scenarios on toxic cyanobacterial bloom in a eutrophic reservoir, Ecohydrology Hydrobiology, 19(3) (2019) 359-369.

[Fadel et al.(2015)]. Fadel, A., Atoui, A., Lemaire, B. J., Vinçon-Leite, B., Slim, K. Environmental factors associated with phytoplankton succession in a Mediterranean reservoir with a highly fluctuating water level. Environmental Monitoring and Assessment, 187(10) (2015) 633.

[Fadel et al.(2021)]. Fadel, A., Kanj, M. Slim, K. Water Quality Index variations in a Mediterranean reservoir: a multivariate statistical analysis relating it to different variables over 8 years. Environ Earth Sci 80, 65 (2021). https://doi.org/10.1007/s12665-020-09364-x.

[Fortunato, S.(2010)]. Fortunato, S. Community detection in graphs, Physics Reports, 486 (2010) 75-174.

[Groppo et al.(2008)]. Groppo, JD., De Moraes JM., Beduschi, CE., Genovez, AM. and Martinelli, LA. Trend analysis of water quality in some rivers with different degrees of development within the São Paulo State, Brazil. River Res. Applic., 24(8) (2008) doi: 1056-1067. doi:10.1002/rra.1091.

[Esterby (1993)]. Esterby, S.R. Trend analysis methods for environmental data, Environmetrics, 4 (1993) 459-481, doi:10.1002/env.3170040407.

[Hayek et al.(2021)]. Hayek, A. , Andaloussi, S. , Tabaja, N. , Toufaily, J. , Garnie-Zarli, E. and Hamieh, T. Multivariate Spatial and Temporal Analysis to Study the Variation of Physico-Chemical Parameters in Litani River, Lebanon. American Journal of Analytical Chemistry, 12, (2021), 373-391. doi: 10.4236/ajac.2021.1211023.

[Holland et al.(1983)]. Holland, P. W., Laskey, K. B., and Leinhardt S. Stochastic blockmodels: First steps, Social Networks, 5 (1983) 109-137.

[Jolliffe (1986)]. I. Jolliffe. Principal Component Analysis. Springer-Verlag, New York, 1986.

[Hubert et al.(1985)]. Hubert, L. and Arabie, P. Comparing partitions, Journal of Classification, 2 (1985) 193-218.

[Latouche et al.(2011)]. Latouche, P., Birmelé, E., and Ambroise, C. Overlapping stochastic block models with application to the French political blogosphere, Annals of Applied Statistics, 5 (2011) 309-336.

[Latouche et al.(2012)]. Latouche, P., Birmelé, E., and Ambroise, C. Variational Bayesian inference and complexity control for stochastic block models, Statistical Modelling, 12, (2012) 93-115.

[Mariadassou et al.(2010)]. Mariadassou, M., Robin, S., and Vacher, C. Uncovering latent structure in valued graphs: a variational approach. Annals of Applied Statistics, 4 (2010) 715-742.

[Mark et al.(2012)]. Mark S., Lucy S., Nabil A. Physicochemical Evaluation of Upper Litani River Watershed, Lebanon, Scientific World Journal, (2012) 1-8.

[Matias and Miele (2017)]. Matias, C., and Miele, V. Statistical clustering of temporal networks through a dynamic stochastic block model, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79 (2017) 1119-1141.

[Ng and Murphy (2021)]. Ng, T.L.J., Murphy, T.B. Weighted stochastic block model. Stat Methods Appl 30, 2021, 1365–1398, https://doi.org/10.1007/s10260-021-00590-6

[Nowicki and Snijders (2001)]. Nowicki, K., and Snijders, T. A. B. Estimation and prediction for stochastic block structure, Journal of the American Statistical Association, 96 (2001) 1077-1087.

[Qiang et al.(2019)]. Qiang He, Brian R. Silliman, Climate Change, Human Impacts, and Coastal Ecosystems in the Anthropocene, Current Biology, Volume 29, Issue 19, 2019, Pages R1021-R1035, ISSN 0960-9822, https://doi.org/10.1016/j.cub.2019.08.042.

[Ridzuan et al.(2020)]. Ridzuan, N.H.A.M., Marwan, N.F., Norlin, K., Ali, M.H., Tseng, M.-L. Resources, Conservation Recycling Effects of agriculture, renewable energy, and economic growth on carbon dioxide emissions: evidence of the environmental Kuznets curve.

Resour. Conserv. Recycl. 160 (2020) (January), 104879. https://doi.org/10.1016/j.resconrec.2020.104879.

[Simpi et al.(2011)]. Simpi B., Hiremath SM., Murthy KNS. Analysis of Water Quality Using physico-Chemical Parameters Hosahalli Tank in Shimoga District, Karnataka, India, Global Journal of Science Frontier Research, 3 (2011) 31-34.

[Snijders and Nowicki (1997)]. Snijders, T. A., and Nowicki, K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure, Journal of Classification, 14 (1997) 75-100.

[Zreik et al.(2017)]. Zreik, R., Latouche, P., and Bouveyron, C. The dynamic random subgraph model for the clustering of evolving networks, Computational Statistics, 32 (2017) 501-533.