Assessment of water quality using stochastic block model method

Alya ATOUI

Supevisors: Régis Moilleron (UPEC), Zaher Khraibani (LU)
Co-supervisors: Samir Abbad Andalousi (UPEC), Kamal Slim (CNRSL)

ASMSA 2020-Poitiers

December 11, 2020

laboratoire eau environnement systemes urbains

Why do we monitor water quality?

Monitoring water quality is important for:

- The assessment of water pollution.
- Determining the proper use of the available water.
- Protecting water resources from deterioration.

What causes water pollution?

Pollution of water has many sources:

- Wastewater.
- Industrial waste.
- Stormwater discharge.
- Pesticides and fertilizers used in agriculture.

Previously Used Methods

- Data analysis (PCA), *(Hayek et al. 2020)*.
- Descriptive & Inferential statistics, *(Diab, W. 2018)*.
- Classical Cluster analysis *(k-means, Hierarchical clustering)*.

Stochastic Block Model

Definition [Nowicki and Snijders (2001)]

The stochastic block model is a random probabilistic graph model which aims to produce classes, called blocks, or more generally clusters in networks.

It takes the following parameters:

- The number of nodes n.
- A partition of the set of nodes $\{1, ..., n\}$ into Q subsets disjoint $C_1, ..., C_Q$ called "Communities"
- A probability matrix of edges of dimension $Q \times Q$.

Clustering Methodology

Notation

Let X be the symmetric weighted matrix of dimensions $n \times n$ encoding the intensity of the observed interactions between nodes.

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

Where n is the number of weighed nodes.

Clustering Methodology

Notation

We denote by Z the binary indicator matrix labeling the assignment of the physicochemical parameters into groups.

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

Where Q is the number of clusters.

Mixture Model With Latent Classes

We propose to generate the stochastic block model 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 $\sum_{q=1}^{Q} \alpha_q = 1$.

Mixture Model With Latent Classes

- 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:

$$X_{ij}|Z_{iq}Z_{jl} = 1 \sim \mathcal{N}(\mu_{ql}, \sigma^2_{ql}),$$

where μ_{ql} and σ^2_{ql} denotes respectively the mean and the covariance parameters associated to the Gaussian distribution.

Inference

Estimate $\theta = (\alpha, \mu, \Sigma)$.
The log-likelihood of the incomplete data:

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

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

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

Inference

where

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

and

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

Variational Expectation Maximization (VEM) algorithm

By using VEM we obtain:

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

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

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

Choice of The Number of cluster

- The number of groups is unknown.
- Integrated Classification Likelihood (ICL) is used to estimate the most adequate number of groups.

The ICL is of the form:

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

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

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

The Litani River

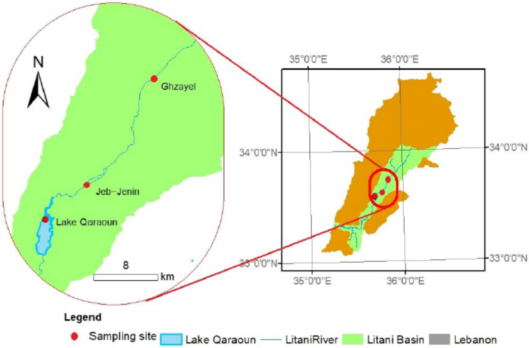

Figure 1: Location of the stations.

Litani River Data

- Samples were collected from three different stations *(Qaraoun, Ghzayel, Jeb-jenine)*.
- Monthly measurements over a period of 10 years (2008-2018), *data dimension* $(12 \times 10, 11)$.
- 11 physicochemical parameters were measured and recorded in each stations.

The physicochemical parameters are: Temperature, pH, TDS, Salinity, Conductivity, Ammonia, Nitrite, Nitrate, Sulfate, Phosphate.

Clusters

By applying the Gaussian SBM, we obtained the following clusters:

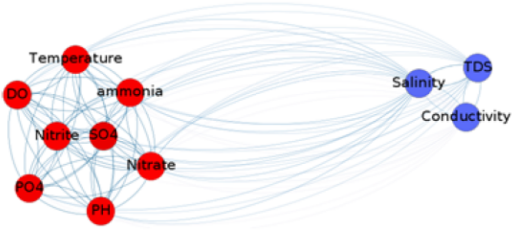

Figure 2: Grouping the physicochemical parameters into clusters.

Clusters

- In the three stations, two clusters are obtained
- TDS, salinity, and conductivity form the first cluster
- The rest of the parameters form the second one

Weight Matrix

The difference between the three stations is in the weight matrix

Jeb-Jenine	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

Figure 3: Weight matrix for the Jeb-Jenine station.

Weight Matrix

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

Figure 4: Weight matrix for the Qaraoun station.

Weight Matrix

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

Figure 5: Weight matrix for the Ghzayel station.

Importance of The SBM Method

- Group the parameters into clusters.
- Describe the relationship between the deduced groups.
- Create and describe a variety of different structures.
- Cover a wide range of data.

Analysis of The Results

- The parameters are divided into clusters depending on the natural interaction between them.
- The magnitude of the weight matrix is a result of the type of pollution within the water body.

Analysis of The Results

The relationship between the parameters depends on two factors:

- The natural interaction between the parameters.
- The type of pollution in the station.

Enhancing Water Quality Based On The Results

- Treating the parameters as groups instead of elements.
- Understand the relationship between the parameters.
- Identify the element with the greatest impact on the others.

References

1. El Haj, A. et al.(2020). Estimation in a Binomial Stochastic Blockmodel for a Weighted Graph by a Variational Expectation Maximization Algorithm. Communication in Statistics Simulation and Computation.
2. Anderson, C. J. et al. (1992). Building stochastic blockmodels. Social Networks,14, 137–161.
3. Celisse, A.et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics, 6, 1847–1899.
4. Diab, W. (2018) étude des propriétés physico-chimiques et colloïdales du bassin de la rivière Litani, Liban.
5. Hayek et al. (2020). Evaluation of the Physico-Chemical Properties of the Waters on the Litani River Station Quaraoun. American Journal of Analytical Chemistry, February 2020.
6. Holland, P.et al. (1983). Stochastic blockmodels: First steps.