Modeling of the environmental pollution : applications to the Litani River and the air pollution in Bekaa Valley (Lebanon)

Alya Atoui

Thèse de doctorat de l'Université Paris-Est Créteil

École doctorale Sciences, Ingénierie et Environnement (SIE)

Modeling of the environmental pollution: applications to the Litani River and the air pollution in Bekaa Valley (Lebanon)

Alya ATOUI

en vue de l'obtention du grade de Docteur de l'Université Paris-Est Créteil

Spécialité : Sciences et Techniques de l'Environnement

Thèse préparée au Laboratoire, Eau, Environnement et Systèmes Urbains

&

Rammal Hassan Rammal Research Laboratory, Environmental Physio-Toxicity group , Faculty of Sciences, Lebanese University

Thèse soutenue le
18 Janvier 2023

Composition du Jury :

Mireille BATTON-HUBERT	Professeure, Ecole des Mines de Saint-Etienne	Rapporteure
Sarah DORNER	Professeure, Polytechnique Montréal	Rapporteure
Sophie AYRAULT	Directrice de Recherche, LSCE	Examinatrice
Ouafae BENRABAH	M.C.F., Université du Littoral côte d'Opale	Examinatrice
Régis MOILLERON	Professeur, Université Paris-Est Créteil	Directeur de thèse
Zaher KHRAIBANI	Professeur, Université Libanaise	Directeur de thèse
Samir ABBAD ANDALOUSSI	M.C.F, Université Paris-Est Créteil	Co-encadrant
Kamal SLIM	Professeur-CNRS-Libanais	Co-encadrant
Anis HOAYEK	M.C.F, Ecole des Mines de Saint-Etienne	Invité

« Man's Role in Changing the Face of the Earth »

List of publications

Articles:

1. Atoui, A., El Haj, A., Slaoui, Y., Fadel, A., Slim, K., Abbad Andaloussi, S., Moilleron, R. & Khraibani, Z. (2022). Spatial assessment of water river pollution using the stochastic block model: Application in the different station in the Litani river, Lebanon. *Statistics, Optimization & Information Computing*, vol. 10 (September 2022), 1204-122. https://doi.org/10.19139/soic-2310-5070-1547.
2. Atoui, A, Abbad Andaloussi, S. A., Slim, K., Moilleron, R. & Khraibani, Z. (Appear-2023). Prediction and analysis of the Extreme and Records values of air pollution data in Bekaa Valley in Lebanon. *International Journal of Environmental Science and Development*, vol.14(1) preprint submitted to Elsevier **(Accepted).**
3. Atoui, A, Abbad Andaloussi, S., Slim, K., Moilleron, R. & Khraibani, Z. (2022). Assessment and modeling of the chaotic variation of the physicochemical parameters at Qaraoun Lake, Lebanon, *Chaotic Modeling and Simulation* (CMSIM), *An International Journal of Nonlinear Science*, 4: 121–138, 2022**.**
4. Atoui, A, Abbad Andaloussi, S., Slim, K., Moilleron, R. & Khraibani, Z. (2022). Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air pollutants in Zahleh, Lebanon, *Atmospheric and Climate Sciences* (ACS), vol. 12(4), 728-749. https://doi.org/10.4236/acs.2022.124040.

Conferences and seminars:

1. **CHAOS2022 International conference:**

Atoui, A, Abbad Andaloussi, S., Slim, K., Moilleron, R. & Khraibani, Z. Assessment and modeling of the chaotic variation of the physicochemical parameters at Qaraoun Lake, Lebanon. *15th CHAOS2022*, 14-17 June 2022, Athens, Greece. (http://www.cmsim.org/images/CHAOS2022_program-1.pdf).

2. **ASMDA2021 International conference:**

Atoui, A, Moilleron, R., Khraibani, Z., Abbad Andaloussi, S. & Slim, K. Prediction and analysis of the Extreme and Records values of air pollution data. *19th conference of the Applied Stochastic Models and Data Analysis International Society and the Demographics Workshop* [Athens- Greece, June 2021] Oral presentation (http://www.asmda.es/images/ASMDA2021_program-final-.pdf).

3. **ASMSA2020: Algorithmes Stochastiques, Modélisation Statistiques et Applications**

Atoui, A, Moilleron, R., Khraibani, Z., Abbad Andaloussi, S. & Slim, K. Application of the Stochastic Block Model to study the effect of the pollution in the Litani river, Lebanon, *Bordeaux-Poitiers (Scientific journey)* [10/12/2020 – 11/12/2020] – Oral Presentation.

Acknowledgments

First and foremost, I would like to thank my thesis advisors, professor Regis Moilleron and Zaher Khraibani for their continuous guidance throughout my research. I express all my gratitude to their presence and for always sharing ideas for improvements. I couldn't have accomplished this work without their support.

I would also like to thank my co-advisors, professor Samir Abbad Andaloussi and Kamal Slim for their follow up and for providing me with useful sources of information.

Besides my advisors, I would like to thank the Litani River Authority and the Ministry of Environment for providing me with the data on air and water quality analysis.

Also, I would like to thank the jury members, Professor, Mireille Batton-Hubert, Sarah Dorner, Sophie Ayrault and Ouafae Benrabah for accepting to evaluate and discuss this work in my defense.

Last but not least, I acknowledge my family, especially my parents, for their support throughout my life and my studies.

Résumé

La pollution de l'air, de l'eau et des sols, les effets du changement climatique, etc., sont des problèmes majeurs auxquels l'humanité est confrontée ces dernières décennies avec l'essor de l'anthropisation. L'objectif de cette thèse est de développer de nouvelles méthodes afin d'apporter des éléments de réponses pour faire face aux enjeux environnementaux, d'abord sur la qualité de la ressource en eau, qui constitue un véritable enjeu de société dans de nombreux pays, ensuite sur la qualité de l'air et ses conséquences en termes de pollution atmosphérique et donc de santé publique. Différentes approches ont ainsi été évaluées afin de proposer de nouveaux outils, qui se veulent efficaces et accessibles aux gestionnaires. Ces outils conduisent à la création de systèmes d'information basés sur la modélisation statistique qui permettent (i) d'étudier la dynamique des polluants environnementaux et (ii) de prédire leur évolution à court et moyen termes.

Ainsi, la pollution de la rivière Litani qui représente, à la fois, un risque sanitaire (en raison de la dégradation de la ressource en eau pour la production d'eau potable) et un risque environnemental (perte de biodiversité) au Liban, a été suivie. Pour cela les paramètres physicochimiques disponibles sur plusieurs stations réparties sur la rivière Litani ont été collectés et exploités. Une méthode de classification dite « *Stochastic Block Model* », nouvelle en sciences de l'environnement, a été explorée pour classer les paramètres physicochimiques selon des approches inter- et intra- stations de surveillance. Ensuite, en utilisant la théorie du chaos sur une période de 11 ans (2008-2018), le comportement ainsi que l'évolution des paramètres physicochimiques ont été étudiés.

Cette thèse a également considéré la pollution atmosphérique. Cette dernière constitue une des principales causes de dégradation de l'environnement au Liban. Les principales sources de polluants atmosphériques de la région de la Vallée de la Bekaa (sources de combustion, activités industrielles - raffineries, cimenteries, etc.) ont été considérées. Le but était, là encore, d'implémenter de nouveaux modèles statistiques en utilisant la théorie des valeurs extrêmes et la théorie des records pour mettre en place des stratégies de réduction des principaux polluants. Enfin, la prédiction de l'indice de qualité de l'air sur une courte période a été investiguée en utilisant différentes méthodes de prévisions.

Mots clés : Pollution de l'eau, Paramètres physicochimiques, Pollution atmosphérique, Théorie des valeurs extrêmes, Théorie de records, Théorie Chaotique, Série temporelle, Litani, Liban.

Abstract

Air, water and soil pollution, the effects of climate change, etc., are major problems that humanity is facing in recent decades with the rise of anthropization. The objective of this thesis is to develop new methods in order to bring elements of answers to face the environmental stakes, first on the quality of the water resource, which constitutes a real stake of society in many countries, then on the quality of the air and its consequences in terms of air pollution and thus of public health. Different approaches have been evaluated in order to propose new tools that are effective and accessible to managers. These tools lead to the creation of information systems based on statistical modeling which allow (i) to study the dynamics of environmental pollutants and (ii) to predict their evolution in the short and medium term.

Thus, the pollution of the Litani River which represents, at the same time, a sanitary risk (because of the degradation of the water resource for the production of drinking water) and an environmental risk (loss of biodiversity) in Lebanon, was followed. For this purpose, the physicochemical parameters available on several stations distributed on the Litani River were collected and exploited. A classification method called "Stochastic Block Model", new in environmental sciences, was explored to classify the physicochemical parameters according to inter- and intra-station monitoring approaches. Then, using chaos theory over a period of 11 years (2008-2018), the behavior as well as the evolution of physicochemical parameters were studied.

This thesis also considered air pollution. The latter is one of the main causes of environmental degradation in Lebanon. The main sources of air pollutants in the Bekaa Valley region (combustion sources, industrial activities - refineries, cement factories, etc.) were considered. The aim was, again, to implement new statistical models using the extreme value theory and the record theory in order to set up reduction strategies for the main pollutants. Finally, the prediction of the air quality index over a short period of time was investigated using different forecasting methods.

Keywords: Water pollution, Physicochemical parameters, Air pollution, Extreme value theory, Record theory, Chaos theory, Time series, Litani, Lebanon.

List of abbreviations

ACF	Autocorrelation Function
AIC	Akaike Information Criterion
AMI	Average Mutual Information
ANOVA	Analyse Of Variance
AQI	Air Quality Index
ARI	Adjusted Rand Index
BLUP	Best Linear Unbiased Predictor
BM	Block Maxima
CA	Cluster Analysis
CCA	Canonical Correlation Analysis
CDM	Correlation Dimension Method
CI	Correlation Integer
CO_2	Carbon Dioxide
DA	Discriminant Analysis
DO	Dissolved Oxygen
E. coli	Escherichia coli
ERML	Environmental Resources Monitoring in Lebanon
EVT	Extreme Value Theory
FA	Factor Analysis
FNN	False Nearest Neighbor
GEV	Generalized Extreme Value
GHG	Greenhouse Gas
GPD	Generalized Pareto Distribution
HAC	Hierarchical Agglomerative Clustering
ICL	Integrated Classification Likelihood
MAE	Mean Absolute Error
MANOVA	Multivariate Analysis Of Variance
MEF	Mean Excess Function

MLE	Maximum Likelihood Estimation
MoE	Ministry of Environment
MSE	Mean Square Error
NA	Missing Data
NH_3	Ammonia
NH_4	Ammoniaque
NO_2	Nitrite
NO_3	Nitrate
NO_X	Nitrogen Oxides
O_2	Oxygen
OOB	Out-Of-Bag
PCA	Principal component analysis
PO4	Phosphate
POT	Peak Over Threshold
RE	Relative Error
RF	Random forest
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SBM	Stochastic Block Model
SD	Standard Deviation
SO_4^{2-}	Sulfate
TBATS	Trigonometric seasonality, Box-Cox Transformation, ARMA Errors, Trend and Seasonal Components
TDS	Total Dissolve Solides
TOE	Tonnes of Oil Equivalent
UNDP	United Nations Development Program
UNEP	United Nations Environment Program
VEM	Variational Expectation Maximization
VOC	Volatile Organic Compound
WHO	World Health Organization

List of Figures

List of Tables

Table of Contents

General Introduction

Pollution in its simplest definition is the introduction of contaminants to the environment, causing some sort of imbalance in its natural cycles or functioning. It can be caused by natural phenomena such as volcanoes and wildfires, or by human activities which are the main contributors to pollution worldwide.

Pollution is generally classified as originating from point sources and non-point sources. According to the US Environmental Protection Agency, "point source" refers to "*any discernible, confined and discrete conveyance, including but not limited to any pipe, ditch, channel, tunnel, conduit, well, discrete fissure, container, rolling stock, concentrated animal feeding operation, or vessel or other floating craft, from which pollutants are or may be discharged. This term does not include agricultural storm water discharges and return flows from irrigated agriculture.*" Therefore point source pollution is when there is one easily defined source of pollution. For example, the pollution of air from an industrial source, or water pollution from a power plant or a water treatment plant. According to the Organisation for Economic Co-operation and Development (OECD), "*Non-point sources of pollution are pollution sources that are diffused and without a single point of origin or not introduced into a receiving stream from a specific outlet.*" The pollutants are generally carried off the land by storm water run- off. In the case of non-point source pollution, it's difficult to trace pollution back to a single source, and is therefore, harder to control because it's the outcome of the daily activities of many different people. Urbanization is one major cause of non-point source pollution. It is altering the ecosystem on a high scale and leading to many changes in the environmental composition.

The effects of pollution are numerous. Some of these effects are climate change, deforestation, reducing biodiversity, and many effects on human health that lead to premature death. In 2015, pollution killed 9 million people (Beil, 2017). The United Nations predict an increase of the number of people living on the planet by 2050 from 7.7 billion to 9.7 billion. During this period, the world's population is expected to become increasingly urbanized. This can only mean more energy usage, more demand for natural resources, and of course more pollution.

All human activities can directly affect the environment. Transportation, agriculture, and the use of electricity in daily life are only some examples of how humans are increasing the levels of pollution daily. Today, most human activities involve burning fossil fuels. Whether it's to generate electricity, to run a vehicle, or simply for heating, the result is more fossil fuel burning with every slight increase in population. According to the BP statistics, humans were burning 7.1 Gtoe or billion "tonnes of oil equivalent" in fossil fuels per year back in 1992. In 2020, this number increased to 11.7 Gtoe (Saxifrage, 2019). This means not only that the damaging energy use is rising, but also that each year, more clean energy from environmentally friendly alternatives needs to be added to replace fossil burning. One direct consequence of burning fossil fuels is climate change due to the emission of carbon dioxide

(CO_2), a greenhouse gas that contributes to trapping heat in the atmosphere and warming the planet. Global CO_2 emissions from energy use increased from 31.2 billion metric tons in 2010 to 36.4 billion metric tons in 2020, and it's expected to reach 45.5 billion metric tons in 2040. (Saxifrage, 2019).

Climate change is a serious environmental problem that menaces human beings' existence, biodiversity, marine life, and the planet as a whole. With climate change comes more frequent and intense drought, storms, heat waves, rising sea levels, melting glaciers, and warming oceans. Another consequence of climate change is the increase in the rate of wildfires. Many regions suffered in the last decade from the spread of enormous and unstoppable wildfires. More recently, the summer of 2022 was a cruel illustration of the spread of these wildfires around the world. This phenomenon is not only a result of climate change that's initially caused by pollution, but also a burning process that leads to more toxic emissions and to losing more green spaces leading to deforestation. With more deforestation, the natural air purification process is disrupted. Hence, we can look at this process as a continuous cycle; burning fossil fuels results in CO_2 emission, which is one of the Greenhouse Gas "GHG" responsible for global warming, and with global warming comes more environmental impacts like wildfires that again release more CO_2 and particulate matter (PM) to the atmosphere resulting in more pollution. A recent study showed that if the CO_2 levels doubled from the preindustrial time, the global temperature will increase by 2.6 to 4.1 degrees Celsius (Sherwood, et al., 2020). A study was conducted in 2009 to show the effect of climate change on surface water quality, there was a clear link between the rising temperature and the decrease in water quality. It was found that lower flows and reduced flow velocity caused by high temperatures mean higher water residence times in rivers and lakes, which will increase the potential for toxic algal blooms which by its turn have a high oxygen consumption, thus dissolved oxygen levels in the water will be decreasing (Whitehead, et al., 2009). In a more recent study, climate change and frequent drought were linked to an increase in water salinity and a decrease in nutrients and water turbidity (Mosley, 2015).

As for the relationship between climate change and air quality, several studies explained the impact of rising temperature on the atmosphere (Gonzalez-Abraham, et al., 2015) (Fuzzi, et al., 2015). It was proved that climate change alone can increase summertime surface ozone levels in polluted regions by 1–10 ppb over the coming decades (Jacob, et al., 2009). The increase in ozone concentration can impact human health since it aggravates asthma and causes inflammation in the respiratory tract. It can also reduce agricultural crops and commercial forest yield.

Furthermore, the demand for natural resources is increasing due to population growth and higher living conditions. The transition to an urbanizing world is altering the ecosystem and disrupting its services (Huang, et al., 2010). We are now uncertain if the ecosystem can sustain food production for the growing population and maintain freshwater and forest resources, especially when more and more land

is being used for urbanization activities. This calls for an urgent assessment and monitoring of the available resources concerning the population in each region. It is also important to know that even if some resources are classified as renewable, we need to pay attention to the quality of these resources. The world health organization (WHO) sets clear and strict limitations for water usage according to its quality, also for the quality of breathable air and the duration of exposure to polluted air which can lead to many severe health problems or even death. Therefore, technically speaking, the availability of natural resources doesn't necessarily mean that humans can benefit from them. In fact, the presence of polluted natural resources like air and water has a serious effect on human existence.

Starting from these facts, we aim in this study to evaluate the available natural resources, especially air and water to make a bigger understanding of the actual situation of these two important resources. We took the Litani River and the Bekaa region in Lebanon as a case study. The Litani river is the most important river in Lebanon; it contains the largest artificial lake, Qaraoun Lake. Also, it has a variety of human activities at its banks. It is where the majority of agricultural activities in Lebanon are practiced, in addition to some industrial activities, touristic attractions, and residential areas. Studying air and water quality in the Litani River and the Bekaa region was an essential part of understanding how growing population, urbanization, and different forms of human activities are affecting the environment and the quality of both air and surface water. We implemented different statistical methods to model air and water pollution in the area. We also conducted a deep environmental analysis to show the relevance of our models and to clearly describe the relationship between human activities and pollution, and the relationship between the pollutants and how they affect one another.

This study aims to identify the quality of both air and surface water in Lebanon focusing on six objectives:

- Modeling water pollution in different stations in the biggest river in Lebanon (the Litani River)
- Showing the relationship between water pollution parameters
- Identifying the main sources of water pollution in Lebanon
- Modeling the behavior of pollution parameters in the Qaraoun Lake
- Modeling extreme levels of air pollution in the Bekaa region
- Calculating the air quality index (AQI) in the Bekaa region and predicting future pollution levels.

The first chapter is a review of bibliographic synthesis which highlights the various environmental problems related to climate change and atmospheric and aquatic pollution in the Mediterranean basin and particularly in Lebanon. This chapter presents the various methodological approaches presented in the literature to evaluate the quality and the pollution of the atmosphere on one hand, and the quality and the pollution of the aquatic resources on the other hand. Then, we discuss the situation of the Litani

River in Lebanon as well as that of Lake Quaroun and the various factors that influence water quality in different monitoring stations. In the rest of this chapter, we will describe the different physicochemical parameters necessary to evaluate pollution. Finally, we provide an overview of air pollution and its consequences on ecosystems. A synthesis of the known results as well as a description of the study area and the examined air pollutants are presented.

In chapter 2, we present a new environmental classification method Stochastic block model "SBM" which aims to classify the physicochemical parameters in three stations of the Litani River. The analysis was carried out based on the annual averages of 11 parameters obtained during the sampling at the three stations Jeb-Jennine, Ghzayel, and Qaraoun lake from 2008 to 2018.

The SBM method differs from other methods based on a generative approach. The SBM method does not only consist in performing partitions but rather provides information on the relationships between the various groups obtained. This method allows to group and identifies the evolution of physicochemical parameters between and within stations. To do this, we calculated the connection matrix between the clusters obtained and the probability that the physicochemical parameters belong to each of the clusters in the different stations. For each of the three stations, the same clusters were obtained, the difference being in the connection matrices and the probability vectors of their belonging to the clusters. The performance of the presented SBM method is proven by simulations and by comparison with classical methods. We first compare the SBM method with the principal component analysis (PCA) method, then with the hierarchical clustering method and the K-means method. The results indicate that the classical methods provide the same two clusters as the proposed method. However, in contrast to our SBM method, the classical approaches do not reveal the block structure of the three stations.

In chapter 3, we propose a new approach for the assessment of the water quality based on chaos theory to investigate the nonlinear time series behavior of physicochemical parameters in Qaraoun Lake. Chaos theory is an approach that emphasizes elements of interactivity and unpredictability instead of using a linear approximation model that focuses on causality. We study the physicochemical parameters for 11 years (2008-2018), where we introduce the integration dimension and the lag time for each parameter. Then we focus on the correct phase space construction for the non-linear physicochemical parameters. We introduce the correlation dimension method to detect the existence of chaos. The estimated parameter dimensions indicate the possibility of chaotic behavior in the physicochemical time series. The occurrence of the chaotic behavior would provide a favorable result for accurate prediction of physiochemical parameter variability due to seasonal variations, climate change, human and industrial factors. In addition, different variations in environmental factors at the Litani River variations make the physicochemical parameters non-linear, which complicates the understanding of Qaraoun Lake state.

Chapter 4 describes the extreme values and observed records of different air pollutants in the Bekaa region of Lebanon. Air pollution is a serious problem that not only threatens public health, but also affects economic growth, agriculture, and the ecosystem. Due to a large number of personal vehicles and the increasing consumption of fossil fuels, the air quality in Lebanon is decreasing. Extreme pollutants exceeding a certain threshold can result in a very high annual loss of life and damage to infrastructure and agriculture. It is therefore important to know the occurrences and probabilities of such extreme events. In the last decades, the modeling of extreme events has been of particular interest in various fields related to earth and environmental sciences, such as hydrology, seismology, and applications related to climate change or air pollution monitoring. We use the two classical approaches Block model (BM) and Peak Over Threshold (POT) to study the behavior of extreme pollutants in Zahle city in the Bekaa region of Lebanon. Based on these two approaches we predict the return level of pollutant concentrations over several years (15, 30, 40 and 50 years) to minimize their harmful impacts and provide information to policymakers, government agencies, and civil society. Second, we use a new approach in the theory of extremes to predict future records based on the theory of records where we calculate the probability of the return of a new record after a specific period.

In chapter 5 we are interested in the study of the air quality in the Bekaa region of Lebanon. The choice of the Memchiyeh station in Zahle comes down to the fact that this region has several characteristics, either at the level of tourism, agriculture, or its geographical location, that affect directly its air quality. Few studies have been done to investigate the air quality in this region. This chapter deals with air pollution to predict air quality by calculating the air quality index. The data was collected from the Ministry of Environment dated from June 2017 to June 2018. We determine a relationship between changes in ambient particulate matter concentrations over a short period. In addition, we compare several air quality indexes (AQI) forecasting models, such as Naive, Exponential Smoothing, TBATS (a forecasting method for modeling time series data), and Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Then to study the performance of these models to obtain a better prediction of the air quality, we calculate several indicators such as; the mean absolute error (MAE), the root mean square error (RMSE), and the relative error (RE). We obtain that the SARIMA model is the most reliable model for the prediction of the AQI. In this chapter, we conclude that the SARIMA model can be applied to other cities to evaluate the level of air pollution to warn the authorities responsible for air quality control to measure the level of air pollution shortly and establish a mechanism to detect the air pollution peaks.

Finally, a general conclusion sums up all the findings obtained in this work. It shows the utility of the applied methods and explains the originality of the conducted studies where some concepts were introduced for the first time to the field of environmental studies and pollution monitoring. Some

possibilities for future work are explained at the end of the manuscript where we suggest comparing different models for more reliable predictions in the future.

Chapter 1

Literature review

1.1 Natural resources

Natural resources are materials that nature offers to living organisms and that human beings can use directly or with few modifications. They are central to humanity's well-being and their life's continuity. They can simply be breathable air, water, or plants. Or more complex like petroleum, mineral ore, and coal. At a fundamental level, every man-made product is made of natural resources. Therefore, besides being essential for humans' existence, natural resources make the principal component in every single human activity. It is obvious that with a growing population, the demand for natural resources is increasing, but the ecosystem's ability to sustain these resources and continuously provide enough for the global population isn't certain. Many resources used by humans are classified as non-renewable resources, which means they can be depleted with time and with uncontrolled human consumption. Other resources are renewable, but even with their abundance, their quality is important and must be monitored.

1.2 Types of natural resources

Natural resources are mainly classified into two categories: renewable and non-renewable resources. A renewable resource is a resource that can be regenerated relatively quickly. This means that the regeneration of such resources can be faster than their consumption by humans. Trees and plants are good examples of renewable resources.

Non-renewable resources, on the opposite, take much more time to regenerate, and we can reach a point where they are depleted and no more available. Oil, gas, and minerals are all natural resources that take so much time to generate, and humans depend on them widely in all types of activities. These resources are used in construction, all forms of industry, transportation, and energy generation. Air and water, being the most used in daily life and the most important of all resources, were always considered renewable. Unfortunately, this is not very accurate since many regions across the globe are suffering from a shortage of water, especially during the hot and dry seasons. Air quality is also degrading worldwide. WHO states that 9 out of 10 people in the world's population are exposed to polluted air, which causes around 4.2 million premature deaths yearly (WHO, 2021).

1.3 Impact of population growth on natural resources

Population growth is the increase in the number of living people on a specific scale, it could be observed over a country, a region, or globally. Population growth is affected by many factors including medical

advances, education rates, and women's employment rates. The world human population has been growing since the XIVth century and is expected to keep increasing. Today, the global population is nearly 8 billion people. However, the United Nations estimations state that the global population might reach 9.8 billion by 2030, and 11.9 billion by 2100 (United Nations, 2017).

With population growth, comes more demand for natural resources. Thus, the increase in the extraction of natural resources from the environment, which will eventually lead to the depletion of non-renewable resources. Also, the extraction process itself often releases harmful pollutants and produces wastes that affect air and water quality. The increase in freshwater use for drinking and irrigation is also caused by population growth, putting more pressure on the available water resources. The most obvious impact of the increase in population is the disturbance of forests, green spaces, and other habitats to construct or expand urban areas. Again, this is affecting more than one natural resource since the building is a highly polluting process that releases CO_2 into the atmosphere, and the absence of trees is depriving the area of natural carbon capture and air purification, thus reducing air quality significantly.

1.4 Impact of climate change on natural resources

Climate change is a phenomenon caused mainly by human activities, especially burning fossil fuels, coal, and gases, which releases temperature-trapping gases known as greenhouse gases. This phenomenon refers to long-term shifts in global temperature and significant changes in weather patterns in different regions of the world.

Figure 1 Impacts of climate change[1]

This change in the climate of Earth is altering the ecosystem and threatening biodiversity. The figure 1 shows the different impacts of climate change represented in heat waves, wildfires, and the accelerated melting of glaciers. Firstly, climate change is represented by a significant rise in the global temperature, which is contributing to faster melting of glaciers and evaporation of water resulting in more precipitation and disturbance in groundwater recharge (Kushawaha, et al., 2021). Secondly, drought

[1] ***https://www.noaa.gov/education/resource-collections/climate/climate-change-impacts***

and heat waves are becoming more frequent (Steffen, et al., 2014) and leading to the degradation of the quality of the available water (Blagrave, et al., 2022). Also, extreme temperature events are resulting in the spread of wildfires (Vitolo, et al., 2019), ending massive spaces of forests and affecting their wildlife, and causing more atmospheric pollution and more degradation in air quality (Oregon, 2021).

1.5 Water Resources

1.5.1 Importance of water as a natural resource

Water is one of the most important natural resources (Aulenbach, 1968). It sustains life on Earth and covers 71% of its surface (Water Facts, 2020). However, only less than 3% of the available water is fresh water, and the remaining amount (97%) exists in the oceans and is considered too salty for human consumption, irrigation, and most industrial uses (Water Facts, 2020). Figure 2 shows the detailed distribution of Earth's water.

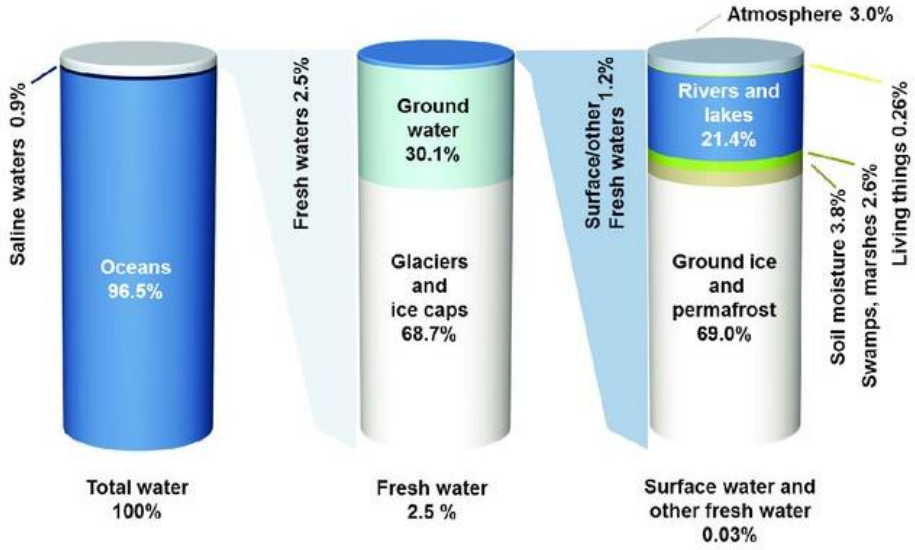


Figure 2 Distribution of water on Earth[2]

In recent years, the availability of water has become the main concern for researchers, governments, and organizations who focus on the abundance and quality of resources. Comparing the average precipitation with the use and consumption per person in diverse human activities in different regions of the world, findings were that the amount of water is droppingin many regions across the globe (Pimentel, et al., 2004). Other studies pointed at the quality of the available water and compared the concentrations of pollutants in water with the WHO guidelines, the results showed that around four billion people around the world, including some developed countries, especially in South Asia, do not have access to drinkable water (Biswas, et al., 2019).

[2] ***https://www.worldbank.org/en/topic/water***

1.5.2 Definition of water pollution

Water pollution is defined as the change in the natural composition of water resources (Kılıç, 2020), where the physical, biological, and chemical properties of water are modified (Hayek, 2021). These modifications are generally induced by human activities such as industrialization and agricultural practices (Owa, 2013).

The WHO sets clear standards for water characteristics, and any exceed of these limits makes water unsuitable for domestic and industrial uses. With more demographic development and more use of chemicals, more contaminants are being introduced to waterbodies making the available water highly polluted and less suitable for the majority of desired uses.

1.5.3 Sources of water pollution

Human activities are without a doubt the main cause of water pollution. Sewage, pesticides and fertilizers used in agriculture, and industrial wastes are directly impacting the quality of stream water. Water subjected to these forms of pollution becomes contaminated with extremely harmful chemical and microbial substances which makes it toxic for humans and all living species within the contaminated water body.

Urbanization

Because of population growth, more land is being occupied by human residential areas in a process called urbanization. This process includes paving new roads, constructing more buildings, and creating new water supply systems. All of this means more wastewater discharge, which, with poor planning as in Lebanon for instance, will invade surface and groundwater and lead to the deterioration of water quality. A very recent study showed a high correlation between population growth and water quality, it was proven that urbanization has affected severely the water resources and subjected the investigated watershed of Alto Atoyac to critical transformations (Estrada-Rivera, et al., 2022).

Sewage

Figure 3 Raw sewage discharge in the rivers[3]

[3] ***https://www.netsolwater.com***

Sewage is the wastewater coming from household, industrial and agricultural use (Figure 3). It contains different substances that contribute to water pollution. Firstly, nutrients like phosphorus and nitrogen which are present in sewage can stimulate the growth of algae and phytoplankton when discharged into rivers and seas. Also, sewage may contain harmful bacteria such as salmonella and E. coli, which are highly poisoning and may even lead to death when consumed by humans. Other substances present in sewage are trace metals, mercury for example, which can have toxic effects on human health and the environment when it enters the food chain.

Agriculture

Pollution from agriculture is a global problem linked to the increasing uses of pesticides and fertilizers since the 50s (Figure 4), that has resulted in higher loads of contaminants in the environment including rivers, aquifers or lakes (Evans, et al., 2019). The use of pesticides and fertilizers has increased to optimize food production and meet the needs of the growing population. However, (Kusum, 2021) explains the serious effects of pesticides and fertilizers on human health. In 2002, pesticide residues were found in samples from bottled drinking water in Delhi, India (Agrawal, et al., 2010), which means that humans are already subjected directly to the hazardous effects of pesticides contaminated water.

Figure 4 Water pollution from agriculture[4]

On another hand, when these chemicals are introduced to a water body, they promote algal growth and cause oxygen depletion. Further, they can poison fish and destroy marine habitats (EPA, 2005). Agriculture is also responsible for a large proportion of surface-water pollution and is responsible almost exclusively for groundwater pollution by nitrogen.

Industrial waste

The industrial waste includes chemical waste, solid waste, and toxic and hazardous waste. When dumped untreated into a water body such as a river or a lake, they can enter the food chain causing a direct threat to local wildlife. They may also affect water acidity, which will eventually harm living species within. Also, many of the hazardous substances present in the industrial waste cannot be easily

[4] ***A brief overview on current environmental issues in Iran***

biodegraded, therefore, they accumulate in water sediments causing illness to fish and other creatures that may lead eventually to their death or extinction. Even drinking water is being contaminated with industrial waste (Jebin, et al., 2021).

1.5.4 The physicochemical parameters

The physicochemical parameters are the factors or elements that represent the overall quality of water. To determine the proper use of the available water, we need to measure these parameters and compare them with international guidelines. These guidelines vary for the river depending on the usage of its water, irrigation, or production of potable water. Besides, they can vary from one country to another while being as close as possible on WHO recommendations. In this study, we will be handling 11 physicochemical parameters since these parameters are those included in the surface water monitoring programs in Lebanon.

Temperature (°C)

The temperature of the water is affected by the ambient temperature. It is an important factor that influences several other parameters and can alter the physical and chemical properties of water. The temperature of the water has a direct effect on aquatic organisms and their metabolic activity. Also, the solubility of some chemicals in water increases with temperature, with the risk to reach or to enhance their toxicity levels. The temperature must be measured both *in situ* and in the laboratory at the sample arriving.

pH

pH affects most chemical and biological processes in water. Indeed, water pH controls the solubility and biological availability of some chemicals such as nutrients (carbon, phosphorus, and nitrogen) and trace metals (copper, chromium, cadmium, etc.). It is one of the most important environmental factors limiting species distributions in aquatic habitats. EPA water quality criteria for pH in freshwater suggest a range of 6.5 to 9 because the majority of aquatic species prefer this pH range. The pH must be measured on the field using a pH meter.

TDS: Total Dissolved Solids (mg/L)

It represents the concentration of the dissolved solids and minerals in the water, which originated from natural sources and/or from human activities. Some dissolved solids in water are not harmful, and they may improve the taste and function of water, but an elevated level of TDS affects negatively the quality of water. The standard method for determining the TDS is to evaporate a known amount of a water sample by heating it to 180 ° C, and just weighing the obtained solid residue.

DO: Dissolved Oxygen (mgO_2/L)

Dissolved oxygen refers to the level of oxygen present in water. It is an important parameter in assessing water quality because it influenced the organisms living within a water body. A dissolved oxygen level that is too low (less than 2 mgO_2/L) can harm aquatic life and affect water quality. DO levels depend on the temperature and the salinity of the water, the atmospheric pressure and the wind speed and the biodiversity within it.

Conductivity (µS/cm)

Conductivity is a measure of water's capability to pass electrical flow. This ability is directly related to the concentration of ions in the water which come from dissolved salts and inorganic materials. As the US EPA reminds, "*Conductivity is useful as a general measure of water quality. Each water body tends to have a relatively constant range of conductivity that, once established, can be used as a baseline for comparison with regular conductivity measurements. Significant changes in conductivity could then be an indicator that a discharge or some other source of pollution has entered the aquatic resource*".

Salinity (mg/L)

Salinity is the concentration of dissolved salt in water, which comes from rainfalls or the erosion of rocks. Salinity directly affects the density of water, the higher the salinity, the denser the water. Also, it's directly linked to the Earth's overall water cycle. It is important to measure the amount of freshwater entering and leaving the oceans every year to probe the water cycle.

Ammonia: NH_3 (mg/L)

Ammonia (NH_3) is a common toxicant derived from wastes, fertilizers and natural processes. Ammonia nitrogen includes both the ionized form (ammonium, NH_4^+) and the unionized form (ammonia, NH_3). Environmental factors, such as pH and temperature, can affect ammonia toxicity to aquatic animals. As a matter of fact, ammonia concentration and toxicity increases as pH increases. At pH higher than pKa (*i.e.*, 9.2), ammonia (NH_3), the more toxic unionized form is predominant. However, when ammonia levels are high enough, toxicants within the aquatic organisms build up in the inner tissues leading to death.

Nitrate and nitrite, NO_2^- and NO_3^- (mg/L)

Nitrate (NO_3^-) and nitrite (NO_2^-) are naturally occurring ions that are part of the nitrogen cycle. High nitrate concentrations are a result of agricultural fertilizers, effluents domestic, and industrial discharges, and wastewater. In addition to the risk of cancer, significant exposure to nitrates and nitrites could cause a rare disease in the blood called methemoglobinemia. This disease destroys red blood cells. The most toxic species is nitrite. Nitrate and nitrite are nutrients with indirect adverse effects on

aquatic communities; they affect the primary production and the growth and accumulation of biomass in the water column.

Orthophosphate, PO_4^{3-} (mg/L)

Orthophosphate ions (PO_4^{3-}) come from a naturally occurring element named phosphorus which is one of the key elements necessary for the growth of plants and animals in lake ecosystems. Phosphorus will stimulate the growth of plankton which represents the base of the food chain. Thus, phosphorus is an essential element that contributes to the environmental cycle. Phosphorus is not harmful to humans unless it is present at very high levels, in that case, digestive problems could occur.

In the table below, we summarize the physicochemical parameters and their measurement methods:

Parameters	Measurement method
pH	pH meter, La Motte
Conductivity/TDS/Salinity	Conductivity/TDS/Sal Portable meter, La Motte
Dissolved Oxygen (% or mgO_2/L)	Dissolved Oxygen Gauge, La Motte
Phosphate as PO_4^{3-} (0 to 5 mg/L)	Spectrophotometry. Method H3035-QNT, Phosphorous, reactive, PhosVer 3, Test N tube procedure. Hach 8048 Cat number 2106069
Nitrate as $N\text{-}NO_3^-$ (0 to 0.5 mgN/L)	Spectrophotometry. Method H2520. QNT Nitrate, Mid-range, Cadmium reduction method, powder pillows. Hach method 8171 Nitra Ver 5.
Nitrite as $N\text{-}NO_2^-$ (0 to 0.3 mgN/L)	Spectrophotometry. Method H2610. FXD Nitrite, low range, Diazotisation method, powder pillows. Hach 8507. NitriVer 3.

Table 1 Physicochemical parameters and their measurement methods

1.6 Air resources

Air is a critical resource for life's sustainability. It is essential for breathing, photosynthesis, maintaining the earth's temperature, and can be used as an energy supplier. Air is normally considered a renewable resource because it can be naturally restored faster than its consumption. However, the abundance of air doesn't necessarily mean the existing air is breathable or has the ultimate efficiency to maintain life.

Air quality is measured by the Air Quality Index (AQI) developed by the U.S. Environmental Protection Agency (EPA). A yardstick that ranges from 0 to 500. The higher the AQI, the worse is air quality. This means that higher levels of AQI indicate poor air quality and requires people to be cautious in their outings and to try to avoid prolonged times outdoors. Poor air quality affects primarily children and people with respiratory diseases such as asthma. Moderate air quality can cause breathing discomfort for initially ill people. Poorer air quality may escalate irritation and may even cause respiratory issues for healthy people if pollution levels are very high.

In recent decades, air pollution has become a serious concern and a leading cause of excess mortality (Lelieveld, et al., 2020). A 2013 assessment by WHO's International Agency for Research on Cancer

(IARC) concluded that outdoor air pollution is carcinogenic to humans, with the particulate matter component of air pollution most closely associated with increased cancer incidence, especially lung cancer. The figure below represents the estimated excess deaths due to ambient air pollution per year in Europe with the contributing disease category.

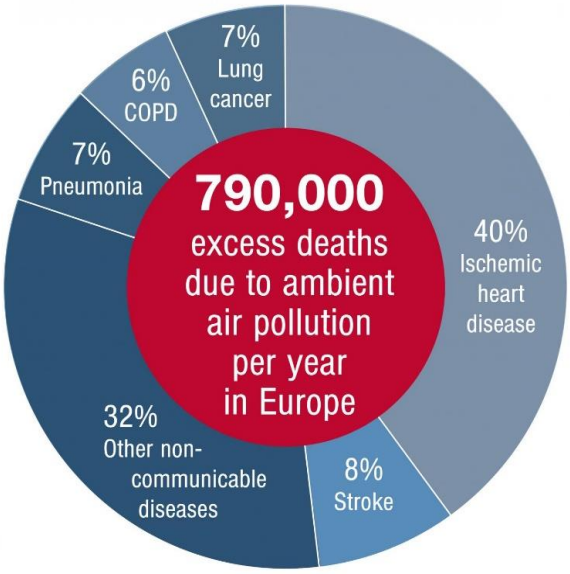


Figure 5 Estimated excess mortality attributed to air pollution in Europe, and the contributing disease category

Air pollution has harmful effects on human health and the environment as a whole. It can cause respiratory and skin diseases, and contribute to poor visibility. Also, greenhouse gases spread in the atmosphere like methane and carbon dioxide are the leading cause of human-made global warming because they trap the heat in the atmosphere. Another consequence of atmospheric pollution is acid rain (Kumar, 2017); when compounds like sulfur dioxide and nitrogen oxides are released into the atmosphere, they undergo a chemical reaction with oxygen and water to form acidic rain, which is another environmental issue of great concern because of its serious large-scale effects on the ecosystems and its transboundary nature (Grennfelt, et al., 2019).

1.6.1 Sources of air pollution

Air pollution is a global problem that began with the industrialization era (Sivaramanan, 2014). It is characterized by the introduction of chemicals, particulates, or biological materials into the atmosphere, causing serious health and environmental issues to humans, living organisms, and the ecosystem (Choudhary, et al., 2013). Air pollution can be caused by natural phenomena, such as volcanic eruptions, dust storms, and wildfires. However, the main reason for air pollution is human activities, especially burning fossil fuels for transportation, energy generation, and industry.

Sources of air pollution are divided into four categories: mobile, stationary, area, and natural sources.

- Mobile sources are the global transportation systems, including all vehicles, ships, and airplanes. Together, they account for almost 25% of the world's energy-related CO_2 emissions, and they emit pollutants while moving from one place to another.
- Stationary sources are sources with a fixed location like power plants, oil refineries, industrial facilities, and factories. They can have single or multiple pollutants-emitting exhausts, and even though this pollution source is fixed, pollution from these facilities can go a long way in the atmosphere.
- Area sources are generally agricultural areas. Livestock and burning agricultural waste are highly polluting. Together, they produce methane, ammonia, and black carbon. It should be noted that methane is a greenhouse gas that contributes to more global warming than carbon dioxide, also it contributes to ground-level ozone, which causes respiratory illnesses such as asthma.
- Natural sources include natural phenomena, smoke from wildfires, and dust from sandstorms which can also carry a lot of pathogens and travel for a long distance, spreading pollution. Although these phenomena can happen naturally, but man-made climate change is making such phenomena more frequent and more severe.

1.6.2 Air pollution parameters

[5]Generally, air quality is defined by 8 main parameters, which are: particulate matter (PM_{10}, $PM_{2.5}$), Ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), lead (Pb) and ammonia (NH_3). These parameters are divided into two categories: primary pollutants and secondary pollutants. Primary pollutants are pollutants emitted directly from a specific source, and they are harmful in the form in which they are produced. Whereas secondary pollutants are produced when primary pollutants react together to form a new toxic compound that was not emitted from a source such as a vehicle or a power plant.

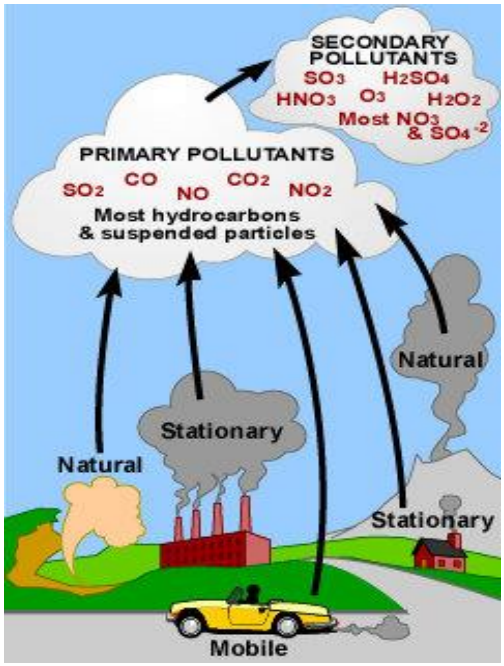


Figure 6 Primary and secondary pollutants

Air pollutants have a defined maximum average concentration over a specific period. When the concentration of these parameters exceeds the limits, air quality is said to be poor and exposure to outdoor air should be limited or avoided depending on the degree of pollution.

[5] ***GUIDELINE Air Quality Assessment Based on Road Traffic Pollutants Dispersion Modeling***

Particulate matter (PM_{10}, $PM_{2.5}$) ($\mu g/m^3$)

Particulate matter (PM), also known as particle pollution, are a mixture of solid particles such as dust, pollen, and liquid droplets suspended in the air. There are two types of particulate matter: PM_{10}, and $PM_{2.5}$. PM_{10} is a particulate matter of a diameter of 10 micro or less, some of them are dark or large enough to be seen with the naked eye. PM_{10} is the heavier type of particulate matter, they can stay in the air for minutes or hours, whereas the smaller particles of 2.5 micro diameters or smaller ($PM_{2.5}$) can stay in the air for days or even weeks. Both types of PM can be produced by natural phenomena or human activities. Together, they affect both the environment and human health and are considered the deadliest form of air pollution. They can penetrate deeply into the lungs causing respiratory illnesses. PM can also cause climate change and affect weather patterns.

Ozone (O_3) ($\mu g/m^3$)

Ozone, also known as trioxygen, is a molecule formed of three Oxygen atoms. When nitrogen oxides (NO_X) and volatile organic compounds (VOCs) are exposed to sunlight, they react and produce this colorless highly reactive gas (Figure 7). When we study air pollution, we refer to O_3 as the ground-level ozone or the ozone present in the troposphere.

Figure 7 Formation of ground-level Ozone[6]

Although ozone in the stratosphere is beneficial since it protects the earth from ultraviolet radiation, ground-level ozone is very harmful to human health because it is present where humans breathe. Ozone has a damaging effect on the lungs. Even when low amounts of ozone are inhaled, they can cause many symptoms of respiratory problems such as chest pain, coughing, shortness of breath, and, throat irritation.

[6] https://raqc.org/localgov/what-causes-air-pollution/

Sulfur dioxide (SO_2) ($\mu g/m^3$)

Sulfur dioxide is a toxic gas with a choking smell that is emitted from the combustion of sulfur-bearing fossil fuels like natural gas, coal, and crude oil. Although sulfur dioxide is not considered a GHG, it is a major pollutant that can cause serious damage to the environment and human health. This gas can easily combine with other atmospheric molecules to form Sulphur-containing particles leading to the formation of haze, smog, and acid rain. All of these environmental hazards contribute to severe respiratory complications.

Nitrogen dioxides (NO_x) ($\mu g/m^3$)

Nitrogen oxide and nitrogen dioxide belong to a family of poisonous and highly reactive gases called nitrogen oxides. They are produced when nitrogen-containing fuels are burnt in the atmosphere. Hence, the main source of nitrogen oxides is road traffic, followed by energy production and distribution. These gases, like all forms of air pollution, are harmful to both human health and the environment; at elevated levels, they increase the severity of respiratory infections and asthma. They also contribute to the formation of smog and acid rain.

Carbon monoxide (CO) (mg/m^3)

Carbon monoxide is an odorless flammable gas; this toxic gas is a by-product of the incomplete combustion of fossil fuels. A small concentration of carbon monoxide can cause poisoning that is characterized mainly by headache, dizziness, and breathing difficulties. Carbon Monoxide poisoning is extremely fatal and may lead to death even before its diagnosis.

The table below sums the WHO guidelines for different the studied air pollutants for a specific duration of exposure.

Pollutant	Duration of exposure	WHO guidelines
$PM_{2.5}$	1 year	$35\mu g/m^3$
PM_{10}	1 year	$20\mu g/m^3$
NO_2	1 year	$40\mu g/m^3$
O_3	8 hours	$100\mu g/m^3$
SO_2	24 hours	$20\mu g/m^3$
CO	8 hours	$10mg/m^3$

Table 2 WHO guidelines for air pollutants

1.6.3 Air Quality Index (AQI)

Air quality index is a tool used to refer to the overall daily air quality for a specific area. It is a simplified way to communicate to the public the air quality in the area in which they are present. Air quality index is a dimensionless number that represents, according to the principal air pollutants (Kowalska, et al.,

2009), the degree of air pollution. A high air quality index means more pollution. For every range of AQI, specific recommendations are given for different categories of people (EPA, 2014). Air quality ranges from 0 to 500, divided into sections from good to hazardous. It is communicated to the public mainly by the color of the range, where each range includes a health message to be followed by the citizens (Table 3) (USEPA, 2017).

AQI Value	Health Message	AQI Color
0 - 50	None	Green
51 - 100	Unusually sensitive people should reduce prolonged or heavy exertion	Yellow
101 - 150	Sensitive groups should reduce prolonged or heavy exertion	Orange
151 - 200	Sensitive groups should avoid prolonged or heavy exertion; general public should reduce prolonged or heavy exertion	Red
201 - 300	Sensitive groups should avoid all physical activity outdoors; general public should avoid prolonged or heavy exertion	Purple
301 - 500	Everyone should avoid all physical activity outdoors	Maroon

Table 3 AQI ranges with the associated health messages

1.7 Study Area: General representation of Lebanon

The Lebanese Republic is a small country with a total area of 10,452 km^2, located on the shores of the western Mediterranean Sea, over a distance of up to 220 km. It is mainly composed of a coastal plain and two mountain ranges (Mount Lebanon and Anti-Lebanon) which form a wide valley, that of the Bekaa, located at an altitude of 950 m (Figure 8).

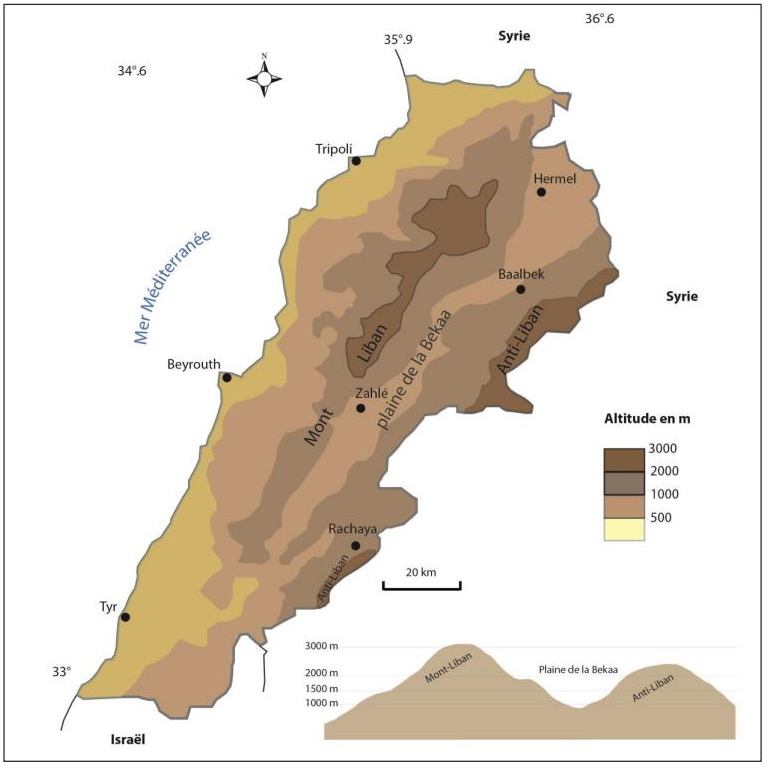


Figure 8 Map of Lebanon: Geolocation of the two mountain ranges

The presence of the two mountain ranges causes climatic variation within the country (Abou El-Enin, 1980), (Fadlallah, 2001). Most of the Lebanese territory is influenced by the Mediterranean climate, which is characterized by intense rainfall during the winter months and very dry summers. We generally recognize 4 varieties of classic Mediterranean climates: 1) Maritime-type climate on the coast, accompanied by high rainfall (830 mm/year on average) during 4 to 5 winter months. The average temperature varies between 13.3°C (January) and 26.7°C (August). 2) At altitudes above 1000 m, the climate is rather moderate, with frequent rains (2 to 4 months). The average annual temperature is around 10°C. 3) A tropical-type climate on the western part of the coast, especially on the central and lower parts of the Bekaa Valley, characterized by rains distributed over 4 winter months (650 mm). The average temperature varies between 5.8°C in January and 24°C in August. 4) A desert climate north of the Bekaa plain due to its opening onto the Syrian desert. This area is characterized by the great dryness of its atmosphere, favored by the weakness and the rarity of the rains which do not exceed 250 mm. The average temperature is between 7°C (January) and 31°C (August).

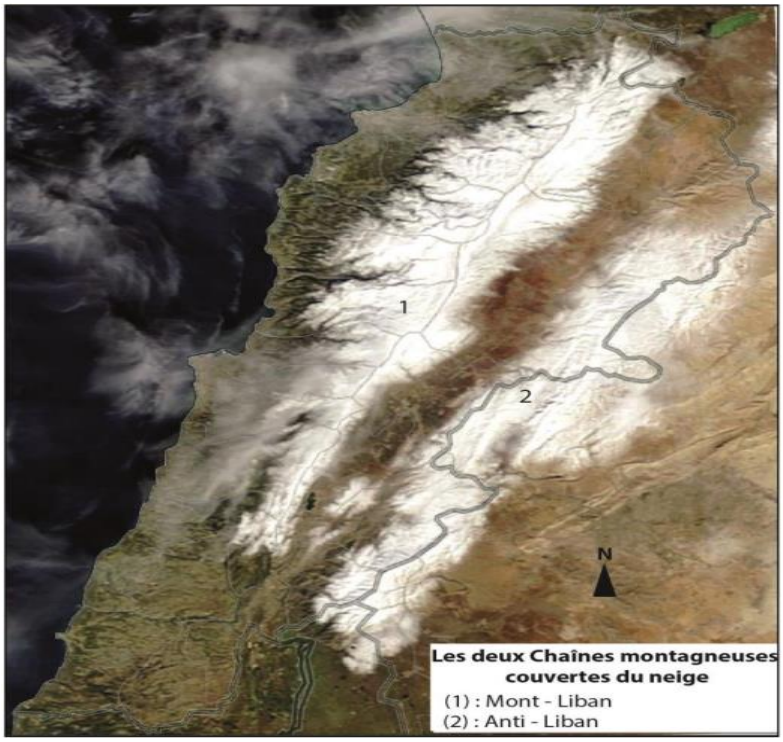


Figure 9 Mount Lebanon and The Anti-Lebanon, water reservoirs. Source MODIS, 2016. (http://modis.gsfc.nasa.gov/)

Despite the precipitation rate for each region, the duration of the dry season is a real challenge for Lebanon. However, the mountains located at high altitudes play a real role in reserves allowing an accumulation of precipitation (rain and snow) and good integration of these to the groundwater that serves to meet the water needs of farmers during the dry period, especially in the Bekaa Valley. Figure 9 shows the two mountain ranges that are covered by snow during winter, which allows the passage of water into the rivers following the melting of snow.

1.8 Water status in Lebanon

Lebanon is a developing middle eastern country located in western Asia. With a unique topography and numerous rivers, Lebanon is considered rich in water and known for its moderate climate. However, several studies showed that the quality of water in Lebanon is degrading, and the country is affected by climate change and a decrease in precipitation (Khraibani, et al., 2020). Lebanon is suffering from many environmental problems. On one hand, the increasing population is contributing to more demand for water, on the other hand, the pollution of rivers caused by human activities is depriving the country of important water resources. Add on, the country is witnessing many changes in its climate. Increased temperature, decreased precipitation, and shifts between the wet and dry seasons are all factors affecting the quantity and quality of water. One of the most important water resources in Lebanon is the Litani River. It is the only river that rises and ends within the Lebanese borders. The Litani River rises from the Bekaa Valley in the northwest of Lebanon and empties into the Mediterranean Sea in the east south of Lebanon north of Tyre. The estimated length of the river is 170

km, making it the longest river in Lebanon. Water from the river was a major source of water supply, irrigation, and hydroelectricity.

1.8.1 Litani River

The Litani River is the longest and largest Lebanese river accounting for about 30% of the total surface water flow of all rivers flowing in the Lebanese territories. Rising in the Bekaa Valley, west of Baalbek, and emptying into the Mediterranean Sea north of Tyre, the Litani River is considered one of the most important water resources in Lebanon as it provides water for irrigation and hydroelectricity for the whole country, especially within the Bekaa Valley where agriculture is commonly practiced. The total drainage area of the basin is about 2175 km² with a length of channels of about 170 km and an annual flow of 750 million m³. The Bekaa plains, where the Upper Litani flows, are located in the anti-Lebanon mountains. The total drainage area of the basin is about 2175 km² with a length of channels of about 170 km and an annual flow of 750 million m³.

The Litani River is separated into two parts: the upper Litani basin and the lower Litani basin. These basins are separated by the Qaraoun lake, initially created to generate hydropower, domestic water supply, and irrigation.

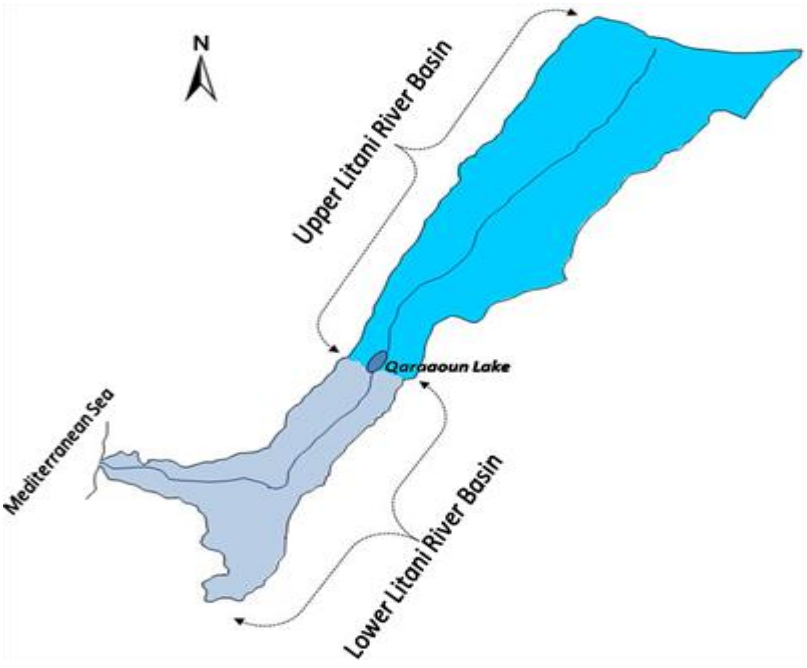


Figure 10 The upper and lower parts of Litani River separated by the Qaraoun lake

Both basins and the Qaraoun Lake are facing a huge environmental crisis since pollution within the river has reached high levels due to many factors that will be explained further. The effect of the Litani River pollution on The Lebanese people life cannot be ignored since water from the river is used to irrigate the plants that we consume daily. Making the majority of the Lebanese population directly subjected to cancerous materials.

1.8.2 Qaraoun Lake

Situated in the west of Qaraoun village in the southern region of the Bekaa Valley, Qaraoun Lake is the largest lake in Lebanon. It's an artificial lake of an 11.9 km² area created in 1959 initially to harvest water from the Litani River for irrigation and to generate electricity. The annual surface water flow received at Lake Qaraoun from the Litani River is 420×106 m^3 (15×109 cu ft). this flow is enough to generate hydroelectric power of 600 GWh at three different power stations located in nearby villages. Unfortunately, the lake is facing several environmental problems limiting its uses. First, the level of water in the lake drops significantly during the dry season (Fadel, et al., 2014). Also, the pollution in the lake is increasing and reaching dangerous levels. In fact, the lake has witnessed two massive fish death in recent years where tons of dead fish floated to the surface (Figure 11). The first one happened in July 2016, and the most recent, which is shown Figure 11, happened at the end of April 2021. This phenomenon was directly associated with elevated levels of pollution, and in 2018, fishing in Qaraoun Lake was banned permanently.

Figure 11 Dead fish are seen floating in Lake Qaraoun on the Litani River, Lebanon, April 29, 2021

In 2005, the Litani River Authority created the Environmental Unit and tasked it with the continuous monitoring of surface water quality. Many studies were carried out for the assessment of pollution in the Litani River. In the first few years, most of the studied parameters were within permissible standards set for human consumption, irrigation, and river quality indicators (Saade, et al., 2012). Thereafter, pollution evolved significantly and water quality deteriorated (Haydar, et al., 2014)

1.8.3 Sources of water pollution in the studied area

The sources of pollution in the Litani River and Qaraoun Lake are numerous, and the level of pollution varies from one zone of the river to another. The most common sources of pollution are sewage, pesticides and fertilizers used in agriculture, and industrial waste. Wastewater from the towns and the villages on the banks of the rivers is discharged directly into the river without any treatment. The

situation is aggravating with the increasing number of Syrian refugees who settle in unplanned camps around the river, increasing the volume of wastewater and different types of waste discharged in the river and the Qaraoun Lake.

Figure 12 Syrian refugees camps around the Litani River[7]

The Bekaa region is known to be an agricultural area. The lands surrounding the river from the Bekaa Valley towards Qaraoun Lake are mostly cultivated areas where the use of pesticides and fertilizers is very common and uncontrolled. These products contain harmful chemicals that can easily reach the river. When the concentration of these chemicals increases, the quality of water is damaged and so is the ecosystem. Most fertilizers include nitrogen, phosphorus, and sulfur are usually monitored to track variations in water quality.

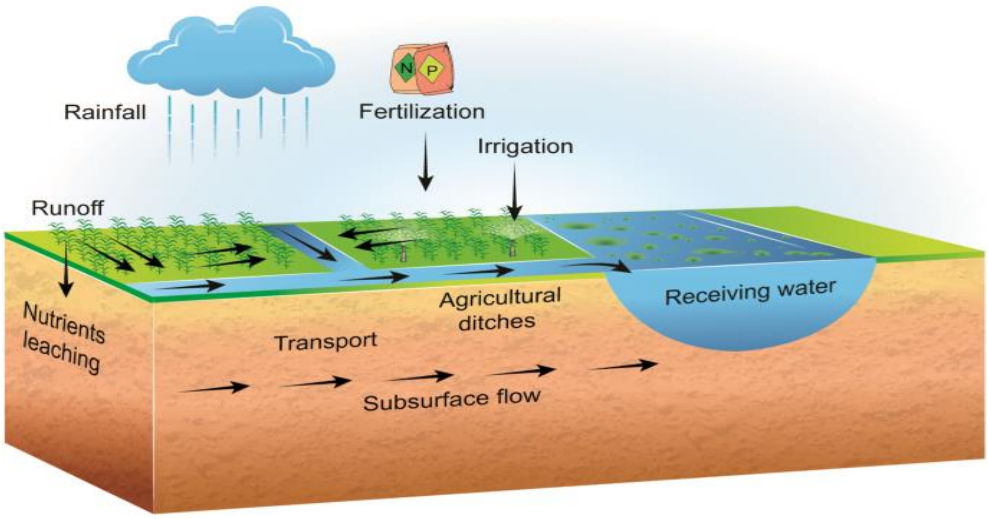


Figure 13 Nutrients leaching and transport to river water

These chemicals must stay within a specific range defined by the WHO guidelines to have water of a good quality that can be used for irrigation and domestic needs. However, the concentration of nitrogen,

[7] ***Litani River Authority***

phosphorus, and sulfate ions in the Litani River is generally high and often exceeds the safety guidelines, especially near agricultural areas.

227 factories and industrial plants are spread around the river. The industrial waste from these plants is discharged into the river causing more pollution. The industrial waste can be solid waste like dirt and gravel or chemical waste that is more toxic and hazardous. It is important to identify the source of pollution to make a good treatment plan. Besides, the quality of water used for irrigation can have a serious effect on the contamination of agricultural crop. (Mcheik, et al., 2018) state that people consuming fruits and vegetables from contaminated lands have higher chances of developing gastrointestinal diseases, they also explain that the increase in cancer cases in these regions is the result of a high level of water pollution.

1.9 Air Quality in Lebanon

Air quality monitoring in Lebanon began in 2013 as part of the Environmental Resources Monitoring in Lebanon (ERML) project supported by the United Nations Environment Program (UNEP) and of the United Nations Development Program (UNDP). The objective was to establish a real-time air quality monitoring network that aimed to monitor population exposure to industries, power plants, and road traffic, in addition to urban sources of pollution. For this purpose, the Lebanese Ministry of Environment (MoE) installed air quality monitoring stations (AQMS) in different locations in the main cities of Lebanon (Hadath, Beirut, Baalbeck, Zahle, Saida). These stations measure and analyze the principal air pollutants: CO, NOx, O_3, SO_2, PM_{10}, $PM_{2.5}$, and meteorological parameters (temperature, humidity, wind speed, wind direction, pressure, global radiation, and rain volume). Using data from these stations, many studies pointed at the air quality in Lebanon. (Abdallah, et al., 2016) made a comparison between different models used in air quality monitoring, The increasing population in Lebanon also menaced air quality. With the massive construction taking place over the green spaces, lack of public transport, uncontrolled industries, and the private generators located in residential areas, air quality in most Lebanese cities are non-satisfactory. It is confirmed that air pollution is contributing to permanent respiratory illnesses, the development of severe symptoms in asthmatic people, and leading to lung cancer and in many cases, death.

We were interested in our study to model air quality in the Bekaa region. Hence, we used data from the monitoring station located at Memchiyeh garden in Zahle city with coordinates 33°59'52.44"N 36°12'16.43"E.

1.9.1 Zahle district

For air quality modeling, we chose the Zahle district located between the two mountain ranges of Mount Lebanon and Anti-Lebanon. Zahle is the capital of the Bekaa governorate. It lies among the eastern foothills of Mount Sannine, and it includes 43 communes, 51 villages, and 4,575 farms.

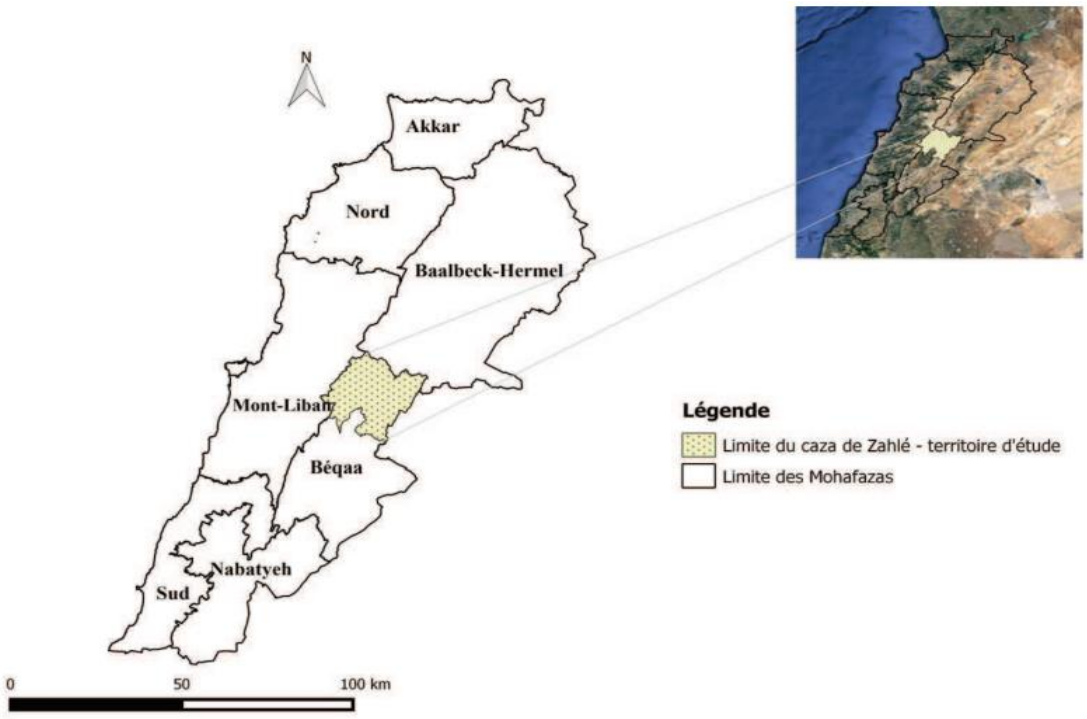


Figure 14 Location of Zahle on the map of Lebanon

The total area of Zahle is 22,153 ha (425 km$^{2)}$ which represents almost 4% of the total area of Lebanon and 10% of the area of the Bekaa (4,280 km^2). The exploited agricultural area is 20,844 ha of which 10,213.56 ha are irrigated. Zahle region is one of the 5 geographical regions belonging to the Bekaa governorate. It is located 40 kilometers from Beirut and lies between Mount Lebanon and the Syrian borders in south-east Lebanon. Zahle district comprises 43 communes, only 27 of which have a municipality. The others do not have a municipal council and are attached to the other municipalities or are directly under the management of the governorate for their common facilities.

1.9.2 Bekaa Valley

The Bekaa Valley has long been known for its agriculture. However, it is recently experiencing many environmental problems, either in terms of water or air pollution that affect more or less directly the soil and then the agricultural sector, which is one of the main activities of the inhabitants of this area. The demographic growth of the last decades, associated with the development of agricultural activities, contribute to an increase in the need for water resources. The majority of these resources, both underground and surface, are used for fertilization of agricultural land in the region during the vegetation period that begins in March. From a water management perspective, the region is an interesting source of data on the national context for several reasons:

- The meteorological factor: this climatic constraint, the semi-humid zone, makes it possible to highlight the problem of droughts and their evolution in a geographical area where the risks of lack of water resources are present in a systematic way.

- The presence of water resources: the presence of surface water is characterized by the Litani River which runs through the entire region. The surface area of the hydrological basin of this river covers about 20% of the total area of the Lebanese territory. In addition, this area has a large amount of groundwater and many artesian wells supplied by farmers.
- Agriculture: it makes a major and essential sector for the economy of the Bekaa Valley. Vegetable, fruit, and vineyard farms are found in the area.

1.10 Principle of modeling

Modeling is justified by the need to be able to describe and predict temporal phenomena. Indeed, the variations of the variables studied, called dependent variables or explained variables, are explained by the variation of several other variables, called explanatory variables or independent variables. The latter sometimes represents time itself (trends, seasonal variations) or temporal variations (meteorological variables). Modeling allows to describe of the development and the behavior of the physical quantity using formulas including a deterministic component and a random component. The deterministic part is the one that allows us to describe the average behavior of the phenomenon (the average behavior). The random part corresponds to the shift between the true magnitude of the observed variable and the deterministic value. In this thesis, we are interested in the statistical modeling of air pollutants on the one hand to detect the extreme pollutants using the theory of extreme values and the theory of records which provides the necessary information to limit the pollution peaks and on the other hand to predict the pollutants in the short term using the forecasting models which allow evolving the quality of the air and which help the authorities to supervise the quality of the air to which the population and the natural mediums are exposed.

In addition, numerical modeling of air pollution has undergone significant development in recent decades. Air quality modeling is based on numerical tools that allow the simulation of chemical and physical processes necessary for the development of air pollutant concentration. The models are fed by different forms and sources of data, in particular by the emissions of pollutants from several activities (road traffic, industrial and agricultural activities, domestic heating, air transport, etc.), as well as by climatic and meteorological conditions that have an impact on the diffusion of toxic substances and the course of chemical processes, and finally by extreme conditions that provide indications of pollution from external sources. Many models are widely used to predict short- and long-term pollutant concentrations in urban areas. Different approaches to modeling air pollutant concentrations are discussed in the literature, including deterministic, statistical, and mixed approaches. A deterministic model is a classical approach for the prediction of air pollutant concentrations. These models refer to physically based models that are constructed under assumptions of causality and are defined in terms of one or more mathematical functions (Taylor, 1915), (Scrase, 1930), (Sutton, 1932) and (Pasquill, et al., 1983).

Although the development of deterministic models has always been a priority to be able to describe the atmospheric dispersion of air pollutants, it is important to consider the statistical properties of air pollutants because of the difficulty of physical analyses related to natural phenomena such as atmospheric disturbances. In general, the concentration of air pollutants is measured sequentially as a function of time, which consists in establishing time series. Air pollution modeling from historical emission-based data has been a key technique for air pollution prediction and detection until now. However, these techniques have several drawbacks, in particular the need to rely on emission concentration factors and flux estimates, which may be inaccurate and have large uncertainty. In addition, modeling requires a thorough knowledge of the different phenomena and interactions that characterize such an ecosystem. In this framework, our objective is to improve the air pollution modeling work in Lebanon by introducing the main air pollutants, their characteristics, sources, and measurement techniques. Then we present the air quality modeling and especially the use of the theory of extremes. Finally, we calculate the air quality index to predict the level of pollution using time series models.

1.10.1 Modeling extreme pollutants values

The theory of extreme values and records is a branch of statistics dealing with extreme deviations from the maxima or minima of observations exceeding a certain threshold. It seeks to evaluate, from a given ordered sample of a given random variable, the probability of events more extreme than anything previously observed. Extreme value analysis is widely used in many disciplines, such as structural engineering, finance, earth sciences, traffic forecasting, and geological engineering (De haan, et al., 2010) (Castillo, 1988) (Abarbanel, et al., 1992). The extreme value theory can be divided into two parts, which are however related to each other. On one hand, the Fischer-Tippett theorem allows us to know the asymptotic distribution of the maximum of n random variables. This theorem is the analog of the central limit theorem, which is interested in the asymptotic distribution of the sum: while the central limit theorem shows a normal distribution, the Fischer-Tippett theorem shows the GEV (Generalized Extreme Value) distribution. On the other hand, the study of "excess function" (law of excesses) reveals the generalized Pareto distribution (GPD).

Historically, various methods have been adopted to examine the behavior of the tail and the associated risk, including the maximum block approach, or copula models. The Generalized Pareto Distribution (GPD) model is considered a method to provide relevant information regarding extreme losses, especially in times of crisis. In this PhD thesis, the extreme value theory is applied to the new case whose extreme values constitute a real change in the state of air pollution in the Bekaa region in Lebanon to detect the level of air pollution in this region whose objective is to create an alert system in case the pollutants exceed a determined threshold. In addition, we use the theory of records to predict future records of air pollutants and calculate the probability of observing a new record based on the

number of records observed. This theory is new in the field of the environment and especially in the prediction of the future records pollutants which is based on the calculation of the laws of exact probability like the probability distribution of a number of records, from where this theory is more effective than that of the extreme values because on the one hand, it treats the exact laws of a certain statistic which do not depend on the law of observations, on the other hand, it takes into account the temporal structure of the pollutants insofar as we study the instants of records contrary to the theory of the extreme values where we are interested only in the study of the limit law of the maximum (Khraibani, et al., 2011), (Hayek, et al., 2020).

1.10.2 Forecasting and Time Series

Air quality forecasting is a crucial topic in monitoring. Today, forecasting is carried out using a variety of computational tools and modeling approaches. Forecasting emerged in the 1960s, when work on air pollution monitoring and tracking began, following the catastrophic events of the late 1950s. These forecasts were meteorological forecasts of adverse air quality conditions using weather forecasting models (Niemeyer, 1960). In the 1970s, air quality forecasting models, in the broadest sense, emerged. There are two main types of models: statistical models, which aim to predict future measurements in a time series based on previous data, and mechanistic or deterministic models, which allow numerical simulation of knowledge models related to atmospheric phenomena.

1.11 Statistical model

In our case, we use several advanced statistical approaches to characterize, on one hand, the behavior of the quality of the Litani River in several sites or stations and, on the other hand, to describe the properties of the air quality in the Bekaa region. For each of the methods, we present the most important aspects concerning the reliability and the robustness of the obtained results through an analysis and an implementation of these methods on the simulated data, and the real data. Different stages are mentioned. First of all, we are interested in processing and filtering the data; it should be noted that difficulties were encountered during the collection of data due to technical issues and the lack of database registers. However, for the physicochemical parameters, we rely on data provided in the thesis (Hayek, 2021) and on those provided by the national authorities of the Litani River. As for the air quality data, we collected these through the department of the Ministry of the Environment and we completed the missing information using high-performance statistical techniques. Once our database was established, we used the obtained data to calibrate and validate the statistical model to minimize the error in the obtained results.

Secondly, it is necessary to design the survey and identify the most suitable statistical method to evaluate environmental pollution. Various tests were performed for different water quality concepts. However, a water quality model can be established from limited or non-existent data. In such cases, it

is difficult to determine which processes to integrate into a model, especially for forecasts over a short lead time with a low risk of error. Then, after selecting the appropriate model, it is required to confirm the statistical model from the data simulated according to the SBM method, and from the collected data by considering the different criteria for calculating the errors between the observed and adjusted series using the chaotic model and time series.

Thirdly, we should also evaluate the performance of the model. Once we select and validate the model, it is possible to carry out more in-depth analyzes and comparisons between different approaches. The simulation consists in restoring the model according to its temporal evolution. Performance verification consists of evaluating the results achieved by the model.

1.12 Methodology and objectives

As already discussed in the introduction of statistical analysis, modeling and monitoring of the water quality parameters are important approaches to studying the water quality in river basins. Few statistical and ecological studies are found in the literature to study the water quality in the Litani River basin in Lebanon, which is considered the most polluted study area due to industrial effluents from nearby industries, disposal of agricultural waste, and municipal sewage effluent to the river. Three stations were selected in the river and monthly water quality data were collected from 2008 to 2018 for water quality analysis and modeling. The specific research objectives of this thesis are:

1. A new classification method "Stochastic Block Model" (SBM) in environmental sciences environmental sciences has been developed to classify the physicochemical parameters of several monitoring stations in Lebanon. This approach is innovative in environmental research and it differs from classical clustering methods, particularly when processing big data.
2. The SBM method gives interesting advanced results on inter and intra station parameters by calculating their estimated connectivity matrix between the obtained clusters. The strength of this method was validated by numerical simulations and by comparison with classical clustering methods classical classification methods (PCA, K-means, Dendograms). These classical approaches do not make possible to show the structure of studied stations as opposed to the suggested method.
3. The problem of the water quality is a challenging issue since water is a complex physical, biochemical and ecological system. A new approach based on Chaos theory to interpret complex arrays of water quality data and identify possible sources or factors that influence water quality.

1.13 Statistical modeling and multivariate analysis of water quality

The concept of statistical modeling is based on the collection of data and their filtration, as well as on their analyses, and the presentation of the expected results as well as their interpretation. It also relates to a set of methods used to process high-dimensional data. It covers all data characteristics, including

planning for information collection. In this thesis, a statistical analysis can be divided into different phases as follows: (a) Describe the different types of environmental data that will be analyzed and presented, (b) Examine the correlation between the different categories of physicochemical parameters, (c) Elaborate and present a new SBM model allowing to understand the existing relationships between the variables as well as between the stations (d) Check the validity of the model, (e) Make a predictive analysis of the future trend based on the chaotic theory.

Note that multivariate statistical analysis consists of studying and analyzing several variables at a given time. This analysis concerns the various objectives of each of the different categories of the multivariate analysis as well as the context in which they are situated and how they evolve between them. Indeed, the multivariate analysis includes both univariate and multivariate analysis allowing a better synthesis between the different variables and their adequacy.

There are several methods in the literature to perform statistical analyzes of physicochemical parameters to assess pollution in a river, each of them involves its type of analysis depending on the type of address problem. The methods are:

(i) Analysis of variance (ANOVA) is a particular type of statistical testing of hypotheses, commonly applied in the analysis of experimental data from a sample. A statistical test is a method for making decisions using information from different samples. The multivariate variance analysis (MANOVA) makes it possible to generalize the analysis of variance to consider the presence of several independent variables to be studied simultaneously. MANOVA examines the difference between the three stations and the seasonal variations and makes it possible to classify the different stations according to common spatial characteristics (Eneji et al., 2012) (Hayek, et al., 2021). Principal component analysis (PCA) is a mathematical tool that uses an orthogonal transformation to transform a series of observed data from correlated variables, called principal components. (Haydar, et al., 2014)

(ii) Factorial analysis (FA) is close to the PCA. It reduces the variability between correlated and observed variables in terms of factors, i.e. a reduced number of variables. (Hayek, et al., 2020), (Nehme, et al., 2021).

(iii) Discriminant analysis (DA) or canonical analysis of variables is a method of statistical analysis allowing the prediction of the behavior of dependent categorical variables or groups of variables based on one or more binary factors independent of continuous variables called predictive variables. The purpose of this analysis is to test whether a set of variables can be used to differentiate between two or more different class groups. This analysis also resembles a cluster analysis which is used to assess spatial and temporal variations of different water quality parameters. DA and PCA will reduce the dimensionality of high-

dimensional data and will indicate the few relevant parameters that represent the greatest variation in water quality (Mustapha, et al., 2012).

(iv) Classification is the grouping of variables where elements in a group, named cluster, are more similar to each other than elements in other groups, also called dissimilar clusters (Kowalkowski, et al., 2006).

(v) The canonical correlation method (CCA) is a method that establishes a linear correlation between two or more variables. When applied to two sets of variables, it represents a general and canonical version of the bivariate correlation. It establishes the two criteria, for each variable, that are optimal in terms of their correlation and at the same time determines the associated correlation. This CCA technique is used to identify the relationship between two sets of data such as air pollution and weather (Statheropoulos, et al., 1998).

After a general outline of the main statistical methods existing in the literature to treat the physicochemical parameters and to evaluate the water pollution in the Litani River in Lebanon, we implement new probabilistic statistical methods for the first time in this study for water quality modeling and monitoring.

1.13.1 Modeling by chaotic theory

Water quality models are powerful tools for determining the concentration of pollutants in aquatic environments and reducing labor and material costs for many chemical analyses. As part of this thesis, a conceptual model based on chaotic theory and collected data is applied. The data used in this model are adapted to the degree of complexity and the initial conditions of the samples taken from Qaraoun Lake. Several approaches are used in the literature to develop prediction models based on original data. These methods raise difficulties, especially when the original data are not sufficiently reliable or simply chaotic. Indeed, the term "chaotic" means that it is a complex system, sensitive to initial conditions and presenting important recurrent characteristics. Chaos theory is generally evoked in the literature to describe situations in which complex random behaviors appear in simple deterministic nonlinear processes, strongly dependent on their initial conditions. (Lorenz, 1963), (Wilks, 1991).

The implementation of this theory in the context of water quality studies makes it possible to develop an original approach compared to the deterministic and stochastic methods previously used in this field. The fact that the different physicochemical parameters have a certain degree of dependence on the initial conditions, called chaotic, makes it possible to model these parameters in a simpler way to assess the quality of the water in Qaraoun Lake and to propose more effective pollution reduction strategies. We present non-linear approaches which allow for direct or indirect highlighting of chaotic behavior. Several nonlinear dynamic approaches are used: first, we address the reconstruction of phase spaces; then, we use the false nearest neighbor (FNN) method; and finally, we apply a method based on the correlation dimensions. The presence of chaotic behavior could provide favorable results for a reliable

prediction of the variation of physicochemical parameters. The advantage of this theory comes when a complex model with a large number of data, the correlation, and dependencies of the parameters it is not possible to estimate the required parameters from the collected data. In addition, some of the physicochemical parameters are modified or sensitive to initial conditions due to climate change or human activities.

1.13.2 Classification: Stochastic block method (SBM)

Statistical modeling is an effective tool for understanding, analyzing, and proposing solutions to different types of problems related to the management of natural resources. It is recognized that networks are used in the context of modeling interactions between groups of entities. Networks are among the most powerful modern tools for data analysis. Many researchers have developed models and algorithmic approaches to analyze and develop networks. Among these models, we find the stochastic block model (SBM) which is proposed by Anderson et al (1992) and Holland et al (1983). This probabilistic model of random graphs aims to generate groups, called blocks, and more generally clusters in networks. The SBM model has been applied in many fields such as networks and biology (Fortunato (2010), Porter et al. (2009)) as well as in the field of natural sciences and machine learning (Goldenberg et al. (2010)). We used the SBM method to characterize the physicochemical parameters of water in several stations of the Litani River to determine the water's quality and the degree of pollution in the three stations (Lake Qaraoun, Jeb-Jenin, and Ghzayel). Several techniques were used to reduce the overall dimension of the data and solve high-dimensional problems. Classical classification techniques are the most frequently mentioned in the literature. However, the PCA method is a classical type classification method, which does not consider dependencies between variables. In addition, there are the most used classification techniques, such as the k-means method and hierarchical classification. These methods provide a graphical representation allowing to summarize the useful information of the database and they also allow for management and making an interpretation in a geometric way of the distribution of the physicochemical parameters due to the projection of the parameters on two main axes. This graphical representation, about the various factors or principal components, is presented in the form of several axes.

We discuss these different approaches in this study and compare them with the SBM method. These methods are simple to implement, flexible, applicable to all types of data, and allow for managing all forms of similarities and distances. Their disadvantages compared to the SBM method lie in their high cost, time, and the fact that their results are generally unsatisfactory.

The SBM method aims to highlight the relationship between the physicochemical parameters to properly assess the quality of water in the three stations of the Litani River. One of the essential aspects of this method consists in identifying the observations relating to water quality parameters having similar properties. The SBM method is not only classification but also a contribution to understanding

the relationship between the different sets obtained. We are particularly interested in the estimation of the parameters of the SBM model and the classification of the nodes of the network concerned. We present applications of the proposed method using simulated data series, then real series.

Chapter 2

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

This chapter refers to the article that was published in the Journal Statistics, Optimization and Information Computing.

2.1 Summary

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.

2.2 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, (2020) (January)), and recreational activities (Simpi B., 2011). Unfortunately, they are among the most seriously threatened ecosystems due to anthropogenic activities over the past century (Dudgeon D, (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 (Fadel, et al., 2014), (Qiang He, 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, 2021), (Hayek, et al., 2020), (Mark S., 2012), (Chaden M.H., 2014), trend analysis method, (Groppo, 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, 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, 1992) and (Holland, 1983) and aims to produce classes, called blocks, or more generally clusters in networks. Several authors have proposed some generalization of this model (Mariadassou, 2010) have treated the case of weighted networks by using the SBM model (Airoldi, 2008), (Ng, 2021) and then (Latouche, 2011) have focused on the SBM model with overlapping clusters. More recently, Barbillon et al. (Barbillon, 2017) have dealt with the case of multiplex networks, where several edges of different types can exist between a pair of nodes and (Zreik, 2017) and (Matias, 2017) have

extended the model to deal with dynamic networks. However, (El Haj, 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, 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, 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, 2008) has proposed a variational approach by introducing the variational expectation maximization (VEM) algorithm while (Latouche, 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, 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, 2012).

2.3.Environmental concept

2.3.1. 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 Figure 15).

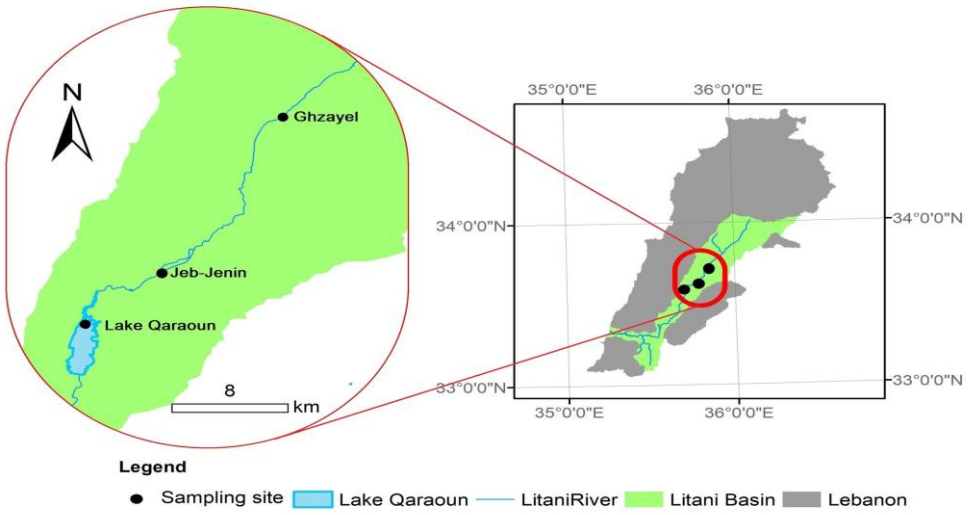


Figure 15 Litani River and located three stations.

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: $N\ 33^0 43'56'' N$ (Latitude), $E35^0\ 56'\ 52'' E$ (Longitude). Noted that the description of the sites Jeb-Jenin (Upper Litani River), with the coordinate: $N33^\circ\ 38^0\ 331^{00}$ $E35^0 46'79''$ 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: $N33^0 46'42''$ (Latitude) $E35^0 53'26''$ (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.

2.3.2. Sampling test

Several parameters such as Temperature, DO, pH, conductivity, total dissolved solids (TDS), Salinity, Ammonia, Nitrite, Nitrate, SO_4, PO_4, were analyzed for each water sample collected during the 2008-2018 periods in three stations using spectrophotometry (Table 4) (Fadel, 2019), (Fadel, 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.

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

Table 4 Methods and quality control used to measure the physicochemical parameters

2.4. Proposed statistical method

2.4.1. 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 $\{1,...,n\}$ 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 $i,j \in \{1,...,n\}$,

$$X_{ij} = \begin{cases} m_{ij} & \text{if the nodes } i \text{ and } j \text{ interact with an interaction weight} \\ 0 & otherwise \end{cases} \quad (2.1)$$

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 $i \in \{1,...,n\}$ and $q \in \{1,...,Q\}$,

$$Z_{iq} = \begin{cases} 1 & \text{if vertex i belongs to cluster q} \\ 0 & otherwise \end{cases} \quad \forall \, i \in \{1, \dots, n\}, \forall \, q \in \{1, \dots, Q\} \quad (2.2)$$

The indicator variables $\{Z_{iq}\}$ for $(i,j) \in \{1,\dots,n\} \times \{1,\dots,Q\}$ are independent.

2.4.2. Mixture model with latent classes

We propose to generate the stochastic block model as follows:

- The (latent) vectors Z_i, for $i \in \{1,...,n\}$, are independents and sampled from a multinomial distribution as follows

$$Z_i \sim \mathcal{M}(1,\alpha = (\alpha_1,...,\alpha_Q)),$$

where $\alpha = (\alpha_1,...,\alpha_Q)$ is the vector of class proportions of dimension $1 \times Q$ such as;

$$\sum_{q=1}^{Q} \alpha_q = 1$$

- The weighted matrix $X = \left(X_{i,j}\right)_{i,j\in\{1,\dots,n\}}$ associated to the network contains the weights of the edges between each two nodes in the network so that $X_{i,j} = l$ if there is an edge joining the nodes i and j and is weighted by the value l. The (observed) variables $\{X_{ij}, i,j\in[n], i < j\}$ are independent conditionally on the sigma-field generated by $\{Z_i, i \in [n]\}$, and are sampled from a Gaussian distribution as follows

$$X_{ij}|Z_{iq}Z_{jl} = 1 \sim N(\mu_{ql}, \sigma_{ql}^2), \qquad (2.3)$$

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

2.4.3. 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

$$\log \mathbb{P}_\theta(X) = log \sum_Z \mathbb{P}_\theta(X, Z) \qquad (2.4)$$

where $P_\theta(X, Z)$ is the joint distribution such that

$$P_\theta(X, Z) = P_{\mu,\sigma}(X|Z)P_\alpha(Z),$$

where

$$\begin{aligned} \mathbb{P}_{\mu,\sigma}(X \mid Z) &= \prod_{i<j}^{n}\prod_{q,l}^{Q} \mathbb{P}_{\mu_{ql},\sigma_{ql}}(X_{i,j} \mid Z_i = q, Z_j = l) \\ &= \prod_{i<j}^{n}\prod_{q,l}^{Q}\left(\frac{1}{(2\pi)^{1/2}\sigma_{ql}} e^{-\frac{1}{2}\frac{(X_{ij}-\mu_{ql})^2}{\sigma_{ql}^2}}\right)^{Z_{iq}Z_{jl}} \end{aligned} \qquad (2.5)$$

and

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

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 developed by (Jordan, 1999) and (Jaakkola, 2000). 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 (2.4) as follows

$$log\ \mathbb{P}_\theta(X) = J_\theta(R_X(Z))\ +\ KL(R_X(Z)\ ||\ \mathbb{P}_\theta(Z \mid X)), \qquad (2.6)$$

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

$$KL(R(.)\ ||\ \mathbb{P}_\theta(.|Z)) = -\sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(Z|X)}{R_X(Z)} \qquad (2.7)$$

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

$$J_\theta(R_X(.)) = \sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(X,Z)}{R_X(Z)} \qquad (2.8)$$

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, 2003), the approximate distribution $R_X(Z)$ can be factorized over the latent variables Z_i as follows

$$R_X(Z) = \prod_{i=1}^n R_{X,i}(Z_i) = \prod_{i=1}^n h(Z_i; \tau_i), \quad (2.9)$$

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

By combining equations (2.4), (2.8) and (2.9), we obtain

$$J_\theta(R_X(Z)) = -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_{i<j} \sum_{q,l} \tau_{iq}\tau_{jl} \left(-\log\left((2\pi)^{\frac{1}{2}}\sigma_{ql}\right) - \frac{1}{2}\frac{(X_{ij}-\mu_{ql})^2}{\sigma_{ql}^2} \right) \qquad (2.10)$$

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 $i \in \{1,...,n\}$, we obtain $\hat{\tau}$ by a fixed point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_l \left(\frac{1}{(2\pi)^{1/2}\sigma_{ql}} e^{-\frac{1}{2}\frac{(X_{ij}-\mu_{ql})^2}{\sigma_{ql}^2}} \right)^{\hat{\tau}_{jl}} \quad (2.11)$$

The estimation of τ is obtained from (2.11) 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

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

2.5. 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 groups $\hat{Q}$. Roughly, this criterion is proposed by (Daudin, 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

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

2.6. Numerical experiments

The purpose of the present section is to outline the major characteristics of the newly developed inference algorithm and to demonstrate the validity of the proposed approach through some simulated data and then by applying it to a real dataset.

2.6.1 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:

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

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

and

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

Using the Gephi software and the "Force Atlas" layout algorithm, the graph shown in Figure 16(a) gives the structure of the simulated data. We have three communities that appear. Notice that throughout the graphs below, the line width used to represent the edges is proportional to their weight, so a large line between two vertices indicate a strong relationship between the vertices.

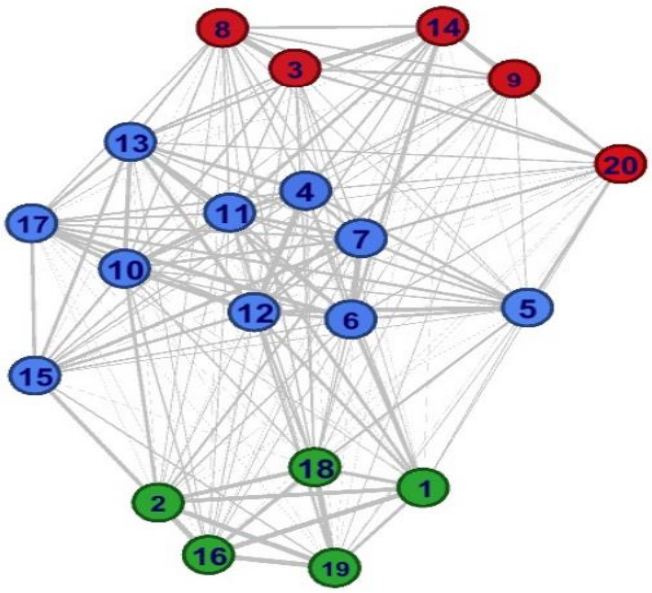


Figure (a) Simulated data graph visualization with Gephi

Clusters	*Vertices*
1	***3 8 9 14 20***
2	***4 5 6 7 10 11 12 13 15 17***
3	***1 2 16 18 19***

Table (a) Grouping simulated graph vertices into clusters

By using the developed algorithm in R software, the vertices are grouped into three clusters as shown in Table (a). Table (a) shows that the nodes in the simulated data graph are broken into three clusters which are the same as in Figure (a). This further validates the performance of this approach.

Now we sample S = 100 random graphs according to the same mixture model. Then we calculate in Table 5, Table 6 and Table 7, for each parameter, the estimated Root Mean Square Error (RMSE) defined by:

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

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

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

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

Table 5 Root Mean Square Error of the parameter $\bar{\alpha}_q$ for the simulated data.

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

Table 6 Root Mean Square Error of the parameter $\bar{\mu}_{qr}$ for the simulated data.

According to Table 5, Table 6 and Table 7, 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.

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

Table 7 Root Mean Square Error of the parameter $\bar{\sigma}_{qr}$ for the simulated data.

In order to compare the estimated clustering results to the simulated ones, we propose to calculate the Adjusted Rand Index (ARI) proposed by (Hubert, 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.

2.7.Clustering in environnemental 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 to compare the clustering of the related networks at the three different stations (Qaraoun, Jeb-Jenin, and Ghzayel) within the same time frame. By applying the Gaussian SBM, we show that it reveals two clusters for each station as shown in Figure 16 by using "Gephi" software with the layout algorithm "Force Atlas".

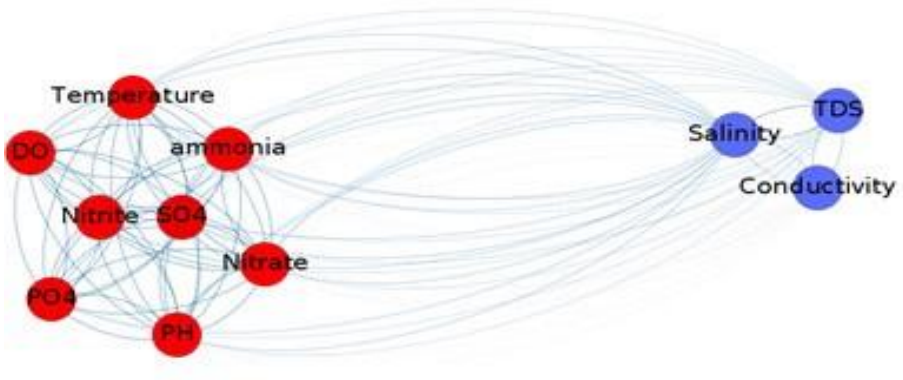


Figure 16 Grouping the physicochemical parameters of each station into clusters.

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.

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 8 Weight matrix for the Jeb-Jenin 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

Table 9 Weight matrix for the Qaraoun 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

Table 10 Weight matrix for the Ghzayel station.

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 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, 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.

2.8.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,

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

Table 11 Comparison of the estimated parameters obtained by applying the proposed method for the three stations: Qaraoun, Jeb-Jenin and Ghzayel.

$\widehat{N}_k = \{i : 1 \leq i \leq n, \hat{g}_i = k\}$ the set of all the physicochemical parameters that belong to the cluster K and $\hat{n}_k = \left|\widehat{N}_k\right|$ 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:

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

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

The number of clusters is the same for the three station (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.

2.9.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 Qaraoun station as an example.

2.9.1. 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.

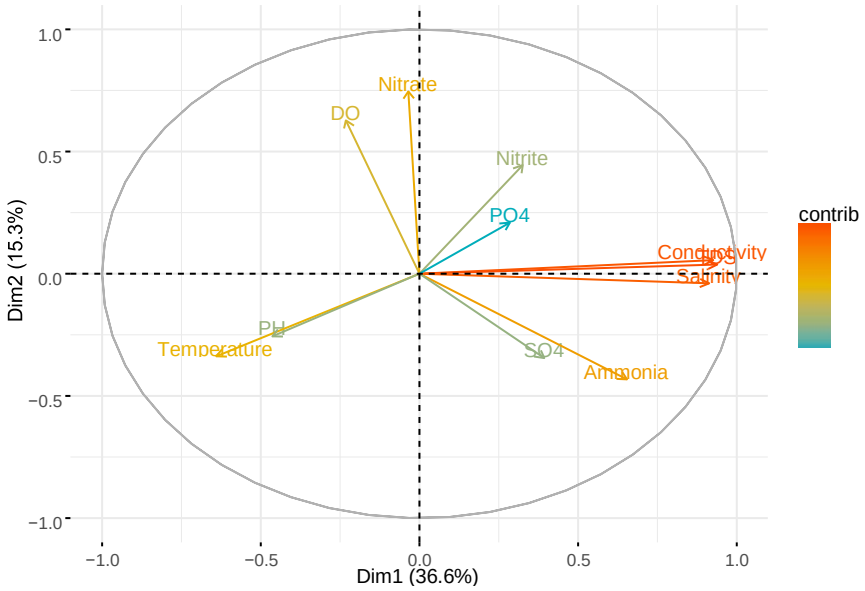


Figure 17 Correlation circle and contribution of the physicochemical parameters of the Qaraoun station.

Figure 17 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. PO_4 and SO_4 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 18.

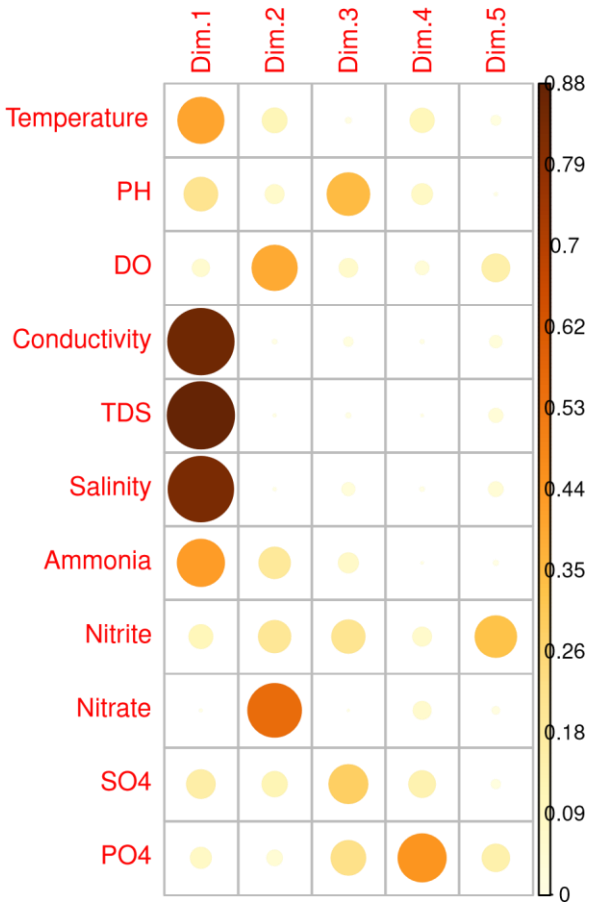


Figure 18 Contribution of the physicochemical parameters on different five axis.

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, SO_4, PO_4).

2.9.2. Hierarchical Cluster

In this subsection, we present the classical hierarchical method to classify the physicochemical parameters of the Qaraoun 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.

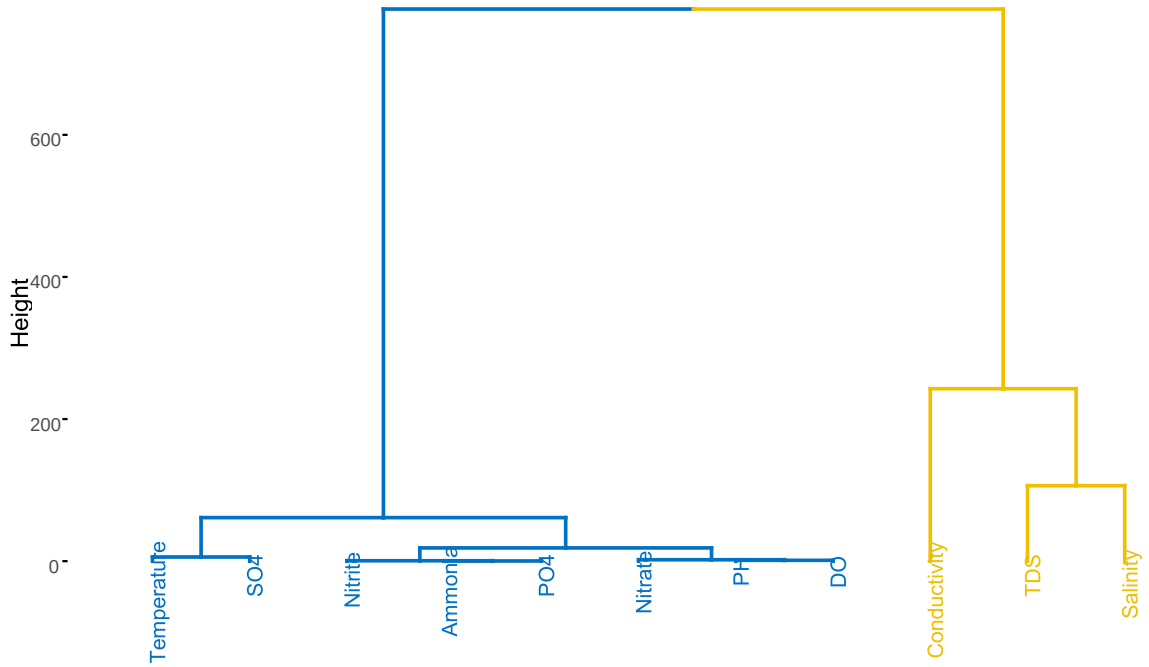


Figure 19 Dendogram of the physicochemical parameters.

Based on the Figure 19, 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.

2.9.3. 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, 1998).

According to the Figure 20, 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.

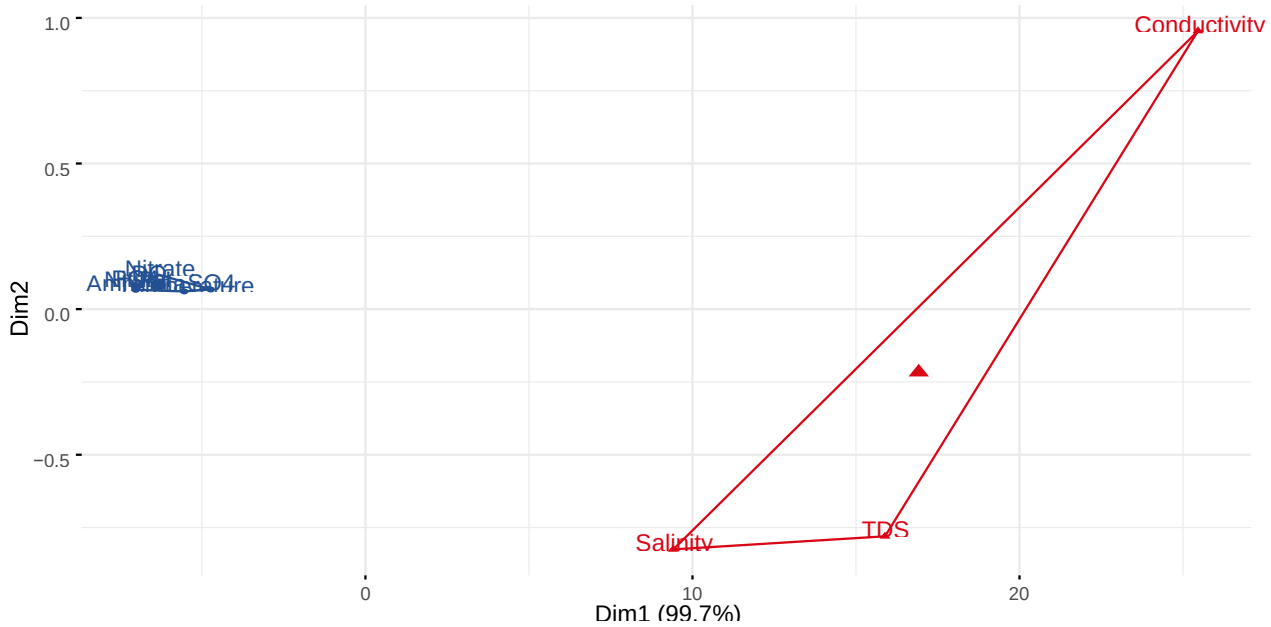


Figure 20 K-means cluster of the physicochemical parameters.

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 Qaraoun 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.

2.10. Conclusion

In this chapter, 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.

Chapter 3

Assessment and modeling of the chaotic variation of the physicochemical parameters at Qaraoun Lake, Lebanon

In the previous chapter, we presented some classification methods allowing us to classify physicochemical parameters among three different stations in the Litani River to be able to assess water pollution. However, it is worthwhile to study the behaviors of these parameters according to the chaotic theory to provide new prospects for water management in order to limit the pollution sources. The present chapter represents a published article in an International Journal "Chaotic Modeling and Simulation (CMSIM)".

3.1. Summary

The world around us often seems unpredictable, disordered, random, and chaotic. The question of water quality is a difficult problem because water is a complex physical, biochemical and ecological system. We propose in this article a new approach for water quality assessment based on the chaotic theory to study the behavior of the nonlinear time series of the physicochemical parameters at Qaraoun Lake in Lebanon. The quality of groundwater and surface water in Lebanon, in most cases, does not meet national, European, and international standards. The impact of human activities in the direction of exhaustion, modification, and the alteration of natural ecosystems is becoming more and more threatening. This impact affects both terrestrial and aquatic ecosystems. Therefore, globally, the environment. One effective solution in analyzing and solving various issues in marine systems is methods based on the mathematical modeling of these systems. In this article, we first study the physicochemical parameters for 11 years (2008-2018), where we present the dimension of integration and the lag time for each setting. Then we are interested in constructing the correct phase space for the nonlinear physicochemical parameters. We present the correlation dimension method to detect the existence of chaos. The estimated dimensions of the parameters indicate the possibility of chaotic behavior in the time series of physicochemical parameters. The presence of chaotic behavior could provide favorable results for a reliable prediction of the variation of physicochemical parameters.

3.2.Introduction

Water covers three-quarters of the surface of our planet. This precious treasure is essential for human, animal, and plant life. This resource meets basic human needs in various areas such as agriculture, power generation, industries, and domestic uses. Due to the increasing human population, the need for

water is gradually increasing (Bavel., 2013). This is aggravated by the decreased amount of potable water available due to reduced precipitation (climate change) on one side, and water pollution on the other (Postel, et al., 1996). Climate change, in recent decades, has had an impact on water quality in lakes and rivers in the Mediterranean region, as well as in different parts of the world. Rising temperatures disrupt climatic conditions and disrupt the natural balance. This situation presents many risks for humans and all other life forms on Earth. In addition, climatic extremes in terms of temperature and precipitation are menacing agriculture, health, energy, the ecosystem, and society (Schoof, et al., 2016). There exist several approaches in the literature to develop an explanatory model from the initial data. But we often have difficulties using these methods, especially when the initial data are not reliable, or we have struggles obtaining them, or they are simply chaotic. Many systems in nature are unpredictable that it seems reasonable to find chaos. When a model of an ecological system behaves chaotically, following compelling conditions on time scales, the chaotic effects become noticeable. A chaotic system depends so significantly on initial conditions that it seems unpredictable even though it is deterministic. A chaotic system is a simple or complex system, sensitive to initial conditions, and presents a repetitive character, and a strong recurrence. A minimal disturbance can lead to considerable instability or unbalance that we cannot predict in the long term. A chaotic system is the opposite of a perfectly regular system. These systems have infinitely complex behavior, which is represented in a space called phase space. The dimensions of the space are related to the number of initial conditions (variables) that act on the system (For example, the speed that describes the system, temperature, pressure, flow...). The theory of chaos is used in the literature in general to designate situations where complex and random behaviors arise from simple non-linear deterministic systems with sensitive dependence on the initial conditions (Lorenz, 1963), (Wilks, 1991). There are fundamental properties in chaotic theory like: non-linear interdependence; hidden determinism; sensitivity to initial conditions, which are very relevant in dynamic systems. Over the past two decades, chaos theory has been applied in various fields including engineering, environmental sciences, economics, medicine, biology social sciences (Holyst, et al., 1996), (Piotrowski, 2020), (Ritchie, 2004), (Seeger, 2002), (Skiadas, et al., 2020). Chaos theory is an approach that emphasizes elements of interactivity and unpredictability instead of using a linear approximation model that focuses on causality (Tobin, 2016). However, chaos theory is not often applied in the environmental firm, especially in water quality resources. Chaos theory was developed to deal with dynamical systems with randomness over time-based on mathematical concepts of recursion (Williams., 1997), (Fadel, et al., 2014). Note that Lorenz (Lorenz, 1963) was the first to report chaotic characteristics of the atmosphere and climate according to numerical models. Hence, applying the chaotic theory to study the physicochemical parameters of water quality gives a new track in the research field compared to the deterministic and stochastic approaches which are already established in this field. The presence

of a determinism order with sensitive dependence on certain initial conditions between the different physicochemical parameters called chaos invokes the possibility of simplified modeling of these parameters to assess water pollution in the Litani River in Lebanon to propose an effective strategy to reduce the sources of pollution. Sources of pollution at Lake Qaraoun and the Litani River in Lebanon are numerous (A.Hayek, 2020). For example, the waste dumped by factories manufacturing agrifood products, pesticides and herbicides, and the residual waters of around a hundred villages and urban towns. In addition, different variations in environmental factors at the Litani River variations make the physicochemical parameters non-linear, which complicates the understanding of Qaraoun Lake state. In this article, we present a new method to study the evolution of the water quality of Lake Qaraoun. This method is chaotic time series given the lack of water quality monitoring most of the time in the lake, measurement error, and other factors such as business activities, tourism, and the existence of chaotic features influenced by climate (precipitation, runoff, seasonal variation, temperature, etc.) There are different methods to study the series of physicochemical parameters: hydrological models, stochastic time series, artificial neural networks, inferential methods (Taheri, et al., 2014), (A.Shaban, 2014), (Daniel, et al., 2018), (Isiyaka, et al., 2019). However, there are not any studies carried out to verify the existence of chaotic characteristics of the physicochemical parameters. Due to seasonal variations, climate change, human and industrial factors, the behavior of physicochemical parameters at Qaraoun Lake is chaotic. We are interested in studying the chaotic behavior in the time series of physicochemical parameters observed at Qaraoun Lake in Lebanon. We present nonlinear dynamic methods that allow direct or indirect identification of chaotic behaviors. Several nonlinear dynamical methods were used: first, we deal with the phase space reconstruction; second, the false nearest neighbor (FNN) algorithm; third, the correlation dimension method. These methods either allow direct identification of the chaotic behavior or at least facilitate the identification by reconstruction of the system, determination of the complexity (especially in terms of dimensionality), and prediction.

3.3. Study Area

Lebanon is well known for its water resources and particularly its rivers. It has 16 permanent rivers and 23 seasonal ones. The Litani is the greatest river in Lebanon and the most important given its length (170 km) and its numerous tributaries including 30 main sources and 140 wells managed by the administration and a very large number of small sources as well as more than 1200 wells which are also exploited by farmers. The Qaraoun lake was selected as the zone of study for this article. It is an artificial lake located in the southern region of the Bekaa Valley near Qaraoun village (Figure 21). The Bekaa region is an area of diverse forms of human activities. From agriculture to tourism and residential neighborhoods, this vibrant area is crowded with a heterogeneous population, a high percentage of which practice agriculture for a living, others choose to work in the touristic field.

However, these activities are mainly practiced near the Litany river, where people can benefit from water for irrigation and the moderate climate at the river.

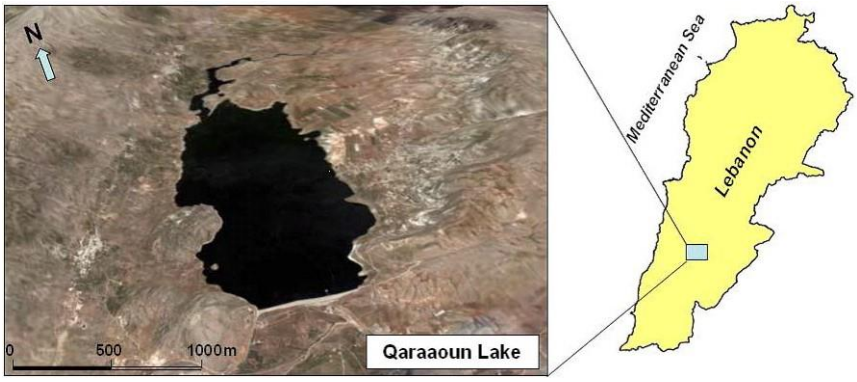


Figure 21 Qaraoun Lake location: Google Earth, Central coordinates: 35◦ 41.38 ' /33◦ 34.54 ' North.

The Qaraoun lake separates the river into an upper and lower parts. The upper part is located in the Bekaa region, and the lower part links the lake to the Mediterranean Sea. Due to its location, Qaraoun lake is an important station to assess and evaluate water quality, that's because the lake is subjected to different types of pollution from the surrounding villages (Hayek, et al., 2020). And still, the water from this lake is used for several purposes like irrigation and hydropower generation. Also, the quality of water in the lake is affected by the upper part of the river where all the water in the lake comes from. The main cause of pollution is the population growth in the Bekaa region, along with population migration from surrounding countries that settle in unplanned camps. This region is also influenced by industrial, tourist, and agricultural activities, as well as by the large chaotic agglomerations built along the rivers.

3.4.Methodology

In this article, we present a non-linear method to study the evolution of the physicochemical parameters at Qaraoun Lake. The methods considered makes it possible to identify the chaotic behaviors of these parameters. We first define the "phase space" reconstruction method, then the "correlation dimension" method.

3.4.1. Phase space reconstruction

The nature of such a dynamic system is described by a number of time dependent quantities $X_1(t)$, $X_2(t),...,X_n(t)$. It can be stochastic, deterministic, or intermediate. Note that the method of identifying the phase space of a time series was presented by Takens (Takens, 1981). Instead of studying the $\{X_n(t)\}_n$ variables separately, it is preferable to represent the system by a single point in a space of n dimensions called: the phase space. Two parameters are necessary for the reconstruction of the phase space: delay time or lag-time noted by T, integration of the dimension noted by m. Since the characters

of systems can be identified as a preliminary indicator, chaos theory has developed a method of identifying the phase space of a time series (Ikeguchi, et al., 2000). This series is an appropriate state vector denoted by:

$$X^m(t) = [X(t), X(t + T), X(t + 2T), \ldots, X(t + (m - 1)T)],$$

where T is the delay time and m is the integration dimension (attractor dimension), and note that $X(t)$ represents the time series of physicochemical parameters observed at time t.

We can find several methods in the literature for determining an appropriate delay time. Some of these methods include: Average Mutual Information (AMI), (Fraser, et al., 1986); Autocorrelation function, ACF (Holzfuss, et al., 1986); Correlation Integer CI (Leibert, et al., 1989). However, we used the AMI in this article because it is the most reliable method since the ACF only reflects linear properties while the CI generally requires more data. The AMI method connects the measure of $X(t)$ (state at time t) to the measure of $X(t + T)$ (state at time $t + T$) (Abarbanel, 1996). The average mutual information method AMI presented by the following formula:

$$I(T) := \sum_{X(i),X(i+T)} P(X(i), X(i+T)) \log\left(\frac{P(X(i), X(i+T))}{P(X(i)) \times P(X(i+T))}\right) \quad (1)$$

Where i is the total number of samples; $P(X(i))$, $P(X(i + T))$ are the probabilities for the measures of $X(i)$, $X(i + T)$, and $P(X(i), X(i + T))$ are the joint probability density for the both measurements $P(X(i))$, $P(X(i + T))$. Whereas, the appropriate delay T is defined as the first minimum of $I(T)$. To represent the system dynamics and the information on the optimal dimension of the phase space, we use the False Nearest Neighbor FNN algorithm proposed by Kennel (Kennel, et al., 1992) whose objective is to provide information on the optimal dimension of embedding of the phase space to properly represent the dynamics of the system. It examines, the nearest neighbor to each vector Y_j as it behaves in dimension $m + 1$, with:

$$Y_j := (X_j, X_j+T, \ldots, X_j+(m-1)T),$$

Y_j is a vector of dimension m which describes the state of the system in position j.

3.4.2. Correlation dimension method

The correlation dimension method (CDM) is a fundamental notion in studying nonlinear dynamical systems, making it possible to distinguish a stochastic process from a deterministic one where the interest is to project on the dimension of the attractor of a system. Note that when the points are attracted towards an equilibrium point or (limit curve), these boundary curves or line are called attractors. The (CDM) allows knowing the necessary information to specify the position of a point on the attractor and the number of variables required to describe the behavior of the dynamic system

inside the attractor. The (CDM) shows significance on the chaotic behavior of the dimension. If the system has a fractal dimension, the physicochemical parameters are assumed to be chaotic. The (CDM) is one of the most effective methods to determine the presence of chaos. In this article, we are interested in the correlation dimension, introduced by Theiler (Theiler, 1986):

$$C(\epsilon) := \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i\neq j=1}^{N} H(r - |Y_i - Y_j| \quad (2)$$

Where N is the number of points on the reconstructed attractor, ϵ is the radius of the sphere centered on Y_i and Y_j, where Y_i and Y_j are two vectors of dimension *m (the choice of m is shown in the section 3.5.3)*; H is the Heaviside function defined by:

$$H(u) = \begin{cases} 1, & \text{if } u > 0 \\ 0, & \text{if } u < 0 \end{cases},$$

We denote by D_2 the correlation exponent (correlation dimension), which can be calculated by the following equation:

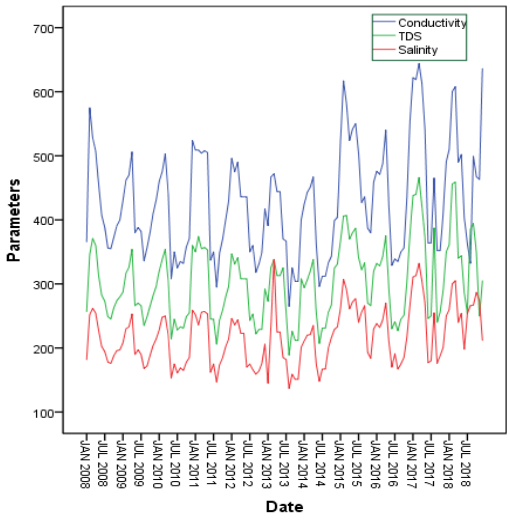
$$D_2 := \lim_{\epsilon\to 0} \frac{\ln C(\epsilon)}{\ln(\epsilon)} \quad (3)$$

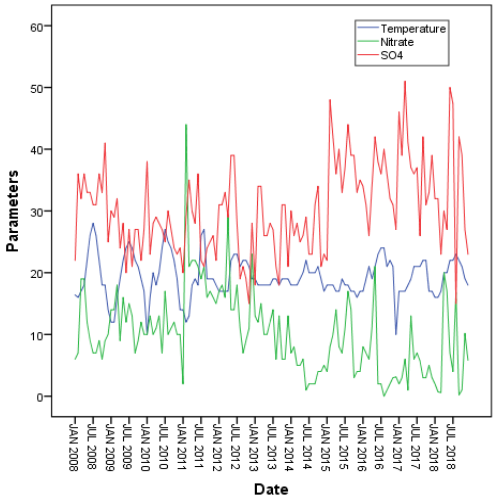
We are based on this method below to identify the variation in the *m*-dimensional correlation exponent. However, if the finite value of the physicochemical parameters curves and the fractal dimension of reconstructed trajectory indicates the possible presence of chaotic behavior. This means that there is an existence of a deterministic dynamic after a certain dimension *m* of the studied parameters.

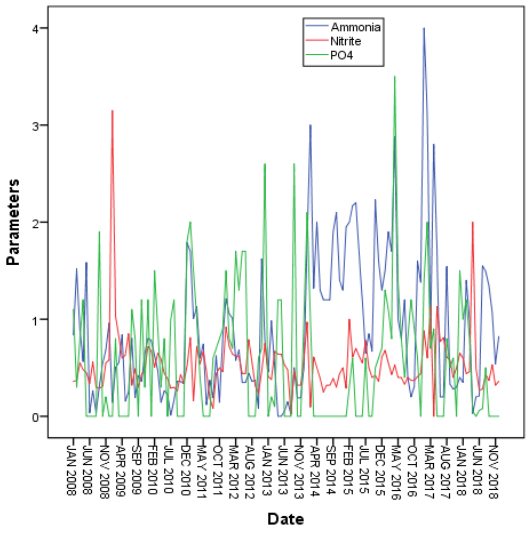
3.5. Application

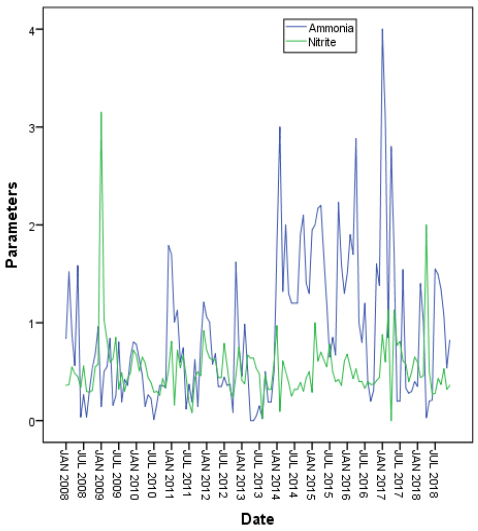
3.5.1. Data Description

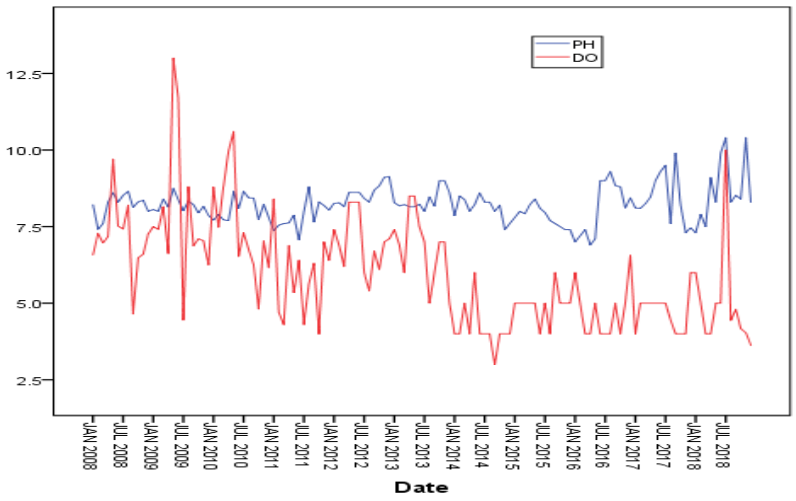
The parameters that define water quality are the physicochemical parameters. These factors are subjected to many changes during the year due to several reasons such as the weather and human activities. The parameters that we studied in this article with respect to the monitoring strategy of the Litani authorities are the following: temperature, pH, Dissolved Oxygen (DO), Conductivity, Total Dissolved Solids (TDS), Salinity, Ammonium (NH_4), Nitrite (NO_2), Nitrate (NO_3), Sulfate (SO_4), Orthophosphate (PO_4). These parameters are affected by seasonal changes, and they are also related to each other. The values of these parameters were collected and recorded monthly over 11 years (2008-2018) from the same location at the Qaraoun Lake are presented in the following figures.









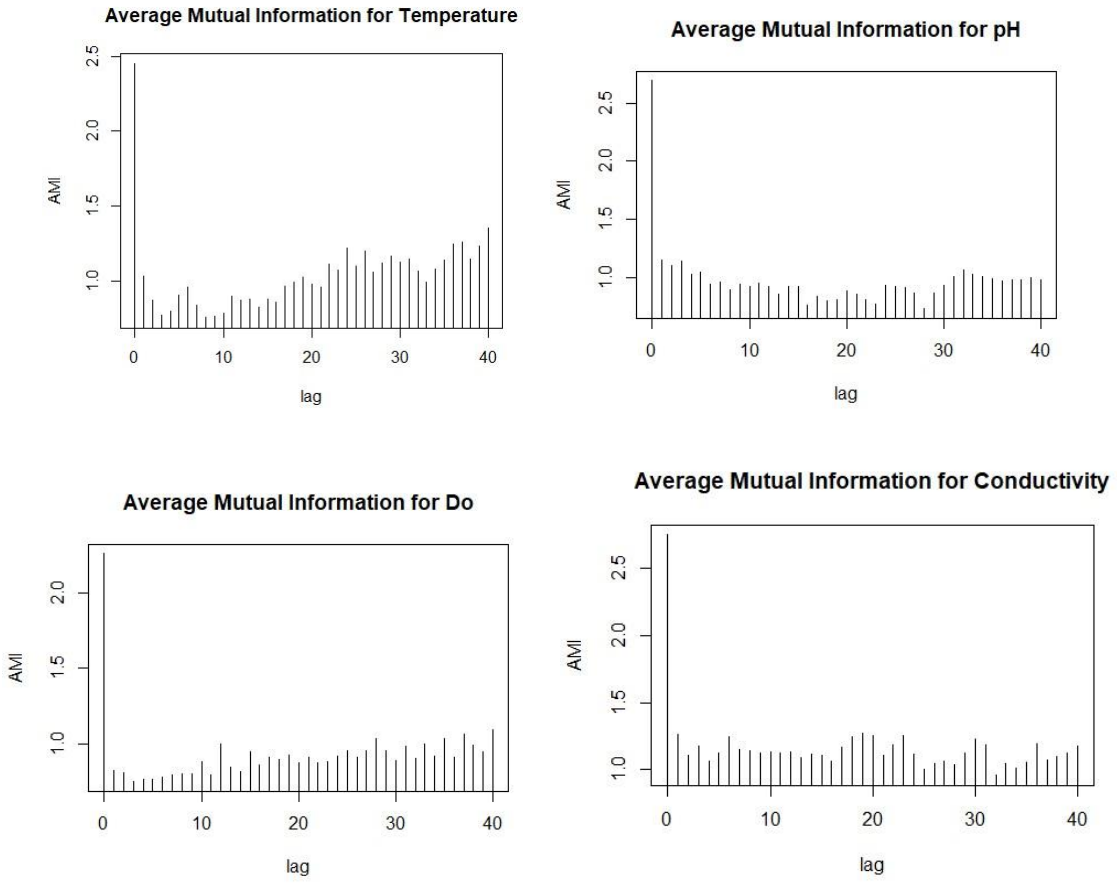


The evolution of all parameters reveals an unbalance in terms of decreases and increases in their values. But overall, along the trend slope, an increase in all parameters is observed. That confirms the growth of the pollution source in the river's neighborhood. From the above figures, it is possible to identify

three zones with a very high level of pollution. The first area is from 2010 to 2013, where a sharp increase in phosphate, nitrate and sulfate concentration and increase in the pH level of water was observed. A peak occurred during 2011, then the values of those parameters progressively decreased till 2013. Based on the above figures, we noticed that some parameters like TDS, Salinity, and Conductivity have predictable behavior, others change chaotically and are harder to predict such as pH, DO, and ammonium. The reason behind the chaotic behavior is that these parameters are more sensitive to the surrounding environment and are dependent on many other factors. So, we notice that they don't follow a specific pattern through the years. Thus, it is important to detect the parameters having a chaotic behavior to make a better understanding of the variation in water quality.

3.5.2. Average mutual information

Average mutual information to construct the phase space is necessary to find the optimal value using the nearest neighbor algorithm, which examines in dimension *m* the closest neighbor to each vector Y_j. Figure 22 represents the nearest neighbor algorithm as a function of the integration dimension for each physicochemical parameter.



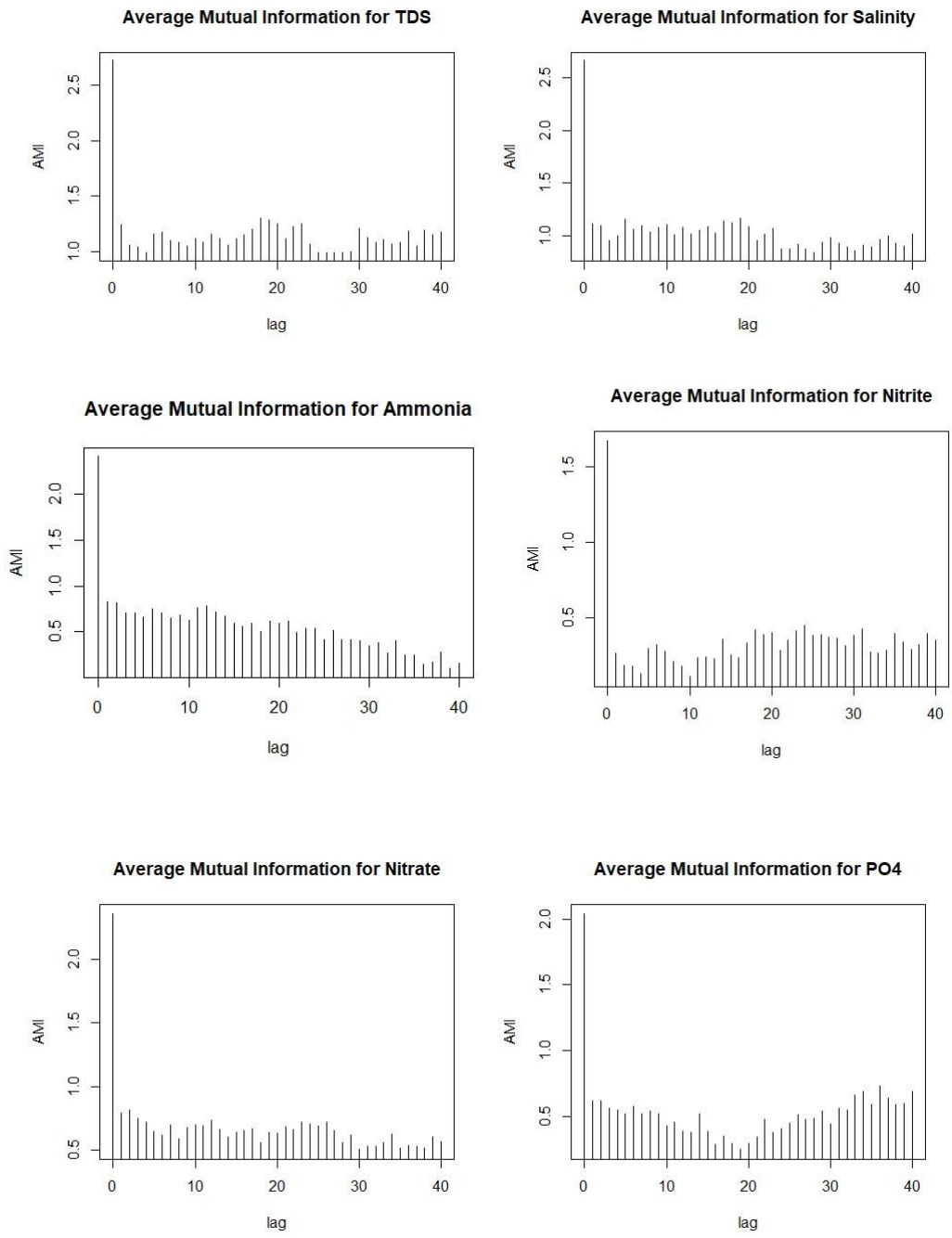


Figure 22 Average mutual information for the different physicochemical parameters for 40 months of measurements.

According to Table 12, the delay time T of each physicochemical parameters is between $1 \leq T \leq 5$. The choice of T is based on the first minimum of AMI.

Parameters	Temperature	pH	Do	Conductivity	TDS	Salinity	Ammonium	Nitrite	Nitrate	SO_4	PO_4
AMI	0.77	1.10	0.74	1.101	0.98	0.95	0.66	0.13	0.79	1.005	0.61
T	3	2	3	2	4	3	5	4	1	1	1

Table 12 Optimal lag-time and minimal value of AMI for the physicochemical parameters.

To find the most suitable lag-time *T* for phase space reconstruction for 40 months of measurement, we calculate AMI for the different physicochemical parameters. Note the Figure 22 provide the different values of *I*(*t*), the first minimum value of *I*(*t*) is given for 40 months represented in the Table 12. For example, the average mutual information for the temperature is equal to 0.77, so the most appropriate lag-time of reconstruction phase space *T* = 3 months.

3.5.3. Optimal dimension of phase space

To represent the dynamic system and determine the information about the optimal dimension *m* of the phase space, we use the nearest neighbor algorithm which examines in the *m* dimension of the closest neighbor at each physicochemical parameter. This algorithm calculates the percent of false neighbors for each dimension and shows which dimension is required to expand the attractor (Kennel, et al., 1992).

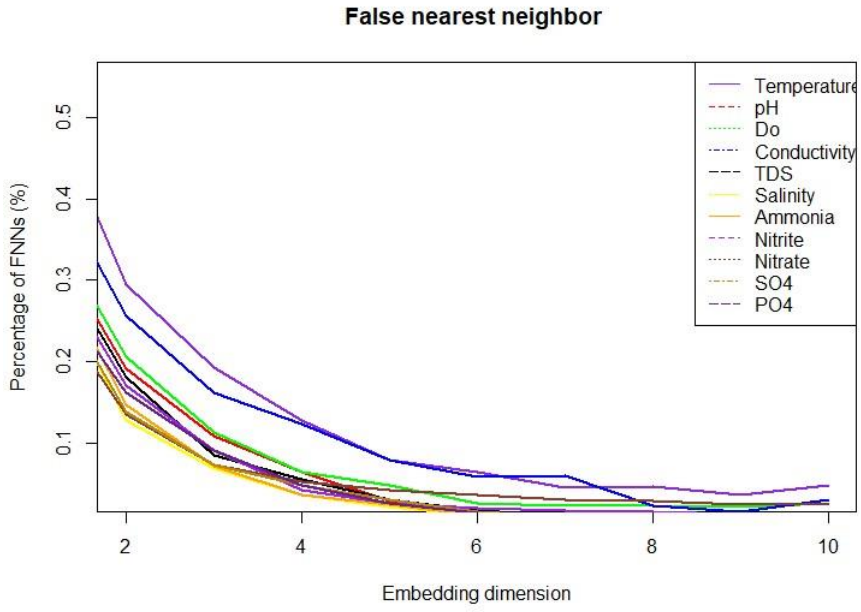


Figure 23 Percentage change of nearest neighbors as a function of integration dimension.

Figure 23 shows the nearest neighbor algorithm as a function of integration dimension.

Parameters	Temperature	pH	Do	Conductivity	TDS	Salinity	Ammonium	Nitrite	Nitrate	SO_4	PO_4
FNN	0.037	0.00125	0.022	0.0158	0	0.0022	0.0036	0.0109	0.0253	0.0057	0.0053
m	9	10	9	9	10	10	9	10	9	10	10

Table 13 Optimal dimension m of phase space for each physicochemical parameters.

Note that, for example, the minimum percentage value for pH is equal to 0.00125 for an integration dimension *m* = 10, then this dimension is the optimal dimension to construct the phase space of pH.

3.5.4. Correlation Dimension method

In this subsection, we use the correlation dimension method. This method was used to calculate the most appropriate integration dimension and identify the presence of chaotic behavior of the physicochemical parameters. We use the Grassberger-Procaccia algorithm (Grassberger P., 1983) to calculate the correlation dimension which has been most widely used in studies of time series hydrology. If the correlation exponent saturates with an increase in the inclusion dimension, then the system is generally considered to have chaos. The saturation value of the correlation exponent is defined as the correlation dimension of the attractor. Now, if the correlation exponent increases unbounded with the increase in the dimension of integration, then the data are generally considered to be stochastic.

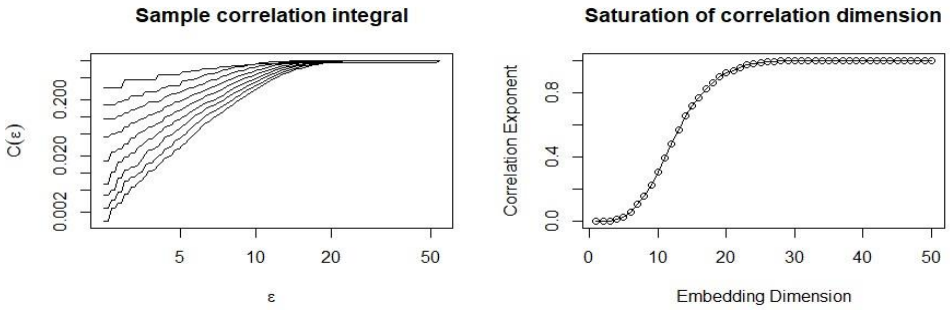


Figure 24 C(ϵ) and correlation exponent for temperature

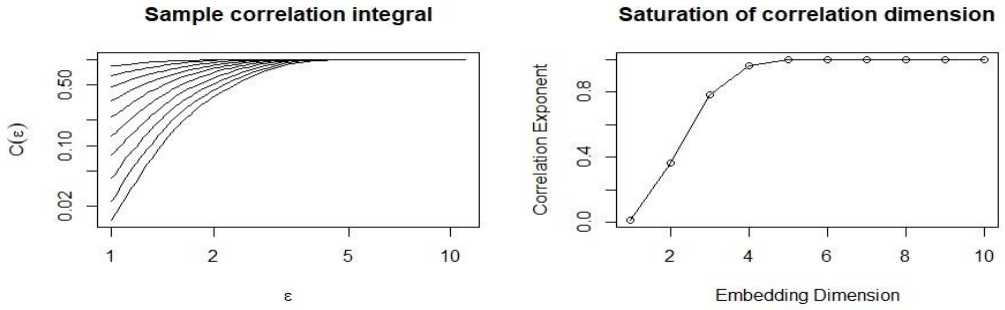


Figure 25 C(ϵ) and correlation exponent for pH

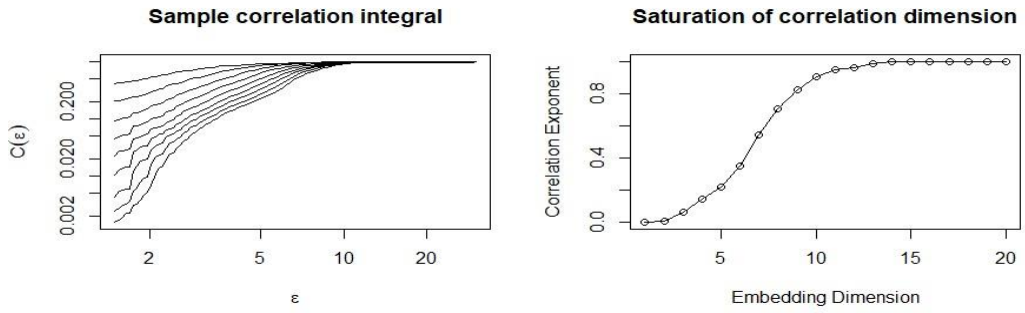


Figure 26 C(ϵ) and correlation exponent for DO

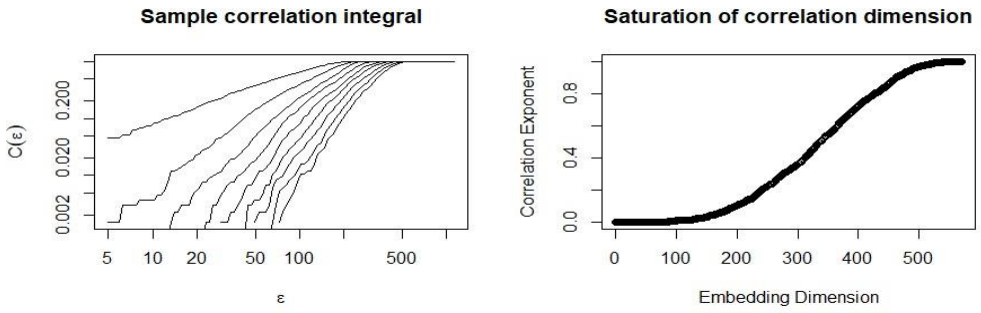


Figure 27 C(ϵ) and correlation exponent for Conductivity

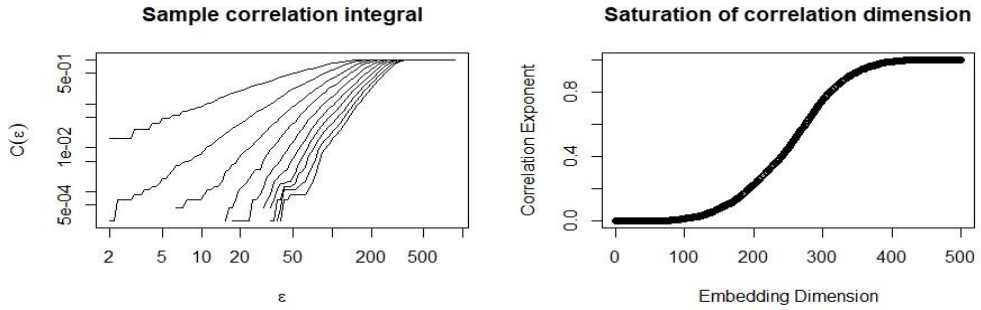


Figure 28 C(ϵ) and correlation exponent for TDS

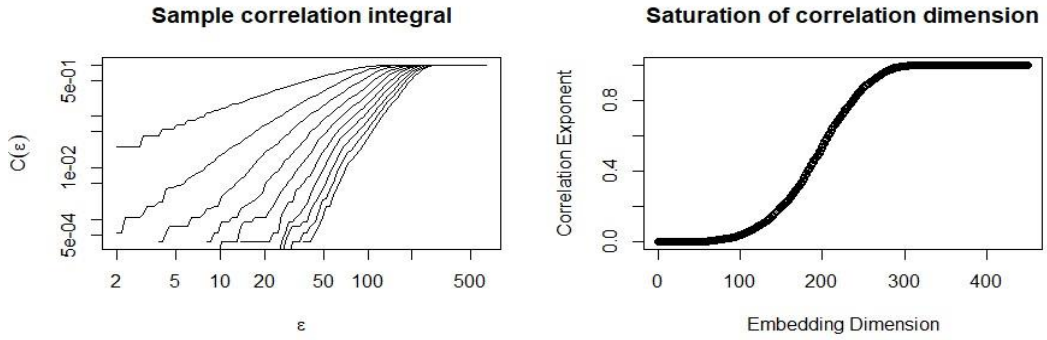


Figure 29 C(ϵ) and correlation exponent for Salinity

There are several methods in the literature to identify chaotic systems from stochastic systems. In this article, we use the correlation dimension method, which is probably the most fundamental. Note that if the correlation exponent saturates with an increase in dimension *m*, the physicochemical parameters will be considered chaotic. If the correlation exponent increases unbounded with increasing embedding dimension, the system is generally considered stochastic. The correlation dimension method was used to identify the dynamic behavior (chaotic or stochastic) for each of the physicochemical parameters of Qaraoun Lake. We identify the presence of chaos in the physicochemical parameter series by plotting the values of the correlation exponents against the corresponding integration dimension values. Figure 24 to Figure 34 shows that the relationship between the correlation function $C(\epsilon)$ and the radius ϵ for the different correlation dimension *m* for

each physicochemical parameters, then the relationship between the correlation dimension values $D_2(m)$ and the dimension values of m. We notice that the values of the correlation exponent increases with the dimension m then it stabilizes. Based on these figures, we note that the saturation of the correlation exponent indicates the existence of deterministic dynamics. The value of the correlation dimension also suggests the possible presence of

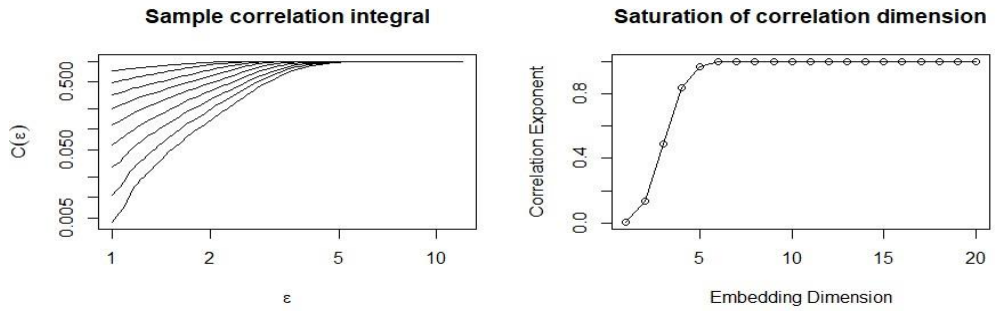


Figure 30 C(ϵ) and correlation exponent for ammonium

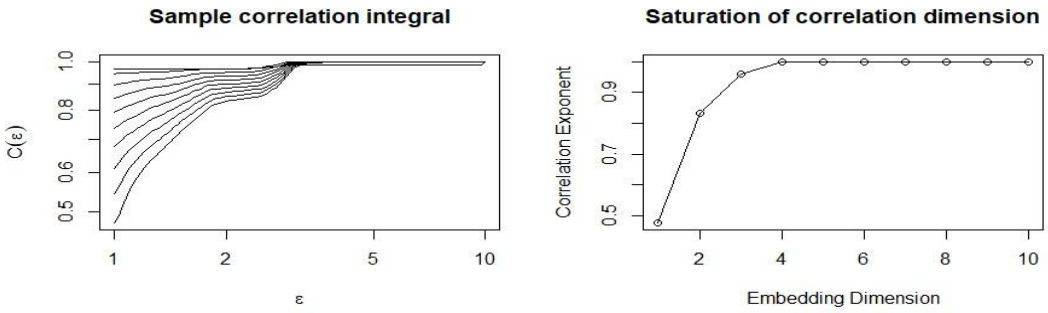


Figure 31 C(ϵ) and correlation exponent for Nitrite

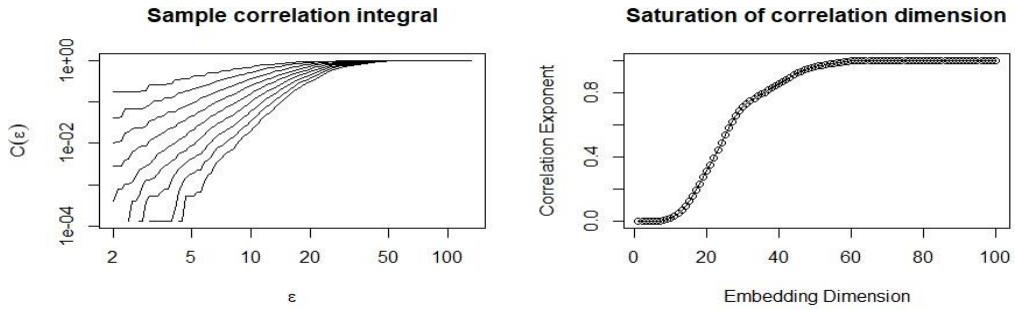


Figure 32 C(ϵ) and correlation exponent for Nitrate

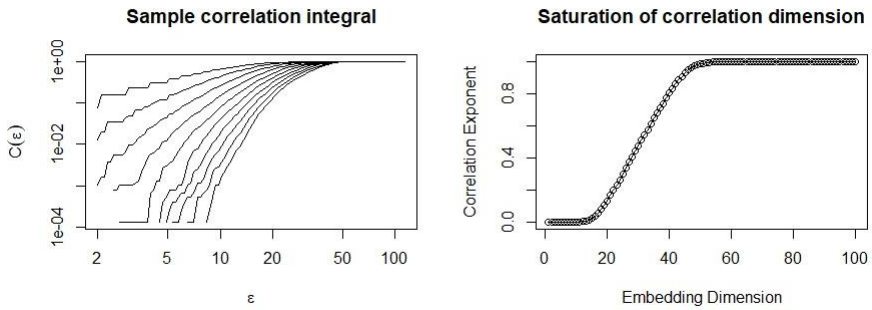


Figure 33 C(ϵ) and correlation exponent for SO_4

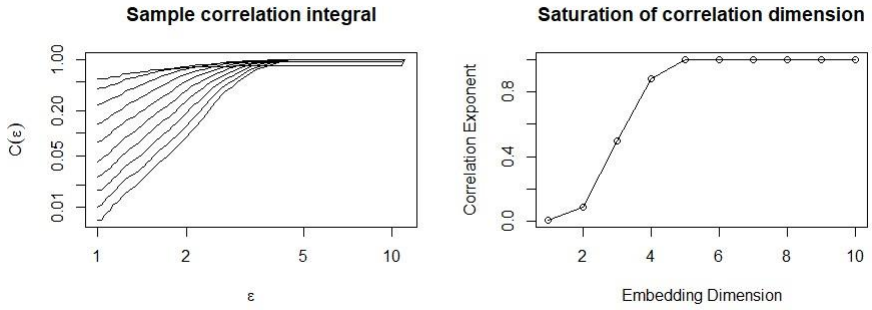


Figure 34 C(ϵ) and correlation exponent for PO_4

chaotic behavior in the data set. The finite values of these curves and the fractal dimension of the reconstructed trajectory indicate the possible presence of chaotic behavior. That means there exists a deterministic dynamic after this dimension *m*. Based on the figures above we can see for the different physicochemical parameters that the value of the correlation exponent increases with the dimension of embedding up to a certain value, and after that it stabilizes for the higher dimensions. We consider that these parameters have a low dimensional chaotic behavior when the value of the correlation exponent is finite, weak and not integer. Note that the exponent correlation saturation value is defined as the attractor correlation dimension. The finite dimensions and correlation obtained for the different series of physicochemical parameters seem to indicate the possible presence of a chaotic behavior of low dimension in the dynamics of each of these components for temperature, pH, DO, ammonium, Nitrite orthophosphate, and a chaotic behavior of large dimension in the dynamics of the other parameters Conductivity, TDS, Salinity, Nitrate, sulfate. The presence of a chaotic system in the series of physicochemical parameters is further confirmed by the correlation dimension method. The finite value of these curves and the fractal dimension of the reconstructed trajectory indicate the possible presence of chaotic behavior. This means that there is an existence of a deterministic dynamic after this dimension *m*.

To understand these results, we must look at the origin of these parameters. In fact, the parameters that are related to living species in the lake seem to have a chaotic behavior. This can be explained by understanding that the pollution is threatening the biodiversity in the lake. Due to the sudden death of

thousands of fish in the lake which happened in July 2016 and the random growth of algae since 2002 (Fadel, et al., 2014), the nitrogen cycle (transition from ammonium to nitrite and nitrate) that controls the levels of DO and pH is becoming very disorganised, and the orthophosphate levels are varying rapidly and arbitrarily. Also, when dead bodies reach the surface of the water, the released sulphur oxidises into sulfate then the concentration of sulfate in the lake increases with massive fish losses that are happening often in Qaraoun lake. The other parameters (TDS, conductivity, and salinity) are a result of the debris and eroded rocks and minerals that end up in the lake due to rainfalls, these parameters follow a rhythmic pattern where they increase during the wet season and decrease in the dry season.

3.5.5. Reconstruction of the phase space

Recall that the chaos theory is a method to analyze nonlinear time series based on reconstructing the phase space of a process, originally introduced by Takens (Takens, 1981). In this section we are interested in constructing the phase space of nonlinear physicochemical parameters. One strategy to evolve the time series of physicochemical parameters is to apply the theory of chaos using the "phase-space diagram." Note that a point in the phase space represents the system dynamics on a given instant. The system dynamics of each physicochemical parameter is represented geometrically by a single point moving on a trajectory where each of its points corresponds to a state of the system. Phase space is viewed as a coordinate system, the coordinates of which represent the variables necessary to fully describe the state of the system at all times. The utility of phase space reconstruction of physicochemical parameters is to show that a nonlinear system can be characterized by self-interaction, and when a time series of a single physicochemical parameter can provide information on the dynamics of the whole multivariate system and this may be due to the correlation between the different physicochemical parameters which varies according to the pollution of Lake Qaraoun. We notice that the phase space diagram presents attractors in well-defined regions, with some dispersion of the parameters, which suggests that the dynamic properties of the system can be explained by deterministic chaos instead of a stochastic model. Note that the correlation functions and the exponents are calculated using the Grassberger Procaccia algorithm. When selecting the delay time for the phase space reconstruction, the autocorrelation function is calculated for different delay times for each of the physicochemical parameters and which is chosen as the delay time at which the autocorrelation function exceeds the zero line (Holyst, et al., 1996).

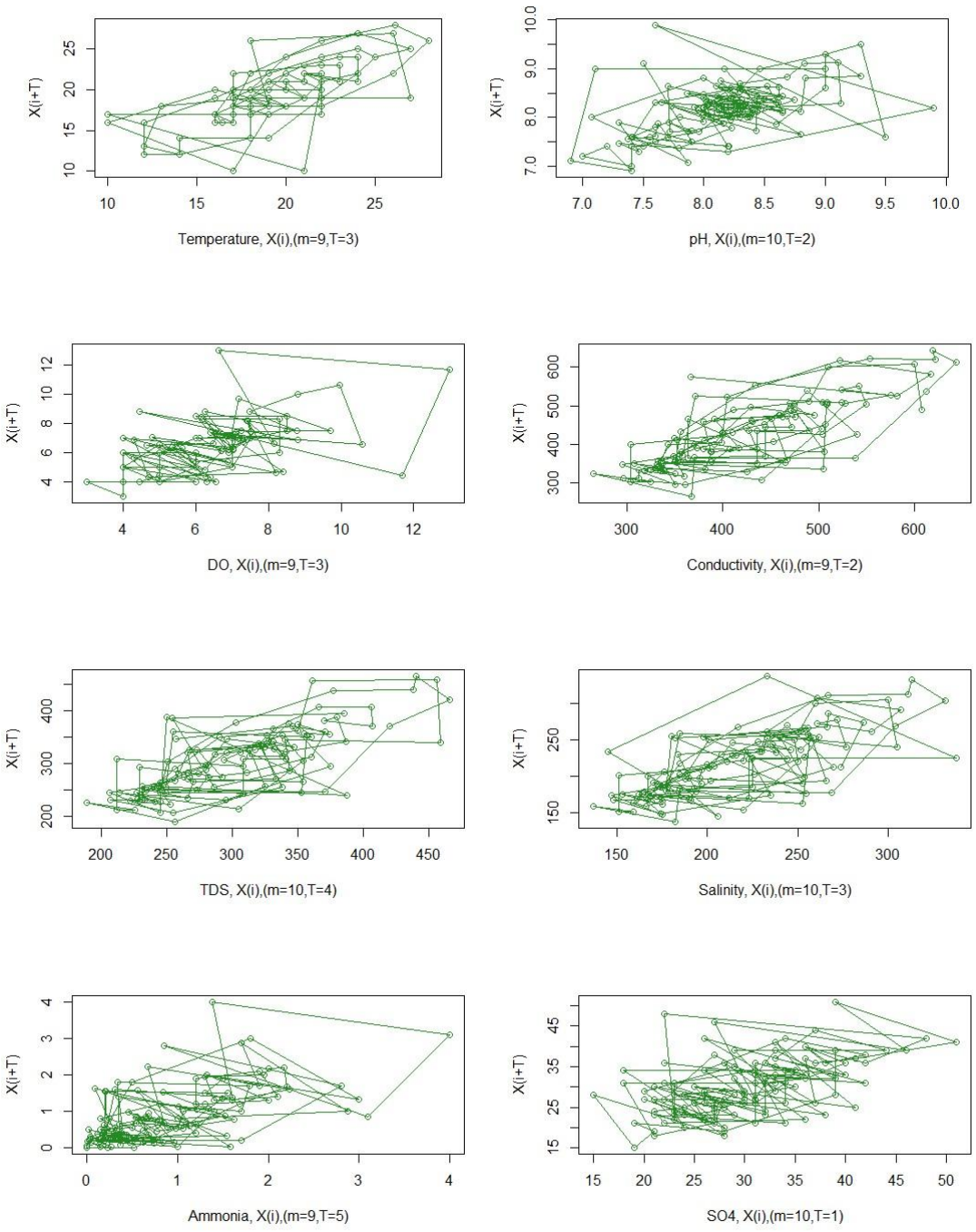


Figure 35 Trajectories in two-dimensional phase space by obtained time lag from AMI.

Figure 35 shows the two-dimensional phase space reconstruction for the time series of physicochemical parameters at Lake Qaraoun. These graphs correspond to different lag-times for the different physicochemical parameters according to table 12, which means that the projection of the attractor differs from one parameter to another; for example, we consider $\{X_i, X_{i+1}\}$ for a delay time $T = 1$. Based on the above figures, chaos theory allows us to explain the presence of noise in the data (scattered band), while the narrow dark band means the existence of a strong determinism which means the convergence of the trajectories of the phase space with a fractal dimension towards the data

attractor. In addition, we notice that the reconstruction of the phase space gives reasonably well-defined attractors for the different physicochemical parameters, although the attractor of the series of temperature, pH, Conductivity, TDS, sulfate is much better than those of the other series, where very few outliers correspond to very high values are clearly evident. Finally, the dynamics of the various physicochemical parameters can be reduced to a set of deterministic laws which allow us to predict its future states.

3.6. Conclusion

This study aimed to seek better understanding of the dynamics of the time series of the physicochemical parameters at Qaraoun Lake in Lebanon using a new approach based on chaotic processes. We studied the 11 physicochemical parameters over a period of 11 years (2008-2018). The correlation dimension method was used to detect the existence of chaos. The estimated dimensions of the parameters indicate the possibility of the existence of chaos in the time series of the following parameters: temperature, pH, DO, ammonium, nitrite, nitrate, sulfate, and orthophosphate. TDS, Salinity and conductivity shows a stationary behavior without any chaotic effect. This present chapter was the first study to understand the behavior of time series of physicochemical parameters at Qaraoun Lake in Lebanon using chaos theory. The characterization of the individual dynamic behavior (chaotic or stochastic) of the 11 physicochemical parameters was treated. For this purpose, the monthly data observed in Qaraoun Lake were analysed using the method of correlation dimensions. The parameters that showed chaotic behavior are mainly the parameters that are connected to the biodiversity and the marine life in the lake. This is a clear sign that pollution is altering the ecosystem and menacing its order. The presence of chaotic behavior could provide interesting results, for a reliable prediction of the process of physicochemical parameters. Finally, it should be noted that it is interesting to see additional chaos identification methods, such as the nonlinear prediction method that can be used to complete the analysis to confirm the current results.

Chapter 4

Extreme and Records Value Analysis for Evaluating Air Quality in Bekaa Valley, Lebanon

This Chapter discusses the investigation of extreme atmospheric pollutants. This work has been published in the International Journal of Environmental Science and Development (IJESD).

4.1.Summary

Air pollution is a major public health problem that affects economic development, agriculture, and the ecosystem. Lebanon is one of the most polluted countries in the Middle East region due to the increase in the concentrations of atmospheric air pollution which exceed the required levels according to the global guidelines. The main objective of this paper is to identify the extreme concentrations of air pollutants in order to minimize their adverse effects. The peak of concentration of the pollution which is measured within a specific period could be described by using the extreme value theory mainly as one kind of the three different types of the extreme value theory and the record theory as a second kind. These two approaches will be applied to predict the expected extreme concentration in the future and the probability of occurrence of a new record. Whereas daily measurements of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ throughout the period 2016-2019 in the Bekaa Valley. The findings indicate that the concentrations of the 15, 30, 40, and 50-year return pollutant levels are continuously increasing. The percentage of the change of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ after 50 years is 64%, 5%, 3%, 29%, 20% and 12%, respectively. The records have been observed at the beginning of a time series and an interval point prediction was given for each measure. The future record values of SO_2, CO, NO, PM_{10}, $PM_{2.5}$ were increased by 0.2%, 0.7%, 1%, 0.5%, 6%, and 5.7%, respectively, over one year. There was a 66% chance to observe a new record-breaking pollutant level that exceeded the guidelines after two years.

*4.2.*Introduction

Atmospheric air quality has emerged as a significant issue in the last century and a half, with effects on people varying according to the level of pollution and duration of the exposure, from respiratory illness to heart attack to death. The effects of air pollution on humans vary depending on the level of air quality and the length of time of exposure, from respiratory problems to heart failure or death. Many

studies have been initiated to control the adverse effects of industrial pollution, deforestation, hazardous disposal, land degradation ,and other urgent environmental issues. Atmospheric air pollution has increased to an alarming level and premature mortality from exposure to air pollution in the world increases each year (WHO) (World Health Organization, 2021) . There are numerous people who are already being exposed to air pollutants, and pollution is still rising due to a number of different factors, among them the burning of fossil combustibles. The landscape in the Middle East is varied and consists of high ground, deserts and the Sahara, coastal zones, and large areas of plain lands. Climate differs across regions, it experiences strong episodes of heavy pollution, high levels of particulates, and heavy levels of acid deposits (Saliba, et al., 2007) in several regions due to large industrialized areas, lack of any effective public transportation system, dense traffic areas, and high population densities (Lelieveld, et al., 2009). Located on the Mediterranean Sea, Lebanon is a relatively compact state with an increasing population, especially in the urban areas, where there is no system of public transit (MoE) available altogether (MoE, 2020). Moreover, the country is exposed to steady winds from eastern European countries and to Saharan desert windstorms, which make Lebanon among the dirtiest countries in the region of the Middle East. This makes it among the most environmentally polluted in the Middle East region. The major environmental challenges are caused by a failure to develop appropriate policies and environmental facilities. Atmospheric air pollution from numerous different sources was identified as being one of the most pressing health issues the country confronts, particularly in urban zones (Saliba, et al., 2006,). In recent several decades, atmospheric air pollution has been considered a significant problem in Lebanon, which affects public health and the environment generally. Consequently, surveillance and analyses of the extreme pollutant concentrations will be critical in order to control environmental danger on one hand and to reduce the human exposition and the health hazards involved on the other. Lebanon's urban environment is affected by high levels of air pollutants from industrial and transportation sectors in the urban areas, as well as high levels of air quality (Abdallah, et al., 2018). Climate change, especially increased environmental temperature, is projected to severely impact biodiversity (Ochoa-Hueso, et al., 2017). The average number of deaths due to air pollution in Lebanon is noted to be 2700 cases in 2018, or an average of 4 deaths for every 10000 people. This is considered to be the largest in the Mediterranean region (A. Farrow, June 2020). Research showed that the increasing number of cars, the uncontrolled industry, and the privately operated generator which are located in the residential areas are mainly the cause of the persistence of air pollution (Farah, et al., 2014). Regional traffic is an important factor in the high concentrations of ozone, the diesel engines generate large quantities of particulate pollutants, and pollutant concentrations are above WHO guidelines in most Lebanese areas (Abdallah, et al., 2016) thus requiring emergency measures. So far, few research have been conducted to investigate air pollution in Lebanon. Consequently, atmospheric air pollution in Lebanon should be studied and controlled through statistical

prediction approaches. Numerous researchers have investigated temporal distributions of atmospheric pollution through the use of probabilistic models (Seinfeld, 1976), (Lee, et al., 2003), and statistical descriptive methods (Voigt, et al., 2004). Time series analysis methods (Marcus, 1990), (Alvim, et al., 1999) where the aim is to demonstrate trends in concentrations of the pollutants over time, and hence impact of air pollutants on health, have also been employed (Pope, 1991), (Touloumi, et al., 2004). Some other researchers used statistical approaches such as Principal Component Analysis (PCA) and Hierarchical Agglomerative Clustering (HAC) (Afif, et al., 2009). Consequently, this study, being the first in Lebanon, investigates pollutant performance through record-based analysis of pollutant concentrations extremes and the theory of the most important ones applied to air pollutant records. This article seeks an introduction to a new extreme values pattern that is more appropriate for record-based environment surveys. The theory allows efficient problem recognition of air pollutants. The presented findings may be used as an air quality control management instrument providing policymakers the means to identify actions necessary for pollution abatement and create a warning mechanism for very high peaks. Results suggest that both extremes and especially record theory are easily applied and can reach an even greater precision than other models in case of a restricted observational database. The extreme phenomena of atmospheric air quality are of special relevance. They are a risk to human beings and may also affect the ecosystems of a city, a nation, and even a planet. The Extreme Value Theory (EVT) offers a valuable framework to analyze the air pollution quality data and predict the future of extreme events while detecting the level and return period of the extremes. The main purpose of the study (EVT) is to develop strategies to reduce criteria air pollutants. Numerous research investigations have used the theory of EVT and records for such environmental purposes as extreme temperature, extreme precipitation, extreme wavelength, etc. (Hayek, et al., 2020), (Kibria, 2019). This approach is advantageous as it shows extreme values and their return period, which in turn assists with creating a system of warnings about the future. Furthermore, it is different from the other statistics approaches that look at the averages of data, as when trying to analyse the environmental data, the focus must be on extreme values that are alarming. It should be noticed that record keeping has been applied in several fields. From Climate change to the environmental sector, to finance and sports. This approach has proven to be relevant to the temporal evolution of a data set over time (Wergen, et al., 2010), (Hoayek, et al., 2017). The application of these extreme approaches to modeling the air quality will assist to model a pattern of pollution extreme or peak values, to predict future peak pollution, and identify the time of year when pollution level is the greatest. In this article, we are interested in the study of air pollutants like carbon monoxide ($CO\ mg|m^3$), nitrogen dioxide ($NO_2\ \mu g|m^3$), and sulfur dioxide ($SO_2\ \mu g|m^3$) which are now recognized as indicators of anthropogenic emissions (vehicles, industry, construction, ports, airports, road networks, etc.), and particulate matter ($PM_{10}\ \mu g|m^3$ and $PM_{2.5}\ \mu g|m^3$) which can be very harmful to human health. The study was made to encompass Bekaa

region, which is an elaborate area with a variety of human activities. The Bekaa region is witnessing a catastrophic environmental scenario affecting the air and water quality in the region, and thus the public health and the ecosystem in general. Signs of climate change such as storms, floods, and heat waves are becoming usual and periodic (Hayek, et al., 2020). These extreme events which are a result of high pollution are affecting the Lebanese economy, especially in the Bekaa region which depends primarily on agricultural activity (Khraibani, et al., 2020).

*4.3.*Background

In general, the theory of extremes corresponds to the study of the occurrence of events with a low probability whose objective is to obtain reliable estimates of the probabilities of occurrence of rare events. This theory is based on the theory of probability to study the tails of the distribution of a sequence of independently and identically distributed random variables. This paper aims to introduce a new model of extremes that is based on the theory of records to predict a new record of pollutant concentrations. Record theory is related to extreme value theory. This theory involves studying the limiting distribution of the maximum $M_n = max(X_1, \ldots, X_n)$ of a series of independent and identically distributed random variables. Record theory deals with the exact distribution of the number of records N_n and record times L_n). These two distributions do not depend on the distribution of observations, so it takes into account the temporal structure of $\{X_1, \ldots\ldots, X_n\}$ insofar as we study the instants of the L_n jumps of M_n that correspond to the instants of records.

4.3.1. Extreme Value Analysis

Studies of extremes began in the 1920s (Fisher and Tippett, 1928 (Tippett, 1928)), and developed rapidly (Gumbel, 1942 (Gumbel, 1942)); Leadbetter et al., 1983 (M. R. Leadbetter, 1983). Two main methods are developed in the study of extreme values. The first is the block maxima method based on the limit distribution of the maximums (Fisher-Tippett theorem): Assume that $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables and $M_n = \max(X_1, \ldots, X_n)$. If a sequence of pairs of real numbers (a_n, b_n) exists such that $a_n > 0$ and $\lim_{x \to +\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x)$, where G is a non degenerate distribution function. The limit distribution G belongs to one of the three distributions below called the generalized extreme value distributions (GEV) which defined respectively by Gumbel, Frechet, Weibull:

- $\Omega(x) = \exp[-\exp(-x)]$, where $x \in \mathbb{R}$
- $\phi_\theta(x) = \exp\left[-x^{-\theta}\right]$, where $x > 0$ and $\theta > 0$
- $\psi_\theta(x) = \exp[-(-x)^\theta]$, where $x \leq 0$ and $\theta > 0$.

The general form of (GEV) distribution is given by:

$$G_{\gamma,\Gamma,\sigma}(x) = G_\gamma\left(\frac{x-\Gamma}{\sigma}\right) = \exp\left[-\left(1+\gamma\left(\frac{x-\Gamma}{\sigma}\right)\right)^{-\frac{1}{\gamma}}\right] \quad (1)$$

$1+\gamma\left(\frac{x-\Gamma}{\sigma}\right) > 0, \gamma \in \mathbb{R}, \Gamma \in \mathbb{R}, \sigma > 0$, Γ the location parameter, σ the scale parameter and γ the shape parameter.

The second approach is the peak over threshold (POT) method, where we select a threshold u, and every value exceeding this threshold is considered an extreme value.

The extreme value theory consists in studying the observations which exceed a threshold u. The excess Y of the variable X above the threshold is given by $X-u$ if $X > u$. The distribution function F_u of the observations above the threshold u is given for $y > 0$:

$$F_u(y) = \frac{F(u+y)-F(u)}{1-F(u)} \qquad (2)$$

For a large enough threshold u, the survival function is approximated by a Generalized Pareto Distribution (GPD), (Pickands theorem 1975 (Pickands, 1975), Balkema and de Haan 1974 (Haan, 1974)):

$$\bar{H}_{\gamma,\sigma}(y) = \begin{cases} \left(1+\gamma\frac{y}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{if } \gamma \neq 0 \\ exp\left(\frac{-y}{\sigma}\right) & \text{if } \gamma = 0 \end{cases}$$

The convergence between the distribution of the maximum to a (GEV) and that of the threshold exceedances to a (GPD) :

$$\lim_{n\to\infty} P\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right) = G_\gamma(x)$$

$$\rightarrow \lim_{u\to x_F} \sup_{[0,x_F-u]} \left|\bar{F}_u(y) - \bar{H}_{\gamma,\sigma(y)}(y)\right| = 0 \qquad (3)$$

where x_F is the endpoint of the cumulative distribution function F, $X_{n,n} = Max(X_1,\dots,X_n), a_n > 0$ and $b_n \in \mathbb{R}$, $\bar{F}_u$: survival function. Both methods provide the $T-year$ return period of a given variable, with a probability of $1/T$.

4.3.2. Threshold Selection

There is no precise method for choosing the threshold. In general, the threshold chosen should be small so that the number of observations that exceed the threshold provides an accurate estimate of the model parameters. In the literature, the "Mean Excess Function" (MEF) is used to choose the threshold u (Resnick, 2010).

$$MEF(u) = E(X-u|X>u),$$

If the adjustment of exceedance with a valid (GPD) with a certain threshold u_0 and for $\gamma < 1$ then:

$$E(X-u_0|X>u_0)=\frac{\sigma_{u_0}}{1-\gamma}$$

For every $u>u_0$:

$$E(X-u|X>u)=\frac{\sigma_u}{1-\gamma}=\frac{\gamma}{1-\gamma}u+\frac{\sigma_{u_0}-\gamma u_0}{1-\gamma}$$

MEF is a linear function in terms of u. The smallest value of u is given such that the (MEF) becomes linear for $\gamma<1$.

4.3.3. Probability and Return Period

The probability of occurrence of pollutant concentrations exceeding a given threshold as well as the return period is calculated based on the limits recommended by the World Health Organization (WHO, 2021). The return period is represented by the time of occurrence of high peaks of pollutant concentrations that exceed the thresholds in the station "Memshieh Garden". In the following, the probability and the return period are calculated in both cases (GEV) and (BM).

$$p=1-\exp\left\{-\left[1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}}\right\} \qquad (4)$$

0 p is the probability and $T=1/p$ is the return period.

Z_p the return level of the $1/p$ observation. will be exceeded every $1/p$ observations.

$$Z_p=F^{-1}(p)$$

$F^{-1}(\mathrm{p})$ is the approximation of the tail survival function:

$$F^{-1}(p)\approx\begin{cases}u+\dfrac{\sigma}{\gamma}\left(\left[\dfrac{p}{F(u)}\right]^{-\gamma}-1\right) & \text{if } \gamma\neq 0\\ u-\sigma\log\left(\dfrac{p}{F(u)}\right) & \text{if } \gamma=0\end{cases}$$

Where p is small enough in order to have $Z_p>u$.

*4.4.*Records Theory

Records theory began in (1952) with Chalender (Chandler, 1952) who introduced a basic properties of records. The main results were given in the period 1952-1983. Assume a sequence of observations $(X_n)_{n\geq 1}$ independent and identically distributed random variables observed successively from a cumulative distribution function F. The record process is composed by (R_n, L_n). The record-breaking index $\{L_n, n\geq 0\}$ is defined by:

$$L_0=1 \text{ with probability } 1$$

for $n>1$,

$$L_n = min\{j: X_j > X_{L_{n-1}}\}$$

The record value $\{R_n\}$ is then defined by:

$$R_n = X_{L_n}, n = 0,1,2,\ldots$$

R_0 is the record trivial. The number of records N_n is defined by:

$$N_n = \{\text{number of records among } X_1, \ldots, X_n\}.$$

Where $N_1 = 1$, X_1 is a trivial record. The number of record is defined by: $N_n = \sum_{i=1}^{n} \delta_i$; with δ_i is a sequence of record indicator:

$$\delta_i = \begin{cases} 1 & \text{if i} = 1 \\ \mathbb{I}\{X_i > max(X_1, \ldots, X_{i-1})\} & \text{if i} > 1 \end{cases}$$

By symmetry the δ_i has a Bernoulli distribution, $Ber\left(\frac{1}{i}\right)$.

4.4.1. Extreme Records Value

It is necessary to know the distribution of extreme values to study the records.. Based on the subsection A, the (GEV) distribution is defined by:

$$F(x) = 1 - exp\left[-e^{\left(\frac{x-\mu}{\sigma}\right)}\right],$$

where $\mu \in \mathbb{R}$, $\sigma > 0$. The upper records values be observed from an extreme value distribution with density function:

$$f(x, \mu, \sigma) = \frac{1}{\sigma} e^{(x-\mu)/\sigma} e^{-e^{(x-\mu)/\sigma}}, x \in \mathbb{R} \quad (5)$$

For such a random variable we have

$$X \overset{d}{=} \mu + \sigma \log X^*,$$

where $X^* \sim Exp(1)$. The corresponding record value sequence can be described by

$$R_n \overset{d}{=} \mu + \sigma \log\left(\sum_{i=0}^{n} X_i^*\right).$$

Arnold (Arnold, et al., 2011) give the expected value and the variance of R_n:

$$E(R_n) = \mu - \sigma\gamma + \sigma \sum_{j=1}^{n} \frac{1}{j},$$

$$V(R_n) = \sigma^2 \left(\frac{\Pi^2}{6} - \sum_{j=1}^{n} \frac{1}{j^2}\right).$$

The expected value and the variance of the number of records are given by:

$$E(N_n) = \sum_{j=1}^{n} \frac{1}{j} \approx \ln(n) + \gamma, \qquad (6)$$

$$V(N_n) = \sum_{j=1}^{n} \frac{1}{j}\left(1 - \frac{1}{j}\right) \approx \ln(n) + \gamma - \frac{\pi^2}{6}. \qquad (7)$$

where γ is Euler's constant 0.5772. The joint probability distribution, $f(r_0, \ldots, r_n)$ of the record values $R_0, R_1, \ldots, R_n$ from a continuous cumulative distribution function $F(r)$, is defined by:

$$f_{R_0,\ldots,R_n}(r_0, \ldots, r_n) = f(r_n) \prod_{i=0}^{n-1} h(r_i),$$

$$for \ -\infty < r_0 < r_1 < \ldots < r_n \qquad (8)$$

where $h(r) = f(r)/1 - F(r)$ is the hazard rate function.

*4.5.*Records Inference

In this section, the statistical inference, point estimation, interval estimation, and prediction of air pollutant concentrations are discussed. The maximum likelihood estimation (MLE) of the distribution parameters is discussed. In addition, the best linear unbiased parameter estimate is shown. The point prediction of the future records in which the best linear unbiased prediction and the best linear invariant prediction are described.

4.5.1. Maximum Likelihood Estimation

Assume that $R_0, R_1, \ldots, R_n$ are the upper record values observed from any location-scale distribution with cumulative distribution function $F(x, \mu, \sigma) = F\left(\frac{x-\mu}{\sigma}\right)$.

The joint probability distribution of $R_0, R_1, \ldots, R_n$ is:

$$f(r_0, r_1, \ldots, r_n; \mu, \sigma) = \frac{1}{\sigma^{n+1}} f\left(\frac{r_n - \mu}{\sigma}\right) \prod_{i=0}^{n-1} \left[\frac{f\left(\frac{r_i-\mu}{\sigma}\right)}{1-F\left(\frac{r_i-\mu}{\sigma}\right)}\right] \qquad (9)$$

From equation (8), the likelihood function is given by:

$$L = \frac{1}{\sigma^{n+1}} \prod_{i=0}^{n-1} \left(\frac{f(r_i^*)}{1-F(r_i^*)}\right) f(r_n^*), \qquad (10)$$

$$-\infty < r_0^* < r_1^* < \ldots < r_n^* < \infty$$

where $r_i^* = \frac{r_i - \mu}{\sigma}$ for $i = 0,1,\ldots,n$.

The log-likelihood function is obtained from equation (10) (Shing, 1998):

$$\log L = -(n+1)\log\sigma - \sum_{i=0}^{n-1} \log(1 - F(r_i^*)) + \sum_{i=0}^{n} \log f(r_i^*), \qquad (11)$$

The Best Linear Unbiased Estimator (BLUE) of μ and σ is given by (DAVID, 1981) and (Balakrishnan, 1991), (Cohen, 1991):

$$\mu^* = R_n - \frac{\alpha_n}{n}\sum_{i=0}^{n-1}(R_n - R_i) \qquad (12)$$

$$\sigma_n^* = \frac{1}{n}\sum_{i=0}^{n-1}(R_n - R_i) \qquad (13)$$

where $\alpha_n = -\gamma + \sum_{i=0}^{n-1}\frac{1}{i}$ with γ is Euler's constant.

4.5.2. Future Records Prediction

We focus on the prediction of record values of pollutants to evaluate the extreme records. We first introduce the point prediction of a future record and then present the conditional prediction intervals for future records. Based on the BLUE's defined above (μ^*, σ^*), the Best Linear Unbiased Predictor (BLUP) R_{n+1}^* of the records values $R_n$in global case do to Goldberger (Goldberger, 1962):

$$R_{n+1}^* = \mu^* + \alpha_m\sigma^* + \omega^T\Sigma^{-1}(R - \mu^*\mathbf{1} - \sigma^*\alpha) \qquad (14)$$

where R is the vector of observed record values, **1** is a vector of **1**'s, α is the vector of means records values, Σ is the variance-covariance matrix of the standard record values and $\omega^T = (\sigma_{0,n}, \sigma_{1,n}, \ldots, \sigma_{n,m})$. The BLUE of σ_n^* is given by the equation (13). Similarly, the BLUE of σ based on $(R_0, R_1, \ldots, R_{n+1})$is given by:

$$\sigma_{n+1}^* = \frac{1}{n+1}\sum_{i=0}^{n}(R_{n+1} - R_i) \qquad (15)$$

From the two equations (13), and (15), we get the point prediction of the future record

$$\begin{aligned} R_{n+1}^* &= R_n - \frac{1}{n}\sum_{i=0}^{n-1} R_i + \frac{1}{n+1}\sum_{i=0}^{n} R_i \\ &= R_n + \frac{1}{n(n+1)}\sum_{i=0}^{n-1}(R_n - R_i). \end{aligned} \qquad (16)$$

Let us take $EV(\mu, \sigma)$ distribution considered in section E, the $100(1-\alpha)\%$ conditional prediction interval for R_{n+1} is given by Chan (Shing, 1998):

$$[R_n + \sigma^* z_{3,\alpha/2}, R_n + \sigma^* z_{3,1-\alpha/2}] \qquad (17)$$

where $z_{3,\alpha}$ the $\alpha - th$ quantile of the conditional distribution:

$$z_{3,\alpha} = \left(\frac{n}{n+1}\right)(1-\alpha)^{-1/n} - 1 + \frac{1}{n+1}.$$

By (Chandler, 1952), the waiting time T_2^* to observe a new record increases proportionally to the number of records. The probability to observe a new record in the coming n_2 years given observations for n_1 years is defined by:

$$P(T_2^* > n_2) = \frac{n_1}{n_1+n_2} \qquad (18)$$

5. Surveillance Station and Data Collection

The Environmental Resources Monitoring in Lebanon (ERML) project began in 2013 with the

support of the United Nations Environment Program (UNEP) and the United Nations Development Program (UNDP). Under the project, the Lebanese Ministry of Environment has installed Air Quality Surveillance Stations (AQMS) in various sites in major Lebanese cities (Hadath, Beirut, Baalbeck, Zahleh, Saida), it was the first real-time air monitoring system aimed at tracking the exposure of the population from industries, electric plants, and traffic on the road in addition to urban sources of air pollutants. The monitoring stations include web-based analyzers for monitoring criteria pollutants: $SO_2\ \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, $PM_{2.5}\ \mu g|m^3$. These settings were recorded hourly between 01/01/2016 and 19/06/2019 in the station "Memshieh Garden - Zahleh" located in Bekaa valley at an altitude of 1150 m with a latitude of $33°\ 51'3.01"$ N(34,270)and a longitude of $35°53'45.54"$ E (Figure 36). The dataset is composed of a matrix of size (30370 × 6). It should be emphasized that this study is made on one specific site which is located close to Litani River, therefore results may differ from other stations or data of different time periods.

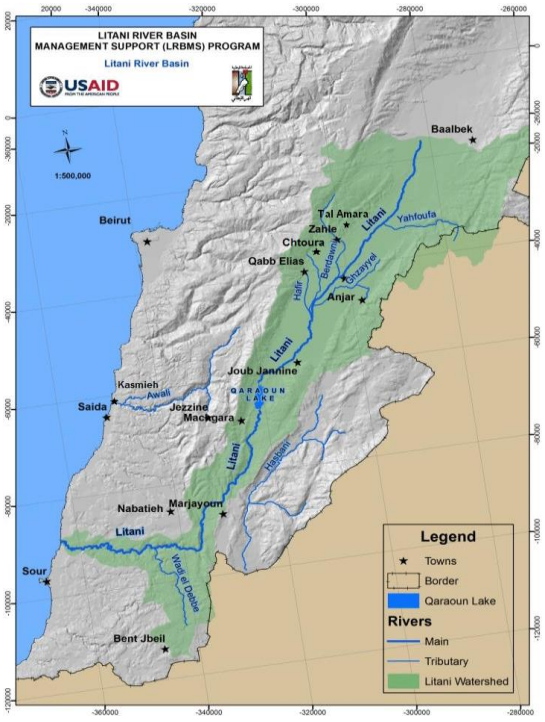


Figure 36 Litani River and Memshieh -Zahleh station.

Within the framework of the monitoring of environmental resources in Lebanon the monitoring of air quality, through the design and implementation of a national air quality monitoring network several monitoring stations are settled in Lebanon. Figure 36 shows the Memshieh Garden station, the choice of this station to study the air quality in the Bekaa region which is characterized by agriculture and tourism and its geographical location. In addition, there are no previous studies to study the air quality and the extreme concentrations of these measurements in this region.

6. Results and Discussion

The following section presents an application of the extreme values theory and the records theory on the complete data of six pollution parameters using the statistical software “R”. The principal objective is to predict the extreme quantiles and calculate the return level of each pollution parameter to forecast future levels of atmospheric pollution.

*6.1.*The Data

Data collected from the Lebanese Ministry from 2016 to 2019 are registered on each hour (MoE: National Environmental Action Plan, 2020). The initial sample size equals n=30370 observations. We encountered missing data (NA). This was caused due to either technical issues, the monitoring station being out of order for many months, or mistakes in data input. The measurements are registered hourly and data is validated on a daily, a monthly, and a yearly basis as a final validation step. The data validation is generally carried out with the use of software that automatically checks the data transmission and the possibility of a technical problem. If there is a reported issue, a service technician will check the monitoring sites and either replace or correct the monitoring equipment. Nevertheless, some data gaps must be completed for use in the survey. To address the missing data problem, we use the R software package "missForest" which provides an imputing algorithm for the missing value that is random forest (RF) method based (Bühlmann, 2012). We use (RF) method since it can deal with high-dimensional datasets and provides robust results. It is a nonparametric approach, meaning it can deal with complex and non-linear data interaction and does not consider probability distributions of data. In addition, (RF) is able to approximate the out-of-bag (OOB) error without requiring a testing package. An explanation for (OOB) is described in the article (Gareth, 2013). Some discrepancy is registered for the interquartile range, the median, and the mean and also for the measure of kurtosis. The OOB error rate is 3.46%. Nevertheless, in this article, we are interested in the study of extreme pollutants. The maximum 24-hour maximum was used for the air pollutant. The maximum sampling size to which we are applying the methods of the extremes and the records is n=1265 observations. A summary of missing data is shown in the table that follows.

Table 14 represents the missing value for each parameter of the initial data set. The percentage rate of missing values in the entire data set is 14.013%. This percentage is relatively small compared to the available data, but it varies from one parameter to another, we notice the presence of high percentages for SO_2, PM_{10} and $PM_{2.5}$. To avoid bias in the analysis processes, the data are supplemented by using the "missForest" algorithm. We proceed in the following with some classical descriptive statistics to get information about the air quality parameters. It should be noted that the presence of high variability in the air quality parameters affects some descriptive statistics (Alvim, et al., 1999). The following table is the summary of the new dataset produced by the " missForest " algorithm. The descriptions of the maximum daily observations are given in the table that follows:

Table 14 Number and percentage of missing values for each parameters

Parameters	$SO_2\ \mu g\|m^3$	$NO\ \mu g\|m^3$	$NO_2\ \mu g\|m^3$	$CO\ mg\|m^3$	$PM_{10}\ \mu g\|m^3$	$PM_{2.5}\ \mu g\|m^3$
NA	7765	2631	2398	589	6593	6597
%NA	22.57%	7.66%	6.9%	7.52%	19.71%	19.72%

Parameters	$SO_2\ \mu g\|m^3$	$NO\ \mu g\|m^3$	$NO_2\ \mu g\|m^3$	$CO\ mg\|m^3$	$PM_{10}\ \mu g\|m^3$	$PM_{2.5}\ \mu g\|m^3$
Min	0.19	2.02	11.88	0	0	0
1st Qu.	5.836	20.3	55.29	1.72	43.52	33.42
Median	12.22	46.78	77.07	3.44	53.6	43.10
Mean	19.18	67.36	81.84	35.86	84.81	69.32
3st Qu.	22.25	93.67	100.9	76.33	88.10	72.10
Max.	318.58	493.64	283.0	313.120	1553	1454.8
C.V.	149.0	94.32	41.08	133.89	132.14	144.82

Table 15 Descriptive Statistics for the Maximum Hourly during the 24 Hours for Air Pollution

The coefficient of variation (C.V.) is a measure of variation over time (Lee, 2002), which allows us a comparison of pollutants in which $SO_2 \mu g|m^3$ has the highest variation over time while NO_2 has the lowest variation. The variations of these parameters can be due to several factors a meteorological conditions $SO_2\ \mu g|m^3 < PM_{2.5} \mu g|m^3 < PM_{10} \mu g|m^3 < CO\ mg|m^3 < NO\ \mu g|m^3 < NO_2 \mu g|m^3$. The values Table 15 represent the statistical description of the maximum values of each 24h over three years of hourly measurements. With respect to WHO guidelines, the yearly average values of all the pollutants exceed the permissible limits, with the exception of $SO_2\ \mu g|m^3$, which is still within the permissible range even though it is very close to the upper limit.

Parameters	$SO_2\ \mu g\|m^3$	$NO\ \mu g\|m^3$	$NO_2\ \mu g\|m^3$	$CO\ mg\|m^3$	$PM_{10}\ \mu g\|m^3$	$PM_{2.5}\ \mu g\|m^3$
SO_2	1					
NO	0.23	1				
NO_2	0.218	0.67	1			
CO	0.43	-0.24	-0.25	1		
PM_{10}	0.29	0.35	0.41	-0.025	1	
$PM_{2.5}$	0.23	0.39	0.44	-0.09	0.91	1

Table 16 Correlation Coefficient between the Different Pollutants

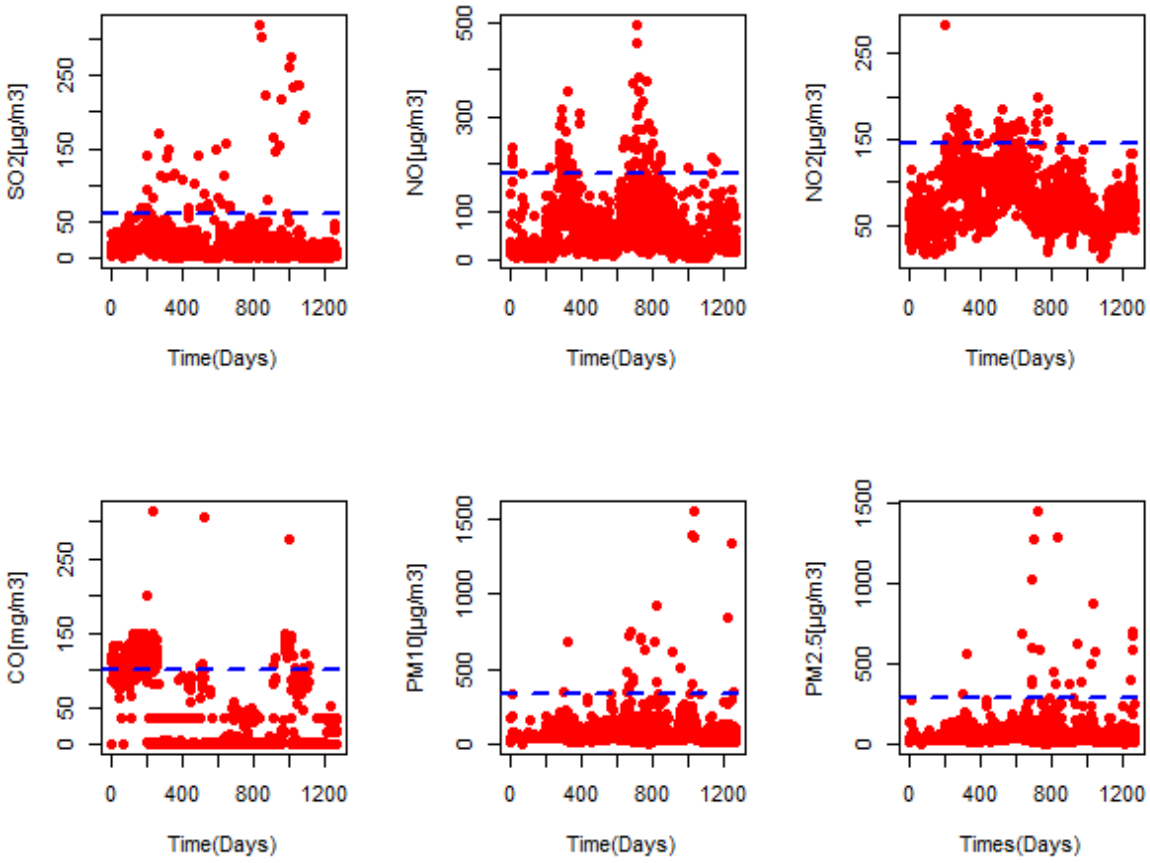


Figure 37 Maximum hourly per day for air pollution parameters in Memshieh garden station.

Also we present in Table 16 the matrix of correlation coefficients between air pollutants parameters, which shows us that most of the air pollutants are positively correlated with the exception correlation coefficient between $PM_{10}\ \mu g|m^3$ and respectively $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$and $PM_{2.5}\mu g|m^3$. The positive correlation indicates coincident time fluctuations of air pollutants. The higher positive correlations between $NO\mu g|m^3$, $NO_2\mu g|m^3$, $CO\ mg|m^3$ were detected.

In Figure 37, the time series for the maximum air pollution concentration recorded per day between 2016 and 2019 is presented. The concentrations of both $CO\ mg|m^3$ and $NO\ \mu g|m^3$ seem to trend. Also

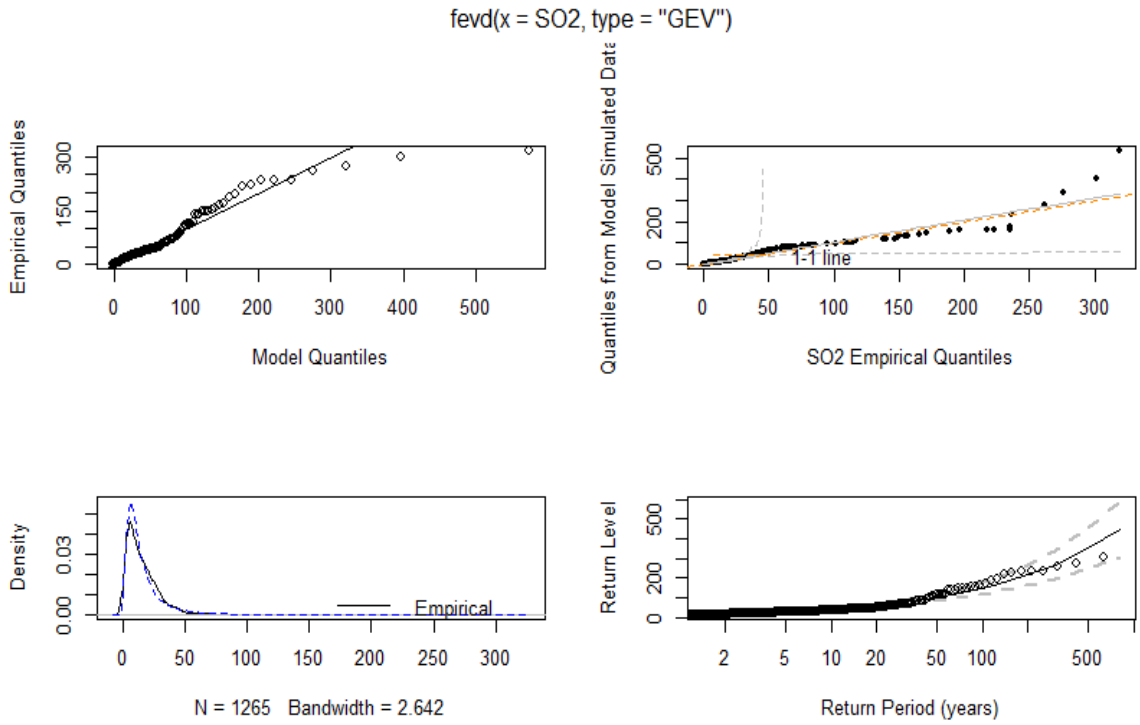
evident is the fact that $SO_2\ \mu g|m^3$ and $NO\ \mu g|m^3$ do not have a consistent significant time trend. The peak $PM_{2.5}\ \mu g|m^3$ shows a trend of increase with respect to time. The blue dashed line is the threshold u, which determines the "average exceedance plot". In the following section, we explore the pollutant exceedance concentration over this threshold by using the (POT) approach.

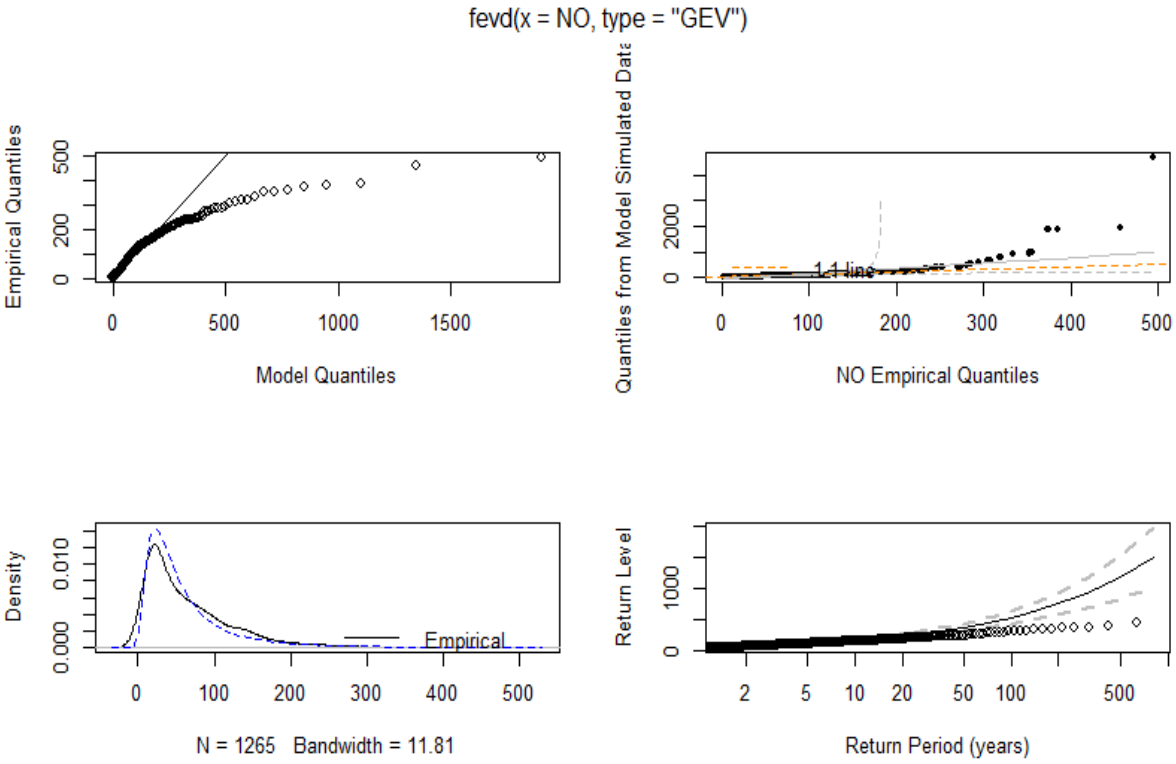
*6.2.*Univariate Extreme Values

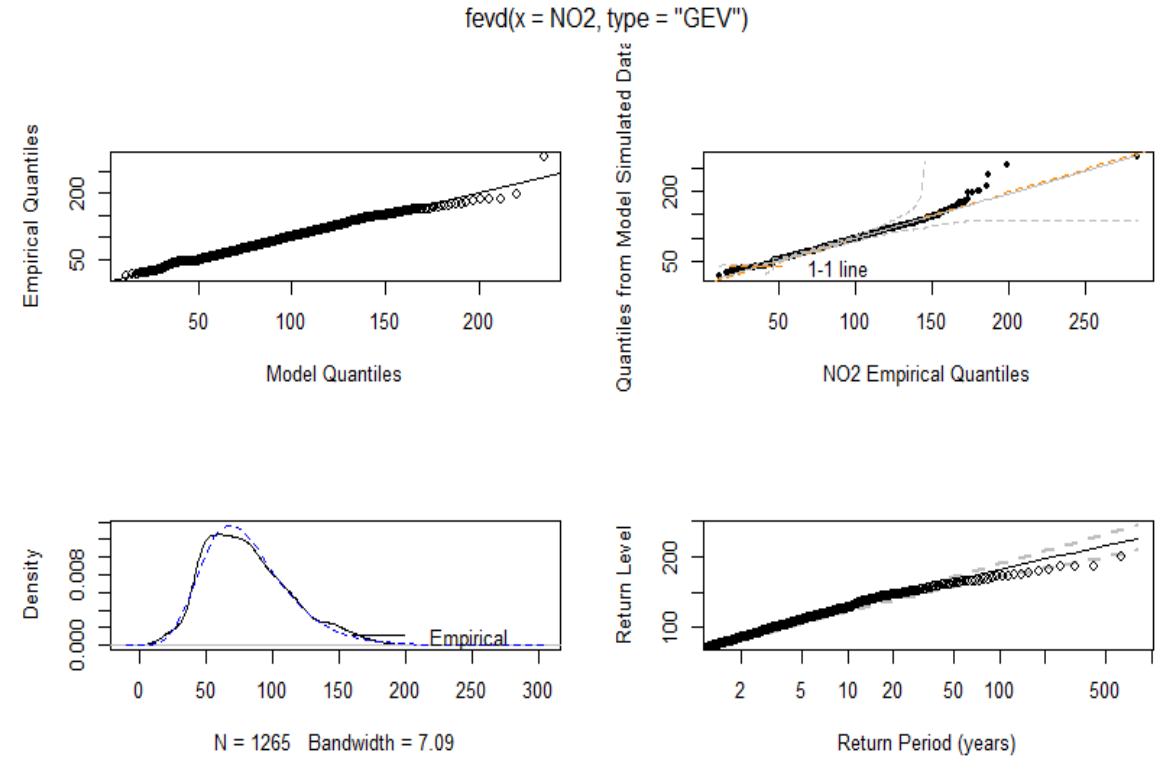
This subsection describes the theory of extreme values for the air quality monitoring data. In order to adjust the extreme value distribution, in general, two methods are provided: Block Maxima (BM) and the Peak Over Threshold (POT). The statistical results were carried out with the R software.

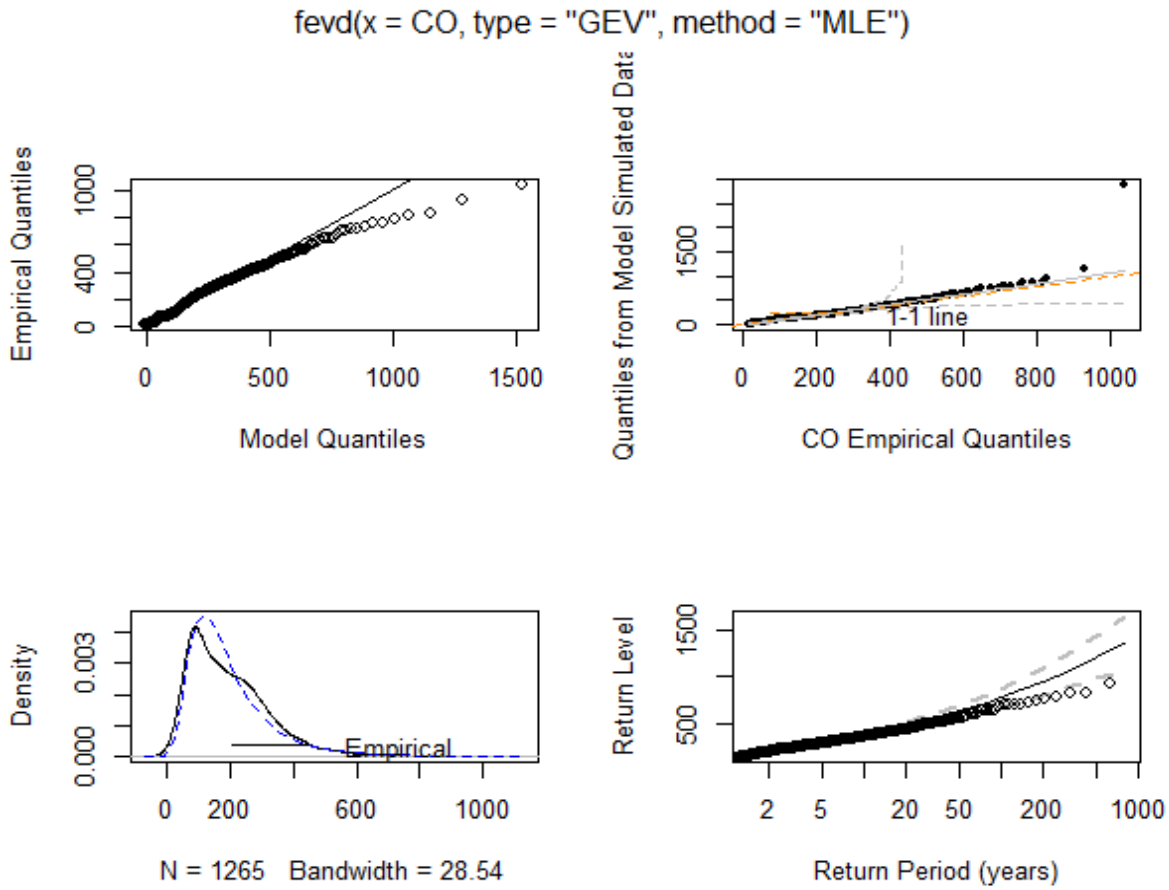
The (BM) involves breaking down data into a number of blocks (monthly blocks) and selecting the peak of each one to find a distribution that fit the peak of the data.

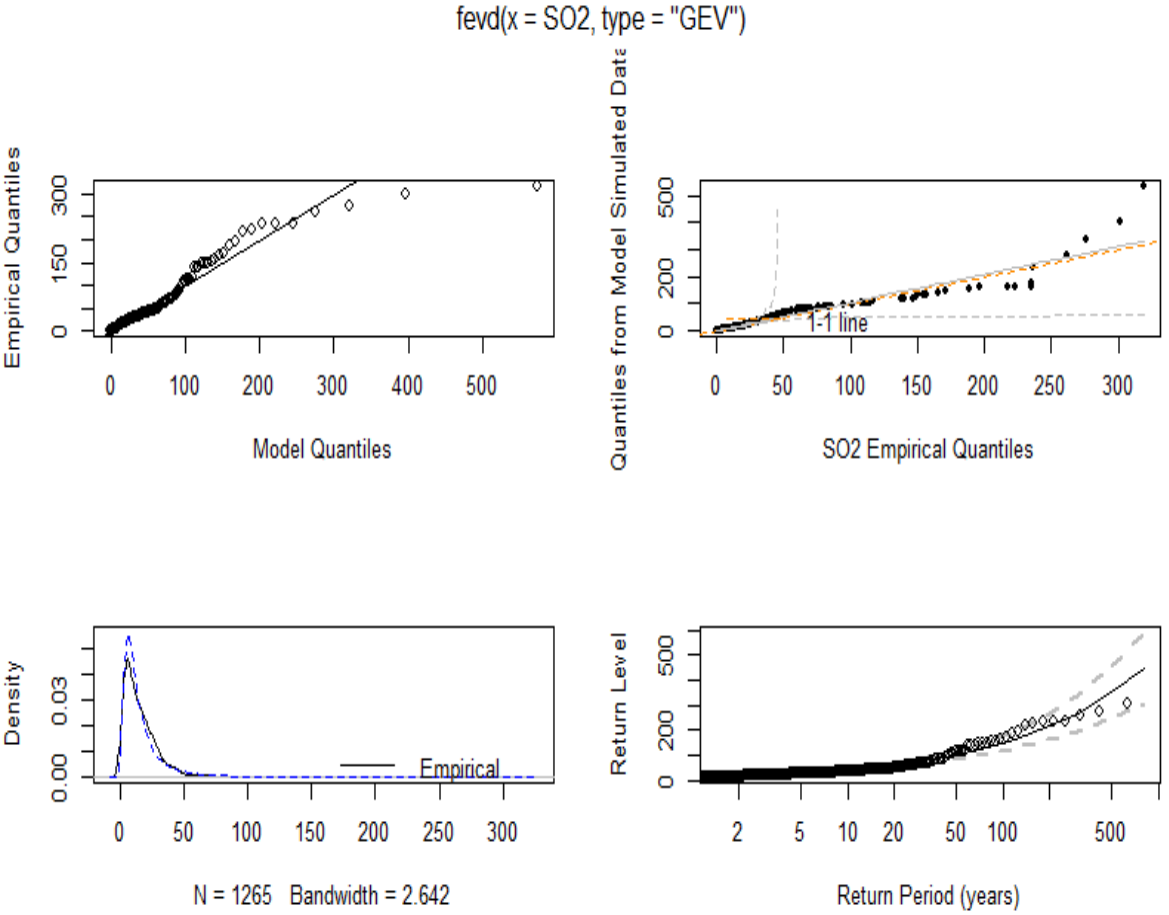
The distribution adjusted (GEV) (equation (1)) for the data is shown as follows:




fevd(x = NO, type = "GEV")
Empirical Quantiles
Model Quantiles
Quantiles from Model Simulated Data
NO Empirical Quantiles
1-1 line
Density
Empirical
N = 1265 Bandwidth = 11.81
Return Level
Return Period (years)


fevd(x = NO2, type = "GEV")
Empirical Quantiles
Model Quantiles
Quantiles from Model Simulated Data
NO2 Empirical Quantiles
1-1 line
Density
Empirical
N = 1265 Bandwidth = 7.09
Return Level
Return Period (years)


fevd(x = CO, type = "GEV", method = "MLE")
Empirical Quantiles
Model Quantiles
Quantiles from Model Simulated Data
CO Empirical Quantiles
1-1 line
Density
Empirical
N = 1265 Bandwidth = 28.54
Return Level
Return Period (years)


fevd(x = SO2, type = "GEV")
Empirical Quantiles
Model Quantiles
Quantiles from Model Simulated Data
SO2 Empirical Quantiles
1-1 line
Density
Empirical
N = 1265 Bandwidth = 2.642
Return Level
Return Period (years)

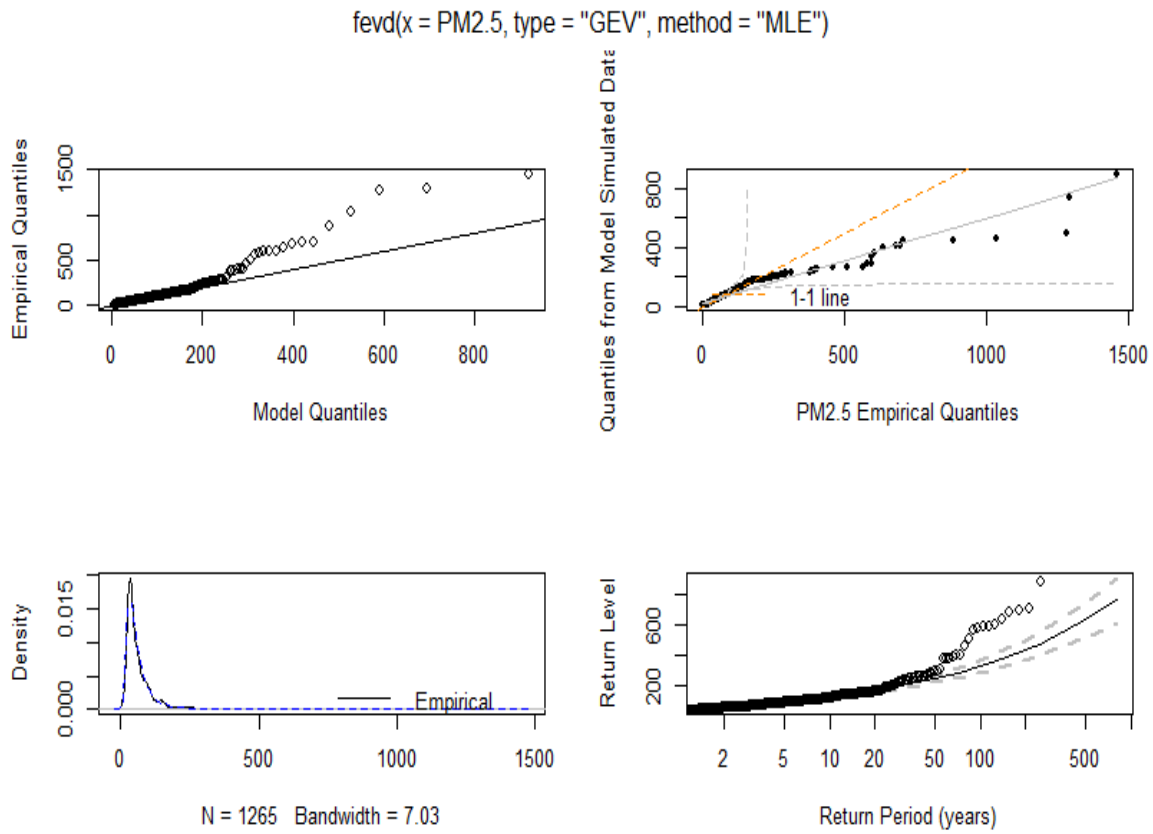


Figure 38 Diagnostic plot for the fitted GEV model for each maximum pollutant.

We note that the probability plot has a linear pattern (close to straight line) and that the two curves of the estimated and the empirical densities seems to be coherent and fitting to each other, indicating that the adjusted GEV gives a reasonable fit to data.

However the quantiles plot exhibits some deviation from the linearity, which is probably related to the higher uncertainty level. Furthermore, empirical estimations in the return level graph is close to the model -based linear line, which is nearly straight line with about 95% of confidence interval level, as the estimated form of distribution parameters (GEV) is near to zero for pollutants concentration. It should also be noted that, although the estimates of the return levels appear to be convincing, the biases in confidence intervals for large return periods also indicate that uncertainty affects this model to high levels. As illustrated inFigure 38, the diagnostics for the maximum hours per day of pollutant concentrations shows that they both yield a strong good fit to a Fréchet distribution that has a positive shape parameter γ. The observations were generated from a (GEV) distribution. The estimation of this parameter can be carried out with two distinct approaches: the method of maximum likelihood and the weighted moments method (Ana, 2007). The parameters of location, scale and shape of the (GEV) were estimated from the data. The table of the parameters estimates for each concentration of the pollutant :

From the Table 17, the probability distribution of the pollutants as a function of the form parameter (γ) sign is specified. All the parameters (γ) are positive, thus the (GEV) tends towards the Frechet distribution, with the exception that for $NO_2\ \mu g|m^3$ where (γ) is negative, and it shows that Weibull is the best

distribution to adjust $NO_2\ \mu g|m^3$. Now let us verify different future return values for pollutant concentration in next 15, 30, 40 and 50 years.

Variables	Parameters	Estimate (Std error)	95% C.I.
$SO_2 \mu g|m^3$	Location(μ)	8.42 (0.24)	[7.94 ; 8.90]
	Scale(σ)	7.32 (0.23)	[6.86 ; 7.79]
	Shape(γ)	0.52 (0.032)	[0.45 ; 0.58]
$NO \mu g|m^3$	Location(μ)	33.15 (0.95)	[31.28 ; 35.02]
	Scale(σ)	28.36 (0.9)	[26.58 ; 30.14]
	Shape(γ)	0.49 (0.033)	[0.42 ; 0.55]
$NO_2 \mu g|m^3$	Location(μ)	67.24 (0.84)	[65.58 ; 68.9]
	Scale(σ)	26.98 (0.61)	[25.78 ; 28.18]
	Shape(γ)	-0.03 (0.019)	[-0.08 ; -0.005]
$CO\ mg|m^3$	Location(μ)	3.36 (0.15)	[3.06 ; 3.65]
	Scale(σ)	5.04 (0.27)	[4.51 ; 5.57]
	Shape(γ)	1.41 (0.037)	[1.34 ; 1.48]
$PM_{10} \mu g|m^3$	Location(μ)	0.36 (0.017)	[0.33 ; 0.4]
	Scale(σ)	0.58 (0.033)	[0.52 ; 0.65]
	Shape(γ)	1.52 (0.039)	[1.44 ; 1.6]
$PM_{2.5} \mu g|m^3$	Location(μ)	5 (0.08)	[4.84 ; 5.16]
	Scale(σ)	2.67 (0.07)	[2.53 ; 2.81]
	Shape(γ)	0.34 (0.017)	[0.31 ; 0.38]

Table 17 Estimated Parameters of the (GEV) Distribution for Each Pollutant Concentration.

*6.3.*Return Level

We noted that data are broken into different daily blocks which are modeled by the (GEV) distribution. The accuracy of the return level estimates generated from the (GEV) distribution is investigated. The return levels may be valuable for evaluating the trends of the pollutant parameters and for predicting their future pollutant extremes. However, the extreme quantiles of order 1-q of the maximum distribution over a given period of time is of particular concern in this context. Our interest is to find x_q where (Hayek, et al., 2020):

$$P(Max(X_1,\ldots,X_n) \leq x) \approx G_{\mu,\gamma,\sigma}(x_q) = 1 - q$$

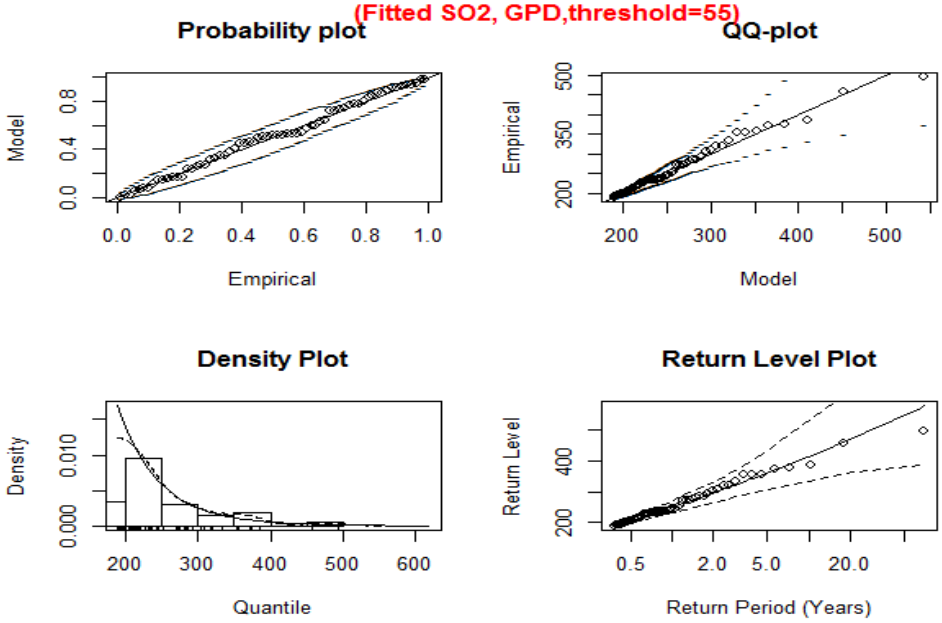
		$SO_2\ \mu g\|m^3$	$NO\ \mu g\|m^3$	$NO_2\ \mu g\|m^3$	$CO\ mg\|m^3$	$PM_{10}\ \mu g\|m^3$	$PM_{2.5}\ \mu g\|m^3$
	15	50.9	190	135.8	431	227.8	1471.4
RETURN	30	76.2	279.5	152.5	543.5	674.1	1516.7
PERIOD	40	89.6	326.4	159.4	597.8	1052.5	1527
	50	101.5	367.6	164.6	636.8	1485.3	1532.9

Table 18 Returning Period for Each Concentration of the Pollutants According to the GEV-Approach.

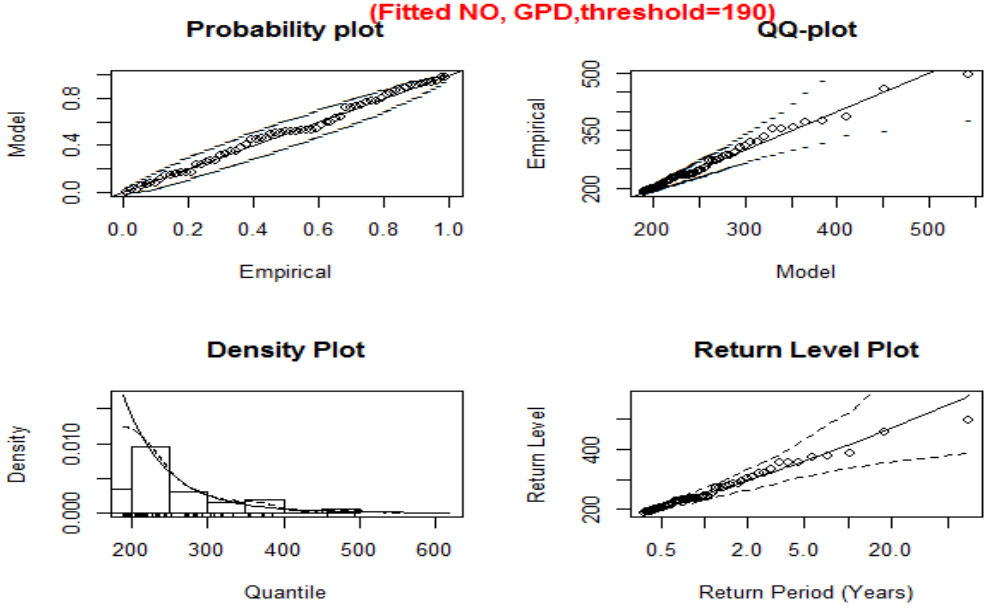
According to Figure 38, the return value of the fitted (GEV) to the maximum likelihood approach data (solid black line) and the associated 95% confidence interval (dotted black line) are shown. The prediction for each peak concentration of the pollutants over 15, 30, 40, and 50 years is provided in Table 18. Below, the approach (POT) for return period computation will be discussed, such that we fit the concentrations of the pollutants through the distribution (GPD). The thresholds for the parameters $SO_2\ \mu g|m^3$,$NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, $PM_{2.5}\ \mu g|m^3$ after being calculated by the MEF method are respectively equal to 55, 140, 150, 100, 400, 350.

The probability plots show the empirical versus observed probabilities for the model. It is expected that the model should fit linearly as shown in the above figures. We note that the (GPD) fit of the actual data exhibits a strong agreement with the approach (POT). Figures also shows return periods (in years) for the concentrations of pollutants. We used the (POT) approach and the generalized Pareto distribution (GPD) introduced by (Pickands, 1975), which is widely applied. The advantage of this approach is that it studies the concentration that exceed some threshold level, which makes possible to investigate every possible available data exceeding the threshold, instead of selecting a maximum set of data, like in the (GEV) theory.

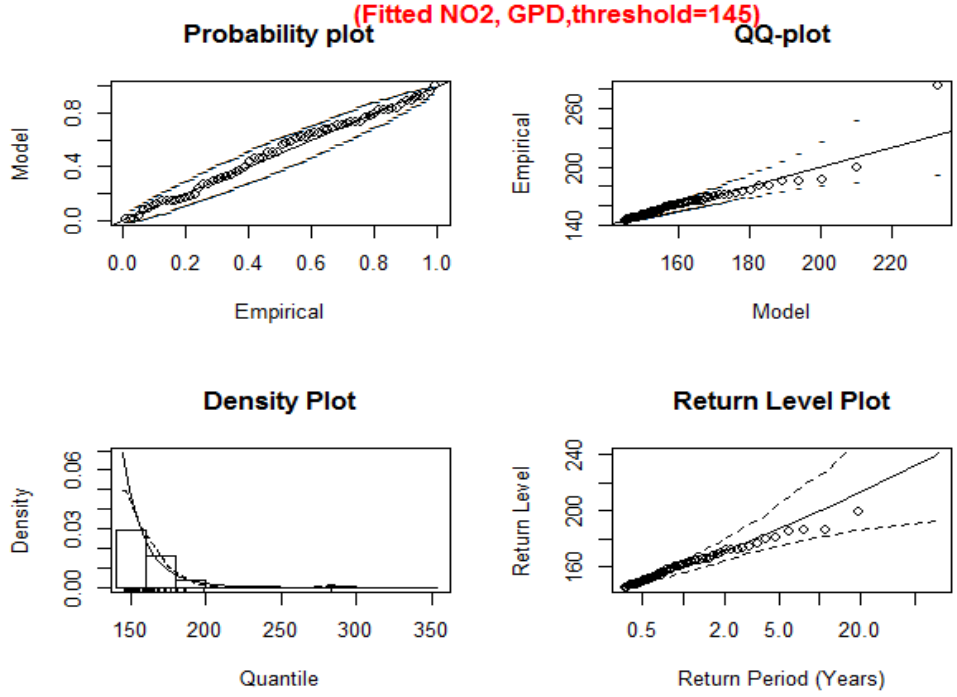
(Fitted SO2, GPD,threshold=55)



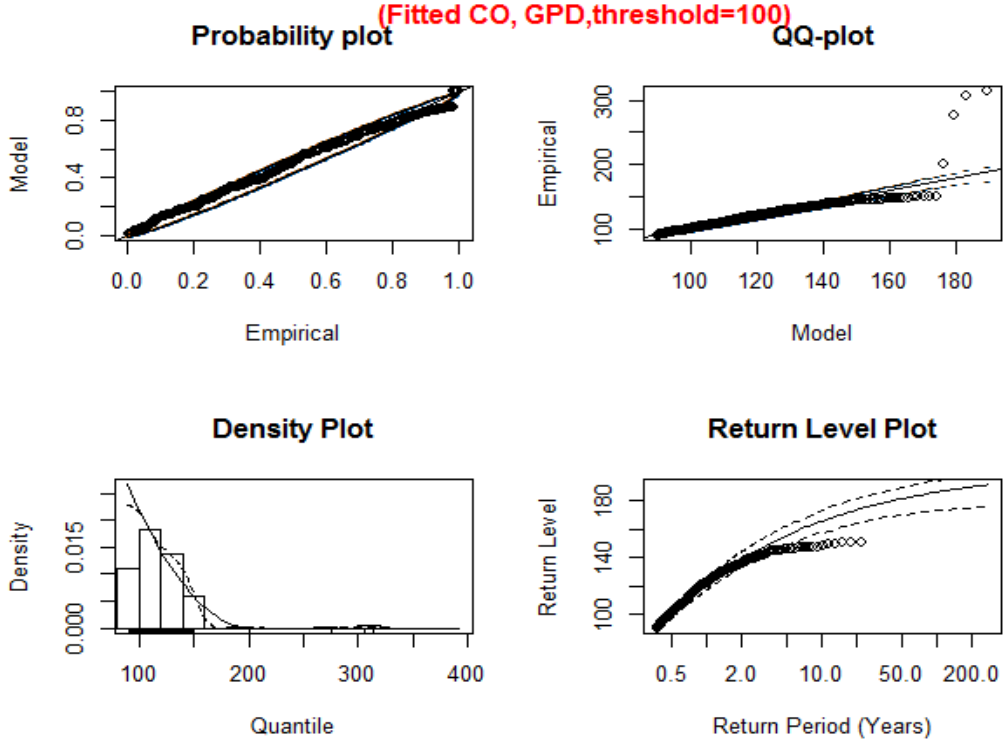
(Fitted NO, GPD,threshold=190)



(Fitted NO2, GPD,threshold=145)



(Fitted CO, GPD,threshold=100)



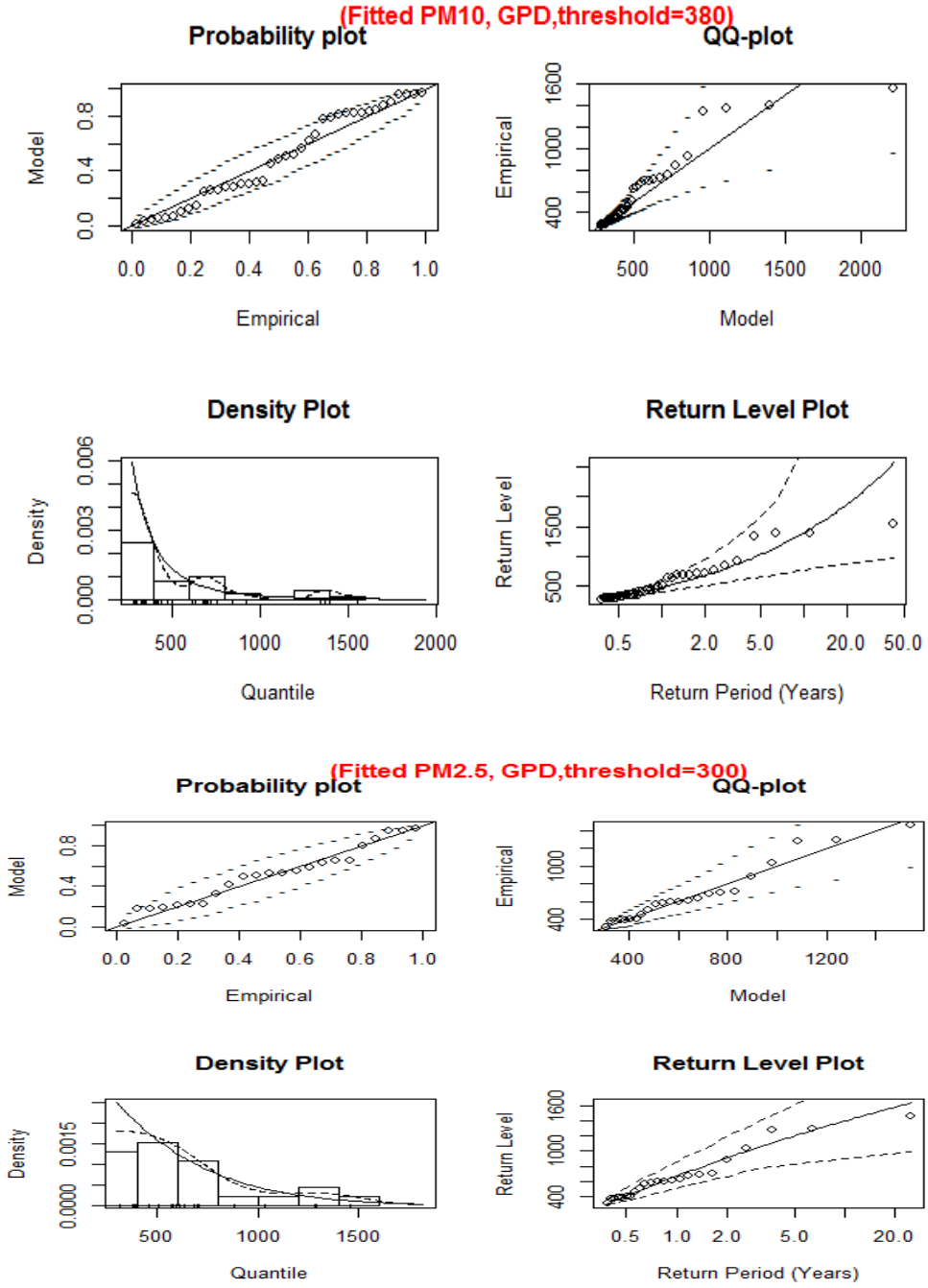


Figure 39 A diagnostic graph for the adjusted POT pattern for each peak air pollutant.

		$SO_2\ \mu g\|m^3$	$NO\ \mu g\|m^3$	$NO_2\ \mu g\|m^3$	$CO\ mg\|m^3$	$PM_{10}\ \mu g\|m^3$	$PM_{2.5}\ \mu g\|m^3$
	15	440.3	500.8	279.3	389.5	1933.8	1588.8
RETURN	30	487.7	517.3	3228	455.2	2159.1	1702.7
PERIOD	40	507.4	523.3	343.3	482.2	2252.5	1745.6
	50	522.7	527.6	361.5	503.1	2325.1	1777.3

Table 19 The Return Period for Each Concentration of Pollutant According to the (POT) Approach

With the POT approach, return levels for the 6 parameters were computed at 15, 30, 40 and 50 years. It can be noticed that concentrations for all the parameters continuously increased. The percentage of change of $SO_2 \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2 \mu g|m^3$, $CO\ mg|m^3$, $PM_{10} \mu g|m^3$, and $PM_{2.5} \mu g|m^3$ after 50 years is respectively 64%, 5%, 3%, 29%, 20%, and 12%. The greatest change in concentration is expected to happen after 15 years, then the change concerning time will be smaller. SO_2 has the highest percentage of change, followed by CO and the particulate matter.

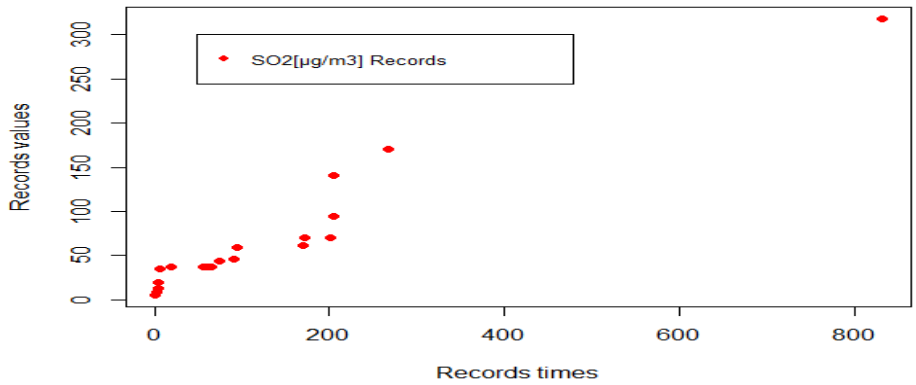
The results of this research indicate that extreme values theory involves investigating a maximum of a random sequence of variables. These results yield insights on the future extreme impacts of the pollutants on Lebanon. The results of (POT) are more coherent with the historic pollutants concentrations compared with the (BM) and the (POT) approaches. We further remark that all return value of the (GEV) approach are lower than the maximal pollutant concentration shown in Table 15 Descriptive Statistics for the Maximum Hourly during the 24 Hours for Air Pollution. Consequently, the (POT) approach is most adequate.

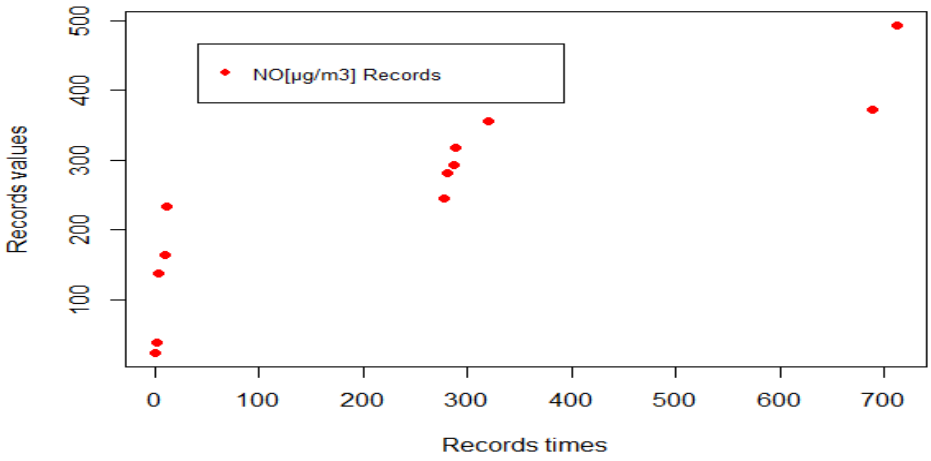
6.4. Extreme Records

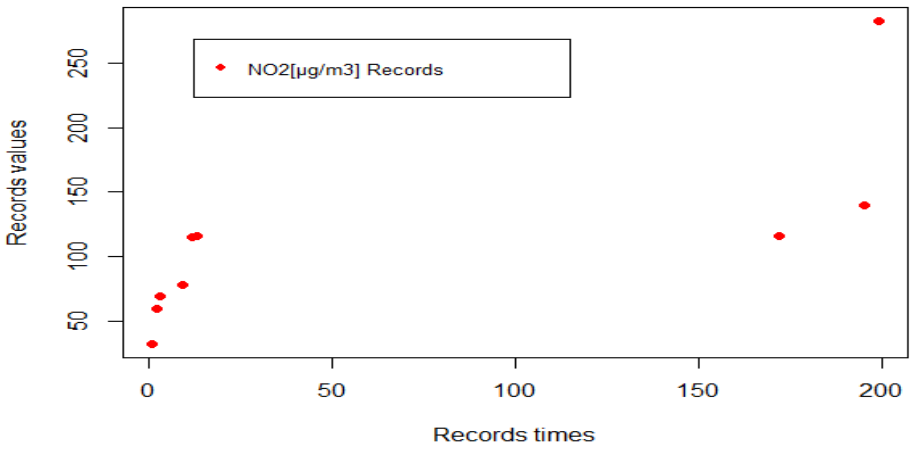
An analysis of record occurrence values for daily pollutant parameters is provided herein. The aim was to gain insight and to quantify statistical records for these parameters within the framework of climate change and atmospheric pollution in Lebanon. Similar to the extremes approach, a simple mathematical pattern of identically distributed random variables has been used to forecast increases in the peak values of the atmospheric pollutants. The statistics of these parameters are investigated further based on a new significant contribution from record theory. Within this framework, however, we also consider the record values and times, which will be required for the prediction of the future observed records.

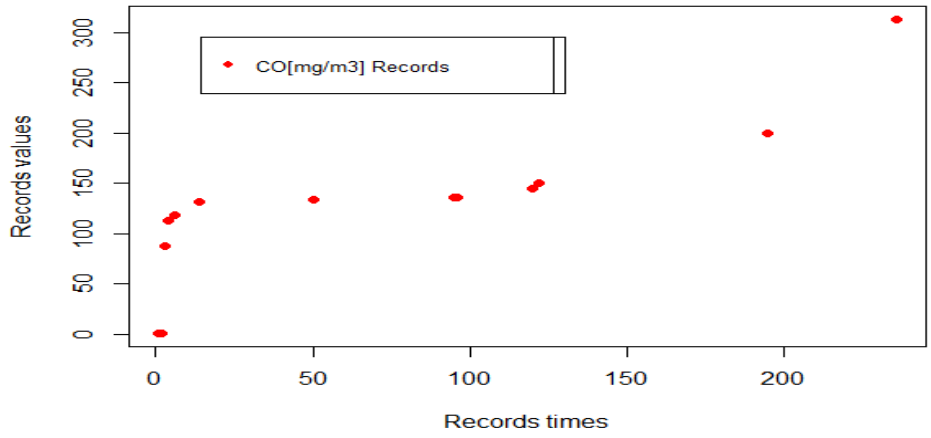
6.4.1. Records results

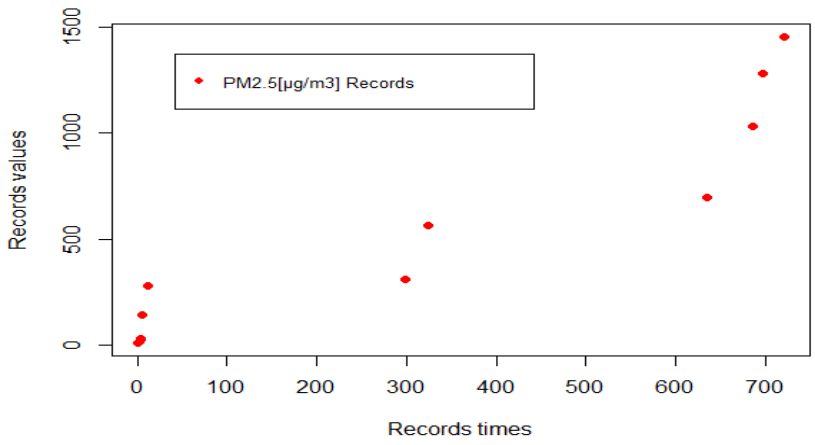
As a baseline, a plot of the upper record values for pollutant parameters throughout 2016 to 2019 is given in the figures as shown below:




NO[µg/m3] Records
Records values
Records times
0
100
200
300
400
500
600
700


NO2[µg/m3] Records
Records values
Records times
50
100
150
200
250


CO[mg/m3] Records
Records values
Records times
0
50
100
150
200
250
300


PM2.5[µg/m3] Records
Records values
Records times
0
500
1000
1500

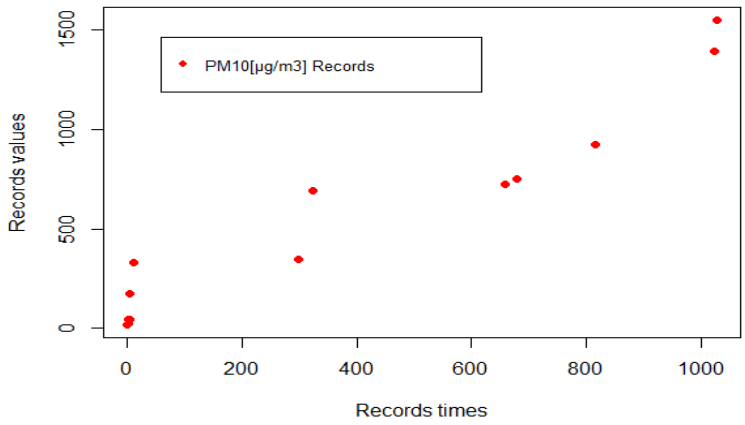


Figure 40 Record values for the various parameters of each pollutant.

From Figure 40, we remark that records are clustered amongst the first observations. The record rate asymptotically converges towards to zero $P(\delta_i) = 1/i$. Likewise $E(N_n) \approx \ln(n)$ which implies that the records have a tendency to get further away over time. Finally, the record values R_n, the record times L_n and the number of observed records N_n for the pollutant parameters are derived. Results obtained are shown in the table below:

$SO_2\ \mu g\|m^3$			$NO\ \mu g\|m^3$			$NO_2\ \mu g\|m^3$		
Record Dates	L_n	R_n	Record Dates	L_n	R_n	Record Dates	L_n	R_n
01/01/2016	1	5.11	01/01/2016	1	23.73	01/01/2016	1	31.79
03/01/2016	3	8.85	02/01/2016	2	38.79	02/01/2016	2	59.75
04/01/2016	4	12.65	03/01/2016	3	137.58	03/01/2016	3	68.66
05/01/2016	5	19.56	10/01/2016	10	164.73	09/01/2016	9	77.8
06/01/2016	6	34.76	12/01/2016	12	234.44	12/01/2016	12	114.53
13/01/2016	19	36.83	04/10/2016	278	245.91	13/01/2016	13	115.83
28/02/2016	56	36.97	07/10/2016	281	281.03	20/06/2016	172	116.10
05/03/2016	59	37.01	13/10/2016	287	293.73	13/07/2016	195	140.03
11/03/2016	65	37.42	14/10/2016	320	317.96	18/07/2016	199	283
21/03/2016	74	43.47	18/11/2017	688	372.67			
10/04/2016	91	45.87	12/12/2017	712	493.64			
14/04/2016	95	58.91						
15/05/2016	169	61.2						
17/05/2016	171	69.95						
16/06/2016	201	70.03						
19/06/2016	204	94.99						
20/06/2016	205	140.64						
24/08/2016	268	170.96						
09/03/2018	832	318.58						
$N_n(SO_2) = 19$			$N_n(NO) = 11$			$N_n(NO_2) = 9$		

Table 20 Record times, record values and number of observed records SO_2, NO, and NO_2.

$CO\ mg\|m^3$			$PM_{2.5}\mu g\|m^3$			$PM_{10}\mu g\|m^3$		
Record Dates	L_n	R_n	Record Dates	L_n	R_n	Record Dates	L_n	R_n
01/01/2016	1	0.92	01/01/2016	1	11	01/01/2016	1	16
02/01/2016	2	1.20	02/01/2016	2	17.6	02/01/2016	2	21.20
03/01/2016	3	87.30	03/01/2016	3	28.8	03/01/2016	3	42.344
04/01/2016	4	113.32	04/01/2016	4	32.67	04/01/2016/	4	44.963
06/01/2016	6	118.14	05/01/2016	5	145.6	05/01/2016	5	174.6
01/02/2016	14	131.93	12/01/2016	12	280.6	12/01/2016	12	332.3
19/02/2016	50	133.76	25/10/2016	299	313.5	25/10/2016	299	347.8
04/04/2016	95	135.6	19/11/2016	324	565.3	19/11/2016	324	690.1
05/04/2016	96	135.86	18/09/2017	634	659.9	20/10/2017	659	722.4
29/04/2016	120	145.24	16/11/2017	686	1032.1	09/11/2017	679	754.2
01/05/2016	122	149.97	28/11/2017	697	1279.3	25/03/2018	815	923.8
13/07/2016	195	200	21/12/2017	721	1454.8	17/10/2018	1021	1396.2
23/08/2016	236	313.12				23/10/2018	1027	1553
$N_n(CO) = 13$			$N_n(PM_{2.5}) = 12$			$N_n(PM_{10}) = 13$		

Table 21 Record times, values and number of observed records CO, $PM_{2.5}$, and PM_{10}

From the above tables, we observe that records occur early in the time series and the high occurrence of the observed records of the pollutant parameters occurs in the year 2016. While the pollution would be exceeded during summer because of the high

temperature, the findings indicate that record values are mostly achieved during the colder weather months. This is explained the fact that some of these pollutants like $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, and $SO_2\ \mu g|m^3$ are a result of incomplete combustion of fossil fuels. In the cold season, the combustion process becomes much more unlikely to be completed due to low temperature, which leads to increased incomplete products of combustion. The probability of a new record can be computed for every pollutant and any nearby time by using $P(\delta_i)$ for an independent identical distribution, e.g., the probability of a new record being observed in 2020 is equal to 0.018. Therefore, according to the formula (6), there is a direct relationship between the number of expected records theoretically derived by $E(N_n)$ and observed record values of concentration pollutants. Results are shown in more detail in Table 22 below.

Parameters	SO_2 $\mu g\|m^3$	NO $\mu g\|m^3$	NO_2 $\mu g\|m^3$	CO $mg\|m^3$	PM_{10} $\mu g\|m^3$	$PM_{2.5}$ $\mu g\|m^3$
Obs. Records Val.	19	11	9	13	12	13
Exp. Nb. Records	17.21	10.62	10.08	12.2	11.62	12.15
Exact relative error	9.42%	3.4%	12%	6.76%	3.16%	6.53%

Table 22 Measurement of Prediction Error of Each Parameters

Observed record values closely approximate expected values, which indicates that the record patterns being considered were well attributed to the concentration pollutants. Furthermore, the minimal accurate relative error is recorded across all parameters, which indicates a good fit of the recording model. Moreover, we can notice that, when n increases, the records have a tendency to be further away in time. The majority of the records may be observed during the first several years, however, when n grows sufficiently large, then the number of records will behave as ln(n).

6.4.2. Records prediction

We will predict the next record value for the pollutant parameters based on the record values collected in previous years. Using equation (16) we predict the record point values and equation (17) gives the confidence interval of prediction for the pollutants.

Parameters	R_n^*	C.I.(95%)
$SO_2 \mu g\|m^3$	319.42	[318.51,328.83]
$NO\ \mu g\|m^3$	497.37	[493.52,537.91]
$NO_2\ \mu g\|m^3$	286.13	[282.84,310.97]
$CO\ mg\|m^3$	314.84	[312.99,332.22]
$PM_{10}\ \mu g\|m^3$	1464.12	[1454.18,1501.62]
$PM_{2.5}\ \mu g\|m^3$	1561.53	[1552.39,1597.68]

Table 23 Predict Records Values for Each Parameters

Table 23 gives point prediction and the upper 95% confidence limit prediction for the (BLUP) of each parameter. After one year of monitoring, the values of records of $SO_2\ \mu g|m^3$, $NO\ \mu g|m^3$, $NO_2\ \mu g|m^3$, $CO\ mg|m^3$, $PM_{10}\ \mu g|m^3$, and $PM_{2.5}\ \mu g|m^3$ increased by 0.2%, 0.7%, 1%, 0.5%, 6%, and 5.7% respectively. Thus, pollution is expected to become more of a problem with time, particularly with regard to particulates, which show the largest increases after year one, and which are known to cause a number of effects to human health in the short and long term. The Bekaa region is a complex region with a variety of human activities. This region is witnessing a catastrophic environmental scenario affecting air quality. These extreme events, which result from heavy pollution, affect the Lebanese economy, particularly in the Bekaa region, which depends mainly on agricultural activity. Atmospheric factors such as carbon monoxide ($CO\ mg|m^3$)nitrogen dioxide ($NO_2\ \mu g|m^3$), and sulfur dioxide ($SO_2\ \mu g|m^3$) are expected to reach future high records. They are emitted by vehicle engines and larger high-temperature combustion plants, industries, and road networks. Similarly, fine particles $PM_{10}\ \mu g|m^3$ and $PM_{2.5}\ \mu g|m^3$ also reached record high values mainly because of heating and road fuels which can be very harmful to human health. Nitrogen oxides are pollutants mostly linked to emissions from road traffic.

Now we compute the probability to expect this record by using the formula (18). Hence, for $n_1 = 4$ years (number of years in the database), the probability that we will have to wait for more than 2 years for example before observing a new record is $\frac{4}{4+2} = 0.66$.

7.4.Conclusion

In this article, extreme and record theories have been used to analyze the air data. The fundamentals of extreme value and record theories were reviewed. Those theories have been applied on six different pollutants. The data were sufficient to investigate the extreme value and make an accurate prediction of the return level of the pollutant concentration after 15, 30, 40 and 50 years using the peak-on-threshold approach of the extreme value theory. The extreme records, the values and the time interval between them can also be followed according to the theory of extremes. We found that extreme pollutant concentration levels occur throughout the observation, but the record extremes are mainly concentrated at the beginning of the observation, and then move farther away from each other over time. This implies that pollutant concentration should be increasing, but the time between records is widening. In addition, most surveys have been performed during the winter season, when there is heavy fossil fuel burning and incomplete combustion as a result of the low temperatures, which leads to more harmful by-products being generated. Note that results on extreme pollutant prediction may differ among various geographies since the pollution concentration is affected by many factors like emission levels, meteorological conditions as well as the geography. Eventually, extremes and record theories may be used as a managerial device for detecting the pollution levels in order to establish a warning system for the future. Another point that should be emphasized is the record values which are higher than the WHO guidelines. These elevated levels may endanger population health when frequently reoccurring.

Chapter 5

Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air pollutants in Zahleh, Lebanon

After studying the extreme atmospheric pollutants in the previous chapter, it is useful to provide the air quality index that will be the purpose of the present chapter. this chapter is published in the journal

Summary: During the last decades, air pollution has become a serious environmental hazard. Its impact on public health and safety, as well as on the ecosystem, has been dramatic. Forecasting the levels of air pollution to maintain the climatic conditions and environmental protection becomes crucial for government authorities to develop strategies for the prevention of pollution. This study aims to evaluate the atmospheric air pollution of the city Zahleh located in the geographic zone of Bekaa. The study aims to determine a relationship between variations in ambient particulate concentrations during a short time. The data was collected from June 2017 to June 2018. In order to predict the air quality index (AQI), Naive, Exponential Smoothing, TBATS (a forecasting method to model time series data), and Seasonal Autoregressive Integrated Moving Average (SARIMA) model were implemented. The performance of these models for predicting air quality is measured using the mean absolute error (MAE), the root mean square error (RMSE), and the relative error (RE). SARIMA model is the most accurate in prediction of AQI (RMSE= 38.04, MAE= 22.52 and RE= 0.16). The results reveal that SARIMA can be applied to cities like Zahleh to assess the level of air pollution and to prevent harmful impacts on health. Furthermore, the authorities responsible for controlling the air quality may use this model to measure the level of air pollution in the nearest future and establish a mechanism to identify the high peaks of air pollution.

5.1. Introduction

Air pollution is one of the most important public health problems, specifically in urban areas where the majority of industries and traffic circulation. It has impacts on human health as well as on the environment. Nowadays, air pollution does not threaten only humanity, but also Earth's ecosystem (Kahraman, 2009). Air pollution is a growing public health problem, and mortality due to air pollution is expected to double by 2050 (Hamanaka, 2018). According to recent estimates reported by the World Health Organization (WHO), more than 90% of the world's population is exposed to unhealthy air quality, which exceeds the guideline limits, which is the main cause of high incidence of premature mortality and morbidity. In addition, each year, 4.6 million people die from causes related directly to air pollution (WHO, 2021). Health problems increase depending on the time of exposure and the type of pollutants. Hence, we can define several types of pollutants which can be classified into two groups based on their sources: primary and secondary. The primary category covers the pollutants which are directly transmitted from an atmospheric source of air pollutants. The Secondary group consists of the pollutants which are formed from the chemical reaction of the Primary pollutants within the atmosphere (Lyon, 1994). The Middle East region is characterized by specific weather, hot and dry in summer and moderate and humid in winter. This region is composed of an almost surrounded sea where the air constitutes a crossroads of polluted and natural emissions (Lelieveld, 2002). The Mediterranean basin is especially sensitive to atmospheric air quality due to cloudless weather and intense summer sunshine. Since Lebanon belongs to the MENA region, the assessment of air quality in this country has been of special importance in recent years. Lebanon is a small country in the Middle East, located at approximately $34^{o}N$,$35^{o}E$ surface 10452 km^2, its widest point is 88 kilometers, and its narrowest is 32 kilometers; the average width is about 56 kilometers (AbuKhalil, 1989). Air pollution in this country is very high, in proportion to its area and number of inhabitants. The increase in signs of pollution along with the number of people suffering from respiratory illness made it very important to monitor and control air pollution. According to the World Health Organization, the air quality in Lebanon is considered very dangerous to the ecosystem. Recently the annual average concentration of $PM_{2.5}$ in the country is 31 $g|m^3$, which

exceeds the recommended maximum of 10 $g|m^3$. Several factors that contribute to poor air quality in Lebanon include cement industries, food processing, minerals and chemicals, petroleum refining, and vehicle emissions. Note that there are seasonal variations in Lebanon with the highest levels of air pollution in winter (December to March) due to heating. The most polluted cities in Lebanon are Baalbak, Beirut, Saida, and Zahleh have high levels of air pollution. Zahleh is a heavily populated city that is affected by several sources of air pollution. Several pollutants were identified that have constantly increasing levels in diesel generator sector (Baalbaki R., et al., 2016). In addition, the significant crisis in solid waste treatment has increased the concentration of pollutants in the air, as well as the risks of short-term cancer and respiratory diseases. Several works were carried out to study the air quality in the capital Beirut, these studies showed serious air pollution which has an impact on the health of the citizens. However, there are no recent studies to assess the air quality in the Bekaa region, especially since this region differs from the capital in geographical characteristics, its rivers, and its valleys. In this article, we are interested in modeling pollution in Zahleh, Lebanon, where we will be focusing on three main objectives; identifying the relationship between the parameters of air pollution and the meteorological elements, calculating the Air Quality Index (AQI), and making an air quality forecasting to predict future levels of pollution. Previously, time series analysis has been used to study the daily average concentration of air pollutants, compare the fluctuation of time series, and investigate the variation of air pollutants overtime on a weekly, monthly, and yearly basis in Lebanon (Farah, 2014). This method was also used worldwide in big cities in very crowded countries such as India (Kumar, 2011) and (Xia X, 2017), (Wang, 2020). The time series method was helpful in showing the variation of the daily and seasonally averages in order to identify the time of the year, and of the day, where some parameters have peak levels (Saliba, 2006). Also, the correlation between air pollution parameters and the meteorological elements was investigated many times in the literature, using different models such as the Random Forest regression (Kaminska, 2018), simple linear regression (Arnaudo, et al., 2020), and multi-linear and nonlinear regressions (Kayes, 2019). Findings were different in each study area and for different parameters, where some parameters showed a highly negative correlation with meteorological elements, and others were affected positively by these elements. Many statistical approaches are

used to study air pollution and its health effects in order to analyze the air quality and to predict the future values (Lee, 2002). Note that descriptive statistics are limited in terms of understanding behavior and air quality variability (Voigt, 2004). In addition to probability models, (Bencala, 1976) to study the temporal distribution of air pollutants, while other studies have relied on time series analysis, which facilitates a better understanding of the cause and effect relationship in environmental pollution (Schwartz, 1990). The ARIMA and SARIMA models were found very useful in different studies conducted in Tehran and Peninsular Malaysia respectively (Kumar, 2021). Then, combined with the ARIMA model, the principal component regression was chosen as the best model to calculate and forecast future air quality index (AQI) in DelhiIndia (Kumar, 2011).

This article is organized into several sections, firstly the description of a monitoring station and the parameters of pollutants in order to illustrate the evolution of the pollution in this region. Secondly, a description is given regarding the statistical method, the linear model, and the performance of the model to model and forecast air quality for the near future. Then, in the next section, we present results, starting with descriptive statistics and graphic presentation per hour/week/month in order to assess the variation of atmospheric pollutants. Afterward, to investigate the correlation between different measurements, we compute the coefficient of correlation for two through a general formula that presents the pollutant parameter to correctly forecast the atmospheric pollution and compare several forecast models together with selecting the SARIMA model as the most suitable one according to the selection guideline provided. In the end, an overall summary is presented whose aim is to evaluate the levels of air pollution to prevent harmful impacts on health.

5.2. Study area

Monitoring of air quality in Lebanon dates back to 2013 from the installation of five monitoring stations as part of the Environmental Resources Monitoring Project in Lebanon (ERML). These stations use online analyzers connected to a control and data acquisition system (DAS) located at the MoE. All the analyzers installed within the air quality monitoring stations (AQMS) are based on reference methods meeting the

requirements of the European directive on air quality 2008/50/EC. The monitoring stations were located in Memchiyeh garden with coordinates of 33°59'52.44"N and 36°12'16.43"E. Zahleh is the largest city in BeKaa and the fourth largest city in Lebanon with an area equal to 8 km^2. It is located 54 *km* to the east of the capital Beirut. It is located on the Eastern foothills of Sannine mountain and surrounded by the Lebanese mountains and the Bekaa valley (Atoui A., 2022).

5.3. Data monitoring

In this paper, we focus on the more prevalent atmospheric pollutants: Nitrogen oxides (NO_x), which contain nitrogen and oxygen in varying levels, such as nitric oxide (NO) and nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and particulate matter (PM) of 10 and 2.5 microns in diameter. The pollutant concentrations are compared to the WHO Guidelines that are summarized in the following table:

Pollutant	Duration of exposure	WHO Guidelines
$PM_{2.5}$	1 year	35 $\mu g/m^3$
PM_{10}	1 year	20 $\mu g/m^3$
NO_2	1 year	40 $\mu g/m^3$
O_3	8 hours	100 $\mu g/m^3$
SO_2	24 hours	20 $\mu g/m^3$

Table 24 WHO guidelines for air pollutants.

The pollutant data is registered by the Department of Air Quality Ministry of Environment-Lebanon, which is monitored at Memchiyeh Garden Station located in Zahleh-Bekaa region. The data were hourly observed from June 2017 to June 2018 (13895 recorded observations). The variables available were NO, NO_2, NO_x, CO, SO_2, $PM_{2.5}$, PM_{10}, *Temperature*, *Humidity*, *Wind Direction*, *Wind Speed* at Zahleh. All

pollutants were measured in $\mu g|m^3$ except for *CO* which was measured in $mg|m^3$. The missing values were observed to be filled by various statistical approaches, such as replacing them with the neighbor's most similar value, daily Memchiyeh station, encounters missing values consecutively. Therefore, the Multivariate Imputation by Chained Equation (MICE) algorithm was suggested as the most suitable method to impute these missing values. The MICE is a package **R**, which assumes that data are randomly missing. It prepares multiple values to be replaced by the missing values by designing an appropriate model, such as regression. Each variable is considered a dependent variable to be treated in this approach. The MICE process involves the prediction of the missing values for the selected variables by using the available data of the other variables. Then the missing values will be replaced by the predicted values to make a new data set called imputed data set, and through iterations, multiple imputed data sets are generated (Kumar, 2011), (Sanchez Lasheras, 2020). There are 22485 missing values from a total of 152895 observations (about 15% of the data are missed) in Memchiyeh station. The detection of outliers is not important in time series analysis, but error detection is significant for analysis. Pollutants can have a high level due to numerous factors that serve the purpose of this study. However, for meteorological variables, there are some outliers that were considered to be seasonal errors. The errors detected have been replaced with the median of the time period they are included.

5.4. Methodology

The initial step prior to applying any models is to treat the raw data. For this purpose, we used the Multivariate Imputation by Chained Equation (MICE) algorithm to fill in the missing data, which was an R package that assumes data are missing randomly. There were 22485 missing values out of a total of 152895 observations (about 15 percent of the data are missing). MICE prepares several values to replace with the missing values through a suitable model design such as regression and then replaces the missing data with predicted data to make a new dataset called imputed dataset, and through iteration, multiply imputed datasets will be generated (Kumar, 2011). Finally, each dataset will be analyzed. Subsequently, the outliers were detected and replaced by the median of the period in which they are included. This is especially important

because, in meteorological parameters, outliers are considered errors depending on the season. It is advisable to use numerical values for data aggregation (e.g. calculation of daily mean values or monthly mean values). We selected the daily average values for the calculation. Once the raw data was processed, R software was used for data analysis.

5.4.1. Statistical Models

The goal of the statistical approach is to develop the most accurate model of the phenomenon, not based on the information that we know about it but on information that can be derived from a dataset describing this phenomenon. Therefore, there is no necessity to formalize the operation of the phenomenon being studied and we do not have biases caused by imprecision's in the models we develop. However, we need to have a data set allowing the model to adequately represent the studied phenomenon, and we need to be able to exploit the information present in this data set. Furthermore, the inaccuracy of the data used will affect the forecast. An important aspect of modeling air pollution is statistical analysis, which involves the prediction of the future development of a data set from the observed data set up to the current observation time. To forecast univariate time series, we use Naïve, TBATS, exponential smoothing, and SARIMA models.

5.4.2. Performance measures

The objective of the assessment is to evaluate the ability of the model to perform well. Several measures are used to validate the regression model. Firstly, we focus on the computation of the mean absolute error (MAE) that uses the absolute value of the difference between the observed and the predicted, thus measuring systematic error and random error. The MAE can be used to evaluate the efficiency of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Secondly, the root mean square error (RMSE) is computed by using the difference between the observed and the predicted rather than absolute values. As a result, this index gives more bias towards greater differences between observed and predicted,

which is desirable for air quality prediction purposes as peak levels of pollution are most important for forecasting.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

It is preferred to provide RMSE results as its square root, the mean square error (MSE), which has the same information but has the same magnitude as the predicted variable. RMSE is most widely used due to its legibility and accurate precision estimate.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Finally, the relative error (RE) is a precision measure, which describes the accuracy of the measurement. RE identifies the relative precision between two or more measurements.

$$RE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n} y_i} \times 100$$

5.5. Results and discussion

Following data filtration, the daily averages were computed as well as descriptive statistics of the measured values in the monitoring station. The following step was to make hourly, weekly, and monthly variation graphs using temporal correlation. Numerical values are recommended for data aggregation (e.g., calculation of daily average values or monthly average values). Here, daily average values have been selected to be computed. Once the pollutants data were processed, R software was used to perform the data analysis. The observations were taken from June 1, 2017 to December 31, 2018 in Zahleh. In the following we represent some graphical descriptions.

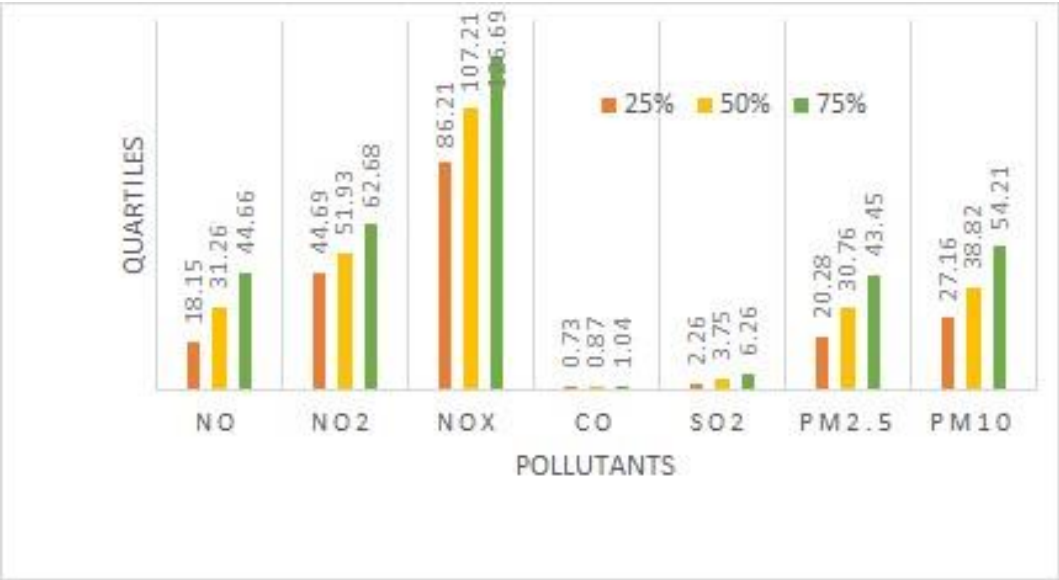


Figure 41 Q1(25%),Q2(50%),Q3(75%) for the pollutants in Memchiyeh station.

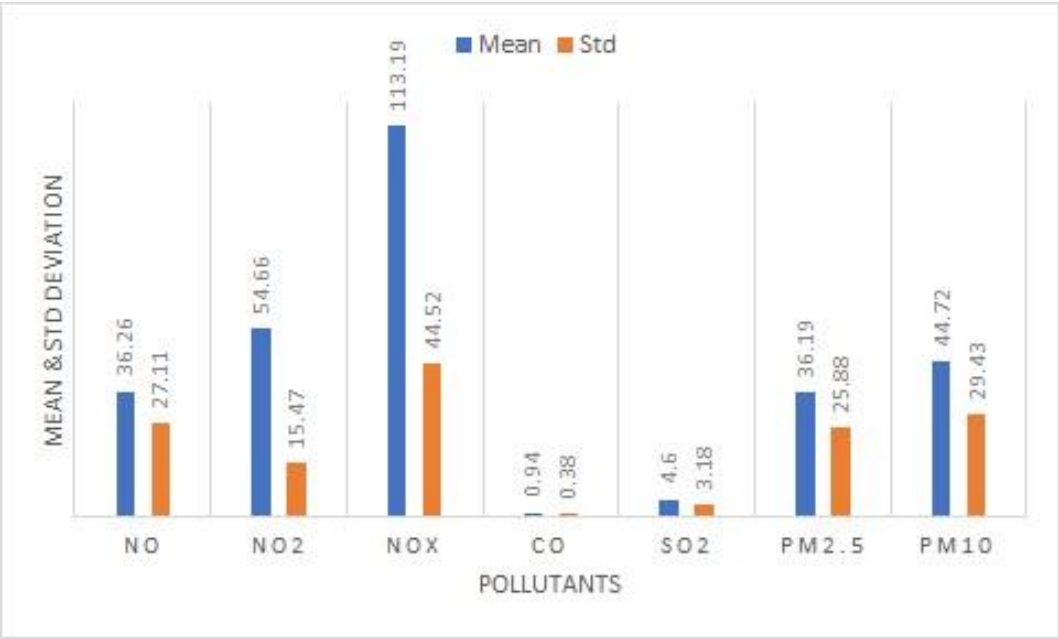


Figure 42 Mean and standard deviation for the air pollutants in the Memchiyeh monitoring station

Hourly variation.

The hourly variation results indicated that there were some times of the day when the concentration of certain air pollution reached its peak, whereas others have no significant variation throughout the day. As an example, *NO*, NO_2, NO_x, and *CO* have two peak time periods, the first recorded between 6 parameters are almost identical

during the day. The following figures illustrate the hourly variation of air pollutants in Zahleh.

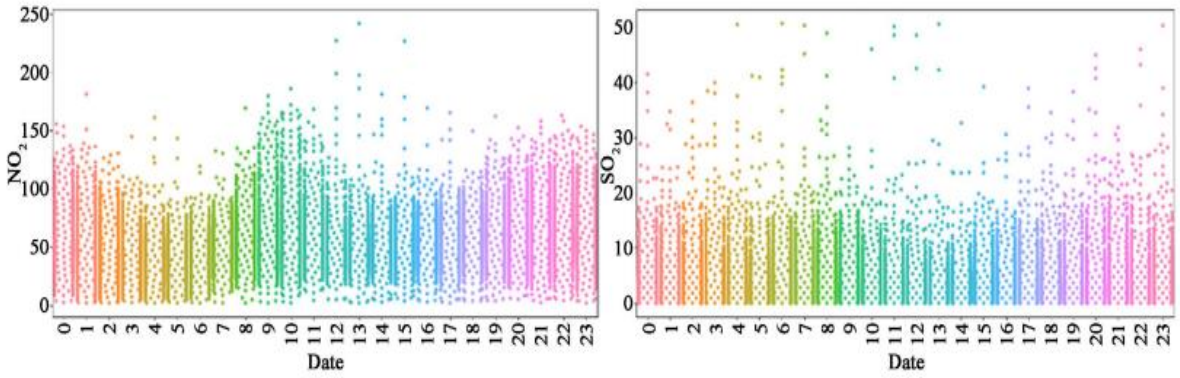


Figure 43 Hourly variation of SO2 and NO2.

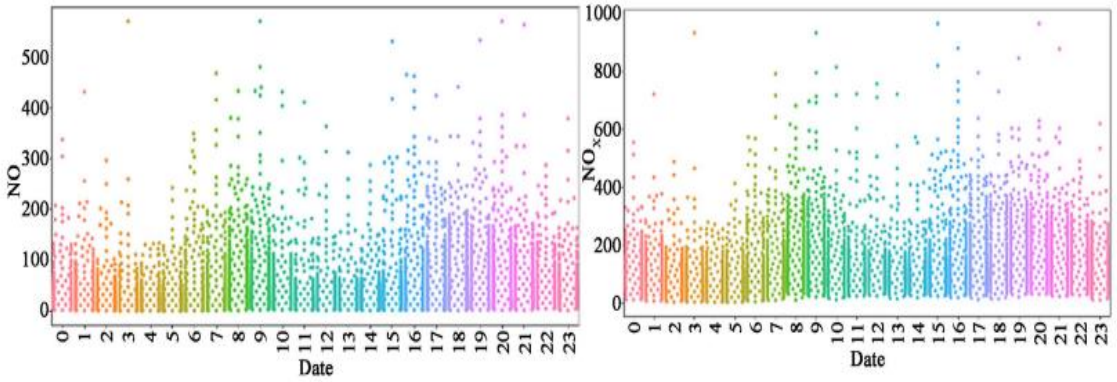


Figure 44 Hourly variation of NO and NO_X .

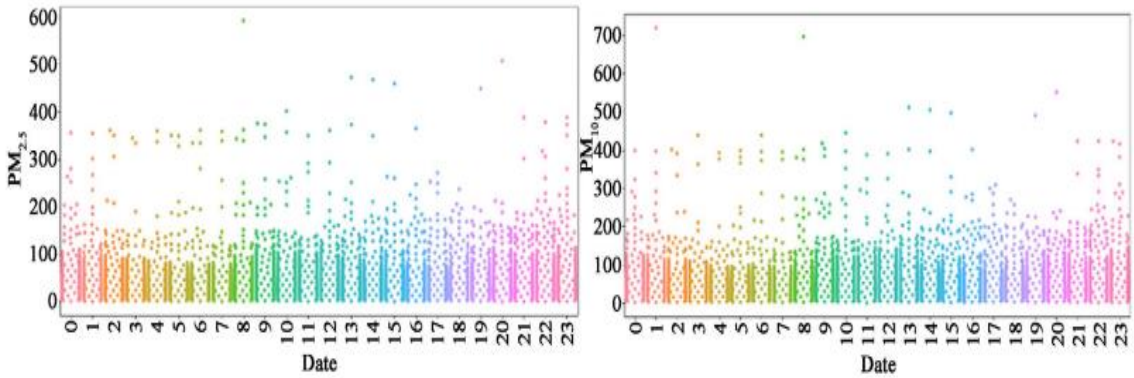


Figure 45 Hourly variation of PM_{10} and $PM_{2.5}$.

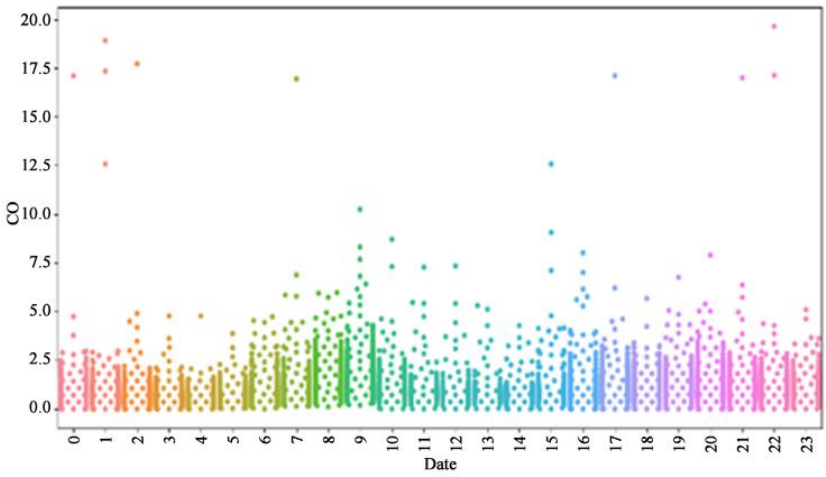


Figure 46 Hourly variation of CO.

The above figures indicate that the concentrations of *NO*, NO_2, NO_x, and *CO* have their first peaks between 6 and 10 AM. The second highest peaks are detected between 16:00 and 1:00 AM. SO_2, $PM_{2.5}$, and PM_{10} pollutants do not exhibit any remarkable hourly variation. High daytime levels are correlated with increasing traffic and urban activities, and night-time concentrations are coincident with an increase in the number of recreation vehicles and outdoor activities. A previous study indicated that peak traffic hours occur between 7 and 11 AM and between 5 and 7 PM from Monday through Friday, although traffic remains relatively high throughout each day (Otayek, 2019).

5.5.1. Weekly variation

This subsection investigates the weekly variation of the pollutants to identify the dependency of their concentration on their source activities which may affect the weekly atmospheric conditions in the Bekaa region, particularly in Zahleh.

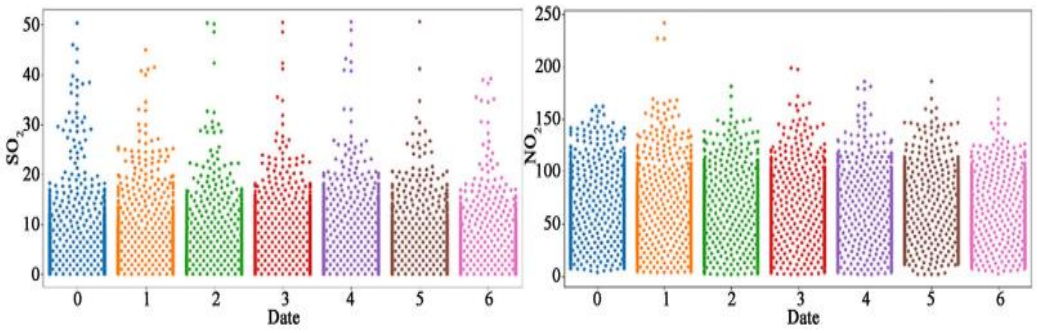


Figure 47 Weekly variation of SO_2 and NO_2.

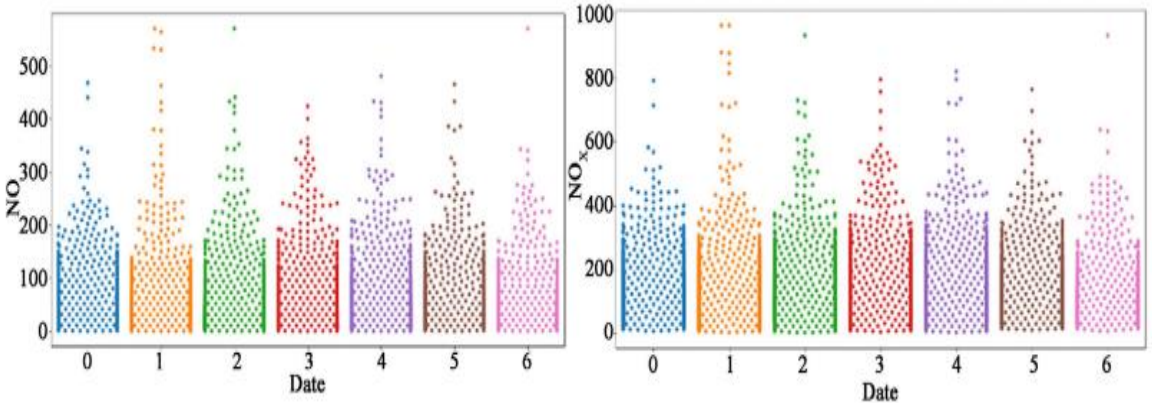


Figure 48 Weekly variation of NO and NO_X.

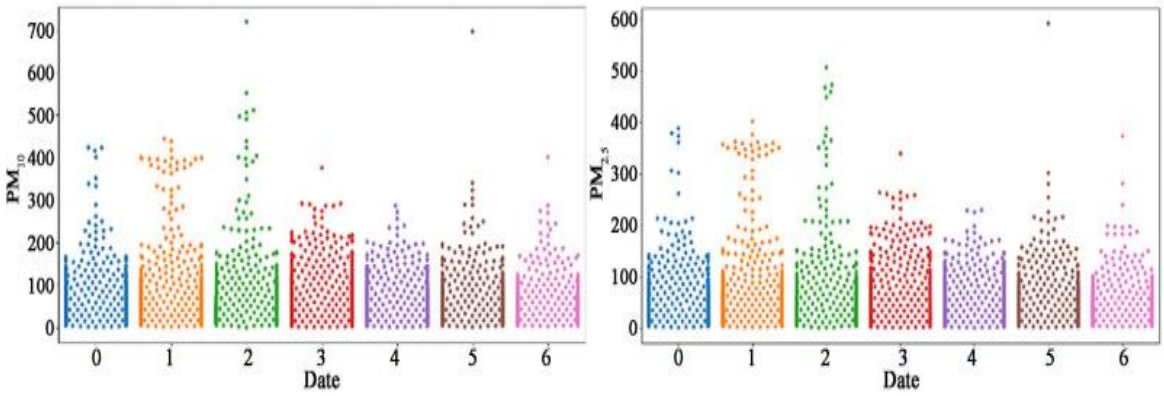


Figure 49 Weekly variation of PM_{10} and $PM_{2.5}$.

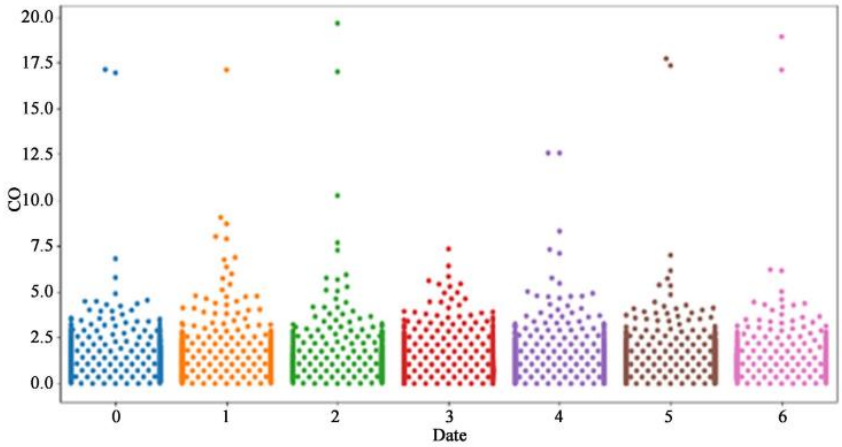


Figure 50 Weekly variation of CO.

Regarding weekly variation, we noticed no significant changes in pollutant levels which remain high throughout the entire week as the city of Zahleh is highly populated and has very frequented roads and driveways during the weekends as well as the weekdays. The principal reason for the absence of weekly variation is that concentrations of pollutants change quickly in urban areas and persist for several days before they disappear. Nevertheless, Zahleh is heavily inhabited. That led to a variety of issues, in particular in the transportation sector, where most transportation methods used in the area were private vehicles, and almost complete absence of any public transport mode, which caused a noticeable increase in pollution in this area.

5.5.2. Monthly variation

Monthly graphs indicate that the concentrations of *NO*, NO_x, SO_2, and *CO* increase in the winter and decrease in the summer. In contrast, NO_2 concentration is highest in the summer. However, $PM_{2.5}$ and PM_{10} have no remarkable variation over the year.

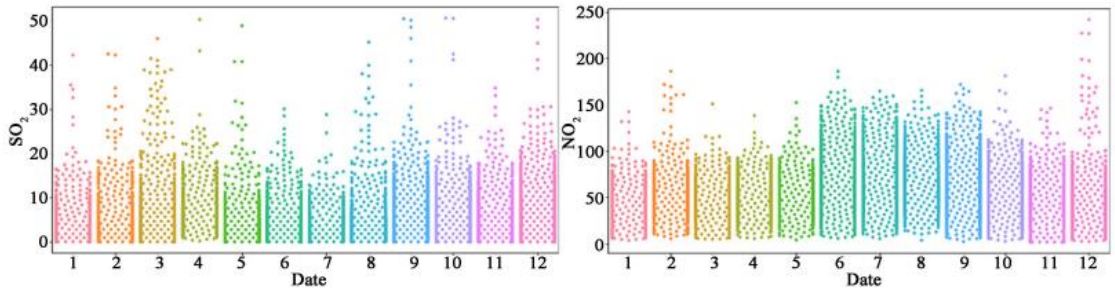


Figure 51 Monthly variation of SO_2 and NO_2.

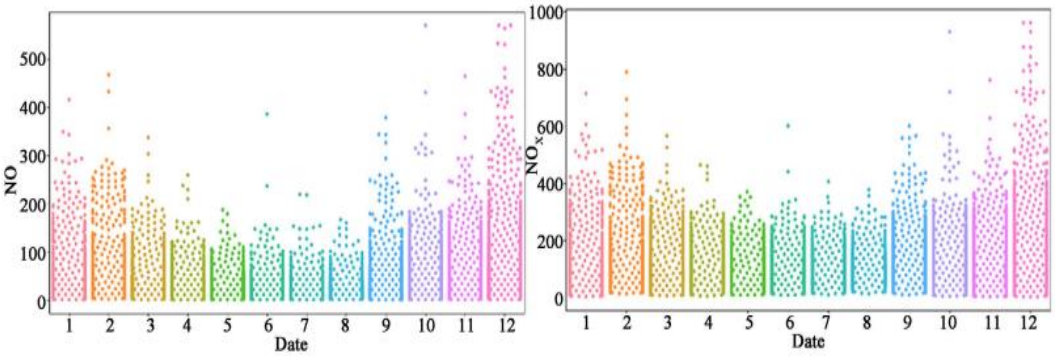


Figure 52 Monthly variation of NO and NO_X.

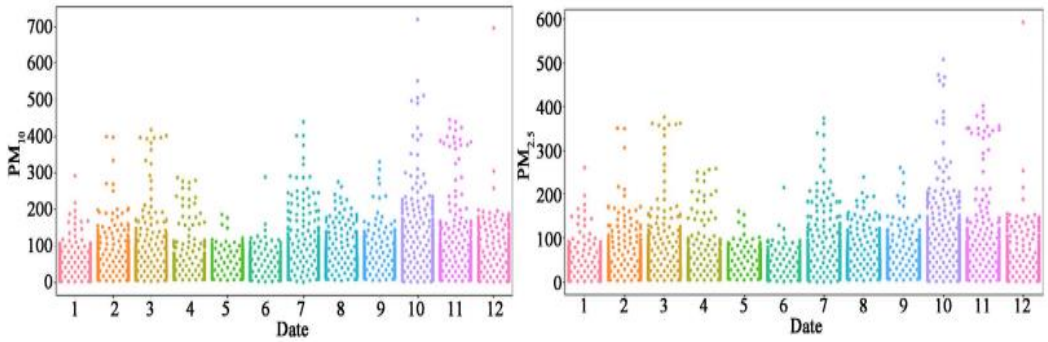


Figure 53 Monthly variation of PM_{10} and $PM_{2.5}$.

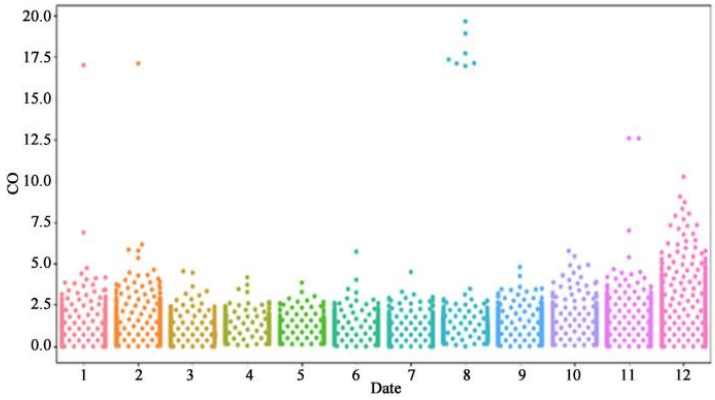


Figure 54 Monthly variation of CO.

High emissions of SO_2 during the winter may be caused by some factors including the development of a reversal atmosphere and delayed wintertime oxidation of SO_2. These pollutants are emitted from diesel vehicles and from central combustion engines. These operations are increased during the winter season, which causes an increase in the levels of pollutants during that season. The traffic levels increase during the winter season and reduce during the summer season as a result of the schools, the universities, and certain work activities that occur seasonally. Whereas during summer, activities of the schools and the universities come to a halt; furthermore, numerous people move out of the country for vacation. Also, heating equipment is not used during summer. Those factors reduce pollutant emissions. Pollutants that do not vary monthly are caused by their sources which do not vary throughout the year.

5.5.2. Environmental factor variation

Considering each parameter independently, using the descriptive statistics, it can be noted that the concentrations of *NO*, NO_2, NO_X, *CO*, $PM_{2.5}$, and PM_{10} are extremely high and that NO_2, $PM_{2.5}$, and PM_{10} are above the guidelines, whereas the SO_2 concentration is almost certainly not above the limits. It is also to be noted that NO_2 concentration is highest during the summer season. In order to justify these levels, it should be mentioned the presence of an electric power station in Zahleh, 5 km away from the monitoring station, which causes 40% of the NO_x emissions. Regarding the elevated *CO* levels, these are the consequence of heating for long periods, since the temperature is usually low and significantly decreases during winter in this city.

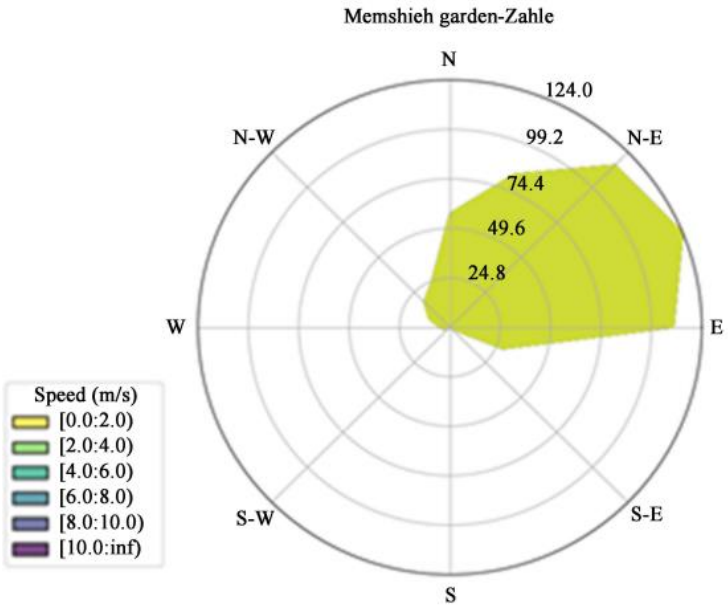


Figure 55 Observed wind speed and direction at Zahleh 06-2017/2018.

One further factor which can affect the level of pollution is the wind direction, as can be shown in Figure 55, where the wind is directed to the northeast, which means that it is moved towards Zahleh, and as this city is located between the Lebanese mountain chains to the east and the west, rather than being an entirely wide area, the pollution transmitted by the wind from other cities can be captured at Zahleh for longer time periods, which explains the high pollution levels around the year. Furthermore, the temperature in Zahleh is much warmer than in other cities like Beirut all year round,

leading to greater usage of gas heaters, fireplaces, and ovens in Zahleh, thus leading to increased *CO* levels.

5.5.3. Pollutants concentration correlation

The Pearson seasonal correlation was used to identify a correlation between the air pollution parameters and the meteorological components. Firstly, the data were divided into two seasons: the summer season including the months from March to August, and the winter season including the months from September to February. Results are shown in the following tables:

	NO	NO_2	NO_X	CO	SO_2	$PM_{2.5}$	PM_{10}	Temp	Hum	WD	WS
NO	1	0.62	0.97	0.89	0.56	0.21	0.23	-0.20	-0.07	0.29	0.00
NO_2		1	0.53	0.75	0.62	0.27	0.29	0.46	-0.59	0.29	-0.16
NO_X			1	0.89	0.55	0.25	0.27	-0.08	-0.20	0.32	-0.03
CO				1	0.59	0.25	0.26	-0.12	-0.18	0.33	-0.02
SO_2					1	0.29	0.31	0.32	-0.57	0.41	-0.05
$PM_{2.5}$						1	1.00	0.23	-0.15	0.16	-0.17
PM_{10}							1	0.24	-0.16	0.18	-0.18
Temp								1	-0.74	0.11	-0.39
Hum									1	-0.27	0.16
WD										1	0.14
WS											1

Table 25 Winter correlation matrix in Zahleh

	NO	NO_2	NOx	CO	SO_2	$PM_{2.5}$	PM_{10}	Temp	Hum	WD	WS
NO	1	0.64	0.77	0.38	0.66	0.15	0.12	-0.45	-0.08	0.21	0.27
NO_2		1	0.53	0.44	0.67	0.11	0.14	0.60	-0.58	0.36	-0.24
NOx			1	0.41	0.62	0.18	0.18	0.02	-0.43	0.41	0.06
CO				1	0.31	0.22	0.23	0.15	-0.25	0.21	-0.09
SO_2					1	0.18	0.16	-0.16	-0.27	0.26	0.27
$PM_{2.5}$						1	1	0.11	-0.23	0.11	0.09
PM_{10}							1	0.16	-0.24	0.12	0.07
Temp								1	-0.58	0.11	-0.45
Hum									1	-0.33	0.00
WD										1	0.32
WS											1

Table 26 Summer correlation matrix in Zahleh.

These Pearson correlation matrices were built and found that, during winter, the temperature has a moderate and positive correlation with NO_2 and SO_2 at Zahleh, and weak correlation with the remaining pollutants. The humidity correlates strongly negatively with NO_2 and SO_2 at Zahleh, and correlates negatively with NO_2, NO_x, and SO_2 moderately. The wind direction shows a moderate positive correlation with NO_x, CO and SO_2 at Zahleh. The wind speed has a moderate negative correlation with NO, NO_X and CO, and a strong negative correlation with NO_2.

During summer, the temperature has a high positive correlation with NO_2, but a moderate negative correlation with NO at Zahleh. Humidity correlates negatively and strongly with NO_2 but moderately with NO_x at Zahleh, negatively correlates moderately with NO and SO_2, and negatively correlates strongly with NO_2 and NO_X. The wind direction has a positive and medium correlation with NO_2 and NO_X. The wind speed

has a weak correlation with all the air pollutants during the entire year at Zahleh. This is attributed to Zahleh being enclosed in two hills which reduces the wind speed. As all the pollutants share some common sources and some pollutants are in the same cluster, we could identify some correlations between them. *NO* and NO_2 belong to the NO_X group, and they strongly correlated and similarly correlate to other pollutants. This is also the case for $PM_{2.5}$ and PM_{10}. During the winter, NO_X and their group are strongly positively correlated with *CO* and SO_2, and are weakly positively correlated with *PM*. *PM* is weakly positively correlated with all the pollutants. SO_2 and *CO* are strongly positively correlated. During the summer season, NO_x and their group show a high, moderate, and low positive correlation with SO_2, CO, and PM, respectively. PM has a weak positive correlation with the other pollutants. CO and SO_2 have a moderate and positive correlation. It is noted that the correlation between the different pollutants, and between the various pollutants and meteorological factors, has not significantly varied among the different seasons. Therefore, in the following process, the seasonal variation is not considered and the results will be treated annually.

5.6. Application of AQI

The principal purpose of AQI is to evaluate the concentration of atmospheric pollutants to study air quality. AQI is a dimension-free measure. Firstly, the sub-AQI of the six principal pollutants ($PM_{2.5}$, PM_{10}, SO_2, *CO*, NO_2, and O_3) were computed using the observed concentrations. Secondly, AQI is derived based on the maxima of the sub-IAQs of all pollutants, as illustrated in equation 2. It should be noted that once the AQI is greater than 50, the maximum sub-AQI pollutant is defined as the primary pollutant occurring that day (Shen, 2017). The higher AQI indicates that air pollution is serious and produces major health damage. The AQI formula is shown as follows:

$$IAQI_P = \frac{I_{High} - I_{Low}}{C_{High} - C_{Low}}(C_p - C_{Low}) + I_{Low} \quad (1)$$

$$AQI = max(IAQI_1, IAQI_2, ..., IAQI_n) \quad (2)$$

The AQI can be divided into six categories of air quality assessment, presented in Table 27. When the values of the AQI are below 100, the quality of air is adequate. Where

the AQI value is almost 100, pollutant measurements are within legal guidelines. In contrast, when the AQI values exceed 100, the air quality deteriorates. The U.S. Environmental Protection Agency (EPA) has proven the existence of one national standard for the quality of the air in order the protect public health (Hamanaka, 2018).

$PM_{2.5}(\mu g\|m^3)$	$PM_{10}(\mu g\|m^3)$	CO	SO_2	NO_2	NO_x	AQI	AQI Category
$C_{low} - C_{high}(24hr)$	$C_{low} - C_{high}(24hr)$	$C_{low} - C_{high}$ $(24hr)$	$C_{low} - C_{high}$ $(24hr)$	$C_{low} - C_{high}$ $(24hr)$	$C_{low} - C_{high}$ $(24hr)$	$I_{low} - I_{high}$	
0.0 - 12.0	0 - 54	0.0 - 4.4	0 – 35	0 – 53	0-40	0 – 50	Good
12.1 - 35.4	55 - 154	4.5 - 9.4	36 – 75	54-100	81-180	51 - 100	Moderate
35.5 - 55.4	155 - 254	9.5 - 12.4	76 – 185	101-360	41-80	101 - 150	Unhealthy for Sensitive Groups
55.5 - 150.4	255 - 354	12.5 - 15.4	186 – 304	361-649	181-280	151 -200	Unhealthy
150.5 - 250.4	355 - 424	15.5 - 30.4	305 – 604	650- 1249	281-400	201 -300	Very Unhealthy
250.5 - 350.4	425 - 504	30.5 - 40.4	605 – 804	1250-1649	$\succ$ 400	$\succ$ 300	Hazardous

Table 27 EPA's AQI values.

All results obtained are shown and plotted below.

Table 28 shows some characteristics of AQI from June 2017 to 31st December 2018. The average AQI score for the Zahleh cities is equal to 139.13.

Mean	Std	Min	Q_1	Q_2	Q_3	Max
139.13	53.89	52	110	129	151	429

Table 28 Descriptive data of measured air quality.

Although Memchiyeh monitoring recorded the highest average AQI score and the most dispersed measure of air pollution concentration (SD= 53.89; range= 377). These results ensure our previous ones, and that Zahleh is affected by air pollution and suffers from serious health problems.

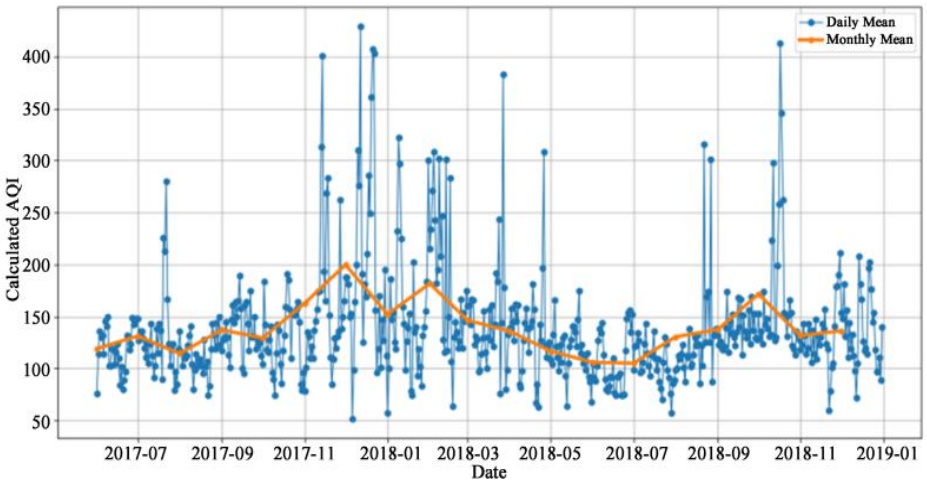


Figure 56 Daily and Monthly average of calculated AQI in Zahleh.

The Figure 56 shows the average of daily and monthly calculated AQI. The monthly AQI reaches its first maxima (199.839) in December and the second one (181.75) in February, then reaches its minima (105.161) in July and the second one (106.033) in June. These results ensure our previous analysis, and that the pollution increase in the winter and decreases in the summer.

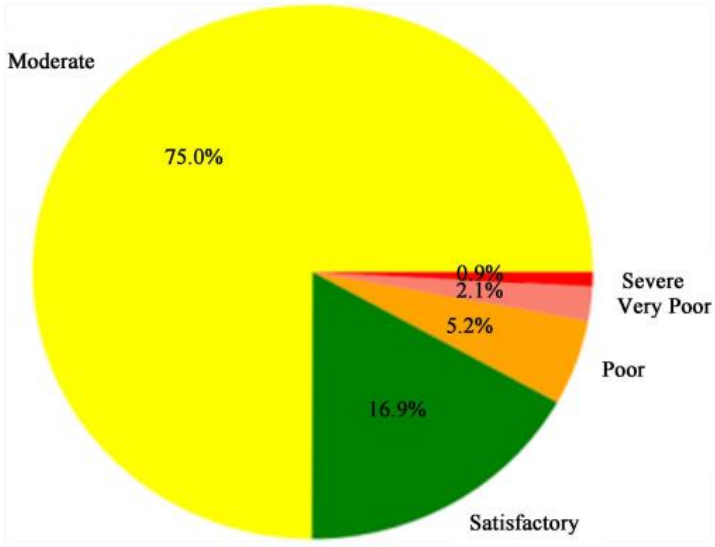


Figure 57 Daily AQI in Zahleh 06-2017 to 12-2018.

We notice that 98 days (16.9%) out of total days are leveled as satisfactory, 434 (75%) leveled as moderate, 30 (5.2%) as poor, and 12 (2.1%) as very poor, and 5 (0.9%) as

severe. This means that the people suffering from lungs, asthma, and heart disease are breathing uncomfortably 75% of the days of the year in Zahleh, 5.2% of the days, people exposed to air pollution will breathe discomfort, 2.1% of the days, exposed people will start developing respiratory illnesses, and 0.9% of the days, healthy people will be affected by the pollution, and people with diseases will be subjected to serious impacts. Generally, Zahleh is affected by pollution, and the quality of the air in this region is generally not satisfactory, which causes many health problems and diseases.

5.6.1. AQI Prediction and Seasonal decomposition

Prior to beginning the predictive models, we should first decompose the univariate time series. After testing both additive and multiplicative air quality models by using the seasonal decompose function, we noticed that the multiplicative model fitted significantly more efficiently to the data set. The seasonally decompose function yields the following results:

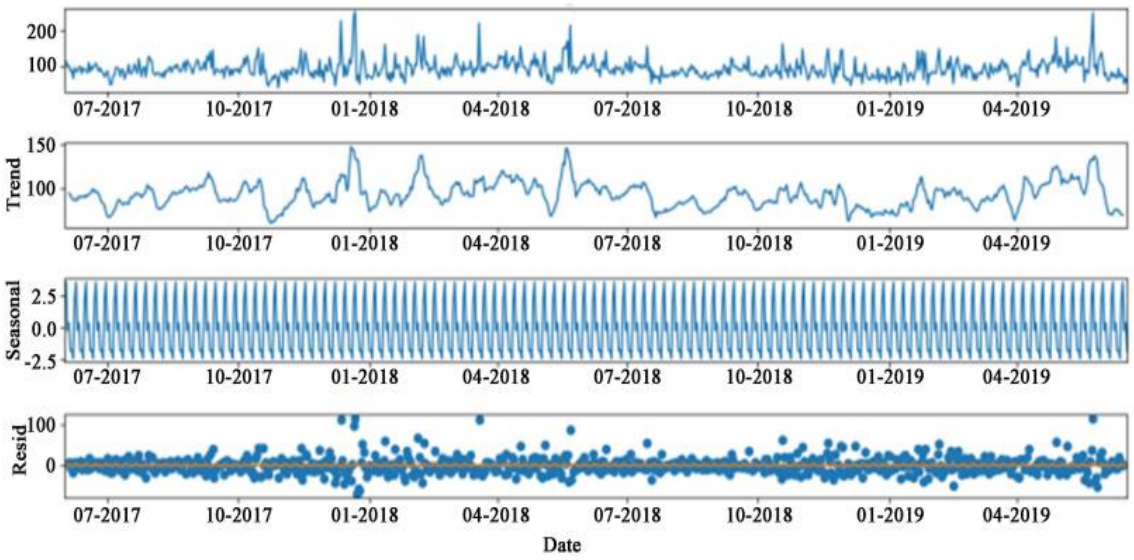


Figure 58 Daily AQI in Zahleh 06-2017 to 12-2018.

These figures above show a variation of the data over time (weekly). In addition, the residuals vary around zero, which verifies the reliability of the predictions and of the fitted model.

5.6.2. Univariate time series forecasting

Air quality forecasting models have two main purposes. First, they validate theoretical knowledge of the atmospheric mechanisms that govern the evolution of air quality and are therefore necessary for research. They also help authorities to monitor the state of the air to which the population and the natural environment are exposed, for the protection of public health and the environment. The predictive models provide the authorities with the necessary information in time to take punctual measures to limit or mitigate pollution peaks in the short term. These measures mainly consist in limiting pollutant emissions through restrictions on traffic (automotive, maritime, airport), industrial activities (production of goods, energy...), or domestic activities (notably heating). Forecasting models also allow permanent monitoring that can guide public policies in terms of land use planning to consider the air quality and improve it. New development projects are being designed to avoid localized concentrations of pollutants. The predictive models can provide an efficient tool to evaluate various scenarios. The Naïve, ETS, TBATS, and SARIMA models have been evaluated using the above measures: RMSE, MAE, and RE.

Measurements	RMSE	MAE	RE
Naïve model	40.44	26.81	19%
Exponential Smoothing	67.80	49.95	36%
TBATS	45.27	24.54	18%
SARIMA	38.04	22.52	16%

Table 29 Summary univariate forecasting models.

In order to evaluate the accuracy, we divided the data into two sets: Training set and test set. Then the test set has been compared with the predicted set, and the following graphs have been obtained:

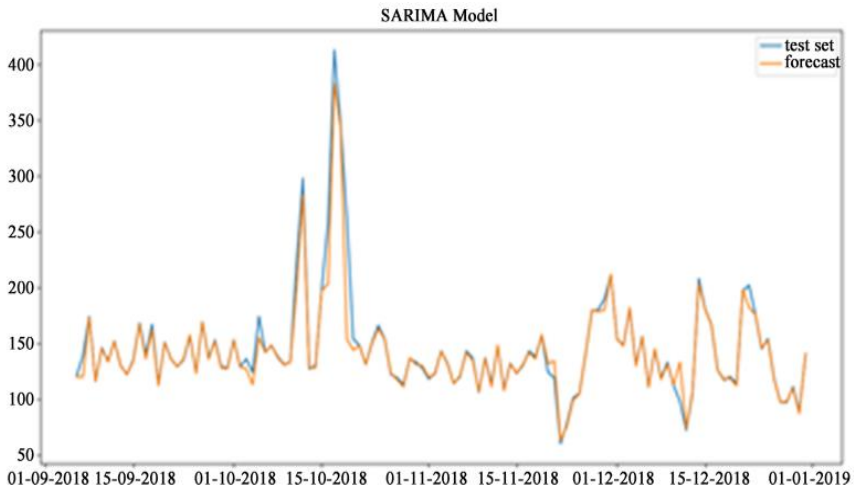


Figure 59 SARIMA model (test set) in Memchiyeh station.

Altogether, the SARIMA model has been shown to be a very appropriate model to study the air pollution level and predict air quality. To construct the SARIMA model, it is necessary to select the hyperparameters for trend and seasonal components of the time series. The identification of hyperparameters can be done either by selecting the parameters directly or by testing several parameters and then selecting the one with the lowest AIC. Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots can be analyzed to find the correlations between the different lag times. An alternative approach for selecting the parameters is to make a loop using and varying *p,d,q,P,D*, and *Q* between two numbers to get the hyperparameters that have the lowest AIC. In this article, we have used the second approach, which turned out to have more accuracy than the former. Also, the results obtained have been $SARIMA(1,1,1)(1,1,1)_{52}$ for two sets of data. Finally, the AQI is predicted by using the SARIMA model, and the results have now been plotted and summarized below.

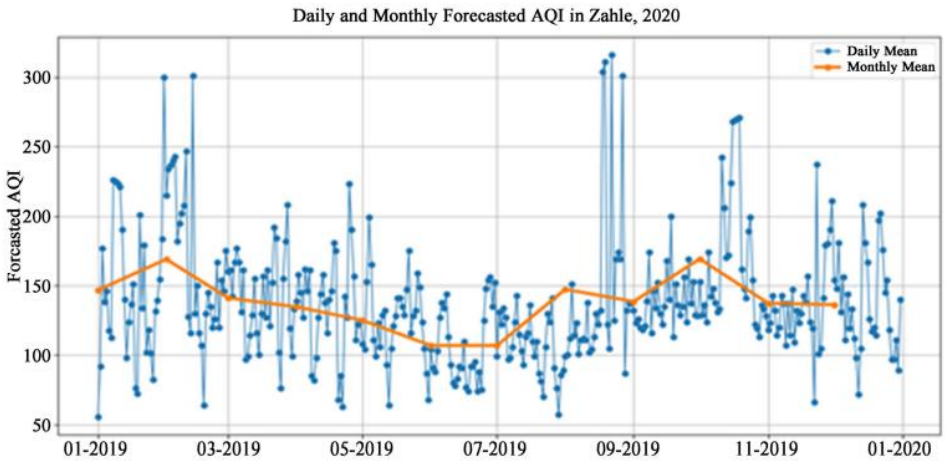


Figure 60 Daily and Monthly forecasted AQI in Memchiyeh station.

For the forecasted air quality index, the results are shown in the following table:

Statistics	Mean	Std	Min	Q_1	Q_2	Q_3	Max
Value	138.07	44.52	55.79	111	131	154	316

Table 30 Data descriptive of forecasted air quality.

It can also be noted that the average air quality value is 138.07 in Zahleh. The lowest value of AQI is 55.79 and the highest is 316 respectively. These findings confirm the earlier ones, that Zahleh is affected by air pollution and has serious problems of health.

5.7. Conclusion

The time series approach was implemented for the air pollutants in Zahleh. The aim was to investigate the correlation between air pollution, pollutants, and certain meteorological conditions, during a specific period. Then the air quality index was calculated and a prediction of the air pollutants and AQI was performed by using the most suitable multivariate time series prediction methods. The findings indicated that the air pollutants frequently exceed the guideline levels, particularly *NO*, NO_2, and NO_X, which reach their maximum at peak periods. Hence, Zahleh is considered to be a very polluted city and it is generally most highly polluted during winter when *NO*, NO_X, SO_2, and *CO* concentrations are higher than during summer. In terms of correlation

with the meteorological components, the Pearson coefficients revealed a strong, moderate, and weak correlation with atmospheric pollution. The air quality index showed that air quality in Zahleh is unhealthy, becoming extremely worse during winter. To forecast the future air quality values, numerous univariate forecast and time series models were tested. The SARIMA model was shown to have the most accuracy for the data available. Note that several factors affect air pollution, including human activities, weather, and topography. Therefore, the predictions can provide some basic information about the air quality in this region but can be susceptible to some change. However, the results showed that there is an urgency to make some decisions concerning the air quality in this city. To begin the process, the removal of manufacturing and burning installations from the residential area could reduce pollution levels and exposure to a highly polluted atmosphere.

General conclusion and Prospects

Air and water are two major natural resources essential for human existence and the continuity of all forms of life on Earth. It is of great importance to understand the consequences of air and water pollution. Addressing such problems is a key element to starting rescue plans. The objectives of this work were to study the physicochemical parameters used to monitor water quality in three stations in the Litany River/Basin, then to model the behavior of these parameters over a period of 10 years from 2008 to 2018 in Qaraoun Lake. Concerning air quality, we aimed to point at extreme pollution events and calculate the probability of future records. Then, we tried to explain the relationship between air pollutants and meteorological factors. Finally, we modeled the AQI in Zahle city from 06-2017 to 12-2018 based on hourly observations during that period which allowed us to predict future values of AQI.

In this work, we introduced new and advanced statistical models to answer several environmental questions related to air and water quality in Lebanon. We discussed the impact of human activities on the quality of air and water, and eventually on the environment, the ecosystem, marine life, and its contribution to climate change. All the modeling and data treatment in this study was done using R-studio and Python. As a first step, we considered different stations at the Litani River which are: Jeb-Jennine, Ghzayel, and Qaraoun Lake. At these stations, 11 physicochemical parameters were monitored in water for 10 years from 2008 to 2018. This data was used to study the relationship between the physicochemical properties of water at the three stations by applying the SBM method. We have noticed that the parameters can be separated into two groups or communities where the parameters that belong to the same group are directly related and can affect one another significantly. However, the SBM can also show the degree of correlation between parameters that belong to different groups, and this is the originality of this model. After observing the matrix of correlation between all the parameters, we noticed that it is affected by the type and degree of pollution in the study area. Hence, the correlation between the parameters was not the same at the three stations. The next step in water quality monitoring was to observe the behavior of the physicochemical parameters in Qaraoun Lake during the same period. We applied the correlation dimension method to detect the existence of chaos. Some parameters showed chaotic behavior, which means that they are random and disordered. The importance of this study was to reconsider

the methods of prediction of the physicochemical parameters and to consider the chaotic behavior of the parameters to reduce the prediction error.

For air quality analysis, the data was collected from a station in Zahle city where seven air quality parameters and four meteorological factors were monitored on an hourly basis from 06-2017 to 12-2018 (13895 observations). Then we applied the extreme records theory to model extreme pollution events and to create an alert system for future extreme records. The next step in air quality analysis was to model the relationship between each pollutant and the other variables. Finally, we predicted the future air quality index. For this purpose, we tested different univariate and multivariate statistical models. The univariate OLS regression was used to model the relationship between the parameters, and the univariate time series forecasting SARIMA was the most suitable for predicting future AQI values.

This study showed the importance of statistical modeling in air and water quality monitoring. The SBM method was used for the first time in water quality monitoring and it contributed to a better understanding of the correlation between the physicochemical parameters and how this correlation changes depending on the main source of pollution at the monitored station. The chaotic model was used to describe the behavior of the physicochemical parameters. Again, it was the first application of the chaotic model in an environmental field. The goal was to determine the parameters that do not follow a specific pattern, which makes them harder to predict. 6 out of 11 parameters showed a chaotic behavior, meaning that they do not obey a general rule; e.g., Season, temperature, etc... Extreme records theory was introduced in this study to model atmospheric pollution. The aim was to point out the importance of extreme levels of pollution and predict the possibility of its occurrence in the future. All the parameters of air pollution were showing extreme events. The extreme records were concentrated at the beginning of the study period and they become much more distant with time. However, there was a high possibility of having new records in the future. Modeling levels of air pollutants using time series in the Zahle region was helpful to understand the situation of air quality in this region. In fact, it was evident air pollutants exceeded the guideline limits in this city. Also, peaks in air pollution were related to daily and seasonal human activities where NO, NO_2, NO_X, and CO concentrations reach their peaks in rush hours. NO, NO_X, SO_2, and CO concentrations increase in the winter and decrease in the summer. But the weekly variation for all pollutants was negligible. AQI of Zahle was modeled to communicate to the public the situation of air quality in the city. It was proof of the previous conclusion where the same results were shown

after plotting the time series of AQI. However, using the AQI is an easier way to inform non-experts about the ambient air quality to apply all the safety measures. The prevision of AQI in Zahle showed no possible improvement in the air quality if no serious measures were taken. However, many factors may impact air pollution, such as human activities, meteorology, and topography. This study provided an important basis for regional air pollution control in Zahle. The presented work has led to important conclusions on the environmental situation in Lebanon. It also opened new questions that may be interesting as future research topics. Three statistical methods in this work were introduced in environmental studies. These methods can be applied to data from different regions and the results can be compared with the present study to validate the model. For example, we can apply the SBM and the chaotic model to data from the river "Seine" in Paris, which is subjected to different types of pollution, but also undergoing an enormous cleaning process. Tracking the progress during this initiative would be very interesting. The results obtained from the chaotic model lead us to think of new ways to predict future levels of pollution. In fact, making a prevision using classical models can result in a considerable error since they are affected by extreme values, contrary to the chaotic model where we can set initial conditions for the parameters (attractors) and make a more reasonable prevision. The chaotic model can be used to predict future levels of pollution in Qaraoun Lake and the results can be compared with the classical models. Air quality monitoring is mandatory for citizens' health. The AQI should be calculated in different regions in Lebanon to set clear safety measures that respect every city's air quality status.

This thesis showed, from the bibliographic study and the newly obtained results, that the impact of human activities, the modification, and the deterioration of the natural ecosystems, become more and more dangerous. Those impacts concern both terrestrial and aquatic ecosystems, therefore globally, the environment. Following this work, several prospects appear, which may provide additional complementary tools for the study of water and air quality. There are several sources of pollution, such as waste discharged from food manufacturing factories, agricultural pesticides and herbicides, and above all wastewater from about 100 villages and cities. It is interesting to treat as well biological parameters like bacterial and phytoplankton studies and other parameters like the concentrations of organic pollutants (dyes, pesticides, drugs...) and heavy metals. These aspects would allow a better understanding of the health risks linked to all these pollutions.

Other studies could be carried out concerning the compositions and the physicochemical properties of the systems which are in contact with the surface water, and the air, where exchanges

occur at the level of the interfaces, like the air and the sediments. In this thesis, we considered the one-dimensional case of extremes and records localized to the Bekaa region to study atmospheric extreme pollution, but it would also be useful to consider a multidimensional and spatial case (several stations) whose interest would be to better understand the variations more globally. This approach would be worth investigating to compare the diversity of pollution between several regions and even between several countries. However, the statistical modeling that we have developed in this research work has helped us to understand the evolution of physicochemical parameters and atmospheric pollutants, but this is not sufficient to build reliable models and compare them. Hence the usefulness to think, in future works, to compare these methods with other physical models such as Navier Stokes.

APPENDIX

####Physicochemical parameters: Classical Cluster Method: PCA, K-means, HAC####

```
install.packages("FactoMineR")
library(FactoMineR)
data=STATION_QUARAOUN_seasonly[1:13,-1:-2]
semi_season=data[1:11,]
semi_season
library(corrplot)
M <- cor(semi_season)
corrplot(M, method = "number", type="upper")
cor(semi_season)
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(M, method="color", col=col(200),  type="upper",   addCoef.col = "black")
corrplot(M, method="color", col=col(200) ,   type="upper",   addCoef.col = "black", order="hclust")
library(ggplot2)
boxplot(semi_season[1:3],vertical = TRUE)
library("FactoMineR")
library("factoextra")
library("FactoMineR")
res.pca <- PCA(semi_season, graph = FALSE)
print(res.pca)
eig.val <- get_eigenvalue(res.pca)
eig.val
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))
var <- get_pca_var(res.pca)
var
```

```
# Coordinates
head(var$coord,12)
# Cos2: quality of representation
head(var$cos2,12)
# Contribution Of PCA
head(var$contrib,12)
# Coordinates of the variables
head(var$coord, 12)
fviz_pca_var(res.pca, col.var = "black")
head(var$cos2, 12)
library("corrplot")
corrplot(var$cos2, is.corr=FALSE)
#Cos2 total of the variables on Dim.1 and Dim.2
fviz_cos2(res.pca, choice = "var", axes = 1:2)
# Coloring according to cos2: quality of representation
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  repel = TRUE)
library("corrplot")
corrplot(var$contrib, is.corr=FALSE)
# Contributions of variables to PC1
fviz_contrib(res.pca, choice = "var", axes = 1, top = 12)
#  Contributions of variables to PC2
fviz_contrib(res.pca, choice = "var", axes = 2, top = 12)
fviz_contrib(res.pca, choice = "var", axes = 1:2, top = 12)
fviz_pca_var(res.pca, col.var = "contrib",  gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"))
fviz_pca_var(res.pca, alpha.var = "contrib")
# Create a continuous random variable of length 10
set.seed (123)
my.cont.var <- rnorm (11)
# Coloring the variables according to the continuous variable
```

```
fviz_pca_var(res.pca, col.var = my.cont.var,  gradient.cols = c("blue", "yellow",
"red"),  legend.title = "Cont.Var")

# Create a cluster variable with kmeans

# Create 3 groups of variables (centers = 3)

set.seed(123)

res.km <- kmeans(var$coord, centers = 3, nstart = 25)

grp <- as.factor(res.km$cluster)

# Coloring the variables by groups

fviz_pca_var(res.pca, col.var = grp, palette = c("#0073C2FF", "#EFC000FF",
"#868686FF"),legend.title = "Cluster")

ind <- get_pca_ind(res.pca)

ind

# Coordinates of the individuals

head(ind$coord)

# Individual Quality

head(ind$cos2)

# Individual contributions

head(ind$contrib)

fviz_pca_ind (res.pca, col.ind = "cos2",  gradient.cols = c("#00AFBB", "#E7B800",
"#FC4E07"), repel = TRUE )

fviz_pca_ind (res.pca, pointsize = "cos2",  pointshape = 21, fill = "#E7B800",  repel =
TRUE )

fviz_pca_ind(res.pca, col.ind = "cos2", pointsize = "cos2",   gradient.cols =
c("#00AFBB", "#E7B800", "#FC4E07"),  repel = TRUE )

fviz_cos2(res.pca, choice = "ind")

# Create a continuous random variable of length 23,

# Same length as the number of active individuals in the PCA

set.seed (123)

my.cont.var <- rnorm(11)

#Coloring of the individuals with the continuous variable

fviz_pca_ind(res.pca, col.ind = my.cont.var,   gradient.cols = c("blue", "yellow",
"red"),  legend.title = "Cont.Var")
```

```
fviz_pca_ind (res.pca)

library(FactoMineR)

library(factoextra)

library(ggpubr)

#Data reduction centering

#To avoid that variables with high variance unduly influence the results

data.cr <- scale(t(newQuaroun[,2:12]),center=T,scale=T)

res.pca <- PCA(data.cr, ncp = 2, graph = FALSE)

res.hcpc <- HCPC(res.pca, graph = FALSE,max=2)

fviz_dend(res.hcpc,  cex = 0.8,  palette = "jco",   rect = TRUE, rect_fill = TRUE,
rect_border = "jco", labels_track_height = 50)

fviz_cluster(res.hcpc,  repel = TRUE,  show.clust.cent = TRUE, palette = "jco",
ggtheme =theme_minimal(), main = "Factor map")

#CAH - Ward's criterion

#Method = « ward.D2 » corresponds to the true Ward's criterion

#Using the square of the distance

cah.ward <- hclust(data.cr,method="ward.D2")

#Diplay dendrog

plot(cah.ward)

#Dendrogram with group materialization

rect.hclust(cah.ward,k=2)

#Split into 2 groups

groupes.cah <- cutree(cah.ward,k=2)

#List of groups

print(sort(groupes.cah))

library(tidyverse)  # data manipulation

library(cluster)    # clustering algorithms

library(factoextra) # clustering algorithms & visualization

distance <- get_dist(data.cr)

fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high =
"#FC4E07"))
```

```
k2 <- kmeans(data.cr, centers = 2, nstart = 25)

str(k2)

k2

fviz_cluster(k2, data = data.cr)

fviz_cluster(k2, data=data.cr,  repel = TRUE,  show.clust.cent = TRUE,  palette =
"jco",  ggtheme = theme_minimal(), main = "Factor map")

set.seed(123)

# function to compute total within-cluster sum of square

wss <- function(k) {

  kmeans(data.cr, k, nstart = 2 )$tot.withinss

}

# Compute and plot wss for k = 1 to k = 15

k.values <- 2:132

# extract wss for 2-15 clusters

wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values, type="b", pch = 19, frame = FALSE, xlab="Number of
clusters K", ylab="Total within-clusters sum of squares")
```

#######CHAOTIC THEORY-PHYSICOCHEMICAL PARAMETERS######

```
library(forecast)

library(itsmr)

ts.quaroun<-ts(data = newQuaroun$PH, start = c(2013, 01), frequency = 12)

require(itsmr)

plot(ts.quaroun, type = "o", col = "forestgreen", main = "time series of pH ", ylab =
"pH", xlab = "periode(month)")

library(scatterplot3d)

quaroun.ts <-ts(newQuaroun[,1],start = c(2008,01),frequency=12)

library(recipes)

library(dplyr)

library(embed)

q<-embed(quaroun.ts)

q
```

```
#####Phase Space######

library(scatterplot3d)

quaroun.ts <-ts(newQuaroun$Temperature,start = c(2008,01),frequency=12)

quaroun.ts

q<-embed(quaroun.ts,3)

#q<-embed(quaroun.ts,m=2,d=1)

q

dim(q)

plot(q, type = "o", col = "forestgreen", xlab = "PO4, X(i),(m=10,T=1)",
ylab="X(i+T)")

scatterplot3d(q, type="l", xlab = "pH",ylab = "pH", zlab = "pH", col.axis =
"blue",main= " Phase space reconstruction in 3 dimensions ")

series <- cbind(seq(1,50),seq(101,150))

head(embedd(series, m=6, d=8))

library(tseriesChaos)

q<-d2(quaroun.ts, m=6, d=8, t=5, eps.min =1)

plot(q)

par(mfrow=c(1,1))

l=0;

for (m in 1:9) {

 l[m]=C2(quaroun.ts,m,d=3,t=5,eps=1)

}

l

revd=rev(l)

revd

plot(revd, type = "o", col = "forestgreen")

head(p)

p<-C2(quaroun.ts, m=6, d=8, t=5, eps = 1)

head(p)

q<-d2(quaroun.ts, m=6, d=8, t=5, eps.min =1)

head(q)
```

```
plot(q)

quaroun.ts <-ts(newQuaroun$Temperature,start = c(2013,01),frequency=12)

fn.out<-false.nearest(quaroun.ts, m= 10, d=1, t=3, eps=3,rt=4)

fn.out[1,]

plot(fn.out[1,],type = "o", col = "forestgreen", main = "False nearest neighbor ", ylab
= "Percentage of FNNs (%)",    xlab = "Embedding dimension")

min(fn.out,na.rm = FALSE)

################################################

#quaroun.ts <-ts(newQuaroun$PO4,start = c(2013,1),frequency=12)

quaroun.ts <-ts(newQuaroun$Temperature[1:109])

quaroun.ts

length(quaroun.ts)

corr<-d2(quaroun.ts, m=9, d=1, t=1, eps.min =1)

head(corr)

length(corr[,1])

plot(corr[1,])

par(mfrow=c(1,2))

l=0;

for (e in 1:50) {

  l[e]=C2(quaroun.ts,m=10,d=1,t=1,eps=e)

}

l

plot(corr, type="o", col="forestgreen",      xlab ="r",ylab = "C(r)" )

plot(l, type="o",main="Saturation of correlation dimension", xlab ="Embedding
Dimension",ylab = "Correlation Exponent" )

output <-lyap_k(quaroun.ts, m=3, d=2, s=200, t=40, ref=1700, k=2, eps=4)

plot(output)

lyap(output, 0.73, 2.47)

#################################################################

#################Coefficient of Correlation################

rcor=cor(corr)
```

```
x=1:9
y=rcor[2:10]
y
plot(x,y,type = "o", col = "forestgreen", xlab="Embedding Dimension", ylab = "Correlation coefficient")
#############################################################
#############Prediction#########################
p=0;
for (i in 1:10) {
  p[i+1]=predict(chaotic[,2],i)
}
p
head(p)
library(forecast)

head(embedd(quaroun.ts, m=10, d=1))
x=corr[,2]
y=1:length(x)
y
cor(x,y,method = "pearson")
plot(x)
cor(t,)
cor(t)
plot(fn.out)
quaroun.ts1 <-ts(newQuaroun$Ammonia[1:109])
quaroun.ts1
quaroun.ts2 <-ts(newQuaroun$Temperature[1:109])
quaroun.ts2
mutual.quaroun<-mutual(quaroun.ts2, partitions = 20, lag.max = 40 ,plot = TRUE)
mutual.quaroun
min(mutual.quaroun,na.rm = FALSE)
```

```
plot(mutual.quaroun, ylab="Percentage of FNNs (%)",main = "False nearest
neighbours")

min(fn.out,na.rm = FALSE)

m=1:length(fn.out1)

fn.out1<-false.nearest(quaroun.ts1, m= 12, d=5, t=1, eps=1, rt=3)

fn.out1

fn.out2<-false.nearest(quaroun.ts2, m= 12, d=3, t=1, eps=1, rt=3)

fn.out2

plot(m, fn.out1, type = "l", col = "red")

plot(m, fn.out2, type = "l", col = "green")

library(tseriesChaos)

scatterplot3d(q, type="l", xlab = "pH",ylab = "pH", zlab = "pH", col.axis =
"blue",main

        = " Phase space reconstruction in 3 dimensions ")

series <- cbind(seq(1,50),seq(101,150))

head(q)

quaroun.ts <-ts(newQuaroun[,2],start = c(2008,01),frequency=12)

plot(quaroun.ts,main="Data original of pH",ylab="pH", xlab="period(month)",ylim

   = c(0,max(newQuaroun[,2])))

quaroun.ts <-ts(newQuaroun[,2],start = c(2013,01),frequency=12)

readline(prompt = "press [enter] to continue")

plot(quaroun.ts, main=" Data in the form of a time series ",ylab="pH",
xlab="temps(mois)",ylim = c(0,max(quaroun.ts)))

print(quaroun.ts)

head(embedd(quaroun.ts, m=6, d=1))

mutual.quaroun<-tseriesChaos(quaroun.ts, partitions = 20, lag.max = 50)

plot(newQuaroun[,3],main="Original data of pH",ylab="pH",
xlab="period(month)",ylim  = c(0, max(newQuaroun[,2])))

readline(prompt = "press [enter] to continue")

quaroun.ts <-ts(newQuaroun[,2],start = c(2008,01),frequency=12)

quaroun.ts

library(tseriesChaos)
```

```
mutual.quaroun<-tseriesChaos::(quaroun.ts, partitions = 20, lag.max = 50 ,plot = TRUE)

library(scatterplot3d)

x <- window(quaroun.ts, start=2008)

xyz <- embedd(x, m=8, d=1)

scatterplot3d(xyz, type="l")

series <- cbind(seq(1,50),seq(101,150))

head(embedd(series, m=6, d=1))

(fn.out <- false.nearest(ts.quaroun, m=8, d=1, t=10, eps=1, rt=3))

plot(fn.out)

library(latex2exp)

library(nonlinearTseries)

library(plot3D)

n.sample=3000

n.sample

h = henon(n.sample= 3000,n.transient= 100, a = 1.4, b = 0.3, start = c(0.73954883, 0.04772637), do.plot = FALSE)

h

h = henon(quaroun.ts,n.transient= 100, a = 1.4, b = 0.3,    start = c(16,30), do.plot = FALSE)

plot( h$x, h$y, pch=".", main="The Henon Attractor")

####Impute missing values###

library(VIM)

library(mice)

impute<-mice(Zahleh_R[,2:8],seed = 500)

summary(impute)

impute$imp$NO

impute$imp$NO2

impute$imp$NOx

impute$imp$CO

impute$imp$SO2

impute$imp$PM2.5
```

```
impute$imp$PM10
Zahleh<-complete(impute,1)
```

####Using Python####

####Importing the packages####

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt  #for plotting
import seaborn as sns  #for swarm plot
from sklearn.metrics import mean_squared_error  #to calculate MSE
from sklearn.metrics import mean_absolute_error  #to calculate MAE
from windrose import WindroseAxes  # to plot WD+WS
import matplotlib.cm as cm   # to plot WD+WS
from sklearn.tree import DecisionTreeRegressor   # for the Decision Tree Regression
model
import statsmodels.api as sm  # for OLS model
from sklearn.ensemble import RandomForestRegressor   # for the Random Forest
Regression model
from xgboost import XGBRegressor   # for the XGBoost model
from statsmodels.tsa.seasonal import seasonal_decompose   # for the seasonal
decomposition
from statsmodels.tsa.holtwinters import ExponentialSmoothing   # for the
Exponential Smoothing model
from tbats import TBATS  # for the tbats model
import itertools
import warnings
warnings.filterwarnings("ignore")
```

####Importing the data####

```
Zahleh = pd.read_excel("C:\THESIS\DATA\qualite dair\Zahleh\hourly-Zahleh.xlsx")
Zahleh = Zahleh.set_index(['Date'])  #set the date as the index for the time series analysis'
```

####Calculate daily average####

```
ZD = Zahleh.resample('1D').mean()
ZD.to_excel("C:\THESIS\DATA\qualite dair\Zahleh\daily avg.xlsx")
```

####Data descriptive####

```
Zdesc = Zahleh.describe()
Zdesc.to_excel("C:\THESIS\DATA\qualite dair\Zahleh\hourly desc.xlsx")
```

####Hourly+weekly+monthly variation####

```
#for each pollutant
sns.swarmplot(data=Zahleh, x=Zahleh.index.hour, y='NO')  #hourly plot
sns.swarmplot(data=Zahleh, x=Zahleh.index.dayofweek, y='NO')  #daily plot
sns.swarmplot(data=Zahleh, x=Zahleh.index.month, y='NO')  #monthly plot
```

####Pearson correlation####

```
SumCorr = summer.corr(method='pearson')
SumCorr.to_excel("C:\THESIS\DATA\qualite dair\Zahleh\SumCorr.xlsx")
WinCorr = winter.corr(method='pearson')
WinCorr.to_excel("C:\THESIS\DATA\qualite dair\Zahleh\WinCorr.xlsx")
SumCorr['NO'].sort_values().to_frame().drop(['NO', 'NO2', 'NOx', 'CO', 'SO2', 'O3', 'PM2.5', 'PM10']).plot.barh()
SumCorr['NO'].sort_values().to_frame().drop(['NO', 'Temp.', ' Wind Direct', 'Wind Speed', 'Pressure', 'Humidity', ' Rain Volume']).plot.barh()
WinCorr['NO'].sort_values().to_frame().drop(['NO', 'NO2', 'NOx', 'CO', 'SO2', 'O3', 'PM2.5', 'PM10']).plot.barh()
WinCorr['NO'].sort_values().to_frame().drop(['NO', 'Temp.', ' Wind Direct', 'Wind
```

```
Speed', 'Pressure', 'Humidity', ' Rain Volume']).plot.barh()   #to plot the correlation between each pollutant and the other variables (once with the other pollutants and once with the meteorological elements) sorting from the highest value to the lowest one
```

####Dividing data into train test and test set####

```
# for each pollutant, we split the data by 80% train set and 20% test set
X11_train = X11[0:463]
X11_test = X11[463:]
y11_train = y11[:463]
y11_test = y11[463:]
```

####Calculate AQI####

```
# AQI data were prepared using Excel
AQI = pd.read_excel("C:\THESIS\DATA\qualite dair\Zahleh\AQI\AQI DATA.xlsx")
AQI = AQI.set_index(['Date'])
# calculating PM2.5 Sub-Index
def get_PM25_subindex(x):
    if x <= 30:
        return x * 50 / 30
    elif x <= 60:
        return 50 + (x - 30) * 50 / 30
    elif x <= 90:
        return 100 + (x - 60) * 100 / 30
    elif x <= 120:
        return 200 + (x - 90) * 100 / 30
    elif x <= 250:
        return 300 + (x - 120) * 100 / 130
    elif x > 250:
        return 400 + (x - 250) * 100 / 130
```

```
    else:
        return 0
AQI["PM2.5_SubIndex"] = AQI['PM2.5'].apply(lambda x: get_PM25_subindex(x))

# Calculating PM10 Sub-Index
def get_PM10_subindex(x):
    if x <= 50:
        return x
    elif x <= 100:
        return x
    elif x <= 250:
        return 100 + (x - 100) * 100 / 150
    elif x <= 350:
        return 200 + (x - 250)
    elif x <= 430:
        return 300 + (x - 350) * 100 / 80
    elif x > 430:
        return 400 + (x - 430) * 100 / 80
    else:
        return 0
AQI["PM10_SubIndex"] = AQI["PM10"].apply(lambda x: get_PM10_subindex(x))

# SO2 Sub-Index calculation
def get_SO2_subindex(x):
    if x <= 40:
        return x * 50 / 40
    elif x <= 80:
        return 50 + (x - 40) * 50 / 40
    elif x <= 380:
        return 100 + (x - 80) * 100 / 300
    elif x <= 800:
        return 200 + (x - 380) * 100 / 420
```

```
    elif x <= 1600:
        return 300 + (x - 800) * 100 / 800
    elif x > 1600:
        return 400 + (x - 1600) * 100 / 800
    else:
        return 0
AQI["SO2_SubIndex"] = AQI["SO2"].apply(lambda x: get_SO2_subindex(x))

#Calculating NOx Sub-Index
def get_NOx_subindex(x):
    if x <= 40:
        return x * 50 / 40
    elif x <= 80:
        return 50 + (x - 40) * 50 / 40
    elif x <= 180:
        return 100 + (x - 80) * 100 / 100
    elif x <= 280:
        return 200 + (x - 180) * 100 / 100
    elif x <= 400:
        return 300 + (x - 280) * 100 / 120
    elif x > 400:
        return 400 + (x - 400) * 100 / 120
    else:
        return 0
AQI["NOx_SubIndex"] = AQI["NOx"].apply(lambda x: get_NOx_subindex(x))

#Calculating CO Sub-Index
def get_CO_subindex(x):
    if x <= 1:
        return x * 50 / 1
    elif x <= 2:
        return 50 + (x - 1) * 50 / 1
```

```
    elif x <= 10:
        return 100 + (x - 2) * 100 / 8
    elif x <= 17:
        return 200 + (x - 10) * 100 / 7
    elif x <= 34:
        return 300 + (x - 17) * 100 / 17
    elif x > 34:
        return 400 + (x - 34) * 100 / 17
    else:
        return 0
AQI["CO_SubIndex"] = AQI["CO"].apply(lambda x: get_CO_subindex(x))

#Calculating NO2 Sub-Index
def get_NO2_subindex(x):
    if x <= 100:
        return x / 2
    elif x <= 200:
        return 50 + ((x-100) / 4)
    elif x > 200:
        return 75 + ((x-200) / 8)
    else:
        return 0
AQI["NO2_SubIndex"] = AQI["NO2"].apply(lambda x: get_NO2_subindex(x))
#AQI bucketing
def get_AQI_bucket(x):
    if x <= 50:
        return "Good"
    elif x <= 100:
        return "Satisfactory"
    elif x <= 200:
        return "Moderate"
    elif x <= 300:
```

```
        return "Poor"
    elif x <= 400:
        return "Very Poor"
    elif x > 400:
        return "Severe"
    else:
        return np.NaN

AQI["Checks"] = (AQI["PM2.5_SubIndex"] > 0).astype(int) + \
                (AQI["PM10_SubIndex"] > 0).astype(int) + \
                (AQI["SO2_SubIndex"] > 0).astype(int) + \
                (AQI["NOx_SubIndex"] > 0).astype(int) + \
                (AQI["NO2_SubIndex"] > 0).astype(int) + \
                (AQI["CO_SubIndex"] > 0).astype(int)

AQI["AQI_calculated"] = round(AQI[["PM2.5_SubIndex", "PM10_SubIndex",
"SO2_SubIndex", "NOx_SubIndex", "NO2_SubIndex", "CO_SubIndex"]].max(axis
= 1))
AQI.loc[AQI["PM2.5_SubIndex"] + AQI["PM10_SubIndex"] <= 0,
"AQI_calculated"] = np.NaN
AQI.loc[AQI.Checks < 3, "AQI_calculated"] = np.NaN

AQI["AQI_bucket_calculated"] = AQI["AQI_calculated"].apply(lambda x:
get_AQI_bucket(x))
AQI[~AQI.AQI_calculated.isna()].head(13)
AQI22 = AQI[~AQI.AQI_calculated.isna()]
AQI22.to_excel("C:\THESIS\DATA\qualite dair\Zahleh\AQI\Zahleh AQI.xlsx")
#Group by AQI label
for_pie = AQI22['AQI_bucket_calculated'].value_counts()
type(for_pie)
for_pie
Name: AQI_bucket_calculated, dtype: int64
```

```
# pie plot
plt.pie(for_pie,
        labels=list(for_pie.index),
        colors=['green', 'yellow', 'limegreen', 'orange'], autopct='%1.1f%%')
plt.title('AQI in Beirut, 06-2017/06-2019')
```

####Plotting daily+monthly variation for AQI####

```
plt.title('Daily and Monthly Calculated AQI in Zahleh, 06-2017 to 2018', fontsize=15,
y=1.05)
plt.xlabel("Date")
plt.ylabel("Calculated AQI")
plt.plot(ZAQI['Date'], ZAQI['AQI_calculated'], '-o', ms=5, lw=1, alpha=1,
label='Daily Mean')
plt.plot(monthly_ZAQI['Date'], monthly_ZAQI['AQI_calculated'], '-o', ms=5, lw=3,
alpha=1, label='Monthly Mean')
plt.legend()
plt.grid()
plt.show()
Seasonal Decompose:
SD1 = seasonal_decompose(y1)
SD1.plot()
plt.show()
Forecasting AQI:
train1 = y1[:463]
test1 = y1[463:]
```

####Naïve model####

```
dd = np.asarray(train1)
y1_hat = train1.copy()
y1_hat['naive'] = train1[len(dd)-1]
```

####ETS model####

```
ETS_model1 = ExponentialSmoothing(train1, trend='add', seasonal='add',
seasonal_periods=52, damped_trend=True)
ETS_fit1 = ETS_model1.fit()
ETS_pred1 = ETS_fit1.forecast(116)
```

####TBATS model####

```
tbats_model1 = TBATS(seasonal_periods=[7, 365])
tbats_fit1 = tbats_model1.fit(train1)
tbats_pred1 = tbats_fit1.forecast(116)
```

####SARIMA model####

```
#Finding the most accurate parameters
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 52) for x in list(itertools.product(p, d, q))]
print('Examples of parameter for SARIMA...')
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[4]))
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod =
sm.tsa.statespace.SARIMAX(y1,order=param,seasonal_order=param_seasonal,enforc
e_stationarity=False,enforce_invertibility=False)
            results = mod.fit()
            print('ARIMA{}x{}52 - AIC:{}'.format(param, param_seasonal, results.aic))
        except:
            continue
#Using the optimal option
sarima_model1 = sm.tsa.statespace.SARIMAX(y1, order=(1, 1, 1),
seasonal_order=(1, 1, 1, 52),
```

```
                    enforce_stationarity=False, enforce_invertibility=False)
resul1 = sarima_model1.fit()
print(resul1.summary().tables[1])
#Diagnostic
resul1.plot_diagnostics(figsize=(18, 8))
plt.show()
#Prediction
sarima_pred1 = resul1.get_prediction(start=pd.to_datetime('2018-09-06'),
end=pd.to_datetime('2018-12-31') , dynamic=False)
```

####Plotting wind direction + wind speed####

```
#To visualize wind speed and wind direction of Zahleh
fig = plt.figure(figsize=(12,6.5))
wd2 = ZD['Wind Direct']
ws2 = ZD['Wind Speed']
ax = fig.add_subplot(121, projection="windrose", )
bins = np.arange(0,12,2)
ax.bar(wd2, ws2, normed=False, blowto=True, edgecolor='white', bins=bins,
cmap=cm.viridis_r)
ax.set_legend()
#To present the graphs side-by-side
fig = plt.figure(figsize=(12,6.5))
ax = fig.add_subplot(121, projection="windrose", )
ax.contourf(wd, ws, bins=np.arange(0, 12, 2), cmap=cm.viridis_r, alpha=0.7,
blowto=True)
ax.legend(bbox_to_anchor=(1.2, 0), title='Speed (m/s)')
ax1 = fig.add_subplot(122, projection="windrose")
ax1.contourf(wd2, ws2, bins=np.arange(0, 12, 2), cmap=cm.viridis_r, alpha=0.7,
blowto=True)
ax1.set_title('Memshieh garden-Zahleh', y=1.1)
```

#########AIR QUALITY INDEX-GEV-POT######

```
library(missForest)
library(foreach)
library(itertools)
library(iterators)
library(gapminder)
library(tidyverse)
library(skimr)
library(Hmisc)
library(ggplot2)
library(Formula)
library(survival)
library(lattice)
library(dplyr)
aligning_plots_packages <- c("gridExtra", "grid")
data_exploration_packages <- c("tidyverse", "plotly", "openxlsx")
table_formating_packages <- c("knitr","kableExtra")
descriptive_statistics_packages <- c("table1","arsenal","pastecs")
air_quality_packages <- c("openair", "worldmet")
NA_treatment_packages <- c("mice", "VIM")
if (!require(install.load)) {install.packages("install.load")}
install.load::install_load(c(aligning_plots_packages,data_exploration_packages,table_
formating_packages,
descriptive_statistics_packages,air_quality_packages,NA_treatment_packages))
summary(dataairquality)
matrixplot(dataairquality, sortby = 2, ylim = c(0,900), font.axis = 4)
#Number of missing values by columns
colSums(is.na(dataairquality))
# Rate of missing values in data set
sum(is.na(dataairquality))/(nrow(dataairquality)*ncol(dataairquality))
library(VIM)
```

```
#a scatterplot with additional information on the missing values
par(
  # Change the colors
  col.main = "#336600", col.lab = "#0033FF", col.axis = "#333000",
  # Titles in italic and bold
  font.main = 4, font.lab = 4, font.axis = 4,
  # Change font size
  cex.main = 1.2, cex.lab = 1, cex.axis = 1
)
marginplot(dataairquality[,c("PM10","PM2.5")],pch = 16 , cex = 1.5 ,numbers = T,
xlim = c(0,800), ylim = c(0,400), main = "Scatterplot with missing values
information", xlab = "Hourly PM10 Concentration ", ylab = "Hourly PM2.5
Concentration ")
# Rate of missing values for PM10
sum(is.na(dataairquality$PM10))/(nrow(dataairquality)*ncol(dataairquality))
sum(is.na(dataairquality$PM2.5))/(nrow(dataairquality)*ncol(dataairquality))
vis_miss(dataairquality)
gg_miss_upset(dataairquality)
gg_miss_var(dataairquality)
dat_miss<- data.frame(dataairquality$Date,
dataairquality$SO2,dataairquality$NO,dataairquality$NO2,dataairquality$NOx,dataai
rquality$O,dataairquality$PM10,dataairquality$PM2.5)
dat_miss<- data.frame( dataairquality$SO2,dataairquality$NO,dataairquality$NO2)
dat_miss
dat_impo <- missForest(dat_miss)
class(dat_impo)
new_data <- dat_impo$ximp
new_data
summary(new_data)
set.seed(123)
library(extRemes)
library(xts)
```

```
plot(new_data$dataairquality.SO2,
    pch=12,

    col="#69b3a2",
    xlab="Time/Hours", ylab="SO2"
)
plot(density(new_data$dataairquality.SO2)
fit_mle <- fevd(new_data$dataairquality.SO2, method = "MLE", type="GEV")
fit_mle
ci(fit_mle,type="parameter")
plot(fit_mle)
rl_mle <- return.level(fit_mle, conf = 0.05, return.period= c(15,30,40,50))
rl_mle
plot(rl_mle)
#########################################################################
```

References

A. Farrow K. A. Miller, and L. Myllyvirta Toxic air: The price of fossil fuels [Book] / ed. Africa Greenpeace Middle East and North. - June 2020. - p. 35.

A.Hayek N. Tabaja, S.A. Andaloussi, J. Toufaily, E. GarnieZarli, A. El Toufaili and T. Hamieh Evaluation of the Physico-Chemical Properties of the Waters on the Litani River Station Quaraoun [Journal] // American Journal of Analytical Chemistry. - 2020. - Vol. 11. - pp. 90-103.

A.Shaban L.Telesca, T.Darwich Analysis of long-term fluctuations in stream flow time series: An application to Litani River, Lebanon [Journal] // Acta Geophys.. - 2014. - Vol. 62. - pp. 164–179.

Abarbanel H. [et al.] Statistics of Extreme Events with Application to Climate [Book]. - [s.l.] : JASON, JSR, 1992.

Abarbanel H.D.I. Analysis of Observed Chaotic Data [Book]. - New York : Springer-Verlag, 1996. - p. 272.

Abdallah C. [et al.] A first annual assessment of air quality modeling over Lebanon using WRF/Polyphemus [Journal] // Atmos. Pollut. Res.. - 2018. - Vol. 9. - pp. 643-654.

Abdallah C., Sartelet K. and Afif C. Influence of boundary conditions and anthropogenic emission inventories on simulated O3 and PM2.5 concentrations over Lebanon [Journal] // Atmospheric Pollution Research. - 2016. - ISSN 1309-1042. - pp. 971-979.

Abdallah Charbel, Sartelet K and Afif Charbel Influence of boundary conditions and anthropogenic emission [Journal]. - [s.l.] : Atmospheric pollution research, 2016.

Abou El-Enin H.S. Essay on the geomorphology of Lebanon [Journal]. - [s.l.] : Hanno, 1980.

AbuKhalil Asaad Geography [Book]. - Washington, D.C : In Collelo, Thomas (ed.) Lebanon: a country study, 1989. - pp. 42-48.

Afif C., Dutot A.L. and Jambert C. Statistical approach for the characterization of NO2 concentrations in Beirut [Journal] // Air Qual Atmos Health. - 2009. - Vol. 2. - pp. 57-67.

Agrawal Anju, Pandey Ravi S and Sharma Bechan Water Pollution with Special Reference to Pesticide Contamination in India [Journal]. - [s.l.] : Journal of Water Resource and Protection, 2010. - Vol. 2.

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

Alvim R.L.R. [et al.] Time series analysis of air pollution data [Journal] // Salcedo Atmos. Environ.. - 1999. - Vol. 33. - pp. 2361-2372.

Ana H. L. F. Extreme Value Theory: An Introduction [Book]. - [s.l.] : Springer, 2007.

Anderson C. J., Wasserman, S., and Faust, K. Building stochastic block models [Journal] // Social Networks. - 1992. - Vol. 14. - pp. 137-161.

Arnaudo E., Farasin A. and Rossi C. A Comparative Analysis for Air Quality Estimation from Traffic and Meteorological [Journal] // Data. Appl. Sci.. - 2020. - 4587 : Vol. 10.

Arnold N., Balakrishnan N. and Nagaraja H.N. Records [Book]. - New York : Wiley, 2011.

Atoui A. Sami A.A., Slim K., Moilleron R., Khraibani Z. Prediction and analysis of the Extreme and Records values of air pollution data in Bekaa valley in Lebanon [Journal] // Preprint submitted to Elsevier. International Journal of Environmental Science and Development. - 2022. - ISSN: 2010-0264.

Aulenbach Donald B Water—Our Second Most Important Natural Resource [Journal]. - [s.l.] : Boston college law review, 1968. - 3 : Vol. 9.

Baalbaki R. El Hage R., Nassar J., Gerard J and Sabila N. B. Zaarour R. Abboud M., Fara W., Khalaf L. K., Shihadeh A. L., Saliba N. A. Exposure to Atmospheric PMs, PAHs, PCDD/Fs and Metals near an Open Air Waste Burning Site in Beirut [Journal] // Lebanese Science Journal. - 2016. - Vols. 17.10.22453/LSJ-017.2.091103.

Baayoun Abdelkader [et al.] Emission inventory of key sources of air pollution in Lebanon [Journal]. - 2019.

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

Bavel. J.Van The world population explosion: causes, backgrounds and projections for the future [Journal] // Facts, views vision in ObGyn. - 2013. - 4 : Vol. 5. - p. 281.

Beil Laura Pollution killed 9 million people in 2015 [Report]. - [s.l.] : ScinceNews, 2017.

Bencala K. E., Seinfeld, J. H. On frequency distribution of air pollutant concentrations [Journal] // Atmospheric Environment. - 1976. - Vol. 10. - pp. 941–950.

Biswas Asit K and Tortajada Cecilia Water quality management: a globally neglected issue [Journal]. - [s.l.] : INTERNATIONAL JOURNAL OF WATER RESOURCES DEVELOPMENT, 2019.

Blagrave Kevin and Sharma Sapna Heatwaves and storms contribute to degraded water quality conditions in the nearshore of Lake Ontario [Journal]. - [s.l.] : Journal of Great Lakes Research, 2022.

Blei D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation [Journal] // The Journal of Machine Learning Research. - 2003. - Vol. 3. - pp. 993-1022.

Bradley P.S., Fayyad, U.M., Reina, C. Scaling clustering algorithms to large databases [Journal] // KDD'98 proceedings of the fourth international conference on knowledge discovery and data mining. - 1998.

Bühlmann D. J. Stekhoven and P. MissForest—non-parametric missing value imputation for mixed-type data [Journal] // Bioinformatics. - 2012. - Vol. 28. - pp. 112-118.

Castillo Enrique Extreme value theory in engineering [Book]. - New York : Academic Press, Inc., 1988.

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

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 [Journal] // Physics Procedia. - 2014. - Vol. 55. - pp. 279-284.

Chandler K. N. The distribution and frequency of record values [Journal] // Journal of the Royal Statistical Society. - 1952. - Vol. 14. - pp. 220-228.

Choudhary Mahendra Pratap and Garg Vaibhaw Causes, Consequences and Control of Air Pollution [Conference] // All India Seminar on Methodologies for Air Pollution Control. - [s.l.] : Malviya National Institute of Technology, 2013.

Cohen N. Balakrishnan and A. C. Order Statistics and Inference: Estimation Methods [Book]. - San Diego : Academic Press, 1991.

Daniel G., Gerhard W. and Krzysztof S. Integrating river hydromorphology and water quality into ecological status modelling by artificial neural networks [Journal] // Water Research. - 2018. - ISSN 0043-1354 : Vol. 139. - pp. 395-405.

Daudin J., Picard, F., and Robin, S. A mixture model for random graph [Journal] // Statistics and Computing. - 2008. - Vol. 18. - pp. 1-36.

DAVID H. A. Order Statistics, Second edition [Book]. - New York : John Wiley L·Sons, 1981.

De haan Lauren and Ferreira Ana Extreme Value Theory: An Introduction [Book]. - [s.l.] : Springer Series in Operations Research and Financial Engineering, 2010.

Dudgeon D Angela HA, Gessner MO, Kawabata ZI, Knowler DJ, Leveque C, Naiman RJ, Prieur RAH, Soto D, Stiassny MLJ, Sullivan CA Freshwater biodiversity: importance, threats, status, and conservation challenges [Journal]. - Biological Reviews of the Cambridge Philosophical Society : [s.n.], (2006) . - 2 : Vol. 81. - pp. 163-181.

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 [Journal] // Communication in Statistics Simulation and Computation. - 2020.

EPA Air Quality Index, A guide to air quality and your health [Report]. - 2014.

EPA Protecting water quality from agricultural runoff [Report]. - Washington : U.S. Environmental Protection Agency, 2005.

Esterby S.R. Trend analysis methods for environmental data [Journal] // Environmetrics. - 1993. - Vol. 4. - pp. 459-481.

Estrada-Rivera Andrés [et al.] The Impact of Urbanization on Water Quality: Case Study on the Alto Atoyac Basin in Puebla, Mexico [Report]. - [s.l.] : Sustainability, 2022.

Evans Alexandra EV and Ippolito Alessio Agricultural water pollution: key knowledge gaps and research needs [Journal]. - [s.l.] : Current Opinion in Environmental Sustainability, 2019. - Vol. 36.

Fadel A., Atoui, A., Lemaire, B. J., Vinc¸on-Leite, B., Slim, K. Environmental factors associated with phytoplankton succession in a Mediterranean reservoir with a highly fluctuating water level [Journal] // Environmental Monitoring and Assessment. - 2015. - 10 : Vol. 187. - p. 633.

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 [Journal] // Environ Earth Sci . - 2021. - 65 : Vol. 80.

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 [Journal] // Ecohydrology Hydrobiology. - 2019. - 3 : Vol. 19. - pp. 359-369.

Fadel Ali [et al.] First assessment of the ecological status of Karaoun reservoir, Lebanon. [Journal] // Lakes and Reservoirs: Research & Management. - 2014. - pp. 142-157.

Fadlallah A.R. Lebanon geographical study [Journal]. - Beirut : The Renaissance House, 2001.

Farah W., Nakhle M. M. and Abboud M. Time series analysis of air pollutants in Beirut, Lebanon [Journal] // Environ Monit Assess.. - 2014. - Vol. 186. - pp. 8203-8213.

Farah W., Nakhlé, M.M., Abboud, M. et al. Time series analysis of air pollutants in Beirut, Lebanon [Journal] // Environ Monit Assess.. - 2014. - Vol. 186. - pp. 8203–8213.

Fortunato S. Community detection in graphs [Journal] // Physics Reports,. - 2010. - Vol. 486. - pp. 75-174.

Fraser A.M. and Swinney H.L. Independent coordinates for strange attractors from mutual information [Journal] // Phys. Rev. A. - 1986. - 2 : Vol. 33. - pp. 1134-1140.

Fuzzi S [et al.] Particulate matter, air quality and climate: lessons learned and future needs [Journal] // Amospheric Chemistry and Physics. - 2015.

Gareth J. An Introduction to Statistical Learning with Applications in R [Book]. - 2013.

Ghanem D.A. Energy, the city and everyday life: living with power outages in post-war Lebanon [Journal] // Energy Res. Soc. Sci. . - 2018. - Vol. 36. - pp. 36-43.

Gonzalez-Abraham R., Chung, S. H., Avise, J., Lamb, B., Salathé Jr., E. P., Nolte, C. G., Loughlin, D., Guenther, A., Wiedinmyer, C., Duhl, T., Zhang, Y., and Streets, D. G. [et al.] The effects of global change upon United States air quality [Journal] // Atmospheric Chemistry and Physics. - 2015. - pp. 12645-12665.

Grassberger P. Procaccia I. Measuring the strangeness of strange attractors. [Journal] // Physica D: Nonlinear Phenomena,. - 1983. - 1-2 : Vol. 9. - pp. 189-208.

Grennfelt Peringe [et al.] Acid rain and air pollution: 50 years of progress in environmental science and policy [Journal]. - [s.l.] : AMBIO A Journal of the Human Environment, 2019.

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 Sao Paulo State, Brazil. [Journal] // River Res. Applic.. - 2008. - 8 : Vol. 24.

Gumbel E. J. On the frequency distribution of extreme values in meteorological data [Journal] // Bull. Amer. Meteor. Soc.. - 1942. - Vol. 23. - pp. 96-105.

Haan A. A. Balkema and L. Residual life time at great age [Journal] // The Annals of Probability. - 1974. - Vol. 2. - pp. 792-804.

Hamanaka R. B., Mutlu, G. M. Particulate matter air pollution: effects on the cardiovascular system [Book]. - 2018. - Vol. 9 : p. 680.

Haydar C.M., Nehme N. and Hamieh T. Water Quality of the Upper Litani River Basin, Lebanon. [Journal]. - [s.l.] : Physics Procedia, 2014.

Haydar Chaden Moussa [et al.] Water Quality of the Upper Litani River Basin, Lebanon [Conference] // Eighth International Conference on Material Sciences, CSM8-ISM5. - 2014.

Hayek A. [et al.] Evaluation of the Physico-Chemical Properties of the Waters on the Litani River Station Quaraoun. [Journal]. - [s.l.] : American Journal of Analytical Chemistry, 2020.

Hayek A. [et al.] Multivariate Spatial and Temporal Analysis to Study the Variation of Physico-Chemical Parameters in Litani River, Lebanon. American Journal of Analytical Chem [Journal]. - [s.l.] : American Journal of Analytical Chemistry, 2021.

Hayek Ali [et al.] Analysis of the extreme and records values for temperature and precipitation in Lebanon [Journal]. - [s.l.] : Communications in Statistics: Case Studies, Data Analysis and Applications, 2020.

Hayek Ali L'impact environnemental du changement climatique, une corrélation avec le cycle hydrologique et la concentration de la pollution dans la région Litani [Report]. - 2021.

Hayek Ali L'impact environnemental du changement climatique, une corrélation avec le cycle hydrologique et la concentration de la pollution dans la région Litani [Report]. - Paris : Thèse de doctorat en Sciences et Techniques de l'Environnement, 2021.

Hoayek A.S., Ducharme G.R. and Khraibani Z. Distribution-free inference in record series [Journal] // Extremes. - 2017. - Vol. 20. - pp. 585-603.

Holland P. W., Laskey, K. B., and Leinhardt S. Stochastic blockmodels: First steps [Journal] // Social Networks. - 1983. - Vol. 5. - pp. 109-137.

Holyst J.A. [et al.] How to control chaotic economy? [Journal] // Journal of Evolutionary Economics. - 1996. - Vol. 6. - pp. 31-42.

Holzfuss J. and Mayer-Kress G. An approach to error-estimation in the application of dimension algorithms [Journal] / ed. Systems. Springer Dimensions and Entropies in Chaotic. - New York : In: Mayer-Kress, G. (Ed.), 1986. - pp. 114-122.

Huang Shu-Li, Yeh Chia-Tsung and Chang Li-Fang The transition to an urbanizing world and the demand for natural resources [Journal]. - [s.l.] : Current Opinion in Environmental Sustainability, 2010. - 3 : Vol. 2. - pp. 136-143.

Hubert L. and Arabie, P. Comparing partitions [Journal] // Journal of Classification. - 1985. - Vol. 2. - pp. 193-218.

Ikeguchi T., Iokibe T. and Aihara K. Chaos and Time Series Analysis. In: Liu ZQ., Miyamoto S. (eds) Soft Computing and Human-Centered Machines [Book] / ed. Science Workbench Computer. - Tokyo : Springer, 2000.

Isiyaka H.A., Mustapha A. and Juahir H. Water quality modelling using artificial neural network and multivariate statistical techniques [Journal] // Model. Earth Syst. Environ.. - 2019. - Vol. 5. - pp. 583-593.

Jaakkola T. S., and Jordan, M. I. Bayesian parameter estimation via variational methods. [Journal] // Statistics and Computing. - 2000. - Vol. 10. - pp. 25–37.

Jacob Daniel J. and Winner Darrel A. Effect of climate change on air quality [Journal] // Atmospheric Environmentd. - 2009. - pp. 51-63.

Jebin Ahmed, Thakur Abhijeet and Goyal Arun Industrial wastewater and its toxic effects [Journal]. - [s.l.] : royal society of chemistry, 2021.

Jolliffe I. Principal Component Analysis [Book]. - New York : Springer-Verlag, 1986.

Jordan M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models [Journal] // Machine learning. - 1999. - Vol. 37. - pp. 183–233.

Kahraman Cengiz Risk analysis and crisis response [Journal]. - 2009. - 4 : Vol. 23. - pp. 413-414.

Kaminska J. A The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wroc law [Journal] // Journal of environmental management. - 2018. - Vol. 217. - pp. 164-174.

Kayes I., Shahriar, S. A., Hasan, K., Akhter, M., Kabir, M. M., Salam, M. A. The relationships between meteorological parameters and air pollutants in an urban environment [Journal] // Global Journal of Environmental Science and Management. - 2019. - 3 : Vol. 5. - pp. 265-278.

Kennel M., Brown R. and Abarbanel H.DD.I. Determining embedding dimension for phase-space reconstruction using a geometrical construction [Journal] // Phys. Rev. A. - 1992. - Vol. 45. - pp. 3403-3411.

Khraibani Z. [et al.] Climate change, agriculture and economic growth in Lebanon: A VAR approach [Journal] // American Journal of Economics. - 2020. - Vol. 10. - pp. 126-131.

Khraibani Zaher [et al.] Climate Change, Agriculture and Economic Growth in Lebanon: A VAR Approach [Journal]. - [s.l.] : American Journal of Economics, 2020.

Khraibani Zaher, Badran Hussein M. and Khraibani Hussein Records Method for the Natural Disasters Application to the Storm Events [Journal]. - [s.l.] : Journal of environmental sciences, 2011.

Kibria S. Gulati and F. G. B. M. G. Analysis of hurricane extremes and record values in the Atlantic [Journal] // Communications in Statistics: Case Studies, Data Analysis and Applications. - 2019. - 2 : Vol. 5. - pp. 101-110.

Kılıç Zeyneb The importance of water and conscious use of [Journal]. - [s.l.] : MedCrave, 2020. - 5 : Vol. 4.

Kowalkowski T. [et al.] Application of chemometrics in river water classification [Journal]. - [s.l.] : Water research, 2006.

Kowalska Małgorzata [et al.] Air quality index and its significance in environmental health risk communication [Journal]. - [s.l.] : Archives of Environmental Protection, 2009.

Kumar A., Goyal, P. Forecasting of daily air quality index in Delhi [Journal] // Science of the Total Environment. - 2011. - 24 : Vol. 409. - pp. 5517-5523.

Kumar Subodh Acid Rain-The Major Cause of Pollution:Its Causes, Effects [Journal]. - [s.l.] : International Journal of Applied Chemistry, 2017.

Kumar T. S., Das, H. S., Choudhary, U., Dutta, P. E., Guha, D. Laskar, Y. Analysis and Prediction of Air Pollution in AssamUsing ARIMA/SARIMA and Machine Learning [Book] / ed. Springer. - Singapore : In Innovations in Sustainable Energy and Technology, 2021. - pp. 317-330.

Kushawaha Jyoti and Sivaiah Borra Abhishek Kumar Kushawaha,Gurudatta Singh4 and Pardeep Singh Climate change and its impact on natural resources [Journal]. - [s.l.] : Water Conservation in the Era of Global Climate Change, 2021.

Kusum Farswan Current Opinion in Environmental Sustainability [Journal]. - [s.l.] : Asian Journal of Research in Social Sciences and Humanities, 2021. - 12 : Vol. 11.

Latouche P., Birmele, E., and Ambroise, C. Overlapping stochastic block models with application to the French political blogosphere [Journal] // Annals of Applied Statistics. - 2011. - Vol. 5. - pp. 309-336.

Latouche P., Birmele, E., and Ambroise, C. Variational Bayesian inference and complexity control for stochastic block models [Journal] // Statistical Modelling. - 2012. - Vol. 12. - pp. 93-115.

Lee C. K. Multifractal characteristics in air pollutant concentration time series [Journal] // Water, Air, and Soil Pollution. - 2002. - Vol. 135. - pp. 389–409.

Lee C.K. [et al.] Simple multifractal cascade model for air pollutant concentration (APC) time series [Journal] // Environmetrics. - 2003. - 2 : Vol. 14. - pp. 255-269.

Lee K. C. Multifractal characteristics in air pollutant concentration time series [Journal] // Water Air Soil Poll.. - 2002. - Vol. 135. - pp. 389-409.

Leibert W. and Schuster H.G. Proper choice of the time delay for the analysis of chaotic time series [Journal] // Phys. Lett.,. - 1989. - Vol. 141. - pp. 386-390.

Lelieveld J. [et al.] Severe ozone air pollution in the Persian Gulf region [Journal] // Atmos. Chem. Phys.. - 2009. - Vol. 9. - pp. 1393-1406.

Lelieveld J., Berresheim, H., Borrmann, S., Crutzen, P. J., Dentener, F. J., Fischer, H., Feichter, J., Flatau, P. J., Heland, J., Holzinger, R., Korrmann, R., Lawrence, M. G., Levin, Z., Markowicz, K. M., Global air pollution crossroads over the Mediterranean [Journal] // Science. - 2002. - Vol. 298. - pp. 794-799.

Lelieveld Jos [et al.] Loss of life expectancy from air pollution compared to other risk factors: aworldwide perspective [Journal]. - [s.l.] : European society of cardiology, 2020.

Lorenz E.N. Deterministic nonperiodic flow [Journal] // Journal of the Atmospheric Sciences. - 1963. - Vol. 20. - pp. 130-141.

Lorenz E.N. Deterministic nonperiodic flow [Journal] // Journal of the Atmospheric Sciences. - 1963. - pp. 130-141.

Lorenz E.N. Deterministic nonperiodic flow. [Journal]. - [s.l.] : Journal of the Atmospheric Sciences, 1963.

Lyon F IARC monographs on the evaluation of carcinogenic risks to humans [Journal] // Some industrial chemicals. - 1994. - Vol. 60. - pp. 389-433.

M. R. Leadbetter G. Lindgren, and H. Rootzén Extremes and Related Properties of Random Sequences and Processes [Book]. - New York : Springer Verlag, 1983.

Marcus J. Schwartz and A. Mortality and air pollution in London: A time series analysis [Journal] // American Journal of Epidemiology. - 1990. - Vol. 31. - pp. 85-194.

Mariadassou M., Robin, S., and Vacher, C. Uncovering latent structure in valued graphs: a variational approach [Journal] // Annals of Applied Statistics. - 2010. - Vol. 4. - pp. 715-742.

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

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

Mcheik Amale [et al.] Effect of Irrigation Water Quality on the Microbial Contamination of Fresh Vegetables in the Bekaa Valley, Lebanon [Report]. - 2018.

MoE National environmental action plan [Report]. - Beirut, Lebanon : Ministry of Environment, 2020.

Mosley Luke M. Drought impacts on the water quality of freshwater system; review and integration [Journal] // Earth-Science Reviews. - 2015. - pp. 203-214.

Mustapha Adamu [et al.] Temporal Aspects of Surface Water Quality Variation Using Robust Statistical Tools [Journal]. - [s.l.] : The Scientific World Journal, 2012. - Vol. 2012.

Nassif Nadine, Maatouk Imane and Saliba Rachad Environmental Assessment and Legal Approach of Qaraoun Artificial Lake, Lebanon [Journal]. - 2015.

Nehme Nada [et al.] Assessment of the physicochemical and microbiological water quality of Al-Zahrani River Basin, Lebanon. [Journal]. - [s.l.] : Jordan Journal of Earth and Environmental Sciences, Hashemite University, 2021.

Ng T.L.J., Murphy, T.B. Weighted stochastic block model [Journal] // Stat Methods Appl . - 2021. - Vol. 30. - pp. 1365-1398.

Niemeyer L.E. Forecasting air pollution potentia [Report]. - [s.l.] : Monthly Weather Review, 1960.

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

Ochoa-Hueso R. [et al.] Ecological impacts of atmospheric pollution and interactions with climate change in terrestrial ecosystems of the Mediterranean Basin: current research and future directions [Journal] // Environ. Pollut.. - 2017. - Vol. 227. - pp. 194-206.

Oregon Department of Environmental Quality Wildfire Smoke Trends and the Air [Report]. - 2021.

Otayek Alain Saroufim and Elie Analysis and interpret road traffic congestion costs in Lebanon [Journal] // MATEC Web Conf.. - 2019. - 02007 : Vol. 295.

Owa F. W. Water pollution: sources, effects, control and management [Journal]. - [s.l.] : International Letters of Natural Sciences, 2013.

Pasquill F.C. and Smith F.B. Atmospheric Diffusion, 3rd edition [Journal]. - [s.l.] : Ellis Horwood, Chichester., 1983.

Pickands J. I. Statistical inference using extreme value order statistics [Journal] // The Annals of Statistics. - 1975. - Vol. 3. - pp. 119-131.

Pimentel David and BONNIE BERGER DAVID FILIBERTO, MICHELLE NEWTON, BENJAMIN WOLFE, ELIZABETH Water Resources: Agricultural and Environmental Issues [Journal]. - [s.l.] : Bioscience, 2004.

Piotrowski C. Covid-19 and Chaos Theory: Applications based on a bibliometric analysis [Journal] // Journal of Projective Psychology Mental Health. - 2020. - 2 : Vol. 25. - pp. 1-5.

Pope C. A. Respiratory health and PM10 pollution-a daily time series analysis [Journal] // American Review of Respiratory Disease. - 1991. - Vol. 144. - pp. 668-674.

Postel S.L. and Ehrlich P.R. Human Appropriation of Renewable Fresh Water [Journal] // American Association for the Advancement of Science. - 1996. - Vol. 271. - pp. 785-788.

Qiang He Brian R. Silliman Climate Change, Human Impacts, and Coastal Ecosystems in the Anthropocene [Journal] // Current Biology. - 2019. - 19 : Vol. 29. - pp. R1021-R1035.

Resnick S. Ghosh and S. A discussion on mean excess plots [Journal] // Stochastic Processes and their Applications. - 2010. - ISSN 0304-4149 : Vol. 120. - pp. 1492-1517.

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. [Journal] // Resour. Conserv. Recycl. . - (2020) (January). - 104879 : Vol. 160 .

Ritchie B.W. Chaos, crises and disasters: A strategic approach to crisis manage ment in the tourism industry [Journal] // Tourism Management. - 2004. - Vol. 25. - pp. 669-683.

Saade Mark, Semerjian Lucy and Amacha Nabil Physicochemical Evaluation of the Upper Litani River Watershed, Lebanon [Journal]. - 2012. - doi:10.1100/2012/462467 : Vol. 2012.

Saliba N. A., Moussa, S., Salame, H., El-Fadel, M. Variation of selected air quality indicators over the city of Beirut, Lebanon: assess ment of emission sources [Journal] // Atmospheric Environment. - 40 : [s.n.], 2006. - 18 : Vol. 40. - pp. 3263-3268.

Saliba N.A. [et al.] Variation of selected air quality indicators over the city of Beirut, Lebanon: Assessment of emission sources [Journal] // Atmos. Environ. - 2006,. - Vol. 40. - pp. 3263-3268.

Saliba N.A., Kouyoumdjian H. and Roumie M. Effect of local and long-range transport emissions on the elemental composition of PM10 and PM2.5 in Beirut [Journal] // Atmos. Environ.. - 2007. - Vol. 41. - pp. 6497-6509.

Sanchez Lasheras F., Garc´ıa Nieto, P.J., Garc´ıa Gonzalo, E. et al. Evolution and forecasting of PM10 concentration at the Port of Gijon [Journal] // Sci Rep. - Spain : [s.n.], 2020. - 11716 : Vol. 10.

Saxifrage Barry Fossil fuel burning leaps to new record, crushing clean energy and climate efforts [Journal]. - [s.l.] : Canada's National Observer, 2019.

Schoof J.T and Robeson S.M. [Journal] // Weather Clim Extremes. - 2016. - Vol. 11. - pp. 28-40.

Schwartz J., Marcus, A. Mortality and air pollution in London: a time series analysis [Journal] // American Journal of Epidemiology. - 1990. - Vol. 31. - pp. 85–194.

Scrase F.J.). Some characteristics of eddy motion in the atmosphere [Journal]. - [s.l.] : Meteorological office geophysycal memoirs, 1930.

Seeger M.W. Chaos and crisis: Propositions for a general theory of crisis communication [Journal] // Public Relations Review. - 2002. - Vol. 28. - pp. 329-337.

Seinfeld K. E. Bencala and J. H. On frequency distribution of air pollutant concentrations [Journal] // Atmospheric Environment. - 1976. - Vol. 10. - pp. 941-950.

Shen Kunrong, Gang Jin, and Xian Fang. Does environmental regulation cause pollution to transfer nearby [Journal] // Econ. Res. J.. - 2017. - Vol. 52. - pp. 44-59.

Sherwood S. C [et al.] An assessment of Earth's climate sensitivity using multiple lines of evidence [Journal]. - 2020.

Shing C. P. Interval estimation of location and scale parameters based on record values [Journal] // Statistics Probability Letters. - 1998. - Vol. 37. - pp. 49-58.

Shing C. P. Interval estimation of location and scale parameters based on record values [Journal] // Statistics Probability Letters. - 1998. - Vol. 37. - pp. 49-58.

Simpi B. Hiremath SM., Murthy KNS. A Analysis of Water Quality Using physico-Chemical Parameters Hosahalli Tank in Shimoga District [Journal] // Global Journal of Science Frontier Research. - Karnataka, India : [s.n.], 2011. - Vol. 3. - pp. 31-34.

Sivaramanan Sivakumaran Air Pollution sources, pollutants and mitigation measures [Journal]. - 2014.

Skiadas C.H. and Dimotikalis Y. 12th Chaotic Modeling and Simulation International Conference [Conference] // Series ISSN 2213-8684 / ed. Complexity Springer Proceedings in. - 2020. - p. 306.

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

Statheropoulos M., Vassiliadis N. and Pappa A. Principal and canonical correlation analysis for examining air pollution and meteorological data [Journal]. - [s.l.] : Atmospheric Environment, 1998.

Steffen Will, Hughes Lesley and Perkins Sarah HEATWAVES: HOTTER, LONGER, MORE OFTEN [Journal]. - [s.l.] : Climate Council of Australia Limited, 2014.

Sutton O.G. A Theory of Eddy Diffusion in the Atmosphere [Journal]. - [s.l.] : Environmental science, 1932.

Taheri T. [et al.] Time series analysis of water quality parameters [Journal] // Journal of Applied Research in Water and Wastewater. - 2014. - 1 : Vol. 1. - pp. 40-50.

Takens F. Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.S.(Eds.), [Journal] // Lectures Notes in Mathematics / ed. Springer-Verlag. - New York : [s.n.], 1981. - Vol. 898. - pp. 366-381.

Taylor G.I. Eddy motion in the atmosphere [Journal]. - [s.l.] : Royal society, 1915.

Theiler J. Spurious Dimension from Correlation Algorithms Applied to Limited Time-Series Data [Journal] // Physical Review A. - 1986. - Vol. 34. - pp. 2427-2432.

Tippett R. A. Fisher and L. H. C. Limiting Forms of the frequency distribution of the largest or smallest member of a sample [Journal] // Mathematical Proceedings of the Cambridge Philosophical Society. - 1928. - Vol. 24. - pp. 180-190.

Tobin P. An introduction to chaos theory [Conference] / ed. Conference Proceedings. - 2016.

Touloumi G., Atkinson R. and Terte A.L. Analysis of health outcome time series data in epidemiological studies [Journal] // Environmetrics. - 2004. - Vol. 15. - pp. 101-117.

United Nations Department of Economic and Social affairs United Nations Department of Economic and Social Affairs, Population Division. [Journal]. - 2017.

USEPA [Online] // Air quality index calculator. - 2017. - www. airnow.gov.

Van Buuren Stef, and Karin Groothuis-Oudshoorn Multivariate imputation by chained equations in R [Journal] // Journal of statistical software . - 2011. - Vol. 45. - pp. 1-67.

Vitolo Claudia, Di Napoli Claudia and Francesca Di Giuseppe, Hannah L. Cloke, Florian Pappenberger Mapping combined wildfire and heat stress hazards to improve evidencebased decision making [Journal]. - [s.l.] : Environment International, 2019.

Voigt K., Welzl G. and Bruggemann R. Data analysis of environmental air pollutant monitoring systems in Europe [Journal] // Environmetrics. - 2004. - Vol. 15. - pp. 577-596.

Voigt K., Welzl, G., Bruggemann, R. Data analysis of environmental air pollutant monitoring systems in Europe [Journal] // Environmetrics. - 2004. - Vol. 15. - pp. 577–596.

Wang L., Wang, J., Tan, X., Fang, C Analysis of NOx pollution characteristics in the atmospheric environment in Changchun City [Journal] // Atmosphere. - 2020. - 30 : Vol. 11(1).

Water Facts Central California Area Office [Online] // Bureau of reclamation. - 2020. - https://www.usbr.gov/mp/arwec/water-facts-ww-water-sup.html#:~:text=Water%20covers%20about%2071%25%20of%20the%20earth's%20surface..

Wehbeh Farah [et al.] Time series analysis of air pollutants in Beirut, Lebanon [Journal]. - [s.l.] : Springer International Publishing Switzerland, 2014.

Wergen G. and Krug J. Record-Breaking Temperatures Reveal a Warming Climate [Journal] // Europhys Lett . - 2010. - 30008 : Vol. 92.

Whitehead P. G. [et al.] A review of the potential impacts of climate change on surface water quality [Journal] // Hydrological Sciences-??Journal-des Sciences Hydrologiques. - 2009.

WHO Ambient (outdoor) air pollution [Online] // WHO. - November 2021. - 09 07, 2022. - https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

Wilks D.S. Representing serial correlation of meteorological events and forecast in dynamic decision-analytic models [Journal]. - [s.l.] : Monthly Weather Review, 1991.

Wilks D.S. Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models [Journal]. - [s.l.] : Monthly Weather Review.

Wilks D.S. Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models [Journal] // Monthly Weather Review. - 1991. - Vol. 119. - pp. 1640-1662.

Williams. G. Chaos Theory [Book]. - [s.l.] : Tamed. Joseph Henry Press, 1997.

World Health Organization (WHO) Air Quality Guidelines [Report]. - Copenhagen, Denmark : WHO Regional Office for Europe, 2021.

Xia X Qi Q, Liang H, Zhang A, Jiang L, Ye Y, Liu C, Huang Y. Pattern of Spatial Distribution and Temporal Variation of Atmospheric Pollutants during 2013 in Shenzhen [Journal] // ISPRS International Journal of Geo-Information. - China : [s.n.], 2017. - 2 : Vol. 6(1).

Zreik R., Latouche, P., and Bouveyron, C. The dynamic random subgraph model for the clustering of evolving networks [Journal] // Computational Statistics. - 2017. - Vol. 32. - pp. 501-533.