Group Project Development Final Report

Statistical Analysis of Water Quality Parameters of the Red River (Red River Delta), 2013

Members

Trương Quốc Cường	BI11-049
Nguyễn Tường Quang Hải	BI11-073
Lê Trần Khương Duy	BI11-069
Bùi Tuấn Minh	BI11-167
Nguyễn Hoàng Nam	BI11-196
Phan Minh Quang	BI11-232

Supervisors

Dr. Nghiêm Thị Phương

Dr. Lê Phương Thu

ICT Lab

February 2023

Abstract

Water is an all-important substance that sustains life here on earth. A poor quality source of water is a concern to not only residential health but the country's economy as well. The rivers in the Northern part of Vietnam are generally of poor quality due to a number of factors. In this project, we aim to find the source of river pollution in depth by analyzing chemical components in river water and atmospheric statics. We have applied various methods to figure out aspects that influence river water quality in the northern part of Vietnam, including one-way ANOVA t-test, paired-sample t-test, Pearson’s correlation, Spearman’s correlation, and principal analysis components (PCA). We used a one-way ANOVA t-test, a paired sample t-test, and some other types of t-test (discussed later in this report) to find out whether or not there is a significant difference in the means of our datasets. Then, Pearson’s and Spearman’s correlation is utilized to inspect how one dataset can affect the others. Lastly, we used PCA to increase the interpretability of the dataset, by reducing its dimensionality and transforming it into a new coordinate system where the variation of the analyzed components can be explained with fewer dimensions.

1. Introduction

Rational and preservative utilization of water resources represent one of the main problems of the 21st century. Water is valued for its quantity, quality, and geographic location as a resource. Water quality is determined by the state of the water system as reflected by physical-chemical, chemical, and biological indicators. The important aspects taken into consideration when examining the top-priority problems of water quality are economic influence, the influence on human health, the influence on the ecosystem, and the influence of the geographic area, as well as the duration of the influence.

In addition to anthropogenic factors, natural processes such as hydrological conditions, topography, lithology, climate, precipitation inputs, catchment area, tectonic, edaphic factors, erosion, weathering of crustal materials, bedrock geology, and environmental influence also interact to affect changes in the metal characteristics of water quality.

It should be also taken into account that Vietnam is striving to become one of the world's leading agricultural countries, and reaching good water quality represents an important challenge.

As such, it can be said that Vietnamese river systems are crucial to the sustained growth of their entire environment, especially when they pass through densely populated areas. The existence and quality of water are essential for maintaining ecological equilibrium, and more research has been conducted based on observations of water quality. Waterbody pollution is a result of human activities on the one hand and heavy urbanization development on the other; at the same time, anthropogenic influences can have detrimental effects on water quality in just a short amount of time. A great indication of the effects of anthropogenic activities on water is the load of organic solids along with the dynamics of their degradation.

In urban areas, wastewater generated by industry and local residents is increasingly being disposed of in rivers and river channels over the past few decades. One of the major issues facing water courses today is the alarmingly high degree of wastewater discharge into surface water. The government of many industrialized nations has implemented stringent legal regulations that require all wastewater to be treated carefully (scrubbed) before being released into a waterway in order to avoid pollution. Despite all legal restrictions, many rivers possess pollution levels higher than allowed since the catchment region is unable to absorb all of the dumped wastewater.

As a result, one of the most crucial elements to be considered when evaluating the sustainability of a region's development is its water quality. While the economy is growing more quickly than ever, the overall quality of water sources has significantly declined. In addition to impeding sustainable development, poor water quality and environmental degradation may endanger public health. The availability and quality of water are crucial for maintaining ecological equilibrium. This has led to an increase in research over the past ten years that is centered around the monitoring of water quality.

2. Material and Methods

2.1 Sampling Areas

In the mountains south of Dali in the Yunnan region of China, the Red River originates. Leqiu, Xi, and Juli rivers' principal headstreams converge at Nanjian to form the Lishe River. At Hongtupo, Chuxiong Prefecture, the Lishe River joins the Yijie River, another headstream. Before exiting China through the Honghe Autonomous Prefecture in Yunnan, it flows mainly southeast, passing through Yi and Dai ethnic minority regions. It constitutes a stretch of the international border between China and Vietnam and enters Vietnam at the province of Lao Cai. Before emerging from the highlands to reach the midlands, the river, known as Thao River for this upper portion, continues its southeasterly path across northwest Vietnam.

Downstream from Việt Trì, the river and its main distributaries, the Đuống River, Kinh Thầy River, Bạch Đằng River, and the Thái Bình river system spread out to form the Red River Delta. The Red River flows past the Vietnamese capital Hanoi before emptying into the Gulf of Tonkin.

In this study, the water quality status as well as the spatial and temporal trends over the 12-month period in 2013 were assessed at 5 different hydrometeorological forecasting stations on the Red River:

Station	Longitude	Latitude	Address	River
Vụ Quang	E 105.25108	N 21.57780	Vụ Quang, Đoan Hùng, Phú Thọ	Lô River (upstream)
Yên Bái	E 104.86972	N 21.70484	Nguyễn Phúc, Yên Bái city, Yên Bái	Red River (upstream)
Hòa Bình	E 105.34298	N 20.83748	Hòa Bình city, Hòa Bình	Đà River (upstream)
Sơn Tây	E 105.511168	N 21.15848	Viễn Sơn, Sơn Tây, Hà Nội	Red River (downstream)
Hồng Hà Nội	E 105.85310	N 21.04001	Đông Hà, Phúc Tân, Hoàn Kiếm, Hà Nội	Red River (downstream, after Sơn Tây)

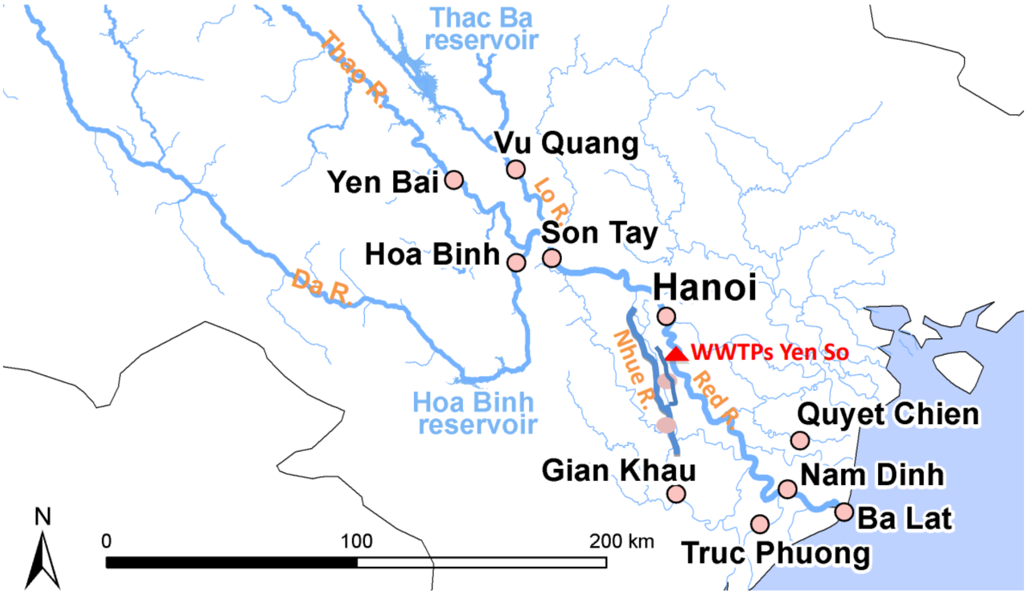


Figure. 1. Geographical location of Hydrometeorological stations on the Red River

2.2 Data and Methods

2.2.1 Database for Weather and Water quality parameters

The Red River's water quality status in 2013 was deduced using the database provided by the ICT lab. Over the course of a year, from January to December, 12 measurements of the water's chemical properties were made at 5 distinct locations: Yên Bái, Vụ Quang, Hòa Bình, Sơn Tây, and Hà Nội. On the other hand, with the exclusion of Sơn Tây and a measurement date in February that did not coincide with the chemical property, the metal properties of the river were only measured 10 times from January to October.

For the purpose of maintaining data adequacy, we made the decision to disregard the measurement date discrepancy and continue with the data processing stage with the lack of metal properties measurements in November and December and in Sơn Tây taking into consideration.

In addition, meteorological data was collected as well, matching the measurement date of the chemical properties. In order to fully understand the dataset, a more thorough investigation of the hydrometeorology of the Red River data is needed. As it was extremely difficult to obtain direct observations of the meteorological data at the five measuring locations in 2013, we were obliged to collect the information we needed from external websites (tcktcktck.org & openweather.org). The meteorological information obtained from the aforementioned websites is not a firsthand observation but rather an estimation made using forecasting algorithms, the source of which is unknown to us. Nevertheless, we must make due for the sake of advancement.

The obtained data are analyzed in the SPSS statistical program. Presented results were obtained according to several different statistical analyses: descriptive statistical analysis, one-way analysis of variance (ANOVA), t-test analysis for independent samples, paired samples t-test, and principal component analysis (PCA). These analyses require particular sets of assumptions that the dataset must “pass” in order to provide valid results; therefore, many more tests are introduced and conducted, such as the Shapiro-Wilk Normality Test, the Kruskal-Wallis H Test, the Leneve Test, Welch’s T-Test, and the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy to alleviate the many violations of assumptions occur in the analysis process. Pearson’s correlation and Spearman’s correlation test are used to measure the correlation ratio and evaluate the linear relationships between the variables. Across 27 water quality variables, only one is measured in nominal (Sal) so it will be exempted from being input in any statistical test.

The water quality parameters analyzed in this assessment include:

2.2.1.1. Average Atmospheric Temperature (ºC)

Atmospheric temperature is a measure of temperature at different levels of the Earth's atmosphere. It is governed by many factors, including incoming solar radiation, humidity, and altitude. When discussing surface air temperature, the annual atmospheric temperature range at any geographical location depends largely upon the type of biome, as measured by the Köppen climate classification. An increase in the air temperature will cause water temperatures to increase. A decrease in the air temperature will cause water temperatures to decrease. The temperature has a much greater influence on the amount of oxygen dissolved in water than air pressure, and if the temperature is low then the amount of dissolved oxygen is high and vice versa.[3]

2.2.1.2. Humidity (%RH)

Humidity is a measure of the amount of water vapor in the air. Relative humidity measures the amount of water in the air in relation to the maximum amount of water vapor (moisture). The higher the temperature, the more water vapor the air can hold. The two basic factors which determine the concentration of humidity in the atmosphere are temperature and pressure. The absolute humidity will rise with a rise in temperature. The relative humidity is usually higher in areas of low pressure. [4]

2.2.1.3. Pressure (hPa)

Atmospheric or air pressure is the force per unit of area exerted on the Earth’s surface by the weight of the air above the surface. The force exerted by an air mass is created by the molecules that make it up and their size, motion, and number present in the air. These factors are important because they determine the temperature and density of the air and, thus, its pressure. The number of air molecules above a surface determines air pressure. The magnitude of air pressure is related to altitude, temperature, and air density. The higher the ground, the lower the air pressure. The higher the temperature, the lower the air density and the lower the air pressure. Air pressure determines the amount of dissolved oxygen in water, and the higher the air pressure, the higher the amount of dissolved oxygen. Higher pressure counteracts the gravity effect of the sun and moon so reduces the high tide level, and presumably a lower low tide as well. [5]

2.2.1.4. Wind speed (km/h)

Wind speed is typically judged as the velocity of the wind. Most measurements of air movement are taken of outside air, and there are several factors that can affect it. For the most part, air moves along these pressure gradients from high pressure to low pressure. The movement is the major force that creates wind on Earth. The greater the difference in pressure, the greater the wind speed. When the wind speed increases, the turbidity of water will be increased because TDS and TS in water are rising. [6]

2.2.1.5. pH

pH is the measure of the acidity of a solution of water. The pH scale commonly ranges from 0 to 14. The scale is not linear but rather it is logarithmic.

For example, a solution with a pH of 6 is ten times more acidic than a solution with a pH of 7. Pure water is said to be neutral, with a pH of 7. Water with a pH below 7.0 is considered acidic while water with a pH greater than 7.0 is considered basic or alkaline. pH is changed because of many factors: CO2 concentration, temperature, presence of chemical detergents and cleaning agents, and dissolved minerals. CO2 content increases result in low pH and CO2 content decreases result in high pH. When there is a high temperature the pH is low and when there is a low temperature the pH is high. The presence of chemical detergents and cleaning agents that are improperly disposed of can increase the pH. pH changed because of dissolved minerals in the water. [7]

2.2.1.6. Total Dissolved Solids (mg/l)

Total dissolved solids (often abbreviated as TDS) is a measurement of the number of dissolved particles in your water. A dissolved particle is anything that is not a water molecule that can pass through a filter with pores of two microns in size. Measurement of total dissolved solids in water includes a number of different types of contaminants, some of which are more harmful than others. Some of the most common dissolved solids in water include calcium, chlorides, THMs, nitrates, phosphorus, iron, sulfur, and bacteria. This means TDS decreases because of fewer organic and inorganic materials in water and vice versa. [8]

2.2.1.7. Conductivity (µs/cm)

Conductivity is a numerical expression of an aqueous solution's capacity to carry an electric current. This ability depends on the presence of ions, their total concentration, mobility, valence, and relative concentrations, and on the temperature of the liquid. Electrical conductivity in water invariably increases with an increase in temperature, as opposed to metals. Dissolved ions enhance salinity as well as increase conductivity. Solutions of most inorganic acids, bases, and salts are relatively good conductors. In contrast, the conductivity of distilled water is less than 1 µmhos/cm. Because conductivity is the inverse of resistance, the unit of conductance is the mho (ohm spelled backward), or in low-conductivity natural waters, the micromho. [9]

2.2.1.8. Sal (‰)

The amount of dissolved salt in a body of water is known as salinity. It makes a significant contribution to conductivity and influences many aspects of the chemistry of natural waters as well as the biological activities that take place there. Temperature, pressure, and salinity all influence the physical properties of water, including its density and heat capacity. [10]

2.2.1.9. Nitrogen

Nitrogen is an essential plant nutrient found in fertilizers, human and animal wastes, yard waste, and the air. About 80% of the atmosphere is nitrogen gas. Nitrogen gas diffuses into the water where it can be "fixed" (converted) by blue-green algae to ammonia for algal use. Nitrogen can also enter lakes and streams as inorganic nitrogen and ammonia. Because nitrogen can enter aquatic systems in many forms, there is an abundant supply of available nitrogen in these systems. Temperature can affect total-N as when temperature increases, total-N decreases. The more nitrogen, the more aquatic plants and algae will grow, creating more CO2 in water which will make the pH decrease. [11]

2.2.1.9.1. Nitrite (mgN/l)

Nitrite (chemical formula NO2-) is a compound of nitrogen and oxygen, commonly found in soil and water. Depending on the presence of nitrogen compounds, we can know the level of water pollution. If the water is mainly NO2- then the water has been contaminated for a fairly long time and is perilous. [12]

2.2.1.9.2. Nitrate (mgN/l)

Similar to nitrites, the presence of nitrates in water indicates that the water source has been contaminated from fertilizer use in agriculture, septic tanks, wastewater treatment systems, animal waste, industrial waste or from the food processing industry. High nitrate concentrations can affect the health and development of aquatic products (shrimp, freshwater fish) and other aquatic organisms. If the water is mostly NO3-, then the oxidation is over. [12]

2.2.1.9.2. Ammonium (mgN/l)

Ammonium in water is a contaminant due to animal waste, sewage, and the potential for bacterial contamination. Ammonium is a source of nutrients for algae to grow, microorganisms that grow in pipes cause corrosion, leakage, and loss of eye-catchiness. Ammonium along with trace substances in water (organic compounds, phosphorus, iron, manganese...) will create conditions for bacteria to grow, affecting the quality of water after treatment. The water can be cloudy, dregs in the pipeline, and contain water. Water is degraded, reducing sensory factors.

Besides, when the concentration of ammonium in water is high, it is easy to form nitrite (NO2-), and nitrate (NO3-). If the water contains NH3 and organic nitrogen, the water is considered contaminated and dangerous. [13]

2.2.1.10. Phosphate (mgP/l)

Phosphates exist in three forms: orthophosphate, metaphosphate (or polyphosphate), and organically-bound phosphate; each compound contains phosphorus in a different chemical arrangement.

These forms of phosphate occur in living and decaying plant and animal remains. High phosphate levels can come from man-made sources such as septic systems, fertilizer runoff, and improperly treated wastewater. The phosphates enter the water as the result of surface run-off and bank erosion. [14]

2.2.1.11. Phosphorus total (mgP/l)

Phosphorus, like nitrogen, is a critical nutrient required for all life and a common ingredient in commercial fertilizers.

Total phosphorus (TP) is a measure of all phosphorus found in a sample, whether that phosphorus is dissolved or particulate. This is commonly used when sampling in wastewater treatment and is notably used to determine the health of waterways.

When there is too much Phosphorus in water, it can speed up eutrophication (a reduction in dissolved oxygen in water bodies caused by an increase of mineral and organic nutrients) of rivers and lakes, which may result in excess algae. [15]

2.2.1.12. Dissolved Organic Carbon (mgC/l)

Dissolved Organic Carbon or DOC is a measurement of the amount of organic matter in water that can be passed through a filter, (filters generally range in size between 0.7 and 0.22 um) (Bruckner)

Total organic carbon occurs in untreated water, such as lakes and rivers. The TOC in a body of water is impacted by vegetation in the area, the climate, and even treated sewage released into the water. [16]

2.2.1.13. Particulate Organic Carbon (mg/l)

Opposite to Dissolved Organic Carbon (DOC), particulate organic carbon (POC) is the carbon that is too large and is filtered out of a sample. If you have ever seen a body of water that appears straw, tea, or brownish in color, it likely has a high organic carbon load. This color comes from the leaching of humic substances from plant and soil organic matter.

Organic carbon can be allochthonous or sourced from outside the system (e.g. by atmospheric deposition or transported long distances via stream flow) or it can be autochthonous, or sourced from the immediate surroundings of the system (e.g. plant and microbial matter and sediments/soils within the catchment). [17]

2.2.1.14. Monthly water discharge (m^3/s)

River discharge is the volume of water passing through a measuring point or gauging station in a river at a given time. It is measured in cubic meters per second. The overall discharge from the drainage basin depends on the relationship between precipitation and storage factors and this can be analyzed by the following formula:

Drainage basin discharge = precipitation - evapotranspiration ± changes in storage

The factors affecting the river discharge are as follows: Rock and soil type, Weather conditions, Type and amount of Rainfall, Type of land, Slope of area. [18]

2.2.1.15. Daily Suspended Solid (mg/l)

Suspended solids refer to small solid particles which remain in suspension in water as a colloid or due to the motion of the water. Therefore, solid materials that are insoluble in water and particle size that are larger than 2 microns are responsible for the amount of SS in water.

Increased erosion of banks of rivers and streams can increase the TSS level in the water. The suspended particles released from dirt and soil can settle out across the water and give it a murky appearance. Runoff — when water flows through eroding soil — may also produce similar results.

Rainfall is the major driver of soil erosion. Excessive phosphorus in surface water can cause explosive growth of aquatic plants and algae. Together with phosphorus, nitrates in excess amounts can accelerate eutrophication, causing dramatic increases in aquatic plant growth and changes in the types of plants and animals that live in the stream. [19]

2.2.1.16. Water temperature (°C)

Water temperature is a physical property expressing how hot or cold water is. As hot and cold are both arbitrary terms, the temperature can further be defined as a measurement of the average thermal energy of a substance 5. Thermal energy is the kinetic energy of atoms and molecules, so the temperature in turn measures the average kinetic energy of the atoms and molecules 5. This energy can be transferred between substances as the flow of heat. Heat transfer, whether from the air, sunlight, another water source, or thermal pollution can change the temperature of the water. [20]

2.2.1.17. Manganese (μg/l)

Manganese may be present in water in the environment from natural sources (rock and soil weathering) or as a result of human activities (such as mining, industrial discharges and landfill leaching). Manganese is found naturally in many surface water (lake and river water) and groundwater (underground water) sources. Water passing through soil and rock can dissolve minerals containing manganese. [21]

2.2.1.18. Chromium(μg/l)

Chromium occurs naturally in small amounts in rocks and soils, some of which are released into the aquatic environment through weathering and erosion processes. [22]

2.2.1.19. Zinc (μg/l)

Zinc enters the air, water, and soil as a result of both natural processes and human activities. Most zinc enters the environment as the result of mining, purifying of zinc, lead, and cadmium ores, steel production, coal burning, and burning of wastes. These activities can increase zinc levels in the atmosphere. Waste streams from zinc and other metal manufacturing and zinc chemical industries, domestic wastewater, and run-off from soil containing zinc can discharge zinc into waterways. [23]

2.2.1.20. Lead (μg/l)

Lead can enter drinking water when plumbing materials that contain lead corrode, especially where the water has high acidity or low mineral content that corrodes pipes and fixtures. [24]

2.2.1.21. Cadmium (μg/l)

Cadmium contamination of a water source can be through natural erosion of cadmium-containing rocks, industrial dust/waste (impurity in zinc), fertilizer (contaminant in phosphate rock), pigment production, mine tailings or spoils, smelting, and plasticizers production. [25]

2.2.1.22. Iron (μg/l)

The most common sources of iron in groundwater are naturally occurring, for example from weathering of iron-bearing minerals and rocks. Industrial effluent, acid-mine drainage, sewage, and landfill leachate may also contribute iron to local groundwater. As rain falls or snow melts on the land surface and water seeps through iron-bearing soil and rock, iron can be dissolved into the water. In some cases, iron can also result from corrosion of iron or steel well casings or water pipes. [26]

2.2.1.23. Copper (µg/l)

Copper is commonly found in aquatic systems as a result of both natural and anthropogenic sources. Natural sources of copper in aquatic systems include geological deposits, volcanic activity, and weathering and erosion of rocks and soils. Anthropogenic sources of copper include mining activities, agriculture, and metal and electrical manufacturing. [27]

2.2.1.24. Silica (mg/l)

Silica, also known as silicon dioxide (SiO2) or mineral quartz, is a derivative compound of the chemical element silicon (Si). Silica is found as dissolved silica or suspended silicate particles in most natural waters. The presence of most silica in natural waters comes from the gradual degradation of silica-containing minerals.

Most water contains an abundance of silicon and its derivatives, silica (SiO2) or silicates (SiO4– and SiO32–). Silica concentration in water is common in the 5 to 25 mg/L range.

Silica and silicates are added to water as conditioners, detergents, and corrosion inhibitors. However, silica in water can cause significant problems for industries, primarily in boiler and steam applications. High pressures and high temperatures cause silica deposits on boiler tubes and heat exchangers.

Measuring silica in water is useful when the efficiency of demineralizers is monitored. Testing for silica (one of the first impurities detected when the exchange capacity of a demineralizer is exhausted) provides a sensitivity check of demineralizer performance. [28]

2.2.2 Outliers and Missing data

2.2.2.1 Outliers

Upon collection, the data was checked for outliers, missing values, and possible measurement errors.

Outliers refer to data that falls out of the general or expected data behavior. Such outliers do not only refer to “extreme” values but also missing data, both of which can throw off the results of a study or analysis, making it difficult to draw any conclusions. [29]

The detection of outliers was done with the use of boxploting. In descriptive statistics, a box plot (or boxplot) is a type of chart often used in exploratory data analysis. Box plots visually show the distribution of numerical data and skewness by displaying the data quartiles. In addition to the box on a box plot, there can be lines (which are called whiskers) extending from the box indicating variability outside the upper and lower quartiles. Outliers that differ significantly from the rest of the dataset are labeled and plotted as individual points beyond the whiskers on the boxplot.

A boxplot is a standardized way of displaying the dataset based on the five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles.

- Minimum (Q_0 or 0th percentile): the lowest data point in the data set excluding any outliers
- Maximum (Q_4 or 100th percentile): the highest data point in the data set excluding any outliers
- Median (Q_2 or 50th percentile): the middle value in the data set
- First quartile (Q_1 or 25th percentile): also known as the lower quartile $q_n(0.25)$, it is the median of the lower half of the dataset.
- Third quartile (Q_3 or 75th percentile): also known as the upper quartile $q_n(0.75)$, it is the median of the upper half of the dataset.

In addition to the minimum and maximum values used to construct a boxplot, another important element that can also be employed to obtain a boxplot is the interquartile range (IQR), as denoted below:

- Interquartile range (IQR): the distance between the upper and lower quartiles

$$IQR = Q_3 - Q_1 = q_n(0.75) - q_n(0.25)$$

In SPSS, values more than three IQRs from the end of a box are labeled as extreme outliers [30]. Values more than 1.5 IQRs but less than 3 IQRs from the end of the box are labeled as true outliers. However, in an outlying observation labeling rules research by the American Statistical Association, outliers detection by multiplying IQR by a factor of 1.5 was demonstrated to be inaccurate approximately 50% of the time, and it was suggested that 2.2 is a more valid multiplication factor [31]. Moreover, these rules are all only applicable to data that is normally distributed. After running a Normality test, the detection results and handling of outliers of data detection have been adjusted accordingly for normal and non-normal distributed variables. [32]

Although the common practice is to remove the outliers, we have decided to leave them as is in the dataset. Because in this case, true outliers represent natural variations in the population and are informative about the data collection process. Other extreme outliers are problematic and removed because they represent measurement errors or data entry.

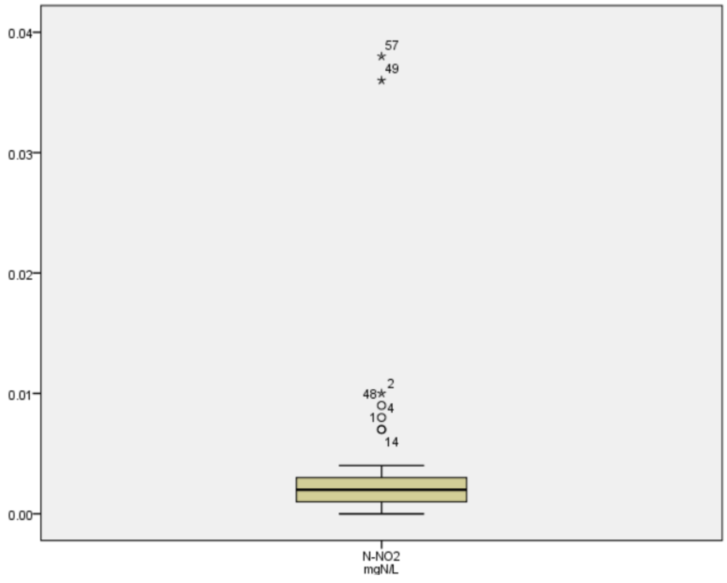


Fig. 2. Example of a boxplot and outliers detection in N-NO2, with true outliers denoted with a circle (o) and extreme outliers denoted as an asterisk (*)

2.2.2.2 Missing Data and Measurement Errors

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Understanding the reasons why data is missing is important for handling the remaining data correctly. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, the analysis may be biased. [33]

After detecting the outliers, the obtained dataset was then manually checked for missing or duplicated values across all 27 parameters. There are 8 missing values not including the exclusion of metal properties data (which remains untouched) which seems missing at random. As the data analysis technique in this study is not robust to the missingness and non-normality of the dataset, these missing values are substituted with the median of the observations of their respective variables. While the number of imputations is smaller than the recommended amount by some scholars, leading to a parameter bias in analyses, a median imputation has the benefit of not changing the sample median for the variables and mitigating the effect of outliers [34]. The small number of imputations also enables the option not to take into account the pattern of missing data that can complicate the imputation process.

We also detected 12 data entry errors, specifically in the POC variable in Hanoi, which seems to have been duplicated from the column next to it. The 12 data entries were excluded from the dataset after a further inspection revealed that they were statistically distinct from the other observations in the same variable.

2.2.3 Independent samples T-Test

An ANOVA ("Analysis of Variance") is used to determine whether or not there is a significant difference between the means of two or more independent groups.

The samples taking part in this test are Yên Bái, Vụ Quang, and Hòa Bình

Assumptions (Requirements) for Independent Samples T-Test:

- The independent variable consists of 2 or more independent groups
- The independence of observations
- The dependent variables are quantitative
- The dependent variables have no significant outliers
- The dependent variables have approximately normally distributed for each category of the independent variable
- The variance within the groups is similar

While the first four assumptions are present in many statistical tests and have been satisfied with the measurement being in five different locations, 26 distinct scaled variables excluding one measured in nominal (Sal), and the outliers being left as is, for the ANOVA's assumption that each of the groups has an **equal variance**, **approximately normal distribution** among categories and **no significant outliers** to be met, several more statistical tests must be conducted. Methods used are as follows:

- Assuming Equal variance
    - ➢ A Levene Test is conducted where it tests the equality of variance of two or more samples
    - ➢ Requirement for Levene Test:
        - ○ Samples are independent
        - ○ Dependent variables are quantitative
- Assuming approximately normal distribution among categories

- ➢ Shapiro-Wilk Normality Test (W-value) is conducted for each independent variables, while the test is robust to outliers, a visual quantitative representation of normal distribution is run alongside it to present distribution with a range of skewness (-+2) and kurtosis(-+7) of a distribution to determine whether or not it is normal. [35]

If the Equal Variance is not assumed a Welch's T-Test is conducted where it was designed for unequal population variances, but the assumption of normality is maintained.

If the Normal Distribution is not assumed, a Kruskal-Wallis H Test is conducted where it tests the medians of samples of two or more independent groups to determine if there is a significant difference. The Kruskal-Wallis H Test has 4 assumptions to be met, with 3 of them being the same as the first three assumptions of the ANOVA's T-Test, and the last assumption being that the distributions in each group (i.e., the distribution of data values for each group of the independent variables) have the same shape [36]. To test this assumption, we use the chi-squared distribution in the Kruskal Wallis Test. If this assumption is violated, only the test's mean rank comparison is used; otherwise, the ANOVA T-Test result is used.

2.2.4 Paired Samples T-Test

The dependent t-test (called the paired-samples t-test in SPSS Statistics) compares the means between two related groups on the same continuous, dependent variable.

Paired samples taking part in this test include the following dataset pairs:

- Yên Bái - Sơn Tây
- Vụ Quang - Sơn Tây
- Hòa Bình - Sơn Tây
- Yên Bái - Hà Nội
- Vụ Quang - Hà Nội
- Hòa Bình - Hà Nội
- Sơn Tây - Hà Nội

With chemical properties data not being presented for Sơn Tây and data entry errors for Hà Nội, any dataset pairs involving missingness will be exempted from testing.

Assumptions (Requirements) for Independent Samples T-Test:

- The dependent variable should be measured on a continuous scale
- The independent variable should consist of two categorical, "related groups" or "matched pairs"
- There should be no significant outliers in the differences between the two related groups
- The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed

  As the paired samples t-test is quite robust to non-normality in distribution, the dataset has met all assumptions and is suitable for testing

2.2.5 Correlation analysis of dataset

As the dataset is non-normally distributed due to extreme outliers, two different correlation tests were conducted. These two tests are Pearson's correlation test and Spearman's correlation test with results of correlation taken from Spearman's for dependent variables in which data is severely non-normally distributed (Iron (Fe), Lead (Pb), and Phosphate (PO4), and Pearson's for the rest of the dependent variables.

2.2.5.1 Pearson's correlation coefficient

Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{(x1,y1),...,(x_n,y_n)\}$ consisting of n pairs, r_{xy} is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- n is the sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean; Also analogous for $\bar{y}$ [37]

2.2.5.2 Spearman’s rank correlation coefficient

In statistics, Spearman's rank correlation coefficient or Spearman's ρ, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables. For a sample of size n, the n raw scores X_i, Y_i are converted to ranks $R(X_i), R(Y_i)$, and r_s is computed as

$$r_x = \rho_{R(X),R(Y)} = \frac{cov(R(X),R(Y)}{\sigma_{R(X)}\sigma_{R(Y)}}$$

Where

- ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,
- $cov(R(X), R(Y)$is the covariance of the rank variables,
- $\sigma_{R(X)}$and $\sigma_{R(Y)}$ are the standard deviations of the rank variables.

Only if all n ranks are distinct integers, it can be computed using the popular formula: $r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$,

Where

- $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks of each observation,
- n is the number of observations. [38]

2.2.5.3 Principal component analysis

2.2.5.3.1 Definition

Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data. Many studies use the first two principal components in order to plot the data in two dimensions and to visually identify clusters of closely related data points.[39]

2.2.5.3.2 Methodology

Step 1: Standardization

Standardization is important prior to PCA, because variables of different degrees range differently, leading to a biased result. As such, we need to standardize and transform all variable values to relatively comparable scales is necessary:

Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

In SPSS 25, there is no need to standardize the data beforehand as the method used in SPSS has standardized the variables before factoring the covariance matrix.

Step 2: Computation of covariance matrix

Computing the covariance matrix helps us understand the relationship between variables, more specifically how they vary from the mean with respect to each other. Additionally, sometimes variables can be highly correlated with one another, containing redundant information. As such, the computation of the said matrix will solve this problem.

The covariance matrix is a symmetric dxd matrix (where d is the number of dimensions) that has entries as the covariances associated with all possible pairs of initial variables.

$$\begin{bmatrix} \mathrm{Cov}(x,x) & \mathrm{Cov}(x,y) & \mathrm{Cov}(x,z) \\ \mathrm{Cov}(y,x) & \mathrm{Cov}(y,y) & \mathrm{Cov}(y,z) \\ \mathrm{Cov}(z,x) & \mathrm{Cov}(z,y) & \mathrm{Cov}(z,z) \end{bmatrix}$$

Since the covariance of a variable with itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. And since the covariance is commutative (Cov(a,b)=Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

Note: The covariance signs suggest the following:

- If positive then: the two variables increase or decrease together (correlated)
- If negative then: one increases when the other decreases (Inversely correlated)

Step 3: Identifying principal components by covariance matrix's eigenvectors and eigenvalues

Principal components are newly created variables constructed as linear combinations or mixtures of the initial variables. It is done in such a way that the principal components are uncorrelated and most of the information is compacted and retained. For example, a 10-dimensional dataset gives you 10 principal components, but PCA puts as much information into the first component, and maximizes the remaining information into the second, and so on. Eventually, we will have enough components to reflect the information of the dataset with less dimensionality, giving way to more flexible forms of visualization and interpretability.

Additionally, to select the optimal numbers of principal components, we use the eigenvectors and eigenvalues. First, they always come in pairs, so for each eigenvector, we will have an eigenvalue. Secondly, the matrix's eigenvectors are actually the directions of the axes where there is the most variance, which we call Principal Components, with the eigenvalues simply the attached coefficients, providing the carried amount of variance. By ranking eigenvectors by eigenvalues, we will get the most optimal principal components.

Eigenvalues of 1.0 or greater are considered significant.

Step 4: Feature vector

With the generated principal components, now we must decide which ones to use for further analysis. We do this by discarding components with lesser significance and form the remaining in a matrix of vectors called a Feature vector. Thus, this is the first step toward dimensionality reduction.

Step 5: Recast data along the principal component axes

With the generated feature vector, we now reorient the data from the initial axes to the ones represented by the principal components. This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedDataSet^T$$ [40]

2.2.5.3.3 Assumptions for PCA

- **Assumption 1:** Variables are quantitative and measured at the continuous level
- **Assumption 2:** There needs to be a linear relationship between all variables
- **Assumption 3:** Sampling adequacy
- **Assumption 4:** Adequate correlations between the variables in order for variables to be reduced to a smaller number of components
- **Assumption 5:** There should be no significant outliers

Assumption 1, 2 and 5 has been met as previously discussed, with linearity only taken into approximate consideration as to accommodate for true outliers.

The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for the overall data set was used to test for sampling adequacy, the statistic is a measure of the proportion of variance among variables that might be common variance.

The formula for the KMO test is:

$$MO_j = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} u}$$

Where:

- R = [r_{ij}] is the correlation matrix,
- U = [u_{ij}] is the partial covariance matrix,
- Σ = summation notation (“add up”)

KMO returns values between 0 and 1, with Kaiser’s rule for interpreting the statistic [41]:

- KMO values between 0.8 and 1 indicate the sampling is adequate.
- KMO values less than 0.5 indicate the sampling is not adequate and that remedial action should be taken.
- KMO Values close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which are a large problem for factor analysis.

This test is not usually calculated by hand, because of the complexity. Therefore, the test has been conducted in SPSS 25, where it returns a KMO value of 0.414. This means that our dataset has not met the assumption of sampling adequacy. With no other way to impute the missing data or neglect the outliers in the dataset, the analysis will have to be undertaken separately with 3 separate datasets of the water quality parameter (the meteorological properties, the chemical properties, and the metal properties)

- The KMO value for the meteorological properties dataset is 0.420
- The KMO value for the chemical properties dataset is 0.531
- The KMO value for the metal properties dataset is 0.513

With a KMO value above 0.5, the dataset for chemical properties and the dataset for metal properties of the Red River has met the assumption of sampling adequacy.

Correlations among chemical and metal variables have been tested with Pearson’s Correlation test and Spearman’s Correlation test. With that, assumption 4 and ultimately all assumptions for the PCA have been met [34].

3. Results and Discussions

3.1 Results

3.1.1 T-Test

- The test statistics in corresponding analyses are calculated as: variation between sample means (for One-way ANOVA T-Test and Welch's T-Test) or mean ranks (Kruskal-Wallis H Test) within the samples
- The higher the test statistics in an analysis, the higher the variation between sample means relative to the variation within the samples
- The higher the test statistics, the lower the corresponding p-value
- If the p-value is below a certain threshold (e.g. α = .05), we can reject the null hypothesis of the analysis and conclude that there is a statistically significant difference between group means.

Table 1. Mean values of Meteorological properties measured across 5 regions

Location		**Average Atm Temp**	**Humidity**	**Pressure**	**Wind Speed**
Yen Bai	**Mean**	25.50	77.25	1010.67	6.17
Vu Quang	**Mean**	25.67	76.92	1010.58	5.92
Hoa Binh	**Mean**	24.68	86.08	1010.73	6.11
Son Tay	**Mean**	27.92	73.83	1064.31	8.43
Ha Noi	**Mean**	25.13	71.66	1009.73	12.19

Table 2. Mean values of Chemical properties measured across 5 regions

Location		**pH**	**TDS**	**Cond**	**N-NO2**	**N-NO3**	**N-NH4**	**Si**	**PO4-P**	**Ptotal-P**	**DOC**	**POC**	**Monthly Water discharge**	**Daily water discharge**	**Daily SS-IMHE**	**T (oC) - IMHE**
Yen Bai	**Mean**	8.01	91.66	188.87	0.002	0.73	0.05	6.63	0.02	0.29	1.54	1.74	540.13	467.67	241.00	24.28
Vu Quang	**Mean**	8.10	100.38	206.17	0.004	0.59	0.09	5.24	0.03	0.09	1.02	1.04	1279.73	1227.75	47.28	24.35
Hoa Binh	**Mean**	7.89	82.38	171.64	0.007	0.40	0.04	6.43	0.01	0.05	1.17	0.57	1487.83	1464.33	11.55	24.79
Son Tay	**Mean**	7.96	88.23	181.13	0.01	0.45	0.08	6.22	0.03	0.12	0.93	0.90	1272.91	1227.17	66.20	24.63
Ha Noi	**Mean**	8.00	92.18	189.22	0.002	0.51	0.05	6.07	0.02	0.13	1.04	1.27	1955.66		68.62	24.96

Table 3. Mean values of metal properties measured across 4 regions

Location		Mn	Cr	Zn	Pb	Cd	Fe	Cu
Yen Bai	**Mean**	12.51	1.64	29.34	8.07	5.36	715.00	33.70
Vu Quang	**Mean**	10.41	1.60	32.34	7.98	5.10	532.00	43.00
Hoa Binh	**Mean**	9.84	1.38	25.27	11.90	5.59	353.00	34.00
Son Tay	**Mean**							
Ha Noi	**Mean**	11.26	1.87	31.00	15.29	6.05	947.60	30.60

3.1.1.1 Independent samples T-Test

Table 4. Independent samples T-Test of 3 upstream regions of the Red River.

	Equal Variance is Assumed	**Normally Distributed**	**Similar shape distribution across categories**	**Test to take results from**	**Test Statistics**	**p-value**
Average Atm Temp	Yes	No	Yes	Anova T-Test	0.095	0.910
Humidity	Yes	Yes	No	Kruskal Wallis Test	8.171	0.180
Pressure	Yes	No	Yes	Anova T-Test	0.001	0.999
Wind Speed	Yes	Yes	Yes	Anova T-Test	0.043	0.958
pH	Yes	Yes	Yes	Anova T-Test	1.847	0.174
TDS	Yes	Yes	No	Anova T-Test	11.121	0.000
Cond	Yes	Yes	No	Anova T-Test	8.140	0.001
N-NO2	No	No	No	Kruskal Wallis Test	19.779	0.000
N-NO3	Yes	Yes	No	Anova T-Test	15.510	0.000
N-NH4	No	No	Yes	Welch's T-Test	1.241	0.313

Si	Yes	Yes	No	Anova T-Test	9.399	0.001
PO4-P	No	No	Yes	Welch's T-Test	1.768	0.197
Ptotal-P	No	No	No	Kruskal Wallis Test	21.527	0.000
DOC	Yes	Yes	Yes	Anova T-Test	1.023	0.371
POC	No	No	No	Kruskal Wallis Test	17.517	0.000
Monthly Water discharge	No	Yes	No	Welch's T-Test	10.369	0.001
Daily water discharge	No	No	No	Kruskal Wallis Test	15.974	0.000
Daily SS-IMHE	No	No	No	Kruskal Wallis Test	29.505	0.000
T_water	Yes	No	Yes	Anova T-Test	0.087	0.917
Mn	Yes	No	Yes	Anova T-Test	0.394	0.678
Cr	Yes	No	Yes	Anova T-Test	0.117	0.890
Zn	Yes	No	Yes	Anova T-Test	0.232	0.795

Pb	Yes	No	Yes	Anova T-Test	0.355	0.704
Cd	Yes	No	Yes	Anova T-Test	0.421	0.661
Fe	Yes	No	Yes	Anova T-Test	1.039	0.368
Cu	No	Yes	Yes	Welch's T-Test	1.201	0.329

3.1.1.2 Paired Samples T-Test

Table 5. Paired samples T-Test of 5 regions across the Red River Delta.

		YB-ST	**VQ-ST**	**HB-ST**	**YB-HN**	**VQ-HN**	**HB-HN**	**ST-HN**
Average Atm Temp	**t-test**	-6.409	-5.937	-8.974	0.645	0.952	-1.472	5.787
	p-value	0.000	0.000	0.000	0.532	0.362	0.169	0.000
Humidity	**t-test**	4.311	3.675	8.407	1.877	1.791	5.441	0.783
	p-value	0.001	0.004	0.000	0.087	0.101	0.000	0.450
Pressure	**t-test**	-254.253	-241.634	-247.549	3.134	2.583	3.378	285.502
	p-value	0.000	0.000	0.000	0.010	0.250	0.006	0.000
Wind	**t-test**	-5.267	-6.210	-5.839	-7.261	-7.097	-5.196	-3.76

Speed								4
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.003
pH	**t-test**	0.670	2.292	-1.682	0.000	1.593	-1.744	-0.476
	p-value	0.517	0.430	0.121	1.000	0.139	0.109	0.643
TDS	**t-test**	1.083	4.038	-2.677	-0.141	2.163	-2.802	-2.118
	p-value	0.302	0.002	0.022	0.890	0.053	0.017	0.058
Cond	**t-test**	1.205	4.315	-2.079	-0.047	2.262	-2.522	-2.193
	p-value	0.253	0.001	0.062	0.964	0.045	0.028	0.051
N-NO2	**t-test**	-0.618	-1.184	-1.588	2.341	0.034	-2.334	1.189
	p-value	0.549	0.261	0.140	0.039	0.973	0.040	0.260
N-NO3	**t-test**	6.687	3.468	-3.061	5.479	2.164	-6.295	-3.338
	p-value	0.000	0.005	0.011	0.000	0.053	0.000	0.007
N-NH4	**t-test**	-0.900	0.040	-1.072	-0.957	1.225	-1.539	0.827
	p-value	0.388	0.969	0.306	0.359	0.246	0.152	0.426
Si	**t-test**	-0.900	0.040	-1.072	-0.957	1.225	-1.539	0.827
	p-value	0.388	0.969	0.306	0.359	0.246	0.152	0.426

PO4-P	**t-test**	-1.415	-0.303	-2.284	-0.916	0.273	-3.000	0.984
	p-value	0.185	0.767	0.043	0.380	0.790	0.012	0.346
Ptotal-P	**t-test**	3.814	-1.279	-3.562	3.479	-1.972	-5.133	-0.367
	p-value	0.003	0.227	0.004	0.005	0.074	0.000	0.720
DOC	**t-test**	2.102	0.418	0.937	1.359	-0.053	0.440	-0.576
	p-value	0.059	0.684	0.369	0.201	0.959	0.669	0.576
POC	**t-test**	2.103	0.525	-1.288	1.888	-1.638	-9.013	-1.224
	p-value	0.059	0.610	0.224	0.086	0.130	0.000	0.246
Monthly Water discharge	**t-test**	-7.210	0.103	2.415	-5.695	-3.238	-2.615	-4.578
	p-value	0.000	0.920	0.034	0.000	0.008	0.024	0.001
Daily water discharge	**t-test**	-4.695	0.007	2.103				
	p-value	0.001	0.995	0.059				
Daily SS-IMHE	**t-test**	5.279	-1.830	-6.988	5.273	-1.833	-6.773	-0.260
	p-value	0.000	0.094	0.000	0.000	0.094	0.000	0.799

T_water	**t-test**	-0.623	-1.238	0.409	-1.754	-6.013	-0.294	-1.558
	p-value	0.546	0.241	0.691	0.107	0.000	0.775	0.148
Mn	**t-test**				0.314	-0.364	-0.604	
	p-value				0.761	0.724	0.561	
Cr	**t-test**				-0.476	-0.774	-0.778	
	p-value				0.645	0.459	0.456	
Zn	**t-test**				-0.299	0.157	-1.052	
	p-value				0.772	0.879	0.320	
Pb	**t-test**				-0.746	-0.700	-0.277	
	p-value				0.475	0.502	0.788	
Cd	**t-test**				-1.130	-1.281	-0.565	
	p-value				0.288	0.232	0.586	
Fe	**t-test**				-0.923	-1.412	-2.363	
	p-value				0.380	0.192	0.042	
Cu	**t-test**				0.924	2.768	0.810	
	p-value				0.380	0.022	0.439	

3.1.2 Pearson's correlation

Table 6. Pearson Correlation Coefficients of meteorological variables with statistical significance of results

		Average Atm Temp (oC)	**Humidity (%RH)**	**Pressure (hPa)**	**Wind Speed (km/h)**
Average Atm Temp (oC)	Pearson Correlation	1	-.532**	-0.115	-0.008
	P-value		0.000	0.380	0.953
Humidity (%RH)	Pearson Correlation	-.532**	1	-0.039	-.390**
	P-value	0.000		0.766	0.002
Pressure (hPa)	Pearson Correlation	-0.115	-0.039	1	0.133
	P-value	0.380	0.766		0.312
Wind Speed (km/h)	Pearson Correlation	-0.008	-.390**	0.133	1
	P-value	0.953	0.002	0.312	
**. Correlation is significant at the 0.01 level (2-tailed).					

Table 7. Pearson Correlation Coefficients of chemical variables with statistical significance of results

		pH	**TDS (mg/l)**	**Cond (µs/cm)**	**N-NO2 mgN/L**	**N-NO3 mgN/L**	**N-NH4 mgN/L**	**Si mg/L**	**PO4-P mgP/L**	**Ptotal-P mgP/L**	**DOC (mg C/l)**	**POC, mg/l**	**Monthly Water discharge, m3/s**	**Daily water discharge, m3/s**	**Daily SS-IMHE, mg/l**	**T (oC) - IMHE**
pH	**Pearson Correlation**	1	0.210	.334**	-0.091	-0.187	0.231	-.379**	-0.005	0.135	-0.011	0.016	0.240	0.033	0.116	0.177
	P-value		0.108	0.009	0.491	0.161	0.076	0.003	0.968	0.304	0.936	0.910	0.065	0.800	0.377	0.176
TDS (mg/l)	**Pearson Correlation**	0.210	1	.941**	0.045	0.063	0.180	-.284*	0.247	0.000	0.088	0.223	-0.139	-0.132	0.004	-.370**
	P-value	0.108		0.000	0.735	0.637	0.168	0.028	0.057	0.997	0.505	0.105	0.289	0.313	0.975	0.004
Cond (µs/cm)	**Pearson Correlation**	.334**	.941**	1	0.025	-0.111	0.210	-.294*	.277*	0.051	0.053	.327*	-0.061	-0.088	0.050	-.288*
	P-value	0.009	0.000		0.848	0.407	0.107	0.022	0.032	0.700	0.688	0.016	0.643	0.506	0.703	0.026
N-NO2 mgN/L	**Pearson Correlation**	-0.091	0.045	0.025	1	0.126	0.064	0.227	0.128	0.108	-0.080	-0.042	-0.076	0.056	0.058	-0.195

	P-value	0.491	0.735	0.848		0.346	0.627	0.081	0.330	0.411	0.541	0.761	0.565	0.670	0.659	0.135
N-NO3 **mgN/L**	**Pearson Correlation**	-0.187	0.063	-0.111	0.126	1	-0.202	0.129	0.007	.349**	-0.071	0.069	-0.187	-0.122	.361**	0.093
	P-value	0.161	0.637	0.407	0.346		0.129	0.336	0.957	0.007	0.596	0.626	0.159	0.361	0.005	0.488
N-NH4 **mgN/L**	**Pearson Correlation**	0.231	0.180	0.210	0.064	-0.202	1	-0.197	0.238	-0.159	0.057	0.001	-.282*	-0.220	-0.127	-.395**
	P-value	0.076	0.168	0.107	0.627	0.129		0.131	0.067	0.226	0.665	0.993	0.029	0.092	0.335	0.002
Si **mg/L**	**Pearson Correlation**	-.379**	-.284*	-.294*	0.227	0.129	-0.197	1	0.055	0.130	-0.201	0.198	-0.179	-0.101	0.220	-0.230
	P-value	0.003	0.028	0.022	0.081	0.336	0.131		0.679	0.324	0.123	0.152	0.171	0.443	0.092	0.077
PO4-P **mgP/L**	**Pearson Correlation**	-0.005	0.247	.277*	0.128	0.007	0.238	0.055	1	0.040	-0.199	0.185	-0.024	-0.106	0.041	-.266*
	P-value	0.968	0.057	0.032	0.330	0.957	0.067	0.679		0.764	0.127	0.180	0.856	0.420	0.753	0.040
Ptotal-P **mgP/L**	**Pearson Correlation**	0.135	0.000	0.051	0.108	.349**	-0.159	0.130	0.040	1	0.090	.431**	-0.055	-0.224	.734**	0.236

	P-value	0.304	0.997	0.700	0.411	0.007	0.226	0.324	0.764		0.495	0.001	0.674	0.086	0.000	0.069
DOC (mgC/l)	**Pearson Correlation**	-0.011	0.088	0.053	-0.080	-0.071	0.057	-0.201	-0.199	0.090	1	-0.017	-0.141	-0.170	0.054	-0.147
	P-value	0.936	0.505	0.688	0.541	0.596	0.665	0.123	0.127	0.495		0.905	0.283	0.193	0.680	0.262
POC, mg/l	**Pearson Correlation**	0.016	0.223	.327*	-0.042	0.069	0.001	0.198	0.185	.431**	-0.017	1	-0.093	-.354**	.551**	-0.067
	P-value	0.910	0.105	0.016	0.761	0.626	0.993	0.152	0.180	0.001	0.905		0.505	0.009	0.000	0.628
Monthly Water discharge, m3/s	**Pearson Correlation**	0.240	-0.139	-0.061	-0.076	-0.187	-.282*	-0.179	-0.024	-0.055	-0.141	-0.093	1	.377**	-0.102	.537**
	P-value	0.065	0.289	0.643	0.565	0.159	0.029	0.171	0.856	0.674	0.283	0.505		0.003	0.439	0.000
Daily water discharge, m3/s	**Pearson Correlation**	0.033	-0.132	-0.088	0.056	-0.122	-0.220	-0.101	-0.106	-0.224	-0.170	-.354**	.377**	1	-0.157	.348**
	P-value	0.800	0.313	0.506	0.670	0.361	0.092	0.443	0.420	0.086	0.193	0.009	0.003		0.232	0.006

Daily SS-IMHE , mg/l	**Pearson Correlation**	0.116	0.004	0.050	0.058	.361**	-0.127	0.220	0.041	.734**	0.054	.551**	-0.102	-0.157	1	0.231
	P-value	0.377	0.975	0.703	0.659	0.005	0.335	0.092	0.753	0.000	0.680	0.000	0.439	0.232		0.075
T (oC) - IMHE	**Pearson Correlation**	0.177	-.370**	-.288*	-0.195	0.093	-.395**	-0.230	-.266*	0.236	-0.147	-0.067	.537**	.348**	0.231	1
	P-value	0.176	0.004	0.026	0.135	0.488	0.002	0.077	0.040	0.069	0.262	0.628	0.000	0.006	0.075	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 8. Pearson Correlation Coefficients of metal variables with statistical significance of results

		Mn (ug/l)	**Cr (ug/l)**	**Zn (ug/l)**	**Pb (ug/l)**	**Cd (ug/l)**	**Fe (ug/l)**	**Cu (ug/l)**
Mn (ug/l)	**Pearson Correlation**	1	-0.095	0.010	-0.060	0.183	0.165	0.106
	P-value		0.559	0.950	0.713	0.258	0.310	0.515
Cr (ug/l)	**Pearson Correlation**	-0.095	1	-.323*	0.173	-.353*	-0.077	0.114
	P-value	0.559		0.042	0.285	0.026	0.636	0.484
Zn (ug/l)	**Pearson Correlation**	0.010	-.323*	1	-0.239	0.184	.320*	-0.129
	P-value	0.950	0.042		0.138	0.255	0.044	0.429
Pb (ug/l)	**Pearson Correlation**	-0.060	0.173	-0.239	1	0.168	-0.220	0.072
	P-value	0.713	0.285	0.138		0.299	0.172	0.657
Cd (ug/l)	**Pearson Correlation**	0.183	-.353*	0.184	0.168	1	0.275	-0.258
	P-value	0.258	0.026	0.255	0.299		0.086	0.107
Fe (ug/l)	**Pearson Correlation**	0.165	-0.077	.320*	-0.220	0.275	1	-0.142
	P-value	0.310	0.636	0.044	0.172	0.086		0.383
Cu (ug/l)	**Pearson Correlation**	0.106	0.114	-0.129	0.072	-0.258	-0.142	1
	P-value	0.515	0.484	0.429	0.657	0.107	0.383	
*. Correlation is significant at the 0.05 level (2-tailed).								

3.1.3 Spearman’s correlation

Table 9. Pearson Correlation Coefficients of chemical variables with statistical significance of results

Correlations		pH	TDS (mg/l)	Cond (µs/cm)	N-NO2 mgN/L	N-NO3 mgN/L	N-NH4 mgN/L	Si mg/L	PO4-P mgP/L	Ptotal-P mgP/L	DOC (mg C/l)	POC, mg/l	Monthly Water discharge, m3/s	Daily water discharge, m3/s	Daily SS-IMHE, mg/l	T (oC) - IMHE
po4	**Spearman Correlation**	-0.014	0.038	0.01	.287	0.115	0.164	0.212	1	0.157	-.335	.354	-0.082	-.269	0.208	-0.127
	p-value	0.915	0.775	0.941	0.026	0.391	0.21	0.104	.	0.23	0.009	0.009	0.532	0.038	0.111	0.334

Table 10. Pearson Correlation Coefficients of metal variables with statistical significance of results

Correlations		**Mn (ug/l)**	**Cr (ug/l)**	**Zn (ug/l)**	**Pb (ug/l)**	**Cd (ug/l)**	**Fe (ug/l)**	**Cu (ug/l)**
pb	**Spearman Correlation**	-0.025	0.296	-.344 *	1	0.032	-.560**	0.058
	p-value	0.88	0.064	0.03	.	0.845	0	0.722
fe	**Spearman Correlation**	0.065	-0.22 3	.426 **	-.560* *	0.113	1	-.329*
	p-value	0.69	0.167	0.00 6	0	0.487	.	0.038

3.1.4 Principal component analysis

3.1.4.1 Total variance table (Eigenvalues > 1)

Table 11 . Total variance table for chemical variables

Principal Component	**Initial Eigenvalues % of Variance**		
	Total	% of Variance	Cumulative %
1	3.411	22.739	22.739
2	2.654	17.696	40.435
3	2.170	14.467	54.902
4	1.486	9.904	64.806
5	1.168	7.785	72.591

Table 12. Total variance table for metal variables

Principal Component	**Initial Eigenvalues**		
	Total	% of Variance	Cumulative %
1	1.983	28.322	28.322
2	1.222	17.453	45.775
3	1.117	15.962	61.737

3.1.4.2 Component matrix

Principal Component matrices have been rotated with Varimax method and Kaiser normalization for better interpretation. [42]

Table 13 . Principal component matrix of chemical variables

	Principal Component				
	1	**2**	**3**	**4**	**5**
pH	0.112	0.097	0.224	0.748	0.207
TDS	-0.117	0.054	0.967	0.049	0.011
Cond	-0.056	0.097	0.927	0.186	0.155
N-NO2	-0.003	0.036	0.110	-0.520	0.320
N-NO3	-0.025	0.553	0.204	-0.322	-0.291
N-NH4	-0.547	-0.214	0.085	0.449	0.344
Si	-0.159	0.224	-0.400	-0.625	0.286
PO4-P	-0.192	0.016	0.331	-0.079	0.663
Ptotal-P	-0.036	0.870	-0.024	0.022	-0.012
DOC	-0.402	-0.064	0.151	0.025	-0.637
POC	-0.279	0.651	0.160	-0.067	0.280
Monthly Water discharge	0.886	-0.260	0.019	0.125	0.151
Daily water discharge	0.890	-0.323	-0.002	-0.008	0.081
Daily SS-IMHE	-0.009	0.916	-0.041	0.038	0.038
T (oC) - IMHE	0.748	0.288	-0.284	0.274	-0.185

Table 14 . Principal component matrix of metal variables

	Principal Component		
	1	**2**	**3**
Mn	0.236	-0.088	0.844
Cr	-0.518	0.366	0.000
Zn	0.296	-0.671	-0.111
Pb	0.282	0.821	0.002
Cd	0.892	0.048	0.086
Fe	0.341	-0.539	0.189
Cu	-0.474	0.129	0.592

3.1.4.3 Principal component plot

Figure 2. Principal component plot for chemical variables

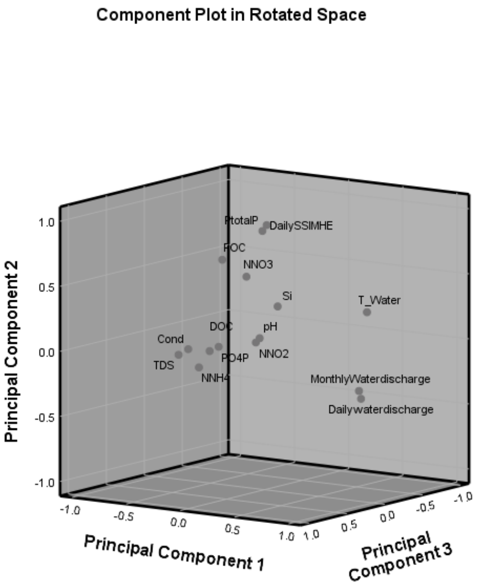
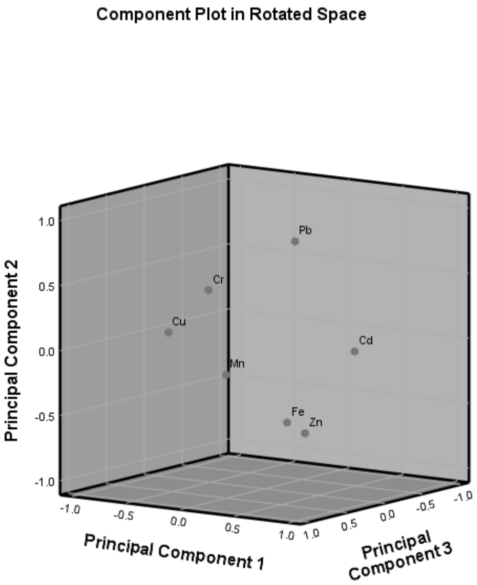


Figure 3. Principal component plot for metal variables



3.2 Discussions

3.2.1 T-Test

Mean values for 26 water quality parameters (Air temperature, air pressure, atmosphere humidity, windspeed, pH, total dissolved solids, conductivity, nitrite, nitrate, ammonium, phosphate, phosphorus total, dissolved organic carbon, particulate organic carbon, monthly water discharge, daily water discharge, daily suspended solid, water temperature, manganese, chromium, zinc, lead, cadmium, iron, copper and silica) for the observed twelve-month period (1/2013 - 12/2013), as well as the results of Independent samples T-Tests and Paired samples T-Test for six regions of the Red River Delta (Yên Bái, Vụ Quang, Hòa Bình, Sơn Tây, Hà Nội) are presented in Table 1, Table 2, Table 3, Table 4 and Table 5.

3.2.1.1 Water Quality Parameters Analysis of upstream regions of the Red River

Independent samples T-Test was used to define if there is a statistically significant correlation between dependent variables (water quality parameters) and independent variables (upstream regions) and test-statistics was applied to represent significantly different variables.

As significance level $p<0.05$ is considered for the significance of variations of means between samples, the dependent variables with statistically significant differences across regions are total dissolved solids, conductivity, nitrite, nitrate, silica, phosphorus total, particulate organic carbon, monthly water discharge, daily water discharge, and daily suspended solid.

The values of metrological properties of three upstream regions show no statistical significance at a significance level of $p<0.05$. This is due to all three regions being mountainous areas.

The pH of surface waters is vital to aquatic life. It affects the ability of aquatic organisms to regulate basic life-sustaining processes, primarily the exchange of respiratory gasses and salts with the water in which they live. Statistical analysis of pH for all three measurement stations doesn’t show significant differences at a significance level $p<0.05$ ($t = 1.847$, $p=0.174$). The highest values of pH were measured in Vụ Quang (8.097), no significant deviations were measured between all three stations.

Total dissolved solids (TDS) are important constituents of water that can affect the environment and human health. In particular, TDS in upstream regions of rivers is of concern because they can contribute to waterborne diseases. The concentrations of TDS in rivers across the Red River Delta vary between 3 upstream regions with a significant difference at the significance level p<0.05(t = 11.121, p=0.000) with the lowest values in Hòa Bình and the highest values in Vụ Quang. It should also be noted that the value of pH in Hòa Bình's river region is also the lowest of all 3 regions, which is the result of the phytoplankton composition of the Hoa Binh reservoir. [43]

Nitrite and nitrate (NO2 and NO3) are two common pollutants that can be found in river water. Nitrite is produced by the decomposition of organic matter and is present in sewage effluent. Nitrate is produced by the natural process of nitrogen fixation and is found in agricultural runoff. Both nitrite and nitrate can be harmful to human health if consumed in large quantities. Results for these variables show statistical significance of variation at a significant level p<0.05. (p=0.000 for both variables) with the values of nitrite and nitrate of Hòa Bình being the lowest (0.0007 mgN/L). This is expected as Hòa Bình has the lowest percentage of land for agricultural use compared to the other two regions. [44]

Phosphorus (P) is essential for plant growth and is found in many rocks and minerals. Phosphate is a major component of fertilizers and is also found in animal bones and manure. In nature, phosphorus usually exists as part of a phosphate molecule (PO4). Both nutrients are important for agriculture and human health. The difference in phosphorus values between regions shows statistical significance (t=21.527, p=0.000) while the difference in phosphate values doesn't (t=1.767, p=0.197). This might be due to the high levels of dissolved solids concentration in certain regions, with values of phosphorus in Yên Bái being the highest (0.29 mgP/L) with significant deviation compared to the lowest values of phosphorus in Hòa Bình (0.05 mgP/L).

Silica (Si) Silica is a major component of soil and rocks, and it is also present in water in the form of dissolved silicon dioxide. As the values of Si across three stations show statistical significance (t=0.399 p=0.001) with the highest value in Yên Bái (6.63 mg/L) and the lowest in Vụ Quang (5.24 mg/L). This might explain why it also has the lowest level of Phosphorus (P) as Vu Quang is not a silicon-rich soil layer, which has the ability to absorb phosphorus. [45]

Particulate organic carbon (POC) in fertilizer river runoff is an important factor in the transport of nutrients and the overall health of aquatic ecosystems. Excess nutrients in rivers can lead to eutrophication, which can degrade water quality and negatively impact aquatic species. In the analysis, the values of POC across regions show significant differences (t = 17.517, p=0.000), with the highest values showing in Yên Bái (1.54 mg/l) and the lowest values showing in Vụ Quang (1.02 mg/l). This is the result of the increased usage of organic fertilizers in the Red River Delta in 2013, which can lead to increased levels of nitrogen and phosphorus in certain areas [46]. This is evident in the values of nitrite, nitrate, and phosphorus in Yên Bái consistently the highest of all three regions.

Suspended solids (SS) are particles in water that are not held in solution and are thus able to float. They are an important water quality concern because they can negatively impact aquatic life and water treatment operations. The suspended solids varied much within three regions: the highest value at Yen Bai (241 mg) and the lowest value at Hoa Binh (11.55 mg). This result is in close relation to the difference in water discharge across the three tributaries of the Red River. The contribution of the tributary that flows through Yên Bái to the discharge of the Red River is the lowest (540.12 m^3/s), while the highest being the Đà river that flows through Hòa Bình (1487.83 m^3/s). Due to the impoundment of two big dams in the Đà tributary, the suspended solids contents decreased remarkably [47].

3.2.1.2 Water Quality Parameters Analysis of the Red River according to river flow

Paired samples T-Test was used to define if there is a statistically significant correlation between a pair of related groups (pairing the upstream regions with the regions after a confluence of tributaries to form the main Red River) on the same continuous, dependent variable (26 water quality parameters) and test-statistics was applied to represent the significance of the difference.

Predicted values for meteorological properties show there is a statistical difference between the pairs of samples, most notably between the upstream regions and the delta region in Sơn Tây at a significance level of $p<0.05$. These results might not be reliable as there are many duplicated predicted observations between regions and there should be a similar difference when comparing the upstream regions to Hà Nội as it is also a delta region. This is the result of Geology in terrain altering the way heat and moisture are exchanged between the atmosphere and the surface. [48]

With a significance level of $p<0.05$ considered, the following variables of chemical properties are statistically significant in the paired samples analysis:

total dissolved solids, conductivity, nitrite, nitrate, phosphorus, phosphate, and water discharge.

As the 3 tributaries merge to form the main Red River, the values of monthly water discharge between upstream regions and delta regions have shown a statistical significance in their differences at a significance level of $p<0.05$.

The values of total dissolved solids, conductivity, nitrite, nitrate, phosphorus, and phosphate have been shown to have statistical significance in their differences at a significance level of $p<0.05$. This is a consequence of the increased use of fertilizers in the Red River Delta in 2013 as mentioned above.

Although the values of metal properties across the three regions show mostly no statistical significance at a significance level of $p<0.005$, it should be noted that the values of all metal properties in Ha Noi are consistently higher than those of the three upstream regions.

Copper (Cu): The Cu concentration at 4 hydrological stations has seasonal differences, the dry season tends to be higher than the rainy season, suggesting that Cu may originate from point discharge sources (mineral mines, industrial wastewater, etc.). …) in the basin

Manganese (Mn): Mn entering the river water environment can be caused by leaching, erosion, metallurgical industrial wastes, batteries, or chemical fertilizer leaching from agricultural land.

Lead (Pb): Pb pollution in Red River water can be caused by industrial production wastewater (for example, wastewater from some battery and chemical factories in Viet Tri, domestic wastewater, activities of means of transportation using leaded gasoline, mining, atmospheric deposition…

Chromium (Cr): The sources of Cr for Red River water can be from point discharge sources such as industrial wastewater (such as paint production, tanning, pigments, dyes, mining, and processing of metals...), and municipal wastes. Cleaning chemicals…

Zinc (Zn): The increase in Zn content in river water in the Red River Delta suggests that the main source of Zinc for Red River water comes from dispersal sources (erosion of soil and rocks in the basin, leaching from agricultural cultivation, etc.)

Cadmium (Cd): This pollution in Cd can be caused by industrial wastewater (for example, the industry producing dyes and paint pigments, glass enamel..,) by domestic wastewater, by washing, by erosion. from agricultural land with excess fertilizer, from the mining of metal mines, or from atmospheric dust deposition in the Red River basin.

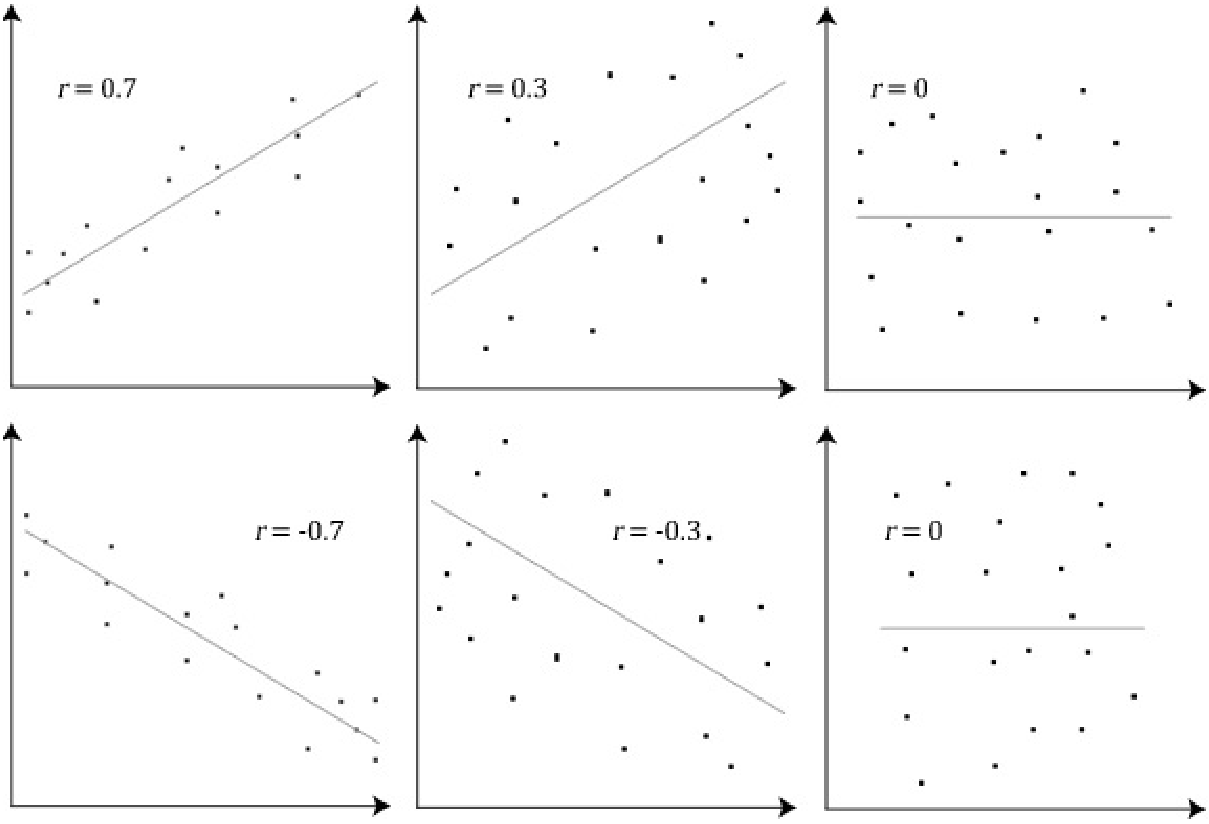
Iron (Fe), Fe content tends to increase in the rainy season, which shows that it is related to point source and emission source from mining minerals, excess fertilizer, and domestic wastewater, due to mechanical erosion from natural geology over the entire river basin. This monitoring result is also consistent with the analysis report of the Center for Environmental Monitoring and the Lao Cai Pollution Control Department: the Red River water in Lao Cai is contaminated with Fe [49].

3.2.2 Correlation analysis of dataset

3.2.1 Pearson's correlation coefficient

Pearson's correlation coefficient ranges from -1 to 1, with the value indicating that the further from 0 it is, the better correlation we get. Here are some figures showing how Pearson's correlation coefficient can represent the strength of a regression model:

Figure 4. Visualization of Pearson's correlation coefficient



and here is the range of correlation coefficient magnitude that show the correlation strength between 2 variables:

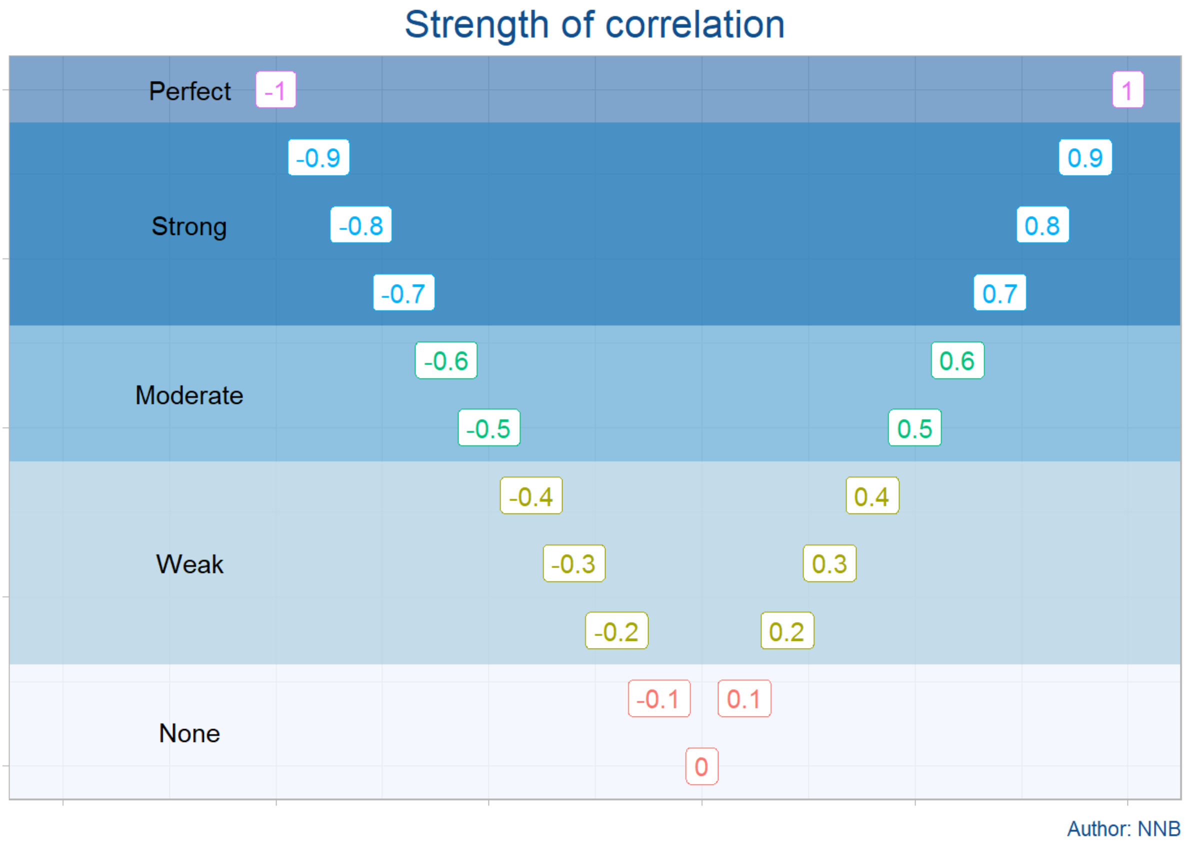


Figure 5. Ranking of Pearson's correlation strength

After Pearson's test, we can conclude which pairs of attributes have an acceptable correlation. Based on the diagram above, we will pick any pairs that have a correlation coefficient equal to 0.45 or above/-0.45 or below as a lower degree could show unsatisfactory results in future analysis [50]. Here are the tables showing attributes with their correlated ones:

Table 15. Showcases of attribute pairs with acceptable Pearson's correlation

Attribute	**Attributes with acceptable correlation**
Average Atm Temp (oC)	Humidity (%RH), Pressure (hPa)
Humidity (%RH)	Wind Speed (km/h), Average Atm Temp (oC)
Pressure (hPa)	Average Atm Temp (oC)
Wind Speed (km/h)	Humidity(%RH)
pH	Si mg/L
TDS (mg/l)	Cond(µs/cm)

Cond (µs/cm)	TDS(mg/l)
N-NO2 mgN/L	N-NH4 mgN/L, Ptotal-P mgP/L, T (oC) - IMHE
N-NO3 mgN/L	Daily SS-IMHE mg/l
N-NH4 mgN/L	N-NO2 mgN/L, T (oC)- IMHE
Ptotal-P mgP/L	N-NO2 mgN/L, N-NO3 mgN/L, POC mg/l, Daily SS-IMHE mg/l
POC, mg/l	Ptotal-P mgP/L, Daily SS-IMHE mg/l
Monthly Water discharge, m3/s	T (oC) - IMHE
Daily SS-IMHE, mg/l	N-NO3 mgN/L, Ptotal-P mgP/L
T (oC) - IMHE	N-NO2 mgN/L, N-NH4 mgN/L, Monthly Water discharge m3/s
Si mg/L, PO4-P mgP/L, DOC(mgC/l), Mn (ug/l), Cr (ug/l), Zn (ug/l), Pb(ug/l), Cd (ug/l), Fe (ug/l), Cu (ug/l)	None

To test if our results are true, the p-value will be used. The lower the p-value, the greater the statistical significance of the observed difference. The decision is made if the result is statistically significant depending on whether or not the p-value is lower than 5%. Upon calculating the p-value for each pair above, the data indicates that no pair has a p-value exceeding 0.05. Therefore, we can conclude that the results are statistically significant.

3.2.2 Spearman's correlation

Spearman's rank correlation **measures the strength and direction of association between two ranked variables** [51]. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function. Here is the table containing how the value of the correlation coefficient created by the Spearman test will indicate the strength of association between two attributes:

Table 16. Spearman's correlation degree according to its value

Grading Standards	**Correlation Degree**
$\rho = 0$	No correlation
$0 < \lvert\rho\rvert \leq 0.19$	Very weak
$0.20 < \lvert\rho\rvert \leq 0.39$	Weak
$0.40 < \lvert\rho\rvert \leq 0.59$	Moderate
$0.60 < \lvert\rho\rvert \leq 0.79$	Strong
$0.80 < \lvert\rho\rvert \leq 1.00$	Very strong
1.00	Monotonic correlation

We will only pick attribute pairs that have the correlation degree of moderate or above as lower degree could show unsatisfactory results in future analysis:

Table 17. Showcases of attribute pairs with acceptable Spearman's correlation

Attribute	**Attributes with acceptable correlation**
PO4-P mgP/L	None
Pb (ug/l)	Fe (ug/l)
Fe (ug/l)	Zn (ug/l), Pb (ug/l)

To test if our results are true, the p-value will be used. The lower the p-value, the greater the statistical significance of the observed difference. The decision is made if the result is statistically significant depending on whether or not the p-value is lower than 5%. Upon calculating the p-value for each pair above, the data indicates that no pair has a p-value exceeding 0.05. Therefore, we can conclude that the results are statistically significant.

3.2.3 Principal component analysis

3.2.3.1 Chemical Variables

According to Table 11, Table 13, and Figure 2, the PCA analysis results accommodate for the 15 chemical variables representing the chemical properties of water quality in the Red River of which 5 principal components return eigenvalues greater than 1 [52]. These 5 principal components explain 72.591% of the variance of the dataset. With a factor loading of 0.3 being the recommended threshold to suppress the coefficient of variables, the variables best represent their principal components are as follows [53]:

Principal Component	**Variables with coefficient greater than +/-0.3**
1	N-NH4, DOC, Monthly Water Discharge, Daily Water Discharge, T - IMHE
2	N-NO3, Ptotal-P, POC, Daily Water Discharge, Daily SS-IMHE
3	TDS, Cond, Si, PO4-P
4	pH, N-NO2, N-NO3, N-NH4, Si
5	N-NO2, N-NH4, PO4-P DOC

The amount of water discharged throughout the various river regions and the concentration of total dissolved solids in the Red River are strongly associated, as was previously discussed, and the characteristics of nitrogen and phosphorus are also highly correlated. The results of the PCA analysis strongly imply that the dataset accurately captured the outcomes of the t-test analysis.

3.2.3.2 Metal Variables

Table 12, Table 14, and Figure 3 show that the PCA analysis results take into account 7 metal variables that indicate the metal characteristics of the Red River's water quality, of which 3 main components return eigenvalues larger than 1 [52]. These five main factors account for 75.089% of the variance in the data.

Principal Component	**Variables with coefficient greater than +/-0.3**
1	Cr, Cd, Fe, Cu
2	Cr, Zn, Pb, Fe
3	Mn, Cu

After the analysis, the correlation between the heavy metals is accurately reflected in the dataset. The results of the PCA analysis strongly imply that the dataset accurately captured the outcomes of the t-test analysis.

4. Conclusion

River water quality and spatial and temporal trends along the 3 big rivers in the northern part of Vietnam were assessed in this project by applying 27 parameters to 2013's rivers' atmospheric, chemical, and metal data. From the parameters of river water quality, the rivers become slightly more polluted as they flow from upstream to downstream. The cause of this phenomenon might come from the plethora of agricultural areas, residential areas, and industrial plants that lie along the rivers. While agricultural areas and residential areas affect river water mostly on chemical components through fertilizers, pesticides, and domestic waste, industrial plants have more influence on metal components through industrial waste. Aquacultural areas might be another cause as well, chemicals used in these areas to cleanse water and promote fish growth also have repercussions on chemical components of river water.

Though the deal of releasing waste into the rivers affects water quality greatly, many activities evolving these actions still have to be done for economics' sake. This could still be under control with proper waste processing procedures but illegal littering is still being done by many people for their own profit. The most important thing we have to do is to raise people's awareness of the impact of river water quality in order to preserve a healthy resource for our generation and the future ones.

5. Acknowledgments

For this project, we would like to express our gratitude to our supervisors Dr. Nghiêm Thị Phương and Dr. Lê Phương Thu for their valuable guidance and advice. We are grateful to them for their help and support throughout this project.

We would like to personally thank Dr. Phương, who served as our main supervisor. Thanks to Dr. Phương's valuable assistance, we were able to understand and fully utilize our technical tools, and motivated us to work better as a team.

As for Dr. Thu, it was because of her incredible knowledge in the science department, that we were given important insights into the analyzed data set. She has made it much easier for us to understand the unfamiliar aspects of the project and helped us realize the connections and correlations of the data.

6. References

1. *Global Historical Weather and Climate Data | Weather and Climate*, https://tcktcktck.org/. Accessed 31 January 2023.
2. *Current weather and forecast - OpenWeatherMap*, https://openweathermap.org/. Accessed 31 January 2023.
3. “Atmospheric temperature.” *Wikipedia*, https://en.wikipedia.org/wiki/Atmospheric_temperature. Accessed 7 February 2023.
4. “What is humidity? Why measure & what your levels mean.” *Airthings*, https://www.airthings.com/en/what-is-humidity. Accessed 7 February 2023.
5. Rosenberg, Matt. “Air Pressure and How It Affects the Weather.” *ThoughtCo*, 4 February 2020, https://www.thoughtco.com/low-and-high-pressure-1434434. Accessed 7 February 2023.
6. “What is Wind Speed? (with pictures).” *AllTheScience*, 11 January 2023, https://www.allthescience.org/what-is-wind-speed.htm. Accessed 7 February 2023.
7. “4 Factors Affecting PH | PT Hyprowira Adhitama.” *Hyprowira Adhitama*, 8 July 2020, https://hyprowira.com/en/blog/factors-affecting-ph. Accessed 7 February 2023.
8. “What are Total Dissolved Solids in Water?” *EPA Water Consultants*, https://epa-water.com/what-are-total-dissolved-solids-in-water/. Accessed 7 February 2023.
9. “Maine Health and Environmental Testing Lab - PH and Conductivity - Division of Public Health Systems | MeCDC.” *Maine.gov*, https://www.maine.gov/dhhs/mecdc/public-health-systems/health-and-environmental-testing/ph-cond.htm. Accessed 7 February 2023.
10. “Indicators: Salinity | US EPA.” *Environmental Protection Agency*, 11 July 2022, https://www.epa.gov/national-aquatic-resource-surveys/indicators-salinity. Accessed 7 February 2023.
11. “Total Nitrogen/Total Phosphorus and Chlorophyll a: Volunteer Data Entry and Information.” *Indiana Clean Lakes Program*, https://clp.indiana.edu/volunteer-data/phosphorus-chlorophyll.html. Accessed 7 February 2023.
12. “Nitrate (NO3) and Nitrite (NO2).” *Maine Environmental Laboratory*, http://maineenvironmentallaboratory.com/?p=1071. Accessed 7 February 2023.
13. “Ammonium NH4 contamination in water: Current status and Solutions.” *Eurofins Scientific*, 8 November 2021, https://www.eurofins.vn/en/news/eurofins-news/ammonium-nh4-contamination-in-water-current-status-and-solutions/. Accessed 7 February 2023.
14. “Surface Water: Phosphate the Fertilizer Promoting Stream Degradation.” *Drinking Water Quality*, https://www.knowyourh2o.com/outdoor-4/phosphate-in-surface-water-streams-lakes-ponds. Accessed 7 February 2023.
15. “Indicators: Phosphorus | US EPA.” *Environmental Protection Agency*, 16 June 2022, https://www.epa.gov/national-aquatic-resource-surveys/indicators-phosphorus. Accessed 7 February 2023.
16. “Tap Water Can Be Deadly. How's Your Total Organic Carbon?” *ATS Innova*, 15 January 2017, https://atsinnovawatertreatment.com/blog/tap-water-deadly-total-organic-carbon/. Accessed 7 February 2023.
17. Bruckner, Monica Z. “Measuring Dissolved and Particulate Organic Carbon (DOC and POC).” *SERC - Carleton*, 16 December 2022,

https://serc.carleton.edu/microbelife/research_methods/biogeochemical/organic_carbon.html. Accessed 7 February 2023.

18. “River Discharge and Factors Affecting River Discharge.” *OMICS International*, 10 February 2015, https://www.omicsonline.org/blog/2015/02/10/3807-River-Discharge-and-Factors-Affecting-River-Discharge.html. Accessed 7 February 2023.
19. Campbell, Brian. “What is Total Suspended Solids (TSS)?” *Water & Wastes Digest*, 9 September 2021, https://www.wwdmag.com/instrumentation/suspended-solids-monitors/article/10939708/what-is-total-suspended-solids-tss. Accessed 7 February 2023.
20. “Water Temperature - Environmental Measurement Systems.” *Fondriest Environmental*, https://www.fondriest.com/environmental-measurements/parameters/water-quality/water-temperature/. Accessed 7 February 2023.
21. “Manganese Fact Sheet.” *Water Quality Association*, https://wqa.org/resources/manganese/. Accessed 7 February 2023.
22. “Chromium in Drinking Water.” *Canada.ca*, 22 July 2015, https://www.canada.ca/en/health-canada/programs/chromium-drinking-water/chromium-drinking-water.html. Accessed 7 February 2023.
23. “PUBLIC HEALTH STATEMENT Zinc.” *Agency for Toxic Substances and Disease Registry*, https://www.atsdr.cdc.gov/ToxProfiles/tp60-c1-b.pdf. Accessed 7 February 2023.
24. “Lead in Drinking Water | Sources of Lead | CDC.” *Centers for Disease Control and Prevention*, https://www.cdc.gov/nceh/lead/prevention/sources/water.htm. Accessed 7 February 2023.
25. “Cadmium in Drinking Water | Contaminant Source | Water Testing Treatment.” *Drinking Water Quality*, https://www.knowyourh2o.com/indoor-6/cadmium. Accessed 7 February 2023.
26. “Iron & Manganese in Groundwater.” *Regional District of Nanaimo*, https://www.rdn.bc.ca/cms/wpattachments/wpID2284atID3808.pdf. Accessed 7 February 2023.
27. “Aquatic Life Criteria - Copper | US EPA.” *Environmental Protection Agency*, 29 August 2022, https://www.epa.gov/wqc/aquatic-life-criteria-copper. Accessed 7 February 2023.
28. Bhatti, Manmeet Singh. “Why Silica in Water is Important to Measure.” *LinkedIn*, 10 August 2022, https://www.linkedin.com/pulse/why-silica-water-important-measure-manmeet-singh-bhatti/?trk=pulse-article_more-articles_related-content-card. Accessed 7 February 2023.
29. Martín, Diego, et al. “Towards Outlier Sensor Detection in Ambient Intelligent Platforms—A Low-Complexity Statistical Approach.” *Sensors*, vol. 20, no. 15, 2020, p. 4217, https://www.mdpi.com/1424-8220/20/15/4217.
30. “PLOT Subcommand (EXAMINE command).” *IBM*, 10 August 2022, https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=examine-plot-subcommand-command. Accessed 7 February 2023.
31. Hoaglin, David C., and Boris Iglewicz. “Fine-Tuning Some Resistant Rules for Outlier Labeling.” *Journal of the American Statistical Association*, vol. 82, pp. 1147-1149.

32. “Shapiro–Wilk test.” *Wikipedia*, https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test. Accessed 31 January 2023.
33. “Missing data.” *Wikipedia*, https://en.wikipedia.org/wiki/Missing_data. Accessed 7 February 2023.
34. Graham, J. W., et al. “How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory.” *Preventative Science*, vol. 8, no. 3, 2007, pp. 208-213.
35. Hair, J. F., et al. “Multivariate data analysis.” *Pearson Education, Upper Saddle River*, vol. 7, 2010.
36. “Kruskal-Wallis H Test in SPSS Statistics | Procedure, output and interpretation of the output using a relevant example.” *Laerd Statistics*, https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php. Accessed 7 February 2023.
37. “Pearson correlation coefficient.” *Wikipedia*, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. Accessed 31 January 2023.
38. “Spearman's rank correlation coefficient.” *Wikipedia*, https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient. Accessed 7 February 2023.
39. “Principal component analysis.” *Wikipedia*, https://en.wikipedia.org/wiki/Principal_component_analysis. Accessed 4 February 2023.
40. Jaadi, Zakaria. “Principal Component Analysis (PCA) Explained.” *Built In*, 8 August 2022, https://builtin.com/data-science/step-step-explanation-principal-component-analysis. Accessed 7 February 2023.
41. Kaiser, Henry F., and John Rice. “Little Jiffy, Mark Iv.” *Educational and Psychological Measurement*, vol. 34, 1974, pp. 111-117, doi:10.1177/001316447403400115
42. Kaiser, H. F. “The varimax criterion for analytic rotation in factor analysis.” *Psychometrika*, vol. 23, 1958, pp. 187-200.
43. Thuy, Thi Duong, et al. “Seasonal variation of phytoplankton assemblage in Hoa Binh reservoir, north of Vietnam.” *J. Viet. Env.*, vol. 6, no. 1, 2014, pp. 22-26.
44. “Quyết định 1467/QĐ-BTNMT 2014 công bố kết quả thống kê diện tích đất đai 2013.” *Thư viện pháp luật*, 21 July 2014, https://thuvienphapluat.vn/van-ban/Bat-dong-san/Quyet-dinh-1467-QD-BTNMT-2014-cong-bo-ket-qua-thong-ke-dien-tich-dat-dai-2013-254339.aspx. Accessed 7 February 2023.
45. Schaller, J., et al. “Silicon as a potential limiting factor for phosphorus availability in paddy soils.” *Scientific Reports*, vol. 12, no. 16329, 2022, https://doi.org/10.1038/s41598-022-20805-4.
46. “Sử dụng phân bón sản xuất lương thực bảo vệ môi trường và giảm phát thải KNK.” *Viện môi trường nông nghiệp*, http://www.iae.vn/Data/upload/files/3_PQHA_Phan%20bon%20voi%20MT_BHH_F.pdf. Accessed 7 February 2023.
47. Le, Quynh Thi Phuong, et al. “Water quality of the Red River system in the period 2012 - 2013.” *J. viet. env.*, vol. 6, no. 3, 2014, https://doi.org/10.13141/jve.vol6.no3.pp191-195.

48. “Tiểu luận Thủy văn Môi trường: Các đặc trưng về địa hình và dòng chảy sông Hồng.” *TaiLieu.VN*, https://tailieu.vn/doc/tieu-luan-thuy-van-moi-truong-cac-dac-trung-ve-dia-hinh-va-dong-chay-song-hong-1682590.html. Accessed 7 February 2023.
49. “Đánh giá thực trạng quan trắc cảnh báo ô nhiễm sông Hồng đoạn chảy qua địa phận tỉnh Lào Cai - 11.” Tailieuthamkhao.com, https://tailieuthamkhao.com/danh-gia-thuc-trang-quan-trac-canh-bao-o-nhiem-song-hong-doan-chay-qua-11-12548. Accessed 7 February 2023.
50. “Correlation Coefficient.” *RPubs*, 14 May 2021, https://rpubs.com/tranquangquy_ictu/769561. Accessed 7 February 2023.
51. Gupta, Aryan. “Spearman’s Rank Correlation: The Definitive Guide To Understand.” *Simplilearn*, 2 February 2023, https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation. Accessed 7 February 2023.
52. Jackson, Donald A. “Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches.” *Ecology*, vol. 74, no. 8, 1993, https://doi.org/10.2307/1939574.
53. Field, A. “Discovering Statistics using SPSS.” pp. 674-675.