PREDICTIVE MODELING OF POLLUTION IN RIVER BASINS USING MACHINE LEARNING TECHNIQUE

HASLINDA BINTI ABDUL SAHAK MCS241004_MCST 1043 CHAPTER 4

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 4

INITIAL RESULTS

4.1 Introduction

This chapter discusses the initial results of applying various machine learning models to predict water quality in river basins. It probably involves collecting data on physical and chemical parameters of water quality, such as Total Suspended Solids (TSS), Ammoniacal Nitrogen (NH3N), and Biochemical Oxygen Demand (BOD). Decision Trees, Artificial Neural Networks, K-nearest neighbors, Naïve Bayes, Support Vector Machine, Random Forest, and Gradient Boosting are then used as the feature engine for machine learning algorithms. The initial results suggested that Gradient Boosting achieved the highest accuracy, sensitivity, and f-measure when predicting water quality and hence, this model seemed to have the best predictive power. With the most significant features highlighted for the model, it will be able to provide the significant parameters that most influence predictions with respect to water quality.

4.2 Data Visualizations

In general, the river quality in 2023 has slightly deteriorated as compared to the previous year, based on the Water Quality Index. In this country, every year, there are 1,353 manual river water quality monitoring stations with a coverage area of 672 rivers. This Department of Environment, Malaysia will continue its monitoring to support the strategies, programs, and activities concerned for the sustainable management of the environment effectively. River water quality monitoring is done to determine the status of river water quality and to detect changes in river water quality. Water samples were collected from designated stations for in-situ measurement and sent to the laboratory as well, for analysis aimed at determining the criteria based on the sciences of physic-chemical and biological. The WQI is used to indicate the level of pollution and the corresponding suitability in terms of water use according to the National Water Quality Standards for Malaysia (NWQS). NWQS is an ambient standard to protect aquatic biodiversity and it is also used as a benchmark for the setting of uses for certain rivers.

WQI is an index range from 0 to 100 where the range is divided into three (3) categories clean, slightly polluted, and polluted. The index is derived based on six (6) parameters, which are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (AN), suspended solids (SS), and pH. Reporting water quality status is reported in two (2) approaches: River water quality status by the river and River water quality status by station. Initial result from the Environmental Quality Report (EQR) 2023, out of the 672 rivers monitored, 25 (4 percent) were polluted, 161 (24 percent) were slightly polluted, and 486 (72 percent) showed good water quality. In 2022, there were six polluted rivers in Malaysia, a decrease from seven polluted rivers in the previous year. The lowest number of polluted rivers in Malaysia was in 2015, with 5 rivers in total. In the same year, there were a greater number of clean rivers in Malaysia.

For water quality status reporting by river, shown in Figure 4.1 the WQI determination is based on the median calculation where the median WQI for the entire data observed at each station to obtain the river WQI. The reporting of water quality

status by river is a reflection of the quality status of the entire river based on the number of stations monitored.

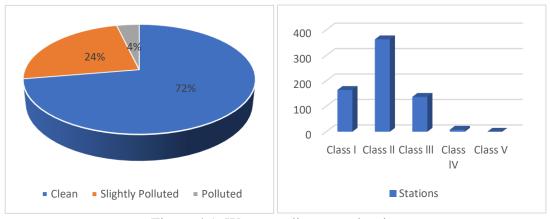


Figure 4.1: Water quality status by river

For water quality status reporting by station, the WQI determination is based on the median calculation of six (6) WQI at the station. The reporting of water quality status by the station is a reflection of the quality status at the station only and is shown in Figure 4.2.

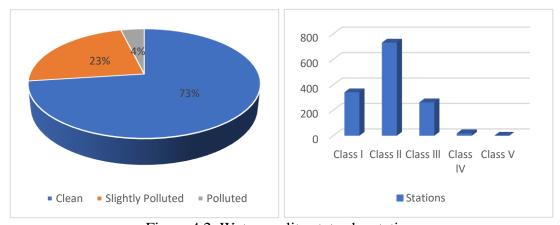


Figure 4.2: Water quality status by station

The water quality status of a total of 672 rivers was monitored every year. Out of the 672 rivers monitored, the water quality of 486 (72%) rivers was indicated as clean, 161 (24%) rivers were indicated as slightly polluted, and 25 (4%) rivers were indicated as polluted. The trend of the monitored river water quality is shown in Figure 4.3. For WQI classification, as many as 164 (25%) rivers are in Class I, 363 (54%)

rivers are in Class II, 137 (20%) rivers are in Class III, and eight (8) (1%) rivers are in Class IV.

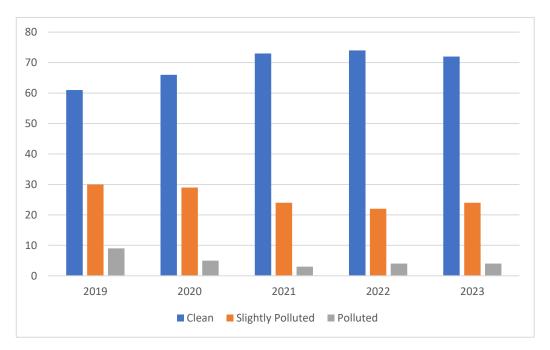


Figure 4.3: River Water Quality Trend from 2019 until 2023

The main indicators were considered to determine river water quality based on the parameters of BOD, AN, and SS. The cause of deterioration of water quality is due to the discharge of the pollution load either from a point source or from a nonpoint source. Point sources can be described as pollution sources with specific identifiable discharge points that are unchanged over time. Sectors such as industry, livestock, and sewage treatment systems fall under this category. Meanwhile, the nonpoint sources such as agricultural activities, earthworks, mining, and sullage (domestic wastewater other than sewage such as kitchen and bathroom wastewater) do not have specific identifiable discharge points and the locations are varied over time. This makes it difficult to estimate the amount of released pollution loads.

4.2.1 Stores Data

The details of dataset construction and preparation steps for the first research objective. Table 4.1: Longitudinal study-the data was collected from 144 river basins over some time, starting from January 1, 2019, and ending on December 31, 2024. The main focus will be on monitoring pollution levels based on different indicators, such as BOD5, NH3N, and SS. The dataset also contained the date of monitoring, status of pollution, and proportion. The above analysis indicated a remarkable improvement in the water quality of river basins, particularly for BOD5 and SS. This study will cover predictive modeling of the Sungai Johor River basin pollution.

Table 4.1: River Basin Pollution Monitoring Dataset

	Date	Basins Monitored	Pollution Indicator	Pollution Status	Number of basins	Proportion
0	1/1/2019	144	bod5	clean	53	36.80555556
1	1/1/2019	144	bod5	slightlypolluted	39	27.08333333
2	1/1/2019	144	bod5	polluted	52	36.11111111
3	1/1/2019	144	nh3n	clean	12	8.333333333
4	1/1/2019	144	nh3n	slightlypolluted	74	51.38888889
5	1/1/2019	144	nh3n	polluted	58	40.27777778
6	1/1/2019	144	SS	clean	78	54.16666667
7	1/1/2019	144	SS	slightlypolluted	36	25
8	1/1/2019	144	SS	polluted	30	20.83333333
••					•••	
16417	31/12/2024	144	bod5	clean	124	86.11111111
16418	31/12/2024	144	bod5	slightly_polluted	11	7.638888889
16419	31/12/2024	144	bod5	polluted	9	6.25
16420	31/12/2024	144	nh3n	clean	58	40.27777778
16421	31/12/2024	144	nh3n	slightly_polluted	37	25.69444444
16422	31/12/2024	144	nh3n	polluted	49	34.02777778
16423	31/12/2024	144	SS	clean	127	88.19444444
16424	31/12/2024	144	SS	slightly_polluted	7	4.861111111
16425	31/12/2024	144	SS	polluted	10	6.94444444
		-				

[16425 rows x 6 columns]

Figure 4.4 Number of main river basins in states of Malaysia. The x-axis is the chart represents states while y-axis represents total number of main river basins. Each of the bars in the chart has been labeled with the corresponding state and the number of its river basins. From the chart, Sabah has the highest number of main river basins,

at 75, followed by Sarawak, which has 40. Johor has 20 main river basins, while Pahang and Terengganu each have 12. Perak is shown to have 11 main river basins, and Negeri Sembilan has 8. Kedah is listed with 7 main river basins, Melaka and Selangor each have 6, and Pulau Pinang has 5. Kelantan has 4 main river basins, WP Kuala Lumpur has 3, and Perlis has the least, with only 1 main river basin. The above information provides good insight into how the main river basins are distributed across various states in Malaysia. Such information is useful in studies related to geography, environment, and resource management.

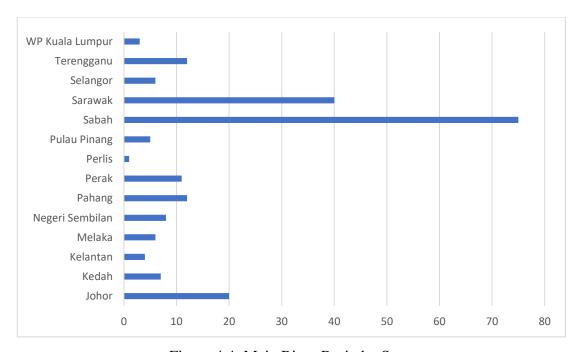


Figure 4.4: Main River Basin by State

Figure 4.5 Summary of water quality assessment for thirteen states of Malaysia. Each state is represented by a set of three bars representing the number of river basins falling into the three categories of clean, slightly polluted, and polluted. The height of each bar directly gives the number of river basins in that category of pollution for the corresponding state. Instantly, the chart visually underlines those states that have clean river basins-for example, Johor, Pahang, and Sabah-and the ones with higher shares of polluted or slightly polluted basins, like Pulau Pinang and WP Kuala Lumpur. Such coloring for each pollution level enhances clarity in visual form, and thus it presents faster comparisons between states.

However, the simplicity of Figure 4.5 limits the depth of the analysis that it can provide. For example, the particular water quality indicators that were used to classify the pollution status of the river basins are not specified (e.g., BOD, dissolved oxygen, nutrient levels). Also, the period over which data has been collected is not specified, so one cannot know whether the data presented reflects conditions of a snapshot or longer-term ones. The presentation of information without any indication of methodologies adopted for collecting data-sampling frequency, locations, and analytical techniques-suggests issues with the reliability and interpretability of the results. In this way, although the chart provides an overview, further analysis would in any case require more contextual detail to reach comprehensive conclusions.

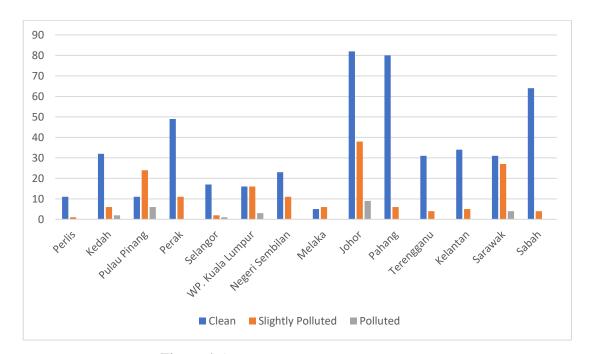


Figure 4.5: Pollution in River Basins by State

4.2.2 Item Data

Table 4.2 gives an overview of the water quality for different states of Malaysia. The first column, "States," refers to the geographical areas within which the respective assessments were done. The column "Clean" represents those river basins in a state that had been classified as "clean," or where the water quality is good. The slightly polluted column represents the number of river basins corresponding to slightly

polluted, which is understood to be a state of moderate levels of contamination. The last column, polluted is the number of river basins that correspond to the category of polluted, meaning serious contamination of water.

The data indicates, based on general observation, that Johor and Pahang contain the largest number of clean river basins, which means these states generally possess good water quality. Sabah also reflects good water quality conditions with a high number of clean basins and very few polluted basins. Pulau Pinang presents the highest number of slightly polluted basins and hence could be considered a potential risk regarding water quality and WP. It is expected that in Kuala Lumpur, high counts for both slightly polluted and polluted basins may suggest a probable impact due to urbanization and industrial activities.

Table 4.2: River Basin Pollution Monitoring by State

State	Clean	Slightly Polluted	Polluted	
Perlis	11	1	0	
Kedah	32	6	2	
Pulau Pinang	11	24	6	
Perak	49 11		0	
Selangor	17	2	1	
WP. Kuala Lumpur	16	16	3	
Negeri Sembilan	23	11	0	
Melaka	5	6	0	
Johor	82	38	9	
Pahang	80	6	0	
Terengganu	31	4	0	
Kelantan	34	5	0	
Sarawak	31	27	4	
Sabah	64	4	0	

Figure 4.6 shows that Johor has the highest proportion of 36% when looking at the river basins that are polluted, signifying the high level of pollution the state has. Pulau Pinang contributes 24%, equally considered as a high contributing to pollution. Sarawak and WP Kuala Lumpur are contributing fairly, at 16% and 12%, respectively. Selangor and Kedah thus appear to have the least problem with only 4% and 8% of the basins polluted, respectively.

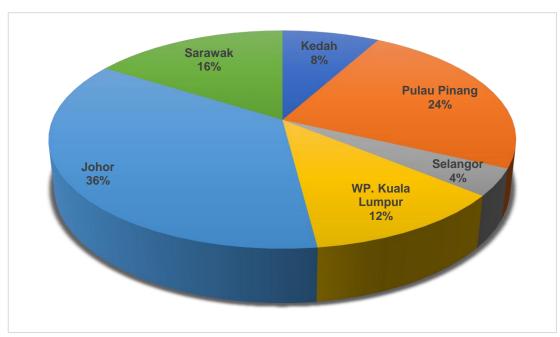


Figure 4.6: Polluted River Basin by State

Waste from manufacturing industries involved in the production of electronics, textiles, and food contaminates the rivers in Johor and Pulau Pinang because of improper waste disposal, which releases hazardous chemicals, heavy metals, and solid wastes into the water. Intensive agriculture, including large-scale animal husbandry and plantations, also pollute rivers through the addition of chemical fertilizers and pesticides and wastes from animals to the water system. This encourages eutrophication-a process reducing oxygen levels in the water and highly detrimental to aquatic life. For Pulau Pinang, tourism adds to pollution through inappropriate household and food waste disposal and possible oil or chemical spills from activities in the water.

Other contributors include construction sites through soil erosion and sediment runoff, as well as chemicals such as cement and paint. Finally, transportation includes shipping and vehicular traffic around rivers introduce oil spills, among other wastes into the water. Again, this is by no means an exhaustive list, and at most locations of river pollution, the actual pollution sources are caused by a combination of some or all of these elements and several others. Specific research will be required in finding out precisely the sources of pollution in each locality.

4.2.3 Merged Data

In this section, the manual river water quality index is concatenated by stacking these data frames horizontally as these data have the same format but with different values. Then, the Data frames are then merged with the manual river water quality index and continuous river water quality index on state, river, and station ID as shown in Figure 4.7 and Figure 4.8



Figure 4.7: Manual River Water Quality Index

Figure 4.7 shows the Water Quality Index (WQI) sampling data for different stations in Johor, Malaysia. The table lists the station number, state, river, station ID, and WQI values for various months. Each month's WQI value is color-coded within the table. There are no questions to answer within the provided image; it's purely a data presentation.

Meanwhile, Figure 4.8 shows a table of hourly water quality readings from four different water intake points in Johor, Malaysia. The table includes the number, state, river, water intake location, station ID, and hourly water quality readings from 00:00 to 14:00. The readings appear to be some type of index, with values ranging from the low 60s to the high 80s. The values are color-coded, with yellow indicating lower values and blue indicating higher values. There are no questions to be answered in the image; it's simply a presentation of data.



Figure 4.8: Continuous River Water Quality Index

The information presented in Table 4.3, shows the average quality index for each state due to river basin pollution. It thus gives a proper overview of different states from the year 2019 until 2023 concerning river basin pollution. Within this table are states such as Perlis and Kedah, Johor having an average quality index for every year across this period. The color-coded values provide an easy-to-view visual indication of the level of pollution, ranging from green for good quality to red for poor quality. A "Status" column has been added to the table, summarizing each state as SP: Slightly Polluted, P: Polluted, or C: Clean based on data available for 2023. For example, in the year 2023, Perlis has the notation SP, which means slight pollution, whereas Johor has P, representing higher severity in terms of pollution.

The table highlights the variations in water quality over time, emphasizing areas that have either improved or deteriorated. For example, Perlis shows a generally increasing trend in its quality index, moving from 71 in 2019 to 81 in 2023, indicating an improvement in water quality. In contrast, Johor has a low and rather consistent quality index, mostly varying within the 30s and 40s, indicating continuous pollution. This series of information is vital to monitor and ease the environmental challenges arising due to the pollution in river basins across various states.

Table 4.3: Average quality index for river basin pollution in every state

State	2019	2020	2021	2022	2023	Status
Perlis	71	75	76	74	81	SP
Kedah	64	70	77	76	73	SP
Pulau Pinang	66	65	63	56	41	P
Perak	85	88	91	86	86	C
Selangor	66	70	71	77	77	SP
WP. Kuala Lumpur	49	71	63	44	57	P
Negeri Sembilan	69	74	73	80	79	SP
Melaka	65	72	71	74	72	SP
Johor	36	29	33	39	33	P
Pahang	89	88	92	91	91	C
Terengganu	87	87	88	89	91	C
Kelantan	81	80	80	85	85	C
Sarawak	89	90	88	91	91	С
Sabah	73	74	80	69	70	SP

Figure 4.9 highlights the trend in Biochemical Oxygen Demand (BOD5) from 2000 to 2021. BOD5 is crucial because it measures the oxygen needed by microorganisms to decompose organic matter in water, with higher values indicating more pollution. From 2000 to 2014, there was a steady increase in BOD5 levels, reflecting a rise in water pollution. This period's increasing pollution could result from heightened industrial activities, urbanization, and insufficient wastewater treatment facilities. The upward trend underscores the escalating stress on water bodies due to human activities.

However, the trend in BOD5 levels decreased between 2014 and 2021, as represented by the graph, reflecting great improvements in water quality. It may be because of many reasons, such as a rise in the stringency of environmental laws, better wastewater treatment processes, and industrial practices with greener methods. This downtrend indicates an effective implementation of pollution control measures and increased awareness of the need to protect this vital resource. This is a good omen for the recovery and gives testimony to effective targeted environmental intervention.

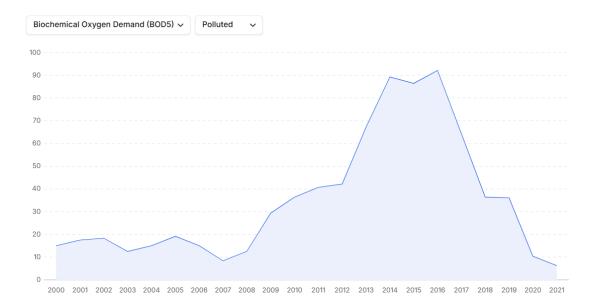


Figure 4.9: River Water Quality Stations Trend on *Biochemical Oxygen Demand* (BOD) Sub-Index

Figure 4.10 presents the trend of Ammoniacal Nitrogen (NH3-N) from the year 2000 to 2021. NH3-N is a certain type of nitrogen pollution usually discharged into water bodies through agricultural runoff, sewage, and industrial waste. Before 2019, NH3-N showed an increasing trend, which indicates a deteriorating situation in water pollution. The probable causes are increasing agricultural activities, urbanization, and inappropriate treatment of waste, leading to increased nitrogenous substances discharged into the water sources.

However, the graph for NH3-N levels showed a decline in the years 2020 and 2021, which is an improvement in water quality. This may be because of a number of factors that include more stringent pollution control regulations, improved wastewater treatment technologies, and changes in agricultural activities to more sustainable farming techniques. This effort seemed to be reducing nitrogen pollution in the water body and proved the efficacy of pinpointed environmental intervention and policies.

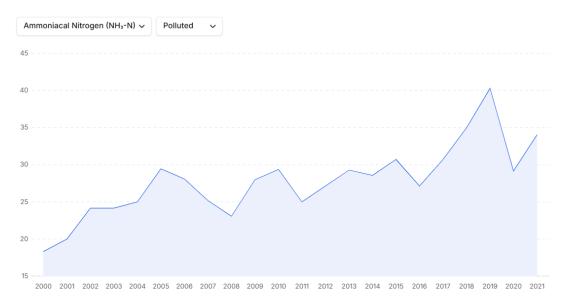


Figure 4.10: River Water Quality Stations Trend on Ammoniacal *Nitrogen (NH₃-N)* Sub-Index

Figure 4.11 presents the trends in Suspended Solids (SS) levels in water from 2000 to 2021. Suspended solids are small particles that float in water, generally comprising soil, organic matter, and man-made products from various industries. High levels of SS result in turbid water, not only harmless to the aquatic organisms but also not fit for drinking and other purposes. Overall, the trend of SS decreases from 2000 to 2014, reflecting an improvement in water quality. This positive impact can be explained by increased pollution control, improved waste water treatment, or improved practices of conservation.

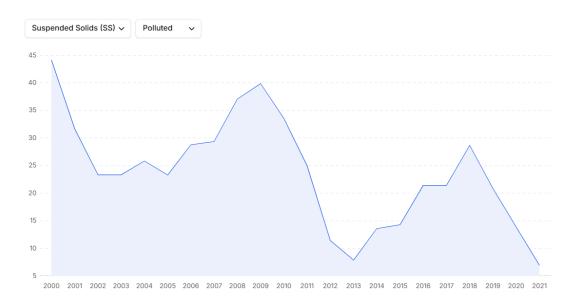


Figure 4.11: River Water Quality Stations Trend on Suspended Solids (SS), Sub-Index

All of these have slight increases in SS levels from 2014 until 2018, showing partial setbacks in keeping the levels of pollution low. This might also be because of increased industrial activities, construction, or changes in land use. From 2018 until 2021, the levels went down once more to reflect new impetus towards pollution control. Despite the progress, the slight increase in SS levels between 2014 and 2018 indicates that continuous and sustained efforts are essential to keep improving water quality and protecting aquatic ecosystems.

The analysis of water quality data for Biochemical Oxygen Demand (BOD5), Ammoniacal Nitrogen (NH3-N), and Suspended Solids (SS) from 2000 to 2021 reveals both challenges and progress in addressing water pollution. The trend of BOD5 shows an upward trend until 2014, reflecting the deteriorating pollution level, but then a sharp decline from 2014 to 2021, reflecting successful pollution control and eventual improvement in water quality. Similarly, NH3-N continued to increase until 2019, reflecting an increasingly serious nitrogen pollution problem; however, a sharp decrease in 2020 and 2021 reflected effective regulatory action and technological gains in treating wastewater.

The SS trend, in general, improved from the year 2000 up to 2014, then had a setback from 2014 to 2018, and continued to decline until 2021. This reflects the fact that, although the efforts put into the reduction of SS have generally been effective,

further vigilance and sustained actions are needed to maintain and further improve the water quality. Put together, these results again bring into focus the need for unceasing environmental interventions, regulatory measures, and improved technologies in pollution control for the health and safety of water resources in the long term.

4.3 Trend of River Water Quality Monitoring Stations

Figure 4.12, illustrates water quality trends from 2019 to 2023, categorizing water bodies as Clean, Slightly Polluted, or Polluted. Overall, there has been an improvement, with a notable increase in the percentage of clean water bodies and a decrease in polluted ones. In 2019, the percentage of clean water bodies was 62%, followed by 67% in 2020 and then 75% in 2021. Then it came down to 74% in 2022 and to 73% in 2023. Similarly, polluted water bodies decreased from 9% in 2019 to just 3% in 2021, but in subsequent years it slightly increased to 4%.

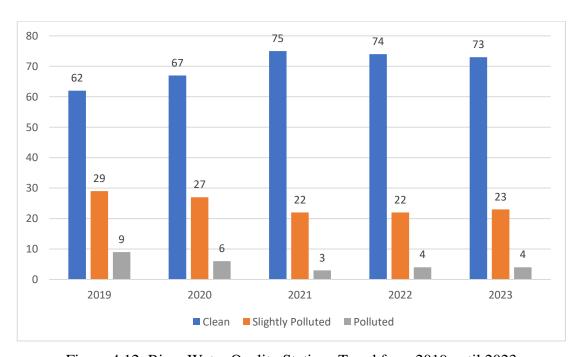


Figure 4.12: River Water Quality Stations Trend from 2019 until 2023

Although the trend is upward, there is a slight fluctuation every year. For example, water bodies that were slightly polluted decreased from 29% in 2019 to 22% in 2021 and rose slightly to 23% in 2023. Such fluctuations reveal that it is still challenging to ensure improvement in water quality constantly and that further vigilance with effective measures against pollution is called for.

The water quality could have been greatly improved with pollution control legislation becoming stricter, methods of treating wastes becoming superior, and good practices by industries and farmers. This increased awareness and additional conservation efforts. However, the limitations in the data are that sources are not specified, and factors such as climate change and population growth are not considered; thus, while there is progress, more work needs to be done for sustained improvements to take place, with all influences on water quality tackled.

4.4 Trend of River Quality Monitoring on Sub-index

Figure 4.13, Trend of river water quality stations in BOD sub-index, 2019-2023, presents large variations on the trend of river water quality during the period under consideration. In 2019, only 16% of the stations representing the water quality of rivers were classified as clean, while a further 44% were slightly polluted and another 40% had been marked as polluted. In 2020, the percentage of clean stations reached as high as 34%, increased to 53% in slightly polluted stations, and drastically dropped to 13% for polluted stations. This trend was continued in 2021, with 78% of stations falling in the clean category, 13% being slightly polluted, and only 9% polluted-the biggest increase within these years.

This growth in 2021 was almost stable during the following years, considering only minor fluctuations. In fact, in 2022, all stations at 75%, slightly polluted at 14%, and 11% were polluted, showing a slight deterioration compared with the previous year. However, in 2023, the percentage of clean stations managed to rise to 78%, even

if those slightly polluted went down to 12% and those polluted remained stable at 10%. This would go to imply that efforts at pollution control and quality management of water continue to pay off, though vigilance and further improvements must be made to keep such gains.

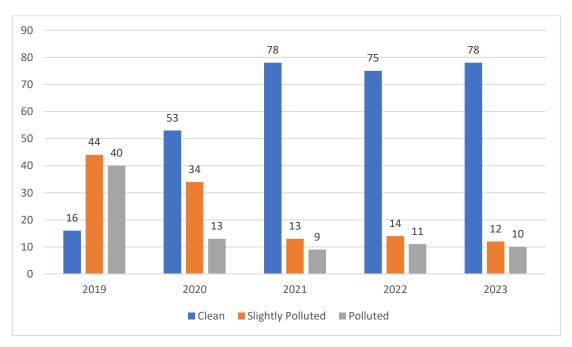


Figure 4.13: River Water Quality Station Trend on BOD Sub-index from 2019 until 2023

Figure 4.14 River water quality station trend on AN Sub-index from 2019 up to 2023, describes the share of river water quality stations in categories as clean, slightly polluted, and polluted within these years. In 2019, 34% of the stations were clean while 28% were slightly polluted and 38% polluted. In 2020, the figure dropped to 32% for clean stations, while Slightly Polluted stations went up to 30%, and those that were considered as Polluted stood at 38%. In the year 2021, the situation has considerably improved with a rise in clean stations to 51%, Slightly Polluted to 20%, and Polluted stations to 29%.

During 2022, it repeated with 49%, 24% more stations being Slightly Polluted, and the % coming down to 27% regarding the presence of a 'Polluted' level. For the year 2023, the numbers were 51%, those in slightly polluted decreased to 23%, while polluted went further to 26%. These changes indicate a general improvement in the water quality of the province, ups and downs in the percentage variations of Slightly

Polluted and Polluted stations, while the number of clean stations is growing across the years. The improvement points to effective environmental measures and better efforts on pollution control, though there is a need for continued attention in this respect.

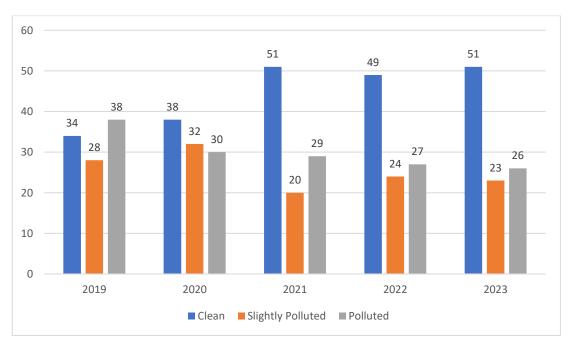


Figure 4.14: River Water Quality Station Trend on AN Sub-index from 2019 until 2023

Figure 4.15 Trend of SS Sub-index for River Water Quality Station from 2019 up to 2023. The percentage share of river water quality stations classified as clean, slightly polluted and polluted are given for the mentioned years. For example, in the year 2019, the percentage share of the classified station as a clean category was 62%, slightly polluted was 12% while polluted ones reached 26%. The following year, 2020, saw a slight improvement with 63% clean stations, 11% slightly polluted stations, and 26% polluted stations remaining constant. There was a significant improvement in 2021, with 77% clean stations, a reduction to 9% for slightly polluted stations, and a further drop in polluted stations to 14%.

In 2022, this percentage decreased to 74%, while those considered slightly polluted remained at 8%, and those considered polluted increased to 18%. In 2023, this trend continued at 74% for clean stations, 8% for slightly polluted, and polluted ones at 18% stability. This shows the overall development of the water quality in these

five years, with distinct improvement between 2019 and 2021. Some minor fluctuations afterward underline that it is of utmost importance to continue maintaining and further enhancing the quality by systematic monitoring and adequate measures for pollution control.

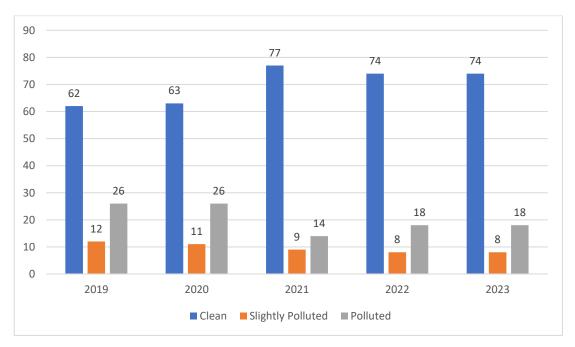


Figure 4.15: River Water Quality Station Trend on SS Sub-index from 2019 until 2023

River water quality, on the other hand, for 2019 and projected to 2023, presented some improvements and challenges on BOD, Ammoniacal Nitrogen, and Suspended Solids. In contrast, BOD had a greater improvement of higher clean and lesser polluted stations. This change became more visible in the year 2021 and after that. Correspondingly, its AN sub-index reported progress and, in general, an increase of a high rate in clean stations for 2021, maintaining positive tendencies with some oscillation during the following years.

The SS sub-index also improved, with more stations becoming clearer and less polluted, although there was a moderate fluctuation year by year. In general, these trends reflect the positive direction in water quality and effective measures for pollution control. However, this requires further efforts and adaptive strategies to maintain the gained pace and to overcome various challenges in water quality management.