



Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil

Marlon Valentini¹ · Gabriel Borges dos Santos¹ · Bruno Muller Vieira²

Received: 5 October 2020 / Accepted: 21 December 2020 / Published online: 11 January 2021
© The Author(s) 2021

Abstract

Accurate assessment of the type and extent of water pollution is a difficult and complicated task. Therefore, the use of the Water Quality Index (WQI) proves to be a useful tool, as this index has the advantage of resulting in a single number that is easy to communicate and understand. One of the statistical methods that can be used to develop a new WQI equation for a given water body is Multiple Linear Regression (MLR). Therefore, this work aims to develop a new WQI equation for Mirim Lagoon through MLR and to evaluate the suitability of that equation. For this, 154 samples collected at 7 monitoring points of Mirim Lagoon over three years were analyzed, from which the original WQI results were obtained. With these results, MLR was applied, generating a new WQI equation with only three variables: phosphorus, dissolved oxygen (DO) and thermotolerant coliforms. This MLR model obtained an R^2 coefficient equal to 0.728, demonstrating that 72.8% of the data variability was explained by the equation generated by this model. Through the paired t test it was possible to demonstrate that the results of the new WQI did not differ significantly from the results of the original WQI, obtaining significance values greater than 0.05 for both absolute WQI values and quality ranges (p value 0.886 and 0.467, respectively), attesting to the adequacy of the new equation. The reduction in the number of variables for monitoring water quality has several advantages, from financial ones, as it makes monitoring less costly, even in reducing the eclipse effect. Therefore, it is concluded that it is of great importance to generate an equation with fewer variables for monitoring the quality of the waters of Mirim Lagoon.

Keywords Statistical model · Statistical analysis · Correlation · Significance test

1 Introduction

Along the entire length of the water bodies, there are several processes that can influence the quality of its waters [28]. From a historical point of view, having urban centers developed around freshwater rivers and lakes, it can be said that these water resources play a great role in the development of populations [29]. Currently, the

population faces a series of environmental, economic and social problems, which causes concerns about the environment to increase [28]. Influenced by the new standards of modern life, as well as the ever-increasing population growth rate, the demand for water is increasing. This and other factors, such as the inappropriate disposal of solid waste and industrial and urban effluents, end up influencing a worrying increase in the pollution of water resources

Supplementary material The online version of this article (<https://doi.org/10.1007/s42452-020-04005-1>) contains supplementary material, which is available to authorized users.

✉ Marlon Valentini, marlon.valentini@hotmail.com | ¹Postgraduate Program in Water Resources, Federal University of Pelotas, Pelotas, RS, Brazil. ²Environmental and Sanitary Engineering, Universidade Federal de Pelotas, Pelotas, RS, Brazil.



SN Applied Sciences (2021) 3:70 | <https://doi.org/10.1007/s42452-020-04005-1>

[41]. Therefore, it is of great interest that the pollution of these natural resources is controlled, reduced and monitored [31].

According to Piedras et al. [30], the Mirim Lagoon River Basin is one of the main transboundary river basins in South America. Located on the Atlantic coast of South America, more specifically between the extreme south of Brazil and the north of Uruguay, Mirim Lagoon is one of the most important freshwater lagoons in the Brazilian territory, being of great importance for the State of Rio Grande do Sul, both economically and ecologically. This important natural resource is the second largest water body in the country, occupying a territorial extension of 375 thousand hectares of water surface, shared between the Brazilian and Uruguayan territories [27, 42].

As a large part of the surrounding land is used for agricultural activities, Mirim Lagoon presents several sources of pollution. These outbreaks can also be attributed to the natural releases of industrial effluents and domestic sewage in the São Gonçalo channel, the channel responsible for connecting Mirim Lagoon to Lagoa dos Patos [27]. Also, due to population growth in the world, surface waters, such as rivers, lakes and ponds are susceptible to contamination from a diverse variety of polluting sources, which causes a deterioration in their quality [13, 14]. In this context, the quality of the waters of Mirim Lagoon has been suffering degradation due to the anthropic activities developed in its surroundings [12, 42]. Therefore, its conservation is necessary, as well as the use of reliable techniques for monitoring the water quality of this lagoon.

Accurate assessment of the type and extent of water pollution is a difficult and complicated task, and one of the main problems faced in monitoring water quality is the complexity of analyzing a large number of variables [21, 23]. According to Alves et al. [2], the Water Quality Index (WQI), proves to be a useful tool to represent the quality of water bodies. Developed by the National Sanitation Foundation (NSF), the WQI was introduced in Brazil primarily through the Environmental Company of the State of São Paulo (CETESB), which adapted it to the country's specificities [3, 15]. This index, adapted by CETESB, consists of nine variables; however, in virtue of the different local characteristics of each region of Brazil, it can be adjusted to better portray the reality of each water body, or better, hydrographic basin [2, 43].

More specifically, the WQI can be modeled for a specific situation, that is, it can be simplified by taking into account only some of the environmental variables that really affect the quality of a given source. This reduction in the number of variables, in addition to facilitate the construction and elaboration of a quality index, makes it more economically attractive, enabling its use in several hydrographic basins in the country. In addition, another positive point

of using an WQI with few environmental variables is the reduction in the eclipse effect. This effect is characterized by the attenuation, for example, of the negative impact of one of the variables due to the stability of the others [38].

Another important tool that can be used very well for monitoring water resources is the use of statistical methods. These methods have the advantage of providing a simple and objective interpretation from a previously complex set of data. The application of these methods can facilitate the manipulation of this data, which helps to better understand the effects of the processes that occur in the water body [18, 49]. In addition, the application of advanced statistical methods is of great use to extract important information without significant loss of accuracy [48].

Several studies around the world have been carried out in order to model quality indexes that best suit the conditions of each region. Al-Janabi et al. [1] assessed the water quality of the Tigre River in Baghdad, using a Canadian water quality index.

One of the statistical methods that can be used thinking of developing a new equation for the WQI of a given water body is the Multiple Linear Regression (MLR). Coming from Linear Regression, MLR is a statistical model related to the treatment of time series of data, where its analysis is a forecast of values of one or more response variables to which it uses a set of explanatory variables [26].

Therefore, this work aims to develop a new WQI equation for Mirim Lagoon through multiple linear regression analysis models and to evaluate the suitability of these models for monitoring the water quality of this water resource, as well as assessing whether the results of WQI generated by this new equation can be used with precision to monitor the quality of the waters of this lagoon.

2 Methodology

2.1 Study area

The Mirim Lagoon hydrographic basin is located on the Atlantic coast of South America, on the coastal plain of Rio Grande do Sul in the extreme south of Brazil, and occupies 47,362 km² of direct contribution and 55,110 km² when including the coastal strip, in which they are located. The main baths of the region are inserted, including Lagoa Mangueira, connected exclusively to Mirim Lagoon by the Taim hydrological system. This important lagoon constitutes the second largest water body with lake characteristics in the country, connecting Lagoa dos Patos through the São Gonçalo channel [27, 42].

With an average depth of 5 meters and 180 km long along its central axis, northeast-southwest, which is almost

parallel to the direction of the most frequent wind, the Mirim Lagoon in the past was affected by the brackish waters of the São Gonçalo Channel coming from of Patos’ Lagoon, which harmed irrigated rice crops in the region. So, in 1977 the Eclusa’s dam was built in the São Gonçalo Canal in order to avoid this [37]. Furthermore, with an estimated average overflow to Patos’ Lagoon of 700 m³/s, the Mirim Lagoon can also connect to Mangueira Lagoon through swamps during flood events, which ends up forming the largest lagoon complex in South America [20, 37]. According to Munar et al. [25], the main tributaries of Mirim Lagoon are the Jaguarão, Taquari and Cebollati rivers, the first on the Brazilian side and the last two on the Uruguayan side. It is also worth noting that the region in question has a subtropical climate, with an average annual

rainfall ranging from 1250 to 2000 mm and an average annual temperature between 14 and 18 °C [8].

The main use of its waters is through direct extraction for agricultural use in rice fields in Brazilian and Uruguayan territory and, in addition, the waters of this lagoon are also intended, after treatment, for human consumption mainly for the municipalities of Pelotas and Rio Grande [22, 36]

2.2 Collections, monitoring points and laboratory analysis

For the purposes of this study, a total of 22 sampling campaigns were carried out, divided into 7 monitoring points over the years 2015, 2016, 2017, totaling 154 samples. The monitoring points of this study can be seen in Table 1. These points comprise the two sides of the lagoon, these being in the directions Pelotas—Jaguarão and Pelotas—Santa Vitória do Palmar, as shown in Fig. 1.

All sample campaigns were performed on the banks of the Mirim Lagoon at the collections points.

For the collection of samples, in each campaign and at each monitoring point, the following flasks were used: two different plastic flasks previously sanitized and sterilized, a 2-L flask for samples intended for analysis of total solids, turbidity and pH and a 1-L flask for samples referring to nitrogen and phosphorus analysis. For the determinations of DO and BOD, two 250-mL glass vials containing a

Table 1 Mirim Lagoon monitoring points

Identification		Geographic coordinates	
LM1	Praia Pontal	32°20'52"	52°49'21.5"
LM2	Fazenda Bretanha	32°29'14"	52°58'14.9"
LM3	Fazenda São Francisco	32°38'25.6"	53°08'56.8"
LM4	Capilha	32°29'23"	052°35'33"
LM5	Curral Alto	32°44'47.41"	52°40'35.99"
LM6	Vila Anselmi	32°54'31"	52°48'08"
LM7	Porto Santa Vitória	33°29'51"	53°26'09"

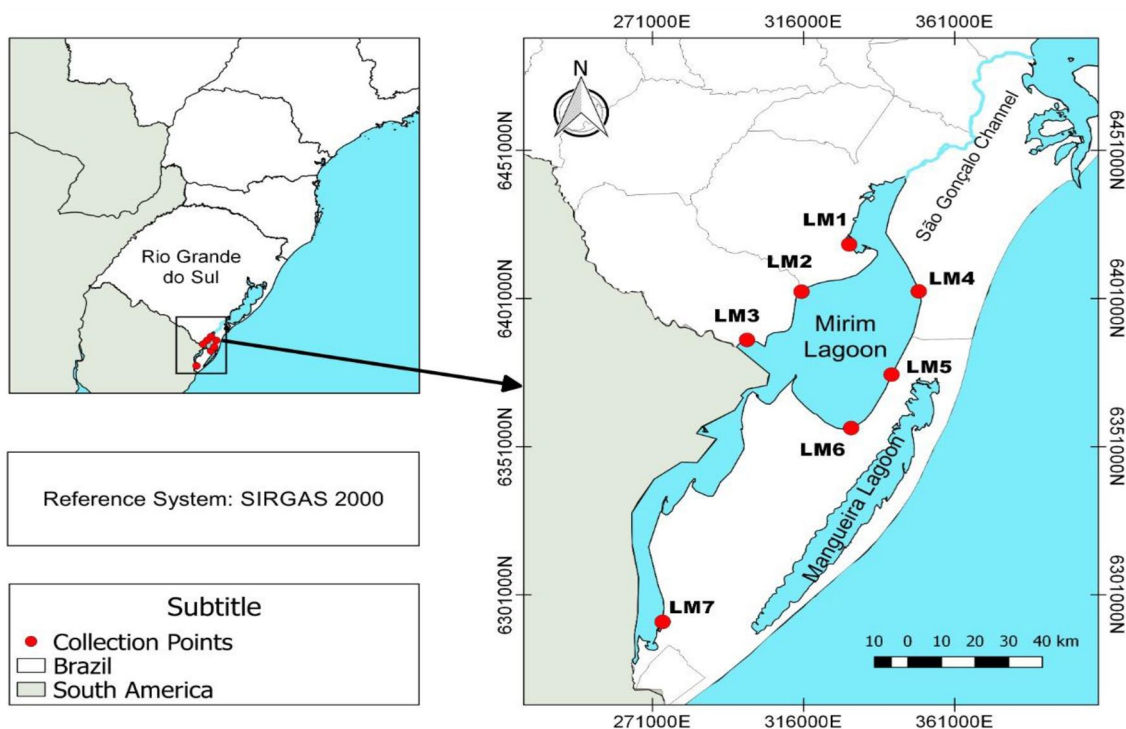


Fig. 1 Mirim Lagoon monitoring points. *Source:* Author

sealing system were used. For the microbiological analysis (thermotolerant coliforms), a 250-mL glass flask, properly sterilized, was used.

All collected samples were duly stored and sent to the water and effluent laboratory of the Mirim Lagoon Hydrographic Basin Development Agency (LMA), where the physical, chemical and microbiological analyzes considered within the WQI calculation (CETESB) were carried out for the study of the water quality of this water body. All analyzes were performed according to the methodology described by APHA [4]. The results of these analyzes were taken from the LMA website, which were publicly accessible during the period covered in this study.

2.3 Original WQI

As previously mentioned, this index appeared in the United States, at the National Sanitation Foundation, due to the need to monitor the quality of water resources and was adapted in Brazil by CETESB in order to better adapt to national water characteristics [15]. Both in its form developed by NSF and in its adaptation by CETESB, the WQI is calculated using the weighted product of each of its parameters, as shown in Eq. 1. This calculation results in a number from 0 to 100, with 100 being better quality.

$$WQI = \prod_{i=1}^n q_i^{w_i} \quad (1)$$

where q_i is the quality of the i th parameter, also a number between 0 and 100, which is obtained through the quality graph as a function of the measured concentration. w_i is the weight of the i th parameter represented by a number between 0 and 1.

To perform the WQI calculation, 9 determinations are performed according to the adaptation made by CETESB of the parameters defined by the NSF. In Brazil, the nitrate parameter was replaced by total nitrogen and the total phosphate was replaced by total phosphorus, maintaining the same weights and quality curves established by NSF. Therefore, the parameters that constitute the WQI in Brazil are: temperature, dissolved oxygen (DO), thermotolerant coliforms, pH, biochemical oxygen demand (BOD), total nitrogen (N), total phosphorus, turbidity and total solids [3].

Each parameter considered within the WQI has a weight for the calculation of this index, as fixed in relation to its relevance to the overall conformation of water quality. These weights are shown in Table 2.

The results obtained by calculating the WQI can be represented in quality ranges, as shown in Table 3. These

Table 2 Relative weights of parameters for calculating the WQI. Source: ANA, 2007

Parameters	Weights
<i>WQI parameters and their respective weights</i>	
Dissolved oxygen (DO)	0.17
Thermotolerant coliforms (Therm. Coliforms)	0.15
Hydrogenionic potential (pH)	0.12
Biochemical oxygen demand (BOD)	0.10
Temperature (Temp.)	0.10
Total nitrogen (N)	0.10
Total phosphor (P)	0.10
Turbidity (Turb)	0.08
Total solids (TS)	0.08

Table 3 Classification of the WQI for the state of Rio Grande do Sul. Source: ANA [3]

WQI value	Water quality
91–100	Great
71–90	Good
51–70	Acceptable
26–50	Bad
0–25	Terrible

ranges vary according to the state to which the studied water body is inserted, with the range for the state of Rio de Janeiro being adopted here. Grande do Sul.

For Centeno and Ceconello [6], the calculation of the WQI consists of the use of variables that reflect the changes that occurred in the region of the micro-basin of the water resource, be they anthropic or natural. For these, the WQI has the advantage of resulting in a value that is easy to communicate and understand, since in this index the interactions and changes in the analyzed parameters, having different units of measurement, result in a single number.

2.4 Statistical analysis

First, the sample distribution of the data covered in this study was analyzed. For this, normality tests were performed. To prove or reject the normality of the data, the Kolmogorov–Smirnov (K–S) and Shapiro–Wilk (S–W) tests were used at a significance level of 0.05. For these tests, the null hypothesis (H_0) considers the distribution to be normal and for p value < 0.05 this hypothesis is rejected. The K–S test is based on the largest difference between the theoretical and the empirical cumulative probabilities, in absolute value. Regarding the S–W test, its calculation can be visualized in Eq. 2.

$$W = \frac{(\sum_{i=1}^n a_{n-i+1}(x_{n-i+1} - x_1)^2)}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

where $i = 1, 2, \dots, n$ is the sample size, x_i = ordered values of the sample, \bar{x} = mean value of the sample, x_{n-i+1} constants generated from mean, variances, and covariates of statistical order of a sample size N and a normal distribution, and W statistical variable of SW.

After the analysis of the sample distribution, a correlation matrix with its coefficients and significance was used, in order to assess the correlation between the chemical and microbiological variables studied and those variables with the WQI. The correlation coefficient to be used depends on the sample distribution of the data, which is why normality tests were previously performed.

For data that follows a normal distribution, for example, Pearson's coefficient is the most commonly used, whereas for data that do not have sample normality, nonparametric coefficients are recommended, such as the Spearman Rhó coefficient, also known as Spearman's coefficient [17].

For the modeling of a new WQI equation, the multiple regression analysis (MLR) methodology was used. This analysis typically uses a methodology called ordinary least squares (OLS) that fits a straight line in the scatter plot of the variables, which minimizes the sum of the squared residuals. In other words, using the OLS method, the linear regression analysis traces a line with a more accurate fit in the center of the data distribution, minimizing its residuals. Through this method, regression analysis allows to quantitatively assess the relationship that specific variables have with a result of interest. [47].

First, regression models will be evaluated using all variables that have a significant correlation according to the aforementioned correlation matrix and, later, the model equation that presents the best adjustments will be extracted. This equation, then, will be used in the calculation of a new water quality index for the water resource studied here, using only the variables established by the regression analysis.

With the new WQI equation obtained by the regression analysis and this new index for each sample already calculated, we proceed with the Student's t test to assess whether there was a significant variation between the original WQI and the new WQI. This test will be done for the numerical results of WQI and for their quality ranges.

The t test for paired samples measures the significance of the variation between pairs of dependent samples. Since the results of the original WQI were used to perform the regression analysis, the results of the new WQI are not independent of these, so the use of the paired t test.

This test, like all significance tests, is a hypothesis test that aims to evaluate a null hypothesis and an alternative.

According to this test, if p value < 0.05 , the null hypothesis is rejected, which says that there is no significant variation between the original WQI and the new WQI, with 95% confidence. That is, if $p > 0.05$ the indices—original and new—are not significantly different.

3 Results and discussion

3.1 Original WQI

As already discussed, the WQI developed by NSF in the USA and adapted in Brazil by CETESB, hereinafter referred to as the original WQI, is calculated by the weighted product of each of its nine parameters. For this study, the WQI of 7 monitoring points in 22 collections was calculated, generating 154 WQI results, as shown in Table 4.

In Brazil, the original WQI adapted by CETESB widely is used. Recent studies have used this WQI in assessing the quality of Mirim Lagoon, which found a likely influence of anthropic activities on its quality, such as agricultural activities developed in its surroundings [39, 46]. Santos et al. [34, 35], also studying a water body located in southern Brazil, used this WQI and found that the water quality index decreased as the water body studied approached the urban area, also indicating pollution probably from activities anthropic.

3.2 Normality test

All variables related to the original WQI, as well as the result of this index, were evaluated for their sample distribution. According to the K-S and S-W tests, most variables do not follow a normal distribution, as they obtained a p value < 0.05 . Therefore, the correlation coefficient used in the next stage of the study will be Spearman's coefficient.

3.3 Correlation matrix

Once the sampling distribution of the data has been evaluated, it follows with the correlation matrix. This matrix, with its coefficients and p values of significance, can be seen in Table 5. The focus is given on the correlations of the parameters with the final result of the original WQI, as these are the correlations that will be used to define the regression models later.

As can be seen in Table 5, the variables phosphorus, thermotolerant coliforms, DO, turbidity and total solids have a significant correlation with the result of the original WQI. Therefore, these are the variables that will be evaluated in the linear regression models later. Regarding the intensity of the correlations, the threshold defined by [19] is used, which says that correlations with a coefficient

Table 4 Original WQI of Mirim Lagoon

Collect	ORIGINAL WQI per monitoring point						
	LM1	LM2	LM3	LM4	LM5	LM6	LM7
1	62	65	63	65	58	59	63
2	72	68	63	89	54	85	71
3	72	61	75	61	63	84	72
4	64	75	83	71	71	76	83
5	69	81	77	82	90	63	81
6	70	63	69	70	68	70	73
7	68	64	61	71	60	79	83
8	75	78	67	59	52	61	69
9	89	79	89	76	60	79	81
10	64	56	65	71	63	69	70
11	60	67	73	41	37	66	72
12	70	56	63	68	63	53	62
13	61	54	52	41	44	39	43
14	63	60	66	72	72	53	76
15	58	56	62	52	52	75	78
16	69	57	56	62	71	46	71
17	63	53	61	63	53	53	65
18	57	54	50	51	48	49	59
19	62	69	54	60	58	59	75
20	61	62	58	59	58	52	63
21	64	72	64	67	73	71	91
22	63	66	78	61	73	68	79

Table 5 Parameter correlation coefficients—original WQI and their respective significance

WQI	Temp.	P	N	Therm. Coliforms	BOD	DO	Turbidity	pH	TS
C.	0.032	-0.601	0.120	-0.710	0.127	0.484	-0.326	0.069	-0.483
Sig	0.692	0.000	0.137	0.000	0.115	0.000	0.000	0.394	0.000

Those in bold are those that had significance according to the statistical test used

equal to or greater than 0.5, in absolute value, are strong correlations. Thus, the phosphorus—WQI and thermotolerant coliforms—WQI correlations are strong, in addition to being significant. It should be noted that the intensity of the correlations is evaluated in absolute value because negative correlations are inversely proportional, while positive correlations have direct proportionality. In other words, the increase in the concentrations of phosphorus and thermotolerant coliforms decreases the final result of the WQI while the increase in the DO concentration increases the WQI, for example.

Fia et al. [16] found high concentrations of phosphorus in their study on Mirim Lagoon, which corroborates the fact that this variable has a strong and significant correlation with the WQI of this water resource. The variables phosphorus, turbidity and total solids are associated with pollution from agricultural activities, with phosphorus being one of the main components of many chemical additives used in plantations [32]. As for the

thermotolerant coliforms and DO variables, both can be related to the presence of pollution by organic matter, also causing a deterioration in water quality [45].

Also, with regard to Thermotolerant Coliforms, their presence in surface waters may be related, according to Santos et al. [34, 35], to sources of agricultural pollution and domestic sewage. Tormam et al. [44] also in a study of Mirim Lagoon, corroborates this inference, showing that coliforms are associated with degradation of water quality in this lagoon, mainly associated with the discharge of domestic sewage. Furthermore, the work developed by Droese et al. [10] also shows that Mirim Lagoon may be influenced by sources of anthropogenic pollution such as effluent discharge. This same study associates turbidity with physical and mineral pollution in Mirim Lagoon.

3.4 Linear regression analysis

Knowing which variables have significant correlation with the original WQI, these variables were used to construct linear regression models. Analyzing the residual statistics of the regression models using these 5 variables, it was possible to infer that the model that had the best fit was the one that used only the variables phosphorus, DO and thermotolerant coliforms, thus only these three parameters will be used for the following analyzes. Since phosphorus is representative of pollution from agricultural activities [32] and OD and thermotolerant coliforms representative of pollution by organic matter [45], these three variables may well be considered to encompass the processes that influence the quality of Mirim Lagoon.

The summary of the models with the 5 variables that have significant correlation with the original WQI can be seen in Table 6. As can be seen, the addition of predictors in the models 1, 2 e 3 causes significant variation, which means that they add more information to the regression models. Although the addition of the turbidity variable has a significant effect on the regression models, this variable will not be used for the following analyzes due to the results obtained by the analysis of the waste statistics, as stated above. The TS variable has no significant effect in this MLR analysis, as can be seen in Table 6.

Still on Table 6, according to Wheelan [47], R^2 is a measure of the total size of the variation, which is explained by the regression equation of this model. For model 3 of the regression analysis in this study, we have an R^2 equal to 0.728, which means that this model explains 72.8% of the WQI variation in this regression.

Siqueira et al. [40], in a study developed in Parauebas River, located in the northern region of Brazil, also used linear regression, in addition to principal component analysis, to assess the influence of the variables that of the WQI. In this study, the authors found that the variable that most contributed to the variability of

Table 6 Linear regression analysis—summary of the models

Model	R	R ²	Change of R ²	Sig. change of F
2	0.707	0.500	0.093	0.000
3	0.853	0.728	0.228	0.000
4	0.885	0.783	0.055	0.000
5	0.886	0.785	0.002	0.227
Predictors: (constant), phosphor				
Predictors: (constant), phosphor, DO				
Predictors: (constant), phosphor, DO, thermotolerant coliforms				
Predictors: (constant), phosphor, DO, thermotolerant coliforms, turbidity				
Predictors: (constant), phosphor, DO, thermotolerant coliforms, turbidity, TS				

water quality was the thermotolerant coliforms, indicating anthropogenic pollution, such as the discharge of organic effluents into the Parauebas River. In addition, demonstrating the importance of statistical methods in the assessment of water quality, several recent studies use methods such as linear regression, principal component analysis, correlation and variance tests and multiple comparisons, among others, for the evaluation of water quality [9, 34, 35, 45, 48, 49].

Table 7 shows the ANOVA of these models. In this table, it is possible to observe the decrease of the residuals and the average squares along the models and the significance of the variation from one model to another. This significance, confirmed by the ANOVA test, is measured with a 95% confidence level and a 0.05 significance level. Therefore, those models that obtain a p value of less than 0.05 have significant variability among themselves [5]. Model 3 presents the lowest residues, demonstrating that this is the best model to be used. The ANOVA test confirms the significant variation between the models, that is, it confirms that there is information gain with the addition of predictors up to model 3, as already noted in the summary.

The coefficients of the linear regression equation, using model 3, can be seen in Table 8. From this table was extracted Eq. 3, which will henceforth be the new WQI equation for this study. It is worth emphasizing that, although the regression method can be used for

Table 7 ANOVA for regression models 1, 2 and 3

Model		Sum of squares	Medium square	Sig.
1	Regression	7098	7098	0.000
	Residue	10,327	68	
	Total	17,425		
2	Regression	8713	4357	0.000
	Residue	8711	58	
	Total	17,425		
3	Regression	12,693	4231	0.000
	Residue	4731	31	
	Total	17,425		

Table 8 Coefficients of model 3 of the linear regression analysis

Model	Coefficients		SE	Significance (p value)
3	(Constant)	60.2	2.9	0.000
	Phosphor	-6.2	0.5	0.000
	OD	1.9	0.4	0.000
	Thermotolerant Coliforms	-0.008	0.001	0.000

Table 9 Waste statistics

	Minimum	Maximum	Average	SD	N
Predicted value	35.818	80.522	65.344	9.108	154
Residue	-12.549	15.764	0.000	5.561	154
Predicted value standard	-3.242	1.666	0.000	1.000	154
Standard residue	-2.234	2.807	0.000	0.990	154

different types of data from different water resources, the equation it generates is specific to the object of study of these data, that is, Eq. 3 generated here should not be used for another water resource other than Mirim Lagoon.

$$Y = 60,2 - 6,2 * Xa + 1,9 * Xb - 0,008 * Xc \quad (3)$$

where Y is the value of the new WQI, Xa is the concentration of phosphorus, Xb is the concentration of dissolved oxygen (DO) and Xc is the concentration of total coliforms.

As can be seen in Table 9, the residuals and the predicted standard values are within a range of approximately three standard deviations, corroborating the adequacy

Fig. 2 Distribution of standardized residuals of regression

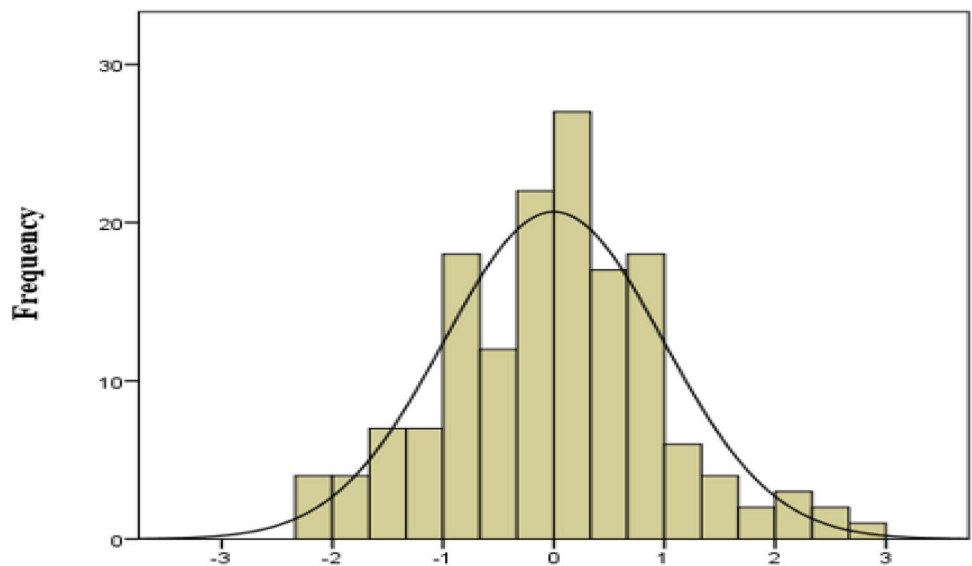
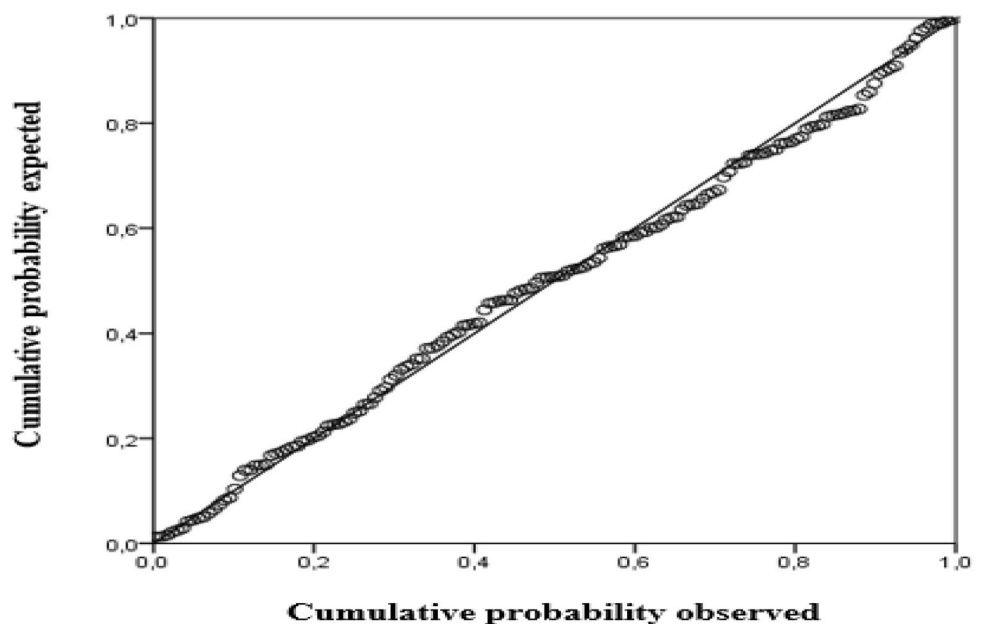


Fig. 3 Graphic distribution of the expected probability versus the observed probability



of the generated equation. The distribution of standardized residuals of regression the graph of the expected and observed cumulative probabilities are presented in Figs. 2 and 3, which demonstrate the adjustment of the regression.

3.5 New WQI

Using Eq. 1, the new WQI results for Mirim Lagoon were calculated. These results can be seen in Table 10. It is noteworthy that the original WQI uses nine parameters, while the new WQI, only 3.

Using the quality ranges described in Table 3 to classify the results of the new WQI and comparing these ranges with those relating to the original WQI, only in 19.5% of the results did the quality range change between the original WQI and the new WQI. In other

words, in more than 80% of the 154 samples, there was no change in the quality range for the Mirim Lagoon WQI.

3.6 T test

Finally, having the results of the original WQI and the new WQI, it is assessed whether there was significant variability between the two, for their numerical results and their quality ranges. This analysis is done by the Student's t test for paired samples and its results can be seen in Table 11. The paired samples t test was used because, unlike the normal t test, it considers that the samples are dependent.

The T test for paired samples considers a null hypothesis in which there is no significant variability between the components of the evaluated pair. Therefore, for p value < 0.05 , the null hypothesis is rejected and the

Table 10 New WQI from Mirim Lagoon

Collect	WQI NEW per monitoring point						
	LM1	LM2	LM3	LM4	LM5	LM6	LM7
1	60	66	67	68	57	64	67
2	73	72	68	76	66	74	71
3	72	67	71	63	59	74	71
4	59	70	71	72	71	77	77
5	67	73	72	75	75	66	77
6	69	59	69	72	70	70	71
7	77	76	72	75	70	75	77
8	75	73	70	61	63	60	65
9	76	74	80	75	62	74	76
10	69	65	70	74	72	73	74
11	69	73	73	39	36	66	72
12	68	56	65	67	60	44	58
13	56	48	39	39	43	45	48
14	64	61	67	72	71	59	71
15	60	61	59	53	61	73	76
16	73	56	56	62	71	57	72
17	70	59	66	68	50	50	71
18	63	62	58	52	50	47	56
19	66	71	66	64	63	54	68
20	58	60	58	59	54	49	65
21	67	72	64	67	70	64	75
22	69	65	73	69	71	65	70

Table 11 T test for paired samples

	Average	SD	Average SE	Degree of freedom	Significance (p value)
Original WQI—New WQI	-0.064	5.561	0.448	153	0.886
Original WQI band—New WQI band ^a	0.026	0.441	0.035	153	0.467

^aThe WQI band is presented in Table 3

significance of this variation is confirmed. For the data in this study, the *t* test obtained a *p* value > 0.05, retaining *H*₀ and then confirming that there is no significant variation between the original WQI and the new WQI for their absolute values and for their quality ranges.

The fact that there was no significant difference between the original WQI and the new WQI was precisely what was intended by this study, since it aimed at building a new WQI using fewer variables without significant loss of information. Reducing the number of variables analyzed for the WQI can generate a number of advantages, such as, for example, financial savings by river basin managers. Another example of the benefit caused by the reduction in the number of independent variables in the calculation of the WQI is the reduction in the eclipse effect, characterized, as already mentioned, by the attenuation of the impact of the variation of one of the parameters due to the stability of the others.

Ewaid et al. [13], in his study about the Tigre River in Baghdad, using this multiple linear regression (MLR) methodology, also obtained results from a new WQI that can be compared to the original WQI, demonstrating that the new equation modeled by them can be used to predict and monitor the quality of the waters of this river. Other studies, such as those by Kangabam et al. [24], Eregno [11], Chenini and Khemiri [7] and Saleem et al. [33] also used a similar methodology for assessing surface and underground water quality, attesting to the importance of the method and corroborating its viability.

4 Conclusion

The new WQI equation, generated through the MLR, in this study uses only three variables, phosphorus, DO and thermotolerant coliforms, facilitating the monitoring of this water resource, as it makes monitoring less financially costly. Still, it is concluded that the use of statistical methods is a very useful tool in the monitoring of water quality. After all, it was through these methods that it was possible to model a WQI equation with fewer variables for this water resource.

Thus, the new WQI equation modeled by this study can be used to predict and monitor the quality of the waters of Mirim Lagoon with considerable precision. It should also be noted that this method can be widely applied to other water resources as long as they have prior historical monitoring.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Al-janabi ZZ, Al-kubaisi AR, Al-obaidy AHMJ (2012) Assessment of water quality of Tigris River by using water quality index (CCME WQI). *Al-Nahrain J Sci* 15(1):119–126
2. Alves ICC, El-Robrini M, Santos MLS, Monteiro SM, Barbosa LPF, Guimarães JTF (2012) Qualidade de águas superficiais e avaliação do estado trófico do Rio Arari (Ilha de Marajó, norte do Brasil). *Acta Amaz* 42(1):115–124
3. ANA - Agência Nacional de Águas (2017) Indicadores de qualidade - Índice de qualidade das águas (IQA). <https://www.ana.gov.br/panorama-das-aguas/qualidade-da-agua/indicadores-de-qualidade>. Accessed 18 May 2020
4. APHA (2017) Standard methods for the examination of water and wastewater, 23rd edn. American Public Health Association (APHA), the American Water Works Association (AWWA), and the Water Environment Federation (WEF), Washington, DC, EUA
5. Bilgin A (2015) An assessment of water quality in the Coruh Basin (Turkey) using multivariate statistical techniques. *Environ Monit Assess* 187:721
6. Centeno LN, Ceconello ST (2016) Modificação de um índice de qualidade da água. *Rev Cient Rural* 18(1):65–82
7. Chenini I, Khemiri S (2009) Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int J Environ Sci Technol* 6(3):509–519
8. Costi J, Marques WC, Kirinus EP, Duarte RF, Arigony-Neto J (2018) Water level variability of the Mirim - São Gonçalo system, a large, subtropical, semi-enclosed coastal complex. *Adv Water Resour* 117:75–86
9. Dantas MS, Oliveira JC, Pinto CC, Oliveira SC (2020) Impact of fecal contamination on surface water quality in the São Francisco River hydrographic basin in Minas Gerais, Brazil. *J Water Health* 18(1):48–59
10. Drose A, Valentini MHK, Duarte V, Santos G, Nadaleti WC, Vieira BM (2020) Utilização de Métodos Estatísticos Multivariados no Monitoramento da Lagoa Mirim. *Meio Ambiente (Brasil)* 2(4):58–67
11. Eregno FE (2014) Multiple linear regression models for estimating microbial load in a drinking water source case from the Glomma river, Norway. MS thesis. Norwegian University of Life Sciences

12. Embrapa. Sustentabilidade Socioambiental da Bacia da Mirim Lagoon. Embrapa Clima Temperado Pelotas, RS. 2010. <https://www.infoteca.cnptia.embrapa.br>. Accessed 18 May 2020
13. Ewaid SH, Abed SA, Kadhum SA (2018) Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. *Environ Technol Innov* 11:390–398
14. Ewaid SH, Salwan AA (2017) Water quality index for AlGharraf River, southern Iraq. *Egypt J Aquat Res* 43(2):117–122
15. Ferreira KCD, Lopes FB, Andrade EM, Miireles ACM, Silva GS (2015) Adaptação do índice de qualidade da água da National Sanitation foundation ao semiárido brasileiro. *Rev Ciênc Agron* 46(2):277–286
16. Fia R et al (2009) Estado trófico da água na bacia hidrográfica da Mirim Lagoon, RS, Brasil. *Ambiente Água Interdiscip J Appl Sci* 4(1):132–141
17. Guimarães PRB (2017) Análise de Correlação e medidas de associação. Universidade Federal do Paraná. <https://docs.ufpr.br/~jomarc/correlacao.pdf>. Accessed 20 May 2020
18. Hair JF et al (2009) Análise multivariada de dados. Bookman Editora, Porto Alegre
19. Helena B, Pardo R, Vega M, Barrado E, Fernández JM, Fernández L (2000) Temporal evolution of groundwater composition in the alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Res* 34:807–816
20. Hirata FE, Möller OO, Mata MM (2010) Regime shifts, trends and interannual variations of water level in Mirim Lagoon, southern Brazil. *Pan Am J Aquat Sci* 5:254–266
21. Huang F, Wang XQ, Lou LP, Zhou ZQ, Wu JP (2010) Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques. *Water Res* 44:1562–1572
22. IRGA - Instituto Rio Grandense do Arroz (2006) Arroz irrigado – safra 2005/2006 – produção municipal. <http://www.irga.rs.gov.br/arquivos/20070117104152.pdf>. Accessed 20 May 2020
23. Iscen CF, Emiroglu O, Ilhan S, Arslan N, Yilmaz V, Ahiska S (2008) Application of multivariate statistical techniques in the assessment of surface water quality in Ulubat Lake, Turkey. *Environ Monit Assess* 144(1–3):269–276
24. Kangabam R, Sarojini DB, Suganthi K, Munisamy G (2017) Development of a water quality index (WQI) for the Loktak Lake in India. *Appl Water Sci* 7(6):2907–2918
25. Munar AM, Cavalcanti JR, Bravo JM, Fan M, Motta-Marques D, Fragoso CR Jr (2018) Coupling large-scale hydrological and hydrodynamic modeling: toward a better comprehension of watershed- shallow lake processes. *J Hydrol* 564:424–441
26. Naghettini M, Andrade Pinto EJ (2007) Hidrologia estatística. Serviço Geológico do Brasil, Belo Horizonte
27. Oliveira HA, Fernandes EH, Möller OO Jr, Collares GL (2015) Processos Hidrológicos e Hidrodinâmicos da Mirim Lagoon. *RBRH-Revista Brasileira de Recursos Hídricos* 20(1):34–45
28. Oliveira LN, Silva CE (2014) Qualidade da água do rio poti e suas implicações para atividade de lazer em teresina-pi. *Revista Equador* 3(1):128–147
29. Oliveira MLVM (2017) Gestão de águas, territórios e desenvolvimento econômico. *ACTA Geográfica, Boa Vista* 11(27):42–61
30. Piedras SRN, Santos JD, Fernandes JM, Tavares RA, Souza DM, Pouey JLOF (2012) Caracterização da atividade pesqueira na Mirim Lagoon, Rio Grande do Sul - Brasil. *Revista Brasileira de Agrociência* 18(2–4):107–116
31. Pimenta RHO, Reis SP, Fonseca M (2016) Diagnóstico ambiental em três trechos distintos do córrego Capão Regional Venda Nova, município de Belo Horizonte, MG. *Revista PETRA* 2(1):153–173
32. Rocha CHB, Pereira AM (2016) Análise multivariada para seleção de parâmetros de monitoramento em manancial de Juiz de Fora, Minas Gerais. *Revista Ambiente & Água* 11(1):176–187
33. Saleem A, Dandigi MN, Kumar KV (2012) Correlation-regression model for physico-chemical quality of groundwater in the South Indian city of Gulbarga. *Afr J Environ Sci Technol* 6(9):353–364
34. Santos GB et al (2020) Avaliação dos parâmetros e do índice de qualidade de água para o Arroio Moreira/Fragata, Pelotas/RS. *Revista Ibero-Americana de Ciências Ambientais* 11(4):287–299
35. Santos GB et al (2020) Análise da qualidade das águas do Arroio Moreira/Fragata (RS) através de métodos estatísticos. *Revista Ibero Americana de Ciências Ambientais* 11(4):217–226
36. SEMA – Secretaria do Meio Ambiente do Estado do Rio Grande do Sul (2006) Porto Alegre: Secretaria do Meio Ambiente do Estado do Rio Grande do SUL. <http://www.sema.rs.gov.br/sema/jsp/rhcommrim.jsp>. Accessed 21 May 2020
37. Silva DV, Oleinik PH, Costi J, de Paula KE, Marques WC (2019) Residence time patterns of Mirim Lagoon (Brazil) derived from two-dimensional hydrodynamic simulations. *Environ Earth Sci* 78(5):163
38. Silva GS, Jardim WF (2006) Um novo índice de qualidade das águas para proteção da vida aquática aplicado ao Rio Atibaia, região de Campinas/Paulínia-SP. *Química Nova, São Paulo* 29(4):689–694
39. Silveira VR et al (2020) Análise da qualidade hídrica da Lagoa Mirim e do Canal São Gonçalo. *Evolução do conhecimento científico na engenharia ambiental e sanitária*, Editora Atena
40. Siqueira GW, Aprile F, Miguéis AM (2012) Diagnóstico da qualidade da água do rio Parauapebas (Pará-Brasil). *Acta Amaz* 42(3):413–422
41. Souza JR, Moraes MEB, Sonoda SL, Santos HCRG (2014) A Importância da Qualidade da Água e os seus Múltiplos Usos: Caso do Rio Almada, Sul da Bahia, Brasil. *Revista Eletrônica do Prodema* 8(1):26–45
42. Steinke VA, Saito CH (2008) Exportação de carga poluidora para identificação de áreas úmidas sob risco ambiental na bacia hidrográfica da Mirim Lagoon. *Soc Nat* 20(2):43–67
43. Sun W, Xia C, Xu M, Guo J, Sun G (2016) Application of modified water quality indices as indicators to assess the spatial and temporal trends of water quality in the Dongjiang River. *Ecol Ind* 66:306–312
44. Tormam MF, Bork CK, Guedes HAS, Manzke J, Ferrão AL (2017) Variabilidade sazonal da qualidade da água na Lagoa Mirim, RS, Brasil. *Revista Brasileira de Engenharia e Sustentabilidade* 4(2):54–59
45. Valentini MHK et al (2020) Monitoramento e identificação de grupos de poluentes da Lagoa Mirim. *Revista Ibero-Americana de Ciências Ambientais* 11:4
46. Vieira BM, Nadaleti WC, Valentini MHK, Santos G, Viana FV, Corrêa MG (2019) Avaliação e comparação da qualidade das águas entre o canal São Gonçalo e a Lagoa Mirim. *Revista Ibero Americana de Ciências Ambientais* 10(2):185–196
47. Wheelan C (2016) Estatística: o que é, para que serve, como funciona. Editora Schwarcz-Companhia das Letras, São Paulo
48. Zhao J, Fu G, Lei K, Li Y (2011) Multivariate analysis of surface water quality in the Three Gorges area of China and implications for water management. *J Environ Sci* 23(9):1460–1471
49. Zhao Y, Xia XH, Yang ZF, Wang F (2012) Assessment of water quality in Baiyangdian Lake using multivariate statistical techniques. *Procedia Environ Sci* 13:1213–1226

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.