Environment and Ecology Research 10(5): 561-571, 2022 http://www.hrpub.org
DOI: 10.13189/eer.2022.100504

Trend Analysis on Water Quality Index Using the Least Squares Regression Models

Iskandar Shah Mohd Zawawi[1,*], Mohd Ridza Mohd Haniffah[2], Hazleen Aris[3]

[1]Faculty of Computer & Mathematical Sciences, Kompleks Al-Khawarizmi, Universiti Teknologi MARA 40450 Shah Alam, Selangor, Malaysia
[2]Department of Water and Environmental Engineering, School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[3]Institute of Informatics and Computing in Energy, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia

Received April 14, 2022; Revised August 11, 2022; Accepted September 19, 2022

Cite This Paper in the Following Citation Styles

(a): *[1] Iskandar Shah Mohd Zawawi, Mohd Ridza Mohd Haniffah, Hazleen Aris , "Trend Analysis on Water Quality Index Using the Least Squares Regression Models," Environment and Ecology Research, Vol. 10, No. 5, pp. 561 - 571, 2022. DOI: 10.13189/eer.2022.100504.*

(b): *Iskandar Shah Mohd Zawawi, Mohd Ridza Mohd Haniffah, Hazleen Aris (2022). Trend Analysis on Water Quality Index Using the Least Squares Regression Models. Environment and Ecology Research, 10(5), 561 - 571. DOI: 10.13189/eer.2022.100504.*

Abstract River water pollution requires continuous water quality monitoring that promotes the improvement of water resources. Therefore, the trend analysis on water quality data using mathematical model is an important task to determine whether the measured data increase or decrease during the time period. This paper is intended to highlight the applicability of the least squares regression models to fit the WQI data of the Skudai River, Tebrau River and Segget River located in Johor, Malaysia. As per the 12 years of trend analysis, the data of WQI are collected from the Environmental Quality Reports 2009-2020. The least squares method is utilized to estimate the unknown constants of the linear, quadratic, cubic, polynomial of degree four and degree five regression models. The advantage of using proposed models is that it can be implemented easily even on relatively low computational power systems. The results show that the higher degree polynomial model fits the data reasonably well, in which the polynomials of degree 4 and 5 have lowest average error. Assessment of actual and predictable values of WQI shows that the trends in WQI for all study areas are downward year after year.

Keywords Water Quality Index, Trend Analysis, Regression Models, Least Squares Method

1. Introduction

With rapid economic and demographic development in Johor, Malaysia have put pressure on the quality of its river water and consequently on its water supplies. There are fears that the shortage of clean water supply will affect the economy and livelihood. The main problem caused by blue water decreased quality is that it induces physical water scarcity [1] in terms of use [2]. The climate change such as rainfall variability will also significantly affect the determinants of pollutants in river water [3]. As degradation of river water quality becomes the serious issue in most of the countries all over the world, the water quality management and monitoring are necessary in achieving the healthy life and ecosystem sustainability. The Malaysian Government in its Vision for Water 2025 states that rivers should achieve at least Class II (treatment for drinking purposes and safe for swimming) as measured by Malaysia's Water Quality Index. Water Quality Index (WQI) is a classification of water quality status that has been used by the Malaysia Department of Environment to summarize the large amount of water quality data into a single index. The WQI serves as the basis for the assessment of environment water quality [4], in which a process or technique [7] may rely on the particular water use for standardization such as for drinking water [8] or recreational use [9].

Many researchers have used the WQI to monitor and assess the water quality of the river system [10-12]. In Malaysia, the Department of Environment (DOE) uses WQI and National Water Quality Standards for Malaysia (NWQS) for about 30 years to evaluate the status of the river water quality. The NWQS defined six classes (I, IIA, IIB, III, IV, and V) for river water classification based on the descending order of water quality. Class I being the best and Class V being the worst water quality [11]. The WQI that has been introduced by DOE is known as WQI-DOE, an opinion-poll where the experts are consulted on the choice of pollutants and weightage to each pollutant. It is used to simplify the extensive amount of data collected coherent to the pollutants listed in the NWQS. Over 120 physico-chemical and biological parameters were reviewed, and six parameters were decided to govern the WQI-DOE. The study considered local conditions due to differences in environmental characteristics and climatology when compared to other WQI. For example, in Malaysia, oxygen solubility is limited by the equatorial climate in which cool climate countries tend to have higher oxygen solubility. Also, the study orientation was on the beneficial uses of water which focused on water for domestic water supply, fisheries and aquatic propagation and livestock drinking [13].

Many analyses that explore the water quality trends are likely focusing on the effect of human activities and particular pollutant concentration on WQI or may be restricted to a particular river. For instance, the spatiotemporal trends in water quality conditions in 92 major rivers in Japan were evaluated by [14] using the Mann Kendall test based on the available water quality data recorded from 1992 to 2005. The QUAL2K model has been employed to simulate the effect of BOD and Ammoniacal Nitrogen (AN) from industrial areas to Tebrau River [15]. The first order interaction multiple regression models were used in forecasting the WQI in Manjung River and its tributaries [16]. Recent study developed an InfoWorks simulation to simulate the load reduction for COD at Malacca River, Malaysia [17].

In the more recent study, the multiple regression model with an interaction term was used to assess the stream water quality based on the combined effect of landscape composition and buffer strip width [18]. A typical multiple regression model involves one dependent variable and more than one independent variables, which includes a technique called polynomial regression [19,20]. Since it is the interaction between independent variables, the effect of one independent variable is depending on the value of other independent variables. Although it is important to monitor all the pollutants concentration or water quality parameters with adequate manpower and laboratory facilities, the analysis of WQI using trend gives useful insights into the potential changes of water quality at different rivers over year. Therefore, the aim of this study is to analyze the trend of WQI and evaluate the status of river water quality in Johor, Malaysia. Several types of mathematical models such as linear, quadratic, cubic and higher degree of polynomial regression models are derived using least squares method, while absolute error is determined to ascertain models that fit the data.

2. Material and Data Preparation

2.1. Water Quality Index Determination

In Malaysia, six primary pollutants or parameters used for the evaluation of WQI are biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), pH, suspended solids (SS) and AN. The DO is one of the main parameters in WQI because it shows the amount of free oxygen dissolved in water that is essential for biochemical activities. The rate of removal of oxygen from water due to biochemical activities i.e bacteria and other microorganism decomposing organic compounds under aerobic conditions at a specific temperature is defined as the BOD. The COD is equivalent to BOD but uses strong chemical oxidizing agents instead of bacteria and microorganism to oxidize the organic compounds. It is mostly used to measure organic pollutants in surface water and wastewater. The AN is the total amount of ammonia in which it has direct toxic effects on aquatic life and humans. It is often contributed by waste products such as sewages, landfill leachates, animal excretions and runoff from agricultural activities. The SS are small solid particles including sand, sediment and particles that suspend instead of dissolving in water. These SS must be removed in order to obtain clear and good water quality. The pH measures the acidity or basicity of water and is important to protect fish life and control the availability of nutrients, biological functions, microbial activities, and the behavior of chemicals in water. The SS, DO and BOD are clearly dominant in determining the quality of the water but any change in pH, COD and NA does not have any significant impact.

There are four common steps to develop the WQI formula, including the selection of parameter, establishment of weights (equally or unequally), obtaining the sub-index value and sub-index aggregation to produce the final index value, which can be expressed in a simple additive aggregation function [7,20]:

$$\mathrm{WQI} = \sum_{i=1}^{n} w_i s_i \qquad (1)$$

where s_i is the sub-index value for parameter i, w_i is the corresponding parameter weight value and n is the total number of parameters. The values of WQI can be classified into five categories, as shown in Table 1 [22].

The Malaysian WQI introduced by the DOE is obtained by the panel of experts, which was consulted on the correlation of weightage parameter (coefficients or constants for each parameter):

$$WQI = 0.22SIDO + 0.19SIBOD + 0.16SICOD + 0.16SISS + 0.15SIAN + 0.12SIpH, \quad (2)$$

where SIDO, SIBOD, SICOD, SIAN, SIpH and SISS are the sub-indices of six parameters. From (2), DO carries maximum weightage of 0.22 and pH carries the minimum of 0.12. The value of sub-index can be determined based on the best-fit equations given in Table 2 [12].

Water quality categories are highly affected by varying characteristics in the surrounding areas. The decrease in WQI may be due to an increase in the number of polluting sources such as sewage treatment plants, manufacturing industries and palm oil mills which contributed to a high pollution loading. The standard classification of WQI set by DOE involves five classes of water quality (Class I – V) and individual pollutant [23].

The DOE-WQI scale classifies the level of water pollution as clean, slightly polluted and polluted if the DOE-WQI falls within the range of 81 to 100, 60 to 80 and 0 to 59 respectively. This study adopted the DOE-WQI tool to evaluate the water quality of the Johor rivers. In addition, the beneficial use of the water was also compared with the DOE water quality classification based on index range such as clean, slightly polluted, and polluted [24].

2.2. Study Area

Malaysia is a tropical country with temperature ranging from 23 to 30 °C and average annual rainfall of about 2800 mm. As shown in Figure 1, all the three river basins, Skudai River, Tebrau River and Segget River in Johor, Malaysia have the same features to above.

The land use of Skudai and Tebrau Rivers is urban at the downstream part and some agriculture for the upstream part. As for Segget River, it is a small catchment and almost fully developed. Skudai River is one of the major rivers with the length and the catchment area being 40 km and 325 km^2 respectively [11]. The river traverses through the boundaries of three local authorities from the furthest northern parts of Johor into the Johor Straits at the south. This river has been a source of income for the fishermen plying the river, an alternative mode of transportation and a source of raw water for the water treatment plant supplying large parts of the Johor Bahru district and Singapore [25]. Other major rivers in Johor include Tebrau River, which is located at the south of Johor with approximately 35 km long. The river flows from Senai Industrial Area and drains into the Straits of Johor. It stretches over 35 km and covers 225 km^2 of catchment area. Tebrau river basin consists of five major tributaries including the Bala River, Pandan River, Sembulung River, Tampoi River and Plentong River [15]. Segget River has a catchment area of 3.6 km^2 and the main river length is approximately 3.6 km. The river system flows through Kampung Wadi Hana, Kebun Teh and Taman Century areas and subsequently passes through Johor Bahru City Centre. It discharges into the Straits of Johor, just 200 m west of the causeway and is controlled by tidal gates. The whole catchment is located within the highly built-up area and the land use is fully urban [25].

Table 1. Standard water quality classification based on WQI value

WQI range	Ratings of water
91-100	Excellent
71-90	Good
51-70	Medium
26-50	Bad
0-25	Very bad

Table 2. Sub-index calculation [12]

Sub-index parameter	Value	Conditions
SIDO	0	$DO \leq 8$
	100	$DO \geq 92$
	-0.395 + 0.030DO2 - 0.00020DO3	$8 < DO < 92$
SIBOD	100.4 - 4.23BOD	$BOD \leq 5$
	$108e^{0.055BOD}$ - 0.1BOD	$BOD > 5$
SIAN	100.5 - 105AN	$AN \leq 0.3$
	$94e^{0.573AN}$ - 5 |AN - 2|	$0.3 < AN < 4$
	0	$AN \geq 4$
SICOD	-1.33COD + 99.1	$COD \leq 20$
	$103e^{0.0157COD}$ - 0.04COD	$COD > 20$
SIpH	$17.2 - 17.2pH + 5.02pH^2$	$pH < 5.5$
	$-242 + 95.5pH - 6.67pH^2$	$5.5 \leq pH < 7$
	$-181 + 82.4pH - 6.05pH^2$	$7 \leq pH < 8.75$
	$536 - 77.0pH + 2.76pH^2$	$pH \geq 8.75$
SISS	$97.5e^{0.00676SS}$ + 0.05SS	$SS \leq 100$
	$71e^{0.0016SS}$ - 0.015SS	$100 < SS < 1000$
	0	$SS \geq 1000$

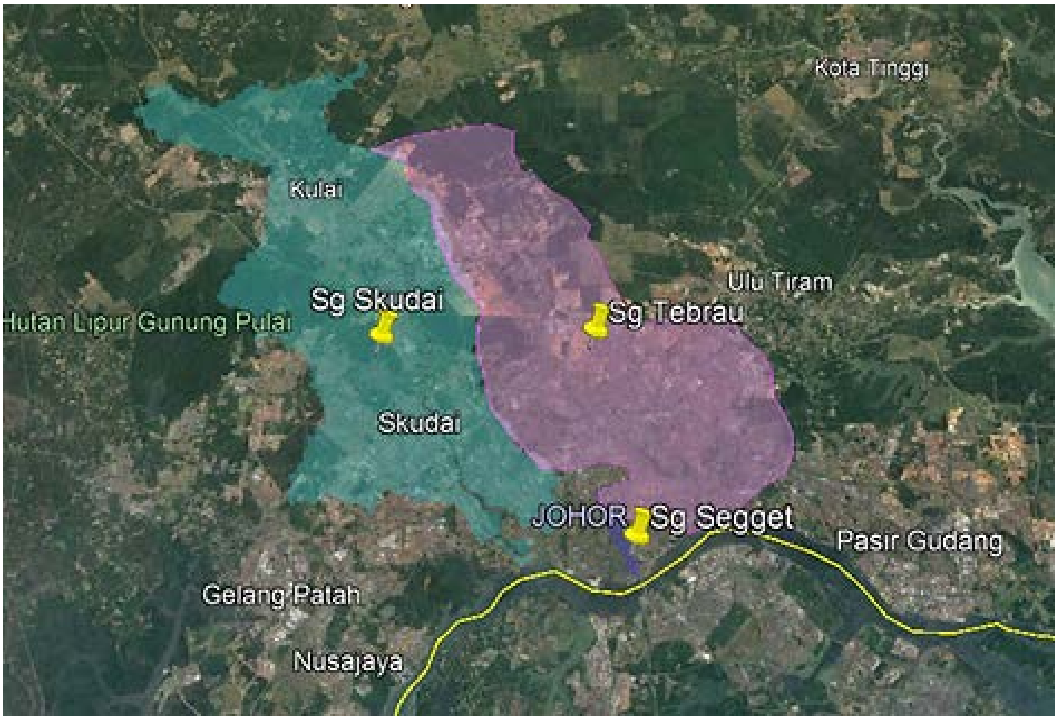

Figure 1. Location of Skudai River, Tebrau River and Segget River catchment in Johor, Malaysia

2.3. Data Collection

The WQI data of Skudai River, Tebrau River and Segget River are collected from the Environmental Quality Report (EQR) 2009-2020 [26-32] as summarized in Table 3.

Table 3. The WQI data of Skudai River, Tebrau River and Segget River in Johor, Malaysia from 2009 to 2020

Year	x-year	Skudai River		Tebrau River		Segget River	
		WQI	Class	WQI	Class	WQI	Class
2009	1	64	III	69	III	50	IV
2010	2	68	III	66	III	51	IV
2011	3	68	III	56	III	53	III
2012	4	69	III	66	III	58	III
2013	5	71	III	64	III	52	III
2014	6	66	III	59	III	48	IV
2015	7	65	III	70	III	49	IV
2016	8	60	III	35	IV	47	IV
2017	9	59	III	41	IV	46	IV
2018	10	63	III	51	IV	59	III
2019	11	57	III	56	III	52	III
2020	12	66	III	58	III	65	III

3. Methodology

The mathematical modelling of least squares regression model begins with the review of least squares method.

3.1. Least Squares Method

The least squares method is one of the most common approaches in regression analysis to approximate the solution of the system [33]. The term "least squares" refers to the overall solution that minimizes the summation of the squares of the errors, which are brought by the results of every single equation. Such inconsistent systems, $A\mathbf{x} = \mathbf{y}$ often arise in applications and many physical problems with a large coefficient matrix, A. Thus, we need to find $\mathbf{x}$ such that $A\mathbf{x}$ is close as possible to $\mathbf{y}$, $\|\mathbf{y} - A\mathbf{x}\|$. For any linear system $A\mathbf{x} = \mathbf{y}$, the associated normal system, $A^T A\mathbf{x} = A^T\mathbf{y}$ is consistent and all solutions of this normal systems are least squares solutions of the system $A\mathbf{x} = \mathbf{y}$. If A is an m x n matrix with linearly independent column vectors, then for every m x 1 matrix $\mathbf{y}$, the linear system $A\mathbf{x} = \mathbf{y}$ has a unique least squares solution which is given as follows:

$$\mathbf{x} = \left(A^T A\right)^{-1} A^T \mathbf{y} \ . \tag{3}$$

3.2. Least Squares Regression Models

A common problem in engineering practice and experimental work is to characterize a relationship between two or more variables mathematically. Therefore, an analysis needs to be carried out to observe the behavior or the trend of the dependent variable, which depends on the changes in the independent variable. The procedure of finding a function to represent this relationship is known as curve fitting or regression. Simple regression is the most widely used method for investigating the linear relationship between two continuous variables, x and y. Regression works by creating the trend line that best fits all the available data points that can be used to examine how the response variable changes as the value of x changes and predicts the value of y for any value of x [34].

Given some data $(x_1, y_1), (x_2, y_2)\ldots..(x_n, y_n)$, it consists of n paired observation of the predictor variables, $x_1, x_2, ..., x_n$ and response variables, $y_1, y_2, ..., y_n$. Suppose that we have the system of linear equations in the matrix form, $M\mathbf{v} = \mathbf{y}$, where $M = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} a \\ b \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$. The size of M is $n \times 2$, where the first column is always 1 and the second column contains the actual observations of x . We have this apparently redundant first column because when we multiply $\mathbf{v}$ by M , there are more equations than unknowns among the set of equations:

$$M\mathbf{v} = \begin{pmatrix} a + bx_1 \\ a + bx_2 \\ \vdots \\ a + bx_n \end{pmatrix}. \tag{4}$$

Since we expect a linear relationship, all these points should lie on a single straight line. The slope of this line will be a , and the intercept is b . The equation of the linear least squares regression model can be expressed as the least squares straight line fits, $\mathbf{y} = a + b\mathbf{x}$ where $\mathbf{v} = \begin{pmatrix} a \\ b \end{pmatrix}$ to the data points. The values of constants a and b must be determined so that they minimize the error. From (3), the least squares equation of the system (4) is obtained as follows:

$$\mathbf{v} = \left(M^T M\right)^{-1} M^T \mathbf{y}, \tag{5}$$

which expresses the fact that $\mathbf{v}$ is the unique least squares solution of the normal equations, $M^T M\,\mathbf{v} = M^T \mathbf{y}$. The matrix M^T is the matrix transpose of M while the product of $M^T M$ is an invertible matrix. However, we discover that the data points do not always lie on the line. In some cases, it is essential to deal with flexible curve fitting techniques whose relationship between input and output variables is non-linear and complex. Another model to consider instead of using linear regression model is polynomial regression [35]. Table 4 shows the general form of linear, quadratic, cubic, degree four and degree five least squares polynomial regression models.

Table 4. The general form of least squares regression models

Model	Least squares regression equation
Linear	$\mathbf{y} = a + b\mathbf{x}$
Quadratic	$\mathbf{y} = a + b\mathbf{x} + c\mathbf{x}^2$
Cubic	$\mathbf{y} = a + b\mathbf{x} + c\mathbf{x}^2 + d\mathbf{x}^3$
Polynomial of degree four	$\mathbf{y} = a + b\mathbf{x} + c\mathbf{x}^2 + d\mathbf{x}^3 + e\mathbf{x}^4$
Polynomial of degree five:	$\mathbf{y} = a + b\mathbf{x} + c\mathbf{x}^2 + d\mathbf{x}^3 + e\mathbf{x}^4 + f\mathbf{x}^5$

To solve (5), we consider M and $\mathbf{v}$ for each model as follows:

$$\text{Quadratic: } M = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

$$\text{Cubic: } M = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

Polynomial of degree four:

$$M = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^3 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix}.$$

Polynomial of degree five:

$$M = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^3 & x_n^5 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \end{pmatrix}.$$

where a, b, c, d, e and f are estimated values by least squares method for parameter $\mathbf{v}$ while variables, $x_1, \cdots, x_n$ consist of successive power terms. The aim is to find the constants of the linear and polynomial regression models that best fit the data. However, it is difficult to compute the least squares solutions analytically due to the large size of matrices. Thus all the computations of (5) are generated using Maple software in order to obtain their least squares solutions and trend line equations.

Table 5. The trend line equations of least square regression models for each river

Skudai River
Linear: $\mathbf{y}=\frac{2281}{33}-\frac{98}{143}\mathbf{x}$
Quadratic: $\mathbf{y}=\frac{1495}{22}-\frac{53}{286}\mathbf{x}-\frac{1}{26}\mathbf{x}^2$
Qubic: $\mathbf{y}=\frac{505}{9}+\frac{34627}{3861}\mathbf{x}-\frac{2227}{1287}\mathbf{x}^2+\frac{335}{3861}\mathbf{x}^3$
Polynomial of degree four: $\mathbf{y}=\frac{5933}{99}+\frac{6497}{1404}\mathbf{x}-\frac{287}{792}\mathbf{x}^2-\frac{1117}{15444}\mathbf{x}^3+\frac{7}{1144}\mathbf{x}^4$
Polynomial of degree five: $\mathbf{y}=\frac{658}{11}+\frac{127099}{26520}\mathbf{x}-\frac{3931}{8976}\mathbf{x}^2+\frac{1687}{29172}\mathbf{x}^3+\frac{571}{116688}\mathbf{x}^4+\frac{1}{26520}\mathbf{x}^5$
Tebrau River
Linear: $\mathbf{y}=\frac{4471}{66}-\frac{447}{286}\mathbf{x}$
Quadratic: $\mathbf{y}=\frac{3293}{44}-\frac{18439}{4004}\mathbf{x}-\frac{937}{4004}\mathbf{x}^2$
Qubic: $\mathbf{y}=\frac{535}{9}+\frac{394307}{54054}\mathbf{x}-\frac{17707}{9009}\mathbf{x}^2+\frac{67}{594}\mathbf{x}^3$
Polynomial of degree four: $\mathbf{y}=\frac{30515}{396}-\frac{1572859}{123552}\mathbf{x}+\frac{357911}{82368}\mathbf{x}^2-\frac{5903}{9504}\mathbf{x}^3+\frac{775}{27456}\mathbf{x}^4$
Polynomial of degree five: $\mathbf{y}=\frac{2593}{22}-\frac{14266181}{194480}\mathbf{x}+\frac{2497049}{77792}\mathbf{x}^2-\frac{106723}{17952}\mathbf{x}^3+\frac{37207}{77792}\mathbf{x}^4-\frac{113}{8160}\mathbf{x}^5$
Segget River
Linear: $\mathbf{y}=\frac{542}{11}+\frac{71}{143}\mathbf{x}$
Quadratic: $\mathbf{y}=\frac{1247}{22}-\frac{5363}{2002}\mathbf{x}+\frac{489}{2002}\mathbf{x}^2$
Qubic: $\mathbf{y}=\frac{4292}{99}+\frac{206014}{27027}\mathbf{x}-\frac{14953}{9009}\mathbf{x}^2+\frac{29}{297}\mathbf{x}^3$
Polynomial of degree four: $\mathbf{y}=\frac{7999}{198}+\frac{678397}{61776}\mathbf{x}-\frac{111953}{41184}\mathbf{x}^2+\frac{1049}{4752}\mathbf{x}^3-\frac{5}{1056}\mathbf{x}^4$
Polynomial of degree five: $\mathbf{y}=\frac{1033}{22}+\frac{724583}{583440}\mathbf{x}+\frac{406163}{233376}\mathbf{x}^2-\frac{24681}{38896}\mathbf{x}^3+\frac{1213}{17952}\mathbf{x}^4-\frac{59}{26520}\mathbf{x}^5$

The accuracy of each model can be evaluated by computing the distance between the data points and the regression line, which can be interpreted as residuals or error, $e = \|\mathbf{y} - M\mathbf{v}\|$. A large error indicates a lot of residual variation, which signifies a poor fit. If the error is small, there is a minimal amount of residual variation, thus the fit is good. According to the least squares method, the best fitting curve has the property that the error incurred is minimum.

4. Results

The process of using the pattern of the data in predicting the future based on current and previous situation is called trend analysis. It is started by considering the time series comprising discrete 12 datasets, $(x_1, y_1), (x_2, y_2), \ldots, (x_{12}, y_{12})$ where $x_1, x_2, \cdots x_{12}$ are the number of years ($n = 12$) and $y_1, y_2 \cdots y_{12}$ are the measured WQI data of Skudai River, Tebrau River and Segget River. Table 5 shows the trend line equations of linear, quadratic, cubic, polynomial of degree four and degree five regression models, which are obtained from the least squares method.

Then we apply the trend line equations of least squares regression models as presented in Table 5 to plot the graph of WQI over the period of 12 years for Skudai River, Tebrau River and Segget River. By using Maple software, the least squares lines fit to WQI data for each river are illustrated in Figures 2 – 4.

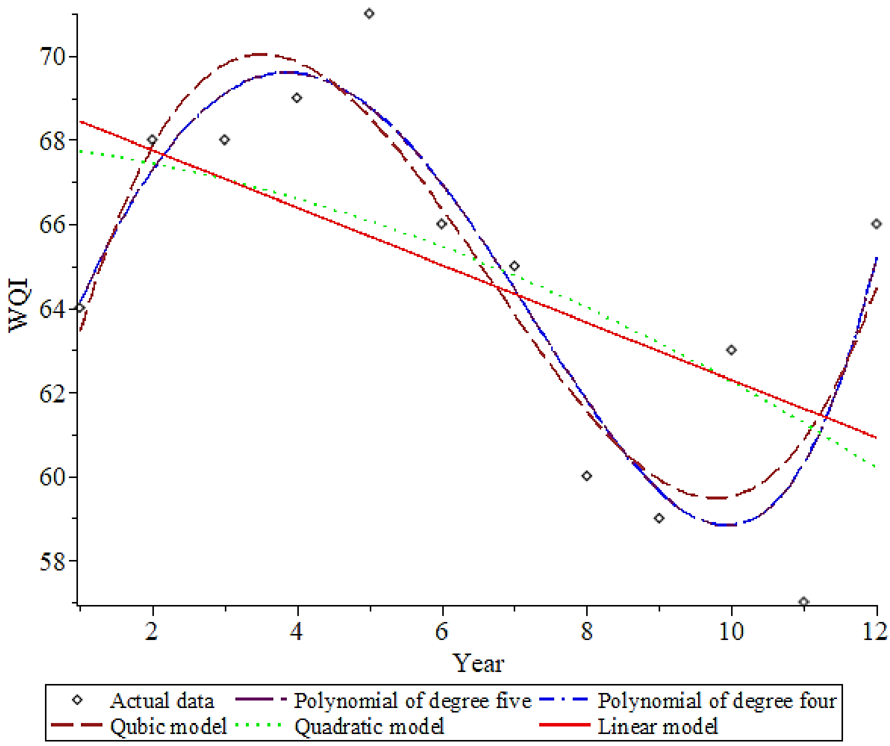

Figure 2. The trend in WQI of Skudai River

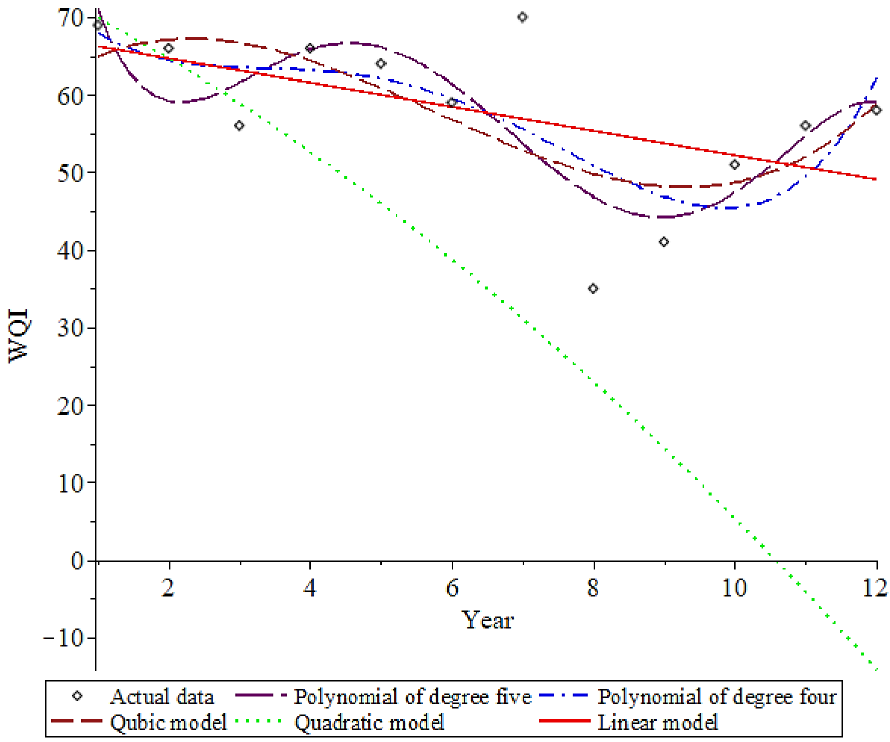

Figure 3. The trend in WQI of Tebrau River

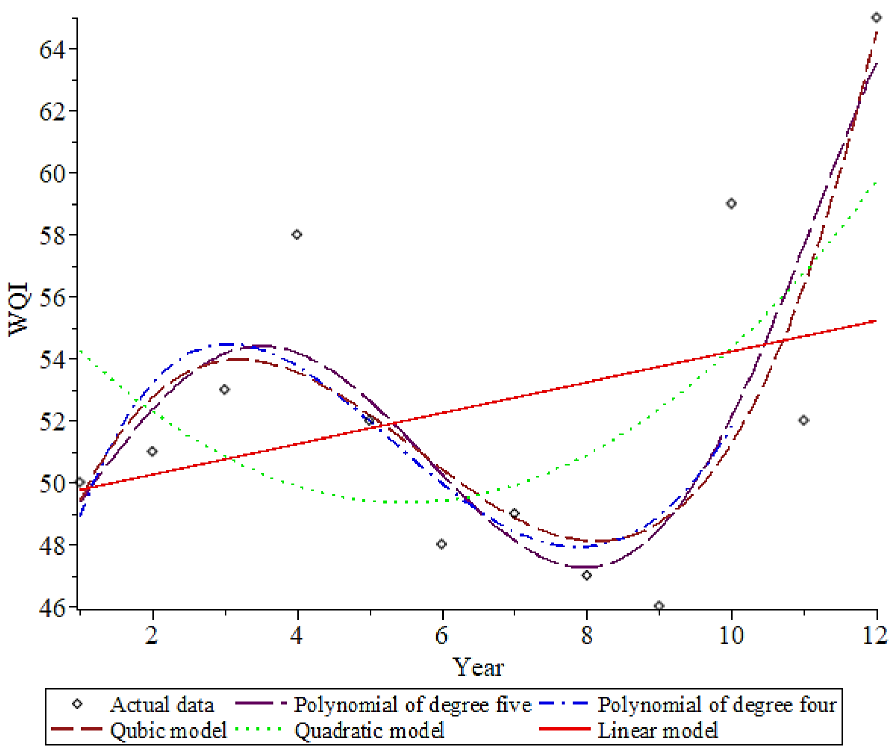

Figure 4. The trend in WQI of Segget River

Table 6. The average error for each model

Model	Magnitude of error		
	Skudai River	Tebrau River	Segget River
Linear	$\frac{2276}{429} \approx 5.31$	$\frac{17365}{858} \approx 20.24$	$\frac{127}{13} \approx 9.77$
Quadratic	$\frac{151}{26} \approx 5.81$	$\frac{71979}{4004} \approx 17.98$	$\frac{16267}{2002} \approx 8.13$
Cubic	$\frac{151}{39} \approx 3.87$	$\frac{22024}{1287} \approx 17.11$	$\frac{23293}{3003} \approx 7.76$
Polynomial of degree four	$\frac{1789}{429} \approx 4.17$	$\frac{40987}{2574} \approx 15.92$	$\frac{12391}{1716} \approx 7.22$
Polynomial of degree five	$\frac{20301}{4862} \approx 4.18$	$\frac{39502}{2431} \approx 16.25$	$\frac{67187}{9724} \approx 6.91$

5. Discussion

To discuss the trend of WQI in Skudai River, we concentrate on the polynomial of degree five regression line. Figure 2 shows an upward trend in WQI of Skudai River in the first five years. It shows that the water quality in the Skudai River has been getting better since 2009 until 2013 due to strict rules and regulations on pollution and water quality control for industries and commercial sectors taken by the Skudai River management team and the DOE. This finding is similar to the water quality trend of the Skudai River analyzed by [11] in 2016. The Skudai River Rejuvenation Plan (SRRP) with river management and monitoring tools (RMMT) have been initiated in 2015 with the goal to improve Skudai River into a sustainable river. One of the strategies is proposed in SRRP in order to classify the Skudai River as Class II including zoning area, for examples, zone 1 is Sedenak - Senai (water quality: class II, III) and zone 2 is Skudai - Teluk Danga (water quality: class III, IV, V). However, the effect was not the same from fifth year to ninth year as can be seen in Figure 2. As a matter of fact, it is observed from Table 2 that WQI classification of Skudai River still remained in Class III for 10 years consecutively.

From 2006 to 2010, under the Ninth Malaysia Plan (9MP), the Federal Government has allocated a total of RM89.60 million from the national budget for river cleaning and drainage programme to be undertaken in Tebrau River and Skudai River [13]. In 2021, the allocation RM4.5 million under Johor Budget 2022 was given by the state government for both rivers sustainability. However, the vice-president of Johor Malaysian Nature Society (MNS) stressed out that at least RM200 million was needed to rehabilitate and rejuvenate the rivers [36]. Therefore, the allocation under the 9MP to improve the water quality of Tebrau River and Skudai River was insufficient. As a result, the linear lines as shown in Figure 3-4 indicate the downward trend through the data of WQI in both rivers.

The Johor Bahru Transformation Programme commenced in 2010 aspired to remake Johor Bahru into a vibrant heritage and cultural city. Among the significant initiatives under this programme is the revitalisation of Sungai Segget, one of Malaysia's most polluted rivers. It can be seen in Figure 4 that the trend in WQI of Segget River is upward slowly in the first four years. However, the trend decreases more slowly in the next five years. The effects of the Segget River revitalisation could be felt in around 2017. With time passed, we observe that the trend in WQI of Segget River increases dramatically after 2017, but the water quality still remains in Class III [13]. Based on the fifth degree of polynomial regression models, the overall trend analysis shows that the WQI of all the three rivers have worsening trend in most of the years. The results coincide with the linear regression models of reducing quality of rivers. With respect to the DOE water quality classification, the Skudai River and Tebrau River are slightly polluted (Class III) while the status of water quality for Segget River is polluted (Class IV), which requires intensive water treatment. It is suggested that monitoring and water treatment should be conducted efficiently for better water quality of Johor rivers. It is advised to always plot a simple scatter diagram before using any polynomial regression model in order to identify the type of relationship between the variable of interest [20]. It can be said that a simple linear model is a sufficient summary of the nature trend although it may not be the best model to fit the data compared to non-linear models. This can be justified by the comparable average error obtained by the linear model and higher degree of polynomial as presented in Table 6. This finding is aligned with the statement made by [37], in which better WQI prediction could be obtained with non-linear models.

6. Conclusions

One of the greatest motivations for this paper is to provide right information of river water quality in a simple mathematical manner. The trend analysis in water quality index based on mathematical modelling makes it competitive for water quality assessment and river basin management. The least squares regression model has earned its place as the primary tool for analyzing the trend of measured data. In fact, the validity of the polynomial model has been justified in the literature. The trend and error analysis reveal that the least squares polynomial regression models of degree four and degree five fit the WQI data better than the linear, quadratic and cubic regression models. On surface, this study provides evidence that the least squares polynomial regression model with higher degree is suitable for analyzing and predicting the long term trend in WQI of any rivers in Malaysia. Although the higher degree of polynomial regression models gives better performance in terms of accuracy, it tends to overfit the model. Thus, it is crucial to choose an appropriate polynomial model to give better representation of the overall behavior of pollution in the river.

Most of the data in science and engineering are modelled and analyzed using linear regression models because either the data are inherently linear or nonlinear, any point of data can be fitted approximately. The estimation of the unknown constants using the least squares method is efficient to obtain the trend line equations. The theory associated with the least squares regression model is well-understood, easy-to-use, less expensive and allows clear trend prediction to DOE-WQI primarily. Although the least squares method is claimed to give an optimal estimation, it is not well fitted to the unusual data points. The presence of multiple outliers in the data can sometimes affect the accuracy of the models. In addition, there are unfortunately fewer model validation tools to detect outliers in the polynomial trend than there are for the linear trend. The significance of polynomial trend means that no specific models are available to predict exactly at what point in time the WQI of a river increases or decreases. This makes critical validation of accurate polynomial models. Practically, there is no so-called best model to fit data.

7. Recommendation

The hypothesis and conclusion of the study may not reflect the real situation with an exposure of pollution history of a large number of rivers in Johor, Malaysia. Thus, the impact of trend assumption would be rather limited. Furthermore, there is a lack of data which are needed for detailed analysis, in which the trend analysis can be further improved when more data such as six parameters concentration for each river become available in the EQR. After reviewing many references, we realize that six parameters such as DO, BOD, COD, SS, AN and pH are important for the trend analysis. The identification of the factors, with its data, which contribute to the six parameters concentration is also important because this is the real deal that leads to water quality of rivers. No such data are available in the EQR furnished by the DOE. As such, we have to depend on the WQI data to analyze the trend in water quality of the river. If the DOE is planning to revise the contents of EQR, it is highly recommended that these data would be included in the EQR and available to the public for comprehensive and better assessment. More importantly the least squares method is an effective mathematical modelling, which can be utilized easily by the responsible authorities and departments involved in river management. However, the authorities must be aware of the limitations of NWQS and WQI, so that the sustainable conservation of river water and environment efforts can be implemented continuously and effectively.

Acknowledgements

The authors would like to acknowledge the support received from UiTM students (Nazreen Awanis Nazri, Anis Syahirah Rasidi and Nurul Nabilah Mohd Fadzli) and the Department of Environment Malaysia for providing the water quality index data. The research work is self-funded under the Universiti Teknologi Malaysia-Collaborative Research Grant (CRG-UTM), registered in UiTM with RMC file number: 600-TNCPI/PBT 5/3 (016/2020)). The research is also supported by the research grant under the CRG-UTM with vote number PY/2019/02850 (cost no. 08G58), PY/2020/03455 (cost no. 4B551) and 20200105CRGJ.

REFERENCES

[1] K. Zisopoulou and D. Panagoulia, "An in-depth analysis of physical blue and green water scarcity in agriculture in terms of causes and events and perceived amenability to economic interpretation," *Water (Switzerland)*, vol. 13, no. 12, 2021, doi: 10.3390/w13121693.

[2] D. Wang, K. Hubacek, Y. Shan, W. Gerbens-Leenes, and J. Liu, "A review of water stress and water footprint accounting," *Water (Switzerland)*, vol. 13, no. 2, pp. 1–15, 2021, doi: 10.3390/w13020201.

[3] R. Damania, S. Desbureaux, A.-S. Rodella, J. Russ, and E. Zaveri, *Quality Unknown*. 2019.

[4] United Nations Statistics Division, "Guidelines for the Compilation of Water Accounts and Statistics (Draft)," no. June, p. 205, 2014.

[5] M. Calmuc *et al.*, "A comparative approach to a series of physico-chemical quality indices used in assessing water quality in the lower danube," *Water (Switzerland)*, vol. 12, no. 11, 2020, doi: 10.3390/w12113239.

[6] A. D. Sutadian, N. Muttil, A. G. Yilmaz, and B. J. C. Perera, "Development of river water quality indices—a review," *Environmental Monitoring and Assessment*, vol. 188, no. 1, pp. 1–29, 2016, doi: 10.1007/s10661-015-5050-0.

[7] N. Akhtar *et al.*, "Modification of the water quality index (WQI) process for simple calculation using the multi-criteria decision-making (MCDM) method: A review," *Water (Switzerland)*, vol. 13, no. 7, 2021, doi: 10.3390/w13070905.

[8] "Global Drinking Water Quality Index Development and Sensitivity Analysis Report," N.Y., U.S.A., 2007.

[9] EPA, "Recreational Water Quality Criteria," *U. S. Environmental Protection Agency*, pp. 1–69, 2012.

[10] A. Al-Ani, "Mathematical computation of water quality index for the assessment of Al-Hilla river ecosystem," *International Journal of Civil Engineering and Technology*, vol. 10, no. 1, pp. 1862–1869, 2019.

[11] Naubi, N. H. Zardari, S. M. Shirazi, N. Farahen, B. Ibrahim, and L. Baloo, "Effectiveness of Water Quality Index for Monitoring Malaysian River Water Quality," *Polish Journal of Environmental Studies*, vol. 25, no. 1, pp. 231–239, 2016, doi: 10.15244/pjoes/60109.

[12] F. Othman, M. E. Eldin, and I. Mohamed, "Trend analysis of a tropical urban river water quality in Malaysia," *Journal of Environmental Monitoring*, vol. 14, pp. 3164–3173, 2012, doi: 10.1039/c2em30676j.

[13] Z. Zaki, "Benchmarking River Water Quality in Malaysia," *JURUTERA*, no. February, pp. 12–15, Feb. 2010.

[14] P. Luo, B. He, K. Takara, B. H. N. Razafindrabe, D. Nover, and Y. Yamashiki, "Spatiotemporal trend analysis of recent river water quality conditions in Japan," *Journal of Environmental Monitoring*, vol. 13, no. 10, pp. 2819–2829, 2011, doi: 10.1039/c1em10339c.

[15] Z. Zainudin, N. A. Rahman, N. Abdullah, and N. F. Mazlan, "Development of Water Quality Model for Sungai Tebrau using QUAL2K," *Journal of Applied Sciences*, vol. 10, no. 21, pp. 2748–2750, 2010.

[16] A. H. Yahaya, M. A. Sutarsono, and F. Ahmad, "First-order interaction multiple regressions model on water quality index in manjung river and its tributaries," *ARPN Journal of Engineering and Applied Sciences*, vol. 11, no. 20, pp. 11919–11924, 2016.

[17] S. A. Che Osmi, W. F. Wan Ishak, M. A. Azman, A. S. Ismail, and N. A. Samah, "Development of water quality modelling using InfoWork river simulation in Malacca River, Malaysia and contribution towards Total Maximum Daily Load approach," *IOP Conference Series: Earth and Environmental Science*, vol. 169, no. 1, 2018, doi: 10.1088/1755-1315/169/1/012073.

[18] T. C. T. Pissarra *et al.*, "A regression model of stream water quality based on interactions between landscape composition and riparian buffer width in small catchments," *Water (Switzerland)*, vol. 11, no. 9, 2019, doi: 10.3390/w11091757.

[19] A. K. R, A. C. R, V. Murali, and A. P. C, "Water Quality Modeling and Regression Analysis of Vembanad Lake and its primary Inflows; Water Quality Modeling and Regression Analysis of Vembanad Lake and its primary Inflows." [Online]. Available: www.ijert.org

[20] E. Ostertagová, "Modelling using polynomial regression," *Procedia Engineering*, vol. 48, pp. 500–506, 2012, doi: 10.1016/j.proeng.2012.09.545.

[21] M. G. Uddin, S. Nash, and A. I. Olbert, "A review of water quality index models and their use for assessing surface water quality," *Ecological Indicators*, vol. 122, p. 107218, 2021, doi: 10.1016/j.ecolind.2020.107218.

[22] O. E. Ruth Olubukola Ajoke Adelagun, Emmanuel Edet Etim and Godwin, *Application of Water Quality Index for the Assessment of Water from Different Sources in Nigeria*. 2021. [Online]. Available: http://dx.doi.org/10.5772/intechopen.98696

[23] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2006," 2006.

[24] S. Suratman, M. I. Mohd. Sailan, Y. Y. Hee, and E. A. Bedurus, "A Preliminary Study of Water Quality Index in Terengganu River Basin , Malaysia," *Sains Malays*, vol. 44, no. 1, pp. 67–73, 2015.

[25] C. Choong, "Designing Urban Observatory Governance and Data Analytics in Iskandar Malaysia," *UNDP Malaysia, Singapore & Brunei Darussalam*. https://www.my.undp.org/content/malaysia/en/home/operations/projects/democratic_governance/designing-urban-observatory-governance-and-data-analytics-in-isk.html

[26] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2009," 2009.

[27] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2011," 2011.

[28] Department of Environment Malaysia, "Malaysia Environmental Quality Report 2012," 2012.

[29] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2015," 2015.

[30] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2016," 2016.

[31] Department of Environment Malaysia., "Malaysia Environmental Quality Report 2018," 2018.

[32] Department of Environment Malaysia, "Malaysia Environmental Quality Report 2020," 2020.

[33] H. Anton and C. Rorres, *Elementary Linear Algebra with Supplemental Applications*, 10th ed. New York: John Wiley & Sons, 2011.

[34] R. El-korashey, "Using Regression Analysis to Estimate Water Quality Constituents in Bahr El Baqar Drain," *Journal of Applied Sciences Research*, vol. 5, no. 8, pp. 1067–1076, 2009.

[35] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, "Efficient Water Quality Prediction Using Supervised," *Water (Basel)*, vol. 11, no. 1, pp. 1–14, 2019.

[36] Z. Musa, "Insufficient funds given for river sustainability," *The Star Online*, 2021.

[37] B. Yan, D. Liu, Z. Gao, Q. Cao, and J. Chai, "Water Quality Evaluation and Variation Trend Analysis of Rivers Upstream of the Dahuofang Reservoir in China," *Water (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/w14091398.