

Study on Correlative Factors of Seawater Temperature Based on Regression Analysis

Yingting Chen^{1, †, *}, Runyi Zhao^{2, †}

¹ Faculty of Science, University of Strathclyde, Glasgow, United Kingdom

² Faculty of Statistics and Mathematics, Shandong University of Finance and Economy, Jinan, China

* Corresponding author: yingting.chen.2022@uni.strath.ac.uk

[†]These authors contributed equally.

Abstract. The seawater temperature is important that affects marine ecosystems and marine life. This paper aimed to analyze the main factors that affect seawater temperature using Bottle Database collected at California Cooperative Oceanic Fisheries Investigations stations. The dataset contained seawater samples collected using bottles. Linear regression was performed using temperature as the dependent variable and depth, salinity, oxygen, and potential density as independent variables. Stepwise methods were used to select variables for regression models. To determine the most influential factor affecting temperature, the random forest was used to rank four independent variables. The plot generated by this analysis showed that potential density strongly influenced seawater temperature, followed by salinity, oxygen, and depth. Then four models were built, including Ordinary Multilinear regression, Linear regression considering the interaction, Linear regression with variable transformation, and Linear regression combining interactions and transformation. The 4th Model, had the highest adjusted R-squared and the lowest prediction sum of squares, making it the most useful model. This model indicated that 4 predictor factors all have significant effects on temperature, and their interactions also play a crucial role. Finally, model 4 can be used to predict water temperature based on depth, salinity, oxygen, and potential density, showing that seawater temperature could be predicted efficiently using these relevant variables. This study provides valuable insights into the main factors affecting seawater temperature and demonstrates the usefulness of linear regression and random forest approaches.

Keywords: Regression analysis, random forest, Correlative Factors, Seawater Temperature.

1. Introduction

The ocean is a vital part of the Earth and serves as a habitat for numerous species of marine life. Seawater temperature is affected by a variety of factors. Changes in seawater temperature can have broader implications for the Earth's climate system. In this context, understanding the relationship between temperature and its influencing factors is of great importance.

The seawater temperature varies with depth. As ocean depth increases, the amount of sunlight diminishes, resulting in a drop in temperature and an exponential increase in pressure [1]. Temperature also affects the salinity of the water. As temperature increases, the rate of evaporation also increases, which can lead to a higher concentration of salt in the remaining seawater. The cool water was mostly relatively saline [2].

According to Truesdale, Downing, & Lowden, the rate of solution of oxygen in water, at any time is directly proportional to the difference between its solubility at the temperature of the water and the actual concentration at that time [3]. So, it is reasonable to infer that water temperature also affects the solubility of gases in seawater, such as oxygen.

Water temperature is an important factor that affects marine life. There is evidence in Danovaro, Dell'Anno, & Pusceddu to suggest that the deep-sea fauna is extremely sensitive to environmental changes, with even small temperature shifts having a significant and rapid impact on the diversity of deep-sea life. Furthermore, a decrease in temperature might lead to reduced functional diversity and species evenness in this ecosystem [4].

There is also a relationship between potential density and seawater temperature. Potential density is the characteristic that governs the stability, mixing, and mesoscale movements of particles in seawater [5]. As seawater temperature decreases, its density increases. However, the relationship between potential density and temperature is not linear because the density of seawater is also affected by other factors such as salinity and pressure.

Additionally, changes in seawater temperature can have broader impacts on the Earth's climate system. From previous research, oceans have absorbed approximately 90 percent of the total heat which was added to the climate system in the past 50 years [6]. Rising levels of greenhouse gases and global warming are likely to result in increases in sea surface temperatures.

In summary, the ocean temperature is a vital observational indicator for the ocean, which is influenced by many factors and is closely related to depth, salinity, gas solubility, biodiversity, and global warming.

An attempt is made to find the relationship between temperature and its influencing factors. Based on the observed ocean sample data, it is curious whether there is a relationship between the observed values of Depth, Salinity, oxygen, Potential Density, and water temperature. What's more, it is also worth exploring which factor is the most influential.

In the following articles, the data from the Bottle Database will be used to conduct the study. Then constructed a regression model for temperature and obtained a regression equation for predicting seawater temperature. Through analyzing ocean sample data and constructing regression models, we can gain insights into predicting seawater temperature and understanding its broader implications for the planet.

2. Methodology

2.1. Source of data

The research used the Bottle Database, collected at California Cooperative Oceanic Fisheries Investigations (CalCOFI) stations. The data set represents the longest and most complete time series of oceanographic data in the world, from the 1949 to present, and contained more than 50,000 sampling stations. They collected seawater samples when cruises using bottles. Oceanographic data includes parameters such as temperature, salinity, dissolved oxygen, depth, silicate, potential density, and many more.

2.2. Data processing and visualization

First, exploratory data analysis was conducted on a sample of 1000 rows and 5 columns of oceanographic data collected by the bottle to understand the relationships between these variables. Since the dataset contains data collected from millions of bottles and some of them include missing values. To better analyze the seawater sample and get a clearer result, 1000 sites with relatively complete data were chosen after deleting the missing value.

Then, to better understand the data, do the data visualization. Patterns, trends, and relationships in data which may not be readily apparent when reviewing raw numbers can be identified easily through viewing visualization.

2.3. Models

2.3.1. Variables choosing and selection

Chose temperature (degrees Celsius) as the dependent variable and depth (meters), salinity (practical salinity scale), oxygen (milliliters per liter), and potential density (Sigma Theta: kilograms per cubic meter) as the independent variables.

After choosing the variables, use stepwise methods to select variables for our regression models. AIC is an important judgment indicator in this method. Bozdogan's research has demonstrated that selecting the model with the minimum expected information loss is asymptotically equal to selecting

the model with the minimum AIC value [7]. So, choose the smallest AIC value to select suitable variables.

2.3.2. Random Forest

Random forests are a modified version of bagged trees that enhance their performance by decreasing the correlation between individual trees. The algorithm restricts the available predictors for each tree split, which may seem counterintuitive but addresses a specific problem in many datasets with one strong predictor and several moderately strong ones. Bagged trees using all predictors use the strong one in the top split, creating highly correlated predictions. By limiting the predictors considered at each split, random forests reduce correlation and improve the model's reliability. This process decorrelates the trees, ensuring only a fraction of the predictors are used in each tree, leading to a less variable and more reliable average of the resulting trees [8].

Use the random forest approach in R to rank four independent variables.

2.3.3. Model building and choosing

Built four models based on linear regression.

In model 1 we do Ordinary Multilinear regression.

In model 2 we consider the interaction between the independent variables.

In model 3, we do Linear regression with variable transformation.

In model 4, we do Linear regression with interactions and transformation.

Then choose the model which performed well. Compared four models using two related quantities: adjusted R-squared and prediction sum of squares (PRESS).

The R^2 statistic is a proportion, ranging from 0 to 1, that represents the amount of variance explained and is not affected by the scale of Y. A high R^2 value indicates that the regression explains a significant portion of the response variability. On the other hand, the adjusted R^2 considers the impact of including irrelevant variables in the model. A higher adjusted R^2 suggests a model with a smaller test error. The model with the highest adjusted R^2 will contain only relevant variables and no extraneous variables [9].

The predictive sum of squares (PRESS) is a useful tool for assessing and comparing patterns. To evaluate a model's predictability, it uses the normal or unauthorized cross-validation technique. Unlike certain model parameters, the PRESS is model-dependent, making it a reliable indicator of predictivity. Therefore, PRESS should be used to identify the most suitable model, while the minimization of the cost function provides the parameters for a specific model [10].

Therefore, the model with the highest adjusted R-squared and lowest prediction sum of squares (PRESS) will be the “best” model.

3. Results And Discussion

3.1. Data Processing

Use a scatterplot matrix to visualize the relationships between the variables. The relationships between any two variables can be seen in this diagram. Use R to plot the matrix.

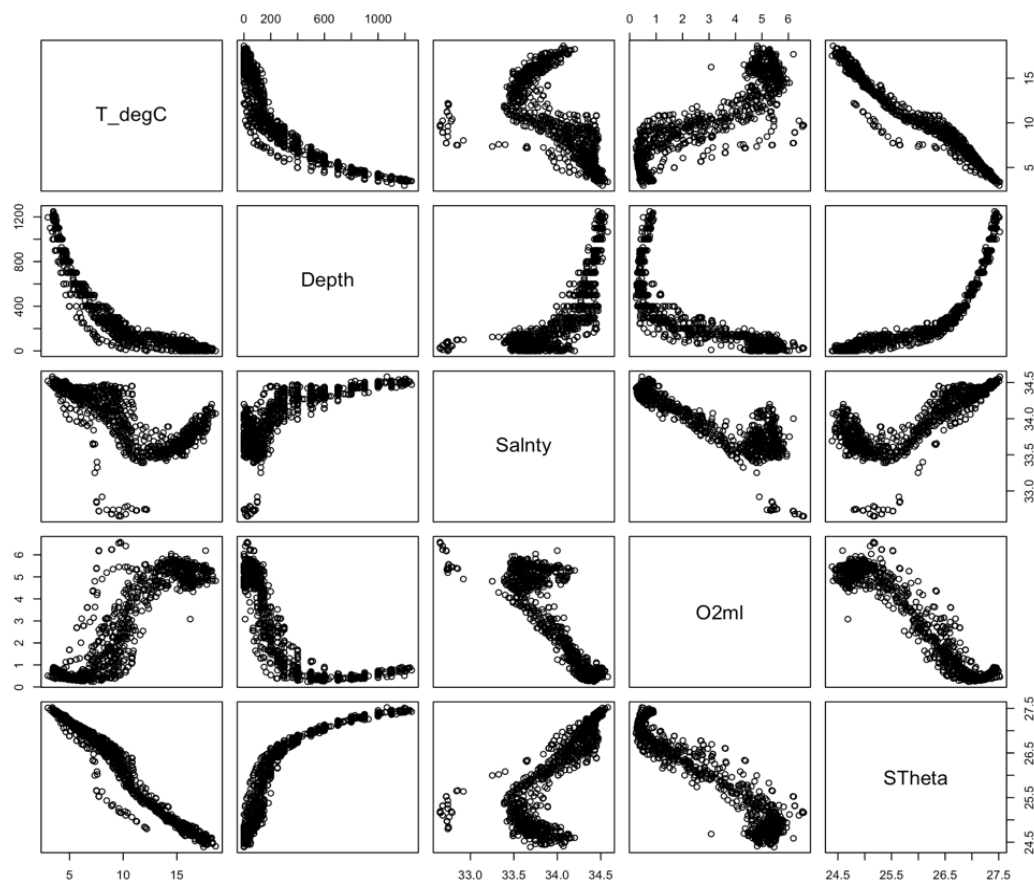


Figure 1. Scatterplot Matrix (Photo credit: Original)

It can determine from Figure 1 that depth, salinity, oxygen, or potential density has a linear relationship with the temperatures.

Then use R to calculate Pearson correlation coefficients to quantify these relationships which are also shown in Table 1.

Table 1. Pearson correlation coefficients

	T_degC	Depth	Salnty	O2ml	STheta
T_degC	1.0000000	-0.8647375	-0.6676037	0.8966448	-0.9800781
Depth	-0.8647375	1.0000000	0.7689906	-0.8083789	0.8704972
Salnty	-0.6676037	0.7689906	1.0000000	-0.8682809	0.792267
O2ml	0.8966448	-0.8083789	-0.8682809	1.0000000	-0.9562838
STheta	-0.9800781	0.8704972	0.792267	-0.9562838	1.0000000

It can be found that the Pearson correlation coefficient between oxygen and potential density is -0.9562838. This shows that there is a strong negative correlation between oxygen and potential density (Sigma Theta). And the Pearson correlation coefficient between oxygen and salinity is -0.8682809. This shows that there is also a strong negative relationship between oxygen and salinity.

Use the descriptive statistics tables for 5 variables to show minimums and means etc. in tabular form in Table 2.

Table 2. descriptive statistics tables

	T_degC	Depth	Salnty	O2ml	STheta
Min	2.980	0.0	32.65	0.240	24.39
1st Qu.	6.055	50.0	33.68	0.600	24.92
Median	9.615	183.5	34.02	2.640	26.23
Mean	10.251	324.7	34.00	2.852	26.03
3rd Qu.	14.990	537.5	34.36	5.070	27.00
Max	18.610	1250.0	34.58	6.580	27.52

It is apparent that in Table 2. Among the four independent variable factors, the data distribution is more dispersed for oxygen and more concentrated for salinity. On the other hand, depth and potential density are more evenly distributed.

3.2. Single factor analysis

Use ggplot to generally view the trend of the effect of depth, salinity, oxygen, and potential density on the seawater temperature respectively. The result is shown in Figure 2. Next, use ggplot to see the relationship between every two factors as shown in Figure 3.

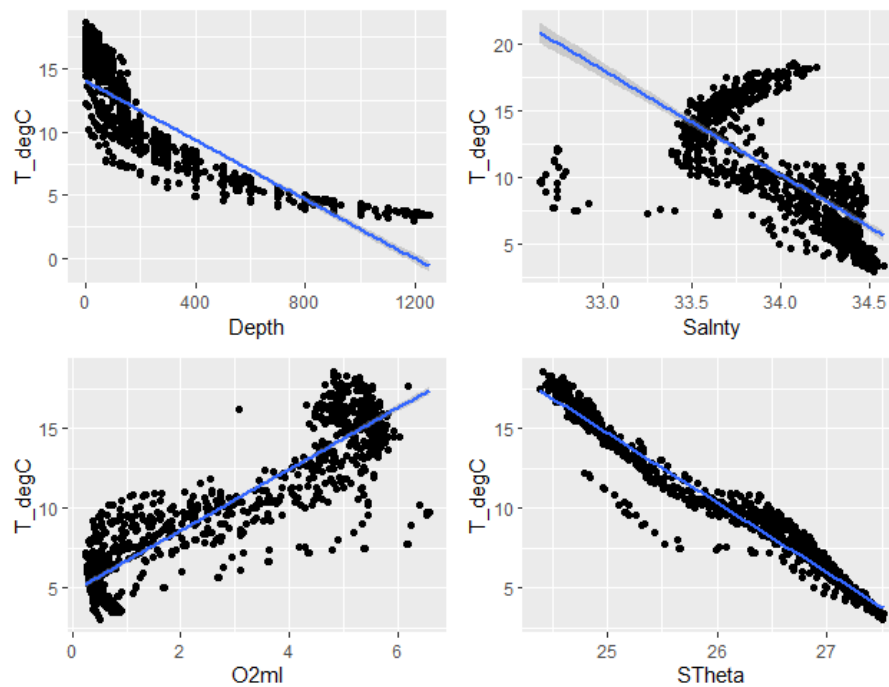


Figure 2. visualize the relationship between independent variables and dependent variable (Photo credit: Original)

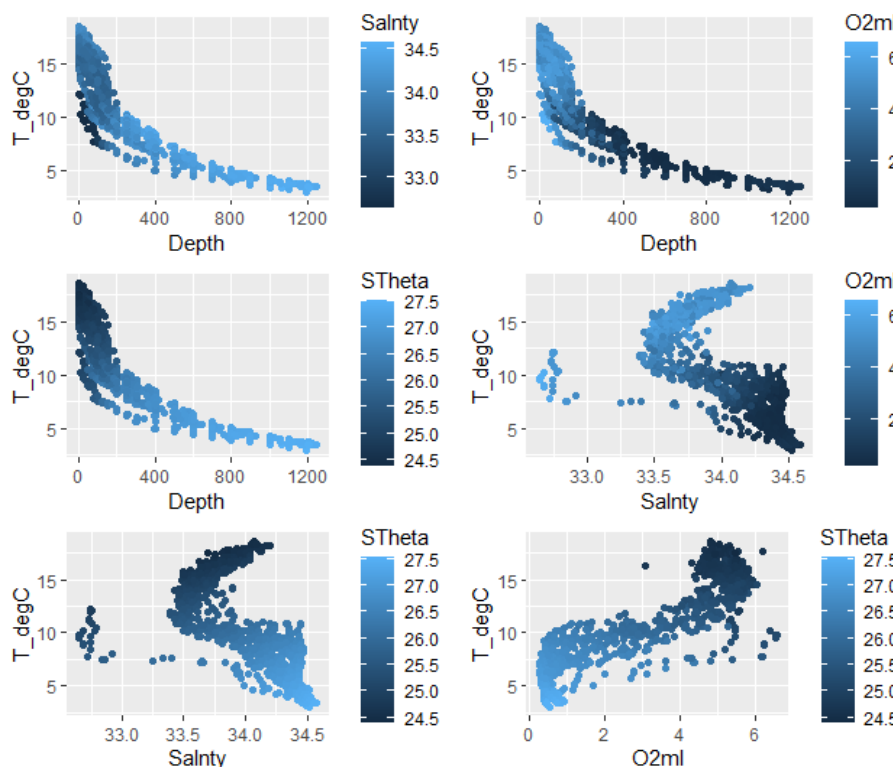


Figure 3. visualize the relationship between factors (Photo credit: Original)

It can be deduced from Figure 2 that there is a strong negative relationship between potential density and temperature. But a relatively less obvious relation between salinity and temperature.

The plots in Figure 3 indicate that:

With oxygen increasing, potential density decreases. With oxygen decreasing, salinity decreases. The relationship between salinity and oxygen is not the obvious one compared to others.

3.3. Model result

First, there is a detailed explanation of the model building of each model. Based on the results of the analysis of the above variables, this section completes the specific construction of the four models, which are designed as follows:

Do Ordinary Multilinear regression in model 1. Because smaller values of AIC indicate better models. The R result is that 4 factors can be independent variables. Then conduct the simple regression with 4 independent variables.

Considering the interaction between the independent variables in model 2. To include interactions among predictor variables in it, using the results of the correlation coefficient calculation in Table 1. Pearson correlation coefficients show that there is a strong relationship between oxygen and potential density, and between oxygen and salinity.

The relationship between salinity and oxygen is not the obvious one in Figure 3. Salt and temperature are generally believed to have a negative correlation, but their relationship plots alone cannot guarantee negative modeling results. Therefore, to increase the accuracy of the model, the interaction between oxygen and salinity was considered during the regression model construction. This approach helped to achieve more precise results in terms of the relationship between salinity and temperature. So, involve variables O2ml interact STheta and O2ml interact salinity in model 2.

Involving the variable transformation in model 3. Setschenow relation was proposed by Garcia, & Gordon [11], which introduces the relationship between temperature and salinity and oxygen solubility. And implies a logarithmic relationship between the solubility of oxygen and temperature. So, we do a logarithmic transformation to variable O2ml.

At last, make a combination of model 2 and model 3 to get model 4.

Next, import the dataset and built the linear model using R. We can get 4 models:

Model 1: Multilinear regression.

$$T_{degC} = 14.39 - 0.00203 \times Depth + 3.739 \times Salnty - 0.04420 \times O2ml - 5.013 \times STheta \quad (1)$$

Model 2: Linear regression with interactions.

$$T_{degC} = -14.64 - 0.001286 \times Depth + 5.547 \times Salnty - 4.211 \times O2ml - 6.26 \times STheta + 0.2666 \times O2ml: STheta - 0.3285 \times O2ml: Salnty \quad (2)$$

Model 3: Linear regression with variable transformation.

$$T_{degC} = 8.407 - 0.002087 \times Depth + 3.84 \times Salnty - 0.01995 \times \ln(O2ml) - 4.92 \times STheta \quad (3)$$

Model 4: Linear regression with interactions and transformation.

$$T_{degC} = 1.683 - 0.001011 \times Depth + 5.324 \times Salnty - 0.3641 \times \ln(O2ml) - 6.601 \times STheta + 0.3532 \times O2ml: STheta - 0.2659 \times O2ml: Salnty \quad (4)$$

3.4. Random forest result

Which can be seen in Figure 4:

potential density is the most influential variable.

Salinity is the second most influential factor.

Oxygen and depth both showed a very indistinctive effect on the seawater temperature.

The importance index of potential density is three or even ten times higher than any other index.

Then conclude that Potential density strongly affected the seawater temperature compared to the other three independent variables. While other factors also affect seawater temperature, they do not have as significant an effect as potential density.

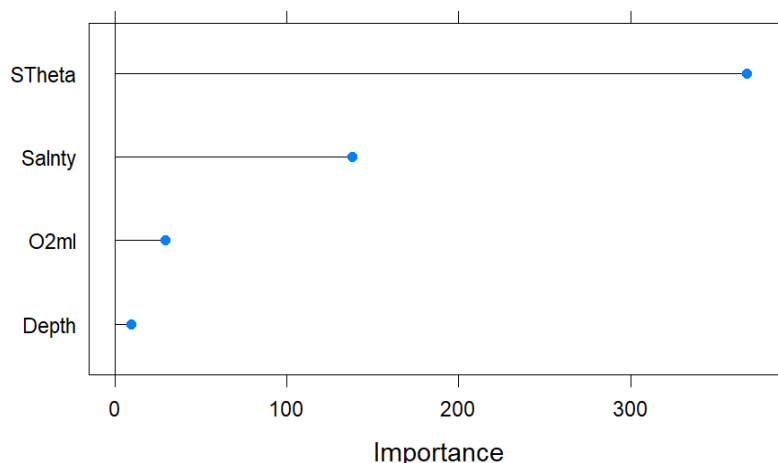


Figure 4. Ranking of the importance of influencing factors (Photo credit: Original)

3.5. Discussion

3.5.1. Models choosing

Compared four models using adjusted R-squared and calculate prediction sum of squares (PRESS). To better observe the differences between the various models, Table 2 was created.

Table 3. Model comparison table

Model	adjR2	PRESS
model 1	0.9978	46.74541
model 2	0.9986	30.54439
model 3	0.9978	47.05163
model 4	0.9987	28.31587

In Table 3, it can be found that model 4, which considers linear regression with interactions and transformation, had the highest adjusted R-squared and the lowest prediction sum of squares. Low values of the predictive summation of the squares correspond to the regression model will perform better in predicting new data. It can conclude that model 4 is the most useful one.

3.5.2. Regression assumptions

To find whether the model that had fitted is appropriate, conducted regression assumption on model 4.

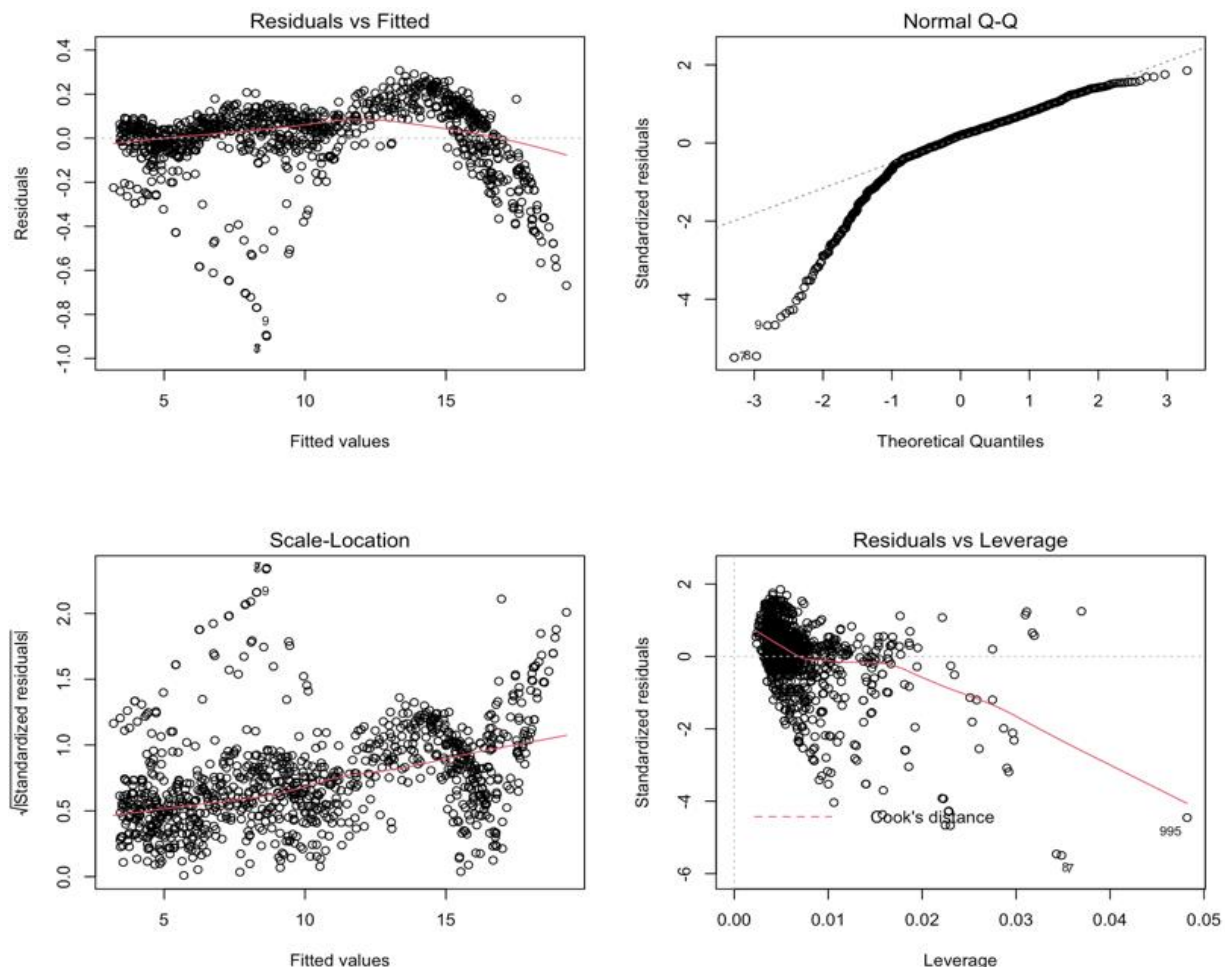


Figure 5. Regression assumptions of model 4 (Photo credit: Original)

This set of plots in Figure 5 illustrates the regression fits well with the linearity hypothesis, residue normality and constant residual variance. Model assumptions appear to be well satisfied [12].

Finally, model 4 can be used to predict water temperature based on depth, salinity, oxygen, and potential density. In conclusion, the program and regression models have shown us that Depth, Salinity, oxygen, and Potential Density are critical factors that influence water temperature. Water temperature is essential in understanding marine ecosystems and the Earth's climate system. We must continue to monitor these changes and make use of the data.

3.5.3. Most influential factor

Potential density strongly affected the seawater temperature compared to the other three independent variables.

Some studies also indicate that the potential density may be reversed from acoustic reflection using accuracy to learn how it affects seawater temperature. To do the potential analysis, the relationship between temperature and salinity also plays an important role [5].

3.6. Limitation

First, the data set is limited. The analysis is based on a specific data set collected at the California Cooperative Oceanic Fisheries Investigations stations. This may limit the generalizability of the findings to other locations or different sampling methods.

Second, the models are simplified in this study. While the model developed for this project includes four important predictors and their interactions, there may be other factors that may also affect seawater temperature that is not included in the analysis. For example, seasonal variations or changes in ocean currents may affect seawater temperature but are not considered.

Furthermore, the predictions have many uncertainties. Although the models developed show promise for predicting seawater temperatures based on the variables of interest, there is still some uncertainty in the predictions. Other factors not considered in the model may still affect temperature, and errors in data collection or measurement may also affect the accuracy of predictions.

Although it is difficult to find the perfect model, it is important to acknowledge these limitations and take them into account when interpreting the results.

4. Conclusion

In conclusion, the analysis of the Bottle Database provided valuable insights into the main factors that affect seawater temperature. Linear regression models that were built showed that depth, salinity, oxygen, and potential density are all significant predictors of seawater temperature, and their interactions also play an important role.

Through a stepwise approach to model selection and identified Model 4 as the most useful model for predicting seawater temperature. This model incorporating interactions and transformation had the highest adjusted R-squared and the lowest prediction sum of squares, making it the most useful model for predicting seawater temperature. Furthermore, our random forest analysis showed that potential density was the most influential factor affecting seawater temperature, followed by salinity, oxygen, and depth.

The study has important implications for improving our understanding of marine ecosystems, as accurate predictions of seawater temperature are critical for it. In addition, the approach to modeling seawater temperature can be extended to other datasets and locations, providing a valuable tool for researchers and policymakers.

Despite the promising findings, there are limitations to this study, including the limited dataset, simplified model, and uncertainty in predictions. Therefore, caution should be taken when interpreting the results and applying them to other locations or scenarios.

By shedding light on the main factors that affect seawater temperature and providing a useful predictive model, this research has the potential to inform management decisions and advance our knowledge of marine ecosystems. Future studies could expand on these results by exploring other factors that may impact seawater temperature. It is hoped that continued research on the primary factors influencing seawater temperature will yield improved predictions and enable better management of marine ecosystems.

References

- [1] M. B. Nirwana, & D. Wulandari (2021). Comparison of Simple and Segmented Linear Regression Models on the Effect of Sea Depth toward the Sea Temperature. *Enthusiastic: International Journal of Applied Statistics and Data Science*, 68 - 75.
- [2] J. N. Heine, B. Semmens (2019). California Cooperative Oceanic Fisheries Investigations Report.
- [3] G. A. Truesdale, A. L. Downing, & G. F. Lowden (2007). The solubility of oxygen in pure water and seawater. *Journal of Applied Chemistry*, 5 (2), 53 - 62.
- [4] R. Danovaro, A. Dell'Anno, & A. Pusceddu (2004). Biodiversity response to climate change in a warm deep sea. *Ecology Letters*, 7 (9), 821 - 828.
- [5] B. Biescas, B. R. Ruddick, M. R. Nedimović, V. Sallarès, G. Bornstein, & J. F. Mojica (2014). Recovery of temperature, salinity, and potential density from ocean reflectivity. *Journal of Geophysical Research: Oceans*, 119 (5), 3171 - 3184.
- [6] M. A. Balmaseda, K. E. Trenberth, & E. Källén (2013). Distinctive climate signals in reanalysis of global ocean heat content. *Geophysical Research Letters*, 40 (9), 1754 - 1759.
- [7] H. Bozdogan (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52 (3), 345 - 370.

- [8] G. James, D. Witten, T. Hastie, & R. Tibshirani, (2013). An introduction to statistical learning (Vol. 112, p. 303-332). New York: springer.
- [9] G. James, D. Witten, T. Hastie, & R. Tibshirani, (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed., pp. 234-235). New York: Springer.
- [10] A. Bartoli (2009). On computing the prediction sum of squares statistic in linear least squares problems with multiple parameter or measurement sets. International journal of computer vision, 85, 133 - 142.
- [11] H. E. Garcia, & L. I. Gordon (1992). Oxygen solubility in seawater: Better fitting equations. Limnology and oceanography, 37 (6), 1307 - 1312.
- [12] Kabacoff, R. (2022). R in Action: Data Analysis and Graphics with R and Tidy verse. Simon and Schuster.