PASIG RIVER WATER QUALITY ESTIMATION USING AN EMPIRICAL ORDINARY LEAST SQUARES REGRESSION MODEL OF SENTINEL-2 SATELLITE IMAGES

J.E. Escoto[1*], A.C. Blanco[1,2], R.J. Argamosa[1], J.M. Medina[1]

[1]Training Center for Applied Geodesy and Photogrammetry, University of the Philippines, Diliman, Quezon City 1101
jimescoto@outlook.com, acblanco@up.edu.ph, rlargamosa@up.edu.ph, jmmedina@up.edu.ph
[2] Dept. of Geodetic Engineering, University of the Philippines, Diliman, Quezon City 1101, Philippines

KEY WORDS: Water Pollution, Remote Sensing, DAO-2016-08, Statistical Models

ABSTRACT:

This study entails generation of empirical ordinary least squares regression models to estimate water parameters. It uses remote sensing for environmental monitoring of Pasig River located in the Philippines. This uses measurements of primary water quality (WQ) parameters defined on Department of Environment and Natural Resources Administrative Order 2016-08 recorded on the Pasig River Unified Monitoring Stations (PRUMS) report from January to June of 2019. Sentinel-2 images are utilized to estimate biological oxygen demand (BOD), Chloride, Color, Dissolved Oxygen (DO), Fecal Coliform, Nitrate, pH, Phosphate, Temperature, and Total suspended solids (TSS). Feature generation involved calculation of different band reflectances from the satellite image. Exhaustive feature selection through application of a Pearson Correlation threshold was applied to limit number of independent variables. The box-cox transformations of water quality parameters (except for Temperature) were used as dependent variables and the selected features are used as dependent variables for the ordinary least squares regression model. The root mean square error (RMSE) values for the models which are computed using the k-fold cross validation technique showed outliers, especially for the TSS model (>547000 mg/L), which made its average negative RMSE so large. Tests for multicollinearity, autocorrelation, and homoscedasticity indicated problems in models created. However, normality of residuals indicates that models allow us to roughly estimate water quality for the river as a whole with the advantages of remote sensing, enabling a better perspective for its spatial distribution.

1. INTRODUCTION

Pasig River connects Laguna de Bay to Manila Bay. The river stretches up to 27 kilometers with an average depth of 50 meters. Through the years, it has served as an important means of transport. However, today, it suffers from high levels of water pollution (Meijer, et. al, 2021). According to the Pasig River Rehabilitation Program Case Study (2004), it dates back after World War 2 when there was a massive population growth, construction of lots of infrastructures and a dispersal of economic activities. It was observed during the 1930s that there was a significant increase of pollution, diminishing fish migration from Laguna de Bay, and decrease of ferry transports. Foul smells began in 1970s and in the 1980s. During the 1990s, its water quality failed to meet Class C standards, a classification suited for fishery water for propagation and growth of fish and other aquatic resources. It was also then declared biologically dead by the Pasig River Rehabilitation Commission (PRRC).

A more recent study by Gorme, et al. (2010) stated that Pasig River was very polluted and failed to meet the Department of Environment and Natural Resources (DENR) standards for dissolved oxygen (DO) and BOD. Water quality in the river improved from the time when the Pasig River Rehabilitation Commission (PRRC) was established in 1999, but continued to deteriorate through the years. According to American Association for the Advancement of Science (2021), Pasig River is considered the world's most polluting river when it comes to plastic waste. The 27-kilometer Pasig River which runs through Metro Manila, accounting for 63,000 tons of plastic entering oceans from rivers per year.

A problem which this paper aims to solve is the big gap in the retrieval of water quality using remote sensing methods. Although remote sensing provides a more cost efficient and faster complementary approach for a more comprehensive assessment of water bodies compared to conventional water quality monitoring methods (i.e. sampling and lab analysis), it is still limited to the retrieval of water clarity, turbidity, water color, and the concentrations of optically active constituents (Wisconsin DNR., n.d.). A study by Márquez, et. al. (2018) generated an empirical model to estimate Temperature, PO_4, Total suspended solids (TSS), Turbidity, pH and Electrical conductivity (EC) using Landsat 8 images tested a multitude of different independent variables of the reflectance values such as individual bands, combinations of bands, square roots, reciprocals, square, cubic, powers, sums, subtractions, logarithms, and band ratios for their linear regression model. This study aims to enhance environmental monitoring of Pasig River using remote sensing methods.

This study uses different Sentinel-2 image band combinations to generate empirical ordinary least squares (OLS) models to estimate different water quality (WQ) parameters established by the Department of Environment and Natural Resources Administrative Order (DAO-2016-08), namely, Biological oxygen demand (BOD), Chloride, Color, Dissolved Oxygen (DO), Fecal Coliform, Nitrate, pH, Phosphate, Temperature, and Total suspended solids (TSS). It aims to analyze the water quality of the river through time based on the Pasig River Unified Monitoring Stations (PRUMS) data from January to June of 2019. It also aims to estimate the water quality parameters using Sentinel-2 satellite images from the same date.

2. METHODS

Table 1 describes the datasets used for this study.

Data	Source	Type	Date
PRUMS WQ Report (Monthly Primary WQ Parameters Readings)	PRRC Environmental Management Division	PDF	2019
Sentinel-2 Level 2-A	Google Earth Engine	TIF	2019
Fourteen PRUMS Station Points	Derived from PRUMS Report	Shapefile	2019

Table 1. Summary of datasets used including source, type and date

Figure 1 describes the general methodology.

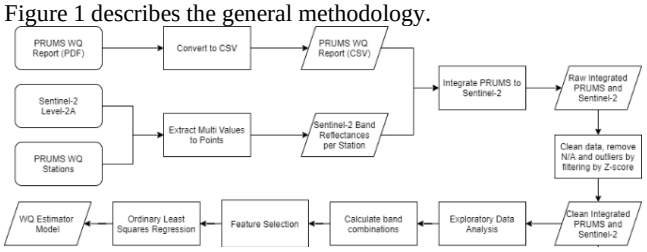


Figure 1. Summary of datasets used including source, type and date

The PRUMS report is manually converted into a CSV format per station per WQ parameter. The shapefile for the PRUMS WQ stations will be used in ArcMap to extract multi values, which are the band reflectances of each calibrated Sentinel-2 Level-2A image. Band reflectance per station is integrated to the PRUMS report with each corresponding date. Cloud-obstructed stations are removed in the image, and PRUMS data is filtered by Z-score, removing entries which go beyond the threshold of 3 standard deviations, to remove outliers. Exploratory data analysis is then implemented to potentially apply a box-cox transformation, or other any necessary transformation, which allows a non-normal dependent variable to be transformed into a normal shape. Different band combinations are calculated (Figure 2) in the next step, then an exhaustive feature selection using different Pearson Correlation thresholds is done to calculate the final empirical OLS model per water quality parameter. These thresholds aim to limit number of features per model between 14 to 16.

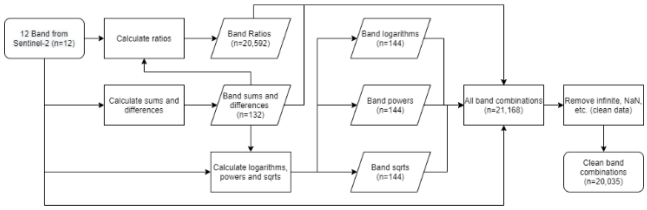


Figure 2. Methodology workflow for generating new features based from band combinations

The exhaustive feature selection described above will be based on a determined threshold for the magnitude of each parameter's Pearson correlation. It allows thousands of features computed (Figure 2) to be filtered quickly. The features which will be selected must have a balance in high positive and high negative correlations with the water quality parameters for a better model performance.

OLS regression estimates the relationship between one or more independent variables (Sentinel-2 bands, 2019 Q1-Q2) and a dependent variable (WQ parameters in PRUMS report, 2019 Q1-Q2), as described in Equation 1.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

where:

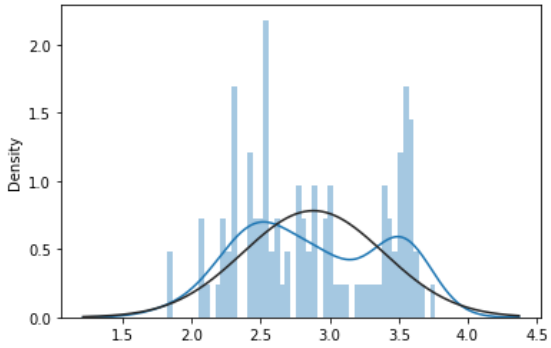
Y = Dependent variable (WQ parameter)
β_0 = Constant
$\beta_1, \ldots, \beta_n$ = Coefficients
$X_1, \ldots, X_n$ = Independent variables (Sentinel-2 bands)
ε = Error

It is a statistical method of analysis which minimizes the sum of the squares in the difference between the observed and predicted values of the dependent values configured as a straight line. Producing one model per WQ parameter, there will be a total of ten models using data from 2019 Q1 to Q2. These models are tested by calculating their RMSE, and by evaluating the normality of their residuals

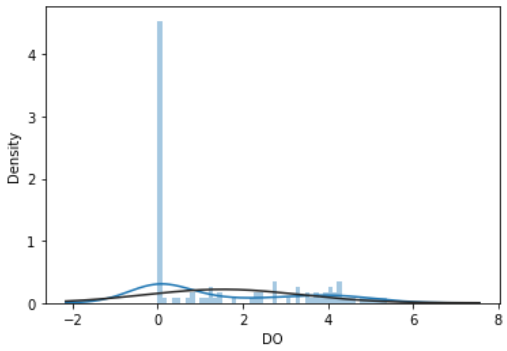
3. RESULTS AND DISCUSSION

Exploratory data analyses (EDA) yielded the following results for the *distplot* (Figure 3 (i) to (x)), which is a combination of the histogram and kernel density estimate. It shows both the distribution of the data in bars and as a comparison to the standard distribution. Box-cox transformation was applied on BOD, Fecal Coliform, Chloride, TSS, Nitrate, Phosphate, Color, and pH on all 106 data point observations. Temperature was not included because it already has near normal distribution.

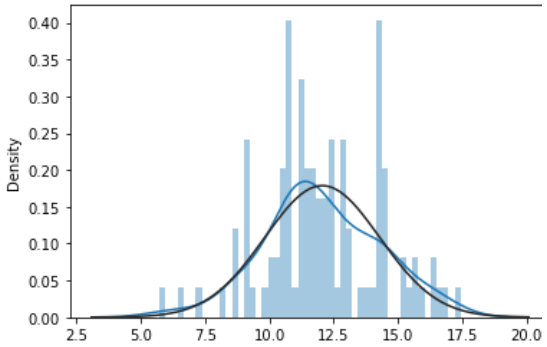
i. BOD



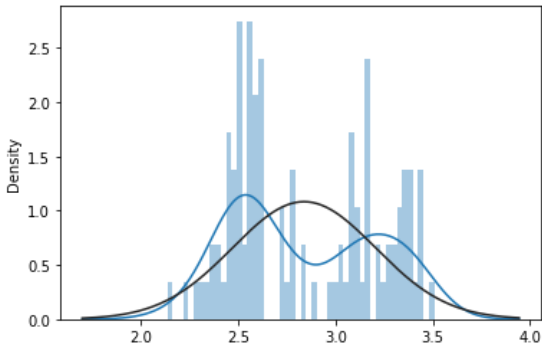
ii. DO



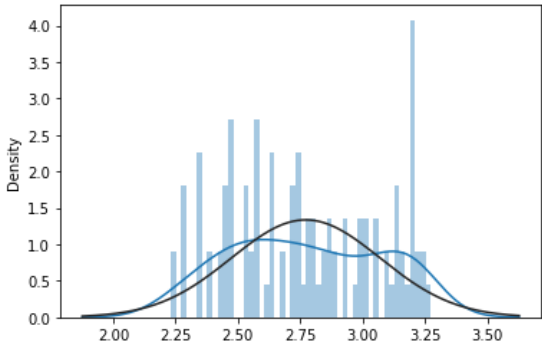
iii. Fecal Coliform



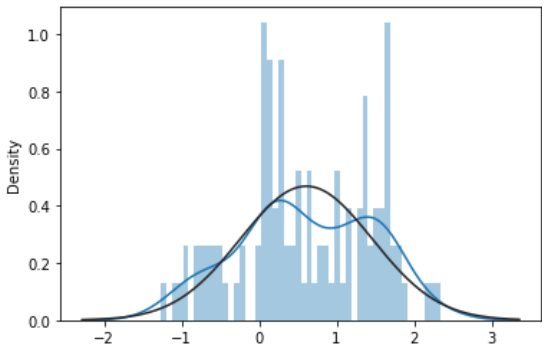
iv. Chloride



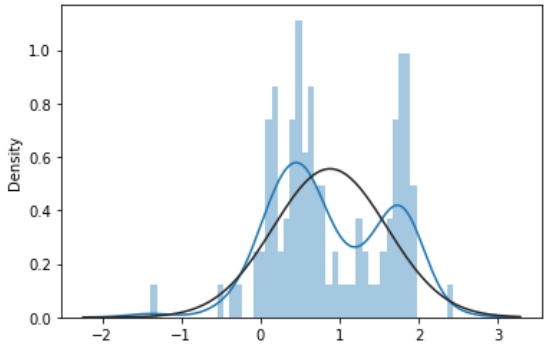
v. TSS



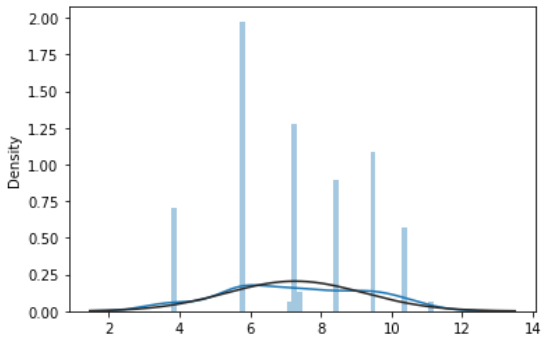
vi. Nitrate



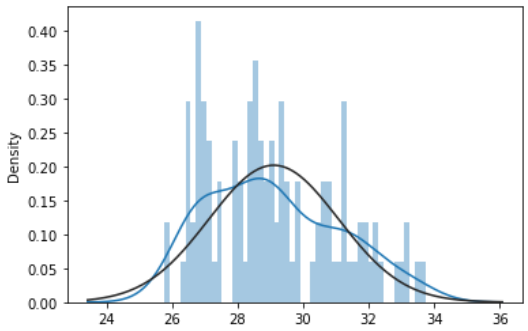
vii. Phosphate



viii. Color



ix. Temperature



x. pH

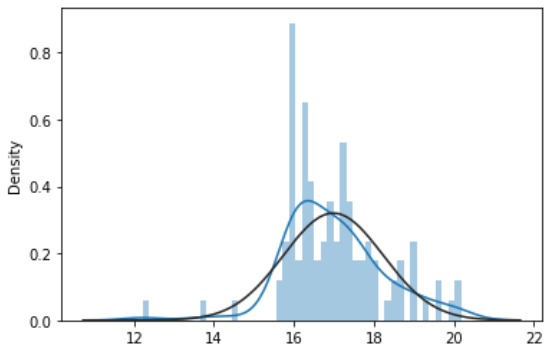


Figure 3. From (i) to (x): BOD, DO, Fecal coliform, Chloride, TSS, Nitrate, Phosphate, Color, Temperature, pH: Shows the *distplot*

The resulting positive and negative thresholds which limits the number of features from 14 to 16 is described in Figure 4 (a). The three WQ parameters with the lowest magnitude of Pearson Correlation threshold includes Fecal Coliform (0.31 & 0.27),

DO (0.29 & 0.31) and Nitrate (0.33 & 0.31), which means that the threshold is lenient because the calculated features has less overall correlation to these parameters. The three highest includes temperature (0.53 & 0.52), color (0.45 & 0.43) and BOD (0.45 & 0.43) on the other hand has a strict threshold since for a feature to be included in the model, it needs to have a very high correlation. The final number of features per WQ model are shown in Figure 4 (b).

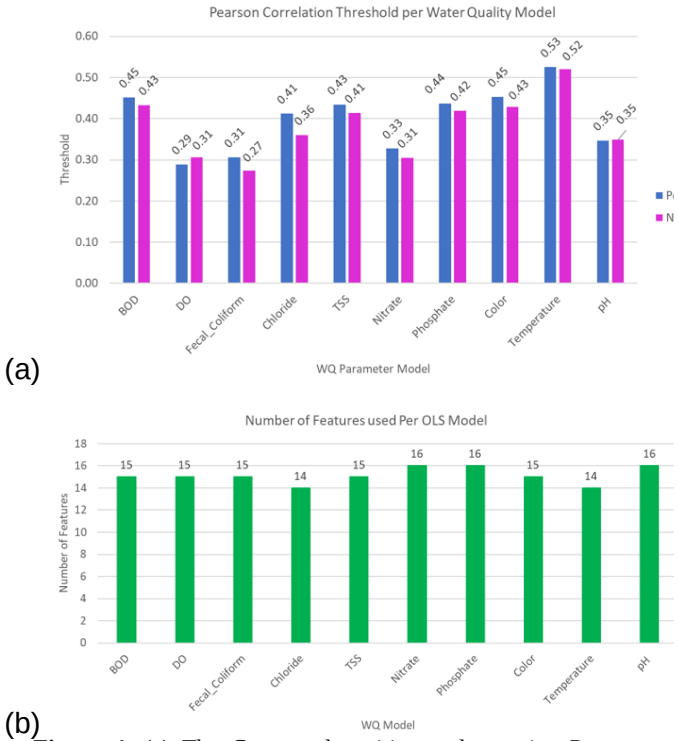


Figure 4. (a) The Computed positive and negative Pearson Correlation Threshold for each of the water quality parameter model after feature selection and (b) Number of features per model

The RMSE values for the models which are computed using the k-fold cross validation technique are shown in Figure 5. It is observed that there are outliers, especially for the TSS model (>547000 mg/L), which made its average negative RMSE so large. This indicates the problem of underfitting for these models.

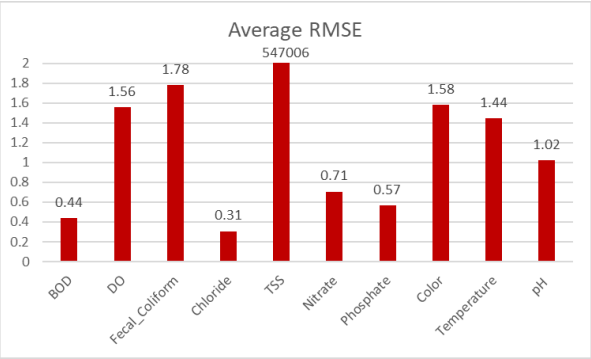
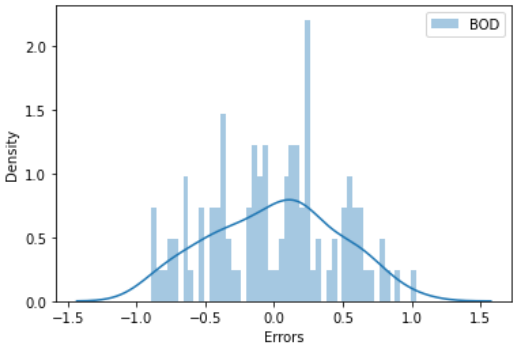


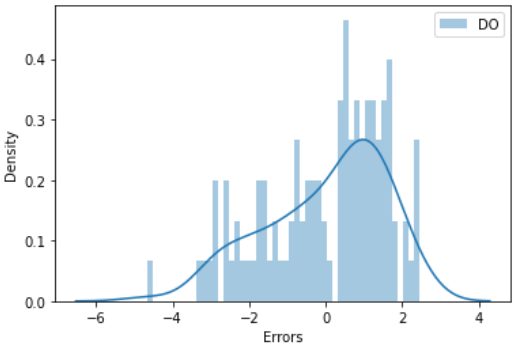
Figure 5. Average RMSE per water quality model

Based from the RMSE values in figure 5, it is observed that the Chloride (0.31), BOD (0.44), Phosphate (0.57) and Nitrate (0.71) models has the least amount of discrepancy in terms of predicted value compared to other models. Meanwhile, the TSS is considered to be an outlier because of its massive RMSE value (>547000).

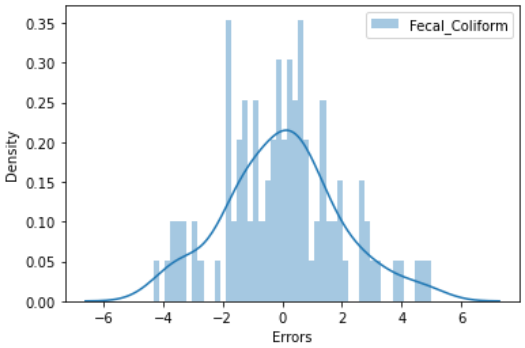
i. BOD



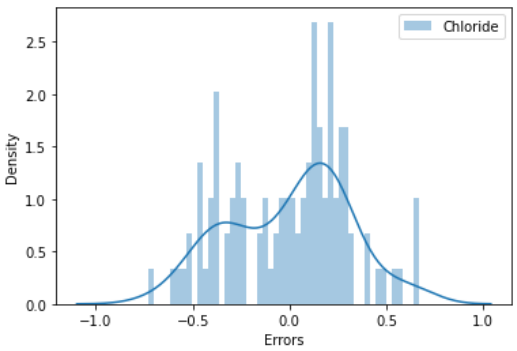
ii. DO



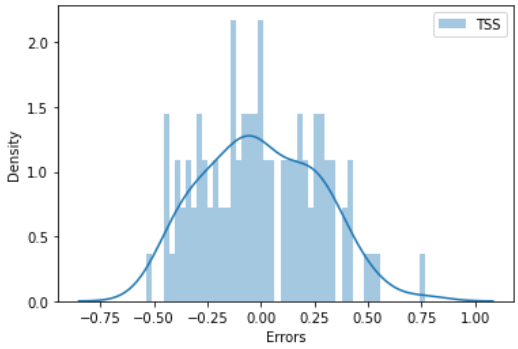
iii. Fecal Coliform



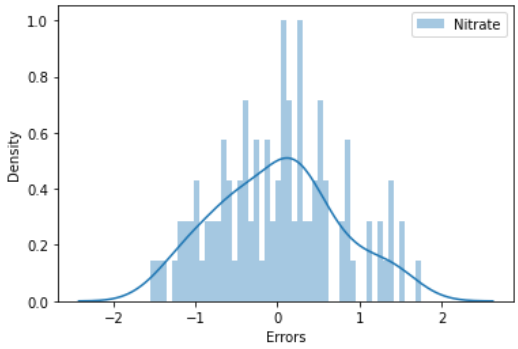
iv. Chloride



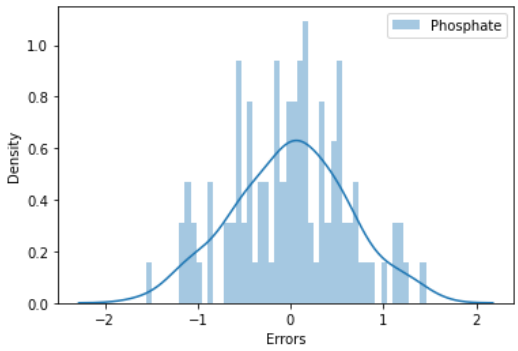
v. TSS



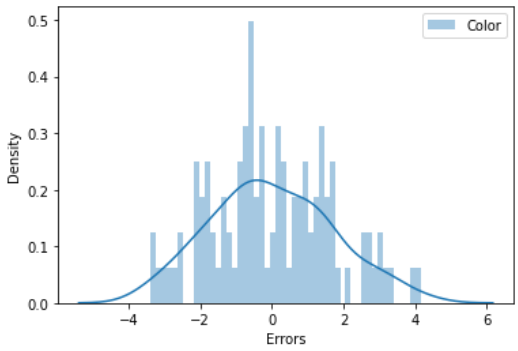
vi. Nitrate



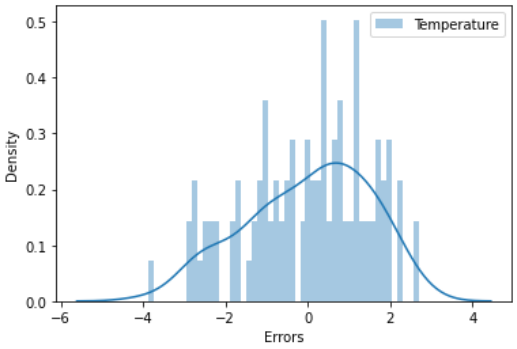
vii. Phosphate



viii. Color



ix. Temperature



x. pH

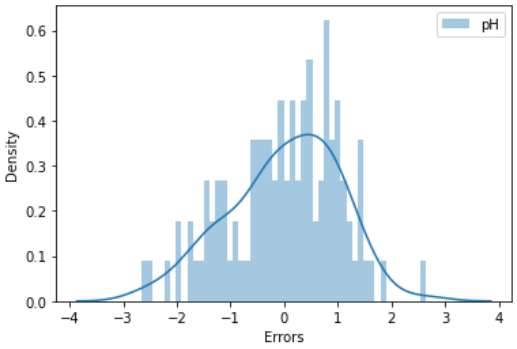


Figure 6. From (i) to (x): BOD, DO, Fecal coliform, Chloride, TSS, Nitrate, Phosphate, Color, Temperature, pH: Shows the *hisogram* of residuals per model

Figure 6 describes the distribution of the residuals through a histogram. The Shapiro-Wilk test which is a reliable test of normality was applied to each model and yielded the following results:

Model	Statistics	*p*	Gaussian Distribution?
BOD	0.98	0.15	Yes
DO	0.94	0	No
Fecal Coliform	0.98	0.35	Yes
Chloride	0.97	0.02	No
TSS	0.98	0.25	Yes
Nitrate	0.98	0.4	Yes
Phosphate	0.99	0.77	Yes
Color	0.98	0.45	Yes
Temperature	0.97	0.01	No
pH	0.98	0.2	Yes

Table 2. Shapiro-Wilk results for water quality models

Results shown in Table 2 indicate that all models except DO, chloride, and temperature have normal residual distribution. Residuals for DO and temperature is skewed towards the right, while a bimodal distribution is present in the residuals for Chloride. A potential cause for this is that some of the predictors are significantly non-normal, which might affect the confidence intervals for these three models.

Table 3 summarizes the results for the following tests: Linear Relationship between the Target and the Feature (Appendix A), Little to no multicollinearity among predictors (Appendix B), Autocorrelation, and Homoscedasticity of Error Terms:

Model	Linear rel. between target & features	Multicollinearity	Autocorrelation	Homoscedasticity
BOD	Biased towards higher values	Yes	Little to no autocorrelation	Yes
DO	Biased towards lower values	Yes	Little to no autocorrelation	No
Fecal Coliform	Slightly biased towards higher values	Yes	Signs of positive autocorrelation	Yes
Chloride	Slightly Biased towards higher values	Yes	Signs of positive autocorrelation	Yes
TSS	Biased towards higher values	Yes	Little to no autocorrelation	Yes
Nitrate	Relationship is non-linear	Yes	Little to no autocorrelation	Yes
Phosphate	Relationship is non-linear	Yes	Little to no autocorrelation	Yes
Color	Relationship is non-linear	Yes	Little to no autocorrelation	Yes
Temperature	Linear relationship	Yes	Signs of positive autocorrelation	No
pH	Biased towards higher values	Yes	Little to no autocorrelation	No

Table 3. Summarized test results for water quality models

Based from the tests, the ones with the most linear relationship between target and features are the temperature (linear), Fecal Coliform (slightly biased), and Chloride (slightly biased), this means that all other models might indicate underfitting. All models failed the multicollinearity test, which becomes a problem since it adds to the overall standard errors to its predictions. Signs of positive autocorrelation appears on the Fecal Coliform, Chloride, and temperature models, which can impact model estimates. The test for homoscedasticity failed for DO, Temperature and pH only.

Even though most of the models exhibit normal distribution of residuals, it has not performed well based on the test results. It can roughly estimate the relative values for the water quality parameters; however, it cannot be concluded that the values are accurate enough to know the precise values in each spot of the river. Possible reasons include: One, that the methodology in feature selection described in Section 2 lacks checking correlation between other features during the process. Two, the dataset is very limited (106 observations) with discrepancies between satellite passing dates and dates of data reading, and the exact locations of the observation point and the sample point picked in the image.

4. CONCLUSION AND RECOMMENDATION

This study generated and tested empirical OLS models for estimating water quality parameters such as BOD, DO, Fecal coliform, Chloride, TSS, Nitrate, Phosphate, Color, Temperature, pH on Pasig River using Sentinel-2 satellite images. Although the exhaustive feature selection process did not produce excellent models based from the test results, they can still be used to roughly estimate water quality from the river because of the normal distribution of residuals which can be performed more quickly and with less cost for fast estimation purposes. The recommendation for the methodology is to include steps to prevent multicollinearity between the features while currently being filtered. This can potentially improve model performance significantly, since this is the test which all models failed.

REFERENCES:

González-Márquez, L. C., Torres-Bejarano, F. M., Rodríguez-Cuevas, C., Torregroza-Espinosa, A. C., & Sandoval-Romero, J. A. (2018). Estimation of water quality parameters USING Landsat 8 images: Application to Playa Colorada Bay, Sinaloa, Mexico. Applied Geomatics, 10(2), 147–158. https://doi.org/10.1007/s12518-018-0211-9

Gorme, J. B., Maniquiz, M. C., Song, P., & Kim, L.-H. (2010). The Water Quality of the Pasig River in the City of Manila, Philippines: Current Status, Management and Future Recovery. Environmental Engineering Research, 15(3), 173–179. https://doi.org/10.4491/eer.2010.15.3.173

Helmer, Richard, Hespanhol, Ivanildo & World Health Organization. (1997). Water pollution control : a guide to the use of water quality management principles. London : E & FN Spon. https://apps.who.int/iris/handle/10665/41967

Meijer, L. J. J., van Emmerik, T., van der Ent, R., Schmidt, C., & Lebreton, L. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. Science Advances, 7(18), eaaz5803. https://doi.org/10.1126/sciadv.aaz5803

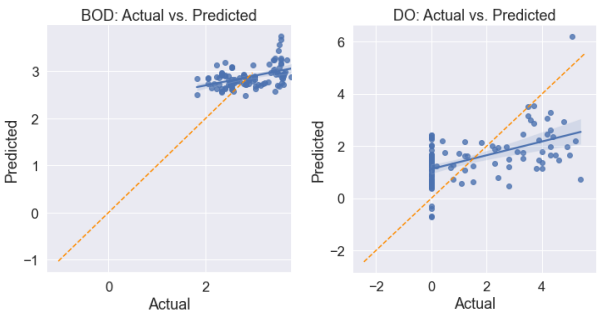
Pasig River Timeline. (n.d.). Retrieved October 22, 2020, from https://prezi.com/cp9rc-egdlta/pasig-river-timeline/

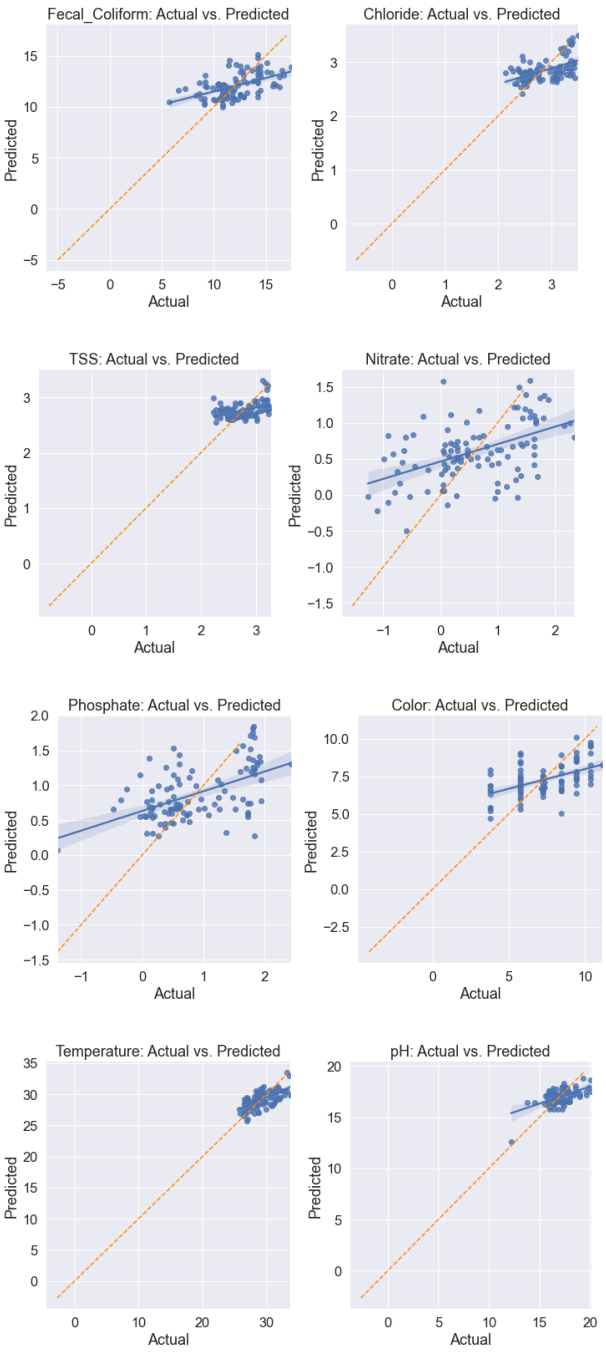
Pollution of the Pasig River. (2020, June 28). Retrieved October 22, 2020, from https://en.wikipedia.org/wiki/Pollution_of_the_Pasig_River

Remote sensing of water quality. Remote Sensing of Water Quality | Wisconsin DNR. (n.d.). https://dnr.wisconsin.gov/topic/lakes/clmn/remotesensing#:~:text=The%20remote%20sensing%20of%20water,dissolved%20organic%20matter%20(CDOM).

Urban Poor Associates, Philippines Case study. Pasig River rehabilitation program. (2004). https://web.archive.org/web/20090605154623/http://www.hic-net.org/document.asp?PID=197.

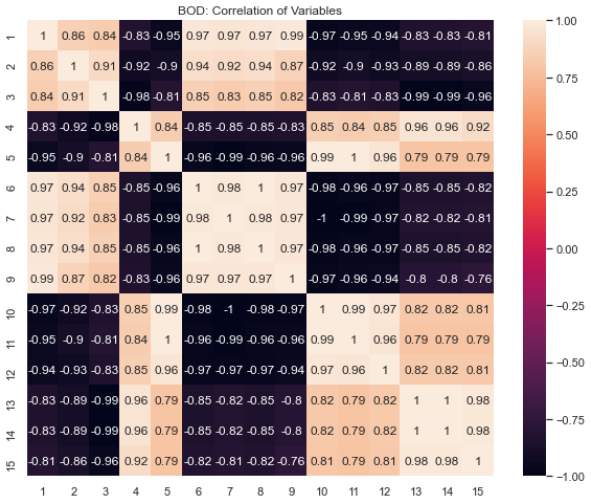
APPENDIX:



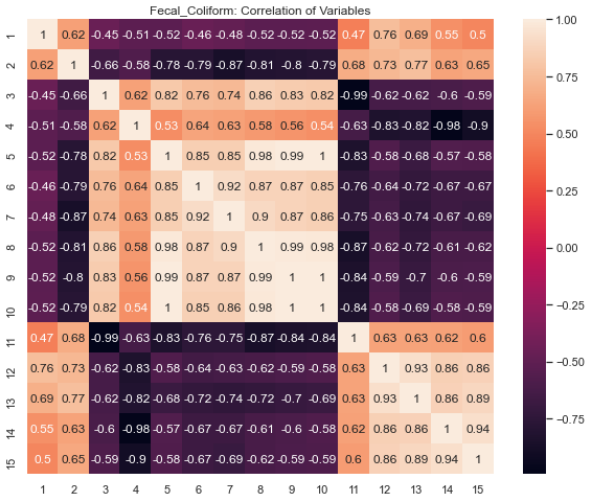


Appendix A. From upper-left to lower-right: BOD, DO, Fecal coliform, Chloride, TSS, Nitrate, Phosphate, Color, Temperature, pH: Shows the *actual vs. predicted* values of model in a scatterplot

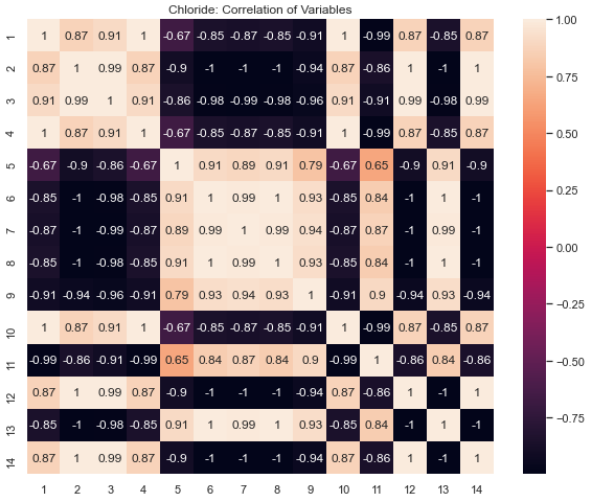
i. BOD



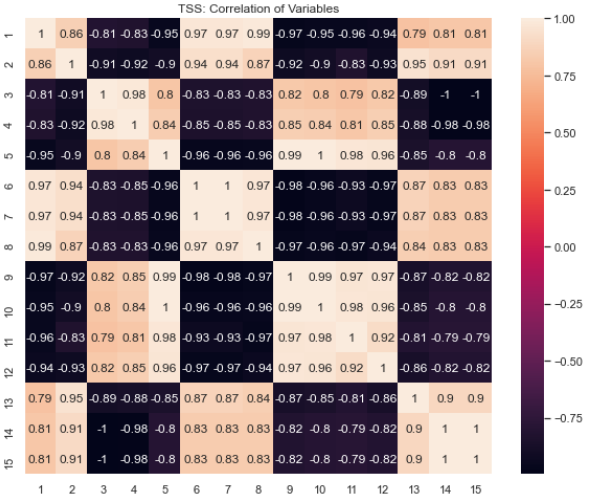
ii. Fecal Coliform



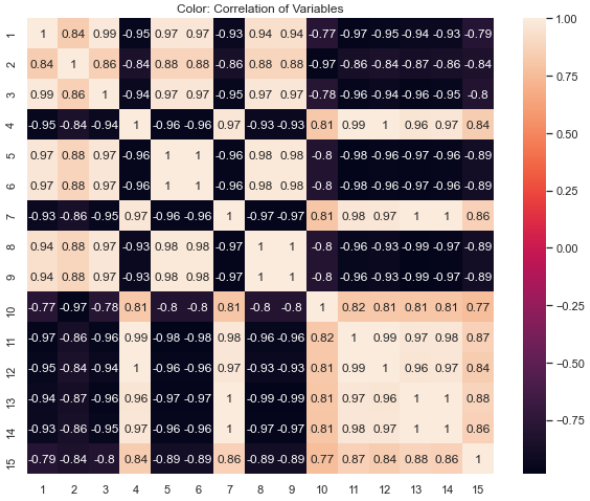
iii. Chloride



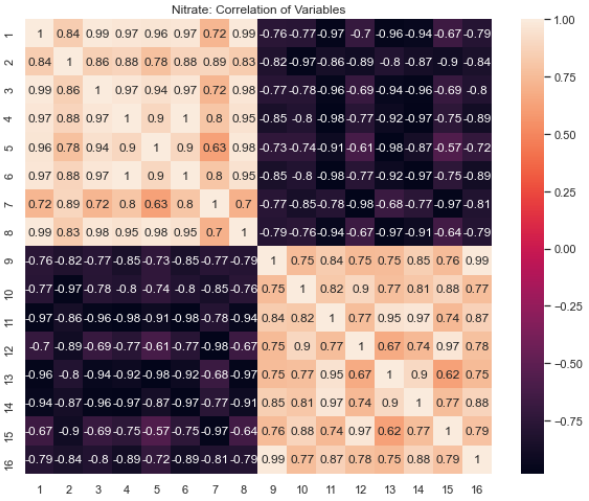
iv. TSS



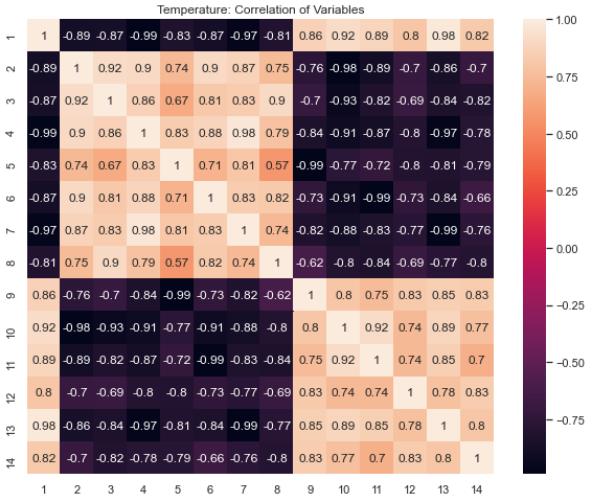
vii. Color



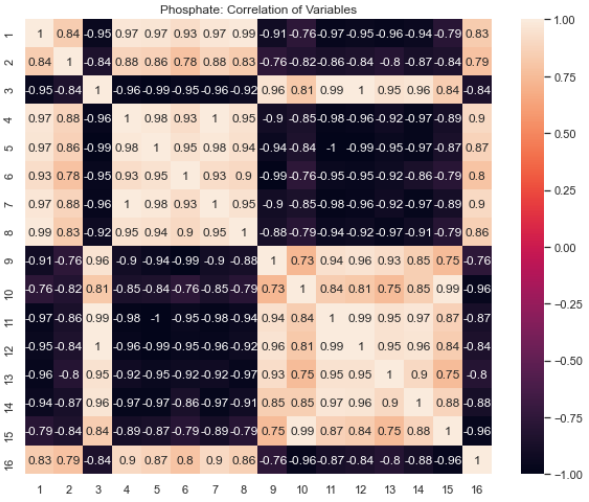
v. Nitrate



viii. Temperature



vi. Phosphate



ix. pH

Appendix B. From (i) to (ix): BOD, Fecal coliform, Chloride, TSS, Nitrate, Phosphate, Color, Temperature, pH: Shows the *correlation of variables* used in model.