A STATISTICAL APPROACH TO UNDERSTAND DIFFERENT BIOFILM PROCESSES

THESIS SUBMITTED

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTERS IN TECHNOLOGY (M. TECH)

IN

ENVIRONMENTAL BIOTECHNOLOGY

PREPARED BY: AVIJIT MOHANTA

EXAMINATION ROLL NO: M4EBT22010

UNIVERSITY ROLL NO: 002030904010

UNIVERSITY REGISTRATION NO: 154531 OF 2020 - 21

UNDER THE GUIDANCE OF

Dr. JOYDEEP MUKHERJEE

&

Dr. RESHMI DAS

School of Environmental Studies under the Faculty of Interdisciplinary Studies Law & Management

JADAVPUR UNIVERSITY, Kolkata-700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled “**A Statistical Approach to Understand Different Biofilm Processes**” is a bonafide work carried out by **Avijit Mohanta (University Roll Number: 002030904010)** under my supervision and guidance for partial fulfilment of the requirement of **Master of Technology** degree in **Environmental Biotechnology** from School of Environmental Studies, Jadavpur University, during the academic session 2020-2022.

……………………………………

THESIS SUPERVISOR

Dr. Reshmi Das
UGC Assistant Professor
School of Environmental Studies
Jadavpur University
Kolkata- 700032

…………………………………….

THESIS CO SUPERVISOR

Dr. Joydeep Mukherjee
Professor & Director
School of Environmental Studies
Jadavpur University
Kolkata- 700032

……………………………………………

DEAN – FISLM
Jadavpur University
Kolkata- 700032

CERTIFICATE OF APPROVAL **

This foregoing thesis is hereby approved as a credible study of an engineering subject carried out and presented in a manner satisfactorily to warranty its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not endorse or approve my statement made or opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

…………………………………………..

Examiner

…………………………………………

Dr. Reshmi Das

Thesis supervisor

** Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undesigned candidate, as part of his Master of Technology degree in Environmental Biotechnology during the academic session 2020-2022.

All Information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by the rules and conduct, I have fully cited and referred all the material and results that are not original to this work.

NAME: AVIJIT MOHANTA

UNIVERSITY ROLL NUMBER: 002030904010

THESIS TITLE: A STATISTICAL APPROACH TO UNDERSTAND DIFFERENT BIOFILM PROCESSES.

SIGNATURE: DATE:

ACKNOWLEDGEMENT

The success of this project required lot of guidance and assistance from my guides Dr. Joydeep Mukherjee and Dr. Reshmi das and I am extremely privileged to get this opportunity to do the project work in their enlightenment. I would like to convey my gratitude to them for giving me all the support and motivation required all along the completion of my project.

I owe my deep gratitude to Dr. Subhajit Datta, who took keen interest on my project work and guided me to learn Python needed for the completion of this project.

I am thankful to Dr. Goutam Kumar Sen. I have learnt basics of statistical modeling approaches from him and deep regards for suggestions to improve project modeling.

I would like to thank my seniors, scholars of School of Environmental Studies for their constant support.

I am fortunate enough to get constant support, encouragement and guidance from my parents and sister in this tough period of Covid 19 pandemic which helped me in successfully complete my project works.

I would not forget to remember some of my close friends Sayantan Singha and Sourav Chakraborty for their encouragement and support till the end of my project.

Last but not the least, to my department School of Environmental Studies and my university, Jadavpur University for providing me all the opportunity to successfully complete my project and thesis work.

This study has indeed helped me to gain more knowledge and explore my potential.

……………………………………………………

AVIJIT MOHANTA

ABSTRACT

In modern days biofilms are extensively used as a tool to assess aquatic environmental conditions. Biofilms are the multilayer assemblage of cells on rocks or solid support surfaces. The support surface can be biologically active or inert. Biofilm carriers are also used in biological wastewater treatments like moving bed biofilm reactors. The interaction among cells and binding forces between the cell support materials may be very complicated. In other words, biofilms are sediment attached multilayer growth of bacteria, protozoa, algae, fungi, diatoms embedded in extracellular polymeric substances. In stagnant biofilm, nutrients diffuse into the biofilm and products diffuse out into liquid nutrient medium. Nutrient and product profiles inside the biofilm are significant factors affecting metabolism, structural and functional endpoints. So, this biofilms in aquatic ecosystem are the effective stress indicator or bio indicator as biofilm communities are the primary ones that are exposed to all the discharge of toxicants through surface run off. When biofilm community structure and growth pattern changes one can identify pollution and overall health of lotic ecosystem. Aquatic biofilm growth can be depended upon several environmental factors and medium characteristics such as suspended matters, dissolved organic carbon, ammonium, nitrate, nitrite and phosphate concentration, water temperature, flow regime, discharge of toxic pesticides and herbicides from the agricultural field through run off. These factors directly or indirectly affect the biofilm characteristics such as dry weight and photosynthetic yield. The purpose of this project was to understand the influence of independent variables on dependent variables through a statistically significant model. This type of modeling can be used for diagnostic task to understand about the various sensitivities as introduced by each of the independent variables in natural networks like biofilm study, wastewater treatment such as moving bed biofilm reactors. In the present study multiple regression analysis approach has been used to understand two different cases of biofilm processes. First study was mainly focused on a statistical approach to understand the effect of temperature, nutrients, herbicide diuron on freshwater riverine biofilms in Morcille river. The second study aimed to understand the inhibitory effect of reject water on nitrite oxidizing bacteria at different biofilm thickness through statistical approaches. Both approaches are statistically significant ($p < 0.01$). The model is built in Python.

INDEX

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION: Pollution is increasing day by day through different means or different pathways. According to Environmental Protection Act (EPA), 1986: a) 'environment' includes water, air and land and human beings and other living creatures, plants, micro-organism and property; b)"environmental pollutant" means any liquid, solid or gaseous substance present in such concentration as may be, or tend to be, injurious to environment; c)"environmental pollution" means the presence in the environment of any environmental pollutant ("CPCB | Central Pollution Control Board" n.d.).

In a natural environmental system, each and every parameter is related to each other. Some parameters have independent effect on some other parameters. Statistical modeling approach is widely accepted to understand or assess the relationship between parameters within a specific natural environmental system.

Previous research studies on statistical modeling approaches are quite helpful to resolve such specific problems to some extent. Regression modelling techniques have many applications in several fields like engineering, data science, trend analysis, forecasting problems, drug response modeling etc.

In data analysis and statistics, we often require to establish a relationship between the various parameters in a data set. This relationship is very important for prediction and analysis. Regression analysis is such a technique. This work mainly highlights the different regression analysis models used nowadays and how they are used in context of different data sets. Selecting the correct model for analysis is commonly the most difficult task and therefore, these models are looked upon closely in this research. While a linear regression analysis model is used to fit linear data, a polynomial regression analysis model focuses on a data set representing polynomial relationship between data variables. Logistic regression model is used in a scenario where someone need a binary type of prediction. When the data set becomes complex, models may suffer from issues like overfitting and underfitting. Ridge and Lasso Regression models are considered the best to deal with such type of situation. Ridge regression is basically used when data suffers from multicollinearity, that is independent variables are highly correlated. Lasso regression uses absolute values in the penalty function, instead of squares. So, in this way Lasso regression differs from Ridge regression. Using these models in the correct way and with the correct data set, Data Analysis and Prediction can produce the most accurate results (Aviral Gupta et al. 2017)

Nowadays, the safety problem in traffic has become the focus of attention and building an effective traffic rules and regulations is of utmost importance. The application of multiple linear regression for the performance analysis of traffic rules mainly discusses the overtaking and lane-changing problem in highway, the differences between traveling on the left and on the right lane. So as to promote smooth traffic effectively, overtaking model is constructed. The changing rule of left lane considers its influence on human heart, and then multiple linear regression models. In the nation with the cars driving on the left, especially that model was suitable for the near S shaped highway. By examining the influence of joint in the radius of curvature, steering angle and friction coefficient on the driving speed of the curve and the obtained speed was within the speed limit. Henceforth, after improving the overtaking model, the left lane vehicle also applied to the lane-changing rule (Qu et al. 2014).

Construction expense of bridges was foreseen by multiple linear regression model, based on data of bridges from Maharashtra state, India. Cost / unit area is taken as a relevant dependent variable. By applying double logarithm as well as conventional regression techniques, models for logarithm of cost/m^2 and cost/m^2 are developed. Total six independent variables, which include both qualitative and quantitative variables, were used to develop the model. Elevation or height of bridge, average span length and depth of foundation were used as quantitative variables. Construction zone, deck type and foundation type were used as qualitative variables in developing model. Strength of these independent variables with dependent variable was found out using Pearson’s correlation approach. Model was then verified using Leave One Out Cross Validation (LOOCV) technique. The best regression model obtained from the experimental data was double log regression with R^2 of 0.850 and a Mean Absolute Percentage Error (MAPE) of 17.74%, as compared to twenty five percent MAPE observed in past for studies related to traditional cost prediction ("Bridge Construction Cost Prediction Using Multiple Linear Regression" 2019).

Multiple regression models are so effective to predict values using predictor variables. The objective of another study was to forecast the global solar radiation in the year of 2019 in the area of East Lima in Peru. Three continuous quantitative predictor variables were analyzed: temperature, humidity, wind speed and the dependent variable was global solar radiation, resulted in a model with excellent significance p less than 0.001 that shows the prediction was significant and effective. The multiple linear regression method was used, found a mean global radiation of 175 Watt/m^2 and explanatory variables with mean temperature of 19.2 °C, humidity 23.9% and wind speed 1.77 m/s, with the maximum temperature in summer recorded at 24.6°C, the highest humidity of 51.2% in autumn, maximum wind speed in summer at 2.63 m/s and the highest global solar radiation in spring with 183 W/m^2 (Soria et al. 2022).

Yet another research attempted to decide and model the performance of short columns under different ground levels and storey number. This behavior includes varying axial forced, shear forces, and bending moments under gravity loads. Results have been taken at the ground level for the column when the process of the short column has appeared. To study the effectiveness of short columns under different ground levels and the number of stories, the three-dimensional finite element by the software sap2000, and multiple linear regression analysis have been used in this study. The findings showed that the axial force for all types of short columns (internal, edge, corner) increases as the number of stories in the building increases at the same ground level. The effect of changing the number of floors with the same value of ground level is not significant in shear and moment for all types of columns except for the one-story and two-storey buildings. The moment and shear for a building composed of one storey was high compared to the two-storey building where the value was significantly low; then became higher again with a 3-storey building and with more stories (Abdulwahid et al. 2020).

Another study was aimed to establish a predictive model to determine the rate of generation of municipal solid waste in the municipalities of the Sumidero canyon in Mexico. Multiple linear regression was used with demographic and social explanatory variables. The compiled database consisted of nine variables with one hundred and eighteen specific data per variable, which were analyzed using a multicollinearity test to select the most important ones. Initially, different regression models were generated, but only two of them were considered useful, because they used few predictors that were statistically significant. The most significant variables to predict the rate of waste generation in the study area were the population of each municipality, the migration and the population density. But some other variables, such as daily per capita income and average schooling were so crucial, they do not seem to have an effect on the response variable in this study. The model with the biggest closeness resulted in an adjusted coefficient of 0.975, 7.70% average absolute percentage error, an average absolute deviation of 0.16 and an average root square error of 0.19, showed a high influence on the phenomenon studied and a good predictive capacity.(Araiza-Aguilar, Rojas-Valencia, and Aguilar-Vera 2020)

Another study showed that exposure to atmospheric air pollution has been linked to a number of health outcomes, starting from modest transient conversions in the respiratory tract and impaired pulmonary function, continued to restrict performance and to the increase emergency rooms visits, hospital admissions or mortality. Increase in the allergenic symptoms has been connected with air contaminants such as ozone, particulate matter, fungal spores and pollen. In view of the potential importance of crossed effects of non-biological pollutants and airborne pollens and fungal spores on allergy worsening. This work was aimed to study and evaluate the influence of non-biological pollutants like O_3 and PM10 and meteorological parameters on the concentrations of pollen and fungal spores using multiple linear regressions. The information considered in this study were collected in Oporto, the second largest Portuguese city, located in the North. Daily mean of O3, PM10, pollen and fungal spore concentrations, temperature, precipitation, relative humidity, wind velocity, pollen and fungal spore concentrations, for 2003 to 2005 were considered. Results showed that the 90th percentile of the adjusted coefficient of determination, P90 (R^2adj), of the multiple regressions varied from 0.613 to 0.916 for pollen and from 0.275 to 0.512 for fungal spores. Ozone and PM10 showed to have some influence on the biological pollutants. Among the meteorological parameters analyzed, temperature was the one that most influenced the pollen and fungal spores' airborne concentrations. Also, the relative humidity showed to have some influence on the fungal spore dispersion. Nevertheless, the models for each pollen and fungal spore were different depending on the analyzed period, which means that the correlations identified as statistically significant cannot be, even so, consistent enough (Sousa et al. 2010).

Forecast of academic performance of the students helps the professors to build the clear understanding of student's community and take the healthy measures to make their learning smooth, comfortable and understandable. Many data analytics techniques can be used to forecast the graph of performance of the students in academics. But the main objective of the study was to use linear regression approaches to build a model which forecasts the performance of the engineering students. The independent variables of the model contained how many hours spent on the internet in some activities based on the data collected. The target or dependent variable is the prediction of end semester examination grades i.e., CGPA (Cumulative Grade Points). Multiple measures were used to estimate and substantiate the models that were predicted along with the percentage of good predictions. The results show that the predicted model gives the better accuracy in prediction (R R et al. 2019).

Another research has done to estimate of irrigation demand which is an important component for managing the water significantly in the canal command area. The irrigation water requirement of a command area mostly depends on climate, the nature of the crop, and the soil where the crop is grown. Detailed calculation of water demand at the Water User's Association and minor canal level holds the pointer for effective management of the water in the command area. In the command area of Wazirabad, farmers undertook rice cultivation in two crop seasons Kharif and Rabi. The usage of water in Rabi Season is considerably high because of dry weather situations. Irrigation requirements were formulated and projected from the daily water demand to ten-day periods. The estimated demand in two crop years for Rabi season in the year 2007 and 2009 for rice crop reviewing the daily meteorological data and the results were used as an associate input to regression model to formulate an equation. Multivariate regression analysis was run for all kinds of soils in the command area. The model is useful to compute the demand for rice crop mostly for Rabi crop under semi-arid conditions under different soil textures (Mittapalli and Chalumuri 2014).

Large pharmaceutical companies require to innovate by registering new artificial intelligence and machine learning approaches to understand the large datasets produced by high-output technologies. In addition to reduce development costs for these industries, regression and classification models of drug response was needed for the final quest of delivering personalized treatment for cancer. In this contribution they presented results obtained by symbolic regression.

They employed a public domain data of drug responses on a huge cancer cell line panel and compared with a previous method based on equivalent of the response data and the use of integer linear programming and found logic models. Presented derived models of drug response for the drugs Dactolisib (BEZ235), Afatinib, Paclitaxel, Cytarabine, Paclitaxel as well as for JQ12, AZD6244, KIN001-102, and PLX4720. The interpretability with a biological analysis of the results for Afatnib and Dactolisib was provided, showing that the models introduced variables that point at known mechanisms of action of these drugs (Fitzsimmons and Moscato 2018).

From ages Cape Town has been affected by water shortages and it was be assumed that the scenario will be aggravated in the coming decades by a growing population, economic development and climatic changes as additional stress factors. In order to disarm the situation, Cape Town has commissioned varieties of feasibility studies concerning the implementation of alternative water sources, with as yet unpublished conclusions. Since sustainable water resource planning requires a comprehensive understanding of the water demand, the objective of the study was to forecast the future demand by the city of Cape Town by analyzing its significant drivers. For this objective, a multiple linear regression analysis was applied on parameters which influence water demand, namely: economy, population, losses of water and water restrictions. In order to generate the linear multiple linear regression model and its regression coefficients, documented historic data was used for the period 2001 to 2012. The result of the analysis showed that the water demand of the city of Cape Town was only influenced by population and water losses. Additionally, the model suggested that a new source would be required by 2021. Thus, water conservation and water supply strategies can be adapted accordingly to ultimately enable a sustainable management of the water sources in the city of Cape Town (Lawens and Mutsvangwa 2018).

Heterogeneous nature of soils and difference in its hydraulic conductivity over several orders of magnitude for multiple soil types from coarse-grained to fine-grained soils, predictive models to calculate hydraulic conductivity of such soils from different properties considered more easily obtainable have now been given an appropriate consideration. Another study formulated the execution of artificial neural network (ANN) being one of the powerful computational intelligence methods in forecasting hydraulic conductivity of varied range of soil types and compared with the conventional multiple linear regression (MLR). MLR and ANN models were developed using six input variables. Results showed that three input variables were statistically significant in multiple regression model development. Evaluation of the performance of these developed models using coefficient of determination and mean square error showed that the forecast capability of ANN is far better than MLR. Additionally, relative study with available existing models showed that the developed MLR and ANN in this study performed relatively better (Williams and Ojuri 2021).

Another research introduced soil strength model by multiple linear regression developed from electrical resistivity and seismic refraction tomography. The multiple linear regression approach was used to calculate the value of dependent variables of soil strength formulated on the basis of the value of two independent variables, namely and velocity and resistivity values. These parameters were regressed using regression analysis technique for generating multiple linear regression model. The analyzed model results of MLR model which was based on estimation of model dependent parameters (resistivity and velocity) calculated for P-value at significance level of 0.05 is 0.01572 and 0.01163, for soil's cohesion parameter and 0.01966 and 0.02534, for soil's angle of shearing resistance or friction angle parameter. 2 multiple regression model equations were formulated from that statistical analysis. The prediction accuracy of the multiple regression model was conducted for verification on the second stage. Depending on these statistical analysis results, a new soil's strength model from geophysical data set for near surface study was developed. The soil's strength models developed using MLR is reliable to image the subsurface in two-dimensional form, which covered more region compared to traditional method (Bery 2021).

Riverine biofilm communities are the primary ones to be exposed to any or all harmful toxic discharges received via break out from agricultural fields. Hence, changes in river biofilm community structure and growth pattern are thought of as indicator of overall health of lotic ecosystem. Toxicants have effect on biofilm biomass, photosynthetic potency and chlorophyll *a* concentration. Statistical models may be applied to estimate the overall health of riverine ecosystems considering biofilms as indicators. Here, previous empirical data of Ricart et al. 2009 on long run effects of environmentally relevant concentrations of diuron on biofilm communities of the stream Llobregat, Kingdom of Spain was considered as model inputs. The objective was to understand the influence of diuron, chlorophyll *a* concentrations and photosynthetic potency or efficiency on biovolume using a statistical model. The non-linear relationships between biovolume (dependent variable) and diuron, chlorophyll *a* concentrations and photosynthetic efficiency (independent variables) were delineated by constructing 3 separate basis functions based on day eight empirical data. Biovolume, attributable to nonlinear influence as yielded by the basis functions were used in a multiple linear regression model to estimate the net biovolume. Model validation was done based on day twenty-nine empirical data. Experimentally observed biovolume and the model estimated biovolume showed similar inclination. Also, diuron and photosynthetic potency had significant ($p < 0.05$) influence on biovolume. Since, the predominance of diatoms as biofilms within periphytic layers is so common in lotic systems, estimation of changes in diatom biovolume was significant to assess the effect of herbicides. Diatom biovolume of any day (for example day twenty-two) mentioned in the experimental study may be determined by this model, without the requirement of monotonous and tedious manual biovolume calculation. This model will be useful in number of other studies undertaken on the toxic effect of pollutants on biofilms to quickly and accurately estimate the biofilm biovolume (Bhowmick et al. 2021).

In (Bhowmick et al. 2021) study, model is based on experiment where the light intensity, irradiance level and temperature were fixed all along the study period, but practically it may not be fixed due to seasonal variations. So, for the deeper insight to this biofilm model, we propose here a statistical model to assess how changes in temperature, nutrient levels, and diuron concentrations affected the riverine biofilm dry weight and photosynthetic yield. Where we use a multiple linear regression model followed by polynomial basis. Polynomial basis functions were introduced to improve the linearity and accommodate the actual or most probable non-linear relationship between independent and dependent variables.

Chapter II

Basic concepts of Regression Analysis and Modeling

Linear regression models the relationships between at least one explanatory variable and one target variable. These variables are known as independent variable and dependent variable respectively. When the number of independent variables become more than 'one' then the regression analysis is termed as multiple linear regression.

Linear regression analysis has done to draw the relationships among dependent and independent variables and to understand it. Another purpose of this regression is to forecast the dependent variable.

Scatter Diagram: When data relating to the simultaneous measurement on two variables are available, each pair of data can be geometrically represented by a point on the graph paper. The values of one variable being shown along X-axis and another along Y-axis. If there are n-pairs of observation then we get n points in the graph paper. This types of diagrammatic representation of bivariate data are known as Scattered Diagram or Scattered Plot.

Scatter Plots provide a visual way of understanding if there is a correlation between two variables or not, is the correlation is weak or strong? And is it positive or negative?("Total Quality Management for Custodial Operations: A Guide to Understanding and Applying the Key Elements of Total Quality Management" n.d.).

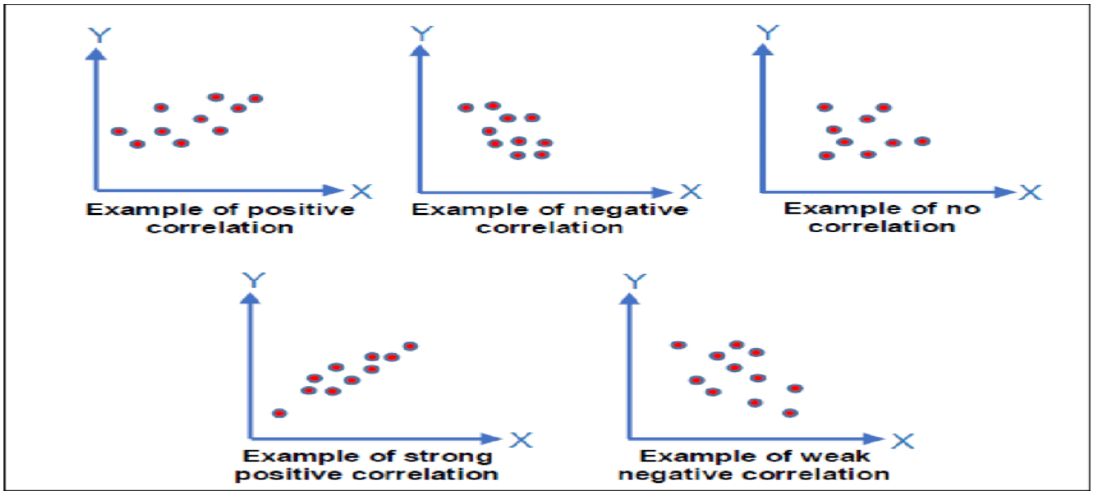

Y
X
Example of positive correlation
Y
X
Example of negative correlation
Y
X
Example of no correlation
Y
X
Example of strong positive correlation
Y
X
Example of weak negative correlation

Figure1: The interpretation of the scattered plots.(Charantimath 2011).

Correlation coefficients:(Mun 2014)

a) Pearson's product moment linear correlation: When we typically use the term correlation, we normally mean a linear correlation. The values of correlation coefficient ranges from -1 to +1. It signifies that correlation coefficient has a sign or direction and a value or magnitude. If the correlation coefficient(r) closer to the value of +1, then we can say that strong positive correlation exists between two variables. That means increase in the value of X increases the value of Y. If the correlation coefficient closer to the value of -1, then we can say there is strong negative correlation exists. That means, increase in the value of X, decreases the value of Y.
r = +1(if positively sloped), indicates perfect positive correlation.
r = -1(if negatively sloped), indicates perfect negative correlation.
r = 0, where the correlation curve is perfectly flat indicates the zero correlation.

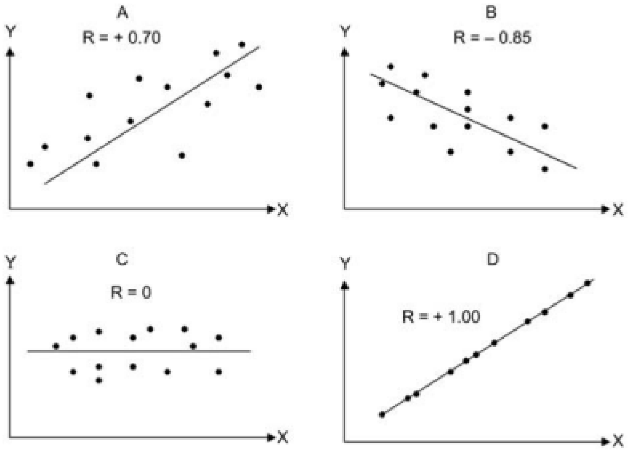


Figure 2: graphical interpretation of Pearson's correlation coefficient. (A): strong positive correlation between X and Y. (B): strong negative correlation between X and Y. (C): Zero correlation between X and Y. (D): perfect positive correlation between X and Y.

b) Spearman's non-linear rank correlation: In many real-life situations the relationship between two variable is not linear. For example, if two variables have a exponential relationship but we calculate Pearson's correlation then this gives us wrong interpretation. Again, if we visualize sinusoidal relationship between two variables with

Pearson's coefficient it gives us flat correlation curve i.e., zero correlation. To visualize the non-linear relationship, Spearman's non-linear rank correlation is used. Therefore, we must distinguish between linear and non-linear correlations.

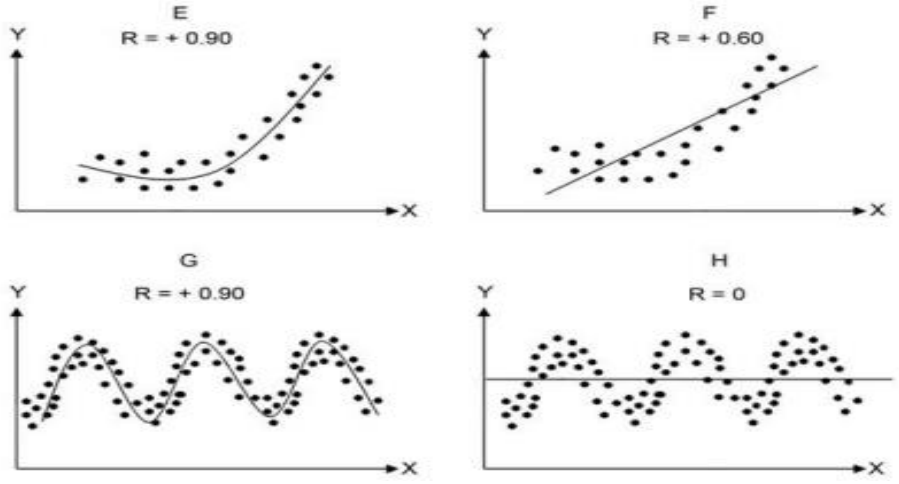


Figure 3: linear vs. nonlinear correlations.

Before modelling a regression problem, we have to visualize the relationships through correlation coefficients.

Table 1: Level of correlation between variables.(Şanli 2019)

Modulus value of 'r'	Level of correlation
0.00 – 0.25	Very low
0.26 – 0.49	Low
0.50 – 0.69	Moderate
0.70 – 0.89	High
0.90 – 1.00	Very high

Different types of Regression problems: (Aviral Gupta et al. 2017)

a) Linear regression: The simplest regression technique is linear regression.
Linear regression equation in its simplest form (i.e., one dependent variable and one independent variable):

$Y = a*X + b + \varepsilon$ (1)

Here Y is the dependent variable, X is the independent variable, ‘a’ is the coefficient of X, ‘b’ is the intercept and ‘ε’ is the error term.

Intercept ‘b’ indicates the value of dependent variable when the independent variables become zero.(Sarstedt and Mooi 2014)

To find ‘a’ and ‘b’ the equation of line formed satisfies the scattered points efficiently.

Mean Absolute Percentage Error of Prediction = modulus of $(\frac{Yactual-Ypredicted}{Yactual}) * 100$

b) Logistic regression: In this type of regression, the dependent variable is binary or it has two values. It has values like True/False or 0/1 or Yes/No. This type of modeling approach is used to determine the chance whether an outcome is depends on one or more independent variables. It uses the logistic function to search the association between the dependent and independent variables. It has similarity with linear regression and it is the exceptional case of normal linear regression model. But there are some significant differences between these two: a) In logistic regression conditional distribution follows Bernoulii’s distribution not the Gaussian distribution. b) Here predicted outcomes are probabilities determined through logistic functions and value ranges between 0 ad 1.

General formula of logistic regression is:

Odds = P/(1-P) = Probability that an event occurs/Probability that an event does not occur.

$\ln (Odds)= b_0 + b_1*X_1 + b_2*X_2 + \ldots\ldots\ldots\ldots + b_n*X_n$ (2)

c) Polynomial Regression: Here the relationship between the independent variable X and dependent variables Y is modelled as an n^{th} degree polynomial of X. Polynomial regression fits a non-linear model to the statistical data.

The nth order polynomial regression equation in one variable is as follows:

$Y = a_0 + a_1*X + a_2*X^2 + \ldots\ldots\ldots\ldots + a_n*X^n + \varepsilon$ (3)

Here a_0 is the intercept, a_1, a_2, …….., a_n are the regression coefficients. ‘ε’ is the error term.

d) Ridge regression: Linear Regression or polynomial regression modeling might lead to an overfitting condition; therefore, Ridge Regression is used to minimize the overfitting condition by adding a new factor to the least square objective of Linear Regression.

e) Lasso regression: Lasso Regression is used to minimize the overfitting condition by adding a different factor to the least square objective of Linear or Polynomial Regression, the factor added is proportional to the absolute sum of the coefficients.

Overfitting and Underfitting:(Minhas 2021)

Underfitting occurs when the selected model has a very high bias and is unable to highlight the complex patterns in the data. This leads to higher training and validation errors because the model is not complex enough to classify the underlying data. So, the prediction error is greater in this case. In this scenario one need to increase the increase the model complexity either by increasing the order or by increasing the number of parameters.

Overfitting condition arises when the selected model becomes very complex and even captures the noise in the data. The model prediction error is too low but it cannot accommodate or generalize new types of data or additional data types.

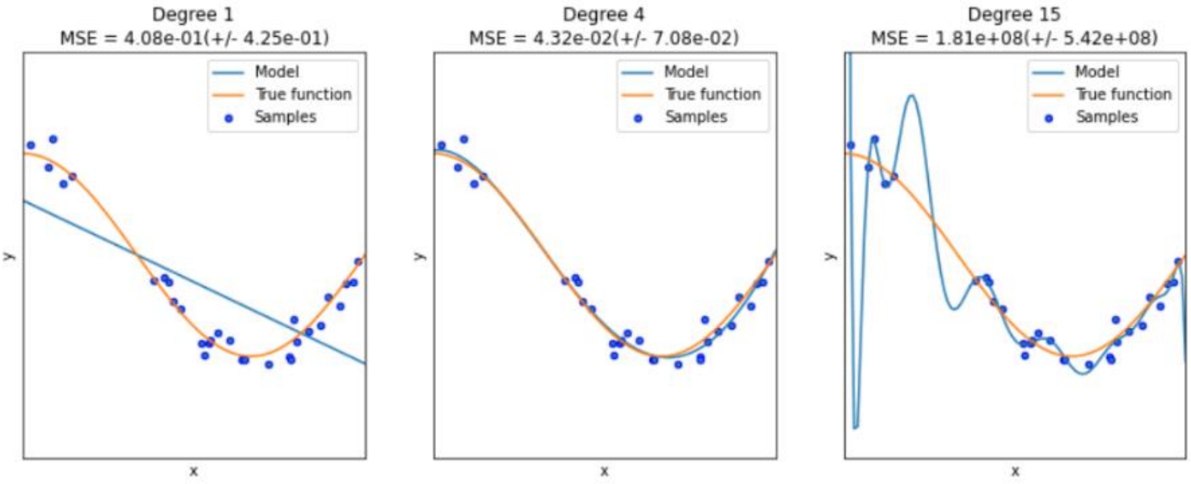


Figure 4: Degree vs Curve fitting. Degree 1 model is underfitted; Degree 4 model is good fitted and Degree 15 model is overfitted. ("Underfitting and Overfitting in Machine Learning" n.d.)

So, the performance of the selected model is poor either due to underfitting or overfitting. Therefore, we have to find out the best or good fit condition of modeling. (Aviral Gupta et al. 2017)

Least Square Method: Method of least squares provides an elegant procedure for fitting a unique curve to a given data set. The curve of best fit is that where the sum of the square of the residuals is minimum. However, the principal of least squares does not help us to determine the form of the appropriate curve. It only determines the best possible values of constants in the regression equation when the form of the curve is known beforehand.

Optimum order of the model: The order of the polynomial model should be kept as low as possible. If the lower order model is not satisfactory the higher order model can be applied but arbitrary choosing the higher order polynomial can be a great abuse of regression analysis. Forward selection approach or backward elimination approach can be used to find the optimal degree of polynomial. One possible method is to successively fit the models in increasing order and test the significance of regression coefficients at each step of model fitting. Keep the order increasing until t -test for the highest order term is nonsignificant. This is called a forward selection approach. The second approach is to fit approximate highest order model and used to reduce the highest order terms one at a time. This process is continued until the higher order remaining term has a significant t-statistic. This is termed as backward elimination approach ("MTH 416 : Regression Analysis" n.d.).

Scoring measures of Regression problem:

We can visualize the overall model fit through coefficient of determination (R^2) value and significance of F-statistics. The coefficient of determination indicates the degree to which the model justifies the observed variation in the dependent variable relative to the mean.

$$R^2 = \quad 1 - \frac{residual\ sum\ of\ squares}{total\ sum\ of\ squares}$$

The value of coefficient of determination lies between 0 to 1. Higher value of coefficient of determination indicates better fit. The adjusted coefficient of determination is a relative measure and it is used to compare different models. The p-value of the F-statistic of below 0.01 or 0.05 does not directly mean that each and every variable is statistically significant. But if the F-stat is significant then it is strongly likely that at least one or more regression coefficients are significant. When we interpret the model F-test is most critical because if the F-stat is insignificant, we cannot interpret the model further. After that we have to interpret the individual independent variable. First, we look at the t-values reported for each individual parameter. If regression coefficients p-value is below 0.05 then we can say that independent variable relates significantly with the dependent variable (Sarstedt and Mooi 2014).

Let S_j, j = 1, 2, ………., N denote the 'N' validation data prediction scores derived for some prediction method (e.g., multiple regression or artificial neural network) built using the training data set. Further, let A_j, j = 1, 2, ……., N denote the ‘N’ actual values of the target variable contained in the validation data set. Let $E_j = A_{j\ -}S_j$, j = 1, 2, ………., N denote the ‘N’ scoring (prediction) errors (over the validation data set) associated with the chosen prediction method.

The following are common scoring measurement technique used to calculate the prediction error:

- Mean absolute error (MAE):

$$\frac{\sum_{j=1}^{N} |Ej|}{N}$$

- Mean square error (MSE):

$$\frac{\sum_{j=1}^{N} Ej\hat{}2}{N}$$

- Mean absolute percentage error (MAPE):

$$\frac{\sum_{j=1}^{N} \frac{|Ej|}{Aj}}{N}$$

- Root mean square error (RMSE):

  RMSE = (MSE)^0.5

- Mean square percentage error (MSPE):

$$\frac{\sum_{j=1}^{N} (\frac{Ej}{Aj})\hat{}2}{N}$$

- Root mean square percentage error (RMSPE):

  RMSPE = (MSPE)^0.5

  In case of percentage errors, one has to be multiplied those fractional results with hundred.

The mean absolute percentage error (MAPE) also known as the mean absolute percentage deviation (MAPD) is a measure of prediction accuracy of the forecasting method in statistics.

The residual standard deviation or residual standard error is a measure used to assess how good will a linear regression model fit the data. Square root of sum of square of residuals divided by the residual degrees of freedom. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response.

Chapter III

Case Studies

Case I: This study is mainly focusses on a statistical approach to understand the effect of temperature, nutrients, diuron on freshwater riverine biofilms. Biofilms are surface-adhered, structured microbial communities composed of sessile cells (bacteria and/or fungi) embedded in a self-produced extracellular polymeric matrix contained DNA, polysaccharide and other components.(Coenye 2013). In other words, biofilms are sediment attached multilayer growth of bacteria, protozoa, fungi, algae, diatoms embedded in extracellular polymeric substances.

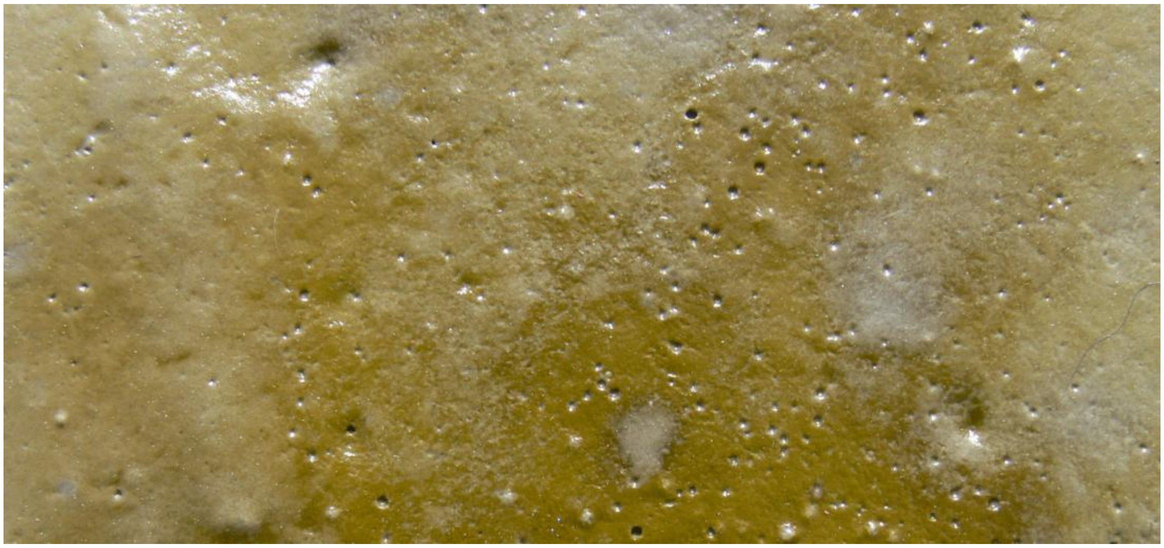
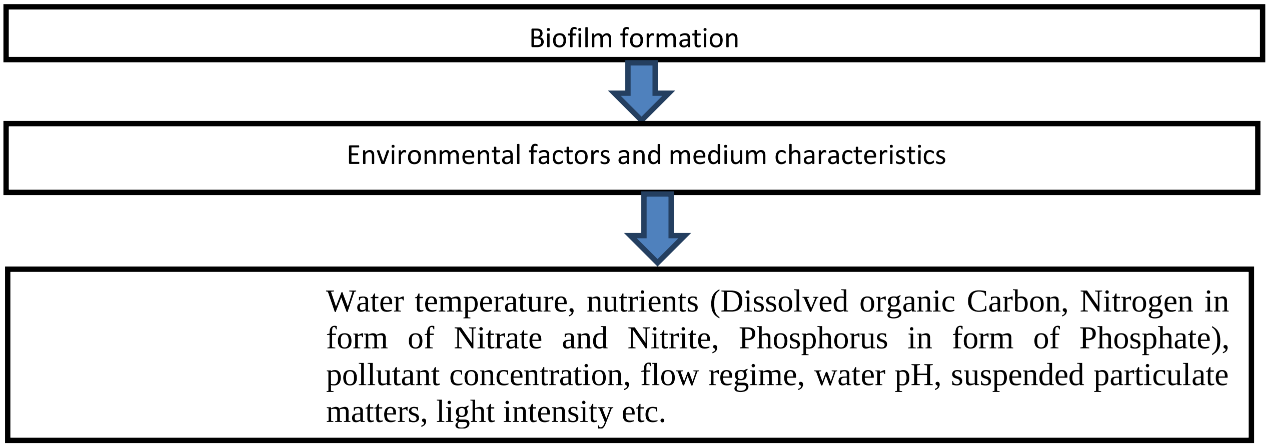


Figure 5 : Marine biofilm (“Marine Organisms Enlisted in Battle against Bacterial Sheets | Research and Innovation” n.d.)

Now the growth of riverine freshwater biofilms is influenced by several biotic and abiotic factors such as water temperature, nutrients, pollutant concentrations etc.



Biofilm formation

Environmental factors and medium characteristics

Water temperature, nutrients (Dissolved organic Carbon, Nitrogen in form of Nitrate and Nitrite, Phosphorus in form of Phosphate), pollutant concentration, flow regime, water pH, suspended particulate matters, light intensity etc.

River biofilms are the effective environmental stress indicator in the aquatic ecosystem and several endpoints have been considered to count the biofilm adverse condition such as photosynthetic potency, biomass growth quantified by dry weight, chlorophyll content and species diversity (Sgier et al. 2018). The biofilms involved in nutrient fluxes as well as dynamics of contaminants. Biofilms can accumulate organic contaminants and metals transported by the flow of water and/or adsorbed onto substrates. Again, due to biofilms high metabolic activity and their role in aquatic food chains, microbial biofilms are also likely to stimulate contaminant fate in aquatic ecosystems (Bonnineau et al. 2021).

This study describes a field-data oriented model where these physicochemical factors were considered to evaluate the effect on overall growth or physiological condition of river biofilms. For the statistical modeling nine independent parameters including water temperature, suspended matter, dissolved organic carbon, nutrients (NH_4, NO_2, NO_3, PO_4, Si) and toxicant (herbicide diuron) concentration and two dependent parameters such as photosynthetic yield and dry weight of biofilm were considered.

Microbial biofilm characteristics are influenced by a) physicochemical conditions such as nutrient availability (DeLorenzo, Scott, and Ross 2001), temperature (Diaz Villanueva et al. 2011), light regime (Guasch and Sabater 2002), stream current or flow velocities (Villeneuve, Montuelle, and Bouchez 2010), toxicity of pollutant exposure and b) biotic factors such as interaction between microorganisms or predation by grazers (Muñoz et al. 2001).

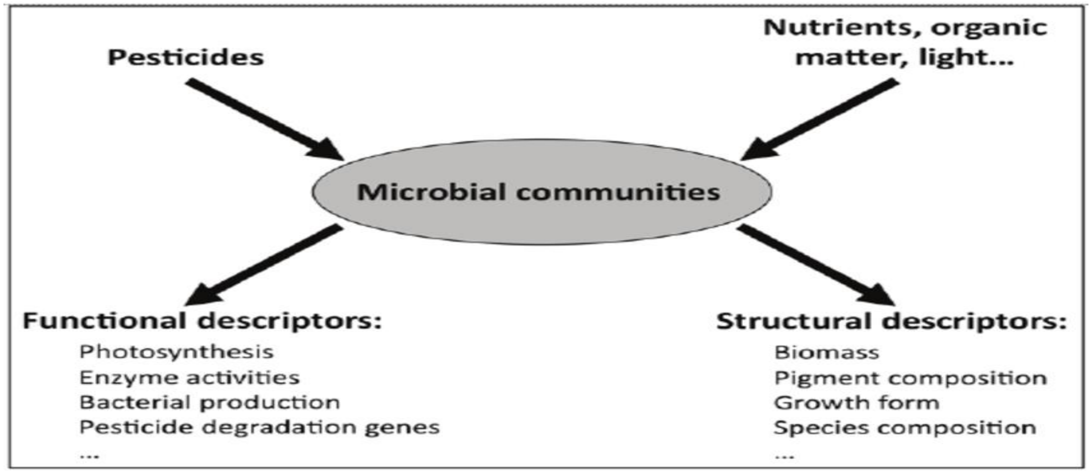


Figure 6: Flow chart to depict some causes of microbial communities change. (Sabater et al. 2007)

Influences of several independent variables on biofilms in aquatic ecosystems:

- Water temperature is vital factor affecting bacterial abundance and film growth (White et al. 1991). In general, an elevated temperature should result in higher metabolic activity of biofilm and thereby increase the growth rate within a specific range (Russell et al. 2013). Warmer temperature also resulted in an increase of biofilm biomass in rocky intertidal regions. Bacterial and algal species of biofilm showed growth with a rise in ambient water temperature and nutrient availability. (Melo and Pinheiro 1992) demonstrated a substantial increase in film thickness with increasing temperature till 45˚C after which there was a certain decline.
- Light intensity is a vital factor as limiting light decreases biofilm growth even in presence of nutrient media and higher intensity of light generally results an increase in film biomass and chlorophyll a concentration and bacterial count (Wagner et al. 2015).
- Evidences suggests that rivers may be N or P limited when not impacted by human activities (ORD US EPA 2015). Considering a typical molar C: N: P – 50: 10: 1 in bacteria. The community responses to nutrients are varied with seasons (Chénier et al. 2003). C (334) : N (28) : P (5.6) ratio enhance the potentiality to grow more complex biofilms and significantly showed higher counts of attached cells (Thompson et al. 2006).
- Water pH is also playing a vital role. Global average sea surface pH is 8.21 (OW US EPA 2016). Attachments of diatoms to hard surfaces was inhibited by lowered pH, however, it was favored in alkaline conditions having pH above 7 (Sekar et al. 2004).
- Toxicant - Herbicide diuron (Liu 2010) :

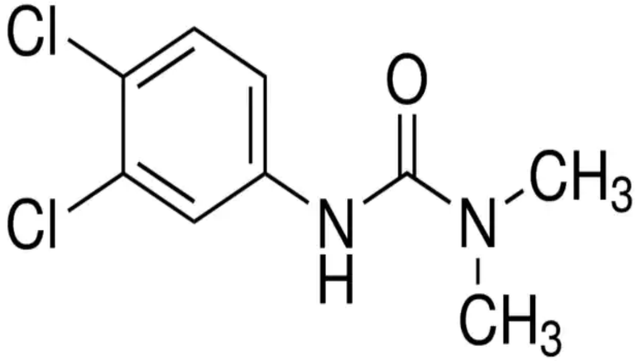


Figure 7: Molecular structure of Diuron (Fernández-Cori et al. 2015).

Commercial names are DCMU, Duran, Dynex, Herbatox, Vonduron, Dichlorofenidim, Karmex, Duirol, Unidron etc.

Molecular formula is $C_9H_{10}Cl_2N_2O$

Molecular Weight: 233.09

Diuron is a white crystalline solid or whitish powder which is odorless as well. Its boiling point is 356 to 374˚F at 760mm Hg. It decomposes at 180-190˚C. It is soluble in water, solubility: 42mg/L at 25˚ C. Diuron is a very common herbicide and it is used globally. Diuron or DCMU (3-(3,4-dichlorophenyl)-1,1-dimethylurea) is an algicide and herbicide, inhibits photosynthesis and frequently detected in freshwater ecosystems (Sgier et al. 2018). Diuron is included in the list of major pollutants in the European Union Water Framework Directive (WFD, 2000/60/EC) (Ricart et al. 2009). Before its ban in France in December 2008, diuron was commonly used in agricultural and urban environments causing acute contamination of surface water, predominantly in small streams draining wine-growing areas (Louchart et al. 2001).

Previously a statistical model was developed to assess the effects of diuron on river biofilm community (Bhowmick et al. 2021) from empirical data of (Ricart et al. 2009). Light intensity, temperature, and nutrient were kept constant throughout the study but not applicable in the real sense. Hence in this study the seasonal variations were not considered. To bring a new insight, we propose here a statistical model to assess how changes in temperature, nutrient levels, and diuron concentrations affects the riverine biofilm dry weight and photosynthetic yield. This model is based on the published (S, C, and B 2010; Pesce, Margoum, and Foulquier 2016) and associated unpublished metadata from a field survey conducted form September 2008 to December 2011 in the river Morcille of France. Monthly measurements of water temperature and other physicochemical parameters of river water and biofilm parameters were carried out that time.

This statistical model explained the interaction between toxicant concentrations and physicochemical parameters of the stream environment leading to the understanding of the dynamics of biofilm functional and structural status.

Methodology:

1. **Study area**: The Morcille river is the study site, located in the French Beaujolais area of eastern France (Latitude 46˚15' N; Longitude 4˚60' E). The river watershed is characterized by the presence of vineyards, which are significant sources of water toxicants, especially pesticides and their residues.

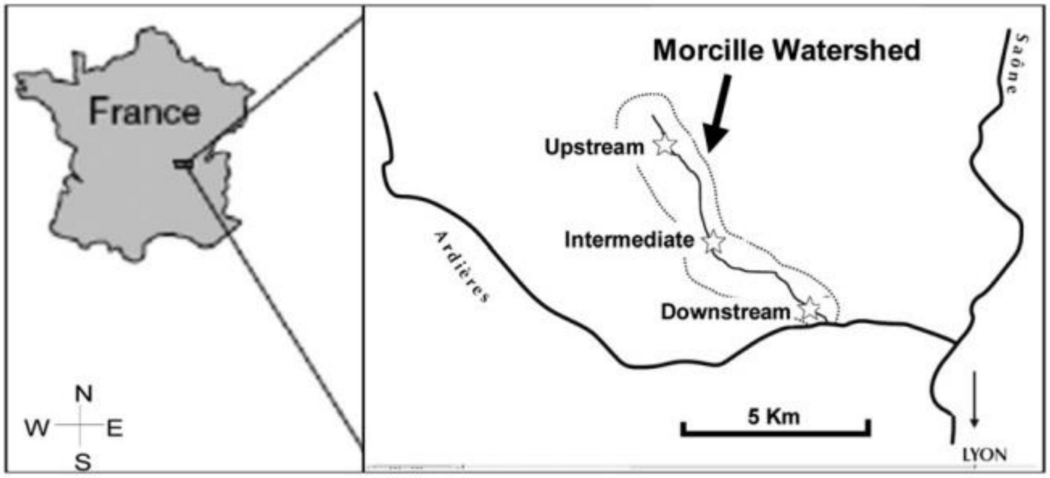


Figure 8 : location of sampling station along the river Morcille (S, C, and B 2010).

Biofilms and water samples were collected from intermediate and downstream stations draining watershed areas of which 52% and 72% of the total watershed area (approximately 8.5 square kilometers) were vineyards respectively. These two stations are 4 km. apart and characterized by similar physical conditions such as depth, light, climate etc. Temperature was measured twice a month at 10:00 a.m. at two sampling area. At the same time water sample was collected and dissolved organic carbon, nitrate, nitrite, ammonium, phosphate concentrations were measured by the standard procedures. Diuron concentrations in water phase was determined by the solid phase extraction on Oasis HLB cartridges followed by liquid chromatography tandem mass spectrometry. At the lab, biofilms were cautiously removed from each replicate slide by razor blade and then suspended in mineral water to give an ultimate concentration of 1cm^2 of biofilm/ millimeter of suspension. 25ml of aliquot sample was used to measure the dry weight after two hours of drying at 105˚C (S, C, and B 2010).

Biofilm parameters such as chlorophyll a, PS yield, dry weight and tolerance of diuron were measured every month. Water temperature, suspended matter (SM), dissolved organic carbon (DOC), nutrients such as ammonium, nitrate, nitrite, phosphate, silicon and diuron concentrations in water were measured every two weeks and were expressed as monthly average. However due to high flow conditions, biofilm collection was not possible in September 2008, December 2008, December 2010 and November 2011 at both intermediate and downstream stations.

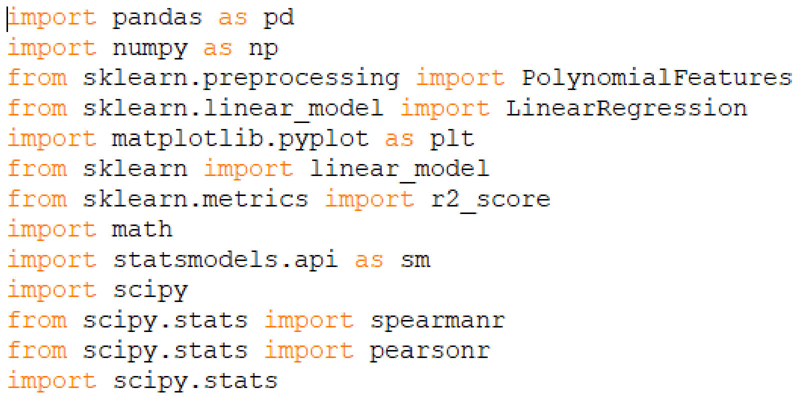
Model Development: The objective of this study was to assess the influence of certain independent environmental parameters (namely water temperature, pollutant concentrations, Suspended matters i.e., SM, dissolved organic carbon i.e., DOC and nutrients) on biofilm growth. The developed model is primarily a polynomial regression model, where parameters of the model were estimated using least square method. In this modelling approach biofilm growth was quantified by two parameters, PS yield and dry weight of the biofilm (dependent variables). Water temperature, diuron concentration, nutrient concentrations (NH4, NO2, NO3, PO4, Si), DOC and SM were considered as independent variables. Data collected from the field survey demonstrated non-linear relationships between variables. To reduce the effects of non-linearity, polynomial basis functions were introduced before fitting all of them into a linear regression model. This model was tested separately for intermediate and downstream station data. Basis functions were expressed using 5th degree polynomial (quintic function). Representation of these non-linear contributions by creating basis function will lead to better understanding of the relationship between these variables. These basis functions for intermediate and downstream station are shown in Figures 10 to 13 in the results and discussion section. Coefficients of these basis functions for dry weight and PS yield for the stations are listed in Table 2 and 3 respectively in the result and discussion section. The general equation of basis functions is expressed as follows:

$$y(n) = \sum_{i=0}^{5} (a_i) x^i \qquad \text{Eq (1)}$$

Where y(n) represents dry weight or PS yield (dependent variable) and 'n' ranges from 1 to 9 for nine different independent variables. Dry weight or PS yield as a function of nitrate concentration [y(1)], phosphate concentration [y(2)], Si [y(3)], DOC [y(4)], SM [y(5)], NH4 [y(6)], NO2 [y(7)], diuron concentration [y(8)] and temperature [y(9)] measured in intermediate and downstream stations were fitted into a multiple linear regression model (Equation 2) to obtain the net dry weight or PS yield (y).

$$y = A + B\ y(1) + C\ y(2) + D\ y(3) + E\ y(4) + Fy(5) + Gy(6) + Hy(7) + Iy(8) + Jy(9) \quad \text{Eq.(2)}$$

Applying this two-step approach, we derived our final results using Python 3.7.6. To code in Python, some basic libraries like pandas were used for wrangling of data and to read the comma separated value data files, Matplotlib for data visualization and generating the graphs, scikit-learn to introduce polynomial features and conduct linear regression, statsmodels for model summary generation etc.



```
import pandas as pd
import numpy as np
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.metrics import r2_score
import math
import statsmodels.api as sm
import scipy
from scipy.stats import spearmanr
from scipy.stats import pearsonr
import scipy.stats
```

Figure 9: Used standard libraries of Python for this modeling.

When dry weight of biofilm was considered as the dependent variable, coefficients of determination (R^2) of the multiple linear regression model for intermediate and downstream station data were 0.601 and 0.531 respectively. When PS yield of biofilm was considered as the dependent variable, coefficients of determination (R^2) of the multiple linear regression model for intermediate and downstream station data were 0.714 and 0.629 respectively. The p-values for the two data sets were less than 0.01 indicating the goodness of fit. (Table 2-5 in the results and discussion section)

Results and Discussion:

The model (basis function followed by linear regression) was developed using the data from September 2008 to December 2011. Model validation for downstream and intermediate stations were done by comparing measured dry weight and PS yield from the field study and model estimated dry weight and PS yield for the year 2011 (Figure 14). The mean absolute percentage error (MAPE) between field measured and model estimated PS yield was 13.2% for intermediate station and 18.5% for downstream station. MAPE between field measured and model estimated dry weight was 25.8% for intermediate and 20.3% for downstream station. The t-test results indicated that the nutrients (NO_3, NO_2, PO_4 and DOC) had statistically significant ($p < 0.05$) on PS yield and dry weight. The coefficient of determination of the multiple linear regression of PS yield ranged from $R^2 = 0.63-0.71$ and same for the dry weight ranged from $R^2 =0.53-0.60$ for two sampled station location. As per our model results and depending upon the measured data, temperature did not have significant influence on the biofilm growth in the two stations. (Table 2-7)

In the present study, the nutrients required for biofilm growth such as ammonium, nitrate, nitrite, phosphate, Silicon and Dissolved organic carbon were measured every two weeks. These nutrients are the main sources of heterotrophic nutrition. Nitrogen and Phosphorus are nutrient supplements to autotrophic nutrition. Silicon is the main constituent for the formation of the diatom frustules. Our model shows significant ($p < 0.05$) positive influence of nitrite, nitrate, phosphate and DOC on the biofilm endpoints in the two stations. During the study period, the average diuron concentrations in the two stations varied from 0.0047 ug/l to 3.2 ug/l. Within this concentration range neither of the measured functional endpoints of the biofilms were significantly affected by diuron. (Table 2-7)

FIGURES:

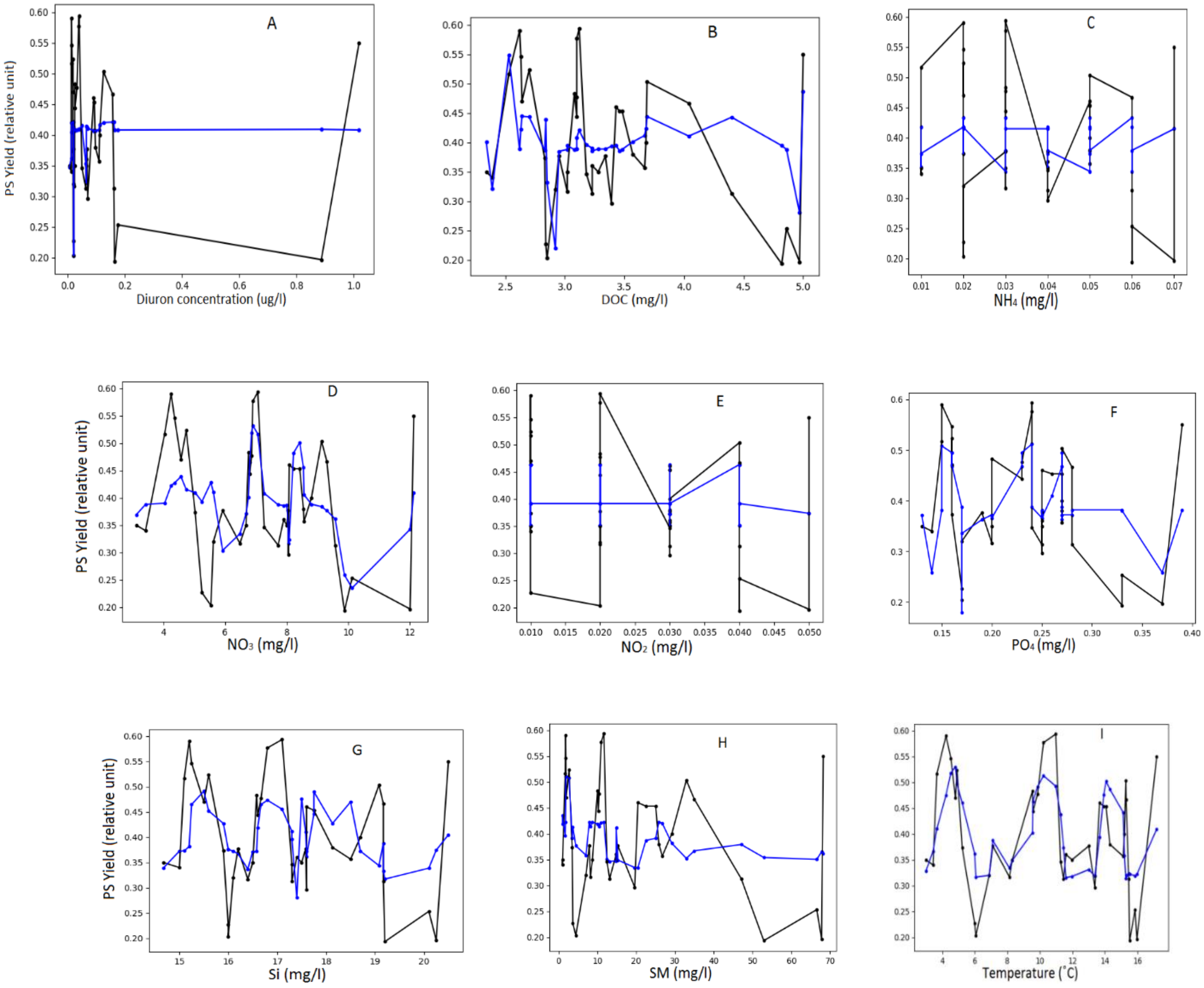


Figure 10: Dependent variable (PS Yield) of intermediate station plotted against the independent variables A. Diuron concentration, B. Dissolved Organic Carbon (DOC), C. NH_4 concentration, D. NO_3 concentration, E. NO_2 concentration, F. PO_4 concentration, G. Si concentration, H. Suspended Matter (SM) and I. Temperature as black lines. Blue lines are plotted after applying basis functions to reduce the effects of non-linearity and understand the relationship between the dependent variable and independent variables.

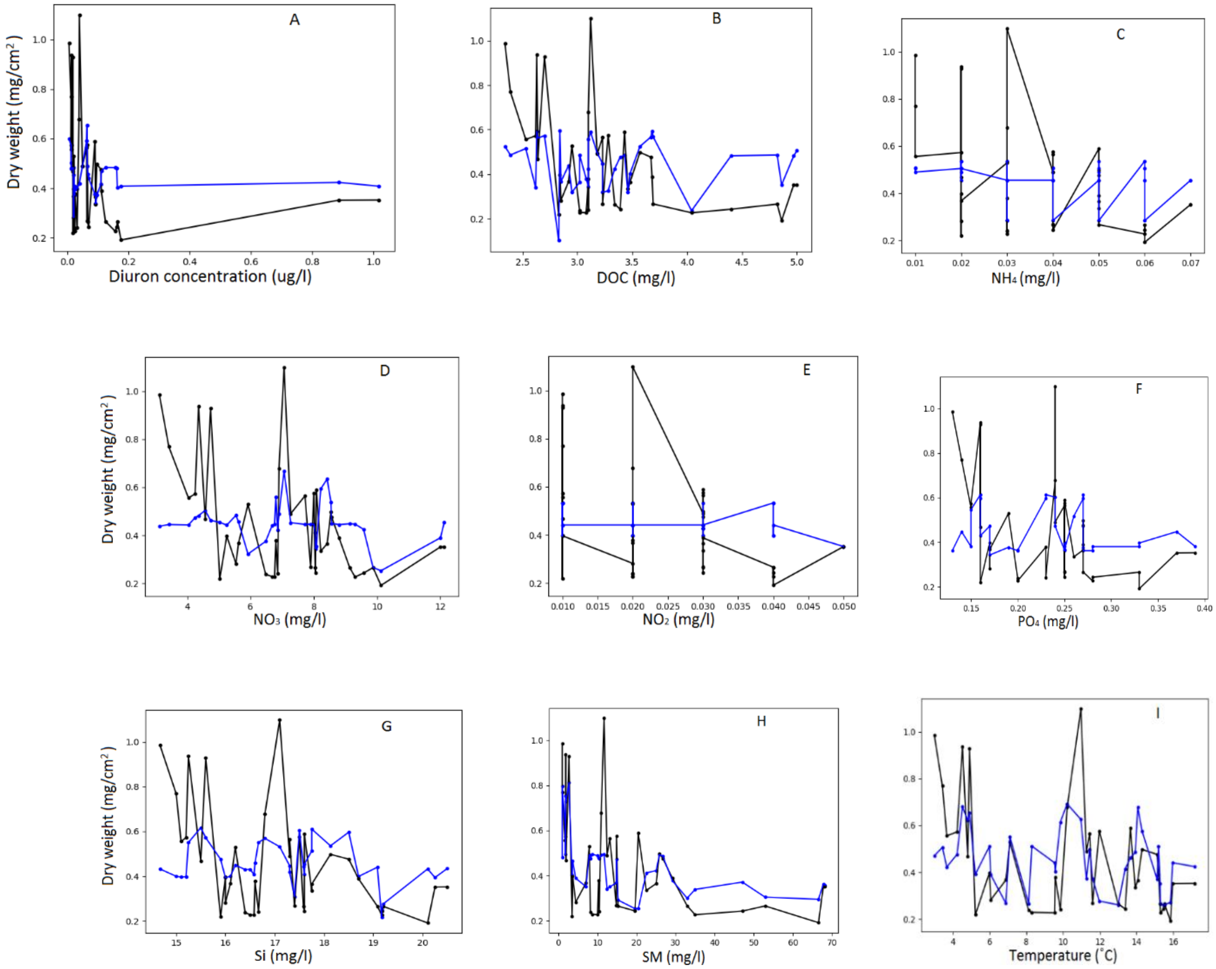


Figure 11: Dependent variable (dry weight) of intermediate station plotted against the independent variables A. Diuron concentration, B. Dissolved Organic Carbon (DOC), C. NH_4 concentration, D. NO_3 concentration, E. NO_2 concentration, F. PO_4 concentration, G. Si concentration, H. Suspended Matter (SM) and I. Temperature as black lines. Blue lines are plotted after applying basis functions to reduce the effects of non-linearity and understand the relationship between the dependent variable and independent variables.

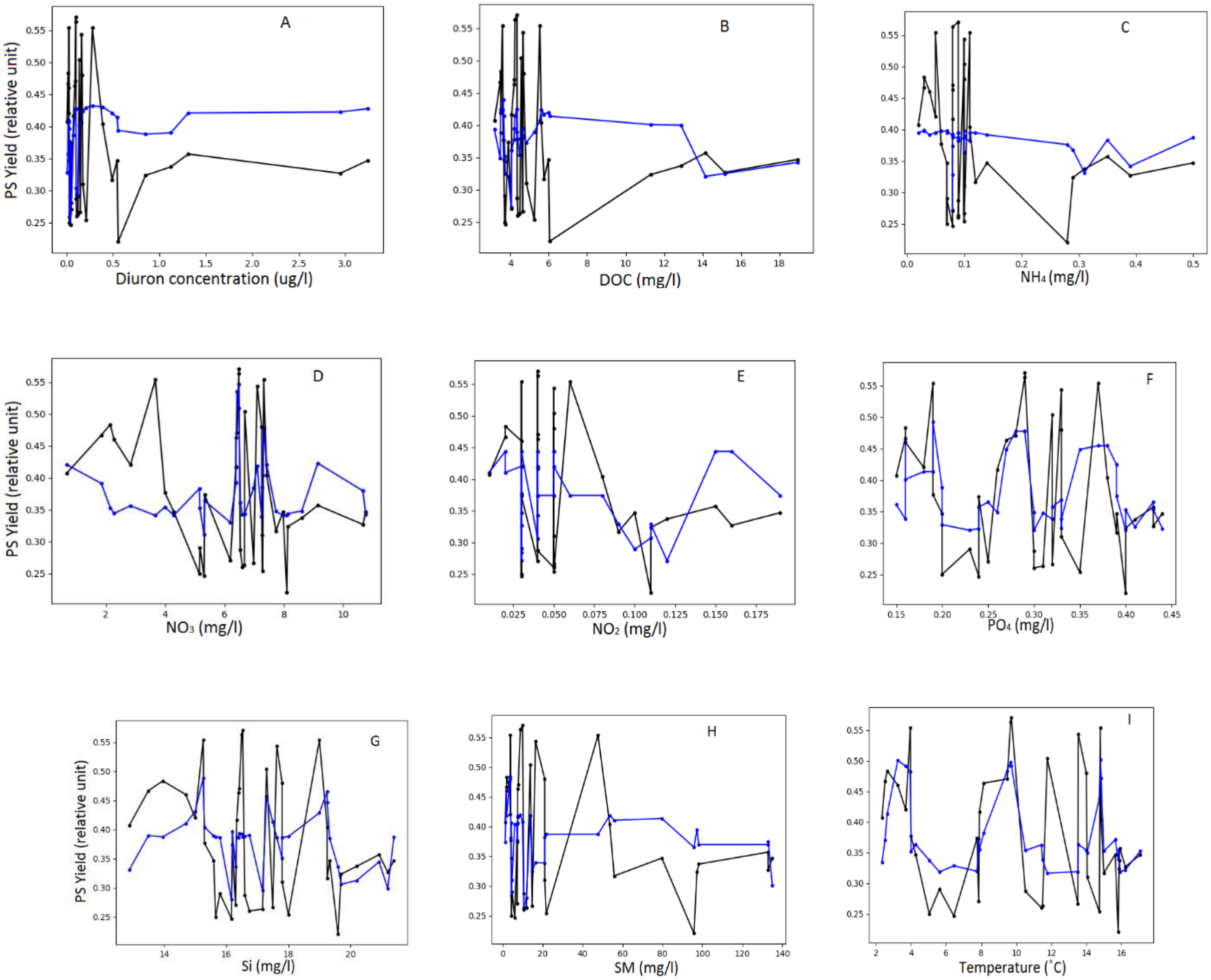


Figure 12: Dependent variable (PS Yield) of downstream station plotted against the independent variables A. Diuron concentration, B. Dissolved Organic Carbon (DOC), C. NH_4 concentration, D. NO_3 concentration, E. NO_2 concentration, F. PO_4 concentration, G. Si concentration, H. Suspended Matter (SM) and I. Temperature as black lines. Blue lines are plotted after applying basis functions to reduce the effects of non-linearity and understand the relationship between the dependent variable and independent variables.

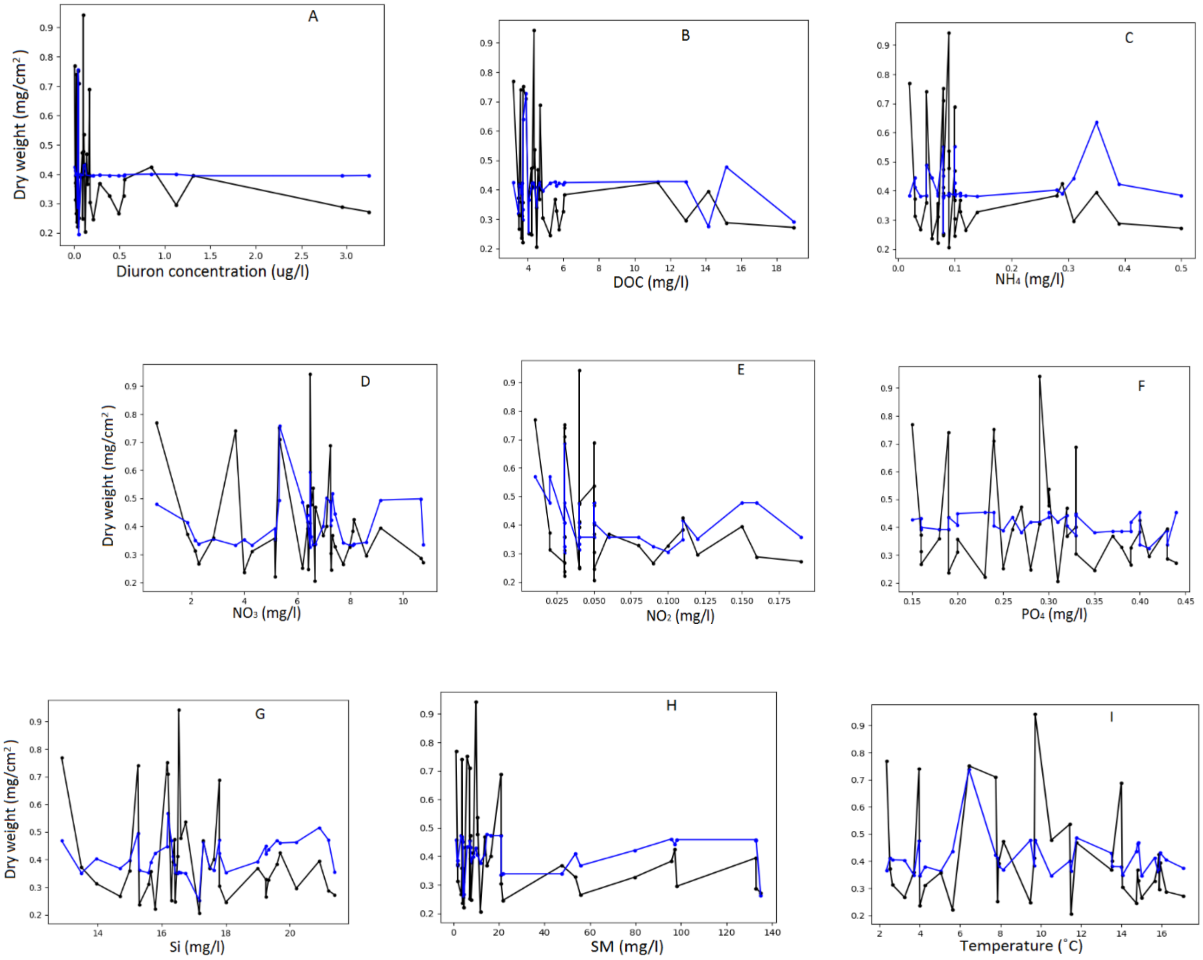
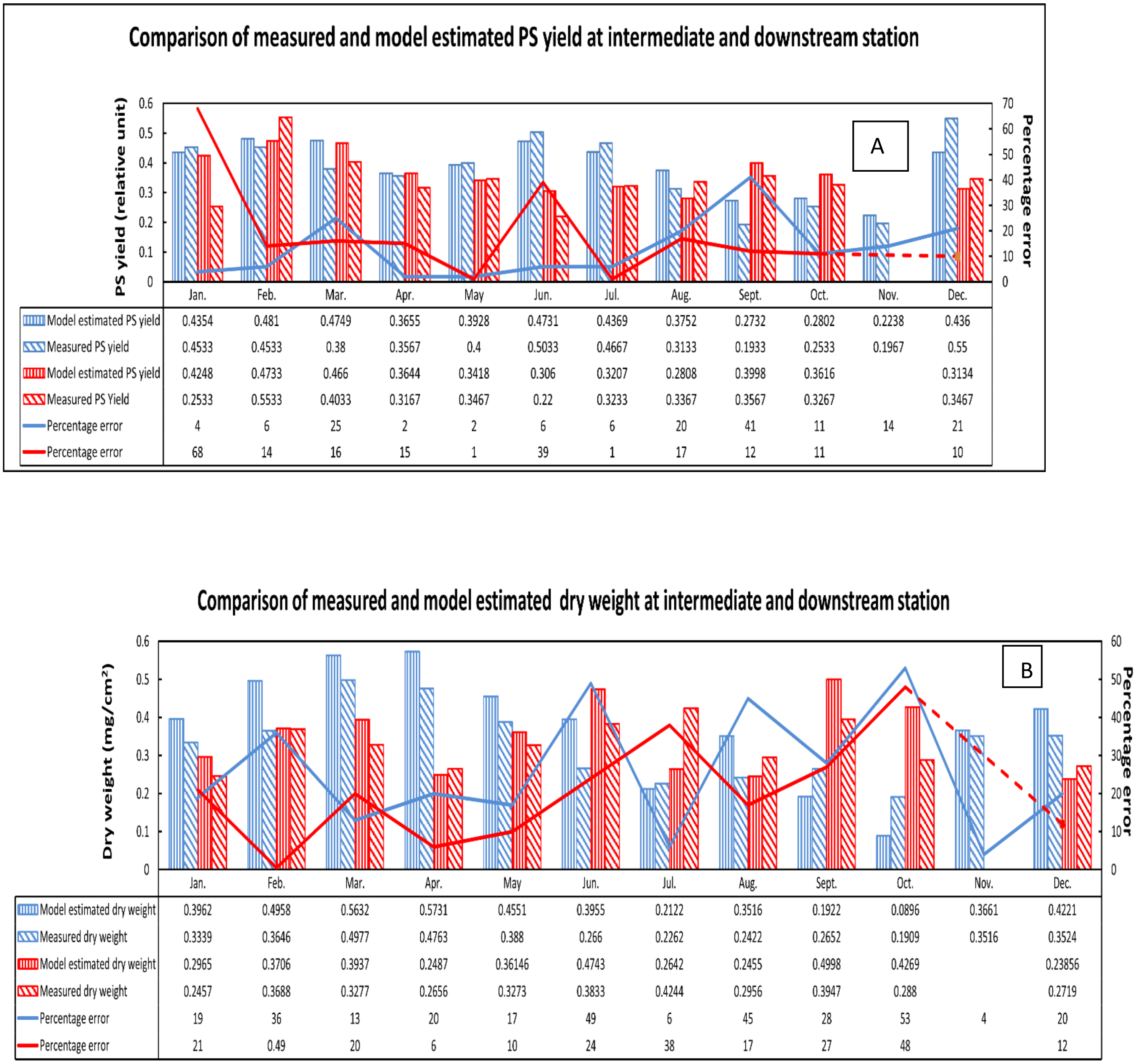


Figure 13: Dependent variable (dry weight) of downstream station plotted against the independent variables A. Diuron concentration, B. Dissolved Organic Carbon (DOC), C. NH_4 concentration, D. NO_3 concentration, E. NO_2 concentration, F. PO_4 concentration, G. Si concentration, H. Suspended Matter (SM) and I. Temperature as black lines. Blue lines are plotted after applying basis functions to reduce the effects of non-linearity and understand the relationship between the dependent variable and independent variables.



Comparison of measured and model estimated PS yield at intermediate and downstream station

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
▥ Model estimated PS yield	0.4354	0.481	0.4749	0.3655	0.3928	0.4731	0.4369	0.3752	0.2732	0.2802	0.2238	0.436
▧ Measured PS yield	0.4533	0.4533	0.38	0.3567	0.4	0.5033	0.4667	0.3133	0.1933	0.2533	0.1967	0.55
▥ Model estimated PS yield	0.4248	0.4733	0.466	0.3644	0.3418	0.306	0.3207	0.2808	0.3998	0.3616		0.3134
▧ Measured PS Yield	0.2533	0.5533	0.4033	0.3167	0.3467	0.22	0.3233	0.3367	0.3567	0.3267		0.3467
— Percentage error	4	6	25	2	2	6	6	20	41	11	14	21
— Percentage error	68	14	16	15	1	39	1	17	12	11		10

Comparison of measured and model estimated dry weight at intermediate and downstream station

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
▥ Model estimated dry weight	0.3962	0.4958	0.5632	0.5731	0.4551	0.3955	0.2122	0.3516	0.1922	0.0896	0.3661	0.4221
▧ Measured dry weight	0.3339	0.3646	0.4977	0.4763	0.388	0.266	0.2262	0.2422	0.2652	0.1909	0.3516	0.3524
▥ Model estimated dry weight	0.2965	0.3706	0.3937	0.2487	0.36146	0.4743	0.2642	0.2455	0.4998	0.4269		0.23856
▧ Measured dry weight	0.2457	0.3688	0.3277	0.2656	0.3273	0.3833	0.4244	0.2956	0.3947	0.288		0.2719
— Percentage error	19	36	13	20	17	49	6	45	28	53	4	20
— Percentage error	21	0.49	20	6	10	24	38	17	27	48		12

Figure 14: Comparison of measured and model estimated (A) photosynthetic yield and (B) dry weight at intermediate (blue) and downstream (red) stations for the survey year 2011. Biofilm were unable to collect in November 2011 in downstream station due to major high flow event. The lines indicate (respective colour) percent error between the measured and model estimated PS yield and dry weight.

TABLES:

Table 2: Coefficients of basis functions (dependent variable: dry weight)

Intermediate station									
Coefficients	y(1)	y(2)	y(3)	y(4)	y(5)	y(6)	y(7)	y(8)	y(9)
a_0	3.4018	-21.184	-7370.63	-154.52	0.191451	2.7752	-0.7155	0.308018	-2.6842
a_1	-3.07e+00	455.47	2.07e+03	2.08e+02	6.617e-02	-4.37e+02	1.95e+02	6.5941	2.52e+00
a_2	1.10e+00	-3643.36	-2.32e+02	-1.10e+02	-4.61e-03	2.70e+04	-1.06e+04	-88.533	-6.79e-01
a_3	-1.81e-01	13942.62	1.30e+01	2.85e+01	1.085e-04	-7.34e+05	2.34e+05	473.033	8.13e-02
a_4	1.39e-02	-25694.98	-3.62e-01	-3.64e+00	-7.34e-07	9.13e+06	-1.81e+06	-733.60	-4.45e-03
a_5	-4.06e-04	18334.64	4.02e-03	1.83e-01	-6.95e-10	-4.25e+07	-2.74e+05	341.98	9.08e-05
Downstream station									
Coefficients	y(1)	y(2)	y(3)	y(4)	y(5)	y(6)	y(7)	y(8)	y(9)
a_0	1.2864	0.38986	-996.76	-1.3602	0.4873	0.8732	0.8219	0.39736	-1.117
a_1	-1.09e+00	10.668	2.948e+02	1.188e+00	-7.41e-03	-14.725	-1.46e+01	-0.0977	1.205e+00
a_2	5.525e-01	-151.36	-3.45e+01	-2.73e-01	-2.56e-05	153.033	9.60e+01	0.91326	-3.30e-01
a_3	-1.25e-01	765.26	2.014e+00	2.58e-02	2.33e-06	-666.487	5.47e+02	-1.4842	4.051e-02
a_4	1.26e-02	-1628.75	-5.82e-02	-1.028e-03	-1.40e-08	1287.328	-7.48e+03	0.7107	-2.28e-03
a_5	-4.61e-04	1237.23	6.68e-04	1.38e-05	6.25e-12	-917.194	1.98e+04	-0.1030	4.845e-05

Table 3: Coefficients of basis functions (dependent variable: PS yield)

Intermediate station									
Coefficients	y(1)	y(2)	y(3)	y(4)	y(5)	y(6)	y(7)	y(8)	y(9)
$\mathbf{a_0}$	0.38276	-15.3727	-4769.31	-81.308	0.3109	0.3482	-0.4467	0.4072	0.7184
$\mathbf{a_1}$	-3.70e-01	365.056	1.35e+03	123.16	2.55e-02	-3.57e+00	1.37e+02	-1.51e-01	-1.12e-01
$\mathbf{a_2}$	1.90e-01	-3224.589	-1.54e+02	-73.209	-1.89e-03	3.76e+02	-7.168e+03	1.16e+01	2.23e-02
$\mathbf{a_3}$	-3.65e-02	13684.17	8.683e+00	21.462	5.061e-05	-5.79e+03	1.502e+05	-1.07e+02	-2.21e-03
$\mathbf{a_4}$	3.06e-03	-28036.67	-2.447e-01	-3.1037	-5.435e-07	-2.55e+04	-1.098e+06	1.86e+02	8.90e-05
$\mathbf{a_5}$	-9.38e-05	22241.97	2.753e-03	0.177	2.004e-09	5.99e+05	-1.66e+05	-9.02e+01	-1.00e-06
Downstream station									
Coefficients	y(1)	y(2)	y(3)	y(4)	y(5)	y(6)	y(7)	y(8)	y(9)
$\mathbf{a_0}$	0.6131	4.6599	-723.611	0.6405	0.2988	0.3611	-0.0356	0.4335	0.0637
$\mathbf{a_1}$	-5.61e-01	-85.30	2.15e+02	-7.705e-02	2.376e-02	1.3911	4.17e+01	-0.24177	3.82e-01
$\mathbf{a_2}$	3.368e-01	-681.18	-2.54e+01	3.196e-03	-1.47e-03	-15.38	-1.24e+03	-0.2514	-1.14e-01
$\mathbf{a_3}$	-7.98e-02	-2646.66	1.49e+00	4.369e-04	3.11e-05	41.25	1.50e+04	0.5380	1.38e-02
$\mathbf{a_4}$	8.116e-03	4943.646	-4.36e-02	-3.94e-05	-2.61e-07	-2.71	-7.97e+04	-0.2579	-7.54e-04
$\mathbf{a_5}$	-2.93e-04	-3548.46	5.061e-04	8.417e-07	7.54e-10	-61.65	1.54e+05	0.0377	1.505e-05

Table 4: Multiple linear regression model parameters for intermediate station data (dependent variable: dry weight).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	-0.9910	0.380	-2.607	0.014
B (NO_3)	0.3661	0.437	0.837	0.410
C (PO_4)	-0.2816	0.509	-0.553	0.585
D (Si)	0.4538	0.414	1.096	0.282
E (DOC)	0.6047	0.280	2.157	**0.040**
F (SM)	0.4058	0.300	1.351	0.188
G (NH_4)	0.5150	0.346	1.489	0.148
H (NO_2)	0.0258	0.569	0.045	0.964
I (Diuron)	0.8068	0.442	1.827	0.078
J (Temp.)	0.3134	0.356	0.880	0.387
dF Residuals	28			
Multiple R squared	0.601			
Adjusted R squared	0.472			
Residual standard error	0.1724			
F statistics	4.678			
p value	0.000765			

Table 5: Multiple linear regression model parameters for downstream station data (dependent variable: dry weight).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	-1.4242	0.453	-3.142	0.004
B (NO_3)	0.6483	0.312	2.076	**0.048**
C (PO_4)	1.8391	0.819	2.246	**0.033**
D (Si)	0.2286	0.492	0.465	0.646
E (DOC)	0.2887	0.745	0.387	0.701
F (SM)	0.8679	0.483	1.796	0.084
G (NH_4)	0.1028	0.569	0.180	0.858
H (NO_2)	0.2745	0.375	0.731	0.471
I (Diuron)	-0.0205	0.770	-0.027	0.979
J (Temp.)	0.2443	0.417	0.586	0.563
dF Residuals	27			
Multiple R squared	0.531			
Adjusted R squared	0.375			
Residual standard error	0.1421			
F statistics	3.396			
p value	0.00656			

Table 6: Multiple linear regression model parameters for intermediate station data (dependent variable: PS yield).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	-0.6479	0.247	-2.624	0.014
B (NO_3)	0.7198	0.320	2.246	**0.033**
C (PO_4)	0.4172	0.320	1.305	0.203
D (Si)	-0.4817	0.437	-1.103	0.279
E (DOC)	0.5038	0.251	2.004	0.055
F (SM)	-0.1507	0.336	-0.448	0.657
G (NH_4)	0.1421	0.538	0.264	0.793
H (NO_2)	0.8462	0.286	2.958	**0.006**
I (Diuron)	0.3746	0.430	0.872	0.391
J (Temp.)	0.2594	0.390	0.665	0.511
dF Residuals	28			
Multiple R squared	0.714			
Adjusted R squared	0.622			
Residual standard error	0.0681			
F statistics	7.774			
p value	1.21e-05			

Table 7: Multiple linear regression model parameters for downstream station data (dependent variable: PS yield).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	-0.5283	0.206	-2.563	0.016
B (NO_3)	0.5727	0.324	1.769	0.088
C (PO_4)	0.2116	0.383	-0.553	0.585
D (Si)	0.2680	0.335	0.800	0.431
E (DOC)	0.1210	0.546	0.222	0.826
F (SM)	0.5933	0.294	2.017	0.054
G (NH_4)	0.2020	0.622	0.325	0.748
H (NO_2)	0.4753	0.290	1.637	0.113
I (Diuron)	0.0887	0.337	0.263	0.795
J (Temp.)	0.2746	0.338	0.813	0.424
dF Residuals	27			
Multiple R squared	0.629			
Adjusted R squared	0.505			
Residual standard error	0.0734			
F statistics	5.087			
p value	0.000464			

Case II

Case II: This study aimed to understand or assess the inhibitory effect of reject water on nitrite oxidizing bacteria at different biofilm thickness via statistical modeling approach. Moving bed biofilm reactor (MBBR) is a type of wastewater treatment process, which was first invented by Prof. Hallvard Odegaard at Norwegian University of Science and Technology in the late 1980s. In the year of 1985, there are strong political debates among the North Sea countries to significantly reduce about 50% of the nutrient (nitrogen and phosphorus) loads to the North Sea during the period of 1985-95. Hence there was an urgent need for the modification of existing wastewater treatment plants as well as installing new treatment plant. In early 90's MBBR system was first introduced (Ødegaard, Rusten, and Westrum 1994) Prof. Hallvard Odegaard in Norway for enhancing the nutrient removal efficiency of conventional ASP reactor. Series of pilot scale studies were performed, all of which showed a very good removal in terms of COD and nutrients (Goswami and Mazumder 2019) .In last two decades, MBBR is proven as a simple, robust, flexible and compact wastewater technology for both municipal and industrial wastewater treatment. The MBBR is particularly advantageous for slow processes like nitrification, where ammonia is oxidized to nitrate by autotrophic bacteria. Due to slow growth of autotrophic biomass, nitrification is often the most critical step in biological nitrogen removal. Although nitrification is usually considered as a one step process in wastewater treatment, it is known to consist of two steps. These two steps involve 1) Ammonia oxidizing bacteria (AOB) converts ammonia to nitrite. 2) nitrite-oxidizing bacteria (NOB) convert nitrite to nitrate. Under normal condition the first step is rate limiting and nitrite rarely build up in wastewater treatment process. However at elevated temperature above 20˚C the oxidation rate of ammonia exceeds the oxidation rate of nitrate and nitrate accumulates in the system (Piculell, Welander, et al. 2016). The oxidation of ammonium to nitrate by autotrophic ammonia-oxidizing bacteria (AOB) i.e., the first step of nitrification, called nitritation is the key for achieving energy efficient nitrogen removal processes such as partial nitritation and annamox (PNA). In conventional process of nitrogen removal, AOB and NOB co-exist in the system and ammonium nitrogen (NH_4^+-N) is completely oxidized to nitrate and then reduced to nitrogen gas in the denitrification stage. But in PNA process, the oxidation of ammonium is halted at nitrite which is then converted to nitrogen (S. Wang et al. 2019)gas by annamox bacteria. For that reason PNA requires less oxygen and carbon compared to conventional process hence energy efficient (Daigger 2014).

MBBR process has also been developed for ammonia removal through both traditional nitrification and de-nitrification process, ammonium ion is oxidized to nitrate by complete Nitrification, and subsequently nitrate is reduced to nitrogen gas by pre or post de-nitrification. nitrogen removal is usually carried out in two different reactors. Inorganic carbon as alkalinity is normally supplied to perform ammonium oxidation. De-nitrification requires easily degradable organic such as methanol as electro-acceptor. Partial nitrification, called nitritation and anaerobic ammonium oxidation can also be achieved to remove nitrogen from wastewater in one reactor by manipulating dissolved oxygen concentration into the biofilm that means oxidation of nitrite to nitrate is suppressed and de-nitrification can occur (S. Wang et al. 2019).

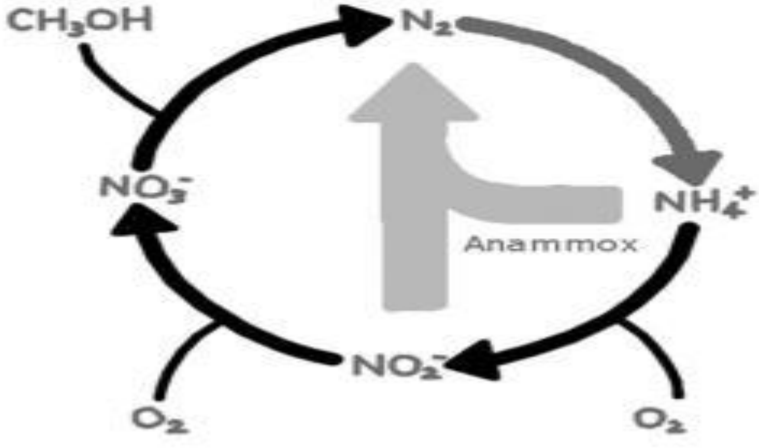


Figure 15: Annamox process flow diagram ("Anammox | SSWM - Find Tools for Sustainable Sanitation and Water Management!" n.d.)

The activity of NOB is detrimental in PNA systems, as NOB will compete with annamox bacteria for nitrite and with AOB for oxygen. So, effective inhibition of NOB activity is crucial to achieve efficient PNA processes in wastewater treatment (Al-Omari et al. 2015; Xu et al. 2015). There are some approaches to achieve NOB suppression at mainstream condition: Aeration with low DO to limit NOB growth, thereby provide the advantage to annamox bacteria to form a partnership with AOB (Ma et al. 2015). Operation can be conducted at low DO and high effluent ammonium concentrations to ensure higher growth rate of AOB over NOB (Isanta et al. 2015). NOB are more sensitive to low DO concentrations than AOB due to their different oxygen affinities, and activity can be inhibited under low DO conditions (Hao, Heijnen, and Van Loosdrecht 2002). An alternate approach to suppress NOB in mainstream operation is to regularly alternate between mainstream and reject operation, either by moving the biomass or by switching the feed (Al-Omari et al. 2015; Q. Wang et al. 2014). NOB suppression could be achieved at high DO in a nitrifying MBBR but biofilm thickness limited to 300micron (Piculell, Suarez, et al. 2016).

In the present study, the statistical model has been developed based on published (Piculell, Suarez, et al. 2016) data. This model has been developed to understand the influence of reject exposure conditions in NOB inhibition for thinner as well as thicker biofilm thickness in nitrifying MBBR with PNA system.

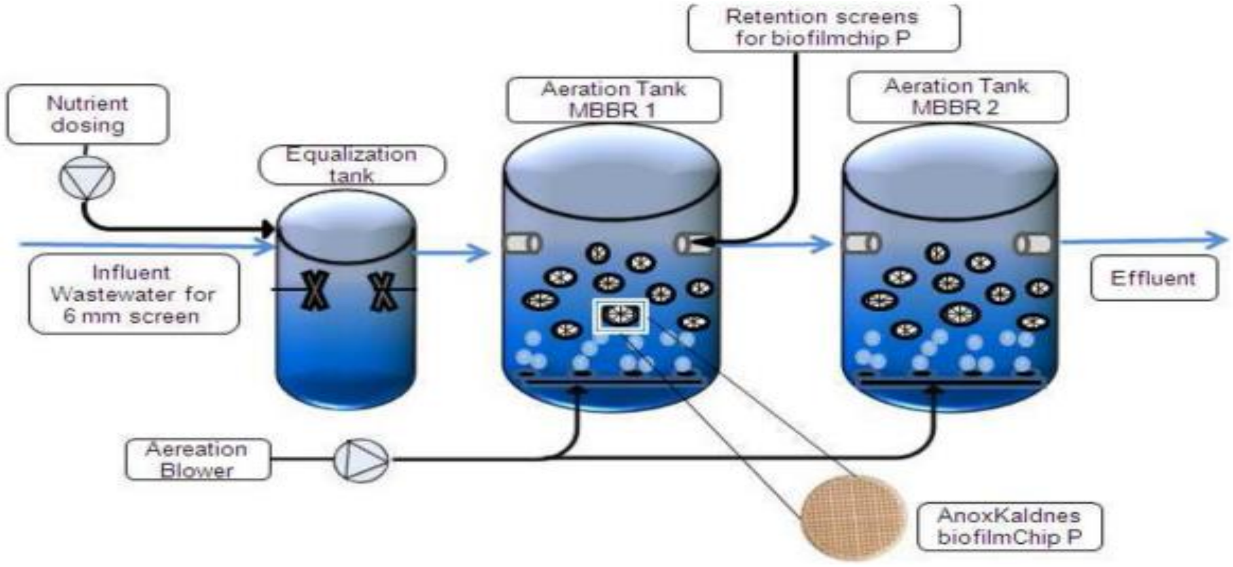

Retention screens for biofilmchip P
Nutrient dosing
Aeration Tank MBBR 1
Aeration Tank MBBR 2
Equalization tank
Influent Wastewater for 6 mm screen
Effluent
Aeration Blower
AnoxKaldnes biofilmChip P

Figure 16: Process flow diagram of MBBR Plant (Revilla, Galán, and Viguri 2016).

Data collection:

The data collected from the published paper (Piculell, Suarez, et al. 2016). They studied a fully nitrifying MBBR system, fed with main-stream wastewater was temporally switched or exposed to reject water at different loading rates from sludge treatment for inhibition of NOB. For each inhibition trial they were used 100 pieces of sample carriers collected from pilot reactor and placed in 1L lab-scale MBBR, either Z400 (0.0013m^2/carrier) or Z50 (0.0011m^2/carrier). Pilot plant of 0.5m^3 of capacity MBBR reactor, located at Sjolunda wastewater treatment plant (Malmo, Sweden). During operation, the pilot reactor fed with municipal high rate activated sludge plant. In this study we used the Table 1 data (Piculell, Suarez, et al. 2016), displaying the different reject exposure conditions and Fig. 7 to read the relative change in NOB activity as observed one day after reject exposure for each inhibition trial.

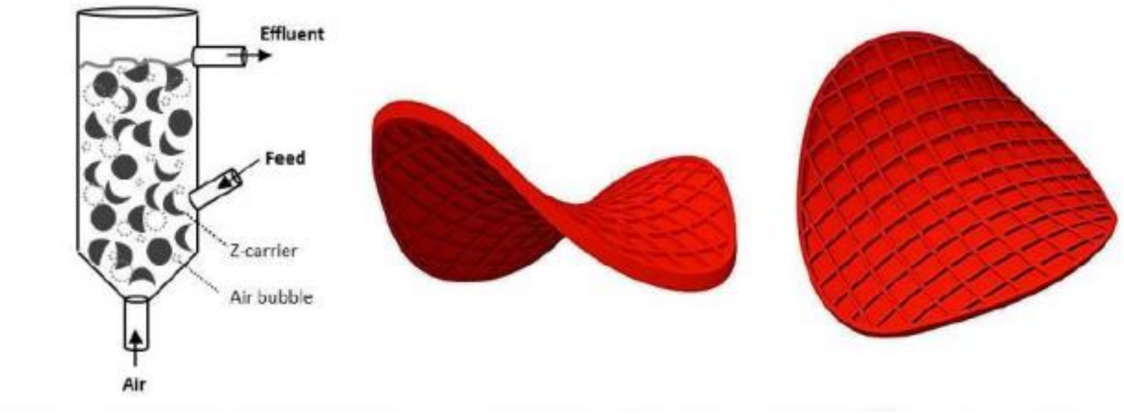


Figure 17: Lab scale MBBR and biofilm carriers (Piculell, Welander, et al. 2016).

Model development: The objective of this study was to assess the sensitivities of certain independent reject exposure conditions [namely total reject exposure (RE_T), loading rate, maximum concentration of free ammonia (max. FA), hydraulic retention time (HRT), maximum concentration of free nitrous acid (max. FNA) and exposure time] on relative change in AOB and NOB activity or inhibition. The developed model is primarily a polynomial regression model where the parameters of the model were estimated using least square method. In this modeling approach relative change in NOB and AOB activity was taken as dependent variables. Whereas total reject exposure (RE_T), loading rate, maximum concentration of free ammonia (max. FA), hydraulic retention time (HRT), maximum concentration of free nitrous acid (max. FNA) and exposure time were considered as independent variables. Data collected from the lab demonstrated non-linear relationships between variables. To reduce the non-linearity, polynomial basis functions were introduced before fitting all them into a linear regression model. Basis functions were expressed using 6th degree polynomial (sextic function). Representation of these non-linear contributions by creating basis function will lead to better understanding of the relationship between these variables.

The general equation of basis functions is expressed as follows:

$$y(n) = \sum_{i=0}^{6} (a_i) x^i \qquad \text{Eq (1)}$$

Where y(n) represents relative changes in AOB or NOB activity and 'n' ranges from 1 to 6 for six independent variables. Relative changes in AOB or NOB activity as a function of total reject exposure (RE_T) [y(1)], loading rate [y(2)], maximum concentration of free ammonia (max. FA) [y(3)], hydraulic retention time (HRT) [y(4)], maximum concentration of free nitrous acid (max. FNA) [y(5)] and exposure time[y(6)] measured for both the biofilm carriers were fitted into a multiple linear regression model (Equation 2) to obtain the relative changes in AOB or NOB activity (y).

$$y = A + B\,y(1) + C\,y(2) + D\,y(3) + E\,y(4) + Fy(5) + Gy(6) \qquad \text{Eq (2)}$$

Applying this two-step approach, we derived our final results using Python 3.7.6.

Results and Discussions: When relative changes in NOB activity was considered as the dependent variable and thinner (Z50) biofilm carriers are used, coefficients of determination (R^2) of the multiple linear regression model is 0.996 (Table 8). When relative changes in NOB activity was considered as the dependent variable and thicker (Z400) biofilm carriers are used, coefficients of determination (R^2) of the multiple linear regression model 0.987 (Table 9). The p values for two data sets were less than 0.01 indicating the goodness of fit. In present study, t-test results indicate that total reject exposure (RE_T), maximum concentration of free ammonia (max. FA) and maximum concentration of free nitrous acid (max. FNA) had significant influence (p<

0.05) on NOB inhibition when Z50 biofilm carriers are used. No such sensitivities observed in case of NOB suppression in thicker biofilms i.e., Z400 biofilm carriers. (Table 8-9)

Table 8: Multiple linear regression model parameters for Z50 (dependent variable: Relative change in NOB activity).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	0.017	0.024	0.714	0.527
B	0.6489	0.175	3.718	0.034
C	0.1953	0.261	0.749	0.508
D	0.5771	0.135	4.264	0.024
E	-0.9319	0.307	-3.032	0.056
F	0.5458	0.098	5.590	0.011
G	0.0142	0.082	0.173	0.874
dF Residuals	3			
Multiple R squared	0.996			
Adjusted R squared	0.987			
Residual standard error	0.028			
F statistics	116.6			
p value	0.000765			

Table 9: Multiple linear regression model parameters for Z400 (dependent variable: Relative change in NOB activity).

Coefficients	**Estimate**	**Standard Error**	**t Value**	**Pr(>|t|)**
A	0.0104	0.017	0.599	0.591
B	1.9431	0.998	1.948	0.147
C	0.3008	0.365	0.824	0.470
D	0.0420	0.139	0.302	0.783
E	-0.0061	1.113	-0.005	0.996
F	-1.2577	0.912	-1.379	0.262
G	0.0992	0.180	0.550	0.620
dF Residuals	3			
Multiple R squared	0.987			
Adjusted R squared	0.962			
Residual standard error	0.0363			
F statistics	39.1			
p value	0.00611			

Chapter IV

Model Significance

This type of modeling can be used for diagnostic task to understand about the various sensitivities as introduced by each of the independent variables.

Case I: This statistical model provides deeper insight into the sensitivities of different biological endpoints towards nutrients and toxicants. Growth and response of lotic biofilms depends on several factors and such multiple linear regression models are able to assess large number of variables. interrelations between them, and are therefore efficient in defining biological processes. Our model shows that within the given diuron concentration limits (maximum concentration 3.2 µg/L) used in the model the herbicide did not influence PS yield or the dry weight of the biofilms. Instead, the various nutrients and DOC played a significant role in enhancing biofilm growth (positive correlation) in the Morcille river.

Case II: Different reject exposure conditions like total reject exposure (RE_T), loading rate, maximum concentration of free ammonia (max. FA), hydraulic retention time (HRT), maximum concentration of free nitrous acid (max. FNA) and exposure time were the probable cause for relative change in AOB and NOB activity or inhibition. From this modeling approach we can easily identify and understand which independent variables had significant influence on the dependent variable among a set of independent variables. So, such multiple linear regression models were able to assess large number of variables and interrelations between them. Therefore, this type of modeling can be fitted to define a biological wastewater treatment process like MBBR.

In that study, total reject exposure (RE_T), maximum concentration of free ammonia (max. FA) and maximum concentration of free nitrous acid (max. FNA) had significant influence ($p< 0.05$) on NOB inhibition and max. FA had significant ($p<0.05$) influence on relative changes in the AOB activity when Z50 biofilm carriers were used. So, this model shows that max. FNA also had a significant impact on NOB inhibition in the thin biofilms along with total reject exposure and max. FA. In thicker biofilms (Z400 carriers) relative changes in the AOB activity was significantly ($p<0.05$) influenced by total reject exposure. No such sensitivities observed in case of NOB suppression in thicker biofilms i.e., Z400 biofilm carriers.

Chapter V

Conclusion

Case I: The linear regression model explained 63-71% of the variance in the PS yield and 53-60% of the variance in the biomass dry weight due to the influence of the different dependent variables. Nutrients (NO_3, NO_2, PO_4) significantly contributed towards biofilm functional (PS yield) and structural (dry weight) endpoints. Dry weight in the intermediate station was positively influenced by dissolved organic carbon (DOC) which is the nutrient needed for heterotrophic metabolism of biofilm. Diuron in the modelled concentration range did not affect two functional endpoints. Therefore, our results all converge to reveal that this model can efficiently assess the effect of co-occurring factors on river biofilm community.

Case II:

- The linear regression model explained 98.7% of the variance in relative change in NOB activity in Z400 biofilm carriers and 99.6% of the variance in relative change in NOB activity in Z50 biofilm carriers due to the influence of different independent reject exposure conditions.
- Total reject exposure, maximum free ammonia, maximum free nitrous acid had significant influence on the relative change in NOB activity or inhibition in Z40 biofilm. No such sensitivities were observed in case of NOB suppression in thicker biofilms i.e., Z400 biofilm carriers.
- Therefore, our results all converge to reveal that this statistical model approach efficiently assess the effect of co-occuring factors on NOB inhibition in PNA systems
- But to understand the process we need more data. In the second study we were facing the problems of lack of data points to study. So, from this study we have learnt that this type of two step model can be applied even if there are ample data points to run the programs or codes and get the proper insight.

Thus, it is concluded that our two-step statistical model approach can be used to understand different biofilm processes. Biofilm processes and modeling is an active field of research worldwide. This model will contribute in this vast field of research to some extent.

References:

Abdulwahid, Mohanad Y., Tamer Haddad, Hend Tubaila, and Imad A. Al-Qasem. 2020. “Multiple Linear Regression Modelling for Predicting Building’s Short Columns Loads under Gravity.” *International Journal of Engineering Research and Technology* 13 (7): 1671. https://doi.org/10.37624/IJERT/13.7.2020.1671-1685.

Al-Omari, Ahmed, Mofei Han, Romain Lemaire, Nicolas Morales, Jose Vazquez-Padin, Heather Stewart, Ángeles Val del Río, and Bernhard Wett. 2015. “Mainstream Deammonification.” In , 107–28.

“Anammox | SSWM - Find Tools for Sustainable Sanitation and Water Management!” n.d. Accessed June 17, 2022. https://sswm.info/water-nutrient-cycle/wastewater-treatment/hardwares/semi-centralised-wastewater-treatments/anammox.

Araiza-Aguilar, J.A., M.N. Rojas-Valencia, and R.A. Aguilar-Vera. 2020. “Forecast Generation Model of Municipal Solid Waste Using Multiple Linear Regression.” *Global Journal of Environmental Science and Management* 6 (1). https://doi.org/10.22034/GJESM.2020.01.01.

Aviral Gupta, Akshay Sharma, Dr. Amita Goel, and Maharaja Agrasen Institute of Technology/Guru Gobind Singh Indraprastha University. 2017. “Review of Regression Analysis Models.” *International Journal of Engineering Research And* V6 (08): IJERTV6IS080060. https://doi.org/10.17577/IJERTV6IS080060.

Bery, Andy Anderson. 2021. “Modeling of Soil Shear Strength Using Multiple Linear Regression (MLR) at Penang, Malaysia.” *Journal of Engineering Research* 9 (3A). https://doi.org/10.36909/jer.v9i3A.7675.

Bhowmick, Tanaya, Goutam Sen, Joydeep Mukherjee, and Reshmi Das. 2021. “Assessing the Effect of Herbicide Diuron on River Biofilm: A Statistical Model.” *Chemosphere* 282 (November): 131104. https://doi.org/10.1016/j.chemosphere.2021.131104.

Bonnineau, Chloé, Joan Artigas, Betty Chaumet, Aymeric Dabrin, Juliette Faburé, Benoît J. D. Ferrari, Jérémie D. Lebrun, et al. 2021. “Role of Biofilms in Contaminant Bioaccumulation and Trophic Transfer in Aquatic Ecosystems: Current State of Knowledge and Future Challenges.” *Reviews of Environmental Contamination and Toxicology* 253: 115–53. https://doi.org/10.1007/398_2019_39.

“Bridge Construction Cost Prediction Using Multiple Linear Regression.” 2019. *International Journal of Innovative Technology and Exploring Engineering* 8 (9): 3115–21. https://doi.org/10.35940/ijitee.I8916.078919.

Charantimath, Poorinma M. 2011. *Total Quality Management*. Pearson Education India.

Chénier, Martin R., Danielle Beaumier, Réal Roy, Brian T. Driscoll, John R. Lawrence, and Charles W. Greer. 2003. “Impact of Seasonal Variations and Nutrient Inputs on Nitrogen Cycling and Degradation of Hexadecane by Replicated River Biofilms.” *Applied and Environmental Microbiology* 69 (9): 5170–77. https://doi.org/10.1128/AEM.69.9.5170-5177.2003.

Coenye, T. 2013. “Biofilms.” In *Brenner’s Encyclopedia of Genetics (Second Edition)*, edited by Stanley Maloy and Kelly Hughes, 335–37. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-374984-0.00154-6.

“CPCB | Central Pollution Control Board.” n.d. Accessed June 17, 2022. https://cpcb.nic.in/env-protection-act/.

Daigger, Glen. 2014. “Oxygen and Carbon Requirements for Biological Nitrogen Removal Processes Accomplishing Nitrification, Nitritation, and Anammox.” *Water Environment Research : A Research Publication of the Water Environment Federation* 86 (March): 204–9. https://doi.org/10.2175/106143013X13807328849459.

DeLorenzo, M. E., G. I. Scott, and P. E. Ross. 2001. “Toxicity of Pesticides to Aquatic Microorganisms: A Review.” *Environmental Toxicology and Chemistry* 20 (1): 84–98. https://doi.org/10.1897/1551-5028(2001)020<0084:toptam>2.0.co;2.

Diaz Villanueva, Veronica, Jordi Font, Thomas Schwartz, and Anna M. Romani. 2011. “Biofilm Formation at Warming Temperature: Acceleration of Microbial Colonization and Microbial Interactive Effects.” *Biofouling* 27 (1): 59–71. https://doi.org/10.1080/08927014.2010.538841.

Fernández-Cori, R.A., Juan Morales Gomero, B. Huayhuas-Chipana, Maria Sotomayor, and José Ruiz-Montoya. 2015. “Nanostructured Sensors for Determination of 3-(3,4-Dichlorophenyl)-1,1-Dimethylurea Based in Molecularly Imprinted Polymers (MIPs) Deposited in Screen Printed Carbon Nanotubes.” *ECS Transactions* 66 (July): 33–41. https://doi.org/10.1149/06637.0033ecst.

Fitzsimmons, Jake, and Pablo Moscato. 2018. “Symbolic Regression Modeling of Drug Responses.” In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, 52–59. https://doi.org/10.1109/AI4I.2018.8665684.

Goswami, Supriyo, and Debabrata Mazumder. 2019. “Modelling and Process Design of Moving Bed Bioreactor (MBBR) for Wastewater Treatment — A Review.” *J. Indian Chem. Soc.* 96: 16.

Guasch, Helena, and Sergi Sabater. 2002. “Light History Influences the Sensitivity to Atrazine in Periphytic Algae.” *Journal of Phycology* 34 (September): 233–41. https://doi.org/10.1046/j.1529-8817.1998.340233.x.

Hao, Xiaodi, Joseph J Heijnen, and Mark C. M Van Loosdrecht. 2002. “Model-Based Evaluation of Temperature and Inflow Variations on a Partial Nitrification–ANAMMOX Biofilm Process.” *Water Research* 36 (19): 4839–49. https://doi.org/10.1016/S0043-1354(02)00219-1.

Isanta, Eduardo, Clara Reino, Julián Carrera, and Julio Perez. 2015. “Stable Partial Nitritation for Low Strength Wastewater at Low Temperature in an Aerobic Granular Reactor.” *Water Research* 80 (May). https://doi.org/10.1016/j.watres.2015.04.028.

Lawens, M., and C. Mutsvangwa. 2018. “Application of Multiple Regression Analysis in Projecting the Water Demand for the City of Cape Town.” *Water Practice and Technology* 13 (3): 705–11. https://doi.org/10.2166/wpt.2018.082.

Liu, Jing. 2010. “Chapter 80 - Phenylurea Herbicides.” In *Hayes’ Handbook of Pesticide Toxicology (Third Edition)*, edited by Robert Krieger, 1725–31. New York: Academic Press. https://doi.org/10.1016/B978-0-12-374367-1.00080-X.

Louchart, X., M. Voltz, P. Andrieux, and R. Moussa. 2001. “Herbicide Transport to Surface Waters at Field and Watershed Scales in a Mediterranean Vineyard Area.” *Journal of Environmental Quality* 30 (3): 982–91. https://doi.org/10.2134/jeq2001.303982x.

Ma, Bin, Peng Bao, Yan Wei, Guibing Zhu, Zhiguo Yuan, and Yongzhen Peng. 2015. “Suppressing Nitrite-Oxidizing Bacteria Growth to Achieve Nitrogen Removal from Domestic Wastewater via Anammox Using Intermittent Aeration with Low Dissolved Oxygen.” *Scientific Reports* 5 (1): 13048. https://doi.org/10.1038/srep13048.

“Marine Organisms Enlisted in Battle against Bacterial Sheets | Research and Innovation.” n.d. Accessed June 15, 2022. https://ec.europa.eu/research-and-innovation/en/horizon-magazine/marine-organisms-enlisted-battle-against-bacterial-sheets.

Melo, L. F., and M. M. Pinheiro. 1992. “Biofouling in Heat Exchangers.” In *Biofilms — Science and Technology*, edited by L. F. Melo, T. R. Bott, M. Fletcher, and B. Capdeville, 499–509. NATO ASI Series. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-1824-8_44.

Minhas, Manpreet Singh. 2021. “Techniques for Handling Underfitting and Overfitting in Machine Learning.” Medium. June 5, 2021. https://towardsdatascience.com/techniques-for-handling-underfitting-and-overfitting-in-machine-learning-348daa2380b9.

Mittapalli, Giridhar, and Dr Ramesh Chalumuri. 2014. “REGRESSION ANALYSIS TO CALCULATE IRRIGATION WATER REQUIREMENTS OF WAZIRABAD COMMAND AREA.” In .

"MTH 416 : Regression Analysis." n.d. Accessed June 24, 2022. https://home.iitk.ac.in/~shalab/course5.htm.

Mun, Johnathan. 2014. "Chapter 3 - A Primer on Quantitative Risk Analysis." In *Multi-Asset Risk Modeling*, edited by Morton Glantz and Robert Kissell, 63–118. San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-401690-3.00003-2.

Muñoz, I., M. Real, H. Guasch, E. Navarro, and S. Sabater. 2001. "Effects of Atrazine on Periphyton under Grazing Pressure." *Aquatic Toxicology (Amsterdam, Netherlands)* 55 (3–4): 239–49. https://doi.org/10.1016/s0166-445x(01)00179-5.

Ødegaard, H., B. Rusten, and T. Westrum. 1994. "A New Moving Bed Biofilm Reactor - Applications and Results." *Water Science and Technology* 29 (10–11): 157–65. https://doi.org/10.2166/wst.1994.0757.

Pesce, Stéphane, Christelle Margoum, and Arnaud Foulquier. 2016. "Pollution-Induced Community Tolerance for in Situ Assessment of Recovery in River Microbial Communities Following the Ban of the Herbicide Diuron." *Agriculture, Ecosystems & Environment* 221 (April): 79–86. https://doi.org/10.1016/j.agee.2016.01.009.

Piculell, Maria, Carolina Suarez, Chunyan Li, Magnus Christensson, Frank Persson, Michael Wagner, Malte Hermansson, Karin Jönsson, and Thomas Welander. 2016. "The Inhibitory Effects of Reject Water on Nitrifying Populations Grown at Different Biofilm Thickness." *Water Research* 104 (November): 292–302. https://doi.org/10.1016/j.watres.2016.08.027.

Piculell, Maria, Pia Welander, Karin Jönsson, and Thomas Welander. 2016. "Evaluating the Effect of Biofilm Thickness on Nitrification in Moving Bed Biofilm Reactors." *Environmental Technology* 37 (6): 732–43. https://doi.org/10.1080/09593330.2015.1080308.

Qu, Jingguo, Yuhuan Cui, Guanchen Zhou, and Qingpeng Ding. 2014. "Application of Multiple Linear Regression Model in the Performance Analysis of Traffic Rules," 6.

R R, Rajalaxmi, P Natesan, N. Krishnamoorthy, and S Ponni. 2019. "Regression Model for Predicting Engineering Students Academic Performance" 7 (April): 71–75.

Revilla, Marta, Berta Galán, and Javier R. Viguri. 2016. "An Integrated Mathematical Model for Chemical Oxygen Demand (COD) Removal in Moving Bed Biofilm Reactors (MBBR) Including Predation and Hydrolysis." *Water Research* 98 (July): 84–97. https://doi.org/10.1016/j.watres.2016.04.003.

Ricart, Marta, Damià Barceló, Anita Geiszinger, Helena Guasch, Miren López de Alda, Anna M. Romaní, Gemma Vidal, Marta Villagrasa, and Sergi Sabater. 2009. "Effects of Low Concentrations of the Phenylurea Herbicide Diuron on Biofilm Algae and Bacteria." *Chemosphere* 76 (10): 1392–1401. https://doi.org/10.1016/j.chemosphere.2009.06.017.

Russell, Bayden D., Sean D. Connell, Helen S. Findlay, Karen Tait, Stephen Widdicombe, and Nova Mieszkowska. 2013. "Ocean Acidification and Rising Temperatures May Increase Biofilm Primary Productivity but Decrease Grazer Consumption." *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1627): 20120438. https://doi.org/10.1098/rstb.2012.0438.

S, Pesce, Margoum C, and Montuelle B. 2010. "In Situ Relationships between Spatio-Temporal Variations in Diuron Concentrations and Phototrophic Biofilm Tolerance in a Contaminated River." *Water Research* 44 (6). https://doi.org/10.1016/j.watres.2009.11.053.

Sabater, Sergi, Helena Guasch, Marta Ricart, Anna Romaní, Gemma Vidal, Christina Klünder, and Mechthild Schmitt-Jansen. 2007. "Monitoring the Effect of Chemicals on Biological Communities. The Biofilm as an Interface." *Analytical and Bioanalytical Chemistry* 387 (4): 1425–34. https://doi.org/10.1007/s00216-006-1051-8.

Şanli, Önder. 2019. "Examining the Effect of Teachers' Perception of Psychological Empowerment on the Stress Level They Perceive." *Journal of Education and Training Studies* 7 (July): 98. https://doi.org/10.11114/jets.v7i8.4283.

Sarstedt, Marko, and Erik Mooi. 2014. “Regression Analysis.” In , 193–233. https://doi.org/10.1007/978-3-642-53965-7_7.

Sekar, R., V.P. Venugopalan, K.K. Satpathy, K.V.K. Nair, and V.N.R. Rao. 2004. “Laboratory Studies on Adhesion of Microalgae to Hard Substrates.” *Hydrobiologia* 512 (1): 109–16. https://doi.org/10.1023/B:HYDR.0000020315.40349.38.

Sgier, Linn, Renata Behra, René Schönenberger, Alexandra Kroll, and Anze Zupanic. 2018. “Evaluation of Phototrophic Stream Biofilms Under Stress: Comparing Traditional and Novel Ecotoxicological Endpoints After Exposure to Diuron.” *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.02974.

Soria, Juan J., Orlando Poma, David A. Sumire, Joel Hugo Fernandez Rojas, and Sulamita Marinela Ramos Chipa. 2022. “Multiple Linear Regression Model of Environmental Variables, Predictors of Global Solar Radiation in the Area of East Lima, Peru.” *IOP Conference Series: Earth and Environmental Science* 1006 (1): 012009. https://doi.org/10.1088/1755-1315/1006/1/012009.

Sousa, Sofia, Fernando Martins, M.C. Pereira, M.C.M. Alvim-Ferraz, Helena Ribeiro, Manuela Oliveira, and Ilva Abreu. 2010. “Use of Multiple Linear Regressions to Evaluate the Influence of O 3 and PM 10 on Biological Pollutants.” *World Academy of Science, Engineering and Technology* 37 (January): 935–40.

Thompson, L.j., V. Gray, D. Lindsay, and A. Von Holy. 2006. “Carbon : Nitrogen : Phosphorus Ratios Influence Biofilm Formation by Enterobacter Cloacae and Citrobacter Freundii.” *Journal of Applied Microbiology* 101 (5): 1105–13. https://doi.org/10.1111/j.1365-2672.2006.03003.x.

“Total Quality Management for Custodial Operations: A Guide to Understanding and Applying the Key Elements of Total Quality Management.” n.d. Routledge & CRC Press. Accessed June 13, 2022. https://www.routledge.com/Total-Quality-Management-for-Custodial-Operations-A-Guide-to-Understanding/audreau/p/book/9781884015519.

“Underfitting and Overfitting in Machine Learning.” n.d. Accessed June 14, 2022. https://datascience.foundation/sciencewhitepaper/underfitting-and-overfitting-in-machine-learning.

US EPA, ORD. 2015. “Nutrients.” Data and Tools. November 4, 2015. https://www.epa.gov/caddis-vol2/nutrients.

US EPA, OW. 2016. “Understanding the Science of Ocean and Coastal Acidification.” Overviews and Factsheets. September 8, 2016. https://www.epa.gov/ocean-acidification/understanding-science-ocean-and-coastal-acidification.

Villeneuve, A., Bernard Montuelle, and Agnès Bouchez. 2010. “Influence of Slight Differences in Environmental Conditions (Light, Hydrodynamics) on the Structure and Function of Periphyton.” https://hal.inrae.fr/hal-02594603.

Wagner, Karoline, Katharina Besemer, Nancy R. Burns, Tom J. Battin, and Mia M. Bengtsson. 2015. “Light Availability Affects Stream Biofilm Bacterial Community Composition and Function, but Not Diversity.” *Environmental Microbiology* 17 (12): 5036–47. https://doi.org/10.1111/1462-2920.12913.

Wang, Qilin, Liu Ye, Guangming Jiang, Shihu Hu, and Zhiguo Yuan. 2014. “Side-Stream Sludge Treatment Using Free Nitrous Acid Selectively Eliminates Nitrite Oxidizing Bacteria and Achieves the Nitrite Pathway.” *Water Research* 55 (May): 245–55. https://doi.org/10.1016/j.watres.2014.02.029.

Wang, Shuai, Sudeep Parajuli, Vasan Sivalingam, and Rune Bakke. 2019. *Biofilm in Moving Bed Biofilm Process for Wastewater Treatment*. *Bacterial Biofilms*. IntechOpen. https://doi.org/10.5772/intechopen.88520.

White, Paul A., Jacob Kalff, Joseph B. Rasmussen, and Josep M. Gasol. 1991. “The Effect of Temperature and Algal Biomass on Bacterial Production and Specific Growth Rate in Freshwater and Marine Habitats.” *Microbial Ecology* 21 (1): 99–118. https://doi.org/10.1007/BF02539147.

Williams, Charles Gbenga, and Oluwapelumi O. Ojuri. 2021. “Predictive Modelling of Soils’ Hydraulic Conductivity Using Artificial Neural Network and Multiple Linear Regression.” *SN Applied Sciences* 3 (2): 152. https://doi.org/10.1007/s42452-020-03974-7.

Xu, Guangjing, Yan Zhou, Qin Yang, Zarraz May-Ping Lee, Jun Gu, Winson Lay, Yeshi Cao, and Yu Liu. 2015. “The Challenges of Mainstream Deammonification Process for Municipal Used Water Treatment.” *Applied Microbiology and Biotechnology* 99 (6): 2485–90. https://doi.org/10.1007/s00253-015-6423-6.