Submitted to the Annals of Applied Statistics

STAT 444 FINAL REPORT

BY SHAKIR RAHMAN[1,a]

[1] University of Waterloo, [a] s232rahm@uwaterloo.ca

1. Introduction.

This report shall determine whether a model that outperforms SeaWif's satellite estimates of chlorophyll concentration levels at various depths in the ocean is possible. Using the 'chl' dataset from the 'gamair' package, this investigation will commence with an exploratory data analysis (EDA) to identify the significance of various features, their potential correlations, and to prepare the dataset for modeling. Then, we aim to compare the efficacy of various modeling techniques, and after selecting the best model we aim to determine whether this model surpasses the predictive accuracy of SeaWif's chlorophyll estimates.

2. Data.

We leverage the 13840 observations of the 'chl' dataset stemming from the 'gamair' package. The features are as follows: 'lon' and 'lat' are the longitude and latitude the chlorophyll data is collected from(in degrees); 'jul.day' is the day of the year the concentration was collected; 'bath' is the depth that the concentration is sampled from; and finally 'chl' and 'chl.sw' are the concentration from a direct sample and from the satellite imaging respectively.

This data was originally collected to compare SeaWif's satellite estimates to the true chlorophyll concentration at various depths of the ocean. The idea was that if the satellite predictions were accurate enough, researchers who required chlorophyll concentration data would not need to collect direct bottle samples, as such a process is laborious and costly. In the end, SeaWif concluded that their satellite estimates were not accurate enough to warrant their use, but they posted all the data collected for public use. The reason being that they believed statistical methods could create a model that outperforms their satellite estimates. This report will address that belief with a conclusive answer.

2.1. *Exploratory Data Analysis.*

We proceed with a brief exploratory data analysis to ensure our data is fit for modelling. Firstly, we note that our data is spatially referenced, and due to the variation in 'lon' and 'lat' values, some of these correspond to land mass areas. In these instances our model will predict the chlorophyll concentration assuming the spacial data refers to the ocean, which is obviously inaccurate, hence a major drawback. Consequently, we will constrain our dataset to the observations with $-60 \leq$ 'lon' ≤ -10 and $30 \leq$ 'lat' ≤ 65 leaving us with 3559 observations. As we aim to determine the efficacy of our model in relation to SeaWif's estimates, we will not consider the satellite estimates in our analysis(chl.sw).

2.1.1. *Output.*

By creating a histogram of 'chl' values, we see the plot fails to follow a normal distribution. Raising the 'chl' values to the power of 0.2 fixes this as desired. We also create a box-plot of our modified 'chl' values, and see many outliers as shown in Figure 2.1.1.1. We remove these outliers and end up with 3490 observations.

2.1.2. *Features.*
Similarly, for the features we create histograms and box-plots. From our histograms, all of them follow a normal distribution as desired. Regarding the box-plots, both 'jul.day' and 'latitude' have outliers and so removing these observations leave us with 3209 observations remaining.

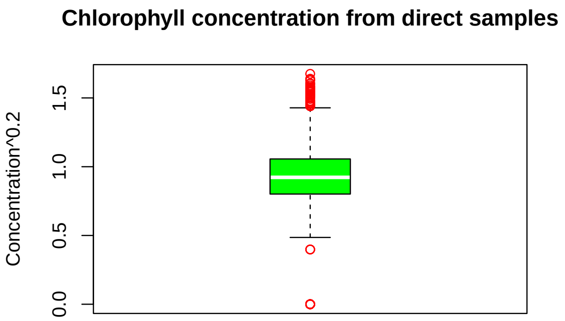

Figure 2.1.1.1: *Boxplot of modified 'chl' values*

Furthermore, we need to examine whether these features seem to have any effect on our output. We proceed by way of scatter-plots and the superimposed linear model on top. We have all scatter-plots are similar to Figure 2.1.2.1, suggesting that all features seem to have some significance regarding the true chlorophyll concentration. Checking for multi-collinearity in figure 2.1.2.2, it seems that 'lon' and 'bath' are quite related, but we do not remove either of these features as there does not seem to be extreme correlation between the two.

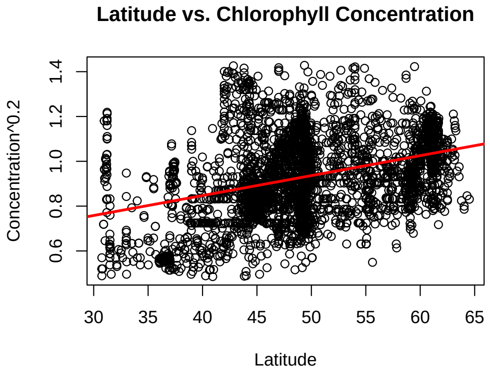

Figure 2.1.2.1: *Scatterplot of Depth vs. chlorophyll Concentration*

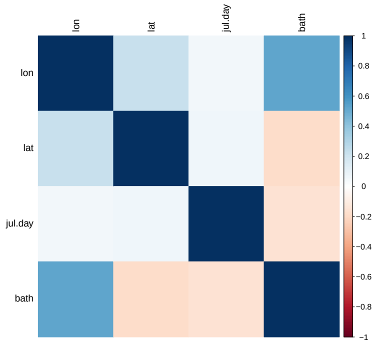

Figure 2.1.2.2: *Correlogram of features*

As we've accounted for any issues with our data, we are now ready to proceed to our methodology.

3. Methodology. We would like to cover the various models we will fit to our data, why we chose them, their specifications, and how we fit them. We note that all coding is done in R.

3.1. *Ordinary Least Squares.* The Ordinary Least Squares model is a linear regression model where the relationship between the response variable y and the predictor variables $\mathbf{X}$ is modeled as $y = \mathbf{X}\beta + \epsilon$ such that y is the $n \times 1$ vector of observed response values., $\mathbf{X}$ is the $n \times p$ matrix of predictor variables, β is the $p \times 1$ vector of regression coefficients, and ϵ is the $n \times 1$ vector of errors such that $\epsilon \sim N(0, \sigma^2)$.

Furthermore, the Ordinary Least Squares estimates of the regression coefficients are obtained by minimizing the sum of squared residuals $\hat{\beta} = \arg\min_\beta (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)$ with the solution being $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$

3.1.1. *Assumptions.* We also have that Ordinary Least Squares relies on four key assumptions, those being: the relationship between predictors and response is linear; the observations are independent from each other; the variance of error terms are constant; and the error terms are normally distributed.

3.1.2. *Fitting.* To fit this model we use the 'lm' function to fit all possible combinations of features in regards to the response variable(chl), and compare their GCV values. The model we select at the end of this process will have the lowest GCV and the fewest features possible.

3.2. *Additive.* Additive models are a type of regression model where the relationship between the response variable y and the predictors $x_1, x_2, \ldots, x_p$ is expressed as the sum of smooth functions of the predictors. If we let $Y =$ 'chl', $\mathbf{x_1} =$ 'lon', $\mathbf{x_2} =$ 'lat', $\mathbf{x_3} =$ 'bath', $\mathbf{x_4} =$ 'jul.day' we have the following.

$$\mathbb{E}(Y \mid x_1, x_2, x_3, x_4) = \alpha + \mathbf{f_1}(\mathbf{x_1}) + \mathbf{f_2}(\mathbf{x_2}) + \mathbf{f_3}(\mathbf{x_3}) + \mathbf{f_4}(\mathbf{x_4}) + \epsilon$$

where $f_i(x_i)$ are the smooth terms.

3.2.1. *Smoothing Terms.* We will estimate the smoothing terms $f_i(x_i)$ by way of penalized regression splines. These penalized splines will find terms that minimize the penalized residual sum of squares as defined below.

$$\hat{f}_i = \arg\min_{f_i} \sum_{i=1}^{n} (y_i - f_1(x_{i1}) - \ldots - f_p(x_{ip}))^2 + \sum_{j=1}^{p} \lambda_j \int (f_j'')^2 dx$$

From the above, we have that $\sum_{i=1}^{n} (y_i - f_1(x_{i1}) - \ldots - f_p(x_{ip}))^2$ is the residual sum of squares, $\sum_{j=1}^{p} \lambda_j \int (f_j'')^2 dx$ is the smoothness penalty for each function f_j and the λ_j such that $j \in [1, \ldots, p]$ are smoothing parameters that control the trade-off between goodness of fit and smoothness of the functions.

3.2.2. *Constraints.* We also have that our Additive model must satisfy some underlying constraints.

The smoothness constraint ensures that the roughness of each f_j term is minimized as much as possible, where the roughness is measured by $\int (f_j'')^2 dx$. We note that the magnitude of the smoothness is controlled by a λ_j parameter as shown in the smoothness penalty.

For identifiability, we impose the sum-to-zero constraint on the smooth terms. This means that $\sum_{i=1}^{n} f(x_i) = 0$, and this constraint ensures the mean of the smoothing terms is zero over the range of data. This prevents the smoothing terms from having any arbitrary offset, ensuring that our model avoids ambiguity.

3.2.3. *Fitting.* We'll fit our additive model using the 'gam' function on all features. Luckily, the 'gam' function automatically selects the optimal smoothing parameters but we need to find the optimal knot values manually. We do so by a coarse into fine grid search, in which we begin with a large range of k values per parameter and find the model that performs the best. The best model will have the lowest penalized GCV score where the penalized GCV is obtained by the formula $GCV \cdot (1 + \lambda \frac{k'}{n})$ where $\lambda = 2$, k' is the sum of all knots in the model and n is the number of observations used to train the model. Once we find the best penalized GCV value, we repeat the search with a finer grid, until the results are stagnant or have very little deviation to warrant any further iterations.

3.3. *Spatial.* Spatial models are extremely similar to additive models, so much so that they can be considered as an extension of an additive model with the following caveat.

$$\mathbb{E}(Y \mid x_1, x_2, x_3, x_4) = \alpha + \mathbf{f_1}(\mathbf{x_1}, \mathbf{x_2}) + \mathbf{f_3}(\mathbf{x_3}) + \mathbf{f_4}(\mathbf{x_4}) + \epsilon$$

Essentially, a Spatial model considers the pairwise combination of longitude and latitude values and their effect on the response, while in an Additive model the longitude and latitude values are separated into their own smoothing terms. Specifics regarding the smooth terms, constraints and assumptions all remain the same as the Additive model.

3.3.1. *Fitting.* We will fit our Spatial model in the same way as our 'gam' model. We still need to find the optimal knot values, but this time we only need to find knot values for three smoothing terms as 'lon' and 'lat' are combined into one smoothing term. We do this by the same grid search approach, with the same penalized GCV calculation for comparison.

4. Results. We now proceed with the results of our model fitting.

4.1. *Ordinary Least Squares.* We fit all possible combinations of features to our response variable for a total of 16 different models. We decide to choose the model with only 'lat' and 'jul.day' for its features, as it has the best GCV score(0.02723227) with the least parameters. We examine the qqplot and residuals vs. fitted values in Figures 4.1.1 and 4.1.2 respectively. The residuals vs. fitted values plot seems to not be evenly scattered around zero, and so this model may not be the best choice for us.

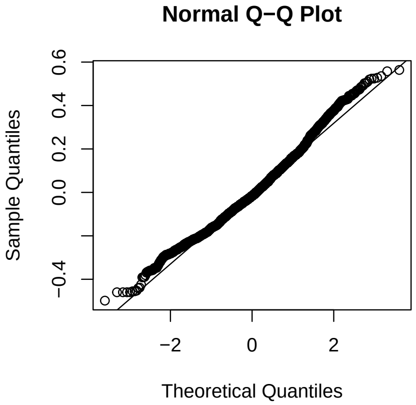

Figure 4.1.1: *QQ Plot of OLS Model*

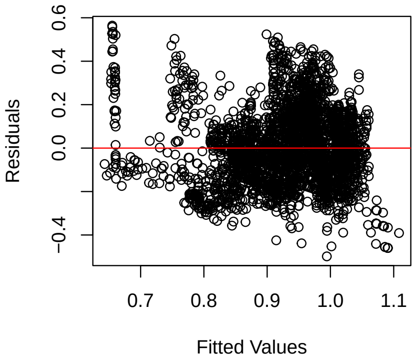

Figure 4.1.2: *OLS Residuals vs. Fitted Values*

4.2. *Additive Model.* Regarding the Additive model, we first discuss the results of our grid search. As this is the first instance of our grid search in this report, we shall list our iteration results below for better understanding. Further grid searches will not include a figure like below.

Iteration	'lon' k	'lat' k	'bath' k	'jul.day' k	Best subset
1	[100, 150]	[100, 150]	[10, 20, 30]	[30, 40, 50]	[100, 100, 10, 50]
2	[90, 100, 110]	[90, 100, 110]	[8, 9, 10, 11, 12]	[50, 60]	[90, 90, 8, 60]
3	[70, 80, 90]	[70, 80, 90]	[5, 6, 7, 8]	[60, 70, 80]	[80, 70, 5, 70]
4	[75, 77, 80, 83, 85]	[68, 70, 72]	[5]	[65, 70, 75]	[80, 70, 5, 75]

Figure 4.2.1: *Grid Search Results*

Figure 4.2.1 leaves us with the optimal knots for 'lon', 'lat', 'bath', and 'jul.day' as 80, 70, 5, and 75 as desired. If we fit the additive model to these values, we have that the QQ plot is similar to that of the linear model, which is good, and the Residual vs. Fitted Values plot is better than in the OLS plot as the values are more evenly scattered around zero, but not perfectly even. Below we plot some of the smoothing terms from the additive model.

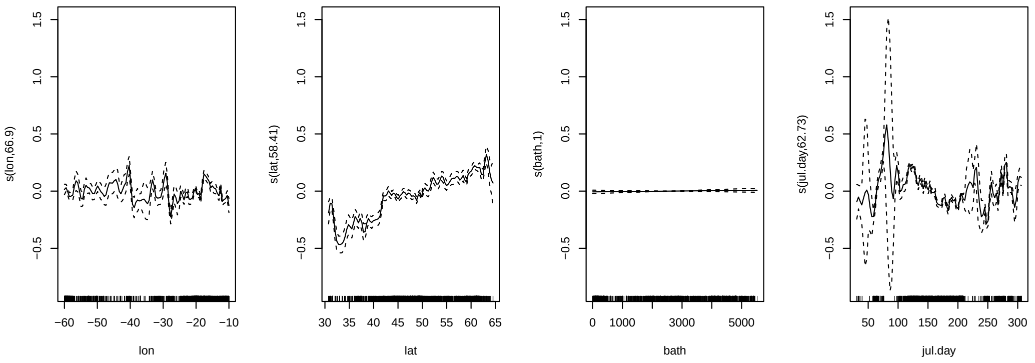

Figure 4.2.2: *Additive Model Smooth Terms*

We can see from the above plots, that clearly 'bath' does not have any significant effect on the response. As a result we repeat our grid search with the 'bath' feature excluded from the model. The search yields the optimal knots as $80, 70$, and 75 for 'lon', 'lat', and 'jul.day' respectively. Plotting the model diagnostic yields plots that all indicate a significant relationship to the response, and our QQ and Residual plots are similar as the first Additive model. Our GCV score is 0.012712, with an adjusted R^2 of 0.647 which largely improves on the OLS model.

4.3. *Spatial Model.* Similarly to the Additive model, we first discuss the grid search results. The optimal knots for 'lon', 'lat', 'bath', and 'jul.day' are found to be 300, 5, and 52 by the grid search. If we fit the Spatial model to these values, and plot the smoothing terms, we obtain plots extremely similar to those of Figure 4.2.2. Thus, once again 'bath' has practically no effect on the response variable and so we repeat our grid search with the 'bath' feature excluded from the model. The search yields 304 and 50 for '(lon, lat)', and 'jul.day' as desired. Diagnostics follow below.

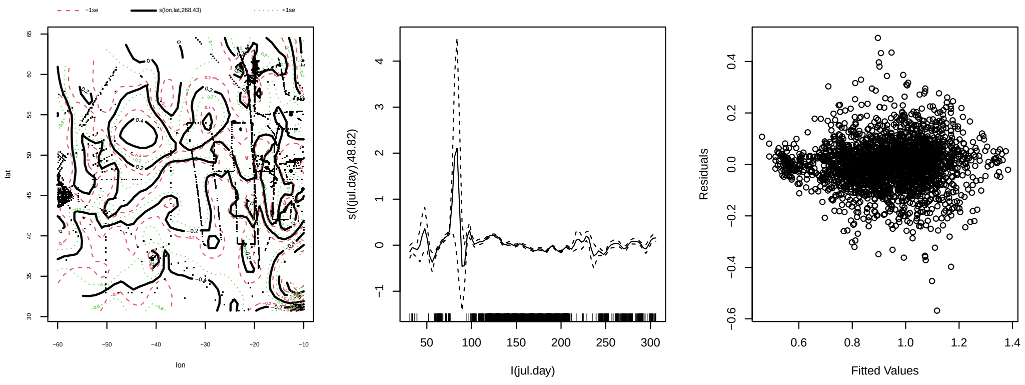

Figure 4.3.1: *Spatial Model Diagnostics*

The diagnostics in Figure 4.3.1 show that chlorophyll concentrations differ based on longitude and latitude, that the day of the year has a significant relationship regarding the concentration, and our Residual vs. Fitted Values plot is better than all other models. Our GCV score is 0.010223, with an adjusted R^2 of 0.73. We see how the Spatial model is a significant improvement upon the Additive.

4.4. *Top Model.* We summarize the results of our models into Figure 4.4.1.

Model	**Covariates**	**Adj. R^2**	**GCV**
OLS	Lon, Lat, Depth, Day	0.1934	0.02723227
GAM	(Lon, Lat), Day	0.647	0.012712
Spatial	(Lon, Lat), Day	0.73	0.010223

Figure 4.4.1: *Model Performance Comparison*

From the summarized results above and our discussion regarding the plots surrounding the various models, the best model is our Spatial model.

4.5. *Comparison to SeaWif's estimates.* We now use the Spatial model and assess its predictions regarding the 3209 total observations in comparison to the 'chl.sw' estimates of these observations by way of K-fold Cross Validation.

Predictions	**MSE**	**RMSE**	**MAE**
chl.sw	0.2093198	0.4561507	0.3603347
Spatial	0.0108181	0.1039104	0.0752064

Figure 4.5.1: *Prediction Comparison*

5. Conclusion. Through careful consideration of environmental factors, and numerous modelling techniques in our analysis, we have determined that it is indeed possible to create a statistical model that outperforms SeaWif's satellite estimates for chlorophyll concentration. While the current Spatial model performs significantly better than SeaWif's satellite estimates, there is room for further exploration through alternate models and more sophisticated modelling techniques. Most notably was our conclusion that the depth a bottle sample is taken at has no effect on the chlorophyll concentration. One would think depth, which is a prominent factor in marine ecosystems, would have a notable effect on chlorophyll concentration levels. The lack of a significant relationship in our analysis might suggest that our Spatial model may not capture all the relationships fully. Perhaps modelling techniques like Random Forests or Generalized Additive Mixed Models could expose the relationship between the depth and chlorophyll concentrations. On the other hand, there is also the fact that in our Spatial model, the most accurate methods were not used due to computational constraints. Thin Plate splines immediately come to mind in this regard as they are very useful for capturing spatial relationships, however they are extremely computationally expensive. In all, while we have created a model that vastly outperforms SeaWif's satellite estimates, there is still lots of analysis to be done to determine if the model we created truly captures the entirety of the relationship between chlorophyll concentration levels and its various influences.

REFERENCES

FELDMAN, G. (2003). *SeaWiFS Project*, Code 970.2. Greenbelt, Maryland.