Open Journal of Marine Science, 2014, 4, 25-30
Published Online January 2014 (http://www.scirp.org/journal/ojms)
http://dx.doi.org/10.4236/ojms.2014.41004

Modeling Ocean Chlorophyll Distributions by Penalizing the Blending Technique

Mathias A. Onabid[1], Simon Wood[2]
[1]Department of Maths-Computer Science, Faculty of Sciences, University of Dschang, Dschang, Cameroon
[2]Department of Mathematical Sciences, University of Bath, Bath, UK
Email: mathakong@yahoo.fr, s.wood@bath.ac.uk

Received September 6, 2013; revised October 19, 2013; accepted November 12, 2013

Copyright © 2014 Mathias A. Onabid, Simon Wood. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. In accordance of the Creative Commons Attribution License all Copyrights © 2014 are reserved for SCIRP and the owner of the intellectual property Mathias A. Onabid, Simon Wood. All Copyright © 2014 are guarded by law and by SCIRP as a guardian.

ABSTRACT

Disparities between the *in situ* and satellite values at the positions where *in situ* values are obtained have been the main handicap to the smooth modeling of the distribution of ocean chlorophyll. The blending technique and the thin plate regression spline have so far been the main methods used in an attempt to calibrate ocean chlorophyll at positions where the *in situ* field could not provide value. In this paper, a combination of the two techniques has been used in order to provide improved and reliable estimates from the satellite field. The thin plate regression spline is applied to the blending technique by imposing a penalty on the differences between the satellite and *in situ* fields at positions where they both have observations. The objective of maximizing the use of the satellite field for prediction was outstanding in a validation study where the penalized blending method showed a remarkable improvement in its estimation potentials. It is hoped that most analysis on primary productivity and management in the ocean environment will be greatly affected by this result, since chlorophyll is one of the most important components in the formation of the ocean life cycle.

KEYWORDS

***In Situ*, Satellite; Ship and Buoy; Penalized Regression Spline; Penalty; Penalized Blending**

1. Introduction

A detailed study of the ocean environment and its constituent elements are of utmost importance in guiding decision-makers on policies regarding marine activities such as fishing and their consequences for human life and society as a whole. In the ocean food chain, phytoplankton, which are found in the upper layer of the ocean, are of extreme importance. Indeed, aquatic life and production revolve about the distribution and biomass of these unicellular algae. Thus, to better understand the ocean food chain, it is necessary to track their existence and monitor their population distribution in the ocean environment. To measure their population by cell counts is very difficult, because of their resemblance to other non-algae carbon rich particles. An alternative method of doing this is in terms of their photosynthetic pigment content, *chlorophyll*, which is endemic across all taxonomic groups of algae [1]. In fact, an appealing method of estimating primary productivity in the ocean is determined by the concentration of ocean chlorophyll [2] and also emphasized by [3]. Therefore, to better monitor and predict the abundance of this phytoplankton, it is important that the distribution of chlorophyll concentration in this environment be determined as accurately as possible. The blending technique described by [4] was successfully used to analyze sea surface temperature [5]. The pioneers of the use of this technique in the calibration of ocean chlorophyll expressed the need for further work to be done in order to improve ocean chlorophyll predictions in areas where observations could not be obtained by ship and buoy [6]. One problem faced when using the technique in ocean chlorophyll calibration is distortion of the blended field as one approaches the coastal land. This distortion is due to the sparseness of

data obtained by ship and buoy (*in situ*) and the noisiness of the satellite field [7]. These factors have been the main handicap to the smooth calibration of ocean chlorophyll estimates from the satellite observations. The penalized regression splines are a technique that could be used to model noisy data [8]. The use of this statistical technique in the calibration of ocean chlorophyll was also suggested by [1].

The objective of this article, therefore, is to demonstrate how the principles of penalized regression could be applied to the blending process in order to obtain better estimate of ocean chlorophyll from the satellite data field. The approach would mainly address the noisiness of the data fields by introducing a penalty on the differences between their observations at positions where both fields have values. The belief is that, by penalizing the differences between the satellite and the *in situ* fields, the satellite will become closer to the *in situ* field and can thus be used to sufficiently estimate ocean chlorophyll values at positions where the *in situ* field could not provide values. Since the process of penalization involves smoothing, the efficiency of the technique will depend on the choices of the smoothing parameters.

Inspiration to this was drawn from the interpolation equation

$$f_{\text{blend}}(x,y) = f_{\text{sat}}(x,y)\sum_{k=1}^{n} g_k(x,y) \qquad (1)$$

found in Onabid (2011) in the section dealing with the proof as to why results from the corrector factor and the smooth in-fill methods should coincide. From this equation, the term of interest is

$$\sum_{k=1}^{n} g_k(x,y)$$

which is the sum of the solution to the partial differential equation obtained at each boundary point k where there is a difference between the satellite and *in situ* value. In order to penalize these differences, the interpolation Equation (1) had to be represented using basis functions. Consider the equation

$$f_{\text{blend}}(x,y) = f_{\text{sat}}(x,y)\sum_{k=1}^{n} g_k(x,y),$$

where g_k is actually the solution to

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = 0, \qquad (3)$$

subject to the boundary conditions

$$\{0; x_i; y_i : i = 1;\cdots; k-1; k+1;\cdots; n; \Delta_k; x_k; y_k\}.$$

This Equation (2) can be re-written with each of the g_k separately as

$$f_{\text{blend}}(x,y) = f_{\text{sat}}(x,y)\sum_{k=1}^{n} \beta_k \Delta_k(x,y), \qquad (4)$$

where β_k is set to the difference between the *in situ* and satellite values at boundary point k and $\Delta_k(x; y)$ representing the *basis function* is the solution to

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = 0 \qquad (5)$$

with external boundary points set to zero and the internal boundary points set to zero everywhere except at the k^{th} position where it is set to 1.0, that is the *knot* of the basis.

What this means is that, for each internal boundary point (*knot*), the blending process is performed to estimate the entire blended field with that particular boundary point acting as the only boundary point for the process. During the process, the value of this boundary point equals 1.0 and the resulting field is the basis for this *knot*. The blended field corresponding to this particular *knot* is obtained by multiplying the original *knot* value with its basis. Blended fields obtained from each of the knots are summed up. This sum is then added to the satellite field to obtain the final blended field which we call the *basis blend*.

2. Penalizing the Blending Process

For the penalized regression spline to be applied, it was necessary to represent the term of interest in the blending process as a regression equation.

Representing Blending as a Regression Equation

Considering the Equation (4) which is the interpolation form of the blending process represented using basis functions, also consider the fact that the objective is to control the differences between the satellite and the *in situ* fields, it is obvious that focus here should be on the term

$$\sum_{k=1}^{n} \beta_k \Delta_k(x,y),$$

from where the $\beta_k s'$ could be estimated by penalized least squares in order to minimize the effect of these differences and consequently maximize the use of the satellite field as estimate to ocean chlorophyll at points where *in situ* could not provide observations

Let

$$Z_k = \log(\text{insitu}_k) - \log(\text{satellite}_k);$$

be calculated for each point it *K*, where satellite and *in situ* have observations. This can be written as a regression equation of the form,

$$Z_k = \sum_{j=1}^{n} \beta_j \Delta_j(x_j, y_j) + \varepsilon_k \qquad (6)$$

where the $\beta_j s'$ are unknown parameters to be estimated

and ε_k the error term. This expression is equal to

$$Z_k = \beta_k \Delta_k (x_k, y_k) + \varepsilon_k$$

Thus if Z_k is expressed using the basis space, one obtains this model;

$$Z_k = \beta_1 \Delta_1 (x_1 y_1) + \beta_2 \Delta_2 (x_2 y_2) + \beta_3 \Delta_3 (x_3 y_3) + \cdots + \beta_n \Delta_n (x_n y_n) + \varepsilon_k$$

Fitting this model by least squares will simply result in the interpolation scheme since there is exactly one parameter per datum, thus nonparametric techniques were then explored. The thin plate regression spline was then used to introduce a penalty to this blending regression equation.

3. Penalizing the Blending Regression Equation

From Equation (6), the control of the smoothness of the differences can be achieved by either altering the basis dimension, that is changing the number of selected knots or keeping the basis dimension fixed and then adding a *penalty term* to the least squares objective. The later was used. Therefore the penalized least squares objective will be to minimize

$$\sum_{k=1}^{n} \left[Z_k - \beta_k \Delta_k (x_k, y_k) \right]^2 + \lambda \mathbf{J} \left(\beta_k \Delta_k (x_k, y_k) \right) \quad (7)$$

where $\mathbf{J}$ is a penalty function which penalizes model wiggliness while model smoothness is controlled by the smoothing parameter λ, as described by [9]. As a first step in estimating the penalized least squares objective, the simple penalized least squares technique of ridge regression was used. In this process, the intention is to penalize each of the parameters separately by introducing a penalty to each of the estimated parameters. Following this method, the penalized least squares objective will be to minimize

$$Q_p(\beta) = \sum_{k=1}^{n} \left[Z_k - \beta_k \Delta_k (x_k, y_k) \right]^2 + \lambda_k \sum_{k=1}^{n} \beta_k^2 \quad (8)$$

with respect to $\beta_k s'$: The penalty is represented by the term $\lambda_k \sum_{k=1}^{n} \beta_k^2$ with λ being the smoothing parameter to control the trade off between model fit and model smoothness. Thus the problem of estimating the degree of smoothness of the model is now the problem of estimating the smoothing parameter λ.

Assuming that the smoothing parameter is given, how then can the $\beta_k s'$ be estimated in this penalized least squares objective?

From Equation (8), the term $\Delta_k(x_k; y_k)$ reduces to a $n \times n$ identity matrix. Now, define an augmented Z, say $\mathbf{Z}$; as $\mathbf{Z} = \begin{bmatrix} Z_1 & \cdots & Z_n & 0 & \cdots & 0 \end{bmatrix}^{\mathrm{T}}$ (with n zeroes) which can also be augmented directly in the objective.

When this is done, Equation (8) could now be written as

$$Q_p(\beta) = \left\| \begin{bmatrix} Z \\ 0 \end{bmatrix} - \begin{bmatrix} I_n \\ \sqrt{\lambda} I_n \end{bmatrix} \beta \right\|^2$$

From here, $\hat{\beta}$ can be calculated as follows:

$$\hat{\beta} = \left(\begin{bmatrix} I_n & \sqrt{\lambda} I_n \end{bmatrix} \begin{bmatrix} I_n \\ \sqrt{\lambda} I_n \end{bmatrix} \right)^{-1} \begin{bmatrix} I_n & \sqrt{\lambda} I_n \end{bmatrix} \begin{bmatrix} Z \\ 0 \end{bmatrix}$$
$$= \left[I_n (1+\lambda) \right]^{-1} Z$$

with $\hat{\beta} = Z \dfrac{1}{1+\lambda}$; this implies that,

$$\left\| Z - I\hat{\beta} \right\|^2 = \left\| Z - Z \frac{1}{1+\lambda} \right\|^2 = \left\| Z \frac{\lambda}{1+\lambda} \right\|^2$$
$$= \|Z\|^2 \left(\frac{\lambda}{1+\lambda} \right)^2$$

3.1. Choosing How Much to Smooth

This refers to the selection of the smoothing parameter λ. This must be done with care such that the selected value should be suitable, so much so that if the true smooth function is f the estimated smooth function $\hat{f}$, should be as close as possible to it. The reason being if λ is too high, the data will be over-smoothed and if it is too low, the data will be under-smoothed hence the resulting estimate will not be close to the true function. The aim as described by [9] will be to select a λ which will minimize the difference between $\hat{f}$ and f that is to say if $\mathbf{M}$ is the difference, then λ should minimize

$$\mathbf{M} = \frac{1}{n} \sum_{i=0}^{n} \left(\hat{f}_i - f_i \right)^2$$

This could have been easier if the true values for f existed already. Because this is not the case, the problem was approached by deriving estimates of $\mathbf{M}$ plus some variation. This was achieved by making use of the ordinary cross validation (OCV) technique. In this technique, a model is fitted to the rest of the data, when a datum is left out. The squared difference between the datum and its predicted value from the fitted model is calculated. This is done for all the points and the mean taken over all the data. Thus the ordinary cross validation criterion is written as

$$\mathbf{V}_0 = \frac{1}{n} \sum_{i=0}^{n} \left(\hat{f}_i^{[-i]} - Z_i \right)^2$$

where $\hat{f}_i^{[-i]}$ is the estimate from the model fitted to all data except Z_i. The idea of calculating V_0 each time leaving out a datum has been proven not to be efficient as described by [9]. It can be shown that

$$\mathbf{V}_0 = \frac{1}{n}\sum_{i=0}^{n}\frac{\left(\hat{f}_i - Z_i\right)^2}{(1-A_{ii})^2}$$

where $\hat{f}$ is the estimate from fitting to all the data and **A** is the corresponding influence matrix.

[9] Emphasizes the fact that though OCV is a reasonable way of estimating smoothing parameters, it has the drawbacks of being computationally expensive to minimize in the case of additive models where there could be many smoothing parameters and secondly it has a slightly disturbing lack of invariance. Thus in practice, the weights $1-A_{ii}$ are often replaced by the mean weight $tr(\mathbf{I} - \mathbf{A})/n$ in order to arrive at the *generalized cross validation* (gcv) score given as

$$\mathbf{V}_g = \frac{n\sum_{i=0}^{n}\left(Z_i - \hat{f}_i\right)^2}{\left[tr\left(\mathrm{I}-\mathrm{A}\right)\right]^2}$$

This has the computational advantage over OCV and can also be argued to have some advantages in terms of invariance. Therefore, an easy way to look for the best smoothing parameter would be to search through a sequence of λ's, each time fitting a penalized regression model with the new λ value and calculating the gcv score. At the end, the λ value corresponding to the lowest gcv score will be the optimal smoothing parameter.

3.2. Calculating the gcv Score

Amongst the techniques of ridge regression, integrated least squares, integrated squared derivatives and efficient method used in computing the gcv score, only the efficient method herein described provided and better estimate for the gcv score.

Efficient Calculation of the gcv Score

The idea here is to provide a means of obtaining optimum values for the gcv score, the degree of freedom $tr(\mathbf{A})$ and the smoothing parameter λ which will minimize the gcv score. These will be very important since the objective is to build a model that will produce estimates in the blended field which are as close as possible to the true field. The QR decomposition described in [10] will be used because it is believed that QR is more stable than the Cholesky decomposition. This was achieved as follows.

The objective is to minimize

$$\|y - X\beta\|^2 + \lambda\beta^{\mathrm{T}} S\beta$$

with respect to β.

$$\Rightarrow X\hat{\beta} = Ay$$

where

$$A = X\left(X^{\mathrm{T}}X + \lambda S\right)^{-1}X^{\mathrm{T}}.$$

The corresponding gcv score for the given λ is then given as

$$V(\lambda) = \frac{n\|y - Ay\|^2}{\left[n - tr(A)\right]^2}$$

In order to calculate the efficient gcv score, let $X = QR$ where R is the upper triangle and Q consist of the columns of an orthogonal matrix such that $Q^{\mathrm{T}}Q = I$ but $QQ^{\mathrm{T}} \neq I$

$$\Rightarrow A = QR\left(R^{\mathrm{T}}R + \lambda S\right)^{-1}R^{\mathrm{T}}Q^{\mathrm{T}} = Q\left(I + \lambda R^{-\mathrm{T}}SR^{-1}\right)^{-1}Q^{\mathrm{T}}$$

From an eigen-decomposition

$$R^{-\mathrm{T}}SR^{-1} = UDU^{\mathrm{T}}$$

where D is a diagonal matrix of eigen values, the columns of U are eigenvectors and U is orthogonal.

$$\begin{aligned}\Rightarrow A &= QU\left(I + \lambda D\right)^{-1}U^{\mathrm{T}}Q^{\mathrm{T}}\\ \Rightarrow tr(A) &= \sum_i \frac{1}{1+\lambda D_{ii}}\end{aligned} \tag{9}$$

$$\begin{aligned}\|y - Ay\|^2 &= y^{\mathrm{T}}y - 2y^{\mathrm{T}}Ay + y^{\mathrm{T}}AAy\\ &= y^{\mathrm{T}}y - 2\hat{y}^{\mathrm{T}}\left(I + \lambda D\right)^{-1}\hat{y} + \hat{y}^{\mathrm{T}}\left(I + \lambda D\right)^{-2}\hat{y}\end{aligned} \tag{10}$$

where $\hat{y} = U^{\mathrm{T}}Q^{\mathrm{T}}y$.

From Equations (9) and (10) it follows that $V(\lambda)$ could be evaluated very cheaply for each new λ since the QR and eigen-decompositions are only needed once.

The smoothing parameter (λ) corresponding to each of these lowest gcv scores were then use in fitting the penalized regression models. The results obtained are then compared to those from the other techniques.

4. Validating the Blended Fields Obtained from the Various Blending Methods

The strength of this method in predicting existing *in situ* observation was compared to that of the normal blending method. Because penalizing the blending method had to make use of the basis function, the blended field obtained from the basis function method was also compared. Since the basis function method works by the use of a basis set (*knots*), after the selection of the validation data set the remaining *in situ* observations were then used as *knots* for the basis function blending method. The penalized blended field was obtained by using parameters obtained from the efficient method of calculating gcv as described in Section 3.2.1. Randomly selected validation data sets each containing 175 observations from the observed *in situ* data for the month of May were used in a validation study. May was selected because it had the highest num-

ber of observations in the *in situ* field. The mean squared differences between the predicted and the observed *in situ* values were computed and plotted (**Figure 1**).

There was not so much difference between predictions from the basis and the penalized. Though most of the times the differences are visible only after the third or fourth decimal place, there are a few times where the differences appear very distinct between the two in favour of the penalized blending method. Penalized models are always expected to perform better than non penalized ones. The poor performance here could have been caused by the choice of the smoothing parameter which is being obtained in this case by cross validation.

5. Discussions

We have been able to successfully establish a procedure for implementing smoothing on the blending process by making use of corrector factor blending technique model of [7]. This was achieved by expressing the interpolation formula used by the corrector factor blending technique in a form making use of the basis function. The aim of expressing the blending process using basis functions was to pave the way to implement penalization. This was implemented by adding a penalty term to the least squares objective. This term contained the penalty function which penalizes the model and a smoothing parameter to control the smoothness of the model. The main issue here was to be able to choose the right smoothing parameter such that the estimated smooth function should be as close as possible to the true function. Cross validation technique was used to obtain the smoothing parameter. To obtain the cross validation score three techniques were used, namely ridge regression, integrated least squares and the integrated squared derivative.

Calculating the cross validation score using ridge regression failed because the final expression for calculating the score did not depend on the smoothing parameter.

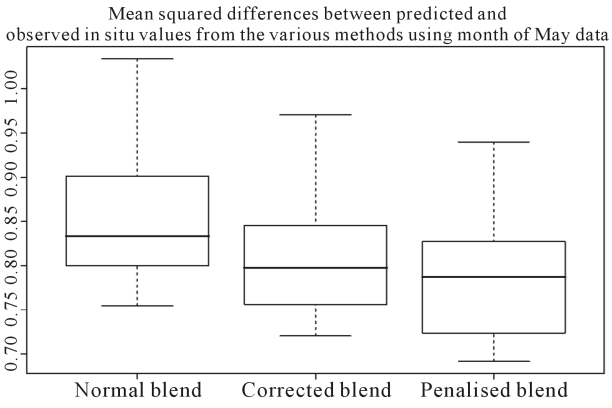

Figure 1. A box plot of the mean squared differences between predicted and observed *in situ* values from the different blending methods.

As described by [9], this is not surprising since if a Z_k is dropped from the model sum of squares term in equation (8), the only thing influencing the estimate of β_k would be the penalty term, which will be minimized by setting $\beta_k = 0$, whatever positive value the smoothing parameter takes. This complete decoupling will cause cross-validation to fail. Thus, if a datum is left out, its corresponding estimate will always be zero since no other data has influence on it. This behavior occurs for any possible value of the smoothing parameter.

Making use of the cross-validation score calculated from the integrated least squares did not improve on the results in this research. This again, according to [9], is not surprising because if one considers any three equally spaced points x_1, x_2, x_3 with corresponding $f(x_i)$ values to be μ_1, μ_2 and μ_3. Also, if $\mu_1 = \mu_3 = \mu^*$ then in order to minimize

$$\int_{x_1}^{x_3} \mu(x)^2 \, \mathrm{d}x$$

one should set $\mu_2 = \mu^*/2$. This condition does not hold for the data fields used in this research since the data fields were sparse, and the missing values were replaced by pseudo zeroes, so it was not uncommon to find a set of three adjacent points with similar values. In a situation like this, [9] states that, if the middle point is omitted from the fitting, the action of the penalty will send its estimate to the other side of zero from its neighbors. Meaning that a better prediction of the omitted datum will only be possible with a high smoothing parameter and this will be closer to zero since the high smoothing parameter will tend to shrink the values of the other included points towards zero and hence the omitted point. With this, cross validation will also have the tendency to always select an estimate for the omitted points closer to zero from the model. This could have been the cause of the poor results obtained. The integrated squared derivative penalty is not expected to suffer from the same problems faced by the previous methods. This is because the action of the penalty is simply to try and flatten the smooth function around the vicinity of the omitted datum. If the smoothing parameter is large, it will increase the flattening and consequently pulls the estimate far away from the omitted datum. The penalty obtained by this technique had very little or no effect on the smoothing function hence the equality in results from the penalized and the basis function model.

6. Conclusion

It is expected that a penalized model would be able to perform better than a non penalized model in a situation where penalization is necessary. Three techniques have been used to obtain penalty matrices in this research with the intention of improving the results from normal

blending method. The penalized model was obtained by first representing the blending method by making use of the basis function which was also considered as a model on its own. Even though the results from the basis function and penalized model were relatively identical since most differences occurred at the third or fourth decimal place, it is important to know that the difference between these methods and the normal blending method is quite alarming (**Figure 1**) and therefore should be encouraged especially if more data could be obtained from ship and buoy. With the emergence of this result, it is hoped that most of the analysis on primary productivity and management in the ocean environment will be greatly affected, since chlorophyll is one of the most important components in the formation of the ocean life cycle.

Future Work

The failure of the penalized blending regression models to perform better than the basis function model could have been because the right penalty was not obtained. Therefore, more work could be done towards obtaining other penalties. Maybe, an integrated squared second derivative could be tried or one could try a combination of the first and second derivatives (double penalization). To enable the blending process to be very close to reality, the possibility of extending it to three dimensions could be looked into.

REFERENCES

[1] E. Clarke, D. Speirs, M. Heath, S. Wood, W. Gurney and S. Holmes, "Calibrating Remotely Sensed Chlorophyll-a Data by Using Penalized Regression Splines," *Journal of Royal Statistics Society*, *Series C*, Vol. 55, No. 3, 2006, pp. 331-353.

[2] R. W. Eppley, E. Stewart, M. R. Abbott and U. Heyman, "Estimating Ocean Primary Production from Satellite Chlorophyll. Introduction to Regional Differences and Statistics for the Southern California Bight," *Journal of Plankton Research*, Vol. 7, No. 1, 1985, pp. 57-70. http://dx.doi.org/10.1093/plankt/7.1.57

[3] D. A. Flemer, "Chlorophyll Analysis as a Method of Evaluating the Standing Crop Phytoplankton and Primary Productivity," *Chesapeake Science*, Vol. 10, No. 3-4, 1969, pp. 301-306. http://dx.doi.org/10.2307/1350474

[4] A. H. Oort, "Global Atmospheric Circulation Statistics," NOAA Prof Paper 14 180pp Nat. Oceanic and Atmospheric Administration Silver Spring, Maryland, 1983.

[5] R. W. Reynolds, "A Real-Time Global Sea Surface Temperature Analysis," *Journal of Climate*, Vol. 1, No. 1, 1988, pp. 75-87. http://dx.doi.org/10.1175/1520-0442(1988)001<0075:ARTGSS>2.0.CO;2

[6] W. W. Gregg and M. E. Conkright, "Global Seasonal Climatologies of Ocean Chlorophyll: Blending *in Situ* and Satellite Data for Coastal Zone Colour Scanner Era," *Journal of Geophysical Research*, Vol. 106, No. C2, 2001, pp. 2499-2515. http://dx.doi.org/10.1029/1999JC000028

[7] M. A. Onabid, "Improved Ocean Chlorophyll Estimate from Remote Sensed Data: The Modified Blending Technique," *African Journal of Environmental Science and Technology*, Vol. 5, No. 9, 2001, pp. 732-747.

[8] S. N. Wood, "Thin Plate Regression Splines," *Royal Statistical Society*, *Series B*, Vol. 65, No. 1, 2003, pp. 95-114. http://dx.doi.org/10.1111/1467-9868.00374

[9] S. N. Wood, "Generalized Additive Models: An Introduction with R," Chapman and Hall/CRC, London, 2006.

[10] R. Scraton, "Further Numerical Methods in Basic," Edward Arnold Ltd, Kent, 1987.