

Autologistic Models With Autocorrelated Errors For an Aquatic Phenomenon

Bernadette S. Fermin and Erniel B. Barrios*

Received July, 2002; revised September, 2002

ABSTRACT

Spatial and autocorrelation structures are hypothesized to be important components in addition to the hydrological variables to describe the state of an aquatic phenomenon indexed by a binary response variable. Two models that incorporate spatial and autocorrelation parameters to an ordinary logistic model (autologistic) are proposed. An algorithm similar to the backfitting routine is proposed. The ordinary logistic model is used as a benchmark for comparison. Results indicate that the autologistic model is superior to the other two models in terms of its ability to recognize the state of a binary event. Inclusion of the spatial parameter in a logistic model is a significant improvement in the prediction of binary events.

KEY WORDS AND PHRASES: autocorrelation parameter, backfitting algorithm, binary response variable, harmful algal bloom (HAB), maximum pseudolikelihood, sampling stations, spatial parameter

1. INTRODUCTION

Harmful algal bloom (HAB) is the sudden increase in population of algae that produces toxins harmful to human as well as health of other organisms. Red tide (bloom of *Pyrodinium Sp.*) is a type of HAB very common in tropical waters. Incidence of red tide has drawn attention of researchers due to the hazards that it posed on the consumers of marine products. Monitoring is an essential component of a mitigation scheme for this phenomenon. There are three objectives of such monitoring: (1) to develop an early warning system on the presence of harmful algae; (2) to increase scientific understanding of the phytoplankton community; and (3) to assure that only safe shellfish are harvested to avoid health hazard for consumers.

This paper focuses on the development of a stochastic model to describe red-tide phenomenon and other HAB conditions. Existing methods for modeling aquatic species use the density or counts as the dependent variable. The behavior of the algal community is so volatile that their number at any point becomes unpredictable. From barely nil count, the number of cells could suddenly increase tremendously and just as suddenly drop to low level afterwards. These ridges in the frequency distribution of cell counts prompt the question of aptness of count data in modeling algal bloom. An indicator of "bloom" or "no bloom" of the organism could be a better indicator of the state of the algal community. The use of a binary response variable takes into account the abrupt changes in count data. Moreover, because of circulation and mixing in an aquatic environment, sampling stations are not necessarily independent. Hence, structure of dependence between sample stations should be integrated into the model to better explain the phenomenon. Since data are taken over a specified period of time, autocorrelation structure should also be considered.

* Ms. Fermin is an Instructor at the Mindanao State University-Iligan Institute of Technology while Dr. Barrios is an Associate Professor, School of Statistics, UP Diliman. Email Address: ernielb@yahoo.com

Several kinds of statistical models have been used to explain a binary response variable. For example, binary responses, denoted by y , can be related to some covariates through a logistic regression model. Given a set of k explanatory variables (x_1, x_2, \dots, x_k) , the logistic regression model utilizes the relationship $y = \text{logit}(p) = f(\mathbf{X}; \boldsymbol{\beta}) + \varepsilon$ as the description of the systematic component of the response y . Here p is $P(y=1)$, the ε 's are zero mean uncorrelated random variables with a common variance and $f(\mathbf{X}; \boldsymbol{\beta})$ is a known function f of the points x_i 's and parameters $\boldsymbol{\beta}$. Bonney(1987) fitted the logistic regression model for dependent binary observations and used regressive logistic model. Autologistic model (Besag, 1972) was proposed as a generalization of the standard logistic model for dependent binary data. Wu and Huffer (1997) used autologistic regression for modeling the presence or absence of plant species in terms of climate variables such as temperature and rainfall and taking into account the terms involving first-order neighborhood system, that is, using only the values at the four sites to predict the values of the central site. Investigation was focused on the performance of three estimation methods, the coding method (COD), maximum pseudo-likelihood method (MPL) and Markov Chain Monte Carlo method (MCMC). Gumpertz, et al. (1997) investigated the autologistic model in predicting the presence or absence of a disease in an agricultural field based on soil variables. The parameters were estimated using maximum pseudo-likelihood method.

Models and analyses that account for spatial heterogeneity have gained much attention in analyzing field data. Cressie (1993) explained that data close together in space are most likely to be correlated; thus, statistical independence can no longer be invoked. For example, nearest-neighbor methods for analyzing agricultural field trials indirectly attempted to take spatial dependence into account by using the residual from neighboring plot as covariates (Cullis and Gleeson, 1991).

In business and economics, many regression applications involve time series. For such data, the assumption of uncorrelated or independent error terms is often not appropriate. Furthermore, it is a known fact in time-series analysis that data close together in time usually exhibit higher dependence than those farther apart. When dependence over time prevails, models for conditional distribution of y_t given $y_{t-1}, y_{t-2}, \dots, y_1$ may be more appropriate (Liang and Zeger, 1991). In this same context, this paper formulates a binary model that incorporates dependence in terms of spatially correlated errors and other environmental covariates. Grobbelaar (1990) postulated that a deterministic model is not appropriate for biological phenomena especially for phytoplankton productivity in turbid water.

The determination of a suitable methodology for the estimation of the parameters is another issue discussed in this paper. Specifically, the performance of a procedure similar to the backfitting algorithm on the proposed models will be explored with phytoplankton blooms in mind.

In the next section we describe the proposed models. In Section 3, we briefly discuss the proposed estimation procedure. Section 4 discusses how the red-tide data was simulated. Section 5 demonstrates the predictive ability of the model; and section 6 presents the conclusions.

2. THE PROPOSED MODELS

Let X denote the $n \times p$ matrix of the hydrological variables used as covariates for the HAB phenomenon, $y_{ijt} = 1$ and $y_{ijt} = 0$ denote the response variable for “bloom” and “no bloom” at time t in site ij , respectively, and $y_{i-1,j,t}, y_{i+1,j,t}, y_{i,j-1,t}, y_{i,j+1,t}$ denote the first-order neighborhood sites.

Suppose the location of the sampling stations can be mapped into a rectangular lattice where each site has coordinate (i,j) that specifies the row and the column of the lattice at any time t (Figure 1). Each circle represents the location where the binary response variable y_{ijt} was measured.

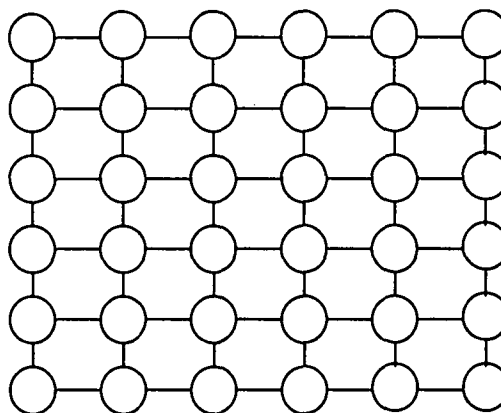



Figure 1. Rectangular Lattice Design

Given a specific site ij at time t , the first-order neighbors can be illustrated by Figure 2.  represents the neighbor of a given specific response variable.

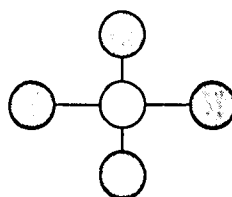


Figure 2. First-Order Neighborhood System

For binary response y_{ijt} , suppose it depends on the observed states of the first-order neighboring sites, on p environmental covariates and on random error following either uncorrelated or a first-order autoregressive process (AR(1)). We may express the response variable as $y_{ijt} = \mathfrak{R}(f(x_{ijt}, \beta_r), g(y_N, \theta), \varepsilon_t)$, i.e., y_{ijt} is a function of $f(X, \beta), g(y_N, \theta)$ and ε_t . The error terms ε_t can be uncorrelated or autocorrelated. For this paper, only autoregressive model of order 1 will be explored. $f(X, \beta)$ is a known function of the p environmental covariates, $g(y_N, \theta)$ is also a known function of the first-order neighboring sites, β and θ are vectors of unknown parameters and $y_{N_l}^{ijt}$, $l = 1, 2, 3, 4$ is an indicator

function (bloom/no bloom) of the first-order neighborhood of site ij at time t . In this paper, the neighborhood function is $y_N^{ijt} = \sum_{l=1}^4 y_{N_l}^{ijt}$ the sum of all first-order neighborhood values.

We shall investigate two models and an estimation procedure using the simulated data on phytoplankton bloom. The proposed models are as follows:

Model 1: Logistic Model with Spatial Parameter

$$y_{ijt} = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)} + y_N^{ijt}\theta + \varepsilon_{ijt} \quad (1)$$

Model 2: Auto-Logistic Model

$$y_{ijt} = \frac{\exp(\mathbf{X}\beta + y_N^{ijt}\theta)}{1 + \exp(\mathbf{X}\beta + y_N^{ijt}\theta)} + \varepsilon_{ijt} \quad (2)$$

For model 1, y depends on the environmental variables X through a logistic function but is a linear function of its neighborhood values. The second model is an autologistic model (Besag, 1972) wherein y is a logistic function in X and the neighborhood values.

To describe the state of red tide phenomenon, a binary response is more appropriate. It takes into account the abrupt changes in the data on cell counts. The proposed models are different from the ordinary logistic model in that spatial structure is incorporated through the neighborhood values. Recall the ordinary logistic model given by the equation :

$$y_{ij} = \frac{\exp(X\beta)}{1 + \exp(X\beta)} + \varepsilon_{ij}.$$

It is speculated that incorporating the spread of the blooming episode or including a spatial structure in the model would increase its predictive ability. Since data are taken over a specified period of time, autocorrelation structure should also be considered. The proposed models may also incorporate AR(1) error terms. Model 1 was formulated to be a simpler alternative model over Model 2 for easier interpretation of the spatial dimension of the phenomenon.

3. THE PROPOSED ESTIMATION PROCEDURE

An iterative fitting procedure is proposed in the estimation of the parameters on the two models. The procedure is similar to that of Speckman's approach as cited by Heckman (1988). Speckman's approach to estimating the parameter β for a semiparametric model defined by $y_i = X_i\beta + g(t_i) + \varepsilon_i$ uses a combination of smoothing and regression. Speckman used a fixed width normal kernel smoother.

This procedure is also similar to the backfitting algorithm for additive models of Hastie and Tibshirani (1990). In Hastie and Tibshirani (1990), the error terms are assumed to be uncorrelated with mean 0 and common variance σ^2 . In the proposed models, we also investigate models with AR(1) error terms, i.e., $\varepsilon_i = \phi\varepsilon_{i-1} + a_i$ where a_i is the usual white noise term and given the location (i,j) . This condition will pose no problem since the paper will deal only with the computational aspects of the algorithm and will not make inferences regarding statistical properties of the estimates.

Suppose the binary response variable for all sites ij at time t , environmental covariates X and the sum of the first-order neighborhood system y_{ij}^{ij} in (Equation 1) are given. The general idea of the estimation procedure for Model 1 is to alternately estimate the parameters corresponding to the environmental covariates (the β 's) and the parameters corresponding to the spatial dimension (θ) and the time dimension (ϕ), as defined in the model with AR(1) error above. The parameters β are estimated using the logistic regression model $y_{ij} = \frac{\exp(X\beta)}{1 + \exp(X\beta)} + \varepsilon_{ij}^*$ where the component ε_t^* will consist of both the autocorrelated error terms and the component attributable to the neighborhood values. Define the resulting residuals as $e_{ij} = y_{ij} - \hat{y}_{ij}$, where $\hat{y}_{ij} = \frac{\exp(X\hat{\beta})}{1 + \exp(X\hat{\beta})}$. Note that these residuals will comprise both the error components ε_t and the component attributable to the neighborhood values.

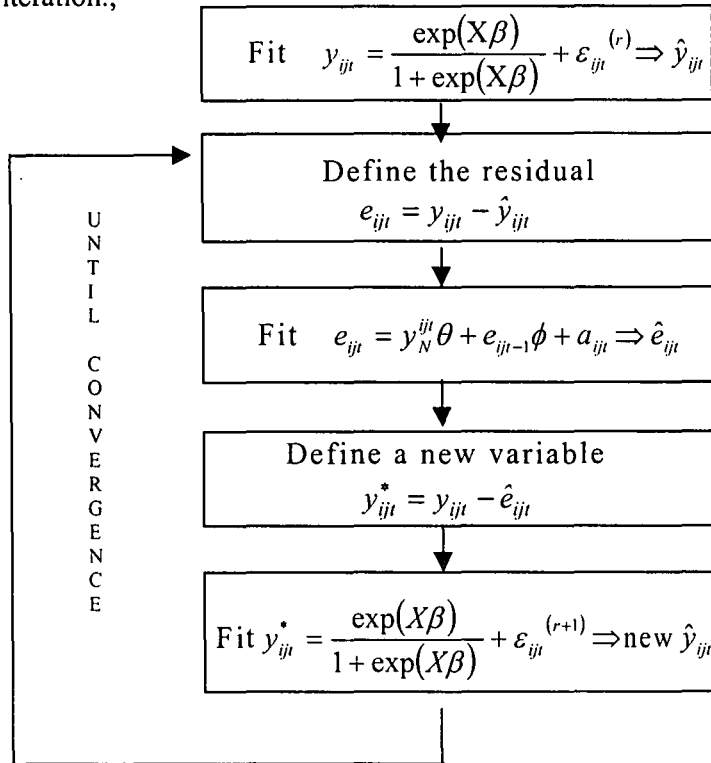
Another possible contribution to the observed residuals may come from some other systematic inadequacies of the predictor variables X and a misspecification of the functional link connecting the response variable to the independent variables. However, suppose that the e_t 's are composed of neighborhood values and ε_t only. The spatial and autocorrelation parameters are then estimated by regressing the partial residuals e_{ij} on the neighborhood values and on AR(1) error term, that is a model given by $e_{ij} = y_{ij}^N \theta + \varepsilon_{ij}$ or $e_{ij} = y_{ij}^N \theta + \phi \varepsilon_{ij-1} + a_{ij}$ is fitted. Estimates of the parameters θ and ϕ are obtained and these in turn yield estimates of ε_t^* and is denoted by $\hat{e}_{ij} = y_{ij}^N \hat{\theta} + \hat{\phi} e_{ij-1}$.

The variable $y_{ij}^* = y_{ij} - \hat{e}_{ij}$ is then defined. This new variable y_t^* , which is no longer dichotomous or binary, is then regressed on the components of X using non-linear regression routine until convergence. Convergence is achieved when $\frac{\beta^{c+1} - \beta^c}{\beta^c} \leq 0.001$,

$\frac{\theta^{c+1} - \theta^c}{\theta^c} \leq 0.001$ and $\frac{\phi^{c+1} - \phi^c}{\phi^c} \leq 0.001$ where c is the iteration number. Figure 3 summarizes the steps of the algorithm.

Figure 3. Estimation Algorithm of Model 1: $y_{ijt} = \frac{\exp(X\beta)}{1 + \exp(X\beta)} + y_{ijN}^{\theta} + \varepsilon_{ijt}$

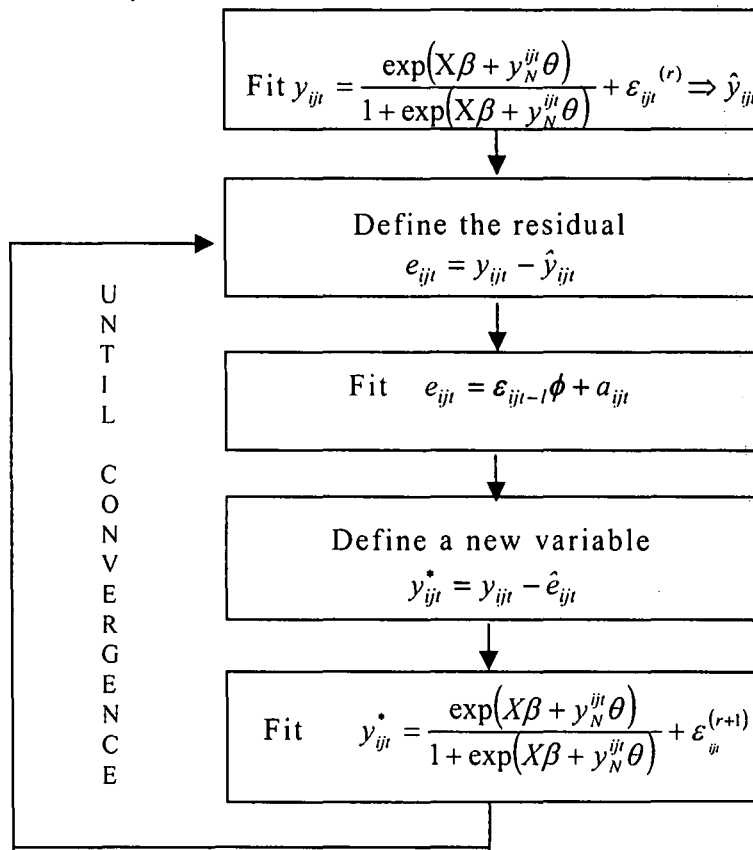
For the r^{th} iteration.,



The estimation procedure for Model 2 follows the same idea as Model 1 except that the effects of both the environmental and neighborhood variables are removed simultaneously. For Model 2, the initial environmental parameters β , and spatial parameter θ are estimated using a method of maximum pseudolikelihood. Maximization of the pseudolikelihood function as were the true likelihood function involves fitting the autologistic model. As cited by Gumpertz, et. al. (1997), the standard errors that are printed out by ordinary logistic regression software are not appropriate for correlated data, and the usual likelihood ratio-type statistics do not have asymptotic chi-square distributions. Again, this would not be a cause for concern, since the focus of the paper is on the computational aspects of the algorithm and not on the statistical properties of the estimates. Figure 4 summarizes the estimation algorithm for Model 2.

Figure 4. Estimation Algorithm of Model 2: $y_{ijt} = \frac{\exp(X\beta + y_{Nt}^{ijt}\theta)}{1 + \exp(X\beta + y_{Nt}^{ijt}\theta)} + \varepsilon_{ijt}$

In the r^{th} iteration,



4. DATA SIMULATION

Most time series on red tide phenomenon are short. This is because collection of indicators reflecting its state and measurement of different hydrological variables are expensive and time consuming. There is an actual data on the red tide phenomenon but due to substantial number of missing data points, simulation was resorted in verifying the models and the estimation procedure. The 4-year data collected by a red tide monitoring and research project in Carigara Bay, Leyte, Philippines (Abuso, et. al., 1997) was used as a priori information for this study. Different spatial and autocorrelation structures were incorporated during the generation of data.

In this section, we shall discuss the procedure on how the data was simulated with the Carigara Bay red tide data set serving as the *a priori* information.

The procedure consists of the following steps:

- Step 1. Compute the mean and standard deviation of the Carigara Bay red tide hydrological variables (X 's) and check for seasonality.
- Step 2. Fit a logistic regression model on the given data and use the parameter estimates (β 's) as initial inputs for the data simulation.
- Step 3. Generate the X 's and the error terms (e 's). Introduce seasonality to those X 's identified in step 1.
- Step 4. Given values for β 's, X 's and e 's generate the response variable y_{ijt}^o .

Step 5. Recode y_{ijt}^o to zero or one depending on the choice of bloom scenario.

Step 6. Assign the first-order neighborhood values y_{N_i} at each site.

Step 7. Given values for β 's, X 's, ε 's and y_{N_i} , generate another set of response variable denoted by y_{ijt} .

Step 8. Recode y_{ijt} to zero or one depending again on the choice of bloom scenario.

The idea in steps 1 to 8 is to generate a binary variable that is related to the X 's through Model 1 or Model 2. Certain spatial and autocorrelation structure for y_{ijt} and the associated error terms are also considered in the simulation.

5. RESULTS AND DISCUSSIONS

Three blooming scenarios were considered for the two models (10%, 20% and 50% of the data points are in bloom status). Three sets of values for the environmental parameters (A, B and C from Table 1) were used as inputs in the simulation. Thirty-six (36) cases were investigated to represent different values of the spatial parameter ($\theta = 0.8, 0.5$ and 0.2) and autocorrelation parameter ($\phi = 0.6, 0.4, 0.1$ including uncorrelated error term) for each of the parameter values in Table 1.

Table 1. Environmental Parameter Values Used

Parameters	A*	B*	C*
β_0	-30.8835	-30.8835	-30.8835
β_1	0.1960	0.0571	0.3349
β_2	0.3072	0.1108	0.5036
β_3	1.3673	1.0427	1.6919
β_4	0.9702	0.5328	1.4076
β_5	0.7708	0.08	1.4616

*A – based on logistic regression from actual red tide data

*B – standard error subtracted from the estimates in A

*C – standard error added to the estimates in A

Model sensitivity and specificity were examined to assess the models' predictive efficiency. Given a "bloom" event, sensitivity is the conditional probability that the model will correctly classify the event. On the other hand, specificity is the conditional probability that the model will correctly classify a "no bloom" event. Sensitivity and specificity should preferably be close to 1 or 100% but inherent relationship between the two types of errors would push the decision-maker to compromise. Depending on the object of the study, one may favor sensitivity at the expense of specificity or vice versa.

Table 2 summarizes the specificity and sensitivity for the logistic model, Model 1 and Model 2 across all different input values of the environmental, spatial, autocorrelation and cut-off values. The cut-off values used for the prediction of the response variable are 0.4, 0.5 and 0.6.

Observe that on the average, the specificity of the ordinary logistic model increases as the bloom episode becomes more frequent while for Models 1 and 2, specificity decreases. On the average, models 1 and 2 exhibits higher specificity in case of a less frequent bloom episode. As the bloom episodes occur more frequently, ordinary logistic model becomes comparable to models 1 and 2. This is to be expected since environmental parameters change

significantly during the blooming episodes. Thus given the X 's, it is not difficult to identify non-bloom events if almost every other point represents a "no bloom" event.

Table 2. Specificity and Sensitivity Rates

Model	Statistics	Specificity			Sensitivity			Correct Classification (%)		
		Bloom Scenario			Bloom Scenario			Bloom Scenario		
		10%	20%	50%	10%	20%	50%	10%	20%	50%
Logistic	Min	30	33	50	56	54	44	34	39	48
	Max	38	41	52	68	63	52	40	44	50
	Mean	34	37	50	62	58	48	37	37	49
	N	27	27	27	27	27	27	27	27	27
Model 1	Min	57	55	41	28	32	42	56	53	48
	Max	72	69	64	49	51	56	66	63	55
	Mean	65	62	54	39	41	49	62	58	51
	N	108	108	108	108	108	108	108	108	108
Model 2	Min	60	57	31	25	28	44	58	54	48
	Max	70	69	57	45	47	72	66	62	55
	Mean	65	62	50	34	39	51	62	57	51
	N	108	106	108	108	106	108	108	108	108

Min: Minimum Proportion

Max: Maximum Proportion

N: Number of Simulated Data Set

On the other hand, notice that on the average, the sensitivity of Models 1 and 2 increases as the frequency of bloom episode becomes abundant. However, the ordinary logistic model experiences a decreasing sensitivity property as bloom episode become more frequent. The result implies that Models 1 and 2 exhibit higher sensitivity as bloom episodes becomes more frequent. Therefore, as far as sensitivity property is concerned, the two proposed models are superior over the ordinary logistic model.

The autologistic model (Model 2) and the logistic with spatial parameter (Model 1) exhibit superiority in terms of its ability to classify binary events. On the average (across all bloom scenarios), the ordinary logistic model can correctly classify events up to 45%. The autologistic model can correctly classify events by about 61%. Similar observation is noticed for Model 1. Thus inclusion of spatial parameter to the logistic model results improvement in its ability to classify binary events.

6. CONCLUSIONS

An estimation procedure is applied to two models, which are the logistic model with spatial parameter (Model 1) and the autologistic model (Model 2). Different scenarios were generated and different set of parameter input values were used in the simulation of data to investigate the two models.

The model performance or the predictive efficiency of the models is examined. Predictive efficiency was measured through the model's ability to correctly classify a "no bloom" event (specificity) and a "bloom" event (sensitivity).

The simulated data gave evidence that the autologistic model exhibits higher specificity in case of a less frequent bloom episode and higher sensitivity as bloom episode becomes more frequent. Model 1 showed similar specificity and sensitivity properties as Model 2. The sensitivity property of the ordinary logistic model decreases as bloom episodes become more frequent. This result makes the two proposed model superior compared to an ordinary

logistic model. Thus inclusion of spatial parameter to the logistic model results improvement in its ability to classify binary events.

ACKNOWLEDGMENT

The authors wish to acknowledge the anonymous referee for this paper who has given comments and suggestions for its improvement.

References

- ABUSO, Z.V., CABELLO, L.T., TUAZON, L.C. AND BARRIOS, E. (1997). "RedTide Monitoring in Carigara Bay Central Philippines". pp. III(38) – III(66), in *ASEAN Marine Environmental Management: Quality Criteria and Monitoring for Aquatic Life and Human Health Protection*, ed. by Vigers, G., Ong, K.S., McPherson, C., Millson, N., Watson, I., and Tang, A. EVS Environment Consultants, North Vancouver and Department of Fisheries Malaysia.
- BESAG, J. (1972). "Nearest-Neighbor Systems and Auto-Logistic Model for Binary Data", *Journal of the Royal Statistical Society. Series B.* 34, 75-83.
- BONNEY, G.E. (1987). "Logistic Regression for Dependent Binary Observations". *Biometrics.* 43, 951-973.
- CRESSIE, N.A.C. (1993). *Statistic for Spatial Data*. New York: Wiley.
- CULLIS, B.R., GLEESON, A.C. (1991). "Spatial Analysis of Field Experiments: An Extension to Two Dimensions", *Biometrics.* 47, 1449-1460.
- GROBBELAAR (1990) "Modeling Phytoplankton Productivity in Turbid Waters with Small Euphotic to Mixing Depth Ratios", *Journal of Plankton Research.* 12(5), 923-931.
- GUMPERTZ, M.L., GRAHAM, J.M., AND RISTAINO, J.B. (1997) "Autologistic Model of Spatial Pattern of Phytophthora Epidemic in Bell Pepper: Effects of Soil Variables on Disease Presence", *Journal of Agricultural, Biological, and Environmental Statistics.* 2(2), 131-156.
- HASTIE, T.J. AND TIBSHIRANI, R.J. (1990). *Generalized Additive Models* Great Britain: St Edmundsbury Press Ltd., Bury St Edmunds, Suffolk.
- HECKMAN, N. E. (1988) "Minimax Estimates in a Semiparametric Model", *Journal of the American Statistical Association*, 83(404), 1090-1096.
- HUFFER, F.W. AND WU, H. (1998) "Markov Chain Monte Carlo for Autologistic Regression Models with Applications to the Distribution of Plant Species", *Biometrics.* 54, 509-524.
- LIANG, K.Y. AND ZEGER, S.L. (1986) "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika.* 73, 13-22.
- WU, H. AND HUFFER, F.W., (1997) "Modeling the Distribution of Plant Species Using the Autologistic Regression Model", *Environmental and Ecological Statistics.* 4, 49-64.