Chapter 5

Generalized Linear Models Applied to Satellite Based Change Detection

Introduction

This chapter represents the theoretical heart of the dissertation. Again, the thesis is that ***Generalized Linear Models (GLMs) can be used to enhance satellite based land cover change detection***. In this chapter we start with a discussion of what we are attempting to model when reflectance values from satellite data are used to monitor changes in land cover. We then present the general framework for Generalized Linear Models (GLMs) and the assumptions associated with these models. This is followed by a description of how these models can be used to enhance satellite-based change detection as well as some caveats of using the models.

Modeling Land Cover Change with Satellite Data

In using satellite data to monitor changes in land cover, the modeling assumption is that the reflectance values from different images can be used to indicate where changes in land cover have occurred. Radiance values are used to model change. The different algorithms described in the preceding chapter can be considered different models. In particular, those models which require the setting of a threshold – and to which GLMs can be applied – can be written as mathematical models. In these models, change is the output variable on one side of the equation and the reflectance values from two different times are input variables on the other side of the equation.

The general form of change detection can therefore be written as the equation:

$$\boldsymbol{Change} = \boldsymbol{f}(\boldsymbol{x}) \qquad \textbf{(eq. 5.1)}$$

where $\boldsymbol{x}$ is a vector of radiance values, the ***change*** variable may be a "0-1" binary response where 0 represents "no change" and 1 represents "change". The ***change*** variable could also be a "0,1,2,…,N" multinomial response where each number represents a particular type of change. The form of $\boldsymbol{f}$ will depend on the change detection algorithm being used.

We will now give particular definitions of the function ***f(.)*** for the "Image Algebra", "Binary Mask", and "Spectral Change Vector" algorithms. Our functional definitions will assume six-band imagery and a binary response. Generalizations of more or fewer bands are obvious.

As an example of image algebra, consider:

$$\boldsymbol{Change} = \begin{cases} \boldsymbol{0\ if}\ \ \boldsymbol{x_{i,Tb} - x_{i,Tb-1} \le T} \\ \boldsymbol{1\ if}\ \ \boldsymbol{x_{i,Tb} - x_{i,Tb-1} > T} \end{cases} \qquad \textbf{(eq. 5.2)}$$

where $\boldsymbol{x}$ represents the normalized radiance value for a given pixel, $\boldsymbol{i}$ represents a band number, $\boldsymbol{T_b}$ and $\boldsymbol{T_{b-1}}$ represent the two time periods, and $\boldsymbol{T}$ represents the threshold value.

As an example of spectral change vector analysis, consider:

$$\boldsymbol{Change} = \begin{cases} \boldsymbol{0\ if}\ \ \boldsymbol{\sqrt{\sum_{i=1}^{6} (x_{i,Tb} - x_{i,Tb-1})^2} \le T} \\ \boldsymbol{1\ if}\ \ \boldsymbol{\sqrt{\sum_{i=1}^{6} (x_{i,Tb} - x_{i,Tb-1})^2} > T} \end{cases} \qquad \textbf{(eq. 5.3)}$$

where the variables follow the same definitions as in the previous equation.

An example of "Multi-date Change Detection Using a Binary Mask" could use either of

these two functions to set up a binary mask (see Chapter 4) and then classify those pixels with functional values beyond the threshold ***T***.

Existing change detection methods require the analyst to choose one of the above methods; and the choice is either 1) somewhat arbitrarily made after some initial investigation or 2) "after-the-fact" based on which method produces the highest accuracy. Once a method is selected, the analyst is faced with the additional task of choosing the best function of reflectance values to use within the given method. For example, if image differencing is to be used, is it the difference in band 4, difference in band 5, or a combination of these two?

If these change detection algorithms can be written as mathematical models, then statistical techniques can be used to fit these models. In particular, reflectance values can be treated as independent variables and judged for their significance in predicting changes in land cover -- the dependent variable. As with standard linear regression, a sample can be selected from the study area. Contained in this sample are the reflectance values from the satellite imagery and the "ground truth" or "reference data" of change/no-change. Statistical analysis can then determine which function of the reflectance values provides the best estimate of change. However, *unlike* standard linear regression, the dependent variable "change" is not typically modeled as a continuous variable. It is either a binary variable (Cox and Snell, 1989) of change/no-change or it is a multinomial response (Cassela and Berger, 1990) representing several discrete and unique types of change.

Generalized Linear Models are a set of statistical models that can be used to model a binary or multinomial response (Agresti, 1990). Several existing studies in remote sensing have incorporated GLMs into the analysis of satellite data. Within this existing literature, studies generally fit into two categories: 1) species or habitat modeling and 2) error or accuracy assessment analysis. With respect to habitat modeling, Narumalani *et al*. (1997) used logistic regression to model aquatic macrophyte communities with air-photo-derived topographic maps and other GIS layers. Pereira and Itami (1991) used logistic regression to model Red Squirrel habitat. With respect to accuracy assessment, log linear models

were introduced by Congalton et al (1983) as an analytical technique to compare the accuracy of different land-cover classifications. A more recent paper by Arora and Foody (1995) use log linear models to compare factors affecting classification accuracy.

While these examples show how GLMs can enhance satellite data analysis, to date there do not appear to be any existing studies that apply GLMs to change detection analysis. Before we go on to describe the particulars of how GLM can be applied to satellite based change detection we will describe the theory behind GLMs.

Generalized Linear Models

Nelder and Wedderburn (1972) introduced Generalized Linear Models over two decades ago and relatively recent texts (e.g. Agresti, 1990 and McCullagh and Nelder, 1989) and statistical software (SAS™, 1990) have become available. These have helped bring these models to a wide range of scientists. The following discussion will give a very brief introduction to the formulation of these models to facilitate the discussion on our application of these models to change detection. For a detailed description of these models the reader is referred to Agresti (1990) and/or McCullagh and Nelder (1989).

The Generalized Linear Model (GLM) can be written as:

$$g(\mu) = \sum_{j=0}^{p} \beta_j \, x_{ij} \quad , \quad i = 1, \ldots N \; ; \qquad \textbf{(eq. 5.4)}$$

where N is the number of observations from the sample.

The model is described, or defined, by distinguishing three elements of the model: namely 1) the *random component*, 2) the *systematic component*, and 3) the link between the two, known as the *link function*.

Random component of Generalized Linear Models

The random component accounts for the independent observations from a known (or assumed) probability density function of the form:

$$f(y_i;\theta,\phi) = exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i,\phi)\right\} \qquad \textbf{(eq. 5.5)}$$

The parameter θ_i is called the *natural parameter* and ϕ_i is known as the *dispersion parameter*.

The "Y" variable is generally the variable of interest; the variable you are trying to model and/or predict. The set of *N* observations is denoted by the vector $\mathbf{Y} = (Y_1, ..., Y_N)'$.

The systematic component of Generalized Linear Models

The Systematic component is that part containing the "X", controlled, variables and the associated unknown parameters, which need to be fit or predicted, and are used to assess the influence and significance of their related X variables. (The parameters in the systematic component will be referred to as the model parameters so to avoid confusion with the parameters in the probability density function.) The systematic component can be expressed as:

$$\eta_i = \sum_{p=0}^{p} \beta_j \, x_{ij} \quad , \quad i = 1,\ldots N \qquad \textbf{(eq. 5.6)}$$

or, in matrix notation: $\eta = \mathbf{X}\beta$; where **X** represents the *N* x *(P+1)* design matrix in which the rows correspond to the N observations and the P+1 columns match the different values/levels of the control variables. The particular values in **X** represent different values of numeric control variables and/or different levels of nominal control variables represented by dummy variables. The vector β is a *(P+1)* x 1 vector containing the model parameters to be fit. Note that η_i is a linear function of the x_{ij} and β_j terms so that it is a

linear predictor.

The link function component of Generalized Linear Models

The final element in the GLM is the link function. The link describes a functional relationship between the linear predictor, η_i, and the expected value of a datum y_i. This is expressed as:

$$\eta_i = g(E[Y_i]) \qquad \textbf{(eq. 5.7)}$$

or

$$\eta_i = g(\mu_i); \textit{ where } \mu_i = E[Y_i] \ . \qquad \textbf{(eq. 5.8)}$$

Assumptions associated with Generalized Linear Models

In discussing the assumptions of GLMs, it is important to note that each form of a GLM will have its own assumption. For example, if the GLM is specified as a standard linear regression then the Y's are assumed to be independent and normally distributed with mean μ and constant variance and the link function is set as $\eta = \mu$. However, GLMs in general do not assume the observed data are normally distributed nor do they assume constant variance. That is, GLM have less restrictive assumptions than standard linear regression equations. However, GLMs do contain certain assumptions.

Similar to standard regression, with GLMs there is still a need to assume that the observed Y's are independent. The significance of this assumption is that it allows the joint probability density function (pdf) to be expressed as a simple product of the single pdfs. Since the joint pdf is the foundation of the maximum likelihood (ML) equations, and since the ML equations are used to solve for the unknown parameters, solutions based on ML estimates are effected by this assumption. Non-independence could lead to a different joint pdf and different ML equations. The typical result is that the error term associated

with the model is less than the true variance. That is, the variance estimate derived from the model will underestimate the true variance (McCullagh and Nelder, 1989, Section 4.5; SAS, 1995, p. 81). This will affect the inference derived from the model. Namely, the p-value associated with the model parameters will be lower than they would be if the actual variance were used.

For any GLM the pdf of the observed data is assumed to take on the form specified in equation 5.5. This assumption is significant mainly in that it is much more general than having to assume a normal distribution. That is, this "assumption" provides much more flexibility with GLMs than there is with standard regression. Most of the common distributions are exponential class: Normal, Poisson, Binomial, Gamma, or Inverse Gaussian (McCullagh and Nelder, 1989, table 2.1). The general nature of the data makes it fairly reasonable to assume one of these distributions.

GLMs assume that the link function is a monotonic and differentiable function. There are many different functional forms that meet these criteria. The assumption, while it must be observed, is not of practical concern since all of the standard link functions (available in statistical software packages) meet these criteria.

Generalized Linear Models Applied to Satellite Based Change Detection

This section will be arranged to match the previous section. We describe each of the three components of a GLM in terms of how they relate to change detection. We then describe how the assumptions affect the use of GLMs for land cover change detection.

Random component applied to change detection

The random component will be the change response variable. For many studies this is simply the binary "change/no-change" determination. The type of response can be modeled with a binomial distribution. The binomial distribution fits into the form needed

for GLMs. Another possibility for the random component in change modeling is that there are several discrete types of change. For example, one may be interested in particular "from-to" changes. This could be modeled with a multinomial distribution. We will pursue modeling the more simple "change/no-change" but note that GLMs may be applied to model more complex change determinations.

The random component can be thought of as the "Y" variable or the variable you are interested in modeling. GLMs require a sample or set of points where the random component has been observed. For change detection, we are interested in using satellite data to determine if change has occurred. So, we will need to have a sample of points where we know if change has occurred or not. In remote sensing studies, such "ground truth" or reference data is typically acquired from some source other than satellite data, generally either higher resolution photos or *in situ* measurements. In applying GLMs to change detection, we need to acquire a set of "ground truth" or reference points to make up the sample that will comprise the data used as the random component of the model. This will require determining land cover changes for a set of ground sample points. In our example we will use higher resolution air photos from the 1988 and 1994 time period. The sample and data collect for our example is described in Chapter 6.

The systematic component applied to change detection

In standard linear regression terms, the systematic component is the explanatory or independent variables. In change detection this can be any function of the reflectance values from the two images as well as ancillary, covariate, data. Each right-hand-side of the example equations given in the first section of this chapter provides an example of a systematic component that can be used in a GLM. The statistical analysis of the models can be used to determine which function of the reflectance values are the most significant and do the best at modeling the change/no-change response. In our change detection study we will have a set of paired data for each point in our sample. There is the reference data interpreted from the reference data. This represents the random component and relates to whether the sample point has change or not. Coupled with each of these points

are the radiance values from the image data at that point. Different functions of the radiance values represent different systematic components that can be used to model change.

The link function component of Generalized Linear Models

The link function mainly depends on the form of the expected value for the random component. For a binomial response, some possible link functions include the "log odds" link used in logistic regression (Agresti, 1990, section 4.4.2), inverse CDF link such as the "probit" model (Agresti, 1990, section 4.5), and the log-log and complementary log-log link (Agresti, 1990, section 4.5.2). Other link functions are available for modeling multinomial random variable (Agresti, 1990, Chapter 9). Prudence also dictates that the link function is one that can be modeled with existing statistical software. The LOGISTIC procedure in SAS ™ statistical software gives maximum likelihood fitting of logit, probit and extreme-value models (Agresti, 1990, p. 485). These models are defined by the following equations. Note that the $\beta' \boldsymbol{x}$ term is not limited to one variable but β can be *vector* of unknow parameters while $\boldsymbol{x}$ can be *vector* of different explanatory variables.

The logistic link functions uses the link:

$$\boldsymbol{Prob.\ of\ \ change} = \frac{\boldsymbol{exp}(\alpha + \beta' \boldsymbol{x})}{\boldsymbol{1} + \boldsymbol{exp}(\alpha + \beta' \boldsymbol{x})} \qquad \textbf{(eq. 5.9)}$$

The "probit" model uses the Cumulative Distribution Function from the standard normal distribution, $\boldsymbol{F(.)}$ and is expressed as:

$$\boldsymbol{Prob.\ of\ \ change} = \boldsymbol{F}(\alpha + \beta' \boldsymbol{x}) \qquad \textbf{(eq. 5.10)}$$

The complementary log-log is defined as:

$$\boldsymbol{Prob.\ of\ \ change} = \boldsymbol{1} - \boldsymbol{exp}\left[-\boldsymbol{exp}(\alpha + \beta' \boldsymbol{x})\right] \qquad \textbf{(eq. 5.11)}$$

Figure 5.1 shows an example of these models. The nature of these models will limit the predicted probability of change to values between zero and one. In figure 5.1, as the X variable increases so does the probability of change. This corresponds to a positive value for the parameter **b**. The relationship is reversed if **b** is negative. That is, for negative **b** the probability of change *decreases* as the X variable increases. Generally the logit and probit models are similar. The Complementary log-log model departs from zero more slowly and approaches one more quickly. Which model is best will depend on the data. In the following chapter we will fit our data testing each of these three link functions.

There is a critical difference from the link functions presented in the previous three equations and the equations given in the beginning of this chapter. By using GLM, although the random component is a binary variable, the models result in a predicted value that is continuous on the interval [0, 1]. The output from the model is a "probability of change" (POC) as opposed to a binary estimate. This can be though of as a fuzzy classification or probability mapping of change/no-change areas (Zadeh, 1965; Foody and Trodd, 1990). In Chapter 8 we discuss the use of POC images.

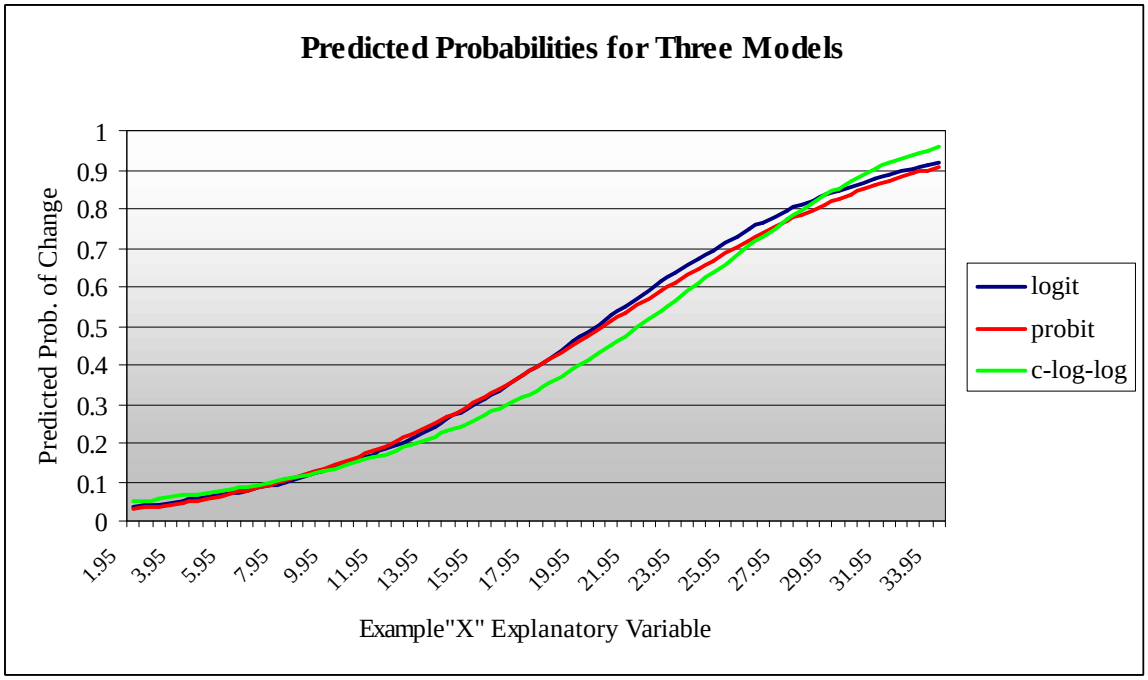

Figure 5.1: Predicted probability curves for the three models to be used in the example

Assumptions associated with Generalized Linear Models

Again, each form of a GLM will have its own assumption. For example, modeling just "change/no-change" will assume a binomial distribution and modeling different types of change will assume a multinomial distribution. For our analysis, since our interest is in the change/no-change response, we need to assume a binomial distribution. The form of the probability density function for either binomial or multinomial response data fit into the form needed for GLM, and so, the assumption on the form of the pdf (eq. 5.5) is rather minor and safe for the application of GLMs to change detection. Each of the link functions listed above is a monotonic and differentiable function and, so, this assumption, too, is met.

The independence assumption is something that needs to be considered when working

with satellite and land cover data. Since the estimates and inference derived from the model depend on the data being independent, and since most land cover phenomena tend to be autocorrelated in space, this assumption should be checked. Note, however, the model assumes the observations of the *random* component are independent. So, the checks for independence need to be on the response variable. The general cross-product statistic (Upton and Fingelton, 1985) can be used to check for spatial independence of a binary or multinomial response variable. This should be done before investigating different models and different functions of the radiance values. If there is autocorrelation present in the response variable, this should be kept in mind when considering the p-values and variance terms used in model selection. However, even with autocorrelation present in the response variable, GLMs can be used to guide model selection. Once a model is chosen, the residuals from that model can be investigated by calculating and plotting variogram values for the residuals. (Variograms will be described in the following chapter.) If the residuals from the model do *not* show correlation then the variance term and inference from the standard model are appropriate. If the variogram of the residuals indicates the presents of autocorrelation then the variance term and inference should be used cautiously. Diggle *et al*. (1995, equation A.6.1) give a robust variance estimate that could be used in this situation.

It should also be noted that there is recent work being conducted within Statistics to handle spatially autocorrelated data. For example, Albert and McShane (1995) have used generalized estimating equations in their application of GLMs with the logit link function to spatially correlated binary data. The theory grows out of the analysis of longitudinal data (Diggle, Lang, and Seger, 1995). Haining (1990, pp. 99-100) describes "autologistic" models for modeling spatially correlated discrete random variables. Also, Version 6.10 and later versions of SAS™ have modeling options within its LOGISTIC procedure to correct for overdispersion caused by positive correlation between binary responses (SAS, 1995, pp. 81 - 82). Here we simply point to this body of work. Our main objective is to introduce GLMs as a form of analysis suitable for change detection studies.

Further research can and should address the issue of using GLMs on spatially correlated change data.

Summary of GLMs Applied to Change Detection

Many change detection algorithms can be viewed as a mathematical model and these mathematical models can be assessed with GLMs. GLMs can be used to analyze binary and multinomial responses such as change/no-change. Typically, air photos or *in situ* data are used a "reference data". In the framework of GLMs, this data comprises the response variable. Different functions of radiance values can be used as different explanatory variable to be explored within the GLM framework. GLMs can be used to regress the binary reference data on the satellite image radiance values. Due to the possibility of spatial correlation in land cover change detection, the assumption of independent response variable needs to be checked. In the following chapters we present an example of how GLMs can be applied to change detection for the Raleigh and coastal areas.